

**TOPIC IDENTIFICATION USING FILTERING AND RULE
GENERATION ALGORITHM FOR TEXTUAL DOCUMENT**

NURUL SYAFIDAH BINTI JAMIL

**MASTER OF SCIENCE (INFORMATION TECHNOLOGY)
UNIVERSITY UTARA MALAYSIA
(2015)**

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Abstrak

Maklumat yang disimpan secara digital dalam dokumen teks jarang disusun mengikut tajuk yang spesifik. Keperluan untuk membaca keseluruhan dokumen akan mengurangkan minat dalam pencarian maklumat. Kebanyakkan kaedah pengenalpastian tajuk bergantung kepada kekerapan perkataan yang muncul di dalam teks. Namun, bukan semua perkataan yang mempunyai kekerapan yang tinggi adalah relevan. Fasa penyarian perkataan dalam kaedah pengenalpastian tajuk menghasilkan perkataan yang mungkin mempunyai maksud yang sama yang dikenali sebagai masalah sinonim. Algoritma penyarian dan algoritma penjanaan peraturan diperkenalkan dalam kajian ini untuk mengenalpasti tajuk di dalam dokumen teks. Algoritma penyarian diperkenalkan (PFA) telah menyari perkataan yang paling berkaitan dari teks dan menyelesaikan masalah sinonim dalam kalangan perkataan yang di keluarkan. Algoritma penjanaan peraturan (TopId) diperkenalkan untuk mengenalpasti tajuk bagi setiap ayat berdasarkan perkataan yang dikeluarkan. PFA akan memproses dan menapis setiap ayat berdasarkan kata nama dan kata kunci yang telah ditetapkan untuk menghasilkan perkataan yang bersesuaian untuk tajuk. Peraturan kemudiannya dihasilkan dari perkataan yang dikeluarkan menggunakan pengkelasan berdasarkan peraturan. Rekabentuk eksperimen telah dijalankan ke atas 224 ayat Bahasa Inggeris daripada terjemahan Al-Quran yang berkaitan dengan isu wanita. Tajuk yang dikenalpasti oleh TopId dan teknik Set Kasar dibandingkan dan kemudian disahkan oleh pakar. PFA telah berjaya menyari perkataan yang berkaitan berbanding dengan teknik penapisan yang lain. TopId telah mengenalpasti tajuk yang hampir sama dengan tajuk dari pakar dengan ketepatan 70%. Kedua-dua algoritma yang dicadangkan berupaya mengeluarkan perkataan yang berkaitan tanpa kehilangan perkataan yang penting dan mengenalpasti tajuk dalam ayat.

Kata kunci: Pengenalpastian tajuk, Algoritma penyarian, Algoritma penjanaan peraturan, Set Kasar, Ayat Al-Quran.

Abstract

Information stored digitally in text documents are seldom arranged according to specific topics. The necessity to read whole documents is time-consuming and decreases the interest for searching information. Most existing topic identification methods depend on occurrence of terms in the text. However, not all frequent occurrence terms are relevant. The term extraction phase in topic identification method has resulted in extracted terms that might have similar meaning which is known as synonymy problem. Filtering and rule generation algorithms are introduced in this study to identify topic in textual documents. The proposed filtering algorithm (PFA) will extract the most relevant terms from text and solve synonym problem amongst the extracted terms. The rule generation algorithm (TopId) is proposed to identify topic for each verse based on the extracted terms. The PFA will process and filter each sentence based on nouns and predefined keywords to produce suitable terms for the topic. Rules are then generated from the extracted terms using the rule-based classifier. An experimental design was performed on 224 English translated Quran verses which are related to female issues. Topics identified by both TopId and Rough Set technique were compared and later verified by experts. PFA has successfully extracted more relevant terms compared to other filtering techniques. TopId has identified topics that are closer to the topics from experts with an accuracy of 70%. The proposed algorithms were able to extract relevant terms without losing important terms and identify topic in the verse.

Keyword: Topic identification, Filtering algorithm, Rule generation algorithm, Rough Set, Al-Quran verses.

Acknowledgement

All praise to Allah SWT who gave me strength and patience to finish this study. Alhamdulillah.

Firstly, I would like to express my gratitude to my first supervisor Prof. Dr. Ku Ruhana Binti Ku Mahamud for her support and her willingness to guide me on such an interesting research for my master thesis. I have learned so much from you, Prof. Secondly; I would like to thank my second supervisor Miss Aniza Binti Mohamed Din for providing me with great input and constant encouragement. Thank you for not losing hope on me from the beginning until the end of my study.

Furthermore, my sincere appreciation also goes to Associate Prof. Dr. Faudziah Binti Ahmad, Dr. Siti Sakira Binti Kamaruddin, Dr. Nooraini Binti Yusof, Dr. Yuhanis Binti Yusof and to Dr. Massudi Bin Mahmuddin who assisted and continuously helped me to understand and deepen my knowledge, especially in text mining and text classification related work. I am also truly indebted to Prof. Dr. Rosna Binti Awang-Hashim, Associate Prof. Dr. Norhafezah Binti Yusof, Prof. Dr. Nena Valdez and Associate Prof. Dr. Arminda Santiago for giving me moral support and continuously inspiring me to love my research and make this study even more interesting and joyful.

Not forgotten, gratitude to my beloved parents who never gave up in providing me with love and support to keep me strong. To my respected father, Jamil Bin Yusuff, you are the biggest reason why I am pursuing my study. Thank you for your valuable advice and motivation. To my cheerful mother, Midah Binti Musa, thank you Mama for supporting me to achieve my dream. Without your prayers, none of this work would have been accomplished.

Last but not least, thank you so much to all my lab-mates for our discussions and knowledge sharing session. It has been a great experience to know all of you. I will never forget our moments of tears and laughter in the years of our study. I wish all of you good luck in the future. To those people whom I have not mentioned here, thank you very much.

TABLE OF CONTENTS

Permission to use	ii
Abstrak.....	iii
Abstract.....	iv
Acknowledgement	v
List of tables	vii
List of figures.....	ix
List of appendices	x
List of abbreviation.....	xi
CHAPTER 1: INTRODUCTION	1
1.1 Problem statement	4
1.2 Research objective.....	6
1.3 Research scope.....	6
1.4 Significance of study	7
1.5 The application domain	8
1.6 Summary.....	10
CHAPTER 2: LITERATURE REVIEW	11
2.1 Introduction	11
2.2 Overview of Text Mining	11
2.3 Text Classification Phase.....	12
2.3.1 Statistical text pre-processing	13
2.3.2 Computational Linguistic text pre-processing	14
2.3.3 Text representation.....	18
2.3.4 Dimensionality reduction.....	21
2.3.5 Feature weighting	25
2.4 Rule-based Classification Techniques	28
2.5 Topic Identification Methods	30
2.5.1 Statistical approach	31
2.5.2 Ontological approach	32
2.5.3 Rule-based approach.....	34
2.6 The Quran as a Case Study	36
2.6.1 Women Issues in the Quran	36
2.6.2 Knowledge extraction from Quran	37

2.6 Summary.....	39
CHAPTER 3: RESEARCH METHODOLOGY	40
3.1 Introduction	40
3.2 Research Framework	40
3.2.1 Phase one: Text pre-processing and term extraction	42
3.2.2 Phase two: Term ranking	45
3.2.3 Phase three: Rule generation.....	45
3.2.4 Phase four: Evaluation	46
3.3 Summary.....	46
CHAPTER 4: TOPIC IDENTIFICATION METHOD	48
4.1 Introduction	48
4.2 The Proposed Topic Identification Method.....	48
4.3 Text Pre-processing and Term Extraction	51
4.3.1 Text pre-processing.....	50
4.3.2 Term extraction.....	59
4.4 Term Ranking	60
4.5 The Proposed Rule Generation Algorithm (TopId)	64
4.5.1 Rule generation using Rough Set technique	66
4.5.2 Comparison of topics with experts	71
4.6 Summary.....	73
CHAPTER 5: EXPERIMENT AND PERFORMANCE EVALUATION	75
5.1 Introduction	75
5.2 Experimental Design	75
5.3 The Proposed Filtering Algorithm Result.....	77
5.4 The Rule Generation Algorithm Result.....	80
5.5 The Comparison of Results with Experts	86
5.6 Summary.....	94
CHAPTER 6: CONCLUSION	95
6.1 Introduction	95
6.2 Contribution of the research	95
6.3 Future work.....	96

List of Tables

Table 1.1: Various female topics in the Quran verse	10
Table 2.1: Summarization of text pre-processing approaches	17
Table 2.2: Summarization of text representation	20
Table 2.3: Summarization of feature weighting	27
Table 2.4: Summarization of rule-based classification techniques	29
Table 3.1: Sample of Part-of-Speech tag set	44
Table 4.1: Sample of relevant terms	62
Table 4.2: Sample of ranked terms and tf , idf & $tf - idf$ score	63
Table 4.3: Sample of decision table used for training with Rough Set technique	68
Table 4.4: Data discretization	69
Table 4.5: Split factor and data division	70
Table 4.6: The trained models divided in Rosetta application	71
Table 4.7: Sample of form for experts	72
Table 4.8: Comparison of topic between expert and TopId	73
Table 5.1: Sample of the extracted terms	78
Table 5.2: Sample of the ranked terms	81
Table 5.3: Sample of the identified topic for each verse by TopId	82
Table 5.4: The result of 10-Fold Cross Validation on Rough Set models	84
Table 5.5: Sample of produced rules and identified topics from Rosetta-Rough Set application ..	85
Table 5.6: Sample of topic comparison for Expert 1 and TopId	87
Table 5.7: Sample of topic comparison for Expert 1 and Rough Set technique	88
Table 5.8: Accuracy for comparison of TopId and Rough with Expert 1	89
Table 5.9: Accuracy for comparison of TopId and Rough with Expert 2	90
Table 5.10: Accuracy for comparison of TopId and Rough with Expert 3	92

List of Figures

Figure 2.1: Basic topic identification system	30
Figure 3.1: Research framework	41
Figure 3.2: Phase 1: Text pre-processing and term extraction	42
Figure 3.3: Phase Two: Term ranking.....	45
Figure 4.1: The proposed topic identification method	49
Figure 4.2: The flowchart of text pre-processing and term extraction	52
Figure 4.3: The pseudo code of text pre-processing and term extraction.....	53
Figure 4.4: The pseudo code of text pre-processing	54
Figure 4.5: Sample of tokenized text	55
Figure 4.6: Flowchart of case folding process	56
Figure 4.7: List of noise words	56
Figure 4.8: Stemming using NLTK Demo.....	57
Figure 4.9: The interface of tagging application	58
Figure 4.10: The produced tagging output	58
Figure 4.11: Flowchart of the proposed filtering algorithm (PFA)	59
Figure 4.12: The proposed rule generation algorithm (TopId).....	64
Figure 4.13: The general scheme of Rough Set technique for topic identification	67
Figure 4.14: Data for the experiment in Rosetta application.....	70
Figure 5.1: Experimental design	77
Figure 5.2: Total matched topics and accuracies for comparison with Expert 1	89
Figure 5.3: Total matched topics and accuracies for comparison with Expert 2.....	91
Figure 5.4: Total matched topics and accuracies for comparison with Expert 2.....	93
Figure 5.5 The accuracies for comparison of TopId and Rough Set with Expert 1, Expert 2 and Expert 3	93

List of Appendices

Appendix A: The verses used as dataset.....	107
Appendix B: The keywords	124
Appendix C: The extracted terms	125
Appendix D: The produced rules and the identified topics by TopId.....	148
Appendix E: The produced rules and the identified topics by Rough Set	159
Appendix F: The identified topics by experts	165
Appendix G: Topic comparison of TopId and experts	170
Appendix H: Topic comparison of Rough Set and experts	176

List of Abbreviations

RSAR	Rough Set Attribute Reduction
PFA	Proposed Filtering Algorithm
TopId	Topic Identification
NLP	Natural Language Processing
POS	Part-of-Speech
VSM	Vector Space Model
TF-IDF	Term Frequency-Inverse Diverse Frequency
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
LSI	Latent Semantic Indexing
PbuH	Praise Be Upon Him
IR	Information Retrieval
HMM	Hidden Markov Model

CHAPTER ONE

INTRODUCTION

The trend to store data in electronic format has increased and requires an efficient effort to organize these important yet beneficial documents (Sumathy & Chidambaram, 2013). Data can be stored in several formats such as image, video and audio. However, text is commonly used to store knowledge and exchange information (Sumathy & Chidambaram, 2013; Jusoh & Alfawaref, 2012; Mahender, 2012 & Aery, Ramamurthy, & Aslandogan, 2003). Text is usually unstructured and not restricted to any specific format (Jusoh & Alfawaref, 2012; Kamaruddin, 2011). The exploration of hidden information from this unstructured text is useful because interesting patterns and valuable knowledge can be discovered from the text (Sumathy & Chidambaram, 2013).

The issue in text mining research domain is to organize a vast amount of textual documents that have no specific topic or category describing the content (Aggarwal & Zhai, 2012; Jusoh & Alfawareh, 2012; Hotto, Nurnbeger & Paab, 2005). Text classification concerns on determining which decision is made on the information. There are many underlying classifiers for text classification such as Decision Tree, Rule-Based, Neural Network, K-Nearest Neighbour and Support Vector Machine (Aggarwal & Zhai, 2012). Topic identification is a method that lies under text mining domain and aims to analyze the text data and assign a correct label as a topic for the text documents (Baghdadi & Ranaivo-Malancon, 2011). However, analyzing texts is tedious, in which the complexity of natural language and misinterpretation that leads to misunderstanding might occur (Sumathy & Chidambaram, 2013; Jusoh &

Alfawareh, 2012). In order to classify text documents with appropriate topics, the assurance to extract relevant features are needed (Ko, Park & Seo, 2011).

The term ‘topic’ refers to a concept or a subject that can be used to categorize a document. It indicates a certain subject which is discussed in the whole text (Jain & Pareek, 2010) and usually represented by a single term. A topic usually assists people to search and understand the whole sentence in a text (Baghdadi & Ranaivo-Malancon, 2011; Jain & Pareek, 2010) and can be determined by looking at the accuracy of terms from the sentence (Butarbutar & McRoy, 2004). The most repeated terms show the potential of important concepts or topics in the text. In some cases, the topic label does not appear in its exact form inside the document (Natarajan, Prasad, Subramanian, Saleem, Choi & Schwartz, 2007). The challenge is to find the relevance of terms and novel information from set (Allam, Carterette & Lewis, 2005; Baghdadi & Ranaivo-Malancon, 2011).

Terms in a text may be irrelevant, or other words, it might have no effect on the processing which could give impact on the processing performance (Ventura & Silva, 2008). Feature selection is needed in the pre-processing stage in order to retrieve the quality of features since not all features are relevant for classifying the text (Bong & Wong, 2005). Several methods have been widely applied for feature selection in text classification such as Information Gain (IG), Mutual Information feature selection (MIFS) (Bakus & Kamel, 2006), Document Frequency-based Selection, Term Strength, and Entropy-based Ranking, but these statistical-based methods only provide information access instead of analyzing information to locate patterns.

For topic identification, the extraction of terms from texts may influence the quality and usefulness of the final results of the analysis. Some of the information extraction systems rely on nouns as their features for entity identification such as in Berkowitz (2010). Noun carries a role to interpret the structure of sentence, mostly in particular, fields such as machine translation, text retrieval, information extraction and text classification (Sagar, Shobba, & Kumar, 2009). Nouns in text may portray the topic that has been discussed (Protaziuk, Kryszkiewicz, Rybinski, & Delteil, 2007). In addition, classifying texts that are based on a single linguistic expression can be effective since nouns can represent specific incidents and general events in the sentence and likely produce good topics in the sentence (Dong, Schall, O'Mahony & Smyth, 2013; Riloff, 1995). By this means, mastering noun leads to understanding the main meaning denoted in the text.

To overcome this problem, statistical based approach feature selection combined with the computational linguistics technique is proposed. The nature form of unstructured textual data is the challenge to ensure that the selected relevant terms from the dataset may contribute to the classification of topics. This study focuses on noun as a term candidate; thus Part-of-Speech tagging technique is implemented for identifying the syntactical parts in the text document.

1.1 Problem Statement

Topic identification method assigns one or several topic labels on a flow of textual data. However, the problem of topic identification is it depends on the occurrence number of terms in the document (Fuddoly, Jaafar, & Zamin, 2013; Hassan, Karray & Kamel, 2012; Baghdadi & Ranaivo-Malancon, 2011; Brun, Smaili & Haton, 2002). Though the topic depends on the occurrences of terms in the text, not all terms with high occurrence numbers are relevant to represent the topic (Ventura & da Silva, 2008). Hence, retrieving the most relevant term is necessary. However, retrieving relevant terms from text is a non-trivial task because of the high dimensionality nature in text and lack of formal structure in text document (Kamaruddin, 2011). In other words, getting a handle on what is important and what is not important from textual data is not easy. Due to the complexity of the text, dimensionality reduction technique is needed to reduce the irrelevant terms in the text and make them easier to handle (Khan, Baharudin, Lee & Khan, 2010).

Another issue is the synonymy of the terms. Synonym is natural linguistic phenomena which Computational Linguistic and Information Retrieval researchers commonly find difficult to cope with (Sheeba & Vivekanandan, 2012). Several attempts which are based on semantic to solve synonymy problem by using WordNet (Mc Crae, 2009), computed semantic relatedness from Wikipedia (Gabrilovich & Markovitch, 2007) and extracting word similarity from website (Phan, Nguyen, Le, Nguyen, Horiguchi & Ha, 2011). Unfortunately, WordNet is not effective as it did not include any specific terms that is needed (Mc Crae, 2009). Contents in Wikipedia are not consistent because some articles are missing and anyone can change the content (Hasan, 2013). Different websites might use different word similarity and this can

cause inconsistency in solving synonymy problem. Another different solution has been implemented which is to use rule-based term extraction as a controlled environment for synonym distribution such as in Wang & Hirst (2009). Therefore, a rule-based filtering algorithm is proposed and a keywords library is also developed to solve synonymy of the extracted terms.

In ranking issue, Term Frequency-Inverse Document Frequency technique or also known as TF-IDF is used to evaluate the relevancy of a term in the document (Hong, Lin, Yang & Wang, 2013; Zhang, Wang, Wu & Hu, 2012; Zhang, Yoshida & Tang, 2011). But, a term that has the highest score has the potential to give vague information to identify one specific topic (Zhang, Wang, Wu & Hu, 2012; Zhang, Yoshida & Tang, 2011). In fact, TF-IDF cannot solve synonymy problems because it ignores the relationship between words (Badr, Chbeir, Abraham & Hassanien, 2010; Ramos, 2003). This matter can be resolved by solving synonymy problem during text pre-processing phase.

Most topic identification methods are based on statistical, ontological (Skorkovska, Ircing, Prazak & Lehecka, 2011; Stein & Eissen, 2004) and rule-based (Zhang & Zhao, 2010; Stoyanov & Cardie, 2008; Yeh & Chen, 2007; Liu, Chin & Ng, 2003 & Clifton & Cooley, 2000). Amongst these approaches, rule-based approach for topic identification has gained attention for this study on several reasons. Rule-based can be tuned on uncertainty data (Qin, Xia, Prabhakar & Tu, 2009) and relatively easy for people to understand the rules (Qin et al., 2009). Unlike rule-based, other rule generation technique such as Rough Set usually chaotic and not stable (Bazan, Nguyen, Nguyen, Synak & Wroblewski, 2000) and not always sufficient for

extracting information from data. Rough Set depends on the information table that needs to be provided before starting the training and testing processes. Decision Tree technique for rule generation is also crucial since a slight change can result in drastically different trees and cause overfitting to happen (Chizi, Rokach & Maimon, 2009). Meanwhile, Association rule tends to produce huge numbers of rules which can cause noisy information and mislead the classification process (Korde, 2012; Rajan & Dhas, 2012; Garcia, Romero, Venture & Calders, 2007).

1.2 Research Objective

The main objective of this study is to develop a topic identification method for textual document. Specific objectives to support the main objective are as follows:

1. To develop a filtering algorithm to extract relevant terms from text.
2. To propose a technique in solving synonymy problem during term extraction.
3. To construct a rule generation algorithm to identify topic from the extracted terms.
4. To evaluate the produced rules and topics.

1.3 Research Scope

The domain of this study is Text Mining and the data that has been used is English translated Quran. The selected verses are retrieved from Surah.my website because it has been actively used in Malaysia (Ku-Mahamud, Ahmad, Mohamed Din, Ahmad, Din, & Che Pa, 2012). The extraction is based on the occurrences of female-related terms in the verse. Knowledge that has been extracted on female-related terms is further explored by constructing rules for identifying the topic. The study only focuses

on the verses that are related to main female topics in the Quran and the topics are; marriage, divorce, and inheritance.

This study employed the technique from both Computational Linguistic and statistical based approach. The Computational Linguistic technique which is Part-of-Speech tagging is applied to identify and collect nouns as the extracted terms. However, this study does not consider semantic value within text, since the objective of this study is to classify text for topic identification instead of understanding the content of the Quran.

1.4 Significance of the Study

The significance of this study is twofold; which are to the body of knowledge and to the society. First, the proposed topic identification method is different with other existing topic identification methods because the proposed filtering algorithm is designed to solve synonymy of the extracted terms. A keywords library that contained specific terms such as ‘zihar’ and ‘iddat’ is also developed to ensure that important terms are not eliminated during extraction phase. Besides, most of the existing rule generation techniques tend to produce too many rules, hence; the proposed rule generation algorithm is simpler and able to produce topics especially for topic Marriage, Inheritance and Divorce. The proposed method can be used in identifying topic from other textual data such as meeting transcript, news articles, and social networking posts like Twitter, Facebook and Tumblr. In addition, the method can be also implemented for other Holy Book such as the Bible because the method can identify topic in sentence level which is by each verse.

Second, The Quran is not only referred by Muslim people to discover knowledge. It is also been studied by Non-Muslim people who has interest to learn from the content of Quran. The proposed topic identification method can identify topic of each verses from the English translated Quran and can ease the people to search for useful and interesting information such as topic Marriage, Inheritance and Divorce.

1.5 The Application Domain

A large amount of English translated Quran sources are available through the Internet. However, these sources are not well structured for accurate search and efficient delivery through the web (Manacer & Arbaoui, 2013). The Quran is the major source of knowledge in Islam and a number of automated applications have been proposed to ease the retrieval of knowledge (Yauri, Kadir, Azman & Murad, 2013a & Yauri, Kadir, Azman & Murad, 2013b). Discovering knowledge that is hidden underneath the Quran has challenged researchers to represent the knowledge wisely. The Quran is revealed to Prophet Muhammad (praise be upon Him) and it is used as the guidance for Muslims worldwide. The Quran is a concise dataset, a text of 77,000 words and sequenced by chapters and 6200 verses. The Quran contains information that can be used as the solutions to many questions and doubts (Ku-Mahamud et al., 2012 & Sharaf, 2009).

Any issue on today's life has been stated in its chapters and verses. Various topics are categorically exposed through the verses in the Quran (Yauri et al., 2013a, 2013b). One example is the information on women issues. The main issues are women rights in marriage, inheritance, and divorce (Abdullah & Sudiro, 2010). Meanwhile, according to a Muslim scholar writer, Badawi (1980) and Badawi (2000), the Holy

Quran has included issues on women in terms of spiritual, social, economical, and political aspects.

In terms of dataset form, the Quran can be considered as a semi-structured dataset because the chapters and verses are arranged accordingly by the structure of numbers (Ahmad, Hyder, Iqbal, Murad, Mustapha,...Mansoor, 2013). However, topics that are related with female are mentioned in several verses in the Quran. In addition to this, the content of female topics needs to be reported in a very accurate manner due the sensitive nature of the Quran. This is because The Quran is the Holy Book of Muslim and it is not only referred by Muslim people as their main reference, but also studied by Non-Muslim people who have interest on it. Hence, identifying topic from the Quran should be done meticulously and any misinterpretation should be avoided.

Most of the studies that retrieve information from the Quran have been based from keyword matching (Yauri et al., 2013a). Nevertheless, keywords or terms without any significant meaning provide inaccurate result and lack of intelligent features in Information Retrieval system (Baghdadi & Renaivo-Malancon, 2011; Edward, James & Raymond, 2011). Hence, the use of term as a keyword has been the main measurement since every single term may have different importance which is based on the weight to the information domain (Edward et al., 2011). There are many terms related to female from the Quran verses and they are; consorts, damsel, divorce, mother, queen, sister, wife, and so on.

Based from this explanation, Table 1.1 is constructed from an initial analysis in the first stage of this study. Table 1.1 shows that female topics on marriage occurs in

separated places, which are the verses from Surah Al-Baqarah, An-Nisa' and Al-Ahzab. Besides that, the table also outlines that various female topics such as inheritance, marriage, and divorce occur in the same verse.

Table 1.1

Various female topics in the Quran verse

Surah	Female Term	Topic
Al-Baqarah (2:236)	women	marriage
An-Nisa' (4:23)	mother, daughter, sister, wife	marriage
Al-Ahzab (33:50)	wife, aunt, daughter, woman	marriage
An-Nisa' (4:12)	wife	inheritance
An-Nisa' (4:128)	wife	divorce
An-Nisa' (4:24)	women	marriage

1.6 Summary

Textual documents are high dimensional data. With the intention to the mentioned problems, this study proposed to design an automatic text processing that is topic identification in text document. Since the retrieval tasks are based from terms, it should be properly chosen and relevant in order to get the topic of the verses. For this reason, the aim of this study is to select noun as a term candidate instead of other linguistic expressions. This is due to the concern that noun is not only able to express the specific incidents in the sentence, but it is reasonable to produce good topics in the sentence as well. Over the years, both Muslims and non-Muslims are interested in Islamic knowledge especially on the Quran. The Quran contains various hidden knowledge. Hence, the English translated Quran has been used as the dataset in this study and the topics from it are explored.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents the literature review of the study. The main objective of this study is to develop topic identification method for textual document. Hence, this chapter aims to identify the limitation of the current work and disclose the gap. The discussion in this chapter is arranged accordingly based from the identified problems and objectives that were explained in Chapter 1.

This study involves text domain study; therefore, the discussion starts by presenting the overview of text mining in Section 2.2. Each of the processes involved in text classification is explored in Section 2.3. Next, rule-based classification techniques are discussed in Section 2.4. The previous studies on topic identification methods are presented in Section 2.5. The background of the domain which is the Quran as a case study is explored in Section 2.6. This chapter ends with a summary in Section 2.7.

2.2 The Overview of Text Mining

Text mining is the process to extract interesting patterns from large textual data for which is useful for discovering knowledge (Gupta & Lehal, 2009; Stavrianou, Andritsos & Nicoloyannis, 2007; Ben-Dov & Feldman, 2005). Text mining is also referred as text data mining, intelligent text analysis or knowledge discovery in text (KDT) (Gupta & Lehal, 2009; Ramanathan & Meyyappan, 2013).

Tasks that can be implemented from text mining are text classification, information extraction and summarization (Aggarwal & Chai, 2012; Gupta & Lehal, 2009). Amongst all these tasks, text classification is known as topic spotting or topic identification because it can automatically sort a set of documents into categories, classes or topics from a predefined set (Sadiq & Abdullah, 2013). Patel (2012) stated that text classification can be a medium for topic tracking in order to classify document by topic and make the process faster.

Based from the previous studies by Yuan (2010) and Kamaruddin (2011), text classification gives benefit to many applications especially for retrieving information from high-dimensional data. Therefore, text classification is used as a benchmark method for topic identification in textual document as in the study by Sadiq and Abdullah (2013). The explanation of the involved processes in text classification is described in the following section.

2.3 Text Classification Phase

Text classification is the task of assigning documents to predefined categories (Bhumika, Sehra, Nayyar, 2013; Ko, 2012; Gupta & Lehal, 2009; Ko, Park & Seo, 2002). Text classification process starts by collecting documents such as .html, .doc, or .pdf. The texts from these documents are pre-processed to eliminate noisy and irrelevant terms (Bhumika, Sehra, Nayyar, 2013; Dalal & Zaveri, 2011; Gupta & Lehal, 2009). The pre-processing step, such as tokenization, stop word removal and stemming, ensures to avoid any repetition and noise data exists in the text (Dalal & Zaveri, 2011). Next, dimensional reduction techniques are performed to extract

discriminative features from the high dimensionality of the text. These features are next weighted and presented to the classifier in order to build a classification model.

2.3.1 Statistical Text Pre-Processing

Text pre-processing is needed before any work of text analyzing can be started in order to present text document into clear words (Ramanathan & Meyyapan, 2013; Sagayam, Srinivasan & Roshini, 2012; Patel & Soni, 2012; Manne & Fatima, 2011 & Hotto et., 2005). Text pre-processing's basic idea is to transform the unstructured textual data into a structured form that can be easily processed and it should perform before representing it into formal structures (Berry & Kogan, 2010; Elder, 2012; Zhong, Li, & Wu, 2012; Gupta & Lehal, 2009; Feldman & Sanger, 2007; Hotto et al., 2005). Text pre-processing involves tasks to reduce the size of text in high dimensional data (Feldman & Sanger, 2006).

In statistical text pre-processing, processes such as removing stop words and stemming (Kamaruddin, 2011) are involved. Stop words is removed since its repetition in the text do not provide any meaningful pattern and can affect the performance of classification. Stemming converts words into their common root word by dismissing any suffixes and prefixes from the words in order to avoid redundancy of terms (Ali & Ibrahim, 2012; Bakar & Rahman, 2003; Rahman, Bakar, & Mohamed Hussein, 2012; Serrano, Del Castillo, Oliva & Iglesias, 2012; Sharma, 2012). The words are then listed according to the document frequency that is based on the occurrences of the words in the document and sorted according to the highest count. Most studies that applied statistical approach during text processing chose the top-ranked words in the list and used them as the keyword to represent the text document.

The most widely used stemming algorithm for English language is Porter's algorithm, for French is Savoy's algorithm, and for Malay language is Fatimah's et al (T. Sembok, Abu Bakar & Ahmad, 2011). Stemming is usually applied in the indexing process since it is able to reduce the size of index terms and improving the degree of relevancy in retrieving documents (T. Sembok, Abu Ata, & Abu Bakar, 2011 & T. Sembok et al., 2011). According to Bakar and Rahman (2003), removing the suffixes using stemmers for English, Slovene and French languages are sufficient for information retrieval purposes.

In contrast with that, removing suffixes, prefixes and infixes properly is needed for stemming Malay texts (Hamzah & T. Sembok, 2006). Statistical approach for text pre-processing is assumed as the simplest approach that is used by researchers such as in the study by Hamzah and T. Sembok (2006), who applied stemming to retrieve BaseNP from text and the retrieval results is significantly proven. Besides that, Ghairabeh and Ghairabeh (2012) also applied stop word removal and stemming process during the pre-processing session and proved that pre-processing helps to clean the raw text and improve the accuracy of classification results. Bakar and Rahman (2003) used both Malay stemming and thesaurus in searching and retrieving relevant Malay translated Quran documents that are based on user query.

2.3.2 Computational Linguistic Text Pre-Processing

Another approach that can be applied for text pre-processing is Computational Linguistic. By using Computational Linguistic or Natural Language Processing (NLP) approach, text documents are processed using general knowledge about natural language. NLP helps human language to be understandable by computers (Mendes &

Antunes, 2009; Hotto et al., 2005). As compared to statistical approach, Computational Linguistic approach is more costly as it requires a deep analysis of text using tagging, syntactic parsing and semantic analyzer (Kamaruddin, 2011). The techniques included in this approach such as tokenization, Part-of-Speech tagging, morphological analysis, syntactic parsing either deep or shallow level capture sentence structure and the meaning it represents.

Generally, datasets are downloaded from the Internet either in DOC or PDF formats and converted into plain text. Hence, tokenization process is required to identify the words in sentences from the text. The stream of characters in a text is broken up into smaller and meaningful units which are called token before any language processing can be performed (Laboreiro, Sarmento, Teixeira & Oliveira, 2010). The tokenization process for English text is easier since white spaces and punctuation marks separate words. However, the tokenization process is complicated on languages such as Chinese, Japanese, and Korean because the spaces do not separate the words (Basit & Jarzabek, 2007; Kaplan, 2005; Laboreiro et al., 2010).

Another technique under Computational Linguistic is Part-of-Speech tagging. It also known as POS tagging that identifies each word's part in the sentence such as noun, verb, pronoun and other grammatical tags according to its context (Collobert, Weston, Bottou, Kavukcuoglu & Kuksa, 2011; Nadkarni, Ohno-Machado & Chapman, 2011; Okhovvat & Bidgoli, 2010; Petrov, Das & McDonald, 2011). POS tagging is a basic tool and needed in many fields of linguistic processing. Beside that, POS tagging is also implemented in the first stage for analyzing text. Many tagging programs have been developed for different languages and used to build various kinds of applications, such

as Persian POS tagger (Okhovvat & Bidgolib, 2010), English POS tagger Stanford tagger (Manning, 2011), Arabic POS tagger (I. Ghairabeh & N. Ghairabeh, 2012; Salem, 2009; Sawalha, Brierly & Atwell, 2012).

Another Computational Linguistic technique that has been normally used in text pre-processing is parsing technique. Parsing is considered with larger syntactic units such as phrases, clauses and sentences. It involves the process of assigning a syntactic analysis to a sentence according to some grammars by using a computer program known as parser. The parsing of natural languages is rather complicated and more advanced than Part-of-Speech tagging, as it involves a greater number of structures that are not just words, but whole phrases and clauses in the sentence (Dhonnchadha, 2008). In other words, a parser is used to analyze the whole syntactic and semantic values of words in the sentence. Various parsing programs are available such as rule based tagger, i.e Brills tagger (Amrani, Aze, Heitz, Kodratoff & Roche, 2004), and dependency treebank, i.e Penn Tree Bank (Gamon, Gao, Brockett, Klementiev, Dolan, Belenko & Vanderwende, 2009).

Table 2.1 summarizes the techniques in both statistical approach and computational linguistic approach that have been discussed in this section. The advantages and disadvantages of the approaches are also listed.

Table 2.1

Summarization of text pre-processing approaches

Approach	Techniques	Advantages	Disadvantages
Statistical	Stop word removal, stemming, word frequency	Remove unimportant words that caused inaccurate classification	Ignore the relationship between words
Computational linguistic	Tokenization, morphological analysis, POS tagging, syntactic parser	Deeper analysis in analyzing texts such as the syntactic and semantic values within the text	Language dependent and high cost in terms of implementation of programs, tools

In brief, the advantage of statistical text pre-processing is simple in terms of its implementation and practicality. However, this approach omitted the relationship between words since the words are treated individually (Kamaruddin, 2011). Referring to the brief elaboration in this section, Computational Linguistic text pre-processing provides the understanding of a text document as deepest as possible. This is due to the fact that all techniques from Computational Linguistic analyzes the structure and the semantic of sentences. However, Computational Linguistic text pre-processing requires extra efforts as it is computationally expensive and also language dependent. Based from gaps mentioned here, this study used statistical text pre-processing because it does not concern on the semantic and syntactic values between words. Only one technique from Computational Linguistic which is POS tagging. This technique is chosen because it is able to give grammatical tags to each terms and can be easily implemented using NLTK tools.

2.3.3 Text representation

Textual data usually consists of strings of characters and cannot be directly interpreted by a classifier (Aggarwal & Zhai, 2012; Srivastava & Sahami, 2010; Zhang et al., 2011). Hence, a text should be transformed into suitable and meaningful representations of text units in order to allow learning algorithm and classification to perform its task (Feldman & Sanger, 2007; Hotto et al., 2005; Zhong et al., 2012). These meaningful representations of text units is referred as terms and reflect the meaning of the text units itself (Hassan, 2013).

Text representation concerns on what should a term be, which focuses on the specific terms in the document (Kamaruddin, 2011). In addition, text representation also demonstrates on the occurrences of frequencies the computation of term weight. Most text representation techniques that are widely used in text classification domain are Vector Space Model, N-gram, Ontology, Single term, and Phrase approach (Khan et al., 2010; Kamaruddin, 2011; Keller & Bengio, 2005).

Apparently, most studies on text classification use the Bag of Words representation since it is simpler for classification purposes (Hassan, 2013; Aggarwal & Zhai, 2012; Harish, Guru & Manjunath, 2010; Hotto et al., 2005). This model was first introduced by Salton and Buckley (1988), also referred as Vector Space Model (VSM). VSM represents a set of documents as term feature and the weight of feature (Korde & Mahender, 2012 & Zhang et al., 2012). The weight of feature refers to the importance of the feature when it describes the content of text. The importance of features depends of how heavy the weight is and usually word frequency is used in order to weight the feature (Turney & Pantel, 2010; Zhao, Chen, Fan, Yan & Li, 2012).

However, VSM representation is high dimensional and causes a loss of correlation with adjacent words (Turney & Pantel, 2010; Zhao et al., 2012). A lack of correlation with adjacent words might also lead to a loss of semantic relationship amongst the terms in the document (Aggarwal & Zhai, 2012). The solution to overcome the correlation problem is to use term weighting techniques (Zhao et al., 2012). An appropriate weight that is assigned to the terms can improve the performance of text classification (Harish et al., 2012). The commonly used term frequency technique is Term Frequency-Inverse Document Frequency (TF-IDF) (Xu et al., 2012). TF is the word frequency that holds the value of the number of appearances of feature entry in the underlying texts. IDF is the inverse frequency that shows the statistical frequency of a feature in a text set. The advantages of TF-IDF are it is easy to implement and is well studied in the text mining domain. In fact, many enhancements have been made on TF-IDF and diverse variations are available.

Another technique that differs from VSM is N-gram. In this technique, a term is decomposed into a unit fragment of size n . A matching algorithm is designed to compare these fragments to identify the similarity and dissimilarity. N-gram is a technique that is robust to noise (Giannakopoulos, Mavridi, Paliouras, Papadakis & Tserpes, 2012) and used as a solution to avoid noise in key terms such as spelling errors (Kamaruddin, 2011). The set of characters in n -gram of a word comprises all substrings of length n of the original text. Compared to VSM, the frequency of n -gram is used to quantify the information. The value for n usually are 2 (bigrams), 3 (trigrams) and 4 (four-grams).

N-gram technique has been applied in several studies for topic identification study. Wang and Wang (2013) applied *N*-gram technique to represent the sentences that have the keywords and they only extract *n*-gram items with keywords from the sentences instead of taking all *n*-gram items. This study has proven that *n*-gram is able to detect and provide potential keywords from the sentence and contribute to their proposed topic detection method. Na, Cai and Zhao (2009) also made use of *n*-gram in their proposed sentiment topic method. They experimented both unigram and bigram to represent the parameters and approved that bigram model performed better than unigram model. Table 2.2 summarizes the techniques that have been reviewed for text representation.

Table 2.2

Summarization of text representation

Techniques	Advantages	Disadvantages
Vector Space Model	Simple	High dimensional features, lack of semantic correlation
<i>N</i> -gram	Noise tolerance	Fail to capture semantic if <i>n</i> size is too small and too large
TF-IDF	Easy to implement	Ignore relationship between terms

Amongst all text representations, VSM is the most applicable and simple in text representation. However, VSM comes with high dimensional features and not all features are relevant. In some situations, a text should be represented semantically to find meaningful patterns in it (Aggarwal & Zhai, 2012). The language processing methods are still not robust enough to represent the accurate semantic of texts. Therefore, the Bag of Words representation is still reliable (Aggarwal & Zhai, 2012). Meanwhile, *N*-gram is noise tolerance since it decomposes terms in units of *n* and computation is faster because *n*-gram sequences have uniform length (Kamaruddin,

2011). However, N -grams with too small n sizes and too large n sizes fail to capture the semantic similarity between terms. Based on the mentioned gaps, TF-IDF technique is chosen for text representation in this study because it is straightforward and easy to be implemented as it does not have to consider semantic relationship amongst the terms.

2.3.4 Dimensionality reduction

In text mining domain, dimensionality reduction is also referred as text pre-processing. In Information Retrieval domain and Machine Learning domain, dimensionality reduction is divided into two parts which are feature extraction and feature selection (Khan et al., 2010). However, dimensionality reduction based on Information Retrieval and Machine Learning domain carries the same goals as in text mining domain.

The major problem in text classification is to handle the high dimensionality of the text data and space features (Nuipian & Meesad, 2013; Khan et al., 2010). Most of the text domains have a number of features that are irrelevant and not beneficial for classification task (Khan et al., 2010). Apart from that, noise features may reduce the classification accuracy as well. Hence, a standard procedure to reduce the feature dimensionality is needed.

To overcome the mentioned issue, dimensionality reduction, or also known as feature reduction, is a process to avoid the overfitting problem where the classifiers fit the training data in the sense that it fails to classify new unseen data (Sebastiani, 2000). Dimensionality reduction is usually used for eliminating noisy and irrelevant terms

(Said, 2007). Based from the review of Harrag, El-Qawasmah, Al-Salman (2011), Khan et al. (2010), and Said (2007), there are two approaches that have been identified which are mainly used for dimensionality reduction, namely; feature selection and feature extraction.

Feature selection is commonly used in text classification to reduce the dimensionality of features and improve the efficiency and accuracy of classifiers. This method aims at selecting some of features or words that have the highest score according to the predetermined measure of the importance of the word (Khan et al., 2010). The selection process is performed by applying either the filter approach or the wrapper approach.

To begin with, the filtering approach is employed most of the time for feature selection stage (Uysal, S.Gunal, Ergin, E.Gunal, 2012). The approach is based on applying a scoring method to evaluate the features. The filtering approach is based from the document frequency in finding and retaining the terms that occur in the highest number of documents. The commonly used filter methods are such as document frequency, mutual information, information gain, chi-square, and Gini index (Aggarwal & Zhai, 2012; Uysal et al., 2012).

The advantages of filter approach are that it is easily scaled to high dimensional datasets, computationally simple and fast. In addition, the filter approach is independent as it only has to be performed only once (Beniwal & Arora, 2012). However, according to Sadiq and Abdullah (2013), the size of the feature space is not reduced by implementing methods in feature selection since the size of the full feature

set is reduced and time consuming. In view of this, the drawback of filter approach is that it tends to ignore the effect of the selected feature set on the classifier algorithm. To add to this matter, this is also supported by Beniwal and Arora (2012) who claimed that most of the filter approaches tend to ignore the interaction with the classifier and since the techniques are univariate, this means that each feature is considered separately.

The wrapper approach wraps the features around the classifiers to be used to anticipate the benefit of adding or removing a certain feature from the training set (Said, 2007). Unlike filter approach, wrapper selects the features that lead to an improvement in the performance of the classifier algorithm. The quality of an attribute subset is directly measured by the performance of the data mining algorithm that is applied to that attribute subset (Beniwal & Arora, 2012). This means, the wrapper approach might be much slower than the filter approach because data mining algorithm is applied to each attribute subset considered by the search. In some cases, several data mining algorithms need to be applied to the data which makes the wrapper approach become computationally expensive.

The wrapper approaches is able to include the interaction between feature subset search and model selection and considers the feature dependencies. However, is risky due to the overfitting issue (Beniwal & Arora, 2012; Silva & Ribeiro, 2010).

As an alternative for feature selection, feature extraction (FE) methods are one of the approaches in dimensionality reduction method. Feature extraction aims to transform the vector space representation of the document into one of a lower dimensionality

(Said, 2007). According to Uysal et al., (2013), stage in feature extraction utilizes the vector space model that makes use of the bag-of-words approach. Many methods have been developed such as Principal Component Analysis, Latent Semantic Indexing and Linear Discriminant Analysis.

The first technique is Principle Component Analysis (PCA). It is a statistical analysis technique for dimensionality reduction method which is designed for linear datasets (Kramer, 2013; Li & Han, 2011; Uguz, 2011). PCA is popular of its simplicity and interpretability (Kamaruddin, 2011), and due to its potential to reduce the dimension of a dataset by projecting onto a lower dimensional subspace (Debruyne & Verdonck, 2010). This linear method combines indicators that are relevant to each other to a few independent composite indicators. These indicators gain all principal messages, and then become the principal component (Li & Han, 2011).

The second feature extraction technique is Latent Semantic Indexing (LSI). This technique is based from Principle Components Analysis (PCA) which uses Singular Value Decomposition to transform the original higher dimension into a lower one where features are ranked by their importance within the document by looking at their patterns and co-occurrence (Aggarwal & Zhai, 2012; Said, 2007). The advantage of Latent Semantic Indexing is the implicit higher-order structure among the terms with documents which is the semantic structure to improve the detection of relevant documents. This technique also addresses problems such as the use of synonymous and polysemous words in the documents (Silva, 2010). However, the new dimensions representing the documents tend to be not intuitively interpretable (Silva, 2010) because this unsupervised technique is blind to the underlying class distribution which

means the features found might not belong to the right direction (Aggarwal & Zhai, 2012).

Third is the Linear Discriminant Analysis (LDA). Unlike LSI, this technique works in a supervised manner which ignores the class labels (Sharma & Kuldip, 2014). Hence, LSI is more suitable for text clustering rather than text classification. Generally, LDA technique finds an orientation that reduces the high dimensionality features vector that belongs to different classes. Conversely with LSI, LDA searches for features that best discriminate among classes. From this, LDA will construct a linear combination of these features that maximizes the margin among the desired class.

Amongst these techniques for dimensionlity reduction, feature selection has the most suitable criteria to be implemented in this study. This is because feature selection only selects the potential terms with the highest score from text instead of reducing the number of terms in text as in feature extraction. Filtering approach is selected instead of wrapper approach because filtering approach is simple, independent and can be performed only once. Meanwhile, wrapper approach is risky because it has overfitting issue.

2.3.5 Feature weighting

After features are obtained during dimensionality reduction phase, it should be weighted before being presented to the classifier. Feature weighting techniques are used in Information Retrieval to identify the most relevant terms in the documents (Xu & Xu, 2010). By computing the feature weight for each feature of document, it helps the IR system to rank the retrieved documents depending on the expected

relevance for the users, (Xu & Xu, 2010; Liu, Li, Li & Li, 2004; Moschitti & Basili, 2004). Six feature weighting techniques have been listed such as Boolean weighting, word frequency, Term Frequency-Inverse Diverse Frequency (TF-IDF), term frequency collection (TFC) and Entropy weighting. Amongst these techniques, only the most widely used methods in text classification task are discussed.

First of all, the basic weighting technique is word frequency or term frequency (TF). This method implies the weight is equal to the frequency of the feature measure on the importance levels of terms in a document (Huynh, Tran, Ma & Sharma, 2011; Liu et al., 2004). The more terms encountered in a certain context, the more they contribute to the meaning of the context. Although term frequency seems to be very intuitive, it does not consider the frequency if the feature from the documents is in the collection. For this reason, TF carries no information about the semantics of a document.

The next technique is Term Frequency-Inverse Diverse Frequency (TF-IDF). This technique is widely studied in text classification tasks. TF represents the importance of a feature in a document, and IDF represents the discrimination of a feature for all (Granitzer, 2003; Huynh et al., 2011; Syam, Fayed, Habib, 2006). According to Xu and Xu (2010), TF-IDF is borrowed from IR and IDF and was introduced to prevent retrieving most documents. The TF-IDF technique is based on statistical approach. For this reason, it does not directly reflect the term's category membership. In other words, it ignores the features that have different discrimination for distinct category labels. Another feature weighting technique is the Boolean weighting. The basic idea of Boolean weighting is that it lets the weight become 1 if the word occurs in the

document, and 0 otherwise (Syiam et al., 2006). Table 2.3 summarizes the reviewed feature weighting techniques.

Table 2.3

Summarization of feature weighting

Techniques	Advantages	Disadvantages
Term Frequency (TF)	Easy, simple	Consider the frequency of terms in the document only instead of across the collection. Ignore common and rare words.
Term Frequency-Inverse Diverse Frequency (TF-IDF)	Direct measure of term frequency in a document and also in the whole collection	Ignore the semantic value amongst terms.
Boolean weighting	Simplest	Ignore the frequency of terms

As a summary for feature weighting techniques, Term Frequency is easy and simple to be implemented. But, the limitation of Term Frequency is this technique only gives weight to represent the frequency of terms in a document without considering that the terms might also appear in the whole text collection. Unlike Term Frequency, TF-IDF can calculate both frequency of terms in a document and frequency of terms throughout the text collection as well. However, TF-IDF tends to ignore the terms semantically. As compared to Term Frequency and TF-IDF, Boolean weighting is the simplest technique but it ignores the frequency of terms as it does not distinguish the importance of different features. Therefore, TF-IDF selected in this study because it calculates the weight of the terms for both TF and IDF.

2.4 Rule-Based Classification Techniques

This section generally reviews the rule-based classification techniques that have been implemented in text classification task.

The first rule-based classification technique is Decision Tree. Decision Tree is a tree-shape structure that has been widely used for building classification models, that is easy for human to understand and criticize (Beniwal & Arora, 2012; Kotsiantis, 2013; Podgorelec & Zorman, 2009). It is a sequential model and simple in terms of understanding and interpreting, especially for novice users, as it has the tendency to base classification on as few tests as possible. Decision tree has two types of nodes which are the root and internal nodes, and the leaf nodes. Both root and internal nodes are always associated with attributes, meanwhile leaf nodes are associated with classes (Padhy, Mishra & Panigrahi, 2012). The categorization of training documents using the Decision Tree technique is by producing true or false queries in the form of a tree structure. Each of the leaves represents the category of documents; meanwhile, the branches represent the conjunctions of features that lead to those categories (Harish et al., 2010 & Khan et al., 2010). Unfortunately, Decision Tree can cause errors in classification by producing many classes (Govindagan, 2007). Besides, the process to grow the tree is computationally expensive because Decision Tree requires as many examples as possible for training purpose.

Second rule-based classification technique is Rough Set. Pawlak introduced Rough Set Theory (RST) in 1982 (Nguyen & Skowron, 2013; Li & Wang, 2004; Liu et al., 2012) as a mathematical tool for data analysis and knowledge discovery such as to analyze ambiguous data and discover hidden important facts from data (Huang, Tseng, Fan & Hsu, 2013; Li & Wang, 2004). Approximately, additional information of the data is not

required since it can work with imprecise values or uncertain data (Liu et al., 2012). Rough Set is also useful in reducing noise attributes such as redundancy and irrelevant attributes. However, Rough Set depends on the preparation of the decision table before starting the training and testing the data. The reduction is also chaotic and not stable (Bazan et al., 2000) because the extracted rules can be countless. Table 2.4 summarizes both rule-based classification techniques.

Table 2.4

Summarization of rule-based classification techniques

Techniques	Advantages	Disadvantages
Decision Tree	Sequential and simple to understand and interpret.	Produce many classes that can lead to classification error. Computationally expensive – require many examples for training. Produce large number of extracted rules.
Rough Set	Does not require additional information – can work with uncertainty.	Depends on the preparation of decision table. The reduction is chaotic and unstable. Produce large number of extracted rules.

The limitation of both Decision Tree and Rough Set is the tendency to produce large number of extracted rules, requires training and testing on data and computationally expensive. Therefore, an algorithm which is based on rule-based classifier is chosen to be employed in this study. This is to avoid error classification since the data for this study is English translated Quran. The rules can be seen as unit of knowledge as it is highly expressive and allows multiple rules to be triggered for a given record and the interpretation is understandable. The works that implement rule-based classification for topic identification are elaborated in Section 2.5.3, page 34.

2.5 Topic Identification Methods

Topic identification is a classification problem where the task is the assignment of the correct topic label (Sadiq & Abdullah, 2013; Baghdadi and Ranaivo-Malancon, 2011; Stein & Eissen, 2004; Bigi, Brun, Haton, Smayli & Zitouni, 2001). Furthermore, topic identification is also known as topic spotting or topic detection and tracking, because it can automatically sort a set of documents into categories or classes or topics from a predefined set (Sadiq & Abdullah, 2013).

The motivation of this study is based on the fact that the numbers of studies that focus on topic identification is limited (Hassan, 2013; Ozmutlu, Cavdur, Ozmutlu & Spink, 2004; Stein & Eissen, 2004). Therefore, this section describes the methods that were proposed by previous researches for the task of topic identification in text document. In the earlier works of topic identification, McDonough, Ng, Jeanrenaud, Gish, and Rohlicek (1994), outlined three essential steps of topic identification work as shown in Figure 2.1. Though their work was based on speech messages, the theory of topic identification is adaptable.

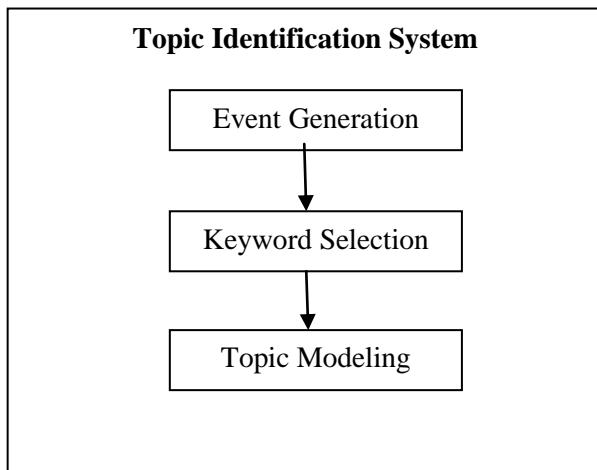


Figure 2.1. Basic topic identification system

Basically, topic identification system consists of three steps, such as Event Generation, Keyword Selection and Topic Modeling. The basic idea of Event Generation stage is to extract relevant features from the text. This is followed by Keyword Selection stage which selects the most discriminant vocabulary between different topics. The purpose of Topic Modeling is to create the probability model of the topic identification work. In this Literature Review, topic identification methods are classified into three separate groups, namely statistical-based approach, ontological approach and rule-based approach.

2.5.1 Statistical based approach

Several techniques based on statistical approach have been taken into consideration for the work of topic identification. The earliest topic identification work that was based on statistical approach is found back in 1994 by McDonough et al. (1994). This work is followed by Lin (1997) who also implements a statistical based technique, which is TF-IDF technique.

Topic Unigram Model and TF-IDF classifier are the techniques used by several researchers such as Aery et al. (2003) and Bigi et al. (2001). These methods are used to weight each unique term in a document to represent whether the topic really defines the word to its respective document. Identifying a topic based on statistics approach was also presented by Berkowitz (2005), who developed a method to identify topics of a text by counting the frequency of noun in the sentence. But he did not state any pre-processing step especially the dimensionality reduction phase.

Besides that, Van Zaanen and Kanters (2010) studied on classifying song lyrics into their mood classes. In order to find the important words in the song lyrics, they make use of TF-IDF technique to compute the relevance of the potential features. Though these researchers claimed that their study provided positive result, those approaches did not apply the text classification principle, which is to reduce the high dimensionality of the textual data at the first stage of their work.

In another research, Natarajan, Prasad, Subramaniam, Saleem and Choiet (2007) proposed to use Hidden Markov Model (HMM) based engine which is OnTopic for topic spotting from a large collection of textual document that consist of thousands of topic labels. Wang and Wang (2013) implemented statistical N -Gram model to capture the pattern of co-occurrence of contiguous words to find the importance of words for identifying topics in a blog. In another similar research, Dalal and Zaveri (2012) also implemented TF-IDF technique and combined them with Bayesian classification for automatically assigning categories for a sports blog. Though their results are positive, the proposed approaches are not suitable for larger feature sets. This is because the statistical language model requires a large amount of training data to ensure the robust estimation of the parameters (Skorkovsk'a, Ircing, Prazak & Lehecka, 2011).

2.5.2 Ontological approach

Ontology is another approach that is used in most topic identification works such as by Coursey, Mihalcea and Moen (2009), Hasan (2013), Jain and Pareek (2010), Schonhofen (2009), Stein & Eissen (2004), Syed, Finin and Joshi (2008) and Tiun, Abdullah & Kong (2001). In this approach, ontology hierarchy is implemented by

mapping the extracted keywords to the corresponding concepts in ontology for identifying major topics in textual document.

Some researchers applied WordNet as an external knowledge for classifying the ontology concept (Janik & Kochut, 2008). Tiun, Abdullah & Kong (2001) solved the vocabulary problem which is to represent all document keywords by extending the ontology through linking each concept with WordNet in order to find the semantic relationships such as synonym, hyponym and meronym. However, the mapping of the keywords is not well represented. Moreover, the common mistake of using hierarchical ontology which is to recover the wrong selection of node is not solved.

Instead of using WordNet in finding the concept for building ontology, Wikipedia also has been widely applied in several proposed ontological-based approaches for topic identification works. For example, studies by Coursey et al. (2009), Hassan (2013), Janik and Kochut (2008), Schonhofen (2009) and Syed et al. (2007) exploited Wikipedia for detecting the concept of the documents in order to develop the ontology. These studies depend on the titles or categories in Wikipedia without considering the information within the article text.

Although Wikipedia contains up-to-date information about the world, it comes with several disadvantages. Firstly, the content of Wikipedia is not consistent as the concept evolves rapidly and the articles can be changed by anyone. Secondly, the density of the category net is uneven because some topics are detailed than others. In addition to this, some categories in Wikipedia have no articles attached to them (Hasan, 2013). Thirdly, Wikipedia combines semantically unrelated concepts which

might provide error results (Schonhofen, 2009). Based from the review of previous works (Tiun et al., 2001; Jain & Pareek, 2010), ontology approach is very good in terms of semantic value within the text. However, this approach is costly because it is time consuming as it requires extra work in designing and implementing it. Once it is developed, then it has to be maintained and modified due to the conceptual changes (Syed et al., 2007).

2.5.3 Rule-based approach

Topic identification works have also been implemented using rule-based approach by performing rule-based classification technique such as Rough Set and other rule-based algorithm.

Devasenal and Hemalatha (2012) proposed a rule reduction algorithm to create token, identify feature to summarize the text and identify the topic. This work focuses on the semantic value in the sentences in order to summarize the whole text. Zhang & Zhao (2010) implement Rough Set rule generation to identify topic that relates to threat. Yeh and Chen (2007) exploited rule-based algorithm for identifying anaphora from Chinese text. In the study of Clifton and Cooley (2000) and Liu, Chin and Ng (2003), an algorithm which is based on association rules is used to identify topic of documents. In another work, Massey and Wong (2011) proposed a rule-based algorithm for topic identification which uses single terms from text and single terms extracted from Yahoo web page to determine the topics. However, the proposed method is purely based on statistical approach which did not employ any linguistic techniques such as name entity recognition and tagging.

In a different case, topic identification has also been based on clustering algorithm such as by Fuddoly, Jaafar & Zamin (2013) who make uses Bracewell's algorithm to find the similarity of keywords in order to identify topic for Indonesian news documents. Baghdadi and Ranaivo-Malancon (2011) also proposed a method to discover topic employed based on clustering algorithm. They have exploited Chen's algorithm to calculate the IDF which is a weight for each noun and verb identify topic and modified the algorithm by selecting the topic with the highest weight. Anaya-Sanchez, Pons-Porrata and Berlanga-Llavori (2008) proposed an algorithm to obtained label document cluster to identify topic of text collection. Next, an algorithm also proposed by Stoyanov and Cordie (2008) for topic identification of fine-grained opinion analysis. Butarbutar and McRoy (2004) employed indexing algorithm to find the importance of topics in the text documents. TF-IDF weighting scheme is implemented to count the frequency of topics in texts. They also exploited technique from Computational Linguitisc which is POS tagging in order to find the topic candidates syntactic parts. Though these proposed algorithms have proven good result, however; clustering approach is not chosen for this study because there are target topics that have been identified which are Marriage, Inheritance and Divorce.

There are three approaches for topic identification which are statistical, ontological and rule-based. Based from the review, rule-based topic identification is able to use set of rules in algorithm to make decision. As statistical approach is too robust and ontological approach is computationally expensive, rule-based topic identification is chosen as a suitable approach for the proposed topic identification method.

2.6 The Quran as a Case Study

The Quran is the *mukjizat* given to Prophet Muhammad (pbuh) and it was sent down to him by Angel Gabriel (Atwell et al., 2010; Sharaf & Atwell, 2009; Yauri et al., 2012). Denffer (1983) stated that the Quran was sent down by stages and not as a complete book in order to strengthen the heart of the Prophet Muhammad (pbuh) whenever the need for guidance arose. The overview of women in the Quran is presented in this section and also discussed on the existing knowledge extraction method that has been made on the Quran.

2.6.1 Women Issues in the Quran

According to the study by Ku-Mahamud et al. (2012), female is one of the popular terms in the Quran. However, verses on female in the Quran are mentioned separately within many different verses in different chapters (Sharaf et al., 2009).

Abdullah and Sudiro (2010) studied on the women issue from Surah An-Nisa' and they identified twelve issues that are related to women. Some of the issues are polygamy, dowry, women who married Christians, concubine, *mut'ah*, *talaq*, *nusyuz*, inheritance and so on. However, they only studied on theory perspectives and covered too many women issues to be explored. In addition, their study only focused on one Surah which is Surah An-Nisa', whereas there are other available Surahs which cover on women issue as well.

There are several studies on extracting knowledge from the Quran, such as Ain and Basharat (2012), Baqai, Basharat, Khalid, Hassan & Zafar (2009), Noordin and Othman (2006), Sharaf (2009), Sharaf and Atwell (2009), Yauri, Kadir, Azman, &

Murad (2013). However, studies to extract knowledge about women from the Quran are limited. There are only two studies that have been identified to extract women knowledge from the Quran, namely by Abdullah, Kassim and Saad (2009) and Ku-Mahamud et al. (2012).

2.6.2 Knowledge Extraction from the Quran

Previously, there is not much computational research on the Quran. Since the Quran has been a major source of reference for all types of problems (Ku-Mahamud et al., 2012; Mukhtar, Afzal, Majeed, 2012), the research to extract knowledge from the Quran has started to evolve. To add to this matter, Sharaf (2009) claimed that the Quran is an attractive target for finding hidden information, relationships, patterns, coincidences and associations. Due to this concern, the numbers of automatic applications are being developed to ease the retrieval knowledge from holy books (Baqai et al., 2009; Yauri et al., 2013).

As a dataset, some researchers use Arabic Quran (Al-Yahya, 2010; Sharaf, 2009; Sharaf & Atwell, 2009), English translated Quran (Ku-Mahamud et al., 2012), and Malay translated Quran (Hanum, Abu Bakar, Ismail, 2013). Each of these datasets requires different approaches due to the factor of language structure which differs with each other.

The pioneers of extracting knowledge from the Quran are Sharaf and Atwell (2009), but their main concern is to develop a knowledge representation model for the Quranic concept. Another similar research also has been proposed by Ku-Mahamud et al. (2012) who implemented semantic network method to represent verses from the Quran which are related to women issues. Meanwhile, Sharaf (2009) and Baqai et al.

(2009) focused on the annotation and analysis of the Quran based on the knowledge in the verses. Another similar research is by Al-Yahya (2010) who studied Quranic lexicons which are the nouns from the Quran. Noordin and Othman (2006) proposed a web-based system designed to retrieve texts from the Quran and knowledge or topics derived from it on the web. However, this work depends on the keywords rather than relating it with the content from the Quran.

Another method for extracting knowledge from the Quran is ontology. Saad, Salim and Zainal (2008) proposed the ontological methodology for extracting knowledge from the Quran. Ain and Basharat (2012) proposed DataQuest as a framework for modeling and retrieving knowledge. Yauri et al. (2012) used Web Ontology Language (OWL) to define the concepts and relationships. However, these methods are based from Arabic Quran domains, in which making the lexical analysis is more complicated as compared to using English translated Quran. Besides that, the development of the ontology technology is time consuming and needs further maintenance.

However, existing works on Quran extraction focus on interpreting the meanings rather than classifying them into specific topics. Based on this gap, this study aims to explore the specific topics of female in the Quran, namely inheritance, divorce and marriage. Topic identification method will be used in order to identify topic of each verse.

2.6 Summary

As a conclusion, it is necessary to follow each of the phases of text classification. The phase starts with text pre-processing which helps to convert raw text into a more understandable form by the classifier. During text pre-processing, the unstructured form of textual data has been converted into a structured textual data.

Dimensionality reduction consists of two approaches, namely feature selection and feature extraction. Feature selection aims to find relevant features by selecting the highest score of terms. Feature selection can be implemented either by filter or wrapper approaches. Meanwhile, feature extraction is based on combining data to make smaller sets of features. In order to find the importance of terms or features in texts, feature weighting is essential for this task.

Topic identification methods can be classified into three groups, i.e. statistical approach, ontological approach and hybrid approach. There are limited studies to extract knowledge and identify topics from the Quran, especially topics relating to female. Since most of the studies focused on interpreting the Quran, this study aims to classify the existing English translated Quran and identify its potential topics, especially those related with inheritance, marriage and divorce.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the phases required for achieving all objectives as stated in Chapter 1. The research framework is presented in Section 3.2. This section is broken down into subsections that describe on the phases involved in the proposed topic identification method. This chapter ends with a summary presented in Section 3.3.

3.2 Research Framework

There are four phases in the proposed topic identification method. The first phase is text pre processing which aims to clean the text from unimportant terms and term extraction which is to extract the important terms from the cleaned text. Filtering algorithm is proposed in term extraction phase to ensure that the deletion of important terms will not occur. The second phase is term ranking, which is to discover the meaningful terms amongst the extracted terms. The third phase is rule generation. An algorithm for rule generation to identify topics is also proposed. The fourth phase is the evaluation to test the proposed filtering algorithm and rule generation algorithm. The overall research framework is pictured in Figure 3.1.

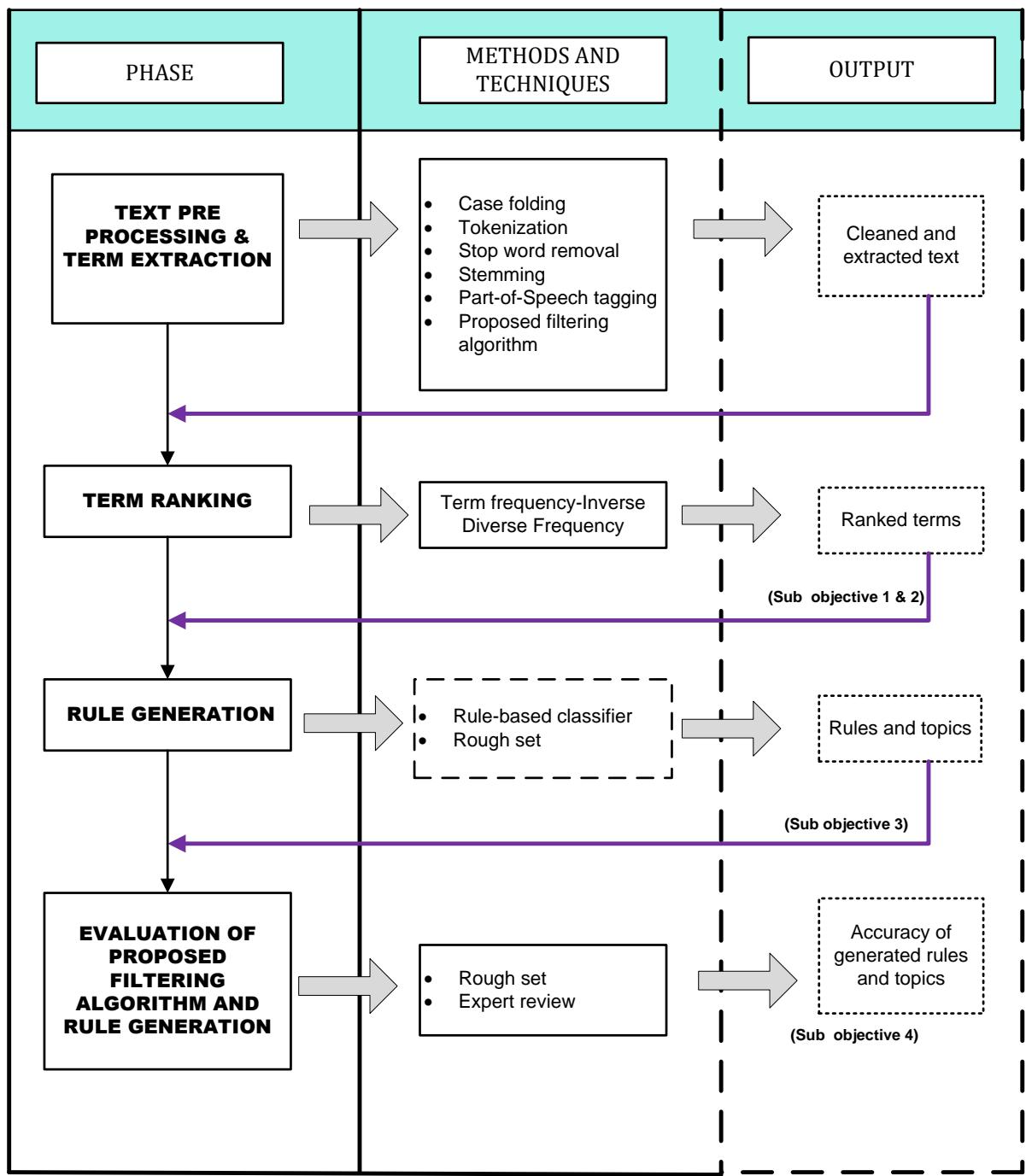


Figure 3.1. Research framework

3.2.1 Phase one: Text pre processing and term extraction

Text pre processing is the initial step in any text mining task such as text classification or topic identification. This phase is mainly performed to remove features that are unimportant for topic identification purposes and present text documents into a clear word format. Figure 3.2 shows the processes in text pre-processing phase and term extraction phase.

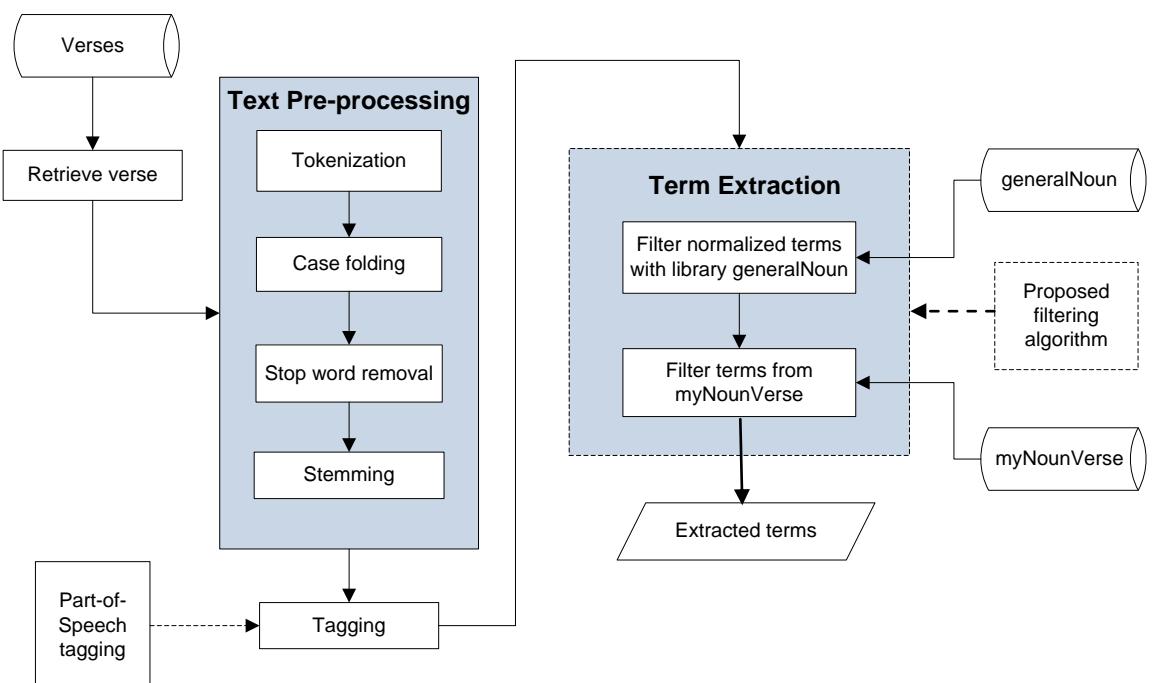


Figure 3.2. Phase 1: Text pre processing and term extraction

In this study, an English translated Quran is retrieved from Surah.my (<http://www.surah.my/>) and computerized into Microsoft Excel and has been used as a dataset. Texts from Surah.my website are chosen since traffic report from bizinformation.com.my shows that the source is frequently accessed and most referable in Malaysia (Ku-Mahamud et al., 2012). In this study, the technique that has been used is adapted from Manne and Fatima (2012), Mahar, Shaikh and Memon, (2012), and Nomponkrang and Woraratpanya (2010).

The first stage in text pre processing phase is tokenization where text is broken up into smaller and meaningful units known as tokens before any language processing can be performed.

Second, all terms in the sentences are converted to lowercase by using case folding step. Case folding is used to avoid the same term to be counted as a different meaning. For instance, term ‘Mother’ and ‘mother’, which carries the same meaning but the first term contains a capital letter and the other term is in lowercase. Hence, case folding is needed in order to solve this problem.

Third stage is to eliminate the noise words in the selected text by using stop word removal. The stop words such as ‘the’, ‘a’, ‘and’ frequently occur and they are assumed as the insignificant words needed to be removed because it is not useful for classification (Dalal, 2011). Next stage is stemming which is to convert different word forms into similar canonical forms, or in other words, it is the process of conflating tokens to their root form.

Once the root word is achieved, the next stage is to tag each of the term into specific values using Part-of-Speech Tagging. POS tagging is the process of assigning grammatical value for each word in the sentence. The purpose of using POS tagging is to identify the potential terms from the text, especially nouns. The expected outcome from this stage is the cleaned texts that are free from any noise. Examples of POS tags are shown in Table 3.1.

Table 3.1

Sample of Part-of-Speech tag set

Tag	Set
AV0 – General adverb	carefully, lovely, intentionally
AVQ – <i>wh</i> -adverb	when, why, how, wherever
NN – Common noun	man, woman, girl, mother, town
NP – Proper noun	Yusuf, Mariam, Miss, Mother

In term extraction, the tagged terms are filtered. Though only noun terms are taken, there are several exclusive noun terms that have been determined as keywords, such as ‘wed’ (verb), ‘marry’ (verb) and ‘will’ (future tense). Some of single terms carry similar meaning and eliminating these terms should be avoided. Therefore, the filtering algorithm is designed in term extraction phase to solve the synonymy of the extracted terms and these important terms are not eliminated during the extraction process.

There are two libraries of terms developed for term extraction stage. The first library is generalNoun that contains collection of nouns and compared with the terms extracted. The second library is called as myNounVerse which consists of important terms that relate to the target topics. In the first filtration, all terms are compared and checked with generalNoun library in order to ensure that all terms are not mistagged. The second filtration filters and removes those terms that do not belong to myNounVerse and not tagged as nouns. Only matched nouns are listed for further process.

3.2.2 Phase two: Term ranking

Term ranking phase processes the extracted terms in order to identify the respected topics of the text. Accurate topics can be found by looking at the number of terms related to female terms in the dataset. This phase applies term weighting which is known as Term Frequency-Inverse Diverse Frequency technique (*tf-idf*). The process for the ranking step is illustrated in Figure 3.3.

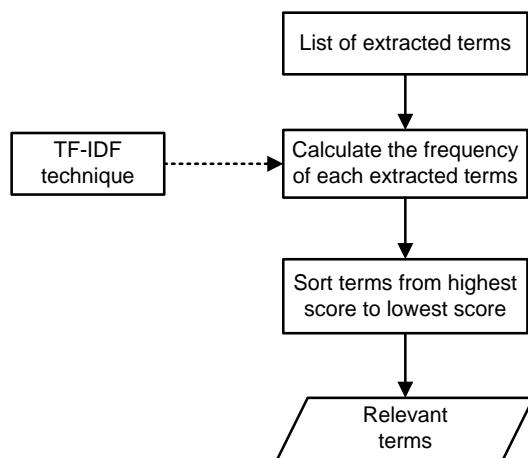


Figure 3.3. Phase Two: Term ranking

The results obtained from the term weighting steps determine which term has the potential to become relevant terms for further steps which include constructing decision rules for topic identification.

3.2.3 Phase three: Rule generation

The goal of this study is to assist human understanding on a good solution to identify topics in a way that makes sense to a person. From the highest ranked terms, a rule can determine whether that document belongs to which topic, normally only one specific topic.

Therefore, the proposed rule generation algorithm is employed to identify topics based on the output obtained from the ranking phase which is the relevant terms. Rough set technique is also proposed to be implemented in this study because this technique is not only able to reduce attributes, but it is also able to generate decision rules. In fact, Rough set technique is able to analyze vague data and discover decision rules from the training dataset. The important part in conducting Rough set technique is to select the appropriate number of attributes and which attributes to select.

3.2.4 Phase four: Evaluation

The purpose of evaluation is to measure the accuracy of the topics produced by the proposed topic identification method. The produced topics are evaluated to determine that they are reasonable and represent the topic of the sentences correctly. The evaluation process to measure topic identification also involves expert opinion since the data for this study is the English translated Quran. Both the experts in Islamic knowledge and text mining domain validate whether the produced rules and topics are correct and reasonable. The proposed filtering algorithm is evaluated with another two filtering techniques, which are Rough Set Attribute Reduction technique and Information Retrieval technique. The accuracy of topics identified by the proposed rule generation algorithm is compared with Rule Generation Rough Set technique and Expert Opinion.

3.3 Summary

As a conclusion, the overall planning in the whole research framework is needed to be followed in achieving all research objectives. Each of the phases depends on one another. The input for text pre processing is the raw texts that are later extracted by

implementing the proposed filtering algorithm. The expected output from this phase is the extracted terms. The second phase which is term ranking aims to find the most relevant terms from the collection of the extracted terms. Based on the highest ranked terms as an input, a rule generation algorithm is constructed for topic identification. The last phase is the evaluation of the produced rules and topics, which aims to find the accuracy of the topics. The techniques involved in each of the phases are provided in this chapter.

CHAPTER FOUR

TOPIC IDENTIFICATION METHOD

4.1 Introduction

This chapter discusses deeply on the proposed topic identification method for textual document. The phases involved in the proposed topic identification method are explained in details in Section 4.2. The experiment for text pre processing and term extraction is discussed in Section 4.3. The proposed filtering algorithm (PFA) is also presented in the term extraction section. Next, Section 4.4 presents the term ranking. The proposed rule generation algorithm (TopId) is provided in Section 4.5. The implementation of Rough Set technique for rule generation is also presented in this section. This chapter ends with a summary in Section 4.6.

4.2 The Proposed Topic Identification Method

This section introduces the description of the proposed topic identification method and the importance of each component. The discussion on how each of the components intends to provide the required contribution in topic identification task is also provided. In the context of this study, the topic identification method aims to provide relevant terms that indicate topics for each verse. A decision rule is also designed in order to allow the TopId to assign a suitable topic for each verse.

Figure 4.1 illustrates the proposed topic identification method. There are three designated phases in this proposed method and each phase is carried out one after another to achieve the objectives of this study.

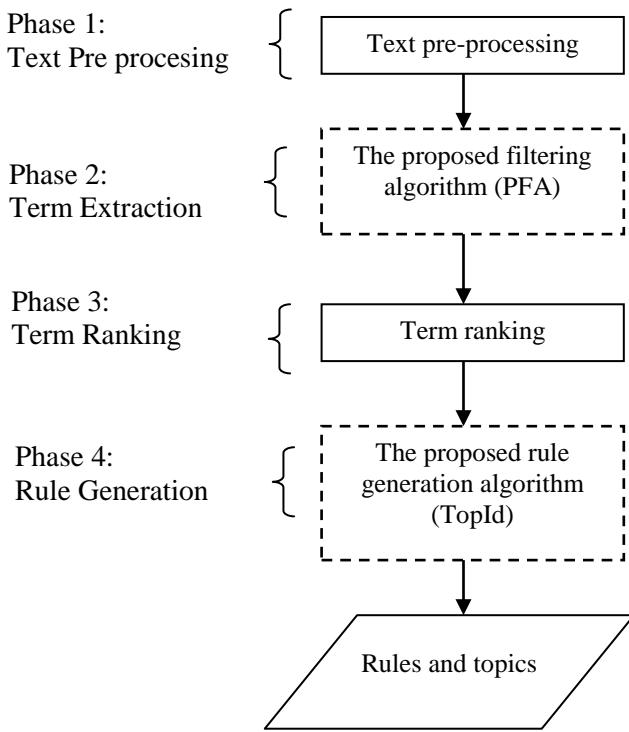


Figure 4.1. The proposed topic identification method

The proposed topic identification method is different from the other methods since it employs the PFA to solve synonymy amongst the extracted terms and to ensure the right terms are extracted. The reason PFA is introduced is because this study uses 224 English translated Quran verses and aims to identify topics for each verse. By considering the Quran as a holy book of Islam, the PFA ensures that the choice of relevant terms can produce suitable topics for the selected verses.

The method starts with Text pre-processing. Text contain a variety of non-standard token types such as digit sequences, words, acronyms, letter sequences in all capitals, mixed case words, abbreviations and etc. Hence, the process to convert this unstructured text from the raw file into more meaningful units is the most crucial task in the proposed method. Text pre-processing is necessary to reduce the high

dimensionality problem of processing textual data. With this phase, the large volume of textual documents is filtered to facilitate the searching for the relevant information.

In term extraction phase, terms which are nouns are taken as relevant terms and the other terms are categorized as noise word and irrelevant terms. Without this phase, it is costly for the next phase to rank relevant terms according to its importance. To add to this matter, too many irrelevant terms might affect the accuracy of the ranking results and determination of topic. A filtering technique from dimensionality reduction approach was adopted where the specific extraction rules are created. The expected outcome resulting from this phase is the extracted terms which are called relevant terms.

The term ranking phase is to rank the relevant features. Ranking technique is effective in Information Retrieval due to the results being ranked based on the co-occurrence of the terms. The purpose of this phase is to measure the importance of terms in a document. The important terms are calculated based on the most repetitive words with high scores.

The last phase is the rule generation, where an algorithm is designed for identifying topics based on the ranked relevant terms. This algorithm is based on the rule-based classifier which classifies records using a collection of IF-THEN rules.

4.3 Text Pre-processing and Term Extraction

There are various available methods that have been introduced and developed in order to extract and filter valuable information from texts. The Computational Linguistic method integrated with rule based process is chosen because it is capable to produce promising results compared to automated shallow methods such as statistical based approach alone.

The Computational Linguistic methods are usually employed in most systems for the purpose to extract terms meticulously and also to incorporate the business event. For instance, if certain text fragments needed to be identified individually or together with its contextual information, composite computational linguistics methods are required. In conducting this study, relevant terms were extracted from the English translated Quran by adapting the aforementioned methods.

Figure 4.2 shows the flowchart of text pre-processing and term extraction, and Figure 4.3 shows the pseudo code of text pre-processing and term extraction. The details processes in text pre-processing and term extraction are described in the following sub sections.

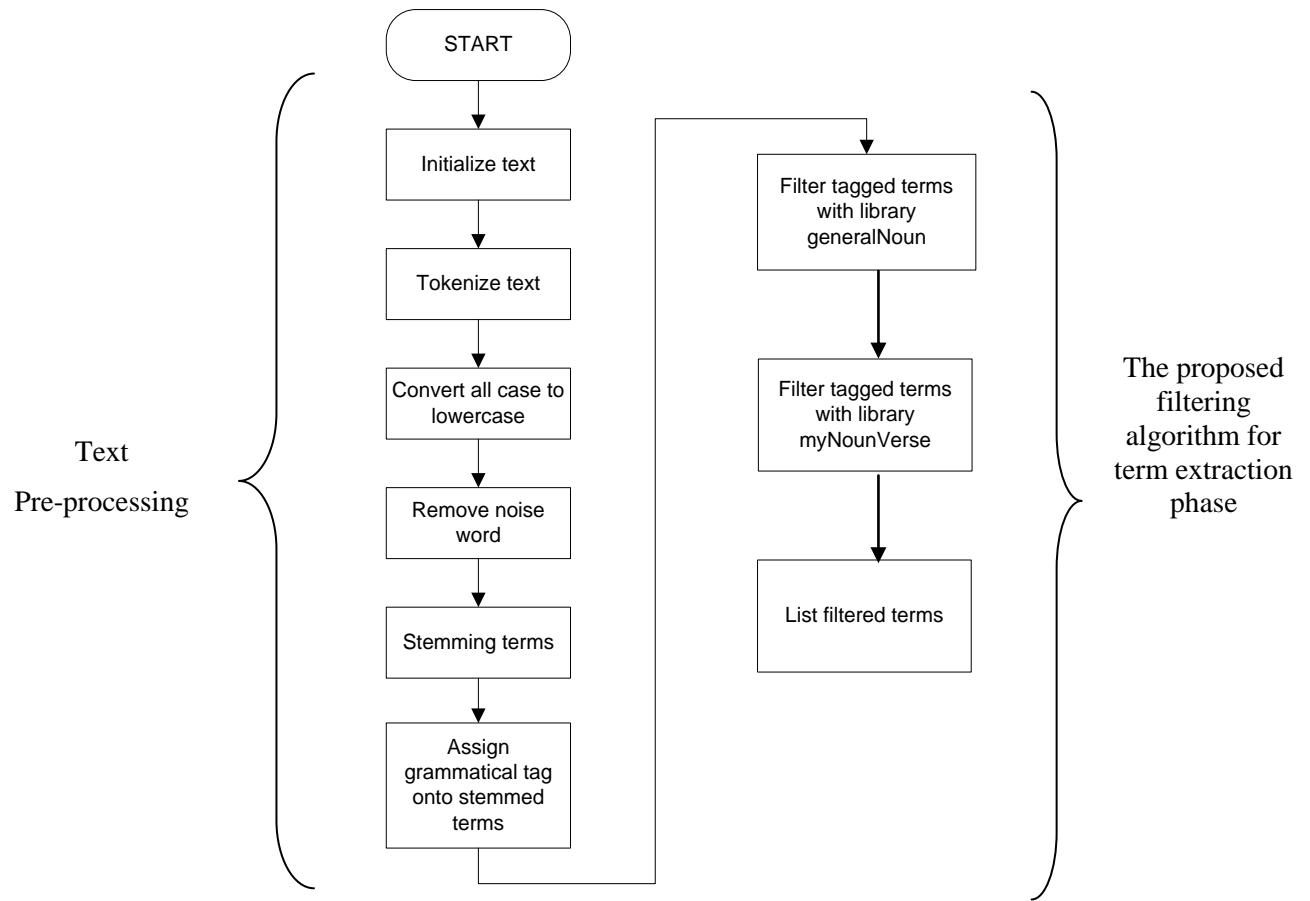


Figure 4.2. The flowchart of text pre-processing and term extraction

```



---


Start
  Execute text pre-processing
    Initialize verse
    Tokenize verse
    Transform all case to lowercase
    Remove noise word using stop word removal
    Perform stemming verse
    Assign grammatical tag using POS tagging to each term
  End text pre-processing
  Start term extraction
    For all tagged terms
      Filter tagged terms with library generalNoun
      Filter tagged terms with library myNounVerse
      List matched terms from filtration
    End for
  End term extraction
End

```

Figure 4.3. The pseudo code of text pre-processing and term extraction

4.3.1 Text pre-processing

The dataset used for the experiment is 224 English translated Quran verses. Only selected verses from the Quran are taken. Not all chapters are involved in this study and the selection of sample is purposeful since an information rich sample needs to be chosen, particularly the one that is known to have topics which are related to female, especially the topics of inheritance, marriage and divorce. The document prepared for the next step in topic identification is represented by a great amount of features. The selection of this amount of verses is due to the scope of this study, which is to identify topics that are related to feminism. The verses used are presented in Appendix A.

In the first step, the text needs to be cleaned and split into tokens. Here, ‘clean’ means the extraction of a text that has content and the removal of auxiliary in the text such as punctuation and numbers. For the execution of text pre-processing, processes such as

tokenizing, case folding and removing noise words were carried out by applying Python code using Natural Language Toolkit (NLTK) application. In order to prevent spending too much time and effort on developing the system from scratch, a commercial integrated development environment (IDE) for text analysis such as NLTK stemming demo, Word Write for word counter has also been used. The algorithm for text pre-processing is presented in Figure 4.4.

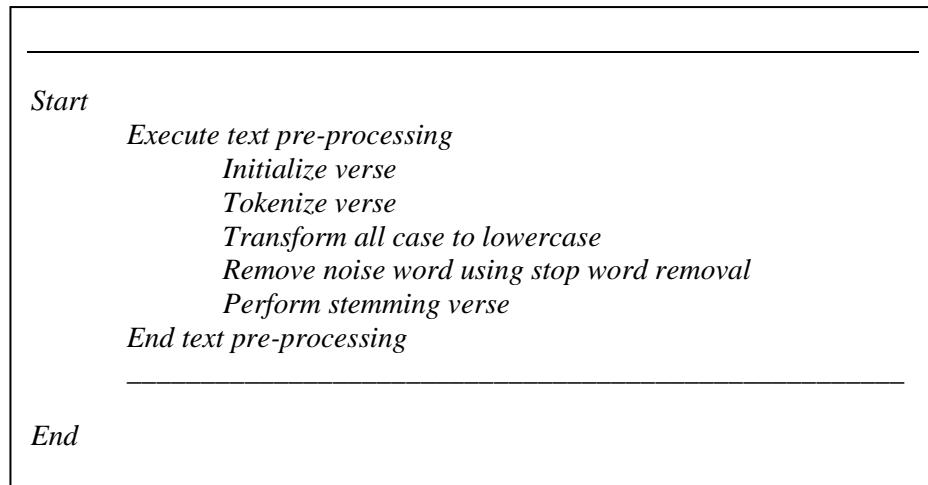


Figure 4.4. The pseudo code of text pre-processing

Next is separating or tokenizing task, which the most fundamental task in processing and analyzing textual data. Tokenizing is the process to split the raw text into individual units called tokens. The tokenizing task is performed by implementing tokenizer function in NLTK tool. This is a standard pass; therefore, no rules are required to be generated. In this phase, the raw text is chunked into text units of alphabetic, numeric, and punctuation. Figure 4.5 shows the sample result after the text has been tokenized. All terms in the sentence are separated by whitespace and comma.

```

>>> from nltk import word_tokenize, wordpunct_tokenize
>>> s = ("And give the women (on marriage) their dower as a free gift; but if they, of their own good
pleasure, remit any part of it to you, Take it and enjoy it with right good cheer.")
>>>
>>>
>>> word_tokenize(s)
[['And', 'give', 'the', 'women', '(', 'on', 'marriage', ')', 'their', 'dower', 'as', 'a', 'free', 'gift',
',', 'but', 'if', 'they', ',', 'of', 'their', 'own', 'good', 'pleasure', ',', 'remit', 'any', 'part',
',', 'of', 'it', 'to', 'you', ',', 'Take', 'it', 'and', 'enjoy', 'it', 'with', 'right', 'good', 'cheer',
',']]

```

Figure 4.5. Sample of tokenized text

Case folding is the process of removing differences between capital and lowercase words and intends to avoid the same words to be counted as different words. This operation is essential on the account that it is necessary to ensure the exact number of repeated terms in the text.

Common words like ‘a’, ‘the’, ‘an’ and ‘this’ are often filtered out to improve the performance, and these words are also considered as noise words. There are two approaches to perform this process. One basic approach is to remove all words that appear frequently in the document. The repeated words may create ‘noise’ that make records less distinguishable. Thus, this approach is not applicable in this study since the high frequency of word appearance could be the best term for the target topic. Hence, the second approach is chosen which is to expel all less relevant words in the list of noise words without losing other important words. The list of noise words can be changed according to the context and accessibility as it is distributed as a free open source material on websites.

After the text has been tokenized and converted to lowercase, Python code is implemented to remove noise words from the sentence. Stop words collection used for this experiment is from NLTK Corpus Stopword by Porter et al compiler. This

corpus contains 2,400 stopwords for 11 languages. The execution process for case folding and stop word removal is depicted in Figure 4.6. The list of noise words is presented in Figure 4.7.

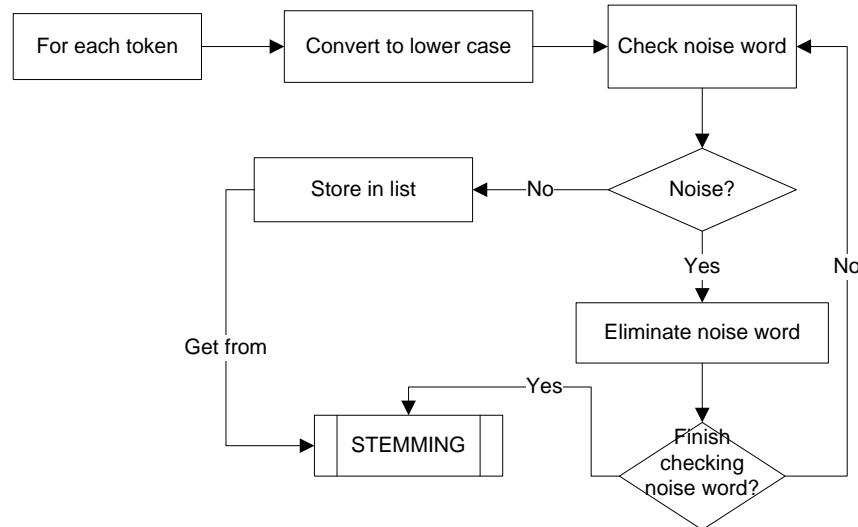


Figure 4.6. Flowchart of case folding and stop word removal process

about, after, all, also, an, and, another, any, are, as, at, be, because, been, before, being, between, both, but, by, came, can, come, could, did, do, each, for, from, get, got, has, had, he, have, her, here, him, himself, his, how, if, in, into, is, it, like, make, many, me, might, more, most, much, must, my, never, now, of, on, only, or, other, our, out, over, said, same, see, should, since, some, still, such, take, than, that, the, their, them, then, there, these, they, this, those, through, to, too, under, up, very, was, way, we, well, were, what, where, which, while, who, with, would, you, your, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, \$, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0,

Figure 4.7. List of noise words

Stemming function should be able to convert all words into their root word. In the first attempt to use stemming function, Porter algorithm is chosen. After a few testings, this algorithm seems to produce inaccurate results. An important word such as ‘married’ is converted as ‘marrie’ instead of ‘marry’. Besides that, Porter stemming

is also unable to convert plural words into their singular words, for instance ‘women’ to ‘woman’. In such situation, this study does not have any employment to enhance stemming algorithm. Hence, another stemmer is occupied. An online application called NLTK Stemming Demo (<http://text-processing.com/demo/stem/>) is used for this experiment. WordNet Lemmatizer stemming is able to stem each words correctly.

Figure 4.8 shows the application for stemming.

Stemming and Lemmatization with Python NLTK

This is a demonstration of stemming and lemmatization for the 17 languages supported by the [NLTK 2.0.4 stem](#) package.

Stem Text

Choose stemmer

WordNet Lemmatizer ▾

Enter text

```
give women on marriage dower free gift  
they good pleasure remit part you take  
enjoy right good cheer
```

Enter up to 50000 characters

Stem

Stemmed Text

```
give woman on marriage dower free gift they good pleasure  
remit part you take enjoy right good cheer
```

Figure 4.8. Stemming using NLTK Demo

This is followed by the Part-of-speech tagging, also known as POS tagging. This process reads a document object as input and filters the tokens for that document based on part-of-speech tag information. A document object is taken as input. The token list is retrieved from the document and only those tokens with part-of-speech tags that match at least one value in the selected tag list are retained. In this study, only words tagged with noun (NN) are taken.

For this experiment, several approaches to assign grammatical tags are tested using Java code, Python code, as well as PHP code. However, these attempts are disturbed by errors and bugs in the code. Due to time constraint, an online tagging application called CST Online Tools (<http://cst.dk/tools/index.php#output>) is employed. This application is chosen because it implemented Brill tagger which is widely used in tagging tasks. The processes to assign POS tagging are presented in Figure 4.9, whereas Figure 4.10 shows the output.

The screenshot shows the 'CST's online tools' interface. At the top, there is a header and a brief description: 'Here you can analyse text with a combination of CST's tools. All tools support Danish and some tools also support other languages.' Below this, there are two input fields: 'Language:' with a dropdown menu set to 'English' and 'Bonus code:' with an empty input field. A text area labeled 'Write a few lines ...' contains the input text: 'give woman on marriage dower free gift they good pleasure
remit part you take enjoy right good cheer'. Below the text area, there is a note: '... or specify a text or RTF file. Or view a **demo text!**' followed by a 'Browse...' button and the message 'No file selected.'

Figure 4.9. The interface of tagging application

The screenshot shows the 'Output' section of the application. It displays the input text with parts-of-speech (POS) tags added: 'give/VB woman/NN on/IN marriage/NN dower/NN free/JJ gift/NN they/PRP good/JJ pleasure/NN remit/NN part/NN you/PRP take/VBP enjoy/VB right/JJ good/JJ cheer/NN'.

Figure 4.10. The produced tagging output

4.3.2 Term extraction

In many text mining tasks, the proper identification and extraction of text units may significantly influence the usefulness of the final results of the analysis, where particularly nouns play an essential role. Having only nouns appearing in a text, one can guess a topic discussed in the text. In the classification tasks, a single noun could be even better attributed than a group of several words belonging to other parts of speech (Protaziuk, Kryszkiewicz, Rybinski, & Delteil, 2007). Hence, term extraction phase aims to extract useful terms which are nouns closely related to the scope of this study.

There are many filtering techniques introduced which make use of WordNet as the database for finding terms. However, those techniques may not be applicable for this study because some important terms such as term ‘*iddat*’, ‘*zihar*’ and ‘*talaq*’ are not available from any dictionary especially Wordnet. This is why the PFA is proposed in order to overcome the problem. Figure 4.11 shows the proposed filtering algorithm to extract relevant terms from the normalized texts.

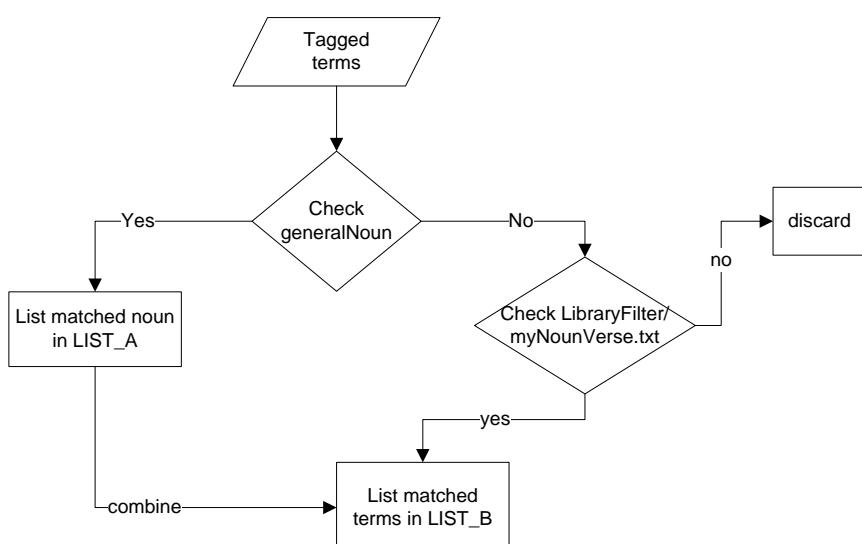


Figure 4.11. Flowchart of the proposed filtering algorithm

This algorithm begins right after the process of assigning POS tagging is finished. Besides, the algorithm allows checking the availability of the matching terms in order to find the most precise terms. From the previous processes, a list of tagged terms is produced and the focus is to collect only noun terms. However, there are important terms which also consist of verb values, such as ‘marry’ and ‘wed’. Therefore, there are two parts of filtering the terms in order to avoid missing important terms. Firstly, the tagged terms are filtered by checking the matching terms with keywords from the library which are library generalNoun and library myNounVerse.

In the first filtering part, the algorithm checks the tagged terms from library generalNoun. Matched terms which are nouns are listed in List A. In the second part, the unmatched terms will be checked again with library myNounVerse. Matched terms from this process are listed in List B and combined with List A. All unmatched terms are discarded. The terms are later ranked in the next phase.

4.4 Term Ranking

Any given textual dataset uses i unique terms. A document can be represented by a vector($t_1, t_2, t_3 \dots \dots \dots, t_n$), where t_1 has a value of 1 if the term i is present, and 0 if term i absent in the document. A term can be assigned a weight to express its importance for a particular document. *Tf-idf* technique is employed to assigns weight to a term based on how frequent the term occurs in the document.

Equation 4.1 shows two important components for calculating *tf-idf*. First is the term frequency (TF) which is to count the number of times a word appears in a document, divided by the total number of words in that document. The length of certain

documents may vary; hence, it is possible that a term would appear more times in long documents than shorter ones. Therefore, term frequency is divided by the document length which is the total number of terms in the document.

$$TF(t) = \frac{\text{(Number of times term } t \text{ appears in a document)}}{\text{(Total number of terms in the document)}} \quad (4.1)$$

Second is inverse document frequency (IDF) which is to compute the importance of a term. During the calculation of tf , all terms are treated as equally important. However, certain terms may appear many times but have insufficient importance or no discriminating power in determining the relevance. Thus, the frequent terms need to be decreased and the rare ones to be scaled up by following Equation 4.2.

$$IDF(t) = \log_e \frac{\text{(Total number of documents)}}{\text{(Number of documents with term } t \text{ in it)}} \quad (4.2)$$

The previous stage has produced a list of filtered terms and the weights are assigned to each term with its frequency score. From the ranking list of results, terms with the highest frequency score are taken. Table 4.1 shows the sample of the relevant terms obtained after the term extraction is finished.

Table 4.1

Sample of relevant terms

Surah_Verse No	Verse	Relevant terms
2_237	<p>And if ye divorce them before consummation, but after the fixation of a dower for them, then the half of the dower (Is due to them), unless they remit it or (the man's half) is remitted by him in whose hands is the marriage tie; and the remission (of the man's half) is the nearest to righteousness. And do not forget Liberality between yourselves. For Allah sees well all that ye do.</p>	<div style="display: flex; justify-content: space-between;"> <div>divorce</div> <div>consummation</div> </div> <div style="display: flex; justify-content: space-between;"> <div>fixation</div> <div>dower</div> </div> <div style="display: flex; justify-content: space-between;"> <div>dower</div> <div>man</div> </div> <div style="display: flex; justify-content: space-between;"> <div>hand</div> <div>marriage</div> </div> <div style="display: flex; justify-content: space-between;"> <div>tie</div> <div>remission</div> </div> <div style="display: flex; justify-content: space-between;"> <div>man</div> <div>righteousness</div> </div> <div style="display: flex; justify-content: space-between;"> <div>liberality</div> <div></div> </div>

The next step is to calculate the weight of each term. The documents are generally identified by sets of terms that are used to represent their contents. Therefore, term frequency technique is implemented because it is a common practice in Information Retrieval to provide relevance feedback based on the most important terms in the document or text. Table 4.2 shows the sample of text data and how each term is ranked based on its score. The verse number is assigned as document 2_237. Each term is weighted by the total number of times it appears in the record.

Table 4.2

Sample of ranked terms and tf, idf & tf - idf score

Verse (d)	Total terms in d	Terms (t)	#t occurs in d	tf	idf	tfidf
2_237	14	dower	2	0.1429	0.2350	0.0336
		man	2	0.1429	0.0979	0.0140
		divorce	1	0.0714	0.1567	0.0112
		marriage	1	0.0714	0.2938	0.0210
		tie	1	0.0714	1.1751	0.0839
		fixation	1	0.0714	1.1751	0.0839
		consummation	1	0.0714	1.1751	0.0839
		hand	1	0.0714	0.1382	0.0099
		remission	1	0.0714	1.1751	0.0839
		righteousness	1	0.0714	0.4700	0.0336
		liberality	1	0.0714	2.3502	0.1679

As referred in Table 4.2, the most relevant terms in the verse are top-ranked based on its score. For example, ‘dower’ and ‘man’ appear twice in Verse 2_237 and sorted in ascending order. It follows by ‘divorce’, ‘marriage’ and ‘tie’ which appear only once in Verse 2_237. The other terms are also counted, but they are listed separately by putting them into different lines because it is considered as less important in the verse and those terms are not listed in the keyword database.

At this stage, each term in the verse must be calculated using *tf-idf* formula. For example, to calculate the *tf* for term ‘dower’, the total number ‘dower’ occurs in the document is 2 and divided by the total number of all terms in the document which is 14. The score is 0.1429. To calculate the *idf*, there are 224 documents and the term ‘dower’ occurs only fifteen times in these documents. Then, the *idf* is calculated as $\log (224/15)$ and equivalent to 0.2350. Next, the *tfidf* is computed as 0.01429 multiply

with 0.2350 and is equivalent to 0.0336. This calculation process is applied to the whole 224 documents. There is no threshold to determine the value of score for the relevant terms. The term with highest score is considered as the most relevant term.

4.5 The Proposed Rule Generation Algorithm (TopId)

An algorithm for topic identification is proposed for this study which is also known as TopId. The input needed by TopId is the highest ranked relevant term from the previous phase. Figure 4.12 presents the designated algorithm for topic identification.

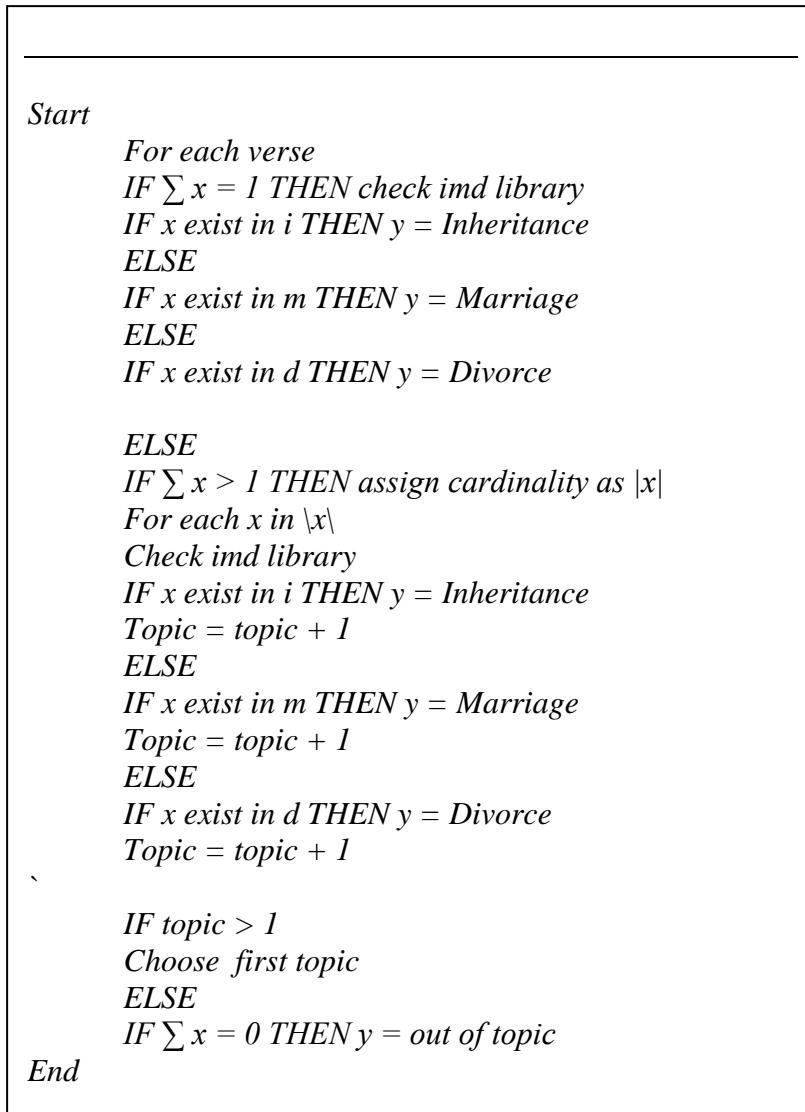


Figure 4.12. The proposed rule generation algorithm (TopId)

Assume x is the term that has the highest frequency score and y is the target topic. There are two conditions based from the number of x . First, if x equal to 1, then it will directly check to match the term with keywords in the library IMD. If the term is matched with keywords in library I, then the topic is Inheritance. If the term is matched with keywords in library M, then the topic is Marriage. Lastly, if the term is matched with keywords in library D, then the topic is Divorce.

Otherwise, in the case where the number of terms with the similar highest frequency score is more than one, then each found term is checked from the library one after another. The number of found terms is classified as cardinality of x ($|x|$). The process is continued until the condition of $|x|$ is equal to zero. If two terms have same score, TopId takes the first topic. Meanwhile, if there is no match between terms and keywords, it is considered as out of topic.

The output from TopId is the rules and the topic of each verse. The condition of the decision rules depends on the availability of certain terms from the keywords database. The rule-based classifier is used to determine the term patterns which are related to the different classes. The produced rules are represented in First Order Predicate Logic (FOPL). For instance, the rule for verse 2_49 is:

$$\forall X \text{term}(son) \wedge \text{belongs_to}(son, \text{libraryMarriage}) \rightarrow \text{is_topic}(son, \text{Marriage})$$

In the set of rule, the left-hand side corresponds to a term pattern, and the right-hand side corresponds to a class label. This rule is used for the purpose of classification. Rule-based classifier is employed in this study because it is highly expressive as it is

almost equivalent to decision tree. Rule-based classifier also allows multiple rules to be triggered for a given record and the interpretation is understandable. The quality of the classification rules can be later evaluated using accuracy measurement based on the produced topics.

In order to compare the rules and topics produced by TopId, an experiment involving Rough Set technique and expert opinions are also employed. Section 4.5.1 presents the experiment using Rough Set technique to produce rules for topic identification, meanwhile Section 4.5.2 presents the comparison of topics based on expert opinions.

4.5.1 Rule generation using Rough Set technique

For the intention to implement Rough set technique in this experiment, a tool known as Rosetta is employed. Rosetta system is a software package completed with algorithms such as discretization, reduct computation and classifier evaluation. In rule generation phase, rough set technique is implemented to generate decision rules for topic identification. Rough set is based on rule learning compress tabular data into IF..THEN rules. The IF part of each rule specifies a minimal pattern needed to detect observations with different labels.

Rough Set technique has been implemented to produce rules and topics for 224 verses. Figure 4.13 illustrates how topic identification task is implemented based on the general scheme of Rough Set technique.

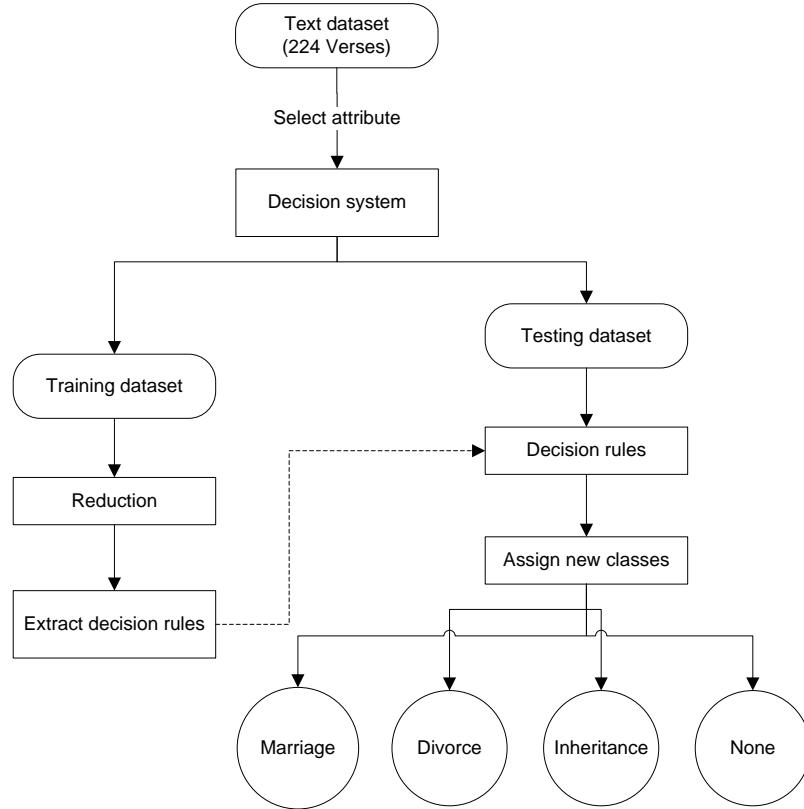


Figure 4.13. The general scheme of Rough Set technique for topic identification

In the general scheme of Rough Set technique, the data is stored in a form of decision table called information system. The decision table should be split into two parts. The first part is known as training dataset, and the second part is the testing dataset. In the decision systems, every object associates with some information set, in which every column is labeled as attributes and the last column is the class or also known as decision. Table 4.3 is the sample of decision table used for this study to perform an experiment using Rough Set technique.

Table 4.3

Sample of decision table used for training with Rough Set technique

Verse	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	Decision Class
2_35	LOW	NULL	NULL	LOW	NULL	NULL	NULL	NULL	NULL	NULL	MARRIAGE
2_49	NULL	NULL	NULL	LOW	MED	NULL	NULL	NULL	NULL	NULL	MARRIAGE
2_102	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	MARRIAGE
2_102	HIGH	HIGH	LOW	NULL	NULL	NULL	NULL	NULL	NULL	NULL	INHERITANCE
2_178	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	LOW	LOW	DIVORCE
2_187	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NONE
2_221	NULL	NULL	NULL	HIGH	NULL	NULL	NULL	NULL	LOW	LOW	DIVORCE
2_222	NULL	MED	NULL	NULL	NULL	MED	NULL	NULL	NULL	NULL	INHERITANCE
2_223	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	MARRIAGE
2_226	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NONE

The first column is the object which is the decision attribute that represents the documents. In this experiment, the documents are referred to as the verses. In every row in the decision condition attributes are $\{t1, t2, t3 \dots t129\}$. t represents the terms and there are 129 identified terms as a keyword and each of these terms is numbered. However, the actual number of attributes in the decision table is too big; thus Table 4.3 only shows a part of the decision table. The last column is the decision classes which represent the topic.

As Rough Set technique requires discrete features, the real valued attribute features are discretized first. Four associated conditions for the attributes are used for the data discretization; namely LOW, MED, HIGH and NULL. These associated conditions value are discretized manually to each object. For example, T1 appears one time in the object. Thus, the associated condition value for T1 is 1 and discretized as LOW. The range for the associated conditions is shown in Table 4.4.

Table 4.4

Data discretization

Range	Category
0	NULL
1 – 2	LOW
3 – 4	MEDIUM
5 – 10	HIGH

In the decision table, there are 224 numbers of records included. Each of these records contains 224 objects that represent the verses or also known as objects in Rough Set technique. There are 130 attributes for each object. 129 attributes represent the existing terms in the verses and the last attribute represents the topic class. The full listing of 129 attributes with their terms and reference as keywords are presented in Appendix B.

The first step implemented is to split the decision table. In this step, the data is divided into two sets, train data and test data using the provided splitting function provided in the Rosetta application. The split factors that have been used for the experiments are 0.2, 0.3, 0.7 and 0.8. The reason for performing experiments on four different splitting factors is to identify the best produced models by Rough Set technique and to avoid bias. Splitting factor 0.2 denotes that 20 % of data are allocated for training and 80% are for testing. Figure 4.14 shows the group of data that have been prepared for the experiments in the Rosetta application. Table 4.5 shows the data division based on the split factors that have been stated.

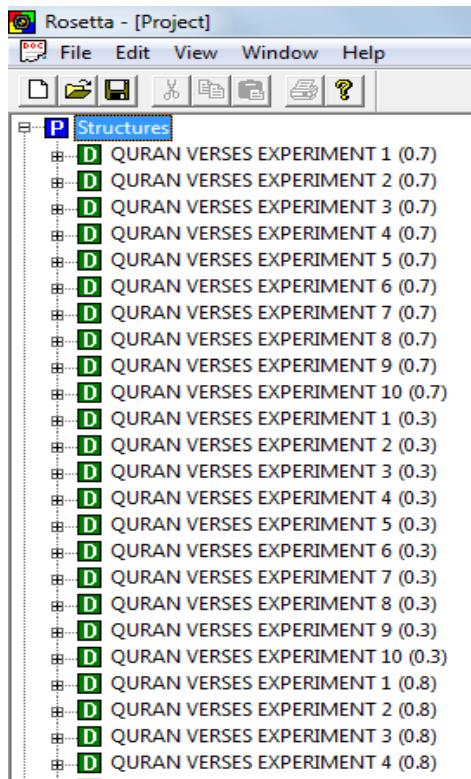


Figure 4.14. Data for the experiment in Rosetta application

Table 4.5

Split factor and data division

Split factor	Train data	Test data
0.2 (20%)	45 objects	179 objects
0.3 (30%)	67 objects	157 objects
0.7 (70%)	157 objects	67 objects
0.8 (80%)	179 objects	45 objects

The training dataset is used to train the rules and extract the decision rules to classify the classes in the testing dataset. In this experiment, Genetic Algorithm reducer has been used onto all training datasets. This is because Genetic Algorithm focused on heuristic search method in solving optimization problems that works well on combinatorial problems such as reduct finding in rough set theory.

The generated rules from the training dataset are used on the testing data. The generated rules are used to match testing objects to compute the strength of the selected rules sets for any decision class. The new object will be assigned to the decision class with maximal strength of the selected rule set.

The next process is to choose which Rough Set model that has the highest accuracy of its classification rules. There are 40 models that have been trained and tested as shown in Table 4.6. Prior to the produced models, a technique called 10-Fold Cross Validation is performed on the 40 models, and the model with the highest accuracy is chosen for the comparison with TopId and experts.

Table 4.6

The trained models divided in Rosetta application

Experiment	Split factor	
	Training	Testing
1 to 10	0.7	0.3
1 to 10	0.3	0.7
1 to 10	0.8	0.2
1 to 10	0.2	0.8

4.5.2 Comparison of topics with experts

On the purpose to evaluate the accurateness of topics produced by TopId and Rough Set technique, an intervention of human experts is required. There are three experts in Quran and Hadith research domains involved. A set of form that has 224 of the selected Quran verses has been given to the experts and they are required to suggest and fill in the appropriate topic based from their knowledge. The experts have been informed that the choice of the topic can be Marriage, Inheritance, Divorce,

Punishment, Moral, History, or Leadership. The sample of the form is shown in Table 4.7.

Table 4.7

Sample of form for experts

Name:			
Expertise:			
Institution:			
No	Surah No/ Verse No	Verse	Suggested Topic
1	2_35	We said: "O Adam! dwell thou and thy wife in the Garden; and eat of the bountiful things therein as (where and when) ye will; but approach not this tree, or ye run into harm and transgression."	
2	2_49	And remember, We delivered you from the people of Pharaoh: They set you hard tasks and punishments, slaughtered your sons and let your women-folk live; therein was a tremendous trial from your Lord.	
3	2_102	They followed what the evil ones gave out (falsely) against the power of Solomon: the blasphemers Were, not Solomon, but the evil ones, teaching men Magic, and such things as came down at babylon to the angels Harut and Marut. But neither of these taught anyone (Such things) without saying: "We are only for trial; so do not blaspheme." They learned from them the means to sow discord between man and wife. But they could not thus harm anyone except by Allah's permission. And they learned what harmed them, not what profited them. And they knew that the buyers of (magic) would have no share in the happiness of the Hereafter. And vile was the price for which they did sell their souls, if they but knew!	

Once all experts have returned the form with the suggested topics, the comparison of topics between topics produced by TopId and the experts is conducted. The identified topics by the three experts are listed in Appendix F. There are two important components for the comparison. Firstly, if both topics from TopId and expert are similar, then the topic is considered as correct. If neither of the topics is similar, then

the topic shall be none. Table 4.8 shows the sample of the comparison between Expert 1 and TopId.

Table 4.8

Comparison of topic between expert and TopId

	Expert		TopId		Topic
IF	M	^	M	→	M
IF	M	^	I	→	NONE
IF	M	^	D	→	NONE
IF	I	^	M	→	NONE
IF	I	^	I	→	I
IF	I	^	D	→	NONE
IF	D	^	M	→	NONE
IF	D	^	I	→	NONE
IF	D	^	D	→	D

Table 4.8 shows the comparison process to evaluate the topics identified by TopId and the experts. The object class, which is the topic, depends on the term patterns between Expert and TopId. At the same time, the topic is classified as ‘NONE’ if both of the conditions are mismatched and considered as false condition.

4.6 Summary

This chapter presented the proposed topic identification method. The stages of the phase were discussed by explaining the reason why it is necessary to be included in the proposed topic identification method. The proposed topic identification method consists of two important phases. First of all is PFA that aims to filter the most relevant terms extracted from the text. PFA depends on the availability of several terms that have been grouped as the keywords in the library. The output from the PFA is the extracted terms. These extracted terms are later being processed in the ranking

phase using *tfidf* technique to find the most important terms amongst the extracted terms. Second is TopId. The rules also depend on the relevant terms obtained from the PFA. That is why term extraction phase is the most crucial task to be implemented in order to avoid TopId to assign wrong topics for the verses. Rough Set technique is also implemented to produce rules and topics from the verses. The experiments are conducted according to the split factor group. All accuracies results obtained from the experiment using Rough Set technique are also compared and the group with the highest accuracy is chosen to be compared with experts in the evaluation phase.

CHAPTER FIVE

EXPERIMENT AND PERFORMANCE EVALUATION

5.1 Introduction

This chapter covers the evaluation on the PFA and the TopId. The experimental design is presented in Section 5.2. The experiments and the results for the PFA are presented in Section 5.3. Section 5.4 provides the experiment and results for the proposed rule generation algorithm TopId. This chapter ends with a summary in Section 5.5.

5.2 Experimental Design

The experimental design is divided into two phases. The first phase consists of the evaluation of the PFA. In this phase, the cleaned texts have been used to test the PFA. The output obtained from the PFA is the collection of relevant terms and these relevant terms are compared with other filtering techniques such as Rough Set Attribute Reduction (RSAR) technique and Information Retrieval technique.

In the first phase of experiment, RSAR technique has been implemented for term filtration by selecting the relevant attributes (terms) out of the larger set of candidate attributes. The relevant attributes are defined as attribute subset that has the same classification capability with the overall attributes. RSAR reduces the dimensionality of the data and enables the learning algorithm to operate effectively. The extraction is based on the calculation of reduct values that have been produced in the Rosetta application. Information Retrieval technique only interprets and ranks its documents according to the relevancy and the importance of the documents itself (Joachims,

Granka, Pan, Hembrooke & Gay, 2005; Ventura & Ferreira da Silva, 2008). In fact, Information Retrieval technique tends to not consider the linguistic criteria unless it is being assigned to (Gelbukh, Espinoza & Galicia-Haro, 2014). Thus, POS tagging has not been applied during the experiment by the Information Retrieval technique in term extraction phase. The PFA is experimented with RSAR is because RSAR is one of the technique in dimensionality reduction and able to reduct the number of features from information system. The PFA has also been compared with IR because it does not employed POS tagging in most term extraction works since IR focuses on the terms with highest score. Meanwhile, PFA chooses noun instead of other regular expression such as verbs and articles.

The second phase in the experimental design is the evaluation of the proposed rule generation algorithm also known as TopId. Before the rule generation phase starts, relevant terms obtained from the PFA are ranked using *tf-idf* technique. The top-ranked relevant terms have been used as the input for rule generation algorithm. The outputs obtained from rule generation algorithm are the rules and the topics. Rule generation using Rough set technique is also employed to obtain the rules and topics. 10-Fold Cross Validation technique has been performed to find the best model of rules that are produced by Rosetta application. The model that has the highest accuracy amongst the 10 models has been selected and the rules are manually extracted. Topics are also assigned on each verse based from the extracted rules by Rosetta application.

Both rules and topics from TopId and Rough Set technique are evaluated with three experts in the Quran and Hadith research domain. The accuracies gained from the comparisons are analyzed. The experimental design is shown in Figure 5.1.

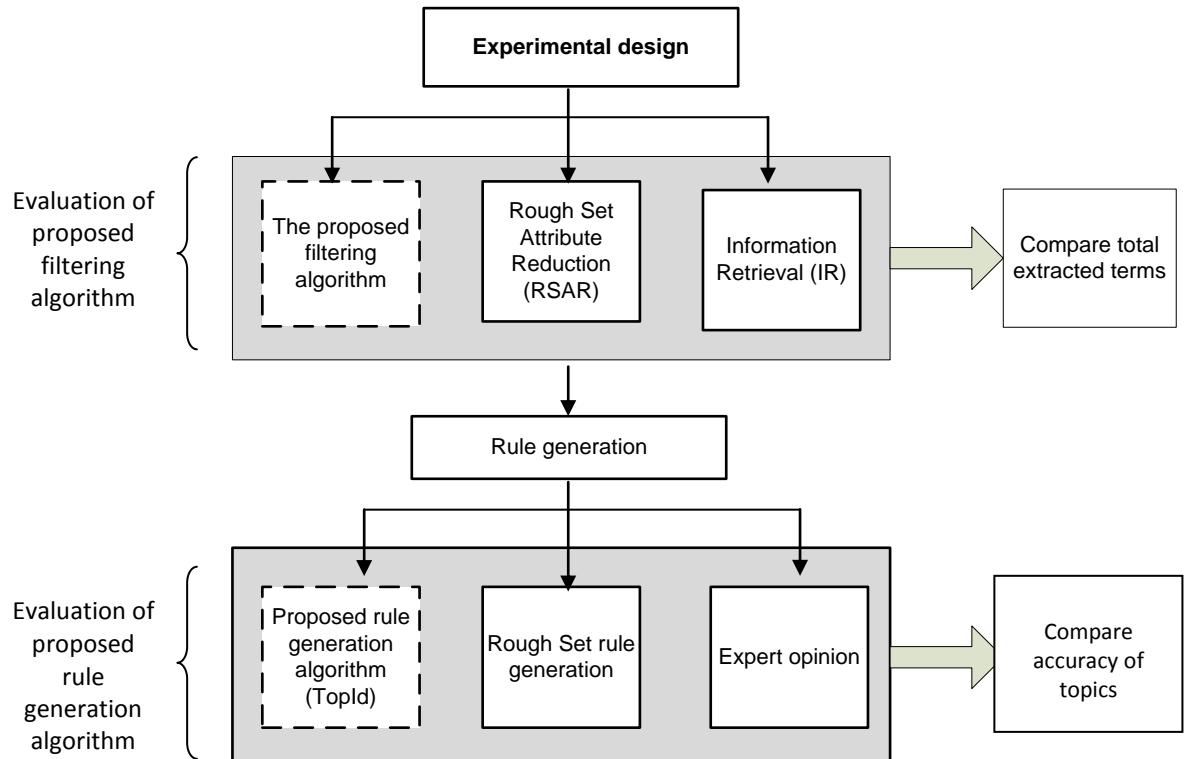


Figure 5.1. Experimental design

5.3 The Proposed Filtering Algorithm Result

The aim of the PFA is to ensure that extractions of terms are relevant enough to be used for rule generation algorithm in the next phase. Besides that, PFA also makes sure that the important terms are not eliminated during term extraction phase.

Table 5.1 shows a sample of the extracted terms that have been produced after the experiments have been conducted. The extracted terms and total number of extracted

terms obtained from these experiments are compared and analyzed. The results of the extracted terms for the 224 verses are presented in Appendix C.

Table 5.1

Sample of the extracted terms

Verse _ No	Extracted terms produced by:					
	Proposed filtering algorithm (PFA)		Rough Set Attribute Reduction technique (RSAR)		Information Retrieval pre-processing (IR)	
	Terms	#terms	Terms	#terms	Terms	#terms
2_35	wife, garden, thing, will, approach, tree, harm, transgression	8	wife, will	2	adam, say, wife, garden, eat, bountiful, thing, approach, tree, run, harm, transgression	12
2_49	people, pharaoh punishment, son, woman, trial, lord,	6	Son, woman	2	remember, deliver, from, pharaoh, set, hard, task, punishment, slaughter, son, let, woman, folk, live, tremendous, trial lord	17

In this experiment, terms in the form of nouns were extracted. From Table 5.1, it can be seen that there are different numbers of the extracted terms from PFA, RSAR and IR.

As an illustration for Verse 2_35, PFA produced 8 extracted terms which consist of nouns and some of the important terms from the keyword library such as the term ‘will’. This is because function words such as verb, adjectives and articles have been removed by the PFA. The terms are filtered with the available keywords in the

libraries. This is due to the fact that PFA does not add complexity for the filtration process in order to keep it simple and fast. The importance of these extracted terms is expressed using a ranking score, indicating which terms is the most meaningful.

The extracted terms obtained for Verse 2_35 by RSAR are only two. The reduction is based from the terms that have a high frequency in the verse and are considered as indiscernible. According to the result, the RSAR consistently has the least number of extracted terms as compared to using PFA and IR. In this case, RSAR removed the irrelevant features from the decision table in order to produce a minimal number of the extracted terms and reduce the processing time for rule generation. However, small numbers of extracted terms do not determine the quality of classification result. This is because a minimal number of extracted terms can be too generalized and may lead to poor coverage for rule generation process. In fact, the information about how much an original attribute contributes is often lost (Janecek & Gansterer, 2008).

RSAR has been used extensively for filtering tasks such as in feature selection and dimensionality reduction. However, Rough Set technique depends on complete information system which is the decision table. The information system should be complete to be processed and objects value must be known. Extra effort and time are required for data preparation and cause to incomplete information system in which missing values often occur in knowledge acquisition as stated by Nabwey (2011).

For IR, the extracted terms for Verse 2_35 are 12 terms. In this case, too many unimportant terms are extracted such as nouns, verbs and adjectives terms, which are considered as irrelevant in this study. This is because these terms do not carry any

meaning to identify the topic of the verses and can cause poor classification result (Janecek & Gansterer, 2008). In addition, important terms such as the term ‘will’ is eliminated since IR tends to group the term ‘will’ as a stop list word, also known as noise word. The term ‘will’ is considered as important in this research context, however, IR neglected the context of the terms and ignored the linguistic properties of the terms. In this situation, IR depends on the linguistic components for it to be able considering the linguistic properties.

5.4 The Rule Generation Algorithm Result

The aim of the TopId is to identify topics for each of the verses. TopId depends on the output obtained by PFA. The relevant terms with the highest ranked score from the previous phase are used as the input for the proposed rule generation algorithm, which is TopId. The output from TopId is the rules and topics. The sample of the extracted terms with the highest rank is presented in Table 5.2.

Table 5.2

Sample of the ranked terms

Verse_No	Term	<i>tf</i>	<i>idf</i>	<i>tf.idf</i>
2_35	will	0.125	0.0480	0.0060
	Wife	0.125	0.0398	0.0050
2_49	Son	0.1429	0.0839	0.0120
	woman	0.1429	0.1022	0.0146
2_102	profit	0.05	1.1751	0.0588
	share	0.05	0.3357	0.0168
	man	0.1	0.0979	0.0098
	wife	0.05	0.0398	0.0020
2_178	equality	0.0556	2.3502	0.1306
	woman	0.1111	0.1022	0.0114
2_187	associate	0.0909	0.7834	0.0712
	approach	0.0909	0.2938	0.0267
	wife	0.0909	0.0398	0.0036
3_35	woman	0.2000	0.1022	0.0204

The frequency scores indicate how frequent the particular terms are mentioned in the verse. The higher the frequency of the terms, the more important the related topic is expected to be. However, terms with a very high frequency of occurrence are usually considered as common words and infrequent terms could not be considered as rare (Ventura & Silva, 2008). The given weight based on the frequency should be checked because it might affect the classification rule by the proposed rule generation algorithm. Nevertheless, this problem has been solved by PFA as it filtered and checked which terms are important to be ranked as relevant terms.

Referring to Verse 2_35 in Table 5.2, both the terms ‘wife’ and ‘will’ have the same score for its *tf* value which is 0.125. However, according to the *tf-idf* score, the term ‘will’ has the highest score which is 0.0060. The TopId takes the highest frequency

ranking score to identify the topic for the verse. Hence, the term ‘will’ has been used by TopId and the identified topic for Verse 2_35 is Inheritance.

On certain cases as Verse 3_35, the highest frequency ranking score is the term ‘woman’. However, TopId could not assign any suitable topic based from the highest frequency ranking score for Verse 3_35 because the term ‘woman’ is too general. As for this solution, TopId identified Verse 3_35 as topic None, which means no topic has been assigned to the verse. A sample of the topic for each verse is shown in Table 5.3. The full result of the identified topics by TopId is presented in Appendix D.

Table 5.3

Sample of the identified topic for each verse by TopId

Verse	Topic
2_227	Divorce
2_231	Divorce
2_178	Inheritance
4_3	Inheritance
2_35	Inheritance
2_49	Marriage
3_35	None
3_42	None

In order to evaluate the rules and topics that have been produced by TopId, Rough Set technique was employed to produce the rules and topics. The rules and topics obtained from both TopId and Rough Set techniques are compared with the three experts.

Prior to the elaboration in Chapter Four, the input to generate rules using Rough Set technique has been set up in the Rosetta application. In the contemplation of avoiding

bias on choosing the best model of rules produced by Rosetta Rough Set application, 10-Fold Cross Validation technique has been performed and the result is presented in Table 5.4. The experiments were divided into four groups according to its splitting factor which are 0.2, 0.3, 0.7 and 0.8. Ten experiments have been carried out for each split factor.

Table 5.4

The result of 10-Fold Cross Validation on Rough Set models

Experiment	Split factor	Training	Testing	% Accuracy
1 (0.7)	70%	157 objects	67 objects	0.67
2 (0.7)	70%	157 objects	67 objects	0.7
3 (0.7)	70%	157 objects	67 objects	0.75
4 (0.7)	70%	157 objects	67 objects	0.6
5 (0.7)	70%	157 objects	67 objects	0.69
6 (0.7)	70%	157 objects	67 objects	0.58
7 (0.7)	70%	157 objects	67 objects	0.66
8 (0.7)	70%	157 objects	67 objects	0.58
9 (0.7)	70%	157 objects	67 objects	0.69
10 (0.7)	70%	157 objects	67 objects	0.66
1 (0.3)	30%	67 objects	157 objects	0.66
2 (0.3)	30%	67 objects	157 objects	0.59
3 (0.3)	30%	67 objects	157 objects	0.7
4 (0.3)	30%	67 objects	157 objects	0.51
5 (0.3)	30%	67 objects	157 objects	0.57
6 (0.3)	30%	67 objects	157 objects	0.68
7 (0.3)	30%	67 objects	157 objects	0.6
8 (0.3)	30%	67 objects	157 objects	0.57
9 (0.3)	30%	67 objects	157 objects	0.63
10 (0.3)	30%	67 objects	157 objects	0.69
1 (0.8)	80%	179 objects	45 objects	0.71
2 (0.8)	80%	179 objects	45 objects	0.6
3 (0.8)	80%	179 objects	45 objects	0.58
4 (0.8)	80%	179 objects	45 objects	0.62
5 (0.8)	80%	179 objects	45 objects	0.56
6 (0.8)	80%	179 objects	45 objects	0.73
7 (0.8)	80%	179 objects	45 objects	0.58
8 (0.8)	80%	179 objects	45 objects	0.67
9 (0.8)	80%	179 objects	45 objects	0.62
10 (0.8)	80%	179 objects	45 objects	0.6
1 (0.2)	20%	45 objects	179 objects	0.66
2 (0.2)	20%	45 objects	179 objects	0.52
3 (0.2)	20%	45 objects	179 objects	0.6
4 (0.2)	20%	45 objects	179 objects	0.69
5 (0.2)	20%	45 objects	179 objects	0.52
6 (0.2)	20%	45 objects	179 objects	0.5
7 (0.2)	20%	45 objects	179 objects	0.55
8 (0.2)	20%	45 objects	179 objects	0.49
9 (0.2)	20%	45 objects	179 objects	0.47
10 (0.2)	20%	45 objects	179 objects	0.66

In this case, the produced model with the highest percentage of accuracy has been chosen. It is shown that the percentage of Experiment 3 for 0.7 split factors has the highest percentage score which is 0.75. There are 541 lines of rules extracted and these rules were tuned manually in order to identify the topics on all 224 verses. The sample of model or known as rules produced from this experiment is shown in Table 5.5. The full result of rules and identified topics based from the Rough Set technique is listed in Appendix E.

Table 5.5

Sample of produced rules and identified topics from Rosetta-Rough Set application

Verse	Rules By Rosetta – Rough Set Technique	Topic
2_35	IF T124(LOW) AND T125(LOW) → NONE	NONE
2_49	IF T107(LOW) AND T127 (LOW) → MARRIAGE	MARRIAGE
2_102	IF T92(LOW) AND T105(LOW) AND T124(LOW) → NONE	NONE
2_178	IF T37(LOW) AND T127 (LOW) → NONE	NONE
2_187	IF T4(LOW) AND T12(LOW) AND T124(LOW) → NONE	NONE
A 2_221	IF T74(LOW) AND T127(HIGH) → INHERITANCE	INHERITANCE
2_222	IF T4(LOW) AND T127(LOW) → NONE	NONE
s 2_223	IF T4(LOW) AND T124(LOW) → NONE	NONE
2_226	IF T124(LOW) → NONE	NONE
2_227	IF T33(LOW) → NONE	NONE
c 2_228	IF T33(LOW) AND T54(LOW) AND T99(LOW) AND T102(HIGH) AND T127(LOW) → INHERITANCE	INHERITANCE
2_229	IF T33(LOW) AND T46(LOW) AND T124(LOW) → NONE	NONE
a	IF T33(LOW) AND T54(LOW) AND T74(LOW) AND T116(LOW) AND T124(LOW) AND T124(LOW) → NONE	NONE
n 2_230	IF T33(LOW) AND T55(LOW) → DIVORCE	DIVORCE
2_232	IF T33(LOW) AND T55(LOW) → DIVORCE	DIVORCE

As can be seen from the table, the rules presented the combination of attribute value.

As an example for Verse 2_35, the value for attribute T124 which is represented as

the term ‘wife’ is low, and the value for attribute T125 which represent the term ‘will’ is also low. The Rough Set decision rule has decided that if this condition happens in any object or verse, then the decision class or topic is NONE. This is because class NONE has been selected as a fallback condition if the rules could not find any suitable class to be assigned on the object during the training session. For Verse 2_49, the attribute value for T107 which is the term ‘son’ is low, and the attribute value for T127 which is the term ‘woman’ is also low. Then, the topic is Marriage. If any objects have the same condition of attribute value as Verse 2_49, the topic shall be Marriage.

The identified topics obtained from both TopId and Rough Set technique have been gathered and compared with the topics given by the experts.

5.5 The Comparison of Results with Experts

To evaluate the correctness of the produced topics by TopId, the topics are compared with the topics given by the experts. The evaluation starts by comparing the topics identified by TopId and topics produced by Rough Set technique with topics given by Expert 1, Expert 2 and Expert 3.

Table 5.6 shows the sample of topic comparison between topics produced by TopId and topics given by Expert 1. Table 5.7 presents the sample of the comparison of topics produced by Rough Set technique and topics from Expert 1. The complete table results for the topic comparison of TopId and experts are presented in Appendix G, and the topic comparison Rough Set and experts are shown in Appendix H.

Table 5.6

Sample of topic comparison for Expert 1 and TopId

No	Verse	Expert 1	TopId	Topic
1	2_35	M	I	-
2	2_49	-	M	-
3	2_178	-	I	-
4	2_221	M	M	M
5	2_222	-	M	-
6	2_229	D	D	D
7	2_230	D	M	-
8	2_237	D	D	D
9	3_35	-	-	-
10	4_7	I	I	I

Based on Table 5.6, both conditions of topics between Expert 1 and TopId should be the same condition. For example, Verse 2_221 has a similar condition of topics from both Expert 1 and TopId, which is M. Thus, it is considered as a matched topic and labeled as M which refers to the topic Marriage. A similar process is applied for Verse 2_229 where both conditions are D, hence it is considered as a matched topic and labeled as D, which refers to the topic Divorce. Diversely for Verse 2_49, neither of the topics is similar and Verse 2_49 is considered as unmatched. For Verse 3_35, it is also labeled as unmatched though both topic conditions are similar because none of it has available topics.

Table 5.7

Sample of topic comparison for Expert 1 and Rough Set technique

No	Verse	Expert 1	Rough Set	Topic
1	2_35	M	-	-
2	2_49	-	M	-
3	2_178	-	-	-
4	2_221	M	I	-
5	2_222	-	-	-
6	2_229	D	-	-
7	2_230	D	D	D
8	2_237	D	M	-
9	3_35	-	-	-
10	4_7	I	-	-

Table 5.7 shows the comparison of topic between Expert 1 and Rough Set technique.

According to Table 5.6, the same procedure as the previous comparison is applied. For example, Verse 2_230 is considered as a matched topic since both Expert 1 and Rough Set produced the same topic which is D or Divorce. For Verse 2_229, neither of the topics are similar, thus the verse is considered as unmatched. Once these comparisons are finished, the calculation to find the accuracies of topics are conducted. The calculation starts by counting the total number of matched topics. It follows by counting the total number of other unrelated topics from the expert. Noted that ‘total other topics from experts’ represent the topics that are not within the scope of this study such as ‘Moral’, ‘Punishment’ and ‘History’.

Based on Table 5.8, there are 37 matched topics between Expert 1 and TopId compared to 11 matched topics for the comparison between Expert 1 and Rough Set technique. The total verses are subtracted with the total number of other topics from the experts and the value is divided with the total number of matched topics. From

that, the accuracy is obtained. The total matched topics and accuracies are depicted in Figure 5.2.

Table 5.8

Accuracy for comparison of TopId and Rough with Expert 1

Criteria	Approach	Expert 1 and TopId	Expert 1 and Rosetta - Rough Set
Total verses		224	224
Total matched topics		37	11
Total other topics from expert		171	171
		37	11
Total matched topics (total verses – total other topics from expert)	$(224 - 171)$	$\frac{37}{(224 - 171)}$	$\frac{11}{(224 - 171)}$
		$\frac{37}{53}$	$\frac{11}{53}$
		0.70	0.21
% accuracy		70%	21%

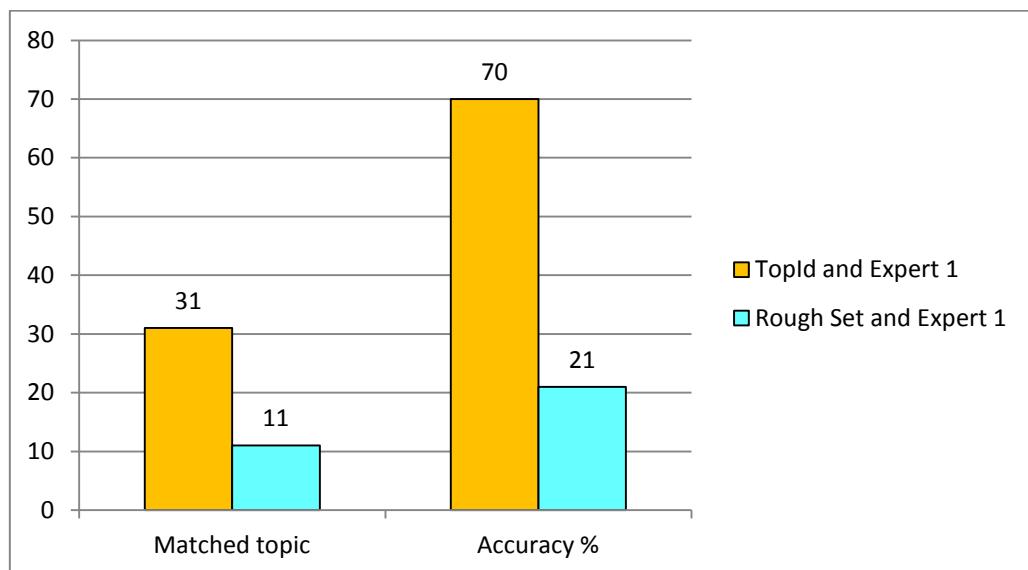


Figure 5.2. Total matched topics and accuracies for comparison with Expert 1

In Figure 5.2, the bar graph clearly states that the total matched topics from TopId and Expert 1 are higher than matched topics from Rough Set technique and Expert 1. The percentage of accuracy for the comparison of TopId and Expert 1 is 70%. The accuracy obtained by the comparison of Rough Set technique and Expert 1 is quite different with TopId and Expert 1 with only 21%.

Table 5.9

Accuracy for comparison of TopId and Rough with Expert 2

Criteria	Approach	Expert 2 and TopId	Expert 2 and Rosetta - Rough Set
Total verses		224	224
Total matched topics		87	38
Other topics from expert		113	113
		87	38
Total matched topics (total verses – total other topics from expert)		$(224 - 113)$	$(224 - 113)$
		87	38
		111	111
		0.78	0.34
% accuracy		78%	34%

Table 5.9 presents the calculation to find the accuracies of the comparison for both TopId and Expert 2 and also for the comparison of Rough Set technique and Expert 2. There are 87 matched topics that have been identified by the topic comparison of TopId and Expert 2. Meanwhile, there are 38 matched topics that have been identified from the comparison of Rough Set and Expert 2. The total matched topics and the accuracy of percentages for both comparisons with Expert 2 are illustrated in Figure 5.3.

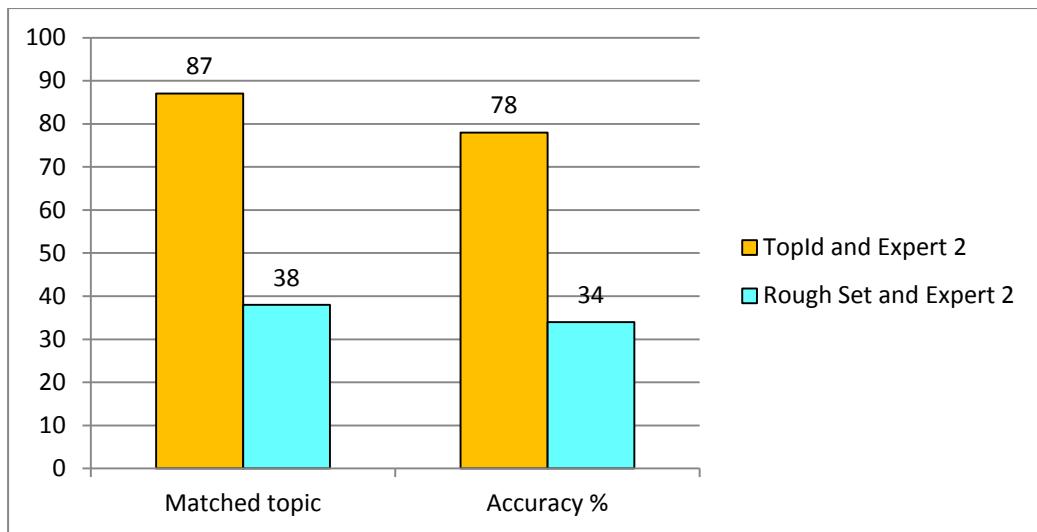


Figure 5.3. Total matched topics and accuracies for comparison with Expert 2

Total matched topics for the comparison of TopId and Expert 2 are higher than the matched topics for the comparison of Rough Set technique and Expert 2. The comparison performed better percentage accuracy for TopId and Expert 2 with 78% of accuracy percentage score. However, the gap of accuracies percentage between TopId and Expert 2 with Rough Set technique and Expert 2 are quite high. The accuracy percentage score obtained by Rough Set technique and Expert 2 is 34%.

Table 5.10 shows the calculation to find the accuracies of the comparison for both TopId and Expert 3 and the comparison of Rough Set technique and Expert 3. There are 48 matched topics that have been recorded after the comparison between TopId and Expert 3. However, only 15 topics are matched from the comparison of Rough Set and Expert 3. The total matched topics and the accuracy of percentages for both comparisons with Expert 3 are depicted in Figure 5.4.

Table 5.10

Accuracy for comparison of TopId and Rough with Expert 3

Criteria \ Approach	Expert 3 and TopId	Expert 3 and Rosetta - Rough Set
Total verses	224	224
Total matched topics	48	15
Other topics from expert	160	160
	48	15
$\frac{\text{Total matched topics}}{\text{(total verses - total other topics from expert)}}$	$\frac{48}{(224 - 160)}$	$\frac{15}{(224 - 160)}$
	48	21
	64	64
	0.75	0.23
% accuracy	75%	23%

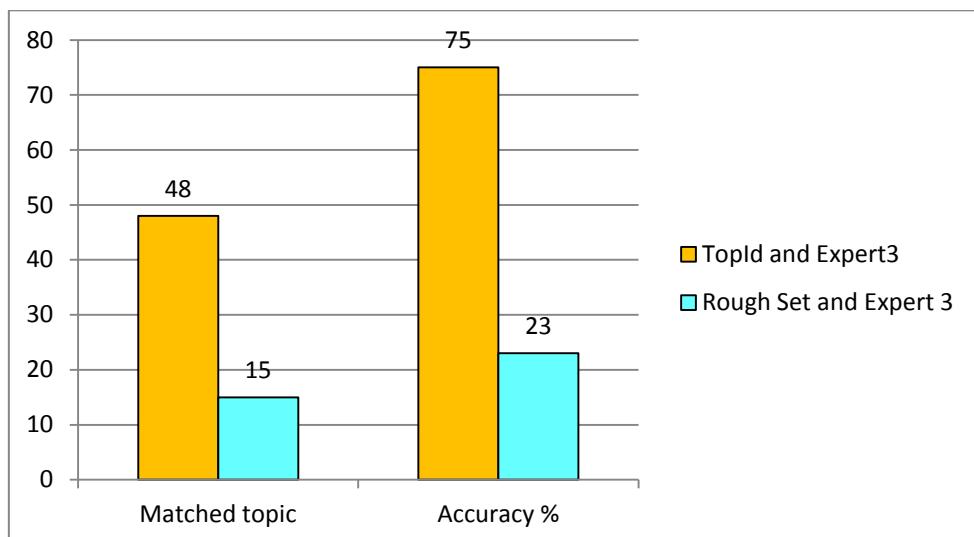


Figure 5.4. Total matched topics and accuracies for comparison with Expert 3

Figure 5.4 illustrates the total matched topics and accuracies obtained from the comparison of both TopId and Rough Set technique with Expert 3. Based from the comparisons, the total matched topics for the comparison of TopId and Expert 3 are higher than the matched topics for the comparison of Rough Set technique and Expert 3. The percentage accuracy for TopId and Expert 3 is 75%. This percentage score is lower compared to the accuracies obtained by the comparison of TopId with Expert 2.

Meanwhile, the accuracy percentage score for Rough Set technique and Expert 3 is 23%. From these three comparisons, Figure 5.5 is depicted to summarize the result of the accuracies.

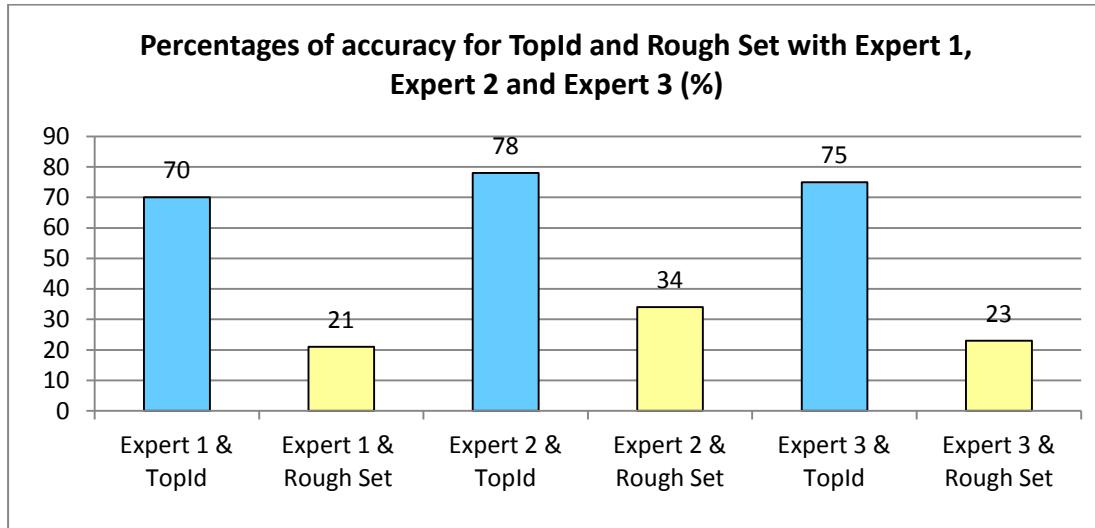


Figure 5.5. The accuracies for comparison of TopId and Rough Set with Expert 1, Expert 2 and Expert 3

The graph illustrated in Figure 5.5 shows the accuracies scores obtained from the comparison with the three experts. It appears that the accuracies obtained by TopId and the three experts increase moderately by 70% to 78% & and slightly decrease to 75%. These results have proven that TopId is able to identify topics as the total number of matched topics between TopId and the three experts are quite high. As a comparison to TopId, the accuracies obtained by Rough Set technique and the three experts are lower with the percentage of 21% increasing to 34% and decreasing to 23%. The result generated is consistently lower to the result obtained by TopId and the three experts. The reason of this poor result is because there are fewer topics that have been identified by Rough Set technique. Apart from that, Rough Set technique assigned any undefined objects as None. This result leads to a high number of

unmatched topics during the comparison of topics from Rough Set technique and the topics given by the three experts.

5.6 Summary

From the experiments, it can be seen that the relevant terms are successfully extracted by the employment of PFA. This is because PFA has its own keyword library that stores the important terms in it. Besides that, PFA ensures that the terms that belong to the keyword library are not eliminated during the filtering phase. The PFA is compared with another two filtering techniques which are Rough Set Attribute Reduction (RSAR) technique and Information Retrieval technique; and it is proven that PFA is able to extract relevant terms from the cleaned texts.

In another matter, the topics given by the experts are used as a benchmark to evaluate the accurateness of the topics identified by the TopId. In order to compare the performance of TopId, Rough Set technique is performed to produce rules for topic identification. Ten experiments have been conducted onto four groups of split factor, which are 0.2, 0.3, 0.7 and 0.8. The rules from the highest accuracy from these experiments were taken for the evaluation with the experts. The experts are selected according to their expertise in Quran and Hadith research domain. The results from all comparisons with the experts are not really consistent since some experts identified several topics that are not included in the scope and study.

CHAPTER SIX

CONCLUSION

6.1 Introduction

This chapter concludes the overall proposed topic identification method. In this thesis, the focuses are on the effective ways to extract the most relevant terms with the application of the PFA and also the TopId to identify topics based from the relevancy of the extracted terms. The contributions of the proposed topic identification method are elaborated in Section 6.2. The future work to enhance the proposed method is also suggested in Section 6.3.

6.2 Contributions of the Research

The major contribution of this research is the proposed method for topic identification in text documents. This major contribution is achieved by the help of the PFA and as well as the TopId.

First is to solve the first sub-objective of this study which is to extract the relevant terms from the text. There are many term extraction methods available, however; these methods cannot be simply embedded in this study. The reason is this study is using verses from English translated Quran and aims to identify topics such as Marriage, Inheritance and Divorce from these verses. In addition, several terms such as ‘will’ might be mistagged as noise word during the pre-processing phase and will be eliminated. In addition, some important terms might be declared as irrelevant such as ‘wed’ because it belongs to the verb class. In addition, certain keywords such as ‘zihar’, and ‘iddat’ may not available in the existing lexical library. Hence, a filtering algorithm is designed to overcome the mentioned problems. In this study, a shallow

technique of computational linguistic is used, namely Part-of-Speech tagging. Shallow techniques are claimed to be a lack of human language understanding. However, this study aims is to identify topics based on the context in the verses instead of interpreting the verses. Hence, it is acceptable. The PFA also reduces the high dimensionality of the text and is able to filter the most important terms. This is because, the PFA has its own keyword libraries whereby all the important terms is stored. In the end, the PFA produces a list of potential terms to be ranked. The effectiveness of the retrieval results by the PFA can be judged by the number of relevant documents retrieved for any particular verses. As shown in the experiments, the PFA has proven that the extracted terms contribute to the performance of the TopId to identify topics which are closer to the topics by the experts.

Second, TopId is also proposed. TopId is designed to identify a topic based on the highest ranked terms and match it with the topic classes. The generated rules also suggest that there is only single term used to represent a topic to each verse. Topics given by the three experts are used as a benchmark to test the produced topics by TopId. Based on the analysis of the experiments, the accuracy of topics from all of three experts is satisfying. Besides, rough set technique is also implemented for topic identification.

6.3 Future work

There are two suggestions to improve the proposed topic identification method. Firstly, in term extraction phase, the PFA only performs the filtration process based on the given keywords in the libraries. The enhancement that can be inserted to help the PFA is to attach the available lexical database such as WordNet. This is because

WordNet can include semantic relations across the concepts of the terms such as the troponymy (the presence of a ‘manner’ relation between two lexemes), antonym (opposite meaning), hyponymy (a word of more specific meaning than a general term is applicable to it) and synonymy (same meaning). However, WordNet lacks of information about the verb syntax. VerbNet is another lexical database which is inspired by Levin’s hierarchy. It contains 237 hierarchically organized classes and 5000 verbs. Each verb in a class is semantically unambiguous and can be explicitly linked to the WordNet. VerbNet provides useful links between the syntax and the semantics of a verb. Hence, a combination of WordNet and VerbNet can be a good enhancement especially in the term extraction phase. This task might be costly as it requires more computational linguistic analysis.

Since this study applies the shallow method for text analysis, an enhancement can be made by applying the deep method in the matter to analyze the syntactic of the terms in the text. Deep natural language processing systems can apply as much linguistic knowledge as possible to analyze texts. Although applying Deep natural language processing can be costly, the extracted terms can be more accurate and meaningful. The references and the appendices for this thesis follow this chapter.

REFERENCES

- Abdullah, Z., Kassim, J. M., & Saad, N. (2009). Pembangunan Perpustakaan Digital: Ayat al-Quran Berkaitan Wanita. *Asia-Pacific Journal of Information Technology and Multimedia*, 6(1).
- Abdullah, A.H.H., & Sudiro, S.R. (2010). Wanita menurut Hamka di dalam Tafsir Al-Azhar: Kajian terhadap Surah An-Nisa'.
- Aery, M., Ramamurthy, N., & Aslandogan, Y. A. (2003). *Topic identification of textual data*. Technical report, The University of Texas at Arlington.
- Aggarwal, C.C., & Zhai, C.X. (2012). A survey of text classification algorithm.
- Ahmad, O., Hyder, I., Iqbal, R., Murad, M. A. A., Mustapha, A., Sharef, N. M., & Mansoor, M. (2013). A Survey of Searching and Information Extraction on a Classical Text Using Ontology-based semantics modeling: A Case of Quran. *Life Science Journal*, 10(4).
- Ain, Q., & Basharat, A. (2011). Ontology driven information extraction from the holy Quran related documents. *26th IEEE Students Seminar 2011. UK*, 41-42.
- Al-Yahya, M., Al-Khalifa, H., Bahanshal, A., Al-Odah, I., & Al-Helwah, N. (2010). An ontological model for representing semantic lexicons: an application on time nouns in the holy Quran. *Arabian Journal for Science and Engineering*, 35(2), 21.
- Ali, N. H., & Ibrahim, N. S. (2012). Porter Stemming Algorithm for Semantic Checking. *ICCIT*.
- Amrani, A., Azé, J., Heitz, T., Kodratoff, Y., & Roche, M. (2004, December). From the texts to the concepts they contain: a chain of linguistic treatments. In *In Proceedings of TREC* (Vol. 4, pp. 712-722).
- Anaya-Sánchez, H., Pons-Porrata, A., & Berlanga-Llavori, R. (2008). A new document clustering algorithm for topic discovering and labeling. In *Progress in Pattern Recognition, Image Analysis and Applications* (pp. 161-168). Springer Berlin Heidelberg.
- Atwell, E., Brierley, C., Dukes, K., Sawalha, M., & Sharaf, A. B. (2011). An Artificial Intelligence approach to Arabic and Islamic content on the internet. In *Proceedings of NITS 3rd National Information Technology Symposium*. Leeds.
- Badawi, J. A. (1980). *Status of women in Islam*. Saudi Arabia Foreigners Guidance Center.
- Badawi, J. A. (2000). The Status of Woman.
- Badr, Y., Chbeir, R., Abraham, A., & Hassanien, A. E. Emergent Web Intelligence: Advanced Semantic Technologies. 2010.
- Baghdadi, H. S., & Ranaivo-Malançon, B. (2011). An Automatic Topic Identification Algorithm. *Journal of Computer Science*, 7(9), 1363.
- Bakar, Z. A., & Rahman, N. A. (2003). Evaluating the effectiveness of thesaurus and stemming methods in retrieving Malay translated Al-Quran documents. In *Digital Libraries:*

Technology and Management of Indigenous Knowledge for Global Access (pp. 653-662). Springer Berlin Heidelberg.

Bakus, J., & Kamel, M.S. (2006). Higher order feature selection for text classification. *Knowledge Information System*, 9,4. 468-491.

Basit, H. A., & Jarzabek, S. (2007, September). Efficient token based clone detection with flexible tokenization. *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering* (pp. 513-516). ACM.

Baqai, S., Basharat, A., Khalid, A., Hassan, A., & Zafar, S. (2009). Leveraging Semantic Web Technologies for Standardized Knowledge Modeling and Retrieval from the Holy Quran and Religious Texts. ACM

Bazan, J. G., Nguyen, H. S., Nguyen, S. H., Synak, P., & Wróblewski, J. (2000). Rough set algorithms in classification problem. In *Rough set methods and applications* (pp. 49-88). Physica-Verlag HD.

Beniwal, S., & Arora, J. (2012, August). Classification and feature selection techniques in data mining. In *International Journal of Engineering Research and Technology* (Vol. 1, No. 6 (August-2012)). ESRSA Publications.

Berkowitz, S. (2010). *U.S. Patent No. 7,805,291*. Washington, DC: U.S. Patent and Trademark Office.

Berry, M. W., & Kogan, J. (Eds.). (2010). Text mining: Application and theory. Wiley.com.

Bigi, B., Brun, A., Haton, J. P., Smaili, K., & Zitouni, I. (2001, November). A comparative study of topic identification on newspaper and e-mail. In *String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings. Eighth International Symposium on* (pp. 238-241). IEEE.

Bolin, H. (2008). Knowledge extraction based on sentence matching and analyzing. *International Symposium on Knowledge Acquisition and Modelling*. 122-126.

Bong, C.H., & Wong, T.K. (2005). An examination of feature selection frameworks in text categorization. *Information Retrieval Technology*. 3689.

Brun, A., Smaïli, K., & Haton, J. P. (2002). Contribution to topic identification by using word similarity. In *INTERSPEECH*.

Bhumika., Sehra, S.S., & Nayyar, A. (2013). A review paper on algorithms used for text classification. *International Journal of Application or Innovation in Engineering & Management (IJAIE)*. 2(3).

Chizi, B., Rokach, L., & Maimon, O. (2009). A Survey of Feature Selection Techniques.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12, 2493-2537.

Coursey, K., Mihalcea, R., & Moen, W. (2009). Using Encyclopedic knowledge for automatic topic identification. *Proceedings of the Thirteenth Conference on Computational Natural*

Language Learning (CoNLL), 210-218, Boulder, Colorado. Association for Computational Linguistics.

Dalal, M. K., & Zaveri, M.A. (2011). Automatic Text Classification : A Technical Review. In *International Journal of Computer Application*. 28(2), 37–40.

Debruyne, M., & Verdonck,T. (2010). Robust kernel principal component analysis and classification.

Devasena, C. L., & Hemalatha, M. (2012, March). Automatic text categorization and summarization using rule reduction. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (pp. 594-598). IEEE.

Dong, R., Schaal, M., O'Mahony, M.P., & Smyth, B. (2013). Topic extraction from online reviews for classification and recommendation. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*.

Edward, H.Y.L., James, N.K.L., & Raymond, S.T.L. (2011). Text information retrieval. In: *Knowledge Seeker-Ontology Modelling for Information Search and Management*. DOI: 10.1007/978-3-642-17916-7_3

Elder, J., Hill, T., Delen, D., & Fast, A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Fuddoly, A., Jaafar, J., & Zamin, N. (2013, November). Keywords Similarity Based Topic Identification for Indonesian News Documents. In *Modelling Symposium (EMS), 2013 European* (pp. 14-20). IEEE.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Belenko, D., & Vanderwende, L. (2009). Using contextual speller techniques and language modeling for ESL error correction. *Urbana*, 51, 61801.

Gelbukh, A., Espinoza, F. C., & Galicia-Haro, S. N. (Eds.). (2014). *Human-Inspired Computing and its Applications: 13th Mexican International Conference on Artificial Intelligence, MICAI2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings* (Vol. 8856). Springer.

Gharaibeh, I.K., & Gharaibeh, N.K. (2012). Towards Arabic Noun Phrase Extractor (ANPE) Using Information Retrieval Techniques. *International Journal of Software Engineering*, 2(2), 36–42. doi:10.5923/j.se.20120202.04

Giannakopoulos, G., Mavridi, P., Palioras, G., Papadakis, G., & Tsipras, K. (2012, June). Representation Models for Text Classification: a comparative analysis over three Web document types. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics* (p. 13). ACM.

Granitzer, M. (2003). *Hierarchical text classification using methods from machine learning* (p. 104).

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60-76.

Hamzah, M. P., & Sembok, T. M. (2006, February). On Retrieval Performance of Malay Textual Documents. In *Artificial Intelligence and Applications* (pp. 156-161).

- Hanum, H. M., Abu Bakar, Z., & Ismail, M. (2013, March). Evaluation of Malay grammar on translation of Al-Quran sentences using Earley algorithm. In *Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on* (pp. 1-4). IEEE.
- Harrag, F., El-Qawasmah, E., & Al-Salman, A. M. S. (2011, April). Stemming as a feature reduction technique for Arabic Text Categorization. In *Programming and Systems (ISPS), 2011 10th International Symposium on* (pp. 128-133). IEEE.
- Harish, B.S., Guru, D.S., & Manjunath, S. (2010). Representation and classification of text documents: A Brief Review. *IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition*.
- Hassan, M. (2013). *Automatic Document Topic Identification Using Hierarchical Ontology Extracted from Human Background Knowledge* (Doctoral dissertation, University of Waterloo).
- Wang, X., & Wang, J. (2013). A Method of Hot Topic Detection in Blogs Using N-gram Model. *Journal of Software*, 8(1), 184-191.
- Hassan, M. M., Karray, F., & Kamel, M. S. (2012, July). Automatic document topic identification using wikipedia hierarchical ontology. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on* (pp. 237-242). IEEE.
- Hong, T. P., Lin, C. W., Yang, K. T., & Wang, S. L. (2013). Using TF-IDF to hide sensitive itemsets. *Applied Intelligence*, 1-9.
- Hotto, H., Nurnberger, A., & Paab, G. (2005). A brief survey of text mining. *Journal for Computational Linguistics and Language Technology*. (20) 1, 9.
- Huang, C. C., Tseng, T. L. B., Fan, Y. N., & Hsu, C. H. (2013). Alternative rule induction methods based on incremental object using rough set theory. *Applied Soft Computing*, 13(1), 372-389.
- Huynh, D., Tran, D., Ma, W., & Sharma, D. (2011, January). A new term ranking method based on relation extraction and graph model for text classification. In *Proceedings of the Thirty-Fourth Australasian Computer Science Conference-Volume 113* (pp. 145-152). Australian Computer Society, Inc.
- Jain, S., & Pareek, J. (2010). Automatic topic(s) identification from learning material: An ontological approach. *2010 Second International Conference On Computer Engineering And Application*. Pp.358-362.
- Janik, M., & Kochut, K. J. (2008, August). Wikipedia in action: Ontological knowledge in text categorization. In *Semantic Computing, 2008 IEEE International Conference on* (pp. 268-275). IEEE.
- Janecek, A., Gansterer, W. N., Demel, M., & Ecker, G. (2008, September). On the Relationship Between Feature Selection and Classification Accuracy. In *FSDM* (pp. 90-105).
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005, August). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 154-161). ACM.

- Jusoh, S., & Alfawareh, H. M. (2012). Techniques, Applications and Challenging Issue in Text Mining ues, Applications and Challenging Issue in Text Mining ues, Applications and Challenging Issue in Text Mining.
- Khan, A., Baharudin, B. Lee, L.H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*. 1(1).
- Kamaruddin, S.S. (2011). *Framework for deviation detection in text*. Universiti Kebangsaan Malaysia, Bangi.
- Kamruzzaman, S. M. (2010). Text Classification using Artificial Intelligence.*arXiv preprint arXiv:1009.4964*.
- Kaplan, R. M. (2005). A method for tokenizing text. *Festschrift in Honor of Kimmo Koskeniemi's 60th anniversary*. CSLI Publications, Stanford, CA.
- Keller, M., & Bengio, S. (2005). A neural network for text representation. In *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005* (pp. 667-672). Springer Berlin Heidelberg.
- Khedikar, K. A., & Lobo, M. L. Data Mining: You've missed it If Not Used Lewis, D. D. (1991). Evaluating text categorization. *Proceedings of the workshop on Speech and Natural Language - HLT '91*, 312–318. doi:10.3115/112405.112471
- Ko, Y., Park, J., & Seo, J. (2002, August). Automatic text categorization using the importance of sentences. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics.
- Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71-82.
- Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261-283.
- Ku-Mahamud, K.R., Ahmad, F., Mohamed Din, A., Wan-Ishak, W.H., Ahmad, F.K., Din, R., & Che Pa, N. (2012). Semantic network representation of female related issues from the Holy Quran. *Knowledge Management International Conference (KMICe)*, Johor Bharu, Malaysia, 4-6 July 2012. Pp.714-718.
- Laboreiro, G., Sarmento, L., Teixeira, J., & Oliveira, E. (2010, October). Tokenizing microblogging messages using a text classification approach. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (pp. 81-88). ACM.
- Li, R., & Wang, Z. O. (2004). Mining classification rules using rough sets and neural networks. *European Journal of Operational Research*, 157(2), 439-448.
- Lin, C. Y. (1997). *Robust automated topic identification* (Doctoral dissertation, University of Southern California).
- Mahar, J. A., Shaikh, H., & Memon, G. Q. (2012). A Model for Sindhi Text Segmentation into Word Tokens. *History*, 3, 37-997.

- Mahender, C. N. (2012). Text Clasification and Classifier: A Survey. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 3(2), 85–99.
- Manacer, M., & Arbaoui, A. (2013). Content extraction of Quran Interpretation (Tafseer) books for digital content creation and distribution.
- McDonough, J., Ng, K., Jeanrenaud, P., Gish, H., & Rohlicek, J. R. (1994, April). Approaches to topic identification on the switchboard corpus. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on* (Vol. 1, pp. I-385). IEEE.
- Mendes, A. C., & Antunes, C. (2009). Pattern mining with natural language processing: An exploratory approach. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 266-279). Springer Berlin Heidelberg.
- Manne, S., & Fatima, S. S. (2011). A novel approach for text categorization of unorganized data based with information extraction. *International Journal on Computer Science and Engineering (IJCSE)*, 3, pp. 2846-2854.
- Manning, C.D. (2011). Part-of-Speech Tagging from 97% to 100%: Is it Time for Some Linguistic?. *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science*. 6608, pp. 171-189.
- Moschitti, A., & Basili, R. (2004). Complex linguistic features for text classification: A comprehensive study. In *Advances in Information Retrieval*(pp. 181-196). Springer Berlin Heidelberg.
- Mukhtar, T., Afzal, H., & Majeed, A. (2012, December). Vocabulary of Quranic Concepts: A semi-automatically created terminology of Holy Quran. In *Multitopic Conference (INMIC), 2012 15th International* (pp. 43-46). IEEE.
- Na, F., Cai, W. D., & Zhao, Y. (2009). A method based on generation models for analyzing sentiment-topic in text.
- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An Introduction. *Journal of the American Medical Informatics Association : JAMIA*, 18(5), 544–51. doi:10.1136/amiajnl-2011-000464
- Natarajan, P., Prasad, R., Subramanian, K., Saleem, S., Choi, F., & Schwartz, R. (2007). Finding structure in noisy text: topic classification and unsupervised clustering. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(3-4), 187-198.
- Nomponkrang, T., & Woraratpanya, K. (2010, September). Thai-sentence classification using conceptual graph. In *Educational and Information Technology (ICEIT), 2010 International Conference on* (Vol. 2, pp. V2-479). IEEE.
- Noordin, M. F., & Othman, R. (2006, January). An information retrieval system for Quranic texts: a proposed system design. In *Information and Communication Technologies, 2006. ICTTA'06. 2nd* (Vol. 1, pp. 1704-1709). IEEE.
- Nguyen, H. S., & Skowron, A. (2013). Rough Sets: From Rudiments to Challenges. In *Rough Sets and Intelligent Systems-Professor Zdzisław Pawlak in Memoriam* (pp. 75-173). Springer Berlin Heidelberg.

- Nuipan, V., & Meesad, P., & Boonrawd, P. (2012). A comparison between keywords and key-phrases in text categorization using feature section technique. *ICT and Knowledge Engineering (ICT & Knowledge Engineering)*. IEEE.
- Okhovvat, M., & Bidgoli, B.M (2011). A hidden Markov model for Persian part-of-speech tagging. *Procedia Computer Science*, 3, 977–981. doi:10.1016/j.procs.2010.12.160
- Padhy, N., Mishra, D., & Panigrahi, R. (2012). The survey of data mining applications and feature scope. *arXiv preprint arXiv:1211.5723*.
- Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*
- Podgorelec, V., & Zorman, M. (2009). *Decision Trees* Decision tree (pp. 1826-1845). Springer New York.
- Protaziuk, G., Kryszkiewicz, M., Rybinski, H., & Delteil, A. (2007). Discovering compound and proper nouns. In *Rough Sets and Intelligent Systems Paradigms* (pp. 505-515). Springer Berlin Heidelberg.
- Ramanathan, V., & Meyyapan, T. (2013). Survey of text mining. *International Conference on Technology and Business Management*.
- Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*
- Riloff, E. (1995, July). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 130-136). ACM.
- Saad, S., Salim, N., & Zainal, H. (2009, November). Islamic knowledge ontology creation. In *Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for* (pp. 1-6). IEEE.
- Said, D. A. (2007). *Dimensionality reduction techniques for enhancing automatic text categorization* (Doctoral dissertation, Faculty of Engineering at Cairo University in Partial Fulfillment of the Requirements for the Degree of MASTER OF SCIENCE in COMPUTER ENGINEERING FACULTY OF ENGINEERING, CAIRO UNIVERSITY GIZA).
- Sadiq, A. T., & Abdullah, S. M. (2012, November). Hybrid Intelligent Technique for Text Categorization. In *Advanced Computer Science Applications and Technologies (ACSAT), 2012 International Conference on* (pp. 238-245). IEEE.
- Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems*, 7(2), 195-207.
- Sagar, B. M., Shobha, G., & Kumar, R. (2009). Solving the Noun Phrase and Verb Phrase Agreement in Kannada Sentences. *International Journal of Computer Theory and Engineering*, 1(3).
- Sagayam, R., Srinivasan, S., & Roshini, S. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. *International Journal of Computational Engineering Research*. 2,5

- Salem, Y. (2009). A generic framework for Arabic to English machine translation of simplex sentences using the Role and Reference Grammar linguistic model.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.
- Sawalha, M., Brierley, C., & Atwell, E. (2012). Predicting Phrase Breaks in Classical and Modern Standard Arabic Text Experimental Dataset : The Quran, 3868–3872.
- Sharaf, A. (2009). Knowledge Representation in the Quran. *Interim Report, University of Leeds*.
- Sharaf, A., & Atwell, E. (2009). Knowledge Representation of the Quran Through Frame Semantics: A Corpus-Based Approach. *Corpus Linguistics-2009*, 12.
- Sebastiani, F. (2005). Text Categorization.
- Serrano, J. I., del Castillo, M. D., Oliva, J., & Iglesias, A. (2011). The influence of stop-words and stemming on human text base comprehension. *Proceedings of the European Perspectives on Cognitive Science*.
- Sharaf, A. B. M., & Atwell, E. (2009). The Qur'an Annotation for Text Mining.
- Silva, C., & Ribeiro, B. (2010). Background on Text Classification. In *Inductive Inference for Large Scale Text Classification* (pp. 3-29). Springer Berlin Heidelberg.
- Skorkovská, L., Irčing, P., Pražák, A., & Lehečka, J. (2011, January). Automatic topic identification for large scale language modeling data filtering. In *Text, Speech and Dialogue* (pp. 64-71). Springer Berlin Heidelberg.
- Srivastava, A., & Sahami, M. (Eds.). (2010). *Text mining: Classification, clustering, and applications*. CRC Press.
- Stein, B., & Zu Eissen, S. M. (2004, June). Topic identification: Framework and application. In *Proc. International Conference on Knowledge Management* (Vol. 400, pp. 522-531).
- Stoyanov, V., & Cardie, C. (2008, August). Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics- Volume 1* (pp. 817-824). Association for Computational Linguistics.
- Sumathy, K. L., & Chidambaram, M. (2013). Text Mining: Concepts, Applications, Tools and Issues—An Overview. *International Journal of Computer Applications*, 80(4), 29-32.
- Syed, Z. S., Finin, T., & Joshi, A. (2008, March). Wikipedia as an Ontology for Describing Documents. In *ICWSM*.
- Syam, M. M., Fayed, Z. T., & Habib, M. B. (2006). An intelligent system for Arabic text categorization. *International Journal of Intelligent Computing and Information Sciences*, 6(1), 1-19.
- T.Sembok, T.M., Abu Bakar, Z., & Ahmad, F. (2011). Experiments in Malay Information Retrieval. 2011 *International Conference on Electrical Engineering and Informatics* 17-19 July. Bandung, Indonesia.

- T.Sembok, T.M., Abu Ata, B.M., & Abu Bakar, Z. (2011). A rule-based Arabic stemming algorithm. *Proceedings of the 5th European Conference on European Computing Conference*. P.392-397, April 28-30,2011, Paris, France.
- Tiun, S., Abdullah, R., & Kong, T. E. (2001). Automatic topic identification using ontology hierarchy. In *Computational Linguistics and Intelligent Text Processing* (pp. 444-453). Springer Berlin Heidelberg.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1), 141-188.
- Uysal, A. K., Gunal, S., Ergin, S., & Sora Gunal, E. (2012). The Impact of Feature Extraction and Selection on SMS Spam Filtering. *Elektronika ir Elektrotechnika*, 19(5), 67-72.
- Patel, F. N., & Soni, N. R. (2012). Text mining: A brief survey. *International Journal of Advanced Computer Research*. (4), 2-7.
- Protaziuk, G., Kryszkiewicz, M., Rybinski, H., & Delteil, A. (2007). Discovering compound and proper nouns. In *Rough Sets and Intelligent Systems Paradigms* (pp. 505-515). Springer Berlin Heidelberg.
- Van Zaanen, M., & Kanters, P. (2010). Automatic mood classification using Tf * IDF based on lyrics. Proceedings of the 11th International Society for Music Information Retrieval Conference, August 9-13, 2010, Utrecht, Netherlands, pp: 75-80.
- Ventura, J., & da Silva, J. F. (2008). *Ranking and extraction of relevant single words in text*. INTECH Open Access Publisher.
- Von Denffer, A. (1983). Ulum al Quran. *The Islamic Foundation*,
- Wang, X., & Wang, J. (2013). A method of hot topic detection in blogs using N-gram model. *Journal of Software*, (8),1.
- Xu, J., Lu, Q., & Liu, Z. (2012, April). Combining classification with clustering for web person disambiguation. In *Proceedings of the 21st international conference companion on World Wide Web* (pp. 637-638). ACM.
- Xu, Y., Zhang, D., & Yang, J. Y. (2010). A feature extraction method for use with bimodal biometrics. *Pattern recognition*, 43(3), 1106-1115.
- Yuan, L. (2010). An Improved Naive Bayes Text Classification Algorithm In Chinese Information Processing. *Science*, 267-269.
- Yauri, A. R., Kadir, R. A., Azman, A., & Murad, M. A. A. (2013a, March). Ontology semantic approach to extraction of knowledge from Holy Quran. In *Computer Science and Information Technology (CSIT), 2013 5th International Conference on* (pp. 19-23). IEEE.
- Yauri, A. R., Kadir, R. A., Azman, A., & Murad, M. A. A. (2013b). Quranic Verse Extraction base on Concepts using OWL-DL Ontology. *Research Journal of Applied Sciences, Engineering and Technology*, 6(23), 4492-4498.
- Zhang, H., Wang, D., Wu, W., & Hu, H. (2012). Term frequency–function of document frequency: a new term weighting scheme for enterprise information retrieval. *Enterprise Information Systems*, 6(4), 433-444.

- Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 38(3), 2758-2765.
- Zhong, N., Li, Y., & Wu, S. T. (2012). Effective pattern discovery for text mining. *Knowledge and Data Engineering, IEEE Transactions on*, 24(1), 30-44.
- Zhao, W. X., Chen, R., Fan, K., Yan, H., & Li, X. (2012, July). A novel burst-based text representation model for scalable event detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 43-47). Association for Computational Linguistics.