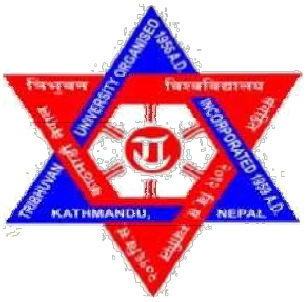


Tribhuvan University
Institute of Science and Technology

A Comparative Study of Naive Bayes and Support Vector Machine Classifier for Nepali News Classification

Dissertation
Submitted To

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

In partial fulfillment of the requirements
for the Degree of Master of Science in Computer Science & Information Technology

Submitted By
Ganesh Bahadur Ayer
Roll No. : 18 (2010-2012)
Date: 28 Dec, 2015

Supervisor
Mr. Nawaraj Paudel

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Date: 28 Dec, 2015

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.

… … … … … … …

Ganesh Bahadur Ayer
M.Sc. CSIT (2010-2012)
Central Department of Computer Science
And Information Technology,
Institute of Science and Technology,
Kirtipur, Kathmandu, Nepal

Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

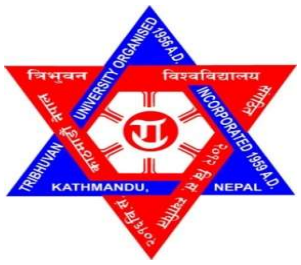
Date: 28 Dec, 2015

Supervisor's Recommendation

I hereby recommend that the dissertation prepared under my supervision by **Mr. Ganesh Bahadur Ayer** entitled "**A Comparative Study of Naive Bayes and Support Vector Machine Classifier for Nepali News Classification**" be accepted as in fulfilling partial requirement for the completion of Master's Degree of Science in Computer Science & Information Technology. In my best knowledge this is an original work in computer science.

Mr. Nawaraj Paudel

Head,
Central Department of Computer Science and Information Technology,
Institute of Science and Technology,
Kirtipur, Kathmandu, Nepal



Tribhuvan University

Institute of Science and Technology

Central Department of Computer Science and Information Technology

Date: 28 Dec 2015

LETTER OF APPROVAL

We certify that we have read this dissertation work and in our opinion it is appreciable for the scope and quality as a dissertation in the partial fulfillment of the requirements of Master's Degree of Science in Computer Science & Information Technology.

Evaluation Committee

Mr. Nawaraj Paudel
Head of Department
Central Department of Computer Science
& Information Technology
Tribhuvan University
Kirtipur

(External Examiner)

Mr. Nawaraj Paudel
Head of Department
Central Department of Computer Science
and Information Technology (T.U)
(Supervisor)

(Internal Examiner)

ACKNOWLEDGEMENTS

Above all, I thank God for his blessing and submit my graduate to my almighty for providing strength and confidence in me to complete this work.

Secondly, I would like to extent my, gratitude and sincerest thanks to my respected supervisor **Mr. Nawaraj Poudel** Head of Computer Science & IT Department Kirtipur, TU (Kathmandu, Nepal) for his impressive tutelage, constructive criticism and intellectual support best owed for me sacrificing his invaluable time. His crucial role to make this report culminate is indescribable.

I would like to express my gratitude to respected teachers Prof. Dr. Shashidhar Ram Joshi, Prof. Dr. Subarna Shakya, Prof. Sudarshan, Karanjeet, Mr. Min Bahadur Khati, Mr.Bishnu Gautam, Mr. Jagdish Bhatt, Mr. Bikash Balami, Mr. Sarbin Syami, Mr. Dhiraj Pandey, Mr. Arjun Singh Saud and others staffs of CDCSIT for granting me broad knowledge and inspirations within the time of period of two years.

I wish to thank to all my colleagues and friends especially **Mr.Tej Bahadur Sahi**,Mr.Ashok Kumar Pant ,Mr.Niranjan Kathayat,Mr.Dipak Prasahd Bhatt and Mr.Rajendra Prashad Joshi for supporting me directly and indirectly in this research work.

From the beginning to end in all count, I would like to thank my family members for their love, support and encouragement.

As we know that, there won't be 100% accuracy and efficiency in any work done by both machine any human, so there may be some errors in my project. But, I have done my best to complete this dissertation, so any suggestion regarding the mistakes of this work will be always welcomed.

Ganesh Bahadur Ayer

ABSTRACT

Automated document classification is the task of assigning the given document into some class of interest. Text classification is the subset of document classification as document can be text, image, music, etc. Document classification has many applications in library science, information science, computer science and others. It can be used for intellectual categorization of documents, indexing of documents, filtering of spams, routing of emails, identification of language, classification of genre, etc.

The problem of automated document classification can be solved in supervised, unsupervised or semi-supervised way. Most of the learning and classification algorithms use document attributes and human inference to learn and classify given documents. In this dissertation work, many Natural Language Processing (NLP) techniques are used for document processing and attribute selection. And, two learning based classification techniques are used namely, Support vector machine (SVM) and Naive Bayes Classifier.

For the evaluation of the system, we have created Nepali text datasets for five classes of documents: Business, Crime, Education, Health and Sports. There are two separate datasets for training and testing of the system. SVM classification system has the average system accuracy rate of 86.34%, precision rate of 84% recall rate of 94.4%. Similarly, Naive Bayes classification system has the average system accuracy rate of 88.8%, precision rate of 92.23% and recall rate of 88.87%.

Keywords:

Automated Document Categorization, Text Classification, Natural language processing, Nepali language, Preprocessing, Feature extraction, Artificial Neural Networks, Support vector machine, Naive Bayes Classifier

Table of Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

AI Artificial Intelligence

ANN Artificial Neural Network

API Application Programming Interface

BPN Back Propagation Network

CDT C/C++ Development Components

IDE Integrated Development Environment

KNN K-Nearest Neighbor

MT Machine Translation

MAP Maximum A Posteriori

NB Naive Bayes

NN Neural Network

NLP Natural Language Processing

PCA Principal Component Analysis

SVM Support Vector Machine

SDK Software Development Kit

TC Text Classification

TF-IDF Term Frequency - Inverse Document Frequency