Faculty of Information and Communication Technology

ENHANCEMENT OF FEATURE SETS FOR SUBJECTIVITY ANALYSIS ON MALAY-ENGLISH CODE-SWITCHING TEXT

Emaliana Binti Kasmuri

Doctor of Philosophy

2023

ENHANCEMENT OF FEATURE SETS FOR SUBJECTIVITY ANALYSIS ON MALAY-ENGLISH CODE-SWITCHING TEXT

EMALIANA BINTI KASMURI

A thesis submitted
in fulfilment of the requirements for the degree of Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DECLARATION

I declare that this thesis entitled “Enhancement Of Feature Sets For Subjectivity Analysis On Malay-English Code-Switching Text” 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 : EMALIANA BINTI KASMURI

Date : 11 January 2023

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 Doctor Philosophy.

Signature : lieza

Supervisor Name : DR. HALIZAH BINTI BASIRON

Date : 11 January 2023

DEDICATION

To my beloved husband, father, mother and sister.

Thank you for all your love, support and patience.

ABSTRACT

A code-switching sentence is a sentence that is constructed using two or more languages. It is a norm for a multi-lingual speaker to use code-switching sentences to share objective and subjective textual information on public platforms such as blogs and social media. Classifying a voluminous code-switching text into subjective and objective classes has posed a new challenge to the current solution of subjectivity analysis. The current solution has limited its design to process only monolingual text. Therefore, the presence of subjective code-switching text is ignored by the current solution. The ignorant limits the capability of the current solution to generate an accurate result of subjectivity analysis on code-switching text. Therefore, this research aims to find a set of solutions for subjectivity analysis on code-switching text. The research process begins by filling in the absence of the subjectivity code-switching corpus. A subjective Malay-English code-switching corpus was built. The corpus contains 35,067 Malay-English code-switching sentences that were harvested from Malay-English blog posts. Each sentence was annotated with either subjective or objective labels. The research process continues with designing the feature sets that represent the subjectivity of the Malay-English code-switching sentences from the corpus. The feature sets were enhanced from the subjective monolingual feature set, that was initially designed to represent subjectivity of English text. The initial subjective monolingual feature sets consist of pronoun, adjective, cardinal number, modal and adverb. The enhanced feature sets consist three feature sets which are embedded code-switching feature set, unified code-switching feature set and stylistic feature set. The embedded code-switching feature used the initial monolingual feature set for English and embeds the feature of Malay language in it. In the unified code-switching feature set, the extracted Malay and English features were unified using an adapted algorithm known as the Malay-English Unified POS. The algorithm predicts the type of each word in a code-switching sentence according to the language of the word. In the stylistic feature set, emoticons, interjections, signs of subjectivity such as exclamation marks and word with exaggerations of spelling were extracted to represent the subjectivity in the code-switching sentences. The effectiveness of the enhanced feature sets was evaluated using the Malay-English code-switching subjectivity corpus as the data set and two machine learning classifiers, which are Naïve-Bayes and Support Vector Machine. The 10-fold cross-validation classification technique was used on different settings of experiments and combinations of feature sets to obtain the performance of the enhanced feature sets. The performance from the combination of unified code-switching and stylistic feature sets has outperformed other feature sets. The combination has consistently performed at the accuracy of 59% using both machine learning classifiers. The consistent performance indicates the combined feature sets are the viable solution for subjectivity analysis on the Malay-English code-switching text.

PENAMBAHBAIKAN SET FITUR UNTUK ANALISA SUBJEKTIVITI KE ATAS TEKS PERALIHAN KOD MELAYU-INGGERIS

ABSTRAK

Sepotong ayat peralihan kod ialah ayat yang dibina menggunakan dua atau lebih bahasa. Adalah menjadi satu kebiasaan bagi seorang yang fasih pelbagai bahasa menggunakan ayat-ayat peralihan kod berkongsi maklumat tekstual objektif dan subjektif di atas platform terbuka seperti blog dan media sosial. Pengelasan teks peralihan kod yang banyak kepada kelas-kelas subjektif dan objektif telah memberikan cabaran baru kepada penyelesaian sedia ada bagi analisa subjektiviti. Rekabentuk penyelesaian sedia ada telah menghadkan rekabentuk penyelesaian hanya untuk memproses teks bahasa mono. Oleh itu, kehadiran teks subjektif peralihan kod diabaikan oleh penyelesaian sedia ada. Pengabaian ini telah menghadkan keupayaan penyelesaian sedia ada untuk menjana keputusan yang tepat bagi analisa subjektiviti ke atas teks peralihan kod. Maka, matlamat kajian ini adalah untuk mendapatkan satu set penyelesaian bagi analisa subjektiviti ke atas teks peralihan kod. Proses kajian ini bermula dengan mengisi kelompongan korpus subjektiviti peralihan kod. Sebuah korpus subjektiviti peralihan kod Melayu-Inggeris telah dibangunkan. Korpus tersebut mengandungi 35,067 ayat-ayat peralihan kod Melayu-Inggeris yang dituai dari hantaran-hantaran blog. Setiap ayat dilabelkan dengan subjektif atau objektif. Proses kajian diteruskan dengan mereka bentuk set-set fitur yang mewakili ciri-ciri subjektif bagi ayat-ayat peralihan kod Melayu-Inggeris. Set-set fitur ini ditambahbaik daripada set fitur subjektif bahasa mono, yang mana pada asalnya mewakili subjektiviti teks Bahasa Inggeris. Set fitur awal subjektif bahasa mono mengandungi kata ganti nama, adjektif, nombor kardinal, modal dan adverba. Set-set fitur yang ditambahbaik mengandungi the set fitur iaitu set fitur peralihan kod tertanam, set fitur peralihan kod kesatuan dan set fitur gayaan. Set fitur peralihan kod tertanam menggunakan set fitur awal bahasa mono untuk Bahasa Inggeris dan ditanam fitur bahasa Melayu di dalamnya. Di dalam set fitur peralihan kod kesatuan, fitur-fitur bahasa Melayu dan Inggeris yang telah disatukan menggunakan algoritma yang telah diadaptasi yang dikenali sebagai POS Kesatuan Melayu-Inggeris. Algoritma ini meramal jenis setiap perkataan di dalam ayat peralihan kod berdasarkan kepada bahasa perkataan tersebut. Di dalam set fitur gayaan, emotikon, celahan, tanda subjektiviti seperti tanda seru dan perkataan dengan ejaaan yang berlebihan digunakan untuk mewakili ciri subjektiviti di dalam ayat-ayat peralihan kod. Keberkesanan penambahbaikkan set-set fitur dinilai menggunakan korpus subjektiviti peralihan kod Melayu-Inggeris sebagai set data dan dua pengelas pembelajaran mesin iaitu Naïve-Bayes dan Support Vector Machine. Teknik klasifikasi bersilang 10-lipatan digunakan ke atas pelbagai tetapan dan kombinasi eksperimen bagi mengukur pencapaian ke atas penambahbaikan set-set fitur. Pencapaian daripada kombanasi set fitur kesatuan dan gayaan telah mendahului set-set fitur yang lain. Kombinasi tersebut telah mencapai ketepatan 59% secara konsisten menggunakan kedua-dua pengelas pembelajaran mesin. Pencapaian yang konsisten menunjukkan kombinasi set-set fitur adalah penyelesaian yang berdaya maju untuk analisa subjektiviti ke atas teks peralihan kod Melayu-Inggeris.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful

I thank Allah for granting me the patience, health, guidance and determination to complete this thesis successfully. This thesis exists because of the assistance, support and inspiration of many people. Firstly, I acknowledge both Universiti Teknikal Malaysia Melaka dan the Government of Malaysia, for supporting me and giving me the opportunity.

I'd want to extend my heartfelt gratefulness to Dr Halizah Basiron for her encouragement and assistance during my work on my thesis. She introduced me to a variety of online and library options. She addressed all of my inquiries and assisted me in narrowing down my search. I appreciate how she constructed her constructive and positive critique and how attentively she reviewed my work. I also wish to thank my second supervisor Dr Mohamed Ishak Desa, who retired during the completion of my studies, for being patient with me and my questions, and Dr Yogan Jayakumar for his comments on my work.

Lastly, for all my friends who support me at Fakulti Teknologi Maklumat dan Komunikasi, I gave them my thanks and pray that Allah will protect them from harm.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AMI	-	Augmented Multi-party Interaction
AUROC	-	Area Under the Curve
BERT	-	Bidirectional Encoder Representations from Transformers
BiLSTM	-	Bidirectional Long Short-Term Memory
BOW	-	Bag-of-Words
CNN	-	Convolutional Neural Network
CRF	-	Conditional Random Field
D-BiLSTM	-	Double Bidirectional Encoder Representations from Transformers
DSE	-	Direct Subjective Expression
EM	-	Expectation Maximization
EN	-	English
ESE	-	Expressive Subjective Element
EWGA	-	Entropy Weighted Genetic Algorithm
GloVe	-	Global Vector
HTML	-	Hypertext Markup Language
IDE	-	Integrated Development Environment
IG	-	Information Gain
LAMP	-	Logistic Average Misclassification Percentage
LDA	-	Latent Dirichlet Allocation

MAP	-	Mean average precision
MPQA	-	Multi Perspective Question Answer
MS-EN	-	Malay-English
MS-EN-CS	-	Malay-English Code-Switching
NLTK	-	Natural Language Toolkit
OQ	-	Opinion Queries
POS	-	Part-of-Speech
PMI	-	Pointwise Mutual Information
RQ	-	Research Question
SLM	-	Subjective Language Model
SVM	-	Support Vector Machine
SWSD	-	Subjective Word Sense Disambiguation
TREC	-	Text Retrieval Conference
TF-IDF	-	Term Frequency-Inverse Document Frequency
UKM	-	Universiti Kebangsaan Malaysia
URL	-	Uniform Resource Locator
U.S	-	United States
VSM	-	Vector Space Model
WMI	-	Web-based Mutual Information

LIST OF PUBLICATIONS

Kasmuri, E. and Basiron, H., 2020. Segregation of Code-Switching Sentences using Rule-Based Technique. *International Journal of Advances in Soft Computing and its Applications*, 12 (1), 49–64.

Kasmuri, E. and Basiron, H., 2019. Building a Malay-English code-switching subjectivity corpus for sentiment analysis. *International Journal of Advances in Soft Computing and its Applications*, 11 (1), 112–130.

Kasmuri, E. and Basiron, H., 2017. Subjectivity analysis in opinion mining - A systematic literature review. *International Journal of Advances in Soft Computing and its Applications*, 9 (3), 132–159.

CHAPTER 1

INTRODUCTION

1.0 Introduction

This chapter gives an overview of the research work carried out for subjectivity analysis on mixed-language documents known as code-switching documents. It begins with a description of the research background. The research problems in the subjective analysis of code-switching documents were highlighted. Research questions were drawn from the problems and objectives were outlined to fulfil the aim of the research. The output of the research is listed. The organization of this thesis is described before the end of the chapter. This chapter is closed with a summary.

1.1 Research Background

Information is easily accessible today as compared to many decades ago. Users create information on various platforms, such as personal websites, blogs and social media. The information can be factual or non-factual, or a mixture of both. Factual information describes truths, details, or particulars about a subject matter related to a text (Banea et al., 2014). For example, facts about a book include the title, the number of pages, the book's author, the synopsis and the book's price. Non-factual information contains evaluative or affective information from some aspect of the described subject matter, such as opinion about the storyline from the perspective of a reader, for example, good or bad, or comments from a reader on the design of the book cover, for example, pleasant or unpleasant. Factual

information is known as objective information, whereas non-factual information is known as subjective information (Banea et al., 2014).

Subjective and objective information has become a piece of important information in decision-making. Both parts of information are used for different purposes. For example, subjectivity analysis was used in an automated question-and-answer system to interpret the type of questions asked, either seeking answers for personal opinions or experiences or verifying information (Li et al., 2008). Subjectivity analysis is also used to analyse the political stance during political campaigns aiding the party in planning its campaign strategy (Jiang and Argamon, 2008). Interestingly, subjectivity analysis was also used in automatic article classification to identify and extract sports betting information that will increase the chances of winning bets in the sports gambling (Hajj et al., 2019). Therefore, this information needs to be segregated into subjective and objective information.

The task of segregating the subjective and the objective information has become burdensome to the decision maker when a voluminous of information is presented to them. The segregation of subjective and objective information becomes more challenging because this kind of information is usually unstructured, where formats are not specified before the creation of the information (Chaturvedi et al., 2018). The information creators are not bound to any specific sentence structure, choice of words or flow of information when it was created. In addition, some bilingual or multilingual speakers have the habit of mixing different languages in the text when expressing their thoughts or opinions.

The habit of mixing different languages in a text has increased the challenge to separate subjective information from objective information (Çetinoğlu et al., 2016). Mixing different languages in a single sentence is known as the code-switching (Muysken, 2000). Code-switching is a common scenario for most bilingual or multilingual speakers and writers. The term code-mixing was also used to describe the mixing of different languages for a