ENGLISH-MALAY CROSS-LINGUAL EMOTION DETECTION IN TWEETS USING WORD EMBEDDING ALIGNMENT

LIM YING HAO

UNIVERSITI SAINS MALAYSIA

2023

ENGLISH-MALAY CROSS-LINGUAL EMOTION DETECTION IN TWEETS USING WORD EMBEDDING ALIGNMENT

by

LIM YING HAO

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

March 2023

ACKNOWLEDGEMENT

To Dr. Jasy Liew Suet Yan, who took me into the master's by research program under her supervision. I appreciate her guidance, her constructive feedback since day one, her fresh perspective that helped me view things in a different light, and most importantly, her constant and firm faith in me that I would eventually complete this program. This thesis would not be in this final shape without her countless hours spent on polishing my drafts. Also, I am incredibly grateful for her financial support from "Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/ICT02/USM/02/3", without which my study would be financially challenging.

To my parents, who have been offering their unwavering support and unconditional love. I am grateful for the bottomless snacks and bread, for the different choices of beverages that never run out, and for the forever warm home where I could study and sleep comfortably. To my siblings, whom I am always blessed to have. I appreciate their invaluable humor and the occasional contribution of snacks and drinks.

To the annotators, Siti Sakinah Ahmad Sanusi, Aqilah Syahirah Shahabudin and Nur Alya Mazlan, who agreed to help me validate the Malay tweets. I appreciate that they squeezed the annotation task into their already busy schedule. To the anonymous reviewers, who provided insightful comments on my manuscripts. I thank them for their time and critical reviews that helped refine my work directly. Finally, to my friend, Yu Syuen, who enrolled for a master's together with me, thanks for making my journey a less lonely one. The last sentence is dedicated to myself, who decided to study for a master’s program. ‘Thank you, your determination and efforts paid off!’

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **ii**

TABLE OF CONTENTS ..... **iii**

LIST OF TABLES ..... **vii**

LIST OF FIGURES ..... **ix**

LIST OF SYMBOLS ..... **xi**

LIST OF ABBREVIATIONS ..... **xii**

LIST OF APPENDICES ..... **xiii**

ABSTRAK ..... **xiv**

ABSTRACT ..... **xvi**

CHAPTER 1 INTRODUCTION **1**

1.1 Background 1

1.2 Problem Definition 4

1.3 Research Objectives 9

1.4 Research Questions 9

1.5 Research Scope 10

1.6 Thesis Overview 11

CHAPTER 2 LITERATURE REVIEW **13**

2.1 Introduction 13

2.2 Direct Translation 14

2.3 Annotation Projection 18

2.4 Joint Learning 23

2.5 Multilingual Language Model 33

2.6 Alignment 39

2.7 Augmentation 47

2.8 Research Gaps 53

CHAPTER 3 METHODOLOGY 57

3.1 Introduction 57

3.2 Phase 1: Data Pre-Processing and Preparation 58

3.2.1 Corpora 59

3.2.1(a) English Training Tweets 59

3.2.1(b) Malay Evaluation Tweets 62

3.2.2 English Monolingual Word Embeddings 67

3.2.3 Malay Monolingual Word Embeddings 67

3.2.3(a) Normalization 67

3.2.3(b) Spell-check 68

3.2.3(c) Noise Removal 69

3.2.4 Bilingual Lexicon 70

3.3 Phase 2: Embedding Alignment and Refinement 71

3.3.1 Monolingual Word Embeddings 72

3.3.2 Embedding Alignment 72

3.3.2(a) Experiment Extensions 73

3.3.3 Embeddings Refinement 75

3.4 Phase 3: Cross-Lingual Emotion Models 78

3.4.1 Hierarchical Attention Model 79

3.4.2 Experimental Setup 86

3.4.2(a) Experiment Set 1: Scaled Dot-Product Attention Mechanism Position Settings 87

3.4.2(b) Experiment Set 2: Effect of the Number of Heads in Scaled Dot-Product Attention Mechanisms 90

3.4.2(c) Experiment Set 3: Effect of the Number of Malay Tweets in Finetuning 91

3.4.2(d) Experiment Set 4: Effect of the Finetuning Layers 92

3.4.2(e) Experiment Set 5: Unweighted Versus Weighted Pre-Training 93

3.4.3 Hyperparameter Tuning 94
3.4.4 Comparison with Baselines 96
3.5 Evaluation Plan 98
3.6 Conclusion 100
CHAPTER 4 RESULTS AND DISCUSSION 101
4.1 Introduction 101
4.2 Bilingual Lexicon Induction 101
4.2.1 Malay Embeddings Coverage Comparison 101
4.2.2 Quality Comparison Between BL-Train and BL-Train-New 103
4.2.3 Augmentation of BL-Train for Bilingual Lexicon Extension 104
4.2.4 Augmentation of BL-Train-New for Bilingual Lexicon Extension 106
4.2.5 Malay Word Coverage Comparison Between the Bilingual Lexicons 107
4.2.6 Nearest Neighbors Analysis 109
4.2.7 Comparison with Other Existing English-Malay Bilingual Lexicons 111
4.3 Development of Cross-Lingual Emotion Detection Model 112
4.3.1 Position of the Scaled Dot-Product Attention Mechanism 112
4.3.2 Effect of the Number of Heads in Scaled Dot-Product Attention 114
4.3.3 Effect of the Number of Malay Tweets in Finetuning 116
4.3.4 Effect of the Finetuning Layers 118
4.3.4(a) Word Attention Visualization 122
4.3.5 Comparison with Baselines 123
4.3.6 Effect of Unweighted Versus Weighted Pre-Training 128
4.4 Summary 133
CHAPTER 5 CONCLUSION 135
5.1 Conclusion 135

5.2 Contributions 136
5.2.1 Language Resources 136
5.2.2 Methodological Contributions 137
5.3 Limitations 138
5.4 Future Work 139
REFERENCES 141
APPENDICES
LIST OF PUBLICATIONS

LIST OF TABLES

Page

Table 2.1 Number of emotion categories used in prior studies on low-resource languages not limited to Malay 55

Table 2.2 Cross-lingual sentiment analysis between English and Malay 55

Table 3.1 English tweet examples with the emotion labels 60

Table 3.2 Emotion distribution of English training tweets 60

Table 3.3 Number of tweets annotated per round 63

Table 3.4 Emotion labels of the Malay tweet examples before and after validation 64

Table 3.5 Inter-annotator agreement statistics 65

Table 3.6 Emotion distribution of EmoTweet-Malay-7 66

Table 3.7 Hyperparameters of the configuration 89

Table 3.8 Emotion distribution of English training tweets and Malay tweets (EmoMalayTweet-7) 93

Table 3.9 Search space of the hyperparameters 95

Table 4.1 P@10 comparison between three Malay embeddings using BL-Train 102

Table 4.2 Alignment precision comparison between BL-Train and BL-Train-New 103

Table 4.3 Alignment precision comparison using BL-Train as the seed lexicon and augmentation using different translation tools 104

Table 4.4 Alignment precision comparison using BL-Train-New as the seed lexicon and augmentation with different filters 106

Table 4.5 Selected English words and their Malay nearest neighbors 109

Table 4.6 Alignment precision comparison of our bilingual lexicon with existing English-Malay bilingual lexicons 111

Table 4.7 Cross-validation macro F1-score of the model for different positions of the scaled-dot product attention 113

Table 4.8 Optimal hyperparameters of our emotion model 115

Table 4.9 Comparison of the macro F1-score on 3426 Malay tweets..... 116

Table 4.10 Comparison of the macro F1-score on Ms-Test when finetuning different layers 118

Table 4.11 Emotion prediction for selected tweets when finetuning different layers 120

Table 4.12 Comparison of our model with the baselines on Ms-Test in zero-shot learning 123

Table 4.13 Comparison of our model with the baselines on Ms-Test in few-shot learning 124

Table 4.14 Our model performance without embedding alignment and refinement 126

Table 4.15 Macro F1-score breakdown of our model using unweighted pre-training 128

Table 4.16 Example of English and Malay tweets containing fear or happiness 130

Table 4.17 Comparison of the macro F1-score when different class weights are applied 131

Table 4.18 Macro F1-score breakdown of our model using ISRCF-weighted pre-training 132

LIST OF FIGURES

Page

Figure 2.1 Approaches in bridging the language gap in cross-lingual sentiment analysis 14

Figure 2.2 Direct translation method 15

Figure 2.3 Annotation projection method 19

Figure 2.4 Example of joint representation learning 24

Figure 2.5 Example of joint model learning 25

Figure 2.6 Example of multilingual language model in zero-shot learning 34

Figure 2.7 Example of word alignment between the English word 'delicious' and the Malay word 'sedap' 40

Figure 2.8 Example of the alignment of English word embeddings to the Malay vector space 40

Figure 2.9 Example of classical co-training 48

Figure 3.1 Overview of the three-phase methodological framework 58

Figure 3.2 Pre-processing steps for English tweets 61

Figure 3.3 Pre-processing steps for Malay tweets 67

Figure 3.4 Overview of Phase 2 72

Figure 3.5 Tweet representation 78

Figure 3.6 Workflow of Phase 3 79

Figure 3.7 Hierarchical attention model architecture 80

Figure 3.8 A long short-term memory unit adopted from Zhang et al. (2020) 81

Figure 3.9 Steps of the experiments 87

Figure 3.10 The position of the scaled dot-product attention in hierarchical attention model (i) word level (ii) sentence level (iii) word and sentence level 90

Figure 4.1 Comparison of changes of Malay words coverage for BL-Train and BL-Train-New 107

Figure 4.2 Changes in the cross-validation macro F1-score for different number of heads 114

Figure 4.3 Attention visualization of the Malay tweets with correctly predicted labels (a) anger and (b) happiness 122

Figure 4.4 Training time comparison across the models in zero-shot learning. 126

Figure 4.5 Training time comparison across the models in few-shot learning.. 127

LIST OF SYMBOLS

κ	Fleiss' kappa
α	Krippendorff's alpha

LIST OF ABBREVIATIONS

ANEW	Affective Norms for English Words
BERT	Bidirectional Encoder Representations from Transformers
Bi-GRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long-Short-Term-Memory
DBP	Dewan Bahasa Dan Pustaka
E-ANEW	Extended Affective Norms for English Words
ELECTRA	Efficiently Learning an Encoder That Classifies Token Replacements Accurately
KNN	K-Nearest Neighbours
LSTM	Long Short-Term Memory
MLM	Masked Language Modelling
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
MLP	Multilayer Perceptron
MUSE	Multilingual Unsupervised and Supervised Embeddings
NSP	Next Sentence Prediction
POS	Part-Of-Speech
SCL	Structural Correspondence Learning
SOTA	State-Of-The-Art
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TPU	Tensor Processing Unit
URL	Uniform Resource Locator
XLM-R	Cross-Lingual Language Model-Robustly Optimized BERT Pretraining Approach

LIST OF APPENDICES

Appendix A Summary of the cross-lingual sentiment analysis study

Appendix B Summary of language pair for cross-lingual sentiment analysis

PENGESANAN EMOSI MERENTAS BAHASA DALAM TWEET ANTARA BAHASA INGGERIS DAN BAHASA MELAYU DENGAN PENJAJARAN PEMBENAMAN PERKATAAN

ABSTRAK

Bahasa yang mempunyai sumber yang kaya seperti bahasa Inggeris mempunyai keistimewaan untuk menggunakan pengesanan emosi. Malangnya, bahasa Melayu tidak mempunyai sumber linguistik yang mencukupi untuk pengesanan emosi dan kajian yang menangani isu ini dengan kaedah merentas bahasa adalah terhad. Kami menggunakan penjajaran pembenaman perkataan untuk melakukan pengesanan emosi merentas bahasa dalam bahasa Melayu yang menggunakan bahasa Inggeris sebagai bahasa sumber. Metodologi tiga fasa untuk mencapai matlamat kajian ini merangkumi penggunaan penjajaran pembenaman perkataan untuk membina pembenaman perkataan rentas-bahasa antara bahasa Inggeris dan bahasa Melayu, pengayaan pembenaman perkataan tersebut dengan maklumat sentiment dan pra-latih model perhatian hierarki hanya dengan tweet dalam bahaasa Inggeris. Model kami dinilai dalam dua senario: pembelajaran syot sifar dan pembelajaran beberapa syot berdasarkan 4176 tweet dalam bahasa Melayu yang telah dianotasi dengan emosi. Kami juga mengkaji bilangan tweet bahasa Melayu optimum yang diperlukan untuk memperhalusi model dan kesan penghalusan lapisan yang berbeza dalam model kami. Hasil kajian kami menunjukkan bahawa penjajaran pembenaman perkataan dan pengayaan sentiment memanfaatkan pengesanan emotion merentas bahasa bagi bahasa Melayu dalam kedua-dua scenario. Kami mendapati bahawa model yang menghasilkan skor F1 yang terbaik yang sebanyak 91.15% dalam senario beberapa syot mempunyai kesemua lapisan telah dihaluskan. Dengan meminjam sumber

daripada bahasa Inggeris, model kami hanya memerlukan 1350 tweet bahasa Melayu untuk mencapai prestasi yang memuaskan. Perbandingan dengan model bahasa terkini mendedahkan bahawa model kami juga berdaya saing dari segi prestasi pengelasan dan masa latihan. Kajian ini menyumbangkan kaedah yang cekap untuk pengesanan emosi merentas bahasa dan menambah sumber linguistik dalam bahasa Melayu.

ENGLISH-MALAY CROSS-LINGUAL EMOTION DETECTION IN TWEETS USING WORD EMBEDDING ALIGNMENT

ABSTRACT

Languages with rich resources like English had the privilege to apply emotion detection. Unfortunately, the Malay language lacks linguistic resources for emotion detection, and limited studies address this issue using the cross-lingual approach. We address this problem by employing cross-lingual emotion detection through word embedding alignment using English as the source language and Malay as the target language. The three-phase methodology to address the goals of this study included the construction of the English-Malay cross-lingual word embedding using word embedding alignment, enrichment of the cross-lingual word embedding with sentiment information, and pre-training of the hierarchical attention model solely on English tweets. We evaluated our model in two scenarios: zero-shot learning and few-shot learning on 4176 Malay tweets annotated with emotion. We also examined the optimal number of Malay tweets required to finetune the model and the effect of finetuning different layers in our model. Our results show that embedding alignment together with sentiment enrichment benefits the cross-lingual emotion detection task for Malay in both scenarios. We found that the model with the best F1-score of 91.15% in the few-shot scenario had all its layers finetuned. By borrowing English resources, the model required only 1350 Malay tweets to achieve satisfactory performance. Comparison with the state-of-the-art language models reveals that our model is also competitive in terms of classification performance and training time. This study contributes an efficient method for cross-lingual emotion detection and expands Malay linguistic resources.

CHAPTER 1

INTRODUCTION

1.1 Background

Sentiment analysis is the task of classifying the sentiment polarity of a given opinionated text unit (Meng et al., 2012). On a coarse-grained level, the task would often be a binary classification problem where the sentiment polarity of the text would fall into either positive or negative (Pang & Lee, 2005). Alternatively, the neutral sentiment would be taken into consideration, making the task a three-class classification problem, as shown in Salameh et al. (2015). Beyond sentiment polarity, the text can be analyzed at a finer-grained level to detect emotions, which is also known as emotion detection. This could help narrow down the broad concepts of sentiment to better capture a person's emotional state (Ahmad et al., 2020). For instance, while anger and fear express negative sentiments, each semantically represents a different emotional state. Anger is perceived as the possible driving force of collective action, whereas fear is viewed as an action inhibitor (Miller et al., 2009).

The tremendous growth of Twitter has led to a specialized area in computational linguistics known as Twitter sentiment analysis (Zimbra et al., 2018). Cortis and Davis (2021) found that Twitter is the most frequently chosen social media platform for sentiment analysis in the social media domain. Twitter allows users to share their thoughts on any topic as a micro-blogging platform. With its broad user base, this provides a desirable platform for researchers to understand the representative sentiment on the subjects of interest, such as sentiment towards COVID-19 vaccines (Marcec & Likic, 2022), presidential elections (Hagemann & Abramova, 2023) and climate change (Lydiri et al., 2022).

However, the application of sentiment analysis, regardless of its granularity level and domain, had only been the privilege of languages with rich resources. The majority of the studies that concentrated on resource-rich languages produced extensively annotated corpora and computational tools exclusive to these languages. We refer to these languages as source languages. However, the rise of cross-lingual sentiment analysis is creating a greater possibility to perform sentiment analysis on resource-poor languages by leveraging the resources in source languages. With cross-lingual sentiment analysis, resource-poor languages are endowed with comparable computational ability in identifying sentiments. Resource-poor language and target language are used interchangeably throughout our study, and they are referred to as a language with no or a small number of annotated corpora and tools for sentiment analysis.

Although more than seven thousand languages are documented worldwide, only approximately 30 languages have been equipped with linguistically annotated resources (Eberhard et al., 2021; Maxwell & Hughes, 2006). Developing these resources for resource-poor languages from scratch is tedious and requires years of persistent effort. Cross-lingual sentiment analysis addresses the lack of language resource problem expeditiously by utilizing the resources from resource-rich languages that are equipped with gold-standard corpora and lexicons proven to be of high quality. It speeds up the development of computational models for sentiment analysis in resource-poor languages without compromising performance. Moreover, the cross-lingual sentiment models could serve as makeshift monolingual models while building the required linguistics resources for resource-poor languages.

In regions where these resource-poor languages are dominantly spoken, the speakers are generally multilingual, of which at least one language is resource-rich (Baumann & Pierrehumbert, 2014). Hence, it is commonplace for these speakers to mentally alternate between resource-rich and resource-poor languages and produce code-switching texts. One example of such interaction could be observed in Malaysia, where the Malay language often interacts with English and sometimes Chinese, resulting in sentences composed of different languages. Monolingual sentiment models would fall short in such a scenario as the models would only be exposed to one language. Therefore, cross-lingual sentiment models provide the perfect complement because it can also deal with multiple languages at once. Cross-lingual sentiment models can better handle code-switching texts particularly when sentiments are conveyed differently using different languages in text. In other words, cross-lingual models can capture sentiments from the perspective of different languages and are more applicable in a multilingual environment.

With cross-lingual sentiment analysis, the application of sentiment analysis previously limited to only resource-rich languages can be extended to resource-poor languages. For example, cross-lingual sentiment analysis can help businesses understand consumer emotional states from customer reviews written in different languages, which often include the customer's most comfortable language that could be resource-poor. As a result, businesses can devise more targeted strategies when dealing with customers who speak different languages. Additionally, cross-lingual sentiment analysis enables businesses to gauge the emotions they attempt to convey in their responses of different languages. For multinational businesses, they can utilize cross-lingual sentiment analysis to monitor the reviews of their regional products in

different languages. They can better identify the emotions behind the negative reviews of their products and resolve them accordingly.

In stock market analysis, investors can leverage cross-lingual sentiment analysis to broaden target foreign markets and diversify the risk. In addition to fundamental and technical analysis, investors can assess public emotions in different countries from stock forums, news, or financial disclosures in other languages. This allows them to make better-rounded decisions when choosing the focus of their investment. Similarly, investors would not be restricted to only one language when assessing public emotions in the local market; they can also take into account those written in other languages and then make a more thorough assessment of public emotions to better forecast the movement and price of the stocks of interest.

Despite its broad applicability, the first step in performing cross-lingual sentiment analysis is to identify the target language. The target language could be chosen from the pool of languages spoken in the targeted community that are still resource-poor. In our study, the target language is Malaysian Malay, also known simply as Malay in Malaysia, and they are used interchangeably in our study. It is the most widely spoken language in Malaysia as every Malaysian begins learning Malay officially in primary school. Our study specifically targets content in Malay tweets that allows us to 'peek' into the population’s norms and emotions from the text.

1.2 Problem Definition

As most of the studies in sentiment analysis focused on resource-rich languages, the distribution of resources such as gold-standard annotated corpora and word embeddings has been substantially imbalanced across languages. Some of the most spoken languages today are still considered resource-poor languages that are

unsuitable for building computational monolingual sentiment models (Farra, 2019). Although not the most spoken language globally, Malay is a dominantly spoken language in Malaysia. Comparably, Malay still lacks linguistic resources for sentiment analysis, particularly at a finer-grained level like emotion detection. This poses a challenge in automatically identifying emotions expressed in Malay text on a large scale, especially on social media platforms where almost everyone takes to share their personal and affective experiences. Emotion detection in Malay would be handy during natural disasters and pandemics or political instability and turmoil in Malaysia as it allows us to assess the emotional states of the affected individuals or evaluate public emotions about the political situation.

Many of the studies adopted machine translation systems to bridge the languages. It is undeniably the most intuitive and efficient approach, although not consistently the most effective. Machine translation systems can accurately translate simple and short text but are inadequate for complex text. In most translations, emotions are not carried over to the translated text (Salameh et al., 2015). Using machine translation also presumes emotions are conveyed similarly in the resulting parallel corpus when the emotion expression, in fact, varies across different languages. The difference in emotion concepts is significant when the source and target languages are from two distant language families (Jackson et al., 2019). Hence, emotions are expected to be perceived differently between English and Malay.

Additionally, current state-of-the-art (SOTA) methods rely on multilingual language models. Although these models pre-trained on massive amounts of corpora are computationally expensive to be employed, the multiplicative attention mechanisms (scaled dot-product attention) used in the models were more

computationally efficient than the additive attention (Britz et al., 2017). These language models also run multiple attention mechanisms in parallel that can capture information from different perspectives. However, no recommendation has been provided on how a hierarchical attention model can benefit from swapping its original additive attention with multiplicative attention. Michel et al. (2019) discovered that most heads were redundant and one head was sufficient enough most of the time based on experiments using non-hierarchical models and on machine translation tasks different from our study.

Previous Malay sentiment analysis methods generally adopted either bag-of-words or term frequency-inverse document frequency to represent words in the corpus (Alfred et al., 2016; Saad et al., 2018; Tiun, 2017; Yin et al., 2021). These methods simplify word representations but suffer from the curse of dimensionality. This problem worsens when the emotion models are trained on parallel corpora in cross-lingual settings. Using these methods in cross-lingual settings also typically involves machine translations that are likely to introduce noise in the translated corpora (Balahur & Turchi, 2014). Unlike word embeddings, these methods representing words using frequency-related vectors also disregard the semantic relation between words. In other words, the sentiment classification task would solely be based on the frequency variation of words in the corpus. Such models would also fail when the training corpus and test corpus have dissimilar word distributions (Hajmohammadi et al., 2014b).

Although word embeddings can capture semantic information, monolingually pre-trained word embeddings are constrained to tasks solely in their own languages as the embedding space is not shared across different languages. To allow cross-lingual

learning, the word embeddings of the source language and target language need to be aligned to the same vector space. Aligning the word embeddings with bilingual supervision requires a seed bilingual lexicon. However, the existing English-Malay bilingual lexicon by Husein (2018) has not been validated. It contains word pairs that are not paired up precisely (e.g., an English word paired up with an English word) and words that could not be encoded, resulting in garbled words. Consequently, the aligned word embeddings would not be desirable. English words could have semantically irrelevant Malay nearest neighbors in the cross-lingual embedding space and vice versa, thus resulting in an inaccurate transfer of the semantic information in the emotion model.

The existing publicly available Malay emotion tweet corpus had been annotated using a rule-based classifier (Husein, 2018). The emotion lexicon used to define the rules to identify emotions is not standardized and exhaustive. Words used to express emotions on Twitter vary over time, and new words are coined from time to time. This makes constructing an exhaustive list of emotion words nearly impossible. In other words, emotion words not on the list would be neglected. Furthermore, it is possible that tweets would be assigned emotions incorrectly for contextually complicated tweets. Tweets with negation and contrast transition would typically exude reversed and more salient emotions. Negations pose another challenge when they are not located next to the emotion words and are thus overlooked by the rule-based classifier. The rule-based classifier relying on keyword matching to assign emotions thus may not capture the overall context in the tweets and likely fail in such a scenario.

The current Malay corpus also contains a combination of Malay and Indonesian tweets. Malay and Indonesian are linguistic variations from the same Malay family, sharing high lexical similarities (Ranaivo-Malancon, 2006) and plenty of interlingual homographs (Lin et al., 2018), but emotions are expressed somewhat differently between the two languages. For instance, *galau* in Malay means 'commotion' or 'confusion', but it means 'uncertain' or 'indecisive' in Indonesian (Chin et al., 2021). It is also more common to use *marah* instead of *berang* to represent 'angry' in Malaysia. Also, Indonesian has more loanwords from Javanese instead of English, producing words not used in Malay (Chin et al., 2021). Hence, most of the colloquial Indonesian (emotion) words would not fit into our Malay language.

The current availability of annotated Malay emotion corpus leads to two scenarios for emotion detection in Malay: zero-shot and few-shot. In the zero-shot scenario, emotion detection in Malay relies entirely on existing annotated corpora in other languages. This scenario thus presents the challenge of effectively exploiting existing resources. On the other hand, the few-shot scenario utilizes a limited number of annotated Malay emotion instances on top of the existing corpora from other languages to perform emotion detection. While Lauscher et al. (2020) showed that few-shot learning is always better than zero-shot learning, it comes with a trade-off between annotation cost and performance gain. Although it is possible to refer to prior studies in estimating the annotation costs, the cost-to-gain conversion rate should be expected to be substantially different across languages and tasks.

To address these challenges, words are represented using embedding vectors. We adopt hybrid language detection to ensure that our Malay emotion corpus has minimal Indonesian tweets. We also manually validate the Malay corpus to serve as

gold-standard test data for our cross-lingual emotion model. As our objective is to develop an emotion model to predict emotions in Malay tweets utilizing English resources, both languages are projected to a shared embedding vector space. This allows our model to learn from the English resources to classify emotions in Malay tweets in both zero-shot and few-shot scenarios.

1.3 Research Objectives

The source language is the resource-rich language (i.e., English) in the study, while the target language is Malay, the resource-poor language. This study aims to enable the Malay language to have comparable advantages to English in applying emotion detection to real-world problems. Specifically, this study addresses the following objectives:

i. To improve the precision of mapping Malay words from the Malay embeddings to the corresponding English words from the English embeddings.

ii. To examine the position and number of heads of the scaled dot-product attention in the hierarchical attention model for cross-lingual emotion detection.

iii. To assess the performance of the English-Malay cross-lingual emotion model in classifying emotion from Malay text.

1.4 Research Questions

As we aim to develop an emotion detection model to predict emotions in Malay tweets by exploiting English resources, the central problem lies in language differences. To resolve this issue, we align the two languages based on semantics and subsequently enrich the embeddings with sentiment information. However, the direct alignment of

English and Malay languages is hampered by the resources in both languages being noisy as tweets are written informally. It is thus essential to first identify the appropriate pre-processing steps to formalize the resources in both languages with maximum consistency. The study aims to answer the following research questions:

- R1: How can English embeddings and Malay embeddings be aligned based on semantics using bilingual lexicons and enriched with sentiment information?
- R2: How do the position and number of heads of the scaled dot-product attention in hierarchical attention models affect the classification performance?
- R3: What is the performance of the cross-lingual emotion model in zero-shot and few-shot learning?

1.5 Research Scope

As we train our emotion model on tweets, we do not expect the model to be generalizable to predict emotions in Malaysian Malay text of other domains. The embeddings pre-trained on tweets (a portion of the Malay training corpus also consists of Instagram posts) captured the particular usage of the words in tweets and Instagram posts. Thus, the embedding vectors predominantly reflect words used in the social media context. As the usage of words may differ across different domains, we only evaluate our models on Malay tweets.

Our study covers only two languages: English as the source language and Malaysian Malay as the target language. English is chosen as the source language because of its abundance of good quality and easily attainable linguistic resources such as annotated corpora and pre-trained embeddings. On the other hand, Malay is chosen

as the target language because it is the national language of Malaysia (thus the most spoken language), which is still a low-resource language.

We also acknowledge that each tweet may express multiple emotions. It is, in fact, natural for the users to have mixed feelings, and while emotions do not cancel each other out, these emotions would be manifested when the tweet is written. However, we restrict each tweet to be only labeled with a single emotion from the pool of 6 emotions (anger, fear, happiness, love, sadness and surprise) and 'none' to reduce the complexity of our model. Also, we do not have multi-label Malay tweets to evaluate the models.

1.6 Thesis Overview

The remainder of this thesis is organized as follows: Chapter 2 surveys related work on cross-lingual sentiment analysis and presents a comprehensive analysis of the techniques used to bridge the gap between languages. We characterize the differentiating properties of each method category and identify several research gaps in cross-lingual sentiment analysis.

Chapter 3 presents our three-phase methodology. Phase 1 prepares and cleans corpora, embeddings and bilingual lexicon needed for subsequent phases. Phase 2 investigates the hypotheses in producing a quality bilingual lexicon to align the monolingual embeddings based on semantics and enriches the aligned embedding with sentiment information. Using the corpora from Phase 1, this embedding is used for the emotion classification experiments in Phase 3.

Chapter 4 begins with the intrinsic evaluation of the embedding alignment. We compare the alignment precision using different augmented bilingual lexicons and

decide on the aligned embedding for sentiment enrichment. The sentiment-enriched word embedding is used to develop the emotion detection model. We then feed the embedding to the emotion detection model for zero-shot and few-shot learning. We also compare the classification performance with the current state-of-the-art (SOTA) models. We demonstrate that our model is at a competitive standing when compared to the SOTA models.

Lastly, Chapter 5 presents the conclusion and contributions of this study. We also highlight the limitations of this study and recommend several directions for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Text-based emotion detection is a task that aims to recognize emotions beyond the coarser-grained sentiments in text. Previous studies performing this task using a categorical emotion model generally followed a particular emotion taxonomy, such as Ekman's six emotions (Ekman, 1992) or Plutchik's wheel of emotions (Plutchik, 1994). However, recent studies like Liew et al. (2016) and Demszky et al. (2020) have expanded the emotion taxonomy to cover a broader range of emotions that allows for capturing richer insights from text.

Emotion detection can be framed as a classification task (Abdul-Mageed & Ungar, 2017; Lyu et al., 2020; Mohammad, 2012; Saad et al., 2018) or a regression task (Kleinberg et al., 2020; Mohammad & Bravo-Marquez, 2017; Strapparava & Mihalcea, 2007). The classification task is more common and is designed to detect the presence of emotions in the text. In contrast, the regression task focuses on identifying the intensities of emotions in the text. Irrespective of the task, emotion detection requires gold-standard resources, which are still limited in most languages today. However, these languages can still perform emotion detection with the aid of resources from another language. Performing emotion detection by 'borrowing resources' from another language is known as cross-lingual emotion detection, or more generally, cross-lingual sentiment analysis (Xu et al., 2022).

This chapter reviews relevant literature in cross-lingual sentiment analysis for both coarser-grained and finer-grained levels. The focus of our survey is on the techniques used to bridge resource-rich languages and low-resource languages. While

there have been a large number of studies in cross-lingual sentiment analysis over the years, to the best of our knowledge, there is no systematic review of the approaches used to bridge languages in the task. For this reason, we develop a typology of methods according to their properties. Next, we categorize and organize methods used in relevant work into one of the methods. Specifically, as shown in Figure 2.1, we have identified these methods as direct translation, annotation projection, joint learning, multilingual language model, alignment, and augmentation. We also provide each method with descriptions that help set it apart from the others. Finally, the last section identifies research gaps, specifically in the areas that led to this study.

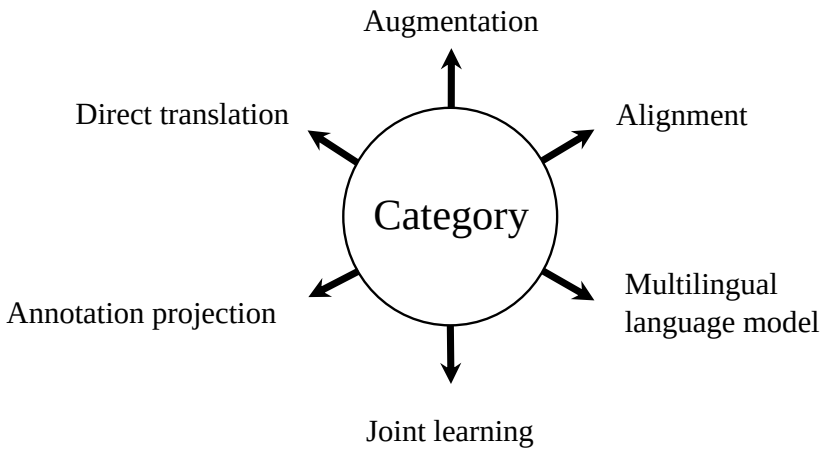

Figure 2.1 Approaches in bridging the language gap in cross-lingual sentiment analysis

2.2 Direct Translation

Direct translation is a straightforward approach that does not involve any sophisticated frameworks. It exploits existing tools such as bilingual lexicons (Ghorbel, 2012; Ghorbel & Jacot, 2011; Rasooli et al., 2018; Yao et al., 2006) or machine translation systems (Das et al., 2012; Esuli et al., 2020; Salameh et al., 2015; Wan, 2012; Wei & Pal, 2010) to translate text from one language to another language. Despite the

laborious process, Das et al. (2012) and Salameh et al. (2015) also adopted human translation in their sentiment analysis study. The translation is usually unidirectional and is of either of these two directions: the first and also the most common direction is to translate text from the source language to the target language; the second direction reverses the first direction by translating text from the target language to the source language. It would then directly employ the translation returned by machine translation tools without alterations for sentiment analysis. Direct translation is efficient but also requires the machine translation tools to be well-developed. In other words, the outcome of the sentiment analysis would be highly dependent upon the quality of the tools (D. Zhou et al., 2012). An example of direct translation is shown in Figure 2.2.

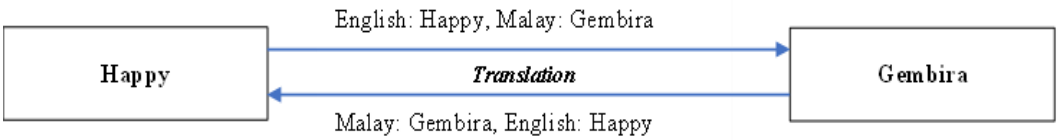

Figure 2.2 Direct translation method

Polarity-annotated words, being the smallest units that carry sentiments, are often limited in the target language (S.-M. Kim & Hovy, 2006). However, these annotated words are essential for rule-based classifiers to perform sentence or document-level sentiment analysis satisfactorily. As one simple method to identify sentiment is through the sentiment polarity of words, Yao et al. (2006) utilized bilingual lexicons to determine the sentiment polarity of Chinese words. To reduce semantic incompleteness, a total of 10 different Chinese-English bilingual lexicons were used in their study. English interpretations of the Chinese words were first extracted from these bilingual lexicons to obtain corresponding English word sequences. This word sequence was used to form the sentiment vectors for each Chinese word by counting the number of positive and negative English interpretations.

Subsequently, a word sentiment classifier was then trained on the sentiment vector to classify the sentiment polarity of Chinese words.

Instead of building polarity lexicons, Ghorbel and Jacot (2011) employed French-to-English word translation using a bilingual lexicon to extract sentiment scores from SentiWordNet directly. These scores were then used to compute the overall sentiment scores for selected part-of-speech (POS) tags as additional training features. In a subsequent study, Ghorbel (2012) improved the translation by selecting French words likely to carry sentiment and translating them to English using a sense-aligned bilingual lexicon (EuroWordNet) to extract the sentiment scores. Similarly, these scores were then used to compute the overall sentiment scores for selected POS tags as additional training features. However, the integration of word polarity in both studies barely improved the classification, and they attributed the problem to the translation error and low coverage of the words in the lexicon. Rasooli et al. (2018) also pointed out that the coverage of the lexicon, especially for languages with morphologically complex words, could affect the performance of direct translation.

In contrast, Rasooli et al. (2018) reported positive results from the bilingual lexicon. The authors adopted a partial translation strategy as one of the techniques to improve the generalization of features beyond source languages. This strategy translated source words in the training corpora that were in the lexicon into target words. Since the coverage of the lexicons was low, the resulting training corpora resembled code-switched text. On average, this approach was shown to improve the performance of sentiment classification in target languages. Nevertheless, Rasooli et al. (2018) stressed that a high-quality bilingual lexicon is indispensable in this approach.

Undoubtedly, a bilingual lexicon is still an expensive resource. The competitively economical alternative for direct translation is to utilize machine translation, such as Google Translate, which is considered a state-of-the-art machine translation system. Wan (2012) demonstrated how machine translation could be adopted readily when only source language resources are available. To fully leverage English resources, Wan (2012) translated Chinese reviews into English using Google Translate and classified them using a rule-based classifier or support vector machine. The resulting outcomes were shown to be promising.

Wan (2012) assumed that the sentiment meaning between different languages was preserved in such translation. However, Salameh et al. (2015) claimed that sentiment, to some extent, might be lost in translation. They translated Arabic social media posts to English using their in-house machine translation system and manual translation. The translated English posts were then classified automatically and manually. They discovered that translation tended to shift the post's sentiment to neutral, which was more significant in machine translated text. Upon further investigation, some sentiment words were replaced with words of opposite sentiment or worse, vanished in the translated texts. Occasionally, the typos in the Arabic posts also caused mistranslation and ambiguity.

On the other hand, Das et al. (2012) found that the translation could preserve emotions satisfactorily. First, they manually translated English news headlines to Bengali and Telugu due to the unavailability of machine translation systems for these languages at that time. Apart from this, these news headlines were also translated automatically to Hindi. Then, they manually annotated the translated headlines and compared them with the original emotions in the English headlines. They achieved

Cohen's kappa coefficients of at least 0.82, indicating that almost all translated headlines still evoked the same emotions as in English. Regardless, they found that the varied conjugated forms in the translated Telugu headlines posed the most challenges to their rule-based classifier, compelling them to derive emotions from the morphemes of the conjugated words.

Wei and Pal (2010) and Esuli et al. (2020) explored an extended approach of direct translation that could minimize the noise in faulty translations. They adapted structural correspondence learning (SCL) originally used in domain adaptation tasks to cross-lingual sentiment analysis. English reviews were translated into Chinese, and vice versa using Google Translate. However, unlike prior studies discussed, the translations were only used to select a set of pivot features. Once pivot features were selected, the original features from the two languages were then linearly projected to a lower shared dimensional vector space. Wei and Pal (2010) observed that using only these pivot features from the translated text resulted in a higher classification accuracy than using all features from the translated texts. This showed that machine translation could introduce some noise in the translated text that would affect sentiment classification performance.

2.3 Annotation Projection

Annotation projection mainly leverages resources in source languages to aid in building sentiment classifiers on target languages. This approach relies on existing parallel corpora to project the annotations from the source language to the target language and train a classifier in the target languages subsequently (Mihalcea et al., 2007; Öhman et al., 2020; Rasooli et al., 2018). Another variation of this method is to use machine translation to translate annotated source language corpus into the target

language and project the annotations to train a classifier in the target language (Abdel-Hady et al., 2014; Balahur & Turchi, 2014; Banea et al., 2008; Wan, 2012). Nonetheless, one could reverse the translation direction, that is, translate annotated target language corpus into the source language and project the annotations from the target language to train a classifier in the source language (Duh et al., 2011; Sazzed, 2020; Sazzed & Jayarathna, 2019). A less common variation translates unannotated target language corpus into the source language, annotates the translated source language corpus, and projects the annotations back to train a classifier in the target language (Banea et al., 2008). An example of annotation projection is visualized in Figure 2.3.

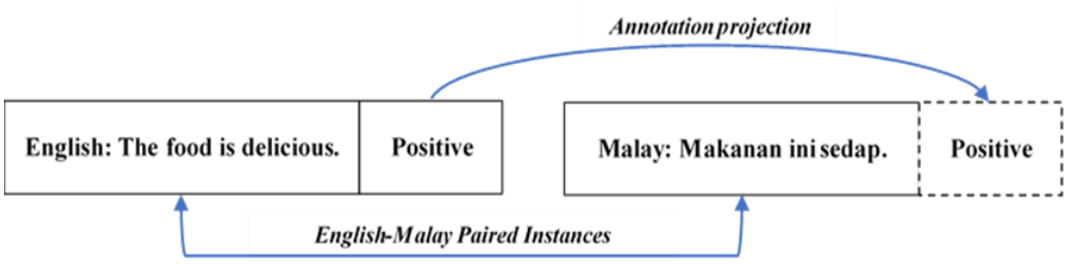

Figure 2.3 Annotation projection method

Annotated corpora required to perform sentiment analysis are limited in many languages. Annotation projection can ease the process of generating annotated corpora in target languages for training. Mihalcea et al. (2007), Rasooli et al. (2018) and Öhman et al. (2020) leveraged existing parallel corpus for annotation projection at the sentence level. Mihalcea et al. (2007) annotated the English side (source language) of the existing English-Romanian parallel corpus automatically using a rule-based classifier and a Naïve Bayes classifier from OpinionFinder separately. The comparison between the projection sources revealed that the Romanian classifier trained on the projection from Naïve Bayes annotations outperformed the rule-based classifier.

Similarly, Rasooli et al. (2018) trained a Naïve Bayes logistic regression classifier on source language corpora to predict text on the source-language side of the parallel corpus. Moreover, they also generalized annotation projection to include multiple sources. In the multi-source setting, majority voting was applied to project the most reliable annotations to the target-language side of the parallel corpus for training. Their study discovered that annotation projection would generally work well when a high-quality parallel corpus was available.

Instead of automatically annotating the English side of the parallel subtitles, Öhman et al. (2020) opted for manual annotations by university students studying sentiment analysis. The annotations were then projected to the corresponding Finnish side to train a Finnish classifier. Additionally, another set of Finnish subtitles was annotated manually to allow for comparison with the projected annotations. The experiment showed that the model trained on manually annotated Finnish subtitles achieved slightly higher F1 score and accuracy than the model trained on projected annotations.

However, like the bilingual lexicon, a parallel corpus is also a resource that is expensive to acquire. Abdel-Hady et al. (2014), Balahur and Turchi (2014), Banea et al. (2008) and Wan (2012) utilized machine translation to obtain the corresponding corpus in target languages. Banea et al. (2008) proposed a similar method to that of Mihalcea et al. (2007) with the difference in the machine translation system to obtain the corresponding corpus in Romanian. Banea et al. (2008) also explored manually annotated English corpus in addition to automatic annotation. The study showed that the Romanian classifier trained on automatic annotations had better performance than that trained on manual annotations. Although there was a slight drop in F1 as directly

compared to the experiment by Mihalcea et al. (2007) who used existing parallel corpus, the use of machine translation to obtain corpus in the target language was deemed a viable approach when no parallel corpus was available.

Balahur and Turchi (2014) experimented with three different machine translation systems: Google Translate, Bing Translator, and Moses, to translate English sentences into French, German, and Spanish. The English-side annotations were then projected to their corresponding translated sentences to train a classifier. They experimented with different feature representations as well, such as the presence of unigrams or bigrams and TF-IDF (term frequency-inverse document frequency) scores of unigrams or bigrams. They discovered that using unigrams of Boolean values was robust in situations where machine translation quality was not up to expectation. Also, they observed that using bagging as a meta-classifier could decrease the noise in the translated training sentences and thus produced satisfactory results.

Similarly, Abdel-Hady et al. (2014) translated annotated English tweets into Spanish and Portuguese using Microsoft Translator. The manual annotations of the English tweets were projected to the translated Spanish tweets and translated Portuguese tweets for subsequent classifiers training in these languages. Comparably, Wan (2012) translated the annotated English product reviews into Chinese and built a Chinese classifier. While Wan (2012) obtained satisfactory results using this approach with a balanced F1 score per class, Abdel-Hady et al. (2014) obtained predictions that seemed to favor the positive class in contrast.

In addition, some studies reversed the translation direction (i.e., translate from the target language to the source language) instead of generating annotated corpus in the target language (Duh et al., 2011; Sazzed, 2020; Sazzed & Jayarathna, 2019). In

such cases, the target language corpus was annotated, and a classifier was trained in the source language together with the projected annotations. Duh et al. (2011) translated Japanese, French and German product reviews into English to train English classifiers. Comparing the performance with the English classifier trained on original English reviews, there were degradations in the accuracy of all three translated English classifiers. Their study found that the vocabulary coverage of machine translation accounted for the degradation in accuracy instead of mistranslation. Nevertheless, the authors have shown that machine translation is sufficiently well-developed for cross-lingual adaptations.

Sazzed and Jayarathna (2019) and Sazzed (2020) performed a similar study to that of Duh et al. (2011) for Bengali comments and reviews, except that the test set was in translated English instead of in original English. Several classifiers were trained on the translated English corpus, and unlike what was observed in Duh et al. (2011), the majority of the classifiers showed some degree of improvement. Despite the fact that the translation was not accurate, machine translation was able to preserve the sentiments in translated English. It gave even better classification performance compared to the classifiers trained on the original Bengali corpus. All in all, their studies demonstrated that such translation direction could be reliable for linguistically complex languages.

A less common scenario of annotation projection could be seen in one of the experiments by Banea et al. (2008). The Romanian corpus, which they translated into English, was initially unannotated. Annotation of the corpus was performed on the translated English corpus using the high-coverage classifier from OpinionFinder. They then projected the annotations back to the initial Romanian corpus to train a classifier.

This approach that trained on the original Romanian corpus without translation was claimed to be more robust and would not suffer from translation errors.

2.4 Joint Learning

Joint learning is an approach that simultaneously performs learning on the corpora in both the source and target languages. In other words, it typically requires the corpus to be represented in pairs. This approach could be further broken down into joint representation learning and joint model learning. Joint representation learning is a unified learning process for bilingual representations that jointly optimize monolingual and bilingual constraints (Z. Wang et al., 2020). It is exemplified in Mogadala and Rettinger (2016), Tang and Wan (2014), G. Zhou et al. (2016), and Ghasemi et al. (2020). We visualize this method in Figure 2.4. However, utilizing paired corpus is not a necessary condition for joint representation learning. We also classify cross-lingual propagation proposed by de Melo (2015), which only requires weighted word pairs and a set of seed word embeddings as joint representation learning. Studies that adopted such an approach include Dong and de Melo (2018a, 2018b) and Giobergia et al. (2020).

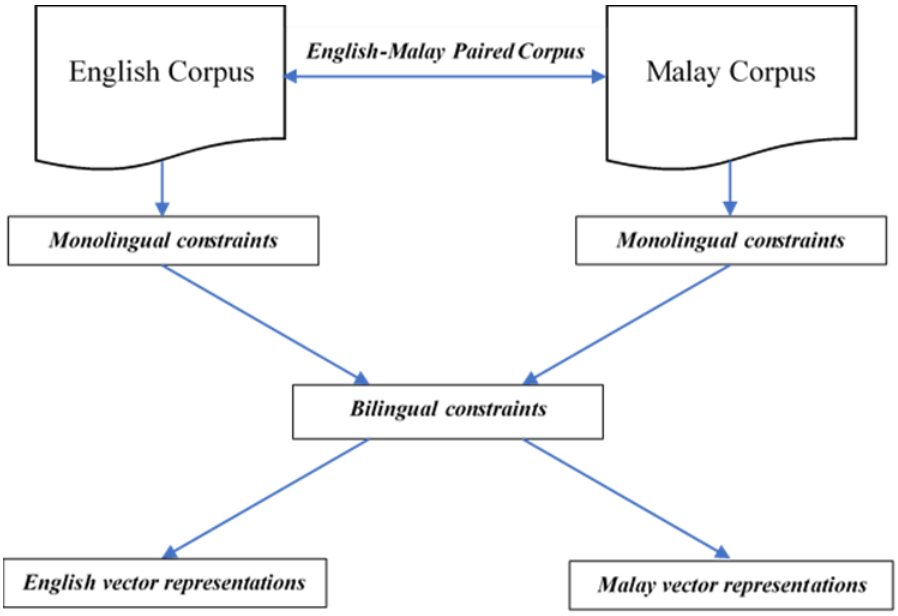

Figure 2.4 Example of joint representation learning

On the other hand, in joint model learning, monolingual representations from the source language and target language would usually be concatenated to form joint representations before training a classifier (Banea et al., 2010; Z. Chen et al., 2019; L. Zhang et al., 2018; X. Zhou et al., 2016b). This method is illustrated in Figure 2.5. In some cases, it is even possible to train a classifier on the synthetic multilingual training corpus, which combines the source corpus and its corresponding translated target corpus or vice versa (Fuadvy & Ibrahim, 2019). However, joint representation learning and joint model learning are not mutually exclusive. Some studies also performed both joint representation learning and joint model learning (Fuadvy & Ibrahim, 2019; H. Zhou et al., 2015; X. Zhou et al., 2016a). Furthermore, sentiment classification that solely relies on ensemble learning is also considered a special type of joint model learning (Wan, 2008).