Curtin University
Malaysia

Low-Resource Natural Language Identification

Tendai Midzi

OCTOBER 2020

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Low-Resource Natural Language Identification

Tendai Midzi

A thesis submitted for the degree of Bachelor of Technology

(Computer Systems and Networking)

October 2020

Contents

List of Figures

List of Tables

Abbreviations

ASR	Automatic Speech Recognition
BOW	Bag Of Words
HRL	High Resource Languages
LRL	Low Resource Languages
MA	Morphological Analysis
NER	Named Entity Recognition
NLG	Natural Language Generation
NLTK	Natural Language Toolkit
NRL	Natural Language Processing
SA	Semantic Analysis
SA	Syntactic Analysis

Chapter 1

INTRODUCTION

1.1 Project Overview

With technology becoming more and more deeply rooted in our day to day culture, there continues to be a growing chasm between the ever changing field of technological innovation and cultural preservation and relevance. As such, there is a concerning rise in the need to preserve certain aspects of society that can find their demise with a rise in technological advancement. One such area is spoken languages. For processing, the industry has moved leaps and bounds beyond what was thought to be possible, however the progress has mainly been in specific international languages such as English, French and Spanish which have a greater reach.

ForLow Resource Languages (LRLs), we find that of the roughly 7 000 languages spoken on earth [1], only 20 are considered High Resource Languages (HRL). This means of all the wonderful things that technology has done, there hasn't been a deep enough impact to preserve other languages. This can be easily attributed to the geography of most technological breakthroughs where if an advancement is made in country A, then there is a high chance that the advancement will be published in languages prevalent and spoken in said country. However, with the power of Natural Language Processing there are now measures to mitigate loss of languages. This not only results in preservation of LRLs but improves the chances of creating relevant and effective educational material, knowledge expansion and even emergency response to name a few [2].

The biggest drawback in working with LRLs is the fact that there are little resources available to build a corpora of words to fully study the languages. As such, issues may arise when collecting resources to fill up said corpora. And once these languages stop being taught in schools due to their lack of official use they risk going extinct. Hence the need to study and create methods and measures that save LRLs.

This study will make use of Malay and Iban languages as a means to add to the work already being done towards creating databases for LRLs. In this case Iban is the low resource with over 700,000 native speakers in Borneo with Malay having over 290 million speakers. The end goal of the project is to create a program/system that can process a given piece of text or document and determine, through various means and processes, whether its in Malay or Iban.

1.2 Motivation

Natural Language Processing (NRL) has given a lifeline to LRLs which may be at the brink of extinction. As such, in order to capitalise on the available technology, NLP methods will be applied to Iban and Malay so as to create a program that can tell whether or not a given word is in Malay or Iban or both.

The similarities between the two languages are evident yet some words when spelled the same have totally different meanings. The purpose of this project will be to try and build a program that can clearly differentiate between the two languages.

1.3 Objectives

This research project has objectives to:

1. form a corpus of documents in the Iban Language
2. study the characteristics of Iban Language for text processing
3. propose an approach to identify the language used in text for LRL, with Iban as a case study.

1.4 Contributions

The contribution in this thesis are as follows

1. contribution 1
2. contribution 2
3. other contribution

1.5 Thesis Structure

The remainder of the document will be structured as follows; Chapter 2 will go in-depth of Literature Review touching on the differences and similarities between Malay and Iban Languages. It will also give an overview of the various steps that a basic NLP System should go through before a final version is made.

Chapter 3 will look into explaining some of the current approaches that have been used to save LRLs in the context of not only Iban but other languages too. This chapter will be heavily focused on the advantages and disadvantages of each approach and solution. Though solutions are limited in the case of Malay and Iban a handful of solutions are available for other languages. From these solutions a lot can be on the feasibility and limitations of using an NLP on Low Resource languages.

Chapter 4 shows the chosen methodologies used to study to create the final system and the types of methods used on the two languages. It will also touch on the importance of the chosen approach and why some steps have to be repeated.

Chapter 5 will look into how the methods discussed in the previous chapter are put into play. This will also look at the issues faced whilst implementing the chosen methods.

Chapter 2

Literature Review

2.1 Overview of Natural Language Processing Techniques

There are multiple processes used in NLP.

2.1.1 Named Entity Recognition

Named Entity Recognition (NER) is identifying Named Entities such as names and places from a given piece of text [3]. This applies in most cases where a response is required by a user. Examples of the uses of NER are in semantic annotation, news categorization and customer support.

With respect to LRLs, NER is useful when it comes to customer service support where, for one reason or the other, a customer might be able to only communicate in one language (Iban) yet the customer service agent can only converse in Malay. With the use of NER, there is reduced ambiguity in what a certain word means as the customer services agent can easily get the context in which the customer is using certain words.

Such a situation can also help in increasing translation accuracy of websites and translation engines when languages are closely related as in the case of Malay and Iban. Some websites such as A and B would treat text in Iban as if its Malay. This leads to greater confusion in terms of how the reader will understand any content being displayed.

2.1.2 Tokenisation

Tokenisation involves the splitting of raw text to sentences, words or characters depending in the use case needed [4]. This allows any NLP to learn the language by breaking it down to tokens that can be analysed further in order or improve accuracy of translation or use. The tokenisation process is used to remove non-meaningful structures such as punctuation marks, brackets and hyphens.

However, this approach introduces errors by not taking into consideration multi word elements such as people's names and city names. This error is one that might result in an inadequate corpus model being made and introduces in the NLP pipeline. Such elements/tokens require special treatment in order to safeguard the authenticity of the project.

Depending on the source of the text, it may contain abbreviations and acronyms that need to be substituted in full. This can present an issue when the tokenising process is started. A filter can be used to catch such instances but it may not filter out all acronyms and abbreviations.

2.1.3 Stemming and Lemmatization

Stemming is the process if reducing inflection in words to their root forms such as mapping a group of words to the same stem [5]. Stemming would mean if given a words like "missing", "misses" and "missed" the stem word would be "miss". These computational procedures are usually based on a specific language, for example; English has Porter and Lancaster stems both of which have differing approaches to the stemming process.

For non-English languages, Python has a Natural Language Toolkit (NLTK) library which contains a Snowball Stemmer. Snowball Stemmer can be used to generate rules for upto thirteen languages depending on the use case as defined by the use [6].

Lemmatization is the process of grouping inflected forms of a word together in order for the to be analysed as a single entity using the words dictionary form (lemma) [7]. For example, such an approach would link ”gone”, ”going” and ”went” to ”go”. This process works well for HRLs as there is a wider base of discovery to build the languages corpus on thus increasing accuracy of the resulting system.

Applications

Stemming and Lemmatization can be used in:

1. Sentimental Analysis
2. Document Clustering
3. Information Retrieval

2.1.4 Bag of Words

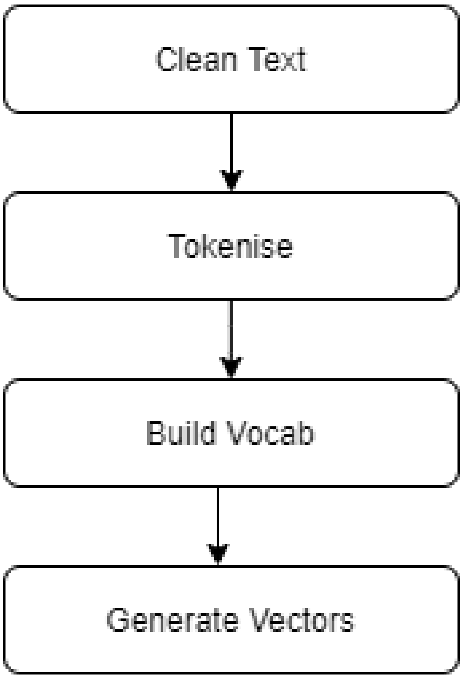


Figure 2.1: Typical Bag of Words Creation Process

The Bag Of Words(BOW) technique is used to simplify natural text or language by removing grammar and word order yet leaving the number of appearances that each word makes. Given text to work with, this method will only count the number of times words appear in the text and then apply various operations depending on the user's.

Given a sentence:

"Tamia has a basketball. Ivan wants a basketball too."

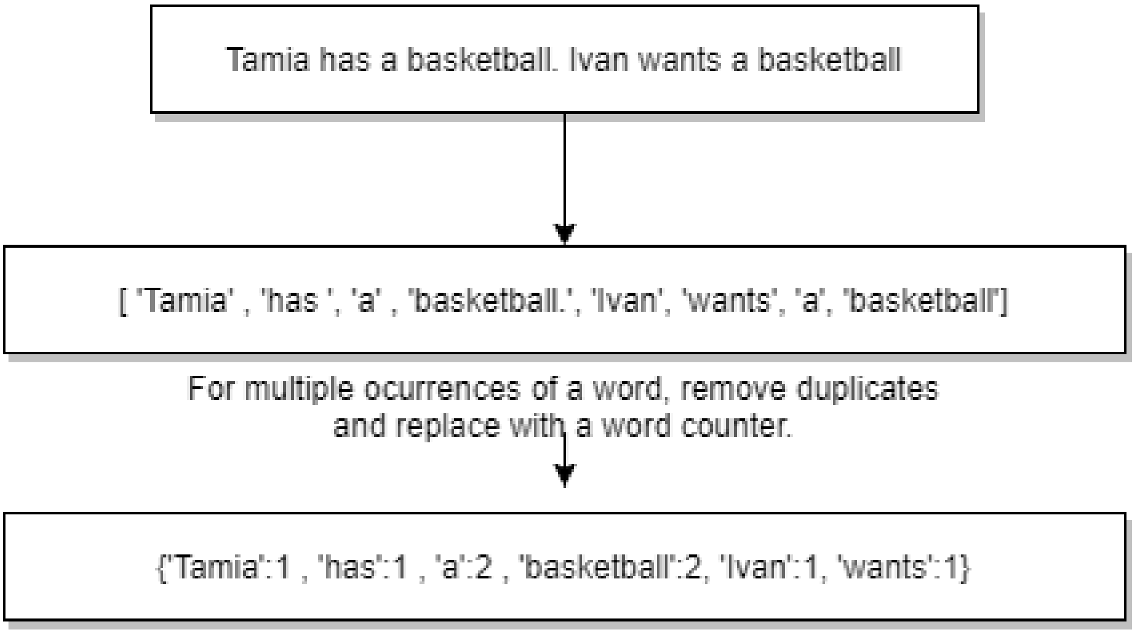

Tamia has a basketball. Ivan wants a basketball
['Tamia' , 'has ', 'a' , 'basketball.', 'Ivan', 'wants', 'a', 'basketball']
For multiple ocurrences of a word, remove duplicates
and replace with a word counter.
{'Tamia':1 , 'has':1 , 'a':2 , 'basketball':2, 'Ivan':1, 'wants':1}

Figure 2.2: Bag of Words Example

The vocabulary of the whole document will be used to create vectors for each sentence, with each vector length being equal to the vocabulary size. Word co-occurrence or word order is not taken into account in the BOW method [8]. However the BOW model can be tailored to take into account pre-determined word order. Such an approach is currently being implemented by WordNet. Another step in the BOW approach is to filter out stock stock words. these are words that appear the most in a language but do not have actually bearing or meaning to the actual words. This step in itself might be skipped if in the case of Iban and Malay the use of filter words can be the determining words of whether or not its Malay or Iban.

For example, word co-occurrence can be seen when using the term ”Kuala Lumpur”; originally this phrase can be classified as 2 different words. However with tailoring, the program can be taught to expect ”Lumpur” soon after ”Kuala”.

2.1.5 Natural Language Generation

Natural Language Generation (NLG) is the process of generating structured text or language from structured Data, This method is often employed in generating reports for companies for clearer human understanding and easier summarisation of key points from a given data. It is used when mainly working with Big Data or readily employed customer service systems including virtual assistants such as Siri or Google Assistant.

The use of either a dictionary based approach or a corpus based approach is wholly left to what suits the user better. A dictionary based approach would mean making use of online resources such as WordNEt and Merriam Webster. This approach provides a more accurate representation of word semantics especially is semantics are of vital importance in the end program. The corpus based would depend on the original corpus created for the specific program with little to no interaction with an online dictionary.

For the purposes of this project a hybrid solution will be used where a general corpus will be created and studied then if there's need to consult the online dictionary then only will it be included. Reason being, for Iban there's little to no proper dictionary source such as WordNet and for Malay a WordNet database does exist. However, due to the languages being closely related, some Iban words may be included in WordNet Bahasa as part of the Malay language due to the lack of proper distinction.

2.2 Natural Language Processing Steps

2.2.1 Morphological Analysis

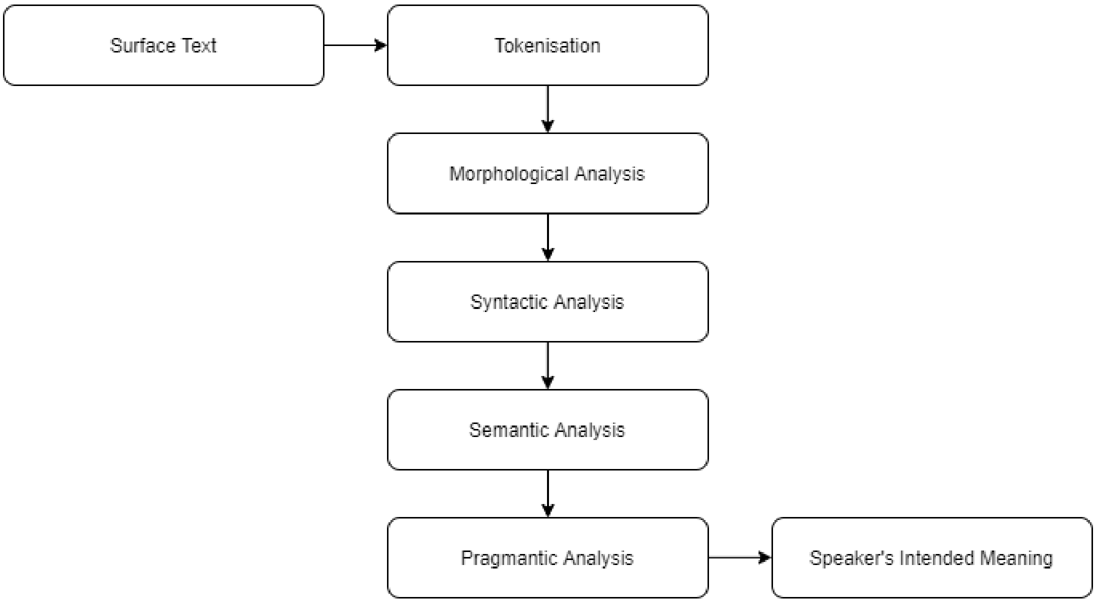


Figure 2.3: Natural Language Processing Steps
[9]

Morphological Analysis (MA) is the process of giving information pertaining to the grammar of a word with regards to the words morpheme [10]. A morpheme is the smallest meaningful part of a word in a given language [11].

For Malay and Iban also, morphemes are either free or bound. The former meaning it can appear on its own as a word with the latter meaning it cannot stand on its own as a word [11], [12]. An for example of a bound morpheme in Malay is the ”ber” in ”bernyala” which means to burn. An identical morpheme also exists in Iban but instead of ”ber” it becomes ”be” [12].

The similarity of the two languages goes beyond just the aforementioned morpheme. There is also a visible similarity in the usage of nominal prefixes such as ”pe”, ”se” and ”ke”. However they differ in what nouns are derived from the prefixes. Using ”pe”, in Iban it derives human, concrete and abstract nouns whilst in Malay, the prefix derives the first two subclasses [12].

Malay and Iban are also separated by the number of suffixes present in each language [12]. For Iban there is only one suffix which is ”-ka” and Malay has three suffixes namely; ”-an”, ”-kan” and ”-i”

These differences have to be fed into the resulting system in order for the system to find which language a given piece of text is in.

2.2.2 Syntactic Analysis

Syntactic Analysis (SA) is the analysis of a sentence or given piece of text with regards to its logical meaning. It can also be called parsing [9]. For the purposes of this project, it will mainly focus on the word construction itself since the main goal is to determine whether a given word or piece of text is in Malay or Iban.

SA often employs the n-gram-based model to determine various differences and similarities between languages. The n-gram model is defined as contiguous grouping of n things from a given example of text or discourse. These can be letters, words, syllables or even phonemes depending on the application. This n-gram model can be difficult or tricky to use once the languages are said to be similar as in the case of Malay and Iban. A study carried out on two closely related languages, Malay and Indonesian [13], reiterated this.

2.2.3 Semantic Analysis

Semantic Analysis (SA) is the process of extracting the meaning of text structures with relation to the text's context. The process entails taking into account the dictionary meaning of given words, phrases or pieces of text. Each word is analysed as a singular item and then as a part of the whole given data. It can also be described as a measure or analysis of a term or piece of text is from being negative or positive [14]. This is a broad generalisation that carries over the basis on which SA is established.

There are various elements to the SA, including; hyponymy, homonymy and polysemy. Hyponymy is the relationship between a term and its instances. Fro example, the term 'colour' (often called a hypernym) and the colour red, brown or green become its hyponyms (the relation).

Homonymy is when words have the same spelling or same form but have different and unrelated meaning.An example is address which can mean to speak to or a location.

Polysemy is almost the same as a homonymy except that a polysemy has a related meaning to the given word.

Semantic Analysis takes into account a number of building blocks such as, Entities, Concepts, Relations and Predicates. Entities represent a particular person, object, location etc. Concepts are the category in which the entities fall such as city, family or group. Relations are simply the connection between the entities and concepts. Predicates are then the verb structures that play a part in the given sentence or piece of text.

Due to the nature and expected results of this project, Semantic Analysis would not play a critical role in determining the outcome of the results. This is mainly because of the amount of time it would take to develop a proper Semantic library for both Iban and Malay. Also, the requirements of this project mainly focus on determining the language of a given piece of text instead of the meaning of it.

2.3 Relation between Malay and Iban

2.3.1 Malay Language

Chapter 3

Existing Solutions

3.1 Definition

3.2 Solution One: Named Entity Recognition Approach

[15]

The approach used by Rayner Alfred [15] was that of using NER based on a Rule-Based Approach. This was done mainly to cater for the Malay language alone without taking into account Iban. This can be considered a solution as it can be used to further review the characteristics of the Malay language and how it appears in text form especially in the NER approach.

However it can be problematic when it comes to telling the difference between Malay and Iban as the NER approach did not take into account the differences between the two languages specifically. As such, one might find that words that are natively Iban may be included in the Approach as Malay words due to the closely relatedness between the two languages. For this project, such a difference is critical to the success of the system.

Some key takeaways from this study would be the use of the NER method to create a corpus of the Malay language which can be helpful in identifying given Malay terms.

3.3 Solution Two: Automatic Speech Recognition for Iban

[16]

This solution is based on work done by [16] the Faculty of Computer Science and Information Technology at the University of Grenoble Alpes. They worked with both Malay and Iban with their focus mainly on Automatic Speech recognition for Iban. Malay was used in a 'reference' capacity due to Iban being a LRL.

Some of the issues faced with implementing the solution was that since their aim wasn't to know the language of the origin via text, they had a smaller corpus to work with when it came to Iban. The pronunciations in both languages can be very similar as such Malay was used to obtained some of the needed information.

The difference between the aims of this solution and the current project are that the case study was done without much regard to the deciphering of whether given information is in Malay or Iban. This project has its main aim being deciphering whether or not the text is in one or the other language.

However this can help to identify key areas that Malay and Iban are similar thereby clearing up the borderline between the two closely related languages.

3.4 Solution Three

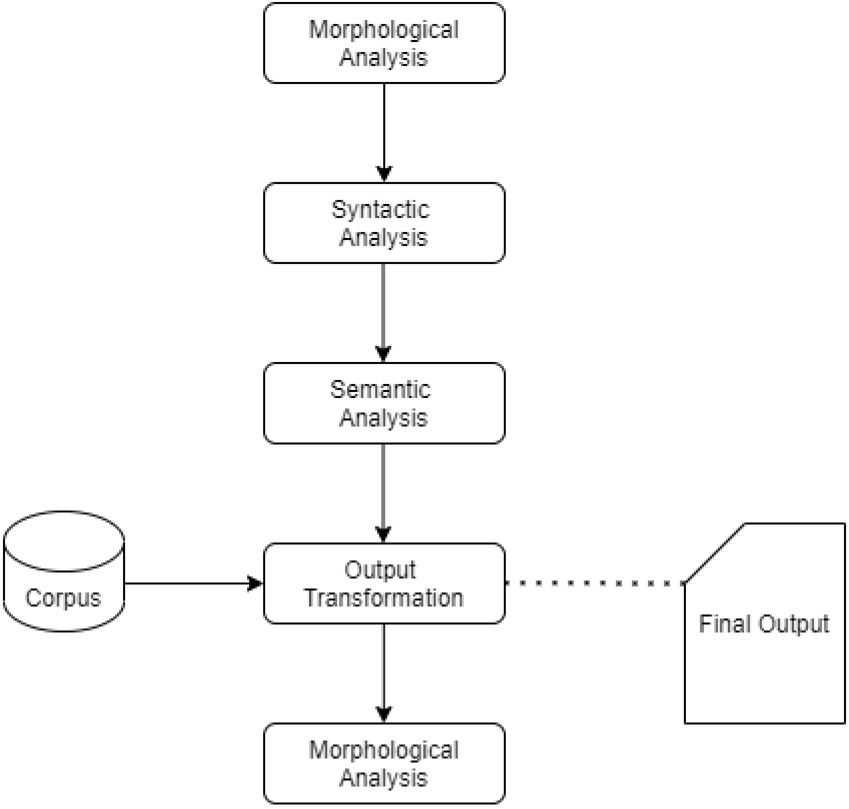
[17]

3.5 Solution Review

Chapter 4

Methodology

Figure 4.1: The basic flow of a NLP system



4.1 Text Collection

For the text collection, articles were obtained from the Utusan Borneo website [18]. Articles topics were mainly news articles with a direct translation from

Malay to Iban. A total of 30 articles were obtained. Each article had two versions; one in Malay and another in Iban. Article topics ranged from Current Affairs to general entertainment.

The content and context of the data is import if the Named Entity Recognition technique is employed. This insures that whenever a person is named in one article, the nature of their title can be easily recognised and analysed.

4.1.1 Text Pre-processing

The articles were originally downloaded in PDF format but this presented problems when it came to text extraction as HTML is not fine tuned to be saved in the PDF format. As such, the HTML webpages had to either be downloaded as is or accessed by a Python webscrapper library.

The library used is Beautiful Soup4 [19]. This permits accessing live data on websites to study and extract the text elements(HTML level) of a website. Once text is extracted, HTML or CSS tags have to removed including any excess data such as images or excess links to external websites. This leaves the text that is required.

For the Utusan Borneo website, the website had a basic structure to how they formatted their articles. This meant a single program could be used to extract all the text from a given set of webpages. The result words are saved in two separate files; one that saves the text in its original order(as sentences that make up part of a whole story) and a CSV file that saves the words as a list regardless of previous order.

Chapter 5

Implementation

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

6.2 Future Work

Future Work content here.........

Bibliography

[1] Enthnologue, “How many languages are there in the world?,” 2020.

[2] Sciforce, “Nlp for low-resource settings,” Oct. 2019.

[3] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbıs, “Named entity recognition: Fallacies, challenges and opportunities,” *Computer Standards & Interfaces*, vol. 35, no. 5, pp. 482–489, 2013.

[4] T. Verma, R. Renu, and D. Gaur, “Tokenization and filtering process in rapidminer,” *International Journal of Applied Information Systems*, vol. 7, no. 2, pp. 16–18, 2014.

[5] J. B. Lovins, “Development of a stemming algorithm,” *Mech. Transl. Comput. Linguistics*, vol. 11, no. 1-2, pp. 22–31, 1968.

[6] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, and I. Mathur, *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016.

[7] O. Dereza, “Lemmatisation for under-resourced languages with sequence-to-sequence learning: A case of early irish,” in *Proceedings of Third Workshop” Compu*, vol. 4, 2019, pp. 113–124.

[8] K. Soumya George and S. Joseph, “Text classification by augmenting bag of words (bow) representation with co-occurrence feature,” *IOSR J. Comput. Eng*, vol. 16, no. 1, pp. 34–38, 2014.

[9] R. Dale, H. Moisl, and H. Somers, *Handbook of natural language processing*. CRC Press, 2000.

[10] N. Mehdi, “Developing a kashmiri morphological analyzer generator,”

[11] H. Abdullah, “The morphology of malay,” Ph.D. dissertation, University of Edinburgh, 1972.

[12] A. H. Omar, “The iban language of sarawak: A grammatical description.,” Ph.D. dissertation, SOAS University of London, 1969.

[13] M. Zampieri, B. G. Gebre, and S. Diwersy, “N-gram language models and pos distribution for the identification of spanish varieties (ngrammes et traits morphosyntaxiques pour la identification de variétés de l’espagnol)[in french],” in *Proceedings of TALN 2013 (Volume 2: Short Papers)*, 2013, pp. 580–587.

[14] A. Al-Saffar, S. Awang, H. Tao, N. Omar, W. Al-Saiagh, and M. Al-bared, “Malay sentiment analysis based on combined classification approaches and senti-lexicon algorithm,” *PloS one*, vol. 13, no. 4, e0194852, 2018.

[15] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, “Malay named entity recognition based on rule-based approach,” 2014.

[16] S. S. Juan, L. Besacier, B. Lecouteux, and M. Dyab, “Using resources from a closely-related language to develop asr for a very under-resourced language: A case study for iban,” 2015.

[17] B. Ranaivo-Malançon, “Automatic identification of close languages-case study: Malay and indonesian,” *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 2, no. 2, pp. 126–134, 2006.

[18] U. Borneo. (), [Online]. Available: `https://www.utusanborneo.com.my/`. (accessed: 14/09/2020).

[19] R. Mitchell, *Web scraping with Python: Collecting more data from the modern web.* ” O’Reilly Media, Inc.”, 2018.