

CATEGORIZED CLASSIFICATION OF TEXTUAL SPAM EMAILS USING DATA MINING

Ms. Sobiavani Gaushikan
169335 H



Dissertation submitted to the Faculty of Information Technology, University of Moratuwa,
Sri Lanka, for the partial fulfillment of the requirements of Degree of Master of Science in
Information Technology

December 2020

Declaration

We declare that this thesis is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text, and a list of references is given.

Ms. Sobiavani Gaushikan

.....

Date:

Supervised by:

Mr. Saminda Premaratne
Senior Lecturer,
Faculty of Information Technology,
University of Moratuwa

.....

Date:

Acknowledgment

First and foremost, glory and gratitude to Lord, the Merciful, for His showers of blessings during my research work to complete this research effectively.

I would like to express my profound and heartfelt thanks to my research supervisor, Mr.S.C. Premaratne, Department of Information Technology, Faculty of Information Technology, University of Moratuwa, to give me the chance to do research and offer support and help during this research. His solidity, optimism, honesty and inspiration have motivated me greatly. He also taught me the techniques to carry out the study and to outlines the research work as briefly as possible. It was a great pleasure and honor for him to work and learn under his supervision. I am deeply thankful for what he has given me to do.

I am most gratitude to my beloved parents for their encouragement, prayers, kindness, and contributions for better preparing me for my career. I am so grateful to my husband for his affection, compassion, prayer, and continuous cooperation in order to complete this research work. I would also like to thank my friends and associates for their help in completing this study in several forms.

I extend my appreciation to the Head and staff of the Center for Information and Communication Technology, Eastern University, for their assistance in my research work.

Finally, my thanks go to all the persons who contributed me in finishing my research study, either knowingly or unknowingly.

Abstract

In the last few years, to boost the connectivity and the safety of the people in the world, the Internet has built various platforms. Across from them, Email is a substantial platform for people's interaction. Email is an automated and efficient message transmission structure used to convey information from one person to another. And also, this appliance preserves plenty of money and time. Apart from that, Emails also have been abused by some assailants. The most commonly known one is Email Spam.

Spam Emails are referred to as unwanted materials, and they are unrequested business Emails or forged Emails forwarded to specific personnel or sent to an Organization or spread among a set of people. For spam Email, users face many difficulties, such as increment of traffic, restricting the data processing duration and the storage area, consuming more time of users, and threatening user protection. Therefore, it is essential to have an appropriate Email filtering approach to secure Email. There has been plenty of general spam Email filtering systems and various research people's various efforts to classify Emails into ham (genuine Emails) or spam (fake Emails) using Machine Learning techniques.

Unfortunately, most of the spam Email filtering solutions proposed so far are focused only on binary classification. However, classifying the already detected spam Emails into different types of categories is not performed yet.

This research explored and proposed an effective system to categorize the Textual Spam Emails into different categories using Data mining. First, The Multi-Nominal Naïve Bayes classifier is applied to distinguish Emails into two groups, such as ham and spam. The Grouping algorithm is used to categorize the spam Emails, which are already obtained from the Naïve Bayes classifier, into distinct categories. (Finance, Health, Marketing, etc...)

Finally, the proposed System's performance was evaluated with the Model Evaluation Techniques: Confusion matrix, Accuracy, Precision, Recall, F1 score.

Table of Contents

Declaration.....	i
Acknowledgment.....	ii
Abstract.....	iii
Table of Contents.....	iv
Abbreviations.....	vii
List of Figures.....	viii
List of Tables.....	x
Chapter 1.....	1
1 Introduction.....	1
1.1 Prolegomena.....	1
1.2 Background.....	1
1.3 Problem Statement.....	3
1.4 Aim and Objectives.....	4
1.4.1 Aim.....	4
1.4.2 Objectives.....	4
1.5 Proposed solution.....	4
1.6 Overview of the Report.....	4
1.7 Summary.....	5
Chapter 2.....	6
2 Literature Review.....	6
2.1 Introduction.....	6
2.2 Strategies to send Spam Emails.....	6
2.3 Approaches to control Spam Emails.....	8
2.4 Machine Learning Techniques to filter spam Emails.....	9
2.5 Summary.....	12
Chapter 3.....	13
3 Technology adapted on Spam Email classification.....	13
3.1 Introduction.....	13
3.2 Data mining.....	13
3.3 Machine Learning in Spam Email classification.....	15
3.4 Involvement of Data mining on Spam Email classification.....	15
3.5 Naïve Bayes algorithm.....	15
3.5.1 Multinomial Naïve Bayes classifier.....	17
3.6 TF-IDF vectorizer.....	17

3.7	Summary	18
Chapter 4	19
4	A novel approach for Spam Emails filtering	19
4.1	Introduction.....	19
4.2	Hypothesis	19
4.3	Input	19
4.4	Process	19
4.5	Output	19
4.6	Summary	20
Chapter 5	21
5	Analysis and Design for the Proposed System	21
5.1	Introduction.....	21
5.2	Framework for Proposed System.....	21
5.3	Data collection	22
5.4	Data Pre-processing	23
5.4.1	Data cleaning	23
5.4.2	Normalization	23
5.5	Feature Extraction.....	23
5.6	Classification.....	24
5.7	Evaluation	24
5.8	Summary	24
Chapter 6	25
6	Implementation of the Proposed System	25
6.1	Introduction.....	25
6.2	Python	25
6.3	PyCharm	25
6.4	NLTK.....	26
6.5	Data collection	26
6.6	Pre-processing.....	27
6.6.1	Importing the libraries	27
6.6.2	Importing the Dataset	28
6.6.3	Cleaning the Data	29
6.6.4	Normalizing the Data	30
6.7	Feature Extraction.....	30
6.8	Split the Dataset into training & testing sets.....	31
6.9	Create and Train the Multinomial Naïve Bayes classifier	31
6.10	Test the Data	31

6.11	Predict the spam Email	32
6.12	Categorize the spam Emails	32
6.13	Summary	33
Chapter 7	34
7	Evaluation	34
7.1	Introduction.....	34
7.2	Evaluation Techniques used for the classification models	34
7.2.1	Accuracy.....	34
7.2.2	Precision	34
7.2.3	Recall.....	35
7.2.4	Confusion matrix	35
7.2.5	F1 score	35
7.3	Evaluate the System on the train Dataset.....	36
7.4	Evaluate the System on the test Dataset.....	37
7.5	Experimental evaluation on spam Email classification and Results	39
7.6	Evaluation on spam Email categorization.....	39
7.7	Experimental evaluation on spam Email categorization and Results	41
7.8	Summary	41
Chapter 8	42
8	Conclusion and Further Work	42
8.1	Introduction.....	42
8.2	Overview of the research	42
8.3	Limitations	42
8.4	Further developments.....	43
8.5	Summary	43
Chapter 9	44
References	44
Appendix A	46
Appendix B	47
Appendix C	49
Appendix D	50

Abbreviations

IP	Internet Protocol
ML	Machine Learning
KDD	Knowledge Discovery in Databases
NLP	Natural Language Processing
TF-IDF	Term Frequency Inverse Document Frequency
IDE	Integrated Development Environment
NLTK	Natural Language Toolkit
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

List of Figures

Figure 1.1: Basic Email structure.....	1
Figure 1.2: Ham Email.....	2
Figure 1.3: Spam Email	2
Figure 1.4: Spam Email filtration	3
Figure 2.1: A taxonomy of Email spam filtering techniques.....	8
Figure 3.1: KDD Process	14
Figure 3.2: Data mining models and tasks.....	14
Figure 3.3: Naïve Bayes algorithm	16
Figure 3.4: Bayes Theorem.....	17
Figure 3.5: TF-IDF Algorithm.....	18
Figure 5.1: System Model for Spam Email classification	21
Figure 5.2: Kaggle Repository Website.....	22
Figure 5.3: HubSpot website.....	22
Figure 6.1: PyCharm IDE	26
Figure 6.2: Email Dataset	27
Figure 6.3: Importing Dataset	28
Figure 6.4: Visualization of Dataset	28
Figure 6.5: Ham Email Word Cloud.....	29
Figure 6.6: Spam Email Word Cloud	29
Figure 6.7: Normalization.....	30
Figure 6.8: Feature extraction	30
Figure 6.9: Data after Feature Extraction	31
Figure 6.10: Train Multinomial Naive Bayes classifier.....	31
Figure 6.11: Test the data.....	31
Figure 6.12: Predict the spam Email.....	32
Figure 6.13: Google spreadsheet with Different types of spam Email keywords	32
Figure 7.1: Accuracy equation	34
Figure 7.2: Precision equation	34
Figure 7.3: Recall equation	35
Figure 7.4: Confusion matrix	35
Figure 7.5: F1 score equation.....	35

Figure 7.6: Python code for evaluation on training Dataset	36
Figure 7.7: Output for the evaluation on training Dataset	36
Figure 7.8: Graph of evaluation on training Dataset.....	37
Figure 7.9: Python code for evaluation on testing Dataset	37
Figure 7.10: Output for the evaluation on testing Dataset	38
Figure 7.11: Graph of evaluation on testing Dataset	38
Figure 7.12: Calculate the Accuracy of the spam Email categorization.....	39
Figure 7.13: Output of the evaluation for spam Email categorization.....	39
Figure 7.14: Graph for spam Email category prediction Accuracy	40
Figure 7.15: Comparison between the human expert and the system in spam Email categorization.	41

List of Tables

Table 2.1: Strategies used by spammers to send spam Emails	7
Table 2.2: Summary of ML Approaches for Spam Email Filtering	12
Table 6.1: Python Libraries.....	27
Table 7.1: Spam Email classifier Accuracy on Data size	39
Table 7.2: Prediction accuracy for different no of keywords	40