UNIVERSITI TEKNOLOGI MARA

PART OF SPEECH TAGGER FOR NON-STANDARDIZED MALAY TEXT IN SOCIAL MEDIA

NUR SYAMIMI BINTI ARDENAN

BACHELOR OF COMPUTER SCIENCE(HONS.)
FACULTY COMPUTER AND MATHEMATICAL SCIENCES

JANUARY 2022

STUDENT DECLARATION

I certify that this thesis and the project to which it refers is the product of my own work and that any idea or quotation from the work of the other people, published or otherwise are fully acknowledged in accordance with the standard referring practices of the discipline.

………………………………………………

NUR SYAMIMI BINTI ARDENAN

2018439746

30 JANUARY, 2022

TABLE OF CONTENTS

CONTENT **PAGE**

CHAPTER 1: Introduction

CHAPTER 2: Literature Review

CHAPTER 3: Methodology

CHAPTER 1

1.1 Introduction

This chapter provide an overview background of this project. It includes defining the problem statement, project objectives, project scope, and significance that led to this project. As an introduction, this project is a web-based application that normalised a non-standardised Malay text in part-of-speech using Hidden Markov Model. A proposed solution was summarized in the problem statement.

1.2 Problem Statement

Part-of-Speech (POS) was defined as a category that divides words on the basis of their use and functions in a sentence. For instance, the major POS like nouns, adjectives, verbs, adverbs, prepositions, conjunctions and interjections. Gimpel et al.(2011) and Antony, Mohan and Soman(2010) stated that POS tagging plays an important role in the linguistic pipeline and is a basic form of syntactic analysis that has numeral applications in natural language processing such as sentiment analysis and named entity recognition(Alshaikhdeeb and Ahmad, 2016). A few POS tagging techniques that is used such as statistical approach (n-gram tagging) done by a Hidden Markov Model(HMM). Text Processing is the automated process of analysing text data for getting structured information. Text Processing is one of the most common tasks in machine learning applications such as language translation, sentiment analysis, spam filtering and others.

Natural Language Processing is the ability of a computer to understand human language in a valuable way. Chowdhury(2003) highlighted that NLP is an area of research and application that deals with the ability of a computer programme to understand and process human language in large amounts of natural language data. NLP rely heavily on the results of text processing. Text processing can only be done on a standard text. However, nowadays, people usually wrote freely without maintaining a formal grammar and correct spelling. They also tend to use a lot of abbreviated words (Java, Song, Finin & Tseng, 2007). Example words that are often used mostly in urban communities like ‘usha’ (perhati/*stare*), ‘skodeng’ (intai /*peek*), ‘cun’(lawa / *beautiful*), and ‘poyo’(buruk/*bad*) (Maslida, 2018). Some of the application of NLP is Automatic Summarization, Machine Translation, Speech Processing, Information Extraction, Opinion Mining and Topic Segmentation. Most of these applications use POS tagging.(Hasan,Uzzaman and Khan, 2007).