Air University Multan campus 

Department of Computer Science and Engineering



**Spam Classifier**

**Project proposal**

Prepared by

|  |  |
| --- | --- |
| Group # | Grade : |
| Group Email: [183159@students.au.edu.pk](mailto:183159@students.au.edu.pk) | |
| Group Members: | |
| Kainat Zahra 183159  (Leader) |  |
| Momina Iqbal 183173  (colaborator) |  |
| M Anas Malik 183183  (colaborator) |  |
| M Furqan Amjad 183205  (colaborator) |  |

## Supervised by

## Ma’am Atka Ali

## 

## Sixth Semester 1442 AH

## Spring 2021

**Revision Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Page #** | **Section #** | **Reviewer** | **Corrected by**  **(Reviewer/Author)** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## 

**TABLE OF CONTENTS**

1. Introduction …………………………………………………………………………………..4

2. The Problem ………………………………………………………………………………….4

3. Project Goals and Objectives ……………………………………………………………...5

3.1. Goals ………………………………………………………………………………...….5

3.2. Objectives …………………………………………………………………………...….5

4. The Solution ………………………………………………………………………………….5

5. Project Scope ………………………………………………………………………………..6

6. Hardware and Software Tools ……………………………………………………………..6

6.1 Hardware …………………………………………………………………………………6

6.2 Software ………………………………………………………………………………….6

7. Cost (if applicable) …………………………………………………………………………..7

8. Timeline……………………………………………………………………………………….7

9. Roles and Responsibilities ………………………………………………………………….7

10. Risk management ………………………………………………………………………….8

References ……………………………………………………………………………………...8

1. INTRODUCTION

In today’s globalized world, email is a primary source of communication. This communication can vary from personal, business, corporate to government. With the rapid increase in email usage, there has also been an increase in the SPAM emails. SPAM emails, also known as junk email, involve nearly identical messages sent to numerous recipients by email. These spam emails are illegitimate and unwanted emails that may contain junk, viruses, malicious codes, advertisements or threat messages to the authenticated account holders.

**Context:**

The Internet has become an important and essential part of human life. The increase in the utilization of the internet has increased the number of account holders over various social sites. Email is the simplest and fastest mode of communication over the internet that is used both personally and professionally. Due to the increase in the number of account holders and increase in the rate of transmission of emails a serious issue of spam emails had aroused.

**Problem:**

Most e-mail readers spend a non-trivial amount of time regularly deleting junk e-mail (spam) messages, even as an expanding volume of such email occupies server storage space and consumes network bandwidth. An ongoing challenge, therefore, rests within the development and refinement of automatic classifiers that can distinguish legitimate e-mail from spam. Apart from being annoying, spam emails can also pose a security threat to computer systems.

**Response:**

This serious issue has generated a need for efficient and effective anti-spam filters that filter the email into spam or ham email. Spam filters prevent the spam emails from getting into the user's inbox. Email spam filters can filter emails on content base or on header base. Various spam filters are labelled into two categorizes machine learning

and non-machine learning techniques.

2. THE PROBLEM

Spamming is one of the major attacks that accumulate the large number of compromised machines by sending unwanted messages, viruses and phishing through emails. We have chosen this project because nowadays there are a lot of people trying to fool you just by sending you fake emails like you have won 1000 dollars, this much amount is deposited in your account once you open this link then they will track you and try to hack your information. Sometimes relevant emails are considered as spam emails. Unwanted email irritating Internet consumers.

• Critical email messages are missed and/or delayed.

• Consumers change ISP's all the time looking for consistent email delivery.

• Loss of Internet performance and bandwidth.

• Millions of compromised computers.

• Billions of dollars lost worldwide.

• Identity Theft.

• Increase in Worms and Trojan Horses.

• Spam can crash mail servers and fill up hard drives.

**3. PROJECT GOALS AND OBJECTIVES**

The objective of classification of Spam emails is To classify the email into spam and non spam. Another objective in this work was to improve the spam detection near 95% of accuracy using Artificial Neural Networks.

The system will fulfil the following features:

.

* Convert the plain text files to files with one word per line.
* Convert each file into a list of words.
* Delete stop words from all the files.
* Replacing all words to their stem which reduces the number of words to be considered for representing a document.
* Create an inverted index file. This file has 3 columns – word, filename and frequency of word in the file.

**4. THE SOLUTION**

* There are many things taken into consideration including the content of your emails, how often you send, and your sending reputation. Spam classifier looks at specific parts of an email, such as the subject line, email content and even the header data. The header data tracks where an email came from, where the email goes and when it arrives at various stops. The Spam classifier looks at this data for anything out of the ordinary and demotes emails to the spam folder if anything is out of place.Moreover, there are global lists of spammers called blacklists. The spam classifier classifies the Email sent by Spammers as Spam and are banned from inbox. Spam classifier checks the messages too. It looks for spamming signs such as hate speech, inappropriate content, Emails with suspicious subject lines, messages with excess of images, or shortened URLs. These are red flags; if users might be unknowingly sending emails like that they are labelled as ‘spam.’
* In addition, if you got mails from 100 senders in a very short period of time, all with the same sender name and subject line, then the spammer could be running these accounts with bot scripts. Spam classifier overcomes this problem by sending these mails into spam.

**5. PROJECT SCOPE**

This project needs a coordinated scope of work. These scopes will help to

Focus on this project. The scopes are:

* Modified existing machine learning algorithm.
* Make use and classify of a data set including data preparation,

Classification and visualization.

* Score of data to determine the accuracy of spam detection.
* It considers a complete message instead of single words with respect to its organization. It can be referred to as the intelligent approach due to its message examining criteria.
* It provides sensitivity to the client and adapts well to the future spam techniques. Even if the spam word is slightly modified, this algorithm still succeeds and notices the spam content.
* It lacks action taking for IT security groups. The degree of business in the real time is one more major drawback that is faced by this framework.
* It does not provide any clarity about how well it is performed in real time for the spam campaigns.
* Spammers can easily develop techniques to meet the preventive measures of Auto RE framework like making legitimate domains fall into the list of illegitimate emails.
* The results that are obtained do not provide any aggregate view of the large groups of emails.
* Moreover, it does not let the network administrator to have an online monitoring system across the network

**6. HARDWARE AND SOFTWARE TOOLS**

**6.1. HARDWARE**

**Processor:** Core i7 4th Generation

**Ram:** 8 GB

**6.2. SOFTWARE**

**Language:** HTML, CSS, Bootstrap, Python

**Framework:** Django

**Database:** SQL

**7. COST (IF APPLICABLE)**

Project doesn’t contain any major cost because we are using our own hardware device and free software for the development of our website

**8. TIMELINE**

1: First of all we will register our website.

2: We will make frontend of a webpage by designing display

3: We will store data respectively.

4: Then we will apply a suitable algorithm.

5: We will use Bootstrap and CSS for styling web pages.

**9. ROLES AND RESPONSIBILITIES**

|  |  |  |
| --- | --- | --- |
| **Role** | **Responsibilities** | **Participants** |
| Project Leader | Overall Management of Web | Kainat Zahra |
| Analysis | Gathering and Comparing information about the web | Kainat, Momina |
| Design | Front End Development--------  Back End Development-------- | Furqan, Anas  Kainat, Momina |
| Implementation | HTML, CSS, Bootstrap,  Django, SQL, Python, Pandas, Numpy | Furqan, Anas  Kainat, Momina |
| Testing | Checks for functionality, usability, bugs, compatibility, performance of the web application | Furqan, Anas |
| Domain Experts | Overall Check and Balance of the Web | Furqan, Anas  Kainat, Momina |

**10. RISK MANAGEMENT**

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Type** | **Probability** | **Strategy** |
| Database failure | crash | 20% | Inspection of database |
| System/Machine Failure | File corruption/ system hack | 20% | Files recovery through internet |
| Server Down | Internet issue | 30% | High availability of Internet |

**REFERENCES**

<https://www.w3schools.com/>

<https://docs.python.org/3/tutorial/>

<https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django>

<https://www.scribd.com/doc/61315817/Intra-Mailing-System>

<http://www.ijesrt.com/issues%20pdf%20file/Archive-2017/April-2017/18.pdf>

<https://www.academia.edu/36614854/Spam_Filtering_Using_ML_Algorithms>