Topic: Spam Email Detection

Prepared by,

GROUP : 08
ABRAR TAHMID GALIB - 19301246
ANNAS MOHD NAZIM - 19301082

Key Terms: Spam, Ham, Machine Learning, Naive Bayes, Logistic Regression, Support Vector Machine.

I. INTRODUCTION

Undesirable and unwanted junk emails are known as spam, and it is sometimes distributed in large quantities to a broad recipient list. Spam is typically transmitted for financial gain. Botnets, or networks of infected computers, are capable of sending it in large quantities.**[1]**Spam emails can be dangerous for a computer system's security in addition to being unpleasant. Businesses were said to have lost almost $100 billion to spam in 2007.**[2]**

In this project, we execute automatic spam filtering using text analysis too. Our goal is to find patterns utilizing data-mining or data-analyzing classification techniques so that we can categorize the emails as either Spam or Ham.

II. METHODOLOGY

A.Dataset Description

We have collected the dataset from ”Kaggle” titled Email Spam Detection Dataset (classification). Top contributors are Shantanu Dhakad, Hosam Wajeeh, Timothy Abwao **[3]** The dataset consists of 5572 data. Among these 4825 data are named as ham and 747 data are marked as spam.Moreover, there are 5 columns. First column, namely “v1”, has the category of emails which are our target features and the second column, namely “v2” has the texts which we will analyze to find target features. In addition, target features are the collection of alphabets, words or other characters.

B.Pre-Processing Techniques applied

The dataset consists of some null and duplicate values. Moreover, it consists of some irrelevant features namely “Unnamed: 2”, “Unnamed: 3” and “Unnamed: 4”. In this stage of data pre-processing we have omitted the null values, removed the duplicates and dropped the columns with all the infelicitous features. In addition, we have extracted the unique values. After the effacing and extracting step, we are left with 5169 data among 5572 data to work on. The histogram, correlation graph and count plot graph of target features are following-

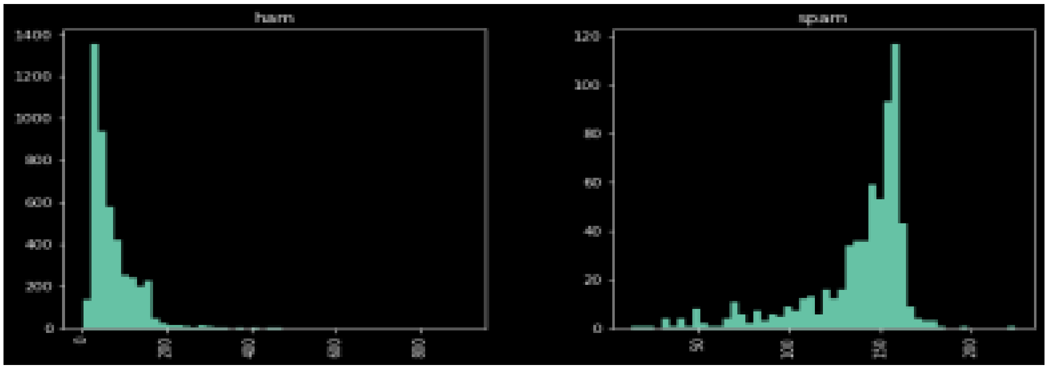
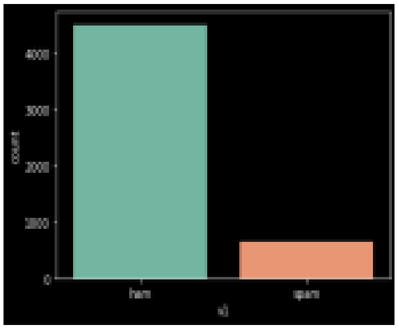


Fig. 1: Histogram for the target features.



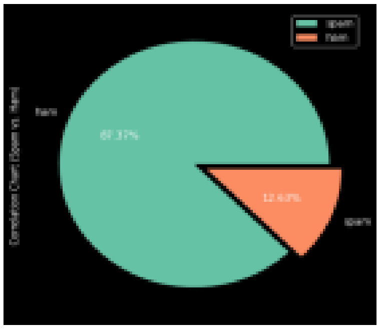


Fig. 2: Count plot graph for the target features. Fig. 3: Pie- chart graph for the target features.

Here, the target value is **“Spam”** is encoded as **“1”** and **“Ham”** is encoded as **“0”**. The data has been splitted into **80%** training and **20%** testing and feature extraction has been done to use outlier sensitive classifiers in terms of training our model also.

C.Models applied

As we are developing a predictive model based on both input and output data, actually we are using “Supervised learning method” which is basically two types and they are- “Classification” and “Regression”. In our model the target feature is classification based so, we will use Classifier Machine-Learning models. Thus, we have decided to use three non-linear models namely-Naive Bayes Classifier and SVM Classifier as well as, we will use logistic regression model. Finally, we will use three models.

1. Naive Bayes:

The Naive Bayes classification algorithm is a probabilistic classifier which is based on probability models that make substantial independent assumptions. The independence presumptions frequently have no bearing on reality.Thus, they are viewed as being naive.With the aid of Bayes' theorem, probability models can be generated. You may be able to train the Naive Bayes algorithm in a supervised learning environment, depending on the characteristics of the probability model. **[4]**

2. Logistic Regression :

Predictive analytics and categorization frequently make use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula.**[5]**

3. Support Vector Machine(SVM):

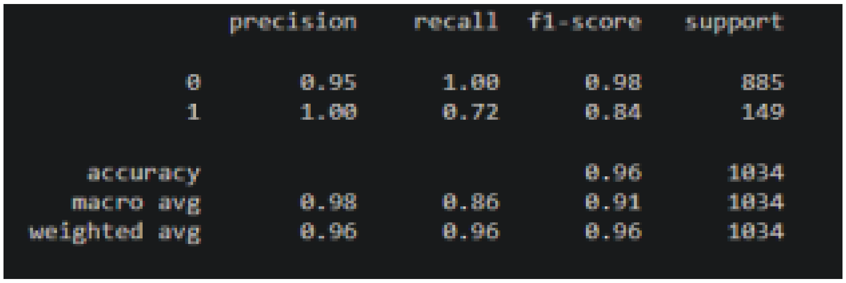
A reliable classification and regression method called Support Vector Machine (SVM) increases a model's predicted accuracy without overfitting the training set. SVM is particularly well suited for data analysis with a very large number of predictor fields, such as thousands. Face and other image identification, bioinformatics, idea extraction from text mining, intrusion detection, protein structure prediction, voice and speech recognition, and many other fields benefit from the use of SVM.[7]

III. RESULT ANALYSIS

A. **Naive Bayes:**

From the test the **Accuracy score** is **95.94%** and **Area Under Curve** is **85.91%.**

Other details-

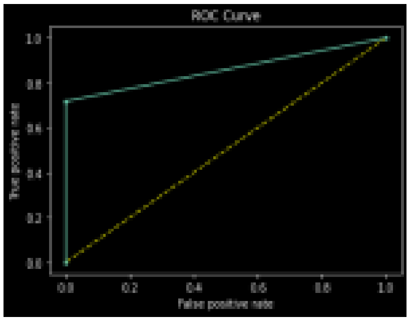


```
              precision    recall  f1-score   support

           0       0.95      1.00      0.98       885
           1       1.00      0.72      0.84       149

    accuracy                           0.96      1034
   macro avg       0.98      0.86      0.91      1034
weighted avg       0.96      0.96      0.96      1034
```

Confusion matrix Graph- ROC Curve



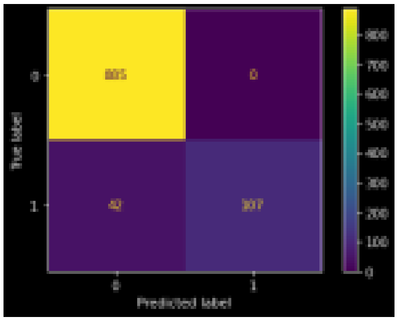
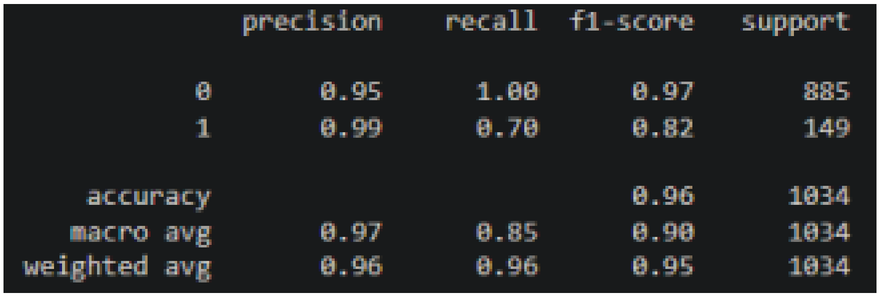


Fig. 4. Confusion Matrix of the prediction. Fig. 5. ROC Curve.

B. **Logistic Regression :**

From the test the **Accuracy score** is **95.55%** and **Area Under Curve** is **84.84%.**

Other details

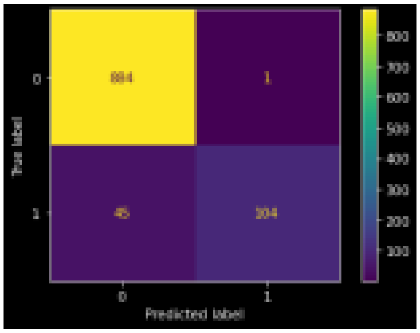


```
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       885
           1       0.99      0.70      0.82       149

    accuracy                           0.96      1034
   macro avg       0.97      0.85      0.90      1034
weighted avg       0.96      0.96      0.95      1034
```

Confusion matrix Graph- ROC Curve



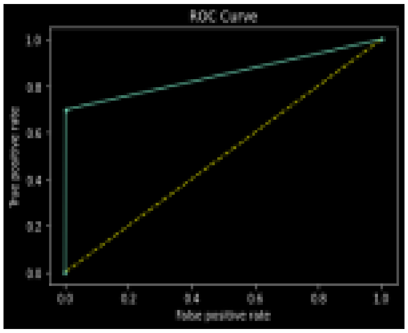
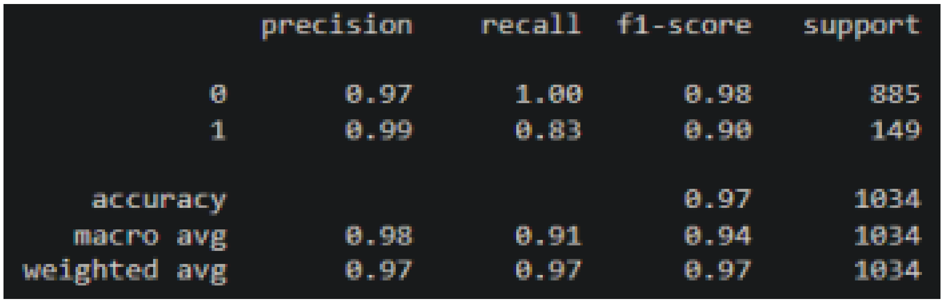


Fig. 6. Confusion Matrix of the prediction. Fig. 7. ROC Curve.

C. **Support Vector Machine(SVM):**

From the test the **Accuracy score** is **97.38%** and **Area Under Curve** is **91.22%.**

Other details-



```
              precision    recall  f1-score   support

           0       0.97      1.00      0.98       885
           1       0.99      0.83      0.90       149

    accuracy                           0.97      1034
   macro avg       0.98      0.91      0.94      1034
weighted avg       0.97      0.97      0.97      1034
```

Confusion matrix Graph- ROC Curve

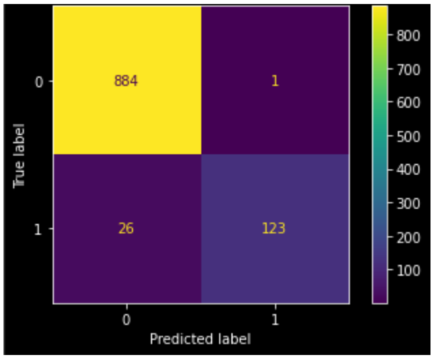


Fig. 10. Confusion Matrix of the prediction

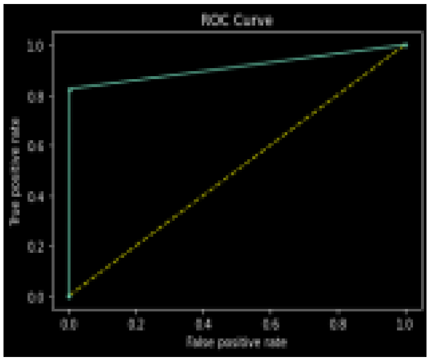


Fig. 11. ROC Curve.

IV. Conclusion

Among the four models which are used on the dataset, **Support Vector Machine(SVM)** has the highest value of **accuracy score (97.38%)** , **f1 score(90%)** , highest **value of AUC (91.22%)** and best **ROC curve** . Therefore, it is concluded that for this dataset, **Support Vector Machine(SVM)** produced the best outcome.

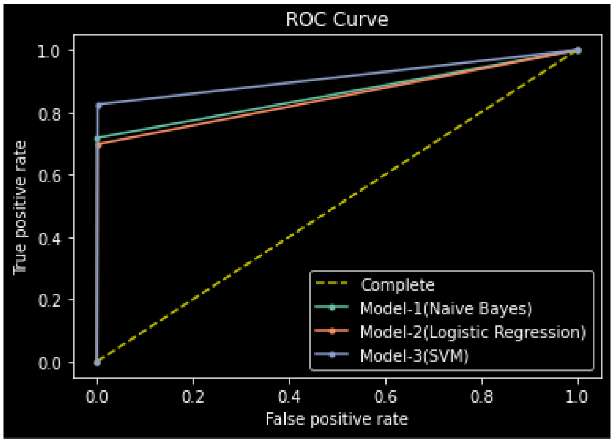


Fig. 12. Comparison of ROC Curves of the four models.

V. REFERENCES

1. Rao, J. M., & Reiley, D. H. (n.d.). The Economics of Spam. *SpamEconomics*. http://www.davidreiley.com/papers/SpamEconomics.pdf

2. CISCO. (n.d.). *What Is Spam Email?* What Is Spam Email? https://www.cisco.com/c/en/us/products/security/email-security/what-is-spam.html 3. https://www.kaggle.com/datasets/shantanudhakadd/email-spam-detection-dataset-classification?resource=download

4. IBM. (n.d.). *Naive Bayes*. IBM Integrated Analytics System. https://www.ibm.com/docs/en/ias?topic=procedures-naive-bayes

5. IBM. (n.d.). *What is logistic regression?* https://www.ibm.com/topics/logistic-regression#:~:text=Logistic%20regression%20estimates%20the%20probability,bounded%20between%200%20and%201.

7. IBM. (n.d.). *About SVM*. SPSS Modeler. https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-about-svm