**Spam Email Detection Report**

**Group No:** 5

**Group Members:**

| **Student ID** | **Student Name** |
| --- | --- |
| 18301268 | Syed Mahathir Md. Lenin |
| 19101147 | Ahmad Zubair |
| 19301074 | Sharmin Akter Keya |
| 19301089 | Prionto Kumar Choudhury |

**Introduction**:

Emailing is one of the most popular ways of communicating online. Thus, there are individuals/organizations who use emails as a way of harming other people. The emails through which people are scammed or the emails which are intended to cause harm to people are usually called spam emails. Spam emails trick potential victims into downloading malwares, sharing data or sending money. On the contrary, the emails which are legitimate are called ham emails. To keep ourselves safe from spam emails we need to be able to detect those emails and so we take the help of Machine Learning to detect the spam emails.

**Methodology:**

The dataset contains two columns. One of which is “Category” and the other is “Message”. Category contains two values such as Spam and Ham (Not Spam). Again, the “Message” column contains all the messages which are categorized into either Spam or Ham.

There were a few techniques that were used for data processing. Firstly, all the null rows were dropped. Secondly, the duplicate rows were dropped from the dataframe. Then, we did Label Encoding on the “Category” column by turning the “Spam” values into 0 and the “Ham” values into 1.

Again, the dataset has been split using train\_test\_split function and the feature extraction has been done using Countvectorizer.

The three models which have been applied are Naive Bayes, Logistic Regression and Support Vector Machine.

**Result:**

The accuracy we got from Naive Bayes is 97.87 % and from the confusion matrix we can see there are 120 true spam emails and 19 false spam emails whereas there are 3 false ham emails and true 890 ham emails.

.

Again, the accuracy we got from Logistic Regression is 97.67 % and from the confusion matrix we see there are 116 true spam emails and 23 false spam emails. On the other hand, there’s only 1 false ham email and 892 true ham emails.

Lastly, the accuracy we got from SVM is 97.48 % and from the confusion matrix we can see there are 113 true spam emails and 26 false spam emails. Again, there are no false ham mails and 893 true ham emails.

**Conclusion:**

All the models have almost the same accuracy. However, Logistic Regression has the accuracy of 97.67 % and from the confusion matrix of this model we can see there’s only 1 false ham email which is really less and the number of false spam emails is 23 only. So, we can say Logistic Regression is doing better comparatively on the dataset.

**Reference:**

Dataset- <https://drive.google.com/file/d/1uzbhec5TW_OjFr4UUZkoMm0rpyvYdhZw/view>