AN ENHANCED MALAY NAMED ENTITY RECOGNITION USING CLUSTERING AND CLASSIFICATION APPROACH FOR CRIME TEXTUAL DATA ANALYSIS

MUHAMMAD SHARILAZLAN BIN SALLEH

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION TECHNOLOGY

2018

Faculty of Information and Communication Technology

AN ENHANCED MALAY NAMED ENTITY RECOGNITION USING CLUSTERING AND CLASSIFICATION APPROACH FOR CRIME TEXTUAL DATA ANALYSIS

Muhammad Sharilazlan bin Salleh

Master of Science in Information and Communication Technology

2018

AN ENHANCED MALAY NAMED ENTITY RECOGNITION USING CLUSTERING AND CLASSIFICATION APPROACH FOR CRIME TEXTUAL DATA ANALYSIS

MUHAMMAD SHARILAZLAN BIN SALLEH

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science
in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2018

DECLARATION

I declare that this thesis entitled "An Enhanced Malay Named Entity Recognition Using Clustering and Classification Approach for Crime Textual Data Analysis" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : …….

Name : …….

Date : …….

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature	:	…….
Supervisor Name	:	…….
Date	:	…….

DEDICATION

This research is dedicated to the Almighty Allah SWT for giving me a good health and strength to implement this research without a sense of hopelessness.

To my beloved parents, thank you for voluntarily providing support and encouragement for my studies. I am very grateful to have both of you in my life. Thank you for giving me the opportunity to improve and change myself to face all the tests and trials in life. Thanks to Allah in the presence of both of you.

To the family and other friends, thank you for always giving help, support and encouragement in this research. I really appreciate it.

To my both respected supervisor and co-supervisor, Dr. Siti Azirah Binti Asmai and Dr. Halizah Binti Basiron, my deepest gratitude goes to both of you for providing me the guidance, understanding and giving great opinions and ideas in helping me to complete this research.

To my fellow friends and seniors, who offered earnest help in solving problems and always there to support all my bittersweet moments throughout this research. I owed all of you my highest appreciation. May Allah bless us all.

ABSTRACT

Named Entity Recognition (NER) is one of the tasks undertaken in the information extraction. NER is used for extracting and classifying words or entities that belong to the proper noun category in text data such as the person's name, location, organization, date, etc. As seen in today's generation, social media such as web pages, blogs, Facebook, Twitter, Instagram and online newspapers are among the major contributors to information extraction. These resources contain various types of unstructured data such as text. However, the amount of works done to process this type of data is limited for Malay Named Entity Recognition (MNER). The deficiency on Malay textual analytic has led to difficulties in extracting information for decision making. This research aims to present a Malay Named Entity Recognition technique that focuses on crime data analysis in the Malay language that extracted from Polis Diraja Malaysia (PDRM) news web page. This Malay Named Entity Recognition (MNER) technique is proposed by using multi-staged of clustering and classification methods. The methods are Fuzzy C-Means and K-Nearest Neighbors Algorithm. The methods involve multi-layer features extraction to recognize entities such as person name, location, organization, date and crime type. This multi-staged technique is obtained 95.24% accuracy in the process of recognizing named entities for text analysis, particularly in Malay. The proposed technique can improve the accuracy performance on named entity recognition of crime data based on the suitability selected features for the Malay language.

ABSTRAK

Pengiktirafan Entiti Dinamakan (NER) adalah salah satu tugas yang dilakukan dalam pengekstrakan maklumat. NER digunakan untuk mengekstrak dan mengklasifikasikan perkataan atau entiti yang dimiliki oleh kategori kata nama yang betul dalam data teks seperti nama, lokasi, organisasi, tarikh, dan sebagainya. Seperti yang dilihat dalam generasi masa kini, media sosial seperti halaman web, blog, Facebook, Twitter, Instagram dan akhbar dalam talian adalah antara penyumbang utama kepada pengekstrakan maklumat. Sumber-sumber ini mengandungi pelbagai jenis data yang tidak berstruktur seperti teks. Walau bagaimanapun, kerja-kerja yang dilakukan untuk memproses jenis data ini terhad kepada Pengiktirafan Entiti Dinamakan Melayu (MNER). Kekurangan analitik tekstual Melayu telah membawa kepada kesulitan dalam mengekstrak maklumat untuk membuat keputusan. Penyelidikan ini bertujuan untuk membentangkan teknik Pengiktirafan Entiti Dinamakan Melayu yang memberi tumpuan kepada analisis data jenayah dalam bahasa Melayu yang diekstrak dari laman web berita Polis Diraja Malaysia (PDRM). Teknik Pengiktirafan Entiti Dinamakan Melayu (MNER) ini dicadangkan dengan menggunakan kaedah kluster dan klasifikasi pelbagai peringkat. Kaedah adalah Fuzzy C-Means dan K-Nearest Neighbors Algorithm. Kaedah ini melibatkan pengekstrakan ciri pelbagai lapisan untuk mengiktiraf entiti seperti nama orang, lokasi, organisasi, tarikh dan jenis jenayah. Teknik multi-tahap ini memperoleh ketepatan 95.24% dalam proses mengenali entiti yang dinamakan untuk analisis teks, terutamanya dalam bahasa Melayu. Teknik yang dicadangkan ini boleh meningkatkan prestasi ketepatan pada pengiktirafan entiti yang dinamakan data jenayah berdasarkan kesesuaian ciri-ciri terpilih untuk bahasa Melayu.

ACKNOWLEDGEMENTS

First and foremost, my greatest indebtedness is to everyone that helps me during the implementation of this research. This research may not be possible without your support. Especially to both my respected supervisor and co-supervisor for this Master by Research, Dr. Siti Azirah Binti Asmai and Dr. Halizah Binti Basiron, a lot of thanks for their valuable helps, advices, opinions, ideas and guidance. They strengthen my spirit to always looking forward to the implementation of research. They are willing to find various ways in every corner to give me an understanding of all the things I need during conducting this research. I also would like to thank them for giving me related information whether in articles, codes or case studies that I need for this research.

Next, I really want to thank my family, which are my parents and siblings for giving me a lot of moral support and always pray for my success in completing the research and also for my future and hereafter. My gratitude to them for always giving me meaningful advice for never lose hope to carry out this research. I would also like to thank my seniors, lab mates and my understanding friends for being there and together with me to solve the problems that occurred during the implementation process of this research. Thank you all for the efforts and patience towards me in completing this research. Last but not least, I would also like to extend my gratitude to the Centre for Graduate Studies of Universiti Teknikal Malaysia Melaka (UTeM), for the fund and support for my study under MyBrain UTeM Scheme.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
CRFs	-	Conditional Random Fields
CRI NKRA	-	Reducing Crime National Key Results Area
FCM	-	Fuzzy C-Means
GTP	-	Government Transformation Programme
HMM	-	Hidden Markov Models
IE	-	Information Extraction
IoT	-	Internet of Things
KNN	-	K-Nearest Neighbors
MEMMs	-	Maximum Entropy Markov Models
MNER	-	Malay Named Entity Recognition
NER	-	Named Entity Recognition
NLP	-	Natural Language Processing
NEs	-	Named Entities
PDRM	-	Polis Diraja Malaysia
POS	-	Part-of-Speech
RP	-	Research Problem
SL	-	Supervised Learning
SSL	-	Semi-supervised Learning

URL	-	Uniform Resource Locator
USL	-	Unsupervised Learning

LIST OF PUBLICATIONS

1. Siti Azirah Asmai, Muhammad Sharilazlan Salleh, Halizah Basiron and Sabrina Ahmad, 2018. An Enhanced Malay Named Entity Recognition using Combination Approach for Crime Textual Data Analysis. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 9(9), pp. 474-483.

2. Salleh, M.S., Asmai, S.A., Basiron, H., and Ahmad, S., 2018. Named Entity Recognition using Fuzzy C-Means Clustering Method for Malay Textual Data Analysis. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(2-7), pp. 121-126. (Scopus).

3. Salleh, M.S., Asmai, S.A., Basiron, H., and Ahmad, S., 2017. A Malay Named Entity Recognition using Conditional Random Fields. *Information and Communication Technology (ICoIC7), 2017 5th International Conference on,* pp. 1-6. (Scopus).

CHAPTER 1

INTRODUCTION

1.1 Overview

This chapter briefly describes the background of this research and the problems of unstructured information from textual data for analysing process especially for crime analysis. Huge data are being used by police forces to analyse in crime prevention purpose. This chapter is divided into several sections includes research background, problem statements, research questions, research objectives, the scope of research, the structure of the thesis, and summary.

1.2 Research Background

Information is one of the important sources in human life that is increasingly rising and technologically. At all times, several types of information have been generated on the Internet and the amount of information is constantly increasing from time to time. Information consisting of diverse types such as text, images, audio, video, data, and so on are increasingly being generated on the Internet which are largely unstructured. The growing number of information that comprises from unstructured data affects the daily lives of people in work, learning and lifestyle. The effective management and organization of this kind of information or data representation is a key strategy for addressing the problem of finding useful information. The appropriate techniques and methods are very necessary to process and extract the essential knowledge contained in this information.

Analysing unstructured textual data has become an active research nowadays and offers a wealth of valuable information into many fields such as business, education, political, healthcare, crime prevention and others. With the advancement in the Internet of Things (IoT) technologies, the masses of unstructured textual data are accessible on the wide web world from various sources such as the online document and newspapers, web journals, Facebook, as well as Twitter or Instagram. However, with no proper cluster solution, the unstructured textual data i.e. crime news may not capable to convey conceivable details to guide appropriate actions taking. As become exciting research field, this text mining is an increasable need in analysing unstructured text data as it contains a lot of valuable information that cannot simply be used by computers for further processing (Behera and Kumar, 2015).

These textual data can be represented in the form of words in the language of human communication. Languages include many aspects that can be comprehended through speech, writing, movement, gestures and others. Language in text or writing contains words and symbols such as letters, digits, and special characters. This analysis of languages has recently been carried out through Natural Language Processing (NLP). Goel (2017) defines that NLP as a research area focusing on the ability of machines to understand and manipulate natural language texts or speeches for useful purposes that involve artificial intelligence, computer science, and computational linguistics as an interaction between human language and computer. Iroju and Olaleke (2015) stated that NLP is defined as the computational linguistic that processes natural languages using the computerized system for human and computer interactions. This problem solving is by using Information Extraction (IE) task. This task used by researchers to convert unstructured data or text into data structures that the machine can understand (Tanwar et

al., 2015). IE is one of the areas of research organized by the NLP that has been involved in many sub-topics related to the field of data restructuring.

IE has been utilized for text analysis process in many tasks and one of them is Named Entity Recognition (NER) task. NER is one of the textual analysis approaches to recognize entities in open-domain text document documents such as person, facility or organization entities. Most of these NER studies are conducted in processing English using various methods that include artificial intelligence and ruled-based methods. As stated by Sulaiman et al. (2017), this named entity recognition task has been carried out in many types of research in identifying named entities in many languages such as English, Arabic, Chinese, and Indian by using different techniques in dealing with these NERs. The suitability of the technique used to perform the NER task is based on the type of language being processed. This is because each language has different presentation and explanation in translating something that meaningful, for example, through vocabulary and grammar use in writing. Similarly, the Malay language also has ways of conveying information through vocabulary and grammar usage. Besides that, the Malay languages have its own morphology. Because of that, the NER research is rarely implemented in the Malay language to obtain valuable information from the Malay language documents (Morsidi et al., 2015). So, the suitable technique needs to apply for the Malay Named Entity Recognition (MNER) to improve the way of recognizing entities in the Malay text document.

Therefore, this research proposes an enhanced Malay Named Entity Recognition (MNER) algorithm based on clustering and classification techniques respectively that used to guide the recognition process of entities from crime unstructured text news in the Malay language. The clustering and classification techniques are proposed in the algorithm to

overcome the multi-representation and uncertainty problem of entity contexts. The proposed NER algorithm is designed in five phases. Firstly, in Phase One, the data acquisitions are conducted by extracting web pages contents. Secondly, in Phase Two, pre-processing data is carried out through several processing parts. Next, Phase Three conducted the process of extracting Malay features as the important phase for Malay Named Entity Recognition (MNER) task. The development of the Malay Named Entity Recognition (MNER) model is developed in Phase Four based on the proposed techniques. Lastly, the evaluation for Malay Named Entity Recognition (MNER) is measured in Phase Five. The research conducted due to tackle the issue of NER in Malay to extract the valuable information from Malay document using the appropriate method.

1.3 Problem Statement

Reducing crime is one of the efforts that are being taken seriously by the Malaysian Government to meet the target of the Reducing Crime National Key Results Area (CRI NKRA). This effort is important to the Malaysia Government in making the country safer and improving the quality of life in line with the requirements of the country to achieve the Government Transformation Programme (GTP). Based on this, the Malaysian government has been progressively making various efforts in reducing crime including improving and diversifying criminal investigations. Among the efforts undertaken were through the analysis of information from the various resources. For example, the Malay crime news in Polis Diraja Malaysia (PDRM) website contains all information in the form of structured and unstructured textual data. Supposedly, each word in Malay crime textual data should be analysed intensively and statistically. This is because it contains the important information that leads to how police investigations and actions can be taken and executed. Although, the crime data can be arranged in predetermined structured recognition, it might

lack to capture all information from natural language data. However, if the information is simply captured without formal structured, the informative entities will be difficult to be extracted and time-consuming which influence the effective police investigation.

There are limited named entity recognition task has been conducted in the Malay language. Malay Named Entity Recognition (MNER) approach by Alfred et al. (2014) is using a set of rules and a list of dictionaries set by the human to identify entities. These rules work to extract the pattern of an entity such as location, organization and other entities based on their basic pattern. The patterns of entities in the Malay language mostly refer to orthographic, grammatical and syntactic features. Additionally, the recognition process is speeded up by using the dictionaries list but these dictionaries types affect the NER system performance. This is because all libraries or dictionaries used should always be updated (Alfred et al. 2014). Therefore, the algorithm made for Malay Named Entity Recognition need some adjustments in rules and dictionaries which have been designated as an improvement effort in recognizing entities. Due to that, an enhanced named entity recognition algorithm should be formulated to recognize the named entities in the Malay language to supply valuable information. Furthermore, currently named entity recognition methods are typically based on supervised expert labelled document, but it consumes a lot of time and resources. Even though the recognition process also can be learned through unsupervised learning but it often performs poorly correspond to the natural entity clustering data.

The current challenge is on how to enhance the classical methods that used the rule-based method that still relying on large collection pre-determined labelled data to recognize the Malay named entities that may consist plentifully multi-representation and uncertainty context of data that cause an ambiguity during recognizing named entity as a

few text data are identified as more than one entity types such as person name entity is used as location name entity. For example, “Jalan Haji Samsuri” should be recognized as a Location entity because of the personal name of “Haji Samsuri” that is commonly understood as a Person entity and this will cause the ambiguity during the entity recognition process. Due to that, the multi-staged of clustering and classification approach can be used to utilize a set of textual data as seeds to start the learning process by using the clustering method at first. Then, the entities cluster is used by diversifying the types of entities in details classes as to be used in classifying textual data for Malay Named Entity Recognition (MNER). The proposed method is implemented because of lack of research have been done for crime entity recognition analysis especially for Malay. Table 1.1 shown the research problem (RP) that have been identified that are used to find the solution for text analysis.

Table 1.1: Research Problem (RP)

RP1	Each of word in Malay crime textual data should be analysed intensively because it contains the important information that leads on how the police investigations and actions can be taken and executed. However, the study of crime-domain are limited and crime-specific information simply reported without formally structured clusters such as trends and statistical analysis of crime. Other than that, crime-specific named entities in the huge of open–textual domain data still difficult to be extracted which influence the police investigation for crime solving process.
RP2	There are limited named entity recognition task has been conducted in the Malay language. The set of rules and list of dictionaries set by the human to identify entities are used for the current approach of NER analysis. However, the algorithm made for Malay Named Entity Recognition need adjustments in rules and dictionaries which have been designated when the domain of studies is changed as an improvement effort in recognizing entities that cause time-consuming. Thus, an enhanced named entity recognition algorithm should be formulated to recognize the named entities in the Malay language.