**Probability & Statistics**

**Project Report**

**Naive Bayes Theorem and**

**Spam Email Detection**

**By**

|  |  |  |
| --- | --- | --- |
| 1 | Sumaira Majeed | Roll No 33 |
| 2 | Rayan Badar | Roll No 18 |



|  |  |  |
| --- | --- | --- |
| Name of Instructor | Faiqa Ali | Lecturer  Department of Mathematics |

**Department of Computer Sciences**

**Namal University Mianwali, Pakistan**

**Submission Date: 16th June, 2025**

**TABLE OF CONTENTS**

[1. INTRODUCTION: 3](#_Toc201007266)

[2. PROJECT STATEMENT: 3](#_Toc201007267)

[3. PURPOSE: 3](#_Toc201007268)

[3.1. MOTIVATION AND OBJECTIVE: 3](#_Toc201007269)

[3.2. DIRECTION AND AIM: 3](#_Toc201007270)

[3.3. APPLICATION AND SIGNIFICANCE: 3](#_Toc201007271)

[4. DATA ANALYSIS: 4](#_Toc201007272)

[5. RESULTS AND GRAPHS: 4](#_Toc201007273)

[5.1. PYTHON BASED RESULTS: 4](#_Toc201007274)

[5.2. PYTHON BASED GRAPHS: 6](#_Toc201007275)

[5.3. EXCEL BASED RESULTS: 8](#_Toc201007276)

[5.4. EXCEL BASED GRAPHS: 9](#_Toc201007277)

[6. CONCLUSION: 10](#_Toc201007278)

[7. REFERENCES: 10](#_Toc201007279)

# INTRODUCTION:

With the growth of the internet, spam emails have become a significant concern for individuals and organizations. Various techniques have been developed to detect spam, and one effective method is the use of machine learning, particularly the Naive Bayes theorem, which is known for its simplicity, efficiency, and high accuracy in text classification problems. The dataset used in this project consists of labeled email messages categorized as either "spam" or "ham" (not spam), where each data entry is an ordered pair: (message, category). The main variables include the message, which is the text content of the email, and the category, which indicates whether the message is spam (1) or ham (0). The tools used for this project include the Python programming language along with libraries such as Pandas, Scikit-learn, NLTK, and Matplotlib, and Tkinter was used to design the graphical user interface.

# PROJECT STATEMENT:

This project aims to develop a machine learning model using the Naive Bayes algorithm to detect and classify spam emails from normal ones. A user-friendly interface is also built to interact with the model by using Python and also we have worked on excel based manual datasets.

# PURPOSE:

### **MOTIVATION AND OBJECTIVE:**

We chose this project because spam emails are a widespread issue, and finding an efficient way to detect them can save time and improve cybersecurity. The goal is to create an accurate and fast spam detection system.

### **DIRECTION AND AIM:**

The project focuses on using statistical learning to identify patterns in emails and classify them accurately as spam or ham. It applies Natural Language Processing (NLP) for preprocessing and vectorization.

### **APPLICATION AND SIGNIFICANCE:**

Spam detection plays a vital role in securing digital communication. With the help of statistical tools like Naive Bayes, the model applies real-world data and probability to solve this issue effectively.

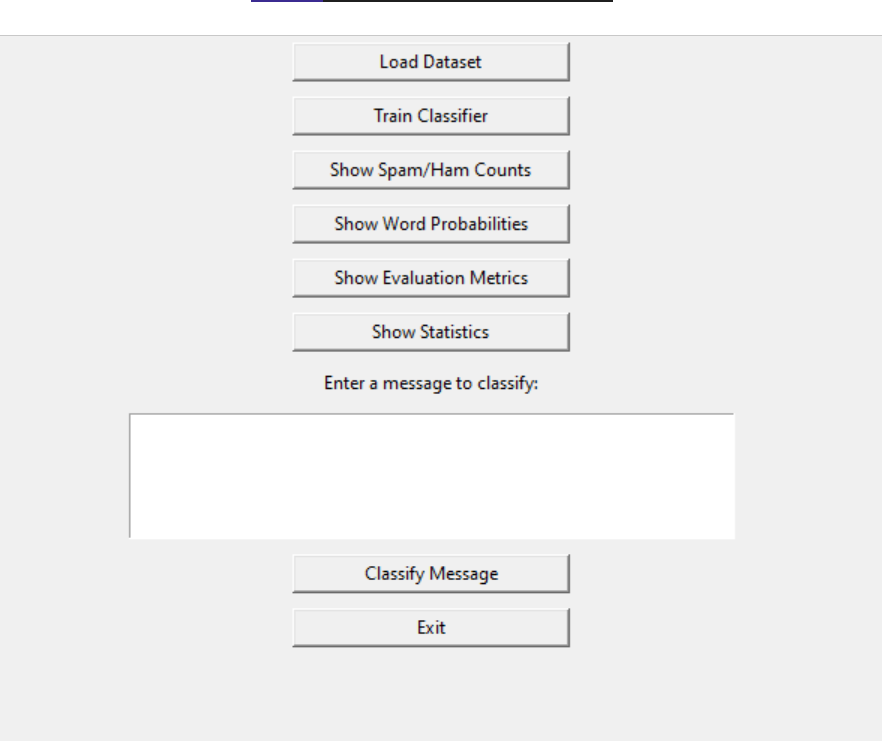
# DATA ANALYSIS:

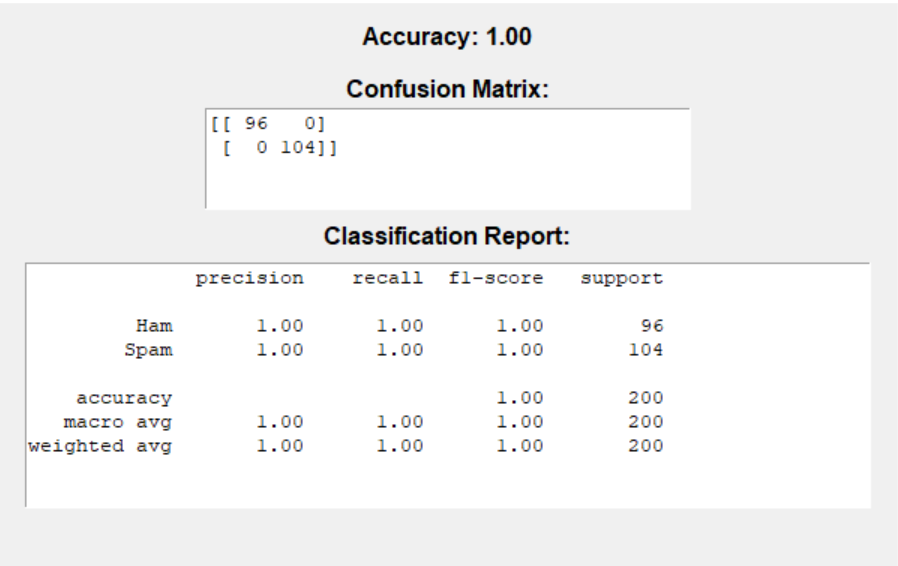
The following statistical analysis was performed on the dataset:

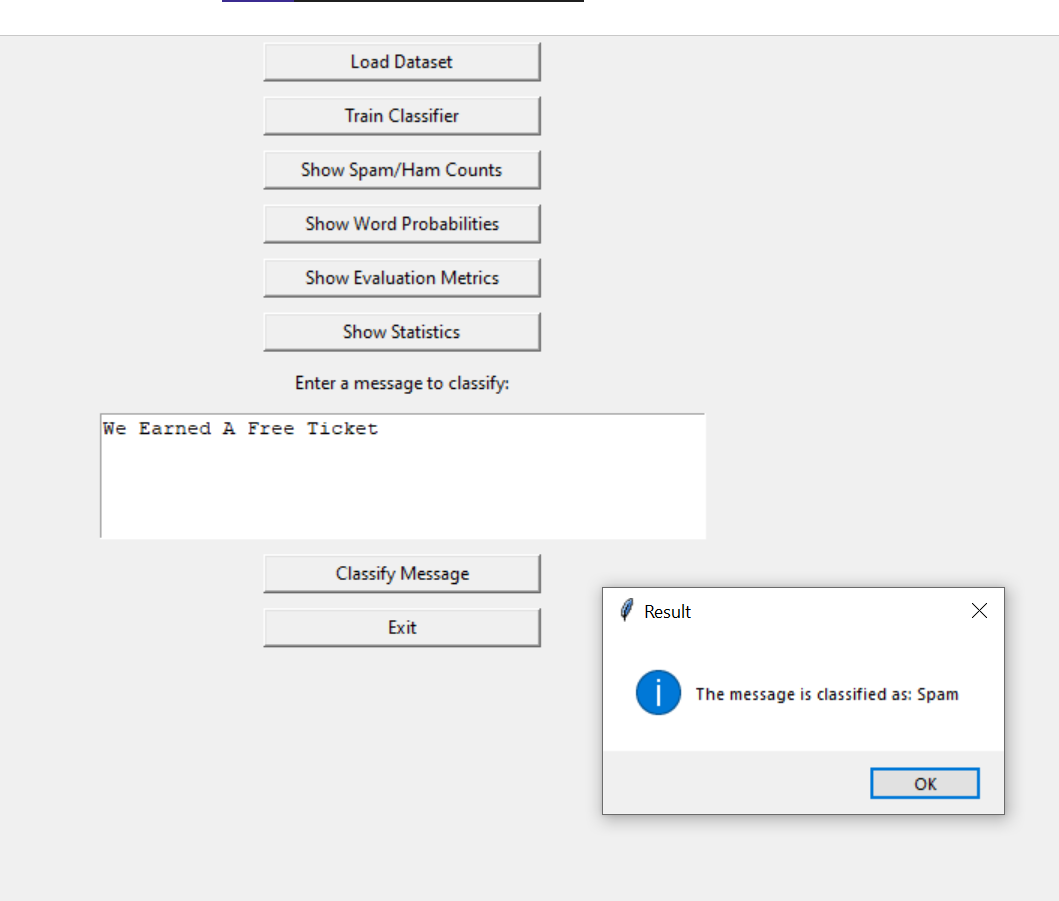
1. **Mean:** Calculated for numeric features such as label frequency.
2. **Median and Mode:** To understand the central tendencies of label distribution.
3. **Quartiles and Percentiles:** Used to study the spread and skewness of data.
4. **Variance and Standard Deviation:** Used to evaluate variability in message characteristics.
5. **Histogram:** A histogram was plotted to check the normal distribution of spam and ham counts.
6. **Correlation Coefficient:** Analyzed correlation between word frequency and label classification.

# RESULTS AND GRAPHS:

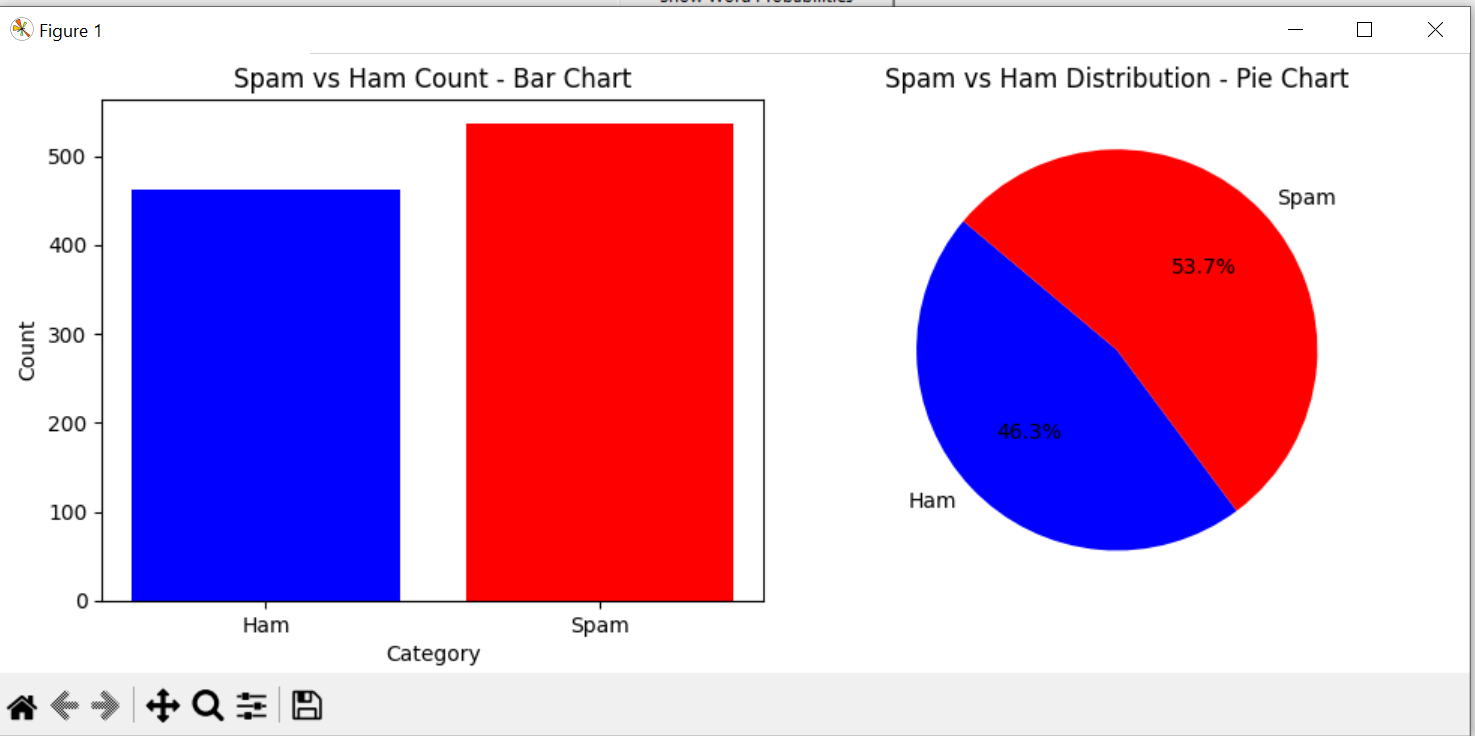
### **PYTHON BASED RESULTS:**

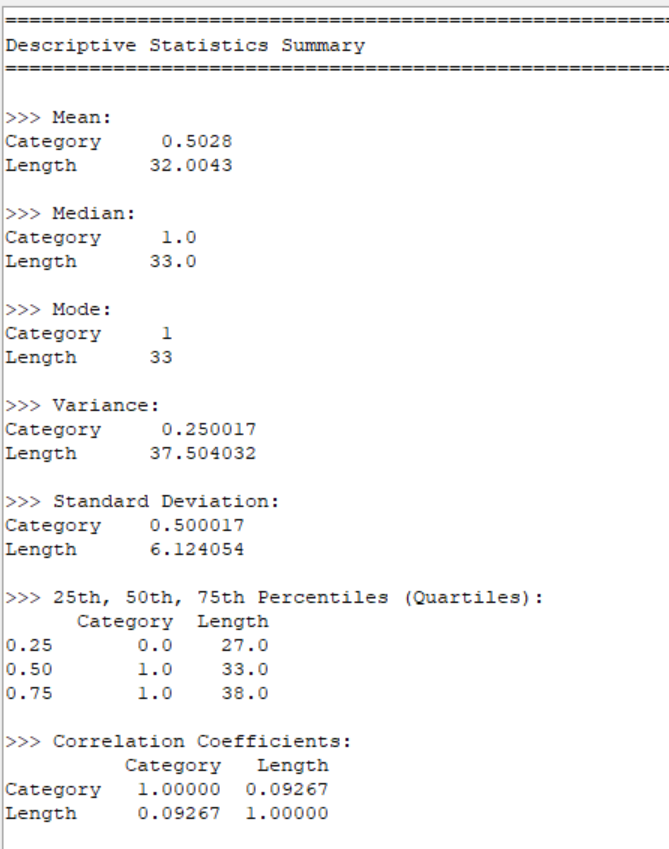


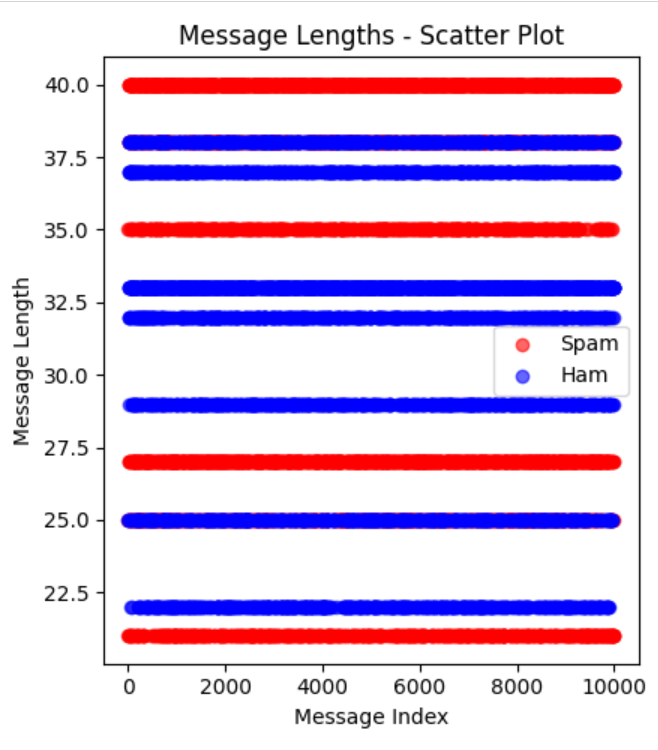




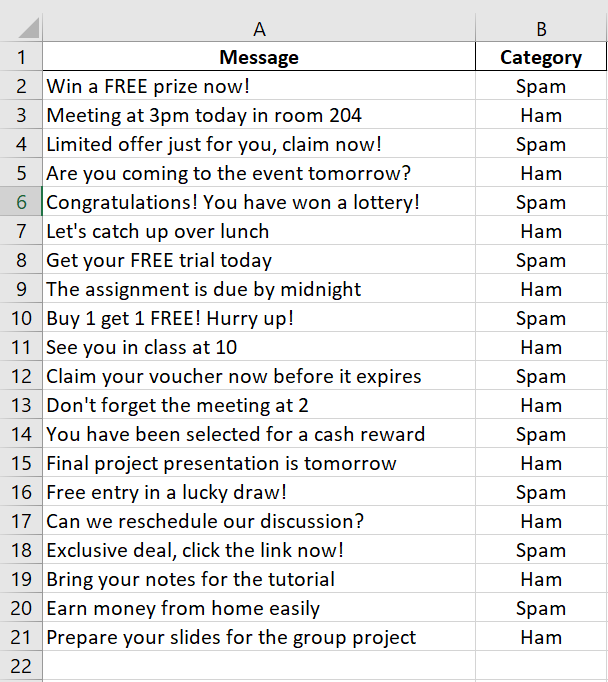
### **PYTHON BASED GRAPHS:**

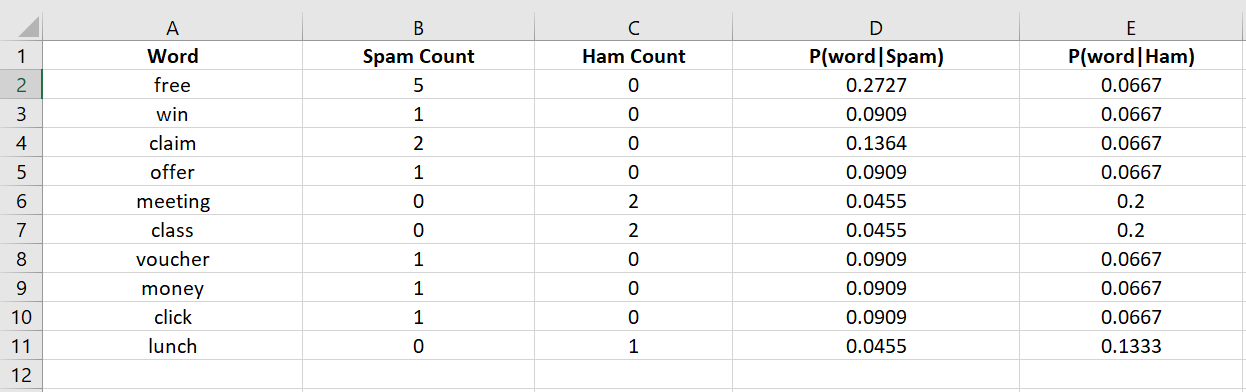


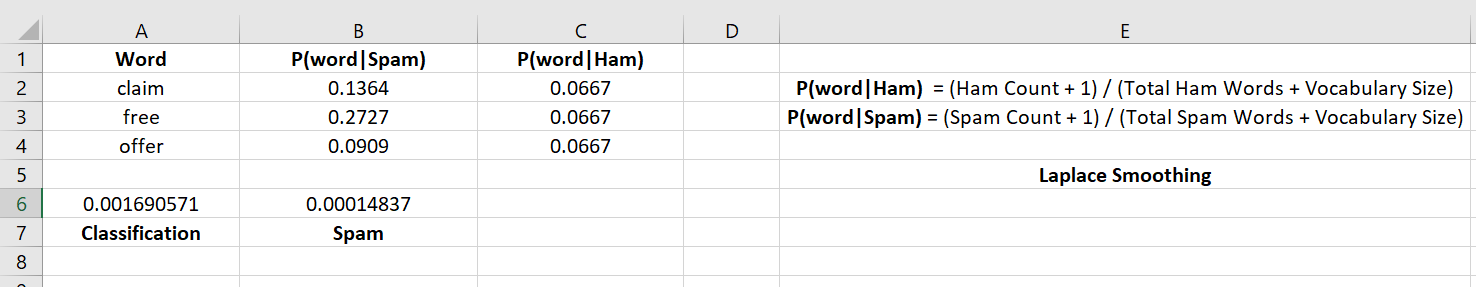




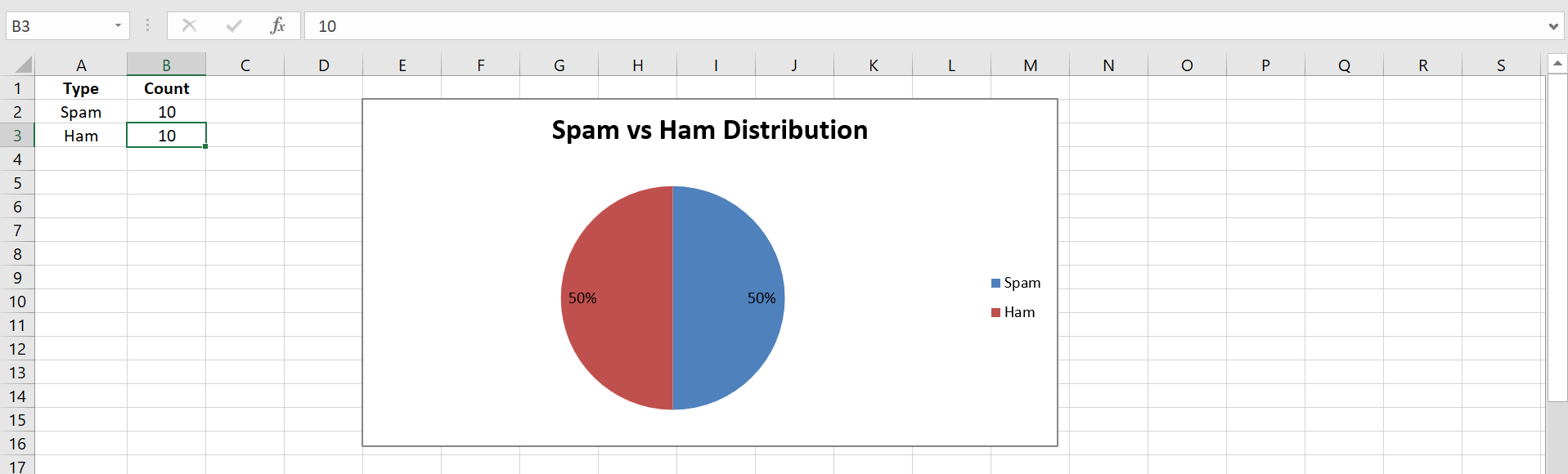
### **EXCEL BASED RESULTS:**

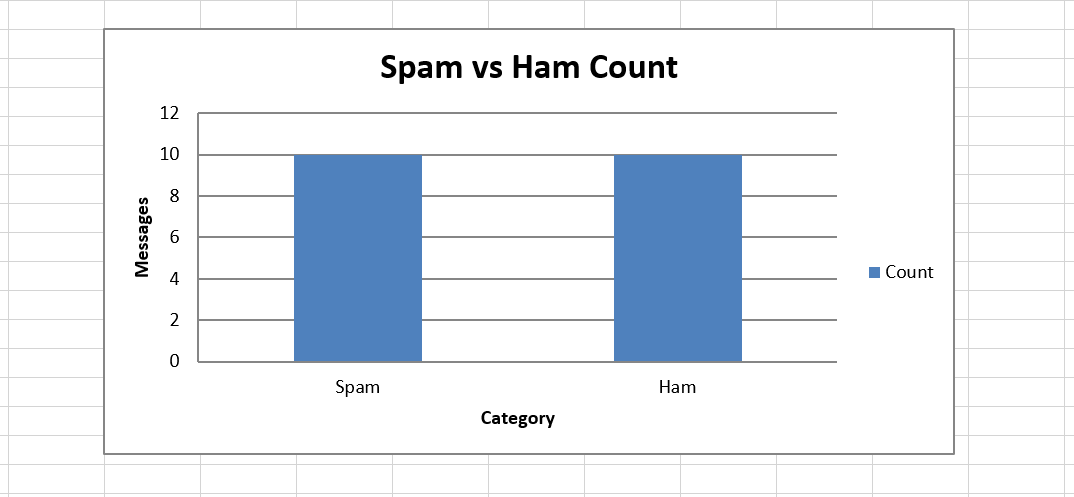






### **EXCEL BASED GRAPHS:**





# CONCLUSION:

The project successfully implemented a spam detection system using the Naive Bayes theorem. The model showed good accuracy and robustness in identifying spam messages. Using GUI features enhanced user experience. For future improvement, more advanced models like ensemble techniques or deep learning can be considered to increase accuracy further. Larger and more balanced datasets may also enhance performance.

# References:

Jurassky, D. &. M. J., 2021. *Speech and Language Processing.* s.l.Pearson.

Manning, C. D. R. P. &. S. H., 2008. *Introduction to Information Retrieval.* s.l.Cambridge University Press.

Rose, S., n.d. *Inroductory Statistics.* s.l.