**UB # 20027555  
Muhammad Shoaib Zulfiqar**

**Research Proposal**

**On**

**Use of artificial intelligence and data analytics to compare the e-mail spam filtering (YAHOO, Gmail, Outlook) depending upon user experiences**

Table of Contents

[1. Introduction 3](#_Toc74296689)

[1.1 Research Problem 4](#_Toc74296690)

[1.2 Rationale of the Study 5](#_Toc74296691)

[1.3 Scope of the study 5](#_Toc74296692)

[1.4 Research Aim 5](#_Toc74296693)

[1.5 Research Objectives 5](#_Toc74296694)

[1.6 Research Questions 6](#_Toc74296695)

[2. Literature Review 7](#_Toc74296696)

[3. Methodology 9](#_Toc74296697)

[3.1 Research Methodology: 9](#_Toc74296698)

[3.2 Research Strategy 9](#_Toc74296699)

[3.3 Research Design 9](#_Toc74296700)

[3.4 Data Collection 10](#_Toc74296701)

[3.5 Data Analysis 10](#_Toc74296702)

[3.6 Research Limitations 10](#_Toc74296703)

[3.7 Ethical Considerations 10](#_Toc74296704)

[3.8 Expected Outcome 11](#_Toc74296705)

[Project Plan/Milestone 13](#_Toc74296706)

[References 15](#_Toc74296707)

# Introduction

Many unwanted emails need to be filtered because they are spam. That's why; there is an immense need for anti-spam filters that are more robust. Our main aim is to identify the spam filters like Google, Yahoo, and Outlook and compare their efficiency. Effective research would be conducted to know the user reviews about spam emails and their performance. These experiences will then be reached using the data analytics technique to find the best among all methods.

Email is a fast and active method of sending and receiving messages among people. It is an inexpensive way of exchanging messages. Any email can be sent or receives, either as it personal or a company email. With the advancement in technology and the internet, people are more focusing on automated means of communication. As internet and web resource usage is increasing day-by-day, with the increase of internet, usage of internet resources is also growing. People utilize different email platforms according to their needs and demands. So, everyone has the liberty to use the email platform of his/her choice. Depending upon the usage, people have different experiences as various advantages have been observed regarding sending and receiving messages. So, other ambiguities and problems still exist, which can cause serious damages to people and society (Bhowmick, 2016).

Email spam is a serious problem because it can cause many damages to companies and individuals. With the increase of sending emails, spam emails have also increased. Due to the rapid rise in the usage of emails, many malicious users are trying to discourage people by sending spam emails. Although many spam filtering tools and techniques have been developed and utilized, the problem of spam still exists. So, people or the users cannot ignore this issue which is increasing day by day. Spam contains different unsolicited emails and fraudulent emails that can mislead people. In this way, people would start disbelieving the emails and their related platforms. So, anti-spam filtering techniques are important because they can stop spam emails from being a part of normal emails. Spam emails also cause security problems because they can cause serious damages to people, companies, and businesses (GV, 2008).

Email sending is a cheap and easier way of communicating and approaching people and customers easily. So, the malicious people use these resources to approach the people. Suppose we see the spam emails at the organizational level. In that case, many problems can be caused by this, i.e., loss of productivity in work, aggressive responses by the employees, wrong usage of the email space or memory, miscommunication among employees and the head, etc. So, these problems still exist and can cause damaging issues for many users. So, this issue must be considered and focused on to get out of this problem (Siponen, 2006).

A lot of people use email as a way of communication to send or receive messages daily. So, if we see from the spammer's point of view, they try to target most people. Even if a small number of people become a part of their mission, it becomes a great advantage for them. Many types of spam still exist which are trying to take benefits from different people. It was analyzed that in the beginning, when the spam emails were sent, the spammers only target the email body. So, the spam filtering techniques were utilized to sort out the spam and actual emails. Furthermore, with the advancement and passage of time, the spammers also use advanced methods to send spam emails (Cranor LF, 1998).

Finally, three chapters will be included in this research proposal. The first chapter gave introductory information on the topic. The second portion analyzes the literature available and proposes a theoretical framework for Email spam and the approaches used to filter it. The concluding part discusses the methodology and provides further information on the study's context and data collection. Finally, this study examines the research's limitations and ethical implications. In the last, project Plan is also illustrated through Gantt chart representation.

## 1.1 Research Problem

Siponen et al., Bhowmick et al. have shown different email spam filtering techniques. None of the studies has focused or narrow-downed the scope to compare the spam filtering performed by other email platforms like Google, YAHOO, and outlook. That's why this research has raised the problem of comparing different email platforms to find the best among the specified.

## 1.2 Rationale of the Study

Different researches of Radicati et al., and Bhowmick et al., have been observed. These researches show that most of them have utilized their efforts to devise new techniques of email spam filtering. It was identified that comparison among different email platforms had not been conducted yet. This research has focused on comparing the other email platforms and analyzing which email platform best filter out spam emails. As none of the studies has been found out the performance evaluation of the cross platforms of emails, this research would be unique in this perspective, providing the comparative evaluation regarding the problem identified.

## Scope of the study

The scope of the study is to gather the user's and customers' experiences regarding email spam filtering services across different platforms. Mostly, this research found that other people use more than one email service provided by various media. So, to find the best email platform, this research will use questionnaires and collect the users' responses about the spam-email services across different platforms. This research would target the students who use email services daily. They would provide the reviews through the questionnaires, and their responses would be used to evaluate the result. Different analysis techniques would be utilized to find the outcomes.

## 1.4 Research Aim

This study aims to determine the use of artificial intelligence and data analytics to compare the e-mail spam filtering (YAHOO, Gmail, Outlook) depending upon user experiences. Also, to reach the spam filtering of different platforms, i.e., YAHOO, Gmail, and Outlook. After comparing the spam filtering of these platforms, other analytics techniques would be utilized to find the best among them.

## 1.5 Research Objectives

The main research objectives of this study are as follows:

* To compare the email platforms through the user experiences
* Find the best among the email platforms regarding spam filtering depending upon user responses.
* A new and advanced way would be sorted out, identifying the methods of different email platforms and services.

## 1.6 Research Questions

Different research questions have been summarized, which would be answered by the research.

1. Gmail is more effective as compared to the YAHOO and outlook services.
2. Yahoo mail is more effective as compared to Gmail and outlook services.
3. Outlook is more effective as compared to the Gmail and Yahoo mail services.

# Literature Review

Email spam is getting worse day by day. So, it is necessary to figure out this problem to secure people from spam emails. Spam emails can cause many security issues, harming people, the company, and its employees. Many techniques have been devised and developed to resolve the problem of email spam filtering. Filtering techniques are very important because they can help people to stay out of danger. Email is used worldwide for communication purposes, and a gradual increase in its users was found at the end of 2016 (Radicati, 2016). The increase in the usage of emails has caused many problems like spam. Many types of spam emails are found and sent to people depending upon the purpose or the spammers.

Email sending is a cheap and easier way of communicating and approaching people and customers easily. So, the malicious people use these resources to approach the people. Suppose we see spam emails at the organizational level. In that case, many problems can be caused by this, i.e., loss of productivity in work, aggressive responses by the employees, wrong usage of the email space or memory, miscommunication among employees and the head, etc. as the usage of spam emails have increased, it was necessary to sort out and find the filters which can stop these spam emails. Many types of spam filtering techniques needed to be devised. Initially, It was thought that introducing legal measures could help them figure out this problem of spam emails. Later, a more effective way was utilized through the usage of techniques (Siponen, 2006).

Different types of spam filtering techniques were suggested and invested, which include anti-spam filters and content-based filters. Most spammers use fake addresses to send spam emails, so content-based spam filtering techniques were utilized to sort spam filtering effectively. But the content-based filtering requires manual help in writing the styles which are used to send spam emails. But the manual method requires a lot of time and effort. Experts are needed to accomplish this task (Cranor LF, 1998).

To resolve the issue of anti-spam filtering, machine learning techniques were introduced. Supervised learning methods were utilized to sort out the spam filtering messages. A thorough analysis was conducted to combine the spam and anti-spam messages. After combining both forms of messages, Naïve Bayes classifier was applied to perform the evaluation. The cost of analyzing both records of messages was analyzed to achieve evaluation. Furthermore, cross-validation of 10 folds was used to find and compare the price, which tells that the cost estimates vary, which is a big issue and has not been addressed yet (Sahami, 1998).

It has been discussed that the bulk of emails are sent daily by spammers. Many spam filters have been introduced, like blacklist analysis, keyword matching, Bayesian analysis, much other content analysis, scanning techniques, etc., but many spam emails are received. This doesn't mean that we don't have strong classifiers for the classification of the emails, but spammers reproduce their methods according to the new techniques. So, to introduce more powerful spam filtering techniques, many classifiers were practiced like C4.5 ID3, a decision tree classifier, Naïve Bayes classifier, and multi-layer perceptron. These classifiers were utilized for the learning processes of the features of the spam emails, and then the model was designed to differentiate the spam and anti-spam emails. A variety of results with different accuracies was obtained, which shows the diversity of the models (Christina, 2010).

Spam email filtering is getting attention day-by-day as the number of spam emails is increasing very rapidly. There is a need to devise an anti-spam filter that can help humans to filter out emails. One of the most used techniques for the representation of spam emails is the Bag of words. But, recent researches show that the hybrid representation outperforms as the neural network model was utilized with an advanced technique called paragraph vector distributed memory. This technique was used to alter the representation of the emails and to filter the emails. Different spam datasets were utilized, i.e., Enron spam and Ling spam dataset. The study results shows that the proposed justifications outperform, and results were measured (Douzi, 2020).

# Methodology

First of all, questionnaires would be made which would be effective for getting responses from the students. The students will be asked about their views regarding the best email platform depending upon the spam filtering. A questionnaire would be prepared using the Likert scale to know about the opinions of the students. After making the questionnaire, we would target students to collect their views about the best email platform. Google forms would be utilized to manage the responses from the users. After getting the responses from the students, these responses would be then analyzed using the Software, i.e., Statistical Package for Social Sciences. Different tests would be performed, i.e., regression analysis, ANOVA, and p-test. Statistical methods could also be utilized to find the statistical measures specifically (Bhuiyan et al., n.d.).

## 3.1 Research Methodology:

Secondary data will be the primary source to resolve this problem since it represents a highly controlled system and is also completely safe to depend on because of time limits. In this research, qualitative data is used, such as journals, articles, and publications, to compare e-mail spam filtering using artificial intelligence and data analytics (Worldwidescience.org, 2011).

## 3.2 Research Strategy

The analysis in this study is accomplished via the use of a questionnaire approach. A questionnaire based on the Likert scale will be utilized to obtain responses from people surveyed in this investigation. A questionnaire is a kind of research instrument in which a series of questions and rating scales are used. It is used to elicit detailed information from respondents. Due to budgetary limits, time restrictions, and the need for measurement accuracy, this methodology is suitable for addressing the study's research goals and objectives (Sanjiban Sekhar Roy and Dr.Madhu Viswanatham V, 2016).

## 3.3 Research Design

The qualitative strategy is used in this study because it is the most effective method for understanding behavior and actions. Hence, it is the most appropriate method for addressing the study's research topic. This study aims to evaluate the e-mail spam filtering capabilities of Yahoo, Gmail, and Outlook using artificial intelligence and data analytics (Information Age, 2019).

## 3.4 Data Collection

For data collection, Google forms are used to get responses from users to get more information on artificial intelligence and data analytics, such as e-mail spam filtering (Yahoo, Gmail, and Outlook). This strategy is more suited to a time frame with limited time since it is more accurate and constant (Gangavarapu, Jaidhar, and Chanduka, 2020). Research is going to target about 200 university students. Sampling method would be random. As the research is going to be conducted purely to know about the responses of students, so, students from different univeristies would be targeted in the research.

## 3.5 Data Analysis

Data will be collected through Google forms and analyzed using Microsoft Excel and SPSS. Different statistical and analysis tests were performed to evaluate the performance. After completing other analysis techniques, the results will be assessed and compared. The comparison evaluation would be performed to get the best method regarding spam filtering emails (Dada et al., 2019).

## 3.6 Research Limitations

Due to the pandemic situation today, the study employs a secondary technique to avoid gathering primary and updated data. Data sampling for the sake of convenience is implausible. This research used a convenient survey. Recognize that chance polls taken at random have a demographic component. These constraints constrain the individual during the procedure. The study's primary concerns include time and financial limits, as well as a lack of reliability.

## 3.7 Ethical Considerations

According to research ethics, it is necessary to maintain the confidentiality and trust of the respondents. Even when the research is conducted, it is essential to hide the identity of the respondents. Moreover, the responses of the respondents should also be kept confidential. Furthermore, the identity of the individuals should also be kept confidential. The respondents should be informed why you are collecting the data, and you should not hide the basics of your research. In this way, respondents would be friendlier and would not hesitate regarding the responses and the analysis.

This research ensures that the data that will be collected would be kept confidential. This data would be only used for relevant purposes and would not be utilized for any other negative purposes. Privacy would be maintained, and their opinions would not be disturbed.

The research also aims to focus on the approval and consent from the respondents. The study also seeks to keep the respondent's reviews confident. This research aims to hide the identity of the individuals to keep them confidential regarding the survey and the analysis. Moreover, the study would be used confidentially to save the environment friendly. The research aims to support the respondents in a trustworthy environment, so that the respondents i.e., students, don't hesitate to fill the forms. This attitude will help them to conduct and complete the research successfully.

## 3.8 Expected Outcome

It is expected that after conducting the survey and collecting the data, a data analytics technique would be used to compare the reviews. The analysis of the data performed would be then utilized to perform the analysis and answer the research questions.

The research results gathered using the questionnaires and data analysis would be utilized to evaluate the performance of the research. The expected outcome is one of the email platforms depending upon the user evaluation. The study done collecting the data through the Google forms would be then utilized in excel and SPSS to perform various tests. Furthermore, Google forms also generate charts regarding the responses of the users. These charts could also be used to observe the users' responses regarding specific questions of the research. These charts could help analyze the overall response of the users. Moreover, the responses will be generated and analyzed through SPSS to get the statistical observations and another test. These test results would be used to interpret the results of the respondents (Bhowmick and Hazarika, n.d.).

The current research strategy is designed to handle research problems successfully and to accomplish current research objectives. The technique is based on diagnostic analysis targets and the causes of how companies respond to scandals; further, the research indicates how companies recover their Legitimacy (Suchman, 2017).

# Project Plan/Milestone

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Task details** | **May**  **– June 2021** | **July 2021** | **July**  **– August 2021** | **August**  **– September 2021** | **September**  **–**  **November 2021** | **December 2021**  **–**  **February 2022** | **February –**  **March 2022** |
| **Identification of Topic and Problem** |  |  |  |  |  |  |  |
| **Aim and Objectives of Research** |  |  |  |  |  |  |  |
| **Literature Review** |  |  |  |  |  |  |  |
| **Selecting Methodologies** |  |  |  |  |  |  |  |
| **Extracting Extensive Literature** |  |  |  |  |  |  |  |
| **Reviewing and Filtering the Literature** |  |  |  |  |  |  |  |
| **Analysing the Data** |  |  |  |  |  |  |  |
| **Conclusion** |  |  |  |  |  |  |  |

The above plan shows the task performance and the milestones in detail monthly. All the tasks would be performed on the monthly basis in order to achieve the overall objectives. These tasks would be performed in order to get the research work done as proposed.

# References

Bhowmick, A. a. (2016). Machine learning for e-mail spam filtering: review, techniques, and trends. *arXiv preprint arXiv:1606.01042.*

Christina, V. S. (2010). Email spam filtering using supervised machine learning techniques. *International Journal on Computer Science and Engineering (IJCSE), 2*(09), 3126-3129.

Cranor LF, L. B. (1998). Spam! *Communications of the ACM.*

Douzi, S. F. (2020). Hybrid email spam detection model using artificial intelligence. *Int. J. Mach. Learn. Comput, 10*(2), 316-322.

GV, C. (2008). Email Spam Filtering: A Systematic Review. . *Foundations and Trends in Information*, pp. 335–455.

Radicati. (2016). *Email Statistics Report.* Executive Summary. Tech. Rep. 650, 2012-2016.

Sahami, M. S. (1998). A Bayesian Approach to Filtering Junk E-Mail. *Learning for Text Categorization, Papers from the AAAI Workshop* (pp. 55–62). Madison Wisconsin. AAAI Technical Report WS-98-05.

Siponen, M. a. (2006). Effective anti-spam strategies in companies: An international study. *In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)* (pp. 127c-127c). IEEE.

Worldwidescience.org. (2011). *spam elimination solution: Topics by WorldWideScience.org*. [online] Available at: https://worldwidescience.org/topicpages/s/spam+elimination+solution.html

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

Sanjiban Sekhar Roy, & Dr.Madhu Viswanatham V. (2016, February). *Classifying Spam Emails Using Artificial Intelligent Techniques*. ResearchGate; Trans Tech Publications, Ltd. https://www.researchgate.net/publication/297607119\_Classifying\_Spam\_Emails\_Using\_Artificial\_Intelligent\_Techniques

Bhuiyan, H., Ashiquzzaman, A., Juthi, T., Biswas, S., & Ara, J. (n.d.). A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques Software & Data Engineering Global Journal of Computer Science and Technology: C Volume 1 Issue 2 Version 1.0 Year 2018 Publisher: Global Journals 8. https://core.ac.uk/download/pdf/231154948.pdf

Gangavarapu, T., Jaidhar, C. D., & Chanduka, B. (2020). Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-020-09814-9

*Can artificial intelligence spot spam quicker than humans?* (2019, April). Information Age. https://www.information-age.com/artificial-intelligence-spam-machine-learning-123481368/

‌Bhowmick, A., & Hazarika, S. (n.d.). *Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends*. https://arxiv.org/pdf/1606.01042.pdf

‌

‌

‌