

E-mail Classification System for Bank Internal mail System

K.T Atapattu

139155D



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

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

April 2016

Declaration


I hereby certify that this project and the all the artefacts associated with it is my own work and it has not been submitted before nor is currently being submitted for any other degree programme.

Full name of the student: Kasun Tharanga Atapattu

Student Number : 139155D

Signature of the student:

Date:

 **Name of the supervisor:** Mrt. Saminda Premaratne
University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Signature of the supervisor:

Date:

Dedication

This Dissertation is dedicated to my loving parents for being part of me and encouraging me always being by my side.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Acknowledgement

Apart from the efforts of me, the success of any project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this Research.

Special mention goes to my supervisor, Mr. Saminda Premarathne not only for his tremendous academic support, but also for giving me opportunity to successfully complete the research.

My next thank goes to Prof. Asoka S Karunanda who taught us Research Methodology and Literature Review and Thesis Writing subjects which were more benefited us throughout the research work

It is my great pleasure to thank all the other Senior lecturers, Lecturers, Instructors, and staff members who helped us in many ways to make this Research. The guidance and support received from all the members who contributed and who are contributing to this project, was vital for the success of the Research. I am grateful for their constant support and help.



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Then I would like to thank my all the batch mate of the M.Sc Information Technology batch 7 in faculty of Information Technology for their various help and support. All so other friend of the faculty as well as friends for outside who gave me supporting and encouraging us with their best wishes.

Abstract

My project is an implementation of 'Naive Based Algorithm' to classify E-mails. The main idea of the project is to implement algorithm in a effective manner to classify bank E-mails in bank internal authenticated mail system. E- mail has the potential to improve efficiency and reduce costs involved in communication. Even after the advent of newer technologies such as instant messaging and VoIP, email remains the number one application for business communication.

With the increasing of information on the internet based communication, our bank internal authentication mail system needs an efficient tool to classify the E-mails into categories. In this way, we can easily classify E-mail from large amount of E-mails available. Automated text categorization is a process that assigning pre-defined category labels to E-mail based on the contents.

Text categorization has many applications. For example, we can classify web pages into different categories to speed up the internet search, which is very useful for some search engines like Yahoo, Google etc. Also E-mail service providers are using those classification techniques for spam filtering and E-mail classification as well.

I have trained the develop algorithm from more than thousand pre - categorized E-mails. I have tested the text categorization algorithm developed based on naive based classifier with several different size data sets. Accuracy is evaluated as well. Experiment results shows my conclusion is efficient.

Contents

DECLARATION	I
DEDICATION.....	II
ACKNOWLEDGEMENT.....	III
ABSTRACT	IV
CHAPTER 1	1
INTRODUCTION.....	1
1.1 Background and Motivation.....	1
1.2 Aims and Objectives	2
1.2.1 Aims of the research.....	2
1.2.2 Objectives of the research.....	3
1.5 Structure of the Thesis.....	3
CHAPTER 2	4
LITERATURE REVIEW	4
2.1 Summary.....	9
CHAPTER 3	10
BANK CUSTOMER E-MAIL CLASSIFICATION.	10
3.1 Introduction	10
3.2 Factor Identification for Customer E-mail Classification.	10
3.3 Data Cleaning and Pre-processing	10
3.4 Extract the set of key words for each attribute in each of the factors.....	10
3.5 Use text classification technique to analyze the data.....	11
3.6 Machine Learning	11
3.6.1 Bayesian Filter	11



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

3.7 WordNet as ontology.....	12
3.8 Summary	12
CHAPTER 4.....	13
TEXT CLASSIFICATION TECHNIQUE FOR E-MAIL CLASSIFICATION.....	13
4.1 Introduction	13
4.2 Methodology Used for System Design and Development	13
4.3 Java	13
4.4 Oracle Database	13
4.5 Wordnet.....	14
4.5.1 Database content	14
4.6 Methodology.....	14
4.6.1 Category identification for the classification.	15
4.6.2 Retrieve the related data e-mail data set for the classification.....	16
4.6.3 Data Cleaning and pre-processing.....	16
4.6.3.1 Stop words.....	16
4.6.3.2 Pre - Processing.....	17
4.6.4 Develop classification using Bayes' Theorem.....	17
4.6.4.1 Bayes' Theorem.....	17
4.6.4.2 Train E-mail classification algorithm using Bayes' Theorem.....	18
4.6.4.3 Get probability of new e-mail from train algorithm.	20
4.7 Summary	21
CHAPTER 5.....	22
SYSTEMS DESIGN	22
5.1 Introduction	22
5.2 Architectural Design of the system.....	22
5.3 Pre - processing layer	23
5.3.1 Architectural Design of the pre - processing layer	24
5.4 Classification layer	25
5.4.1 Architectural Design of the classification layer	26

5.5 Summary	26
CHAPTER 6	27
IMPLEMENTATION	27
6.1 Introduction	27
6.2. Implementation of Pre - Processing Layer	27
6.2.1 Component for read the e-mails retrieved from bank authentication system to Microsoft excel document.	27
6.2.2 Component for remove stop words.	28
6.2.3 component for data pre - processing	29
6.2.4 Component for enter new e-mail to get the category and change the category if required.	30
6.3 Implementation of Classification Layer	32
6.3.1 Use WordNet for avoid repeating words with similar meaning.	32
6.3.2 Train the algorithm with pre - processed data and Calculate positive probability using Bayes' Theorem.	33
6.3.3 Classify new e-mail using trained algorithm and self train the algorithm from provided e-mail.....	36
6.4 Summary	37
CHAPTER 7	38
EVALUATION	38
7.1 Introduction	38
7.2 Evaluation	38
7.2.1 Component testing	38
7.2.1.1 Evaluate 'Read E-mails' Component	38
7.2.1.2 Evaluate ' Remove stop words' Component	39
7.2.1.3 Evaluate ' Data pre - processing' Component.....	39
7.2.1.4 Evaluate ' WordNet' Component	39
7.2.1.5 Evaluate ' Calculate Positive Probability' Component	39
7.2.1.6 Evaluate ' Classify new e-mail using trained algorithm' Component	40
7.2.1.7 Evaluate Self train the classification algorithm Component.....	40
7.2.2 Integration Testing	40
7.2.2.1 Implementation of Integration testing	40
7.3 Summary	41



CHAPTER 8	42
CONCLUSION	42
8.1 Introduction	42
8.2 Analysis of Performance and Accuracy.	42
8.2.1 Algorithm training performance.....	42
8.3 Algorithm drawbacks	43
8.4 Limitations	43
8.5 Future Work	43
8.6 Summary	44
REFERENCES	45



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk