Analyzing the Performance of Different Text Classification Algorithms for “Dhivehi” Documents

Fathimath Rashma Mohamed
MS19806914

A THESIS

SUBMITTED TO

SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION SYSTEMS

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Dr. Prasanna S. Haddela

Approved for MSc. Research Project:

MSc. Programme Co-ordinator, SLIIT

_____

Approved for MSc:

_____

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: *Fathimathrashmamohamed*

Date: 12 November 2024

ABSTRACT

Analyzing the Performance of different Text Classification Algorithms for “Dhivehi” Documents

Fathimath Rashma Mohamed
MSc. in Information System
Supervisor: Dr. Prasanna S. Haddela
December 2024

This research investigates the effectiveness of various machine learning classification algorithms applied to Dhivehi text-based documents. Dhivehi, the official language of the Maldives, presents unique linguistic challenges for text classification due to its limited digital resources and distinct grammatical structure. The study aims to identify the most suitable algorithm for classifying Dhivehi documents and to provide insights into optimizing text classification approaches for less-resourced languages.

The research systematically evaluates the performance of several machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes, Decision Trees, XGboost , Random Forest and Neural Networks. These algorithms are applied to a diverse dataset of Dhivehi text, encompassing various genres and topics. The study employs a rigorous methodology involving data preprocessing, feature extraction, and model training and testing.

Performance metrics such as accuracy, precision, recall, and F1-score are used to compare the efficacy of each algorithm. Additionally, the research explores the impact of different text representation techniques, including bag-of-words, TF-IDF, and word embeddings, on classification performance.

The findings offer valuable insights into optimizing text classification methods for low-resource languages and aim to advance natural language processing tools specifically tailored for “Dhivehi.”

The evaluation highlights that K-Neighbors achieved the highest performance, with an accuracy of 64.7% and F1 scores (macro: 0.640, weighted: 0.642), demonstrating a strong balance between precision and recall. Support Vector Machines (accuracy: 63.9%) and XGBoost (accuracy: 62.8%) also showed competitive results, with SVM slightly outperforming XGBoost in F1 metrics. Decision Tree exhibited the lowest performance across all metrics. By identifying the most effective classification algorithms and representation techniques, this research aims to enhance the accuracy and efficiency of Dhivehi text classification tasks. The results will have practical applications in areas such as sentiment analysis, document categorization, and information retrieval systems tailored for the Dhivehi language.

Furthermore, the dataset is publicly available on Mendeley data under the name “Dhivehi Categories data set” to foster future research and innovation in this domain.

Keywords— Dhivehi Language, Dhivehi Text Classification, Dhivehi dataset, Low-resource languages, Asian languages.

ACKNOWLEDGEMENT

During my time at the Sri Lanka Institute of Information Technology, I was fortunate to work under the guidance of Dr. Prasanna S. Haddela, an exceptional mentor. Despite the diverse perspectives among my advisors, Dr. Haddela stood out for his multifaceted support. He consistently provided insightful advice, offered constructive criticism on my ideas and writing, and generously shared his professional network. Additionally, and granted me the autonomy to pursue my own projects while working under his research account. Dr. Haddela's mentorship was instrumental in shaping my research experience and academic growth.

TABLE OF CONTENTS

DECLARATION ii
ABSTRACT iii
ACKNOWLEDGEMENT v
TABLE OF CONTENTS vi
List of Figures viii
List of Tables ix
Chapter 1 Introduction 1
1.1 Problem Statement 2
1.1.1 .Research Gap 3
1.2 Aim 4
1.3 Objectives 4
1.3.1 Research Question 4
1.3.2 Hypothesis 4
Chapter 2 Related Work 7
2.1 dhivehi_nlp library 8
2.1.1 Expected research Contribution 9
2.2 Document classification life cycle 9
2.3 IS systems and automated document classifications 10
2.4 Document classification- business and organizational impact 10
2.5 Applications of Text categorization 11
2.6 Text Categorization using Machine Learning 12
2.7 Decision tree 12
2.8 Artificial Neural Networks 12
2.9 Linear Support Vector Machine 12
2.10 k-Nearest Neighbor 13
2.11 Naïve Bays 14
2.12 XGBoost 14
2.13 Random Forest 16
Chapter 3 Methodology 17
3.1.1 Data collection and Curating the data set 17
3.1.2 Model Selection 18
3.1.3 Text Pre-processing 19
3.1.4 Feature extraction 22
3.1.5 Extracting Custom Features 24
3.1.6 SMOTE: Synthetic Minority Over-sampling Technique 25
3.1.7 Stratified K-Fold Cross Validation 26

3.1.8 Testing Environment 27
3.1.9 Evaluation 28
3.2 Implementing the classifiers 29
3.3 Training the classifiers 34
Chapter 4 Results 36
4.1 Performance of the Support Vector Machine (SVM) 36
4.1.1 Classification Report 36
4.1.2 Confusion Matrix 37
4.2 Performance of the Decision tree 38
4.2.1 Confusion matrix 39
4.3 Performance of Neural Network 40
4.3.1 Confusion Matrix 41
4.4 Performance of Naïve Bayes 42
4.4.1 Confusion matrix 43
4.5 Performance of K- Nearest neighbor 45
4.5.1 Confusion matrix 46
4.6 Performance Random Forest classifier 47
4.6.1 Confusion matrix 49
4.7 Performance of XGboost classifier 50
4.7.1 Confusion matrix 52
4.8 Overall Classifier performance 54
4.9 Per-class F1 Scores 57
4.10 Per – class precision scores 58
4.11 Per class recall scores 59
4.12 Performance Evaluation- initial vs. post-training rounds 61
Chapter 5 Conclusion 64
5.1.1 Key Findings 64
5.1.2 Implications 65
5.1.3 Future developments 66
Bibliography 69
Appendix 72
Appendix 1: ICARC Paper 72
Appendix 2: Data Article 78
Appendix 3: Important Codes 86

List of Figures

Figure 1: Document Classification life style 9
Figure 2: Diagrammatic Representation of KNN Algorithm 14
Figure 3: Confusion Matrix- SVM 37
Figure 4: Confusion Matrix - Decision tree 39
Figure 5: Confusion Matrix- Neural Network 41
Figure 6: Confusion Matrix- Naive bayes 43
Figure 7: Confusion Matrix - KNN 46
Figure 8: Confusion Matrix- Random Forest 49
Figure 9: Confusion Matrix - XGBoost 52
Figure 10: Overall classifier performance 54
Figure 11:Classifier performance Heat map 54
Figure 12: Per class F1 Score 57
Figure 13: Per Class precision Scores 58
Figure 14: Per class Recall Scores 59
Figure 15: Overall Performance – initial run 61
Figure 16: Overall Performance – post runs 61

List of Tables

Table 1: SVM classifier report 36
Table 2: Performance of Decision tree 38
Table 3: performance of Neural network 40
Table 4: Performance of Naive Bayes 42
Table 5: Performance of KNN 45
Table 6: Performance of Random Forest 47
Table 7: Performance of XGBoost 50
Table 8: Tabular representation of overall classifier performance 55