National University of Computer & Emerging Sciences Karachi Campus

Project Report

Spam Email Classification

Group Members:

Muzammil Arif Zuberi (19k-0342)
Zoha Mobin (19k-1469)
Taha Yousuf (19k-1352)

Introduction

We intended to design a program which classifies email messages into ham or spam using standard classifiers. This is used to predict the categories to which the data should belong. Complete Focus was put into Emails.
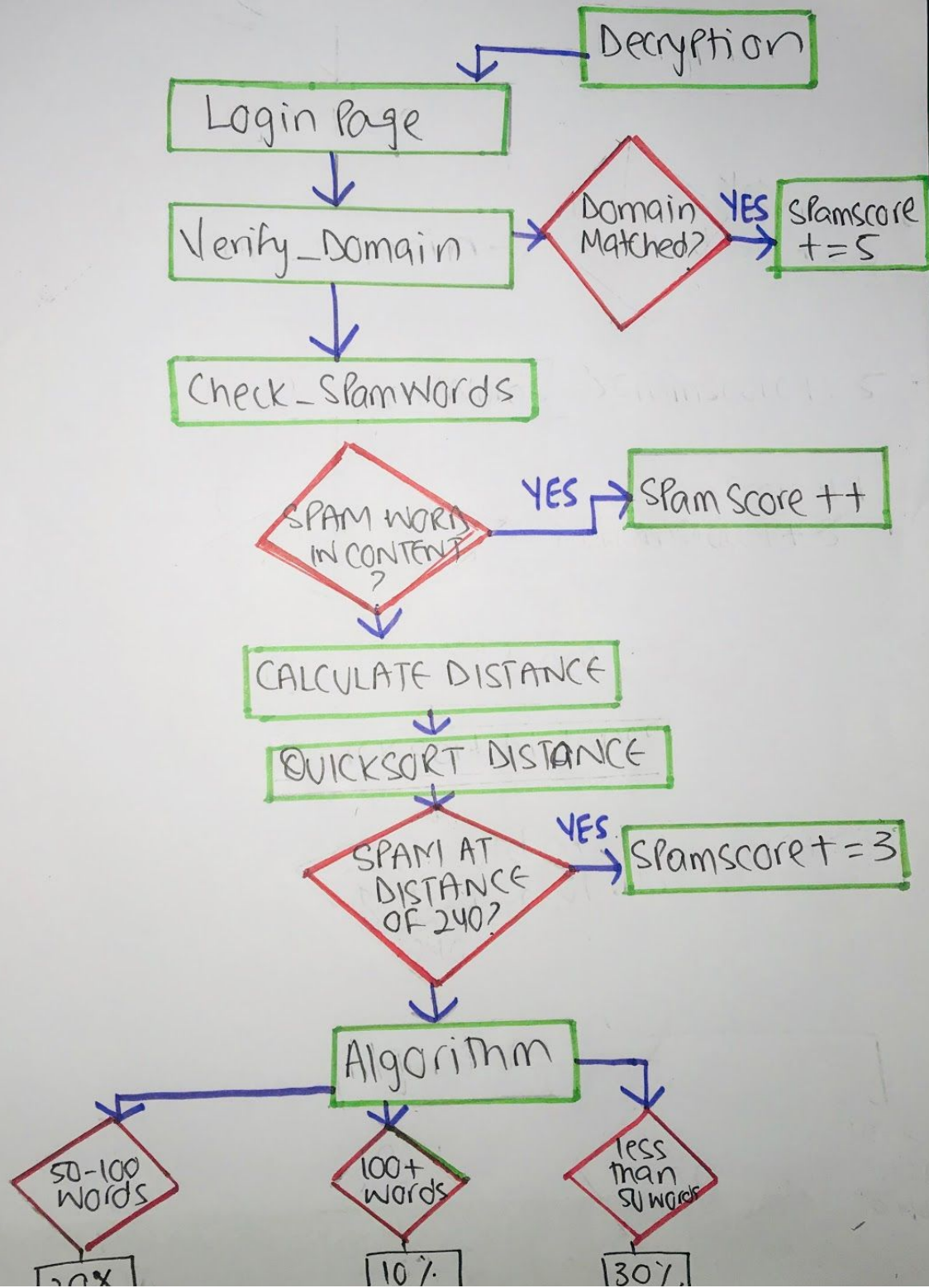
Methodology

Spam emails or commonly known as unsolicited emails that are either generated by computer or sent by unknown individuals to deviate or acquire some personal information ,which is mostly intended for malicious purposes.Therefore protection against them is necessary as they may leak some private information.

We went through certain filters which contributed to the <u>Spam Score Generation</u>

- Login Page
- Decryption of Password using shift cipher
- Domain Verification
- Spam Words Detection
- Checking Substring
- Calculation of the Distance between subsequent spam words.
- Quick Sorting the Distance Array.
- Naive Bayes Algorithm

Flow Diagram

FLOW DIAGRAM

Decryption

Login Page

Verify_Domain → Domain Matched? —YES→ Spamscore += 5

Check_SpamWords

SPAM WORD IN CONTENT? —YES→ Spam Score ++

CALCULATE DISTANCE

QUICKSORT DISTANCE

SPAM AT DISTANCE OF 240? —YES→ Spamscore += 3

Algorithm

50-100 words → 30%

100+ words → 10%

less than 50 words → 30%

Outputs of the System

Login Page


```
----------------------------------------------
          Welcome to Spam Email Classification
----------------------------------------------

              Enter your LOGIN Details

Enter your Email Address:zohamobin
Registered User

Password:fusaf

Access Granted
                           ----------------------------------------------
```

Note: The Emails name weren't hardcoded, they were incremented with respect to the memory block as shown below.


```
email01
email02
email03
email04
email05
email06
email07
email08
email09
email10
New Spam Words
```



```
C:\WINDOWS\system32\cmd.exe                                    —   □   ✕
----------------------------------------------
          Welcome to Spam Email Classification
----------------------------------------------

Enter the Email Number of the email you wish to classify: 04
```

Classified as SPAM

email08 - Notepad

File Edit Format View Help

double your cash now
5 December 2020 at 07:57
muz@nu.edu.pk
suf14@yahoopk
Hi
This is Lucy from London, I've a deal for you to double you cash in a week.
You just need to follow a few steps and claim numerous prizes, sales, rewards,
discounts, fantastic treats full of luxury.
Email me back and win this offer.
Thank you
****

C:\WINDOWS\system32\cmd.exe

```
            ---------------------------------------------
                   Welcome to Spam Email Classification
            ---------------------------------------------


Enter the Email Number of the email you wish to classify: 08

                    SPAM ALERT! Sender's Domain not Matched
                              Spam Score=+5


                        Scanning for Spam Words..
                              Spam Score++

                   Calculating the Distance Between Spam Words..
                              Spam Score=+3


                          Algorithm Calulation..

                   ---------------------------------------------
                   Alert! The Email is SPAM, with SPAM Score: 33
                   ---------------------------------------------
Press any key to continue . . .
```

congratulations _you_have won * !
31 December 2020 at 5:10
mrsmr2021@gmail.com
tahayousuf@nu.edu.pk
Dear congratulations to you and your entire family as we bring to your notice,
the results of the Second Category draw of E-Mail Lottery Promo organized and sponsor by the British Columbia Lottery Canada (BCLC).
We are happy to inform you that your email address attached to Ticket #.: A9564756045011
drew the Winning # 01 02 13 30 35 38 with bonus # 45 have emerged a winner of a total sum of US$1,200,000.00
(One Million Two Hundred Thousand United States Dollars), in cash credited to file CATIS/B16OYQKL/20.
This is from a total cash prize of 150,000,000 Million Dollars,
shared amongst the first One Hundred (100) lucky winners in this category.
Please note that your lucky winning number falls within our Eurasian representative office as indicated in your electronic play coupon.
Our Europe and Asia agent will immediately commence the process to facilitate the release of your funds as soon as you contact our Eurasian Agent's office.
All participants were selected randomly from World Wide Web site through computer draws system and extracted from over 10,000,000 companies and personal emails.
Your winning funds has been Insured with your email address and being deposited to our paying bank.
For security reasons, you are advised to keep your winning information confidential till your claims is processed
and your fund remitted to you in whatever manner you deem fit to claim your prize.
FILL THE BELOW FORM AND SEND IT BY EMAIL TO OUR CLAIM AGENT, FOR THE PAPER WORK OF YOUR PAYMENT
Congratulations, once again our winner.
****

```
--------------------------------------------
         Welcome to Spam Email Classification
--------------------------------------------


Enter the Email Number of the email you wish to classify: 04

             SPAM ALERT! Sender's Domain not Matched
                        Spam Score=+5


                    Scanning for Spam Words..
                         Spam Score++

            Calculating the Distance Between Spam Words..
                         Spam Score=+3


                    Algorithm Calulation..

             --------------------------------------------
             Alert! The Email is SPAM, with SPAM Score: 67
             --------------------------------------------

Press any key to continue . . .
```

<u>Classified as NOT SPAM</u>

email09 - Notepad

File Edit Format View Help

Attendance for 28th September
2 April 2020 at 02:18
zohamobin@nu.edu.pk
aminsadiq@nu.edu.pk
Aslamualikum Sir
I wanted to remind you to mark my presence for 28th September 2020,
with reference to the conversation we had in lab today.
Would be grateful.
Thankyou.
Zoha Mobin
K19-1469
****

C:\WINDOWS\system32\cmd.exe

```
---------------------------------------------
         Welcome to Spam Email Classification
---------------------------------------------


Enter the Email Number of the email you wish to classify: 09
              Sender/Reciever Domains Matched


                  Scanning for Spam Words..
                       Spam Score++

              Calculating the Distance Between Spam Words..
                       Spam Score=+3


                   Algorithm Calulation..


         ---------------------------------------------
                  The Email is NOT Spam
         ---------------------------------------------

Press any key to continue . . .
```

Conclusion

The project we have provided is extremely practical as it serves as a protective wall not only against spam emails but also makes the user identify phishing emails or any other malicious attempts of stealing data.
Our System runs smoothly and includes almost every element taught in the Coal Course.