

The copyright © of this thesis belongs to its rightful author and/or other copyright owner. Copies can be accessed and downloaded for non-commercial or learning purposes without any charge and permission. The thesis cannot be reproduced or quoted as a whole without the permission from its rightful owner. No alteration or changes in format is allowed without permission from its rightful owner.



**MELEX: A NEW LEXICON FOR SENTIMENT ANALYSIS IN MINING
PUBLIC OPINION OF MALAYSIA AFFORDABLE HOUSING
PROJECTS**

NURUL HUSNA MAHADZIR



**DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2020**



Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

NURUL HUSNA MAHADZIR

calon untuk Ijazah
(candidate for the degree of)

PhD

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

**"MELEX: A NEW LEXICON FOR SENTIMENT ANALYSIS IN MINING PUBLIC OPINION
OF MALAYSIA AFFORDABLE HOUSING PROJECTS"**

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : **29 April 2020.**

*That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
April 29, 2020.*

Pengerusi Viva:
(Chairman for VIVA)

Assoc. Prof. Dr. Nerda Zura Zaibidi

Tandatangan
(Signature)

Pemeriksa Luar:
(External Examiner)

Assoc. Prof. Ts. Dr. Mustafa Man

Tandatangan
(Signature)

Pemeriksa Dalam:
(Internal Examiner)

Ts. Dr. Izwan Nizal Mohd Shahrane

Tandatangan
(Signature)

Nama Penyelia/Penyelia-penyelia:
(Name of Supervisor/Supervisors)

Assoc. Prof. Dr. Mohd Faizal Omar

Tandatangan
(Signature)

Nama Penyelia/Penyelia-penyelia:
(Name of Supervisor/Supervisors)

Assoc. Prof. Sr. Dr. Mohd Nasrun Mohd Naw

Tandatangan
(Signature)

Tarikh:

(Date) **April 29, 2020**

Permission to Use

In presenting this thesis in fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University 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

Analisis sentimen berpotensi sebagai alat untuk melakukan analisis bagi memahami kecenderungan awam. Ia telah menjadi satu bidang yang aktif dan semakin popular di dalam pencarian informasi dan perlombongan teks. Walaubagaimanapun, dalam konteks Malaysia, analisis sentimen masih terhad disebabkan kekurangan leksikon sentimen. Oleh itu, fokus kajian ini adalah untuk membangunkan leksikon yang baru dan meningkatkan ketepatan klasifikasi analisis sentimen dalam melombong pendapat umum bagi perumahan mampu milik di Malaysia. Leksikon yang baru untuk analisis sentimen dibangunkan dengan menggunakan pendekatan dwibahasa dan domain khusus leksikon sentimen. Kajian terperinci tentang pendekatan analisis sentimen di dalam kajian terdahulu telah dijalankan dan leksikon sentimen dwibahasa baharu yang dikenali sebagai MELex (Leksikon Bahasa Melayu-Inggeris) telah dihasilkan. Pendekatan yang dibangunkan ini dapat menganalisa teks di dalam dua bahasa utama yang digunakan di Malaysia iaitu Bahasa Melayu dan Bahasa Inggeris, dengan ketepatan yang lebih baik. Proses pembangunan MELex melibatkan tiga aktiviti iaitu pemilihan set perkataan awal, penetapan skor dan penambahan sinonim melalui perlaksanaan empat eksperimen yang berbeza. Penilaian adalah berdasarkan kepada pendekatan eksperimen dan kajian kes di mana PR1MA dan PPAM dipilih sebagai kes projek. Berdasarkan kepada keputusan perbandingan ke atas 2,230 data ujian, kajian ini telah menunjukkan klasifikasi menggunakan MELex adalah lebih baik berbanding pendekatan sedia ada dengan ketepatan yang dicapai adalah sebanyak 90.02% untuk PR1MA dan 89.17% untuk PPAM. Ini menunjukkan keupayaan MELex dalam mengklasifikasi sentimen awam terhadap projek perumahan PR1MA dan PPAM. Kajian ini telah menunjukkan dapatan dan keputusan yang lebih baik di dalam domain hartanah berbanding kajian lepas. Oleh itu, pendekatan berasaskan leksikon yang dilaksanakan dalam kajian ini dapat mencerminkan kebolehpercayaan leksikon sentimen dalam mengklasifikasikan sentimen awam.

Kata Kunci: Analisis sentimen, Leksikon sentimen, Leksikon Melayu-Inggeris, Pelombongan pendapat, Projek perumahan mampu milik

Abstract

Sentiment analysis has the potential as an analytical tool to understand the preferences of the public. It has become one of the most active and progressively popular areas in information retrieval and text mining. However, in the Malaysia context, the sentiment analysis is still limited due to the lack of sentiment lexicon. Thus, the focus of this study is to a new lexicon and enhance the classification accuracy of sentiment analysis in mining public opinion for Malaysia affordable housing project. The new lexicon for sentiment analysis is constructed by using a bilingual and domain-specific sentiment lexicon approach. A detailed review of existing approaches has been conducted and a new bilingual sentiment lexicon known as MELex (Malay-English Lexicon) has been generated. The developed approach is able to analyze text for two most widely used languages in Malaysia, Malay and English, with better accuracy. The process of constructing MELex involves three activities: seed words selection, polarity assignment and synonym expansions, with four different experiments have been implemented. It is evaluated based on the experimentation and case study approaches where PR1MA and PPAM are selected as case projects. Based on the comparative results over 2,230 testing data, the study reveals that the classification using MELex outperforms the existing approaches with the accuracy achieved for PR1MA and PPAM projects are 90.02% and 89.17%, respectively. This indicates the capabilities of MELex in classifying public sentiment towards PR1MA and PPAM housing projects. The study has shown promising and better results in property domain as compared to the previous research. Hence, the lexicon-based approach implemented in this study can reflect the reliability of the sentiment lexicon in classifying public sentiments.

Keywords: Sentiment analysis, Sentiment lexicon, Malay-English lexicon, Opinion mining, Affordable housing projects

Acknowledgment

All praise and thanks go to Allah s.w.t the Almighty, for giving me the strength and patience to complete this study. I want to acknowledge the enthusiastic supervision of my supervisor, Assoc. Prof. Ts. Dr. Mohd Faizal Omar and my co-supervisor, Assoc. Prof. Sr. Dr. Mohd Nasrun Mohd Nawi. Without their inspiration, excellent advice, guidance and active participation throughout the journey of my study, I would never have finished. Not forgetting, my appreciation goes to the Ministry of Higher Education Malaysia and Universiti Utara Malaysia for providing funds and the opportunity to conduct this study.

On a more personal level, I would like to thank my beloved parents and family members, who are always supporting me and praying for me to obtain this degree. I respect their deep faith, unconditional love, and support at each time of my life made me who I am today. Special thanks go to my daughter, Hasya Irdina for giving *Ummi* the best time of my life. Your smile blessed and encouraged me to complete this journey as soon as possible.

Besides, I wish to thank my dearest supporters particularly; Mama Fizah and Papa Joe, Kak Lina, Nik and Razman, Ana and Aizat, Wani and Laina for giving me their unequivocal support throughout this long journey, for which my mere expression of thanks, likewise, does not suffice.

Table of Contents

Permission to Use.....	i
Abstrak	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Tables.....	x
List of Figures	xii
List of Appendices	xiv
List of Abbreviations.....	xv
List of Publications	xvi
List of Awards and Recognitions.....	xvii
CHAPTER ONE INTRODUCTION	1
1.1 Overview	1
1.2 Background	1
1.3 Statement of the Problem	4
1.4 Research Questions	8
1.5 Research Objectives	9
1.6 Case Study.....	10
1.7 Significance of the Study	10
1.8 Scope of the Study	12
1.9 Thesis Structure.....	13
CHAPTER TWO MALAYSIA PROPERTY AND SOCIAL MEDIA.....	16
2.1 Introduction	16
2.2 Overview: Property Industry in Malaysia	16
2.3 Affordable Housing Projects.....	17
2.3.1 PR1MA	17
2.3.2 PPAM.....	18
2.4 The Issue in Property/Affordable Housing	19
2.5 Public Opinion and the Property Issue.....	21

2.6 Issues in Current Studies	22
2.7 Public Opinions on Social Media	22
2.8 Mining Social Media Data	24
2.9 Chapter Summary.....	26
CHAPTER THREE SENTIMENT ANALYSIS	27
3.1 Introduction	27
3.2 Overview of Sentiment Analysis	27
3.2.1 Machine Learning Approach	29
3.2.2 Lexicon-based Approach	29
3.3 Sentiment Analysis Applications	30
3.4 Non-English Sentiment Analysis	31
3.5 Mixed Language Sentiment Analysis	31
3.6 Languages Used in Malaysia	33
3.6.1 Malay Language.....	33
3.6.2 Mixed Language (<i>Bahasa Rojak</i>)	33
3.7 Sentiment Analysis in the Malaysian Context	34
3.7.1 Sentiment Analysis Tasks	35
3.7.2 Sentiment Classification Approaches	36
3.7.3 Sentiment Lexicon	42
3.7.4 Language Covered	44
3.7.5 Domain Applied.....	46
3.8 Research Gaps in Sentiment Analysis for the Malaysian Context.....	47
3.8.1 Domain-Specific Sentiment Lexicon	47
3.8.2 Analysis of Mixed Language	48
3.8.3 Analysis of Property Domain.....	48
3.9 Established Sentiment Lexicon	49
3.9.1 SentiWordNet	49
3.9.2 AFINN	50
3.9.3 SentiStrength.....	50
3.9.4 General Inquirer	50
3.10 Sentiment Lexicon Creation: Prominent Techniques	51
3.11 Word Vector Representation.....	53

3.12 Term Frequency	53
3.13 Sentiment Lexicon Creation for Malaysia Property Domain.....	54
3.14 Evaluation Measures	55
3.15 Manual Annotation Procedure	56
3.16 Chapter Summary.....	57
CHAPTER FOUR RESEARCH METHODOLOGY	58
4.1 Introduction	58
4.2 Research Design.....	58
4.3 Stage I: Theoretical Study	59
4.4 Stage II: Exploratory Study.....	60
4.4.1 The Selection of the Case Study	60
4.4.2 Sampling and Data Collection	61
4.4.3 Data Analysis Procedure.....	62
4.5 Stage III: Experiments.....	62
4.5.1 Identifying the Annotators	63
4.5.2 Determining the Pre-processing Activities	63
4.5.3 Determining the Lexicon Creation Technique.....	64
4.5.4 Determining the Resources to be Used	64
4.5.5 Determining the Classification Process	65
4.6 Stage IV: Performance Evaluation.....	65
4.6.1 Determining the Evaluation Criteria	65
4.6.2 Determining the Baseline Comparison	66
4.7 Chapter Summary.....	67
CHAPTER FIVE PRELIMINARIES.....	68
5.1 Introduction	68
5.2 Sentiment Analysis Framework	68
5.3 Datasets	69
5.4 Data Pre-processing	72
5.4.1 Removal of Re-Tweets.....	73
5.4.2 Removal of URLs, symbols and hashtags	74
5.4.3 Language Identification	74

5.4.4 PoS Tagging.....	75
5.5 Data Annotation	77
5.6 Data Categorization.....	80
5.7 Chapter Summary.....	81
CHAPTER SIX CONSTRUCTION OF MELEX.....	82
6.1 Introduction	82
6.2 MELEX: Overview	82
6.3 Definitions.....	83
6.4 Seed Words Selection	85
6.5 Polarity Assignment	86
6.5.1 Word Vector.....	86
6.5.2 Term Frequency	88
6.6 Synonym Expansion	89
6.6.1 English Resource	90
6.6.2 Malay Resource	91
6.7 Experimental Setup	91
6.7.1 Experiment 1: MELEX_v1.....	92
6.7.2 Experiment 2: MELEX_v2.....	94
6.7.3 Experiment 3: MELEX_v3.....	96
6.7.4 Experiment 4: MELEX_v4.....	98
6.8 Sentiment Classification	100
6.9 Performance Evaluation	104
6.9.1 Evaluation Metrics	105
6.9.2 Baseline Comparisons.....	106
6.9.2.1 General Purpose Lexicon	106
6.9.2.2 Machine Learning Classifiers.....	107
6.10 Chapter Summary.....	111
CHAPTER SEVEN RESULTS AND DISCUSSION	112
7.1 Introduction	112
7.2 Results	112
7.2.1 MELEX	112

7.2.2 Sentiment Classification	115
7.2.2.1 PR1MA.....	117
7.2.2.2 PPAM	121
7.2.2.3 Results Analysis	125
7.2.3 Performance Comparison.....	127
7.2.4 Misclassification	131
7.3 Discussions.....	135
7.3.1 Case Studies	135
7.3.2 MELex	136
7.3.3 Baseline Comparisons.....	139
7.4 Chapter Summary.....	141
CHAPTER EIGHT CONCLUSIONS AND FUTURE WORK	143
8.1 Introduction.....	143
8.2 Objectives of the Study: Revisited.....	143
8.3 Contributions.....	145
8.3.1 Practical Contributions.....	145
8.3.2 Methodological Contribution.....	146
8.3.3 Empirical Contributions.....	147
8.3.4 Dataset Contributions.....	148
8.4 Limitations	148
8.5 Suggestion for Future Research	150
8.6 Conclusion	153
REFERENCES.....	154

List of Tables

Table 3.1 Sentiment Classification	40
Table 3.2 Summary of Lexicon Constructions' Work	44
Table 5.1 Keywords Used to Retrieve Data	70
Table 5.2 PoS Tag and Its Definition	76
Table 5.3 Tweets and It's PoS Tags	76
Table 5.4 Language Annotation	79
Table 5.5 Training and Testing Data	80
Table 5.6 Total number of collected data	81
Table 6.1 Sample of Input	84
Table 6.2 Sample of Output	85
Table 6.3 Sample of Seed Word S	86
Table 6.4 Representation of the Word Vector Model	87
Table 6.5 Polarity Score P	88
Table 6.6 Seed Words and Its Synonyms	89
Table 6.7 Experimental Setup	92
Table 6.8 Samples of MELex_v1	94
Table 6.9 Samples of Seed Word, tf and Polarity Value	96
Table 6.10 Sample of MELex_v4	100
Table 6.11 Polarity Criteria	104
Table 6.12 Confusion Matrix	105
Table 7.1 Number of Words Generated	113
Table 7.2 Sample of the Output: Tweets and Its Polarity	116
Table 7.3 PR1MA: Type of Data	117
Table 7.4 Confusion Matrix: PR1MA - Mixed Language	118
Table 7.5 Evaluation Metrics: PR1MA - Mixed Language	118
Table 7.6 Confusion Matrix: PR1MA - Single Language	119
Table 7.7 Evaluation Metrics: PR1MA – Single Language	119
Table 7.8 Overall Performance - PR1MA	120
Table 7.9 PPAM: Type of Data	121
Table 7.10 Confusion Matrix: PPAM - Mixed Language	122

Table 7.11	Evaluation Metrics: PPAM – Mixed Language.....	122
Table 7.12	Confusion Matrix: PPAM - Single Language.....	123
Table 7.13	Evaluation Metrics: PPAM - Single Language.....	123
Table 7.14	Overall Performance – PPAM	124
Table 7.15	Performance Comparison: Mixed Language	128
Table 7.16	Performance Comparison: Overall Classification.....	129
Table 7.17	Examples of Misclassification Tweets.....	131
Table 7.18	Total No of Misclassification Data	132
Table 7.19	Comparison of Results	140



List of Figures

Figure 1.1. Social media overview: Malaysia. Adapted from “Digital 2020: Malaysia”	4
Figure 1.2. Thesis structure	13
Figure 2.1. The number of unsold property (2016-2018). Adapted from “NAPIC: Overhang units”	19
Figure 2.2. Property performance status in Quarter 3, 2019. Adapted from “NAPIC: Key Statistic’s Report 2019”	20
Figure 2.3. Statistics: Number of Malaysia Internet users. Adapted from “Statista 2020”	24
Figure 3.1. Sentiment analysis approach. Adapted from “Sentiment Analysis Algorithm and Applications: A Survey,” by W. Medhat, A. Hassan, & H. Korashy, 2014, Ain Shams Engineering Journal, 5(4), p. 3.	28
Figure 3.2. The workflow of a machine learning approach	29
Figure 3.3. The workflow of the lexicon-based approach	30
Figure 3.4. Sentiment analysis task	35
Figure 3.5. Language covered	45
Figure 4.1. Research design	59
Figure 5.1. Sentiment analysis framework	69
Figure 5.2. Sample tweets for PR1MA	71
Figure 5.3. Sample of data extraction	71
Figure 5.4. Sample raw data in Excel format	72
Figure 5.5. Pre-processing activities	73
Figure 5.6. Removal of RTs – Sample code	73
Figure 5.7. Removal of symbols - Sample code	74
Figure 5.8. Language identification – Sample code	75
Figure 5.9. PoS tagging – Sample code	75
Figure 5.10. Instructions for data annotation	78
Figure 5.11. Data annotation’s results	79
Figure 6.1. MELex’s architecture	83
Figure 6.2. Sample code of synonym expansion	91

Figure 6.3. List of negation words	102
Figure 6.4. Flowchart to determine the polarity	103
Figure 6.5. Sample of sentiment classification's output	104
Figure 6.6. Flow chart: Machine learning classification.....	107
Figure 6.7. Sample output of feature extraction using TfidfVectorizer	108
Figure 6.8. Sample output for machine learning classifiers.....	110
Figure 7.1. Sample MELex_v1 and MELex_v3	113
Figure 7.2. Sample MELex_v2 and MELex_v4	114
Figure 7.3. Word cloud: Most frequent sentiment words	115
Figure 7.4. Sentiment's result for PR1MA	120
Figure 7.5. Performance comparison	129
Figure 7.6. Sentiment analysis result for PR1MA	135
Figure 7.7. Sentiment analysis result for PPAM.....	136



List of Appendices

Appendix A Data Extraction using Twitterscraper	188
Appendix B Python Packages	189
Appendix C Selected Source Codes of MELex Development.....	190
Appendix D Samples of MELex_v1 / MELex_v3.....	192
Appendix E Samples of MELex_v2 / MELex_v4	193



List of Abbreviations

AIN	Artificial Immune Network
HTML	Hypertext Markup Language
k-NN	k-Nearest Neighbors
MELex	Malay-English Lexicon
MI	Mi-Intelligence
NB	Naïve Bayes
NLP	Natural Language Processing
PoS	Part-of-Speech
PR1MA	Perumahan Rakyat 1Malaysia
PPAM	Perumahan Penjawat Awam Malaysia
SNS	Social Networking Sites
SVM	Support Vector Machine
TF	Term Frequency
URL	Uniform Resource Locator

List of Publications

The work in this thesis has contributed to the following publications:

Mahadzir, N. H., Omar, M. F., & Nawi, M. N. M. (2016). Towards sentiment analysis application in housing projects. *Journal of Telecommunication, Electronic and Computer Engineering*, 8(8), 145-148. (Q3 Scopus)

Mahadzir, N. H., Omar, M. F., & Nawi, M. N. M. (2018). Semantic Similarity Measures for Malay-English Ambiguous Words. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 10(1-11), 109-112. (Q3 Scopus)

Mahadzir, N. H., Omar, M. F., & Nawi, M. N. M. (2018). A sentiment analysis visualization system for the property industry. *International Journal of Technology*, 9(8), 1609-1617. (Q2 Scopus)

Omar, M. F., **Mahadzir, N. H.**, Nawi, M. N. M., & Zulhumadi, F. (2019). Prototype Development and Pre-Commercialization Strategies for Mobile Based Property Analytics. *International Journal of Interactive Mobile Technologies (iJIM)*, 13(10), 198-204. (Q3 Scopus)

List of Awards and Recognitions

Gold Award at The International Invention, Innovation & Technology Exhibition (ITEX) 2018, KL Convention Centre, Kuala Lumpur. (*Project Title: PropertyInsights*)

Gold Award at The Industry Networking and Business Pitching (eREKA), UniMAP 2018, Perlis. (*Project Title: Social Media Analytics for Malaysia Property Industry*)

IP Registered

PropertyInsights

Filing No. MyIPO : LY2018002827

Filing Date : 2 Aug 2018



CHAPTER ONE

INTRODUCTION

“The four most important words in the English language are, ‘What do you think?’

Listen to your people and learn.”

(J. W. Bill Marriot Jr.)

1.1 Overview

This introductory chapter began with the background of the study, followed by a discussion of the problem. Thereafter, research questions are formulated and used to construct the research objectives. Next, the case study used in this research, the significance of the study as well as the scope of the study is presented. Finally, the outline of the remaining chapters of this thesis is also presented.

1.2 Background

Housing affordability is considered as a global issue around the world, including Malaysia. Furthermore, the affordability problem concerning the property industry is one of the most common problems within most developed and developing countries (Salfarina, Malina & Azrina, 2010).

The Malaysian government has addressed the need for affordable housing under the Eleventh Malaysia Plan, especially for the bottom 40% of the household income group (B40) to alleviate the rising cost of living. In fact, the government is targeting to provide 606,000 new affordable houses spanning from 2016 to 2020 (Mottain, 2017).

The agenda would be a continuation of a few initiatives such as Program Perumahan Rakyat 1Malaysia (PR1MA), Perumahan Penjawat Awam Malaysia (PPAM) and Rumah Mesra Rakyat. While it is commendable that the government is trying its best to help the *rakyat* (citizen) to afford a house of their very own such as the implementations of several affordable housing projects as mentioned above, it remains to be seen whether their efforts can be called a success.

In order to measure the success, there is a suggestion by the industry experts, Datuk Steward Labrooy to gather the information on the housing needs and satisfactions from the Malaysian citizens especially the aspiring buyers. Moreover, he pointed out that the property industry is crucially in need of a big data analysis in making better decisions on housing development (Ng, 2019; Rafee & Wai, 2019; Rosli, 2019). Gathering such information from the public is vital to provide insights into the real thoughts of people, and this challenge is the object of research in the discipline called “sentiment analysis” (Liu, 2012).

Sentiment analysis has the potential as an analytical tool to understand the preferences of the public. It is a field of research in Natural Language Processing (NLP) that aims to automatically detect and classify the opinion expressed through text (Cambria, Schuller, Xia, & Havasi, 2013; Liu, 2012; Pang & Lee, 2008). In recent years, there is an increase in interests by many organizations and companies towards the application of sentiment analysis which proved the arisen importance of this field. Furthermore, sentiment analysis has been applied to a wide variety of topics and issues as reported in previous research such as online products reviews (Mukherjee & Bhattacharyya 2012), hotel reviews (Kasper & Vela, 2011), political and financial analysis (Chan &

Chong; 2017; Schumaker, Zhang, Huang, & Chen, 2012; Thanvi, Sontakke, Waghmare, Patel, & Gavhane, 2017).

At present, research work on sentiment analysis has been dominated by two approaches; machine learning and lexicon-based. The machine learning approach aims to build classifiers by extracting features and algorithms from trained data. The other is the lexicon-based approach, which utilizes lexical resources like sentiment lexicons or dictionaries, to determine the polarity (Pang & Lee, 2008). In the latter approach, sentiment lexicon plays a very significant role as it provides information of opinionated words with its associated category such as positive or negative polarity.

To date, a massive volume of studies has been implemented in mining the sentiment written in a single language, especially English. However, to perform sentiment analysis in the Malaysian context, two things need to be considered. First, sentiment analysis should be applied for the Malay language as *Bahasa Melayu* is the national language. Second, Malaysians tend to mix both Malay and English language known as *Bahasa rojak* mainly when they write on social networking sites (SNS). Previous sentiment analysis research is limited in fulfilling these two needs.

In this thesis, sentiment analysis using the lexicon-based approach is applied to two well-known affordable housing projects in Malaysia which are PR1MA and PPAM. A thorough search of the relevant literature yielded that this research is among the first work to apply sentiment analysis for property projects and, hopefully, it will be a valuable mechanism for the government to improve the execution of the project plan in the future.

In demonstrating the performance of the proposed approach, the experimental studies have been conducted. The results obtained in both projects consistently show that the classification using the newly created lexicon called MELex (Malay-English Lexicon) has performed better than the state-of-the-art and increased the performance.

1.3 Statement of the Problem

The digital landscape in Malaysia has evolved over the past few years, and it changes the way Malaysians communicate with each other, how they express their thoughts, and how they make decisions.

In Digital 2020: Malaysia's report as presented in Figure 1.1, the results have shown that 26 million Malaysians are active social media users. In fact, researches indicate that Malaysians allocated approximately 8 hours daily in surfing SNS (Statista, 2019).

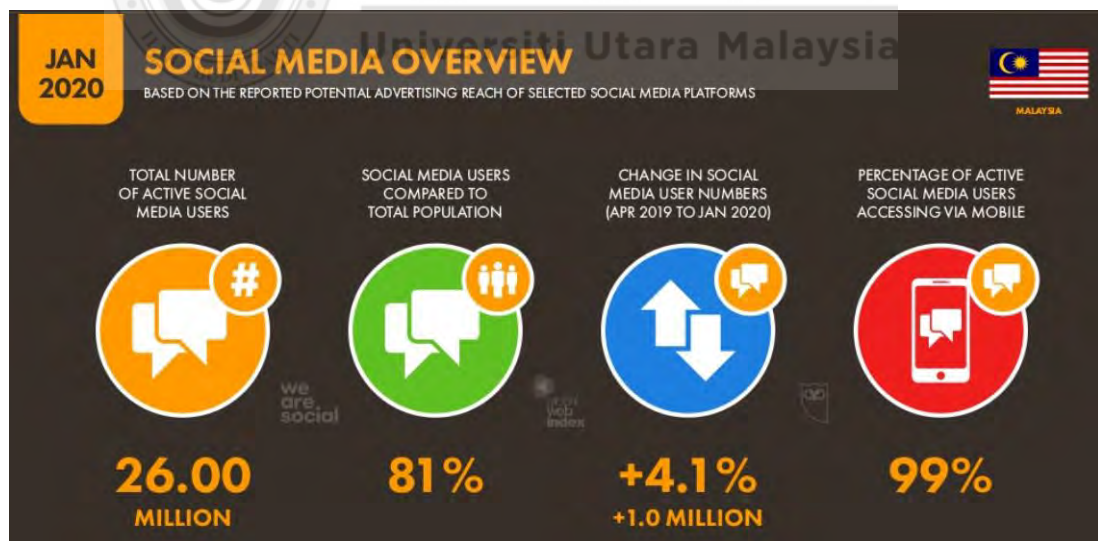


Figure 1.1. Social media overview: Malaysia. Adapted from “Digital 2020: Malaysia”

Based on the statistic given, it can be seen that SNS has created a new way of communications. Besides, this kind of platform does facilitate real-time marketing which takes business one step ahead by enabling brands to engage with their consumers. Moreover, it allows the business to be close to the target audience, enables the companies or organizations to take direct action in satisfying their customers, produce insights that facilitate the decision-making process and engage in driving business results.

As a popular SNS, Twitter has snowballed into an essential communication tool and become one of the most visited websites in the world (Hardwick, 2019). A growing number of people are voluntarily posting their thoughts and reactions through the Twitter platform, which offered a valuable online source for gathering public insights. However, for many industries and businesses in Malaysia, this low-cost, high reach channel that has 24 million active users appears to be missed out and neglected.

In this thesis, there are two issues have been addressed which are:

- 1. The needs of social media analysis to understand public sentiments towards the property industry in Malaysia.**

Currently, there are many concerns regarding the property market situation in Malaysia, for instance, where it is heading in the next three to five years and challenges it may face in the future. Experts have confidence that the market fundamental is still resilient, and people still have the power, but if that element is being put aside, market confidence seems to remain lacking, especially among first-time house buyers (Begum, 2018).

Public reviews on property shared through SNS have become very influential information sources that impact the property industry in many ways. At present, only few researches involving traditional research methods such as interview, questionnaires and surveys which aimed at only specific group of people have been carried out to gather the public reviews on Malaysia property industry (Chan & Lee, 2016; Jamaluddin, Abdullah & Hamdan, 2016; Mustafa, Adnan & Nawayai, 2017).

However, the traditional research methods are known to be restricted to a set of questions that are sometimes forced onto people, who might not give candid and straightforward answers. While the relationship between the sentiment of public reviews, and the growth of user-generated content has been demonstrated to have a connection to the performance of industries, there have been limited studies of this genuine impact in the area of the property (Murphy et al., 2014; Yu, Duan, & Cao, 2013).

Furthermore, the posts and comments shared in SNS are considered as honest responses from the public as they post their thoughts without being asked for it. Since the ease for the public to review property industries can only increase in time, creating a quantifiable impact for the shareholders and decision-makers to understand what the public is saying is necessary and invaluable to the property industry.

However, to date, there is no known research study targeting the SNS platform as well as broad coverage of the audience that has been conducted to gather such information appropriately.

2. The importance of covering both Malay and English languages in the sentiment analysis process.

Since the early 2000s, sentiment analysis has become one of the most active research areas in NLP (Liu, 2012). To date, sentiment analysis has been applied to various domains such as products, movies, sports and political reviews.

The majority of the previous research in this field has concentrated on analyzing a single language only, especially English. Nevertheless, with the need for globalization, it is quite common to see the post written in multiple languages especially in SNS which makes the sentiment analysis process even harder and more challenging. The amount of textual data produced in multiple languages is so massive that it introduces many challenges for researchers wanting to perform sentiment analysis on the data.

Besides, in an unstructured content such as Twitter posts, people tend to mix languages in one single sentence. According to Dashtipour, Poria, Hussain and Cambria (2016), specific information in another language might be left out if the analysis is done for a single language only.

To thoroughly analyze the public sentiments towards the property industry in Malaysia, it is crucial to perform sentiment analysis for both Malay and English languages because Malay or *Bahasa Melayu* is a native language in this country while English is the second used language in conversation by Malaysians. Unlike English, Malay sentiment analysis did not receive much attention in the prior works.

1.4 Research Questions

The following research questions are to be answered at the end of this study:

RQ1: *What are the techniques available in constructing a bilingual and domain-specific sentiment lexicon?*

RQ2: *How to develop a sentiment lexicon that is bilingual and domain-dependent?*

RQ3: *What are the sentiments of affordable housing projects written in single and mixed-language content?*

RQ4: *Does the performance improve by using the developed lexicon as compared to the state-of-the-art sentiment analysis technique?*

The motivation behind the first research questions is the investigation of the prominent technique applied in the previous studies to construct sentiment lexicon. This would help to show the different techniques or methods that can be employed to build the lexicon in a better way. The second research question investigates the possibility of developing a sentiment lexicon that may serve two purposes; bilingual and specific to one particular domain. The third research question arises in order to know the results of sentiment classification for both data types; Malay and *Bahasa rojak*, so that the impact and the significance of analyzing both contents can be further investigated. The last research question concerns the improvement of sentiment analysis performance following the implementation of sentiment classification using the developed lexicon.

1.5 Research Objectives

Based on the considerations mentioned earlier, the main aim of this research is to explore an effective way to perform sentiment analysis for the property domain in the Malaysian context. Therefore, to cater to this aim, there are four objectives itemized as below:

RO1: *To identify the techniques used in developing a bilingual and domain-specific sentiment lexicon.*

RO2: *To construct and develop a new sentiment lexicon for Malay and English languages specifically for the property domain.*

RO3: *To perform sentiment classification for affordable housing projects written in single and mixed language by using the constructed lexicon.*

RO4: *To evaluate the performance of the proposed approach.*

The first objective intends to identify possible techniques that could be used to construct the new sentiment lexicon. Secondly, a Malay-English and domain-specific sentiment lexicon needs to be constructed and is used to determine the polarity. The sentiment lexicon contains the words with their sentiment score. The third objective is to classify the sentiments of affordable housing projects based on the constructed lexicon. The last objective is to report on the performance of sentiment classification using the proposed approach in this research.

1.6 Case Study

This research run in the confines of a case study. Since the focus of this research is to perform sentiment analysis for the Malaysia property domain, two well-known government's affordable housing schemes for Malaysians; PR1MA and PPAM were used as a case study in order to investigate the above research objectives.

PR1MA project is offered to all Malaysians with the household income of between RM2,500 and RM15,000 monthly while the PPAM scheme is supplied only for government servants.

These government's affordable housing projects were a preferred case study due to the current market mismatch of supply and demand issues faced by both projects. Besides, the availability of the data for these two projects through the Twitter platform is considered enough in evaluating sentiment analysis performance as proposed in this study. The discussion on both projects was detailed out in Section 2.3.

1.7 Significance of the Study

With microblogging been growing in popularity worldwide, people have started to voice out their thoughts and opinions on a wide variety of topics and events on these platforms. Hence, sentiment analysis application towards product or service reviews through SNS has provided an effective way of assessing public opinion for business marketing or service improvement.

This study is expected to serve as a blueprint for property companies as well as governments who wanted to know the public sentiments through SNS for their marketing or service improvement purposes. Specifically, this research is aimed at assisting the decision-makers in the property industry. From a practical point of view, the significance of this study lies in the implementation of sentiment analysis using SNS platforms in gaining public insights towards the property domain in Malaysia. This research can provide insight into the relationship between the property business and public reviews in a meaningful and quantifiable way.

Since this is an applied study, industry practitioners could replicate the approach used in this study to gain insight and understanding of their customer base through the application of sentiment analysis. Besides, this research is relevant to the bilingual sentiment analysis research area. Most of the sentiment analysis implemented for the Malaysian context has catered only for a single language which is Malay. However, the analysis of information conveyed in *Bahasa rojak* is not well established. A specific challenge is encountered when attempting to deal with *Bahasa rojak* content found in informal platforms such as Facebook and Twitter. *Bahasa rojak* content in these sites is generally characterized to be written in a highly informal Malay and English language that is used in native speaking.

Theoretically, the findings of this study can be utilized for handling mixed language content. With the ability to classify mixed language contents that frequently appear in unstructured online platforms, it contributes to the accuracy improvement of the sentiment analysis.

The experiment results show the promising classification performance of using the developed bilingual lexicon; MELex. If significant, it could further the development of social media analysis, encourage academic study into public sentiment for the property industry, and encourage the usage of more advanced lexicon-based approach for the usage of complex studies.

1.8 Scope of the Study

This study focuses on the implementation of sentiment analysis for the Malay and English languages only. Malay (officially known as *Bahasa Malaysia*) is the official language of Malaysia and it is the most widely spoken language in the country. Even though there are many races residing in Malaysia, but nearly every Malaysian can speak both Malay and English and they use these two languages in their daily lives; either in speaking or writing.

The data used in this research is extracted from a single platform; Twitter. One of the characteristics of Twitter is that it is limited to 140 characters. This is different from the other platforms such as Facebook or blogs, which are usually long. The number of active Twitter users in Malaysia has increased every year and it nearly reached up to 2.4 million in 2019 (Statista, 2019). In addition, it can be observed that there is high interest by the Malaysian to give their thoughts or opinions through this platform.

The lexicon generated is specific for the property domain as the purpose of this study is to increase the accuracy by developing domain-specific sentiment lexicon. As for the sentiment classification, this study employs word-level classification and involves manual labeling in constructing a bilingual sentiment lexicon.

Even though it is labor-intensive and time-consuming, manual labeling still needs to be done due to the absence of training data and it is renowned that a classifier may perform better in the domain that is trained.

1.9 Thesis Structure

This thesis consists of eight chapters. Figure 1.2 highlights the structure of this thesis as well as the research objectives.

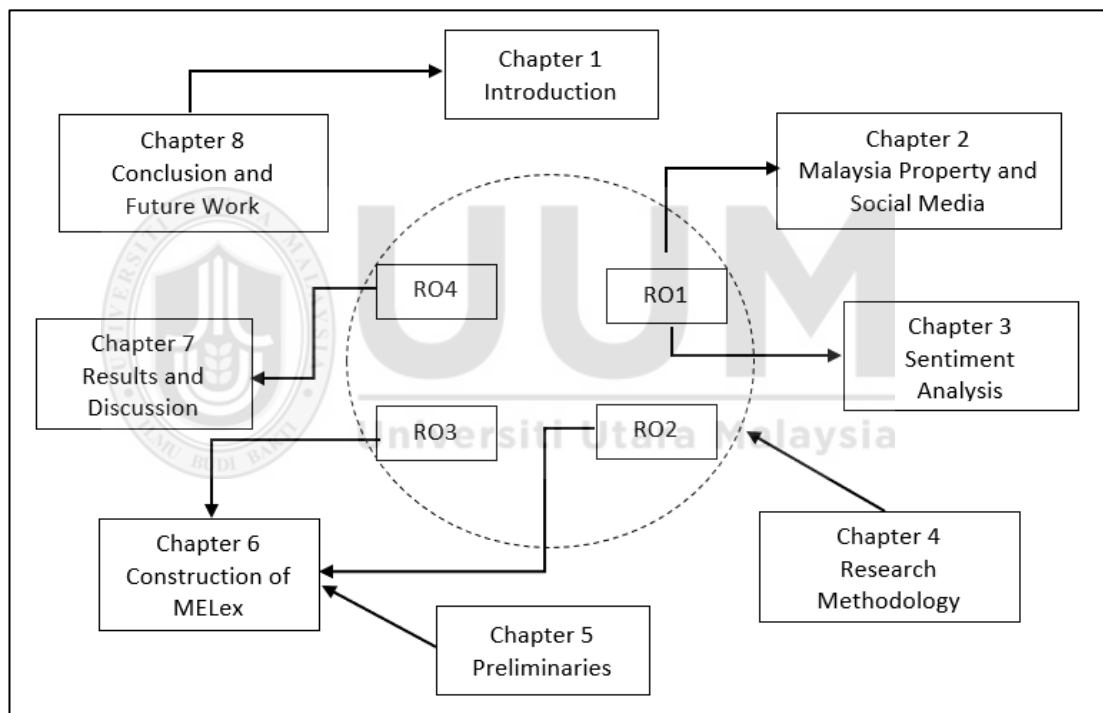


Figure 1.2. Thesis structure

In this chapter, the research background and statement of the problem are presented. The main objectives of this study are also explained. The remaining chapters of this thesis are structured as follows:

Chapter 2 – Malaysia Property and Social Media.

An overview of the property industry in Malaysia and issues related to this domain are presented. In particular, the focus is given to the government's affordable housing projects. The relations between the issue highlighted and social media were elaborated in this chapter.

Chapter 3 – Sentiment Analysis.

This chapter focuses on examining the literature to learn from previous research and give insights on the topic. Specifically, the sentiment analysis research in the Malaysian context is discussed in detail and the research gaps are identified. This chapter mainly addresses Research Objective 1.

Chapter 4 – Research Methodology.

This chapter outlines the work and research activities to be carried out. It discusses the research methodology that was used to achieve the objectives of this study. The four stages conducted in order to carry out the research activities are elaborated in detail.

Chapter 5 – Preliminaries.

In this chapter, the task to be done prior to the construction of sentiment lexicon is thoroughly discussed which includes preprocessing activities, data annotation and data categorization. Datasets used for training and testing sets are described in detail.

Chapter 6 – Construction of MELex.

This chapter details out the activities to develop a new sentiment lexicon, followed by sentiment classification tasks as well as performance evaluation. Different strategies to generate an excellent bilingual sentiment lexicon are discussed and implemented. Research Objective 2 and 3 are addressed in this chapter.

Chapter 7 – Results and Discussion.

The final research objective (Research Objective 4) is investigated in this chapter. The results obtained from the experiments are presented and the performance as well as the evaluation is thoroughly discussed. The proposed approach is compared to some baselines and other state-of-the-art classifiers. The error analysis is also provided in this chapter.

Chapter 8 – Conclusions and Future Work.

The last chapter concludes the thesis and each research objective is revisited and summarized how this research addresses them. The limitations of the study as well as several potential future directions are highlighted.

CHAPTER TWO

MALAYSIA PROPERTY AND SOCIAL MEDIA

2.1 Introduction

This chapter offers an overview of the property industry and affordable housing projects in Malaysia. Specifically, this chapter highlighted the current property issue faced by the Malaysian government and identified the gaps in the previous research concerning the property domain. Besides, this chapter emphasized the criticality of mining social media to obtain public opinion on the subject matter.

2.2 Overview: Property Industry in Malaysia

Owning a home is considered as a basic human need, along with food and water to live a comfortable life. It is stated in Article 13 under the Federal Constitution of Malaysia that everyone has the right to own property (Bari & Shuaib, 2009). Besides, the property industry has been one of the biggest and significant sectors in the Malaysian economic growth. In fact, the Malaysian government always prioritizes this sector to guarantee that all Malaysian at any income level group have equal opportunities to have quality, and affordable housing in this country (Osman, Khalid, & Yusop, 2017).

Generally, the government in any country is responsible for providing affordable, adequate and quality housing for the citizens. In Malaysia, the government focuses mainly on the Bottom 40 (B40) and Middle 40 (M40) income level groups to own

property. For that matter, there are various affordable housing programs that have been implemented to achieve this goal.

2.3 Affordable Housing Projects

Khor (2019) defines affordable housing as a property that is restricted for households with specific income requirements and sufficient in terms of quality and location. Besides, the price of affordable housing is not so high, which enables its occupants to fulfill other essential living needs.

The affordability of housing has always been a worldwide concern, including Malaysia. A report produced by Khazanah Research Institute (KRI) in 2015 indicated that the house price was 4.4 times the median annual household income which makes the Malaysian property market as 'seriously unaffordable' (Suraya, 2015).

The government is obviously doing its part by always prioritizing the lower and middle-income level groups in any policy or initiative, including the housing sector. It is to ensure that every household can afford a shelter of their own. For instance, the Malaysian government had established several affordable housing projects such as the Perumahan Rakyat 1Malaysia (PR1MA) and Perumahan Penjabat Awam 1Malaysia (PPAM) as the catalyst in providing adequate, quality and affordable houses.

2.3.1 PR1MA

The Perumahan Rakyat 1Malaysia (PR1MA) is one of the most well-known government's affordable housing projects in Malaysia. It was initiated in 2011 to offer affordable houses for households in the middle-income group in urban areas.

Precisely, this scheme is provided for Malaysian citizens with monthly income starting from RM2,500 to RM4,000 and houses priced at the range of RM100,000 to RM400,000 in metropolitan areas.

The PR1MA Homes are established not just to serve the national agenda of developing high-quality yet affordable and comfortable houses for the bottom and middle-income citizens, but they are also to assist and encourage homeownership among those interested buyers who are facing challenges in buying a property.

2.3.2 PPAM

PPAM (formerly known as PPA1M) is a government initiative which solely benefits civil servants. It was launched in early 2013 which aimed to help the civil servants in owning assets at an appropriate price. This affordable housing initiative was designed to ensure that low-income and middle-income government servants could afford homes in major urban areas. It operated by encouraging private developers to actively involved in PPAM developments, which are then subsidized by the government.

PPAM projects were designed with the following criteria in mind; high quality yet affordable housing, been built in areas with great interest by civil servants, located in major cities and offered a variety of housing types including high-rise and landed with the price range from RM100,000 to RM400,000 per unit. The government is expected to complete 4,245 housing units under the PPAM project nationwide by 2020 (Bernama, 2020).

2.4 The Issue in Property/Affordable Housing

At present, the property market in Malaysia is facing a crucial supply and demand imbalance. The Central Bank of Malaysia (BNM) has reported that the supply and demand mismatch in Malaysia's property market has started since 2015, with unsold residential properties already at its highest in 10 years (Ling, Almeida, Shukri, & Sze, 2017). The leading property consultants, Knight Frank Malaysia, predicted that Malaysia's property market to be moving slower in 2018 and will remain challenging even in 2019 (Zakariah, 2019).

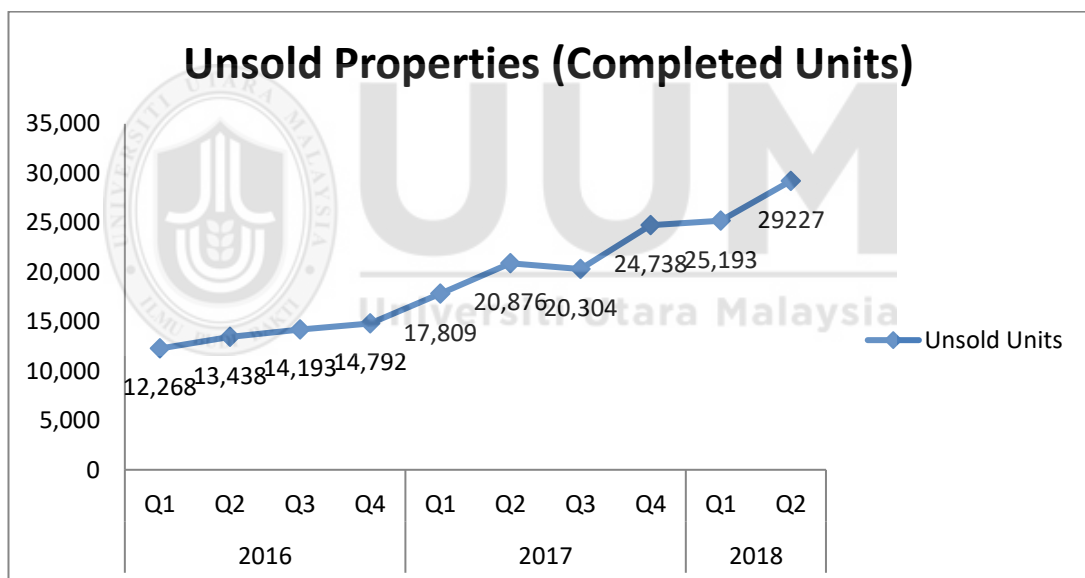


Figure 2.1. The number of unsold property (2016-2018). Adapted from “NAPIC: Overhang units”

Figure 2.1 shows the property overhang data as reported by NAPIC from 2016 until the second quarter of 2018. It can be seen that 27.64% or 29,227 out of 105,753 are the highest numbers of the unsold residence units.

In the latest report produced by NAPIC, the number of unsold properties shows no sign of decreasing. As illustrated in Figure 2.2, there are 124,179 unsold units as in the third quarter (Q3) of 2019. In fact, the property overhang has increased by 1,865 units as compared to the second quarter (Q2) of 2018.

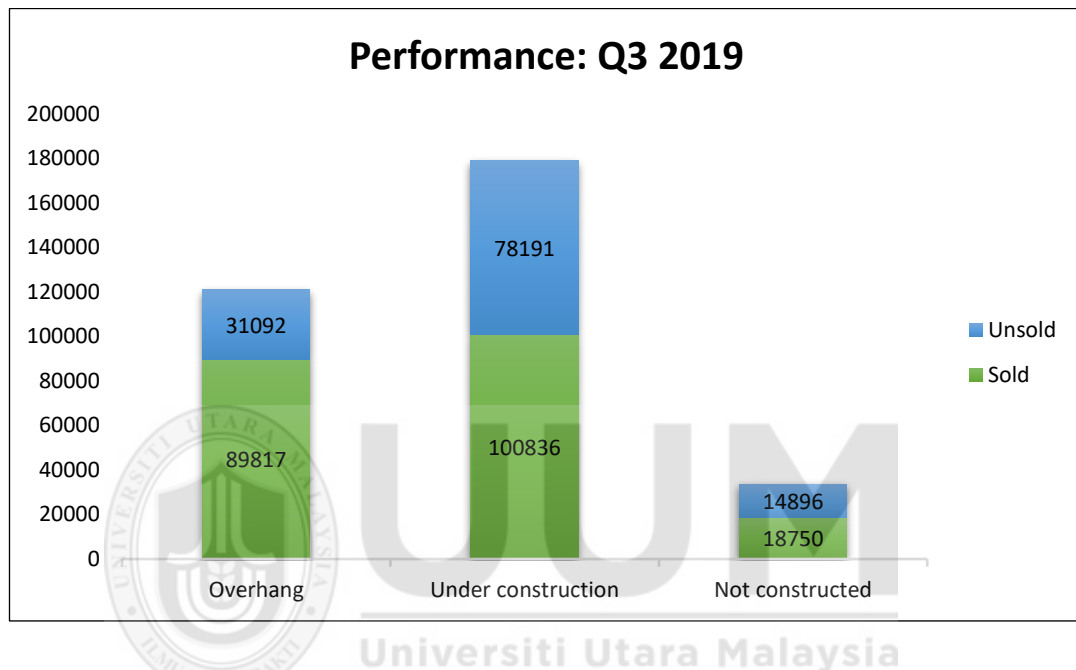


Figure 2.2. Property performance status in Quarter 3, 2019. Adapted from “NAPIC: Key Statistic’s Report 2019”

The issue of property overhang is not only affecting the private property units, but PR1MA homes contribute to the numbers as well. As of November 2017, the highest number of unsold property units in Kedah state was PR1MA homes and Johor stood second with more than 18% of total overhang nationwide (Young, 2017). This statistic shows that the factors and causes of the massive unsold property issue crucially need to be addressed to ensure that this particular initiative can ultimately achieve its goal.

2.5 Public Opinion and the Property Issue

There is a clear indicator that the high amounts of property overhang neither appeal to the target market nor caters to the actual needs and requirements of the property buyers. In fact, the efficiency of the housing delivery system is measured based on how effective public and private housing developers are in regulating their real estate activities to suit the budget, needs and wants of the household (Teck-Hong, 2012). Based on the statement by the Executive Director of property consultancy Jones Lang Wootton, Prem Kumar, this problem shows that the property players are in a severe lack of up-to-date information needed in order to make informed decisions. One of the initiatives as a way forward suggested by him is to seek the public's opinion to ensure their property development is within the demand profiling (Wong, 2018). He continued that the missing piece in the Malaysian property industry is the availability of big data which could provide better information concerning market trends. His statement was supported by the Housing and the Government Minister, Zuraida Kamaruddin who agreed that the lack of a big data system had obstructed the government's ability to understand the local housing needs (Rosli, 2019).

Big data refers to a complex and massive data sets either in a structured or unstructured format (Sivarajah, Kamal, Irani, & Weerakkody, 2017; Taylor-Sakyi, 2016). It is considered a new environment of collecting, storing and processing data which includes the use of social media platforms such as Google, Twitter, Instagram and Facebook where nearly 2.5 quintillion bytes of data are generated daily (Marr, 2018). For that matter, big data has become a new option in gauging and analyzing public opinion towards certain subjects or events.

2.6 Issues in Current Studies

Research on gathering and analyzing public opinion towards the property demand is still limited and incomplete. Salfarina et al. (2010) focused on the urban housing needs and issues in Malaysia while Lim, Olanrewaju, Tan, and Lee (2018) and Maimun et al. (2018) studied the factors influencing the demand for affordable housing. The same line of study has been done by Zainon, Mohd-Rahim, Sulaiman, Abd-Karim, and Hamzah (2017) to find the influential factors behind the decision of the property purchase among middle-income groups in the Klang Valley area. The research methods implemented by all the studies mentioned above are purely based on a quantitative approach; survey (Leh, Mansor, & Musthafa, 2017; Salfarina et al., 2010) and questionnaire (Lim et al., 2018; Zainon et al., 2017).

Hence, none of the studies has utilized social media platforms in gathering information on public preferences. The significant scarcity of using surveys and questionnaires is the respondent might not give a truthful answer in order to protect their privacy and it is hard for them to convey emotions and feelings through a limited set of questions.

2.7 Public Opinions on Social Media

Social media has been an essential platform for most people around the world to voice out their feelings and thoughts. In fact, the information they share through this mechanism has long been considered as truthful feedback as they voice out their opinions without being asked for it.

This phenomenon has a significant impact on governments, corporations, business owners, and decision-makers in obtaining end-user feedback towards their products or

services. It will no longer be necessary to conduct surveys, questionnaires, employ external consultants or organize focus groups to get consumer opinions about a particular matter because the online platform can already give them such information.

The Internet and SNS in Malaysia are seen as vibrant, with the majority of the Malaysian population opting to this digital platform to voice out their opinions. Recent statistics have shown that about 75% of the Malaysian population are active social media users and Facebook, Instagram, as well as Twitter, are among the social media of choice for Malaysians (“Active social”, 2019). Besides, Malaysian SNS users allocated an average of five hours and forty-seven minutes daily across platforms. Figure 2.3 visualizes the number of Internet users in Malaysia starting from 2017 and the number is expected to be continuously increasing in the coming years (Statista, 2020).

Instead of commenting on online businesses, Malaysian citizens also voice their opinion towards government initiatives through SNS, particularly involving property projects. This scenario creates an excellent opportunity for government or organizations to get honest feedback and truly understand people’s thoughts and feelings over any issues.

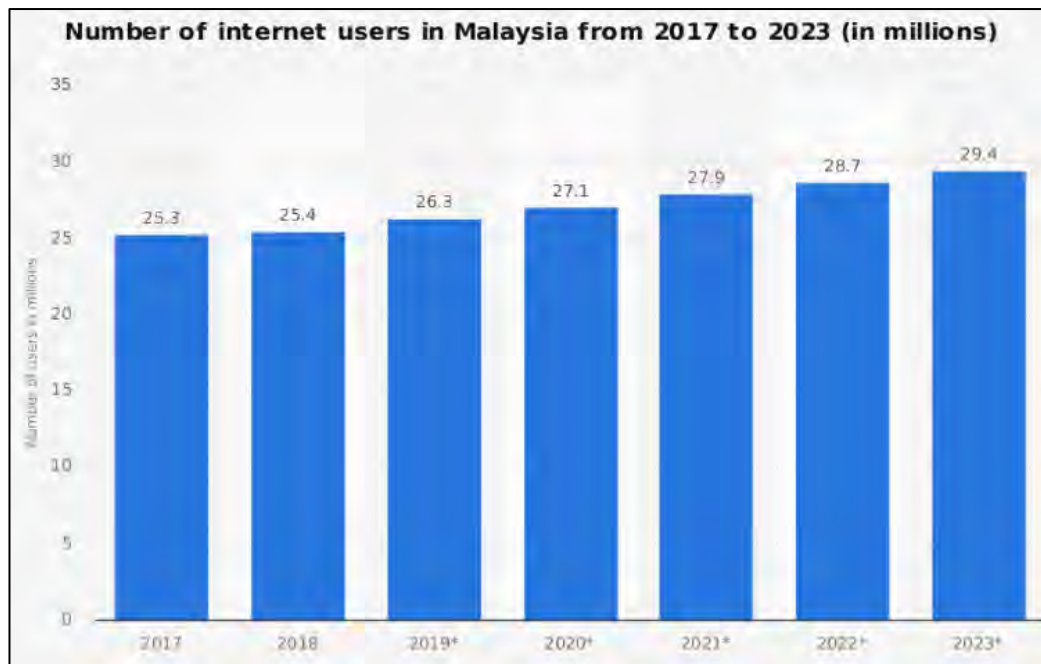


Figure 2.3. Statistics: Number of Malaysia Internet users. Adapted from “Statista 2020”

SNS are increasingly used by the general public as a platform to express their concerns and discuss controversial issues. This information could be utilized for the purposes of solicitation of public opinions towards property issues. However, current studies in property reviews are lacking in terms of analyzing public opinion through broader coverage such as SNS. The next section reviews several well-known kinds of research works that have been done for analyzing social media data.

2.8 Mining Social Media Data

The focus of mining social media data is mainly on the massive user-generated content that is being produced each day by the users. This phenomenon is likely to continue

with exponentially more content in the future. The data generated in this platform are plentiful and diverse, which makes them a relevant source for data science. Unlike traditional methods such as surveys and questionnaires, the analysis of social media content promises powerful new ways of knowing the public, their preferences and capturing what they say and do.

The works devoted to the analysis of social media data falls under the field of data mining and NLP which includes sentiment analysis (Cambria, 2016; Medhat, Hassan, & Korashy, 2014), trending topics detection (Papadopoulos, Corney, & Aiello, 2014; Peng, Tseng, Liang, & Shan, 2018) and events detection (Dong, Mavroeidis, Calabrese, & Frossard, 2015; Zhou & Chen, 2014), to name a few.

Sentiment analysis is a trendy ongoing and well-established field of research that determines people attitude towards particular topics or issues and classify them into positive or negative sentiments (Sapountzi & Psannis, 2016). There are many sophisticated methods and techniques that have been developed to gauge sentiment from the text (Liu, 2012).

The application of sentiment analysis is useful for mining public opinions on products or services through their reviews or online posts. For example, sentiment analysis was shown to complement and inform public opinion polling when several surveys conducted on political opinion as well as consumer confidence in 2009 were found to associate with sentiment term frequencies in Twitter posts over the same period (O'Connor et al., 2010).

Similarly, there is evidence that the moods of the nation, as measured by tweets, correlate with changes in stock prices (Bollen et al., 2011). Also, sentiment analysis has been implemented to predict box-office revenue for movies (Jain, 2013).

Another inclusion in this research work is trending topic detection. It is a task to tell what topic is trending and to know what is currently happening in the real world (Georgiou, El Abbadi, & Yan, 2017). Event detection is more focused on reporting real-life occurrences that unfold over time. For example, Sakaki, Okazaki, and Matsuo (2010) have utilized the Twitter platform to predict earthquake and Benson, Haghighi, and Barzilay (2011) identified new musical events through what has been mentioned by Twitter users. In the case of this study, sentiment analysis is seen as the most relevant because knowing the sentiment of the citizens' posts and comments would provide much actionable knowledge of appropriate interventions and services for the public.

2.9 Chapter Summary

In this chapter, an overview of the property industry in Malaysia, as well as projects related to affordable housing were presented. The discussion focused on the issue related to the property, which leads to rising imbalances of supply and demand. Besides, this chapter addressed the importance of obtaining public opinions as one of the ways forward to resolve the issue. In particular, the need for social media analysis to gain public insights towards property is explained which later can be a useful mechanism for the property players and governments in understanding the factors affecting the slow property demand from the public perspective.

CHAPTER THREE

SENTIMENT ANALYSIS

3.1 Introduction

As explained in the previous chapter, there is a great need for mining public opinions in the property industry as well as the relevancy of sentiment analysis in achieving this goal. This chapter presents a review of the existing literature related to sentiment analysis in general, as well as research specifically for the Malaysian context. The review includes all the focus areas in the previous study and the current research gap as well as the challenges that this research work seeks to address.

3.2 Overview of Sentiment Analysis

Sentiment analysis is a process within the field of NLP to analyze and determine the polarity of the opinion or emotion expressed in a text document especially on the Web (Liu, 2012; Pang & Lee, 2008). For the past decades, tremendous research on sentiment analysis has been conducted for the significant benefit it brings to the development of various domain areas such as economy, marketing and politic. The significance of this research field has been acknowledged by the great number of techniques and approaches proposed in the previous study and become one of the reasons for its rapid development, as well as by the interest of companies and agencies that it raised over the past few years.

One of the key activities in sentiment analysis is sentiment classification where it concentrates on categorizing opinionated text with various polarities such as positive or negative (Montoyo, Martínez-Barco, & Balahur, 2012). Several previous research works focused on categorizing text into positive, negative or neutral (Pak & Paroubek, 2010) while others consider more fine-tuned classification such as highly positive, positive, neutral, negative or highly negative (Ortega, Fonseca, & Montoyo, 2013) in their classification.

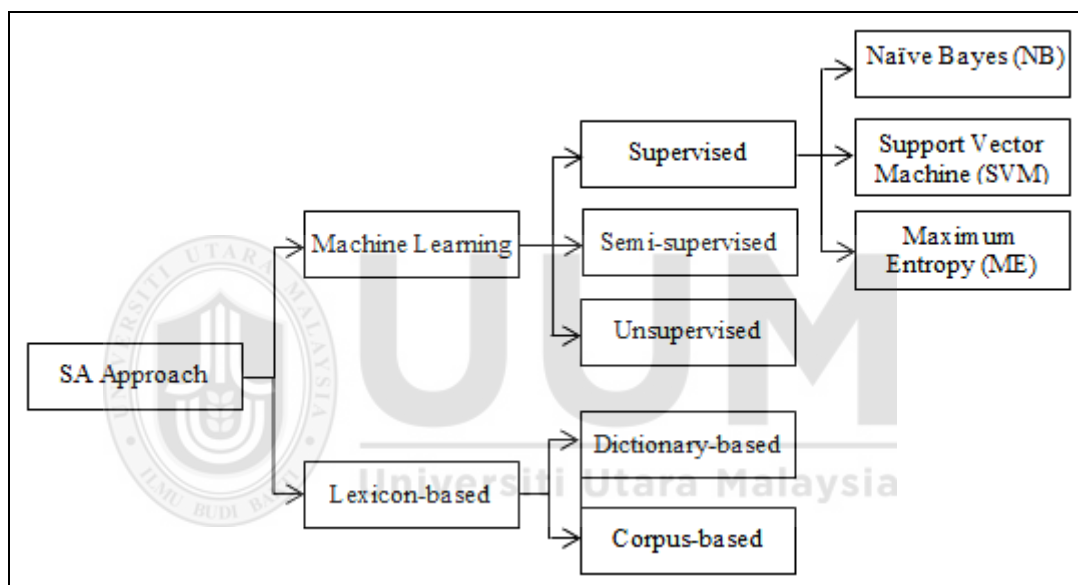


Figure 3.1. Sentiment analysis approach. Adapted from “Sentiment Analysis Algorithm and Applications: A Survey,” by W. Medhat, A. Hassan, & H. Korashy, 2014, *Ain Shams Engineering Journal*, 5(4), p. 3.

As depicted in Figure 3.1, two approaches have been broadly used which are machine learning or a lexicon-based approach in performing sentiment classification and will be elaborated further in the following subsections.

3.2.1 Machine Learning Approach

The machine learning approach has been extensively used for text classification. It can be categorized into supervised, semi-supervised or unsupervised techniques in constructing the model (Ravi & Ravi, 2015). Among these techniques, supervised learning has been widely applied in many sentiment analysis tasks. The application of the machine learning approach requires a large labeled training corpus in building the model and to be learned by the classifiers. Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers are among the frequently applied sentiment classifiers in classifying the data (Liu, 2010). The typical workflow of sentiment analysis using the machine learning approach is shown in Figure 3.2.

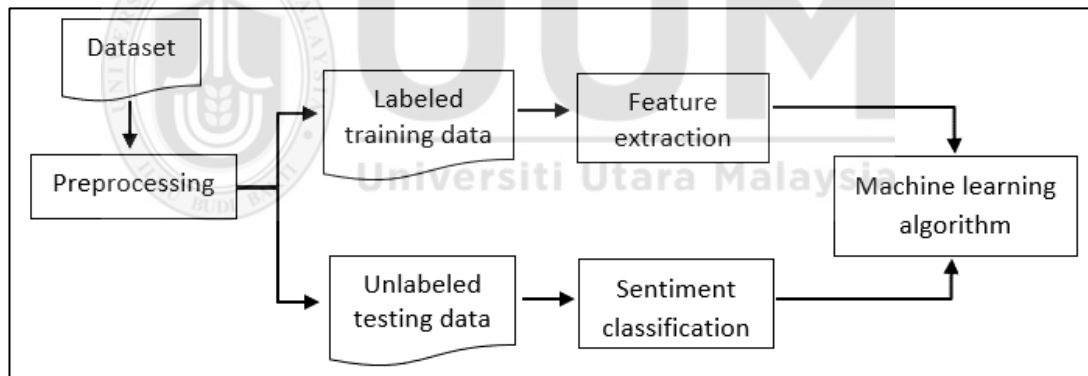


Figure 3.2. The workflow of a machine learning approach

3.2.2 Lexicon-based Approach

The lexicon-based approach uses the lexicon or dictionaries consisting of opinionated words with its polarity. Figure 3.6 illustrates the typical workflow for lexicon-based sentiment analysis.

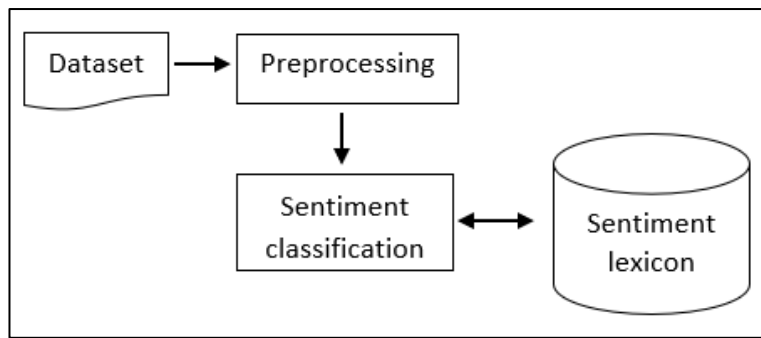


Figure 3.3. The workflow of the lexicon-based approach

The sentiment lexicon is constructed either using a dictionary-based or corpus-based approach. The dictionary-based approach generally relies on available dictionaries such as WordNet in extracting the sentiment words while the corpus-based approach is applied to find opinions words using a large corpus (Feldman, 2013). Precisely, the latter approach mainly depends on sentiment lexicon containing opinionated terms and their associated polarity score to classify sentiments.

3.3 Sentiment Analysis Applications

To date, sentiment analysis has been applied to almost every possible domain such as product, movie, sport and political reviews. Furthermore, with the growth of micro-blogs platforms, most organizations and businesses have applied sentiment analysis to obtain public opinions and thoughts about their products or services. Previous studies have reported that a wide range of issues and topics have been covered such as products reviews (Fang & Zhan, 2015; Zhou, Jiao, & Linsey, 2015), hotel reviews (Valdivia, Luzón, & Herrera, 2017), political and financial analysis (Chan, & Chong, 2017; Ramteke, Shah, Godhia, & Shaikh, 2016).

3.4 Non-English Sentiment Analysis

A considerable amount of prior research in classifying the sentiments written in the English language has been conducted. Although the English language continues to be the primary language applied in majority of research works in this field, there are also ongoing efforts in applying sentiment analysis to other languages such as Malay (Alexander & Omar, 2017), Chinese (Lee & Renganathan, 2011) and Spanish (Miranda & Guzman, 2017). The implementation of sentiment analysis for a specific language generally depends on manually or semi-automatically developed sentiment lexicons found in dictionaries or corpora.

3.5 Mixed Language Sentiment Analysis

Although a lot of work has been focusing on mining data in a single language, there are some recent studies have been conducted to analyze mixed language content as well.

The initial efforts on this subject matter have focused on pre-processing or normalization task which involves the activities like identification of noisy text, correction of spelling and stop words removal (Samsudin et al., 2013; Vyas, Gella, Sharma, Bali, & Choudhury, 2014). Normalization of mix English and Bangla language was studied by Dutta, Saha, Banerjee, and Naskar (2015) and they focused on spelling correction using a noisy channel model. Zhang, Chen, and Huang (2014) introduced two-stage methods to normalize Chinese – English mixed texts which are word translation and word categorization.

For word translation, the neural network language model was used to translate in-vocabulary English words to Chinese, while for out-of-vocabulary words, a graph-based unsupervised model is applied to categorize them.

In a mixed language environment, various methods within both approaches have been applied in judging the sentiments. Sitaram, Murthy, Ray, Sharma, and Dhar (2015) trained a classifier on the 24 mixed English – Hindi language data directly rather than translated to a single language. Raghavi, Chinnakotla, and Shrivastava (2015) learned a basic SVM based question classification system for English - Hindi data. All the data have been translated into English before feature selection and classification were performed. In contrast, Yan, He, Shen, and Tang (2014) proposed a bilingual approach to process review comments written in Chinese and English. Their models are able to analyze sentiments without translation and to process two different languages simultaneously. For a code-mixed (English – Spanish) environment, the result shows that the multilingual model is the best option when Spanish is the majority language. Lo et al. (2016) have constructed a toolkit to analyze polarity for Singlish (Singaporean English) using the semi-supervised approach. Unlike previous research which relying on English knowledge-based such as SenticWordNet (Denecke, 2008) and WordNet (Miller, 1995), Lo has used SenticNet (Cambria et al., 2014) which includes 30,000 common-sense concept, negation and adversative terms handling as the core resource for their polarity detection. They detect ambiguous words during bigram and trigram analysis, and it was treated as Singlish stop words. Another significant work using the lexicon based approach was proposed by Sharma et al. (2016) where they have used various lexicon resources such as WordNet and English

SentiWordNet to classify sentiments. They have obtained a precision of 0.80 in determining the sentiments of the English – Hindi dataset.

3.6 Languages Used in Malaysia

Malaysia is a country that is diverse in terms of cultures and languages. Three main races are residing in Malaysia; Malays, Chinese and Indians. Due to this diversity, various languages and dialects are used in Malaysia. Below subsections describe the official and common language applied in daily communications by Malaysians.

3.6.1 Malay Language

In Malaysia, the Malay language is called *Bahasa Melayu* and is used as an official language. Besides that, the Malay language is also widely spoken in three other countries that include Indonesia, Singapore and Brunei. Almost 77 million people in these countries are considered native speakers of Malay language and it is ranked sixth after Arabic for the most spoken languages on earth (Julian, 2019). However, the Malay language is still known as an under-resourced language in terms of linguistic technologies and the availability of lexical resources even though many essential texts in either formal documents or SNS are written in this language.

3.6.2 Mixed Language (*Bahasa Rojak*)

The use of mixed language arises from the fact that some multilingual speakers or writers feel more comfortable to convey information in their native language compared to English. Mixed language either verbally or in written form is considered

typical, especially in multilingual societies like Malaysia and Singapore. The term mixed language refers to the use of more than one language in the same conversational event either in speaking or writing (Gumperz, 1983; Sharma, Srinivas, & Balabantaray, 2016). The use of mixed language is usually found in social media content such as Facebook, Twitter and forums. In Malaysia, social media users tend to mix Malay and English language known as *Bahasa rojak* in their informal communication (Chuah, 2013). Below are the examples of *Bahasa rojak* posted on a Twitter platform that contains both Malay and English texts:

Example 1: buku ni *brilliant*...*everyone should read*!!

Example 2: tahniah Azizul...*the Keirin World Champion*!

Example 3: *jammed* teruk *from* Tapah *to* Ipoh, dah 2jam *stuck* kat sini...

The statement in the example above is a mixture of two languages; Malay and English. Words in italic belong to the English language, while the rest belongs to the Malay language.

3.7 Sentiment Analysis in the Malaysian Context

A comprehensive and thorough literature search based on the title, abstract and keyword was conducted through three scholarly publications search engines which are Scopus, Dimensions and Google Scholar. The keyword used included ‘sentiment analysis Malay’, ‘sentiment analysis Malaysia’, ‘opinion mining Malay’, ‘opinion mining Malaysia’, ‘sentiment analysis bahasa rojak’ and ‘opinion mining bahasa rojak’. Articles in refereed journals and conference proceedings that included these particular terms in their titles, abstracts or keyword lists covering various focus areas and sentiment analysis tasks were considered.

As a result, eighty-three (83) papers published from 2010 were scanned during this process, and approximately fifty-five (55) articles related to sentiment analysis in the Malaysian context were identified and included in the analysis. Below subsections explain in detail researches according to sentiment analysis task, sentiment classification approaches, construction of sentiment lexicon, language covered, and domain applied.

3.7.1 Sentiment Analysis Tasks

Aside from sentiment classification activity as stated in Section 3.2, there are few other activities involved in performing sentiment analysis such as pre-processing which meant to remove the unnecessary data and subjectivity analysis which focuses on categorizing text into subjective or objective. It has been identified that most efforts for sentiment analysis in the Malaysian context have concentrated on four sentiment analysis tasks. The categories included; i) pre-processing ii) feature selection iii) sentiment classification and iv) lexicon construction. Figure 3.4 illustrates the sentiment analysis tasks covered by previous researchers.

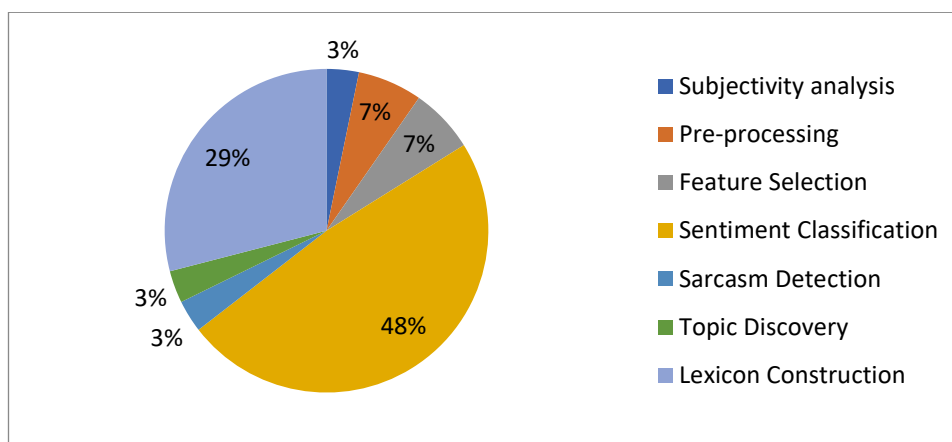


Figure 3.4. Sentiment analysis task

Based on Figure 3.4, sentiment classification and lexicon construction are the most popular sentiment analysis task conducted for the Malaysian context. Sentiment classification or polarity detection is considered as the main task in sentiment analysis. Thus, it explains why most researches are concentrated on this task which decides the sentiment of a given sentence or text is either positive or negative (Al-Moslmi, Omar, Albared & Alshabi, 2017; Alsaffar & Omar, 2015; Chekima & Alfred, 2018; Isa, Puteh, & Kamarudin, 2013; Sadanandan, Osman, Hussain Saifuddin, Ahamad, & Hoe, 2016; Shamsudin, Basiron, & Sa'aya, 2016; Zamani, Abidin, Omar, & Abiden, 2014). Another important task performed in sentiment analysis is the construction of sentiment lexicon. The goal of this activity is mainly to develop a dictionary containing words with its associated sentiment or polarity.

Besides sentiment classification and lexicon's construction, there are five other tasks which include pre-processing (Arif & Mustapha, 2017; Samsudin, Puteh, Hamdan, & Nazri, 2013), feature selection (Samsudin et al., 2013), subjectivity analysis (Kasmuri & Basiron, 2019), sarcasm detection (Suhaimin, Hijazi, Alfred, & Coenen, 2019) and topic discovery (Kannan, Govindasamy, Soon, & Ramakrishnan, 2018) were conducted as well.

3.7.2 Sentiment Classification Approaches

As mentioned in Section 3.2, there are two main sentiment classification approaches commonly used, which are machine learning and lexicon-based. However, there is another increasingly growing approach known as a hybrid that combines both machine learning and a lexicon-based approach. These three fundamental approaches have

been widely applied in Malay sentiment analysis as well (Bakar, Idris, Shuib, & Khamis, 2020). All the related research is highlighted below.

1. Machine Learning Most sentiment analysis researches within the Malaysian context has chosen machine learning techniques for the sentiment classification task. Samsudin, Puteh, and Hamdan (2011) pioneered the research where they have utilized online forums and blogs as their data sources. Isa, Puteh, and Kamarudin (2013) have applied Artificial Immune Network (AIN) in classifying Malay newspaper articles. NB, SVM and k-Nearest Neighbour (k-NN) have been a popular method applied by most of researchers (Alfred, Yee, Lim, & Obit, 2016; Alsaffar & Omar, 2014; Alshalabi, Tiun, Omar, & Albared, 2013; Arif & Mustapha, 2017; Naing, Thwe, Mon, & Naw, 2018; Puteh, Isa, Puteh, & Redzuan, 2013; Suhaimin et al., 2019).

Due to its popularity, Al-Moslmi et al. (2017) conducted a comparative study on these three methods; NB, SVM and k-NN with the use of few ensemble methods and concluded that SVM produced the best result for the particular machine learning technique. In another study, Sadanandan et al. (2016) have applied Mi-Intelligence (MI) method and reported the highest accuracy; 94.34% while Ibrahim and Yusoff (2015) achieved 90% accuracy using NB method in analyzing tweets for Malayan Banking (Maybank).

2. Lexicon-based Lexicon-based technique requires the construction of a new sentiment lexicon or the adoption of an available one. The lexicon serves as a resource from which scores or polarities of sentiment words are obtained to analyze the sentiment of a given text (Muhammad, 2016). Shamsudin et al. (2015) and Zamani

et al. (2014) have manually created their lexicon based on Facebook comments to classify the sentiment. Whereas Hijazi, Libin, Alfred, and Coenen (2016) have modified the existing resource, SentiStrength's score to cater to zero-valued sentiment words found for several Sabah dialects words in the original lexicon. Chekima & Alfred (2018) have developed a new lexicon known as MySentiDic and have been utilized in various domains such as politics, economics and movies. The classification using MySentiDic has achieved 86.5% F1-measure. Rodzman, Rashid, Ismail, Rahman, Aljunid, & Rahman (2019) have created their own lexicon table based on the expert's judgement on the polarity score and their classifications on three different domains have outperformed NB classifier by 20%. On the other hand, Shamsudin et al. (2016) performed classification on Facebook comments and reported the lowest accuracy with 52.12% as compared to other methods and techniques. In more recent work, Bakar, Rahmat, & Othman (2019) have created a new tool known as Malay Polarity Classification Tool (MaCT) based on the AFINN sentiment lexicon.

3. Hybrid The combinations of the first two approaches (hybrid) explained previously is introduced because of the pros and cons of each approach. The machine learning approach is well-known for having a high accuracy given a high-quality training corpus, but the drawback of this technique is it requires a lot of effort in manual data annotation. On the contrary, the lexicon-based approach adapts better for various domains to have ease of use. However, the approach is unable to produce comparable accuracy to machine learning techniques.

The first research using a hybrid method was done by Alsaffar and Omar (2015) where they have translated English WordNet into Malay language and classified the sentiments using the k-NN method. Shuhidan, Hamidi, Kazemian, Shuhidan, and Ismail (2018) analyzed financial news headlines using the Opinion Lexicon developed by Hu and Liu (2004) and the classification was performed using the NB method. In another study, Eshak, Ahmad, and Sarlan (2017) developed a hybrid approach to analyze communications using the Malay language on social media platforms to determine the customer intention to purchase in social commerce. Table 3.1 summarized all the related sentiment classification work conducted in the previous research.



Table 3.1

Sentiment Classification

References	Techniques	Methods	Language Covered	Domain	Data Sources	Dataset	Accuracy (Acc) Reported
Rodzman, et al. (2019)	Lexicon-based	Expert judgement	Malay	Song, politic, product	Students' reviews	N/A	Acc: 70%
Bakar, et al. (2019)	Lexicon-based	AFINN	Malay		Twitter	1000	Acc: 90%
Al-Moslmi, et al. (2017)	Machine learning	NB, SVM, k-NN	Malay	Movie reviews	Web pages, blogs	2000	F1: 85.81%
Alsaffar & Omar (2015)	Hybrid	Bagging, Stacking, Voting, AdaBoost, MetaCost Lexicon: English WordNet Classification: k-NN	Malay	Movie reviews	Online forums, Blogs	2000	F1: 86.43%
Alsaffar & Omar (2014)	Machine learning	SVM, NB, k-NN	Malay	General	Social media, blogs	2000	Acc: 87%
Al-Saffar, Awang, Tao, Omar, Al-Saiagh, & Al-bared (2018)	Hybrid	Lexicon: WordNet (translated) Classification: NB, SVM, Deep Belief Network (DBN)	Malay	Movie reviews		2478	F1: 94.48%
Arif & Mustapha (2017)	Machine learning	NB, SVM, k-NN	Malay	Movie	Online forums, blogs	2000	F1: 75.29%
Chekima & Alfred (2018)	Lexicon-based	MySentiDic	Malay	Politic, Economi, Movie, Fashion	Online forums, Facebook, Twitter	16280 documents	F1: 86.5%

Table 3.1 continued

Nasharuddin, Abdullah, Azman, & Kadir (2017).			Malay, English		News articles (Bernama)	883	
Puteh, et al. (2013).	Machine learning	Artificial Immune System	Malay	News articles	Newspaper	1000	Acc: 88.5%
Sadanandan, et al. (2016)	Machine learning	MI	Malay	General	Social media	1861	Acc: 94.34%
Samsudin, Puteh, & Hamdan, (2011)	Machine learning	NB, SVM, kNN	Malay	Movie	Forums, blogs	1000	Acc: 68.35%
Shamsudin et al. (2016)	Lexicon-based	Term counting, Term counting average (TCAvg)	Malay	General	Facebook	450	Acc: 52.12%
Shamsudin, et al. (2015)	Lexicon-based	Lex: Malay adj score dictionary, Malay-English score dictionary Average on comments, Term counting, Term score summation	Malay	General	Facebook	450	Acc: 63.4%
Suhaimin, et al. (2019)	Machine learning	kNN, non-linear SVM	Malay	Economic	Facebook	1970	F1: 99.4%
Naing, et al. (2018)	Machine learning	SVM					
Alfred, et al. (2016)	Machine learning	SVM, NB, k-NN	Malay	News headlines	Newspaper	600	Acc: 84.8%
Isa et al. (2013)	Machine learning	SCIN	Malay	Newspaper articles	Newspaper –Berita Harian	1080	Acc: 53.67%

3.7.3 Sentiment Lexicon

There are several readily available off-the-shelf sentiment lexicons constructed solely for the English language. However, those lexicons do not perform well when it applies to a foreign language like Malay (Darwich et al, 2016; El-Beltagy, & Ali, 2013). Due to the importance of having a specific sentiment lexicon for one particular language, it can be seen that the works on generating sentiment lexicon have increased as well. There are various sentiment lexicons with varying sizes and formats have been either manually or semi-automatically constructed. Table 3.2 summarizes the studies related to automatic lexicon generation for the Malay and mixed Malay-English language. It shows that most of the sentiment lexicons constructed are based on WordNet Bahasa as it is one of the widely used lexical databases in the Malay language research study (Noor, Sapuan, & Bond, 2011).

WordNet Bahasa is a combination of lexical semantics from various resources and it contains over 45,000 Malay words and 58,000 Indonesian words. The WordNet Bahasa was also developed based on the English WordNet. The latest version of English WordNet known as WordNet 3.0 contains approximately 155,287 words. The mapping technique between WordNet Bahasa and English WordNet version is the common method used to generate lexicon (Alexander & Omar, 2017; Darwich, Noah, & Omar, 2016; Nasharuddin et al., 2017). Even though Darwich et al. (2016) achieved the highest accuracy with 89.4%, but the word extractions in their studies only involved the adjectives word class for the mapping and ignored adverb and verb which probably carries sentiment as well.

To date, most available sentiment lexicons are directed at the English language only and very few are in other languages. Commonly used English language lexicon resources are WordNet, SentiWordNet and SentiStrength. Most researchers when performing sentiment analysis on languages other than English will resort to either build their lexicon manually or by translating English lexicon to their preferred language (Yusoff, Jamaludin, & Yusoff, 2016). Alsaffar and Omar (2015) developed a Malay lexicon to analyze sentiment in Malay text by translating English WordNet to the Malay language.

Polarities are manually annotated to these translated words and values are assigned to these words by Malay linguistic expert manually. The drawback of this technique is, an English word when translated into the Malay language would have a different meaning, which will affect the polarity. Due to the issue with a translation, Anbananthen, Selvaraju, and Krishnan (2017) developed an automated process of lexicon generation based on a set of Malay tweets using a dictionary-based approach. However, the lexicon has not been made publicly available. Tan, Lam, Azlan, and Soo (2016) and Chekima and Alfred (2018) have constructed bilingual (Malay-English) sentiment lexicon to analyze mixed language content. Both works have used the Twitter platform as the source of data collection. RojakLex lexicon developed by Chekima and Alfred (2018) has recorded 79.28% accuracy which is an increase of 27.9% as compared to the baseline. This research makes a comparison in terms of performance with RojakLex as well as SentiLexM and the result is discussed in Section 6.2.

Table 3.2

Summary of Lexicon Constructions' Work

Author	Domain-Dependent	Language Covered	Resources Used
(Darwich et al., 2016)	No	Malay	WordNet Bahasa, English WordNet
(Nasharuddin, et al., 2017)	No	Malay	WordNet Bahasa, English SentiWordNet
(Anbananthan et al., 2017)	No	Malay	Malay lexicon (generated manually based on Malay tweets)
(Alexander & Omar, 2017)	No	Malay	WordNet Bahasa, English WordNet
(Darwich, Noah, & Omar, 2017)	No	Malay	WordNet Bahasa, Kamus Dewan
Tan et al. (2016)	No	Malay-English	AFINN Lexicon (Nielsen, 2011)
Chekima & Alfred (2018)	No	Malay-English	MySentiDic, Emoticon Lexicon, Neologism Lexicon

3.7.4 Language Covered

As shown in Figure 3.5, a great deal of previous work has been focusing on mining data in a single language which is Malay (Alexander & Omar, 2017; Alsaffar & Omar, 2014; Darwich et al., 2015; Isa et al., 2013; Suhaimin et al., 2019).

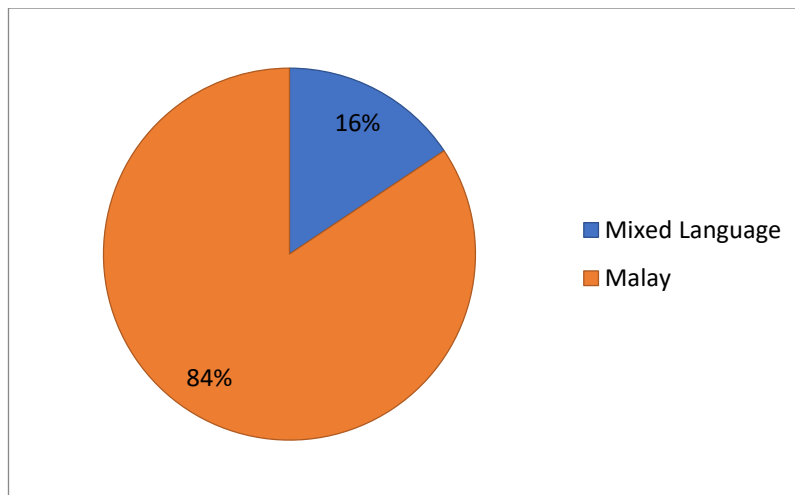


Figure 3.5. Language covered

However, there are some recent studies have been conducted to analyze mixed Malay-English language (known as *Bahasa rojak*) content as well (Chekima & Alfred, 2018; Kasmuri & Basiron, 2019; Samsudin et al., 2013; Tan et al., 2016; Zabha, Ayop, Anawar, Hamid, & Abidin, 2019). The number of researches in *Bahasa rojak* keeps increasing as it is believed that ignoring the mixed language content may lead to an inaccurate analysis result.

The first research conducted in catering for mixed-language content focuses on pre-processing and feature selection task (Samsudin et al., 2013). The introduction of a new normalization approach known as MyTNA (Malay Mixed Text Normalization Approach) and the use of the Immune Network System in selecting the features have achieved an accuracy of 92.25%. The more recent study concentrated on developing subjective corpus which differentiates the factual and opinionated sentences (Kasmuri & Basiron, 2019).

Subjectivity analysis usually is studied before sentiment classification task as it will filter out the objective sentences within a text. Thus, it helps to improve the accuracy at the polarity classification level.

3.7.5 Domain Applied

There is one domain that dominated the sentiment analysis research in Malaysia which is movie reviews (Al-Moslmi et al, 2017; Alsaffar & Omar 2015; Arif & Mustapha, 2017; Chekima & Alfred, 2018; Samsudin et al., 2013). Besides movie reviews, sentiment analysis has been applied to other domain areas as well such as social tension detection (Jamil, Kamaruddin, & Ahmad, 2019), hotel reviews (Alexander & Omar, 2017), telecommunications (Tan et al., 2016), dengue monitoring (Kannan et al., 2018) and GST (Kaur & Balakrishnan, 2016; Zabha et al., 2019). Jaidka, Ahmed, Skoric, and Hilbert (2019) studied elections predictions in three countries; Malaysia, India and Pakistan. Their studies found that machine learning techniques gave the most accurate results for their predictions. Naing et al. (2018) developed a system for analyzing National Educational Rate and Crime Rate that occurred in Malaysia, Singapore and Myanmar using the SVM method. Kaur and Balakrisynan (2017) studied the impact of letter repetition using Sentiment Intensity Calculator (SentI-Cal) in determining the sentiment scores for two major airlines in Malaysia. Since the research concentrated on the Malaysian context is still limited, it is vital for future research work to focus on the other domains as well.

3.8 Research Gaps in Sentiment Analysis for the Malaysian Context

The previous studies that have been done in sentiment analysis for the Malaysian context pose some issues. This research highlights three of the issues; domain-specific sentiment lexicon, the analysis of mixed language and the application of sentiment analysis in the property domain which will be elaborated further in the following subsections.

3.8.1 Domain-Specific Sentiment Lexicon

All known lexicon construction researches for the Malaysia context are dedicated to developing a general-purpose sentiment lexicon which can be applied to any domain (Alexander & Omar, 2017; Anbananthen et al., 2017; Chekima & Alfred, 2018; Darwich et al., 2016; Hijazi et al., 2016). However, a general-purpose sentiment lexicon cannot get the best result of the classification, since the sentiment of some words is domain-dependent.

Another issue regarding the lexicon's construction is related to the language. Even though lexicon construction research is the second largest research conducted (as illustrated in Figure 3.2), their research is limited to Malay language only. The shortfall of the current monolingual lexicon is it may not be able to make an accurate classification of the mixed language sentence if the sentiment words appear in another language.

3.8.2 Analysis of Mixed Language

Despite a growing number of researches on sentiment analysis, little has been done on the analysis of mixed language content due to the complications arising from the difficulty of processing two languages simultaneously. Among the issues associated with the use of mixed language are the grammatical differences and improper switching of languages in one sentence which introduces new challenges in the field of NLP.

Moreover, as mentioned in Section 3.6.2, mixed language posts are commonly found in social media platforms where the opinions expressed in these posts are presented in both monolingual and bilingual ways. According to Lo, Cambria, Chiong, and Cornforth (2017), the inability to analyze mixed language text will affect the accuracy of the overall classification's result. Therefore, different approaches and techniques are needed in order to achieve comparable performance levels to what has been achieved in a single language such as English or Malay.

3.8.3 Analysis of Property Domain

As highlighted in Section 3.7.5, it can be summarized that none of the previous studies have been conducted in analyzing property reviews using social media platforms despite the crucial needs for sentiment analysis in this domain. Research on gathering public sentiments for property domain was circled around surveys or questionnaires only which targeted only a specific number of audiences.

Thereby, this research is aimed to explore a better way of analyzing Malaysian sentiments for the property domain. In order to narrow these research gaps discussed above, this research is motivated to explore a new approach to generate a bilingual sentiment lexicon. A new method is proposed in pursuance of filling these gaps, which is presented in Chapter 6.

3.9 Established Sentiment Lexicon

There are several widely applied sentiment lexicons in the previous researches, which include SentiWordNet (Denecke, 2008), WordNet (Miller, 1995) and SenticNet (Cambria, Olsher, & Rajagopal, 2014). All of these lexicons are elaborated further in the following subsections.

3.9.1 SentiWordNet

SentiWordNet is a widely used English lexicon where each synset is annotated with labels indicating how objective, positive, and negative the terms in the synset are. SentiWordNet 1.0 included 28,431 non-neutral words while SentiWordNet 3.0 includes 38,182 non-neutral words. SentiWordNet 3.0 is an improvement over the original SentiWordNet proposed in Esuli and Sebastiani (2006). It is based on WordNet, the well-known lexical database for English where words are clustered into groups of synonyms known as synsets (Miller, 1995). In SentiWordNet each synset is automatically annotated in the range [0, 1] according to positivity, negativity and neutrality.

3.9.2 AFINN

Inspired by ANEW (Affective Norms for English Words), Nielsen (2011) created the AFINN lexicon of 2,477 English words, which is more concentrated on the language used in microblogging platforms. ANEW was released before the growth of SNS and hence, many slang words commonly used in social media were excluded. Considering that there is empirical evidence about significant differences between microblogging words and the language used in other domains, a new version of ANEW was required. In AFINN, positive words are ranged from 1 to 5 and negative words from -1 to -5.

3.9.3 SentiStrength

This lexicon returns the positive score ranges from 1 (not positive) to 5 (extremely positive) and the negative one ranges from -1 (not negative) to -5 (extremely negative). SentiStrength is manually annotated and includes both formal and informal English words. The scores can also be adapted to a specific domain using machine learning. This lexicon applies linguistic rules for dealing with questions, negations, emoticons, and booster words. These are employed together with the lexicon for calculating the positive and negative outputs.

3.9.4 General Inquirer

General Inquirer has 1,915 and 2,291 positive and negative words respectively. It is a lexicon constructed by Stone, Dunphy, and Smith (1966). The terms included in the lexicon are tagged according to multiple dimensions such as emotions, polarity, and semantics.

3.10 Sentiment Lexicon Creation: Prominent Techniques

In sentiment analysis, one of the primary activities is to classify whether the given text expresses positive or negative sentiments and sentiment lexicon becomes an essential tool in performing this activity. In building such lexical resources, there are plenty of studies that have investigated various methods and techniques. To date, the literature on the creation of sentiment lexicon can be broadly classified into three categories; manual approach, knowledge-based and corpus-based (Liu, 2012; Vania, Moh. Ibrahim, & Adriani, 2014).

The manual approach is known to be costly in terms of human annotation's time and effort (Dragut, Wang, Yu, Sistla, & Meng, 2015). Due to that matter, the manual approach is usually applied in conjunction with the other two approaches to minimizing human intervention and to check the accuracy of the generated lexicons.

The knowledge-based approach utilizes the available lexical resources such as WordNet in extracting synonyms and antonyms information (Agrawal & Siddiqui, 2012; Hu & Liu, 2004). The primary strategy of this approach is to begin with manually collect few initial sets of seed words and their polarities and followed by searching for their synonyms and antonyms in a lexical dictionary in order to expand the set (Hailong, Wenyan, & Bo, 2014). While there are a lot of available lexical resources available for the English language, the resources in Malay are limited.

The corpus-based approach assigns polarity in a large corpus based on the assumption that two sentiment words co-occur together are likely to express the same polarity. The idea behind this method is the sentiment of a word tends to be consistent with the sentiments of its surrounding words (Moreno-Ortiz & Fernández-Cruz, 2015; Peng & Park, 2011). The main advantage of this approach is the ability to obtain domain-specific sentiment words as it utilizes domain-specific corpus in the extraction process. Most studies are adopting this approach in constructing domain-specific lexicon such as mobile shopping domain (Feng, Gong, Li, & Lau, 2018), and product reviews (Liu, Lei, & Wang, 2013).

The above-mentioned research study was mostly focused on the English language. In order to generate sentiment lexicon for non-English, most researchers resulted in translating the readily available lexicon for English to their target languages. (Dehkharghani, Saygin, Yanikoglu, & Oflazer, 2016; Shahid & Kazakov, 2009). The translation method is known to be time-consuming and the translated words may have a different meaning which can affect the polarity of the words (Anbananthen et al., 2017)

In assigning the polarity for each sentiment word in the lexicon, previous approaches have applied either manual annotation (Abdul-Mageed, Diab, & Korayem, 2011; Esuli & Sebastian, 2006) or word mapping by employing the English WordNet (Perez-Rosas, Banea, & Mihalcea, 2012). This research is different in the sense that it presented a new polarity score assignment using word vector representation and term frequency.

3.11 Word Vector Representation

The word vector representation technique is commonly used in various NLP tasks such as PoS tagging (Collobert et al., 2011), machine translation (Devlin et al., 2014; Zou, Socher, Cer, & Manning, 2013) and named entity recognition (Guo, Che, Wang, & Liu, 2014; Turian, Ratinov, & Bengio, 2010).

In brief, word vector is considered as a breakthrough from the traditional technique for representation of words. It uses a vector of numbers to represent the meaning of a term. This technique aims to turn the data into a vector space that facilitates the process of creating features for classification. For example, Ohana and Tierney (2011) have used word vector to perform sentiment classification for a collection of film reviews and they have reported accuracy of 85.39% using this technique. Another research focusing on similarity comparison has trained a corpus of 5.8million reviews using the word vector technique (Liu, Liu, Zhang, Kim, & Gao, 2016). In their study, the word vector is treated as a form of knowledge learned from the historical data. Therefore, given the state-of-the-art above, the word vector representation technique has been successfully applied to various NLP tasks.

3.12 Term Frequency

Term frequency refers to the number of times a term appears within a document. It is a useful method in text mining where the number of word occurrences is an indicator for important terms and these words have a significant impact in identifying the polarity of opinion.

Previously, the term frequency technique has been applied in various text classification tasks. In sentiment analysis, the frequency of terms plays a vital role in terms of identifying essential information (Deng, Luo, & Yu, 2014). Demiroz, Yanikoglu, Tapucu, and Saygin (2012) assigned polarity based on this technique as well as the SentiWordNet lexicon in classifying sentiments.

Other than that, the term frequency technique combined with inverse document frequency (IDF), and count-based probability measures have been used to enhance the weighting scheme for the polarity score in health-related terms (Asghar, Ahmad, Qasim, Zahra, & Kundi, 2016). It was also observed that the term frequency method was used for the feature extraction process in tourism reviews (Menner, Höpken, Fuchs, & Lexhagen, 2016). In the same line of study, Quan and Ren (2014) have combined the term frequency-inverse document frequency (TF-IDF) method with pointwise mutual information (PMI) and they were able to demonstrate better results as compared to the state-of-the-art using those combinations.

This study proposes the use of both word vector representation and term frequency techniques in determining the polarity score for sentiment words in the lexicon. The detail processes of the polarity assignment are explained in Section 6.5.

3.13 Sentiment Lexicon Creation for Malaysia Property Domain

Although there are quite several works done previously in generating domain-specific lexicon, none is explicitly built for the property domain. In the process of generating the sentiment lexicon specifically for the Malaysian property domain, two issues have been taken into consideration: mixed language and domain-specific lexicon.

Since English and Malay are the two most spoken languages in Malaysia, the bilingual lexicon is developed which gives an advantage in classifying mixed language content. Based on the state-of-the-art described in the previous section, this research leverages both corpus and knowledge-based approaches in constructing the lexicon. This research has used Twitter as a corpus and WordNet as a resource to obtain the synonyms. Furthermore, this study improves the traditional approach in assigning the polarity score by utilizing the word vector model to determine the polarity and term frequency to determine the weight of sentiment words.

3.14 Evaluation Measures

In assessing retrieved tweets, some of the common measurements applied to evaluate sentiment analysis results are precision and recall (Castillo, Mendoza, & Poblete, 2011; Hong, Dan, & Davison, 2011). Precision is defined as the ratio of true positive to the summation of true positive and false positive documents. The recall is the ratio of true positive sentences to those that are correctly predicted. In another study carried out by Pak and Paroubek (2010), in order to determine if a tweet is positive, negative or neutral, they used accuracy and F-measure instead of precision and recall, respectively. The F-measure is calculated as the combination of precision and recall while the accuracy is the total correctly classified tweets normalized by the total number of tweets.

Another form of measurement which considered as the most effective way to validate sentiment analysis is to ask different human annotators to assign the values

(Chamlertwat, Bhattarakosol, Rungkasiri, & Haruechaiyasak, 2012; Taboada et al., 2011).

In order to evaluate the quality of the lexicon, Al-Moslmi et al. (2017) and Lu, Castellanos, Dayal, and Zhai (2011) have performed a comparison with other lexicons and several prominent machine learning classifiers as their experimental evaluation.

This research employed accuracy, precision, recall and F-measure as quality metrics to evaluate the performance of sentiment analysis. The results obtained from the proposed approach is compared against the human-annotated data sets. Besides, another comparison with the well-established sentiment lexicon and machine learning classifiers is also implemented.

3.15 Manual Annotation Procedure

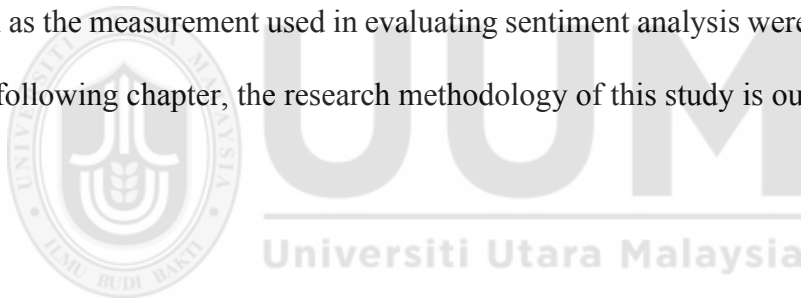
The methodologies of manual data annotations in previous research vary significantly. Strapparava and Mihalcea (2007) instructed their annotators to annotate the given title for sentiment and no further guidance or training was carried out. Similarly, a brief instruction is given by Nakov, Ritter, Rosenthal, Sebastiani, and Stoyanov (2016) to the annotators, but a list of examples annotated sentences were provided.

In regard to number of annotators, Mohammad (2016) and Refaee (2017) have selected two annotators to perform the labelling while other researchers assigned more than two in implementing the data annotations (Alayba, Palade, England, & Iqbal, 2017; Shalunts, Backfried, & Prinz, 2014).

3.16 Chapter Summary

This chapter gave an overview of sentiment analysis in general and sentiment analysis in the Malaysian context in particular. It can be summarized that most of the research efforts into sentiment analysis in the Malaysian context have centered on several sentiment analysis tasks such as pre-processing, feature selection, construction of sentiment lexicon and sentiment classification.

Several issues from the state-of-the-art sentiment analysis system have been pointed out and one of the limitations which are the analysis of mixed language has been highlighted. The common techniques employed in generating the sentiment lexicons as well as the measurement used in evaluating sentiment analysis were also presented. In the following chapter, the research methodology of this study is outlined in detail.



CHAPTER FOUR

RESEARCH METHODOLOGY

4.1 Introduction

The previous chapter provides a review of the literature. It gives an understanding of the issues related to the domain of the study. This chapter presents the methods used and activities involved in answering the research questions and achieving the objectives, as mentioned in Chapter 1. It begins by introducing the research design in Section 4.2, which continues with the four research stages involved in this study from Section 4.3 until Section 4.6. The chapter ended with a summary in Section 4.7.

4.2 Research Design

Figure 4.1 illustrates the research design, which includes the stages, research activities involved and outcomes for each stage. Four stages incorporated in guiding this study, which are; i) Theoretical study, ii) Exploratory Study, iii) Experiments and iv) Performance Evaluation. The explanation of each stage is included in the next subsections of this chapter.

Stage I Theoretical Study	Stage II Exploratory Study	Stage III Experiments	Stage IV Evaluation
Research Activities			
<ul style="list-style-type: none"> Identify the issues in Malaysia property industry Identify gaps in the existing sentiment analysis approaches Review techniques for construction of sentiment lexicon of mixed-language text 	<ul style="list-style-type: none"> Selection of case studies Sampling and data collection Data analysis procedure 	<ul style="list-style-type: none"> Identifying the annotators Determine the data pre-processing activities Determine the lexicon creation techniques Determine the available resources Determine the sentiment classification process 	<ul style="list-style-type: none"> Determine the evaluation criteria Determine the baseline comparison

Figure 4.1. Research design

4.3 Stage I: Theoretical Study

The theoretical study was performed by thoroughly reviewing the literature to identify the issues and gaps related to the domain of the study. Consequently, the main ideas were gained through the literature by reading the printed as well as online references. Among them are journals, proceeding papers, standards documentation, online newspapers, books and thesis.

From the knowledge gained, the problem and case study were defined. Among the activities conducted were identifying the issues in the Malaysian property industry, identifying the gaps in the existing sentiment analysis approach and systematically reviewing the techniques to perform sentiment analysis for both Malay and English text especially on the creation of sentiment lexicon.

4.4 Stage II: Exploratory Study

The second phase of the study is an exploratory study. There are three main activities performed in conducting the exploratory study as laid out in the following subsections.

4.4.1 The Selection of the Case Study

Particular attention was given to the selection of the case study. There are two criteria were considered for case study selection which are:

- i. The appropriateness of the cases with the issue specified.*

Both cases chosen for this research are affected by the issue specified in Section 2.4, which is the imbalances of supply and demand in the Malaysian property industry. Besides, both PR1MA and PPAM projects are among the anticipated affordable housing projects by the majority of Malaysians.

ii. *The availability and accessibility of the data to be collected.*

The availability of data through SNS was also being considered to ensure the number of data collected is sufficient enough to perform the analysis. Other than that, the data need to be publicly available to make sure that there is no security and privacy violation for social media users. Due to that matter, Twitter was chosen as a platform to extract the data because it provides a sufficient amount of publicly shareable data to be used in this research. An additional reason to choose the Twitter platform to collect the data is due to its popularity among Malaysians when it comes to expressing their opinion online. Furthermore, most of the research work in Malay sentiment analysis has used Twitter in collecting the data (Hijazi et al., 2017; Suhaimin et al., 2019; Tan et al., 2016).

4.4.2 Sampling and Data Collection

As mentioned in Section 1.6, there are two datasets were used in this study; PR1MA and PPAM. Python package called Twitterscraper (Taspinar & Schuirmann, 2017) was used to extract the data. The scraper returns tweets for the given query based on the condition that the tweets contain the query keyword. During this phase, strategic decisions about the duration of data collection as well as the search keywords for each case study were determined.

The duration selected for sampling is between January 2016 and Mac 2018. As presented in Figure 2.1, the imbalances of the property's supply and demand are crucial starting from 2016. Hence, it explains the reason why the data was collected

during that time period. This sampling of data allows for the usage of sentiment lexicon's construction, model building and testing set. The data extraction script is presented in Appendix A.

4.4.3 Data Analysis Procedure

The data obtained from the activities in Section 4.4.2 were coded using the Python programming language. All the Python modules, functions or packages applied in this research are described in Appendix B.

The present study first elaborates on the primary sentiment analysis process employed which includes activities such as data pre-processing, data annotation, construction of lexicon and sentiment classification. Data pre-processing typically involve activities such as removing duplicated tweets; removing symbols and punctuation, language identification and tagging words. The next step was data annotation. The annotation was done manually by two bilingual speakers. The third step was the construction of the Malay-English sentiment lexicon. The final step performed was sentiment classification. The sentiment for each dataset was classified into two categories; positive and negative sentiments.

4.5 Stage III: Experiments

The third stage was to conduct the experiments. The outcomes from the theoretical study, such as the problems of the existing sentiment analysis and the techniques for obtaining better accuracy were carefully evaluated. There are five elements in executing the experiments which are; identifying the annotators for manual data

annotation purpose, determine the preprocessing activities to clean the data and remove noise, determine the approach in constructing the Malay-English sentiment lexicon, determine the availability of the resources in constructing the lexicon and finally, determine the sentiment classification process. All of the processes mentioned above are discussed further subsequently.

4.5.1 Identifying the Annotators

Referring to Mohammad (2016) and Refaee (2017), the number of annotators and instruction guidelines was determined. The annotators are chosen among Malaysian who are proficient in both Malay and English languages. They are given a detail explanation of what and how to perform the annotation. Besides, they are given a set of an example of the annotation sentences as a guideline. To measure the reliability of the annotations, the inter-annotator agreement was conducted, and the scheme proposed by Gatti, Guerini, and Turchi (2015) was followed in this study.

4.5.2 Determining the Pre-processing Activities

As preprocessing helps in maximizing the classification's performance, the right tasks and activities were carefully determined. The tasks involved in data preprocessing was adapted from Balahur and Perea-Ortega (2015) and Ghosh, Ghosh, and Das (2017).

There are four tasks executed, which are the removal of re-tweets, URLs, symbols and hashtags, language identification and Part-of-Speech (PoS) tagging. The language identification activity was performed in this research as this task is essential for the bilingual or mixed-language sentiment analysis to detect the correct language for each

word in a sentence (King & Abney, 2013). More details of the pre-processing task are explained in Section 5.4.

4.5.3 Determining the Lexicon Creation Technique

One of the main contributions of this study is the lexicon creation technique that has been improved by incorporating word vector representation and term frequency in assigning the polarity score. The activities and techniques involved in automatically generate sentiment lexicons introduced in the previous research were taken into consideration.

Several experiments with various combinations of techniques were investigated in order to find the best method in generating a bilingual sentiment lexicon. To the end, four experiments involving two different polarity score calculations, as well as with and without synonym expansions were carried out. The detailed construction of the sentiment lexicon is provided in Chapter 6.

4.5.4 Determining the Resources to be Used

In developing a new lexicon, the available resources such as dictionaries and sentiment lexicon which may ease the process of lexicon's development were identified. The resources are meant to be used in tasks such as PoS Tagging and synonym expansion. At this point, English WordNet (Miller, 1995) and WordNet Bahasa (Noor et al., 2011) are the resources selected to perform those activities.

4.5.5 Determining the Classification Process

Since this study employed a lexicon-based approach, several prominent sentiment classification techniques within this approach were evaluated. As a result, the Term Counting approach is applied due to its simplicity and widely used in lexicon-based sentiment analysis (Shamsudin et al., 2016; Turney, 2002). The final score for each case study is classified into two categories; positive and negative. Apart from that, the technique to handle negation is identified and included within the sentiment classification process because it may cause incorrect classification if it is not catered in this research. Further explanation can be obtained from Section 6.8.

4.6 Stage IV: Performance Evaluation

To ensure that the constructed sentiment lexicon and the experiments conducted are producing a promising result, the performance evaluation was performed. It was evaluated through the evaluation criteria and baseline comparison. The sentiment classification results for both case studies are evaluated and compared against human judgment for the testing data as outlined in Section 6.9.

4.6.1 Determining the Evaluation Criteria

In this phase, the evaluation criteria were defined. Various evaluation techniques have been employed for gauging the performance of the sentiment analysis. As explained in Section 3.14, the accuracy, precision, recall and F-measure are the most widely used performance measures in sentiment analysis (Kaur & Mohana, 2015; Refaee, 2017).

Accuracy is the most frequently used evaluation metric and it measures how often the method being evaluated made the correct prediction. It is calculated as the sum of the true predictions divided by the total number of predictions. The recall is a statistical measure that shows the fraction of relevant instances that are retrieved. Precision shows what portions of retrieved instances are relevant. F-measure is a combination of precision and recall which commonly applied by researchers to describe the overall performance of sentiment analysis.

The confusion matrix or error matrix is used to present the result of the classifier for prediction (Alaei, Becken, & Stantic, 2019). It is a unique table to visualize the performance of sentiment analysis. Detail calculation of accuracy, recall, precision and F-measure, as well as detail explanation on confusion matrix, can be found in Section 6.9.1.

4.6.2 Determining the Baseline Comparison

The existing sentiment analysis approach is divided into two main categories; lexicon-based and machine learning, hence, the baseline comparisons were conducted for both approaches. It is determined based on the most widely adopted lexicon and machine learning classifiers in the previous studies. For the sentiment lexicon, there are no publicly available Malay-English lexical resources that can be found. Therefore, the baseline lexicon is developed similar to the baseline used in comparison with SentiLexM constructed by Tan et al., 2016. Likewise, for the machine learning approach, the classifiers are chosen based on the most commonly used classifiers in the previous Malay sentiment analysis research. The selection of activities and

techniques involved in performing machine learning classification were based the available pre-build Python packages. The detail baseline comparison is discussed in Section 6.9.2.

4.7 Chapter Summary

This chapter has elaborated on the methodology used to carry out this research. It consists of four stages; theoretical study, exploratory study, experiments and performance evaluation. The first stage focused on reviewing the state-of-the-art within this domain, while the second stage has detailed out the selection of case studies, data collection and analysis procedure. A particular concentration is given in designing the activities for developing the new bilingual sentiment lexicon during the third stage. The last stage involves the process of determining the evaluation criteria and baseline for the comparison purpose. The next chapter discusses in detail the activities conducted in this study prior to the construction of sentiment lexicon.

CHAPTER FIVE

PRELIMINARIES

5.1 Introduction

This chapter describes the activities required to be done prior to the construction of the sentiment lexicon. Firstly, the overview of the sentiment analysis framework is presented in Section 5.2, whilst the datasets used in this study are explained in Section 5.3. Section 5.4 till Section 5.6 focused on data cleaning, manual annotation and data categorization. Section 5.7 ends the chapter with a summary.

5.2 Sentiment Analysis Framework

Figure 5.1 illustrates the framework of sentiment analysis, which consists of six main activities; data-preprocessing, data annotation, data categorization, construction of sentiment lexicon, sentiment classification and performance evaluation. In this chapter, the first three activities are explained in detail while the other three activities will be discussed in the following chapter. The implementation of the approach was programmed using Python programming language.

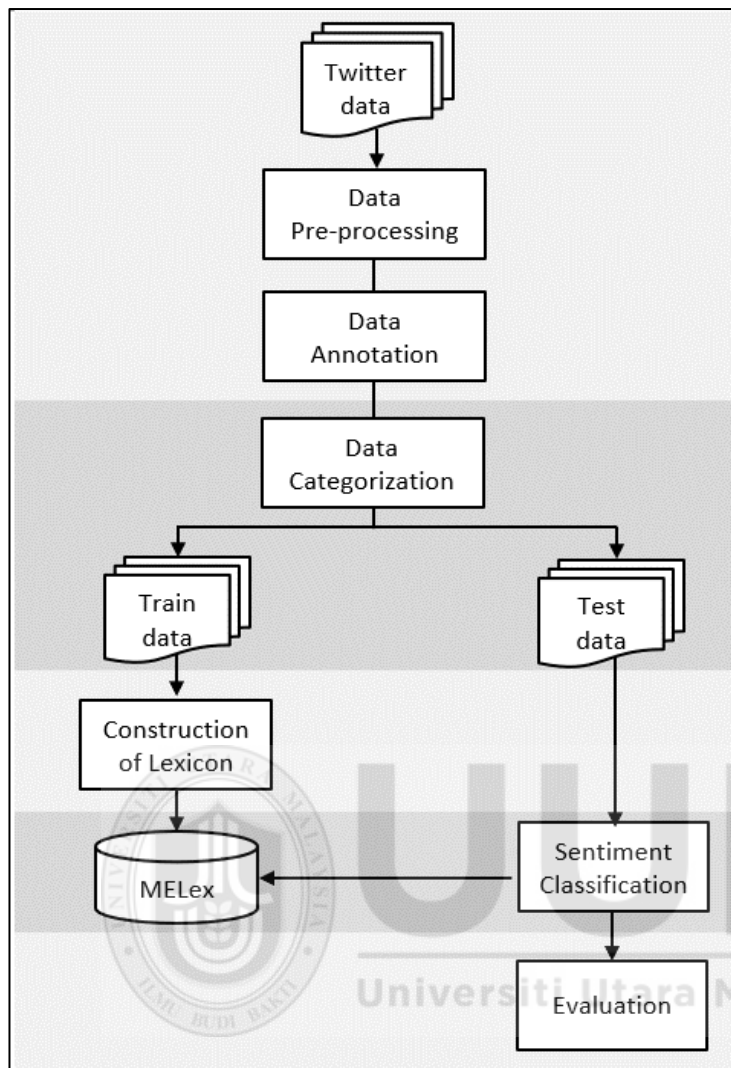


Figure 5.1. Sentiment analysis framework

5.3 Datasets

As mentioned in Section 4.4.2, the data used in this research study was collected from the Twitter platform. There are two datasets collected; PR1MA and PPAM datasets. Table 5.1 shows the keywords used to extract the data and the total number of data extracted for each dataset.

Table 5.1

Keywords Used to Retrieve Data

Datasets	Keywords	No of Data
PR1MA	PR1MA, #pr1ma	16,788
PPAM	PPA1M, #ppa1m	7,049

As stated in Table 5.1, a total of 23,837 tweets were collected for both datasets, and it can be observed that both case studies are varied in the number of data being collected where the difference is 9,739 data. This is expected due to the PPAM project is supplied for government servants only, hence, the number of tweets in regard to this scheme is also limited. This study found that tweets with less than three words were not sufficiently informative to determine the sentiments. Therefore, these tweets were discarded at this phase.

From the initial Twitter search, it has been discovered that most of the affordable housing project-related tweets are shared in Malay and English language. Just a few tweets were found written in other languages such as Mandarin and Tamil. Therefore, tweets written in both Malay and English languages containing keywords specified in Table 5.1 have been collected. Figure 5.2 shows the sample of tweets for the keyword ‘PR1MA’.



Figure 5.2. Sample tweets for PR1MA

The collected tweets are stored in JSON (JavaScript Object Notation) format, and a sample of the extracted data is presented in Figure 5.3.

```
[{"timestamp": "2017-10-01T01:33:42", "text": "betul ke Pr1ma ni murah?", "user": "niladdani", "retweets": "0", "replies": "0", "fullname": "Nil Addani Nizar", "id": "914302268815446016", "likes": "0"}, {"timestamp": "2017-10-01T12:54:17", "text": "boleh simply apply ke for pr1ma houses? ke ada syarat syarat dia", "user": "hanismohdnoor", "retweets": "1", "replies": "0", "fullname": "syahanis mohd noor", "id": "914473544737742848", "likes": "2"}]
```

Figure 5.3. Sample of data extraction

This raw data is then converted to Excel format as depicted in Figure 5.4 and will be preprocessed, which will be explained further in the following section.

timestamp	text	user	retweets	replies	fullname	id	likes
2018-02-25T20:18:40	*KERAJAAN NEGERI HALANG PROJEK KISHimaSam	ShimaSam0	0	0	ShimaSamatOfficia	967856	0
2018-02-25T14:13:11	#alorsetar Kediaman Perumahan Rakyat hmetromy	7	0	0	Harian Metro	967764	4
2018-02-25T11:31:58	PR1MA lah low cost lah sekarang counc warabiyah	1	0	0	Belgrave	967723	0
2018-02-25T09:53:43	c0le pr1ma	lxintergal	1	1	trov	967699	2
2018-02-25T08:49:18	Nak beli rumah PR1MA pun tak lepas	myjoe_	0	2	joe	967682	1
2018-02-25T04:28:59	I'm at Site Pr1ma Alam Damai Cheras ir sayaadala	0	0	0	S.A.K.A	967617	0
2018-02-25T03:17:59	BN GE13: #fighting corruption - Publicly syawal	0	0	1	syawal™ シ	967599	2
2018-02-25T00:33:19	Undi tak undi sama je, br1m tak dapat ngawi_cyb	0	0	0	Ngawi Cyber	967558	0
2018-02-25T00:24:55	Undi tak undi sama je, br1m tak dapat Dee_ayqa	0	0	1	DEE	967555	0
2018-02-24T16:44:09	Applicable ke untuk rumawip, selangor syahredzu	0	0	1	MSR	967440	0
2018-02-24T15:55:42	Yup untuk final year & PSM ataupun d byzulfadh	0	0	1	zulfadhizin	967427	0
2018-02-24T06:05:51	Janji rumah mampu milik 1juta? 500k pgaj_2	4	4	1	G	967279	2
2018-02-24T04:01:49	Yes, Tok pernah dengar kes-kes macam twt_malay	4	0	0	Rakyat: Sarah	967248	2
2018-02-24T03:16:24	Siapa free harini bole mai PR1MA Alam lebah5ml	0	0	0	Faisal Sheckler	967236	0
2018-02-24T01:45:35	PR1MA ni btol2 sejuk bak ang.. lebah5ml	0	0	0	Faisal Sheckler	967213	0
2018-02-23T16:22:20	Ada kenal kwn ni takde hutang ana kec PinkvProf	0	0	1	PinkyProf	967072	0

Figure 5.4. Sample raw data in Excel format

5.4 Data Pre-processing

The extracted data were raw in nature. Therefore, the raw data requires some initial pre-processing prior to the implementation of sentiment analysis to avoid incorrect and misleading results (Awrahman & Alatas, 2017). The pre-processing activity is meant to clean and remove the unnecessary content and symbols from the data.

As demonstrated in Figure 5.5, there are several tasks have been performed such as the removal of repetition tweets (re-tweets), URLs, symbols and hashtags (#), identifying languages either Malay, English or *Bahasa rojak* and PoS tagging. Each task is elaborated further subsequently.

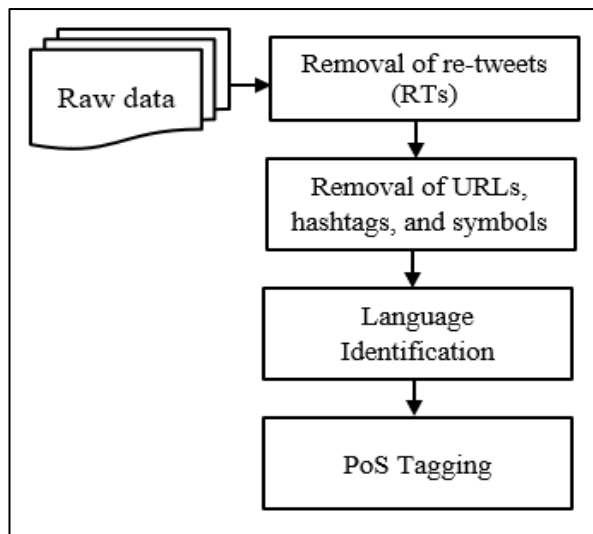


Figure 5.5. Pre-processing activities

5.4.1 Removal of Re-Tweets

The first step in cleaning the raw data was the removal of repetitive or duplicate tweets. This removal can speed up the classification process as re-processing the same data more than once can be avoided and it saves space, in case storage is an issue. The big number of tweets were reduced dramatically during this stage. Figure 5.6 shows the sample code to remove the Re-Tweets (RTs).

```
import io
content = open('duplicate.txt',encoding="utf8").readlines()
content_set = set(content)
clean_data = io.open('duplicate_1.txt','w',encoding="utf-8")

for line in content_set:
    clean_data.write(line)
```

Figure 5.6. Removal of RTs – Sample code

```
= 'Despatch (@Perbadanan PRIMA) https://www.swarmapp.com/c/eDU1A8  
remove_tags(tweet))
```

```
= 'Despatch (@Perbadanan PRIMA) https://www.swarmapp.com/c/eDU1A8  
remove_tags(tweet))
```

```
= 'Despatch (@Perbadanan PRIMA) https://www.swarmapp.com/c/eDU1A8  
remove_tags(tweet))
```

```
= 'Despatch (@Perbadanan PRIMA) https://www.swarmapp.com/c/eDU1A8  
remove_tags(tweet))
```

```
= 'Despatch (@Perbadanan PRIMA) https://www.swarmapp.com/c/eDU1A8  
remove_tags(tweet))
```

```
= 'Despatch (@Perbadanan PRIMA) https://www.swarmapp.com/c/eDU1A8  
remove_tags(tweet))
```

```

from langdetect import detect
import io

with open('identify_language.txt', 'r', encoding="utf8") as f:
    for ln in f:
        a = detect(ln)
        print(a)

```

Figure 5.8. Language identification – Sample code

5.4.4 PoS Tagging

Once the language for each tweet is identified, each sentence in the datasets is tokenized and PoS tagged. PoS tagging is a process of marking up each word in the datasets based on a particular part of speech and it is a necessary process for the purpose of sentiment word extraction (Mohamed, Omar, & Aziz, 2011). Python package called ‘polyglot’ was utilized to assign PoS tags to each word found in the datasets (Al-Rfou, Perozzi, & Skiena, 2013). Figure 5.9 presents the sample code of the PoS tagging process.

```

from polyglot.text import Text

with open('pos_tag.txt', 'r' , encoding="utf8") as f:
    for ln in f:
        text = Text(ln, hint_language_code='en')
        print(text.pos_tags)

```

Figure 5.9. PoS tagging – Sample code

Since there is no available Python package to handle mixed language, tweets identified other than Malay or English were manually tagged as objective or subjective tweets. If it is subjective, the annotators are further annotating those tweets as either positive or negative sentiments. The data annotation process is elaborated in Section 5.5. Table 5.2 gives the meaning of PoS tags used in this research and Table 5.3 shows an example of Malay tweets with its associated PoS tags.

Table 5.2

PoS Tag and Its Definition

Tags	Meaning
ADJ	Adjective
ADV	Adverb
VERB	Verb
NOUN	Noun
PROPN	Proper Noun
X	Other

Table 5.3

Tweets and It's PoS Tags

Tweets	PoS Tagging
rumah ppa1m istimewa untuk penjawat awam	[('rumah', 'PROPN'), ('ppa1m', 'NOUN'), ('istimewa', 'ADJ'), ('untuk', 'ADP'), ('penjawat', 'NOUN'), ('awam', 'NOUN')]
good pr1ma house prices remain unchanged	[('good', 'ADJ'), ('pr1ma', 'NUM'), ('house', 'NOUN'), ('prices', 'NOUN'), ('remain', 'VERB'), ('unchanged', 'ADV')]

After assigning each word with its PoS tags, sentences that match the [adjectives, adverb and verb] rules as proposed by Subrahmanian and Reforgiato (2008) and Shamsudin et al. (2016) are added to the candidate list of the sentiment lexicon. Finally, a total of 7,403 tweets were used for the next process. Further explanation on the candidate list for the sentiment lexicon can be found in Section 6.4.

5.5 Data Annotation

After pre-processing activities, the whole datasets were annotated by two native speakers of Malay language, proficient in English and they are well aware of the case studies used in this research. In this research, detail instructions with few examples were given to the annotators as can be found in Figure 5.10. The annotated datasets were used as a baseline for evaluation purposes during the fifth phase. The annotators were asked to assign each tweet/sentence in the datasets with its polarity. Two polarity types, namely positive and negative, were used in this study.

Instruction:

This data is about PR1MA (Perumahan Rakyat 1Malaysia) and PPAM (Perumahan Penjawat Awam Malaysia) projects. Kindly fill in columns B and C.

Identify language

Column B - fill in the language of text either Malay, English, Mixed or Others.

Mixed means the text contains a mix of English and Malay languages in the same sentence.

Others are referring to languages like Mandarin or Tamil. For language 'Others', you don't have to fill in column C.

Identify polarity

Column C – There are two choices to fill in the polarity column: Positive/Negative

Positive - contains the text contain positive opinion (Example: rumah ni murah, loan aku lulus)

Negative - means the text contain negative opinion (Example: rumah ni mahal, aku tak mampu beli)

Example:

A	B	C
Text	Language	Polarity
PR1MA mampu tengok, SelangorKu mampu milik	Malay	Negative ▼
PR1MA ni memang dah lama aku tak pandang dah. Mahal, mampus milik.	Malay	Negative ▼
Aduh lawa pulak aku tengok rumah pr1ma apartment kat simpang kuala ni	Malay	Positive ▼

Figure 5.10. Instructions for data annotation

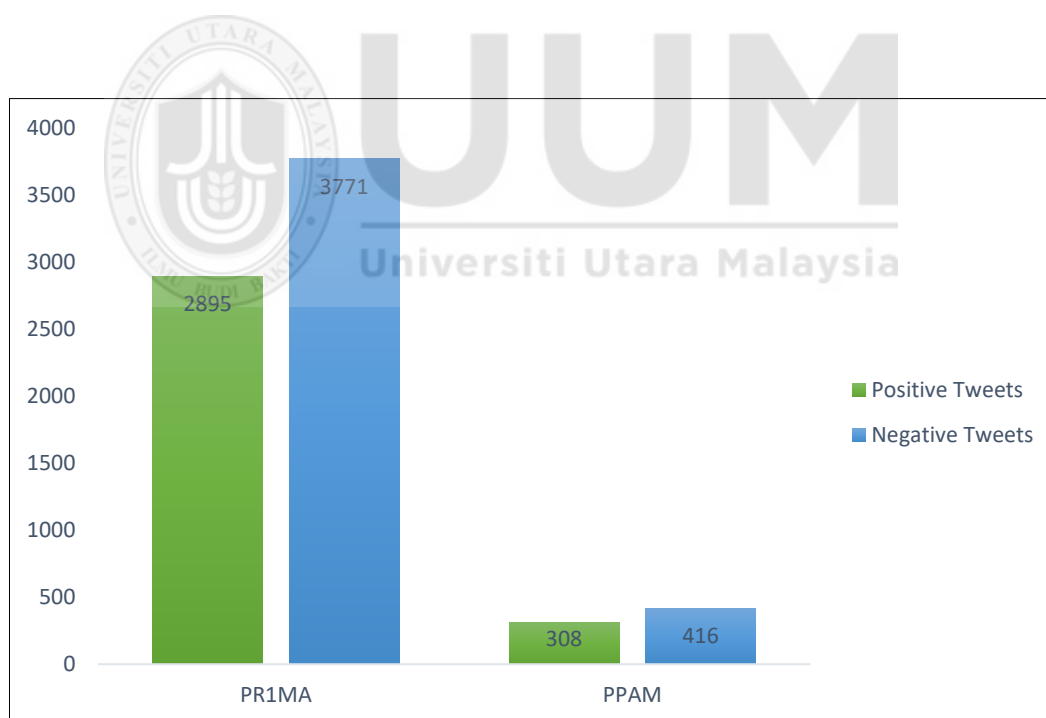
In the end, the total numbers of 7,403 tweets are labeled as either positive or negative sentiments. Table 5.4 presents the language categories annotated as English, Malay or Mixed.

Table 5.4

Language Annotation

Language	No of Data
English	2010
Malay	5113
Mixed	280
Others	-

The results of manual annotation for both case studies are demonstrated in Figure 5.11, where the negative occurrences for PR1MA are higher. In contrast, the positive and negative tweets are relatively balanced for the PPAM project.

*Figure 5.11. Data annotation's results*

5.6 Data Categorization

The next step after the annotation process is to categorize the data into training and testing data. The data is divided into two sets which are:

- i. **Training data** – All the tweets for both case studies are combined to construct a sentiment lexicon. A total of 5,173 data was used to serve this purpose.
- ii. **Testing data** – The data in the test set is used for sentiment classification purposes. A total of 2,230 data is divided into mixed language and single language during the classification process.

The human efforts were needed only up to this stage, the polarity assignment for each sentiment words and synonym expansion in the later stage, which automatically generated, with no human intervention.

Table 5.5

Training and Testing Data

Datasets	Training Data (70%)	Testing Data (30%)
PRIMA	4666	2000
PPAM	507	230
Total	5173	2230

As presented in Table 5.5, there are 7,403 data in total were used for this research study. For a fair split between training and testing data, this study has followed the work by Poria et al. (2017) where 70% of the overall data were used at the training data, and the remaining data were utilized for testing purposes. Table 5.6 summarizes the number of data used for this research.

Table 5.6

Total number of collected data

Dataset	Raw data	No of tweets after pre-processing	No of training data	No of testing data
PR1MA	16788	6666	4666	2000
PPAM	7049	737	507	230
Total	23837	7403	5173	2230

5.7 Chapter Summary

This chapter described the preliminary activities required before the new sentiment lexicon can be generated. There are three tasks have been performed which are; data pre-processing, data annotation and data categorization. Data pre-processing is employed in this study in order to remove the unnecessary symbols and to clean the raw data. Data annotation involves processes like labeling the language and assigning polarity for each tweet. As stated in Section 5.6, the data were categorized as training data, which is used to construct a new sentiment lexicon while testing data is occupied for the sentiment classification process.

In the next chapter, the process of sentiment lexicon generation as well as the evaluation process for the sentiment classification will be detailed out.

CHAPTER SIX

CONSTRUCTION OF MELEX

6.1 Introduction

One of the primary areas of focus within this chapter is the construction of a new bilingual sentiment lexicon specific for the property domain. The first section is started with an overview of the lexicon followed by its architecture. The detail activities and step by step of the lexicon's development are explained in Section 6.3 to Section 6.6. The subsequent chapter outlines the experimental setup, continues with sentiment classification in Section 6.8. The performance evaluation was presented in Section 6.9 and the chapter is ended with a summary in Section 6.10.

6.2 MELEX: Overview

This sentiment lexicon is named as MELEX. The method to construct MELEX consists of three main steps:

- i. Seed words selection: the first step is to choose the sentiment words to be included in MELEX.
- ii. Polarity assignment: the second step is to assign a polarity score for each sentiment word.
- iii. Synonym expansion: the last step is to extend the coverage of MELEX by including synonyms for each sentiment word.

In general, the process of generating MELex is illustrated in Figure 6.1 and several definitions were given as follows.

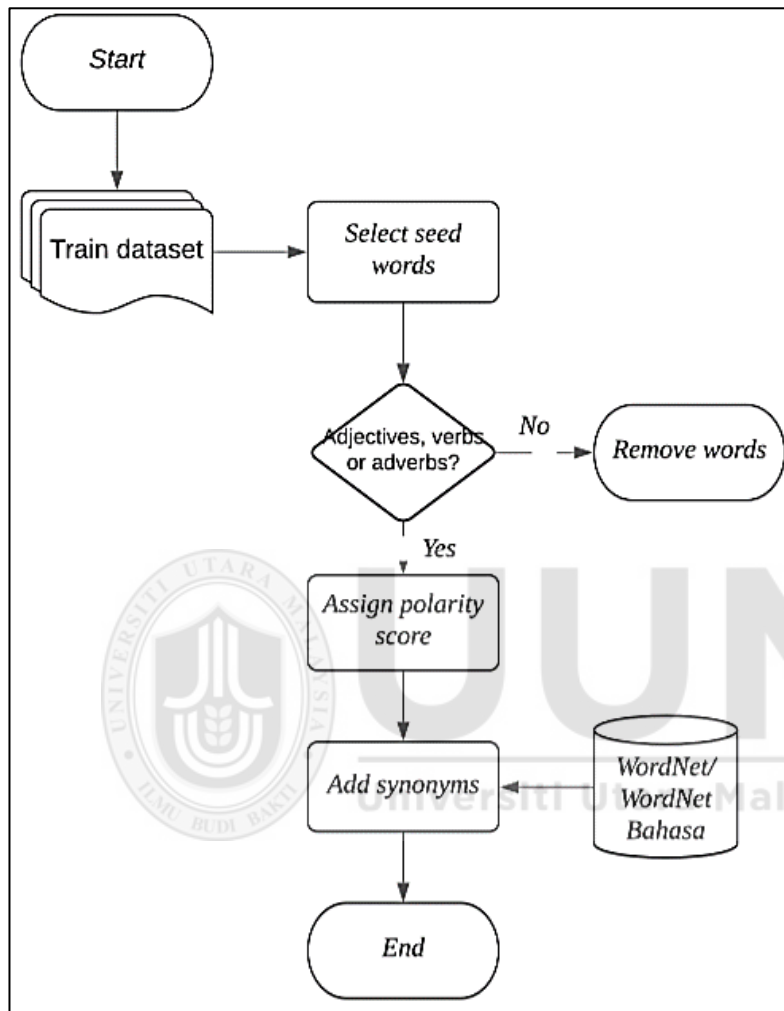


Figure 6.1. MELex's architecture

6.3 Definitions

In the preliminaries chapter, it describes that 70% of the overall manually annotated data for both case studies are allocated for training data and used to develop MELex. This section outlines the definition used throughout this research.

A training data D contains tweets t_i and their manually annotated polarity score $s(t_i)$ was taken as input. As mentioned in the annotation process in Section 5.5, each t_i was annotated as +1 (positive) or -1 (negative). Table 6.1 shows a sample of input data.

Table 6.1

Sample of Input

t_i	$s(t_i)$
ppa1m tu memang rumah idaman tapi terpaksa ditolak (ppa1m was indeed an ideal home but had to be rejected)	-1
i ambik rumah pr1ma je senang kerajaan bagi (I took pr1ma home, easy as it provided by the government)	1
paling selfish pakai kereta mahal duduk apartment pr1ma (someone has a luxury car but stay at pr1ma's apartment, selfish)	-1
mengarut pr1ma starting rm300k (ridiculous pr1ma starting rm300k)	-1
walaupun aku in ppa1m tapi lawyer mahal kecewa (even I got ppa1m but lawyer fees are expensive disappointed)	-1

The output from this process is a bilingual sentiment lexicon MELex containing entries of $S:P$ where the opinion indicator S is each word tagged as adjectives, adverbs or verbs in the training data D and P is its polarity value. The sample of output is presented in Table 6.2, and the next section will describe the selection process of sentiment word S .

Table 6.2

Sample of Output

<i>S</i>	<i>P</i>
Idaman	-0.33
Senang	0.25
Selfish	-0.5
Mahal	-0.5
Berminat	0.75
Disappointed	-0.12

6.4 Seed Words Selection

Using public reviews as seed words is essential in developing a domain-specific lexicon. Therefore, this study proposes the use of Twitter data with specific keywords for the property domain as seed words in generating the lexicon.

The first step in constructing MElex is to detect opinionated words in each t_i by its corresponding PoS Tagging. Following the work presented in Dang, Zhang, and Chen (2010) and Taboada, Brooke, Tofiloski, Voll, and Stede (2011), a combination of adjectives, verbs and adverbs were used as an indicator of subjective/opinionated contents, and hence those words were chosen as a candidate for seed word S in the lexicon MElex. The words tagged other than adjectives, verbs or adverbs were excluded from the lexicon lists as it does not carry any sentiments.

Table 6.3

Sample of Seed Word S

e.g: rumah ni mahal dan jauh (this house is expensive and far away)					
t_i	Rumah	Ni	Mahal	Dan	Jauh
PoS Tag	NOUN		ADJ		ADJ
S	-	-	Mahal	-	Jauh

Table 6.3 displays a sample of seed word selection where the word ‘*mahal*’ (expensive) and ‘*jauh*’ (far) become the only S candidates for t_i ‘*rumah ini mahal dan jauh*’ because both words were tagged as adjectives.

6.5 Polarity Assignment

The next step is to assign polarity value P for each seed word S selected in the previous step. In order to find the most effective way in determining the polarity score P , two experiments involving two techniques; word vector representation v and term frequency tf have been conducted and explained in detail in the following subsections. The discussion on the experiments can be found in Section 6.7.

6.5.1 Word Vector

The first technique, the word vector representation model (Maas, Daly, Pham, Huang, Ng, & Potts, 2011) is employed to calculate the polarity score P . The word vector model returns a vector of length equal to the number of words in a sentence. As shown in Table 6.4, t refers to each tweet in the training data D and w indicates the word in each t where 1 represents the seed word S .

Table 6.4

Representation of the Word Vector Model

	w_1	w_2	w_3	...	w_j
t_1	1	1	-	...	1
t_2	-	1	1	...	-
...
t_i	-	-	1	...	1

To determine the polarity score for each S , it is calculated according to Equation 6.1:

$$S = \sum_{i=1}^T \left(\frac{1}{n} \cdot s(t_i) \right) \quad (6.1)$$

where n is the number of w consist of adjectives, adverbs and verbs in t_i . T indicates the total number of t_i in training data D while $s(t_i)$ is the polarity value manually assigned by the annotators for each t_i . The score P will return a value in the range of -1.0 to +1.0. Table 6.5 shows an example of how polarity score P is calculated for a single t .

Table 6.5

Polarity Score P

e.g: rumah ni mahal dan jauh (this house is expensive and far away)						$s(t_1)$	Total
t_1	Rumah	Ni	Mahal	Dan	jauh	-1	
PoS Tag	Noun		Adjectives		adjectives		
n	0	0	1	0	1	2	2
$n = 2, s(t_i) = -1$							
$p = \sum_{i=1}^T \left(\frac{1}{2} \cdot -1 \right) = -0.5$							
-0.5				-0.5			

As shown in Table 6.5, following the Equation 6.2, the p value for the word ‘*mahal*’ and ‘*jauh*’ is -0.5, respectively. Next, the score of seed word S will be summed up based on its total appearance within document D in order to get the final value P . The P value equals to the total summation of p divide by its term frequency tf as defined in Equation 6.2:

$$P = \frac{\sum p}{tf} \quad (6.2)$$

6.5.2 Term Frequency

Another method introduced in this study is the usage of tf in determining the polarity value P . In this method, the final polarity of P either positive (+) or negative (-) is determined by word vector v discussed in Section 6.5.1.

While the weightage of the polarity which scaled from -3 to +3 is determined by the frequency tf of the seed word S in the training data D . The polarity score P is defined as:

$$P = \begin{cases} +1 \text{ or } -1, & \text{if } tf < 5 \\ +2 \text{ or } -2, & \text{if } 5 \leq tf < 10 \\ +3 \text{ or } -3, & \text{if } tf \geq 10 \end{cases} \quad (6.3)$$

From Equation 6.3, if tf of an S is less than 5, the final value P will be assigned as +1 or -1. If tf is between 5 and 9, then the P -value will be + 2 or -2 and for tf above 9, the P -value is +3 or -3.

6.6 Synonym Expansion

To add more candidate words in MELex, another two experiments were conducted and explained below. The WordNet interface from the NLTK package (Miller, 1995) was used to extract synonyms for each seed word S in the MELex lexicon. Each synonym word was then assigned the same P score as the original seed word S . Table 6.6 demonstrates two examples of seed word S with its synonyms extracted from WordNet.

Table 6.6

Seed Words and Its Synonyms

Language	S	Synonyms
English	Cheap	inexpensive
Malay	Mampu	berdaya, berkaliber, berkebolehan, berkemampuan, berkeupayaan, berupaya

As shown in Table 6.6, the word ‘inexpensive’ is the synonym of seed word ‘cheap’. Hence, the ME_{Lex} will be added with the word ‘inexpensive’ and the score P for ‘inexpensive’ is a similar score assigned for its seed word ‘cheap’. Algorithm 1 depicts the algorithm used for synonym expansion using WordNet.

Algorithm 1 WordNet Synonym

Input: Seed word S in Sentiment Lexicon ME_{Lex} ; WordNet W ; S_i synonym words

Output: Expanded $ME_{Lex} = \{S : P\}$ where S : sentiment word, P : polarity value

Begin

```

1: For each  $S$  in  $ME_{Lex}$  do
2:   If  $S$  exist in  $W$ 
3:     Add  $S_i$  to  $ME_{Lex}$ 
4:   End If
5: End For

```

End

Algorithm 1 shows that WordNet is utilized in order to find a synonym for each seed word in ME_{Lex}. Once the synonym is found, it will be added into the lexicon. There are two resources used in this study to extract the synonyms; English WordNet and WordNet Bahasa as elaborated in the following subsections.

6.6.1 English Resource

As for the English resource, English WordNet was utilized to obtain the synonym for English words (Miller, 1995). This lexical resource group synonym terms in a set called synset that includes a gloss (natural language explanation) for each synset. There are about 117,000 synsets in English WordNet.

6.6.2 Malay Resource

Since the Malay resource is very limited, WordNet Bahasa was applied as part of this study in gathering the synonym for Malay sentiment words in MELex (Noor et al., 2011). Figure 6.2 shows the sample Python code to perform synonym expansion.

```
from nltk.corpus import wordnet as wn
from nltk import word_tokenize as wt

with open('sentiment_words.txt') as a:
    wn_tokens = (a.read())

for token in wt(wn_tokens):
    synonyms = []
    for syn in wn.synsets(token):
        for l in syn.lemmas():
            synonyms.append(l.name())
    print(set(synonyms))
```

Figure 6.2. Sample code of synonym expansion

6.7 Experimental Setup

The purpose of this experimental setup is to determine the words' sentiment weights, expanding the lexicon with new words and the combination of both. Hence, four experiments with different combinations have been conducted and briefly depicted in Table 6.7. The first two experiments; MELex_v1 and MELex_v2 focus on finding the best method to determine the polarity value P and it does not involve any synonym expansion. The other two experiments; MELex_v3 and MELex_v4 involves the synonym expansion process in order to add more candidates to the lexicon. At the end of the experiments, four lexicons are produced.

MELex_v1: The polarity of sentiment words was determined by the word vector representation technique.

MELex_v2: The polarity of sentiment words was determined by the combination of word vector and term frequency technique.

MELex_v3: Similar to MELex_v1 in terms of polarity assignment but involves synonym expansion

MELex_v4: Similar to MELex_v2 in terms of polarity assignment but involves synonym expansion

Table 6.7

Experimental Setup

Experiment	Polarity Assignment		Synonym Expansion
	Word Vector	Term Frequency	
MELex_v1	√	-	-
MELex_v2	√	√	-
MELex_v3	√	-	√
MELex_v4	√	√	√

6.7.1 Experiment 1: MELex_v1

In this experiment, the word vector model as discussed in Section 6.5.1 is employed. Algorithm 2 exhibits the construction process of the MELex_v1.

Algorithm 2 Construction of MELex_v1

Input: Training data $D = \{t_i, s(t_i)\}$

Output: Sentiment lexicon $MELex_v1 = \{S : P\}$ where S : sentiment word, P : polarity value

Initialization: $tf = 0$, where tf : total number of S in D

Begin

```
1: For each word  $w$  in  $t_i$  do
2:   #seed word selection
3:   If  $w = [\text{adj}, \text{adv}, \text{verb}]$  then
4:      $w = S$ 
5:   End If
6:   #polarity assignment
7:   For each  $S$  in  $t_i$  do
8:     assign  $P, tf$ 
9:     #calculate polarity score  $P$ 
10:     $P = s(t_i) / tf$ 
11:   End For
12: End For
```

End

MELex_v1 as outlined in Algorithm 2 shows that the polarity is assigned by dividing the score given by the manual annotators with the frequency of seed words appear in the whole training data. Table 6.8 presents the output for MELex_v1.

Table 6.8

Samples of MELex_v1

1	<i>O</i>	<i>v</i>	<i>TF</i>
2	mahal	-0.32623	79
3	murah	0.142955	57
4	mampu	0.000985	34
5	naik	-0.08276	29
6	lulus	0.14707	26
7	dapat	0.101789	26
8	besar	0.090385	26
9	senang	0.007804	18
10	jauh	-0.03235	17
11	diluluskan	0.000486	9
12	gila	-0.15437	9
13	lancar	0.535714	7
14	tahu	0.006706	6
15	bebankan	-0.8125	4
16	fleksibel	0.044118	4
17	malas	-0.09259	3
18	confirm	-0.75	2
19	lingkup	-0.75	2

6.7.2 Experiment 2: MELex_v2

Experiment 2 is meant to introduce the idea of combining the word vector representation and term frequency technique in order to determine the polarity score. However, the score generated using the word vector model is only to determine the polarity either positive or negative, while the weightage of each individual word S was decided by the term frequency tf within the training data D and the construction process is shown in Algorithm 3.

Algorithm 3 Construction of MELex_v2

Input: Training data $D = \{t_i, s(t_i)\}$

Output: Sentiment lexicon $MELex_v2 = \{S : P\}$ where S : sentiment word, P : polarity value

Initialization: $tf = 0$, where tf : total number of S in D

Begin

```
1: For each word  $w$  in  $t_i$  do
2:   #seed word selection
3:   If  $w = [\text{adj}, \text{adv}, \text{verb}]$  then
4:      $w = S$ 
5:   End If
6:   #polarity assignment
7:   For each  $S$  in  $t_i$  do
8:     assign  $P, tf$ 
9:     #calculate polarity score  $P$ 
10:     $P = s(t_i)/tf$ 
11:   End For
12:   For each  $S$  in  $D$  do
13:     If  $P > 0$  then
14:       If  $(tf < 5)$  then
15:          $P = 1$ 
16:       Else if  $(5 \geq tf < 10)$  then
17:          $P = 2$ 
18:       Else
19:          $P = 3$ 
20:       End if
21:     Else if  $P < 0$  then
22:       If  $(tf < 5)$  then
23:          $P = -1$ 
24:       Else if  $(5 \geq tf < 10)$  then
25:          $P = -2$ 
26:       Else
27:          $P = -3$ 
28:       End if
29:     End if
30:   End For
31: End For
32: End For
End
```

Table 6.9 illustrates the changes in P value depending on its term frequency. For example, for the word ‘mahal’, the final P value is -3 because the frequency of the term ‘mahal’ exceeded 10 and it is assigned a negative value which comes from the v value.

Table 6.9

Samples of Seed Word, tf and Polarity Value

1	O	v	TF		1	O	P
2	mahal	-0.32623	79		2	mahal	-3
3	murah	0.142955	57		3	murah	3
4	mampu	0.000985	34		4	mampu	3
5	naik	-0.08276	29		5	naik	-3
6	lulus	0.14707	26		6	lulus	3
7	dapat	0.101789	26		7	dapat	3
8	besar	0.090385	26		8	besar	3
9	senang	0.007804	18		9	senang	3
10	jauh	-0.03235	17		10	jauh	-3
11	diluluskan	0.000486	9		11	diluluskan	2
12	gila	-0.15437	9		12	gila	-2
13	lancar	0.535714	7		13	lancar	2
14	tahu	0.006706	6		14	tahu	2
15	bebaskan	-0.8125	4		15	bebaskan	-1
16	fleksibel	0.044118	4		16	fleksibel	1
17	malas	-0.09259	3		17	malas	-1
18	confirm	-0.75	2		18	confirm	1
19	lingkup	-0.75	2		19	lingkup	-1

6.7.3 Experiment 3: MELex_v3

In this experiment, the P score was assigned based on the word vector as mentioned in Experiment 1. The purpose of Experiment 3 is to add more sentiment words to the lexicon through the synonym expansion process and the process is demonstrated in Algorithm 4.

Algorithm 4 Construction of MELEX_v3 (word vector as P with synonym expansion)

Input: Training data $D = \{t_i, s(t_i)\}$; *WordNet*: synonym words

Output: Sentiment lexicon $MELEX_v3 = \{S : P\}$ where S : sentiment word, P : polarity value

Initialization: $tf = 0$, where tf : total number of S in D

Begin

```
1: For each word  $w$  in  $t_i$  do
2:   #seed word selection
3:   If  $w = [\text{adj}, \text{adv}, \text{verb}]$  then
4:      $w = S$ 
5:   End If
6:   #polarity assignment
7:   For each  $S$  in  $t_i$  do
8:     assign  $P, tf$ 
9:     #calculate polarity score  $P$ 
10:     $P = s(t_i)/tf$ 
11:   End For
12:   #synonym expansion
13:   For each  $S$  in  $MELEX\_v3$  do
14:     Search synonym  $Y$  in WordNet
15:     Add  $Y$  in  $MELEX\_v3$ 
16:     Assign  $P$  of  $S$  to  $Y$ 
17:   End For
18: End For
End
```

Based on Algorithm 4, the focus is on synonym expansion. A synonym is searched in WordNet for each seed word in order to expand the number of sentiment words covered in MELEX_v3. Once the synonym is found, the score is assigned based on the score for its seed word. The polarity assignment for the third experiment is based on word vector representation as discussed in Section 6.7.1. Table 6.10 shows the sample of output for MELEX_v3.

Table 6.10

Sample of MELex_v3

<i>S</i>	<i>P</i>
Expensive	-0.125
Menarik	0.25
Insulting	-0.09091
Mudah	0.1191
Cepat	0.1985
Jauh	-0.0323
Highest	0.25

6.7.4 Experiment 4: MELex_v4

In this experiment, the *P* score was based on the term frequency as assigned in Experiment 2. As shown in Algorithm 5, the focus for Experiment 4 is to expand the lexicon by including the synonym for each seed word. A synonym is searched in WordNet in order to include more sentiment words covered in MELex_v4. Once the synonym is found, the score is assigned based on the score of its seed word. The polarity assignment for the fourth experiment is based on the integration of word vector and term frequencies as discussed in Section 6.7.2. Table 6.10 shows the sample of output for MELex_v4.

Algorithm 5 Construction of MELex_v4

Input: Training data $D = \{t_i, s(t_i)\}$; *WordNet*: synonym words

Output: Sentiment lexicon $MELex_v4 = \{S : P\}$ where
 S : sentiment word, P : polarity value

Initialization: $tf = 0$, where tf : total number of S in D

Begin

```
1: For each word  $w$  in  $t_i$  do
2:   #seed word selection
3:   If  $w = [\text{adj}, \text{adv}, \text{verb}]$  then
4:      $w = S$ 
5:   End If
6:   #polarity assignment
7:   For each  $S$  in  $t_i$  do
8:     assign  $P, tf$ 
9:     #calculate polarity score  $P$ 
10:     $P = s(t_i) / tf$ 
11:   End For
12:   For each  $S$  in  $D$  do
13:     If  $P > 0$  then
14:       If  $(tf < 5)$  then
15:          $P = 1$ 
16:       Else if  $(5 \geq tf < 10)$  then
17:          $P = 2$ 
18:       Else
19:          $P = 3$ 
20:       End if
21:     Else if  $P < 0$  then
22:       If  $(TF < 5)$  then
23:          $P = -1$ 
24:       Else if  $(5 \geq tf < 10)$  then
25:          $P = -2$ 
26:       Else
27:          $P = -3$ 
28:       End if
29:     End if
30:   End For
31:   #synonym expansion
32:   For each  $S$  in  $MELex\_v4$  do
33:     Search synonym  $Y$  in WordNet
34:     Add  $Y$  in  $MELex\_v4$ 
35:     Assign  $P$  of  $S$  to  $Y$ 
36:   End For
37: End For
End
```

Table 6.10

Sample of MELex_v4

<i>S</i>	<i>P</i>
Expensive	-3
Menarik	1
Insulting	-1
Mampu	3
Murah	3

6.8 Sentiment Classification

To examine the performance of all the four versions of MELex, classification on the testing data was applied. The basic idea is as follows; given a test data, opinion words in that sentence are first identified by matching with the words in the MELex lexicon. Then the score is assigned based on the score stated in the lexicon which then the final score is summed up for the whole sentence. Algorithm 6 shows how the classification process is conducted.

Algorithm 6 Sentiment classification process

Input: Testing data T , Lexicon $MELex$, Negation file N

Output: $S = \{Pos, Neg \text{ or } Neut\}$, where Pos : Positive, Neg : Negative, $Neut$: Neutral

Initialization: Total_Pos and Total_Neg = 0, where

Total_Pos: accumulates the positive polarity t_i in T

Total_Neg: accumulates the negative polarity t_i in T

Begin

```
1: For each  $t_i$  in  $T$  do
2:   For each word  $w$  in  $t_i$  do
3:     Search for  $w$  in  $MELex$ 
4:     If  $w$  exist in  $MELex$  then
5:       assign  $P \leftarrow w$ 
6:       If word before  $w$  exist in  $N$  then
7:         reverse  $P$ 
8:       End If
9:     End If
10:  End For
11:  If  $P > 0$  then
12:    Total_Pos  $\leftarrow$  Total_Pos +  $t_i$ 
13:  Else if  $P < 0$  then
14:    Total_Neg  $\leftarrow$  Total_Neg +  $t_i$ 
15:  End If
16: End For
17: If Total_Pos > |Total_Neg| then
18:    $S = Pos$ 
19: Else if Total_Pos < |Total_Neg| then
20:    $S = Neg$ 
21: Else
22:    $S = Neut$ 
23: End If
```

End

As shown in Algorithm 6, to determine the polarity of sentences in test data, the opinion word is checked in the MELEX lexicon. If the word exists in the MELEX, the polarity score of the specified word is assigned. If the word cannot be found in MELEX, the sentence is assigned as 'neutral'. The final sentiments for each sentence are calculated based on the summation of the score assigned.

Negation handling – In this work, negation was also handled in the classification process. Without negation handling, when negation word is coupled with a positive word, it classifies a sentence as a positive instead of negative. The negation terms in Malay and English language extracted from Lo et al. (2017) and Shamsudin et al. (2016)’s works are added to a negation file *N*. Here, the polarity of sentiment term under the influence of negation is inverted. For instance, “*not bad jugak rumah pr1ma ni*” would return positive sentiment due to the presence of negation term “*not*” prior to sentiment word “*bad*” where the negative polarity for the word “*bad*” as stated in MELEX was reversed to positive polarity. Figure 6.3 listed all the negation words used in this research.



<i>bukan</i>	<i>not</i>
<i>tak</i>	<i>no</i>
<i>xde</i>	<i>without</i>
<i>x</i>	<i>nil</i>
<i>tidak</i>	<i>n't</i>
<i>bkn</i>	<i>never</i>
<i>takde</i>	<i>none</i>
<i>tanpa</i>	<i>neither</i>
<i>tiada</i>	<i>nor</i>
<i>takkan</i>	<i>non</i>
<i>usah</i>	<i>deny</i>
<i>jangan</i>	<i>reject</i>
<i>belum</i>	<i>refuse</i>

Figure 6.3. List of negation words

The flowchart presented in Figure 6.4 has detailed out how the polarity for each sentence is determined.

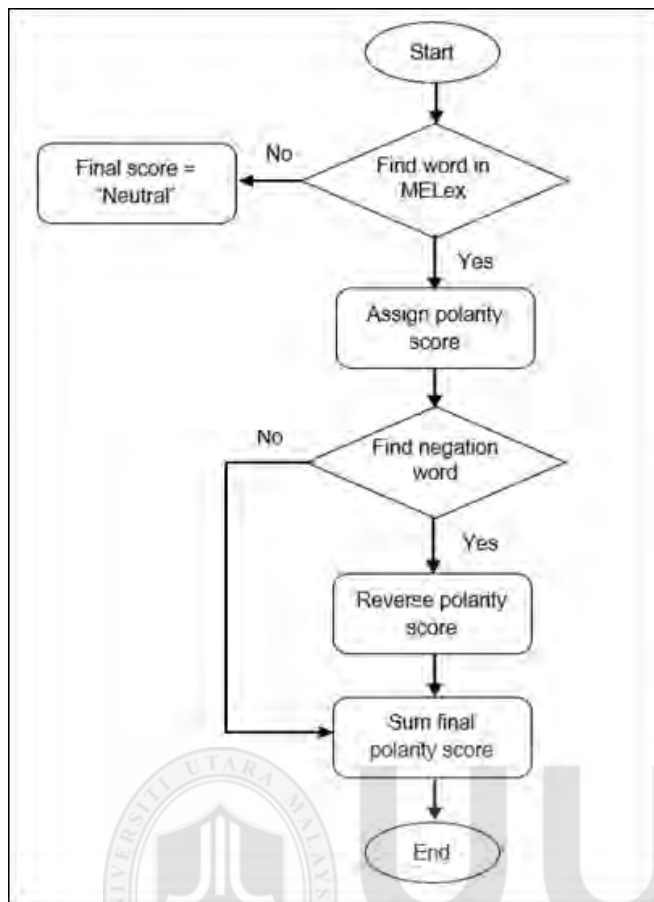


Figure 6.4. Flowchart to determine the polarity

To calculate the final score of each tweet/sentence, the Term Counting method is applied. It is a method introduced by Ohana and Tierney (2009) to classify positive and negative sentiments by counting the positive and negative polarity found in a text. The sentiment score of each review is calculated and categorized based on the score, as mentioned in Table 6.11.

Table 6.11

Polarity Criteria

Condition	Polarity
Sentiment score > 0	Positive
Sentiment score = 0	Neutral
Sentiment score < 0	Negative

Based on Table 6.11, a tweet is categorized as positive if the highest polarity score is positive and vice versa. A tweet turns neutral if the sentiment score is equal to 0. The highest polarity score determines the final polarity. Figure 6.5 illustrates a sample of sentiment classification's output.

```

103 print(lex.main("ini tak mampu milik ni mampu tengok je"))
104 print(lex.main("best prima ni"))
105 print(lex.main("terbaik"))
106
{'score': ['ini tak mampu [-2.0] milik [2.0] ni mampu [2.0] tengok [-3.0] je'], '=': 'Negative'}
{'score': ['best [2.0] prima ni'], '=': 'Positive'}
{'score': ['terbaik [1.0]'], '=': 'Positive'}

```

Figure 6.5. Sample of sentiment classification's output

6.9 Performance Evaluation

There is no existing dataset available to evaluate the quality of the developed lexicon, which is in the form of a sentiment score. Therefore, to evaluate the effectiveness of the proposed approach, two evaluation methods as suggested by Muhammad (2016) have been performed; i) evaluation metrics and ii) baseline comparisons.

6.9.1 Evaluation Metrics

Table 6.12 shows the confusion matrices used for sentiment classification.

Table 6.12

Confusion Matrix

	True (Predicted)	False (Predicted)
Positive (Actual)	True Positive (TP)	False Positive (FP)
Negative (Actual)	True Negative (TN)	False Negative (FN)

Based on the confusion matrix,

True Positive (TP): Number of positive tweets classified correctly.

False Positive (FP): Number of negative tweets classified incorrectly as a positive.

True Negative (TN): Number of negative tweets classified correctly.

False Negative (FN): Number of positive tweets classified incorrectly as a negative.

Accuracy shows how accurate is the true positive and true negative instances compare to all the possible cases. Accuracy A is defined as:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.4)$$

Precision at the opinion level, shows how many of extracted opinions are correct.

Precision is defined as the positive predictive value that is presented in Equation 6.5.

$$P = \frac{TP}{TP + FP} \quad (6.5)$$

A recall is defined as the true positive rate that is presented in Equation 6.6. For example, the recall of opinion extraction shows how many opinions are extracted.

Recall R is defined as:

$$R = \frac{TP}{TP + FN} \quad (6.6)$$

Finally, F-measure F is defined as:

$$F = 2 \cdot \frac{P \cdot R}{P + R} \quad (6.7)$$

6.9.2 Baseline Comparisons

Another evaluation method is to compare with other prominent classifiers from both sentiment analysis approaches; lexicon-based and machine learning. For the lexicon-based approach, the AFINN's lexicon is used while SVM and NB are used as a classifier for the machine learning approach.

6.9.2.1 General Purpose Lexicon

It is worth to note that there is no previous work on Malay-English sentiment lexicon made publicly available, hence the absence of a baseline to allow for a feasible comparison. AFINN Lexicon is chosen to be used as a baseline as the other researchers used this resource in generating their lexicon.

AFINN-111 – AFINN Lexicon is an English based sentiment lexicon developed by Nielsen (2011). This sentiment dictionary used polarity scale ranging from -5 (extremely negative) to +5 (extremely positive) and it includes 3,382 English words. Following the work presented by Tan et al. (2016) in creating the SentiLexM lexicon, all the English words in AFINN are translated into Malay and the sentiment score is assigned similar to its original English words. Overall, the total number of words including English and its Malay translated word is 6,764 sentiment words.

6.9.2.2 Machine Learning Classifiers

The machine learning classification has been reported to perform well in English content, and two well-known machine learning classifiers were identified to be included in this experiment. Figure 6.6 shows a flow chart on the machine learning classification process implemented in this study.

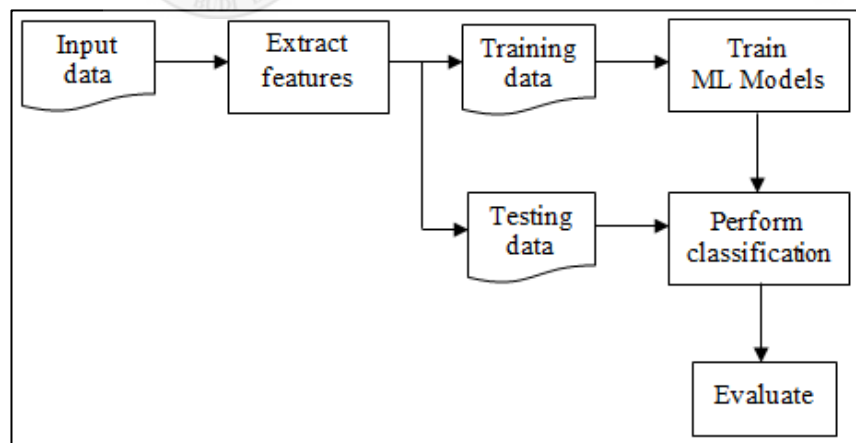


Figure 6.6. Flow chart: Machine learning classification

This research relied on an open-source Python library, the scikit-learn library in performing the machine learning classification (Pedregosa et al., 2011). The library includes several classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM) and Logistic Regression. Furthermore, scikit-learn is user-friendly, which makes it easy to be implemented. For performance evaluation, two commonly used machine learning classification algorithms were considered; NB (Tan, Cheng, Wang, & Xu, 2009) and SVM (Mullen & Collier, 2004) and described further below.

1. Feature extraction – As machine learning classification requires mathematical formats to train the models, the textual data is first needed to be transformed into numeric form. Feature extraction is meant to perform the task of converting the textual data into the numeric form as illustrated in Figure 6.7. The ‘TfidfVectorizer’ class from Python’s scikit-learn library (Hackeling, 2017) was employed to tokenize the documents and convert the most frequently occurring words into a bag of words feature vectors.

```
from sklearn.feature_extraction.text import TfidfVectorizer
text = ["such a waste of money", "beautiful house"]
# create the transform
vectorizer = TfidfVectorizer()
# tokenize and build vocab
vectorizer.fit(text)
# summarize
print(vectorizer.vocabulary_)
print(vectorizer.idf_)
# encode document
vector = vectorizer.transform([text[0]])
print(vector.toarray())

{'such': 4, 'waste': 5, 'of': 3, 'money': 2, 'beautiful': 0, 'house': 1}
[ 1.40546511  1.40546511  1.40546511  1.40546511  1.40546511  1.40546511]
[[ 0.   0.   0.5  0.5  0.5  0.5]]
```

Figure 6.7. Sample output of feature extraction using TfidfVectorizer

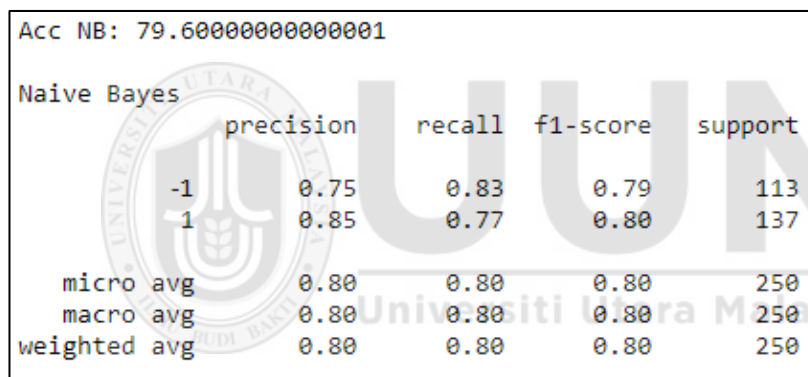
2. Categorizing Data into Training and Testing Sets – Once the data is converted into numeric form, the data is divided into training and testing datasets. This research has used the same training dataset for MELex’s construction to train the machine learning algorithm and validated the performance using a similar testing dataset in evaluating MELex and AFINN lexicons. The ‘train_test_split’ class from the ‘model_selection’ module under the scikit-learn library was used to categorize the data into training and testing data. The percentage of 70/30 was used in splitting the data, similar to the percentage applied in dividing data for MELex’s construction.

3. Train the Model – Once the data is split into training and testing, the machine learning algorithms are applied to learn from the training data. In this research, the NB and SVM algorithms are chosen due to their ability to acquire promising results in the previous research.

- i. **Naïve Bayes (NB)** – The NB classifier relies on Bayesian probability and it is a well-known machine learning technique for sentiment analysis due to its simplicity and effectiveness (Hutto & Gilbert, 2014). The sklearn.naive_bayes module is used to train the model using the NB algorithm.
- ii. **Support Vector Machine (SVM)** – SVM classifier has shown to be highly effective for classification task and it generally outperforms other machine learning classifiers (Shein & Nyunt, 2010). Different from the NB classifier, SVM utilizes hyperplane to separate classes and represented by a support vector that distinguishes the positive and negative training vectors. The sklearn.svm is applied to perform the SVM algorithm in training the model.

4. Machine Learning Classification – Once the model is trained, the next step is to perform sentiment classification on the model. A module called ‘predict’ to do the classification or prediction on the testing datasets.

5. Evaluation – The same evaluation metrics, which are Accuracy, Recall, Precision and F-measure used to evaluate the MELex lexicon is applied to evaluate machine learning classifiers. The modules `classification_report`, `accuracy_score` and `confusion_matrix` under `sklearn.metrics` library were used to obtain the values. The sample of output from the evaluation process is shown in Figure 6.7.



```
Acc NB: 79.60000000000001
Naive Bayes
      precision    recall  f1-score   support
-1      0.75      0.83      0.79       113
 1      0.85      0.77      0.80       137

 micro avg      0.80      0.80      0.80      250
 macro avg      0.80      0.80      0.80      250
weighted avg      0.80      0.80      0.80      250
```

Figure 6.8. Sample output for machine learning classifiers

As presented in Figure 6.8, the accuracy obtained using the NB algorithm is 79.6% for the data size of 250 with 113 labeled as negative (-1) and 137 positives (+1) sentiments.

6.10 Chapter Summary

This chapter has detailed processes involved in generating the new lexicon known as MELEX. The four experiments involving four different combinations have been carried out with the objective of identifying the best technique in assigning scores for sentiment words in the sentiment lexicon. In other words, at the end of this study, four versions of MELEX are produced.

MELEX_v1 and MELEX_v2 are the versions that focused on finding the best techniques in determining the polarity score, while MELEX_v3 and MELEX_v4 concentrated on synonym expansion in order to add more sentiment words in the lexicon. English WordNet and WordNet Bahasa were used to obtain the synonym words.

The sentiment classification using the Term Counting technique was elaborated and the performance evaluation was included in this chapter. The AFINN lexicon, as well as NB and SVM classifiers were applied as a baseline to assess the performance. The next chapter will explain and discuss the results obtained from this experimental study.

CHAPTER SEVEN

RESULTS AND DISCUSSION

7.1 Introduction

In the previous chapter, a bilingual and domain-specific sentiment lexicon known as MELex was developed. There were four experiments have been conducted which produced four versions of MELex. In this chapter, the results obtained from the experiments are presented in Section 7.2 and its performance is discussed in Section 7.3.

7.2 Results

In this section, the results obtained from the development of sentiment lexicon; MELex and the output from the sentiment classification for PR1MA and PPAM test sets are presented. The performance of MELex is compared with the other baselines and the evaluation is elaborated accordingly.

7.2.1 MELex

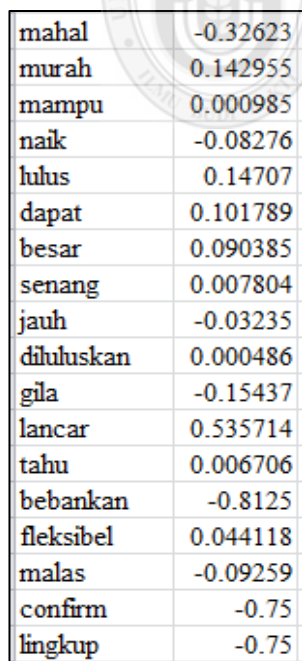
As explained in the previous chapter, MELex is a lexical resource constructed using tweets data as a seed word. Four experiments with different combinations have been implemented. As a result, four versions of MELex have been created; MELex_v1, MELex_v2, MELex_v3 and MELex_v4. The number of words generated for each version is presented in Table 7.1.

Table 7.1

Number of Words Generated

Version	No of Positive Words	No of Negative Words	Total
MELex_v1	1069	1151	2220
MELex_v2	1069	1151	2220
MELex_v3	3521	2611	6132
MELex_v4	3521	2611	6132

Based on Table 7.1, MELex_v1 and MELex_v2 contain 2,220 entries with 1,069 positive and 1,151 negative words. Through the synonym expansion process implemented in MELex_v3 and MELex_v4, a total number of sentiment words in both versions achieved 6,132 words. The samples of the MELex lexicon are shown in Figure 7.1 and Figure 7.2.



mahal	-0.32623
murah	0.142955
mampu	0.000985
naik	-0.08276
lulus	0.14707
dapat	0.101789
besar	0.090385
senang	0.007804
jauh	-0.03235
diluluskan	0.000486
gila	-0.15437
lancar	0.535714
tahu	0.006706
bebankan	-0.8125
fleksibel	0.044118
malas	-0.09259
confirm	-0.75
lingkup	-0.75

Figure 7.1. Sample MELex_v1 and MELex_v3

sentiment word	score
aneh	-1
bad	-1
badai	-1
badly	-1
bagus	2
bahagia	1
anggun	1
annoying	-1
approved	1
aset	1
batal	-1
clear	1
closed	-1
complaint	-1
complementary	1
completed	1
cuai	-1
curi	-1
curiga	-1
dahsyat	-1
daif	-1
dakwa	-1
dalih	-1

Figure 7.2. Sample MELex_v2 and MELex_v4

Other than the number of words generated for each MELex's version, the difference between Figure 7.1 and Figure 7.2 is in the polarity score assigned for each sentiment word. Figure 7.3 illustrates the word cloud represents 50 most frequent sentiment words stored in MELex_v1 and MELex_v3. From that figure, it can be summarized that the word '*mampu*' and '*ada*' are the most frequent terms appear in the training data.

Table 7.2

Sample of the Output: Tweets and Its Polarity

Tweets	Polarity
mohonlah sekarang pr1ma, cerah masa depan. seremban pun dah dekat ngan kuala lumpur <i>(apply now for a bright future. In fact, Seremban is near to Kuala Lumpur)</i>	Positive
padan muka group of women cheated greedy vips in pr1ma housing scam <i>(serve you right, group of women cheated greedy vips in pr1ma housing scam)</i>	Negative
pr1ma skim yang paling popular, dibuat bagi memenuhi keperluan rumah golongan kelas tengah <i>(pr1ma, the famous scheme, developed to meet the needs of the middle-income group)</i>	Positive

To evaluate the performance of each version of MELex, the results obtained were compared with the manually annotated testing data. Using the evaluation measures of accuracy, precision, recall and F-measure as mentioned in Section 6.9.1, the results are calculated and compared.

The reason for the division between mixed and single language is to examine the influence of mixed language's accuracy towards the overall performance. The classification's results for PR1MA and PPAM test sets are presented in the following subsections.

7.2.2.1 PR1MA

A total of 2,000 tweets were used as testing data. Out of 2,000 tweets, 250 tweets consist of mixed language data and the other 1,750 tweets were identified as either Malay or English language only. Table 7.3 shows the type of data, which indicates that mixed language content contributed 12.5% to the total number of PR1MA test data.

Table 7.3

PR1MA: Type of Data

	No of Data
Mixed Language	250
Single Language	1750
Total	2000

The results of sentiment classification are divided into mixed language and single language and presented as follows.

Mixed Language - Table 7.4 presents the confusion matrix of sentiment classification for 250 mixed language tweets extracted from the PR1MA test set. It is notable that the number of sentences predicted as neutral has decreased significantly. The evaluation metrics are calculated based on the overall confusion matrix and presented in Table 7.5.

Table 7.4

Confusion Matrix: PRIMA - Mixed Language

		MELex_v1			MELex_v2			MELex_v3			MELex_v4		
Polarity	No of Sentences	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral
Positive	98	58	24	29	67	20	31	75	21	5	89	7	6
Negative	152	69	70		47	85		46	103		23	125	

Table 7.5

Evaluation Metrics: PRIMA - Mixed Language

Metric	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
MELex_v1	47.94	45.67	45.3	45.5
MELex_v2	69.41	58.77	77	66.67
MELex_v3	72.65	62	78.13	69
MELex_v4	87.7	79.5	92.7	85.6

From Table 7.5, it can be seen that the accuracy increased by 24.71% and 18.29% when synonym expansion applied in MELex_v3 and MELex_v4. In fact, the recall and F-measure are distinguished good which is about 92.7% and 85.5%, respectively. A promising result of F-measure indicated that the performance of both precision and recall are considered balance when classified using MELex_v4.

Single Language – There are 1,750 tweets consist of either Malay or English language, which shows that single language tweets contributed 87.5% towards the overall results. Table 7.6 shows the results obtained from the classification and its evaluation is presented in Table 7.7.

Table 7.6

Confusion Matrix: PR1MA - Single Language

Polarity	No of Sentences	MELex_v1			MELex_v2			MELex_v3			MELex_v4		
		Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral
Positive	697	401	267	62	449	236	59	540	146	43	607	51	40
Negative	1053	480	540		335	671		130	891		114	938	

Table 7.7

Evaluation Metrics: PR1MA – Single Language

Version	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
MELex_v1	55.75	45.5	60	51.78
MELex_v2	66.23	57.3	65.55	61.13
MELex_v3	83.83	80.6	78.7	79.65
MELex_v4	90.35	84.2	92.25	88

As observed from the result presented in Table 7.7, MELex_v4 yielded the best Recall of 92.25% for the classification of a single language. Besides, the accuracy obtained was the best as compared to the first three versions of MELex.

Overall classification – Figure 7.4 illustrates the overall sentiment of the public towards the PR1MA project for 2,000 test data classified using manual annotation as well as the four versions of the MELex lexicon.

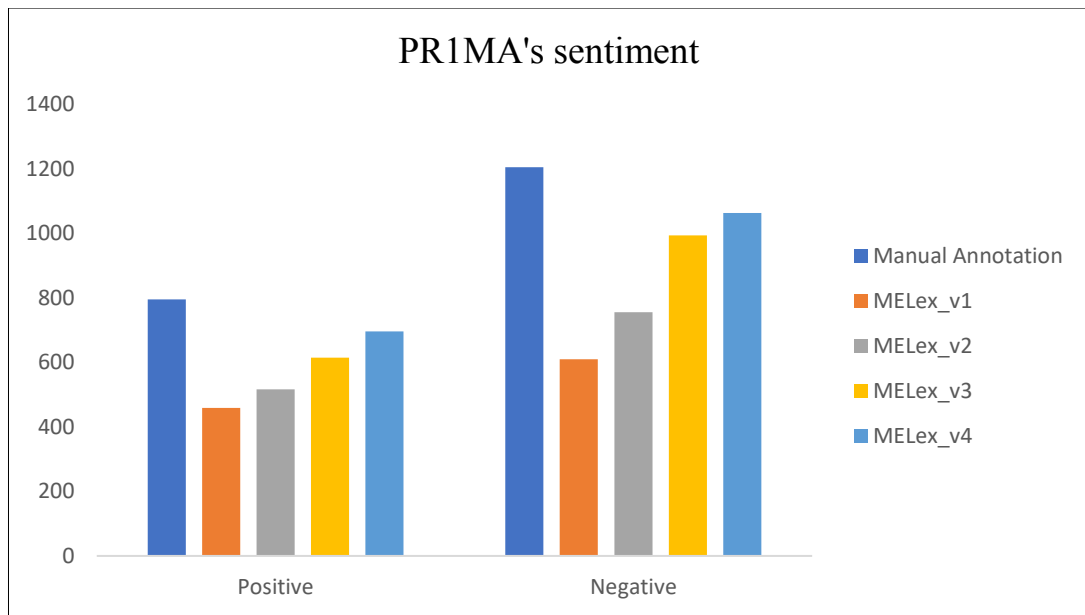


Figure 7.4. Sentiment's result for PR1MA

The results presented in Figure 7.4 have clearly display that MELex_v4 has obtained the closest result as manually annotated data for both polarities; positive and negative. The overall classification result for the PR1MA dataset with regards to the accuracy, precision, recall and F-measure are detailed out in Table 7.8.

Table 7.8

Overall Performance - PR1MA

Version	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
MELex_v1	54.68	45.5	57.66	50.89
MELex_v2	66.6	57.46	66.84	61.8
MELex_v3	82.43	77.75	78.64	78.2
MELex_v4	90.02	83.55	92.3	87.7

As displayed in Table 7.8, all four versions of MELex reported low precision metrics followed by F-measure. The accuracy and recall metrics look good especially for

MELex_v4. The reason of low F-measure is most probably due to the absent of sentiment words in the lexicon, hence the data were classified as ‘neutral’ sentiments.

7.2.2.2 PPAM

Unlike the PR1MA test dataset, the total number of PPAM data to be classified is much smaller. For the PPAM project, a total of 230 tweets have been used for testing purposes, which consists of 30 mixed language and 200 single language data as listed in Table 7.9.

Table 7.9

PPAM: Type of Data

	No of Data
Mixed Language	30
Single Language	200
Total	230

Mixed language – The classification results for 30 mixed language data consist of 13 positive sentences and 17 negative sentences are presented in Table 7.10 and Table 7.11 shows its evaluation metrics.

Table 7.10

Confusion Matrix: PPAM - Mixed Language

Polarity	No of Sentences	MELex_v1			MELex_v2			MELex_v3			MELex_v4		
		Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral
Positive	13	3	1	10	7	3	9	6	2	3	11	2	2
Negative	17	9	7		2	9		9	10		2	13	

Table 7.11

Evaluation Metrics: PPAM – Mixed Language

Version	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
MELex_v1	50	25	75	37.5
MELex_v2	76.2	77.8	70	73.68
MELex_v3	59.26	40	75	52.17
MELex_v4	85.71	84.61	84.61	84.61

As compared to the PR1MA test set, it is noticeable that MELex_v2 performs better than MELex_v3 by approximately 17% in accuracy and 21.51% in F-Measure. However, in terms of recall, MELex_v2 gives a 5% lower performance than MELex_v1 and MELex_v3.

Single Language – There are 200 tweets consist of either Malay or English language, which shows that single language tweets contributed 87% towards the overall results.

Table 7.12 shows the results obtained from the classification using MELex.

Table 7.12

Confusion Matrix: PPAM - Single Language

		MELex_v1			MELex_v2			MELex_v3			MELex_v4		
Polarity	No of Sentences	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral	Predicted Positives	Predicted Negatives	Predicted Neutral
Positive	95	50	37	18	61	29	17	69	21	8	85	5	6
Negative	105	54	41		38	55		39	63		15	89	

Table 7.13

Evaluation Metrics: PPAM - Single Language

Version	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
MELex_v1	50	48.57	57.47	52.36
MELex_v2	63.39	61.6	67.78	64.55
MELex_v3	68.75	63.8	76.6	69.7
MELex_v4	89.7	85	94.44	89.5

For the classification of a single language for the PPAM project, MELex_v4 has shown remarkable results with the highest recall of 94.44%.

Overall Classification – The sentiment results for the PPAM project based on 230 test data is shown in Figure 7.5.

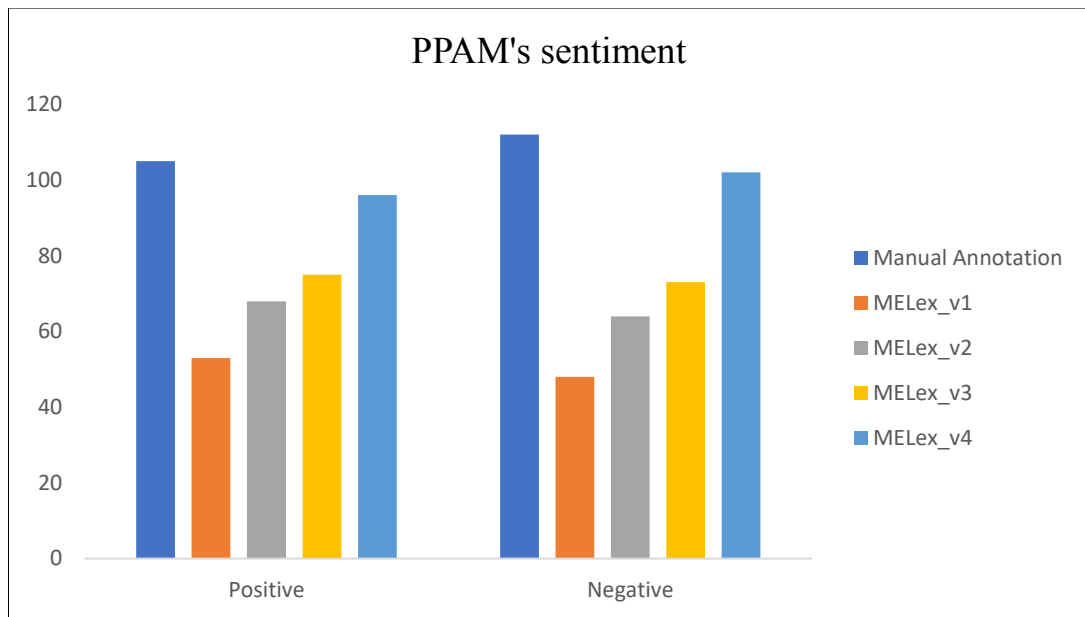


Figure 7.5. Sentiment's result for PPAM

Table 7.14 summarizes the overall performance of PPAM test sets, where 13% of the results are coming from mixed language classification and the remaining percentage is from single language classification.

Table 7.14

Overall Performance – PPAM

Version	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
MELex_v1	50	45.69	58.24	51.2
MELex_v2	64.71	63	68	65.38
MELex_v3	67.58	61	76.5	67.87
MELex_v4	89.19	85	93.2	88.9

As depicted in Table 7.14, the results of the evaluation show that MELex_v4 outperforms the other three versions, with higher accuracy. According to the values of precision and recall, the results show that the MELex_v4 performs well in analyzing either positive or negative reviews.

7.2.2.3 Results Analysis

From the results presented in the previous sections for both datasets, it can be summarized that MELex_v4 has outperformed the other versions in all comparisons conducted; either for a mixed-language or single language.

Both MELex_v1 and MELex_v2 obtained lower accuracy in both test sets; PR1MA and PPAM. This is expected due to the limited number of words covered in both lexicons. As for the MELex_v3 and MELex_v4, the accuracy reported is higher than the first two versions. This is probably due to the synonym expansion that leads to the improvement of the performance for both test sets.

The performance of the classification for mixed-language content using MELex_v4 had achieved an accuracy of more than 80%, while the classification for single language tweets exhibits a consistent result with 90.35% and 89.7% accuracy, respectively.

As for the overall performance, both PR1MA and PPAM using MELex_v4 have obtained low precision and high recall as shown in Table 7.8 and Table 7.14. The low precision achieved indicates that the classification using MELex has misclassified negative reviews more than the positive ones.

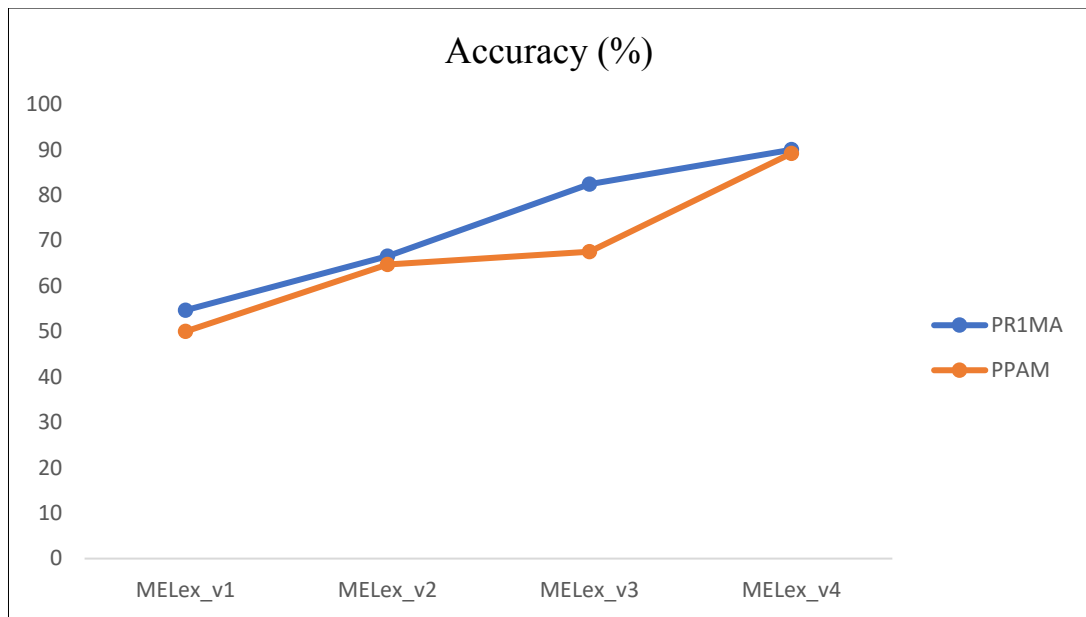


Figure 7.6. Accuracy's results

As demonstrated in Figure 7.6, the accuracy obtained using the ME Lex_v4 lexicon is undoubtedly promising with the overall accuracy of 90.02% and 89.7% for both test sets. The classification of mixed language alone has contributed approximately 11% to the overall accuracy. Even though the classification suffers from a low precision value, the results achieved relatively high values in both recall and F-measure.

ME Lex_v1 gives the lowest performance for both datasets in all measures, be it in mixed language or single language classification. In particular, ME Lex_v4 achieves the highest performance among all other versions. The results obtained indicated that using the integration of word vector representation and term frequency to determine the sentiment's score and the expansion of the lexicon through WordNet synonym do contribute to better classification's performance.

It can be summarized that how the MELex is built and how the polarity score is determined does affect the accuracy of the classification. The modification of the polarity score based on term frequency has increased the accuracy significantly. In fact, the synonym expansion has also improved the classification's performance. Based on the results produced, MELex_v4 performs the best as compared to the other versions. Therefore, in the following section, other experiments involving SVM and NB classifiers as well as AFINN lexicon were conducted and compared with MELex_v4.

7.2.3 Performance Comparison

The main objective of this evaluation is to determine whether the proposed sentiment analysis approach using MELex_v4 can improve sentiment classification accuracy. To this end, another sentiment orientation based on the general-purpose sentiment lexicon AFINN (Nielsen, 2011) and the two most well-known machine learning classifiers; NB and SVM as described in Section 6.9 is applied on the same testing data and comparisons of the results are presented.

There are two experiments have been conducted. The first experiment involves the classification of mixed language and the second one includes the classification of the entire PR1MA and PPAM test sets. Table 7.15 presents the performance's results for the first experiment in terms of overall accuracy, precision, recall and F-measure.

Table 7.15

Performance Comparison: Mixed Language

Dataset	Technique	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
PR1MA	MELex_v4	87.7	79.5	92.7	85.6
	AFINN	82.68	75	83.5	79
	SVM	76.4	77.2	76.3	76.01
	NB	77.88	78.01	77.57	78.3
PPAM	MELex_v4	85.71	84.61	84.61	84.61
	AFINN	75	73.3	78.6	75.9
	SVM	80.4	81.33	80.3	75.97
	NB	76.67	77.11	76.8	77.03

Table 7.15 summarizes the performance measures for mixed-language contents in both test sets, which shows MELex_v4 achieved the highest accuracy with 87.7% for PR1MA and 85.71% for PPAM. As for the SVM classifier, it performs quite well in the PPAM dataset, but not in PR1MA. It can be concluded that SVM might perform better in smaller datasets. In contrast, AFINN performed better in the large test set like PR1MA, but worst in PPAM except for recall prediction. Concerning the PPAM test set, like the PR1MA test set, classification using MELex_v4 is significantly better than other classifiers.

The following table presents the performance comparison for the overall classification, followed by Figure 7.5, which shows the comparison in terms of the accuracy for both test sets.

Table 7.16

Performance Comparison: Overall Classification

Dataset	Technique	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
PR1MA	MELex_v4	90.02	83.55	92.3	87.7
	AFINN	85.29	75.4	90.8	82
	SVM	77.88	80	80	80
	NB	77.68	78	78	78
PPAM	MELex_v4	89.19	85	93.2	88.9
	AFINN	81.31	74	88.8	80.7
	SVM	76.5	77	76	76
	NB	81.11	81	81	81

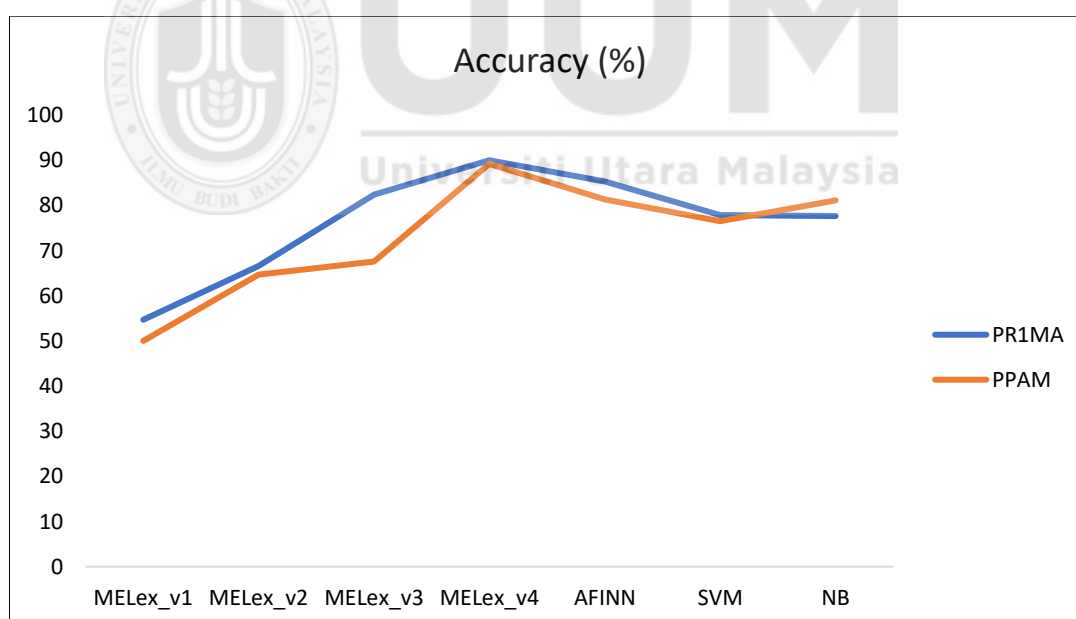


Figure 7.5. Performance comparison

As presented in Table 7.16 and Figure 7.5, the most accurate result was achieved by MElex_v4 for both datasets with an accuracy of 90% for PR1MA and 89.19% for PPAM. MElex_v4 gained a high score on recall which indicates low false negative. The classification using the NB technique has produced the worst accuracy in the PR1MA test set but performed better for the PPAM dataset.

For the PR1MA test set, both lexicons; MElex_v4 and AFINN outweigh machine learning techniques except on precision. The achieved accuracy as well as recall and F-measure using AFINN lexicon are reasonable and close to the accuracy obtained using MElex_v4. However, it posted significantly low scores in terms of precision either for mixed language or the overall classification for both datasets.

As compared to the PR1MA test set, the NB classifier performed better than AFINN in the PPAM test set but still lower than the accuracy achieved by MElex_v4. The precision result obtained by the NB classifier is close to the classification's result obtained by MElex_v4 where the difference is only 0.7% for the PPAM dataset. Furthermore, AFINN has reported the lowest precision result as compared to others.

In summary, the classification performed using MElex_v4 in this study achieved a reasonable classification accuracy for both test sets and outperformed the state-of-the-art baselines.

7.2.4 Misclassification

Although the results show a high accuracy for both datasets, it is worth understanding the inaccurate cases. Table 7.17 provides some examples of test data that MELex_v4 has incorrectly classified. The second column in Table 7.17 shows the results produced by sentiment classification using MELex_v4, while the third column represents the sentiment given by the data annotators.

Table 7.17

Examples of Misclassification Tweets

Tweets	Incorrect Polarity	Manually Labelled Polarity
anak boss aku nak beli rumah rm135,000. tu rumah pr1ma kott. rumah murah. luls. murah celah mana? <i>(my boss's son wants to buy a house for rm135,000. it's pr1ma home. Cheap? luls.)</i>	Positive	Negative
aik reasonable price. previous dapat offer pr1ma, 1k sqf pun from 244-356k <i>(reasonable price? Even pr1ma, 1k sqf is 244-356k)</i>	Positive	Negative
ni kmk tauk nya mesti maok status kahwin. mn bujang sik dapat beli rumah mn pr1ma dpt bujang tapi gaji mau 2.5k ke atas <i>(only married can apply, singles can't apply for pr1ma but the salary should be 2.5k and above)</i>	Neutral	Negative
rumah kotak yg mcmana pprrt pr1ma ppa1m semenyih punya rumah kotak pun dah 300k <i>(the small house of pprrt, pr1ma, ppa1m at semenyih were already rm300k)</i>	Neutral	Negative

A total of 273 tweets, as presented in Table 7.18 have been incorrectly classified for both test sets which means that MELex_v4 has misclassified around 12.24% out of the overall testing data. The highest number of misclassifications is for a single language in the PR1MA test set.

Table 7.18

Total No of Misclassification Data

Datasets		Total
PR1MA	Mixed language	36
	Single language	205
PPAM	Mixed language	6
	Single language	26
Total		273

The misclassifications by the MELex_v4 reflect its limitation of that approach. There are a few weaknesses in MELex_v4, which lead to misclassifications and errors.

Firstly, the possible reasons are due to the number of sentiments word covered in the lexicon is limited. As mentioned in Section 7.2.1, MELex_v1 and MELex_v2 contain 2,220 sentiment words while MELex_v3 and MELex_v4 have 6,132 terms in total. The number of sentiment terms is considered small as compared to the well-established English lexicons such as SentiWordNet (2 million words) and General Inquirer (7, 444 words). Due to the small number of words in the lexical resource, a lack of sentiment words can lead to false matching.

For example, in this sentence, ‘rumah pr1ma apa benda starting price rm300k’ (*why does it call PR1MA with the starting price of RM300k*), none of the word in that sentence could be matched or exist in the lexicon. Hence, this sentence is mistakenly tagged as neutral.

Another reason is probably due to sarcastic words or sentences used in the testing data. The sarcastic expression is referred to as a form of speech in which the writers write the opposite of what they mean. Within the context of sentiment analysis, sarcasm is hard to deal with. When one writes something positive, the system returns the sentiments as positive even though he/she actually means negative, and vice versa. Researches show that sarcastic expressions are not very common in the consumer reviews, but more frequent in the discussion or tweets related to the politics or any government’s agenda (Bakliwal et al., 2013). For instance, in the sentence ‘*pr1ma tu affordable lah sangat. murah? Murah celah mana?*’, the MELEX_v4 was unable to process correctly. The word *affordable* is detected as positive in the lexicon, which returns a positive score for this sentence. However, the exact meaning of this sentence is the affordability of PR1MA is still questionable which should be returned as a negative sentiment.

Another obstacle that may lead to misclassification is the presence of ambiguous words. A word is considered ambiguous when it carries multiple meanings even though it shares similar spelling. For example, the word *mampu* (*afford*) is tagged as positive in the MELEX lexicon. However, when it is used with the word *tengok*; ‘*mampu tengok*’ (*only can see*) it indicates the negativity of the sentence.

The only way to classify ambiguous words is to take the surrounding words into account in order to get a sense and contextual information in identifying the real sentiments of the word (Chanda, Das, & Mazumdar, 2016). The ambiguity issue has become even worst in the mixed-language content where one single word may belong to two or more languages. For example, the word ‘liar’ is belonging to both English and Malay language. In English, it means a person who tells lies, however in Malay, it means wild or uncontrolled. To disambiguate this data, it requires knowledge of the context for the whole sentence in order to obtain the real meaning of the word.

Another notable observation of misclassification case is due to misspelled words or in shortened forms, slangs or dialectal language that commonly used by Malaysians. MELex is not able to recognize those words. For example, ‘aku dh dpt umah prlma sebiji’ (*I have owned a PRIMA house*). The full sentences that MELex would be able to analyze are “aku dah dapat rumah prlma sebiji”. Slang words and dialectal language are hard to be detected in the textual analysis system. For example, Sarawakian will use ‘sik’ which means tidak (*no*) in a formal language. However, MELex_v4 lacks the word ‘sik’ and it leads to the incorrect classification of the data.

Finally, the misclassification is caused by the unigram feature employed in this study. The unigrams are considered as a simple approach where the classification is done by individual words. Besides, unigrams are simply a bag of words feature which is unable to incorporate any contextual information. The other two approaches are bi-grams and trigrams which take two or three consecutive words in the classification. Therefore, the performance may be improved if the bigrams or trigrams are used in the sentiment classification.

7.3 Discussions

The previous section offers empirical results of sentiment classification task for Malaysia property reviews. A detail discussion on the evaluation of its performance is included in this section, which is divided into three subsections; case studies, MELex performance and baseline comparisons.

7.3.1 Case Studies

Figure 7.6 and Figure 7.7 illustrates the final sentiment analysis results for PR1MA and PPAM datasets produced from the aforementioned experiments.

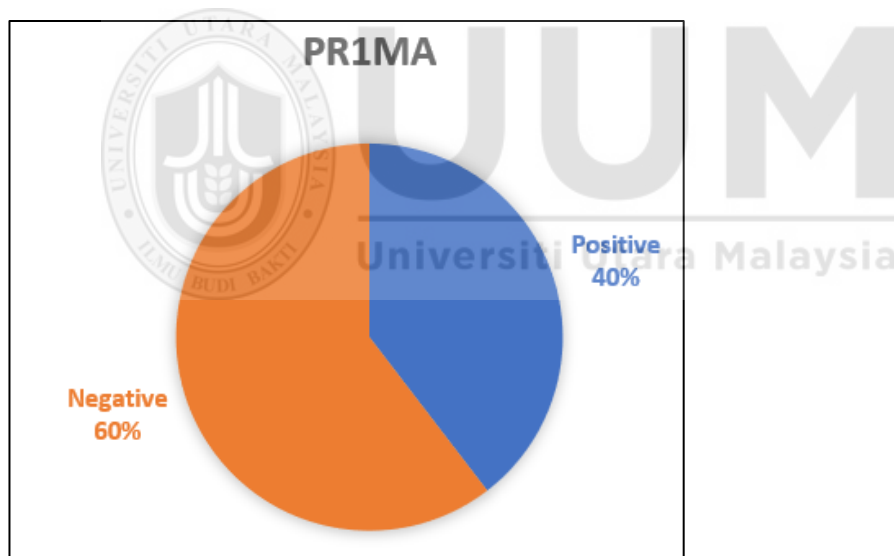


Figure 7.6. Sentiment analysis result for PR1MA

The results presented in Figure 7.6 shows that the negative sentiment towards the PR1MA project has a higher percentage which is 60%. It indicated that people view negatively about this project. As for PPAM projects, a more balanced percentage for both sentiments was obtained as the difference is only 4%.

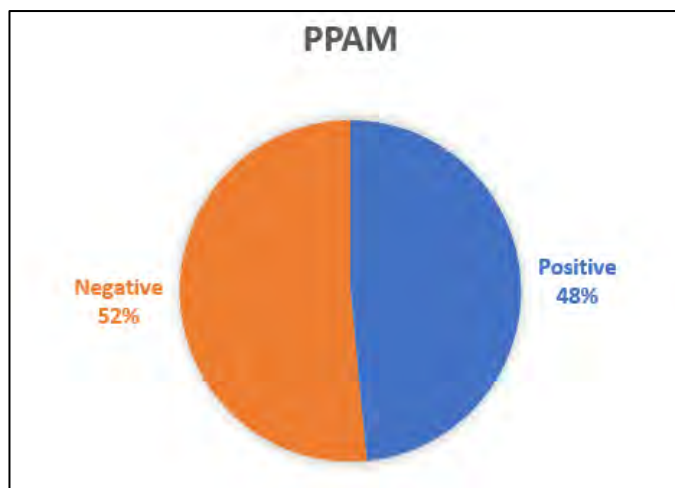


Figure 7.7. Sentiment analysis result for PPAM

Figure 7.6 and 7.7 shown above have answered Research Question 3. By comparing the percentage of positive and negative sentiments, the public's thoughts and satisfaction can be determined. By using this sentiment analysis results, the government may create strategies to improve the services.

7.3.2 MELex

The purpose of the experiments conducted in this research is to classify mixed language and single language content for PR1MA and PPAM test sets. The contribution of mixed-language classification towards the overall performance is observed.

As mentioned in Chapter 6, MELex was developed in three steps; seed words selection, polarity assignment and synonym expansion. The twitter data extracted based on a specific keyword have been used as training data and the words PoS tagged as verbs, adverbs or adjectives were assigned as seed words.

For the polarity assignment, two different techniques were employed; word vector representation and term frequency. The last step is to expand the lexicon by including synonyms extracted from English WordNet and WordNet Bahasa. As a result, four different experiments have been conducted in order to find the best combinations of techniques and steps in constructing sentiment lexicon.

Briefly, the main reason for the success of MELex is that it relies on the bilingual and domain-specific lexicon. In addition, the use of Twitter data which is specific for the Malaysian property domain as the seed words have a massive impact on boosting the results.

Like any other lexicon-based approach, the domain-specific sentiment words as well as the synonym expansion plays a vital role in boosting the results, as the detection of opinionated words in the test data is dependent on the quality and the number of words coverage of the lexical resource.

Besides, to examine whether the accuracy of mixed language classification will improve the overall performance, an experiment solely on mixed language sentences has been carried out. Out of 2,000 test data, 227 data are coming from a mixed language with approximately 11%. It was concluded that the classification of mixed language using MELex had contributed about 10% towards the accuracy of the overall classification. Besides, the incorrect classification of mixed language has only contributed 9.8% towards the misclassification results, which consider as a small portion.

The results reported in Table 7.5 and Table 7.11 supports the claim in this thesis that the proper classification of mixed-language data can have a tremendous improvement in the overall results. In terms of the overall performance, the accurate classification of mixed language has contributed 11.7% towards the total accuracy of the PR1MA test set.

One crucial observation is the performance of MELex in classifying mixed-language sentences which were reasonable and comparable with the other baseline methods. Also, this research is in line with several works reported previously where machine learning techniques are less effective in classifying low resource language (Hutto & Gilbert, 2014).

One of the advantages of the lexicon-based approach is that it works even in such situations where no available labeled data. In terms of lexical resources, only the sentiment lexicon is needed to calculate the overall sentiment of a property review. To generate the domain-specific sentiment lexicon, the target domain's training data has been used and the sentiment classification for the property review was performed using the generated lexicon.

Another advantage of obtaining a domain-specific sentiment lexicon is that it can be reused 'as-is' in other property projects as it shares similar domains since the domain-specific lexicons are similar and the calculation of the overall sentiment is straightforward.

However, the classification using MELex does have a few limitations. The main drawback of this method is it requires a dataset with manual labeling, which is time and effort consuming. The availability of a dataset specifically labeled for sentiment analysis tasks might be the primary challenge for many low resource languages such as Malay.

Another important observation is that the cause of misclassification is due to the absence of sentiment words in MELex. In summary, there are 5.8% of data from both test sets were classified as neutral.

According to Liu (2012), it is considered impossible for any sentiment analysis system to interpret the natural language text perfectly. The implemented sentiment analysis based on MELex has achieved excellent performance with an average accuracy of 87.25% which implies that the proposed approach is notably reliable.

7.3.3 Baseline Comparisons

The performance of MELex was compared with AFINN, which is one of the general-purpose sentiment lexicons widely used in other research and two well-known machine learning methods; NB and SVM classifiers.

Following the work presented in constructing the RojakLex lexicon, in order to compare with AFINN, the sentiment words were first translated into the Malay language. For the machine learning classifiers, the same training data was used to train the model.

As for the evaluation, similar testing data used in this research were employed to perform sentiment classification for these baselines. The accuracy results for the baseline comparisons for both datasets are presented in Table 7.19.

Table 7.19

Comparison of Results

Dataset	Technique	Accuracy (%)
PR1MA	MELex_v4	90.02
	AFINN	85.29
	SVM	77.88
	NB	77.68
PPAM	MELex_v4	89.19
	AFINN	81.31
	SVM	76.5
	NB	81.11

Based on the results presented in Table 7.19, it was found that the classification using MELex_v4 provides encouraging results and remarkably outperforms the baseline approaches either through lexicon-based or machine learning approaches on both datasets. Overall, the proposed sentiment analysis using MELex_v4 has provided an overall accuracy of 90.02% in the PR1MA reviews and 89.19% in the PPAM reviews.

Furthermore, it is noticeable that the accuracy obtained using MELex is even higher than the accuracy reported by the previous works on *Bahasa rojak* such as RojakLex (Chekima & Alfred, 2018) which reported 71.9% accuracy and SentiLexM (Tan et al.,

2016) with 78.5% accuracy. These results are however not directly comparable to this research work, as different datasets, with different levels of difficulty are applied.

Even though the classification using ML approach was not promising, additional experiments which may include various techniques for feature extraction and different percentage for data split are expected to be done in order to make it properly comparable.

7.4 Chapter Summary

In this chapter, the results obtained from the experiments conducted in Chapter 6 are reported and its performance is evaluated. In the course of evaluating the performance of sentiment classification using MELex, a total of 2230 test set from PR1MA and PPAM datasets have been used. There are four versions of MELex that have been produced and the classification using MELex_v4 has achieved the highest accuracy of 90.02% for PR1MA and 89.19% for PPAM. The experiments conducted show the proposed approach in producing MELex_v4 produced promising sentiment classification accuracy. Besides, as noted from the results reported in the experiment, there is a great effect of the synonym expansion on the overall performance.

MELex was then evaluated and compared against a general-purpose sentiment lexicon; AFINN and two well-known machine learning classifiers; NB and SVM using the same datasets. Based on the results, MELex has obtained better results, which indicates the performance of the proposed sentiment analysis approach is effective in this experiment.

Furthermore, the performance of MELex in analyzing mixed language content has been evaluated as well. Additionally, the misclassifications that lead to inaccurate results are also explained in this chapter.



CHAPTER EIGHT

CONCLUSIONS AND FUTURE WORK

8.1 Introduction

In this chapter, the achievements of each research objective are summarized, research contributions are highlighted, research limitations are introduced, and directions for future works related to this research are provided.

8.2 Objectives of the Study: Revisited

The main aim of this study is to propose the application of sentiment analysis in Malaysia's affordable housing projects using a new sentiment lexicon with the capability to classify mixed language content and produce better accuracy. This aim was achieved through four objectives which have been defined earlier in Section 1.5. The achievement of each objective is highlighted accordingly.

Objective 1: To identify the techniques used in developing a bilingual and domain-specific sentiment lexicon.

The research was started with the investigation of the best techniques that can be employed in the construction of MELex. The first objective was achieved through the systematic reviews conducted, as discussed thoroughly in Chapter 3. The current practices of sentiment lexicon creation and the techniques applied were studied. Additionally, the needs of bilingual sentiment lexicon have been revealed to cater to mixed-language content which frequently written by Malaysians.

Objective 2: To construct and develop a new sentiment lexicon for Malay and English languages specifically for property domain

This objective was fulfilled through the construction of MELex. Two main things have been considered in the construction process. First, the lexicon must be specific for the Malaysian property domain and second, it should be able to classify mixed Malay and English language content. Thus, the lexicon constructed has fulfilled these two needs, as discussed thoroughly in Chapter 6. As a result, four versions of MELex have been developed in this study.

Objective 3: To perform sentiment classification for affordable housing projects written in single and mixed language by using the constructed lexicon

The accomplishment of the third objective was reached through the implementation of the sentiment classification tasks, as explained in Section 6.8, and the results of the classification were reported in Section 7.2. In order to achieve this objective, the sentiment classification process was divided into a mixed and single language. Through this process, the percentage of the contribution that comes from mixed language classification can clearly be seen.

Objective 4: To evaluate the performance of the proposed approach

The fourth objective was achieved through two phases of assessment; evaluation metrics and the comparison with other general-purpose sentiment lexicon and two well-known machine learning classifiers as presented in Section 6.9.

This study has successfully improved the accuracy of sentiment classification for both datasets; PR1MA and PPAM. Moreover, the proposed approach in this research has the advantage of classifying mixed language content which contributed to more accurate results of the sentiment classification for the overall datasets.

8.3 Contributions

The findings from this research provide essential practical, methodological, empirical and dataset contributions for both academicians and practitioners. Specifically, this study may benefit scholars who seek to evolve research in sentiment analysis for low resource languages and to practitioners (particularly in the Malaysian property industry) when developing strategies in obtaining public insights and opinions towards their products or services. The following subsections describe each contribution in more detail.

8.3.1 Practical Contributions

Practically, the importance of the findings from this study would be of significant contribution to the Malaysian property industry. Up until now, most of the researches in the sentiment analysis has been carried out on products, real-events, politics, etc. In this research work, sentiment analysis for the property domain has been proposed.

Specifically, this research has analyzed the public sentiments towards the two most prominent affordable housing projects in Malaysia, which are PR1MA and PPAM. The outcome of this study can be utilized by practitioners such as government or real-estate players as a tool to gain insights from the public towards their property projects.

Also, this study fills the research gap between social media platforms and public sentiments. With an increasing number of citizens who prefer to express their thoughts and feelings online, social media is seen to be the best platform to obtain their opinions or sentiments. Hence, instead of using the traditional way of gaining public opinion, the present study provides a new method in gathering public sentiments using the free and available platform; Twitter.

8.3.2 Methodological Contribution

Methodologically, there are three contributions to this research study. A significant contribution is a novel approach in constructing a bilingual sentiment lexicon named MELex which incorporated below elements in developing the lexicon:

- i. Verb-adjective-adverb as sentiment terms: Any words in the training data tagged either as a verb, adjective or adverb were selected as a candidate for MELex.
- ii. Word vector and term frequency in determining the polarity score: Word vector was used to decide either the sentiment term is positive or negative, while term frequency was applied to determine the weight of those terms which is scaled between -3 to +3.
- iii. Malay and English language coverage: This research focuses explicitly on Malay and English language as both are the most communicated languages in Malaysia.

Although lexicon-based approaches have been widely applied in other sentiment analysis research for the Malaysian context, this research brings an essential contribution to the bilingual resource creation task that is proposed. Furthermore, the technique implemented in creating a Malay-English sentiment lexicon can be replicated to construct other bilingual lexical resources as well.

8.3.3 Empirical Contributions

Another novelty and contribution of this research are brought by the experimental evaluation conducted. This study shows that the proposed lexicon-based approach using the MELex lexical resource achieves high scores and more accurate compared to the state-of-the-art lexicon or machine learning classifiers when tested on the same datasets. The evaluation results and insights obtained using the proposed approach offer a vital source for future work in this direction.

This research contributes to the advancement of research in bilingual sentiment analysis, especially in the resource creation task as well as in the investigation of the impact of sentiment classification for mixed-language content. Moreover, this research provides a foundation for further progress on lexicon's construction and it defines a baseline that can be used as a benchmark in future work since MELex is available publicly.

This study proved that neglecting or inaccurate classification of mixed language data does affect the overall performance of sentiment analysis. Multiracial countries like Malaysia or Singapore, where the mixed language usage among the public is

considered typical, covering this type of content in the analysis is essential and may lead to better performance.

8.3.4 Dataset Contributions

The last contribution is the release of publicly available lexicon and datasets via the GitHub¹ repository, which is anticipated to facilitate the expansion of bilingual sentiment analysis research in the future.

- a. Property Dataset: 7,390 of manually annotated tweets of property reviews
- b. MELex_v4: the fourth version of a domain-specific Malay and English sentiment lexicon which contains 6,132 sentiment terms and its polarity score.

8.4 Limitations

Despite the valuable contributions of the research, there are several limitations inherent in the methodology and approach of the study that necessitates additional research and investigation.

First, this study is limited because it focused solely on affordable housing projects in extracting data and constructing the sentiment lexicon. Even though a Malay-English lexicon has been developed in this study, it should not be considered as a complete resource as the results presented in the previous chapter indicates that there are still many words that cannot be recognized by MELex. The total number of words

¹ <https://github.com/nhusna84/MELex>

generated from PR1MA and PPAM projects only is considered not sufficient. It is expected that more sentiment words could be gathered if more property projects, including private property projects are collected. In addition, MElex is a domain-specific lexicon and may or may not perform at the same level on a different domain.

Second, the data obtained was limited to the Twitter platform only. The nature of the Twitter platform forces users to write within 140 characters only. The other social media platforms without this limitation should be considered too in extracted public reviews such as Facebook and Instagram. In fact, there is an increasing number of SNS users in Malaysia who turn into Facebook and Instagram in expressing their opinion.

Third, the sentiment classification presented in this study has cover unigrams only. Based on the observation of the results, it shows that few misclassifications were due to word-by-word classification. The weakness of the word-level classification is the inability to capture the contextual information. Without contextual information that can be obtained through the surrounding words, certain sentences are incorrectly classified.

Fourth, the classification using MElex did not cover informal expressions that commonly used in SNS such as slang words, dialectal languages and misspelled words, and consequently suffers from limitations when analyzing opinions from micro-blogging text. Moreover, there are a variety of dialects used in Malaysia where the people often apply it in their daily life either verbally or in writing. The next section will highlight a number of suggestions for future work in order to improve the overall classification performance.

8.5 Suggestion for Future Research

While the sentiment analysis approaches using MELEX exhibit better performance, there is still room for improvement. The handling of more complex sentence calculations such as bigrams and trigrams using the generated sentiment lexicon will be the target of future study. The unigram's classification employed in this study has shown contextual information was unable to be captured. While bigrams and trigrams analysis have the ability to capture contextual information as it takes the surroundings word into consideration.

Another direction is to include more sentiment words in the lexical resource to enhance the capability of MELEX. To expand the lexicon, slangs, and dialects words commonly used by Malaysians will be considered as lexicon's candidate because this type of words might provide useful information in determining the sentiments. Furthermore, antonyms of sentiment words in MELEX which can be extracted from WordNet should be considered too to expand the lexicon.

To further increase the accuracy, spelling correction should be performed before sentiment classification activities. These misspelled words may be derived from the short-form word, error in typing, or can be influenced by the dialect of the language. Hence, the process of correcting spelling for both Malay and English languages should be done as one of the activities during pre-processing. It is notably necessary to handle misspelling because those misspelled words often resulted in incorrect classification.

This research has dealt with formal expressions only and cannot deal with sarcastic expressions. In user-generated content, it is common to see posts or comments written using a looser style than standard texts and often express sarcasm. For example, when the word “hebat” or “great” appears in a sentence, it can be positive, as in original meaning, or negative if meant ironically. Therefore, further research to detect and classify sarcasm expression is needed.

In this research, the performance of MELex is compared with the general-purpose sentiment lexicon and two commonly applied machine learning techniques. Even though MELex out-perform other compared techniques, it is believed that the results can further be improved. As future work, sentiment classification using a hybrid method by integrating MELex with any machine learning techniques may be beneficial in facilitating improvements of the results.

Apart from that, with respect to negation handling, the rules used can be improved in the future to heuristically detect patterns of negation in the sentence. Besides, the list of negation words should be expanded as well. In short, the improvements of negation handling may lead to better accuracy.

The proposed approach in constructing the lexicon may be improved by considering other word classes like nouns, which may carry sentiment as well. Besides, a better approach to minimize human involvement in labeling the data should be considered.

Another aim for future research is to visualize the sentiment analysis results in order to make it more understandable by the larger audience; the real-estate player in this case. The use of data visualization to present the output of sentiment analysis will be more significant for the decision-maker for enhancing their services or products.

Lastly, this study is a pioneering work in classifying property reviews. To get more valuable information on public insights, it is vital to perform aspect-based sentiment analysis as it looks into the features or aspects of the particular object, which users are commenting about. Aspect based sentiment analysis allows practitioners to gain a better understanding of the sentiments expressed by the public since it provides more detail analysis and the classification is performed by aspects/features such as price, location and design.



8.6 Conclusion

At the end of this study, a new bilingual sentiment lexicon for the property domain was developed. The experiments to prove the effectiveness of the sentiment classification using the newly constructed bilingual and domain-specific sentiment lexicon known as MELex has been conducted. The result obtained from the experiments is improved using the proposed approach. Besides, few experiments involving the general lexicon as well as machine learning approaches were conducted for comparison purposes.

This chapter has highlighted the contributions of the research, the limitations encounter at the end of this study and several potential future works were also pointed out. Although considerable and future works remain, this thesis demonstrates the applicability of sentiment analysis in analyzing public opinion by using the lexicon-based approach. From the findings obtained in this study, there were indications that the newly developed bilingual sentiment lexicon; MELex is significantly applicable to fellow academicians as well as practitioners to adopt into their sentiment analysis process.

REFERENCES

- Abdul-Mageed, M., & Diab, M. T. (2011). Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire. *Proceedings Of the Fifth Law Workshop*, 110–118. Retrieved from <http://www.aclweb.org/anthology/W11-0413>
- Active social media users as percentage of the total population in Malaysia from 2016 to 2019 (2019) *Statista*. Retrieved from <https://www.statista.com/statistics/883712/malaysia-social-media-penetration/>
- Agarwal, A., Sharma, V., Sikka, G., & Dhir, R. (2016). Opinion mining of news headlines using SentiWordNet. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 1-5. doi: 10.1109/CDAN.2016.7570949
- Agrawal, S., & Siddiqui, T. J. (2012). Feature based star rating of reviews: A knowledge-based approach for document sentiment classification. *International Journal of Hybrid Information Technology*, 5(4), 95-110. Retrieved from <https://pdfs.semanticscholar.org/ecf8/5604966f1095509f32d6d2f8cf75e1eaa151.pdf>
- Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 183-192. Retrieved from <https://www.aclweb.org/anthology/W13-3520>.

Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2017). Arabic language sentiment analysis on health services. *2017 1st International Workshop on Arabic Script Analysis and Recognition*, 114-118. doi: 10.1109/ASAR.2017.8067771

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. *Journal of Travel Research*, 58(2), 175-191. doi: 10.1177/0047287517747753

Alexander, N. S., & Omar, N. (2017). Generating a Malay Sentiment Lexicon Based on WordNet. *Asia-Pacific Journal of Information Technology and Multimedia*, 6(1), 126–140. Retrieved from <http://ejournal.ukm.my/apjitm/article/view/21974>

Alfred, R., Yee, W. W., Lim, Y., & Obit, J. H. (2016). Factors affecting sentiment prediction of Malay news headlines using machine learning approaches. *International Conference on Soft Computing in Data Science*, 289-299. doi: 10.1007/978-981-10-2777-2_26

Al-Moslmi, T., Omar, N., Albared, M., & Alshabi, A. (2017). Enhanced Malay sentiment analysis with an ensemble classification machine learning approach. *Journal of Engineering and Applied Sciences*, 12(20), 5226-5232. Retrieved from <http://docsdrive.com/pdfs/medwelljournals/jeasci/2017/5226-5232.pdf>

Al-Saffar, A., Awang, S., Tao, H., Omar, N., Al-Saiagh, W., & Al-bared, M. (2018).

Malay sentiment analysis based on combined classification approaches and

Senti-lexicon algorithm. *PloS one*, 13(4), 1-18. doi:

10.1371/journal.pone.0194852

Alsaffar, A., & Omar, N. (2014). Study on feature selection and machine learning

algorithms for Malay sentiment classification. *Proceedings of the 6th*

International Conference on Information Technology and Multimedia, 270–

275. doi: 10.1109/ICIMU.2014.7066643

Alsaffar, A., & Omar, N. (2015). Integrating a Lexicon based approach and K

nearest neighbour for Malay sentiment analysis. *Journal of Computer*

Science, 11(4), 639-644. doi: 10.3844/jcssp.2015.639.644

Alshalabi, H., Tiun, S., Omar, N., & Albared, M. (2013). Experiments on the use of

feature selection and machine learning methods in automatic Malay text

categorization. *Procedia Technology*, 11, 748-754. doi:

10.1016/j.protcy.2013.12.254

Anbananthen, K. S. M., Selvaraju, S., & Krishnan, J. K. (2017). The generation of

Malay lexicon. *American Journal of Applied Sciences*, 14(4), 503–510.

Retrieved from <http://thescipub.com/PDF/ajassp.2017.503.510.pdf>

Arif, S. M., & Mustapha, M. (2017). The effect of noise elimination and stemming in

sentiment analysis for Malay documents. *Proceedings of the International*

Conference on Computing, Mathematics and Statistics (iCMS 2015), 93-102.

doi: 10.1007/978-981-10-2772-7_10

Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R., & Kundi, F. M. (2016).

SentiHealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, 5(1), 1139-1152. doi: 10.1186/s40064-016-2809-x

Awrahman, B., & Alatas, B. (2017). Sentiment analysis and opinion mining within

social networks using konstan information miner. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(1), 15-22. Retrieved from

<https://journal.utem.edu.my/index.php/jtec/article/view/882>

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., & Hughes, M.

(2013). Sentiment analysis of political tweets: Towards an accurate classifier. *Proceedings of the Workshop on Language Analysis in Social Media*, 49-58.

Retrieved from <https://www.aclweb.org/anthology/W13-1106>

Balahur, A., & Perea-Ortega, J. M. (2015). Sentiment analysis system adaptation for

multilingual processing: The case of tweets. *Information Processing & Management*, 51(4), 547-556. doi: 10.1016/j.ipm.2014.10.004

Bari, A. A., & Shuaib, F. S. (2009). *Constitution of Malaysia: text and commentary*.

(3rd ed.). Retrieved from

https://www.academia.edu/33560884/3_Constitution_of_Malaysia_Text_and

_Commentary_3rd_Ed._Petaling_Jaya_Pearson_Prentice_Hall_2009_476_pa
ges_incl._index_ISBN_978-967-349-027-1

Begum, K. (2018, February 8). Consumer sentiment seen improving. *New Straits Times*. Retrieved from
<https://www.nst.com.my/property/2018/02/333370/consumer-sentiment-seen-improving>

Benson, E., Haghighi, A., & Barzilay, R. (2011). Event discovery in social media feeds. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1*, 389-398.
Retrieved from <https://www.aclweb.org/anthology/P11-1040>

Bernama. (2020, January 2). One million houses in 10 years is achievable – Zuraida. *New Straits Times*. Retrieved from
<https://www.nst.com.my/news/nation/2020/01/552936/one-million-houses-10-years-achievable-zuraida>

BNM report: Unsold houses at decade-high in 2017. (2018, February 15). *New Straits Times*. Retrieved from <https://www.pressreader.com/malaysia/new-straits-times/20180215/282218011257161>

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), 1-8. doi:
10.1080/15427560701381028

- Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2), 102-107. doi: 10.1109/MIS.2016.31
- Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI Conference on Artificial Intelligence*, 1515–1521. Retrieved from <http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8479>
- Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2), 15-21. doi: 10.1109/MIS.2013.30
- Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. *Proceedings of the 20th international conference on World wide web*, 675-684. doi: 10.1145/1963405.1963500
- Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. (2012). Discovering Consumer Insight from Twitter via Sentiment Analysis. *J. UCS*, 18(8), 973-992. doi: 10.3217/jucs-018-08-0973
- Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, 94, 53-64. doi: 10.1016/j.dss.2016.10.006
- Chan, A. P. J., & Lee, B. H. C. (2016). A Study on Factors Causing the Demand-Supply Gap of Affordable Housing. *INTI Journal Special Edition–Built*

Environment, 6-10. Retrieved from

<http://eprints.intimal.edu.my/600/1/EA%20-%201.pdf>

Chanda, A., Das, D., & Mazumdar, C. (2016). Unraveling the English-Bengali code-mixing phenomenon. *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, 80-89. oi: 10.18653/v1/W16-5810

Chekima, K., & Alfred, R. (2018). Non-english sentiment dictionary construction. *Advanced Science Letters*, 24(2), 1416-1420. doi: 10.1166/asl.2011.1261

Chekima, K., & Alfred, R. (2018). Sentiment Analysis of Malay Social Media Text. *International Conference on Computational Science and Technology*, 205-219. doi: 10.1007/978-981-10-8276-4

Chekima, K., Alfred, R., & Chin, K. O. (2017). Rule-Based Model for Malay Text Sentiment Analysis. In *International Conference on Computational Science and Technology* (pp. 172-185). Retrieved from <https://www.springerprofessional.de/en/rule-based-model-for-malay-text-sentiment-analysis/15488048>

Chuah, K. M. (2013). Aplikasi media sosial dalam pembelajaran Bahasa Inggeris: Persepsi pelajar universiti. *Issues in Language Studies*, 2(1), 56-63. Retrieved from <http://www.ils.unimas.my/vol2-no2/11-volume/29-working-in-groups-for-coursework-assignments-the-tertiary-students-perspective>

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P.

(2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12, 2493-2537. Retrieved from <https://dl.acm.org/doi/10.5555/1953048.2078186>

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4), 46-53. doi: 10.1109/MIS.2009.105.

Darwich, M., Noah, S. A. M., & Omar, N. (2015). Inducing a domain-independent sentiment lexicon in Malay. *JAIST Symposium on Advance Science and Technology*. Retrieved from <https://pdfs.semanticscholar.org/dbf5/78417667ae5d1fcf22ca7e01d24a2bbd8287.pdf>

Darwich, M., Noah, S. A. M., & Omar, N. (2016). Automatically Generating A Sentiment Lexicon For The Malay Language. *Asia-Pacific Journal of Information Technology and Multimedia Jurnal Teknologi Maklumat Dan Multimedia Asia-Pasifik*, 5(1), 49–59. Retrieved from <http://www.ftsm.ukm.my/apjitm>

Darwich, M., Noah, S. A. M., & Omar, N. (2017). Minimally-Supervised Sentiment Lexicon Induction Model: A Case Study of Malay Sentiment Analysis. In *International Workshop on Multi-disciplinary Trends in Artificial*

Intelligence (pp. 225-237). Retrieved from

https://link.springer.com/chapter/10.1007/978-3-319-69456-6_19

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4), 757-771. doi: 10.1007/s12559-016-9415-7

Dehkharghani, R., Saygin, Y., Yanikoglu, B., & Oflazer, K. (2016). SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation*, 50(3), 667-685. doi: 10.1007/s10579-015-9307-6

Demiroz, G., Yanikoglu, B., Tapucu, D., & Saygin, Y. (2012). Learning domain-specific polarity lexicons. *2012 IEEE 12th International Conference on Data Mining Workshops*, 674-679. doi: 10.1109/ICDMW.2012.120

Denecke, K. (2008, April). Using sentiwordnet for multilingual sentiment analysis. *Proceedings - International Conference on Data Engineering*, 507-512. doi: 10.1109/ICDEW.2008.4498370

Deng, Z. H., Luo, K. H., & Yu, H. L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, 41(7), 3506-3513. doi: 10.1016/j.eswa.2013.10.056

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., & Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation.

In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1370-1380).

Retrieved from <https://www.aclweb.org/anthology/P14-1129/>

Dong, X., Mavroeidis, D., Calabrese, F., & Frossard, P. (2015). Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5), 1374-1405. doi: 10.1007/s10618-015-0421-2

Dragut, E., Wang, H., Yu, C., Sistla, P., & Meng, W. (2012). Polarity consistency checking for sentiment dictionaries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 997-1005). Retrieved from <https://www.aclweb.org/anthology/P12-1105/>

Dutta, S., Saha, T., Banerjee, S., & Naskar, S. K. (2015). Text normalization in code-mixed social media text. In *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)* (pp. 378-382). Retrieved from <https://ieeexplore.ieee.org/document/7232908>

El-Beltagy, S. R., & Ali, A. (2013, March). Open issues in the sentiment analysis of Arabic social media: A case study. In *2013 9th International Conference on Innovations in Information Technology (IIT)* (pp. 215-220). Retrieved from <https://ieeexplore.ieee.org/abstract/document/6544421>

Eshak, M. I., Ahmad, R., & Sarlan, A. (2017). A preliminary study on hybrid sentiment model for customer purchase intention analysis in socialcommerce.

2017 IEEE Conference on Big Data and Analytics (ICBDA), 61-66).

doi: 10.1109/ICBDAA.2017.8284108

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (pp. 417–422). Retrieved from <https://www.aclweb.org/anthology/L06-1225/>

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2(1), 5. doi: 10.1186/s40537-015-0015-2

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82-89. doi: 10.1145/2436256.2436274

Feng, J., Gong, C., Li, X., & Lau, R. Y. (2018). Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews. *Wireless Communications and Mobile Computing*, 2018. doi: 10.1155/2018/9839432

Gatti, L., Guerini, M., & Turchi, M. (2015). SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4), 409-421. doi: 10.1109/TAFFC.2015.2476456

Georgiou, T., El Abbadi, A., & Yan, X. (2017, July). Privacy-Preserving Community-Aware Trending Topic Detection in Online Social Media. *IFIP Annual Conference on Data and Applications Security and Privacy*, 205-224. doi: 10.1007/978-3-319-61176-1_11

Ghosh, S., Ghosh, S., & Das, D. (2017). Complexity metric for code-mixed social media text. *Computación y Sistemas*, 21(4), 693-701. Retrieved from <http://arxiv.org/abs/1707.01183>

Gumperz, J. J. (1982). Introduction: Language and the communication of social identity. *Language and social identity*, 1-21.
doi:10.1017/CBO9780511620836.003

Guo, J., Che, W., Wang, H., & Liu, T. (2014). Revisiting embedding features for simple semi-supervised learning. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 110-120. doi: 10.3115/v1/D14-1012

Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd.

Hailong, Z., Wenyan, G., & Bo, J. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. *2014 11th Web Information System and Application Conference*, 262-265. doi: 10.1109/WISA.2014.55

Hardwick, J. (2019, June 25). Top 100 most visited websites by search traffic (as of 2019). Retrieved from <https://ahrefs.com/blog/most-visited-websites/>

Hijazi, M. H. A., Libin, L., Alfred, R., & Coenen, F. (2016). Bias aware lexicon-based Sentiment Analysis of Malay dialect on social media data: A study on the Sabah Language. *2016 2nd International Conference on Science in*

Information Technology (ICSITech), 356-361. doi:

10.1109/ICSITech.2016.7852662

Hong, L., Dan, O., & Davison, B. D. (2011). Predicting popular messages in twitter.

Proceedings of the 20th international conference companion on World wide web, 57-58. doi: 10.1145/1963192.1963222

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews.

Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177. doi:

10.1145/1014052.1014073

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media* (pp. 216-225). Retrieved from <https://pdfs.semanticscholar.org/a6e4/a2532510369b8f55c68f049ff11a892fefeb.pdf>

Ibrahim, M. N. M., & Yusoff, M. Z. M. (2015). Twitter sentiment classification using Naive Bayes based on trainer perception. *2015 IEEE Conference on e-Learning, e-Management and e-Services*, 187-189. doi:

10.1109/IC3e.2015.7403510

Isa, N., Puteh, M., & Kamarudin, R. M. H. R. (2013). Sentiment classification of Malay newspaper using immune network (SCIN). *Proceedings of the World Congress on Engineering*, 3, 3-5. Retrieved from

<https://pdfs.semanticscholar.org/9586/878ea483dc4e8c26ac6339847b300b67ca0b.pdf>

Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2019). Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication*, 29(3), 252-273. doi: 10.1080/01292986.2018.1453849

Jain, V. (2013). Prediction of movie success using sentiment analysis of tweets. *The International Journal of Soft Computing and Software Engineering*, 3(3), 308-313. doi: 10.7321/jscse.v3.n3.46

Jamaluddin, N. B., Abdullah, Y. A., & Hamdan, H. (2016). Encapsulating the delivery of affordable housing: An overview of Malaysian practice. In *MATEC Web of Conferences*, 66, 47-55. doi: 10.1051/mateconf/20166IBCC 2016 600047

Julian, G. (2019). What are the Most Spoken Language in the World? *Fluent in 3 Months*. Retrieved from <https://www.fluentin3months.com/most-spoken-languages/>

Kannan, R., Govindasamy, M. A., Soon, L. K., & Ramakrishnan, K. (2018). Social Media Analytics for Dengue Monitoring in Malaysia. *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 105-108. doi: 10.1109/ICCSCE.2018.8685028

Kasmuri, E., & Basiron, H. (2019). Building a Malay-English Code-Switching Subjectivity Corpus for Sentiment Analysis. *International Journal of Advances in Soft Computing & Its Applications*, 11(1), 113-130. Retrieved from <https://pdfs.semanticscholar.org/90ea/7dca64ecb7927ebf98d983575bf410e131e4.pdf>

Kasper, W., & Vela, M. (2011). Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference* (pp. 45-52). Retrieved from https://www.dfki.de/fileadmin/user_upload/import/5601_25.pdf

Kaur, W., & Balakrishnan, V. (2016, April). Bilingual Sentiment Detection- Investigating Impact of Tweet Translation. In *ICADIWT* (pp. 105-111). Retrieved from <http://eprints.um.edu.my/15865/1/0001.pdf>

Kaur, S., & Mohana, R. (2015). A roadmap of sentiment analysis and its research directions. *International Journal of Knowledge and Learning*, 10(3), 296-323. doi: 10.1504/IJKL.2015.073485

Khor, S. C. (2019). The Implementation of a Sustainable Management System for the Delivery of Affordable Housing. *Greening Affordable Housing: An Interactive Approach*, 32-51. doi: 10.1201/b22317-3

Lee, H. Y., & Renganathan, H. (2011). Chinese sentiment analysis using maximum entropy. *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*, 89-93. Retrieved from

<https://pdfs.semanticscholar.org/9a31/4bd5b6d9d6fa45a03a6bf8776a4d1d70f768.pdf>

Leh, O. L. H., Mansor, N. A., & Musthafa, S. N. A. M. (2017). The housing preference of young people in Malaysian urban areas: A case study Subang Jaya, Selangor. *Geografia-Malaysian Journal of Society and Space*, 12(7), 60-74. Retrieved from <http://ejournal.ukm.my/gmjss/article/view/17670>

Lim, X. Y., Olanrewaju, A., Tan, S. Y., & Lee, J. E. (2018). Factors determining the demand for affordable housing. *Journal of the Malaysian Institute of Planners*, 16(2), 109-118. Retrieved from <https://pdfs.semanticscholar.org/d787/66c98123ebd07f3bd891502e62c0550abb86.pdf>

Ling, C. S., Almeida, S., Shukri, M., & Sze, L. L. (2017). Imbalances in the property markets. *BNM Quarterly Bulletin, Quarter*, 3, 26-32. Retrieved from https://www.bnm.gov.my/files/publication/qb/2017/Q3/p3_ba2.pdf

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167. Retrieved from <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>

Liu, L., Lei, M., & Wang, H. (2013). Combining Domain-Specific Sentiment Lexicon with Hownet for Chinese Sentiment Analysis. *Journal of Computers*, 8(4), 878-883. Retrieved from

<https://pdfs.semanticscholar.org/5e6d/fc7ccee4b7c8e197dd2dbf5cde9946fe8318.pdf>

Liu, Q., Liu, B., Zhang, Y., Kim, D. S., & Gao, Z. (2016, March). Improving opinion aspect extraction using semantic similarity and aspect associations. *Thirtieth AAAI Conference on Artificial Intelligence*, 2986–2992, Retrieved from <https://dl.acm.org/doi/10.5555/3016100.3016320>

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4), 499-527. doi: 10.1007/s10462-016-9508-4

Lu, Y., Castellanos, M., Dayal, U., & Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: an optimization approach. *Proceedings of the 20th international conference on World wide web*, 347-356. doi: 10.1145/1963405.1963456

Maimun, N. A., Ismail, S., Junainah, M., Razali, M. N., Tarmidi, M. Z., & Idris, N. H. (2018, June). An integrated framework for affordable housing demand projection and site selection. *IOP Conference Series: Earth and Environmental Science*, 169(1). doi: 10.1088/1755-1315/169/1/012094

Marr, B. (2018, May 21). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. *Forbes*. Retrieved from <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do->

we-create-every-day-the-mind-blowing-stats-everyone-should-read/#477536da60ba

- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142-150. Retrieved from <https://www.aclweb.org/anthology/P11-1015.pdf>
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113. doi: 10.1016/j.asej.2014.04.011
- Menner, T., Höpken, W., Fuchs, M., & Lexhagen, M. (2016). Topic detection: identifying relevant topics in tourism reviews. *Information and Communication Technologies in Tourism 2016*, 411-423. doi: 10.1007/978-3-319-28231-2_30
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41. doi: 10.1145/219717.219748
- Miranda, C. H., & Guzman, J. (2017). A review of Sentiment Analysis in Spanish. *Tecciencia*, 12(22), 35-48. doi: 10.18180/tecciencia.2017.22.5
- Mohamed, H., Omar, N., & Aziz, M. J. (2011). Statistical malay part-of-speech (POS) tagger using Hidden Markov approach. In *Semantic Technology and*

Information Retrieval (STAIR), 2011 International Conference on (pp. 231-236). Retrieved from <https://ukm.pure.elsevier.com/en/publications/statistical-malay-part-of-speech-pos-tagger-using-hidden-markov-a>

Mohammad, S. (2016). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 174-179). Retrieved from <https://www.aclweb.org/anthology/W16-0429/>

Montoyo, A., MartíNez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4), 675-679. doi: 10.1016/j.dss.2012.05.022

Moreno-Ortiz, A., & Fernández-Cruz, J. (2015). Identifying polarity in financial texts for sentiment analysis: a corpus-based approach. *Procedia-Social and Behavioral Sciences*, 198, 330-338. doi: 10.1016/j.sbspro.2015.07.451

Mottain, M. (2017). Affordable housing – all about the supply. *The Star Online*. Retrieved from <https://www.thestar.com.my/business/business-news/2017/07/29/affordable-housing-all-about-the-supply>

Muhammad. A., M. (2016). *Contextual lexicon-based sentiment analysis for social media*. (Doctoral dissertation). Retrieved from

<https://pdfs.semanticscholar.org/b492/11633da7a9dfcad0c11bb8d415b3c2931c76.pdf>

Mukherjee, S., & Bhattacharyya, P. (2012). Feature specific sentiment analysis for product reviews. *International Conference on Intelligent Text Processing and Computational Linguistics*, 475-487. doi: 10.1007/978-3-642-28604-9_39

Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 412-418). Retrieved from <https://www.aclweb.org/anthology/W04-3253.pdf>

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., ... & Harwood, P. (2014). Social media in public opinion research: Executive summary of the aapor task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, 78(4), 788-794. doi: 10.1093/poq/nfu053

Mustafa, A., Adnan, N., Nawayai, M., & Salwana, S. (2017). The Influence of Product Quality and Service Quality on House Buyer's Satisfaction in Prima Home. *Pertanika Journal of Social Sciences & Humanities*, 25(4), 1841 - 1852. Retrieved from <https://pdfs.semanticscholar.org/3a59/a8db580edc7234ae210a5ed7bd6aa5244b90.pdf>

Naing, H. W., Thwe, P., Mon, A. C., & Naw, N. (2018). Analyzing Sentiment Level of Social Media Data Based on SVM and Naïve Bayes Algorithms.

International Conference on Big Data Analysis and Deep Learning Applications, 68-76. doi: 10.1007/978-981-13-0869-7_8

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 1-18). Retrieved from <https://www.aclweb.org/anthology/S16-1001>

Nasharuddin, N. A., Abdullah, M. T., Azman, A., & Kadir, R. A. (2017). English and Malay cross-lingual sentiment lexicon acquisition and analysis. *International Conference on Information Science and Applications*, 467-475. doi: 10.1007/978-981-10-4154-9_54

National Property Information Centre. (2019). *Residential: Overhang Status Q3 2019* [PowerPoint slides]. Retrieved from <http://napic.jp-ph.gov.my/portal>

Ng, S. (2019). Urgent need for a government-led big data system, say industry experts. *The Edge Markets*. Retrieved from <https://www.theedgemarkets.com/article/urgent-need-governmentled-big-data-system-say-industry-experts>

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *CEUR Workshop Proceedings* (pp. 93-98). Retrieved from <https://arxiv.org/pdf/1103.2903.pdf>

Noor, N. H. B. M., Sapuan, S., & Bond, F. (2011). Creating the open Wordnet bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. Retrieved from <https://www.aclweb.org/anthology/Y11-1027>

Number of internet users in Malaysia from 2017 to 2023 (2020) *Statista*. Retrieved from <https://www.statista.com/statistics/553752/number-of-internet-users-in-malaysia/>

Ohana, B., & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet. *9th IT & T Conference*. Doi: 10.13140/2.1.4547.0089

Ohana, B., & Tierney, B. (2011). Supervised learning methods for sentiment classification with RapidMiner. *International Journal for Research in Applied Science & Engineering Technology*, 5(XI), 80-89. Retrieved from <https://www.ijraset.com/fileserve.php?FID=10885>

Ortega, R., Fonseca, A., & Montoyo, A. (2013, June). SSA-UO: unsupervised Twitter sentiment analysis. *Second joint conference on lexical and computational semantic*, 2, 501-507. Retrieved from <https://www.aclweb.org/anthology/S13-2083>

Osman, M. M., Khalid, N., & Yusop, S. W. M. (2017). Housing affordability in the state of Selangor, Malaysia. *Advanced Science Letters*, 23(7), 6118-6122. doi: 10.1166/asl.2017.9218

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010, May).

From tweets to polls: Linking text sentiment to public opinion time series. In
Fourth international AAAI conference on weblogs and social media. Retrieved
from
<https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842>

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and
opinion mining. In *LREc*, 10(2010), 1320-1326. doi:
10.17148/IJARCCE.2016.51274

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations
and Trends in Information Retrieval*, 2(1-2), 1-135. doi:
10.1561/15000000011

Papadopoulos, S., Corney, D., & Aiello, L. M. (2014, April). SNOW 2014 Data
Challenge: Assessing the Performance of News Topic Detection Methods in
Social Media. In *SNOW-DC@ WWW* (pp. 1-8). doi: 10.1.1.662.2750

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... &
Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of
machine learning research*, 12, 2825-2830. Retrieved from
<http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

Peng, W., & Park, D. H. (2011). Generate adjective sentiment dictionary for social
media sentiment analysis using constrained nonnegative matrix factorization.

In *Fifth International AAAI Conference on Weblogs and Social Media*, 273-280. Retrieved from <https://pdfs.semanticscholar.org/843f/fbc6e29a0884cfa08a3d592452407f4990a4.pdf>

Peng, S., Tseng, V. S., Liang, C. W., & Shan, M. K. (2018). Emerging Product Topics Prediction in Social Media without Social Structure Information. *Companion Proceedings of the The Web Conference 2018*, 1661-1668. doi: 10.1145/3184558.3191625

Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish. *LREC*, 12(73), 3077-3081. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/1081_Paper.pdf

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. P. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 873-883). doi: 10.18653/v1/P17-1081

Puteh, M., Isa, N., Puteh, S., & Redzuan, N. A. (2013). Sentiment mining of Malay newspaper (SAMNews) using artificial immune system. In *Proceedings of the World Congress on Engineering* (pp. 1498-1503). Retrieved from http://www.iaeng.org/publication/WCE2013/WCE2013_pp1498-1503.pdf

- Quan, C., & Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, 272, 16-28. doi: 10.1016/j.ins.2014.02.063
- Rafee, H., & Wai, W. K. (2019, June 9). Big data to aid in making better property development decisions. *The Edge Markets*. Retrieved from: <https://www.theedgemarkets.com/article/big-data-aid-making-better-property-development-decisions>
- Raghavi, K. C., Chinnakotla, M. K., & Shrivastava, M. (2015, May). " Answer ka type kya he?" Learning to Classify Questions in Code-Mixed Language. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 853-858). doi: 10.1145/2740908.2743006
- Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. *2016 International Conference on Inventive Computation Technologies (ICICT)*, 1, 1-5. doi: 10.1109/INVENTIVE.2016.7823280
- Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14-46. doi: 10.1016/j.knosys.2015.06.015
- Refae, E. (2017). Sentiment analysis for micro-blogging platforms in Arabic. *International Conference on Social Computing and Social Media*, 275-294. doi: 10.1007/978-3-319-58562-8_22

Rosli, L. (2019, April 2). Big data to help solve property overhang issue: Zuraida.

New Straits Times. Retrieved from

<https://www.nst.com.my/business/2019/04/475335/big-data-help-solve-property-overhang-issue-zuraida>

Sadanandan, A. A., Osman, N. A., Hussain Saifuddin, M. K., Ahamad, D. N. P., &

Hoe, H. (2016) Improving Accuracy in Sentiment Analysis for Malay

Language. In *Proceeding of the 4th International Conference on Artificial*

Intelligence and Computer Science (pp. 28-29). Retrieved from

<https://pdfs.semanticscholar.org/44fd/01a0fc7062e53cf651a85ebc5fe2cc3f9b4c.pdf>

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users:

real-time event detection by social sensors. In *Proceedings of the 19th*

international conference on World wide web (pp. 851-860). doi:

10.1145/1772690.1772777

Salfarina, A. G., Nor Malina, M., & Azrina, H. (2010). Trends, problems and needs

of urban housing in Malaysia. *Malay*, 248, 62. doi: 10.5281/zenodo.1333957

Samsudin, N., Puteh, M., & Hamdan, A. R. (2011). Bess or xbest: Mining the

Malaysian online reviews. In *Data Mining and Optimization (DMO), 2011*

3rd Conference on (pp. 38-43). doi: 10.1109/DMO.2011.5976502

Samsudin, N., Puteh, M., Hamdan, A. R., & Nazri, M. Z. A. (2013). Mining opinion in online messages, *International Journal of Advanced Computer Science and Applications*, 4(8), 19-24. doi: 10.14569/IJACSA.2013.040804

Sapountzi, A., & Psannis, K. E. (2016). Social networking data analysis tools & challenges. *Future Generation Computer Systems*, 86(2018), 893-913. doi: 10.1016/j.future.2016.10.019

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3), 458-464. doi: 10.1016/j.dss.2012.03.001

Shahid, A. R., & Kazakov, D. (2009, January). Automatic Multilingual Lexicon Generation using Wikipedia as a Resource. *ICAART*, 357-360. doi: 10.5220/0001783003570360

Shalunts, G., Backfried, G., & Prinz, P. (2014, May). Sentiment analysis of German social media data for natural disasters. *ISCRAM*. Retrieved from <https://pdfs.semanticscholar.org/1728/24c8de84c98dda9ac65d4350bb207bdb4b9b.pdf>

Shamsudin, N. F., Basiron, H., Saaya, Z., Rahman, A. F. N. A., Zakaria, M. H., & Hassim, N. (2015). Sentiment classification of unstructured data using lexical based techniques. *Jurnal Teknologi*, 77(18), 113-120. doi: 10.11113/jt.v77.6497

- Shamsudin, N. F., Basiron, H., & Sa'aya, Z. (2016). Lexical Based Sentiment Analysis-Verb, Adverb & Negation. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(2), 161-166. Retrieved from <https://journal.utm.edu.my/index.php/jtec/article/view/976>
- Sharma, S., Srinivas, P. Y. K. L., & Balabantaray, R. C. (2015). Sentiment analysis of code-mix script. In *2015 International Conference on Computing and Network Communications (CoCoNet)* (pp. 530-534). doi: 10.1109/CoCoNet.2015.7411238
- Shein, K. P. P., & Nyunt, T. T. S. (2010). Sentiment classification based on Ontology and SVM Classifier. In *2010 Second International Conference on Communication Software and Networks* (pp. 169-172). doi: 10.1109/ICCSN.2010.35
- Shuhidan, S. M., Hamidi, S. R., Kazemian, S., Shuhidan, S. M., & Ismail, M. A. (2018). sentiment analysis for financial news headlines using machine learning algorithm. In *International Conference on Kansei Engineering & Emotion Research* (pp. 64-72). doi: 10.1007/978-981-10-8612-0_8
- Sitaram, D., Murthy, S., Ray, D., Sharma, D., & Dhar, K. (2015). Sentiment analysis of mixed language employing Hindi-English code switching. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 1, pp. 271-276). doi: 10.1109/ICMLC.2015.7340934

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263-286. doi: 10.1016/j.jbusres.2016.08.001

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. *Computer in Behavioral Science*, 7(4), doi: 10.1002/bs.3830070412

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 70-74). Retrieved from <https://www.aclweb.org/anthology/S07-1013>

Subrahmanian, V. S., & Reforgiato, D. (2008). AVA: Adjective-verb-adverb combinations for sentiment analysis. *IEEE Intelligent Systems*, 23(4), 43-50. doi: 10.1109/MIS.2008.57

Suraya, I. (2015). Making Housing Affordable: Khazanah Research Institute. Retrieved from [http://www.krinstitute.org/assets/contentMS/img/template/editor/_FINAL_Full_Draft__KRI__Making_Housing_Affordable__with_hyperlink__220815%20\(1\).pdf](http://www.krinstitute.org/assets/contentMS/img/template/editor/_FINAL_Full_Draft__KRI__Making_Housing_Affordable__with_hyperlink__220815%20(1).pdf)

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307. doi: 10.1162/COLI_a_00049

- Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009). Adapting naive bayes to domain adaptation for sentiment analysis. In *European Conference on Information Retrieval* (pp. 337-349). doi: 10.1007/978-3-642-00958-7_31
- Tan, Y. F., Lam, H. S., Azlan, A., & Soo, W. K. (2016). Sentiment analysis for telco popularity on twitter big data using a novel Malaysian dictionary. In *Frontiers in Artificial Intelligence and Applications*, 282, 112–125. doi: 10.3233/978-1-61499-637-8-112
- Taspinar, A., & Schuirmann, L. (2017). Twitterscraper 0.2. 7: Python Package Index. Retrieved from <https://pypi.org/project/twitterscraper/>
- Taylor-Sakyi, K. (2016). Big data: Understanding big data. *arXiv preprint arXiv:1601.04602*. Retrieved from <https://arxiv.org/ftp/arxiv/papers/1601/1601.04602.pdf>
- Teck-Hong, T. (2012). Housing satisfaction in medium-and high-cost housing: The case of Greater Kuala Lumpur, Malaysia. *Habitat International*, 36(1), 108-116. doi: 10.1016/j.habitatint.2011.06.003
- Thanvi, P. R., Sontakke, N. S., Waghmare, S. R., Patel, Z. S., & Gavhane, S. (2017). Sentiment Analysis for Political Reviews using AAVN Combinations. *International Research Journal of Engineering and Technology*, 5(1). Retrieved from <https://www.irjet.net/archives/V4/i4/IRJET-V4I417.pdf>

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173. doi: 10.1002/asi.21662

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384-394). doi: 10.1.1.301.5840

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424. doi: 10.3115/1073083.1073153

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment analysis in tripadvisor. *IEEE Intelligent Systems*, 32(4), 72-77. doi: 10.1109/MIS.2017.3121555

Vania, C., Moh. Ibrahim, & Adriani, M. (2014). Sentiment Lexicon Generation for an Under-Resourced Language. *Int. J. Comput. Linguistics Appl.*, 5(1), 59-72. Retrieved from <https://pdfs.semanticscholar.org/3331/6dd85eaaa6f0dbddefa823479435caaaed6.pdf>

Vyas, Y., Gella, S., Sharma, J., Bali, K., & Choudhury, M. (2014). Pos tagging of english-hindi code-mixed social media content. *Proceedings of the 2014*

Conference on Empirical Methods in Natural Language Processing (EMNLP), 974-979. doi: 10.3115/v1/D14-1105

Wong, J. (2018). Breaking the lull. *Focus Malaysia*, 263. Retrieved from <http://www.focusmalaysia.my/Property/breaking-the-lull>

Yan, G., He, W., Shen, J., & Tang, C. (2014). A bilingual approach for conducting Chinese and English social media sentiment analysis. *Computer Networks*, 75, 491-503. doi: 10.1016/j.comnet.2014.08.021

Young, E. (2017, November 21). Most unsold homes in Kedah are PR1MA units. *The Sun Daily*. Retrieved from <http://www.thesundaily.my/news/2017/11/21/most-unsold-homes-kedah-are-pr1ma-units>

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4), 919-926. doi: 10.1016/j.dss.2012.12.028

Yusoff, N., Jamaludin, Z., & Yusoff, M. H. (2016). Semantic-based Malay-English Translation using N-Gram Model. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(10), 117-123. Retrieved from <https://journal.utem.edu.my/index.php/jtec/article/view/1382>

Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Abidin, Z. Z. (2019). Developing Cross-lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-

based Approach. *International Journal of Advanced Computer Science and Applications*, 10(1), 346-351. doi: 10.14569/IJACSA.2019.0100146

Zainon, N., Mohd-Rahim, F. A., Sulaiman, S., Abd-Karim, S. B., & Hamzah, A. (2017). Factors affecting the demand of affordable housing among the middle-income groups in Klang Valley Malaysia. *Journal of Design and Built Environment*, 1-10. doi: 10.22452/jdbe.sp2017no1.1

Zakariah, Z. (2019). Property market to remain challenging in 2019: CBRE | WTW. *New Straits Times*. Retrieved from:
<https://www.nst.com.my/business/2019/01/451252/property-market-remain-challenging-2019-cbre-wtw>

Zamani, N. A. M., Abidin, S. Z., Omar, N. A. S. I. R. O. H., & Abiden, M. Z. Z. (2014). Sentiment analysis: determining people's emotions in Facebook. In *Proceedings of the 13th International Conference on Applied Computer and Applied Computational Science* (pp. 111-116). Retrieved from
<https://pdfs.semanticscholar.org/f072/a795575a78f0d884cdc85bec57ccac6d371f.pdf>

Zhang, Q., Chen, H., & Huang, X. (2014). Chinese-English mixed text normalization. In *Proceedings of the 7th ACM international conference on Web search and data mining* (pp. 433-442). doi: 10.1145/2556195.2556228

Zhou, X., & Chen, L. (2014). Event detection over twitter social media streams. *The VLDB Journal—The International Journal on Very Large Data Bases*, 23(3), 381-400. doi: 10.1007/s00778-013-0320-3

Zhou, F., Jiao, R. J., & Linsey, J. S. (2015). Latent customer needs elicitation by use case analogical reasoning from sentiment analysis of online product reviews. *Journal of Mechanical Design*, 137(7), 1-56. doi: 10.1115/1.4030159

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1393-1398. doi: <https://www.aclweb.org/anthology/D13-1141>



Appendix A

Data Extraction using Twitterscraper

```
#install package  
pip install twitterscraper
```

```
#extract 'PR1MA'  
twitterscraper PR1MA%20since%3A2015-01-01%20until%3A2017-12-31 -o  
pr1ma1.json  
twitterscraper #pr1ma%20since%3A2015-01-01%20until%3A2017-12-31 -o  
pr1ma2.json
```

```
#extract 'PPAM'  
twitterscraper PPA1M%20since%3A2015-01-01%20until%3A2017-12-31 -o  
ppam1.json  
twitterscraper #PPA1M%20since%3A2015-01-01%20until%3A2017-12-31 -o  
ppam2.json
```



Appendix B

Python Packages

Module/Functions	Description
twitterscraper()	Used to retrieve historical twitter data
polyglot()	Used to assign part-of-speech tagging
wordnet()	A lexical database for the English language which contain English words with its synonyms
regex()	It is a regular expression's module to perform pre-processing activity
langdetect()	Used to detect languages such as Malay and English
sklearn	It is a Python's library that provide various machine learning algorithms.
sklearn.model_selection	The available modules within sklearn's library
sklearn.naive_bayes	
sklearn.svm	
sklearn.predict	
sklearn.metrics	

Appendix C

Selected Source Codes of MELex Development

Lexicon Creation

```
class lexicon:
    startfrom = 0

    def __init__(self):
        self.wordx = []
        self.scorex = []
        self.occurn = []
        self.loadlexicon()
        self.ofile = open("lexicon.txt", 'w')

    def listtextfiles(self, foldername):
        owd = os.getcwd()
        fld = foldername + "/"
        os.chdir(fld)
        arr = []
        for file in glob.glob("*.txt"):
            arr.append(file)
        os.chdir(owd)
        return arr

    def loadlexicon(self):
        presentfiles = self.listtextfiles('lexicons')
        if 'lexicon.txt' not in presentfiles:
            print "Initializing!"
        else:
            tofrom1 = open('lexicons/status.txt', 'r')
            tofrom2 = tofrom1.read()
            tofrom3 = tofrom2.split()
            self.startfrom = int(tofrom3[1])
            tofrom1.close()
            infile = open("lexicons/lexicon.txt", 'r')
            inlines = infile.readlines()
            lenv = len(inlines)
            for j in range(lenv):
                stuff = inlines[j].split()
                self.wordx.append(stuff[0])
                self.scorex.append(float(stuff[1]))
                self.occurn.append(int(stuff[2]))
            infile.close()
```

```

def w_add(self,word,score):
    lenw = len(self.wordx)
    self.wordx.append(word)
    self.scorex.append(score)
    self.occurn.append(1)

def w_modify(self,word,score,indx):
    prevscr = self.scorex[indx]
    prevocc = self.occurn[indx]
    newocc = prevocc+1
    newscr = (prevscr*prevocc + score)/float(newocc)
    self.scorex[indx] = newscr
    self.occurn[indx] = newocc

def w_process(self,word,score):
    p,q,r = self.checkpresent(word)
    if p==1:
        self.addword(word,score)
    else:
        self.modifyword(word,score,r)

def write_lex(self):
    for j in range(len(self.wordx)):
        self.ofile.write(self.wordx[j]+" "+str(self.scorex[j])+" "+str(self.occurn[j])+"\n")

def getlexicon(self):
    return self.wordx,self.scorex,self.occurn

```



UUM
Universiti Utara Malaysia

Appendix D

Samples of MELex_v1 / MELex_v3

mahal	-0.32623
murah	0.142955
mampu	0.000985
naik	-0.08276
lulus	0.14707
dapat	0.101789
besar	0.090385
senang	0.007804
jauh	-0.03235
diluluskan	0.000486
gila	-0.15437
lancar	0.535714
tahu	0.006706
bebankan	-0.8125
fleksibel	0.044118
malas	-0.09259
confirm	-0.75
lingkup	-0.75



UUM
Universiti Utara Malaysia

Appendix E

Samples of MELex_v2 / MELex_v4

sentiment word	score
aneh	-1
bad	-1
badai	-1
badly	-1
bagus	2
bahagia	1
anggun	1
annoying	-1
approved	1
aset	1
batal	-1
clear	1
closed	-1
complaint	-1
complementary	1
completed	1
cuai	-1
curi	-1
curiga	-1
dahsyat	-1
daif	-1
dakwa	-1
dalih	-1



UUM
Universiti Utara Malaysia