CROSS-LINGUAL TRANSFER LEARNING FOR MALAY SENTIMENT ANALYSIS

LIU MINGYI

UNIVERSITI KEBANGSAAN MALAYSIA

CROSS-LINGUAL TRANSFER LEARNING FOR
MALAY SENTIMENT ANALYSIS

LIU MINGYI

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTER OF DATA SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

PEMBELAJARAN PEMINDAHAN MERENTAS BAHASA UNTUK ANALISIS SENTIMEN MELAYU

LIU MINGYI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT UNTUK MEMPEROLEH IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries, which have been duly acknowledged.

05 February 2025

LIU MINGYI
P137423

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my research supervisor, Associate Professor Dr. Nazlia Omar, for her invaluable guidance, encouragement, and unwavering support throughout the course of my research. Her expertise and dedication have been instrumental in helping me navigate the challenges of this work and achieve its completion. It has been an honor to learn under her mentorship.

I would also like to extend my sincere appreciation to the faculty and staff of the National University of Malaysia (UKM) for providing a nurturing academic environment and the resources needed to pursue this research. Their commitment to excellence has significantly contributed to my learning experience.

To my colleagues and friends, thank you for the stimulating discussions, collaboration, and encouragement throughout this journey. Your insights and companionship have made this experience truly enriching.

I am profoundly grateful to my family for their unconditional love, patience, and constant support. Your belief in me has been my greatest source of strength and motivation.

Finally, I wish to acknowledge any indirect contributors, including those who provided resources, data, or tools that facilitated the progress of my work.

This research would not have been possible without the collective effort, guidance, and encouragement of all those who have been a part of my journey. Thank you for your kindness, generosity, and belief in me.

ABSTRAK

Kajian ini meneroka aplikasi pembelajaran pemindahan silang bahasa untuk analisis sentimen dalam bahasa sumber rendah, memfokuskan kepada bahasa Melayu. Disebabkan kekurangan set data beranotasi yang besar dalam bahasa Melayu, penyelidikan ini bertujuan untuk menangani cabaran ketidakcukupan data dengan memanfaatkan set data bahasa sumber tinggi, khususnya bahasa Inggeris. Objektif utama adalah untuk menyelidik bagaimana model transformer yang telah dilatih, seperti DistilBERT, XLM-R Tiny dan MobileBERT, boleh digunakan dengan berkesan untuk meningkatkan prestasi analisis sentimen dalam bahasa Melayu. Selain itu, kajian ini mengkaji teknik pengoptimuman seperti pemampatan model dan penyulingan pengetahuan untuk meningkatkan kecekapan, ketepatan dan kebolehskalaan model ini dalam persekitaran yang terhad sumber. Dengan memperhalusi model ini dengan data bahasa Inggeris dan memindahkan pengetahuan yang dipelajari kepada bahasa Melayu, penyelidikan menunjukkan bahawa pembelajaran pemindahan silang bahasa boleh merapatkan jurang sumber dalam pemprosean bahasa tabii(PBT) untuk bahasa sumber rendah. Hasilnya menunjukkan bahawa XLM-R Tiny mengatasi model lain dari segi ketepatan analisis sentimen, manakala pemampatan model dan penyulingan menyumbang kepada peningkatan ketara dalam kelajuan inferens dan kecekapan memori. Penyelidikan ini memberi gambaran tentang potensi pembelajaran pemindahan silang bahasa untuk menangani cabaran dalam analisis sentimen untuk bahasa yang kurang diwakili dan menetapkan asas untuk penyelidikan masa depan dalam PBT berbilang bahasa.

ABSTRACT

This study explores the application of cross-lingual transfer learning for sentiment analysis in low-resource languages, focusing on Malay. Due to the scarcity of large, annotated datasets in Malay, the research aims to address the challenge of data insufficiency by leveraging high-resource language datasets, particularly English. The main objective is to investigate how pre-trained transformer models, such as DistilBERT, XLM-R Tiny, and MobileBERT, can be effectively applied to improve sentiment analysis performance in Malay. Additionally, the study examines optimization techniques like model compression and knowledge distillation to enhance the efficiency, accuracy, and scalability of these models in resource-constrained environments. By fine-tuning these models with English data and transferring the learned knowledge to Malay, the research demonstrates that cross-lingual transfer learning can bridge the resource gap in NLP for low-resource languages. The results highlight that XLM-R Tiny outperforms the other models in terms of sentiment analysis accuracy, while model compression and distillation contribute to significant improvements in inference speed and memory efficiency. This work offers valuable insights into the potential of cross-lingual transfer learning to address challenges in sentiment analysis for underrepresented languages and sets the foundation for future research in multilingual NLP.

TABLE OF CONTENTS

LIST OF TABLES

Table No. **Page**

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
XLM-R Tiny	Tiny Cross-Lingual Model - Roberta
MobileBERT	Mobile Bidirectional Encoder Representations from Transformers
UKM	Universiti Kebangsaan Malaysia
IMDB	Internet Movie Database
FT	Fine-Tuning
CNN	Convolutional Neural Network
RNN	Recurrent Neural NetworkZSL
ZSL	Zero-shot learnin

CHAPTER I

INTRODUCTION

1.1 Research Background

Sentiment analysis can also be related to as opinion mining, a branch of NLP that generally aims at the extraction of subjective information from text for opinion classification as positive, negative, or neutral. In the modern digital world, sentiment analysis has grown to be an important tool in making sense of user-generated content across diverse domains. The applications involve anything from market trend predictions and customer feedback assessments to political sentiment analysis and social media monitoring. Challenges in sentiment analysis come from the sophistication in which feelings are expressed apart from classifying into either positive, negative, or neutral, even multi-aspect sentiments come in-for example, when a user feels satisfied but at the same time frustrated with some other issues in a review, and finally sarcasm and irony, which might make an expression not coincident with a real emotional feeling. In particular, the expression of sentiment in social media and informal communication often includes such complexities, making sentiment analysis a far more intricate task.

High-resource languages have seen substantial improvements in sentiment analysis because of the availability of large annotated datasets and pre-trained models.However, the same has not been achieved for low-resource languages, creating an imbalance in global NLP capabilities. Low-resource languages have a peculiar set of difficulties, including but not limited to the scarcity of annotated datasets, limited linguistic resources, and missing pre-trained models. Unlike their high or low-resource language counterparts, large-scale corpora are scarce, making training effective machine learning models a great challenge. At the same time, the characteristic features of this language, namely complex morphology and rich syntactic variations, have their unique cultural expression, which really complicates the whole process of robust sentiment analysis modeling. These barriers not only restrict the usage of NLP tools regarding these languages but also the knowledge about digital interactions of diverse communities. For example, cultural and linguistic features may specify important difficulties in expressing sentiment; the same sentiment could be expressed differently in this or that cultural context. Therefore, models trained basically on English or other high-resource languages lack the ability to capture such subtleties when working with low-resource languages.

In recent years, one of the transformational approaches in machine learning has emerged as transfer learning. Transfer learning, in general, is borrowing knowledge acquired in one task or domain and applying it to a related task or domain. This works especially well in NLP, where large pre-trained language models, such as BERT and GPT, along with their multilingual variants, have shown the ability to generalize across tasks. Cross-lingual transfer learning now builds on this foundation by leveraging pre-trained multilingual models to bridge linguistic gaps between high-resource and low-resource languages. Cross-lingual transfer learning enables the sentiment analysis systems to work under low-resource settings without requiring a large number of annotated data by exploiting shared semantic and syntactic features among languages. Despite the promise of transfer learning, challenges still exist, particularly in cases of far-from-similar source and target languages. For instance, models may find it hard with languages that have very little linguistic overlap, meaning that the technique of transfer learning should be further developed to take such cases into consideration. Another important direction, domain adaptation is also an essential point: a model is unlikely to work properly on formal texts, say news or legal documents, even though it has been trained on social media posts in the same language. The current research focuses on using synergy between high-resource and low-resource languages to solve practical challenges. In this regard, English will be considered the high-resource language and Malay, being among the low-resource languages spoken extensively in Southeast Asia, will be taken as a target language.The reason behind the selection of English is its dominance in research related to NLP due to its abundance in available datasets, benchmarks, and overall tools.

Malay, on the other hand, is a low-resource language even though it boasts a huge speaker base. Such an imbalance needs to be addressed in order to make linguistic inclusivity possible in sentiment analysis. Cross-lingual transfer learning thus comes into play by transferring the linguistic and task-specific knowledge from English to Malay, hence mitigating resource constraints and allowing the development of Malay sentiment analysis tools. It also alignes with the larger vision of democratizing AI and NLP technologies. The work, therefore, tried to develop methods for the sentiment analysis of texts in low-resource languages, furthering the broader objective of decreasing the digital divide across linguistic communities. Furthermore, it is an argumentative piece on developing NLP models which would respect and represent the cultural and linguistic diversity inherently associated with human communication.

For example, enabling the models to comprehend and process culturally subtle expressions in Malay will not only expand the scope of NLP but also further strengthen the representation of regional languages in digital spaces.

This approach is inclusive and ensures that diverse linguistic communities are not left behind in the technological revolution. The integration of cross-lingual transfer learning into this context presents a typical example of its potential to serve as a powerful tool that will ever increase the scalability of NLP applications and their adaptability. This work aims to explore how transfer learning can help to overcome low-resource language processing challenges more effectively, utilizing multilingually pre-trained models like MOBILE BERT, XLM-R TINY, and other transformer-based architectures to drive inclusive language technologies forward. Besides, it also underlines the importance of exploiting multilingual frameworks to find those linguistic similarities that allow, with the support of automatic systems, the understanding and sharing of knowledge across cultures. For the future, this research contributes to developing not only sentiment analysis for low-resource languages but also to new perspectives in NLP technologies.

Data augmentation methods and knowledge distillation are particularly relevant to improving the model performance of low-resource languages by way of solving limitations of data and improving model efficiency.

Also, the integration of multimodal data-images or videos, apart from texts-will enable the construction of more robust Sentiment Analysis, capturing richer contextual information, and this is a very relevant avenue of research when it comes to social media platforms filled with informal information and multimedia contents. Finally, with the evolution of NLP technologies, solving the challenges thrown up by digital content and sentiment analysis across languages and regions is the need of the hour. This research would help in contributing to making NLP technologies more inclusive and adaptable to diversified linguistic contexts so as to contribute toward the key role of their evolution in sentiment analysis as a tool universally applicable.

1.2 Problem Statement

Most works, therefore, have pointed out different challenges that have been doing sentiment analysis in low-resource languages, greatly barring the development of effective solutions. The low-resource languages, such as Malay, are limited in having very scarce annotated datasets, poor lexical resources, and a lack of pre-trained models which handle their specific linguistic characteristics. Besides, Malay contains a number of morphological, syntactic, and semantic patterns that make the direct application of sentiment analysis models designed for high-resource languages, such as English, difficult to perform (Cambria et al.2016). While there has been increasing research on multilingual NLP, most works have been done on high-resource languages like English, Spanish, and Chinese, with relatively little work on languages like Malay (Joshi et al., 2017). For example, Cambria et al. (2017) highlighted the challenges of adapting sentiment analysis models across languages due to differences in syntactic structure and cultural nuances. These challenges are particularly prominent in Malay, where the scarcity of labeled sentiment data and the lack of a robust sentiment lexicon hinder model performance (Nazratul Naziah et al., 2021).

Besides, previous studies have pointed out the challenges of cross-lingual transfer learning. Some scholars have noted that despite the potential of cross-lingual transfer learning, migrating knowledge from high-resource languages to low-resource languages often leads to domain mismatch and suboptimal fine-tuning (Artetxe & Schwenk, 2019). This includes issues such as overfitting patterns specific to high-resource languages and the inability to effectively generalize to low-resource languages (Ruder et al., 2019).

Nguyen et al. (2020) identified a number of important gaps in the evaluation of performance for methods of cross-lingual transfer learning methods in sentiment analysis and, in particular, when multiple converter-based models were compared. Despite the great success of cross-lingual transfer learning models, such as XLM-R and DistilBERT in general multilingual tasks, their relative performances in application-specific tasks, such as Malay sentiment analysis, have not yet been explored.

In the light of the above challenges, this study, therefore, seeks to fill the following gaps: to investigate the efficiency of cross-lingual transfer learning from a high-resource linguistic sentiment dataset, English, to improve the performance of Malay sentiment analysis; and to compare the performance of the converter models DistilBERT, XLM-R Tiny, and MobileBERT for applicability in the sentiment analysis of low resource languages.

We shall also explore fine-tuning possibilities for cross-linguistic models in overcoming the issues of domain mismatch and overfitting that were highlighted by Ruder et al. (2019).

By addressing these challenges, this study will help expand the understanding of sentiment analysis for low-resource languages and provide a systematic framework for overcoming data scarcity through the use of cross-lingual transfer learning techniques.

1.3 Research Motivation

This research is a response to the pressing need felt by the research fraternity to address the challenge of linguistic inclusiveness in NLP. Sentiment analysis has grown from an emerging field of research to an indispensable tool in e-commerce, marketing, opinion research, and policy making. This dominance of high-resource languages, such as English, in NLP research has created huge inequality in the availability and quality of sentiment analysis tools across low-resource languages. This further leads to inequality in the way different language communities are able to benefit from technological advances, thus increasing the digital divide.

The challenges posed by low-resource languages like Malay are at least threefold: very limited annotated datasets, rich diversity in linguistic structure, and underrepresentation in pre-trained models. These limitations ensure that a significant number of low-resource language speakers cannot access or utilize sentiment analysis techniques to address local and regional needs. The research is motivated by the urge to develop effective solutions that can bridge this gap and make sentiment analysis accessible to a greater number of language groups, promoting the democratization of AI and its applications.

Recently, cross-lingual transfer learning has been the most effective line of research that addresses these challenges. Pre-trained multilingual models embed such common semantic and syntactic features and present excellent generalization performance across languages. A language like Malay finds a solution using such models, setting a precedence for the extension of such solutions to other low-resource languages toward sentiment analysis. The motivation for this further coincides with the growth in recognizing that research in AI should be supportive of all linguistic communities, both resource-rich and resource-poor.

From a practical point of view, the integration of sentiment analysis into low-resource languages like Malay opens up a lot of opportunities. Businesses can understand customer feedback from various regions, governments can monitor public sentiment to formulate policies, and social campaigns can understand community needs. This has far-reaching implications for industries and organizations operating in multilingual and multicultural environments, especially in Southeast Asia where Malay is widely spoken.

In addition to this, such a study is generally motivated to investigate how to make the NLP approach more capable. While it holds great potential in multilingual pre-training, further scalability, efficiency, and adaptation to languages for which less training data exists could be highly promising. It makes an attempt at making valuable progress in cross-lingual transfer learning with the study of innovative approaches that include techniques in model optimization and finetuning strategy. These experiences can provide insight into creating better and more effective NLP tools for low-resource languages around the world.

Finally, the personal importance of this research is to realize that language is essentially the most important constituent part of one's identity and communication. That NLP techniques preserve and reflect linguistic diversity is a challenge but at the same time is also a cultural and ethical issue. The given study aims at making a small contribution toward a more inclusive digital environment where all languages, regardless of their resource status, are respected and supported with regard to the advances in technologies.

1.4 Research Objectives

The objectives of this study:

1. To develop a cross-lingual transfer learning framework using a multilingual converter model to improve sentiment analysis for low-resource languages especially Malay.
2. To evaluate multiple cross-lingual transfer learning models for sentiment analysis in the low-resource language Malay.

1.5 Research Scope

This research examines cross-lingual transfer learning techniques to target sentiment analysis in low-resource languages, with significant focus on Malay. Specifically, the study tests the ability of multilingual transformer models, DistilBERT, XLM-R Tiny, and MobileBERT, to transfer knowledge from a high-resource language, namely, English, to a low-resource language.

The research scope of this article includes the following parts.

1. selection and Preparation of DataSet:

Exploiting public English sentiment analysis datasets to fine-tune high-resource models and collecting and preprocessing a Malay sentiment analysis dataset to assess the performance of cross-lingual transfer.

2. Development and optimization of model

English datasets for fine-tuning multilingual pre-trained models.Implementing optimization approaches (including knowledge distillation and model compression) towards enhancing the overall efficiency and accurateness of FCN in low-resource settings.

3. Performance Evaluation

Performance testing of the constructed framework based on accuracy, precision, recall and F1-score using the Malay sentiment analysis data set.Performance comparison of the selected models (DistilBERT, XLM-R Tiny and MobileBERT) for the task of Malay sentiment analysis.

4. Framework Generalization

Introducing a transferable methodology for implementing cross-lingual transfer learning on alternative low-resource languages.

5. Application Context:

showcasing real-world applications of the proposed framework in sentiment analysis use-cases like social media or customer feedback analysis where you have low-resource languages.

These works are limited to sentiment analysis tasks and do not investigate other types of tasks commonly encountered in NLP. It only relates to the Malay language as a case study and guides how the resulting framework can be adapted to other low-resource languages in this work.

1.6 Significance of Project

The work bears immense importance to the field of NLP in general and has broader implications, especially in low-resource languages like Malay, for which the growth of powerful NLP tools has always been restricted by the scarcity of labelled data and linguistic resources. In this paper, the effectiveness of cross-lingual transfer learning is examined in enhancing the performance of Malay sentiment analysis models using knowledge from high-resource languages like English. This approach will, therefore, encourage the application of sentiment analysis to low-resource settings and show ways in which challenges due to limited annotated datasets for underrepresented languages may be overcome by transfer learning. Furthermore, the current study also applies state-of-the-art optimization techniques, including model compression and knowledge refinement, which have been playing an important role in enhancing efficiency and scalability for pre-trained multilingual models.

These techniques can reduce the computational cost of large models by a significant amount, thus making them suitable for resource-constrained settings, such as multilingual environments, mobile devices, or applications with limited computing power. This study improves the efficiency and accuracy of such models, which in turn contributes to the overall goal of creating practical, real-world NLP solutions that organizations, government agencies, and businesses can use across different settings. Another important aspect of the project is that it can enable the development of even more holistic language technologies. As research has shown, cross-lingual learning among migrants may bridge the gap in languages and thus provide equal opportunities for using NLP tools in different languages.

This is particularly important in Southeast Asia, since languages like Malay are still underrepresented in research and application within the field of NLP. Therefore, the successful application of transfer learning here favors not only sentiment analysis for Malay but also opens the door to similar approaches for other low-resourced languages, thus fostering better international communication and knowledge access. The framework and methodology developed in this study will finally be able to act as a bedrock for further research and applications. The successful porting of sentiment analysis models into the Malay language has far-reaching implications, from better customer service tools to improved social media monitoring in languages with poor NLP resources. The current research therefore fills a gap in the literature, but it also paves the way for further progress in multilingual NLP, as this work enables other research on developing tools that are more sensitive to different linguistic features and can meet the needs of different, underrepresented language communities.

1.7 Organization of Project

The methodology section outlines the systematic approach employed in this study. Figure 1.1 demonstrates the stages of the methodology.

Chapter I Introduction： This chapter embodies the study's structure by outlining the background of the research context. It sets forth the issues in the research that are particularly relevant to the low-resource languages, like Malay, with the help of the problems posed by the sentiment analysis. The chapter therefore formulates the research problem and specifies the research gap, highlighting the need for a solution to bridge the linguistic divides in NLP for Malay. In addition,the study's goals are well-defined, and the methodology and approaches are stated to be aimed at overcoming the challenges and consequently improving sentiments analysis for Malay through cross-lingual transfer learning.

Chapter II Literature Literature Review： This chapter looks at what other people have already written about text classification and sentiment analysis in the Malay language. It examines the different ways documents can be classified using natural language processing (NLP) and explains why it is important to use languages with a lot of data to improve sentiment analysis in languages with less data. The chapter also looks at the different ways features can be extracted and machine learning algorithms used in the context of Malay text mining. It also looks at earlier research and points out what hasn't been done yet in the study of how feelings are expressed in the Malay language.

Chapter III Methodology: This chapter describes the methodology of the research that has been conducted in this work. It explains in detail the training and evaluation methodology of the developed sentiment analysis models. It covers the datasets adopted for the experiment, including the high-resource English language dataset and the low-resource Malay dataset. The chapter explores the pre-processing techniques applied on the datasets such as tokenization, stopword removal, and stemming. It also describes the preparation of text data for machine learning models and the used algorithms, such as DistilBERT, XLM-R Tiny, and MobileBERT. The chapter also touches on cross-lingual transfer learning, model optimization techniques, and evaluation metrics.

Chapter IV: Results and Discussion: The chapter presents experimental results done in the research study. Comparing the results from different pre-trained models at Malay sentiment analysis includes DistilBERT, XLM-R Tiny, and MobileBERT. Discussion on performance results has been supported with accuracy, precision, and recall of respective models. Discussions have also been included about transfer learning, which helps improve the models' performance. It also looks into how different techniques aimed at the improvement of the models-such as making them smaller, using them to share knowledge-perform.

Chapter V: Conclusion and Future works: The last chapter summarizes the key results and ideas obtained in the present research. It puts together the findings obtained from the comparison and demonstrates how cross-lingual transfer learning can be done to overcome some of the problems related to Sentiment Analysis, when resource-poor languages are to be dealt with. Further, the chapter goes on to show other areas wherein future research has to be undertaken. It also covers how SA models can further be improved and the use of transfer learning for other low-resource languages. Finally, this work reflects on implications that this study may have for inclusive language technology development and the general domain of multilingual NLP.The chapter content is shown in Figure 1.1.

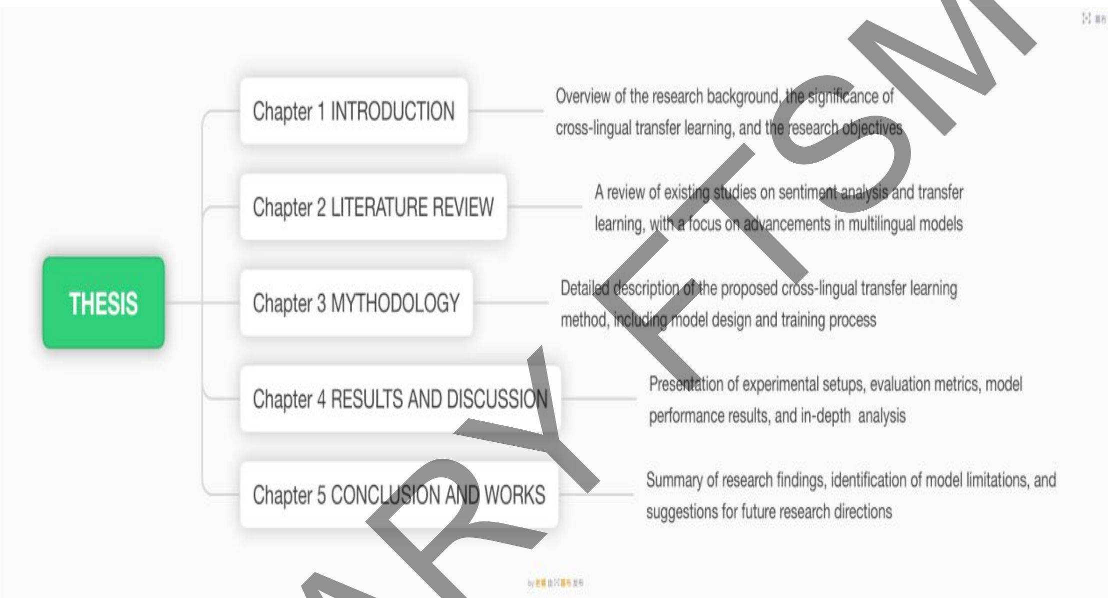


Figure 1.1 Research Methodology Stages

1.8 Summary

Chapter I introduces the background to the study. It has highlighted different challenges that exist in conducting a sentiment analysis of under-resourced languages like Malay. This is attributed to the scarcity of labeled datasets. It indicates where the research gaps are and has identified a problem statement, explaining that we need better means of doing sentiment analysis in these languages. It proposes the aims of the study, which consist in exploring how to transfer learning across languages and how to make our models better, by compressing and reducing the amount of knowledge required by them. All these techniques are targeted to improve the performance of sentiment analysis and further create a framework for the use of these techniques on other low-resource languages. In general, Chapter I introduces the research, stating the challenges, objectives, and methodologies that will be further discussed in the following chapters.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Sentiment analysis could be also entitled as opinion mining and is considered to be an integral part of NLP. This technique involves information collection about what a particular person or individual thinks from texts. This has gained a lot of attention in recent times, as it can be applied to many different industries, including marketing, customer service, and social media monitoring. With more and more people using different languages all over the world, the use of sentiment analysis in low-resource languages is becoming increasingly popular, for which there is not much data available. It presents several challenges, such as poor quality or a lack of labeled data in the languages considered, but cross-lingual transfer learning holds enormous promise: it offers several ways to adapt knowledge in high-resource languages to low-resource languages. Approaches include making use of large pre-trained models, multilingual embeddings, and machine translation. The technique resulted in performance improvement in Sentiment Analysis for Low Resource languages and became a pathway for applications across most languages.

This section discusses what other studies have done, what has recently been achieved, and what problems remain to be solved. This sets the scene for the research that is contained within this thesis.

2.2 Development and Current State of Sentiment Analysis

2.2.1 Dfinition and Application

Sentiment Analysis is a way of opinion identification and classification in text. The sentiment can be positive, negative, or neutral. In the last years, SA became one of the most explored NLP tasks because it is versatile and can be applied in many different topics. For example, it can be used to understand what customers think about products or services (Pang & Lee, 2008), or to track what people are saying about brands on social media like Twitter, Facebook and Instagram (Pak & Paroubek, 2010). In public opinion analysis, the feelings expressed in political discussions are studied to understand voter opinions and public attitudes towards policies (Liu, 2012).

Other applications that are important include reputation management of a company, where companies gauge public opinion with the view to minimizing reputational risks; analysis of customer feedback that enables a company to shape its products and services; and market research, whereby sentiment data will be used to drive marketing strategy and product development (Cambria et al., 2013). Sentiment analysis has recently gained wide applications in health for the purpose of understanding patient feedback and discussions about health. This aids in patient care and decision-making within the realm of medicine (Kumar et al., 2021). All these applications highlighted give meaning as to how much different areas apply the decision made on data.

2.2.2 Methods for Sentiment Analysis

The ways we analyse people's feelings have got better a lot in the last few decades. There are four main types: methods based on the lexicon, machine learning methods, deep learning methods, and pre-trained language models.

1. Lexicon-Based Methods

Early sentiment analysis techniques relied heavily on predefined sentiment lexicons or dictionaries, where words are manually labeled as positive, negative, or neutral. Prominent examples include the SentiWordNet lexicon (Esuli & Sebastiani, 2006) and the AFINN lexicon (Nielsen, 2011). Lexicon-based methods are computationally efficient and straightforward to implement, making them suitable for scenarios with limited computational resources. However, they face several limitations. The most significant drawback is their inability to capture the context in which words are used. For instance, the word "bad" in the context of "not bad" conveys a positive sentiment, but a lexicon-based method would likely misinterpret it as negative. Moreover, lexicon-based approaches struggle with domain adaptation, as they often fail when applied to texts from industries or topics that were not considered during the lexicon's development (Liu, 2012). Despite these limitations, lexicon-based methods remain a useful baseline for sentiment analysis tasks.

2. Machine Learning Methods

The adoption of machine learning techniques marked a paradigm shift in sentiment analysis. Traditional algorithms such as Support Vector Machines (SVM) (Joachims, 1998), Naïve Bayes (McCallum & Nigam, 1998), and decision trees became popular for text classification tasks. These methods typically require feature extraction techniques such as bag-of-words, n-grams, and part-of-speech tagging. They demonstrated superior performance over lexicon-based methods by enabling models to learn patterns in labeled datasets. However, machine learning-based approaches are not without challenges. They require significant manual effort in feature engineering, which can be time-consuming and domain-specific. Additionally, their performance is highly sensitive to the quality of training data. Nevertheless, they laid the groundwork for more advanced sentiment analysis techniques.

3. Deep Learning Methods

With the advent of deep learning, models began to automatically learn hierarchical features from raw text. Techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have demonstrated impressive performance in various sentiment analysis tasks (Socher et al., 2013). Unlike traditional machine learning, deep learning methods eliminate the need for manual feature extraction by automatically learning representations from data. CNNs are very good at understanding the relationships between different parts of a text, while RNNs and LSTMs are particularly good at dealing with data that is in a sequence, which makes them perfect for sentiment analysis. However, training these deep learning models requires large amounts of labelled data and a lot of computing power.

4. Pre-trained Language Models

The other big change for the field of sentiment analysis is introduced by the development of advanced deep learning models. In this paper, we have used three transformer-based models: These are DistilBERT, XLM-R Tiny, and MobileBERT. DistilBERT (Sanh et al., 2019) is a lightweight version of BERT developed by using something called "knowledge distillation". It retains 97% of the performance of BERT but is 60% faster; hence, it will be very suitable for tasks that need quick completions without losing much accuracy. The small version by Bajaj et al. 2021 of XLM-RoBERTa is called the XLM-R Tiny model, and this model tends to understand different languages very well, with just a little consumption in computing. It is particularly effective in analyzing the emotions behind text written in various languages, which is very useful when the computational power is not that great. MobileBERT: a lighter version for mobile and edge devices. This development of BERT by Sun et al. 2020 is efficient and powerful. A specially designed inverted bottleneck structure and teacher-student training contribute to achieving this. Model types best suited to the above-stated balance between accuracy and efficiency for multilingual and low-resource sentiment analysis tasks.

2.2.3 Challenges in Sentiment Analysis

There have been big improvements in the ways we analyse people's feelings, there are still several problems:

Limited data: One of the biggest problems in sentiment analysis, especially for languages that don't have much data, is the lack of labelled datasets. Languages like English have lots of data that has been labelled to show if something is positive or negative, which helps train models. But for languages like Malay, Swahili, and Hindi, there is often not enough data with labels, which means that the models can't work as well.

Domain Adaptability: Another big problem is adapting to different domains. Sentiment analysis models trained in one area (e.g. product reviews) may not work well in another area (e.g. political opinions). This is because language, terminology and how people express their opinions can vary a lot between different areas.

Cultural differences: Sentiments can be expressed very differently across cultures, which makes sentiment analysis more difficult in places where lots of different languages are spoken. What is seen as positive in one language or culture might not be seen that way in another. This makes it even harder for systems that are designed to work in lots of different languages..

2.3 Foundations and Advances in Cross-lingual Transfer Learning

2.3.1 Definition of Cross-lingual Transfer Learning

Cross-lingual transfer learning refers to effective knowledge transfer from the source language to the target, which usually includes a high-resource language for learning and transferring information to one without enough data or resources. That is excellent, as that helps alleviate the problem of not having enough data, which is a big problem in NLP and especially for low-resource languages. The biggest advantage of cross-lingual transfer learning is that it can use large, pre-trained models and multilingual embeddings which have been trained on lots of data from high-resource languages. While this includes multilingual versions of BERT and other transformer-based architectures, it provides much valuable knowledge to the target language for improving the performance bar in different NLP tasks, such as sentiment analysis, for which the classic approaches fail due to a general lack of data.

This, in essence, makes cross-lingual transfer learning quite an excellent means to link the features of various languages. Though languages can be quite different structurally, with vocabularies and grammar, a common semantic space can be generated where words or phrases of similar meanings across different languages are represented comparably. That means models trained on a high-resource language can be effectively utilized in a low-resource language even when there is not much labeled data for the target language. It means we can use this knowledge to help languages that don't have much computing power or information about how their language is used.

Such is an approach that permits zero-shot as well as few-shot learning; that is to say, models trained on the high-resource languages can be made useful in the low-resource setting with very slight retraining. Under zero-shot settings, the model is able to predict the sentiments of texts even in a language it has not seen. To be sure, it can without example texts from a target language only rely on its own acquired knowledge. Few-shot learning enables the model to be fine-tuned with a small number of examples in the target language and thus works even better. Cross-lingual transfer learning will, therefore, be particularly effective in cases where the collection of large sets of examples for a low-resource language is impossible. This innovative approach is important in advancing NLP in multilingual environments, since it allows for the application of sentiment analysis and other language-processing tasks to a much wider range of languages.Figure2.1 shows architecture of the transfer learning model.

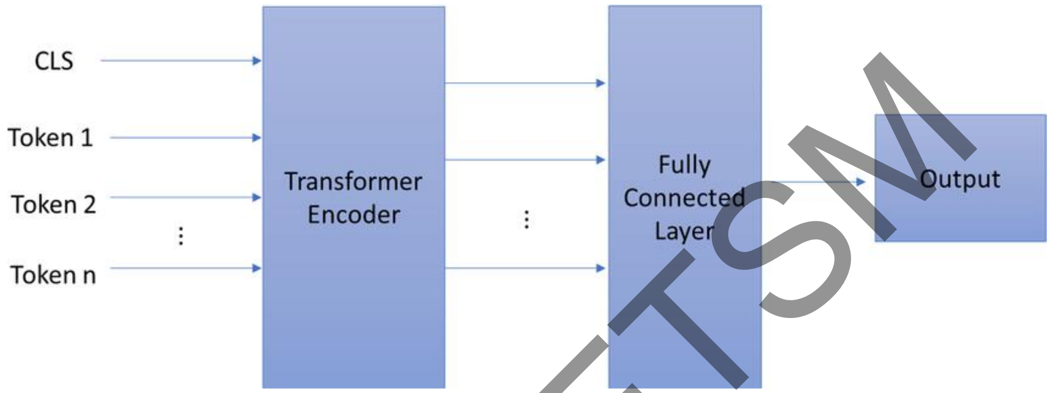


Figure 2.1 Transfer learning model

This figure shows the basic architecture of XLM-RoBERTa, which is based on the Transformer encoder for natural language processing (NLP) tasks. The text is first segmented into tokens and special CLS tags are added, then input into the Transformer encoder to learn contextual information using the Self-Attention mechanism. After being processed by the Fully Connected Layer, the final output is the prediction result, such as sentiment classification or text classification. This architecture is particularly suitable for cross-lingual transfer learning and performs well in low-resource language tasks such as Malay.

2.3.2 Early Approaches to Cross-lingual Learning

Early methods for transferring learning between languages used machine translation and ways of matching words, thus laying the foundation for later developments in this area.

Machine Translation: the first ways of using machine translation to help with translation between different languages involved the use of machine translation systems to convert text in the target language into the source language. Subsequently, the text could then be processed by those models that were trained on big, high-resource datasets, usually in English. This would allow the application of a sentiment analysis model in any language using translation as a bridge, and this would easily solve the scalability problem. As promising as it sounded, however, there was a serious defect with this approach. The most significant issue using this approach lay wholly and completely on the quality of the translation system. A wrong translation can change the meaning of the text. This, in turn, can affect the accuracy of sentiment analysis. For example, idiomatic expressions, cultural nuances, or context-specific terms might not be correctly translated, which could lead to errors in sentiment analysis. This approach also suffered from issues related to languages with complex grammar or poor translation systems and thus was not very effective for many low-resource languages.

Word Alignment Models: The inefficiency of the machine translation methods gave birth to another technique involving cross-lingual word embeddings, such as word2vec and fastText. These models aimed to map words from different languages into a shared vector space, allowing semantically similar words to be aligned across languages. In this way, by aligning words of similar meanings in different languages, these word embeddings allowed the transfer of sentiment analysis models across languages. This represented a significant improvement from earlier machine translation-based methods. Word embeddings had the immediate impact of bypassing some of the problems related to translation errors, since semantic meaning was effectively captured by the word representations even when surface forms across languages differed. However, word alignment models faced a number of challenges with regard to correctly aligning syntactic structures between languages. While they could align words of similar meanings, most often they would not perform well with languages that use different syntactic constructions, word order, or morphological structure. Also, these models were rather limited by the quality and coverage of the word embeddings, as they required huge corpora to provide accurate vector representations. Despite these limitations, the release of cross-lingual word embeddings represented a crucial next step in cross-lingual transfer learning work, finally allowing for significantly more accurate transfer of sentiment analysis models beyond previous methods.The development of Cross-lingual Transfer Learning Technologies is shown in Table 2.1.

Table 2.1 Development of Cross-lingual Transfer Learning Technologies

2013: Word Embeddings	Proposed methods for monolingual word embeddings, with cross-lingual modeling not yet mainstream.
2016: Multilingual Word Embeddings	Introduction of methods using shared vocabularies and joint embeddings (e.g., FastText).
2018: mBERT	Google introduced multilingual BERT, laying the foundation for cross-lingual transfer learning.
2020: LASER and LaBSE	Focused on generating universal multilingual sentence representations, expanding the applicability of cross-lingual models.
2023: Fine-Tuned Cross-Lingual Models	Increased focus on transfer performance for low-resource languages, combining multi-task learning and few-shot learning methods for further improvement.

2.3.3 Multilingual Pre-trained Models

The progress made in recent works on cross-lingual transfer learning has been driven to a great extent by the development of efficient multilingual models with a performance emphasis on sentiment analysis for low-resource languages. These models seek to reduce computational and data demands associated with their traditional pre-trained architectures while retaining or improving performance on cross-lingual tasks. Among these, DistilBERT, XLM-R Tiny, and MobileBERT have emerged as lightweight yet powerful alternatives to larger multilingual models like mBERT and XLM-R.

1. DistilBERT

DistilBERT was introduced by Sanh et al. (2019) as a leaner and speedier variant of the original BERT model, developed by a knowledge distillation technique. The underlying idea of DistilBERT is to train a smaller student model to mimic the performance of a larger teacher model on similar capabilities. Though originally developed for monolingual tasks, multilingual variants of DistilBERT have been adapted for cross-lingual applications.

In cross-lingual sentiment analysis, DistilBERT enjoys quite a few advantages in computational efficiency and memory usage. Because DistilBERT retains about 97% of BERT's language understanding with only 60% of its parameters, it enables faster training and inference, which is ideal for real-time applications and deployment in resource-constrained environments. The compact size of the model is highly useful when handling low-resource languages where computational resources can be limited.

While DistilBERT indeed showed competitive performance on cross-lingual tasks, the smaller architecture may have difficulty capturing such subtle language-specific nuances as larger models like XLM-R. With this limitation, DistilBERT would still be worth using in many multilingual sentiment analysis tasks because of its efficiency and scalability in scenarios that are required.

2. XLM-R Tiny

XLM-R Tiny is a lightweight version of the powerful XLM-R model, which was developed to solve the computational challenge with large-scale multilingual models.

With the same idea as in XLM-R, such as cross-lingual masked language modeling and shared multilingual vocabulary, the performance of XLM-R Tiny has been directed toward high performance with much smaller architecture.

XLM-R Tiny works especially well for low-resource languages in sentiment analysis. This model learns shared representations across many languages and thus will allow for effective transfer learning even in languages with very little training data. Careful design and optimization have allowed the XLM-R Tiny to outperform other lightweight models on cross-lingual benchmarks. It makes it especially apt for low-power applications like smartphones and IoT devices owing to its small size and low computational demands. Using even larger models-such as the XLM-R and mBERT-models in those applications would, therefore, hardly be practical at all.

While tiny in nature, XLM-R Tiny also tends to perform worse on languages that are highly different in syntactic structure or those languages which are very sparsely represented in the training corpus. This model presents an excellent fit for cross-lingual sentiment analysis tasks in general, mainly when computational resources become limited, given the right balance between efficiency and performance.

3. MobileBERT

MobileBERT, proposed by Sun et al. (2020), represents a significant step toward the development of efficient transformer models optimized for mobile and edge devices. MobileBERT adopts the BERT architecture but is enhanced in a number of ways to reduce the size and computation of the model without sacrificing much performance: bottleneck structures and a teacher-student framework.

MobileBERT is a generic and effective model for cross-lingual sentiment analysis. This model, because of its lightweight architecture, runs effectively on less powerful devices like smartphones and tablets. That has made it one of the first choices in real-world applications where accessibility and responsiveness matter. Still compact in size, MobileBERT retains the capability to capture deep semantic and syntactic relations, thus offering robust performances in multiple languages.

Therefore, MobileBERT has the specialty of dealing with low-resource languages, effectively borrowing knowledge from the pre-training high-resource language on a specific task such as sentiment analysis. Fine-tuning it on the datasets of sentiment analysis would generalize well to many languages, including zero- or few-shot learning settings. Similar to the other compact models, this may be impacted by the quality and multilinguality of training data for highly structurally different or poorly represented languages.The comparison of the three models is shown in Table 2.2.

Table 2.2 Comparison of Three Models

Scenario	**DistilBERT**	**XLM-R Tiny**	**MobileBERT**
Sentiment Analysis (Efficient Scenario)	Stable performance in high-efficiency scenarios, suitable for real-time analysis	Better suited for complex low-resource sentiment analysis tasks	Higher efficiency in optimized scenarios and can be deployed on mobile devices
Cross-lingual tasks	Moderate performance, suitable for small cross-language tasks	Excellent, especially in multilingual scenarios	The performance is close to XLM-R, but it focuses on single language sentiment analysis optimization
Low-resource language support	Slightly weak, but relatively stable	Best, designed for low-resource languages	Good performance, suitable for lightweight deployment

Following is the summary of advantages and disadvantages.

1. DistilBERT:

Advantages: high efficiency, fast inference speed, small memory usage, suitable for general NLP tasks.Disadvantages: slightly inferior performance in cross-lingual tasks, especially in the processing of low-resource languages.

2. XLM-R Tiny:

Advantages: specially optimized for cross-lingual tasks, excellent performance, suitable for low-resource languages, fast inference speed.Disadvantages: may rely heavily on the diversity of training data.

3. MobileBERT:

Advantages: designed for mobile devices, fast inference speed, both efficiency and performance.Disadvantages: a little weaker than XLM-R Tiny in support for cross-lingual tasks but better performance in single-language optimization. Balancing Efficiency and Performance.

In fact, DistilBERT, XLM-R Tiny, and MobileBERT represent a new wave in cross-lingual transfer learning-a focus on efficiency without sacrificing effectiveness. These models avoid many of the challenges with larger multilingual models that come at a high computational cost and with heavy memory usage while being able to provide competitive performance in sentiment analysis tasks across languages.

These lightweight models are especially important for low-resource languages because they enable zero-shot and few-shot learning and reduce reliance on large volumes of labeled datasets. However, the main challenges are related to fine-tuning these models for specific linguistic and domain-specific contexts, in particular for languages that are far from the languages represented in their training data. Further innovations can be expected as research in this area goes on, allowing these cross-lingual models for sentiment analysis-and most other tasks in natural language processing-to scale up, broaden, and increase their speed.

It argues that DistilBERT focuses on the efficiency-memory trade-off, whereas XLM-R Tiny focuses on improving the performance of cross-lingual tasks. Lastly, MobileBERT balances both performance and mobile optimization requirements, making it suitable for a wide range of scenarios and resource constraints in sentiment analysis.

2.3.4 Lightweight Transformers and Optimization Techniques

While multilingual pre-trained models have achieved enormous success in cross-lingual transfer learning, their computational cost is excessively high. In general, this may pose serious problems when such models are being deployed in realistic low-resource settings where computational resources are limited or not available at all. More recently, several efforts have placed emphasis on optimizing such models for better efficiency while retaining their effectiveness on tasks such as sentiment analysis in low-resource languages. These are efforts being invested in multilingual models, aiming at their practicality for diverse applications through reasonable computational cost performance.

1. Lite Transformers

Among the main efforts put into optimizing multilingual pre-trained models, Wu and Dredze (2020) presented Lite Transformers. In simple words, such models preserve the effectiveness of self-supervised cross-lingual transfer learning while reducing the computational cost drastically. Lite Transformers do this by optimization of model architecture, reduction of parameter redundancy, layer pruning, and other ways of parameter sharing. These models are designed with a smaller number of transformer layers, compact embedding space, and fitted out for quick training and low memory consumption. Wu and Dredze proved that Lite Transformers can work perfectly for many languages in a sentiment analysis task and at the same time be lightweight enough for use in resource-constrained environments.

Xia et al. (2021) extend this idea further by proposing model compression techniques for multilingual transformers, which guarantee much smaller size with hardly degraded performance. These compressed models are of particular use in mobile devices and other edge computing scenarios where efficiency becomes crucial. In ensuring that Lite Transformers are robust across diverse languages and tasks, some challenges must be overcome. More research is needed in underrepresented languages and specialized domains to ensure that the models perform equally well.

2. Knowledge Distillation

Another popular method for cross-lingual model optimization involves knowledge distillation. Knowledge distillation, pioneered by Hinton et al. (2015), has gained wide application in multilingual NLP. In this context, knowledge distillation usually means that a small "student" model is trained to mimic the performance of a much larger "teacher" model. Zhang et al. (2020) have already demonstrated its effectiveness for cross-lingual sentiment analysis tasks. The team adapted knowledge from very large multilingual models like XLM-RoBERTa into compact models, reducing computational requirements with significant savings while retaining competitive accuracies.

A major advantage of knowledge distillation lies in its capability for adaptation to low-resource languages by leveraging multilingual pretraining from the teacher model. For example, MobileBERT (Sun et al., 2020) is a resource-constrained setting student model distilled from BERT. The combination of compact architecture and teacher-student training has demonstrated great effectiveness in text classification tasks, particularly in low-resource settings. However, there still seems to be some trade-off in performance since the distilled models lack some fine-grained abilities compared to larger models.

Recent breakthroughs in iterative distillation (Jiao et al., 2020) aim to alleviate these challenges. Knowledge is distilled step by step, whereby student models will gradually get closer to teacher models in performance. Besides, Li et al. (2021) proposed multilingual-specific knowledge distillation strategies that align cross-lingual embeddings to enhance generalization across languages.

3. Combined Approaches and Future Directions

In addition to these, hybrid approaches which combine the Lite Transformers with knowledge distillation have been explored. For example, Wang et al. (2022) combined the approaches of parameter reduction with teacher-student training and attained both efficiency and performance gains on cross-lingual sentiment analysis tasks. These hybrid methods put together the advantages of both techniques, thus further increasing the applicability of multilingual pre-trained models in low-resource settings.

While the Lite Transformers and knowledge distillation represent significant steps toward optimizing multilingual models, challenges remain. Balancing model size and accuracy, ensuring robustness across diverse languages, and adapting to evolving data distributions are critical areas for future research. Besides, the development of evaluation benchmarks for low-resource settings will help standardize performance assessment and guide the design of next-generation multilingual models.

2.4 Application of Cross-lingual Transfer Learning in Sentiment Analysis

2.4.1 Methods

Cross-lingual transfer learning has gradually become the most important way to solve the scarcity of data in low-resource languages. These works follow in the wake of improving knowledge extracted from high-resource languages in rich linguistic contexts. Some of the key works are related, especially those focused on lightweight multilingual models that increase computational efficiency without losing the characteristics of high-performance capability.

1. Machine Translation-Based Methods

The most straightforward methods of sentiment analysis in low-resource languages involve translation of the text from the target language into a high-resource language, such as English, and then applying sentiment analysis models that have been trained on the high-resource language. Such methods leverage the robustness of pre-trained models in high-resource languages to overcome the lack of resources for training native models in low-resource setups. The lighter version of the MobileBERT-based models was very efficient in handling a large input dataset with considerable performance related to translated text for sentiment analysis by Chauhan et al. (2022).

Though simple and pragmatic, translation-based methods inherently cannot rise beyond the quality of machine translation systems utilized. Errors introduced during translation—such as misinterpretations of idiomatic expressions, culturally specific phrases, or language structures—can add noise and impact the downstream sentiment classification task (Zhang et al., 2021). For example, studies by Artetxe et al. (2020) and Wu et al. (2022) highlight that translations of complex sentiment-laden texts often fail to capture subtle emotional tones or contextual meanings, leading to inaccurate sentiment predictions.

However, some of these challenges have been greatly overcome by the improvement of machine translation systems. For example, neural machine translation models-based on transformer-based architectures among others (Vaswani et al., 2017)-have widely improved the quality of translation by capturing long-range dependencies and semantic structures. Recent advancements in multilingual NMT systems, such as mBART (Liu et al., 2020) and mT5 (Xue et al., 2021), also allowed better generalization across languages and therefore improve sentiment analysis results even for very underrepresented languages. However, even state-of-the-art NMT systems can perform poorly with low-resource languages, where there is a general scarcity of parallel corpora to train, as pointed out by Goyal et al. (2022).

A recent approach to solving this challenge is tuning translation models along with sentiment analysis tasks. In multi-task learning, the translation system is optimized not only for accurate translation but also towards downstream objectives of sentiment analysis, hence yielding more task-specific outputs (Pires et al., 2022). Other approaches, combining machine translation with domain-specific linguistic rules, have also been promising in reducing translation noise for sentiment-specific expressions (Chen et al., 2021).

Most recently, the combination of powerful pre-trained sentiment models with better machine translation systems has made competitive performance in resource-poor settings possible. In particular, fine-tuned light-weight models like MobileBERT (Sun et al., 2020) and DistilBERT (Sanh et al., 2019) have been demonstrated to be effective in real-time applications where computational power is low for deploying sentiment analysis pipelines. These models can process large-scale data with pretty high efficiency and accuracy if combined with quality translations.

Translation-based approaches have indeed been compared to direct cross-lingual transfers. Indeed, Laubli et al. (2019) talk about how, although the translation methods offer an easy-to-follow pipeline, another line of research that involves direct cross-lingual models, like those based on XLM-R (Conneau et al., 2020), has its benefits from not necessarily needing the translation stage. These models instead rely on shared multilingual embeddings to perform the sentiment analysis directly on the target language. While such methods are not subject to translation errors, they do presuppose large-scale pre-training on multilingual corpora, something that is practically unattainable in a really low-resourced language setup.

Ultimately, machine translation-based methods remain one of the most popular and pragmatic choices to handle the sentiment analysis challenge in low-resource settings. Their efficacy depends on the advancement of translation quality, the robustness of downstream models, and the incorporation of task-specific optimizations. Future research can be channeled into reducing translation noise and investigating hybrid pipelines that bring efficiency and accuracy together, realizing their full potential in multilingual sentiment analysis applications.

2. Multilingual Embeddings and Transfer Learning

Recent breakthroughs in cross-lingual sentiment analysis have been driven by multilingual embeddings such as XLM-R Tiny and DistilBERT. These models learn shared vector representations of words from multiple languages by aligning semantically similar words across languages. This shared embedding space enables the transfer of sentiment analysis knowledge from high-resource languages to low-resource ones, circumventing the need for large annotated datasets in the target language.

XLM-R Tiny is an extremely lightweight version of the original XLM-R, proposed by Conneau et al. (2020), optimized for resource-constrained environments. In practice, as Liu et al. (2020) point out, XLM-R Tiny can achieve a very good balance between computation efficiency and model performance. This makes the optimization particularly apt for tasks that occur in computationally restricted environments, like mobile devices or low-power servers. It also performs well on language pairs that are linguistically similar, such as Spanish and Portuguese, where shared morphological and syntactic features allow the model to transfer knowledge effectively (Ruder et al., 2021).

However, it is more difficult to adapt these models to linguistically distant languages because structural and semantic differences impede effective knowledge transfer. For example, Finnish and Mandarin differ quite a lot in syntax and morphology from English and often show worse sentiment analysis results by several per cent (Artetxe et al., 2021). In these cases, fine-tuning multilingual embeddings on task-specific data may enhance their generalization. Lightweight models like DistilBERT, when trained on large-scale multilingual corpora, have shown significant improvements in sentiment classification for linguistically distant language pairs, such as English and Hindi (Sanh et al., 2020).

Recently, some authors explored how to make MUSE more robust with adversarial training or even alignment-based learning. In general, adversarial approaches aim at training multilingual embeddings by reducing the gap between different language representations; therefore, increasing their alignments (Chen et al., 2021). Among the alignment-based approaches, techniques such as those implemented by Lample et al. (2018) have made use of parallel corpora in a fine-tuning process where semantically comparable words of a pair of languages would come close in space.

Another recent development is the introduction of LAFT. LAFT fine-tunes multilingual models on small amounts of monolingual data before task-specific training and manages to significantly improve their performance on low-resource languages (Pfeiffer et al., 2020). This has proven especially successful for the tasks of sentiment analysis in underrepresented languages such as Swahili and Amharic, as Agrawal et al. (2022) have shown.

Although promising solutions have come from multilingual embeddings, their success usually relies on the availability of pre-trained multilingual corpora. So far, models such as mBERT (Devlin et al., 2019) and XLM-R have been trained on datasets containing more than 100 languages, but these corpora are still imbalanced toward high-resource languages. Low-resource languages usually remain underrepresented in the embedding space, and their performance remains suboptimal

(Kumar et al., 2021). This imbalance can be addressed by augmenting multilingual corpora with additional data from low-resource languages, as suggested by Conneau et al. (2020).

Further, several hybrid approaches that combine multilingual embeddings with rule-based sentiment analysis systems have also reported promising results. For example, Wu et al. (2022) proposed a hybrid framework that coupled DistilBERT embeddings with manually created sentiment lexicons for domain-specific sentiment analysis. This approach leverages the much-improved contextual understanding brought forth by multilingual embeddings and embodies language-specific nuances, hence leading to improved performance on complex sentiment tasks.

The integration of cross-lingual transfer learning with domain-specific pre-training is another area of active research. Models such as mT5 (Xue et al., 2021) and mBART (Liu et al., 2020) have been extended to include domain-specific knowledge, enabling them to perform sentiment analysis with higher accuracy in specialized contexts, such as financial sentiment or healthcare sentiment analysis (Chen et al., 2023). Due to the combination of general and domain-specific knowledge, such models could efficiently cope with the issue of limited labeled data in a target language.

Lightweight and hardware-efficient multilingual models, such as TinyBERT by Jiao et al. (2020) and MobileBERT by Sun et al. (2020), are also finding their applications in real-time sentiment analysis. These prove to be of particular help when one needs to do on-device processing-for example, in mobile phone applications or chatbots. The low computational overhead and very good performance make these models quite suitable for easy deployment in low-resource settings.

3. Zero-Shot Learning

Zero-shot learning is an exceptionally powerful paradigm where models predict performance in new languages without any need for labeled training data in target languages. Indeed, this would constitute an exceptionally novel approach: taking the pre-training of multilingual models like DistilBERT and XLM-R Tiny on large multilingual corpora to easily generalize across languages. These are thus able to perform sentiment classification in completely unseen languages based on common syntactic and semantic patterns in their pre-training.

Xu et al. (2021) showed the effectiveness of zero-shot learning in Swahili sentiment analysis with DistilBERT. The model, pre-trained on a rich multilingual corpus, achieved competitive results without task-specific fine-tuning. This shows how common linguistic features at the level of word order or shared roots are represented in pre-trained embeddings and reused for downstream tasks. Conneau et al. (2020) investigated the zero-shot performance of XLM-R Tiny for a wide range of African languages and achieved very good performance across a wide linguistic variety. Their most outstanding characteristics included being computationally efficient and broadly applicable; as such, it could be used on mobile devices or in other areas with constrained resources.

Despite the promise, zero-shot learning faces challenges, particularly when applied to linguistically distant languages. For instance, indigenous languages with unique grammatical structures or phonetic systems, such as Quechua or Nahuatl, often fall outside the pre-training corpus of popular multilingual models. As a result, models like MobileBERT, which excel in general-purpose tasks, tend to underperform on such languages (Liu et al., 2020). The lack of sufficient representation in pre-training data indicates the need for more inclusive multilingual corpora.

These challenges have, in turn, encouraged a number of improvements in the ZSL paradigm. One such approach involves clustering languages with similar linguistic properties, either morphological or syntactic, during pre-training. Artetxe et al. (2021) demonstrated that such language clustering improves cross-lingual transfer in general but mostly for underrepresented languages. For instance, clustering the Bantu languages during pre-training improved the performance in zero-shot settings for related low-resource languages such as Zulu and Xhosa.

While second, adaptive pre-training methods, such as in Pfeiffer et al. (2020), aim at augmenting multilingual models with a small quantity of target language unlabeled text. This enriches the model with representations of unique linguistic features in the target language without requiring any labeled data. For example, Agrawal et al. (2022) showed that this adaptive pre-training improved the ZSL performance at the Dravidian languages Tamil and Kannada, which are structurally very different from the Indo-European ones.

Prompt-based learning has also been a very effective approach for improving zero-shot performance. In the case of the prompt-based methods, this alignment between pre-training objectives and downstream tasks is much stronger, since these reformulate the task at hand as cloze-style problems. Indeed, Schick et al. (2021) explored an application to sentiment analysis; the authors found that this greatly improves zero-shot performance on DistilBERT and XLM-R Tiny. For example, the sentiment classification might be cast in a form like: "The sentiment of this sentence is [MASK], " where a model predicts the masked word.

Other novel ideas include unsupervised machine translation-based data augmentation. That would involve translating high-resource language-unlabeled data into the target language to act like extra input for fine-tuning and evaluation. For example, Wu et al. (2022) combined machine translation with XLM-R Tiny in developing synthetic labeled datasets targeting low-resource languages, yielding state-of-the-art results on zero-shot sentiment analysis for languages such as Khmer and Lao.

Knowledge distillation for zero-shot learning has also aimed at developing lighter, efficient models. Jiao et al., in 2020, derived TinyBERT, which is the distilled version of BERT, optimized especially to perform the ZSL tasks in constrained resource scenarios. TinyBERT can preserve its performance by shrinking the computational requirements through the transfer of knowledge from larger teacher models; this has been effective, particularly on-device applications ranging from mobile apps to sentiment analysis and conversational agents.

While zero-shot learning reduces the need for labeled data in the target language, its effectiveness is highly dependent on the quality and diversity of the pre-training corpus. For example, Conneau et al. (2020) observed that XLM-R's performance on ZSL tasks was highest for languages that were well-represented in the CommonCrawl dataset but decreased for languages with limited or noisy data. This observation underlines the importance of building balanced multilingual corpora representing low-resource and minority languages.

Lastly, zero-shot learning develops with the evaluation methodologies. The traditional benchmarks of GLUE and SuperGLUE are good for the English language but do not reflect the challenge of ZSL in a low-resource language. Various alternative benchmarks have been suggested in this regard: XTREME (Hu et al., 2020) is a benchmark for multilingual models across a wide variety of languages and tasks. Employing such benchmarks that actually reflect linguistic reality helps researchers more effectively probe and enhance the generalization of zero-shot models.

Zero-shot learning, in return, has shown the most promise as a front for cross-lingual sentiment analysis, which allows scalability on issues of resource scarcity. Some avenues that future work would take include multilingual corpus enrichment, more effective pre-training methods, and further specialized studies of models for low-resource languages. Such efforts will equally extend the usability of the ZSL method in real scenarios: from monitoring social media posts to analyzing customers' feedback in resource-constrained settings.

2.5 Related Work

2.5.1 Sentiment Analysis In Malay

Sentiment analysis is also termed "opinion mining." It is a basic task in NLP, which aims at extracting polarity, which may be positive, negative, or neutral sentiments, or the emotional tone underlying the texts. Applications involve a wide range of fields, from customer review analysis and brand reputation monitoring to the measurement of public opinions on social, economic, and political issues.

1. Early Approaches in Malay Sentiment Analysis

Early systems of sentiment analysis were developed using the rule-based and lexicon-based approaches. These early efforts within the Malay language include using a sentiment lexicon such as SentiBahasa, developed in 2014. The sentiment scores were pre-defined in the lexicon for Malay words. Applications of SentiBahasa have been reported for tweet analysis and product review analysis among others (Ahmad et al., 2014). For example, SentiBahasa assigns the words "baik" (good) or "buruk" (bad) the scores of + or-, respectively.

However, such methods often faced difficulties in catching linguistic complexities, such as sarcasm or negation. For example, "tidak terlalu buruk" (not too bad) may have a positive meaning, but typically the rule-based method will not be able to catch this without more syntactic or contextual analysis.

Furthermore, there were a lot of dialects in which Malay was spoken, including Kelantanese and Sarawakian Malay. As discussed by Aziz et al. (2015), the variation of vocabulary and syntax across dialects reduces the generalizability of the lexicon-based approaches. These pioneering works, however, gave a launching pad for more sophisticated approaches, especially putting forward the urgent need for increased contextual understanding in Malay Sentiment Analysis.

2. The Shift to Machine Learning

Machine learning's arrival significantly developed the area of sentiment analysis. The supervised learning algorithms, like Naive Bayes, Support Vector Machines, and Logistic Regression, became cornerstones in sentiment classification tasks. Researchers such as Hamzah et al. (2016) applied these methods to Malay datasets, including movie and product reviews, achieving higher accuracy compared to lexicon-based approaches. For instance, Hamzah et al. reported 82% accuracy using SVM on a manually labeled dataset of 10,000 Malay sentences.

Common among them are n-grams, Part-of-Speech, and TF-IDF features. For instance, Mustafa et al. (2017) showed that the inclusion of Part-of-Speech tags improves the classification results by structuring the syntactic dependencies, such as adjective-noun word pairs, into the Malay text. A major shortcoming of most these approaches is that they require substantial amounts of labeled data from within the domain of interest. For instance, most movie review-trained models failed miserably when tested on other domains like political tweets or news articles. This again pointed out the inadequacy of good-quality, publicly available Malay sentiment datasets.

3. The Deep Learning Revolution

Deep learning revolutionized the field of sentiment analysis as models became able to learn complicated patterns of texts. The most adopted models include RNN and LSTM. For example, the paper by Zainuddin et al. (2018) discussed the effectiveness of using LSTMs to capture long dependencies in Malay sentences. For example, their model gave 90% accuracy on sentiment classification tasks with complex sentences such as "saya sangat kecewa semalam, tetapi kini saya berpuas hati" (I was very disappointed yesterday, but now I am satisfied).

This was further advanced by the introduction of the Transformer-based models, such as BERT by Devlin et al. (2019) and XLM-R by Conneau et al. (2020). In this regard, multilingual models such as mBERT and XLM-R were first pre-trained on diverse corpora before fine-tuning on Malay sentiment tasks. For instance, Rosli et al. (2021) show that fine-tuning XLM-R on 50,000 Malay social media posts achieved an F1 score of 93% for sentiment classification. These models did an exemplary job in handling contextual nuances such as "bagus sangat" (very good) which, according to the context, can change their sentiment.

Moreover, the Transformer models have supported multitask learning whereby the same model could be used on related tasks like emotion detection, opinion mining, and stance classification. For example, a study by Tan et al. (2022), in that multitask learning improved the performance by an average of 5% in low-resource languages like Malay.

4. Fine-Grained Sentiment Analysis

The nature of sentiment analysis research has lately moved away from simple polarity classifications. Such fine-grained approaches in Malay would include the following.

Sentiment Intensity Analysis: Models like VADER were modified for Malay language research by Khalid et al. (2020); a few examples could be the degrees of sentiment classified, from mild positive to strong negative. Specifically, in social media commentaries, a distinction correctly belonged to "sangat bagus" ("very good") versus "bagus" ("good").

Emotion Detection: Emotion detection is the process of recognizing basic emotional states, such as happiness, anger, or fear. In the work of Ismail et al. (2021), XLM-R was fine-tuned on the EmoMalay dataset, achieving an accuracy of 88% in the identification of emotions in Malay tweets related to COVID-19.

ABSA is one that deals with aspect-specific sentiments. In the case of product reviews, ABSA extracts sentiments related to quality, price, and delivery. Specific works, such as Rahman et al. (2021), conducted ABSA on Malay e-commerce reviews, obtaining an F1 score of 87% through BERT fine-tuning.

5. Domain Adaptation and Cross-Domain Sentiment Analysis

The domain adaptation challenge is still one of the crucial ones. The approaches of adversarial training and fine-tuning on the target domain have looked very promising. For example, Norazman et al. (2022) applied DANN to adapt a sentiment model, which was initially trained on movie reviews, to classify sentiments in political tweets. This approach increased accuracy by 15% compared to the baseline methods.

6. Sentiment Analysis for Low-Resource Languages

Cross-lingual transfer learning therefore benefits a low-resource language such as Malay. Representation shared across languages by models such as mBERT and XLM-R enhances their performance. A study by Abdullah et al. (2022) showed that fine-tuning mBERT on an English-Malay parallel corpus requires no need for large labeled datasets but yielded an F1 score of 91%.

7. Challenges and Future Directions

Sarcasm and Irony: Sarcasm is hard to detect because it usually contains subtle linguistic cues. Works such as Zhang et al. (2021) propose combining contextual embeddings with sentiment lexicons for better detection.

Code-Switching: Most code-switching between Malay and English occurs in Malaysia. The potential of the multilingual embedding techniques, fine-tuned on code-switched corpora, has been realized. For example, Ramli et al. (2023) have proposed a hybrid model that attained an accuracy of 85% on code-switched datasets.

Bias and Fairness: Most of the models are biased on gender or ethnicity. Kiritchenko & Mohammad (2018) discuss how essential model auditing is to uncover bias and how different mitigation techniques should be used.

Explanability: Explanation of model predictions is key in any real-world application. Various techniques such as SHAP and LIME are under research for enhancing the interpretability of Malay sentiment models.

Malay sentiment analysis is a field in its evolutionary stage. With the development of deep learning and NLP, one will be enabled to overcome these challenges and unleash the full potential of sentiment analysis for low-resource languages like Malay.

2.5.2 Transfer Learning

This really has reshaped the landscape of NLP, whereby pre-trained models' knowledge was used successfully to solve downstream tasks in a very effective manner, especially when the data was not so large. Traditional machine learning approaches need one model and lots of labeled data for every single task, which is really not efficient and just doesn't scale. Transfer learning, however, bridges this gap by pretraining models on very large-scale datasets, such as Wikipedia, Common Crawl, and BooksCorpus, then fine-tuning for a particular application. Howard & Ruder, 2018; Devlin et al., 2019 Transfer Learning Foundations.

Pretrained language models have become the cornerstone of transfer learning. Models such as BERT by Devlin et al. (2019), GPT by Radford et al. (2018), and T5 by Raffel et al. (2020) are illustrative of its transformative power. These models are trained on vast amounts of text data to learn linguistic, syntactic, and semantic patterns that can then be transferred to a wide range of tasks.

For instance, BERT uses a masked language modeling task that masks random words and predicts their context, combined with a next sentence prediction task. These pretraining objectives enable BERT to create bidirectional contextual embeddings that work very well in tasks such as sentiment analysis, named entity recognition, and question answering. Fine-tuning BERT on specific tasks requires only small amounts of labeled data, making it particularly advantageous for domains with limited data (Sun et al., 2019).

Design-wise, GPT was based on the autoregressive language model and predicted the next token given the previous ones. While one-directional originally with respect to generating text, this was then extended into its later versions-GPT-2 and GPT-3-their few-shot learning expanded by large-scale datasets with enlarged parameter sizes (Brown et al., 2020). In this way, the studies showed how well transfer learning would scale and provide excellence in lots of summarization, translation, and conversational AI without much finetuning.

While T5 proposed a unified text-to-text frame, all NLP tasks were framed as text generation problems: classification, translation, and summarization. This innovative approach simplified task-specific adaptation and reinforced the flexibility of transfer learning models (Raffel et al., 2020).

1. Fine-Tuning in Transfer Learning

Fine-tuning in transfer learning generally refers to adapting pre-trained models to any specific downstream task. Unlike these pre-trained traditional models on a very large corpus, fine-tuning updates a subset of model parameters with the small labeled dataset of the target domain. In fact, it is one of the most efficient processes in which a minimum computational overhead is wasted to retain knowledge regarding linguistics during pretraining.

For sentiment analysis, fine-tuning has consistently yielded state-of-the-art results. By adapting pretrained models such as BERT or DistilBERT to domain-specific datasets, researchers have achieved higher accuracy and robustness compared to traditional supervised approaches (Mozafari et al., 2020). Fine-tuning also enables models to handle nuanced tasks like sarcasm detection or aspect-based sentiment analysis, where understanding contextual dependencies is critical.

2. Cross-Lingual Transfer Learning

This has further been extended to multilingual and cross-lingual, whereby the challenges because of diversity in languages are put to task. These models, such as DistilBERT (Sanh et al., 2019), when pre-trained on multilingual corpora and aligned in their embeddings across different languages, allow transfer across languages. In this way, zero-shot or few-shot learning becomes possible whereby the models which are trained on high-resource languages are directly applied to low-resource languages.

Other recent works, for example, are on the more robust multilingual variant of BERT, XLM-R, which has illustrated very good performance across different cross-lingual tasks. XLM-R was pre-trained on 100 languages using a masked language modeling objective, thus allowing its applications such as cross-lingual sentiment analysis and machine translation since it generalizes well across languages. Conneau et al. (2020) A different approach would be DistilBERT, lighter and faster BERT, but equally as efficient in handling multilingual tasks while retaining the same level of accuracy as the others. Sanh et al. (2019).

Challenges and Future Directions Successful though it has been, transfer learning still faces a number of challenges. When models are adapted for resource-poor languages, the adaptation often involves domain-specific knowledge and strategies regarding linguistic nuances and cultural differences. Sentiment analysis is one such area that depends much on context-dependent expressions and idiomatic phrases, which can be wide apart between different languages and regions (Chen et al., 2020).

These challenges were tackled in various works concerning domain adaptation, few-shot learning, and parameter-efficient fine-tuning using adapters, Pfeiffer et al. (2020). Among the other promising directions of transfer learning in multilingual NLP, one can distinguish the development of prompt-based learning and continual learning.

In a nutshell, transfer learning has grown as a paradigm in NLP, providing scalable, effective, and robust solutions for a wide range of tasks. The pretraining models set benchmarks for representation, while fine-tuning strategies increased their applicability across domains and languages. Cross-lingual transfer refinement, efficiency enhancement, and bias reduction remain some aspects where future research is needed to exploit the full potential of transfer learning.

2.5.3 Multilingual Processing

Multilingual processing has grown as a cornerstone in natural language processing, especially in tasks that involve sentiment analysis, since it requires models able to operate on digital content that is increasingly becoming globalized. At the core of multilingual processing lies the development of techniques and models that have the capacity to handle diversity among languages at syntax, semantics, morphology, and expressions of cultural sentiment. This section describes some of the major developments in multilingual processing, with a focus on sentiment analysis and cross-lingual applications.

However, this very same task, multilingual sentiment analysis, recently achieved amazing success with a more modern pre-trained multilingual model. The multilingual language models learned from large corpora in several languages share representation across many of those languages. These can be considered as embeddings from high to low-resource languages. For instance, fine-tuning XLM-R on sentiment datasets in the English language has made several successful cross-lingual transfers onto the test datasets for languages in which only very limited amounts of labeled data were available for Conneau et al. 2020.

It has also played a critical role in multilingual processing of multilingual word embeddings. Techniques like word embedding alignment (Mikolov et al., 2013; Ruder et al., 2019) project the embeddings of different languages into one unified vector space, thus enabling cross-lingual tasks without parallel corpora. Further refinements of those embeddings, such as unsupervised machine translation and self-supervised learning, ensure robust multilingual representation learning.

However, with such progress being witnessed, not all challenges regarding cultural and linguistic subtlety are overcome. The expression of sentiment largely differs among different languages and cultures, and several words or phrases could mean extreme contrasts in their meanings under different contexts. Chen et al. (2020) depicted these difficulties and mentioned that not considering this aspect of cultural differences may give rise to bias in a model's performance. The idiomatic and colloquially expressed sentiments generally give a headache to the models, which rely on literal translations.

In the case of the aforementioned issues, domain-specific lexicons and sentiment-aware fine-tuning strategies have been developed. Mozafari et al. (2020) proposed domain-adaptive fine-tuning techniques that adapt pretrained multilingual models to specific domains, such as movie reviews or social media sentiment. Zero-shot and few-shot learning approaches have also been utilized to overcome data scarcity in low-resource languages (Pires et al., 2019).

Another interesting area is the use of adapters in conjunction with mechanisms for efficient parameter transfer. Pfeiffer et al. (2020) introduce multilingual adapters that efficiently fine-tune large models on new tasks and languages without updating the base model parameters. These approaches have shown remarkable scalability and adaptiveness in multilingual NLP tasks.

Sentiment analysis now creates a whole new dimension with pre-trained multilingual models and embeddings that are improved. Cultural nuances, linguistics variability, and low-resource languages remain active research areas. Such gaps can only be filled through innovative solutions which blend linguistic insights, domain adaptation, and scalable model architectures.

2.5.4 Low-resource Language Processing

Most of the low-resource languages are spoken by linguistic minorities or in areas where there is limited technological infrastructure. This makes NLP a tough challenge. Such languages are not endowed with annotated data, standard tools, and linguistics resources, prerequisites for developing robust models. Besides, many low-resource languages are morphologically richer and orthographically nonstandardized, having dialectal variation, which thus requires a different approach to be processed compared to well-studied languages (Joshi et al., 2020).

1. Challenges in Low-Resource Language Processing

One of the main challenges in low-resource language processing is the lack of labeled data. Unlike high-resource languages like English, where large corpora and annotated datasets exist, low-resource languages have no parallel or monolingual corpora, lexicons, and pre-trained embeddings. This scarcity limits the training of language models and downstream applications such as sentiment analysis, machine translation, and information retrieval Blasi et al. (2022).

Another specific challenge is that of linguistic diversity: low-resource languages often represent agglutinative or polysynthetic languages, where in one word, rich ideas can be delivered with the help of morphological inflections. Examples are

Quechua, Basque, and Inuktitut. These languages require specialized tokenization and representation strategies which effectively capture their rich linguistic features (Anastasopoulos & Neubig, 2019).The Role of Transfer Learning.

Transfer learning has recently cropped up to turn the tide in low-resource language processing. These multilingual models have achieved very good performance, transferring knowledge acquired from high-resource languages to low-resource ones. This is possible due to shared representations between different languages learned during training by large multilingual corpora or few-shot-zero-shot learning with respect to the low-resource languages.

For instance, XLM-R, pre-trained on 100 languages, achieved surprising performance on cross-lingual sentiment analysis and question answering even for low-resource languages. Fine-tuning XLM-R on small datasets leads to massive task-specific improvements, even when using very small amounts of labeled data (Hu et al., 2020). Similarly, multilingual models are used in zero-shot machine translation, where a model trained on high-resource language pairs is capable of translating between low-resource languages without explicit training data for those pairs (Artetxe et al., 2019).

2. Data Augmentation and Synthetic Data Generation

Some of the data augmentation techniques go a long way in mitigating this shortage of labeled data for the low-resource languages. Back-translation has widely been in use in enhancing the performance of machine translation and other NLP tasks by generating synthetic parallel data from target language translations to source languages and vice-versa (Edunov et al., 2018).

Another promising approach is cross-lingual embeddings. These embeddings map words from multiple languages into a shared vector space, which allows for the use of annotated datasets of high-resource languages to bootstrap NLP tasks for low-resource languages. For example, researchers have used bilingual dictionaries and cross-lingual embeddings to train sentiment analysis models for low-resource languages (Liang et al., 2020).

3. Semi-supervised and Unsupervised Learning

Generally, semi-supupervised and unsupervised learning methods have contributed much to low-resource language processing. For instance, pretraining models like mBERT and XLM on unsupervised objectives, such as masked language modeling, allowed them to learn rich representations from unlabeled corpora and thus fine-tune on small labeled datasets for specific tasks.

Similarly, unsupervised machine translation has also been pursued and put into practice. It aligns corpora without parallel data. Among the methods are unsupervised neural machine translation, which, with the use of monolingual corpora and iterative back-translation, synthesizes parallel data, thereby bridging the resource gap for low-resource languages (Lample et al., 2018).

Active Learning and Human-in-the-Loop Strategies

It has also been studied as one of the most effective approaches in order to squeeze the utmost utility out of the very small labeled data under a low-resource scenario. Undoubtedly, spotting the most informative samples to label would improve both model performances and cost reduction over annotation. Human-in-the-loop approaches that combine machine learning and human expertise, in particular for low-resource languages, allow linguists and native speakers to provide high value annotations on specialized tasks beyond what state-of-the-art systems can do in Bojar et al. 2021.

4. Future Directions

Where much has been done, there is more yet to be overcome. Most multilingual models are biased toward the disadvantage of low-resource languages, as high-resource languages dominate during training. Future work should aim at balanced multilingual dataset creation by the elimination of biases to ensure fairness in performance across languages. Rust et al., 2021.

Apart from this, community-driven efforts toward the advancement of NLP research for the African languages, such as Masakhane, point out that collaboration is key to improving low-resource language processingAbbott et al. (2021). All these further elaborate on the need for inclusive datasets, culturally aware NLP tools, and teamwork with linguistic communities for innovation on such aspects.

Fundamentally, low-resource language processing is a big area in NLP, having great implications for the ambitions of global inclusivity and access to technology. These three together-transfer learning, data augmentation, and collaboration-help the researchers bridge this linguistic divide and extend the benefit of NLP to underserved communities worldwide.

2.5.5 Fine-Tuning

Currently, fine-tuning is the mainstream paradigm for adapting large-scale PLMs, including BERT, DistilBERT, and XLM-R, to downstream tasks. This naturally follows from the transfer learning paradigm: first train the models on a large corpus to learn general representations of the language and then fine-tune on the dataset of the specific task to optimize the performance. Fine-tuning has already shown very promising results on sentiment analysis and, more precisely, cross-lingual sentiment analysis for languages in which only a few resources are available. Howard and Ruder (2018) have also presented another approach called Universal Language Model Fine-tuning. They note that fine-tuning is very important to be performed in a row:

language model pretraining, task-specific pretraining, and fine-tuning. The ease with which their findings showed FT could be adapted to several tasks paved the way for the use of FT in sentiment analysis. Similarly, multilingual models such as mBERT and XLM-R -Conneau et al. (2020)-have further extended the use of FT to cross-lingual tasks. In these models, different languages share multilingual embeddings, enabling fine-tuning on high-resource languages to improve low-resource languages.

Several works leveraged the effectiveness of FT in cross-lingual sentiment analysis. A study by Xu and Yang (2020) explored multilingual performance on the tasks of mBERT and XLM-R for sentiment analysis. It turned out that fine-tuning these models with English datasets significantly improved their performance in other languages, even without any labeled data available in the target language. This again corroborated what was said by Pires et al. (2019), who attributed this to the cross-lingual transfer capabilities of PLMs in their ability to learn language-agnostic representations.

Moreover, some researchers explored domain adaptation in FT. For example, Mozafari et al. (2020) investigated the fine-tuning of BERT on domain-specific sentiment datasets, such as those from movie reviews and product reviews. From their findings, they concluded that fine-tuning should be done with great care in order to handle domain shifts because a model pre-trained on one domain may degrade in performance upon application to another domain.

Recent years have also witnessed the emergence of lightweight fine-tuning approaches that alleviate some of the computational challenges. Indeed, Houlsby et al. (2019) proposed Adapter modules which add small trainable layers to frozen PLMs and thus drastically reduce the memory and computation required. Pfeiffer et al. (2020) further generalized this method for multilingual tasks and demonstrated that language adapters fine-tuned on English can be successfully transferred to low-resource languages.

In addition to these strengths, fine-tuning has several shortcomings, including overfitting into small datasets and catastrophic forgetting of knowledge acquired during pretraining. To deal with this challenge, Zhao et al. (2021) introduced contrastive fine-tuning that injects data augmentation and contrastive loss during training. These methods can provide better generalization and robustness, particularly on the imbalanced dataset of low-resource languages.

Fine-tuning, in other words, has taken cross-lingual sentiment analysis one step further by developing powerful PLMs on tasks and languages. Though conventional FTs from yesteryears are performing rather well, their applicability is being innovatively enhanced by several methods: domain-specific tuning, lightweight adapters, among others, contrastive approaches to name a few. Indeed, these modifications hint at the beginning of growing importance for this approach under the multilingual natural language processing umbrella.

2.6 Summary and Research

Although the sentiment analysis with cross-lingual transfer learning is achieved with huge success, there is much left to reach for low-resource languages. Languages of this nature typically have a limited amount of available large-scale annotated data and generally create special problems in carrying out efficient sentiment analysis. Although significant improvements over the above challenges have indeed emerged through multilingual pre-trained models, machine translation-based methods, and multilingual word embeddings, most of the current techniques for dealing with this have their own drawbacks. Multilingually pre-trained models have worked commendably to achieve better cross-lingual knowledge transfers. Such models remain computationally expensive owing to the complexity of fine-tuning these huge models, which come with demanding resource requirements when applied in resource-limited deployment scenes. Apart from that, machine translation-based methods help through translation to high-resource languages; they also depend on the quality of the translation model. Poor translations can distort sentiments expressed in the target language, hence giving incorrect sentiment classification.

Another important challenge occurs when trying to apply the concept of multilingual embeddings, which map words of different languages to a common vector space. While that is a good way of aligning semantic features across languages, it may not be good enough to allow fine-grained SA in languages with complex linguistic structures or which express sentiments differently. It often fails in capturing subtleties of sentiment expressions in languages with diverse syntactic and semantic rules.

These challenges make the demand for more effective and efficient techniques to carry out cross-lingual sentiment analysis, especially in low-resource languages, paramount. This thesis, therefore, seeks to fill these gaps by investigating alternative methodologies that aim at optimizing the performance of sentiment analysis models in such challenging environments. This work will investigate to what extent multilingual pre-trained models can be optimized for use in low-resource settings without strong reliance on costly computational resources. Further, optimization techniques will also be discussed, including knowledge distillation and lightweight transformer models, with the intention of reducing computational cost for such models with limited degradation in performance across languages.

These explorations hopefully contribute to increasingly scalable and more efficient solutions for low-resource language sentiment analysis. It is hoped that the work will further extend the scope of languages that sentiment analysis can be effectively applied to, as well as advance NLP more generally in both multilingual and low-resource settings with the increase of effectiveness of techniques for cross-lingual transfer learning.

CHAPTER III

METHODOLOGY

This chapter presents the methodology employed in this research, focusing on the application of cross-lingual transfer learning for sentiment analysis in low-resource languages, with an emphasis on Malay. It outlines the research design, data collection and preparation processes—including dataset sources, preprocessing, and tokenization—along with the fine-tuning of multilingual models such as DistilBERT, XLM-R Tiny, and MobileBERT. Furthermore, it discusses the use of optimization techniques like model compression and knowledge distillation and details the evaluation metrics and procedures employed to ensure the reliability and validity of the findings.Figure3.1 shows the:

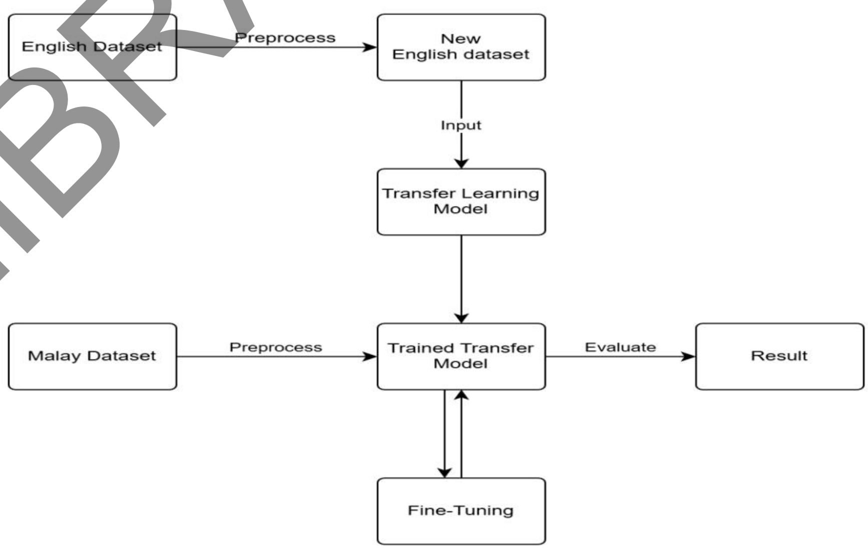


Figure 3.1 Transfer learning process

3.1 Dataset Overview

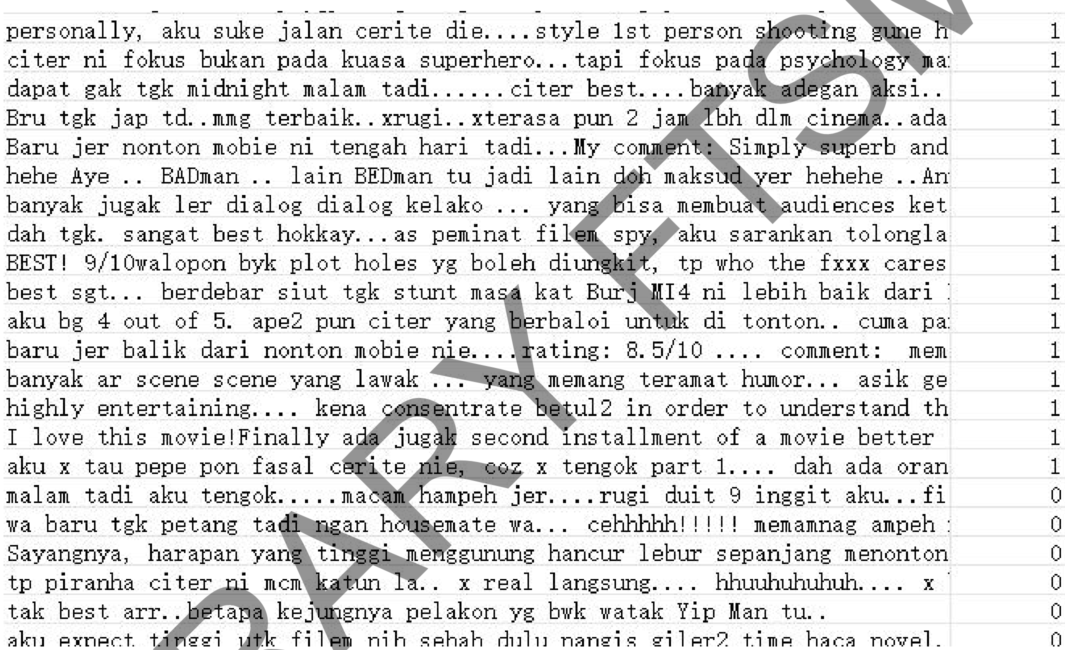
The experimental setup of this study focuses on the evaluation of cross-lingual transfer learning in sentiment analysis. Carefully selected datasets, both of high-resource and low-resource languages, have been used to make sure that the results will be meaningfully indicative of the methodology's effectiveness.

English Dataset (High-Resource Language): The English dataset was sourced from the widely used, publicly available movie review collections like the IMDB dataset. It consists of 25,000 reviews, equally distributed between positive and negative sentiments. This dataset was used to train the model in its entirety. Part of English datasets are shown in Figure 3.1:

Disney's Buena Vista Pictures presents a wonderfully told fact-based s	1
I was blown away by this film. I'm one of those people who just takes	1
This is first of all a good, exciting story, with well developed chara	1
I am not a golf fan by any means. On May 26 about 10:30 PM the movie s	1
THE GREATEST GAME EVER PLAYED (TGGEP, 2005) is an amazingly uplifting,	1
The focus of the key relationship in a young man's life, that of his r	1
I thought that this movie was incredible. I absolutely loved it, even	1
With its ww2 timing, falling in and out of love, and easy on the eye K	1
Oh my god what a story! This movie is very good and it had to be God w	1
At 20 years old, Francis Ouimet (Shia Lebouf) as his whole life ahead	1
I will divide my review into following 5 categories each accounting a	1
if you watch this at home on DVD or Bluray! be sure you have great sou	1
I'm not a sports fan - but I love sports flics! So, why ... what is a	1
As you know "The Greatest Game Ever Played" is about golf. I used to s	1
This movie gives golf a high mark, it was well acted and well directed	1
Wow, here is another great golf movie. That's at least three in the pa	1
I was extraordinarily impressed by this film. It's one of the best spo	1
Although I'm not a golf fan, I attended a sneak preview of this movie	1
From the start of "The Edge Of Love", the viewer is transported to the	1
This movie, with all its complexity and subtlety, makes for one of the	1
I've seen this story before but my kids haven't. Boy with troubled pas	1
Once again Mr. Costner has dragged out a movie for far longer than nec	0
This is an example of why the majority of action films are the same. G	0
First of all I hate those moronic rappers, who could'nt act if they ha	0
Not even the Beatles could write songs everyone liked, and although Wa	0
Brass pictures (movies is not a fitting word for them) really are some	0

Figure 3.2 English Dataset.csv

Low-resource Language Dataset: Low-resource language dataset: This dataset, developed by Al-Saffar et al. (2018), contains 2,000 reviews and serves as a testbed to benchmark the performance of the best models, as seen from training in the high-resource setting of the English dataset. The results will also demonstrate how well sentiment analysis models can be cross-transferred from high-resource languages to low-resource ones without additional training or fine-tuning. Part of the Malay dataset is shown in Figure 3.1. The dataset was introduced in the study "Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm" .



personally, aku suke jalan cerite die....style 1st person shooting gune h	1
citer ni fokus bukan pada kuasa superhero...tapi fokus pada psychology ma	1
dapat gak tgk midnight malam tadi......citer best....banyak adegan aksi..	1
Bru tgk jap td..mmg terbaik..xrugi..xterasa pun 2 jam lbh dlm cinema..ada	1
Baru jer nonton mobie ni tengah hari tadi...My comment: Simply superb and	1
hehe Aye .. BADman .. lain BEDman tu jadi lain doh maksud yer hehehe ..An	1
banyak jugak ler dialog dialog kelako ... yang bisa membuat audiences ket	1
dah tgk. sangat best hokkay...as peminat filem spy, aku sarankan tolongla	1
BEST! 9/10walopon byk plot holes yg boleh diungkit, tp who the fxxx cares	1
best sgt... berdebar siut tgk stunt masa kat Burj MI4 ni lebih baik dari	1
aku bg 4 out of 5. ape2 pun citer yang berbaloi untuk di tonton.. cuma pa	1
baru jer balik dari nonton mobie nie....rating: 8.5/10 .... comment: mem	1
banyak ar scene scene yang lawak ... yang memang teramat humor... asik ge	1
highly entertaining.... kena consentrate betul2 in order to understand th	1
I love this movie!Finally ada jugak second installment of a movie better	1
aku x tau pepe pon fasal cerite nie, coz x tengok part 1.... dah ada oran	1
malam tadi aku tengok.....macam hampeh jer....rugi duit 9 inggit aku...fi	0
wa baru tgk petang tadi ngan housemate wa... cehhhhh!!!!! memamnag ampeh	0
Sayangnya, harapan yang tinggi menggunung hancur lebur sepanjang menonton	0
tp piranha citer ni mcm katun la.. x real langsung.... hhuuhuhuhuh.... x	0
tak best arr..betapa kejungnya pelakon yg bwk watak Yip Man tu..	0
aku expect tinggi utk filem nih sebab dulu nangis giler2 time baca novel.	0

Figure 3.3 Malay Dataset.csv

3.1.1 Data Preprocessing

Several preprocessing steps were done for the compatibility of both datasets with multilingual models, but also to increase the quality of the data:

Text Preprocessing: The text data for each review has been cleaned to remove unnecessary material such as special characters, hyperlinks, and extra whitespace. All text data is lowercased for consistency across the dataset to reduce token variations.

Tokenization: The tokenization process was done using tools specific to their languages for appropriate sentence structure handling. For instance, English reviews were tokenized by libraries such as spacy and NLTK, while Malay reviews utilize MALAYTEXT, catering to the special handling of the Malay language.

Label Encoding: Positive reviews were considered to be of 1 while the negative ones as 0, a method of simple label encoding, making life easy with respect to binary classification for both these datasets.

An example of a preprocessed dataset is shown in Figure 3.4:

```
English review data processing example:
originalEnglish:  I went and saw this movie last night after being coaxed to by a few friends of mine.
PreprocessedEnglish:  i went and saw this movie last night after being coaxed to by a few friends of mi
ToklenizedEnglish:  ['i', 'went', 'and', 'saw', 'this', 'movie', 'last', 'night', 'after', 'being', 'cc
Malay review data processing example:
originalMalay:  okla citer ni dua bintang daripada lima bintang,yusry memang improve..but then citer mc
PreprocessedMalay:  okla citer ni dua bintang daripada lima bintangyusry memang improvebut then citer m
ToklenizedMalay:  ['okla', 'citer', 'ni', 'dua', 'bintang', 'daripada', 'lima', 'bintangyusry', 'memang
```

Figure 3.4 Preprocessed dataset

The whole preprocessing pipeline was implemented in Python, using the libraries pandas and NumPy for general data manipulation and specialized NLP tools. This structured approach ensured that the datasets were clean, consistent, and ready for model training.The data preprocessing flow chart is shown in Table 3.1.

Table 3.1 Data Processing Flow

Layer	**Description**	**Purpose**
Input Layer	Input text data (source language)	Raw text data before preprocessing (reviews, sentences, etc.)
Preprocessing	Tokenization, Lowercasing, Removal of Special Characters	Clean the text data and prepare it for embedding representation

to be continued...

...continuation

Embedding Layer	Pre-trained multilingual embeddings (e.g., mBERT, XLM-R)	Map text to vector space using pretrained embeddings
Encoder Layer	Transformer-based encoder (e.g., BERT, XLM-R)	Encode the text into contextualized embeddings for the model to learn
Cross-Lingual Transfer Layer	Knowledge transfer mechanism (e.g., attention, shared representation)	Facilitate cross-lingual knowledge transfer from high-resource to low-resource languages
Classification Layer	Feed-forward neural network for classification	Classify the sentiment or task at hand (e.g., positive/negative sentiment classification)
Output Layer	Sentiment labels or probabilities	Provide the final sentiment classification (positive/negative) or other classification labels

3.2 Model

3.2.1 Model Fine-Tuning and Hyperparameter Search

In addition to the core fine-tuning procedure already outlined, a hyperparameter search was performed to identify the most optimal configurations for training the models.

Learning Rate: To find the optimum value that can balance the convergence speed and accuracy without leading the model to overfit, different learning rates were tried. A grid search was done to find the best range of values.

Batch Size: Several batches were experimented upon to reach the right configuration, which was to provide computational efficiency and stability for training.

Epochs: The number of epochs was tried so that the model does not get overfitted and can still capture the required pattern in the data.

3.2.2 Model Selection

Successful model selection was crucial to the experiment. For the best trade-off between efficiency of computation and performance, three state-of-the-art lightweight and multilingual models were selected in this paper because of their outstanding performances on cross-lingual tasks and real-world applications.

1. DistilBERT:

DistilBERT is a compact version of the original BERT model. This model is generated by knowledge distillation, a process whereby. It includes the training of a smaller "student" model to imitate the behavior of a much larger "teacher" model. All in all, DistilBERT retains 97% of the performance of the original BERT, is approximately 60% faster, and consumes a lot fewer resources. Its small size and good efficiency make this model perfect in cases where there is a necessity for faster inferences and saving memory. This ability of DistilBERT to handle subword embeddings lets it handle languages with rich morphology and perform well on cross-lingual sentiment analysis tasks in diverse linguistic settings.

2. XLM-R Tiny:

XLM-R Tiny is a lighter version of the XLM-RoBERTa model, optimized for multilingual tasks. It covers more than 100 languages and has been designed to do well in cross-lingual applications, especially for low-resource languages that are underrepresented in existing datasets. Similar to its larger variant, XLM-R, the XLM-R Tiny is trained on a very large and diverse multilingual corpus but much smaller in model size, increasing its suitability for deployment on computational-resource-limited environments. Despite its compact architecture, XLM-R Tiny holds much promise in the capture of multilingual semantic patterns and their generalization across languages, which makes it an ideal candidate for cross-lingual sentiment analysis tasks where efficiency and scalability are at stake.

3. MobileBERT:

MobileBERT is a variant of the original BERT model optimized for edge devices and mobile platforms. Knowledge distillation and bottleneck layers are the hallmarks of its architecture that allow for much better efficiency at almost no cost to accuracy. Specially, MobileBERT is designed to bridge high-performing models and resource-constrained environments, hence versatile in performing on-device sentiment analysis tasks. This is achieved through a trade-off between compactness and robust linguistic relationship understanding across many languages. Considering this aspect, MobileBERT is designed to be lightweight while powerful. Thus, it performs excellently during the deployment of multilingual SA systems in devices with low computation such as smartphones and IoT gadgets.

The architectures of DistilBERT, XLM-R Tiny, and MobileBERT are all based on the Transformer framework, dedicated to providing deep contextual relationships in texts with the use of self-attention mechanisms. These allow these models to easily excel in understanding the subtleties of multilingual text, which is what is essentially required for effective cross-lingual sentiment analysis.

The pre-trained models were adapted for sentiment classification by adding a dense classification head with a softmax activation function. This addition will transform the models into task-specific classifiers capable of classifying textual reviews into sentiment labels such as positive and negative. By nature, these lightweight models strike a great balance between performance and efficiency. Therefore, even when computational resources are low, their results will still be very good. Because of their nature, the enhanced architecture applied in the models assured that their performance would also support their cross-lingual adaptability. Consequently, they have turned out to be quite powerful in the practice of sentiment analysis in diverse languages.

Table 3.2 Transfer Learning Model Architecture

Step	Action	Purpose
1. Input Data (Raw Text)	Collect text data (reviews, comments, etc.) from high-resource and low-resource languages	Gather input data for the model.
2. Data Preprocessing	Tokenize text, lowercase, remove special characters, optional language detection	Prepare the raw text for embedding transformation.
3. Embedding	Apply word embeddings	Convert preprocessed text into vector representations that the model can understand.
4. Cross-Lingual Alignment	Align text from high and low-resource languages in a shared embedding space	Ensure the model can process both high and low-resource languages in the same vector space.
5. Model Input	Feed aligned text data into the model for training/evaluation	Supply the processed data into the model for training or testing.

3.3 Fine-Tuning Procedure

Fine-tuning on the English dataset allowed the pre-trained model to grasp a deeper understanding of sentiment classification in a high-resource language.

Fine-Tuning on Malay Dataset: The model fine-tuned on the English dataset was further fine-tuned using the Malay dataset. This step allowed the model to leverage its knowledge of English to improve performance on Malay sentiment analysis.

In this cross-lingual transfer learning approach, the objective was to exploit this large dataset in English so that the performance of the model on the Malay dataset would get better, requiring less large-scale annotated data in a low-resource language.

Step 1: Training on the English Dataset.The Objective: This is to develop a strong English sentiment classification capability, using English as a proxy high-resource language.The pre-trained DistilBERT model, distilbert-base-multilingual-cased, is loaded and has its classification head replaced by a dense layer for binary classification. The high_resource_data training set in English was preprocessed and tokenized to fit the model's input format. Training for 1 epoch with a learning rate of 2e-5 and a batch size of 8, with the aim of strong basic performance in sentiment analysis, was done.

Step 2: Evaluation on the Malay Dataset The objective here is to evaluate the model's performance on Malay, a low-resource language, using its multilingual capabilities.Process: The model was further tested on the preprocessed Malay dataset, low_resource_data, after it was trained on the English data. The Malay dataset was tokenized similarly to the English to make it compatible with the DistilBERT tokenizer. To assess the generalizability of the model and its adaptiveness to Malay, metrics such as accuracy, precision, recall, and F1 score were calculated. Unified Training Approach:

This implementation unifies the approach towards training and evaluating the model for efficiency in high-resource language training coupled with the ability to generalize to low-resource ones. Here, there is no fine-tuning on the Malay dataset separately.

1. Hyperparameter Settings

The following hyperparameters were utilized to optimize training and evaluation:

Learning Rate: Fixed at 2e-5 across both datasets to provide a good trade-off between convergence speed and overfitting.

Batch Size: Set to 8 for both training and evaluation, ensuring compatibility with the computational environment.

Number of Epochs: The training was done for 1 epoch so that a baseline sentiment analysis model is quickly established.

The optimizer is AdamW, using 0.01 weight decay which is generally favorable for model stability and convergence.

Gradient Accumulation: On with one step, as it will handle smaller batch sizes gradient_accumulation_steps=1.

2. Training Framework:

The Hugging Face Transformers library combined with PyTorch made the training and evaluation of this work easier, ensuring reproducibility, efficiency, and compatibility with state-of-the-art Transformer-based architectures.4.4 Cross-Lingual Transfer Learning WorkflowEvaluation Metrics:

To evaluate the performance of the models, the following metrics were used:

Accuracy: It is the ratio of correctly predicted instances to the total number of predictions.

Precision: It is the ratio of true positives to the sum of true positives and false positives.

Precision-Recall: The ratio of true positives within the actual positive instances;

F1-Score: This is the harmonic mean of precision and recall, hence balancing the tradeoff between these two. The above metrics are standard for classification tasks, especially in the case of sentiment analysis.