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 Information derived from the published or unpublished work of others has been acknowledged in the text, and a list of references is given.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

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
NTLK	Natural Language Toolkit
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

List of Figures

List of Tables