NAMED-ENTITY RECOGNITION FOR NUMERICAL EXPRESSION IN MALAY TEXT-TO-SPEECH SYSTEMS

LIT WEI WERN

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

NAMED-ENTITY RECOGNITION FOR NUMERICAL EXPRESSION IN MALAY TEXT-TO-SPEECH SYSTEMS

LIT WEI WERN

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SOFTWARE ENGINEERING (SOFTWARE TECHNILOGY)

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Lit Wei Wern

Matric No: WOC160013

Name of Degree: Master of Software Engineering

Title of Dissertation: Named-Entity Recognition for Numerical Expression in Malay Text-To-Speech Systems

Field of Study: Human Computer Interaction

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every right in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:
Designation:

NAMED-ENTITY RECOGNITION FOR NUMERICAL EXPRESSION IN MALAY TEXT-TO-SPEECH SYSTEMS

ABSTRACT

Text-to-speech (TTS) system is a system that able to convert text strings into human-like artificial speech. Natural Languages Processing (NLP) of the TTS system is the application of computational techniques to analysis and synthesis of natural language and speech. NLP must be able to convert non-words text material into words before they are synthesized. Some non-words texts like numbers are difficult to be converted to words because numbers have various formats such as calendar, time, currency, address, measurement units and so on. However, many of the existing TTS systems cannot accurately convert numbers which contain several types of numerical formats, leading to reduced intelligibility of the synthetic speech generated by the existing systems. The objective of this research is to automatically classify the numerical format of Malay text, to enable the NLP to perform the appropriate text to speech conversion, which will improve the intelligibility of the synthetic speech generated by the existing TTS systems. This research is categorized as named-entity recognition (NER) for numerical expression as it fits the definition of NER, which is the identification of certain entities in text. This research has proposed a context-based classification technique which classifies the numerical format of numbers, based on contexts of the numbers. Classifying the numbers into its appropriate format can assists in the accurate conversion of the numerical format into relevant text. This research has developed a classification system that able to classify six types of numeric contexts which consists of date, time, phone number, currency, measurement, and percentage. These formats were selected because they are the most commonly used in online news. A Malay text-numbers corpus containing over five hundred numerical formats with their sentences was built. There are four commonly used machine learning techniques were adopted for

developing the classification system, which are the Support Vector Machine (SVM), K-Nearest Neighbors (KNN) Linear Discriminant Analysis (LDA), and Decision Tree (DT). 10-fold cross-validation and listening evaluation by native listeners was used for performance evaluation of the system. The confusion matrix is used to describe the performance of a classification model in more detail. Calculations on classification accuracy, precision, recall, and F-Measure were performed to each classifier. The highest classification mean accuracy achieved is 94.37% by using the context-based model as a features extractor, and DT as a classifier. For classifiers, the mean accuracies for SVM, KNN, and LDA were 93.86%, 91.07%, and 90.39%, respectively. In conclusion, the proposed solution was found to be effective in classifying the number format, and the accuracy of text conversion. From the listening test, this research has increased the intelligibility of the synthetic speech generated by the existing Malay TTS system which includes the numerical formats.

Keywords: Text-to-Speech (TTS), Named Entity Recognition (NER), Machine Learning (ML), Natural Language Processing (NLP).

PENGENALPASTIAN BIDANG-ENTITI UNTUK UNGKAPAN NUMERIK DALAM SISTEM TEKS-KE-UCAPAN BAHASA MALAYSIA

ABSTRAK

Sistem teks-ke-ucapan (TTS) adalah sistem yang dapat menukar rentetan teks ke dalam ucapan buatan manusia seperti. 'Natural Language Processing' (NLP) sistem TTS adalah penerapan teknik pengiraan untuk analisis dan sintesis bahasa semulajadi dan ucapan. NLP mesti dapat menukar kata-kata teks bukan berbentuk kata-kata sebelum mereka disintesis. Beberapa teks bukan kata seperti angka sukar ditukar kepada perkataan kerana nombor mempunyai pelbagai format seperti kalendar, masa, mata wang, alamat, unit ukuran dan sebagainya. Walau bagaimanapun, banyak sistem TTS yang sedia ada tidak dapat mengubah nombor yang mengandungi beberapa jenis format berangka, yang membawa kepada kecerdasan bunyi ucapan sintetik yang dihasilkan oleh sistem yang sedia ada. Objektif penyelidikan ini adalah untuk mengkelaskan secara automatik format berangka teks Melayu, untuk membolehkan NLP melaksanakan teks yang sesuai untuk penukaran pertuturan, yang akan meningkatkan kecerdasan ucapan sintetik yang dihasilkan oleh sistem TTS sedia ada. Kajian ini dikategorikan sebagai pengiktirafan bernama-nama (NER) untuk ungkapan numerik kerana ia sesuai dengan definisi NER, yang merupakan identifikasi entiti tertentu dalam teks. Kajian ini telah mencadangkan teknik pengklasifikasian berdasarkan konteks yang mengklasifikasikan nombor berangka nombor, berdasarkan konteks nombor-nombor. Mengelaskan nombor ke dalam formatnya yang sesuai boleh membantu dalam penukaran tepat format berangka ke dalam teks yang berkaitan. Penyelidikan ini telah membangunkan sistem klasifikasi yang dapat mengklasifikasikan enam jenis konteks numerik yang terdiri daripada tarikh, masa, nombor telefon, mata wang, pengukuran, dan peratusan. Format ini dipilih kerana ia adalah yang paling biasa digunakan dalam berita dalam talian.

Korpus nombor teks Melayu yang mengandungi lebih daripada lima ratus format berangka dengan ayat mereka dibina. Terdapat empat teknik pembelajaran mesin yang biasa digunakan untuk membangunkan sistem klasifikasi, iaitu ‘Support Vector Machine’ (SVM), ‘K-Nearest Neighbour’ (KNN), ‘Linear Discriminant Analysis’ (LDA), dan ‘Decision Tree’ (DT). Teknik pengesahan 10 kali ganda validasi-silang dan penilaian mendengar oleh pendengar asli digunakan untuk penilaian prestasi sistem. Matriks kekeliruan digunakan untuk menggambarkan prestasi model klasifikasi dengan lebih terperinci. Pengiraan ketepatan klasifikasi, ketepatan, ingat, dan F-Ukur dilakukan untuk setiap pengelas. Pengelasan tertinggi ketepatan yang dicapai adalah 94.37% dengan menggunakan model berdasarkan konteks sebagai pengekstraksi ciri, dan DT sebagai pengelas. Bagi pengelas, ketepatan min bagi SVM, KNN, dan LDA masing-masing adalah 93.86%, 91.07% dan 90.39%. Sebagai kesimpulan, penyelesaian yang dicadangkan didapati berkesan dalam mengklasifikasikan format nombor, dan ketepatan penukaran teks. Daripada ujian mendengar, kajian ini telah meningkatkan kecerdasan ucapan sintetik yang dihasilkan oleh sistem TTS Melayu yang sedia ada yang merangkumi format berangka.

Kata Kunci: Teks-ke-Ucapan, Pengenalpastian Bidang-Entiti, Pembelajaran Mesin, Pemprosesan Bahasa Semulajadi.

ACKNOWLEDGEMENTS

First and foremost, I have to thank my research supervisor, Dr. Mumtaz Begun Mustafa of the Faculty of Computer Science and Information Technology at University of Malaya. Without her assistance and dedicated involvement in every step throughout the process, this research would have never been accomplished. I would like to thank you very much for your support and understanding over the past years.

I am grateful to my friends who participated in my presentations and supported me along the way.

Finally, I must express my very profound gratitude to my parents, my siblings, and my girlfriend for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this dissertation. This accomplishment would not have been possible without them. Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI: Artificial Intelligent

ANN: Artificial Neural Network

CNN: Convolutional Neural Network

CRF: Conditional Random Fields

DARPA: Defense Advanced Research Projects Agency

DT: Decision Tree

HMM: Hidden Markov Model

KNN: K-Nearest Neighbors

LDA: Linear Discriminant Analysis

LR: Logistic Regression

LSTM: Long Short-Term Memory

ME: Maximum Entropy

MEMM: Maximum Entropy Markov Model

ML: Machine Learning

MUC: Message Understanding Conference

NB: Naïve Bayes

NER: Named-Entity Recognition

NLP: Natural Language Processing

NSWs: Non-Standard Words

RPOS: Rule-Based Part of Speech

SMO: Sequential Minimal Optimization

SVM: Support Vector Machine

TTS: Text to Speech

VIA: Voice Information Associates

WER: Word Error Rate

CHAPTER 1: INTRODUCTION

1.1 Overview

Computers have become a part of human's life and inseparable in this era. In current technology, most of the daily tasks are highly rely on computer technologies to execute because it's able to complete the given tasks in a short time and has perfect results. Computer technologies are the development of software programs embedded in electronic devices such as personal computers, laptops and smartphones. Those devices are highly relying on the visual display such as pictures and texts to interact with users. The human can only interact with those devices through eyesight. Computers offered a lot of benefits to human but people with poor eyesight are hard to enjoy. Voice-User Interface is another interaction environment that makes use of speech to interact with computers. People suffer in vision problem such as visual impairment is able to use the computer and enjoy all the benefits as the others if the Voice-User Interface technology is widely implemented.

Text-to-speech (TTS) system is a speech synthesis system that was first developed to aid the visually impaired by reading a text to the user. TTS system is able to generate a spoken voice output that is converted from the input texts. This function may assist visually impaired patients to listen to articles or written contents on an electronic device. Users of the TTS system keep on increasing because this system not only works well for the blind, but also appeals to the normal people. TTS system can function as an essential tool for people with hectic lifestyles, especially those who live in the urban area. For instance, users can listen to written contents such as the news while working or driving.

1.2 Research Background

This section introduces the basics of the text-to-speech system as well as the roles and significance of numbers in the text.

1.2.1 Text-to-Speech System

There are a lot of commercially well-known TTS systems which are able to read aloud text but not all of the TTS systems are able to achieve high naturalness and intelligibility. Some of them not only support common standard input text but also other forms of written text input such as PDF files, PowerPoint files, etc. The commercially available TTS systems consist of Festival, INOVA, TTSreader, NaturalReader, iSpeech and TextAloud3. All of TTS systems support multiple languages such as English, Mandarin, Korean, etc. but none of them support Malay language.

Natural language processing (NLP) and digital signal processing are the two main parts of a TTS system. NLP is a process where the text input is converted into linguistic representation, whereas the speech synthesis is a process to generate speech waveforms based on the linguistic representation. NLP refers to the application of computational approaches that function as analysis and synthesis of human-like language and speech. It comprises of three processes and they are text analysis, phonetic analysis, and prosodic analysis (Aida-Zade et al., 2013). The prosodic analysis depicts the emotion of the speaker. On the other hand, the phonetic analysis assigns phonetic transcription to each word. In contrast, the text analysis is a process of computational analysis of texts. On the other hand, speech synthesis refers to the process that generates intelligible and natural human speech by computer or machine on the human language text. There are two attributes that help to identify the quality of a TTS system which are the intelligibility and naturalness (Hinterleitner, Norrenbrock, & Möller, 2013).

Intelligibility refers to the accuracy of a TTS system to convert the text input into speech output, whereas naturalness refers to how close the synthetic speech are to the human sound.

1.2.2 Text Normalization

Text normalization process is one of the processes of text analysis. Text normalization process has a lot of important functions which consists of spellchecking, pre-processing, number expansion, punctuation analysis, non-standard words (NSWs) observation, and disambiguation. The abilities of text normalization process easily affect the intelligibility and naturalness of the TTS system.

According to a previous study (San-Segundo, Montero, Giurgiu, Muresan, & King, 2013), text normalization is an important process which is used for converting NSWs into machine readable format. A text normalization module used to enhance number transcription was created in this research. Numbers are divided into digits and rewritten into standard words when handling NSWs such as number.

1.2.3 Number Expansion and Classification

Number is one of the NSWs that still cannot be converted accurately by most of the existing NLPs. According to Burkhardt & Reichel (2016), it may be due to the inability of these systems to recognize the relevant context of the numbers before expanding them to the corresponding standard words.

According to Black & Lenzo (2000), most of the TTS systems are usually built for a limited domain. For instance, some of the TTS systems are purposely built for reading only time format or just phone number format. These kinds of TTS systems are not

intelligible enough to be commonly used by the public. However, in order to produce accurate speech output for a variety of number formats, numerical classification process should be performed. Numerical classification process can highly improve the accuracy of number to text conversion and produce correct speech output which can improve the intelligibility and usability of the TTS systems.

Numbers exist in a variety of formats, such as date, time, measurement, currency and etc. On the other hand, these formats can be similar to each other and thus difficult for NLP to determine the exact number format. NLP and text normalization processes can be performed more accurately by classifying the context of numbers (Shetake, Patil, & Jadhav, 2014). Moreover, accurate classification of numerical formats can help the TTS systems to generate more intelligible synthetic speeches.

There are some related existing researches have been conducted and present the number classification accuracy of their research. Saleh, Tounsi & van Genabith (2011), conducted a research which extracting temporal and numerical expressions in Arabic language. The developed system achieving state-of-the-art results with an F-score of 88.5% (resp. 96%) for bracketing and 73.1% (resp. 94.4%) for detection. Besides that, Yaser (2016) has developed a model for predicting the numerical format and the model able to achieve 97.16% of accuracy by using Support vector machine classifier.

Proper number handling is an ability that is very important, which expresses the usefulness of the tool. The ability of proper number handling is very important and most of the TTS systems lack this ability. For instance, when numbers that are separated by a dash (-) occurs, most of the TTS systems either totally ignored the dash or performed arithmetic on the numbers string (Tetschner, 2003). These sorts of errors can be very serious and affect the intelligibility of the TTS system.

1.3 Research Problems

TTS system performs well in converting words into speech output but it cannot perfectly convert non-standard words (NSWs) such as abbreviations, acronyms, symbols, dates, numbers, etc. into speech (Sproat et al., 2001). TTS system is unable to convert numbers correctly because the system is unable to classify the format and actual meaning of the numbers accurately. This problem is due to multiple formats of number that exist, where the existing TTS system are unable to classify correctly.

According to Tetschner (2016), numbers is one of the NSWs that has a variety of formats and still cannot be converted correctly. At the same time, the lack of the accuracy in number to text conversion can result in the wrong speech output generated by the existing TTS systems and it will degrade its usefulness and intelligibility.

Numerical classification method has been applied as an additional process to classify the types of numbers in order to perform the number expansion process based on the number type. Numerical classification using machine learning is a solution which classifies the numerical contexts with high accuracy (Yaser, 2016).

Existing researches on Malay TTS system (El-Imam, & Don, 2000; Samsudin, & Kong, 2004; Swee, Hussain, & Salleh, 2008; Khalifa, Ahmad, Hashim, & Gunawan, 2008) focus on generating intelligible and natural speech, but they overlooked the importance of numerical classification which can be a possible solution to enhance the conversion of number into Malay text.

1.4 Research Aim and Objectives

The overall aim of this study is to automatically classify the numerical format of Malay text, so as to enable the NLP to perform the appropriate text to speech conversion, which will improve the intelligibility of the synthetic speech generated by the existing TTS systems. The specific objectives are:

1. To propose suitable technique(s) to automatically classify the numerical formats in Malay text.
2. To develop an automatic numerical format classification system for Malay text.
3. To evaluate the performance of the
    a. proposed technique(s) in classifying various number formats and,
    b. Intelligibility of the synthesized Malay speech which consists of different number formats.

1.5 Research Questions

- RQ1. What are the suitable technique(s) that can be applied for classifying the numerical format? (Objective 1)
- RQ2. How the numerical format in Malay text is classified? (Objective 2)
- RQ3. How are the performance of the proposed technique(s) and the intelligibility of the synthesized Malay speech consisting different number formats measured? (Objective 3)

1.6 Research Scope and Focus

This study focuses on classifying the numerical format of the input text. This process is called as named-entity recognition for numerical expression. The scope of this research is mainly on the classification of number formats for input texts in the Malay language. The numerical formats include calendar dates, time, phone number, currency, and measurement, which appear in writing frequently.

1.7 Significance of Research

The purpose of this study is to automatically classify the numerical format based on their context in Malay language, to produce the correct linguistic representation of numbers. The proposed solution for an automatic numerical classification is able to improve the TTS systems to generate correct semantic representation of numbers, which can highly increase the intelligibility of the synthesized speech. Besides that, the proposed solution simplifies the process of data collection and pre-processing of numbers for improving the classification accuracy. People who have difficulties in reading, such as dyslexia, as well as the illiterates and the blind are able to access online contents with the support of this study.

1.8 Overview of Dissertation

Chapter 2: This chapter reviews the literature related to this research, such as limitation of existing TTS systems in synthesizing text for numerical input, suitable classification techniques, data collection and evaluation methods.

Chapter 3: This chapter describes all the necessary stages of this research that were carried out, such as research problem and solution, dataset collection, design and development, and evaluation method, which help to meet the research goals.

Chapter 4: This chapter describes the processes in collecting the dataset to develop the proposed system. This includes processes such as data collection, data selection, data labeling, applied procedures, and material used.

Chapter 5: This chapter focuses on the evaluation and results of the proposed classification system. Different evaluation methods are used to evaluate the classification accuracy of the proposed system.

Chapter 6: This chapter summarizes and concludes this study. Besides that, this chapter describes the research contributions and limitations, and the future works related to this research.

CHAPTER 2: LITERATURE REVIEW

2.1 Overview

This chapter mainly focuses on the research-related issues in the TTS system, including the limitations of existing commercial TTS system in synthesizing number formats, suitable techniques to automatically classify the numerical formats in Malay text, as well as the evaluation methods. Nowadays, most of the Smartphone have embedded TTS systems that support multiple languages. However, Malay language is not one of the languages supported by the existing TTS systems for commercial use. The desire to develop a Malay language TTS system is on the rise. The limitations of the existing TTS systems should be studied to overcome, so that a Malay TTS system with improved features can be implemented.

2.2 Existing Research on Malay TTS Systems

Over the years, many researches have been conducted to develop a reliable Malay TTS system. El-Imam & Don (2000) used the system that was previously used to synthesize the Arabic language, to synthesize standard Malay language. The synthesis method has been enhanced and modified in order to suit the standard Malay language.

Samsudin & Kong (2004) presented a simple Malay speech synthesizer using syllable concatenation approach. The synthesizer concatenates the Malay sound according to the syllable segment which has been arranged based on the raw text. The quality of the system is reasonably intelligible because it is able to speak all types of Malay words but the output sound is not natural enough.

Swee, Hussain, & Salleh (2008) highlighted that the Malay TTS system has poor quality in generating speech sound. They focused on generating more natural and accurate unit

selection for synthesizing speech by combining both linguistic context and feature distance cost for selecting the best matched unit.

Khalifa, Ahmad, Hashim, & Gunawan (2008) presented a rule-based TTS synthesis system for standard Malay, which uses pre-recorded wave files and sinusoidal method in generating speech. Their work focused on developing a natural sounding and intelligible formant-based speech synthesis system. Table 2.1 shows the existing work on Malay TTS systems.

Table 2.1: List of Existing Researches on Malay TTS Systems

Authors	**Title**	**Research Focus**	**Remark**
(El-Imam, & Don, 2000)	Text-to-Speech Conversion of Standard Malay	Segmental synthesis of Standard Malay speech	Overlooked the issues of NSWs
(Samsudin, & Kong, 2004)	A Simple Malay Speech Synthesizer Using Syllable Concatenation Approach	Malay Speech synthesizer using a concatenative approach with syllables unit	Overlooked the issues of NSWs
(Swee, Hussain, & Salleh, 2008)	Corpus-based Malay text-to-speech synthesis system	A method of combining both linguistic context and feature distance cost for selecting the best match unit	Overlooked the issues of NSWs
(Khalifa, Ahmad, Hashim, & Gunawan, 2008)	SMaTalk: Standard Malay text to speech talk system	Rule-based TTS synthesis system using a sinusoidal method and some pre-recorded wave files in generating speech	Overlooked the issues of NSWs

Existing researches on Malay TTS systems have focused on generating intelligible and natural speech. Researchers have applied different synthesis approaches to generate more intelligible and natural Malay speech, whereas they overlooked the issues of Non-Standard Words (NSWs) which will affect the intelligibility of TTS systems.

2.3 Non-Standard Words

The real text contains not only words but also NSWs such as digits, abbreviations, acronyms, symbols, dates, numbers, etc. The majority of NSWs are difficult to be pronounced because they are directly based on the context (Sproat et al., 2001). Speech technologies like speech translation and TTS systems, which are unable to pronounce NSWs correctly will decrease the usefulness and reflect low intelligibility characteristic of the system. NSWs have different rules and processes for pronouncing the words as compared to standard words. NSWs have to undergo many processes, such as normalization and number expansion processes before the system is able to pronounce. Text normalization is a process to replacing the NSWs with the contextually appropriate ordinary word or sequence of words. On the other hand, number expansion is a process to expand a number to show the value of each digit. For instance, the system has to identify the numbers format, expand digit sequences into words based on the format type, and are spelled as ordinary words. For example, number 1990 can be expanded as nineteen-ninety (in case of years) or one thousand nine hundred and ninety (in case of measurement). For abbreviation such as kg (kilograms), the missing letters of the original word have to be found out in order to spell it correctly.

Sproat et al. (2001) proposed four main ways to read numbers: Read numbers as a string of digits (e.g. phone numbers), a pair of digits (e.g. years), a cardinal (e.g. quantities) and ordinal (e.g. dates). A cardinal number is a number that shows how many of something there are, such as one, two , three, whereas an ordinal number is a number that shows the position of something in a list, such as 1^{st}, 2^{nd},3^{rd} etc. Sproat et al. (2001) also proposed an implemented set of methods for dealing with various classes of NSWs and mentioned that hand-constructed rule is the best approach to handle number expansion for various languages.

Burkhardt & Reichel (2016) presented a taxanomy of problem classes such as Normalization, Foreign Linguistics, Natural Writing, Language Specific, and General. This study discussed the problems of text normalization faced by the TTS systems and discussed each problem class in detail when text normalization involved NSWs, which included abbreviations, acronyms, acronym-abbreviations, addresses, numbers and units, and also special characters. Figure 2.1 shows the numbers of samples per problem class involved in the study and the distribution is far from being equal.

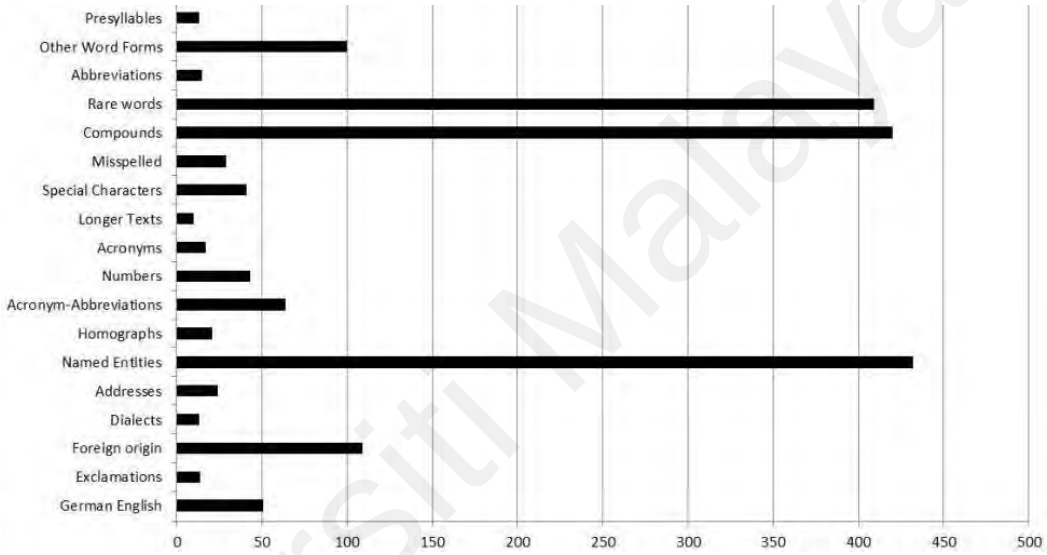


Figure 2.1: Numbers of Samples Per Problem Class

Contact number is one of the problematic numbers for text conversions, according to Pappas (2015). A contact number should be read as a string of digits. For instance, *+60345678900* can be expanded as – *plus six zero three four five six seven eight nine zero zero* - (since it is a phone number) and not - *sixty billion, three hundred and forty-five million, six hundred and seventy-eight thousand, nine hundred* - which is a common mistake currently faced by most of the existing TTS system.

2.4 TTS Systems in Synthesizing Numbers

According to a previous study about testing the accuracy of commercially available TTS products that existed as of 2003 (Tetschner, 2003), 7 TTS systems' accuracy tests have been carried out, which consists of number processing, words of foreign origin processing, acronym processing, abbreviation processing, name processing, address processing and homograph processing.

Table 2.2 shows the TTS accuracy summary presented by Tetschner under funding from Voice Information Associates (VIA). There were a total of 19 commercially available TTS systems involved in these testing and number processing showed the lowest accuracy among the tests with only 55.6% correct, whereas the other areas were stronger with higher average percentage being correct. Homograph processing achieved the highest accuracy with an accuracy of 83.4%.

Table 2.2: TTS Accuracy Summary (Tetschner, 2003)

TTS Accuracy Test	**Average % of Accuracy**
Number Processing	55.6%
Words of Foreign Origin Processing	58.8%
Acronym Processing	74.1%
Abbreviation Processing	72.9%
Name Processing	70.7%
Address Processing	69.0%
Homograph Processing	83.4%

The study (Tetschner, 2003) highlighted that incorrect numbers handling are a common issue that appears in TTS number processing. The most common errors with a TTS product are: minus (-) in a sentence is either called "dash" or ignored directly by the system.

According to a comparative study on comparing commercial and open source TTS system's performance (Burkhardt & Reichel, 2016), a taxonomy of specific classes in the TTS synthesis has been presented. Normalization is one of the problem classes and it consists of abbreviations, acronyms, acronym-abbreviations, addresses and numbers, and units. Based on the findings of this study, both the commercial and open source TTS systems have to be improved to overcome all the listed problem classes to achieve a good TTS system.

A previous study highlighted that the TTS systems have a total of five general attributes, which include intelligibility, naturalness of speech, prosodic quality, disturbances, and calmness (Hinterleitner, Norrenbrock, & Möller, 2013). This study has also mentioned that intelligibility and naturalness are two attributes that are very important and commonly used. Intelligibility describes the ability to convert text input into speech output of a TTS system, whereas naturalness expresses how human-like the speech is.

According to a study conducted by Yaser (2016), three commercially available TTS systems have been selected to compare the ability of number handling in English language. This research has finally chosen Festival, INOVA, and TTSreader TTS systems to evaluate their capability of number handling. Festival speech synthesis system is an open source, free software multilingual speech synthesis workbench which can be used as a research toolkit for speech synthesis (Alan Black, n.d.). On the contrary, the INOVA TTS system is one of the top ten TTS software solutions ranked by eLearning Industry's Network (Pappas, 2015). Lastly, TTSreader was shortlisted because it is a well-known rule-based TTS system (Yaser, 2016). Numbers in different formats have been evaluated, which include phone numbers, dates, fractions, time, and currency format. The result of the TTS systems in handling numbers in various formats shows that Festival has the highest mean classification accuracy among the three TTS

systems, but with only 50%. On the other hand, the second highest is TTSreader with 43.75%, whereas INOVA has the lowest mean classification accuracy with only 37.5%. The results show that these TTS systems lack the ability to synthesize numbers with various formats accurately even in English language.

2.5 Named-Entity Recognition (NER)

Natural Language Processing (NLP), abbreviated as NLP is a subfield of artificial intelligence and computer science deals with the interactions between computers and human languages. NLP is an area of research that discusses how computers can be applied to understand and handling natural language text or speech to perform useful tasks (Chowdhury, 2003). Named Entity Recognition (NER) is one of the subtask of NLP which has many applications mainly in natural language understanding, text to speech synthesis, information extraction, information retrieval, machine translation, question answering etc. These applications need an accurate NER system to provide better level of performance. The purpose of NER is to recognizes named entities in any plain text and classify texts into some predefined categories such as name of persons, name of locations, name of organizations, date, time, quantities, and etc.

NER is a pioneer for many NLP tasks. The aim of NER is to identify names of all the peoples, organizations, geographic locations, time, currency and percentage expressions in a text. The word “Named Entity” that commonly used in NLP is initiated during Sixth Message Understanding Conference (MUC-6). “Named Entity” task is a subtask of the Message Understanding Conference (MUCs) which sponsored by the Defense Advanced Research Projects Agency (DARPA) in December 1993. MUC is concentrated on Information Extraction tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as

newspaper articles at that moment. Participants of MUCs realized that it is necessary to categorize information units such as name of persons, name of locations, name of organizations, date, time and quantities to have a deep understanding for the message (Grishman & Sundheim, 1996). Identifying references to these entities in text was recognized as one of the important sub-tasks of Information Extraction and was called "Named Entity Recognition and Classification" (Kaur & Gupta, 2012). The early systems used handcrafted rule-based algorithms. Over the year, ML is more often applied to the systems due to its flexibility and advantages.

A rule-based NER system uses hand constructed rules are also known as the rule-based technique, which applies all the relevant rules of grammar and imposes linguistic constraints in the classification of Named Entities in unstructured textual content as well as documents such as news and academic articles. Rule-based NER systems are useful in NLP discipline but it has several drawbacks such as it is language dependent and the difficulty to adapt changes.

Machine learning (ML) is a technique where the computer can learn from data and information autonomously by using computer algorithms. It can change and improve the algorithms automatically without changing their program every time. ML is also known as a subset of artificial intelligent (AI) which can learn and make a decision after appropriate training is given. ML has a lot of benefits and advantages, such as low cost, powerful computational processing, and data storage accessibility. ML techniques such as conditional random fields (CRF), Maximum Entropy Markov Model (MEMM), Support Vector Machine (SVM) and Hidden Markov Model (HMM) are commonly used for NER (Morwal, Jahan, & Chopra, 2012).

2.6 Malay Named-Entity Recognition

Over the years, some researchers have studied the NER for Malay language. Alfred, Mujat, & Obit (2013) have proposed a rule-based Part of Speech (RPOS) tagger to perform the tagging operation on Malay text articles. This research using POS tag dictionary and a set of rules to identify the words that are considered parts of speech. As a result, the proposed rule-based tagger has a higher performance accuracy compared to a statistical POS tagger which indicates that a rule-based POS tagger for Malay language can predict unknown word's POS at some promising accuracy.

Alfred, Leong, On, & Anthony, (2014) have proposed a rule-based NER algorithm for Malay articles. The Malay NER system is focuses on identifying three name-entities such as person, location and organizations. This research uses rule-based technique instead of ML technique because the lack of annotated corpus resources for Malay language that can be used as a training data. Collecting and creating a large annotated dataset for Malay language is very time-consuming. The proposed Malay NER algorithms have a reasonable output of 89.47% for the F-measure value, 94.44% of recall and 85% of precision.

Salleh, Asmai, Basiron, & Ahmad, (2017) have proposed a Malay NER using CRF. They highlighted a lack of research has been done to analyze the recognition of Malay entities. This Malay NER system is focuses on recognizes entities from unstructured textual data. The created model was tested with data that consisted of 1995 tokens word that were generated from Bernama news. The evaluation results for NER effectiveness were measured by precision, recall and F1-score (F-measure). The precision for all entities were 75%, recall was 72% and F1-score was 70%. Table 2.3 shows the existing research on Malay NER.

Table 2.3: Existing Researches on Malay NER

Title	Year	Language	Technique	Remarks
A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles	2013	Malay	Rule-based technique	- Using rule-based technique - No training data required
Malay Named Entity Recognition Based on Rule-Based	2014	Malay	Rule-based technique	- Using rule-based technique - No training data required
A Malay Named Entity Recognition Using Conditional Random Fields	2017	Malay	Machine Learning technique	- Using Conditional Random Fields ML technique - Testing data consisted of 1995 tokens word - F1-score was 70%

Table 2.3 shows that there is little exploration on Malay NER. Various ML techniques can be applied in Malay NER to achieve a better result compare to the existing research.

2.6.1 Named-Entity Recognition for Numerical Expression

NER is a precursor for many NLP tasks. Various numerical expressions such as date, time, currency, percentage and measurement are difficult to extract using traditional NLP because they belong to open class of expressions (Kaur & Gupta, 2012). For instance, there is an infinite variety and new expressions are constantly being created from time to time.

Limited exploration has been done on NER for numerical expression. A research that came close to this purpose is done by Saleh, Tounsi, & van Genabith (2011) that present a ML technique applied to identify different forms of numerical expressions in Arabic language and another ML technique to identify Arabic temporal. The classifier used in this research is SVM classifier and evaluation metrics are detection performance and bracketing performance. For this research, the system achieving a F-score of 88.5%. Table 2.4 shows the existing research on NER for Numerical Expression.

Table 2.4: Existing Researches on NER for Numerical Expression

Title	Year	Language	Technique	Remarks
ZamAn and Raqm: Extracting Temporal and Numerical Expressions in Arabic	2011	Arabic	- Machine learning technique	- F1 measure for 88.5% for bracketing performance and 73.1% for detection performance.

Table 2.4 shows that there is no existing research focuses on Malay NER for numerical expression. Since current TTS system has poor performance in reading numerical expression in Malay language therefore research should be done to further improve the performance for current TTS system.

2.7 Techniques in Named-Entity Recognition

The techniques in NER can be classified into two types, which are the rule-based technique and ML technique. The earliest work in NER involved hand constructed rules. Following that, growing number of work is created using ML technique due to its benefits and advantages.

2.7.1 Rule-Based Technique in Named-Entity Recognition

Hand constructed rules are also known as the rule-based technique, which applies all the relevant rules of grammar and imposes linguistic constraints. The rule-based technique is a technique that applies hand constructed rules to make choices. For example, a set of rules is defined to convert numbers into proper text before a system creates an output speech. The rule-based technique is very flexible in incorporating the domain knowledge into linguistic knowledge.

On the other hand, a previous study mentioned that the growing number of rules can lead to a more complex system (Khan, Anwar, & Durrani, 2017). The rule-based

technique can become very complex under some circumstances although it is flexible to use. As numbers can have several formats, more rules are needed to completely cover all types of number formats.

Gorinski, Wu, Grover, Tobin, Talbot, Whalley & Alex (2019), show that the gap between the rule-based and ML approaches of NER for electronic health records is relatively small and ranging between 0.03 and 0.06 in F-score. A rule-based system engineered by domain experts requires a high development cost in terms of time and effort.

Rule-based technique is a traditional technique applied in many researches. However, this technique is not appropriate for this research because a bunch of rules are needed to completely cover a single numerical format. The rules become complex and grow significantly when more numerical formats are involved.

2.7.2 Machine Learning in Named-Entity Recognition

Google Scholar is a freely accessible web search engine that provides a lot of published articles. Based on the search result from Google Scholar on July 2019, a total of 12 English articles were found that published within year 2015 to year 2019 with Machine Learning and Named-Entity Recognition as keyword in the title of article.

Vijay & Sridhar (2016) presented a ML approach to NER for the travel and tourism domain. Conditional random fields (CRF) are the ML technique that has been applied in this research. The Travel and Tourism domain data was collected manually from Wikipedia and TripAdvisor.com. It consists of 6996 sentences and 140481 words. The evaluation shows that the CRF model provides good performance for travel and tourism domain with 82% of recall, 85% of precision, 82% of accuracy and 83% of F-measure.

Khanam, Khudhus & Babu (2016) presented a NER using ML techniques for Telugu language. A rule based named entity recognizer was implemented with some suffix and prefix features as well as dictionary which consists of two hundred thousand words. Furthermore, they applied CRF ML technique to further enhance the system.

Filipiak, Agt-Rickauer, Hentschel, Filipowska, & Sack (2016) presented a case study about quantitative analysis of art market using ontologies, NER and ML. Deep Convolutional Neural Network (CNN) is the ML technique and ImageNet is the well-known dataset that applied in this research. This research predicts a style of a painting to overcome issues with missing style information successfully.

Suryana & Ipnuwati (2016) presented a vector ML method for text mining Indonesian social media NER. Support Vector Machine (SVM) is the ML technique that applied in this research. Large Indonesia social media datasets were used to carry out the experiment in this research. The overall recall measure achieved in this research is ideologies (65.27%), social (56.19%), cultural (55.56%), political (61.99%), economic (48.15%) and Defence and Security (63.03%).

Mohammed & Bagash (2017) presented a biomedical NER using ML classifiers and rich feature set. Three classification techniques such as Naïve Bayes (NB), K Nearest Neighbour (KNN) and Decision Trees (DT) classifiers were compared using different feature sets in order to synthesize a more accurate classification procedure. Result shows that the KNN trained with suitable features is more suitable to recognize named entities of biomedical texts than other models.

Salah & Qadri binti Zakaria (2017) presented a comparative review of ML for Arabic NER. Most of the researches focused on supervised Arabic NER ML and implemented many ML models such as CRF, SVM, DT, Maximum Entropy (ME) Models, and Artificial Neural Network (ANN). On the other hand, only a little research focused on semi supervised whereas no research focused on unsupervised technique for Arabic language yet.

Kanimozhi & Manjula (2017) presented a CRF based ML approach for biomedical NER. CRF is the ML technique that applied in this research. Two datasets were used in these experiments to compared three named entity position encoding scheme. The first dataset contain twenty thousand sentences whereas the second dataset contain only 3655 sentences. The highest results achieved in this research are 82.29% of Recall, 87.93% of Precision and 85.02% of F-measure.

Taşpınar, Ganiz & Acarman (2017) presented a feature based simple ML approach with word embeddings to NER on tweets. Few ML involved in this research such as DT, SVM, KNN, Logistic Regression (LR), ExtraTreeClassifier. MultinomialNB and BernoulliNB. The 2016 twitter dataset used in this research was provided by Rizzo, van Erp, Plu & Troncy (2016). The performance of the study in this research achieved 56% of Recall, 71% of Precision and 58% of F1 metric.

Bhandari, Chowdri, Singh & Qureshi (2017) presented resolving ambiguities in NER using ML. NB is the ML technique that applied in this research and the dataset containing approximately three thousand sentences was collected from Wikipedia. The highest result achieved in this research is 85.59% for F-score.

Wu, Lu, Hyder, Zhang, Quinney, Desta & Li (2018) presented an integrated ML and lexicon mapping NER Method for drug metabolite. Sequential minimal optimization (SMO) is the ML technique that applied in this research. Two datasets involved in this research which are internal and external dataset. Internal dataset achieved 77% of recall, 89% of precision and 83% of F-measure whereas the external dataset achieved 85% of recall, 86% of precision and 86% of F-measure.

Zhang, Wang, Hou & Li (2018) presented a clinical NER from Chinese electronic health records via ML methods. CRF method and bidirectional long short-term memory (LSTM)-CRF are the ML technique that applied in this research. The dataset that used for this research is the benchmark dataset with human annotated Chinese EHRs which provided by the China Conference. The results for CRF achieved in this research are 87.09% of Recall, 92.03% of Precision and 89.49% of F1 score whereas the results for bidirectional LSTM-CRF achieved in this research are 89.74% of Recall, 91.12% of Precision and 90.43% of F1 score.

Gorinski, Wu, Grover, Tobin, Talbot, Whalley & Alex (2019) presented a comparison of rule-based and ML approaches NER for electronic health records. LSTM-CRF is the ML technique that applied in this research. This research consists of three datasets and named as ESS, Tay and TayExt. The results show that the gap between the rule-based and ML approaches NER for electronic health records is relatively small and ranging between 0.03 and 0.06 in F-score. Table 2.5 provides a summary of the existing works of ML in NER.

Table 2.5: Existing Researches of ML in NER

Authors	Research Focus	Technique	Remarks
(Vijay & Sridhar, 2016)	A ML approach to NER for the travel and tourism domain	CRF	Recall 82%, Precision 85%, Accuracy 82%, F-measure 83%
(Khanam, Khudhus & Babu, 2016)	A NER using ML techniques for Telugu language	CRF	No result has been mentioned.
(Filipiak, Agt-Rickauer, Hentschel, Filipowska, & Sack, 2016)	A case study about quantitative analysis of art market using ontologies, NER and ML	CNN	No result has been mentioned.
(Suryana & Ipnuwati, 2016)	A vector ML method for text mining Indonesian social media NER	SVM	Recall measure, ideologies (65.27%), social (56.19%), cultural (55.56%), political (61.99%), economic (48.15%) and Defence and Security (63.03%)
(Mohammed & Bagash, 2017)	Biomedical NER using ML classifiers and rich feature set	NB, KNN and DT	KNN trained with suitable features is more suitable to recognize named entities of biomedical texts than other models.
(Salah & Qadri binti Zakaria, 2017)	Comparative review of ML for Arabic NER	CRF, ME, SVM, DT, HMM, and ANN	Many works focused on supervised Arabic NER ML studies.
(Kanimozhi & Manjula, 2017)	A CRF based ML approach for biomedical NER	CRF	Recall 82.29%, Precision 87.93%, F-measure 85.02%
(Taşpınar, Ganiz & Acarman, 2017)	Feature based simple ML approach with word embedding to NER on tweets	LR	Recall 56%, Precision 71%, F1 0.58%
(Bhandari, Chowdri, Singh & Qureshi, 2017)	Resolving ambiguities in NER using ML	NB	F-score 85.59%
(Wu, Lu, Hyder, Zhang, Quinney, Desta & Li, 2018)	An integrated ML and lexicon mapping NER Method for drug metabolite	SMO	Recall 77%, Precision 89%, and F-measure 83% for the internal testing set
(Zhang, Wang, Hou & Li, 2018)	Clinical NER from Chinese electronic health records via ML methods	CRF and LSTM-CRF	CRF (Recall 87%, Precision 92%, F1 score 89%) LSTM-CRF (Recall 90%, Precision 91%, F1 score 90%)
(Gorinski, Wu, Grover, Tobin, Talbot, Whalley & Alex, 2019)	Comparison of rule-based and ML approaches NER for electronic health records	Rule-based and LSTM-CRF	Ranging between 0.03 and 0.06 points in F1

2.8 Evaluation Methods

When training an ML model, the main concern is about how accurate the ML model can achieve to provide the best classifications. According to a previous study by Olson & Charles (2009), a classification system with accuracy of 80% and above can be considered as acceptable. There are several kinds of metrics available to evaluate the trained model, which consists of classification accuracy, cross-validation technique and confusion matrix with precision, recall and F-Measure.

2.8.1 Classification Accuracy

Classification accuracy is one of the metrics for evaluating classification models. The classification accuracy is calculated by using the number of correct classifications made, divided by the total number of classifications made, multiplied by 100 to turn it into a percentage. This method is simple and commonly used for evaluating the performance of the ML (Vijay & Sridhar, 2016).

2.8.2 Confusion Matrix

Confusion matrix, also known as contingency table, is an unambiguous way to describe the performance of a trained model or a classifier. According to Vijay & Sridhar (2016), performance of a system can be analyzed using confusion matrix with various measures such as precision, recall and F-Measure. Furthermore, they highlighted that the evaluation of the NER task is done by using precision, recall and F-Measure. Confusion matrix is useful for calculating precision, recall and F-measure given the classified labels from a model (Taşpınar, Ganiz & Acarman, 2017). For binary classification issues, the table has two columns and two rows as shown in Table 2.6. Across the top are the classified class labels and down the side are the actual class labels.

Table 2.6: Truth Table Confusion Matrix

	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

2.8.3 Cross-Validation Technique

According to a previous study by Seni & Elder (2010), cross-validation can be a good metric to evaluate the performance of a model. Cross-validation is used to test the model's ability to classify a new set of data that was not used in estimating it. According to a previous study (Ross et al., 2009), 10-fold cross-validation (k = 10) is the most commonly used cross-validation technique, which means the cross-validation process is repeated 10 times and 10% of subsamples are used exactly once as the validation data. Several ML NER systems have been implemented and evaluated using cross-validation technique such as (Kanimozhi & Manjula, 2017), (Wu, Lu, Hyder, Zhang, Quinney, Desta & Li, 2018) and (Zhang, Wang, Hou & Li, 2018).

2.8.4 Listening Evaluation

According to Mustafa, Don, Ainon, Zainuddin & Knowles (2014), a listening test can be used to evaluate the intelligibility and naturalness of the synthetic speech produced from a system. In this research, listening evaluation of the synthetic speech is necessary to show the level of intelligibility achieved by the proposed Malay TTS system. Malay listeners will be involved in this listening evaluation since the proposed TTS focuses on Malay language.

2.9 Summary

This chapter describes comprehensive information about this research through the review of literature, such as limitation of existing TTS system in synthesizing text for numerical input, suitable classification techniques, and evaluation methods. Handling numbers is the weakest ability in the text normalization process of the existing TTS systems. In order to overcome this problem, ML for text classification can be an appropriate solution.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Overview

This chapter describes the methodology utilized in this research in order to meet the entire requirements listed in the research objectives. The main focus of this research is to propose a suitable technique to automatically classify the numerical formats in Malay text. The research methodology consists of identification of research problem, dataset collection, proposed solution, development, evaluation, and discussions of the major findings, in order to fulfill the aim of this research. Table 3.1 shows the details of the process flow.

Table 3.1: Details of the Process Flow

Methodology Key Step	Methodology Process Description
Identification of Research Problem	Review existing journals, reports and thesis to find the limitation of TTS systems → Review the accuracy of the existing TTS systems on their ability to convert numbers into speech → To find a suitable classification technique
Dataset Collection	Extraction of text from online newspaper → Filtration of the extracted database on the selected numerical formats → Build a module for labeling
Design and Development of Classification System	To design a suitable classification technique for classifying the numerical formats → Selection of a suitable classification technique → Development of the proposed classification system that incorporates the technique
Evaluation	Training and testing → Analysis of the result → Test data used for evaluation purpose

3.2 Research Problems and Solutions

Existing researches on Malay NER (Alfred, Mujat, & Obit, 2013; Alfred, Leong, On, & Anthony, 2014; Salleh, Asmai, Basiron, & Ahmad, 2017) focus on identifying name-entities from unstructured textual data. Every word in the text was classified to its entity type based on the word itself, without taking into consideration of the context. However, numerical format in Malay language are in fact highly reliable on the context as according to the result of feature occurrence analysis shown in Table 3.2. Performance of classification is highly improved when context was taking into consideration for numerical format classification.

According to Yaser (2016), numerical classification by using the ML technique is a viable solution to overcome the problems of text normalization. ML technique has the ability to classify the format of numbers through the training process, which can make the system more intelligible compared to traditional rule-based technique. With this reference, the proposed classification system is developed using ML technique instead of the current utilized rule-based technique.

Katz, Shapira, Ofek, Bar-Zev, & Negev, (2014) stated that context-based technique can be used for text classification. Context-based technique was used to identify key terms and contexts to generate context-based features as additional features into their model, effectively classified text with high accuracy. With this reference, context-based technique is a suitable feature extraction technique which can be applied in this research to generate context-based features in order to classify the numerical formats effectively.

According to Alfred, Leong, On, & Anthony, (2014), collecting and creating a large annotated dataset for Malay language is very time-consuming. The lack of annotated corpus resources for Malay language that can be used as a training data is the main reason for their research using rule-based technique instead of ML technique.

Salleh, Asmai, Basiron, & Ahmad, (2017) have proposed a Malay NER using CRF due to a lack of research has been done to analyze the recognition of Malay entities. The created model was tested with data that consisted of 1995 tokens word that were generated from Bernama news. The evaluation results for NER effectiveness were measured by precision, recall and F1-score (F-measure). This F1-score or F-measure that can be overall perspective of evaluation process can be improved by undergoing another experiment by increasing the training dataset for a better result. This is because the rate of increase in accuracy for recognizing entities depending on the model trained and features sets used.

Limited exploration has been done on NER for numerical expression. A research that came close to this purpose is done by Saleh, Tounsi, & van Genabith (2011) that present a ML technique applied to identify different forms of numerical expressions in Arabic language and another ML technique to identify Arabic temporal. The classifier used in this research is SVM classifier and evaluation metrics are detection performance and bracketing performance. For this research, the system achieving a F-score of 88.5%.

3.3 Dataset Collection

A Malay datasets is needed to perform NER task in this research. The development of the proposed system consists of data collection, data processing, feature extraction, classification, and evaluation. Data collection is not easy because there is no existing and compatible dataset that can be used to classify numbers in Malay text, so data collection can be a very challenging task and time-consuming.

A compatible dataset is an important element in the training process of ML. The dataset for supervised ML algorithms is difficult to produce as it is time-consuming in labeling the data. Since there is no previous work that focused on classification of numerical

format for Malay text, there is no existing dataset that can be applied in this research. In other words, a compatible dataset with a proper label should be created before carrying out the development of the proposed classification system.

Pre-processing of the dataset, such as filtering and extracting relevant data through online source is compulsory in order to create a compatible dataset. The extracted data should contain the numerical formats which are selected for this research, whereas the numerical formats must be labeled correctly based on the meaning of the numerical in the sentences. The selected numerical formats consist of date, time, phone number, currency, measurement, and percentage. A complete process of data collection is explained in detail in chapter 4.

3.4 Design and Development of the NER for Numerical Format Classification System

Named-Entity Recognition (NER) for numerical expression addresses the classification of number format in unstructured text. ML is the technique for the task of NER for numerical format. There are three stages in the development of numerical format classification system using ML technique. Stage one extracts the context-based features of numbers from the dataset. Some information is needed in order to train the classifiers, such as features and its labels. Stage two is training of model and classification of dataset. Features of numbers and its labels collected in stage one are used to train a model. The model will learn and recognize the relationship between the features and labels. Stage three is classification and evaluation. The trained model will be used to classify some of the available data in order to evaluate its accuracy and efficiency as shown in Figure 3.1.

There is a variety of suitable classification techniques to be applied in classifying the numerical format. The classification technique with the highest accuracy will be selected as a suitable technique in order to fulfill the requirement of this research.

Standard ML framework has two phases, a training phase and a classification phase. This framework is applied in the proposed classification system which is shown in Figure 3.1. In the training phase, the dataset which comprises numerical data formats and related labels are used to train the classifier. Features are extracted based on the specified context which is meaningful to the numerical format. Labels and features are forwarded to ML techniques in order to train the model.

Classification phase classifies the labels of the new numerical input by using the model that was trained in the training phase. The same feature extractor in the training phase is applied to extract the features from the context. The extracted features will be forwarded to the trained model in order to classify the label that can represent the numerical format.

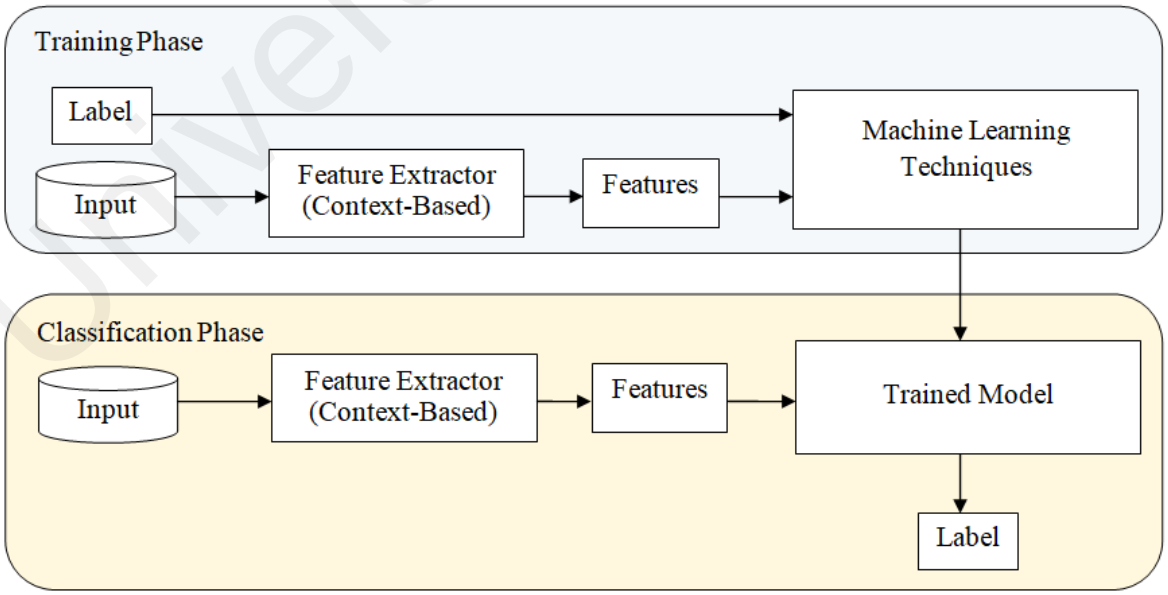


Figure 3.1: NER Framework using Context-Based Numerical Feature Extraction and Machine Learning Techniques

3.4.1 Context-Based Numerical Feature Extraction

In this research, context-based technique was used as the feature extractor. The feature extractor was implemented by using python programming language. The general idea of the context-based technique is that the system extracts the keywords and symbols nearby to the number to find out the feature of the number. The keywords consist of collective nouns, measurement units, currency code, etc., which are usually positioned next to the detected numbers. On the other hand, the symbols involve all punctuation which consists of currency symbols, commas, full stops, etc.

The context-based model is used for text categorization, where the words nearby to the numbers are used as features for training purposes. The feature extraction program will extract two words, before and after the number, to find out the features in the sentence in order to classify the format of the numbers. Two words before and after the numbers are collected because most of the keywords in the Malay language appear in between this range. Figure 3.2 to Figure 3.7 show the samples of a sentence with different format.

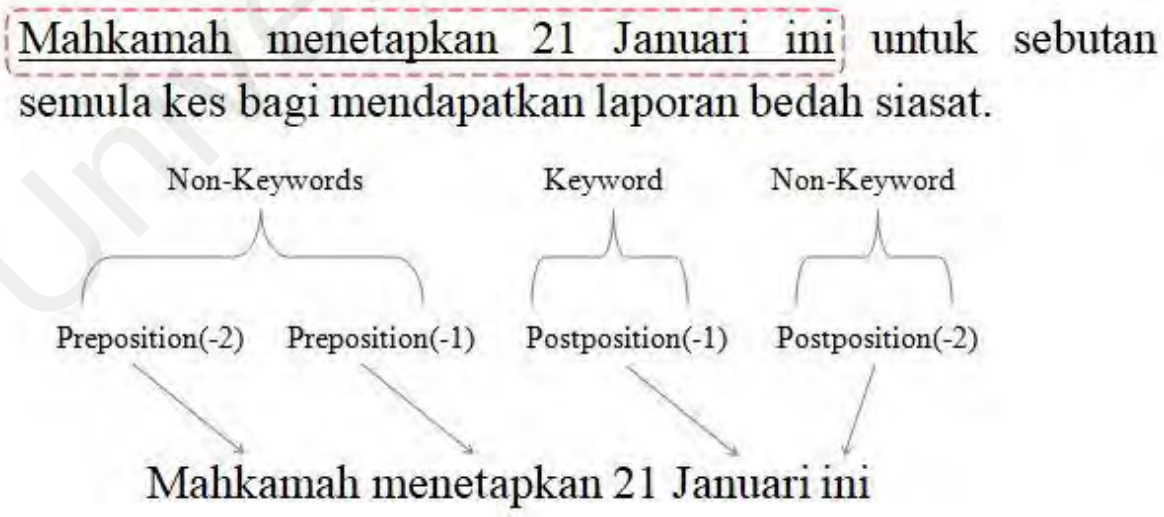


Figure 3.2: Sample of Date Format Sentence

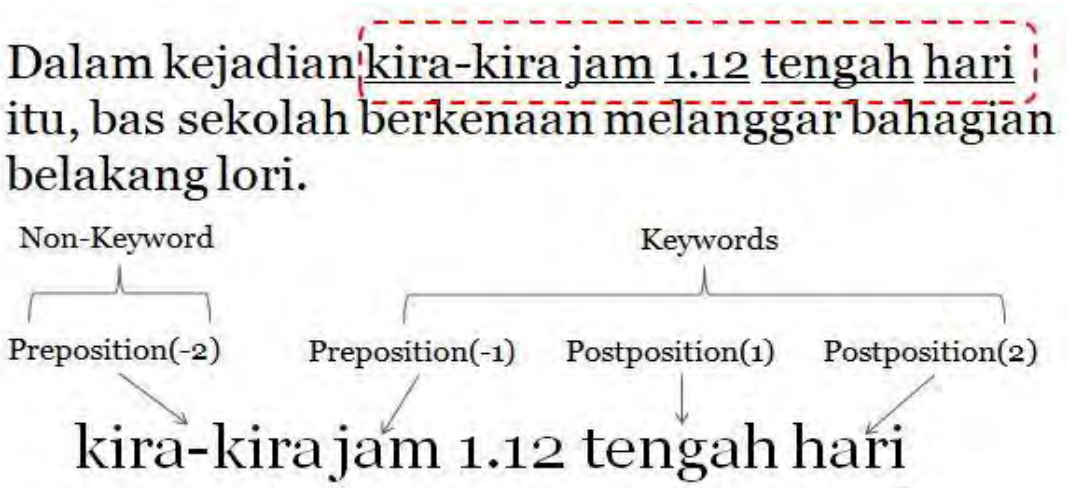


Figure 3.3: Sample of Time Format Sentence

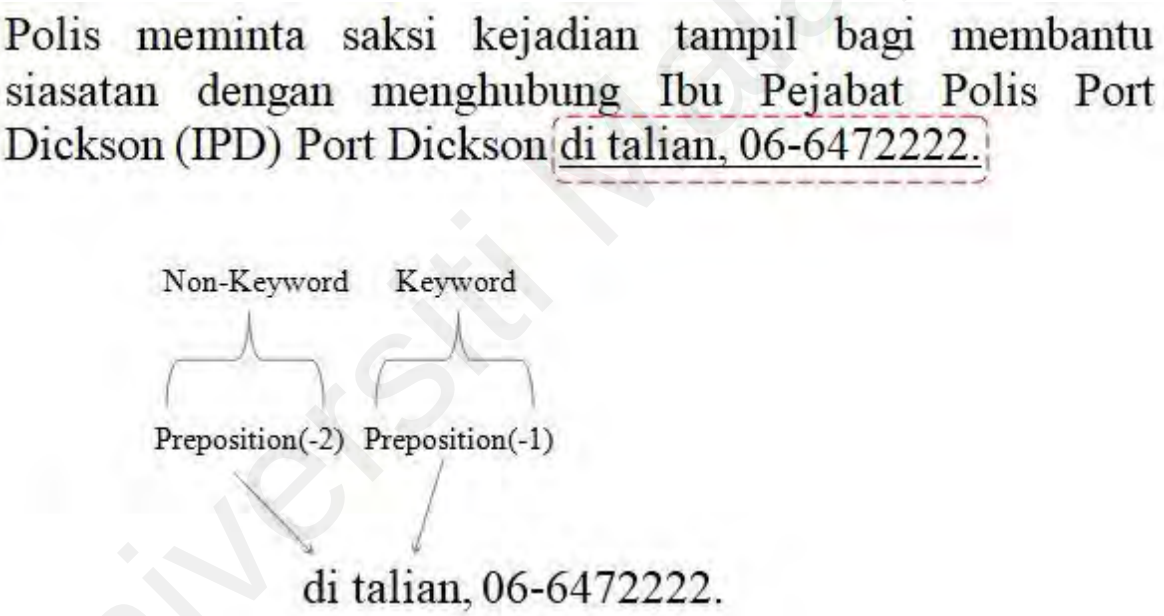


Figure 3.4: Sample of Phone Number Format Sentence

Hanya Menteri Kewangan mempunyai kuasa untuk meluluskan atau tidak pemindahan wang RM19.4 bilion ke dalam Tabung Bayaran Balik GST.

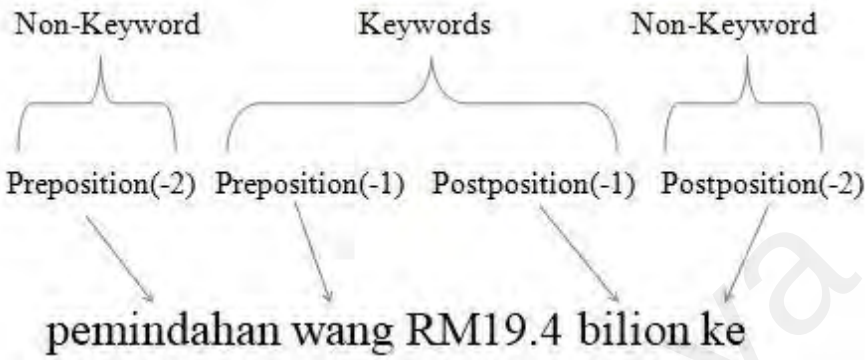


Figure 3.5: Sample of Currency Format Sentence

Angka korban bencana gempa bumi dan tsunami di Indonesia kini mencecah 832 orang serta dijangka terus meningkat, kata agensi bencana hari ini.

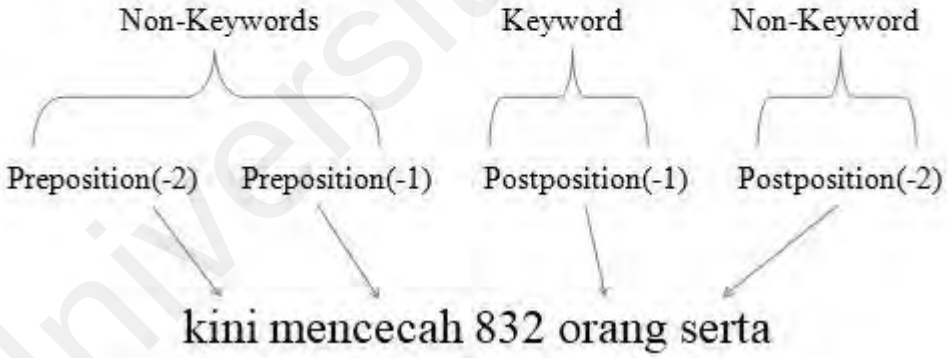


Figure 3.6: Sample of Measurement Format Sentence

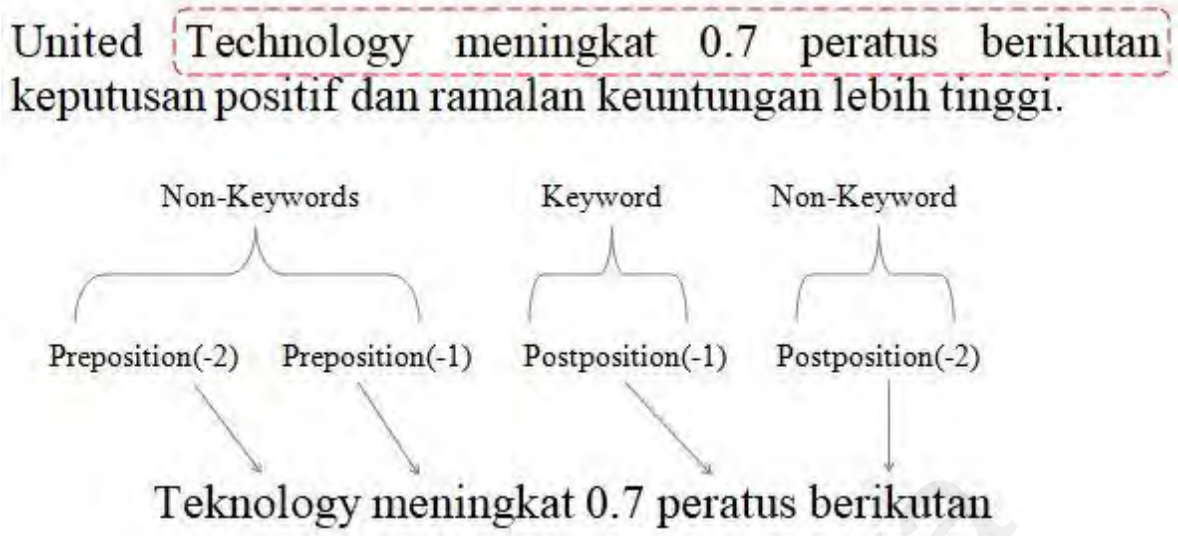


Figure 3.7: Sample of Percentage Format Sentence

There are a total of 571 labeled numerical formats that have been applied in this study. Based on the analysis of the datasets, feature occurrence in preposition 1 (1st word before the number), postposition 1 (1st word after the number) and postposition 2 (2nd word after the number) is significantly high. On the other hand, features occurrence in preposition 2 (2nd word before the number) is lower compared to preposition 1, postposition 1 and postposition 2. Preposition 2 is selected because the features in preposition 2 are highly related to the numbers compared to preposition 3 (3rd word before the number), preposition 4 (4th word before the number), preposition 5 (5th word before the number), preposition 6 (6th word before the number), postposition 3 (3rd word after the number), postposition 4 (4th word after the number), postposition 5 (5th word after the number) and postposition 6 (6th word after the number). Table 3.2 shows the analysis of the feature occurrence in a sentence.

Table 3.2: Analysis of Feature Occurrence in a Sentence

	Preposition							Postposition					
Sequence	6	5	4	3	2	1		1	2	3	4	5	6
No. of Occurrence	7	18	17	13	14	173		361	64	9	16	9	9
Percentage (%)	1.23	3.15	2.98	2.27	2.45	30.3		63.22	11.21	1.58	2.8	1.58	1.58

3.4.2 Machine Learning Classifiers for Numerical Format Classification

Yaser (2016) has proposed and experimented four ML classifiers based on the recommendation from the existing literature for classifying text. These four shortlisted ML classifiers were SVM, KNN, LDA and DT. These four ML classifiers were adopted in this research due to its outstanding function for classifying text. Besides that, the same ML classifiers were used as a benchmark for this research.

3.5 Evaluation Method

3.5.1 Cross Validation

Cross-validation can be a good metric to evaluate the accuracy of a classifier, which has been proposed by Yaser (2016). He has evaluated the performance of the proposed classifiers using cross-validation which is a popular method due to its simplicity to understand and its less biased estimate of the model skill.

Below shows the equations of precision, recall and F-Measure.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

3.5.2 Intelligibility Test

For any research on TTS system, the intelligibility test is crucial just like the recognition accuracy for automatic speech recognition system. The performance and ability of speech synthesis system in correctly synthesize speech is evaluated based on major attribute of the intelligibility. One of the main techniques for evaluating the performance of a speech synthesis system is to have human listeners to listen to synthetic speeches and response to specific questions objectively or subjectively (Zhi-Zheng Wu et al., 2010; Mustafa et al., 2014). The intelligibility of speech synthesis system is usually expressed as a percentage of words, sentences or phonemes correctly identified by a group of listeners. In this research, a listening evaluation was conducted on the synthetic speech generated by the Malay TTS system developed by the University Malaya, Speech Technology research team. More detail related to listening evaluation is discussed in chapter 5.

3.6 Summary

This chapter describes the methodology utilized in this research in order to meet the requirements listed in the research objectives. Existing TTS systems were unable to synthesize numbers as expected and affect the intelligibility of the systems. In order to overcome the limitation of the TTS systems, this research proposed to develop a suitable technique to automatically classify the numerical formats in Malay text.

CHAPTER 4: DEVELOPMENT OF CONTEXT-BASED NUMERICAL FORMAT CLASSIFICATION SYSTEM

4.1 Overview

This chapter describes the processes of developing the dataset, which includes collecting, filtering and labeling data processes. This chapter also depicts the development of the proposed numerical format classification system, which consists of three stages, which are data processing, feature extraction, and classification.

4.2 Data Collection Processes

This section describes the processes involved in developing the dataset that was used for the development of the proposed classification system. Figure 4.1 shows the flowchart of the overall processes for developing the dataset. A complete dataset in this research contains three important elements, which consist of a sentence, number format, and label. The dataset is built by performing several processes, which are data collection, data filtration, and data labeling.

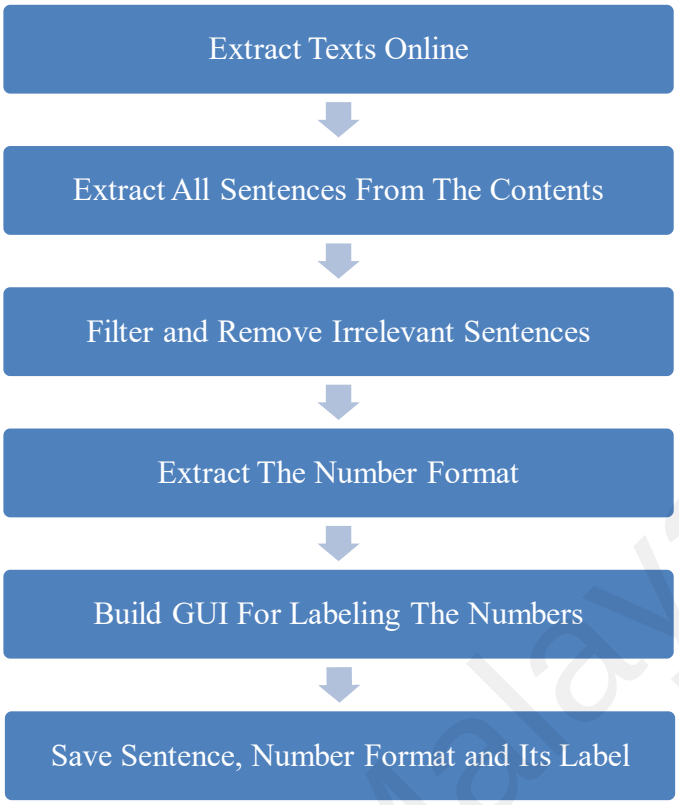


Figure 4.1: The Flowchart of Development of Dataset

4.2.1 Resources and Tools

Several resources and tools have been used in this research to develop the appropriate dataset. In this research, data are collected through online sources, Berita Harian, Utusan Online, MalaysiaKini and Harian Metro. Table 4.1 shows the list of Malay Online News and Newspapers which used in this research.

Table 4.1: List of Malay Online News and Newspapers

Online Media	**Website**
Berita Harian	https://www.bharian.com.my/
Utusan Online	https://www.utusan.com.my/
Malaysiakini	https://www.malaysiakini.com/my
Harian Metro	https://www.hmetro.com.my/

The collected data are all in the Malay language. These online resources provided a lot of text along with numbers that are commonly used by Malaysians. NetBeans IDE is an integrated development environment for Java programming language. It has been used to develop the program for data extraction, data filtration, and data labeling process in this research.

4.2.2 Scope of Dataset

There are a lot of number formats in the real world and it might increase in the future. This research focuses on six categories of number formats only, which are date, time, phone number, currency, measurement, and percentage, since it is not possible to cover the entire number format in the real world. In order to develop an appropriate dataset, the collected data should relate to the six categories of number formats. Besides that, each number format has multiple sub-categories. The dataset should contain enough data or at least three of each sub-category for training and testing purposes. Lastly is to exclude long sentences, which have more than twenty words to prevent too many unnecessary data and affect the performance of the system. Table 4.2 shows the sub-categories of each number format.

Table 4.2: Sub-categories of Each Number Format

Number Type	**Sub-Catrgories**
Date	xx month
	xx month xx
	xxxx-xx-xx
	xx/xx/xxxx
	xxxx-xx
	tahun xxxx
	ke-xx
	xxxxan
	xxan-xxan
	xxan dan xxan
	xxxx
Time	xx pagi/petang/malam/tengah hari
	x.x pagi/petang/malam/tengah hari
	xx.xx pagi/petang/malam/tengah hari
	jam x pagi/petang/malam/tengah hari
	jam x.xx pagi/petang/malam/tengah hari
	jam x:xx pagi/petang/malam/tengah hari
	pukul x pagi/petang/malam/tengah hari
	pukul x.xx pagi/petang/malam/tengah hari
	pukul x:xx pagi/petang/malam/tengah hari
	xx saat/minit
	jam xxx
	xxx jam
	x hingga x jam
	xH
	xx:xx:xx am/pm
Phone Number	xx-xxxxxxx
	xxxx-xx-xxxx
	+xx-xx-xxxxxx
	+(xxx)xxxxxxxx
Currency	RM xx
	RMxx
	$xx
	AS$/ASD$/US$/yen/baht/rupiah/Rp/Swiss francs xx
Measurement	xx (collective nouns, measurement units)
Percentage	peratus
	%
	xx hingga xx peratus
	xx – xx peratus
*Note: xx indicate any number	

4.3 Data Filtration

One thousands of texts have been collected from online sources such as online news in Malaysia. After that, all sentences are split and extracted from the texts using a program built by using Netbeans with the Java programming language. Sentences with or without numbers, have been extracted from the texts. Another filtration program is built to check and filter the sentences. The filtration program will remove irrelevant sentences which do not have any number in the sentence, whereas the sentences which contain numbers have another level of filtration in order to fulfill the scope of the dataset as mentioned above. Finally, a total of 571 sentences are saved into an excel file.

4.4 Data Labeling

The training process of supervised ML requires datasets that have been labeled correctly. Data labeling process has to label extracted numbers based on the actual meaning of the number in the sentence. A program was developed to extract number format from the collected sentences automatically for ease of labeling. On the other hand, the labeling of number format with representative numerical value was done manually. The numerical values serve as label for different number format. Table 4.3 shows the representative label values of each number type, whereas Table 4.4 shows the examples of actual sentences and the number format extracted.

Table 4.3: Number Type and Label

Number Type	Date	Time	Phone Number	Currency	Measurement	Percentage
Label	0	1	2	3	4	5

Table 4.4: Examples of Actual Sentences and the Number Format Extracted

Sentences	Number Extracted	Label	Number Type
Mereka ditahan Agensi Penguatkuasaan Maritim Malaysia (APMM) dan didakwa pada 21 Disember tahun lalu.	21	0	Date
Latihan bermula awal pagi dan pemain biasanya tiba sebelum jam 9.00 pagi.	9.00	1	Time
Polis meminta saksi kejadian tampil bagi membantu siasatan dengan menghubung Ibu Pejabat Polis Port Dickson (IPD) Port Dickson di talian, 06-6472222.	06-6472222.	2	Phone Number
Gaji pokok seorang menteri kabinet ialah RM 14,907.20 sebulan.	14,907.20	3	Currency
Lebih daripada 400 orang disahkan terbunuh, kebanyakannya dihanyutkan ombak besar akibat tsunami yang dicetuskan oleh gempa bumi kuat di Sulawesi.	400	4	Measurement
Lebih 70 peratus pelajar di Sekolah Menengah Agama Persekutuan (SMAP) Kajang diarah pulang ke rumah masing-masing selepas dijangkiti penyakit beguk, baru-baru ini.	70	5	Percentage

4.4.1 Punctuation Marks

Punctuation marks such as full stop, comma, colon, hyphen, brackets and slash in numbers can be used to distinguish number format. For instance, currency symbols such as the dollar sign "$" commonly appear together with the number in a sentence indicate the number is currency format. Punctuation marks act as features of the number in this research in order to increase the accuracy of the proposed system. Table 4.5 shows the list of punctuation marks commonly appear in the extracted numbers in this research.

Table 4.5: List of Punctuation Marks

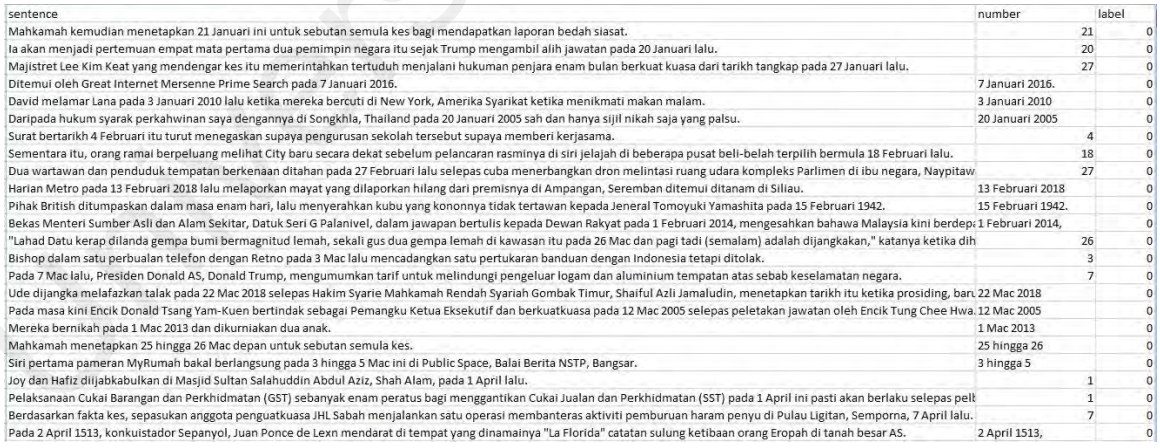
Punctuation Marks	Full Stop	Comma	Colon	Hyphen	Bracket	Slash	Dollar Sign
	.	,	:	-	( )	/	$

4.5 Text-Numbers Dataset in Details

The completed dataset contains three important elements, which are the source sentences, extracted number formats, and representative labels. The dataset remained as 571 of labeled number formats after undergoing the data filtration process. The dataset was segmented into six categories, which include Date, Time, Phone Number, Currency, Measurement, and Percentage. The 571 of well-labeled dataset in excel file is available in Appendix. Table 4.6 shows the number of data that were collected for each number type, along with its representative label, whereas Figure 4.2 shows the well-labeled dataset in excel file.

Table 4.6: Number Formats for Each Number Type

Number Type	Date	Time	Phone Number	Currency	Measurement	Percentage
No. of Data	108	129	18	110	100	106
Label	0	1	2	3	4	5



sentence	number	label
Mahkamah kemudian menetapkan 21 Januari ini untuk sebutan semula kes bagi mendapatkan laporan bedah siasat.	21	0
Ia akan menjadi pertemuan empat mata pertama dua pemimpin negara itu sejak Trump mengambil alih jawatan pada 20 Januari lalu.	20	0
Majistret Lee Kim Keat yang mendengar kes itu memerintahkan tertuduh menjalani hukuman penjara enam bulan berkuat kuasa dari tarikh tangkap pada 27 Januari lalu.	27	0
Ditemui oleh Great Internet Mersenne Prime Search pada 7 Januari 2016.	7 Januari 2016.	0
David melamar Lana pada 3 Januari 2010 lalu ketika mereka bercuti di New York, Amerika Syarikat ketika menikmati makan malam.	3 Januari 2010	0
Daripada hukum syarak perkahwinan saya dengannya di Songkhla, Thailand pada 20 Januari 2005 sah dan hanya sijil nikah saja yang palsu.	20 Januari 2005	0
Surat bertarikh 4 Februari itu turut menegaskan supaya pengurusan sekolah tersebut supaya memberi kerjasama.	4	0
Sementara itu, orang ramai berpeluang melihat City baru secara dekat sebelum pelancaran rasminya di siri jelajah di beberapa pusat beli-belah terpilih bermula 18 Februari lalu.	18	0
Dua wartawan dan penduduk tempatan berkenaan ditahan pada 27 Februari lalu selepas cuba menerbangkan dron melintasi ruang udara kompleks Parlimen di ibu negara, Naypitaw	27	0
Harian Metro pada 13 Februari 2018 lalu melaporkan mayat yang dilaporkan hilang dari premisnya di Ampangan, Seremban ditemui ditanam di Siliau.	13 Februari 2018	0
Pihak British ditumpaskan dalam masa enam hari, lalu menyerahkan kubu yang kononnya tidak tertawan kepada Jeneral Tomoyuki Yamashita pada 15 Februari 1942.	15 Februari 1942.	0
Bekas Menteri Sumber Asli dan Alam Sekitar, Datuk Seri G Palanivel, dalam jawapan bertulis kepada Dewan Rakyat pada 1 Februari 2014, mengesahkan bahawa Malaysia kini berdep	1 Februari 2014,	0
"Lahad Datu kerap dilanda gempa bumi bermagnitud lemah, sekali gus dua gempa lemah di kawasan itu pada 26 Mac dan pagi tadi (semalam) adalah dijangkakan," katanya ketika dih	26	0
Bishop dalam satu perbualan telefon dengan Retno pada 3 Mac lalu mencadangkan satu pertukaran banduan dengan Indonesia tetapi ditolak.	3	0
Pada 7 Mac lalu, Presiden Donald AS, Donald Trump, mengumumkan tarif untuk melindungi pengeluar logam dan aluminium tempatan atas sebab keselamatan negara.	7	0
Ude dijangka melafazkan talak pada 22 Mac 2018 selepas Hakim Syarie Mahkamah Rendah Syariah Gombak Timur, Shaiful Azli Jamaludin, menetapkan tarikh itu ketika prosiding, ban	22 Mac 2018	0
Pada masa kini Encik Donald Tsang Yam-Kuen bertindak sebagai Pemangku Ketua Eksekutif dan berkuatkuasa pada 12 Mac 2005 selepas peletakan jawatan oleh Encik Tung Chee Hwa.	12 Mac 2005	0
Mereka bernikah pada 1 Mac 2013 dan dikurniakan dua anak.	1 Mac 2013	0
Mahkamah menetapkan 25 hingga 26 Mac depan untuk sebutan semula kes.	25 hingga 26	0
Siri pertama pameran MyRumah bakal berlangsung pada 3 hingga 5 Mac ini di Public Space, Balai Berita NSTP, Bangsar.	3 hingga 5	0
Joy dan Hafiz diijabkabulkan di Masjid Sultan Salahuddin Abdul Aziz, Shah Alam, pada 1 April lalu.	1	0
Pelaksanaan Cukai Barangan dan Perkhidmatan (GST) sebanyak enam peratus bagi menggantikan Cukai Jualan dan Perkhidmatan (SST) pada 1 April ini pasti akan berlaku selepas pelt	1	0
Berdasarkan fakta kes, sepasukan anggota penguatkuasa JHL Sabah menjalankan satu operasi membanteras aktiviti pemburuan haram penyu di Pulau Ligitan, Semporna, 7 April lalu.	7	0
Pada 2 April 1513, konkuistador Sepanyol, Juan Ponce de Lexn mendarat di tempat yang dinamainya "La Florida" catatan sulung ketibaan orang Eropah di tanah besar AS.	2 April 1513,	0

Figure 4.2: Well-Labeled Dataset in Excel File

4.6 Development of Classification System

This section discusses the development of the proposed classification system which include feature extraction, data transformation, and classification.

4.6.1 System Requirements and Tools

MATLAB is a multi-paradigm numerical computing environment that able to perform a variety of ML tasks. This research used MATLAB R2017b operated on Windows 7 operating system to implement and evaluate the proposed numerical format classification system. MATLAB was used for training and evaluation of the classifier. MATLAB library provided several classifiers, such as SVM, KNN, LDA, and DT, which was adopted for this research.

4.7 Feature Extraction

Context-based is a feature extraction technique that is used in this research. The general idea of the context-based technique is it is able to extract the keywords and symbols nearby to the number to find out the feature of the number. Figure 4.3 shows the proposed technique for classifying numerical format.

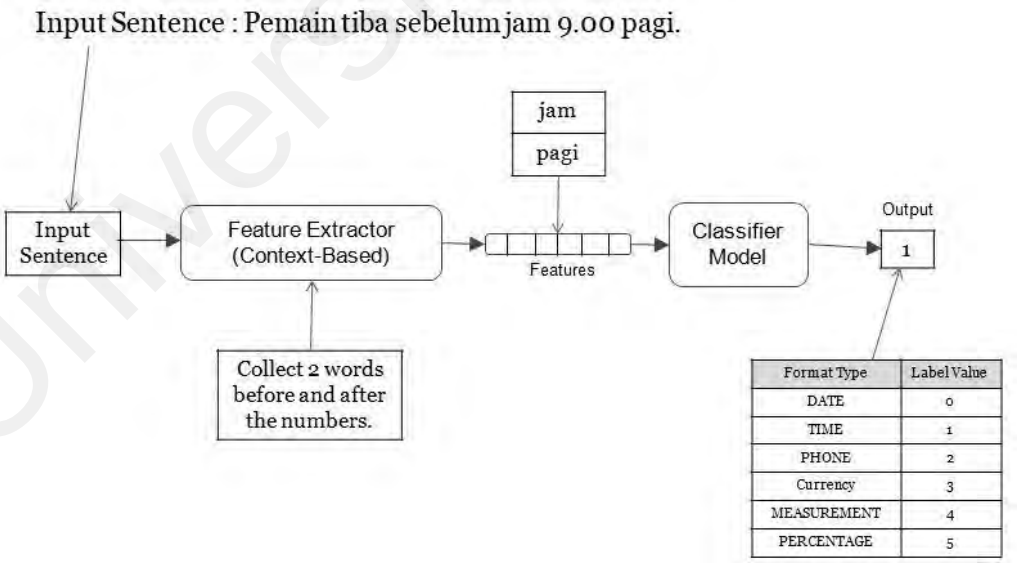


Figure 4.3: The Proposed Technique for Classifying Numerical Format

Python programming language is the tool used to implement the context-based features extractor. First of all, the system will collect the sentences from the uploaded CSV file which is converted from the excel file and then locate the number in the sentence by using the python's find function.

After the system has located the number in the sentence, it will split the sentence into separated words in order to collect two words before and after the located number and move the keywords into the string variable. Table 4.7 shows the list of the keywords, and Figure 4.4 shows the source codes that collect the two words before and after the number. The complete source code is available in Appendix.

Table 4.7: List of the Keywords

Types	**Keywords**
Date	januari, februari, mac, april, mei, jun, julai, ogos, september, oktober, november, disember, tarikh, hari, bulan, tahun, abad
Time	am, pm, masa, pukul, jam, minit, saat, pagi, petang, malam, tengah
Phone Number	tel, telefon, talian, menghubungi
Currency	ringgit, baht, birr, boliviano, dalasi, deutsche, dinar, dobra, dolar, euro, froint, franc, kip, koruna, krone, kroon, lats, lek, leone, lev, lira, litas, manat, metical, paun, penny, peso, rand, real, renminbi, rial, riyal, rubel, ruble, rufiyaa, rupee, rupiah, scheidemunze, somoni, tahil, tenge, tugrik, yen, yuan, zloty
Measurement Units	meter, kilometer, gram, kilogram, ampere, kelvin, mol, kandela, radian, steradian, hertz, newton, pascal, joule, watt, coulomb, volt, farad, ohm, siemens, weber, tesla, henry, celsius, lumen, lux, becquerel, gray, sievert, catal, koulomb
Collective Nouns	angkatan, baris, batang, bentuk, berkas, bidang, biji, bilah, blok, bongkah, botol, buah, buku, bungkus, butir, carik, cawan, cebis, cekak, gemal, colek, cubit, cucuk, das, deret, ekor, gelas, gelung, lingkar, genggam, gerombolan, gugus, gulung, gumpal, helai, hidang, hiris, iris, ikat, jambak, jambangan, jemput, kajang, kaki, kalung, kandang, kapur, kawan, kelompok, kepal, keping, kepul, kerat, ketul, kotak, kumpulan, kuntum, laras, lembar, longgok, mangkuk, naskhah, orang, pangsa, papan, pasang, pasukan, patah, petak, pintu, potong, pucuk, puntung, rangkai, rangkap, rawan, ruas, rumpun, perdu, sikat, sisir, suap, tandan, tangkai, teguk, timbun, tingkat, titik, titis, tongkol, ulas, untai, urat, utas
Percentage	peratus
Place Value	puluh, ratus, ribu, juta, bilion, trilion, kuadrilion, kuintilion
Others	nilai, bernilai, jumlah, berjumlah, harga, berharga, seramai

```
#First two words
pre_sentence = sentence[0:index].translate(translator).strip()
pre_words = pre_sentence.split(" ")
pre_words_2 = pre_words[-2:]
for temp in pre_words_2 :
    temp = temp.lower()
    if temp in prepositions:
        if temp not in F_words:
            F_words.append(temp)

#Last two words
post_sentence = sentence[index + len(numbers) :len(sentence)].translate(translator).strip()
post_words = post_sentence.split(" ")
post_words_2 = post_words[:2]
for temp in post_words_2 :
    temp = temp.lower()
    if temp in prepositions:
        if temp not in F_words:
            F_words.append(temp)
```

Figure 4.4: Collecting of Two Words Before and After the Number

The proposed classification system is focused on extracting keywords and symbols as the feature of the number. The system will extract the symbols within the number after the keywords of the number have been collected. Figure 4.5 shows the source code that extracts symbols and punctuations from the number.

```
#symbols and punctuations
symbols = re.sub("[0-9]|[a-z]|[A-Z]| |\t|\n","",numbers)
for s in symbols:
    if symbols not in F_words:
        F_words.append(symbols)
        F_symbols.append(s)
```

Figure 4.5: Extracting Symbols and Punctuations from the Number

After the keywords and symbols have been extracted, the system will collect the alphabet within the number. For instance, RM20 is a word which contains alphabet. The RM will be extracted from the word and treated as a feature of the number. Figure 4.6 shows the source code that extracts the alphabets within the numbers.

```
numberss = re.sub(r'[^a-zA-Z\s]', '', numbers)
numbersss = numberss.split(" ")
F_numbersss.append(numbersss)

valids= ""
for character in numbersss:
    if character.isalpha():
        valids += character
        valids = valids.lower()
        if valids not in F_words:
            F_words.append(valids)
            F_letters.append(valids)
```

Figure 4.6: Extracting Alphabets within the Numbers

In this research, there are a total of 103 different features that have been extracted from the 571 labeled data, which included keywords and symbols. Table 4.8 shows the extracted features from the labeled data. The outcome of feature extraction has a total of 571 rows which represent the total number of labeled data, and 103 columns which represent the total numbers of distinct features. From Figure 4.7, each column represents the feature that may exist in the sentence and it shows how many times the feature exists in the sentence.

Table 4.8: Extracted Features from the Labeled Data

Keywords	januari	februari	mac	april	mei	jun
	julai	ogos	september	oktober	november	disember
	tarikh	tahun	abad	pagi	petang	malam
	tengah	hari	jam	pukul	minit	saat
	masa	am	pm	tel	talian	nilai
	harga	ribu	juta	bilion	trilion	berharga
	jumlah	berjumlah	nilai	bernilai	rm	yen
	baht	rupiah	orang	meter	gram	kilogram
	butir	bungkus	peratus	seramai	dan	hingga
	ke	an	anan	andan	andanan	asd
	as	us	rp	kg	h	
Symbols	.	,	--,	--.	--	//,
	//.	//	-,	-.	-	(-)
	:	..	::	-,’’	---.	-/
	---	+--	+()	,.	,,	$.
	$	$,	$,,	$..	$,.	$,,
	.,	%	.%.	,%	%.	%,
	.%	.-.				

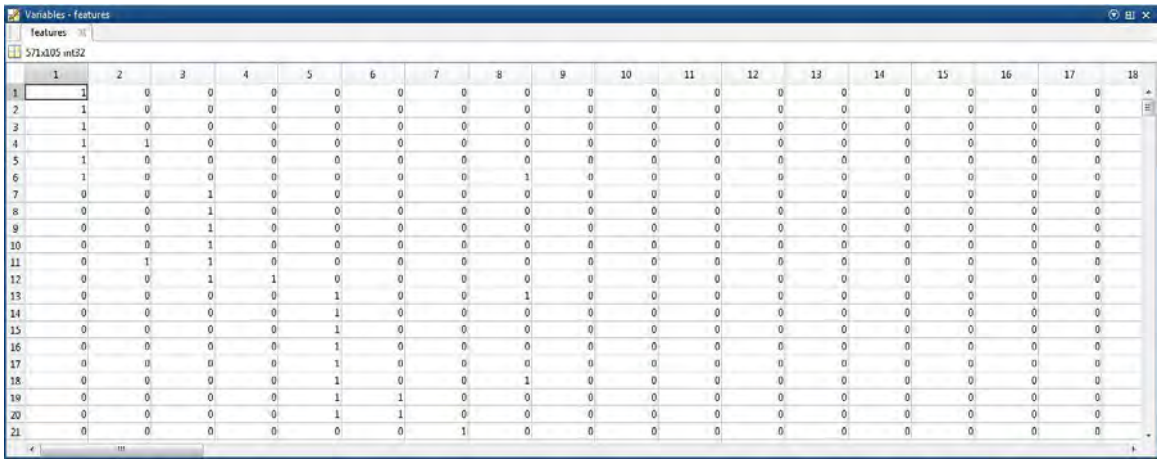


Figure 4.7: Output of Feature Extraction

4.8 Data Transformation

The output of feature extraction was transformed into MATLAB-style (.mat) file, which is compatible with MATLAB in order to proceed to the next step.

4.9 Classification

In this research, there are four main classification techniques that have been applied to classify the numerical format, which are SVM, LDA, KNN, and DT. Besides that, there are different techniques of kernel trick for SVM classifier that was used, such as Polynomial (poly), Linear, and Gaussian (RBF). On the other hand, for KNN, a single nearest neighbor (k=1) was chosen along with three nearest neighbors (k=3).

4.9.1 Training and Testing of the Classifiers

MATLAB was used to train and test each classifier separately in this research. Testing of the classifier was done using 10-fold cross-validation method, where 90 percent of the dataset was used for training purpose and the remaining 10 percent was used for testing purpose. Figure 4.8 shows the source code for testing of the classifier and Figure 4.9 shows the source code that calculates the accuracy as well as mean accuracy of training and testing.

```
k=10;
Indices = crossvalind('Kfold', Y, k); % 10 fold cross validation
for i=1:k,i
    TrainIndex=find(Indices~=i);
    TestIndex=find(Indices==i);
    Xtr=X(TrainIndex,:);
    Ytr=Y(TrainIndex,:);
    Xts=X(TestIndex,:);
    Yts=Y(TestIndex,:);

    [OutputTrain,OutputTest]=MultiClassClassifier(Xtr,Ytr,Xts,1);

    ResultKfold(i).OutputTrain=OutputTrain;
    ResultKfold(i).OutputTest=OutputTest;
    ResultKfold(i).Xtr=Xtr;
    ResultKfold(i).Ytr=Ytr;
    ResultKfold(i).Xts=Xts;
    ResultKfold(i).Yts=Yts;

end
```

Figure 4.8: Testing of the Classifier Using 10-Fold Cross-Validation Method

```
    Ytr= ResultKfold(i).Ytr;
    Yts= ResultKfold(i).Yts;

    eval( ['OutputTrain=ResultKfold(i).OutputTrain.' cell2mat(ClassMethod(EvalClassnum)) ';' ])
    eval( ['OutputTest=ResultKfold(i).OutputTest.'  cell2mat(ClassMethod(EvalClassnum))  ';'])

    AccuracyTrain(i)=sum( OutputTrain==Ytr)/numel(Ytr);
    AccuracyTest(i)=sum( OutputTest==Yts)/numel(Yts);

end

meanAccuracyTrain=mean(AccuracyTrain)
meanAccuracyTest=mean(AccuracyTest)
```

Figure 4.9: Source Code Used to Calculate the Accuracy As Well As Mean Accuracy of Training and Testing

4.10 Summary

This chapter describes in detail the processes that were carried out to complete the dataset needed for developing the proposed classification system. The processes include extracting the number format from each sentence and labeling each number format with a representative label. The completed dataset contains three important elements, which are the source sentences, extracted number formats, and representative labels. The collected corpus contains 571 labeled number formats. The corpus was segmented into six categories, which include Date, Time, Phone Number, Currency, Measurement, and Percentage. Furthermore, this chapter also shows the development of the proposed classification system in detail, which includes feature extraction, data transformation, and classification.

CHAPTER 5: EVALUATION, RESULTS AND DISCUSSIONS

5.1 Overview

This chapter focuses on the evaluation and results of the proposed numerical format classification system and the intelligibility of the synthesized speech which consists of numbers. Performance evaluation of the system was done by using testing data of 10-fold cross-validation and listening evaluation by native listeners. In the test, each classifier performs calculation on classification accuracy, precision, recall and F-Measure. Further, confusion matrix is used to describe the performance of a classification model in more detail.

5.2 Evaluation

This research has evaluated the performance of the proposed technique in classify the number formats as well as the intelligibility of the synthesized Malay speech which consist of different number formats.

5.2.1 Formulae of Classification Accuracy

Classification accuracy and cross-validation are popular performance measurement techniques that were used to evaluate the ML classifiers.

Classification accuracy results of each classifier are calculated using testing data of 10-fold cross-validation with the following classification accuracy formulae.

$$Classification\ Accuracy = \frac{The\ number\ of\ correct\ predictions\ made}{Total\ number\ of\ predictions\ made} \times 100$$

Classification accuracy is a metric for evaluating classification models. It shows the fraction of the correct classifications of model, which is the number of correct

classifications made divided by the total number of classifications made, multiplied by 100 to turn it into a percentage. In these experiments, each classifier is conducted with 10 repetition tests to calculate 10 classification accuracy results. Mean classification accuracy result are calculated for each classifier to compare the usefulness and effectiveness of the classifiers.

Performances of each classifier evaluated using precision, recall and F-Measure from the confusion matrix.

5.2.2 Formulae of Precision, Recall and F-Measure

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

A set of experiments were carried out by using the collected dataset in order to evaluate the performance of the classifiers using precision, recall and F-Measure from the confusion matrix. Precision refers to the percentage of classified results which are relevant, whereas recall refers to the percentage of total relevant results correctly classified by the classifiers. F-Measure is the harmonic mean of precision and recall. In these experiments, 90% of the dataset was used for training purpose, whereas the remaining 10% was used for testing purpose. The performance evaluation was conducted separately for each classifier.

5.2.3 Evaluating the Intelligibility of the Synthesized Malay Speech

In this research, a listening evaluation was conducted on the synthetic speech consisting different number formats generated by the Malay TTS system developed by the University Malaya, Speech Technology research team. Listening evaluation was conducted to determine the intelligibility of the synthetic voices.

The Word Error Rate (WER) is used to measure the intelligibility of the synthetic voices in this research.

$$WER = \frac{S + D + I}{N}$$

Where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions and
- N is the number of words in the reference

5.2.3.1 Procedures

Listening evaluation was carried out with the objective of evaluating the intelligibility of neutral sounding Malay synthetic speeches generated by the Malay TTS system for 6 different types of number formats. The listening evaluation was conducted in computer laboratory, with minimal external noises, involving one evaluator at a time. For conducting the listening evaluation, Dell desktops with Intel Core Duo 2.0 GHz, 4 GB memory and 250GB hard drive, which is equipped with ASIO Echo headphones, were used for the recordings. This research has adopted the test of intelligibility by (Zhi-Zheng Wu et al., 2010; Mustafa et al., 2014).

The listening evaluation in this research involves 20 native listeners of Malay, which comprises 10 male and 10 female evaluators aged between 21 to 25 years who are university students.

Each evaluator was allowed to listen to all the 20 utterances only once before responding, whereby they typed in what they had heard. The intelligibility of the utterances is evaluated at word level. Each correctly typed word is given a score of 1 and incorrect word is 0. Typing errors and spelling mistakes (for correct sound), which does not alter the semantic meaning is not considered as error of intelligibility. For example the word 'sembahyang' (pray), 'genggaman' (grip), or 'menunjukkan' (show) typed as 'sembayang', 'gengaman' or 'menunjukan', is considered as spelling mistake but not considered as intelligibility error because the spelling mistakes does not alter the semantic meaning of the word.

Each evaluator was given a specific set of voice snippets and had to respond to the questions pertaining to the voice snippet using a specially constructed evaluation questionnaire, which is presented in Appendix B.

5.2.3.2 Testing data

In this research, 20 Malay sentences selected from the 571 test data used for classification evaluation have been synthesized using the Malay TTS system developed by the University Malaya, Speech Technology research team for the purpose of listening evaluation. The 20 sentences created for synthesis comprise of 6 different number formats as shown below (the sentences are available in Appendix A):

- Date (4 sentences)
- Time (4 sentences)
- Phone (2 sentences)
- Currency (2 sentences)
- Measurement (5 sentences)
- Percentage (3 sentences)

The sentences were created with length from 8 to 30 words. Table 5.1 provides the detailed information for the 20 sentences created for listening evaluation.

Table 5.1: Details of 20 Sentences Prepared for Synthesis

Sentence Length	No. of Sentences
Words ≤ 10	2
10 < Words ≤ 20	11
20 < Words ≤ 30	7

5.3 Results

5.3.1 Classification Accuracy

Table 5.2 shows the classification accuracy of the classifiers, and Table 5.3 shows the summary of the classification accuracy. Table 5.3 shows two types of important results of each classifier, which is the highest classification accuracy of the 10 repetition tests, and mean classification accuracy of the 10 repetition tests. In this research, DT classifier achieved the highest mean classification accuracy of 94.37%. SVM with linear kernel classifier have the second highest classification accuracy of 93.86%. SVM with polynomial kernel classifier have the third highest classification accuracy of 92.12%. On the other hand, the classification accuracy of LDA, KNN1, KNN3, and SVMrbf classifiers are 94.83%, 96.49%, 96.43%, and 89.29% respectively.

Table 5.2: Classification Accuracy Results of each Classifier using Test Data of 10-fold Cross-Validation

Classifier	Repetition										Mean
	1	2	3	4	5	6	7	8	9	10	
SVMpoly	0.9138	0.8276	0.9298	0.9464	0.9474	0.9649	0.9298	0.9310	0.8727	0.9483	0.9212
SVMrbf	0.7759	0.7193	0.8929	0.6207	0.8276	0.8246	0.8246	0.8036	0.7321	0.7931	0.7814
SVMlinear	0.9483	0.8772	0.8929	0.9298	0.9310	0.9821	1	0.9123	0.9474	0.9649	0.9386
LDA	0.9138	0.9091	0.9455	0.9483	0.8448	0.9474	0.8596	0.9138	0.8621	0.8947	0.9039
KNN1	0.9138	0.9286	0.9483	0.8772	0.9649	0.8421	0.9310	0.8596	0.9273	0.9138	0.9107
KNN3	0.9643	0.9474	0.8596	0.8793	0.7719	0.8947	0.9123	0.9123	0.8947	0.8793	0.8916
DT	0.9655	0.9123	0.9483	0.9821	0.9483	0.9483	0.9825	0.8889	0.9310	0.9298	0.9437

Table 5.3: Summary of the Classification Accuracy Results of the Classifiers

Classifier	Highest Accuracy (%)	Mean Accuracy (%)
SVMpoly	96.49	92.12
SVMrbf	89.29	78.14
SVMlinear	100	93.86
LDA	94.83	90.39
KNN1	96.49	91.07
KNN3	96.43	89.16
DT	98.25	94.37

5.3.2 Confusion Matrix

Tables 5.4 to 5.10 show the confusion matrix for each classifier.

Table 5.4: Confusion Matrix for SVMpoly Classifier

Classifier	**SVMpoly**						
Classify / Actual	Date	Time	Phone	Currency	Measurement	Percentage	Recall (%)
Date	**10**	0	0	0	1	0	90.91
Time	0	**13**	0	0	0	0	100
Phone	0	0	**1**	0	0	0	100
Currency	0	0	0	**11**	0	0	100
Measurement	0	0	0	0	**10**	0	100
Percentage	0	0	0	0	1	**10**	90.91
Precision (%)	100	100	100	100	83.33	100	**96.49**

Table 5.5: Confusion Matrix for SVMrbf Classifier

Classifier	**SVMrbf**						
Classify / Actual	Date	Time	Phone	Currency	Measurement	Percentage	Recall (%)
Date	**6**	0	0	4	0	0	60
Time	0	**13**	0	0	0	0	100
Phone	0	0	**1**	0	0	0	100
Currency	0	0	0	**11**	0	0	100
Measurement	0	0	0	1	**9**	0	90
Percentage	0	0	0	1	0	**10**	90.91
Precision (%)	100	100	100	64.71	100	100	**89.29**

Table 5.6: Confusion Matrix for SVMlinear Classifier

Classifier	**SVMlinear**						
Actual \ Classify	Date	Time	Phone	Currency	Measurement	Percentage	Recall (%)
Date	**11**	0	0	0	0	0	100
Time	0	**13**	0	0	0	0	100
Phone	0	0	**1**	0	0	0	100
Currency	0	0	0	**11**	0	0	100
Measurement	0	0	0	0	**10**	0	100
Percentage	0	0	0	0	0	**11**	100
Precision (%)	100	100	100	100	100	100	**100**

Table 5.7: Confusion Matrix for LDA Classifier

Classifier	**LDA**						
Actual \ Classify	Date	Time	Phone	Currency	Measurement	Percentage	Recall (%)
Date	**11**	0	0	0	0	0	100
Time	0	**12**	0	0	0	1	92.31
Phone	0	0	**2**	0	0	0	100
Currency	0	0	0	**11**	0	0	100
Measurement	0	0	0	0	**10**	0	100
Percentage	0	0	0	0	2	**9**	81.82
Precision (%)	100	100	100	100	83.33	90	**94.83**

Table 5.8: Confusion Matrix for KNN1 Classifier

Classifier	**KNN1**						
Actual \ Classify	Date	Time	Phone	Currency	Measurement	Percentage	Recall (%)
Date	**11**	0	0	0	0	0	100
Time	0	**13**	0	0	0	0	100
Phone	0	1	**1**	0	0	0	50
Currency	0	0	0	**11**	0	0	100
Measurement	0	0	0	1	**9**	0	90
Percentage	0	0	0	0	0	**10**	100
Precision (%)	100	92.86	100	91.67	100	100	**96.49**

Table 5.9: Confusion Matrix for KNN3 Classifier

Classifier	**KNN3**						
Actual \ Classify	Date	Time	Phone	Currency	Measurement	Percentage	Recall (%)
Date	**10**	0	0	0	0	0	100
Time	0	**12**	0	0	0	1	92.31
Phone	0	0	**1**	0	1	0	50
Currency	0	0	0	**11**	0	0	100
Measurement	0	0	0	0	**10**	0	100
Percentage	0	0	0	0	0	**10**	100
Precision (%)	100	100	100	100	90.91	90.91	**96.43**

Table 5.10: Confusion Matrix for DT Classifier

Classifier	**DT**						
Actual \ Classify	Date	Time	Phone	Currency	Measurement	Percentage	Recall (%)
Date	**11**	0	0	0	0	0	100
Time	0	**13**	0	0	0	0	100
Phone	0	0	**2**	0	0	0	100
Currency	1	0	0	**10**	0	0	90.91
Measurement	0	0	0	0	**10**	0	100
Percentage	0	0	0	0	0	**10**	100
Precision (%)	91.67	100	100	100	100	100	**98.25**

In these experiments, confusion matrix is used as a scoring metric in the 10 fold cross validation. The diagonal entries (in bold) demonstrate the number of test units accurately classified, whereas the others are wrongly classified by the classifiers. These tables also provide results of precision and recall of the models in percentage. Tables 5.11 to 5.13 show a summary of recall, precision and F-Measure of each numbers type and classifiers.

Table 5.11: Recall of Each Numbers Type and Classifiers

Classifiers / Numbers	SVMpoly (%)	SVMrbf (%)	SVMlinear (%)	LDA (%)	KNN1 (%)	KNN3 (%)	DT (%)
Date	90.91	60	100	100	100	100	100
Time	100	100	100	92.31	100	92.31	100
Phone	100	100	100	100	50	50	100
Currency	100	100	100	100	100	100	90.91
Measurement	100	90	100	100	90	100	100
Percentage	90.91	90.91	100	81.82	100	100	100

Table 5.12: Precision of Each Numbers Type and Classifiers

Classifiers / Numbers	SVMpoly (%)	SVMrbf (%)	SVMlinear (%)	LDA (%)	KNN1 (%)	KNN3 (%)	DT (%)
Date	100	100	100	100	100	100	91.67
Time	100	100	100	100	92.86	100	100
Phone	100	100	100	100	100	100	100
Currency	100	64.71	100	100	91.67	100	100
Measurement	83.33	100	100	83.33	100	90.91	100
Percentage	100	100	100	90	100	90.91	100

Table 5.13: F-Measure of Each Numbers Type and Classifiers

Classifiers / Numbers	SVMpoly (%)	SVMrbf (%)	SVMlinear (%)	LDA (%)	KNN1 (%)	KNN3 (%)	DT (%)
Date	95.24	75	100	100	100	100	95.65
Time	100	100	100	96	96.30	96	100
Phone	100	100	100	100	66.67	66.67	100
Currency	100	78.57	100	100	95.65	100	95.24
Measurement	90.91	94.74	100	90.91	94.74	95.24	100
Percentage	95.24	95.24	100	85.72	100	90.91	100

5.3.3 Comparative Study

In this research, the proposed classification system that uses the selected classifiers such as SVM, DT, LDA, and KNN were able to perform well compared to the existing TTS systems in generating the correct number formats. Table 5.14 shows the comparison of the mean classification accuracy between the proposed techniques and the existing TTS system (Yaser, 2016).

Table 5.14: Comparison of the Mean Classification Accuracy between the Proposed Techniques and the Existing TTS System

	The Proposed Techniques				**Existing TTS Systems**		
Systems	DT (%)	SVMlinear (%)	KNN1 (%)	LDA (%)	INOVA (%)	TTSreader (%)	FESTIVAL (%)
Mean Classification Accuracy	94.37	93.86	91.07	90.39	50	43.75	37.5

Table 5.14 shows the mean classification accuracy of the proposed techniques and the existing TTS system, such as INOVA, TTSreader, and FESTIVAL. The mean classification accuracy of INOVA was 50%, TTSreader was 43.75%, and FESTIVAL was 37.5%, which are much lower compared to the proposed techniques. The existing TTS systems are poor in generating the correct number format even for English language because they did not perform the classification of the number format.

Table 5.15 shows the comparison of the mean classification accuracy between the proposed context-based techniques and the existing BoW technique experimented by Yaser (2016).

Table 5.15: Comparison of the Mean Classification Accuracy between the Proposed Context-Based Techniques and the Existing BoW Technique

	Mean Classification Accuracy			
Systems	DT (%)	SVMlinear (%)	KNN1 (%)	LDA (%)
Context-Based Technique	94.37	93.86	91.07	90.39
BoW Technique	56.96	61.69	55.91	59.74

This research has used the same Malay dataset and classifiers such as SVM, DT, LDA, and KNN for both context-based and existing BoW experiments to compare the mean classification accuracy. 90% of the dataset was used for training purpose, whereas the remaining 10% was used for testing purpose. The mean classification accuracy of BoW technique using classifiers such as SVM, DT, LDA, and KNN are 56.96%, 61.69%, 55.91% and 59.74% respectively. The mean classification accuracy for BoW technique is much lower than context-based technique.

5.3.4 Intelligibility Test

In this research, the 20 sentences are categorized into 3 categories which are sentence length less than or equal to 10 words, sentence length greater than 10 words and less than or equal to 20 words, and sentence length greater than 20 words and less than or equal to 30 words. The intelligibility of the synthetic Malay speech generated by the Malay TTS system developed by the University Malaya, Speech Technology research team is found to be high. The sentence length less than or equal to 10 words have an intelligibility score of 100% on the whole sentence, the sentence length greater than 10 words and less than or equal to 20 words have an intelligibility score of 98.7% on the whole sentence and the sentence length greater than 20 words and less than or equal to 30 words have an intelligibility score of 98.8% on the whole sentence. Table 5.16 shows

the details of intelligibility tests and Table 5.17 shows the summary of intelligibility tests.

Table 5.16: Details of Intelligibility Tests

Number of Sentence	Number of Words	20 Evaluators			
		N	S	D	I
1	17	340	0	0	0
2	26	520	20	0	0
3	26	520	0	0	0
4	16	320	0	0	0
5	25	500	0	0	0
6	18	360	0	0	0
7	13	260	40	0	0
8	30	600	20	0	0
9	25	500	0	0	0
10	24	480	0	0	0
11	15	300	0	0	0
12	22	440	4	0	0
13	12	240	0	0	0
14	15	300	0	0	0
15	12	240	2	0	0
16	18	360	0	0	0
17	8	160	0	0	0
18	14	280	0	0	0
19	9	180	0	0	0
20	12	240	0	0	0
Total	357	7140	86	0	0

Table 5.17: Summary of Intelligibility Tests

Sentence Length	(S+D+I)/N	WER (%)	Intelligibility (%)
Words ≤ 10	0	0	0
10 < Words ≤ 20	(40+2)/ (340+320+360+260+300+240+300+240+360+280 +240)	1.3	98.7
20 < Words ≤ 30	(20+20+4)/ (520+520+500+600+500+480+440)	1.2	98.8
Total	86/7140	1.2	98.8

This research is solving the problem of classifying the number format, it will be more complete if generalize the solution through a Malay TTS system (developed by UM group) since the most difficult thing for a TTS system is to identify the format before correctly convert it into text and then into speech form.

This research could not compare the results with the intelligibility of TTS without the numerical convertor, whereby the existing TTS system will hang (Intelligibility 0), will not continue reading any text after the numbers.

At the moment there is no single Malay TTS system that can synthesize numbers in Malay speech and there is no any existing Malay TTS system to be compared. As such it is impossible to compare the intelligibility results in this research with previous research.

5.4 Discussions

SVMlinear achieved the best performance among the classifier to classify the number type in this research. It was able to achieved 100% of recall, precision and F-Measure for all number type, which is the highest result compared to the others. A currency was wrongly classified as a date for DT classifier. DT achieved a recall of 90.91% for currency, precision of 91.67% for date and F-Measure of 95.65% and 95.24% for date and currency respectively. A date and a percentage were wrongly classified as measurement for SVMpoly classifier. SVMpoly achieved a recall of 90.91% for date and percentage, precision of 83.33% for measurement and F-Measure of 95.24%, 90.91% and 95.24% for date, measurement and percentage respectively.

For the intelligibility test, the intelligibility of longer utterances is lower than the shorter utterances could be due to evaluator's loss of concentration when listening to longer utterances. The total intelligibility error in terms of WER for 20 sentences is 1.2% on

the whole sentences which indicated the intelligibility score of the synthetic Malay speech is 98.8% on the whole sentence for 20 sentences. The 1.2% of errors for the 20 sentences is due to English words in the sentences. The Malay TTS system unable to generate Malay speech correctly for English words such as Port Dickson and Mcdonald in the sentences.

5.5 Summary

The developed model for classifying the number format shows that DT classifier achieved the highest mean classification accuracy of 94.37% compared to the other selected classifiers. The context-based technique was proven to be an effective technique for feature extraction. The performance of the classification system can be improved by considering the keywords and symbols in a sentence. On the other hand, the synthetic Malay speech generated by the Malay TTS system developed by the University Malaya, Speech Technology research team achieved 98.8% of the intelligibility score on the whole sentence for 20 selected sentences. This intelligibility score shows the effectiveness of the Malay TTS system in generating the Malay speech that able to listen clearly by human.

CHAPTER 6: CONCLUSION AND FUTURE RESEARCH

6.1 Overview

This chapter summarizes the findings of this research as well as the fulfillment of research aims and objectives. Besides that, this chapter shows the research contributions, research limitations, and future research.

6.2 Identification of Problem and Solution

The main aim of this research is to propose a suitable technique to automatically classify the numerical formats in Malay text in order to improve the Malay text-to-speech conversion of the existing TTS system. Intelligibility of the existing TTS system can be greatly enhanced with the development of this research.

For achieving the aims of this research, the issues of non-standard words in the existing TTS system have been discovered through literature review. Furthermore, some of the existing TTS systems have been chosen to evaluate their ability in number processing.

6.3 Research Objective Revisited

The three objectives that have been fulfilled by this research are explained as follows:

6.3.1 Research Objective 1

The first objective of this research was to propose a suitable technique to automatically classify the numerical formats in Malay text. To accomplish this objective, Named Entity Recognition (NER) with some of the existing machine learning classifiers has been discovered and their text categorization ability has been compared.

6.3.2 Research Objective 2

The second objective of this research was to develop an automatic numerical format classification system for Malay text. The few tasks that had to be performed in order to achieve these objectives are as follows:

- This research has identified an appropriate NER framework using context-based numerical feature extraction and supervised classification of machine learning which can be applied in the development of the proposed classification system.
- Collected and accumulated the Text-Numbers corpus that contains 6 types of numerical formats, and labeling them together with their sentences in Malay language from online news websites. A total 571 sentences have been collected and well-labeled as a complete dataset in this research.
- Development of the proposed NER based classification system. Python has been used for context-based features extraction from the data. On the other hand, MATLAB software has been used for training and testing the proposed classification system. Context-based technique which is used to classify the number format will split the sentence into separated words and collect two words before and after the located number in order to find the keywords of the number.

6.3.3 Research Objective 3

The third objective of this research was to evaluate the performance of the a) proposed technique in classifying various number formats and b) intelligibility of the synthesized Malay speech which consists of different number formats.

a) Performance evaluation of the system was done by using testing data of 10-fold cross-validation and listening evaluation by native listeners. On the other hand,

the confusion matrix was used to describe the performance of a classification model in more detail. Classification accuracy, precision, recall and F-Measure are calculated for each classifier in this research.

In conclusion, this research has developed a context-based classification system for classifying the numerical format in Malay text successfully. The proposed classification system was able to achieve a high classification accuracy of numerical format, which has been proven in the previous chapter. As a result, the proposed solution can be applied in the existing NLP in order to improve the quality of number normalization in the existing TTS system.

b) The intelligibility of the synthesized Malay speech which consists of various number formats has been improved due to the improvement in numerical classification.

6.4 Conclusion

This research mainly focuses on one of the most important issues that exist in the current TTS system, which is the grapheme-to-phoneme conversion of the NLP, especially the number to text conversion. Existing TTS systems are not being able to convert the number to text accurately according to the evaluation of the existing TTS systems. This issue decreases the performance of producing the expected speech output and affects its intelligibility.

This research has proposed a solution to overcome the number to text conversion issue, by developing a classification system that is able to identify the numerical format in the sentence. TTS system can convert the number to the corresponding text correctly based on the classified numerical format in order to improve the intelligibility and usability of the TTS system. This research has described the development of the proposed

classification system as well as the tools and resources required to develop the system. The development processes consist of data collection, data processing, training, and testing of the model.

Context-based classification technique is the proposed technique that is applied in the feature extractor of this research. Besides that, four appropriate machine learning techniques that are effective in classifying the numerical format are shortlisted, which comprise of Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Decision Tree (DT). However, DT classifier achieved the highest mean classification accuracy compared to the other selected classifiers.

In conclusion, the proposed classification system is able to classify the format of the number by using the appropriate classification technique. It can improve the function of normalization of numbers, thus it can be incorporated into the existing NLP units and therefore improve the intelligibility of speech synthesized by the Malay TTS system.

6.5 Research Outcomes and Contributions

The contributions of this study can be summarized as follow:

I. A new technique developed to automatically classify the numerical format in Malay text.
II. The proposed work can be generalized into different Natural Language Processing (NLP) applications, which will lead to the improvement of the results for these tasks as well.
III. This study addresses an important aspect of normalization problems of NSWs and fills the research gap.

6.6 Future Research

As highlighted in this research, number has several formats, and these formats can be identified based on the context of the number. Future works can target on collecting more well-labeled data in Malay text in order to improve the accuracy of the classification system as well as cover another numerical format which are not included in this research.

There are many numerical formats in real life and it may increase in the future. The labeled data with different numerical formats should be collected as much as possible to train the classification system, and therefore, improve the performance and intelligibility of the TTS system.

REFERENCES

Aida-Zade, K. R., Ardil, C., & Sharifova, A. M. (2013). The main principles of text-to-speech synthesis system. *International Journal of Signal Processing*, *7*(1), 13–19. Retrieved from http://www.waset.org/journals/ijice/v6/v6-1-3.pdf

Alfred, R., Leong, L. C., On, C. K., & Anthony, P. (2014). Malay named entity recognition based on rule-based approach.

Alfred, R., Mujat, A., & Obit, J. H. (2013, March). A ruled-based part of speech (RPOS) tagger for Malay text articles. In *Asian Conference on Intelligent Information and Database Systems* (pp. 50-59). Springer, Berlin, Heidelberg.

Bhandari, N., Chowdri, R., Singh, H., & Qureshi, S. R. (2017, December). Resolving Ambiguities in Named Entity Recognition Using Machine Learning. In *2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)* (pp. 159-163). IEEE.

Black, A. (n.d.). Festvox: Festival. Retrieved March 30, 2017, from http://festvox.org/festival/

Black, A., & Lenzo, K. (2000). Building voices in the Festival speech synthesis system. Retrieved from http://festvox.org/docs/festvox-1.1/festvox.ps.gz

Burkhardt, F., & Reichel, U. D. (2016). A Taxonomy of Specific Problem Classes in Text-to-Speech Synthesis : Comparing Commercial and Open Source Performance. *In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Portoroz, Pp. 744-749. ISBN 978-2-9517408-9-1*. Retrieved from http://real.mtak.hu/45967/1/BR_LREC2016.pdf

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

El-Imam, Y. A., & Don, Z. M. (2000). Text-to-speech conversion of standard Malay. *International Journal of Speech Technology*, *3*(2), 129-146.

Filipiak, D., Agt-Rickauer, H., Hentschel, C., Filipowska, A., & Sack, H. (2016, July). Quantitative analysis of art market using ontologies, named entity recognition and machine learning: a case study. In *International Conference on Business Information Systems* (pp. 79-90). Springer, Cham.

Gorinski, P. J., Wu, H., Grover, C., Tobin, R., Talbot, C., Whalley, H., ... & Alex, B. (2019). Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches. *arXiv preprint arXiv:1903.03985*.

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Hinterleitner, F., Norrenbrock, C., & Möller, S. (2013). Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech. *Proc. of the 8th ISCA Speech Synthesis Workshop (SSW 2013)*, *2*(ii),

167–171. Retrieved from http://ssw8.talp.cat/papers/ssw8_PS2-1_Hinterleitner.pdf

Ilievski, F., Rizzo, G., van Erp, M., Plu, J., & Troncy, R. (2016, May). Context-enhanced adaptive entity linking. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 541-548).

Kanimozhi, U., & Manjula, D. (2017, February). A CRF Based Machine Learning Approach for Biomedical Named Entity Recognition. In *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)* (pp. 335-342). IEEE.

Katz, G., Shapira, B., Ofek, N., Bar-Zev, Y., & Negev, I. (2014). CoBAn: A Context Based Approach for Text Classification. *Journal Information Sciences: an International Journal archive*, *262*(March), 137-158.

Kaur, K., & Gupta, V. (2012). Name entity recognition for Punjabi language. *IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN*, 2249-9555.

Khalifa, O. O., Ahmad, Z. H., Hashim, A. H. A., & Gunawan, T. S. (2008). SMaTalk: Standard malay text to speech talk system. *Signal Processing: An International Journal*, *2*(5), 1.

Khan, N. J., Anwar, W., & Durrani, N. (2017). Machine Translation Approaches and Survey for Indian Languages. *Cornell University Library*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1701/1701.04290.pdf

Khanam, M. H., Khudhus, M. A., & Babu, M. P. (2016, August). Named entity recognition using machine learning techniques for Telugu language. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (pp. 940-944). IEEE.

Mohammed, A. S. A. H. A., & Bagash, F. O. F. T. (2017). A biomedical named entity recognition using machine learning classifiers and rich feature set. *IJCSNS*, *17*(1), 170.

Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing (IJNLC)*, *1*(4), 15-23.

Mumtaz B. Mustafa, Zuraidah M. Don, Raja N. Ainon, Roziati Zainuddin, Gerry Knowles. 2014. Developing an HMM-based speech synthesis system for Malay: a comparison of iterative and isolated unit training. IEICE Transactions on Information and Systems, vol.E97-D,No.5,1273-1282. (*ISI-Indexed*)

Mustafa, M. B., Don, Z. M., Ainon, R. N., Zainuddin, R., & Knowles, G. (2014). Developing an HMM-based speech synthesis system for Malay: a comparison of iterative and isolated unit training. *IEICE TRANSACTIONS on Information and Systems*, *97*(5), 1273-1282.

Olson, J., & Charles, E. (2009). Is 80 % Accuracy Good Enough ? *In Proceedings of the ASPRS 17th Pecora Conference.*, 18–21.

Pappas, C. (2015). The Best Text To Speech (TTS) Software For eLearning. Retrieved October 11, 2016, from https://elearningindustry.com/top-10-text-to-speech-tts-software-elearning

Ross, K. A., Jensen, C. S., Snodgrass, R., Dyreson, C. E., Jensen, C. S., Snodgrass, R., … Chen, L. (2009). Cross-Validation. In *Encyclopedia of Database Systems* (pp. 532–538). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Salah, R. E., & Qadri binti Zakaria, L. (2017). A Comparative Review of Machine Learning for Arabic Named Entity Recognition. *International Journal on Advanced Science, Engineering and Information Technology*, *7*(2), 511-518.

Saleh, I., Tounsi, L., & van Genabith, J. (2011, December). Zaman and raqm: extracting temporal and numerical expressions in arabic. In *Asia Information Retrieval Symposium* (pp. 562-573). Springer, Berlin, Heidelberg.

Salleh, M. S., Asmai, S. A., Basiron, H., & Ahmad, S. (2017, May). A Malay Named Entity Recognition using conditional random fields. In *2017 5th International Conference on Information and Communication Technology (ICoIC7)* (pp. 1-6). IEEE.

Samsudin, N. H., & Kong, T. E. (2004). A Simple Malay Speech Synthesizer Using Syllable Concatenation Approach. In *MMU International Symposium on Information and Communications Technologies*.

San-Segundo, R., Montero, J. M., Giurgiu, M., Muresan, I., & King, S. (2013). Multilingual Number Transcription for Text-to-Speech Conversion. *8th ISCA Workshop on Speech Synthesis (SSW)*, 65–69. Retrieved from http://ssw8.talp.cat/papers/ssw8_PS1-8_San-Segundo.pdf

Seni, G., & Elder, J. F. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *2*(1), 26–35. https://doi.org/10.2200/S00240ED1V01Y200912DMK002

Shetake, P. S., Patil, S. A., & Jadhav, P. M. (2014). REVIEW OF TEXT TO SPEECH CONVERSION METHODS. *International Journal of Industrial Electronics and Electrical Engineering*, *2*(8), 2347–6982.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. *Computer Speech and Language*, *15*, 287–333. https://doi.org/10.1006/csla.2001.0169

Suryana, A., & Ipnuwati, S. (2016, February). Vector Machine Learning Method for Text Mining Indonesian Social Media Named Entity Recognition. In *The 5th International Conference on Information Technology and Engineering Application (ICIBA2016)*. Bina Darma University.

Swee, T. T., Hussain, S., & Salleh, S. (2008, October). Corpus-based Malay text-to-speech synthesis system. In *2008 14th Asia-Pacific Conference on Communications* (pp. 1-5). IEEE.

Taşpınar, M., Ganiz, M. C., & Acarman, T. (2017, June). A feature based simple machine learning approach with word embeddings to named entity recognition

on tweets. In *International Conference on Applications of Natural Language to Information Systems* (pp. 254-259). Springer, Cham.

Tetschner, W. (2003). Text-to-Speech - Naturalness and Accuracy. *ASR News*. Retrieved from http://nwmlt2013.learnpunjabi.org/talks/TTS_Testing.pdf

Tetschner, W. (2016). *Text-to-Speech Accuracy Testing*. Retrieved from http://www.asrnews.com/TTS Acc asrnews_website_2016.pdf

Vijay, J., & Sridhar, R. (2016). A Machine Learning Approach to Named Entity Recognition for the. *Asian Journal of Information Technology*, *15*(21), 4309-4317.

Wu, H. Y., Lu, D., Hyder, M., Zhang, S., Quinney, S. K., Desta, Z., & Li, L. (2018). DrugMetab: An Integrated Machine Learning and Lexicon Mapping Named Entity Recognition Method for Drug Metabolite. *CPT: Pharmacometrics & Systems Pharmacology*, *7*(11), 709-717.

Yaser (2016). *Automatic Numerical Format Prediction of Web Sources For Text-To-Speech System (Unpublished Academic Project)*. University of Malaya, Kuala Lampur.

Zhang, Y., Wang, X., Hou, Z., & Li, J. (2018). Clinical Named Entity Recognition From Chinese Electronic Health Records via Machine Learning Methods. *JMIR medical informatics*, *6*(4), e50.

Zhi-Zheng Wu, Eng Siong Chng, Haizhou Li. (2010). Development of HMM-based Malay Text-to-Speech System, Proceedings of the Second APSIPA Annual Summit and Conference, pages 494–497.