



**Tribhuvan University**  
**Institute of Science and Technology**

# **A Comparative Study of Naive Bayesian Spam Filtering Using Word Distribution and Trigrams**

**Dissertation**

Submitted to

Central Department of Computer Science & Information Technology  
Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements  
for the Master's Degree in Computer Science and Information Technology

by  
**Pabitra Dangol**

**December, 2011**



**Tribhuvan University**  
**Institute of Science and Technology**

# **A Comparative Study of Naive Bayesian Spam Filtering Using Word Distribution and Trigrams**

**Dissertation**

Submitted to

Central Department of Computer Science & Information Technology  
Kirtipur, Kathmandu, Nepal

In partial fulfillment of the requirements  
for the Master's Degree in Computer Science and Information Technology

by

**Pabitra Dangol**  
**December, 2011**

Supervisor

**Assoc. Prof. Dr. Subarna Shakya**

**Tribhuvan University**  
**Institute of Science and Technology**  
**Central Department of Computer Science & Information**  
**Technology**

**Student's Declaration**

I hereby declare that I am the only author of this work and that no sources other than the listed here have been used in this work.

.....  
**Pabitra Dangol**  
Date: Dec, 2011

**Supervisor's Recommendation**

I hereby recommend that this dissertation prepared under my supervision by **Mr. Pabitra Dangol** entitled “**A Comparative Study of Naive Bayesian Spam Filtering Using Word Distribution and Trigrams**” in partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Information Technology be processed for the evaluation.

.....  
**Assoc. Prof. Dr. Subarna Shakya**  
Department of Electronics and Computer Engineering,  
Institute of Engineering,  
Pulchowk, Nepal

**Date:** ... ..

**Tribhuvan University**  
**Institute of Science and Technology**  
**Central Department of Computer Science & Information**  
**Technology**

**LETTER OF APPROVAL**

We certify that we have read this dissertation and in our opinion it is satisfactory in the scope and quality as a dissertation in the partial fulfillment for the requirement of Masters Degree in Computer Science and Information Technology.

**Evaluation Committee**

.....  
**Assoc. Prof. Dr. Tanka Nath Dhamala**  
Central Department of Computer Science  
and Information Technology,  
Tribhuvan University, Nepal  
**(Head)**

.....  
**Assoc. Prof. Dr. Subarna Shakya**  
Department of Electronics and Computer  
Engineering,  
Institute of Engineering,  
Pulchowk, Nepal  
**(Supervisor)**

.....  
(External Examiner)

.....  
(Internal Examiner)

Date: .....

**Tribhuvan University**  
**Institute of Science and Technology**  
**Central Department of Computer Science & Information**  
**Technology**

**LETTER OF APPROVAL**

We certify that we have read this dissertation and in our opinion it is satisfactory in the scope and quality as a dissertation in the partial fulfillment for the requirement of Masters Degree in Computer Science and Information Technology.

**Evaluation Committee**

.....  
**Assoc. Prof. Dr. Tanka Nath Dhamala**  
Central Department of Computer Science  
and Information Technology,  
Tribhuvan University, Nepal  
**(Head)**

.....  
**Assoc. Prof. Dr. Subarna Shakya**  
Department of Electronics and Computer  
Engineering,  
Institute of Engineering,  
Pulchowk, Nepal  
**(Supervisor)**

.....  
**Mr. Sarbin Sayami**  
**(External Examiner)**

.....  
**Mr. Bishnu Gautam**  
**(Internal Examiner)**

Date: .....

## **Abstract**

A comparative study of Naive Bayesian spam filter is done on the basis of tokenization. The study is focused on the reliability and accuracy of the spam filter between word-based tokenization and trigram-based tokenization. Both of the filters are implemented using the same classifier and trainer. The results of the study is that word-based spam filtering is better when the amount of pre-categorized emails available for training are limited and when the resources available for the classification process were limited as well. For sufficient amount of resources and emails, the results suggest that trigram-based spam filtering is better due to its higher reliability and accuracy.

## **Acknowledgement**

Many people have contributed to complete this dissertation work. I want to acknowledge and thank to all of them for their great effort and help for this thesis. I am especially grateful and want to express heartfelt regards to my respected teacher and dissertation supervisor Assistant Dean Prof. Dr. Subarna Shakya, Institute of Engineering for his boundless and instructive help, encouragement and support for confidently allowing me to perform this work.

I would like to thank my respected teacher Mr. Dinesh Bajracharya for providing me invaluable knowledge during this work. I must thank to the friends who provided me advices and support during this dissertation work too.

Thanks to all authors and researchers for their publications like research papers, journals and books which I have used to broaden my knowledge in order to perform my research work.

My gratitude is also to my family members for their encouragement, support and providing study environment for this thesis work. Finally I want to dedicate this dissertation work to all of the known and unknown people who had helped, supported and encouraged me.

## Table of Contents

Abstract .....	vi
Acknowledgement .....	vii
Table of Contents .....	viii
List of Figures .....	xi
List of Tables .....	xii
Notations .....	xiii
Chapter 1 INTRODUCTION .....	<b>Error! Bookmark not defined.</b>
1.1 Problem of Spam .....	<b>Error! Bookmark not defined.</b>
1.2 Why Naive Bayesian ? .....	<b>Error! Bookmark not defined.</b>
1.3 Research Objectives .....	<b>Error! Bookmark not defined.</b>
Chapter 2 LITERATURE REVIEW .....	<b>Error! Bookmark not defined.</b>
Chapter 3 TECHNIQUES TO ELIMINATE SPAM ..	<b>Error! Bookmark not defined.</b>
3.1 Hiding the e-mail Address .....	<b>Error! Bookmark not defined.</b>
3.2 Pattern Matching, Whitelists and Blacklists .....	<b>Error! Bookmark not defined.</b>
3.3 Rule Based Filters .....	<b>Error! Bookmark not defined.</b>
3.4 Statistical Filters .....	<b>Error! Bookmark not defined.</b>
3.5 E-mail Verification.....	<b>Error! Bookmark not defined.</b>
3.6 Distributed Blacklists of Spam Sources .....	<b>Error! Bookmark not defined.</b>
3.7 Distributed Blacklist of Spam Signatures .....	<b>Error! Bookmark not defined.</b>
3.8 Money e-mail Stamps.....	<b>Error! Bookmark not defined.</b>
3.9 Proof-of-work e-mail Stamps.....	<b>Error! Bookmark not defined.</b>
3.10 Legal Measures .....	<b>Error! Bookmark not defined.</b>
Chapter 4 STATISTICAL CLASSIFIERS .....	<b>Error! Bookmark not defined.</b>
4.1 Features and Classes.....	<b>Error! Bookmark not defined.</b>



4.2 Text Categorization .....	<b>Error! Bookmark not defined.</b>
4.3 Basics about Probability Theory .....	<b>Error! Bookmark not defined.</b>
4.4 Bayes Theorem.....	<b>Error! Bookmark not defined.</b>
4.5 Classical vs. Bayesian Statistics.....	<b>Error! Bookmark not defined.</b>
4.5.1 Using Statistics.....	<b>Error! Bookmark not defined.</b>
4.5.2 Objective and Subjective Probabilities .....	<b>Error! Bookmark not defined.</b>
4.5.3 Inference Differences.....	<b>Error! Bookmark not defined.</b>
4.5.4 Example of Statistical Spam Classification .....	<b>Error! Bookmark not defined.</b>
4.5.4.1 Classical Statistics.....	<b>Error! Bookmark not defined.</b>
4.5.4.2 Bayesian Statistics .....	<b>Error! Bookmark not defined.</b>
Chapter 5    NAIVE BAYESIAN SPAM FILTERING ..	<b>Error! Bookmark not defined.</b>
5.1 The model.....	<b>Error! Bookmark not defined.</b>
5.2 Naive Bayesian Classifier .....	<b>Error! Bookmark not defined.</b>
5.3 $t^2$ Statistics .....	<b>Error! Bookmark not defined.</b>
Chapter 6    IMPLEMENTATION.....	<b>Error! Bookmark not defined.</b>
6.1 Tokenization.....	<b>Error! Bookmark not defined.</b>
6.1.1 Word-based Tokenization.....	<b>Error! Bookmark not defined.</b>
6.1.2 Trigram-based Tokenization.....	<b>Error! Bookmark not defined.</b>
6.2 Datasets .....	<b>Error! Bookmark not defined.</b>
6.3 Training and Classification .....	<b>Error! Bookmark not defined.</b>
Chapter 7    EVALUATION TECHENIQUES .....	<b>Error! Bookmark not defined.</b>
7.1 Precision and Recall.....	<b>Error! Bookmark not defined.</b>
Chapter 8    RESULTS AND ANALYSIS.....	<b>Error! Bookmark not defined.</b>
Chapter 9    CONCLUSION AND FUTURE WORK ....	<b>Error! Bookmark not defined.</b>
9.1 Conclusion.....	<b>Error! Bookmark not defined.</b>
9.2 Future Work .....	<b>Error! Bookmark not defined.</b>
References.....	<b>Error! Bookmark not defined.</b>
Appendix A.....	<b>Error! Bookmark not defined.</b>
Appendix B .....	<b>Error! Bookmark not defined.</b>

Appendix C ..... **Error! Bookmark not defined.**

Appendix D ..... **Error! Bookmark not defined.**

Appendix E ..... **Error! Bookmark not defined.**

Appendix F ..... **Error! Bookmark not defined.**

## List of Figures

1.		Figur
	e 5.1: A model of Naïve Bayesian Spam Filtering .....	20
2.		Figur
	e 8.1: Recall and precision rates of word-based spam filter .....	31
3.		Figur
	e 8.2: Recall and precision rates of trigram-based spam filter .....	32

## List of Tables

1.		Table
	1.2: Classification result based on data size .....	2
2.		Table
	2.1: Classification results using various feature sets (Sahami, et al., 1998)	5
3.		Table
	6.2.1: The amount of emails used in training and testing the classification	26
4.		Table
	6.2.2: The distribution of spam to ham in the training corpus during result analysis .....	27
5.		Table
	8.1: Summary statistics for the recall and precision rates .....	32

## Notations

$\chi^2$

Chi square statistics

$n_{spam \rightarrow ham}$

Number of spam messages classified as good

$n_{spam \rightarrow spam}$

Number of spam messages classified as spam

$N_{ham \rightarrow ham}$

Number of good messages classified as good

$N_{ham \rightarrow spam}$

Number of good messages classified as spam

