Universiti Teknologi MARA

Automatic Text Summarization for Malay News Documents Using Latent Dirichlet Allocation and Sentence Selection

Siti Nur Afiqah Binti Ramlan

Thesis submitted in fulfilment of the requirements for Bachelor of Computer Science (Hons.) Faculty of Computer and Mathematical Sciences

July 2019

ACKNOWLEGMENT

Alhamdulillah, praises and thanks to Allah because of His Almighty and His utmost blessings, I was able to finish this research within the time duration given. Firstly, my special thanks goes to my supervisor, PM Ts Dr Nurazzah Binti Abdul Rahman for her invaluable guidance, assistance and support throughout this project. Under her supervision, many aspects regarding on this project has been explored and with the knowledge, idea and support from her, this thesis can be completed within the given time.

Special appreciation also goes to Madam Haslizatul Fairuz Binti Mohamed Hanum for her idea, guidance and assistance throughout this project and my beloved parents for always encouraging me to complete this project proposal. Last but not least, I would like to give my gratitude to my dearest friend.

ABSTRACT

The proliferation of internet newspapers making an Automatic Text Summarization is now a need to produce a summary that contains most of the important information from the original document. This project focused on the keyword extraction using Latent Dirichlet Allocation and Sentence Selection that used rule based concept approach to produce extractive summary. 100 Malay news documents covering general, sports, health and technology were collected from Utusan Online to evaluate the effectiveness of the system. This project only used a single topic from LDA and top 10 words in the selected topic as the keywords. To evaluate, summary generated by the system was compared to summary generated by human expert using Precision Recall formula. The results showed the effectiveness of the summary generated by the system which is the best score 63.7 % that can help people read the Malay news documents in short time as the summary assist the readers to understand the important parts of the document without reading from the beginning to the end.

TABLE OF CONTENTS

CONTENT **PAGE**

CHAPTER 1

INTRODUCTION

This chapter provides the background and rationale for the study. In the beginning of this chapter, it will explain about the background study of this project, problem statements, project scope and significance of this project.

1.1 Background of Study

The purpose of automatic text summarization is to generate summary from a single document or multiple documents that should express the whole content in minimum number of words without losing its information content (Kumar Meena & Gopalani, 2015).

Keywords can summarize the content of articles and reflect the topic of articles (Lin, Gao, Wang, & Qiu, 2018). Keyword extraction is important in order to identify the relative information in documents that contain some significant terms or words that best express the main point in the document (Hasan, Sanyal, & Chaki, 2018). Automatic keyword extraction is the task to identify a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document (Zhang, 2008). Many text mining applications like automatic indexing, automatic summarization, automatic classification, automatic clustering can take advantage from the automatic keyword extraction (Hasan et al., 2018). Therefore, the automatic text summarization can use the enormous assistance of the automatic keyword extraction.

This project mainly focuses on the role of Latent Dirichlet Allocation (LDA) as topic modelling for extracting the keywords from the documents. The extracted keywords were used in Sentence Selection algorithm to summarize the Malay news documents.