TRANSFORMER-BASED SENTIMENT ANALYSIS CLASSIFICATION IN NATURAL LANGUAGE PROCESSING FOR BAHASA MELAYU

MOHD ASYRAF BIN ZULKALNAIN

MASTER OF SCIENCE IN ELECTRONIC ENGINEERING

2025

Faculty of Electronics and Computer Engineering

TRANSFORMER-BASED SENTIMENT ANALYSIS CLASSIFICATION IN NATURAL LANGUAGE PROCESSING FOR BAHASA MELAYU

Mohd Asyraf Bin Zulkalnain

Master of Science in Electronic Engineering

2025

TRANSFORMER-BASED SENTIMENT ANALYSIS CLASSIFICATION IN NATURAL LANGUAGE PROCESSING FOR BAHASA MELAYU

MOHD ASYRAF BIN ZULKALNAIN

A thesis submitted
in fulfillment of the requirements for the degree of
Master of Science

Faculty of Electronics & Computer Technology and Engineering

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2025

DECLARATION

I declare that this thesis entitled “Transformer-Based Sentiment Analysis Classification In Natural Language Processing For Bahasa Melayu” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:...
Name	: Mohd Asyraf Bin Zulkalnain
Date	: 08.02.2025

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Electronic Engineering

Signature :..

Supervisor Name : Assoc. Prof. Dr Syafeeza binti Ahmad Radzi

Date : 09.07.2024

DEDICATION

I dedicate this thesis to my beautiful parents, Zulkalnain bin Mohd Yussof and Norlela binti Mohd Nordin, and to my family, friends and colleagues. May Allah grant us good health and the strength in remembering Him.

ABSTRACT

Sentiment analysis in Bahasa Melayu leverages Natural Language Processing (NLP) to interpret opinions and emotional tone expressed in Malay texts. This research investigates the application of transformer-based deep learning models—Bidirectional Encoder Representations from Transformers (BERT), DistilBERT, BERT-multilingual, ALBERT, and BERT-CNN—for sentiment classification into positive, negative, and neutral categories. The study addresses challenges in Bahasa Melayu sentiment analysis, including limited annotated resources, linguistic nuances, and common mixed-language usage on platforms like social media.To train and evaluate the models, a large-scale Malay dataset (Malaya dataset) was used. Pretrained models from HuggingFace were fine-tuned using 10-fold cross-validation to improve generalization. Optimization methods such as data augmentation were also implemented. The evaluation considered not just accuracy but also precision, recall, F1 score, and computational efficiency. Among the models, BERT-CNN achieved the best performance, with 96.30% accuracy and consistently high scores across all sentiment classes. BERT also performed well, especially for neutral sentiment, reaching 89.5% accuracy but showed slightly lower recall in the positive class. DistilBERT offered competitive performance (88.96% accuracy) while being faster and more lightweight, making it suitable for deployment in resource-limited environments. BERT-multilingual showed balanced results with a peak accuracy of 89.84%, and ALBERT, despite having fewer parameters, reached 88.76% accuracy but underperformed in positive sentiment recall. The results demonstrate that transformer-based models outperform traditional machine learning and lexicon-based approaches, particularly in handling informal, mixed-language Malay text. The proposed models can support real-world applications such as analyzing consumer sentiment, public opinion, or social response to policies. This study contributes to advancing sentiment analysis for low-resource languages by offering comparative insights and effective model configurations, setting a solid foundation for further research and practical deployment.

PENGELASAN ANALISIS SENTIMENT BERASASKAN TRANSFORMER DALAM PEMPROSESAN SEMULA JADI UNTUK BAHASA MELAYU

ABSTRAK

Analisis sentimen dalam Bahasa Melayu memanfaatkan Pemprosesan Bahasa Asli (NLP) untuk mentafsir pendapat dan nada emosi yang dinyatakan dalam teks Melayu. Penyelidikan ini menyiasat aplikasi model pembelajaran mendalam berasaskan transformer—Perwakilan Pengekod Dwi Arah daripada Transformers (BERT), DistilBERT, BERT-berbilang bahasa, ALBERT dan BERT-CNN—untuk klasifikasi sentimen kepada kategori positif, negatif dan neutral. Kajian ini menangani cabaran dalam analisis sentimen Bahasa Melayu, termasuk sumber beranotasi terhad, nuansa linguistik dan penggunaan bahasa campuran biasa pada platform seperti media sosial. Untuk melatih dan menilai model, set data Melayu berskala besar (dataset Malaya) telah digunakan. Model pra-latihan daripada HuggingFace telah diperhalusi menggunakan pengesahan silang 10 kali ganda untuk meningkatkan generalisasi. Kaedah pengoptimuman seperti penambahan data juga dilaksanakan. Penilaian itu bukan sahaja mempertimbangkan ketepatan tetapi juga ketepatan, ingatan semula, skor F1, dan kecekapan pengiraan. Antara model, BERT-CNN mencapai prestasi terbaik, dengan ketepatan 96.30% dan skor tinggi secara konsisten merentas semua kelas sentimen. BERT juga menunjukkan prestasi yang baik, terutamanya untuk sentimen neutral, mencapai ketepatan 89.5% tetapi menunjukkan ingatan lebih rendah sedikit dalam kelas positif. DistilBERT menawarkan prestasi kompetitif (88.96% ketepatan) sambil lebih pantas dan lebih ringan, menjadikannya sesuai untuk digunakan dalam persekitaran terhad sumber. BERT-berbilang bahasa menunjukkan hasil yang seimbang dengan ketepatan puncak 89.84%, dan ALBERT, walaupun mempunyai parameter yang lebih sedikit, mencapai ketepatan 88.76% tetapi kurang berprestasi dalam ingatan sentimen positif. Keputusan menunjukkan bahawa model berasaskan pengubah mengatasi pembelajaran mesin tradisional dan pendekatan berasaskan leksikon, terutamanya dalam mengendalikan teks bahasa Melayu tidak formal dan bercampur. Model yang dicadangkan boleh menyokong aplikasi dunia nyata seperti menganalisis sentimen pengguna, pendapat umum atau tindak balas sosial terhadap dasar. Kajian ini menyumbang kepada memajukan analisis sentimen untuk bahasa sumber rendah dengan menawarkan cerapan perbandingan dan konfigurasi model yang berkesan, menetapkan asas yang kukuh untuk penyelidikan lanjut dan penggunaan praktikal.

ACKNOWLEDGEMENT

In the Name of Allah, the Most Gracious, the Most Merciful. First and foremost, I would like to take this opportunity to express my sincere acknowledgement to my beautiful parents who have raised me until I grow up to become a responsible person. My mother Norlela binti Mohd Nordin and my father Zulkalnain bin Mohd Yusoff. Also I would like to thank my family members, teachers, friends and colleagues for their teachings as they are my life's teachers. I would also like to express my sincere gratitude and acknowledgement to my supervisor, Dr. Syafeeza Ahmad Radzi for her guidance, support, knowledge and encouragement throughout the completion of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
AP	-	Average Precision
API	-	Application Programmable Interface
ASCII	-	American Standard Code for Information Interchange
AUC-ROC	-	Area Under the Receiver Operating Characteristic Curve
BERT	-	Bidirectional Encoder Representations from Transformers
BPE	-	Byte-Pair Encoding
CNN	-	Convolutional Neural Network
CPU	-	Central Processing Unit
CSV	-	Comma-Separated Values
EDA	-	Easy Data Augmentation
ELMo	-	Embeddings from Language Models
GLUE	-	General Language Understanding Evaluation
GPT	-	Generative Pre-trained Transformer
GPU	-	Graphical Processing Unit
GRU	-	Gated Recurrent Unit
GUI	-	Graphical User Interface

ICA	-	Independent Component Analysis
IDF	-	Inverse Document Frequency
KNN	-	K-Nearest Neighbor
LLMs	-	Large Language Models
LSTM	-	Long Short-Term Memory
M-BERT	-	BERT-multilingual
ML	-	Machine Learning
MLM	-	Masked Language Model
MRC	-	Malay Reviews Corpus
NB	-	Naive Bayes
NLP	-	Natural Language Processing
NLTK	-	Natural Language Tool Kit
NSP	-	Next Sentence Prediction
PCA	-	Principal Component Analysis
PRC	-	Precision-Recall Curve
RF	-	Random Forest
RNN	-	Recurrent Neural Network
SA	-	Sentiment Analysis
SQL	-	Structured Query Language
SQuAD	-	Stanford Question and Answering Dataset
SVM	-	Support Vector Machines
SWAG	-	Situations with Adversarial Generation
TF	-	Term Frequency

TF-IDF - Term Frequency-Inverse Document Frequency

URL - Uniform Resource Locater

LIST OF SYMBOLS

x_t - Input state

h_0 - Zero vector

o_t - Vector of probabilities

$\propto_j$ - Softmax function

h_t - Hidden state

LIST OF APPENDICES

LIST OF PUBLICATIONS

The followings are the list of publications related to the work on this thesis:

Mohd Asyraf Zulkalnain, A.R., Syafeeza, W.H.M. and Rahaman, S., 2025. Evaluation of Transformer-Based Models for Sentiment Analysis in Bahasa Malaysia. *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. X, No. X, pp. XX-XX. Available at: https://doi.org/10.54554/jtec.202x.xx.xx.xxx.

CHAPTER 1

INTRODUCTION

1.1 Background

Sentiment analysis is the process of determining the opinion or emotional tone of a given text and involves the use of Natural Language Processing (NLP). Natural language processing (NLP) focuses on techniques which enable computers to understand, interpret and generate human language. In addition to identifying particular feelings and viewpoints expressed in the text, this can also involve assessing the general sentiment of the text—whether it be positive, negative, or neutral. NLP also makes it possible to intepret words, such as sentiments.

Applications for sentiment analysis are numerous and include evaluating consumer reviews, identifying the tone of news articles, and reviewing the tone of social media posts. For instance, an organization may employ sentiment analysis to automatically categorize customer reviews as neutral, negative, or positive in order to instantly effectively determine the general perspective of its customers. Another example of sentiment analysis is detecting the sentiment of social media posts about a particular topic or brand. This can be useful for businesses, as it can help them understand how people are feeling about their products or services, and identify areas for improvement.