AN ENHANCED FEATURE SELECTION TECHNIQUE FOR CLASSIFICATION OF GROUP-BASED HOLY QURAN VERSES

ADELEKE ABDULLAHI OYEKUNLE

A thesis submitted in
fulfillment of the requirement for the award of the
Degree of Master of Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

January 2018

DEDICATION

This humble work is dedicated to all seekers of knowledge...

ACKNOWLEDGEMENT

All praises, thanks and adorations are due to Allaah – the Supreme Being, the uncreated Creator, the Sustainer, Cherisher, Protector, Beneficient, and most Merciful. None can be compared to Him, the First and the Last. No power nor might except from Him *subhanahu wa ta'ala*. I sincerely appreciate His mercy upon me most especially for the successful completion of this study. May the peace, mercy, and blessings of Allaah be upon the noblest of all souls, the Prophet Muhammad *sallallahu 'alayhi wa sallam*, his guided households, companions, and the generality of Muslims until the day of Accountability. Aameen

I humbly appreciate the scholarly and friendly influence of my supervisors Dr. Noor Azah Samsudin and Assoc. Prof. Aida Mustapha, whom are always available for verifications, clarifications, and suggestions throughout the study. Furthermore, I would also like to take this opportunity to thank Universiti Tun Hussein Onn Malaysia (UTHM), especially Faculty of Computer Science and Information Technology and ORICC for providing me good environment, facilities, and research grant to complete this study.

Finally, an honorable mention goes to my families and friends for their supports and encouragements when most needed.

Adeleke Oyekunle Abdullahi

Dec, 2017

ABSTRACT

This thesis is about proposing an enhanced feature selection technique for text classification applications. Text classification problem is primarily applied in document labeling. However, the major setbacks with the existing feature selection techniques are high computational runtime associated with wrapper-based FS techniques and low classification accuracy performance associated with filter-based FS techniques. Therefore, in this study, a hybrid feature selection technique is proposed. The proposed FS technique is a combination of *filter-based information gain (IG) and wrapper-based CFS algorithms*. The purpose of combining these two FS algorithms is to achieve both high classification *accuracy* performance (wrapper) at *lower computational runtime* (filter). The study also developed a group-based Quran dataset to improve on the understanding and analysis of the textual data (Quranic verses). The group-based dataset is a combination of Holy Quran translation and commentary (tafsir). The Quranic verses were selected from two chapters, Surah Al-Baqarah and Surah Al-Anaam. The verses are classified into three categories: Faith, Worship, and Etiquette. In the experiment, six feature selection algorithms were applied: *Information Gain (IG), Chi-square (CH), Pearson Correlation Coefficient (PCC), ReliefF, Correlation-based (CFS), and the proposed IG-CFS algorithms*. The textual data (Quranic verses) were preprocessed using *StringtoWordVector* with weighted *Term Frequency-Inverse Document Frequency (TF-IDF)*. Meanwhile, the classification phase has involved four algorithms: *Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Support Vector Machine (LibSVM), and Decision Trees (J48)*. The experiment results were evaluated based on two established performance metrics in text classification: *Accuracy* and *Area under Receiver Operating Characteristics (ROC) curve (AUC)*. The proposed hybrid feature selection technique has shown promising results in terms of *Accuracy* and *Area under Receiver Operating Characteristics (ROC) curve (AUC)* by achieving at a *lower computational runtime* (3.89secs) *Accuracy* of 94.5% and *AUC* of 0.944 with the group-based Quran dataset.

ABSTRAK

Dalam tesis ini, teknik pemilihan ciri untuk aplikasi pengkelasan teks telah dicadangkan. Masalah pengkelasan teks diaplikasi terutamanya dalam melabel dokumen. Walau bagaimanapun, kelemahan ketara teknik pemilihan ciri sedia ada melibatkan masa larian pengiraan tinggi dengan teknik pemilihan berasaskan tapisan. Oleh yang demikian, teknik pemilihan ciri hibrid telah dicadangkan. Teknik pemilihan ciri cadangan tersebut merupakan kombinasi *filter-based information gain (IG)* dan algoritma *wrapper-based CFS.* Tujuan kombinasi kedua-dua algoritma pemilihan ciri tersebut adalah untuk mencapai ketepatan keputusan klasifikasi pada masa larian pengiraan yang rendah. Penyelidikan ini turut membangunkan set data Quran berasaskan kumpulan untuk menambah kefahaman dan analisa data teks (ayat Quran). Set data berasaskan kumpulan tersebut adalah kombinasi terjemahan dan tafsir Quran. Ayat Quran telah dipilih daripada dua surah: Surah Al-Baqarah dan Surah Al-Anaam. Ayat-ayat telah diklasifikasi kepada tiga kategori: Tauhid, Ibadah, dan Akhlak. Enam algoritma pemilihan ciri telah diaplikasi dalam eksperimen: *Information Gain (IG), Chi-square (CH), Pearson Correlation Coefficient (PCC), ReliefF, Correlation-based (CFS),* dan algoritma cadangan iaitu *IG-CFS.* Data tekstual (ayat Quran) telah dipraproses menggunakan *StringtoWordVector* dengan *Term Frequency-Inverse Document Frequency (TF-IDF)* berpemberat. Manakala, fasa klasifikasi telah melibatkan empat algoritma: *Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Support Vector Machine (LibSVM),* dan *Decision Trees (J48).* Keputusan eksperimen dinilai berdasarkan dua matrik penilaian yang telah dibangunkan dalam pengkelasan teks: ketepatan dan luas di bawah lengkungan *Receiver Operating Characteristics (ROC).* Teknik hibrid pemilihan ciri cadangan telah menunjukkan keputusan yang memberangsangkan dari segi ketepatan dan luas di bawah lengkungan *Receiver Operating Characteristics (ROC).* Ia mencapai ketepatan 94.5% dan keluasan bernilai 0.944 dengan data Quran berasaskan kumpulan tersebut.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AI	-	Artificial Intelligence
AUC	-	Area Under (ROC) Curve
CFS	-	Correlation based Feature Selection
CH	-	Chi Square
DT	-	Decision Trees
FN	-	False Negative
FP	-	False Positive
FS	-	Feature Selection
GBFS	-	Group Based Feature Selection
IG	-	Information Gain
KDD	-	Knowledge Discovery in Databases
kNN	-	k Nearest Neighbor
ML	-	Machine Learning
NB	-	Naïve Bayes
PCC	-	Pearson Correlation Coefficient
ROC	-	Receiver Operating Characteristic
SVM	-	Support Vector Machines
TC	-	Text Classification
TF-IDF	-	Term Frequency Inverse Document Frequency
TN	-	True Negative
TP	-	True Positive

CHAPTER 1

INTRODUCTION

1.1 Overview of Study

The massive technological growth over the years has made the field of machine learning one of the mainstays of information technology (Ivanovic and Radovanovic, 2015). Machine learning as defined by Arthur Samuel is a field of study that gives computers the ability to learn without being explicitly programmed (Das *et al.*, 2015). The field of machine learning received much attention in recent years with the development of many successful machine learning applications such as data mining programs, information retrieval systems and autonomous vehicles.

In the field of Artificial Intelligence (AI), Machine Learning (ML) can be defined as the capability of an AI system to improve its performance over a time period through acquiring new knowledge and skills as well as its ability to reorganize the existing knowledge based on the newly acquired knowledge (Saloky and Seminsky, 2008). Learning is considered as a parameter for intelligent machines to be able to take decisions in a more optimized form as well as work smoothly (Talwar and Kumar, 2013). The concept of ML is based on training machines to be able to detect patterns and adapt to new circumstances.

Important task in machine learning is classification, also referred to as pattern recognition (Pundir *et* al., 2013). Pattern recognition (PR) is a field concerned with machine recognition of meaningful regularities in noisy or complex environments (Thasleema and Narayanan, 2013). Sharma and Kaur (2013) presented three tasks in a typical Pattern Recognition System (PRS): Data acquisition/preprocessing, feature extraction/selection and decision making.

Data classification is the problem of identifying to which set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known (Tang *et al.*, 2014). Text categorization (also known as text classification) is the task of automatically sorting a set of documents into categories from a predefined set (Faraz, 2015; Bhumika *et al.*, 2013; Nguyen and Shirai, 2013; Dalal and Zaveri, 2011). Important to text classification is feature selection: a dimensionality reduction method used in reducing the high level of dimensionality commonly present in text data (Tang *et al.*, 2014). The curse of dimensionality occurs as a result of large feature space comprising of relevant, irrelevant, and/or redundant features (Ladha and Deepa, 2011). To optimize the performance of the classification algorithms, techniques such as in feature selection are essentially needed in cleansing and selecting the most relevant feature subsets for the experimental work.

This research work is based on implementing text classification algorithms for automating Holy Quran verses labeling. The study focuses on improving the preprocessing phase of the experimental work which involves applying feature selection method and techniques.

1.2 Motivation and Problem Statement

The Holy Quran is the religious text for Muslims. A divine, highly comprehensive and detailed book from Almighty God, considered as an essential reference for over 1.6 billion Muslims in the world (Hilal and Srinivas, 2015; Alhawarat, 2015). The Holy text is the word of God whose meanings are rich, unlimited and of great significance to the Muslim faithful and others. Furthermore, the interesting features in the Quran such as the arrangement of words into verses, chapters grouped into *juz* make the Holy book referred to as 'Golden text' in the field of Data mining. Thus, knowledge discovery from the Holy Quran is highly important and of great benefit for both the learned and common people. For this purpose, religious and Quranic scholars have devoted much attention and efforts in producing classical works such as the Quran commentaries, science of hadith (Prophetic sayings), grammar (Nahw and Sarf), and many more. However, in recent times, with the proliferation and exploration of information technology, as well as the increasingly quest for religious, Quranic knowledge gathering and disseminating within a very short time, automating the techniques of text classification for the Holy scripture analysis in the field of data mining or more broadly machine learning becomes a necessity.

For many centuries, religious scholars, exegetes, readers of the Quran, and intellectuals have focused on the Quranic study with large scholarly works being produced. A chapter in the Holy Scripture is represented by a group of verses. Within a verse or group of verses, messages (or information) on issues such as faith (Iman), worship (Ibadah), etiquettes (Akhlak) can be found therein. These issues could otherwise be categorized as labels (or topics).

Analyzing a chapter (or chapters) of the Holy Quran for better understanding of mentioned issues leads to knowledge discovery. This requires looking into multiple related sources of data. For example, the translation and tafsir (commentary) as two sources of Holy Quran data are not individually sufficient- for the analysis purpose. In most of the existing works on text classification (Jamil *et* al., 2017; Gupta *et al.*, 2016; Zhou *et al.*, 2016; Goudjil *et* al., 2015; Hassan *et* al., 2015; Hilal and Srinivas, 2015; Al-Kabi *et* al., 2015; Prusa *et al.*, 2015), features are generated, selected from individual single sources of textual data. This approach may not be sufficiently applicable in some classification problems such as the Quranic verses labeling. Therefore, in this study, multiple related textual data are combined in analyzing the Quranic verses for the classification task. The resulting grouped data is referred to as group-based Holy Quran verses.

Furthermore, there are two standard methods for feature selection purpose: filter and wrapper FS methods (Ladha and Deepa, 2011). However, these methods have both strength and weaknesses. The filter method is simple, efficient, and less computationally expensive but loses performance accuracy due to its independence from the classification algorithm. On the other side, wrapper FS method has high performance accuracy but due to its complexity, it is computationally expensive (Tang *et al.*, 2014; Ladha and Deepa, 2011). Due to these advantages and limitations, the study proposes an enhanced feature selection technique which aims at achieving both the efficiency of the filter method as well as the accuracy of the wrapper. The proposed FS technique is based on the hybridization approach to feature selection (Aladeemy *et al.*, 2017; Wang and Liu, 2016; Xue *et al.*, 2016).

1.3 Aim and Objectives of Study

This research work aims to improve the feature selection techniques applicable for the classification of Holy Quran verses. To achieve this goal, the study will focus on the following objectives:

1. To develop a group-based dataset applicable for analysis of Holy Quran verses labeling.
2. To propose and develop an improved hybrid IG-CFS algorithm for identifying appropriate features from (1) in the classification task.
3. To apply the features from (2) in the implementation phase using standard classification algorithms namely: Naïve Bayes (NB), Support Vector Machines (LibSVM), k-Nearest Neighbors (*k*-NN), and Decision Trees (J48).
4. To evaluate and compare the performance of the proposed FS technique with the existing feature selection techniques using standard performance metrics namely: Accuracy, Precision, Recall, F-Measure, and Area Under (ROC) Curve (AUC).

1.4 Scope of Study

The scope of the study focused primarily on classifying verses in chapter two (Surah al-Baqarah) and chapter six (Surah al-Anaam) of the Holy Quran into three distinct predefined categories. These categories are the three main branches of Islam and it includes Faith (Iman), Worship (Ibadah), and Etiquettes (Akhlak). Surah al-Baqarah is a Madani chapter (revealed in Madinah) consisting of 286 verses (ayaat) and Surah al-Anaam is a Makki chapter (revealed in Makkah) consisting of 165 verses (ayaat). In addition, the study experiments six feature selection algorithms for the preprocessing phase of the experimental work. These FS algorithms are: Information Gain (IG); Chi-Square (CH); Pearson Correlation Cofficient (PCC); ReliefF; Correlation-based (CFS); and the proposed IG-CFS. Furthermore, four conventional classification algorithms which include: naïve bayes, support vector machines, k-nearest neighbor, and decision trees, were implemented for the classification task.

1.5 Significance of Study

Feature selection may influence the performance of classification algorithms (Ladha and Deepa, 2011). The proposed IG-CFS feature selection technique may be of help in real world classification problems in order to achieve high classification accuracy at lower computational runtime. Furthermore, the group-based Holy Quran data developed may be useful since features are needed to be determined from multi-related textual sources.

1.6 Organization of Thesis

The rest of the chapters are organized as follows: Chapter 2 explores the relevant background information of the study domain. Chapter 3 presents the research steps and framework of the proposed approach with elaborated explanations of design and implementation. Chapter 4 shows and discusses the experimental results as well as the evaluation process of the study. Finally, the study concludes in Chapter 5 with recommendations and direction to future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter covers the overview of important fields related to the research study found in literatures as well as the different techniques and algorithms commonly used in the study domain. The existing related works are reviewed, analyzed and summarized.

2.2 The Holy Quran

The Holy Quran is a divine scripture, a religious text of the Muslims' faithful with a population of about 1.6 billion (Houssain, 2014). The Holy scripture is a collection of the words of Almighty God conveyed to mankind by the Archangel Gabriel through Prophet Muhammad (peace be upon him). It is sacred, most authentic, and unaltered book from the Almighty God since its revelation over 14 centuries ago (Khan and Alginahi, 2013). The Holy Quran has about 78,000 words (Adeleke *et al.*, 2017) divided into 114 chapters of varying sizes with each chapter comprising of verses sum total to about 6,243 verses in the entire Quran.

2.2.1 Quranic Chapters

The Holy Quran chapters (also known as surahs) are classified according to religious scholars into Makki or Madani depending on the time of revelation with respect to the religious migration (hijrah) of the Holy Prophet Muhammad (peace be upon him) and his companions from Makkah to Medinah. In summary, the 114 chapters of the Holy Quran comprises of both Makki and Madani categories with 86 chapters in Makki and 28 chapters in Madani.

2.2.2 Quranic Translations

Due to the significance and importance of the divine scripture to the Muslims' faithful widely spread across various parts of the world, great efforts have been made over the years to translate the holy book from its original divine language- the Arabic language. Today, there numerous translations of the Quran in almost all living languages of the world. Among the most famous of these translations are: Abdullah Yusuf Ali's translation in English, Hatta in Bahasa Melayu, Muhammad Shakir in English, and Ma Jian in Chinese (Adeleke *et al.*, 2017).

2.2.3 Quranic Exegesis (Tafsir)

Over the years, with the distinct, authentic, unaltered, and pure sources of the Quranic study, most importantly, the Prophetic narrations (hadith), great, commendable efforts have been made to produce scholarly commentaries (tafsir) of the Holy Quran. Based on reliable sources of knowledge, religious experts (Imams) have attempted to interpret and analyze the words of Almighty God (Quran). From among the most known classical commentary books of the Holy Quran are: Tafsir of Imam Ismaeel ibn Kathir which was originally in Arabic but has been translated into many languages including English. Others include: the commentary by Abu Ala Mawdudi, Imam Tabari, Muhammad al-Uthaymeen, Muhammad ash-Shinqinti, and many others.

2.2.4 Contemporary Research on Holy Quran

With the proliferation of information technology, there have been extensive and continuous research being carried out in many various fields with numerous applications successfully developed and techniques proposed. The field of Holy Quran study have witnessed quite a number of research works among which are: text classification applications on the Holy Quran (Adeleke *et al.*, 2017; Jamil *et al.*, 2017; Goudjil *et al.*, 2015; Hassan *et al.*, 2015; Hilal and Srinivas, 2015; Al-Kabi *et al.*, 2015; Akour *et al.*, 2014); ontology-based applications (Ibrahim *et al.*, 2017; Alqahtani and Atwell, 2016; Hamed and Ab Aziz, 2016; Abdelnasser *et al.*, 2014; Alrehaili and

Atwell, 2014; Abdelhamid *et al.*, 2013); digitized Holy Quran applications (Akkila and Abu Naser, 2017; Ahmed and Abdo, 2017; Aljaloud *et al.*, 2016).

This research study is based on applying text classification techniques to the Quranic verses labeling. Existing related works in literatures are further reviewed in section 2.5.

2.3 Text Classification: Overview

Data classification technique has been successfully applied to many applications domain such as text classification, imaging, biometric identification, biological classification, credit scoring, and pattern recognition. However, this study is based on applying the concept and techniques of text classification to the Holy Quran verses classification.

Text classifcation (also known as text categorization) is the task of automatically sorting a set of documents into categories from a predefined set. In other words, it is the task of assigning predefined categories to natural language text (Manne *et al.*, 2014). Typically, the task is to label texts as belonging to one of a small number of classes (Kassner and Mitschang, 2016). Text classification (TC) is one of the most widely used and significant methods of supervised learning in data mining (Hassan *et al.*, 2015).

In text classification, the dimensionality of feature vector is huge, making it very difficult to classify large dimensional data. To reduce this difficulty, the feature reduction approaches (Feature extraction and selection) are applied (Guru and Parveen, 2014).

2.3.1 Text Classification Algorithms

A growing number of data mining techniques have been applied to text classification problem, including the Bayes probabilistic approach (Tang *et al.*, 2014), decision trees (Zharmagambetov and Pak, 2015), neural networks (Wang and Wang, 2014), support vector machines (SVM) (Sabbah and Selamat, 2014), and k-nearest neighbor (Towsend *et al.*, 2015). In this study, four conventional classification algorithms

(James and Dimitrijev, 2012): nearest neighbor (*k*-NN), SVM, naïve bayes, and decision trees classifiers are implemented for the labeling task of Holy Quran verses.

The *k*-NN classifier is an instance-based learning algorithm that has shown to be very simple but effective for text classification problem (Gharehchopogh *et al.*, 2015). It is a non-parametric method used in classification and works by calculating the Euclidean distance between points (Dey *et al.*, 2016). In classifying a new document x, the algorithm ranks the document's neighbors in the training set, and then uses the class of *k* most similar neighbors to predict the class of a new document (also known as majority vote). The Euclidean distance is given as:

$$d(x, x_i) = \sqrt{\sum_{i=1}^{n} (x_j - x_{ij})\ _{2}} \tag{2.1}$$

where x is the new point, x_i is the existing point across all input attributes j.

The naïve bayes classifier greatly simplify learning by assuming that features are independent given class and has proven effective in many practical applications, including text classification. The classifier is a simple probabilistic model based on the Bayes rule (Nikam, 2015). Given a class C, the probabilty of a particular document d to belong to C is given as:

$$P(C_i \quad d) = \frac{P(d \quad C_i) * P(C_i)}{P(d)} \tag{2.2}$$

SVM is one of the most widely used and applied classification methods. It has been successfully applied to many application domains. SVMs are typically used for learning classification, regression, or ranking function (Adeleke *et al.*, 2017). The algorithm works by searching a seperating hyperplane to seperate between samples with a maximal margin (Amarappa and Sathyanarayana, 2014). The seperating hyperplane is:

$$w^T x + b = 0 \tag{2.3}$$

To classify an unseen document d, the sign of $w^T x + b$ must be known (Amarappa and Sathyanarayana, 2014). This is further shown as:

$$w^T x_i + b \geq 1 \;\text{ or }\; w^T x_i + b \leq 1 \tag{2.4}$$

Decision tree is one of the most popular and powerful approaches in data mining used to extract knowledge by making decision rules from large amount of

REFERENCES

Abdelhamid, Y., Mahmoud, M., and El-Sakka, T. M. (2013). Using Ontology for Associating Web Multimedia Resources with the Holy Quran. *Taibah University International Conference on Advances in Information Technology for the Holy Quran and its Sciences,* pp. 266-271.

Abdelnasser, H., Mohamed, R., Ragab, M., Mohamed, A., Farouk, B., and El-Makky, N. (2014). Al-Bayan: An Arabic Question Answering System for the Holy Quran. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing,* pp. 57-64.

Adamatti, D. F., Silveira, J. A., and Carvalho, F. A. H. (2016). Analyzing brain signals using decision trees: an approach based on neuroscience. *Revista Eletronica Argentina-Brasil de Technologies da informacao e da Communicacao, 1(5).*

Adeleke, O. A., Samsudin, N. A., Mustapha, A., and Nawi, N. M. (2017). Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses. *International Journal on Advanced Science, Engineering, and Information Technology, 7(4),* pp. 1419-27.

Ahmed, A. H., and Abdo, S. M. (2017). Verification System of Quran Recitation Recordings. *International Journal of Computer Applications, 163(4),* pp. 6-11.

Akkila, A. N., and Abu Naser, S. S. (2017). Teaching the right letter pronunciation in reciting the holy Quran using intelligent tutoring system. *International Journal of Advanced Research and Development, 2(1),* pp. 64-68.

Akour, M., Alsmadi, I., and Alazzam, I. (2014). MQVC: Measuring Quranic verses similarity and Sura classification using N-Gram. *WSEAS Transactions on Computers, 13,* pp. 485-491.

Aladeemy, M., Tutun, S., and Khasawneh, M. T. (2017). A new hybrid approach for feature selection and support vector machine model selection based on self-adaptive cohort intelligence. *Expert Systems with Applications, 88,* pp. 118-131.

Alhawarat, M. (2015). Extracting Topics from the Holy Quran using Generative Models. *International Journal of Advanced Computer Science and Applications, 6(12),* pp. 288-294.

Aljaloud, H. O., Dahab, M., and Kamal, M. (2016). Stemmer Impact on Quranic Mobile Information Retrieval Performance. *International Journal of Advanced Computer Science and Applications, 7(12),* pp. 135-139.

Al-Kabi, M. N., Abu, A. B. M., Wahsheh, H. A., and Alsamadi, I. M. (2013). A Topical Classification of Quranic Arabic Text. *Taibah University International Conference in Advances in Information Technology for the Holy Quran and Its Sciences, 2013.*

Al-Kabi, M. N., Wahsheh, H. A., Alsmadi, I. M., and Al-Akhras, A. A. (2015). Extended Topical Classification of Hadith Arabic Text. *International Journal on Islamic Applications in Computer Science and Technology, 3(3),* pp. 13-23.

Alqahtani, M., and Atwell, E. (2016). Arabic Quranic Search Tool Based on Ontology. *21st International Conference on Applications of Natural Language to Information Systems,* pp. 478-485.

Alrehaili, S. M., and Atwell, E. (2014). Computational Ontologies for Semantic tagging of the Quran: A survey of past approaches. *Ninth International Conference on Language Resources and Evaluation.*

Amarappa, S., and Sathyanarayana, S. V. (2014). Data Classification using Support Vector Machine (SVM), a simplified approach. *International Journal of Electronics and Computer Science Engineering, 3(4),* pp. 435-445

Aphinyanaphongs, Y., Fu, L. D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C. F., and Statnikov, A. (2014). A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization. *Journal of the Association for Information Science and Technology, 65(10),* pp. 1-24.

Arras, L., Horn, F., Montavon, G., Muller, K. R., and Samek, W. (2017). "What is relevant in a text document?": An interpretable machine learning approach. *PLoS ONE, 12(8),* pp. 1-23.

Bhumika, S., Sehra, S., and Nayyar, A. (2013). A Review Paper on Algorithms used for Text Classification. *International Journal of Application or Innovation in Engineering & Management, 2(3),* pp. 90-99.

Bijalwan, V., Kumar, V., Kumari, P., and Pascual, J. (2014). KNN based Machine Learning Approach for Text and Document Mining. *International Journal of Database Theory and Application, 7(1)*, pp. 61–70.

Bleik, S., Mishra, M., Huan, J., and Song, M. (2013). Text Categorization of Biomedical Data Sets using Graph kernels and a controlled vocabulary. *IEEE Transactions on Computational Biology and Bioinformatics.*

Borrajo, L., Romero, R., Iglesias, E. L., and Marey, C. M. R. (2011). Improving imbalanced scientific text classification using sampling strategies and dictionaries. *Journal of Integrative Bioinformatics, 8(3),* pp. 1-15.

Brindha, S., Prabha, K., and Sukumaran, S. (2016). Pattern Document Weight Discovery For Text Classification Mining. *2016 International Conference on Communication and Electronic Systems, IEEE, 2016,* pp. 2–6.

Chen, J., Chen, C., and Liang, Y. (2016). Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word. *Advances in Intelligent Systems Research, 133*, pp. 114–117.

Dalal, M. K., and Zaveri, M. A. (2011). Automatic Text Classification: A Technical Review. *International Journal of Computer Applications, 28(2),* pp. 37-40.

Das, S., Dey, A., Pal, A., and Roy, N. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal of Computer Applications, 115(9),* pp. 31-41.

Dey, L., Chakraborty, S., Biswas, A., Bose, B., and Tiwari, S. (2016). Sentiment Analysis of Review Datasets using Naive Bayes' and k-NN Classifiers. *International Journal of Information Engineering and Electronic Business, 4,* pp. 54-62.

Eid, H. F., Hassanien, A. E., Kim, T., and Banerjee, S. (2013). Linear Correlation-Based Feature Selection For Network Intrusion Detection Model. *Advances in Security of Information and Communication Networks,* pp. 240-248.

Faraz, A. (2015). An Elaboration of Text Categorization and Automatic Text Classification Through Mathematical and Graphical Modelling. *Computer Science & Engineering: An International Journal, 5(213),* pp. 1-11.

Feng, P. M., Ding, H., Chen, W., and Lin, H. (2015). Naive Bayes Classifier with Feature Selection to Identify Phage Viron Proteins. *Computational and Mathematical Methods in Medicine.*

Ghareb, A. S., Abu Bakar, A., and Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications, 49,* pp. 31-47.

Gharehchopogh, F. S., Khaze, S. R., and Maleki, I. (2015). A New Approach in Bloggers Classification with Hybrid of k-Nearest Neighbor and Artificial Neural

Network Algorithms. *Indian Journal of Science and Technology, 8(3),* pp. 237-246.

Goudjil, M., Bedda, M., Koudil, M., and Ghoggali, N. (2015). Using Active Learning in Text Classification of Quranic Sciences. *International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2015,* pp. 209-213.

Goyal, S. (2016). Sentiment analysis of Twitter Data using Text Mining and Hybrid Classification Approach. *International Journal of Advance Research, Ideas and Innovations in Technology, 2(5),* pp. 1-9.

Gupta, A., Vedaldi, A., Zisserman, A. (2016). Synthetic data for text localisation in natural images. *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2315-24.

Guru, A., and Parveen, A. (2014). Classification of Text Data using Feature Clustering Algorithm. *International Journal of Research in Engineering and Technology, 3(3),* pp. 321-323.

Hagenau, M., Liebmann, M., Hedwig, M., and Neumann, D. (2012). Automated News reading: Stock Price Prediction based on Financial News using Context-specific Features," *IEEE 45th Hawaii Int Conf. on Syst Sciences,* pp. 1040-1049.

Hamed, S. K., and Ab Aziz, M. J. (2016). A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification. *Journal of Computer Sciences, 12(3),* pp. 169-177.

Hancer, E., Xue, B., and Zhang, M. (2017). Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems, 000,* pp. 1-17.

Hassan, G. S., Mohammad, S. K., and Alwan, F. M. (2015). Categorization of 'Holy Quran Tafseer' using k-Nearest Neighbour Algorithm. *International Journal of Computer Applications, 129(12),* pp. 1-6.

Hassan, R., Hossain, M. M., Bailey, J., and Ramamohanarao, K. (2008). *Improving k -Nearest Neighbour Classification with Distance Functions Based on Receiver Operating Characteristics*. Springer-Verlag Berlin Heidelberg, pp. 489–504.

Hilal, A., and Srinivas, N. (2015). Analytical of the Initial Holy Quran Letters Based on Data Mining study. *American International Journal of Research in Formal, Applied & Natural Sciences, 10(1),* pp. 1-8.

Hossin, M., and Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5(2),* pp. 1-11.

Houssain, K. (2014). *The World Muslim Population, History & Prospect.* Research Publishing.

Hu, R., Mac, B., and Delany, S. J. (2016). Active learning for text classification with reusability R. *Expert Systems With Applications*, *45*, pp. 438–449.

Ibrahim, E. A. A., Ataelfadiel, M. A. M., and Atwel, E. S. (2017). Provisions of Quran Tajweed Ontology (Articulations Points of Letters, UN Vowel Noon and Tanween). *International Journal of Science and Research, 6(8),* pp. 756-761.

Ivanovic, M., and Radovanovic, M. (2015). Modern Machine Learning Techniques and their Applications. *International Conference on Electronics, Communications and Networks.*

James, A. P., and Dimitrijev, S. (2012). Ranked selection of nearest discriminating features. *Human-centric Computing and Information Sciences.* Springer.

Jamil, N. S., Ku-mahamud, K. R., Din, A. M., Ahmad, F., Chepa, N., Ishak, W. H. W., Din, R., and Ahmad, F. K. (2017). A subject identification method based on term frequency technique. *International Journal of Advanced Computer Research, 7(30)*, pp. 103-110.

Jiang, L., Cai, Z., Wang, D., and Jiang, S. (2007). Survey of improving k-Nearest-Neighbour for Classification. *2007 IEEE Fourth International Conference on Fuzzy Systems and Knowledge Discovery.*

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., and Atteveldt, W. V. (2013). RTextTools: A Supervised Learning Package for Text Classification. *The R Journal, 5(1),* pp. 6-12.

Kassner, L. B., and Mitschang, B. (2016). Exploring Text Classification for Messy Data: An Industry use case for Domain- Specific Analytics. *Industrial and Applications paper.*

Khairnar, J., and Kinikar, M. (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classification. *International Journal of Scientific and Research Publications, 3(6)*, pp. 1–6.

Khan, M. K., and Alginahi, Y. M. (2013). The Holy Quran Digitization: Challenges and Concerns. *Life Science Journal, 10(2),* pp. 156-164.

Ladha, L., and Deepa, T. (2011). Feature Selection methods and algorithms. *International Journal in Computer Science and Engineering, 3(5),* pp. 1787-1797.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of Massive Datasets.* 2nd ed, England: Cambridge University Press.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2016). Feature Selection: A Data Perspective. pp. 1–73.

Liu, H., and Motoda, H. (2007). *Computational Methods of Feature Selection.* CRC Press.

Liu, C. L., Hsaio, W. H., Lee, C. H., Lu, G. C., and Jou, E. (2012). Movie rating and review summarization in mobile environment." *IEEE Transactions on Systems, Man, and Cybernetics.*

Manne, S., Muddana, S., Sohail, A., and Fatima, S. (2014). Features Selection Method for Automatic Text Categorization: A Comparative Study with WEKA and RapidMinor Tools. In ICT and Critical Infrastructure: *Proceedings of the 48th Annual Convention of Computer Society of India,* Springer.

Matias, L. M., Moreira, J. M., Gama, J., and Brazdil, P. (2012). *Text Categorization Using an Ensemble Classifier Based on a Mean Co-association Matrix.* Springer-Verlag Berlin Heidelberg, pp. 525–539.

Menaka, S., and Radha, N. (2013). Text Classification using Keyword Extraction Technique. *International Journal of Advanced Research in Computer Science and Software Engineering, 3(12),* pp. 734–740.

Nguyen, T. H., and Shirai, K. (2013). Text Classification of Technical Papers Based on Text Segmentation. *18th International Conference on Applications of Natural Language to Information Systems,* 2013.

Nikam, S. S. (2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. *An International Open Free Access, Peer Review Research Journal, 8(1),* pp. 13-19.

Nisha, S., Ali, N., and Ali, A. B. M. S. (2014). Searching Quranic Verses: A keyword based Query solution using .Net Platform," *5th Int. Conf. on Information and Communication Technology for the Muslim World,* Malaysia.

Pashaei, E., and Aydin, N. (2017). Binary black hole algorithm for feature selection and classification on biological data. *Applied soft computing, 56,* pp. 94-106.

Prusa, J. D., Khoshgoftaar, T. M., and Dittman, D. J. (2015). Impact of Feature Selection Techniques for Tweet Sentiment Classification. *Proceedings of the Twenty-Eight International Florida Artificial Intelligence Research Society Conference, 2015,* pp. 299–304.

Pundir, P., Gomanse, V., and Krishnamacharya, N. (2013). Classification and Prediction Techniques using Machine Learning for Anomaly Detection. *International Journal of Engineering Research and Applications, 1(4),* pp. 1716-1722.

Radovanovic, M., Ivanovic, M., and Budimac, Z. (2010). Text Categorization and sorting of Web search results. *Computing and Informatics, 28(5),* pp. 861-893.

Rentoumi, V., Raoufian, L., Ahmed, S., Jager, C. A., and Garrard, P. (2014). Features and Machine Learning Classification of Connected Speech Samples from Patients with Autopsy Proven Alzheimer's Disease with and without Additional Vascular Pathology. *Journal of Alzheimer's Disease, 42,* pp. 3-17.

Sabbah, T., and Selamat, A. (2014). Support Vector Machine based approach for Quranic words detection in online textual content. *8th IEEE Malaysian Software Engineering Conference,* Malaysia, pp. 325-330.

Saha, S. S., Sajjanhar, A., Gao, S., Dew, R., and Zhao, Y. (2010). Delivering Categorized News items using RSS feeds and Web services. *10th IEEE International Conference on Computer and Information Technology,* Los Alamitos, Calif, pp. 698-702.

Saikrishna, V., Dowe, D. L., and Ray, S. (2016). Statistical Compression-Based Models for Text Classification. *Fifth International Conference on Eco-Friendly Computing Communication and Systems, IEEE,* pp. 1-6.

Saloky, T. and J. Seminsky (2008). Artificial Intelligence and Machine Learning.

Santra, A. K., and Christy, C. J. (2012). Genetic Algorithm and Confusion Matrix for Document Clustering 1. *International Journal of Computer Science Issues, 9(1),* pp. 322–328.

Sewaiwar, P., and Verma, K. K. (2015). Comparative Study of various Decision Tree Classification Algorithm using WEKA. *International Journal of Emerging Research in Management & Technology, 4(10),* pp. 87-91.

Sharma, P., and Kaur, M. (2013). Classification in Pattern Recognition: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering, 3(4),* pp. 298-306.

Sharma, N., and Singh, M. (2016). Modifying Naive Bayes Classifier for Multinomial Text Classification. *IEEE International Conference on Recent Advances and Innovations in Engineering, 2016,* pp. 1-7.

Siddiqui, M. K., Naahid, S., and Khan, M. N. I. (2014). A Review of Quranic Web Portals through Data Mining. *VAWKUM Transactions on Computer Sciences, 5(2),* pp. 1-7.

Tang, J., Alelyani, S., and Lin, H. (2014). Feature Selection for Classification: A Review. *In Data Classification: Algorithms and Applications.* CRC Press.

Teli, S., and Kanikar, P. (2015). A Survey on Decision Tree Based Approaches in Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering, 5(4),* pp. 613-617.

Thasleema, T. M., and Narayanan, N. K. (2013). Multi Resolution Analysis for Consonant Classification in Noisy Environments. *International Journal of Image, Graphics and Signal Processing, 4(8),* pp. 15-23.

Timsina, P., Liu, J., and El-Gayar, O. (2016). Advanced analytics for the automation of medical systematic reviews. *Information Systems, Frontiers, 18(2),* pp. 237-252.

Tomar, D., and Agarwal, S. (2013). A Survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology, 5(5),* pp. 241-266.

Townsend, K. R., Sun, S., Johson, T., Attia, O. G., Jones, P. H., and Zambreno, J. (2015). k-NN text classification using an FPGA-based sparse matrix vector multiplication accelerator. *IEEE Int. Conf. on Electro/Information Technology,* pp. 257-263.

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications, 43,* pp. 82-92.

Wang, H., and Liu, S. (2016). An Effective Feature Selection Approach Using the Hybrid Filter Wrapper. *International Journal of Hybrid Information Technology, 9(1),* pp. 119-128.

Wang, J. H., and Wang, H. Y. (2014). Incremental Neural Network Construction for Text Classification. *IEEE International Symposium on Computer Consumer and Control,* pp. 970-973.

Xue, B., Zhang, M., Browne, W. N., and Yao, X. (2016). A Survey on Evoluntionary Computation Approaches to Feature Selection. *IEEE Transactions on Evoluntionary Computation, 20(4),* pp. 606-626.

Yang, J., Qu, Z., and Liu, Z. (2014). Improved Feature-Selection Method Considering the Imbalance Problem in Text Categorization. *Scientific World Journal, 2014,* pp. 1-17.

Zewen, C. (2016). Short Text Classification Based on Wikipedia and Word2vec. *2016 2nd IEEE International Conference on Computer and Communications, IEEE, 2016,* pp. 1195–1200.

Zharmagambetov, A. S., and Pak, A. A. (2015). Sentiment analysis of document using deep learning and decision trees. *Twelve IEEE International Conference on Electronics Computer and Computation,* pp. 1-4.

Zhou, P., Qi, Z., Zhang, S., Xu, J., Bao, H., Xu, B. (2016). Text classification improved by integrating Bidirectional LSTM with Two-dimensional Max pooling. *26th International Conference on Computational Linguistics: Technical Papers,* pp. 3485-95.