UNIVERSITI TEKNOLOGI MARA

MULTI-LEVEL TEXT DATA AUGMENTATION FOR MALAY INTENT CLASSIFICATION MODEL

ANIS SYAFIQAH BINTI MAT ZAILAN

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Computer Science)

Faculty of Computer and Mathematical Sciences

February 2026

ABSTRACT

Intent classification for Malay language queries remains a challenge due to limited annotated datasets, severe class imbalance, and rich morphological variation. This research introduces ATISMalay, a validated Malay-language dataset constructed by translating the ATIS benchmark using machine translation and refined through bilingual expert evaluation, with Cohen's Kappa confirming fair agreement. The dataset revealed structural limitations, prompting the development of ICDAMalay, an intent classification model based on BERT and BERT+CRF architectures. Performance comparisons and error analysis highlighted recurring misclassifications, especially in low-resource intent classes. To address these issues, a multi-level text data augmentation strategy was implemented during pre-processing, applied systematically at the character, word, and phrase levels. Eight augmented datasets were generated and evaluated using BLEU and BPRO metrics, with full-tiered augmentation improving accuracy by 95% over the benchmark and 9.99 over the ATISMalay baseline. The study's unique contribution is a systematic, tiered augmentation framework tailored for low-resource languages. This research supports practical applications in Malay-language chatbots, e-government platforms, and educational tools, where accurate intent classification is essential. Future work may explore multi-intent classification, cross-domain scalability, and integration with open-domain conversational systems to further advance Malay NLP.

ACKNOWLEDGEMENT

First and foremost, I am profoundly grateful to Allah the Almighty for granting me the strength, patience, and perseverance to embark on this Master journey and successfully complete this challenging yet fulfilling endeavour. This achievement would not have been possible without His endless guidance and blessings.

I would like to express my sincere appreciation to my esteemed supervisors, Dr. Noor Hasimah Ibrahim Teo and Assoc. Prof. Dr. Nur Atiqah Sia Abdullah, whose invaluable guidance, unwavering support, and insightful expertise have been instrumental throughout this research. Their encouragement and constructive feedback have shaped my academic growth, and I am truly honoured to have been under their mentorship.

Finally, my deepest gratitude goes to my beloved parents and my entire family, whose unconditional love, prayers, and sacrifices have been my greatest source of inspiration. Their unwavering belief in me has been the driving force behind my perseverance and success. This achievement is a testament to their endless support, and I am forever grateful for their presence in my life. Alhamdulillah.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Research Background

Question Answering System (QAS) comprises natural language texts or a pre-structured database which are used to automatically provide the right and accurate response to questions posed by humans in human natural language (Dao et al., 2021; Sweta et al., 2021; Akbari et al., 2023a). QAS is more capable and more efficient in answering user queries than most search engines such as Google, YouTube, Facebook, Amazon, and Microsoft Bing (Davies, 2021). While search engines return a list of related websites and documents, QAS directly detects, recognizes, and classifies the user's intent to provide accurate and relevant answers (Sweta et al., 2021; Tang et al., n.d.).

QAS is fundamentally driven by Intent Classification (IC) and Natural Language Processing (NLP), a core subfield of Artificial Intelligence (AI) that enables machines to understand and generate human language (Refai et al., 2023). Within NLP, intent classification is a critical function of Natural Language Understanding (NLU), focusing on interpreting user utterances to uncover their underlying goals (Zhang et al., 2018; Kavlakoglu, 2020). Intent detection and classification are essential in task-oriented dialogue systems and QAS (Alshahrani et al., 2022; Wu D. et al., 2020; Dao et al., 2021), as they determine how the system processes and responds to user queries.

However, while extensive research has been conducted for English-language intent classification with highly impressive results, Malay remains significantly underexplored. To date, only a few studies have targeted Malay-language queries (Nurnasran et al., 2019; Mustafa & Zakaria, 2024), leaving a critical gap in multilingual SLU (Spoken Language Understanding) research. This gap is largely due to limited research and resource scarcity in low-resource Malay intent classification, making it challenging to develop accurate and reliable models.

One of the foremost issues is the scarcity of publicly available Malay datasets for intent detection and classification. Despite Malay being ranked 71st among the world's most spoken languages with 19.2 million speakers (Eberhard et al., 2022), existing resources are extremely limited. Available datasets are often inaccessible,