BDT Mini Project Report
on
Email Spam Detection

Mini Project Members

Viren Kadam 1032201570 PG-28

Mihir Agarwal 1032200967 PG-11

Hussain Agarwal 1032200990 PG-14

Keval Pambhar 1032201039 PG-15

School of Computer Engineering and Technology
MIT World Peace University, Kothrud,
Pune 411 038, Maharashtra - India
2022-2023

Introduction:

Email has become extremely popular among people nowadays. In fact, it has been reported to be the cheapest, popular and fastest means of communication in recent times. Despite the huge benefits of emails, unfortunately its usage has been bedeviled with the huge presence of unsolicited and sometimes fraudulent emails which must be promptly detected and isolated through what is popularly referred to as spam detection system. Spam detection is highly needed to protect email users and prevents several negative usages to which emails have been subjected to recently. Unfortunately, due to the adaptive nature of unsolicited emails through the use of mailing tools, the effectiveness of the spam detecting tools has often been limited and sometimes rendered ineffective, hence the need for better spam detection tools to achieve better spam detection accuracy. Several spam detection models have been proposed and tested in the literature, but still the reported accuracy indicated that there is still need for more work in this direction in order to achieve better accuracy.

Solutions Available:

Content-Based Filtering Technique: Algorithms analyze words, the occurrence of words, and the distribution of words and phrases inside the content of e-mails and segregate them into spam and non-spam categories.

Case Base Spam Filtering Method: Algorithms trained on well-annotated spam/non-spam marked emails try to classify the incoming mails into two categories.

Heuristic or Rule-Based Spam Filtering Technique: Algorithms use pre-defined rules in the form of a regular expression to give a score to the messages present in the e-mails. Based on the scores generated, they segregate emails into spam and non-spam categories.

Adaptive Spam Filtering Technique: Algorithms classify the incoming mails into various groups and, based on the comparison scores of every group with the defined set of groups, spam, and non-spam emails get segregated.

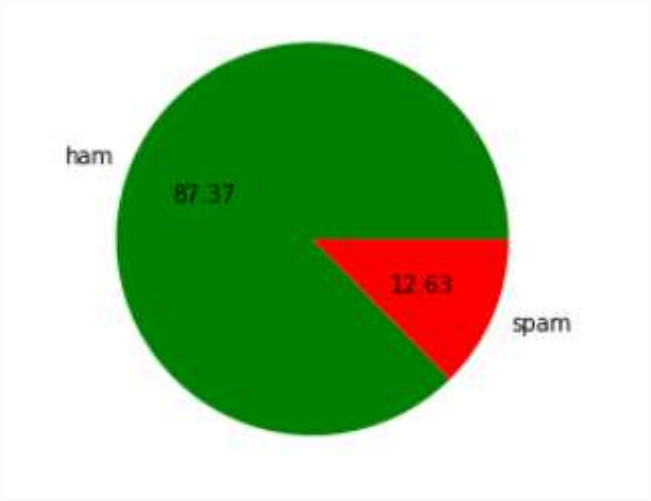
Solution Chosen:

Case Base Spam Filtering Method: Algorithms trained on well-annotated spam/non-spam marked emails try to classify the incoming mails into two categories.

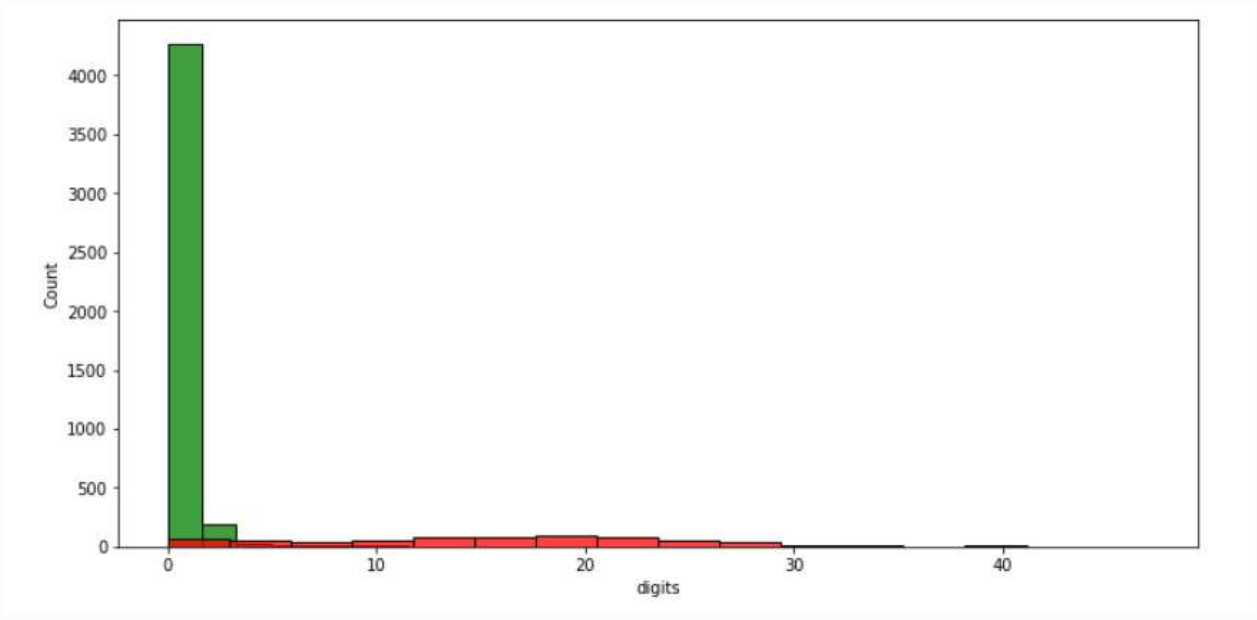
Tech Stack Used:

- Numpy
- Pandas
- Scikit-learn
- NLTK
- Matplotlib and Seaborn

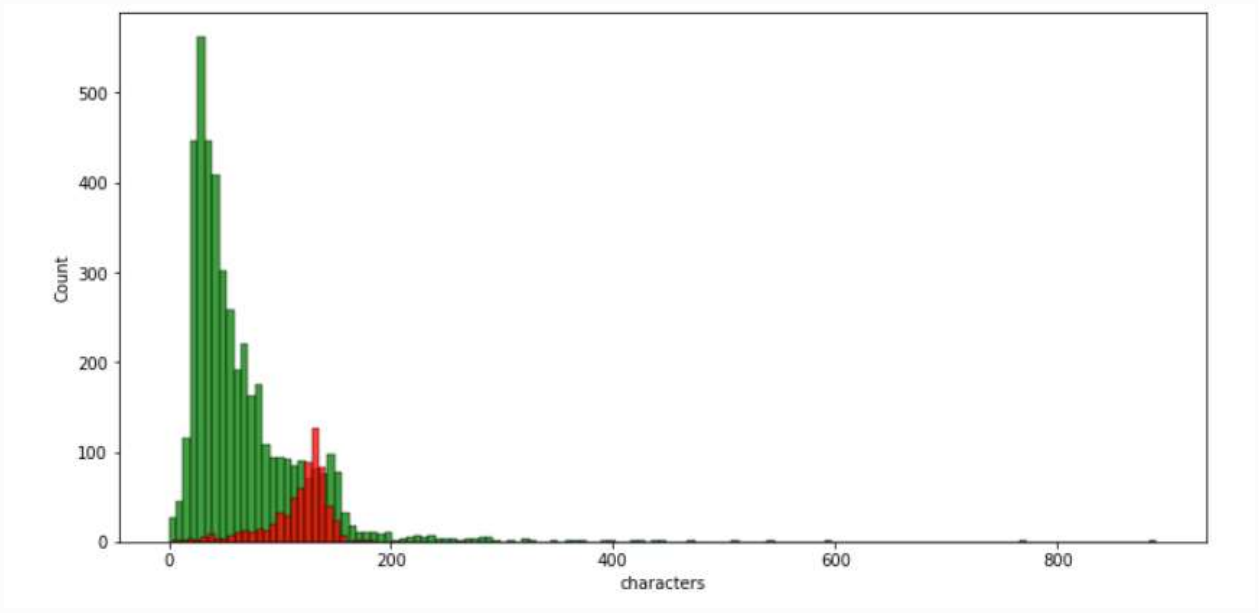
Project Output Images:



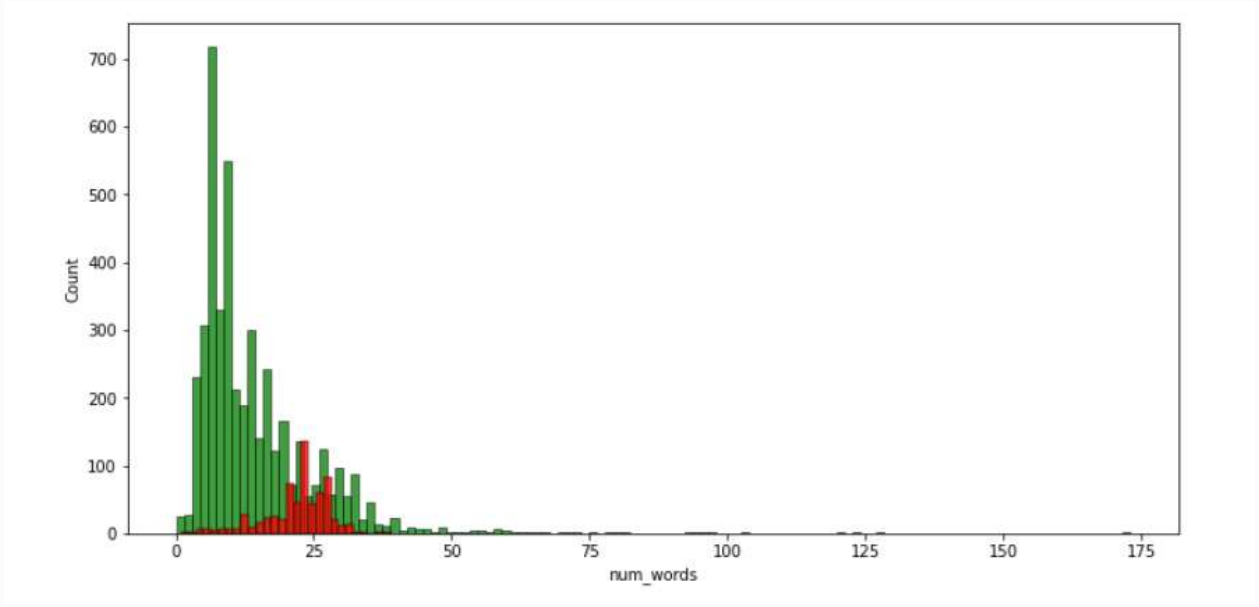
Ham VS Spam



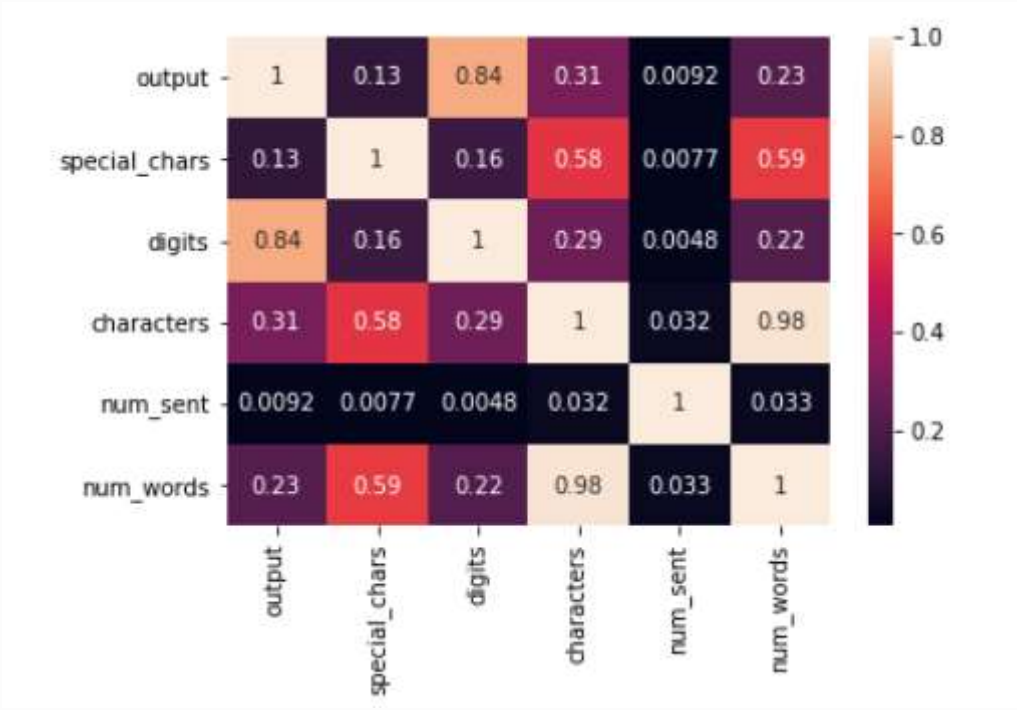
Digits in hams VS Digits in spams



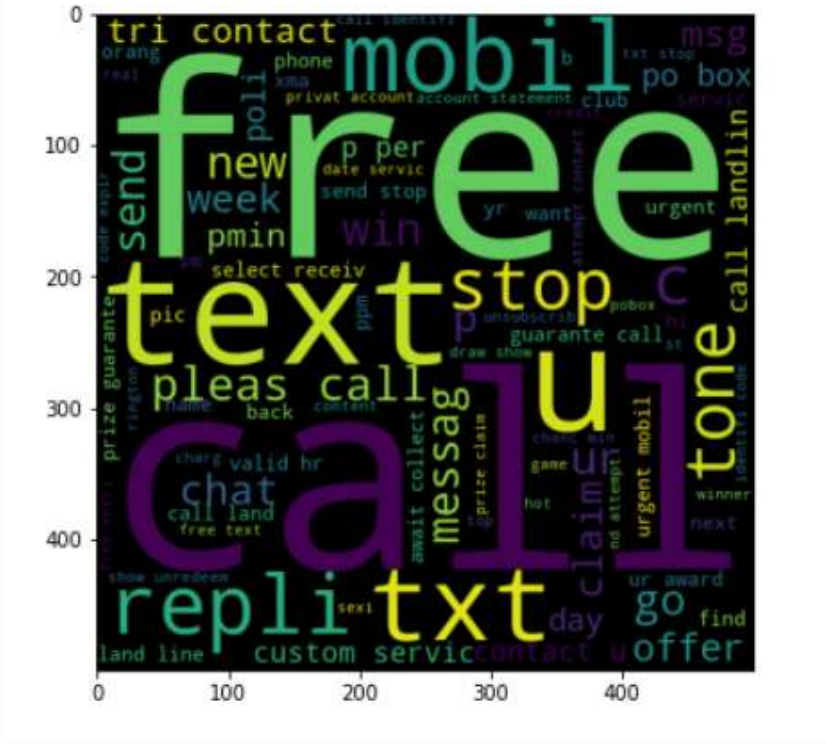
Characters in Hams VS Characters in Spam



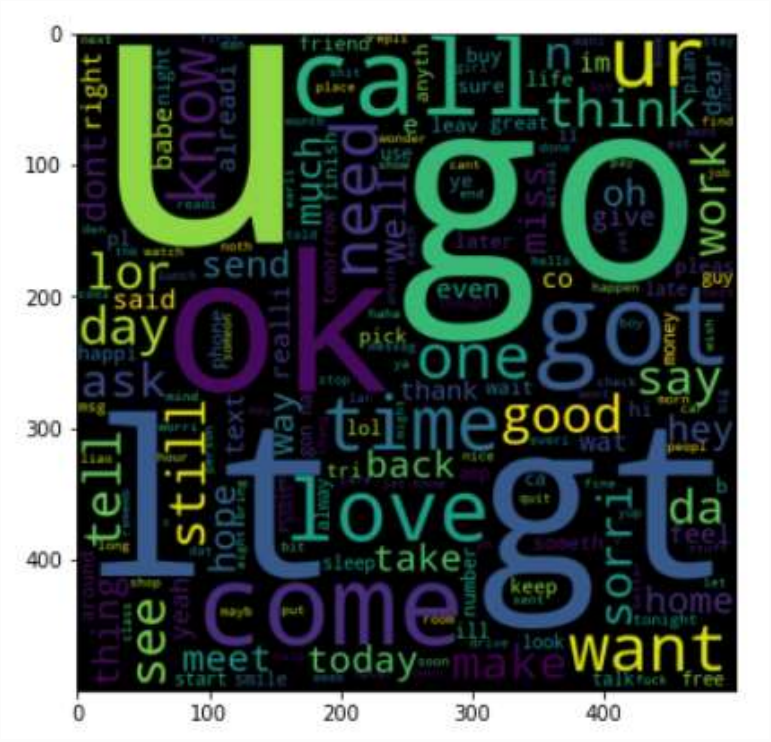
Words in Hams VS Words in Spams



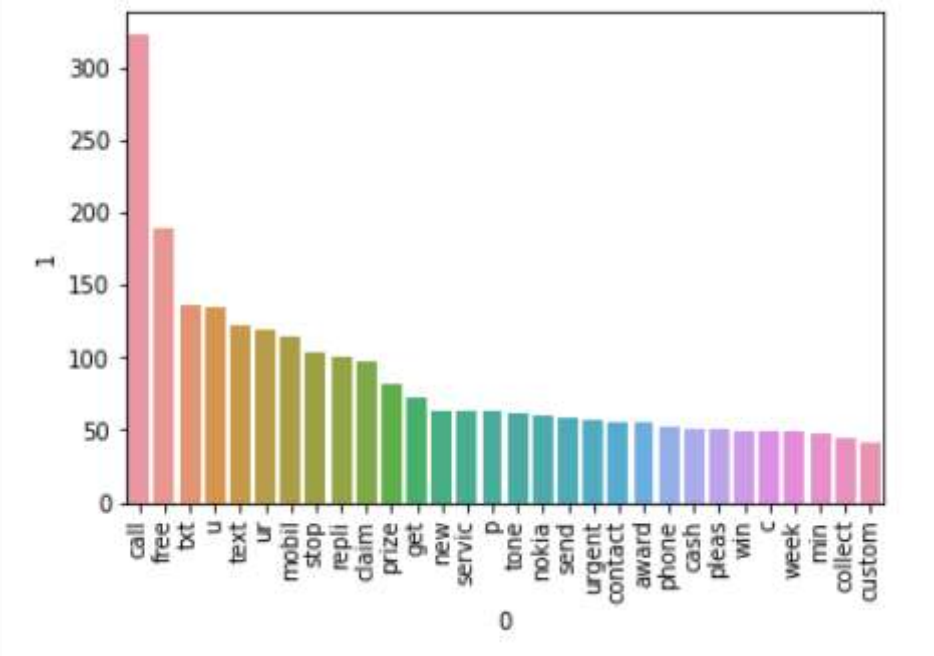
Correlation Matrix



Spam word cloud



Ham wordcloud



Frequency of most common 30 words in spam

```
For  Support Vector Classfiers
Accuracy -  0.9289168278529981
Precision -  0.7219917012448133
Confusion Matrix -
 [[1747   67]
 [  80  174]]
```

- **Results for Support Vector Machines**

```
For  K Nearest Neighbours
Accuracy -  0.9100580270793037
Precision -  1.0
Confusion Matrix -
 [[1814    0]
 [ 186   68]]
```

- **Results for K Nearest Neighbours**

```
For  Multinomial Naive Bayes
Accuracy -  0.9608317214700194
Precision -  0.8145454545454546
Confusion Matrix -
 [[1763   51]
 [  30  224]]
```

- **Results for Multinomial Naive Bayes**

```
For  Gaussian Naive Bayes
Accuracy -  0.8699226305609284
Precision -  0.4823529411764706
Confusion Matrix -
 [[1594  220]
 [  49  205]]
```

- **Results for Gaussian Naive Bayes**

```
For  Binomial Naive Bayes
Accuracy -  0.9656673114119922
Precision -  0.9740932642487047
Confusion Matrix -
 [[1809    5]
 [  66  188]]
```

- **Results for Binomial Naive Bayes**

```
For  Decision Trees
Accuracy -  0.9347195357833655
Precision -  0.9407407407407408
Confusion Matrix -
 [[1806    8]
 [ 127  127]]
```

- **Results for Decision Trees**

```
For  Logistic Regression
Accuracy -  0.9700193423597679
Precision -  0.9848484848484849
Confusion Matrix -
 [[1811    3]
 [  59  195]]
```

- **Results for Logistic Regression**

Learnings:

Challenges Faced: In literature we have studied that many anti- spam strategies have been discovered but still there are some open challenges to these techniques. Some of them are highlighted below:

- In a trust modeling system the user's trust tends to vary over time according to the user's experience and involvement in social networks. Only a few approaches deal with the dynamics of trust by distinguishing between recent and old tags. Future work considering dynamics of trust would lead to better modeling in real world application.
- Most of the existing approaches are based on text information assuming a monolingual environment. However social network services are used by people from various countries, so various languages simultaneously appear in tags and comments. In such cases some text information may be regarded as wrong or considered as spam due to language spam. Therefore incorporating multilingualism in trust modeling would solve this problem.
- It is observed that interaction across social networks has become popular. For e.g. users can use their Facebook accounts to log in to some other social network services. Thus the future challenge is to investigate how trust models across domains can be effectively connected and shared.
- Trust modeling most of the current techniques for noise and spam reduction focus only on textual tag processing and user profile analysis while audio and visual content features of multimedia content can also provide useful information about the relevance of the content and content tag relation. In future challenges could be to combine multimedia content analysis with the conventional tag processing and user profile analysis.

References

- Olatunji, S.O. (2017) *Improved email spam detection model based on support vector machines - neural computing and applications*, *SpringerLink*. Springer London.
- Mujtaba, Ghulam, et al. "Email classification research trends: Review and open issues." IEEE Access 5 (2017)
- Cihan Varol, Hezha M.Tareq Abdulhadi “Comparison of String Matching Algorithms on Spam Email Detection”, International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism Dec, 2018.
- Thashina Sultana , K A Sapnaz , Fathima Sana , Jamedar Najath, 2020, Email based Spam Detection, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 06 (June 2020)