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

List of Figures

List of Tables

Notations

χ^2	Chi square statistics
$n_{spam->ham}$	Number of spam messages classified as good
$n_{spam->spam}$	Number of spam messages classified as spam
$N_{ham->ham}$	Number of good messages classified as good
$N_{ham->spam}$	Number of good messages classified as spam