PROPER NOUN DETECTION USING REGEX ALGORITHM AND RULES FOR MALAY NAMED ENTITY RECOGNITION

FARID BIN MORSIDI

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE (INFORMATION SYSTEM AND MANAGEMENT)
(MASTER BY RESEARCH)

FACULTY OF ART, COMPUTING & CREATIVE INDUSTRY
UNIVERSITI PENDIDIKAN SULTAN IDRIS

2018

ABSTRACT

This study was aimed to develop a Malay proper noun detection method to cluster and classify named entity categories, particularly for major important classes such as person, location, organization, and miscellaneous for Malay newspaper corpus. Regular Expression pattern identification (regex) algorithm and rule were introduced in this study to overcome the limitation of dictionary and gazetteer. Two visualization techniques namely as Decision Tree and Term Document Matrix had been used to evaluate the efficiency of the method. The result obtained 74% of accuracy during the generation of decision tree. Visualization for term document matrix achieves a maximized value of 9.8007403, 9.8718517, and 9.9890683 for Astro Awani, Berita Harian, and Bernama dataset respectively. As a conclusion, the regex algorithm could indicate the presence of Malay proper noun, thus making it an appropriate method for extraction tool to cluster and classify Malay proper noun. The study implicates that the use of Malay proper noun detection method can increase the effectiveness in named entity recognition and beneficial to improve document retrieval for Malay language.

PENGESANAN KATA NAMA MENGGUNAKAN ALGORITHMA REGEX DAN PETUA UNTUK PENGECAMAN ENTITI NAMA MELAYU

ABSTRAK

Tesis penyelidikan ini bertujuan untuk memperkenalkan sistem pengecaman kata nama Melayu yang boleh digunakan untuk mengelompok di samping mengelaskan kategori entiti nama berdasarkan algoritma *regular expression* terutamanya bagi kelompok objek penting seperti manusia, lokasi, organisasi, dan lain-lain. Dua teknik pengvisualan data iaitu Pepohon Keputusan (Decision Tree) dan *Term Document Matrix* telah diadaptasi dalam pendekatan ini untuk menilai keberkesanan teknik algoritma *regular expression*. Ujian di peringkat akhir kajian telah menghasilkan peratusan ketepatan 74% semasa menjana pepohon keputusan. Visualisasi *Term Document Matrix* memperlihatkan capaian nilai maksimum pada 9.8007403, 9.8718517, dan 9.9890683 bagi set data Astro Awani, Berita Harian, dan Bernama. Sebagai kesimpulan, adalah boleh dilihat bahawa algoritma *regular expression* berpotensi untuk mengecam struktur corak kata nama Melayu sekaligus menjadikannya suatu alternatif untuk mendukung usaha mengelompokkan serta mengelaskan kategori kata nama Melayu. Usaha penyelidikan ini membuktikan bahawa penggunaan sistem pengecaman kata nama Melayu mampu menyumbang kepada percambahan bilangan entiti nama Melayu sedia ada, di samping bermanfaat untuk tafsiran dokumen tak bertanda dalam simpanan maklumat bahasa Melayu.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA	Astro Awani
BER	BERNAMA
BH	Berita Harian
CRF	Conditional Random Field
GATE	General Architecture for Text Engineering
IE	Information Extraction
IR	Information Retrieval
KNN	K-Nearest Neighbour
LOC	Location
ME	Maximum Entropy
MISC	Miscellaneous
MUC	Message Understanding Conference
NER	Named Entity Recognition
NERC	Named Entity Recognition and Classification
NLP	Natural Language Processing
ORG	Organization
PER	Person
POS	Part-of-Speech
SVC	Support Vector Classifier
SVM	Support Vector Machine
SVR	Support Vector Regressor
TF-IDF	Term Frequency-Inverse Document Frequency

CHAPTER 1

INTRODUCTION

1.1 Named Entity Recognition

Named Entity Recognition (NER) is a significant field that constitutes a subcomponent of the Natural Language Processing (NLP) field. The term "Named Entity" was proposed during the 6th MUC (Message Understanding Conference) (Grishman et al., 1996). The Message Understanding Conference (MUC) emphasized Information Extraction (IE) tasks, where structured information from various company activities besides defence-related activities is extracted from unstructured text. These unstructured texts were derived from credible academia sources. During the process of categorizing the task, researchers noticed that it is important to categorize information units such as names, including person, organization and location, as well as numerical expressions, including time, date, money, and percent

expressions. The identification references to these entities in text was recognized as "Named Entity Recognition and Classification (NERC)".

The field of NER and NLP which is synonymous with each other are closely associated with each other due to its nature in language process analysis. NER is important for NLP in terms of document retrieval, morphological analysis, and information extraction (Liao et al., 2009; Nothman, 2008). NLP tools can be used to improve the processing of clinical records (Patrick et al., 2011). In a clinical search, named entity extractor may be used to distinguish patient and physician names. Within Information Retrieval (IR), NER improved the detection of relevant documents (Alfred et al., 2014).

The major languages that undergo NER analysis have their own morphologies and techniques to handle issues related to NER research. A good proportion of work for NER research area has been dedicated to the study of English, but a larger section of it addresses language independence and multilingualism problems. Each distinctive language has its own processes in order to determine the classification of named entities according to its most precise categories, which is unique and may only be applicable within its own domain and not to others (Alfred et al., 2014). An example of this includes the Indonesian language which is closely related with Malay domain, which has been extensively studied (Massandy et al., 2014; Rashel et al., 2014; Purwarianti, 2014; Suwarningsih et al., 2015) and Arabic (AbdelRahman et al., 2010; Abu Bakar et al., 2013; Althobaiti et al., 2014; Darwish et al., 2014) has started to receive a lot of attention during large-scale projects such as Global Autonomous Language Exploitation (GALE).

An important question at the inception of the NER task was whether machine learning techniques are considered important in terms of data retrieval in a conventional system, aside from the relevance of simple dictionary lookup as sufficient to produce optimal performance. Some NER techniques available nowadays use utility features to improvise the efficiency of information extraction, such as gazetteers as a storage to people's names, places, organizations, and various other forms of information regarding proper nouns. While problems of coverage and ambiguity prevent straight forward lookup, injection of gazetteer matches as features in machine-learning based approaches is critical for good performance (Aboaoga et al., 2013; Abu Bakar et al., 2013). Problems occur when many names need to be stored in the gazette. However, for some situations, a small gazetteer is enough to provide good accuracy and recall values (Alfred et al., 2014; Althobaiti et al., 2014).

Information Extraction is a field closely related to Natural Language Processing, in terms of data analysing and categorisation. Information Extraction serves to analyse texts based on both sets of theories and technologies. Among the research fields applied the advancement of IE are Information Retrieval (IR) and Question Answering. There are vast resources of linguistic data in the form of text corpora, both including the freedom of accessibility for users and some included restrictions for purposes other than data interpretation. Text items exist in the form of text corpora, such as structured, semi-structured, or free text. (Cao et al., 2012; Zamin et al., 2012). These items have been implemented generally as unstructured input texts and are written to be understood by humans.

However, news articles, which are composed in format that is understandable by humans, are comprised of a textual form that is hard to be understood by machines (Abu Bakar et al., 2013). The processes encased within the identification of named entities enable these text data to be extracted for further processing. However, the massive size of these articles would require a long time for humans to manually process (Abu Bakar et al., 2013; Alfred et al., 2014; Elyasir et al., 2013). This indicates the importance of an automated process to interpret these data into chunks that are post-processed for the ease of humans.

The purpose of this research is to investigate the correlation of the fields in NER and Text Pattern Detection. These data that were accumulated via data collection methods such as web scraping and crawling were stored and categorised according to their morphological features, including their aliases, typing, identifiers, noun structures, and keyword significance. Named entities in a document, whether web content, articles, and text corpora possesses lexical characteristics known as word (Che et al., 2013). Words include those that represent similar traits and characterization. These are referred to as the three main categories for the categorization of Named Entity Recognition. These categories include PEOPLE, ORGANIZATION, and LOCATION.

Although there are several minor categories that have been the accumulating factor of word clusters to be gathered into 1 equal group, these categories remain under the shadow of the 3 major categories. The implementation of NLP (Natural Language Processing) algorithms into enhancing the processing of text chunks is typically affected by the domain of the studies. Specific domains are utilized for

recognizing named entities or specific domain of certain typing, for example linguistic nouns with linguistic Automated Speech Recognition (ASR). Furthermore, different languages require proficiency in varied languages in order to identify the named entity (Sari et al., 2010; Zamin et al., 2011).

Several credible NER systems for major languages exist, these includes major linguistic field such as English, Arabic, German, Chinese, Hindu, Indonesian, and so on. However, no existing credible system would appropriately implement the NER detection of proper nouns in the Malay language. Along with the purpose of the research, it is targeted that classification algorithm concept that is related in terms of data clustering be applied into an algorithm which is dedicated so that a better clustering result could be obtained, along with the purpose of improving the accuracy for clustering. Prior research has indicated the proficiency of applying ontological approach to improve the performance of document clustering due to the nature of expert systems that conceptualizes a domain into an individually-identifiable format, along with the visualization of machine-readable format containing entities, attributes, relationships, and axioms. By interpreting various techniques to classify unstructured documents, a technique depending on document classification has been determined to be a superior alternative to improve the efficiency of data clustering. This research intends to improvise the text pattern recognition concept into clustering algorithms to improve the efficiency of Malay noun allocation and resolve ambiguities within unannotated data classification.

1.2 Problem Statement

NER rule-based algorithms intend to improve the proficiency of element detections once they have been accumulated into similar categories. This approach is directed towards overcoming the overhead costs of corpus annotation, among the automatic creation of standardised corpora and semi-supervised training methods. As there is still no existing detection mechanism that could automatically annotate Malay nouns systematically into respective repository for further reference, there exists a need for a technique to be innovated or improvised into current works in order to enrich the availability of Malay linguistic resource.

There is a problem of a lack of NER clustering algorithms for the purpose of classifying and clustering named entity for Malay language. After the Message Understanding Conference 6 (MUC-6) in 1996 that introduced the concepts of named entity recognition, the fields of Information Retrieval and Information Extraction had begun the effort of mapping text contents to represent knowledge in a more concise form, along with the identification of text features that is derived from the text collection itself (Nothman, 2013).

NER studies have been performed primarily among the major recognised languages of the world, and when both supervised and unsupervised training are pioneered. Top research in NER has been widely circulated among major languages such as Arabic and German (Nothman et al., 2008). The approaches conducted to carefully isolate named entities according to their categories takes into account respective traits in the languages that should be identified by analysis on their

morphologies and sentence structures. However, NER approach for a particular language is deemed appropriate to only that particular domain and not applicable anywhere else.

A problem in addressing Malay computational linguistic analysis is to identify and classify new data into its respective groupings. For this purpose, classification algorithms and machine learning processes are applied. There remains ambiguity of the proper standards for classification algorithms targeted to retrieve and isolate Malay text data (Abu Bakar et al., 2013; Alfred et al., 2013). Various data mining algorithms had been incorporated with the concept of the conventional NER algorithms to optimise the extraction and classification of named entities. Among the algorithms that have been extensively used are Support Vector Machines (Mansouri et al., 2008), Expectation-Maximization (Borman, 2009), K-Means(Jain, 2010), K-Nearest Neighbour (Trstenjak et al., 2014), and A Priori (Indurkhya et al., 2010), which all incorporate different features in categorisation. Although these algorithms could be incorporated into the Malay NER research after improvisations is made, the steps that should be considered in order to obtain optimal named entity classification according to its categories are still considered vague and could not be entirely accustomed to adapt to language resources that are still lacking in availability for Malay itself.

Different NER extraction techniques for specified languages decrease in productivity when applied to a Malay NER framework. This mainly happens due to the variance in structure for the language itself as compared to other major linguistic elements that had been experimented with in numerous NER tagging and extraction

procedures. In this aspect, the NER research would be intricate in the identification of the distinctive feature that distinguishes a particular language syntax with others, as seen in English with writing order from left to right, whilst Arabic writing is vice versa. The lack of suitable NER algorithms to classify Malay text data had become a detrimental factor in the effort of its application on different language structures. For example, English NER procedures were almost not applicable in processing Malay articles due to the variant in its contextual usages (Alfred et al., 2013; Alfred et al., 2014).

To cater to the native Malay noun schema itself, the NER scheme needs to be designed based on a Malay Part-of-Speech (POS) features and contextual features that have been implemented to handle Malay articles (Mohamed et al., 2011; Abu Bakar et al., 2013; Rayner et al., 2013). From these annotated results, proper nouns would be identified or detected for the possible candidates of annotation. Symbols and conjunctions must also be considered in the process of identification of named entity for Malay-related articles. Current NER techniques for Malay often rely on external tools such as dictionaries and gazetteers to assist in improving the extraction and retrieval rate within a conventional system (Alfred et al., 2014). Research for Malay linguistic NER had been steadily seen an increment since 2010, with the focus been given on data clustering purposes such as manual tagging and identification of unannotated text as proven in studies conducted (Mohd Don, 2010; AbdelRahman et al., 2010; Alfred et al., 2013; Alfred et al., 2014). Most of these efforts emphasize the participation of both human and machine learning, as fully-automated supervised approach could not be executed efficiently without a gold data corpus.

As for the methods applied in the detection of named entities for Malay unannotated documents, extraction patterns for the word structure is performed based on external tools as a scaffold to overcome the deficiencies of lack in resource availabilities. Implementations of NER algorithms have relied mostly on rule-based context and dictionary listing that have been predefined and clarified (Alfred et al., 2014). This is mainly due to the objective of applying semi-supervised approach for training Malay text entities to assimilate the entries of new data collection into the existing Malay text data collection. However, in order to increase the quality of the output corpus itself, the dictionaries used in the process and rule schemes applied in the data extraction needs to be constantly updated (Nadeau, 2007; Alfred et al., 2014). This effort requires a vigorous human effort aside from high cost for annotating the new data collection. Therefore, a more efficient algorithm could be proposed to overcome the problems incurred from all the included data training methods.

1.3 Research Objective

As indicated prior, although there already existed an abundance source of NER systems dedicated towards linguistic studies, there almost had been no effort that is emphasized on the relevance of proper Malay noun schematics for the research field itself. The nature of the language itself encourages more current research to explore the possibilities of innovating and embedding closely related NER algorithms into enhancing the classification and identification of Malay naming system. Most available research encases the approach of utilizing supervised training process in order to improve the NER training regime. However, the potential values available

within the unsupervised training itself make it possible to embed more suitable algorithms to improve the process (Hovy, Navigli, & Ponzetto, 2013; Chou et al., 2014). For the context of this research, a few objectives have been identified as a measurement tool towards the efficiency of semi-supervised learning technique.

I. To propose annotated unstructured Malay proper noun newspaper corpus that can be used to classify named entity for Malay text
II. To develop proper noun detection method using regex algorithm to cluster and classify word categories as Named Entity for Malay text
III. To evaluate the performance of the proposed regex algorithm in detecting proper nouns for categorisation into respective named entity classes

1.4 Research Questions

As the process of NER involves the categorisation of entity within provided data corpora, it is a norm for the tagging of new and non-existing entities. The tagging of Named Entities for particular things, classes, and numeric expressions is considered as a vital component technology for various NLP applications (Mohamed et al., 2011; Alfred et al., 2013; Rashel et al., 2014). Fields available in NER include Information Extraction (IE), Question-Answering (Q & A), Summarization, and Information Retrieval (IR). Algorithms implemented during the categorisation of data clustering attempts to normalize or standardize the document contents in a language-neutral way. This is done via mapping text contents to an independent knowledge representation

such linguistic corpora and thesaurus, or by recognition of language-independent text features inside the documents. A few questions have been identified within the scope of this research.

I. Which NER algorithm is suitable to be implemented with the existing learning system to enable a better data clustering and entity identification within Malay text corpus?

Machine translation systems besides manual human recognition were used up until now to resolve the issues of learning process for systems to recognise individual entities among given corpora (AbdelRahman et al., 2010). Named Entity (NE) plays a vital role in news document. Generally the main categories of NE consisted of PERSON, ORGANIZATION, and LOCATION (Alfred et al., 2014; Althobaiti et al., 2014). Although minor categories such as DATE, TIME, or NUMBER co-exist within the domain, they often lead to grouped documents with little content in similarity (Alfred et al., 2014). This research attempts to indicate the availability of suitable classification algorithm to be applied during the pre-processing of data clusters and learning process to optimize the detection and categorisation of Malay noun schematic terminologies.

II. What regex pattern is suitable in detecting specific morphological features such as proper nouns that could be utilized to classify word categories as NE for Malay text?

The establishment of an expert user system may or may not require the participation of human activities in determining the validity and credibility of categorised text data (Aljoumaa, 2012). Most existing clustering algorithms take into account the grouping of a particular data category according to similarity or the distance among the documents. It would be advantageous should an expert system able to properly categorise data clusters according to their own categories with little or minimal human assistance. A pattern detection schematics could be established in order to properly annotate the detection of words that exist in the form of proper noun structure, particularly for language resources that are lacking in abundance such as Malay itself.

III. How did level of supervision affect the efficiency of the precision and recall rates of document retrieval within unannotated Malay articles?

The ability to recognise previously unknown entities is an essential component in NER research. Learning processes were induced to automatically generate rule-based systems or sequence labelling algorithms beginning from a collection of training data accumulated (Abdallah, Shaalan, & Shoaib, 2012; Alfred et al., 2014). The concept behind supervised learning is to investigate the efficiency of positive and negative examples of NE over a huge collection of annotated documents, in addition to design rules that capture instance of a particular linguistic typing. Alternative learning methods to resolve the costly

and difficult to maintain supervised learning process have been innovated (Zafarian & Rokni, 2015).

1.5 Research Significance

Named Entity Recognition, also known as Named Entity Recognition and Classification (NERC), is a component of Text Mining and is heavily involved in the information extraction process. There are already various NER systems for the world's major linguistic approaches, including English, German, and Chinese. As different languages contain various morphologies and syntaxes, this caused them to require distinctive NER processes as well. Useful sorting and categorisation algorithms were also implemented into the system according to compatibility. In the Malay Archipelago, including Malay Straits, Indonesia, and the Philippines, Malay is among the fundamental languages that form the foundation of regional dialects. The lack of research conducted for the purpose of reforming the categorisation and classification of Malay noun system is regrettable, since it is the Archipelago's biggest linguistic similarity point.

For the purpose of the proper implementation of text pattern identification and classification to take place in the Malay NER system, an in-depth understanding towards the specific issues related to the NER system must be developed. As there existed various classification algorithm that is implemented across specific fields of linguistic NER, each positive traits of the respective methods of algorithm paradigm should be identified to further enhance the building of a proper Malay NER schematic

system. This research attempts to cultivate the available prospects of major NER researches into a proper Malay noun system. Although the approach may be considered minute compared to the relatively already extensive availability of Malay linguistic and morphology aspect, it is hoped that the research will aid in raising the credibility of the Malay language among the major languages of the world.

The methodology of supervised and unsupervised learning process is to isolate efficiently the unannotated examples of named entities over a huge deposition of data clusters to fulfil the objective of assimilation of data with their similar typing. Both learning processes have potential and adverse effects for categorizing linguistic data properly. Supervised learning leans more towards an automated approach, which is the requirement of a credible and large collected corpus (Gunawan et al., 2015; Le Nguyen et al., 2014). Although this process has proven to be beneficial, it is difficult to maintain. While semi-supervised learning involves the assimilation of pre-existing training data into deducing statistical inferences towards unannotated document sources, it only involves typically minimal to little levels of human participation in completing the task (Maraziotis, 2012; Poria et al., 2013; Suakkaphong et al., 2013). Therefore, it could be said that both learning process have their own prospects and consequences to be taken into consideration of developing a linguistic NER system. The output for the research attempts to investigate the effect of limited supervised learning process brought to the establishment of an effective Malay NER system.

Good precision and recall of a system requires the participation of both human and machine to enable it to operate at its optimal rate. Algorithms for NER can be classified into three types: rule-based, machine learning, and hybrid approach, which

is the assimilation of both rule-based and machine learning techniques (Mohammed, Omar, & Bangi, 2012). Of these, the rule-based algorithm exists most widely and is seen to be more compatible with both supervised and unsupervised learning process (Rashel et al., 2014). As the limitations of semi-supervised learning generally include reliance upon available lexical resources, search patterns, and statistical computations on an unannotated corpus, the recall and precision rates are also directly influenced.

The problems noted during the retrieval of relevant documents in a linguistic domain are known to reside in three aspects of natural languages, which are polysemy, synonymy, along with variations in word structures. Machine-learning and human annotation techniques were the propelling factors of the current research into NER. Various existing NER schematic systems have already illustrated positive traits encasing each process. NER system involves the classification of text based on the occurrence of keywords in the context, as the accumulation of named entities within the corpora itself involves uncovering textual forms and ontological features that are considered significant to the semantics of the text itself.

The concept for rule-based itself is to associate ambiguity factors into predictable variables for the purpose of determining logical reasoning. Along with several beneficial features that is induced from Information Extraction in order to improve precision and recall rate in systems, for example pre-processing and data clustering, and the likes of implementation of fuzzy logic complimented with rule-based approach is seen as a potential alternative to resolve the current problems aforementioned on the lack in availability of annotated corpus such as Malay itself.

This approach could also assist in overcoming the limitations of relying upon frequently-updated dictionary lists and gazetteers.

1.6 Scope of Research

The detection of proper noun is relevant with the annotation of named entity presence. This hold significance for any analysis on the occurrence of named entity to emphasize on the frequency of proper noun occurrence in a particular text collection. Several aspects have been highlighted in accordance with the research requirement.

1.6.1 Text Corpora

Malay linguistic resources are still lacking for the purpose of further reference for morphology analysis. Computational linguistics also bears the similar stigma. For the purpose of this research, open Malay corpora had been targeted. This includes access to a few online news corpora such as Bernama (http://www.bernama.com) and Astro Awani (http://www.astroawani.com).

1.6.2 Learning Process

As per mentioned, the research generally emphasized on the detection and improvisation of a suitable classification algorithm to alleviate the problems endured by the process for semi-supervised learning procedures. Learning techniques usually rely upon lexical resources, search patterns, along with clustering of data across an unannotated corpus. The semi-supervised learning method is focused on determining the level of efficiency for supervision that would influence the learning procedures of certain linguistic corpora.

1.7 Operational Definition

Within the scope of the research, a few terms have a tendency to appear most frequently throughout the evaluation and development process. It is important that these keywords be highlighted according to their definitions and operational feasibilities.

1.7.1 Information Retrieval (IR)

Information Retrieval is another field in computer science applied within the scope of Natural Language Processing. This field is derived as the activity of locating documents that answer an information need with the aid of indexes. IR embedded systems tend to be classified as “search engines”. The user must read each document

to know the facts reported in it. This contrasts the practice in its closely associated field, Information Extraction, where the aim is to extract crucial information without requiring the end user of the information to read the text. The goal of Information Retrieval is to tabulate the facts reported in a large number of documents in a literature source (Srivastava et al., 2009). IE aspects can be used to support a fact retrieval service, or as a prerequisite for text mining based on conceptually annotated text.

1.7.2 Information Extraction (IE)

Information Extraction is the task of extracting factual assertions from text (Ananiadou et al., 2010). The definition of Information Extraction (IE) is also categorised as follows (Ananiadou et al., 2006):

- *Take a natural language text from a document source, and extract essential facts about one or more predefined fact types.*
- *Represent each fact as a template whose slots are filled on the basis of what is found from the text.*

These templates may include slots for the name, age, place, date, and relevant information depicting the characteristics of the entities to be described. Information Extraction is typically carried out in support of other tasks, and usually forms part of some application or pipeline of processes. The results of IE are typically isolated and stored within databases and subjected to data mining algorithms or querying (Elder et

al., 2012), integrated in knowledge bases to allow reasoning, or presented directly to users who require support in dealing with identification, assembly, and comparison of facts as in data curating tasks.

1.7.3 Linguistic Analysis

Linguistic analysis is a research field related specifically to multilingualism, which is the usage of multiple different languages co-existing in a single domain. As a single document may exist in multiple forms of languages, linguistic analysis seeks to implement the similarities of the traits from any language, be it from the morphological, syntax, sentence structure, or simply guided towards terminologies. The aspects of linguistic analysis usually incorporate several fields into one to enforce data extraction and categorisation, for example text and data mining and web scraping for the purpose of data categorisation in addition to extraction (Ananiadou et al., 2010; Banko et al., 2007).

1.7.3.1 Language Identification

The Web serves as an interpretation medium nowadays, where various materials in different formats were relayed in the form of information display. These information displays are mainly in the form of raw or processed text. As the usage of the Web already exceeds the population of the society, it is clear that various data are relayed in the form of native or standardized formatting. A sentence could exist in few forms,

after the pre and post processing of its information is conducted. As such, language studies emphasize the importance of identifying the language scope before it is further processed.

1.7.3.2 Identification of Grammatical Categories

Language relay exists in natural language form which is understood by humans. Therefore, for the analysis to be carried out the steps applied includes the identification of the nouns, verbs, adjectives, and adverbs in the texts of the corpus. This step requires further grammatical analysis. However, this procedure may be complicated by the presence of homographs. A homograph is a collection of words that exists in the form of normal sentence (Zhang et al., 2012), but may contain a definition that seems distorting rather than normal ordering of a particular sentence.

1.7.3.3 Disambiguation

Naturally, there may be several ambiguities within the source of a natural language text. This may be due to the polysemy of words (the fact of a word may exist in several variety of definitions) to ellipses (in a telegraphic style); homographs (duplication of a single word with similar spelling but different meaning within a sentence, for example “lead me to the lead mine!”); or as far as antiphrasis and irony, even the failure of detecting the presence of a word automatically (Indurkhya et al.,

2010). Anaphora also could lead the disambiguation effort to futile. The existence of anaphora could give rise to ambiguities that must be removed (Tuffery et al., 2011).

1.7.3.4 Recognition of Compound Words

A collection of sentences may contain ordering of words that consisted of mixed characters, be it in the form of alphanumerical of just in the plain alphabet structure. The field of language studies in a NER classification task consisted of the processing for a text and identification of the sequence of words that represent entities such as Person, Location, and Organization (Abdallah et al., 2012; Abdul-hamid et al., 2010; Alfred et al., 2014). Therefore, it is a must for the identification of appropriate categories to take place, where the proper expression of sentence presence is justified. For example, words such as *"1 April 2013"*, *"Perdana Menteri Malaysia"*, and *"Perbadanan Malaysia Berhad*" are groups representing a date, a person, and an organization.

1.7.3.5 Lemmatization

After the appearance of compound words in a corpus is identified, it is compulsory for them to be simplified without altering the definition so that the main themes could be extracted more easily (Abdallah et al., 2012). This step commences by the lemmatizing of the texts, which includes the classification of terms into their

canonical forms. The nouns would be placed in the singular category, while the various forms of the verb would be placed in the infinitive.

1.7.3.6 Variant Grouping

Closely related with lemmatization (1st stage of simplification), this stage involves the grouping of the variants located within the texts of the corpus. This form may exist in the form of graphic variants (words with almost similar spelling, for example "realise" = "realize"), syntactic variants (name of a man=a man's name), semantic variants ("X eats Y from Z" = "Z sells Y to X"), synonyms (similar analogies which represents the same meaning but different spelling in context. For example, US = USA = *United States of America*), para-synonym (words with closely related meaning. For example: *angry, discontent, dissatisfied*), besides full form of abbreviations (€ = EUR = *Euro*, BBC = B.B.C = *British Broadcasting Corporation*).

1.7.3.7 Theme Identification

For this stage, text analysis is completed by grouping all the terms around their respective themes. These themes are identified from the similarities representing the category of the language group itself. The grouping may occur within a tiered-level procedure.

1.7.4 Text Mining

Text mining had been described as the process of discovering and extracting knowledge from unstructured data (Ananiadou et al., 2006). Text mining consists of three subcomponents: Information Retrieval, Information Extraction, and Data Mining. Text mining generally involves a set of techniques and methods used for the automatic processing of natural language text data available in reasonably large quantities in the form of computer files, with the aim of extracting and structuring their contents and themes for the purposes of rapid analysis (non-literary), discovery of hidden data or automatic decision making.

1.7.5 Text Clustering

Text clustering is related to the field of text mining, which is the task that is applied to aid the organization, knowledge extraction, and exploratory search of text collections. Text clustering is one of the priority tasks of text mining, and may be applied for the purpose of organizing documents when there is no set of labelled data available.

1.7.6 Expert System

An expert system is a computer-based system with artificial intellect that emulates the reasoning process of a human expert within a specific domain of knowledge. Expert systems are applied in a variety of specific activities such as consulting, diagnosis, learning, decision support, design, planning, or research. Any expert system includes a knowledge base, a database, an inference engine and user interface for communicating in restricted natural language with the user. For expert systems, the knowledge base usually includes a set of production rules which connects antecedents with consequences, premises with conclusions, or conditions with actions. Each component on the expert system constitutes a specific feature that completes the entire system flow. There are five basic structures of an expert system: *database, inference engine, user interface, knowledge acquisition module*, and *expert system shell.*

1.7.7 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field related with both lexicon analysis and information extraction. NLP is the activity of processing natural language texts by computer to access their meaning (Indurkhya et al., 2010). NLP systems could analyse (or known as parse) well-formed sentences of great complexity, should the grammar of the language have been encoded for the system and the lexical resources (dictionary) cover the vocabulary used.

However, this is only feasible on a smaller scale. For a large body of documents, a significant proportion of sentences will not be fully recognized by the grammatical and lexical resources of any given NLP parser. An NLP system is targeted to resolve the goals of dealing robustly with parts of a text that fall outside available resources.

1.7.8 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a field introduced in 1996 that is related to automatic information retrieval. The objective of NER is to identify the expressions with special meaning such as person names, organizations, times, monetary values, and so on. These named entities categorization could be afterwards successfully implemented in other fields and applications, for example Question Answering (Q and A), Information Filtering, and including Information Extraction (IE) and Information Retrieval (IR). The fundamental of classification groups include Person, Location, and Organization. However, these classes had expanded its sub categories to up until 200 for the general domain (Alfred et al., 2013). There are currently two main paradigms for NER field: knowledge engineering techniques and machine language (ML) approaches.

1.8 Conclusion

This beginning section of the thesis introduces the concept of Named Entity Recognition, along with its importance in the fields of Natural Language Processing and Data Mining. Since its inception during the 6th Message Understanding Conference back in 1996, its pertaining concepts had been used as a guideline to extract raw data entities from a cluster of uncategorised text collection. As the nature of existence for information is widely prevalent on the Web nowadays as compared to traditional text material, the approach focuses mainly on the retrieval of information chunks from the web resource.

In the NER field of research, effort has been made to resolve the most acclaimed problems in IR and IE, which is the identification and classification of unsorted text from their respective categories. As such, there have been numerous attempts to carry out this effort, such as the implementation of classification algorithms, clustering of data according to their main categories, and to the more recent approach by applying rule-based nomenclature to properly synthesize the word root according to their classes. NER techniques apply learning techniques in order for the data to be assimilated and credible towards its usage in the entire system. To resolve the problems of lack in abundance for Malay text resource, the NER approach is applied in the upcoming research approach.

In the next section, discussions on research already conducted for Named Entity Recognition in languages closely related with Malay will provide an overview of the relevance of the approach to be used. Basic concepts that are applicable with

this research field would also be discussed in detail, for example Natural Language Processing, Text Mining, and Data Clustering. In this research, semi-supervised learning technique is going to be used along with implementation of clustering algorithm & rule-based technique in order to resolve the ambiguities of unannotated data in linguistic resource that is lack in availability. The chapter also emphasizes learning approaches in which clustering algorithms & rule-based pattern detection are used as a scaffold to assist in the categorisation of unannotated text data.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Identification of word categories had been an important trait for information retrieval, especially for field such as Knowledge Extraction to classify relevant details into its respective context. Named-Entity Recognition constitutes a component in Knowledge Extraction and is a beneficial procedure for information extraction. Information Extraction is a process that extracts information from unstructured articles to provide more useful information (Alfred et al., 2014) which is closely related to Knowledge Extraction. NLP handles the analysis of texts derived based on a set of theories and technologies (Al-shuaili et al., 2016).

The dexterity of handling applications is distinguished via the massive amounts of prior knowledge in order to perform complicated interdependent decisions. Among the objective of NER research is to detect the availability of named entities in open documents. The data provided may be devised from searches within credible text sources, for example websites, text corpora, and online newspapers. Most named entity refers to a phrase representing a specific class. There are already several major NER research that have successfully classified data clusters according to their own groupings and categories; however, the success and high precision rate depends solely on the domain of the studies (Alfred et al., 2014). The corpus analysed in question must have the criteria in Figure 2.1.

i.	***Must be in a data processing format,***
ii.	***Include a minimum number of texts,***
iii.	***Sufficiently comprehensible and coherent,***
iv.	***Does not have too much different theme in each text, and***
v.	***Avoid as far as possible the use of innuendo, irony, and antiphrasis.***

Figure 2.1. Criteria For Locating Rules in a Number of Predefined Themes Per Single Domain (Tuffery, 2011)

This includes its use in discovering hidden information and in decision making, data and news filtering. Allocation of hidden information via text mining and decision making are classified as forms of information retrieval, whilst quick analysis is a form of information extraction (Srivastava et al., 2009). Information retrieval is related to documents in their entirety and with the themes in which they are closely associated

to. Information retrieval is utilized in the comparison of documents and document type detection. Another field which information extraction is involved in is for the search for specific information located in the documents. This process does not take into consideration any fundamental comparisons between documents, nor the order and proximity of words, to discriminate between different statements which have identical keywords.

Information extraction commenced with data obtained from natural-language data chunks before focusing on the development of a structured database (Banko et al., 2007). Information extraction has a higher complexity, as it requires the usage of lexical and morphological-syntactic analysis to recognize the constituents of the text, their nature and their relationships. This chapter contains an overall review on the characteristics of NER research in computational linguistic, together with comparisons between previous works that is relevant with current ongoing research works in proper noun extraction specifically for the Malay language and its consecutives.

2.2 Named Entity Recognition

The application of NER within the field of Natural Language Processing was considered an essential approach. As NLP generally consisted of raw text data collection accumulated from a single domain, these data still existed without proper sorting and grouping according to its own categories. The incorporation of NER

systems in identifying proper name entities in open or closed domain sources have been crucial to improve the applications of NLP applications.

During the handling of massive amounts of information, NER systems could assist in Information Extraction, Information Retrieval, in addition to Question Answering (QA) task accomplishment (Alfred et al., 2014). This task also bears its own significance from the recognition and classification of defined named entities from large text or in the general context of newswires (Mohit et al., 2012; Althobaiti et al., 2014). Research work during the early days identified the NER problem as the distinguishing process of "proper names" in general (Al-shoukry et al., 2015). As mentioned above, the most studied types of information entity are three specializations of "proper names", including "Persons", "Locations" and "Organizations". These distinguished categories are known as *enamex*. From these *enamex,* further categories could be devised according to the three majority types. The "location" types can be divided into multiple subcomponents such as city, state, country, etc. (Abdallah et al., 2012; Alfred et al., 2014).

Another category that could be divided further into sub-categories like "politician" and "entertainer" is quite common and used in combination with other cues for medication and disease names extraction. As its name suggests, the field of Named Entity Recognition involves the identification and classification of named entities from an open-domain text. The NER research field is considered a significant subfield in the context of Information Retrieval and Information Extraction. The areas where NER research were seen to be most dominant include Question Answering, Machine Translation, and Information Retrieval (Benajiba et al., 2008;

Benajiba et al., 2008). Most of the tasks involved in the NER research field is divided into classification category; that is the identification of named entities from the given domain, and the classification of these named entities according to their set of extracted features into the predefined class sets (Benajiba et al., 2008).

Basically, the NER processes were carried out to assist users to produce a corpus that had its content been classified according to respective groups, in this case the main categories as cited in MUC-6 (PERSON, ORGANIZATION, and LOCATION). Together from these main classes, subclasses that corresponds to the main categories can also be directly derived. Named Entity Recognition is among the key information extraction tasks used for the identification of names for entities such as people, locations, and products. Reliable entity recognition and linking of user-generated content are catalysts for other information extraction task, as well as opinion mining and summarization. NER is usually assigned for the detection of named entities.

The accomplishment of this task is considered a difficult task due to the capitalization of words, besides the words themselves is annotated by Automatic Speech Recognition (ASR) technologies. Optical Character Recognition (OCR) also suffers from similar complication in terms of detecting named entities (Bontcheva et al., 2013). A renowned English NER analysis tool, Stanford NER system also used machine learning-based method for the detection of named entities alongside the distribution of the system with CRF models for English newswire text. During the previous times, researchers utilized handcrafted rules to resolve the available problems. However most recently, research began to implement supervised and

unsupervised machine learning to automatically stimulate rule-based systems on a collection of training examples. In this context, a rule system was preferable when training examples were not available (Rahem et al., 2015).

Message Understanding Conference is responsible for the introduction of the research done in NER. Held in November 1995, the conference had introduced a set of four evaluation tasks to review the performance of information extraction systems applied to a common task, among these include co-reference, template elements, scenario templates (also known as traditional information extraction), and named entity recognition (Grishman et al., 1996). The introduction of these specifications is in accordance with the requirements to push the current information extraction system with a better compatibility with new domains, along with the encouragement for more basic work on natural language analysis via the evaluations of some basic language analysis technologies.

To detail the task of NER, the MUC-6 had involved the recognition of entity names (in accordance with the name for people and organizations), place names, temporal expressions, and specific types of numeric expressions. This process is intended to provide a direct, real value for text annotation to ease the effort of searching within the domain, along with assisting the relating language processing tasks, for example IE and IR. Among research done to recognize named entities in other languages includes English (Nothman, 2008; Prasad et al., 2015), Chinese (Che et al., 2013; Duan et al., 2011; Hu et al., 2016), Arabic (Abdallah et al., 2012; Aboaoga et al., 2013; Al-Saleh et al., 2016) and Indian (Ekbal et al., 2011; Ma et al., 2013).

The following figure is an example of challenges faced in determining the proper detection of common NER schema.

Soccer - Blinker Ban Lifted. London 1996-12-06 Dutch forward Reggie Blinker had his indefinite suspension lifted by FIFA on Friday and was set to make his Sheffield Wednesday comeback against Liverpool on Saturday. Blinker missed his club's last two games after FIFA slapped a worldwide ban on him for appearing to sign contracts for both Wednesday and Udinese while he was playing for Feyenoord.

Output from Stanford Named Entity Tagger

SOCCER - **<PER BLINKER>** BAN LIFTED. **<LOC LONDON>** 1996-12-06 **<MISC DUTCH>** forward **<PER REGGIE BLINKER>** had his indefinite suspension lifted by **<ORG FIFA>** on Friday and was set to make his **<ORG SHEFFIELD WEDNESDAY>** comeback against **<ORG LIVERPOOL>** on Saturday. **<PER BLINKER>** missed his club's last two games after **<ORG FIFA>** slapped a worldwide ban on him for appearing to sign contracts for both **<ORG WEDNESDAY>** and **<ORG UDINESE>** while he was playing for **<ORG FEYENOORD>.**

Figure 2.2. Problem in Systemizing a Proper Named Entity Categorisation.

The research field of NER in terms of text and data mining is targeted with the objective of processing raw text types along with the identification of word sequence which represent entities for the major classes, such as PERSON, ORGANIZATION,

and LOCATION. There are various approaches included in order to efficiently carry out the task of data clustering and classification, where the processes are performed in a few pre-planned steps or simply broken down into sub-problems. Both tasks for recognition and classification were divided into subcategories to cater with this objective. NE recognition is aimed at marking the boundaries of the mention of an entity within a running text context, whereas for NE classification it is scheduled to assign the correct categories to the text span. Recent studies have suggested that existing taxonomies for entity classes have reached 200 for the general domain (Nothman et al., 2013).

The primary training approach used for data training involves the usage of machine learning techniques that had been explored in major evaluation forums such as Conference on Natural Language Learning (CoNLL). Data comparison obtained from CoNLL 2002 and 2003 were utilized for the exploration of NER in various major languages such as Spanish, Chinese, English, and German (Nothman et al., 2013; Chou et al., 2014; Eduardo et al., 2015). As output from NER research was measured in a variety of formulas, among those prominent is F-Measure, Recall and Precision Rate. It is proven that some outstanding supervised approaches had achieved score rate at 72 and 89% in F-Score (Manning et al., 2009). The frequency for the results depended on resources used in order to obtain them, such as NE listings, gazetteers, and POS taggers.

A typical supervised classification task consists of learning the classification rules from a myriad of test sets consisting of labelled examples, in order to later apply the learned rules to previously unseen examples (Witten et al., 2011). Thus, this adds

the available clusters to a specific label. Unsupervised classification task, known as clustering, consists in finding internal structure in a set of data items in order to group them together. The unlabelled data items, or the least a larger amount of such items, were available to the classifier at the training stage and are used for training. This training includes methods for refining regularities in the dataset.

As this progresses, there is no specific predefined inventory of labels is used in unsupervised classification. This task consists of using both a small number of labelled examples and a large number of unlabelled examples in order to learn to assign labels from a predefined inventory to unlabelled examples (Althobaiti et al., 2015). This characteristic enables a few positive traits to emerge from the process. The unsupervised classification (or unsupervised learning process) assigns specific labels using much smaller training labelled dataset. A huge advantage of this process over supervised classification lies in its ability to compensate the lack of information from labelled examples by information extracted from a large set of unlabelled examples (Miner et al., 2012).

The phases of data collections are done in non-parallel stages, in which one level occurs only after certain prerequisites had been fulfilled on a particular process. Among the various pre-mediated stages devised for the purpose of data collection in both text and data mining field includes pre and post processing. These stages begin in the primary stage of data collection accumulated from credible text resources for usable articles. These text resources that were accumulated emphasizes on relevant topics that contain credible text data to be gathered for further classification and

recognition. The predominant steps include pre-processing, data cleaning, and manual tagging.

2.2.1 Text Tokenization

The lexical analysis of available corpora requires an extensive study at how words and other tokens could be recognised, analysed, and formally characterized to enable further processing. This trait is applied towards text corpus that contains a considerable huge amount of raw text data awaits to be categorised via the conventional NER classification, along with the text and data mining processes. Before any linguistic analysis can be implemented, the basic tokens involved such as words, acronyms, abbreviations, numbers, punctuation symbols, and morphological syntax must be identified. Tokenization involves the segmentation of the input text character stream into linguistically plausible units (Abdallah et al., 2012; Ananiadou et al., 2010; Bontcheva et al., 2013).

i. **Abbreviations**

Words are not always kept in a structure separated from other tokens by features such as white space. Even the presence of a period may signal an abbreviation which must be identified from a sentence-delimiting period. This problem becomes more of a hindrance should the character appears at the end of the sentence. This feature may interfere with the process of sentence boundary detection.

ii. Apostrophes

A manifested clitic as an apostrophe added by the sequence of one or more letter need to be isolated from its harbouring word, because it denotes a linguistically meaningful relationship between two entities. For example, the English possessive marker's in IL-10's biology term *cytokine synthesis inhibitory activity*.

iii. Hyphenation

The fact that tokenization should or shouldn't return one or more tokens for hyphenated word (text-based) remains ambiguous, as illustrated when authors are not compliant with the rules of hyphenation. For example, *coexist* versus *co-exist*.

iv. Numbers Existing in Multiple Formats

The appearance of numbering schemes in numerous formats contains ambiguous separators. This dampens the effort to isolate numeric and alphabetical characters. There are a few entities that exist in multiple formatting, such as phone numbers, dates, addresses, and account numbers. This is for example, 505 354 and 505, 354.

v. **Sentence Boundary Detection**

Sentences usually appear at the limit (**demarcated**) by the typical sentence-delimiting punctuation marks. Other punctuation marks serve to indicate the boundaries of sentences. Examples include the colon, semicolon, and M-dash.

2.3 NER for English

In order to appropriately perform categorisation on raw, unannotated datasets in the effort to enable proper grouping for new, named entities, there is a need to devise a suitable strategy. Among the approaches widely utilized in the fields of NER includes Document Clustering (Chen et al., 2011). Languages were divided into appropriate linguistic clusters, as monolingual cluster is composed of documents written in one single language whilst multilingual cluster consisted of documents written in different languages. The strategy that utilizes language-independent representation attemts to normalize or standardize the document categories in a language-neutral way (Liao et al., 2009). These particular tasks range from the mapping of text contents to an independent knowledge representation (such as thesaurus), or via the recognition of language-independent text features within the documents (Liao et al., 2009; Smith et al., 2010; Liao et al., 2011; Ma et al., 2013; Král, 2014).

NER research is performed widely in English, with its concise identity in the world's language hierarchy (Nadeau, 2007). Both supervised and unsupervised training approaches include appropriate NER algorithms during the data extraction

process. The efficiency of the algorithm itself relies upon the precision of the data entity extracted, along with the size of the targeted corpus. NER research was derived from major linguistic categories investigation before being specified into existing languages that are lacking. As such, NER studies were closely associated with Natural Language Processing. Liu applied a K-Nearest Neighbour classifier to assign initial labels to words based on their contexts, before the resulted tags were exposed to a CRF labeler (Abu Bakar et al., 2013).

A few significant works include utilized tools and processing techniques to initiate the named entity recognition process (Ma et al., 2013; Poria et al., 2013; Le Nguyen et al., 2014). They used a gazetteer, stopword list, current, previous, and next words, cross-language capitalization, POS tagging, base-phrase chunking, and adjectives indicating nationality. A simplified feature set that rely based on character level features, including leading and trailing letters is applied in words (Abdul-hamid et al., 2010). The most prominent language research in the NER community that had started after the Message Understanding Conference (MUC-6) in 1996 is English.

The majority of studies are for English stems as it is within the top 5 languages most often used as a communication tool in the world, the others being Chinese, French, Arabic, and Spanish. This factor alleviates a larger proportion of the research outcome to resolve the language independence and multilingualism problems. Earlier studies of NER research in English have utilized tools such as gazetteers and dictionaries to annotate a list of names or common words that were often discovered during the named entity classification. This in turn gives rise to the three major data training approaches: *supervised*, *unsupervised*, and *semi-supervised*. Below is a brief

review on the definitions of the three training approaches that have been discussed prior.

Supervised learning: A system that detects a large annotated corpus, entities list memorization, and disambiguation rules initiation based on discriminative features.

Unsupervised learning: A varied approach from supervised learning which relied heavily on machine participation, unsupervised learning uses data clustering to accumulate named entities from the derived cluster groups based on the similarity in the word context. This approach depends on the derivation of lexical patterns and the existence of huge unannotated text collection.

Semi-supervised learning: Considered a mixture of traits from supervised and unsupervised learning, the techniques for semi-supervised learning includes a considerably minute supervision degree as compared to its counterparts to begin the learning process. The system classifies the new data entities based on the collections of existing annotated datasets, which is also known as seed sets. The learning process is repeated to the newly discovered examples in order to classify the new related contexts.

The baseline procedures of NER extraction from English-based corpus is basically divided into a few phases dedicated to initial data extraction, pre-processing, classification rules, gazetteer listings, and interlink among entities within the similar domain. English named entities in Wikipedia with named entities from

other languages is classified with category keyword heuristics embedded with a training corpus (Nothman et al., 2013). Simile heuristics for the identification and classification of entities is also applied in the context of a given document (Nadeau, 2007). The framework of this research is depicted in Figure 2.2.

A project of exploiting the text and structure of Wikipedia corpus was tested in order to propose a Learning Multilingual NER (Nothman et al., 2013). The research classifies the Wikipedia article into a NE) type and labels approximately 7200 articles manually. This approach is done across nine languages. The links available between articles were transformed into NE annotations by the projection of the classifications from the article's content into the anchor text. From this heuristic approach, the result showed an accuracy of approximately 95%. This technique does not compare the newly obtained data with existing ones. However, modification occurs in the Wikipedia corpora by introducing extraneous links to properly synchronize the automatic annotations to approved standardizations.

Kazama and Torisawa reported an increment of 3% for F-Score when many Wikipedia-derived gazetteer features were included in the NER system that they researched (Althobaiti et al., 2014). However, a gazetteer could only be constructed from pre-existing datasets. A state-of-the-art English CoNLL entity recognizer incorporated 16 Wikipedia-derived gazetteers (Turian et al., 2010). The tools such as gazetteer, however, had their own flaws, in the sense that gazetteers do not allocate the important contextual evidence as available in annotated corpora (Nothman et al., 2013). In the NER research of Nadeau, the scope emphasises on

the development of a NER system for the classifications of rigid designators in text (Ojo et al., 2012). A proof-of-concept semi-supervised system is implemented that improves the capability to recognize four NE types. The capacity is further expanded via the improvement of key technologies, before the concept is applied into an entire hierarchy consisting of varied 100 NE types.

This research contributes in the creation of a proof-of-concept semi-supervised NER system that could isolate the abundance of noise for the purpose of NE listing generation. This procedure also directly developed an acronym detection algorithm. This NER system is proven to possess the advantages of classifying designators in text fields, for example proper names, scientific species classification, besides temporal expressions. The results obtained showed that minute supervision is possible for the development of a complete NER system (Ojo et al., 2012). Research has identified several traits that should be placed into consideration during the categorisation of an unannotated document, for instance the document and corpus features. The next table illustrates some of the features for document elements that could be discovered in a general unannotated retrieved dataset. The features of a single, unannotated dataset within a document are explained as in Table 2.1.

Note: The blank space in **Lists** means the Ambiguous or Unambiguous Entities that make up the corpus.

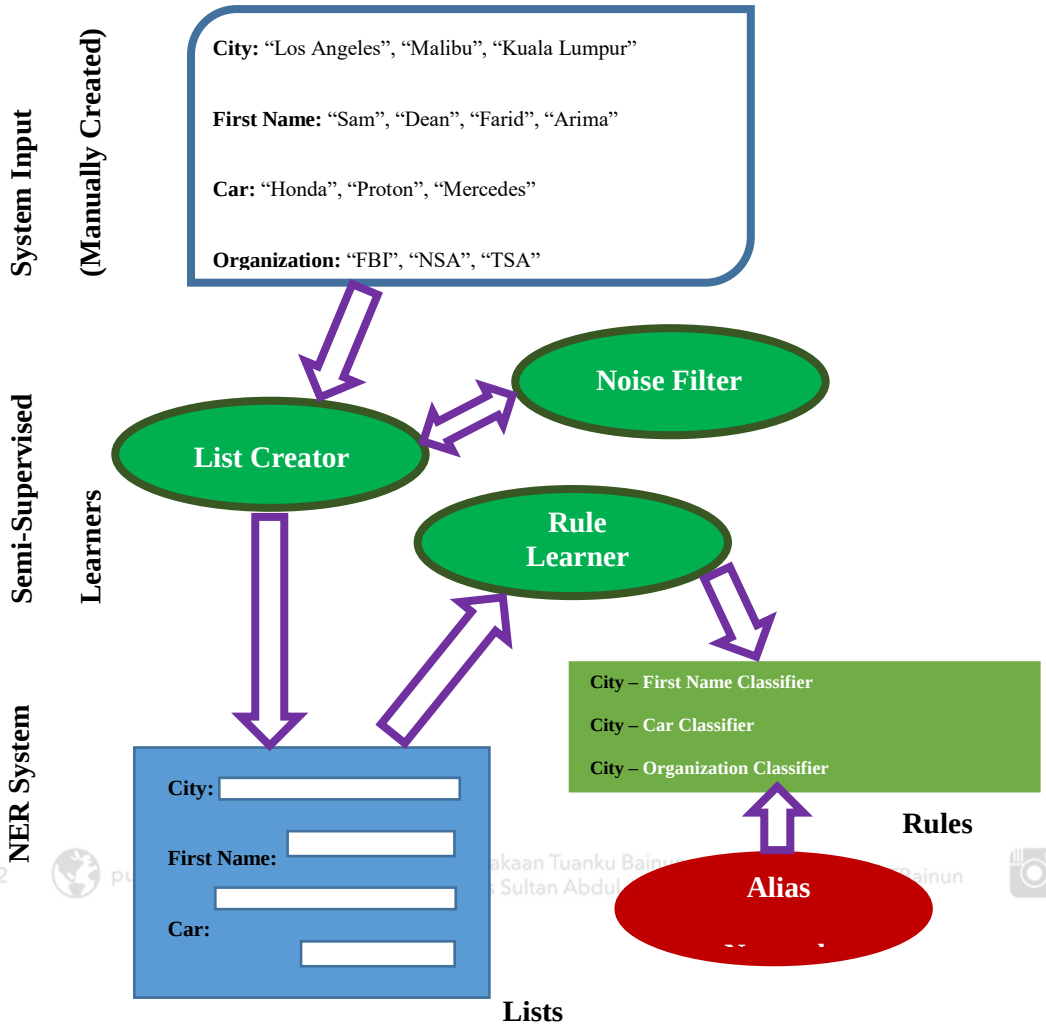


Figure 2.3. Nadeau's Framework of Semi-Supervised NER Extraction for English Word Collection

Within the context of a single document, there is bound to be a few instances of repetition of characters, which is identified within that particular text domain as possessing multiple occurrences. This factor influences the redundancy factor of the categorisation for the datasets within the domain itself. As indicated in a few prior research (Nadeau et al., 2007) had identified words that appear both in upper and lower case form in a single document (Ojo et al., 2012). These characters were

hypothesized as common nouns that appear in both ambiguous (beginning of sentence) and unambiguous position.

Table 2.1

Document Features in an Unannotated Dataset

Features	Examples
Multiple Occurrences	• Other entities in the context • Upper-case and lower-case occurrences • Anaphora, co-reference
Local Syntax	• Enumeration, apposition • Position in sentence, in paragraph, and in document
Meta-information	• URI, email header, XML section • Bulleted / numbered lists, tables, figures
Corpus Frequency	• Word and phrase frequency • Co-occurrences • Multi-word unit permanency

The classification of a word character in a single domain is unique to its own definition and cluster typing. Co-references to particular word include occurrences of a word sequence referring to a given entity within a document (Carlson et al., 2010). In order to appropriately deriving these local syntaxes from co-references in a single domain, the task would involve the exploitation of the context of every occurrence in that particular domain (for example, He **ate** the food; She **ate** the food; They **ate** the food, and so on). Deriving features from aliases are mainly performed via the leveraging of the union of alias words (Carlson et al., 2010).

Meta-information, otherwise also known as the incorporation of HTML tagging is a common prerequisite in every development of web pages nowadays. These include the embedding of metadata in every component of document, according to its typing and classification (Indurkhya et al., 2010). Most metadata embedded in documents could be utilized directly. Web authors implement these features from the provided articles. As such, these metadata carries distinct feature that distinguishes them from other characters' traits (Indurkhya et al., 2010). For example, news always begins with a location name as marker and ends with date where the article is authored. The precision of retrieving certain data entity information from the system that is trained, along with the accuracy in which the data obtained is relevant to the actual source, depends a lot on its appearance within a single corpus, also known as its corpus frequency.

Word and phrase frequency determines the distinctive word's rarity or vice versa, its common existence (Chen et al., 2011). Along with its frequent appearance in the domain, its co-occurrence with other text entities that have similar points among each other is another trait would determine its relevance in the corpus. Although the frequency of appearance of a word in that domain would determine the efficiency of the entire corpus itself, the incorporation of that single word unit in multiple similar meaning (hyponym) is also an element that would contribute to the overall performance of data training (Elyasir et al., 2013).

2.4 NER for Arabic

The Arabic language is used by an estimated amount of 300 million people of the world, and has been made an official language at the United Nations, and were spoken by more than 300 million in the Arab world (Abdallah et al., 2012). Existing studies have indicated that there is a level of scarcity for the NER research dedicated towards the Arabic texts.

For Arabic related text processing, a machine language approach is emphasized based on the neural network technique. This technique is centred on the abilities of learning progress based on the availability and existence of the previous data. Furthermore, this technique could be deployed towards larger database which organizes new information. The named entities of Arabic language consisted of singular and composite types for Arab personal names (Abdallah et al., 2012). The Arabic language also is characterized by a rich vocabulary besides a complicated morphology. As indicated in the research (Aboaoga et al., 2013), among the challenges identified for the field of Arabic NER studies lies with the complexity of the morphological system itself, written text non-standardization, ambiguities within the word character itself, and lack of resource abundance. The derivation of gazetteer via the clustering of words in unstructured text yielded higher gains (Althobaiti et al., 2014).

NE research for a particular language is usually initiated towards the linguistic category in which the language to be researched holds the closest similarity to, in terms of morphological, syntax, and sentence structures. A parallel corpora in

Spanish and Arabic, along with an NE tagger in Spanish is applied as a baseline system to perform name tagging in Arabic corpus (Nothman et al., 2013). They applied sentence pair alignment to a simple mapping scheme to transliterate all the words in the Arabic sentence. The results return paired matching NE in the Spanish sentence with the NE detected in Arabic. The precision for the corpus has been improved by applying a filter to the Arabic words to omit the stop words from the possible transliterated candidates. This approach had been indicated to only be applicable should the type of the corpus is parallel.

Among the prominent studies that have successfully implemented the rule-based method, along with minor semi-supervised approach for the Arabic language are (Abdallah et al., 2012; Aboaoga et al., 2013; Al-shammaa et al., 2015; Elsayed et al., 2015). Relating to the same morphological study, the identical points that can be reviewed from these studies is the manipulation of the Arabic noun scheme which are classified in the Arabic morphology and grammar rules without the use of any gazetteers. Another work also investigated the statistical approach towards NER using probabilistic models (AbdelRahman et al., 2010). In this context Maximum Entropy was used before applying Conditional Random Fields (CRF). The research integrated the corpus that the researchers developed themselves called ANERcorp that consisted of a training and test corpus annotated specifically for the task. The total performance assimilates all features in terms of precision, recall, and F-measure, with average scoring of 86.9%, 72.77%, and 79.21%, respectively. The priorities targeted in the NER field is the attempt of incrementing performance accuracy with respect to the classification and extraction of named entities.

Research applicable in the concept of implementing fuzzy algorithms is done by Naji F. Mohammed and Nazlia Omar in 2012. The research uses machine learning approach to isolate named entities from articles composed in Arab based on neural network technique. This research was done with the factor of resource scarcity for Arabic named entities along with the lack of accuracy in previous NER for Arabic systems. Naji stated that there are two major challenges in the focus emphasized on Arabic NER studies: absence of capital letters in the orthography, and highly inflectional Arabic language (Mohammed et al., 2012).

The absence of capitalization letters in the orthography made the effort to detect the named entities available to be more difficult (AbdelRahman et al., 2010). As the Arabic language is varied as compared to Latin, the Arabic language has no such unique signals that lead to the recognition of named entities. Other languages possess a signal in the orthography which is the capitalization of the first letter. These features are important, as capitalization is often used to point to a word or a string of words. The Arabic language also contains morphologies and terms that are highly inflectional. The word structure consists of more than single affixes. These forms of prefixes can be articles, prepositions, or conjunctions. The suffixes could also be in the form of objects or personal and possessive anaphora (AbdelRahman et al., 2010). This problem is resolved via the segmentation of each word as a pre-processing step (also known as tokenization). This step would make the named entities always appear in the same form, besides reducing the appearance of alternate forms of the contexts in which the named entities would appear.

Elsayed's work with Arabic corpus has two sections: processing Arabic nouns extraction without any assistance of gazetteers, and manipulating the nouns extracted based on gazetteers where the rules is targeted to be applied (Elsayed et al., 2015). The system implements the rule-base aspect to the classified nouns, where the words are delegated to tags with most distinctive features, such as person name, title and country, and date-time. Within Arabic Named Entity classifications, gazetteers play a huge role in terms of entity extraction.

The system functions via applying methods of extracting nouns by the Next Nouns algorithm that is used as Arabic grammar rules, and applying another secondary noun measure in the form of classification nouns that specifies in recognizing and classifying the name of distinctive classes, for example person name, and title. The combination of these approaches together with gazetteers improved the performance, as the system relies on grammar as a guidance to recognize other entities.

Nouns Algorithm

The system is revised to apply NER algorithms to extract Arabic nouns by using iterative algorithms based on Arabic grammar in Modern Standard Arabic.

Classification Nouns

In this stage, the system uses a GATE system as a foundation to apply gazetteer extensions containing lists of person names, titles, cities, and countries. GATE

(General Architecture for Text Engineering) is a language engineering environment which contains various gazetteers useful for named entity recognition tasks.

Figure 2.4 demonstrates the intended approach to extract Arabic noun based on rule pattern approach as explained in the above section. Research dedicated towards the area of Arabic NER is considered to be categorized as still in its premature level of development. Abdul Hamid and Darwish carried out a study where fundamental sets of features were applied to firmly distinguish NER for Arabic without the common necessities for morphological or syntactic analysis or gazetteers (Abdul-hamid & Darwish, 2010). These training and testing data were extracted from experiments. Neural networks were identified to possess a few positive attributes that motivates its implementation towards a variety of system applications (Mohammed et al., 2012). These include learning adaptability, definition derivation from complex or imprecise data, self-organization, and possess different structural designs that produced a diverse range of algorithms.

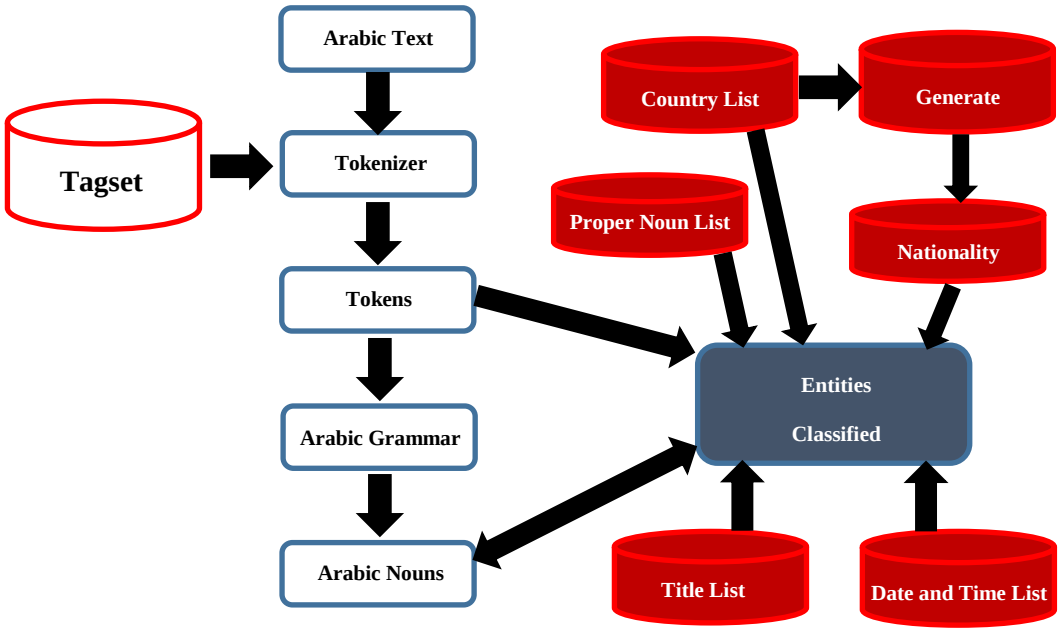


Figure 2.4. System Architecture of Arabic Rule-based NER (Elsayed & Elghazaly, 2015)

An NER system had also been developed based on Arabic text constructed for Maximum Entropy (Abdul-hamid et al., 2010). This research scope developed their own training and test corpora in addition to gazetteers to evaluate and present their system output. Subsidiary research done by the likes of pioneer researchers, such as Nadeau and Sekine, surveyed NER field for a wide range of linguistic aspects along with the exploration of a myriad of features, some of them include contextual features captured by the sequence labelling algorithms, character-level features that include the first or last few letters of words, and part-of-speech and morphological features.

As there is also quite scarce resource for Arabic NER, most of the research initiated try to resolve the ambiguity and lack of abundance of credible and labelled Arabic corpus. A few prior research had indicated about the fact of the lack of gazetteers or reliable NE candidates (Tran et al., 2015). Due to this nature, attempts to establish Arabic corpus were done with resources which are widely applicable, including microblogs and Twitter feeds.

Another NER system trained on news data on Arabic tweets had seen training results retrieved that were far lower than those feeds retrieved on news materials (Darwish & Gao, 2014). This approach used cross-lingual features from English to improve Arabic NER. The recall rate of NER on Arabic Wikipedia using self-training and recall-oriented classification had also been tested (Mohit et al., 2012). Besides that, a hybrid NER system for Arabic is also suggested that integrates rule-base system with a decision tree classifier (Abdallah et al., 2012). The results indicated the increment of F-score by 8% and 14% as compared to the original rule-based system and pure machine learning methods. Oudah and Shaalan also developed a hybrid Arabic NER system that integrated a rule-based approach with three different supervised technique: notably decision trees, Support Vector Machines, and logistic regression (Abdallah et al., 2012). Their experimental hybrid system exceeded some notable state-of-the-art Arabic NER systems on fundamental test sets.

Besides focusing on Arabic morphology NE extraction, Arabic NE research includes a wide array of neural network incorporation. NER system based on a neural network approach is applied to recognize named entities of Arabic texts (Mohammed et al., 2012). The research proposed the usage of a machine learning approach as a

classification of NER from Arabic text based on the neural networks technique. For the task of neural network approach, it is intended that the system could identify component patterns and performs intelligent decisions based on existing data. This could in turn also be applied towards the classification of new information contained within sizable databases. The precision of the system reached 92%, which in turn visualized the potential of neural network approach against the Decision Tree algorithm in terms of performance and accuracy, where Decision Tree algorithm gained a close value of 87% for measurements in precision values.

2.5 NER for Indonesian and Malay Language

The existing of various languages consists of different morphologies. Therefore, different NER processes must be applied to specific ones and not only using singular, unified NER technique or algorithm. There are a few available NER systems dedicated towards major languages, for example English, Indonesia, Hindu, Arabic, and Chinese. For the Malay Archipelago, there is no existing system that is devised for the detection of named entity types for the Malay language.

Different language may require different techniques in order to properly identify the named entities present. As shown in the English NER, the detection of named entity types for written articles could be easily carried out through the identification of the proper nouns. These proper nouns normally start with a capital letter. This technique however loses its efficiency when applied to linguistic corpora with different morphological structure. For example in Arabic language, the

deficiency is present in the fact that it does not contain such unique symbols that could be used for NE detection (Mohammed et al., 2012).

The Malay field of linguistic research is still currently undergoing development in NER studies. Even though Malay corpus still lacks availability, there is already a few related research done which involves the identification of unannotated text and part-of-speech tagging (Abu Bakar et al., 2013; Alfred et al., 2014; Mohammed et al., 2012; Mohd Don, 2010). For example, Rayner had proposed a Rule-Based NER algorithm specifically for Malay articles (Alfred et al., 2014). These algorithms detect the named entity in a particular content via a set of rules and a list of dictionaries that were manually defined by human. Extraction patterns are determined based on the pattern base for location names, organization, and so on. These patterns are made up from grammatical, syntactic, and orthographic features.

This proposed NER was devised based on a Malay Part-of-Speech tagging feature besides contextual traits that were improvised to handle Malay articles. A number of manually accessed dictionaries were created to handle the main Named Entity (NE) attributes, which are Person, Location, and Organization entities. The F-Measure obtained was significantly high, at 89.47%. Having more complete dictionary and correct rules would improve the proposed Malay NER algorithm (Alfred et al., 2014). A Rule-Based Pattern Extractor and a semi-supervised NER approach had been implemented for extracting patterns automatically from a limited scope of corpus (Sari et al., 2010). Extraction patterns were constructed using the Stanford's Part-Of-Speech tagger and Link Grammar Parser for the identification of named entity. This extraction pattern is used as a catalyst for the semi-supervised

NER to further classify entities. The experimental results demonstrated that the precision achieved for the data search is around 50 to 70% on F-Measure, even if only two features are used (Sari et al., 2010).

As relatively Malay corpus availability is still considered scant until recent times, efforts have been made by researchers to alleviate the problem of lack in Malay NER systems. The problems of a lack of Malay resources have been attempted to be overcome via the incorporation of Information Extraction approaches, among these include POS tagging, association rule extraction, and classification of a Malay noun according to their distinctive categories. Few research categories had managed to emphasize the importance of phase-by-phase detection of Malay word structures, particularly by breaking down upon its morpheme and syntax components (Mohd Don, 2010; Abu Bakar et al., 2013).

One particular element of similarity that could be shared among the effort to detect Malay proper noun is illustrated in Rayner's work, where research for Malay noun entities is conducted via emphasis on tokenization and evaluation on existing dictionary and gazetteer. As such, supervised or unsupervised approach is seen as both lacks in fulfilment for carrying out the task of entity categorisation properly. External schematics such as rule-based features could be included during named entity categorisation so that the problem of lacking in annotated corpus resources could be overcome. Among the phases that is seen practical in assisting the discovery of new data entity includes tokenization, part-of-speech tagging, and applying appropriate rule-based features into the class suffixes.

Figure 2.5 show the process of implementing rule-based NER for Malay language, as suggested in prior research (Alfred et al., 2014). The entire extraction phase is carried out based on the rule-based POS tagging process for Malay language, along with contextual features rule. These rule schemas were derived from neighbouring counterparts, such as Indonesian and Iban languages which share the most similar traits to Malay noun morphology. The rule system functions by applying a specific rule to the current word to interpret whether it exists as an entity or other word compounds. The rules were also developed in order to detect the 3 major categories of named entities, which are Person, Organization, and Location. These words undergo tokenization and are evaluated via POS tag dictionary. The classification of the tokens takes into consideration words that are placed under proper nouns category. The function of the framework illustrated in Figure 2.5.

These nouns would be applied with rules, unless they are based on location and person prepositions. The extracted nouns would be screened for suffixes from major NER categories, starting from organization suffixes. After the detection of organization entities, location prepositions are applied to certify the name entity for location. The process continues with the screening of person preposition, organization rules, location rules, person rules, before eventually the next token order is retrieved from the text collection. The entire process relies on pre-mediated rules before the experiment, and the process is repeated on all the other unannotated datasets.

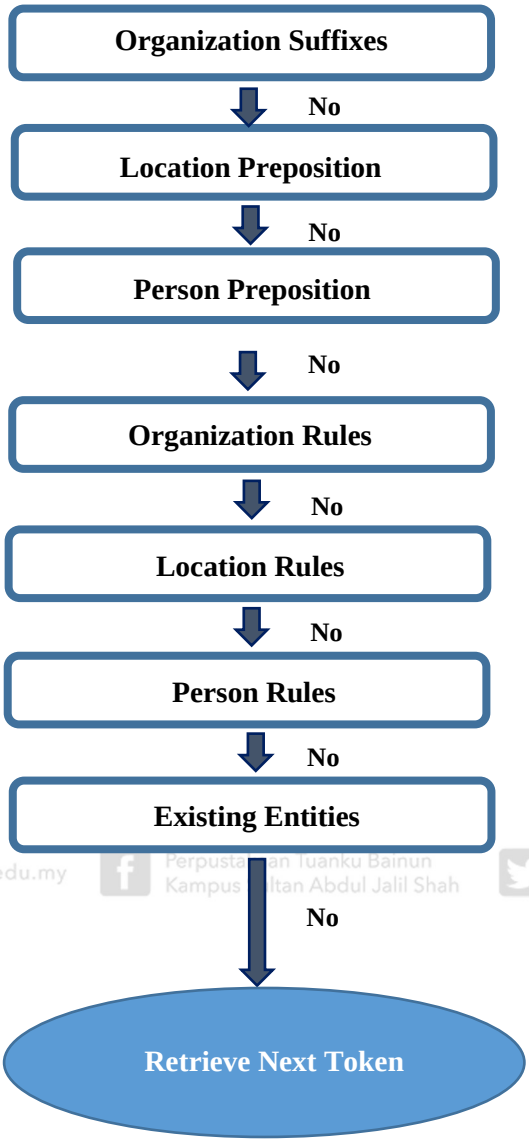


Figure 2.5. Proposed Framework for Malay NER (Alfred et al., 2014)

2.6 Proper Noun Detection and its Relevance in NER Categorization

Categorization of text data can be arduous and considerably meticulous, especially when there are no key guidelines to highlight its existence and how each of them

correlates respectively within a certain linguistic domain. The frequency of proper noun is already a common sight within any textual documents for languages which represent their contextual information in alphabetic characters. The detection of NE is very vague, unless there exist gold data or referral resources to authenticate the credibility of the classified words. NE could present as proper names and numerical expressions, however this magnitude of classification would become complicated as more features is brought into the fray such as class names and ordinal words (Sekine, 2010).

Previous works had seen improvisations done on improving the classification of NE via various learning techniques that focuses on categorizing labels based on predicting the availability of language features that is short of content and lack in context information which were not classified entities in their own right to extract common names, countries, locations, and temporal expressions (Abu Bakar et al., 2013; Alfred et al., 2014; Tran et al., 2015). Proper noun have several major distinctive categories that made them more convenient for detection into NE, such as normally beginning with a capital letter (Alfred et al., 2014). Proper nouns are applied to disambiguate unique named entities for majority language articles in Romanized form, word ordering chronologically from left to right. However certain clitic languages that does not use alphabets in its syntaxes are not subjectable to this rule (Abdul-hamid et al., 2010; Luo et al., 2016).

Certain techniques have been introduced to appropriately detect the presence of NE stemming from proper noun annotation. These includes rule-based NER that would indicate the presence of NE via accordance of specific rule set and a dictionary

list that have been determined by human efforts (Rahem et al., 2015), POS taggers trained from treebanks in the newswire domain (Gimpel, Schneider, O'Connor, & Das, 2011), and Twitter entity disambiguation dataset on proper nouns existing in English (Derczynski et al., 2014). It had been proven that Roman based writing which closely adheres to English character could identify the word features such as acronym that usually is represented by proper noun with capitalization well (Abu Bakar et al., 2013).

The important issue that is attempted to be addressed by Malay NER seen so far for Information Extraction is the lack of resource in which to be referred as a baseline for future data annotations (Sidi, 2011; Hanum et al., 2014; Alsaffar et al., 2015) . Dependence on scaffolding tools such as gazetteer and dictionary listing is still vital for further categorisation of Malay words into its respective NE classes. Learning methods have seen a transition to be less depending on old Malay annotated data, as high quality resources tend to be non-existent or hard to obtain, or the creation itself is rather human dependant (Abu Bakar et al., 2013). Therefore, the annotation of proper noun could assist in overcoming the current problems endured with Malay corpus categorization.

2.7 Review on Learning Techniques for Named Entity Recognition

Being closely associated with the fields of literary studies, NER research had been carried out within various contexts of language fields. A good proportion of research is dedicated towards the English language; however, it is undeniable that the research

purposes also serves to address language independence and multilingualism problems. Currently a few related works have been identified as the encouraging and guidance factor to study the issues addressing the needs to establish a proper Malay noun text collection for further application in information systems. Some of the studies identified adopted a context-based approach that bears significance to the research that has been identified.

As the method of semi-supervised learning was synthesised as the main component of the research, a context-based approach has been identified as it involves the derivation of information learned from the context of the source word and the target word. The approach of currently existing NER systems for other linguistic values could not be implemented into the incorporation of NER system for a Malay corpus, as currently there is still lack of resources available for Malay language to be taken into account for further analysis (Abu Bakar et al., 2013).

For all NER systems, the ability to recognise unknown and new entities is important to signify its efficiency for upholding data scheming procedures. These features also rely on recognition and classification rules associated with positive or negative feedbacks. Most recent NER systems have utilized supervised or unsupervised machine learning to implement rule-based systems or sequence labelling algorithms starting from a collection of training catalysts (Rahem et al., 2015). The methodology of supervised and unsupervised learning includes identifying and researching the features of positive and negative examples of named entities over a large collection of scraped text, annotated documents and design rules that fulfils the intent of a given type.

Although these two learning procedures could generally categorise the instances of the given type, the main disadvantage of supervised learning is the requirement of a credible, large collected corpus (Althobaiti et al., 2014). Due to the insufficient availability of these resources and the cost needed to create them, alternative learning methods have been improvised indirectly, including semi-supervised learning and unsupervised learning (Althobaiti et al., 2013; Carlson et al., 2010). Natural language has a remarkable feature of describing spatial relation information without using numbers and they also offer a variety of lexical terms for describing spatial location of entities. An information extraction system generally converts unstructured text into a form that can be loaded into a database table. Useful information such as the names of people, places, or organization mentioned in the text may then extracted without a deep understanding of the text.

Table 2.2 highlights among the most used clustering algorithms with implementation of classification algorithms that had been identified from prior researches. Within the context of this research, unsupervised learning would be placed in more focus for further implementation.

Table 2.2

Examples of the Most Utilized Clustering Algorithm

Methods	**Vectors**			**Optimization Procedure**
	B	***A***	***X***	
One-side K-Means	$B = I$	$a_{ik} \in \{0, 1\}$, $\sum^{K}_{i=k} a_{ik} = 1$	$X = (A^T A)^{-1} A^T W$	*Alternating least square*
One-side Low Dimensional Clustering	$B = I$	$a_{ik} \in \{0, 1\}$, $\sum^{K}_{i=k} a_{ik} = 1$	Rank $(X) = t$, $t \leq \min(K - 1, m)$	*Low-rank approximation*
Spectral Relaxation	$B = I$	Orthonormal	$X = (A^T A)^{-1} A^T W$	*Trace maximization*
Concept / Non-Negative	$B = I$ Matrix factorisation	Non-negative	$X = RW$ (Linear combination of points in the cluster)	*Constrained optimization*
Double K-Means	$b_{jc} \in \{0, 1\}$ $\sum^{C}_{j=c} b_{jc} = 1$	$a_{ik} \in \{0, 1\}$, $\sum^{K}_{i=k} a_{ik} = 1$	$X_{kc} = [\sum^{n}_{i=1} \sum^{m}_{j=1} a_{ik} b_{jc} w_{ij}] / \sum_{i \in n} \sum_{j \in m} a_{ik} b_{jc}$	*Alternating least square (two-side)*
Iterative Feature Data Clustering	Arbitrary	Arbitrary	$X = I$	*Mutually reinforcing optimization*
Generalized Spectral Relaxation	Orthonormal	Orthonormal	$X = A^T W B$	*Two-side trace maximization*
Subspace Clustering	$B \in R^{m \times K}$	$a_{ik} \in \{0, 1\}$, $\sum^{K}_{i=k} a_{ik} = 1$	$X = (A^T A)^{-1} A^T W$	*Explicit subspace identification*

2.7.1 Supervised Learning

A supervised learning technique is output produced from an approach to initiate rule-based systems or sequence labelling algorithms which is gathered from a collection of training examples. The approach is portrayed based on the investigation of positive and negative feature examples for named entities taken from a huge collection of annotated documents and design rules which captures the instances of a given type (Kanagavalli et al., 2013). Due to the incorporation of human and computer interaction in the aspects of dataset collection and analysis, the supervised learning approach includes the necessity of a large annotated corpus (Senthil et al., 2014).

The absence or lack of availability for said resources, along with the restrictions of cost in creating them had innovated the establishment to semi-supervised and unsupervised learning approaches (Senthil et al., 2014). An algorithm to create a large list of a given type of entity or semantic class had managed to be applied for NER (Ritter et al., 2011) . This technique is enhanced by equalizing a few given seed entities into the query to retrieve web pages that contain similar seed names. The performance of the system's foundation relies on the capabilities of vocabulary transfer that consisted of the proportion of words that neglects the repetition of appearance in both corpuses which are set up for the purpose of testing and training (Mohammed et al., 2012).

Vocabulary transfer determines the frequency of recall rate for the overall baseline system. Machine learning approach were utilized mostly for supervised systems which in turn reduced the need for human effort in terms of data clustering

and categorisation (Sinoara et al., 2014). For supervised systems, machine learning approaches are mostly used, thus reducing the participation of human in the matter of data clustering and categorization (Sinoara et al., 2014). The data as submitted to the machine learning algorithm is fully labelled either been manually tagged or distinctively annotated by a lexicon. Classifiers learn from data, as it generally involves the process of assigning labels to yet unseen instances (Mencar et al., 2011).

The approaches available to automatically classify words that could be used include rule-based methods, probability-based, and transformation-based. Within the fields of NER, the most common supervised learning techniques that have been used includes Support Vector Machines, Maximum Entropy, and Conditional Random Fields.

2.7.1.1 K-Means Algorithm

As a component of fuzzy relational calculus, the K-Means algorithm is among the most commonly implemented NER algorithm in the context of data clustering purposes. The basis behind the usage of K-Means lies within its proportional objective to clustering algorithms. The K-Means algorithm is a basic iterative clustering algorithm that partitions the given dataset into a user-specified number of clusters, known as *k*.

The usage of K-Means algorithm applies to objects that were represented by points in a d-dimensional vector space. This algorithm is a clustering type algorithm that partitions the set of *d*-dimensional vectors, *D* into *k* cluster of points. This means that K-Means algorithm clusters all of the data points in D to the extent that each point x_i is categorised into 1 and only 1 of the *k* partitions. In order to keep track of the points created, each point is assigned a cluster ID. Therefore, points with the similar ID remains in the same cluster, whilst points with varied ID values were isolated into different clusters. This is denoted via the relationship between cluster vectors **m** of length *N*, where the value of m_i is the cluster ID of x_i.

The value of *k* is an input to the base algorithm, that is *k* exist based on criteria such as prior knowledge of how many clusters actually appear in *D*. This is also determined by the number of clusters that is needed for the current application or in the terms of cluster types that is discovered by the exploration with different values of *k*. In the *k* region, each of the k clusters is represented via a single point in k^d. We could address this cluster set as the set $C = \{c_j \mid j = 1\ldots, k\}$. These k cluster representatives are also identified as the cluster means, or cluster centroid.

As elaborated in the past discussions, points in a clustering algorithm were grouped via a notion of distance, or "similarity" with each other. The default measure of closeness in K-Means algorithm is known as the Euclidean distance. In other words, K-Means algorithm is an attempt of minimize the total Euclidean distance among point x_i along with its neighbouring representative, c_j. The algorithm's attempt to minimize any non-negative cost functions could be iterated into the following formula (Wu et al., 2008):

$$\sum_{i=1}^{N}\left(argmin_j \parallel x_i - c_j \parallel_2^2 \right)$$

The algorithm above is initialized by the random selection of data from *k* points. These steps would commence repeatedly until the extent where convergence is achieved. The algorithm achieves convergence should there be no more alterations in its assignments (along with the value for c_j). A way to evaluate whether this step had been achieved is finding a decrease in property whenever there is any change in the assignment or the relocation steps. Together with this, convergence is guaranteed within a finite number of iterations.

Table 2.3

K-Means Algorithm (Wu et al., 2008)

Algorithm 2.1 The K-Means Algorithm

Step 1 Input **:** Dataset *D*, cluster number *k*
Step 2 Output **:** Set of cluster representatives *C*, cluster membership vector **m**
Step 3 /*Initialize cluster representatives *C**/
Step 4 Randomly choose *k* data points from *D*
Step 5 Use these *k* points as the initial set of cluster representatives *C*
repeat
/*Data Assignment*/
Reassign points in *D* to the closest cluster mean
Update **m** such that m_i is the cluster ID of i^{th} point in *D*
/*Relocation of means*/
Update *C* such that c_j is mean of points in j^{th} cluster
until convergence of objective function

Selecting the most suitable value for *K* from the dataset lot might prove to be a difficult task to carry out. The selection depends on a full comprehension towards the dataset. This knowledge includes the number of partition that naturally comprises the dataset itself, or should the partition number is unknown a *model selection* method is used instead. The most primitive way of trial-and-error evaluation to determine the *k* value is via the testing of several different *k* values and choosing the clustering which minimizes the K-Means objective function. The K-Means algorithm produces minimum variance clustering. However, it doesn't guarantee that the global maximum of the criterion function would always be discovered. The results relies heavily on the initial choice of cluster centroids (Zhang et al., 2012).

The clustering obtained may experience variance from run to run due to the fact that the choice is usually made at random. A solution to deal with the problem is to run the algorithm several times with different random number seeds and then select the clustering that maximized the criterion function. Assuming that the number of cluster that the algorithm should produce is known in advance, a divisible partitioning strategy could be planned as implementation due to the fact that the only decision that needs to be made is based on how to split clusters. This also made it more proficient than agglomerative clustering, where all possible candidates for merging would be evaluated (Zhang et al., 2012).

$$\sum_{i=1}^{k} \sum_{d_{1 \in D_i}} sim\,(\,c_i\,,d_l\,)$$

Where:

c = similarity between cluster centroid

d = distance between data points / cluster centre

i = number of clusters (*predetermined*), and

k = number of documents.

K-Means Algorithm was among the partitioning methods that proceed in its data clustering procedure with a few given assumptions (Wu et al., 2008). As collected when *k* individuals were chosen as the initial centres for the cluster, these packages were selected based simply at random, or taking the first *k*, or taking 1 out of *n/k*. The distances between each individual and centre C_i of the preceding step were calculated, where each individual entity would be assigned to the nearest centre. This thus defines the existence for the *k* clusters (Al-shammaa et al., 2015). The *k* centres C_i were replaced with the centres of gravity of the *k* clusters defined in the preceding step. These centres of gravity are not compulsory to be individuals within the population.

After this step is conducted, a check is made to indicate should the centres have remained sufficiently stable via the comparison of their movement to the distances between the initial centres. This is also done if a fixed number of iterations has been concluded. The process comes to a halt should the answer had fulfilled the prerequisite. This step usually continues at least within 10 iterations. If the answer

does not fulfil the premises, the step of calculating the distances between each individual and each centre C_i will be repeated.

Figure 2.6 illustrates the basic concept for agglomerative clustering.

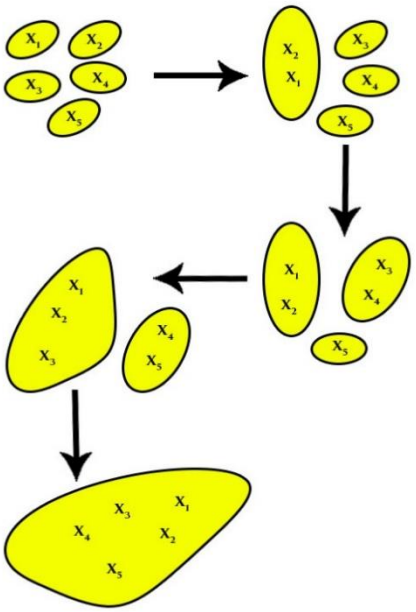


Figure 2.6. Agglomerative Clustering During The Process Of Data Classification

2.7.1.2 Support Vector Machine (SVM)

As almost all clustering partitioning algorithms implement the theories of Support Vector Machines, it is important to be discussed in the context of the amendment of the fuzzy NER algorithm. Support Vector Machines (SVM), including Support Vector Classifier (SVC) and Support Vector Regressor (SVR) were considered among the

most robust and appropriate technique in the field of data mining algorithms. SVM possesses a characteristic of sound theoretical foundation directed towards statistical call learning theory, thus only required a small number of examples for training. SVM also has the trait of insensitivity to the number of dimensions (Wu et al., 2008; Maraziotis, 2012).

As a subcomponent of Support Vector Machines, Support Vector Classifier is targeted to locate a hyperplane that could isolate two classes of provided samples with a maximal margin that had been dictated able to provide the best generalization characteristics. This generalization ability means that the classifier should not only have good classification performance on the training data, however, also guarantees high predictive accuracy of the future data from the similar distribution as the training data itself.

Margin existed from gaps among these hyperplane, which is defined as the amount of space or separation in between the 2 classes as defined by a hyperplane. Predefined in a geometrical factor, the margin corresponds to the shortest distance between the closest data points to any point on the hyperplane (Wu et al., 2008). The equation for optimal hyperplane is derived in the next formula.

$$w^{T} x + b = 0$$

Where:

w = weight vector, and

b = bias.

Figure 2.8 illustrates the integration of the position of elements in the most optimal location for a hyperplane in a Support Vector Classifier. Concurrently, the Support Vector Classifier has the objective of finding the parameters *w* and *b* of an optimal hyperplane, so that the margin of separation that is determined by the shortest geometrical distances *r** is maximized from the two respective classes. This leads the SVC to also be known as the *maximal margin classifier* (Wu et al., 2008).

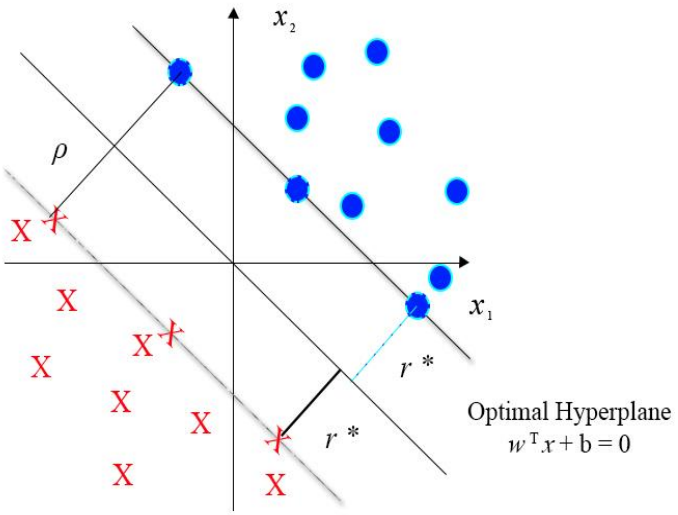


Figure 2.7. Illustration of the Optimal Hyperplane in SVC for A Linearly Separable Case (Wu et al., 2008)

2.7.1.3 K-Nearest Neighbour (KNN) Algorithm

The *k*-Nearest Neighbour Algorithm is closely related to the Rote classifier, a machine learning technique that is based on distance calculation from the training data set. Rote classifier is considered the simplest but rather trivial classifier for text mining datasets, as it memorizes the entire training data and perform classification

only if the prerequisite of the attributes for the test object matches the attributes of one of the training objects. However, one complication arising from the algorithm's implementation is the ignorance of classification for many test records that clashes with the training record, and should two or more training records possess the similar attributes however different class labels.

The K-Nearest Neighbour algorithm is introduced to find a group of *k* objects in the training set that were considered in the vicinity to the test object itself, and bases the assignment of a label on the predominance of a particular class within the confinements of the domain (Kantardzic, 2011). The introduction of *k*-Nearest Neighbour algorithm resolves a few issues that plague the basic NER algorithm. In its most fundamental form, the KNN may involve the assignation of an object the class of its nearest neighbour, or of the majority of its nearest neighbours (Kantardzic, 2011). KNN is considered easy to modify for more complicated classification procedures. Its application is highly compatible for multimodal classes, as well as for applications in which an object could possess many class labels.

Table 2.4

KNN (K-Nearest Neighbour) Algorithm

Algorithm 2.2 Basic *kNN* Algorithm

Input : *D*, the set of training objects, the test object, **z**, which is a vector of attribute values, and *L*, the set of classes used to label the objects

Output : $c_z \in L$, the class of *z*
foreach object **y** $\in$ *D* **do**

| Compute the value of *d* **(z, y),** the distance between **z** and **y**;
End

Select $N \subseteq D$, the set (neighbourhood) of k closest training objects for z;

$$C_z = \underset{v \in L}{\operatorname{argmax}} \sum_{y \in N} I(v = class(c_y))$$

Where I (.) is an indicator function that returns the value 1 if its argument is true and otherwise, 0.

2.7.2 Unsupervised Learning

Here, unsupervised learning is defined as an approach in which machines induces data training from unlabelled inputs, generally related to problems without a target output for further evaluation. As contrast with supervised approach, unsupervised learning were usually applied to data categories that possesses a more ambiguous and vague context; in this terms dataset that have not undergone any crediting or still raw without processed.

Unsupervised approach focuses on the objective of performing data learning without requiring any influence from external factors. Therefore, the algorithms that would be involved calculates the accuracy of propagation for datum groups into their respective classes. The algorithms popular with unsupervised methods include Expectation-Maximization, K-Means, Neural Network, and the more wide-scale Cluster Analysis.

Unsupervised learning is used when there is the problem of lack of labelled data (Etzioni et al., 2005). The objective that is targeted from unsupervised effort is the construction of data representations that could be further used in processing attributes, such as in compressing data, classifications, and decisions. Unsupervised learning is classified as a least popular approach in terms of NER research as its practical usage in systems were usually relevant with the application of other supervised efforts.

2.7.3 Semi-Supervised Learning

Semi-supervised learning is a technique that had been widely explored recently from the efforts to improve the deficiencies of fully supervised techniques. As the usage of semi-supervised techniques embeds already existing data chunks to improve the efficiency besides accuracy for the obtained clustering results, it is considered a good alternative to a machine language method (Alfred et al., 2014). The performance of a semi-supervised clustering approach could normally be improved with a very small amount of prior knowledge as compared with supervised methods. Few label objects are used to improve the performance and does not require the entire category information to be known to grasp the foundation of the classification model.

Another advantage that semi-supervised learning possessed over supervised methods is that it is less time consuming along with the benefits of not requiring additional training process. Yang stated that the existing methods for semi-supervised clustering generally fall into two categories, which are similarity adapting-based and search-based methods (Yan, Chen, & Tjhi, 2013). In other words, this procedure

requires a very small amount of the label objects to be improvised as a catalyst to improve the performance, and does not require the entire category of information to be known to learn the classification model (Yan et al., 2013).

The clustering algorithms were added to further enhance the grouping of named entities according to their distinctive categories, via the categorisation of pattern base for the major categories in NER. However, there are several defunct factors that affect the progressive identification for raw datasets, even with the traits of semi-supervised learning approach that involves the participation of both human and computer intelligence in order to maximize the efficiency of the process itself.

Research from closely related practices had proven the issues that arose within the implementation of semi-supervised learning, among these includes the resource deficiencies, word emergence frequency that is least comparable between corpuses, and corpora that undergo least balanced development in terms of linguistic structures (Ando et al., 2005; Carlson et al., 2010). In order to resolve the lack of resources arising from a lack of corpus availability, unsupervised or semi-supervised approaches are applied to classify the existing data (Baharudin et al., 2010).Simultaneously when unsupervised or semi-supervised techniques were chosen to train existing or incoming data, appropriate techniques that could assist in labelling raw and unsorted documents based on the most minimal resources available (Althobaiti et al., 2013; 2014).

The identified deficiencies of a semi-supervised learning method in NER systems are highlighted in the following table.

Table 2.5

The Difficulties Emerged From Prior Research on Semi-Supervised Learning for Language with Different Morphology and Syntaxes, Including the Platforms Involved (ECST Model=Baseline Word Spelling System, IDT=In-Domain Term)

Author	Year	Main Issue		Approach	
Nadeau	2007	Lack of Resources		Minimally supervised	Locating techniques that assist in annotating raw, unsorted documents from minimal resources
Yunita	2009				
Shaalan	2009				
Don	2010				
Saniati	2014	Less comparable word occurrence frequencies between corpora	Missing target words	ECST	-
Gunawan	2015		Domain stop words		
Martinez	2015				
Nadeau	2007	Less balance content of a particular corpora in terms of word context	Hypernymy, Polysemy, Hyponymy, Synonymy	IDT, more appropriate semi-supervised technique to be implemented across NER algorithm	-
Khan	2010				Suitable NER algorithm to improve classification of unknown entities
Gunawan	2015				

Although there might be possibilities for similar word to appear in a corpora collection within a single domain, there could also be no doubt word occurrence that appears to be less frequent within the domain itself, which cannot be used to compare the entire corpus' traits with other relative corpus. This problem would affect the entire corpus in terms of data evaluation. Some of the subcomponents that persist

from rare occurrence of word frequency within a single domain includes missing target words and lack of domain stop words.

The most common problem in NER research that would cause semi-supervised approach to be preferred is the content of a particular corpus that has a composition that is less balanced in terms of word availability and ambiguity. Issues that may arise from a less balanced corpus content includes polysemy, the coexistence of many possible meanings for a word or phrase; and the correlation between values of hypernyms and hyponym, relationship between the more general terms with the more specific instances of the term itself (Baharudin et al., 2010; Gunawan et al., 2015). The next Table 2.6 highlights all of the discussed prior researches done in major NER fields related with Malay morphology.

Table 2.6

Comparison between Researches Done In NER for Major Linguistic Fields

Language	Characteristics	Advantages
English	• Classification of related linguistic articles into a Named Entity (NE) type, labelling articles manually • Modifying corpora content by introducing extraneous links to properly synchronizing the automatic annotations to approved standardizations • Development of acronym detection algorithm • Classifying designators in text fields • Minimal supervision is required	• Transforms article contents into Named Entity (NE) by the projection of the classifications from the article's content into the anchor text • Properly synchronizing the automatic annotations done to corpora into approved standardizations • Classifying designators in text fields and temporal expressions

(Continued in next page)

Arabic	• Focused on the machine language approach relying on the neural network technique • Adapt to new data clusters, synchronization with the previously accumulated data to introduce new information clusters • Advantageous traits that ease its implementation towards a lot of system applications • High precision and accuracy achievement rate	• Operated based on the availability of a neural network • Learning progress based on the availability and the existence of the previous data • Learning adaptability • Definition derivation from complex or imprecise data • Self-organization • Produced a diverse range of algorithms from inherited from different structural designs
Malay	• Detection of Named Entity existing in data clusters via rule sets and dictionary listing that were defined manually by human • Extraction pattern could	• Extraction pattern possesses high compatibility and adaptability during searches • Unique, distinct extraction pattern base for location names, organizations, and people to improve search and categorization rate • Part-of-Speech (POS) tagging feature that tags independent characters and contextual traits to handle properly local Malay articles • Training processes could further classify entities to improve searching algorithms and to determine the precise character

	be defined for location names, organizations, and people (main categories in NER) based on grammatical, syntactic, and orthographic features • Freedom training for learning process, whether for the semi-supervised or supervised data clusters	• Linking speech patterns and grammar parser to improve the precision achieved for the data search

2.7.4 Data Cleaning

The procedure of data cleaning is included within the pre-processing phase, after the accumulation of raw text data from targeted sources. As the text content consisted of various different topic scopes, there is bound to be foreign character that is present within the data accumulated. These foreign characters include wording and additional characters that were not intended to be incorporated into the text content. Commonly occurring strings that do not help in the processing of natural language data should be removed (Abu Bakar et al., 2013; Mikolov et al., 2013). The removal of function words from the texts is also recommended so as to increase the values of many nouns (Indurkhya et al., 2010). Data cleaning identifies and isolates any English wording and character irrelevant to the process.

2.7.4.1 Manual Tagging

The number of currently existing NLP system incorporates the application of part-of-speech taggers with the sole purpose of categorical syntactic disambiguation. This step is acknowledged as the pre-processing procedure, which is important to assimilate the various forms of lexicon-syntactic ambiguities (Indurkhya et al., 2010). This is shown in the usage of daily words, such as common nouns or verbs (for example, “report” and “support”) depends on its syntactic context. Most taggers operate as representatives of supervised and data-driven approaches that critically relies on training data via syntactically annotated corpora (Ananiadou et al., 2010).

2.8 Evaluation for NER

This section briefly provides an overview of the evaluation techniques applied in the fields of text mining, and the context of this research relating to Named Entity Recognition.

In order to evaluate the efficiency of NER system in its effort to categorise named entity appropriately according to its own grouping, an evaluation measure is utilized. The current methods for NER system evaluation are mostly influenced by the notions of evaluation devised for IR areas, among these includes the accuracy, precision, and recall method (Alfred et al., 2014; Baharudin et al., 2010; Powers et al., 2011). Derived from both precision and recall methods, the average precision-recall score could also be determined using a measure known as the F_1 score. The evaluation methods are described in the following section.

2.8.1 Accuracy

The method of calculating accuracy score within an NER system involves 2 variables, which is representing the percentage of items right for the system itself: to select or vice versa.

$$\textbf{Accuracy} = \frac{\text{tp+tn}}{tp+fp+tn+fn} \quad (2.1)$$

Where:

t*p** = number of cases the system succeeds by selecting the targeted item. This is also known as the ***true positive.

t*n** = number of cases the system succeeds by not taking the wrong item. This is also known as the ***true negative.

fp* =** number of cases the system succeeds to select the targeted item, however the item is wrongly labelled. This is also known as the ***false positive.

fn* =** number of cases the system succeeds to avoid selecting item, however the item is the actual target. This is also known as the ***false negative.

This method is not encouraged due to its emphasis on less important cases that's represented by the huge *t*n value.

2.8.2 Precision and Recall

The precision value in Information Retrieval field to be *"a measure of selected items that the system got right"*, whilst the recall value is derived as *"the proportion of the target items that the system selected"* (Manning et al., 2009).

The precision value, *p* could be determined as follows:

$$\text{Precision}, p = \frac{tp}{tp+fp} \quad (2.2)$$

Where *t*p and *f*p is the values of **true positive** and **false positive** respectively.

The recall value, *r* can be measured as shown.

$$\text{Recall, } r = \frac{tp}{tp+fn} \quad (2.3)$$

From the results obtained, a *contingency matrix* representing the concepts used during the evaluation of the precision and the recall rate may be devised.

Table 2.7

Sample of Confusion Matrix Table (Manning Et Al., 2009)

System	**Actual**	
	Target	**⌐ Target**
Selected	*t*p	*f*p
⌐Selected	*f*n	*t*n

Where *f*n represents the number of cases the system failed to take the target item into account, which is also identified as a *false negative*.

2.8.3 F1 (F-Measure)

The $\boldsymbol{F_1}$ score, or F-measure, is the evaluation method that combines precision and recall values for overall performance. This measure is applied to resolve the dominant trade off issue in between precision and recall values. This is due to the fact of the

values of both precision and recall rate were not concluded as a fixed variable, however might be possible to change according to the output. For example, all items in the dataset could be selected so 100% recall rate would be obtained, however, causing a drop in precision value. The measure is simplified as follows.

$$\text{F-Measure, } F_1 = \frac{2pr}{(p+r)}$$

The scores obtained via $\boldsymbol{F_1}$ possess a similarity to accuracy measures, as they are both prone to certain cases encountered by the system. In the Information Retrieval field of research, the $\boldsymbol{F_1}$ measure is sensitive to the numbers of correct cases whilst the accuracy is only prone to sensitivity induced by the number of errors. The rank of the candidates is an essential trait (Gaussier et al., 2004). Sorting and ranking of the candidate group according to their respective scores are both vital if any of the evaluation measures is to be performed.

2.9 Summary

This chapter begins the discussion on research conducted in named entity recognition, along with the fundamental points of rule-based approach that had been applied in prior researches. Issues related to the research for the respective linguistic field are also been highlighted in accordance with the research in Malay NER to be conducted. As semi-supervised approach is the key method carried out in the process, this chapter relates the similar approach that other researchers have implemented, along with features from their research that could be applied in the context of this research alone.

This chapter also discussed the basic concepts pertaining to the focus of the upcoming study. As a review effort on the key research which possessed similar elements in terms of algorithms and techniques, survey had been done on previous works to illustrate the efforts that could be applied together when the research is carried out.

In English NER, the emphasis of entity classification has been performed by manually labelled articles. The singular corpus domain of English had included external links for its word elements to standardize the internal contents, thus producing automatic annotations for a cluster of word collections. Although English is among the widely utilized language in the community, most NER research still requires a considerable level of supervision in order to function entirely. On the other hand, Arabic language that is considered the most applied language (300 million people in the world) in terms of religious context and dialect spoken which has also been proven to be dependent on proper clustering and processing to detect the presence of named entity in its sentences due to its different sentence alignment and organization.

For the upcoming chapter, the research presents the initial methodologies to be used to carry out the data analysis and exploration, along with a more comprehensive on a few classification algorithms' features that would be implemented in the upcoming data collection. Explanations are also given to elaborate on the benefits of the selection for the algorithms highlighted during the scope of this research.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The previous chapter reviewed several existing studies related to the field of Named Entity Recognition for entity classifications. Most available NER processes implements supervised learning techniques due to its nature of less human participation. Several critical components had been highlighted to be included in the implementation of the incoming suggested learning techniques in terms of extracting proper noun for Malay news articles into their respective domains.

This chapter serves to demonstrate the workings for extracting proper nouns towards contributing to named entity recognition, particularly via the effort of categorisation of unclassified, raw Malay corpus accumulated from targeted online newspaper. Steps implemented in the process involves clustering of unannotated text from news articles along with classifications into their respective groupings, especially for language categories. The research design was devised from several identified components that shape clustering and classification procedures, including linguistic resources, classification algorithms, association measures, analysis formulas, and learning techniques.

In the context of each component, related methods have been discussed in depth. Evaluation technique to measure the performance of the developed system is discussed in detail with accordance with the research scope. Several experiments were devised to test the conceptual research design during the course of study. The framework of the methods involved is also reviewed.

3.2 Research Design

Figure 3.1 illustrates the research design which had been imposed during the initiation of proper noun detection from the Malay news articles.

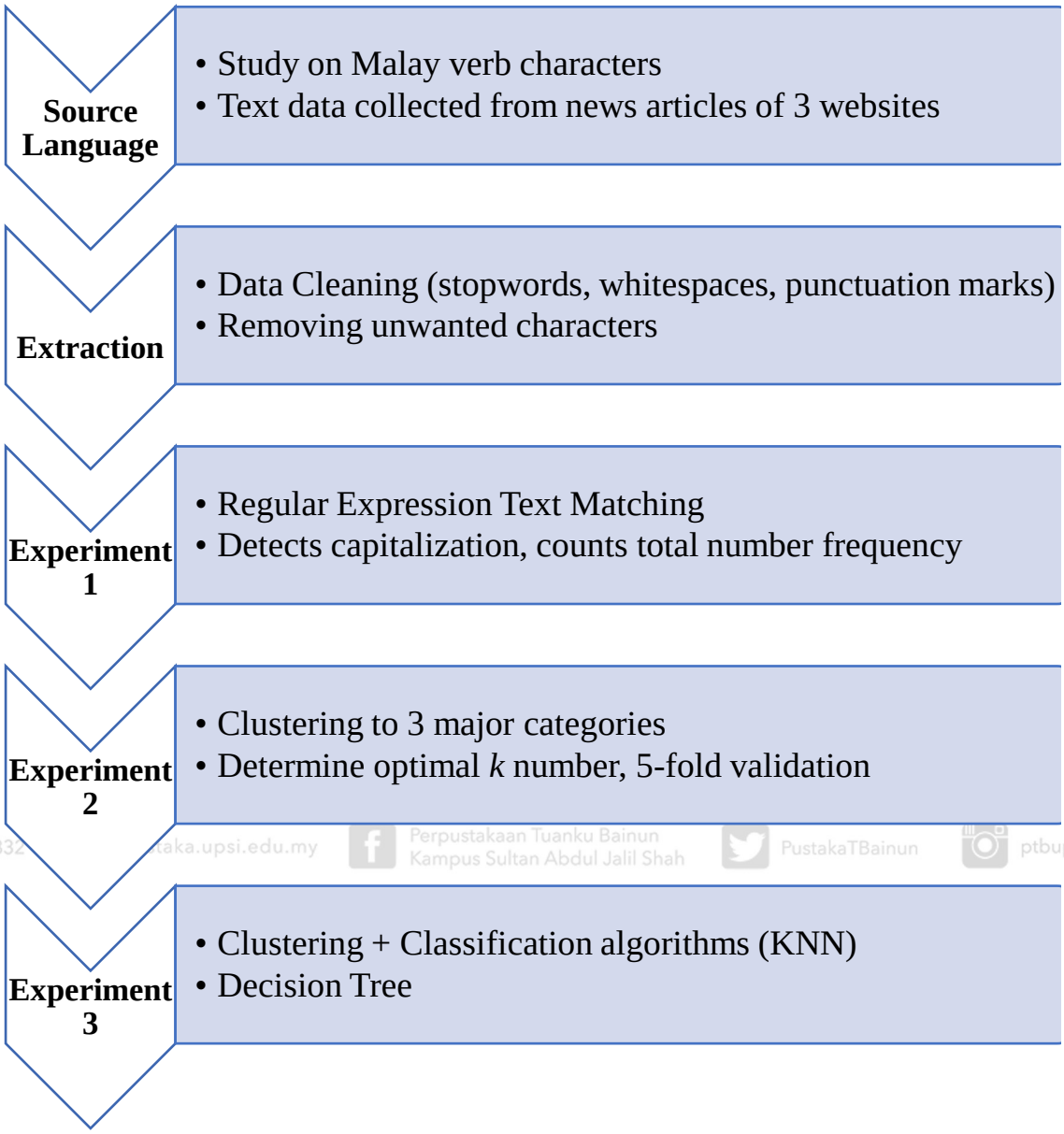


Figure 3.1. Research Design

This study incorporates 3 experiments to testify the theories from the referred past works, where each of them serves to fulfil the research objective and functions as a key performance indicator of NER. The experiments carried out comprises the following:

Experiment 1: Purposed specifically to annotate special regex pattern identification mechanism to locate the presence of proper nouns from the collected web articles. During this process, disparate patterns had been sculpted to differentiate the characteristic of capitalized word, punctuation marks, availability of spaces between paragraphs, and so on. The output generated from this practice comprises of a listing of Malay proper nouns that is perceived to be accurate and have not gone under expert verification. This process generates a list of proper nouns. A side experiment also occurs here, to illustrate the change in performance between the data that had been going through processing, and those that have not. The annotation process with linguistic expert is going to be performed simultaneously after the Experiment 1 concludes.

Experiment 2: After the annotation takes place, the text data needs to be further analysed via clustering and classification algorithms. To this aspect, the text data needs to be converted to numeric form in order to be computationally friendly. Thus, text to number vectorisation takes place. This process then produces a number list translated as a term document matrix. The purpose of this matrix is to produce a numerical representative of each of the word's similarity within a single domain in which they reside. Afterwards, clustering is performed to indicate the classes available and to simulate the most optimal number of clusters they can reproduce. The magnitude in which the performance could be replicated is repeated with 5-fold validation.

Experiment 3: In this stage, similar clustering and classification algorithm is applied again. The only difference is in this stage, fold validation is omitted. The reason for

this practice is to see whether the graph visualized could replicate the same effect or vice versa. Decision Tree algorithm is also embedded during the algorithm commences. From a total of Decision Trees generated, the most optimal one that includes the most nodes taken from the data cluster is chosen to illustrate the co-reference between the text similarities.

3.3 Research Framework

Selection of the targeted Malay resource news articles would take into consideration the elements constituting the language structure, such as its morphological, syntax, or lexical appearance. The research began by performing a survey on existing approaches on linguistic analysis and noun extraction done in other languages, basically linguistic categories where the study is highly accommodated on, such as English, Chinese, Arabic, and Indonesian. After that, further study was performed to identify the main categories for learning approaches, mainly supervised, unsupervised, and semi-supervised. After the steps were done, various classification algorithms were identified and categorised based on their adaptability to properly categorise the nouns in Malay unannotated articles.

3.3.1 Data Collection

The procedure of data collection involves a few processes suggested to retrieve text data from their respective articles, such as web crawling, manual feature extraction

and simple data cleaning. The accumulation of corpuses was done on several identified and credible web resources, such as local online news portals. These collected text data were contained in their unmodified primary state according to certain structures or formatting.

The product of data collection emphasizes on the elimination of certain document formatting, such as document headers, footers, and diagrams (Miner et al., 2012). The traits that need to be eliminated are determined before the procedure is carried out. Before the processes were performed, the basic structure of web pages that contain these articles were identified and scanned. The end product is obtained by isolating these unwanted structures to form a document that contains characters that is desired for the post-processing, such as article title, body content, and date of publication (Banko et al., 2007).

Huge organizations specialized in offering information to the public develops and provides text corpora for further access. Usually this organizations comprised of government sectors that contains a lot of archival materials, however recently private sectors had seen an improvement in resource availability especially news stations and radio broadcasting. In line with the research objective, major open access news portals were selected which serve their content in the Malay language.

i. Astro Awani (*http://www.astroawani.com/)*

An in-house television channel, a subsidiary of the parent broadcasting satellite channel ASTRO. This TV channel provides a website portal that includes current

affairs of news coverage, including news content in Malay. The content from the site is derived directly together with its on-air news content. Astro Awani presents news content ranging from Malaysia's current affairs, lifestyle, documentaries, along with a widespread magazine. The portal offers bilingual information, in English and Malay. The portal is categorized as a bilingual news portal.

ii. Bernama (*http://www.bernama.com/)*

An abbreviation from its Malay dub Pertubuhan Berita Nasional Malaysia, BERNAMA is the news agency of the current government, along with the owning status of an autonomous body placed under the supervision of Ministry of Communication and Multimedia. BERNAMA releases its news content under the pretext of pro-government of the day, and offers news content in multiple languages, including English, Malay, and Mandarin. The portal is categorized as a multilingual news portal.

iii. Berita Harian (*http://www.bharian.com.my/)*

A web portal established to simultaneously relay news content to its audience, together with its daily publications of newspapers. The web portal consists of news content composed entirely based on Malay language and targeted for a more major crowd of Malay audience, particularly in the Peninsular Malaysia region. As seen by other major news corpus that focuses on the deliverance of current affair materials in multiple contexts of audiences, Berita Harian delivers its content in a manner that is precisely focused on the current undergoing nomenclature in the nation's

development, which it had encompassed since its earlier newspaper department establishment during 1957. It is identified that this news corpora contains rich data content in Malay language that could be further explored. The portal could be categorized into monolingual news portal.

Three main collections of Malay articles were collected from the three news portals, where each of these represents 20 news articles for each respective news portals. The articles were collected in a predefined timestamp of 1 month. This process is detailed in Chapter 4. The articles are collected using web scraping technique in both in-house developed Python and PHP scripting, however several techniques had been embedded aside as well such as Python's BeautifulSoup4 library. Each article is named after its respective domains, for example Astro Awani as AA01, Berita Harian as BH01, and Bernama as BER01.

3.3.2 Pre-Processing

Pre-processing, also known as the processing techniques applied after initial data extraction, involves techniques to clear out unwanted residues that remains after the text data is extracted from its original source. The procedure is usually done according to the feature that is desired in the research scope itself. Most pre-processing measures is done to allow the text data to exist in a human-readable form, besides tidying the text data appearance. Type of text available would define the approaches to be performed before performing initial text processing (Abdallah et al., 2012).

The stage of text pre-processing is done to ensure that the text data collected were in the most appropriate form for further analysis. Among the processes defined by previous works to extract text from outliers are removing whitespaces to improve memory space, and elimination of punctuation marks to remove limiters between each sentence.

Tokenization of text further identifies a given word structure and divides the input text into a number of individual tokens (Aboaoga et al., 2013). This task is done so that certain unwanted traits that had been collected during the data accumulation process could be eliminated before any deep cleaning and analysis could be carried out. Among the characters often considered to be placed into the consideration of text splitting includes words, number, symbols, and white spaces between characters (Aboaoga et al., 2013). This procedure utilizes the amount of white space between the words to determine the number of characters that need to be tokenized within that particular sentence structure.

The Malay language could be categorized as sharing the similar structure with most European languages as its word boundaries are indicated by whitespace insertions. Word structures could also belong to agglutinating classes, where words divide into its subsequent units (morphemes) with clear limits between them. Inflectional categories also exist, where the boundaries between morphemes are ambiguous and morpheme components could express more than one grammatical meaning (Althobaiti et al., 2014). Most tokenization problems were indicated based on the usage of punctuation marks, since the similar punctuation mark could be

interpreted into many different functions within a single sentence. For example, see the next excerpt:

> ***"…Majlis penuh bersejarah itu turut disaksikan Presiden Indonesia, Joko Widodo; Perdana Menteri, Datuk Seri Najib Razak dan Pengerusi Proton, Tun Dr Mahathir Mohamad di Kompleks Kecemerlangan Proton di Subang Jaya pada Jumaat. "***
>
> (Extracted from http://www.astroawani.com/berita-bisnes/proton-tandatangan-mou-bangunkan-kereta-tempatan-indonesia-53488)

Figure 3.2. Example of Multiple Proper Noun Presence in Malay Sentence That Is Separated With Punctuation Marks.

The sentence above uses semicolons that are common for Latinate and space-delimited languages, and several commas to indicate the continuation of sentence structures. Tokenizers are an indicator in the usage of punctuation marks aside from differentiating whether the punctuation marks appearing within a sentence structure is part of another token. In many cases, sentence characters such as commas, semicolons, and periods are assumed to be separate tokens. Tokenization process should be assisted by filtering of existing mark-up and headers before the entire sentence undergoes the actual tokenization process.

The most important sentence boundary punctuation marks that have been determined are the question mark, period, and exclamation points. The definition of the sentence is limited to the text sentence (Indurkhya et al., 2010). The sentence in Figure 3.3 serves as an example of performing sentence segmentation. The overall

structure contains several features that indicate halted information chunks, for example a period (.) at the end, a semicolon (;) to indicate the continuation of sentences, and a comma (,) before the end of sentence. The example would be isolated into 4 grammatical sentences within the opening and closing quote alone when semicolons are used to end the sentence (Figure 3.3).

> ***"…Sejak kebelakangan ini, terdapat ramai suara-suara liberal tampil ke hadapan; mereka yang berpura-pura patriotik sedangkan tidak mengetahui apa-apa tentang perjuangan, tidak tahu apa-apa tentang prinsip yang didokong pejuang-pejuang dahulu," katanya pada majlis ulangtahun ke-5 IDEAS serta peringatan Almarhum Tunku Abdul Rahman Putra Al-Haj ke-112.”***
>
> **(**Extracted from http://www.astroawani.com/berita-bisnes/proton-tandatangan-mou-bangunkan-kereta-tempatan-indonesia-53488)

Figure 3.3. Example of Malay Extended Sentence that Requires Segmentation.

The isolation (known as *decomposition*) is considered a good point in sentence analysis, as short sentences are less likely to produce errors on analysis. Sentence segmentation assists the entire recognition process by reducing the error rate and improvement on other aspects, for example recall and precision rate.

3.3.3 Clustering

After data pre-processing is performed on the retrieved text collection, the accumulated text data would undergo generic in-depth model training via the chosen criteria, in order to produce a baseline model that would be used for data evaluation. The gathered text data would be divided into two clusters: one for training statistical model, and the other one is for the objective of data modelling.

For the purpose of entity recognition aspects, the extraction of models from a particular corpus is important so as to determine its relevant matching patterns and identical traits that could appropriately perform classification on the distinctive subcomponents. The resulting model is expected to provide an insight towards the features of the text cluster itself. Most rule-based systems that incorporate classification algorithms locate the distance between distinctive data with its epicentre to maximize the closeness between associating clusters. The distance is determined by the adaptability of the algorithm to compute similarity between vectors. In this purpose, clustering and classification algorithms are considered a priority.

The more similar an object is among each vectors, the similarity score also directly fluctuates (Turian et al., 2010). The closer the distance between objects, the distance score also is lowered. As semi-supervised learning approach involves the combination of rule-based and machine learning-based methods, the annotation of data clusters improvises the newly founded data into existing ones to improve the efficiency of learning and classifying new data categories. The newly founded data entity group is allocated into few sub-phases for the purpose of data training to

investigate the effectiveness of assimilating new and old text entities collected from the targeted Malay corpus.

The unlabelled data retrieved from manual extraction is carefully isolated according to its traits and categories. KNN classification decides the categorisation of the test samples based on the class for one particular training sample, which could turn out labelled or atypical (Kantardzic et al., 2011). The training objective of a KNN classifier is based on determining the values of *k* (Kantardzic et al., 2011), as seen in the objective of K-Means algorithm as well (Wu et al., 2008). The entire collected Malay text corpus is accumulated into 1 large class, where the class itself would undergo further processing to eliminate unwanted traits before evaluation measures is applied. This is to reduce risk of errors (Kantardzic et al., 2011). The product accumulated from this process is labelled the baseline training set.

Figure 3.4 demonstrates the process of classification between unlabelled and labelled text corpus. The purpose of grouping for the unstructured cluster is to identify the correlation between elements in a single random cluster. Classification commences with the accumulation of all random data from targeted dataset, where these data still remain unclassified. The word similarity calculated from every existing cluster would be identified as the most appropriate elements that constitutes a single cluster. This is when labelling comes into hand, where similarly distanced objects would be categorised into 1 cluster, before further processing is done to remove noisy outliers.

The process of applying noisy outlier processing depends on the intricacy of the algorithm itself, along with the evaluation to determine the distance calculation for elements that reside within a certain class. IE processes employs removal of outliers via vectorization and patching of absent variables. Since with the case of unstructured data where there might be presence of true and false positive variables, patching is executed on data space that is undefined. The selection of most confident labelled examples stems from data that agglutinates from a certain epicentre, where the data point that converges most closely to the centre is considered valid.

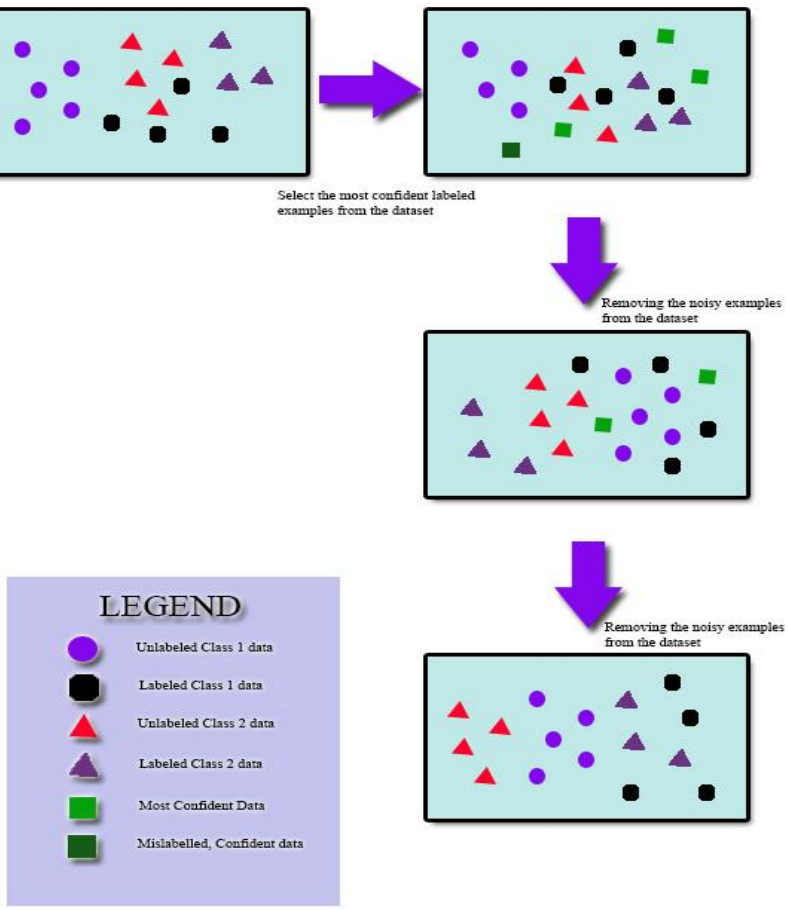


Figure 3.4. The Process of Labelling Unannotated Entities with Editing.

Based on the objective for classifying numerous data categories, the test set that are attempted to be constructed could be devised into 2 sections: development test and test set.

i. Development Test

After the training set has been determined, the data goes through the development test. During the development test, features from the extracted Malay text data are identified and classified manually according to group. The development test is also considered as the step to check the accumulated text data's integrity, as similar to during the process of manual tagging is performed. Besides checking on the text data integrity, the development test also calculates the initial precision and recall score. This provides a baseline reference in the future to compare the efficiency of the entire learning process with the selected evaluation method.

ii. Test Set

During the completion of development test phase, the data gathered is known as the test set. This test set would be used during the actual evaluation measure. In this process, linguistic rule would be applied according to the schematics that had been determined beforehand. For this process, the research processes the detected proper noun into main categories of NER, mainly Person, Location, Organization, and Miscellaneous.

3.3.4 Classification

Related to the previous phase, after a baseline model is formed to evaluate data, the collected text data is further classified from its unannotated state into annotated form so that the data cluster could be used for training. In the process, classification algorithms are applied together for the purpose of evaluation.

The situation varies with the unannotated labels, where the unannotated input would have its feature extracted according to its main classes before it's eventually be applied with the learning algorithm. The focus of the task is to obtain a baseline classifier model in order to justify the annotated labels from the text collection. Similar with the processes for extraction of appropriate labels, the initial input (in the form of unannotated text entries) would had its distinctive features extracted before being moulded into classifier model for training. From the finalized output, the labels of the respective classes would be obtained.

The next algorithm structure further elaborates on the algorithm structure for K-Nearest Neighbour that is applied during the process. The algorithm contains a few basic components, such as the training and test objects, to evaluate the closest distance between objects within certain clusters.

Table 3.1

The Basic Structure Of K-Nearest Neighbour Algorithm In Data Clustering (Wu Et Al., 2008)

Algorithm 3 Basic KNN Algorithm Structure

Input : *D*, the set of training objects, the test object, **z**, which is a vector of attribute values, and *L*, the set of classes used to label the objects

Output : $c_z \in L$, the class of *z*

foreach object **y** $\in D$ **do**

| Compute the value of *d* **(z, y),** the distance between **z** and **y**;

end

Select $N \subseteq D$, the set (neighbourhood) of *k* closest training objects for *z*;

$C_Z = \text{argmax}_{V \in L} \sum_{y \in N} I(v = \text{class}(C_y))$

Where *I* (.) is an indicator function that returns the value 1 if its argument is true and otherwise, 0.

3.4 Evaluation Method

The final process of proper noun identification involves the evaluation measures to review the components of the obtained text collection. The motivation behind this process was to determine whether the unannotated data retrieved previously were correctly classified and trained accordingly to their classes and subsequently, subclasses.

As the corpus gathered for the purpose of the research is categorized as unannotated and could be doubted in its authenticity, these Malay text data were assigned to a linguistic expert to be certified before any further analysis is performed. This procedure is also done to reduce the possibility of redundant error that may emerge after all the evaluation measures had been completed. The final product of the annotated corpus was trained using a classifier according to their respective classes and categories.

The process of named entity recognition includes evaluation formulas that is applied to determine the actual values of precision, recall, and accuracy. As classified by aforementioned past research (Kondrak, 2007; Bilotti et al., 2008; Liao et al., 2009; Ekbal et al., 2012; Kaur et al., 2015), the main characteristics to determine the training and retention rate is via observing the precision and accuracy values in which the corpus returns after it has been properly isolated according to its features. Should it be assumed that the procedure of classifications for new named entities in the accumulated sample had its class label known beforehand (during the pre-processing phase), any clustering of this document could be evaluated in terms of the tagged datasets.

Most studies in Indonesian and Arabic prefers to evaluate the efficiency of data training is via the calculation of three values obtained after the text data undergo training phases, which are precision, accuracy, and recall. The importance for the evaluation of these three classification values is to determine whether a categorised corpus would perform under expectation in terms of retention and precision rate in which the corpus is applied with the context of system usage. Citing a few research

outputs (Abdallah et al., 2012; Althobaiti et al., 2014; Gunawan et al., 2015), these research all induced a moderate rate of precision and accuracy in all of their training corpuses.

In order to appropriately measure the degree of validity for a single variable as compared with its constituents that reside in a similar domain or vice versa, a key determinant is required as a baseline measurement. A comparison is made between newly accumulated data and the existing ones so that their correlation could be proven. For research field such as Information Retrieval, confusion matrix table is devised to annotate the accurately labelled data from the ones are not. This table is divided into four main partitions, consisted by positives (true/false) and negatives (true/false). Explanations regarding these values is described in Table 3.2.

Table 3.2

Labels for Positive and Negative Classes in Common Clustering and Classification Problems.

Item	Description
True positive (**TP**)	Actual positive, and predicted as positive
False positive (**FP**)	Actual negative, but predicted as positive
True negative (**TN**)	Actual negative, and predicted as negative
False negative (**FN**)	Actual positive, but predicted as negative

3.4.1 Accuracy

This measure is first proposed by Shantanu Godbole and Sunita Sarawagi in 2004 which is independent of example-based and label-based accuracy measures. The accuracy measure is used to determine the distance between two points of item in a corpus. The accuracy in a data cluster is also known as the ratio of the size of the union and intersection of the predicted and actual label sets. To obtain accurate results, the value is taken from each example and averaged over the number of examples. It is the number of successful hits relative to the total classification number. Most classifier use the accuracy measurement as a standard tool of determining the contents of a particular data cluster (Ekbal & Saha, 2011; Mencar et al., 2011; Witten et al., 2011).

3.4.2 Precision and Recall

Precision and recall values play a significant role with each other in terms of the NER systems' accuracy and retention rate after the components undergone data training. As aforementioned in Section 3.4, the classes-to-cluster evaluation measure further categorises each text data content into 4 consecutive values, mainly true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The values derived from precision and recall rates produced from the system's training would be used to calculate the entire system's F-Score for further evaluation.

Precision values are applied as the measure of accuracy of predicted positives as compared to the rate of discoveries of real positive values, also known as True Positives. Contrary to recall values, the values for precision are only related to the positive side. The precision value, *p* for a data cluster could be determined by:

$$\text{Precision}, p = \frac{tp}{tp+fp} \quad (3.1)$$

Recall values, or also known as Sensitivity, is defined as the section of actual positive cases that were already predicted as the precise true positive values. The value of recall rate is interpreted by the values of the relevant cases that have been picked up by the true positive values. Mostly, recall rate were indicated at a low percentage as usually the detection of desired corpus from the relevance of the documents that are not returned is ambiguous (Markov et al., 2007).

The purpose of initiating recall values is to classify the values of true positive cases after the entire system has underwent training (Mohit et al., 2012). Recall values is also known to only relate to the positive column in the contingency matrix table (Powers et al., 2011). The recall value, *r* could be measured as the following:

$$\text{Recall}, r = \frac{tp}{tp+fn} \quad (3.2)$$

The following figure illustrates the relationship between positive and negative values obtained from the NER system's training and its correspondence in determining the value for the F-Measure.

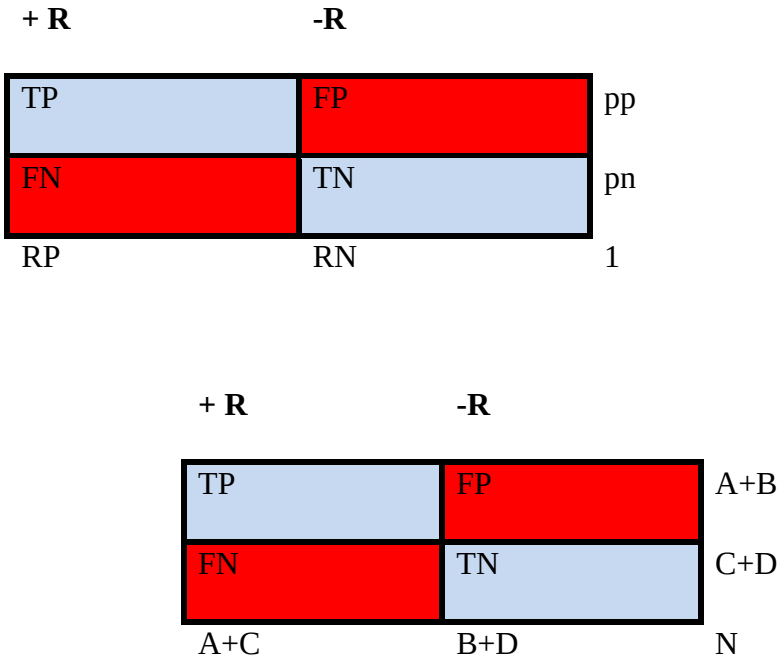


Figure 3.5. Confusion Matrix Representing Positive and Negative Values for Precision and Recall.

Note: The shaded region shows the correct (blue=*correct*) and incorrect (red=*incorrect*) rates.

The two measures, Precision and recall both would display a much more intricate sector when the positive and negative variables is whittled down into prediction scenarios. On the figure above, Recall amounts to only the +R columns while Precision to the +P row. The alphabet **A**, **B**, **C**, **D** are combinations that might materialize after the positive and negative labels are certified. Collectively, **P** and **R** that appears in **pp** and **RP** are abbreviations for Predictive and Real, respectively.

3.4.3 F-Measure

As the evaluation of F-Measure values are derived from both precision and recall rates (Powers et al., 2011), F-Measure efficiently references the entirety of the contingency matrix, which is to take into account both positive and negative values. The idealistic values from the data training's after product is obtained after normalization of both recall and precision's mean values. The basic formula for emergence of F-Measure is as follows:

$$\text{F-Score} = 2 * \frac{(precision * recall)}{(precision + recall)} \qquad (3.3)$$

3.5 NER Process and its Components

This section reviews the process that the research employs within the scope of extracting Malay proper nouns via the implementation of clustering algorithms, including discussions on the relevance for each process to take place.

3.5.1 Data Type

Fundamentally, all corpus resources available consisted of only two base data types, namely lexicons and texts (Miner et al., 2012). Most lexical resources could be represented via a record structure, for example a key plus one or more field. A lexical resource could persist in the form of dictionary or a relatively comparable wordlist.

Along with this, corpora could be formed from a phrasal lexicon, in which the key field is a phrase rather than a single word (Faruqui et al., 2015). Another key element composing a regular corpus is known as a thesaurus.

A corpus is not established based on fully formed materials, however. It involves a careful preparation and input from many people over a collective, extended time period (Abu Bakar et al., 2013). The contents of a corpus are developed from raw text data that needs to be collected, cleaned up, besides being stored in a systematic structure. Despite the complexity and idiosyncrasy surrounding the creation of individual corpora, it is basically a collection of text along with record-structured data. However even when the number of corpus openly available suits the user's expectations, their contents were often not being able to address to the research inquiries (Indurkhya et al., 2010).

A few prominent corpus in the NER research field is seen in English, such as the Brown corpus (Brown University Standard Corpus of Present-Day American English), 100-million word British National Corpus (BNC), and Bank of English corpus with 524-million word collections (Indurkhya et al., 2010). For the purpose of semi-supervised system training that involves linguistic corpus which are lack in availability such as Malay, the size of corpus that is intended to be annotated may be able to perform as catalyst to qualitative research, but this was still not sufficient to give a reliable base for quantification. Therefore, within the research scope study focuses on how the proficiencies of rule-based algorithm affect the training performance.

The retrieval of text structure from a document is divided into two basic structures, mainly lexicon and text. The retrieval of lexicon structure emphasizes more on random text categorisation, and intended to extract certain word structure rather than the entire sentence. This is known as the extraction of fielded record, according to the structure and category of the word itself. The extraction of text leans more on the structure of a singular text, where it constitutes in a word sentence. The order of extraction could be in hierarchical arrangement, depending on the definition context of the word in the sentence segmentation. Text is extracted into named entities that represents their own meanings.

3.5.2 Classification of Text

The NER research tool involves a few web authoring software to develop the main interface, besides programming languages that serve as backend code for the purpose of algorithm construction and corpora extraction. Detecting patterns after extracting the targeted named entities is among the most important process in NLP. As each word entity is unique, there are techniques to recognise the placement of that word, its definition, along with the significance that it carries in the construction of a word structure in the language group itself (Witten et al., 2011). The patterns that are observable in the word structure and frequency happen to coincide with particular aspects of meaning in the article, for example the tense and topic (Miner et al., 2012).

The general definition of classification is defined by the task of choosing the correct class label for a given input. For basic classification tasks, each input is considered in isolation from all other inputs. Furthermore, the set of labels were defined in advance. The basic classification task consists multiple variants. A classifier is known as supervised if it is built based on training corpora containing the correct label for each input (Mencar et al., 2011). The objectives targeted in NER research basically aims at 2 items: (i) *detecting named entities*; and (ii) *extracting those named entities into the form of predefined types* (Abdallah et al., 2012).

As mentioned in the previous sections, three main learning methods were used to fulfil these objectives: rule-based, machine-learning, and hybrid (Abdallah et al., 2012). The procedure of classification training data for a supervised approach is divided into two factions: the training and prediction phase. The process involves six elements: label, input, feature extractor, features, machine learning algorithm, and classifier model. During the training phase, feature extractor is used to convert each input value to a feature set. These feature sets capture the basic information related to each input that should be used to classify it. Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model.

Lexicon
Abstraction: **Fielded Record**

key	field	field	field	field
key	field	field	field	field

Examples for a ***dictionary***:

bangun: ba-ngun [v], a condition that cease to sleep…
jalan: ja-lan [v], progress by lifting and setting down each foot…

Examples for a ***comparative wordlist***:

bangun; wake; awake; awoke
jalan; walk
tulis; write; scribble

Example for ***verb paradigm***:

bangun	terbangun	dibangunkan
jalan	berjalan	menjalani
tulis	menulis	tertulis

Text
Abstraction: **Time Series**

token attrs	token attrs	token attrs	…

time ⟹

Examples of a ***written text***:

Pada suatu masa lalu, Aci dan Muthu tinggal bersama. Mereka merupakan rakan karib. …

Examples of a ***POS-tagged text***:

<adj>Pada<adj><kn>suatu<kn><adj>masa<adj>
<kh>lalu<kh>, <kn>Aci<kn><kh>dan<kh>
<kn>Muthu<kn><adj>tinggal<adj>
<kn>bersama<kn>.

Examples for an ***interlinear text***:

Menanak	me	nanak
Me-nanak	m-e	na-nak

Figure 3.6. Basic Linguistic Data Types, Consisting of Lexicons and Text.

Note: Lexicons are seen to have a record structure, while annotated texts have a temporal organization.

The effort is seen as difficult to accomplish for language collections that are still lacking in abundance (Abdallah et al., 2012; Alfred et al., 2014). Basically, these rules are to match certain patterns related to the list of lookup gazetteers. An example

of rule-based system is demonstrated by Maloney and Niv known as TAGARAB that is seen as an early adapter to resolve the problems aroused in Arabic NER (AbdelRahman et al., 2010). The system applied the pattern matching engine with a morphological tokenizer to identify the main categories in NER, such as Person, Location, Organization, Time, and Date. The approach had seen a result in which morphological tokenizer outperforms individual NE from accuracy terms. PERA developed by Shaalan and Raza also improved accuracy by applying filtration mechanism on the named entities annotated via certain grammatical rules to exempt invalid named entities from being collected (Abdallah et al., 2012).

3.6 Research Tools

Due to the nature of text mining research that emphasized a lot on retrieving text for further analysis, the research scope is going to emphasize on a few programming languages embedded into the development for the internal environment. Among the programming languages that have been chosen in this context are PHP (Hypertext Pre-processor), Python, and Java components. There have been a few authoring tools that have been utilized for these purposes. The primary tools identified during the development process include the following:

3.6.1 Development Platform

In terms of the program codes to be compiled for the purpose of retrieving the text data, along with further pre-processing and analysis, several code editors are introduced in the research.

Python IDLE

IDLE (Python version 3.5) is Python's Integrated Development and Learning Environment. IDLE functions as the primary shell window that operates via command-prompt, with colorizing of code input, output, and error messages. This development environment also operates via multi-window text editor along with multiple text editing features, such as debugger with configurations, browsers, and other dialogs. In IDLE, there are various separate IDEs that provide the service of Python programming that caters to specific features.

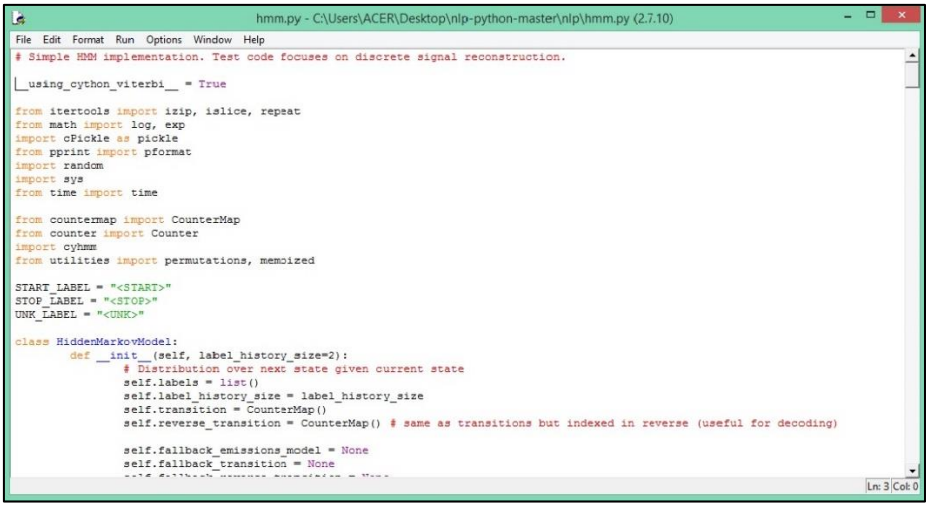


Figure 3.7. A Snippet of Python IDLE.

iPython

A command shell derived from Python IDLE for interactive computing in multiple programming languages. IPython specifically caters to architecture that provides parallel and distributed computing. This development tool operates based on a browser-like notebook component with support for code, text, mathematical expressions, inline plots, and other media as well. An advantage that it possesses is the support for interactive data visualization and use of GUI toolkits.

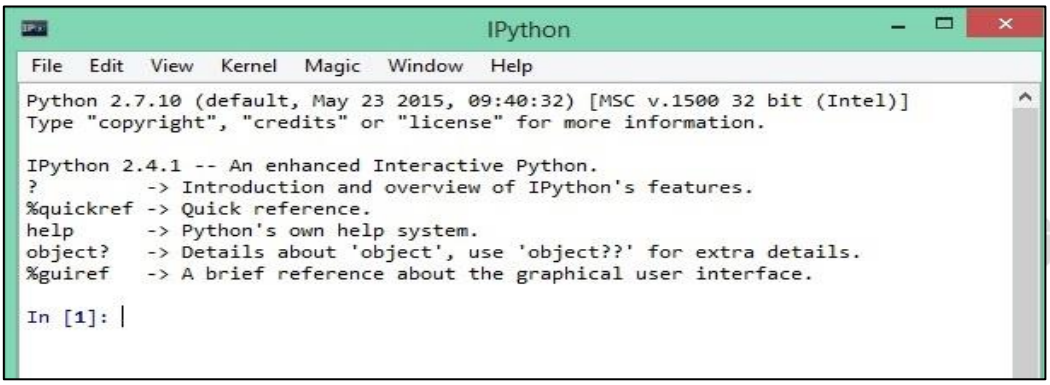


Figure 3.8. A Snippet Section of the Ipython Command-Line Interface.

SciTE

An abbreviation from SCIntilla-based Text Editor, the tool is designed mainly to edit source files, along with to perform syntax highlighting and inline function reference for many scripting languages. Among the features that SciTE included in the service are Replace in Selection, regular expressions are replaced with subgroups, find in files,

code folding, API files, copy formatted, abbreviations, multiple selection, aside from variable width font support.

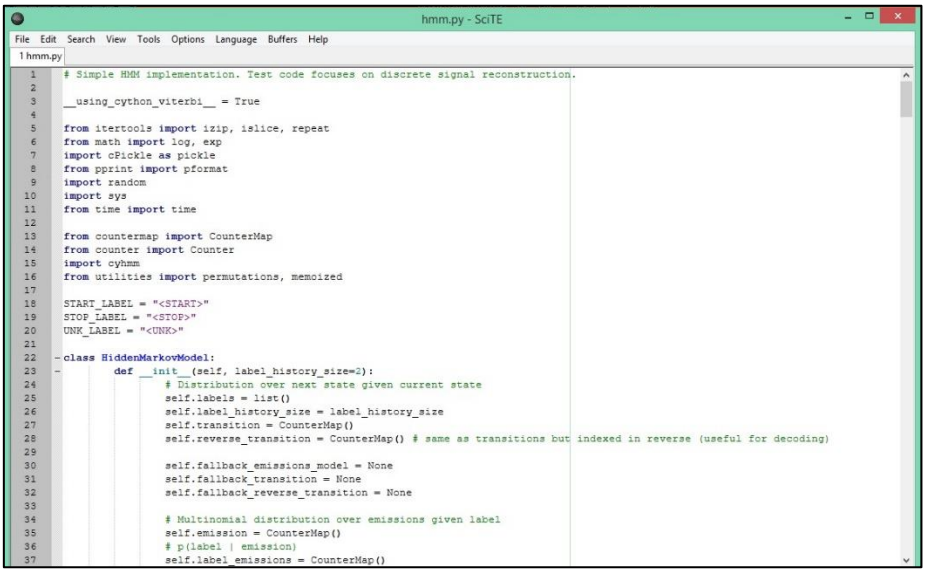


Figure 3.9. A Program Snippet for SciTE interface.

Natural Language Toolkit (NLTK)

Also known as NLTK, this IDE is designed to build Python programs to cooperate with human language data. NLTK provides support and user-friendly interface to over 50 corpora and lexical resources, together with a suite of text processing libraries for pre and post processing features, for example classification, tokenization, stemming, tagging, parsing, and semantic reasoning. NLTK functions by importing corpus and lexical resources to be incorporated with the available source code for further data analysis, process, and evaluation.

3.6.2 Programming Language

In the context of this research, Python is the programming language that is identified as the most dynamic one that offers the most benefits and advantages due to its nature of website development under numerous frameworks, the more prominent ones are Ruby on Rails (RoR) and Django (Python). Python's constant improved library support make it a strong candidate for data manipulation tasks.

i. Python as Incorporation Feature

Most modern computing environments contains a same set of libraries for performing linear algebra, optimization, integration, quick Fourier transforms, and other algorithms. Most programming scripts consisted of minute sections of coding where execution time is spent, along with insignificant glue code. This condition is known as a bottleneck and does not allow algorithms to function efficiently. As used in majority programming languages such as Java and C frameworks, Python reduces the incompetence of these minute sections of snippets to optimize incorporation and fusing of codes among different language environments.

ii. Resolving the "Two-Language" Problem

Most computer science research efforts conduct testing using a more domain-specific computing language before the program is ported onto a larger production system written in more prominent languages, such as Java, C++, and C#. Python is appropriate for performing research and prototyping, along with development of the

production systems. This lessen the issues of converting the coding already done beforehand (legacy systems) into future development projects. The extensive availability of Python repository libraries made it possible for data mining projects to be done in a single programming script, where task delegation for specific features could be reduced.

PHP (Hypertext Pre-processor)

The language itself exists as among the most popular server-side scripting alternative, as its features were also applicable for functionalities toward the client side as well. PHP is used extensively in website authoring, particularly in the executions of functions to be interpreted from the server-side command. Among some of the features that had been identified suited to be applied into data mining systems, including:

i. CURL request

A feature exclusive for PHP scripting, where web page content could be requested using CURL, along with the further access to the web content. CURL could perform some of the basic web scraping concepts, such as conversion of the scraped content into modular objects, sorting of elements from a page based on tags, CSS hooks (class/ID), and integration with server-side content. As the main contents of any particular web page consisted of HTML wireframe, implementation of CURL requests is ideal for rendering and retrieving the requesting content back to the client web server.

ii. Traversing and Navigation through Multiple Page Elements

The elements in CURL contains commands to identify page elements constituting the back-end scripting within a particular web pages, as well as the incorporation of the scraped content with client-side DBMS (database management systems). Some advantages of PHP CURL traits that is applicable in data mining includes the identification of pagination, navigation of crawled content through multiple pages, and association of the data to a client-side database such as MySQL for further access. All of this feature implements basic object-oriented programming (OOP) for the construction of basic classes that could be included in other client-side coding.

3.7 Conclusion

In this chapter, a study on various methodologies that had implemented for the research process is elaborated and reviewed in detail. Together with the nature of research dedicated towards the classification of named entities, the concepts that accompanies the processing features and application of appropriate NER algorithms have been reviewed in accordance with the research scope. Before any further development and analysis is performed on the unannotated corpus, it is considered to be a vital task to elaborate on the approaches that is intended to be included in the process, so as to highlight the steps that would be relevant in terms of text data pre-processing and to apply the suitable sets of rule annotation. This trait also helps in reducing the possibilities of achieving low performance values when the data is fed to the evaluation techniques.

Due to the nature of learning methods which normally utilize the incorporation of machine learning accompanied together by human participation in data aggregation, it is seen that rule-based schematics could be used to improve the significance areas in precision and recall rate retrieved from unannotated text data in Malay documents as these rules were derived by human efforts as well. Contrary to other NER research that utilizes the existing, accessible corpus, this research intends to annotate its own group of Malay text collections to experiment on the entire planned methodology. Therefore, this chapter also discusses in depth on research method that is deemed suitable to be improvised in entirety to support the research motives and target.

CHAPTER 4

DEVELOPMENT AND EVALUATION OF PROPER NOUN DETECTION METHOD FOR MALAY NAMED ENTITY RECOGNITION

4.1 Introduction

This chapter elaborates on the development of proper noun detection method to extract of Malay proper nouns from the unannotated text that can be used to classify named entities in the Malay language.

Further analysis is performed to combine the results obtained from text pattern detection into the implementation of regex rule algorithm in order to prune decision tree that annotates the correlation between word frequencies within an article, compared with a level of urgency for proper noun determination. The regular expression text pattern detection in the previous chapter serves as a foundation to test the theory on how a pre-determined text pattern could locate the desired output from a given text structure as planned. The approach itself could be developed to pinpoint certain features that the end system desires to be achieved, such as linguistic feature detection (capitalization) and outlier character analysis (whitespace, punctuation mark elimination).

The side objective in this chapter is to predict the frequencies of the top word stems extracted from the accumulated Malay news articles. The standard baseline obtained in recent studies sees an average of 50% in the Recall, Precision, and F-Measure values. Experimentations had been done on modification for various parameters to acknowledge the best model with a relatively acceptable tolerance of evaluation values for Precision and Recall after cross-validation on the training dataset. Comparison of different learning models had been done using KNN and Decision Trees to replicate the high average precision and recall rate to improve the accuracy of the learning models. Several pre-processing methods have also been improvised with the custom Malay dataset to further testament the correlation between word similarity and frequency in which they appear. Among the methods applied includes TF-IDF, feature selection, clustering, and term document matrix.

Figure 4.1 represents the involved research flow chart to detect proper nouns based on unannotated text using text pattern identification and clustering methods. Several important stages are highlighted in red, which indicates experiments that had took place. These experiments were planned in phases upon the completion of previous tasks.

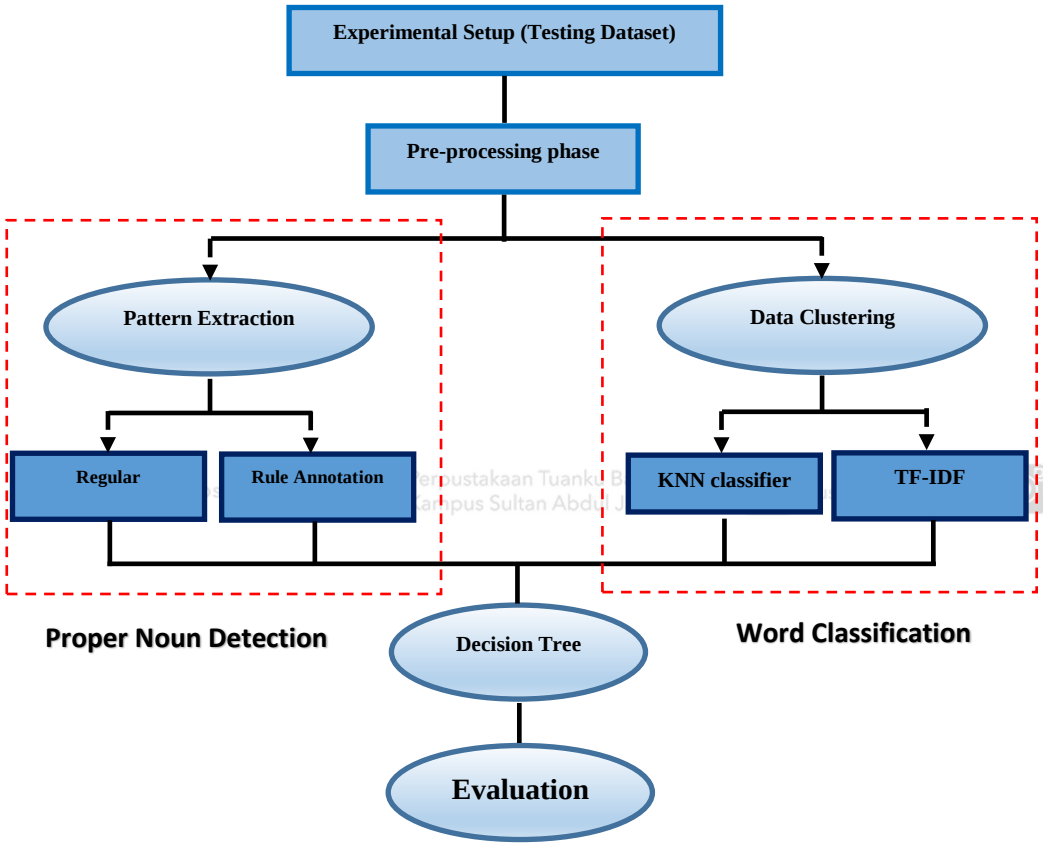


Figure 4.1. An Overview of the Proposed Proper Noun Detection System Process Flow for Malay

Figure 4.1 above summarizes the entire process conducted for this research into 1 flow, culminating from Chapter 4. As in previous explanations, the research has been carried out by accumulating Malay web articles from the chosen newswires, feed the articles through a regex proper noun detection system, data clustering, and eventually evaluation measures and data visualization. During the course of this process, some

approaches may be carried out simultaneously as a proof-of-concept, such as the comparison between the performances of word identification before and after the articles undergo pre-processing. However, all of the reunites at the end during data clustering, where they undergo similar data evaluation & visualization.

The progress of this later research does not always come with the desired end result where there are some unexpected occurrences in the process: most of the techniques suggested to assist the data manipulation fails to support the initial theory. This happened due to the nature of dataset that is highly ambiguous and lack in feature selection after the initial data gleaning is performed. The approaches implemented so far only involves frameworks from Information Retrieval, where future recommendation to overcome this problem is to include more approaches from Information Extraction as the raw Malay dataset is too vague and requires further validation.

4.2 Malay Language Characteristics

For Malay language, the characters are consisted of alphanumeric symbols and hyphens that constitutes a part of Malay word. Hyphens are applied to join the elements of words that is duplicated its usage in a sentence (Malanyon, 2009). For example, the term "*adik-beradik"* that uses the prefix Ber^{+} . Combination of prefixes and borrowed prefix is also applied in the word starting with a capital letter (for example, *anti-Russia*, *se-Indonesia*). In many categories of Malay sentences, different

techniques are innovated to automatically acknowledge phrases. The current Malay spelling system uses the 26 Roman alphabet characters to express the connotation of word structure. Malay word structure also seen a lot of loan words reforming its morphological structure. Different morphological processes are also used to generate new words from a simple or complex base, among these includes affixation where an affix is added to the base.

The Malay language shows a high usage of schwa (the unstressed central vowel in the International Phonetic Alphabet, represented by the symbol ə). The current Malay morphology structure follows the schematics pioneered by Za'aba, also known as the Za'aba style (Zainal Abidin Ahmad, a prominent Malay linguistic expert). This schematic place a diacritic mark on <e> to represent the schwa, and is not user friendly in terms of user writing. The reformation of Malay sentence structure and phonemes also resulted in a few values being added to the rule as well, for example agreements on common letter-values, reduplication of words under the rubric of symmetricity, consonant clusters in loanwords, and word ending schwas for loanwords (Bischoff et al., 1989).

Figure 4.2 illustrates the research design implemented during the entire process.

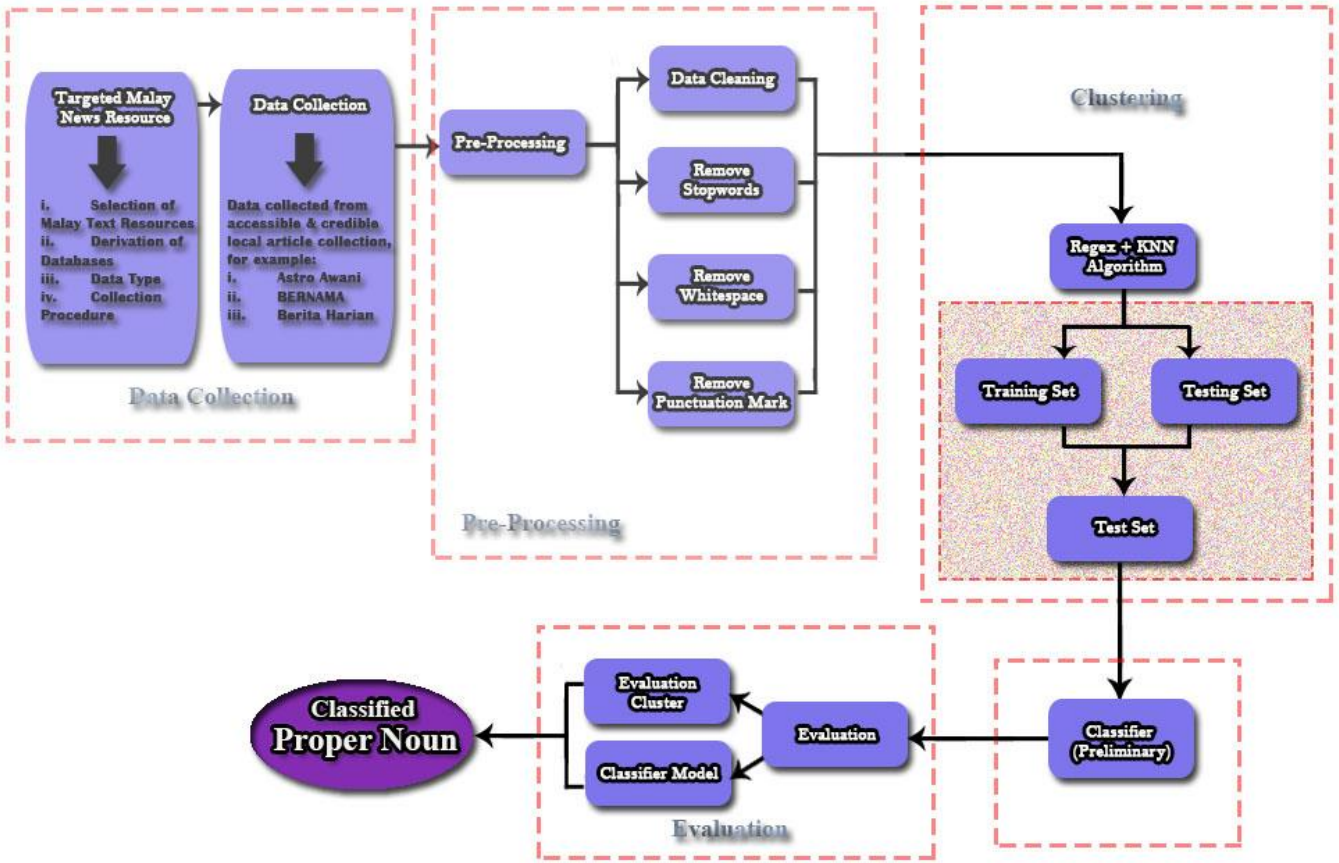


Figure 4.2. Research Design for Proper Noun Extraction from Malay News Article Collection.

For the Malay language, affixes were added to the base words to produce derivatives, such as adjectives and prepositions. The Malay affix rule is broken down into the following ranks.

4.2.1 Prefix Rule

The prefix could be consisted of several other versions, such as *be-*, *bel-*, and *ber-* that changes its structure according to the base word which it is added into. The usage of this prefix rule is to form intransitive verb from other combination of words. The most widely used prefixes in Malay include *Be+*, *meN+*, and *peR+*. Below are examples of the usage of prefix in every day sentences.

I. To enforce action

Example: *Setiap hari Fandy berjalan kaki ke kedai.* (**Fandy walks to the store every day.**)

Here, *beR+* is added to the adjective jalan (definition: walking) to express the action that Fandy did.

II. To reply an action done by others upon oneself

Example: *Kanak-kanak itu berebut untuk mengambil kotak mainan di atas rak kedai.* (**The kids argued among themselves to retrieve the toy box on the top of the shop's shelves.**)

Here, *be-* is added to the adjective rebut (definition: to struggle in order to obtain something) to describe the kids' actions amongst themselves.

4.2.2 Suffix Rule

The suffix is added at the ending structure of a base word. The purpose of this amendment is to produce derivative words from adjectives, conjunctions, and nouns. Suffix usage is limited to only certain compatible words, and only serves to relay state of objects or consequences from inflicted actions. Below are suffixes in situational and contextual inferences.

I. Tools/Appliances

Example: *Kukuran, empangan.*

Usage in sentences: *Empangan Bakun dikatakan mampu menampung banyak air*. (Bakun dam is claimed to be able to contain a huge amount of water.)

II. Places

Example: *Lapangan, bangunan.*

Usage in sentences: *Lapangan Terbang Sibu boleh menampung kira-kira 5 kapal terbang pada satu-satu masa.* (Sibu Airport could support about 5 airplanes in a time.)

III. Variety of types for objects

Example: *Buah-buahan, bunga-bungaan, bijan.*

Usage in sentences: *Datuk memiliki ladang buah-buahan yang besar di Kanowit.* (Grandfather owns a huge fruit orchard in Kanowit.)

IV. Annual sequence of events

Example: *Bulanan, mingguan, tahunan.*

Usage in sentences: *Bank itu dikatakan mempunyai kadar faedah tahunan yang tinggi berbanding dengan bank tempatan lain.* (That bank is said to provide a high interest rate per annum as compared to other local banking institutions.)

4.2.3 Prefix-Suffix Pair Rules

The prefix-suffix pair rules have a slight uniqueness compared to the previous duo, in the terms that the combination of prefix and suffix is used to represent preposition or conjunctions. Therefore, the usage of prepositions is abandoned after prefix-suffix pair rules is used. The order which it is placed is fixed either at the front or back of the base word. Several known pair rules are *beR-…-kan*, *ke-…-an*, *-i*, and *meN-…-i*. The examples in its application are elaborated as follows.

I. Pair rule beR-…-kan

Example: berdasarkan, bermatlamatkan

Usage in sentence: *Dia melakar peta itu berdasarkan panduan peta telefon pintar.* (He sketched the map based on the reference from the map on the smartphone.)

II. Pair rule –kan

Example: melaporkan, diceritakan

Usage in sentence: *Pihak polis melaporkan bahawa terdapat kemalangan di Sektor 9.* (Police reported that there is an accident at Sector 9.)

III. Pair rule ke-…-an

To represent location

Example: kementerian, kediaman

Usage in sentence: *Perdana Menteri menetap di kediaman rasminya di Putrajaya.* (The Prime Minister resides at his official residence in Putrajaya.)

To represent situation

Example: keputusan, kemahiran

Usage in sentence: *Adik memperoleh keputusan yang memuaskan dalam peperiksaan awam lepas,* (Brother obtained an average result during the last public examinations.)

4.2.4 Infix Rule ($^{+}$er^{+})

Infix rules is applied to form proper nouns, conjunctions, and adjectives. The base word is assumed to have the similar characteristics as the infix itself. Several derivative words that could arise from this practice includes the circumfix *–el-*, *-em-*, and *–er-*. However, this rule is not often used in today's Malay sentence structures as other derivatives. Below shows some of the base words that have been merged with the infix rule. Example:

Selenggara (senggara + infix *-el-*)

Kemuning (kuning + infix *-em-*)

Kerontang (kontang + infix *-er-*)

Serabut (sabut + infix *-er-*)

Seruling (suling + infix *-er-*)

4.2.5 Malay Stop Word List

The following list of words are the most predetermined collection of high frequency words that possess a positive tendency of appearing in almost all published news articles every day, ignoring the genre and the word count. These total of 40 words are inputted as a filtering agent during the text pattern detection. The initial listings of stop words have been retrieved from several scholar sites relating to Malay linguistic analysis, before being compared and sorted out eventually into 40 most used words in any given text according to several works.

Table 4.1

Top 40 Imposed Malay stop word (Sidi, 2011)

adakah	akan	amat	antara	apabila
atau	bagi	bahawa	dalam	dan
dari	daripada	dengan	di	hendak
ialah	jangan	jika	juga	kalau
ke	kepada	kerana	mahupun	malah
manakala	masih	masing-masing	oleh	pada
pun	sahaja	sedang	sudah	telah
tentang	tetapi	untuk	walaupun	yang

4.3 Regex Detection of Malay Proper Noun (Experiment 1)

The huge collection of entity extraction processes could be performed via a feature in Information Extraction to carefully identify and retrieve desired word structures from certain sentence context, known as regular expressions (regex). Regex is developed in Pattern Recognition field to process and identify similarities of patterns encompassing object structures. Among the benefits of using regex to identify text patterns is the providence of domain knowledge along with credibility of restricting the search space (Li et al., 2008). Primarily, the regex provides a natural mechanism for exploration to provide a further insight knowledge about the structure of the extracted entity. Secondly, the space of output regex that had been predetermined could be restricted down to the minimum via the appropriate refining of the relationship among text entities to the input expressions (Li et al., 2008).

4.3.1 Regular Expressions (Regex)

Regular expressions are defined as a specific kind of text pattern that could be applied towards many programming languages to extract certain features in the text cluster. The features could be based on text that matches the pattern within a larger text collection, as a substitute for other text or chunks of the matched text, and to divide the text block into a listing of subsequent texts. Regular expression consists of a string of combination of normal characters and special meta-characters. Meta-characters are characters or sequences of characters that represents objects, for example temporal

expressions and character types. In order to define a regular expression that matches the character which is desired to be discovered, pattern matching is performed extensively.

This process consisted of locating a section of text that matched by a regular expression. The following section addresses the pattern matching features used by several programming languages compatible with the research scope, along with rule schematics derived to properly detect Malay word nouns.

For the regular expressions applied in the extraction of Malay proper nouns, only specific features that are needed would be placed into consideration for retrieving. As the regex patterns are developed specifically to extract certain features from Malay sentences, the process of determining text entities rely on how accurate the word structures that is retrieved would attend to the main classes in Named Entity Recognition (in this scope PER, LOC and ORG).

- **LOCATION*:*** Extraction of any location names, such as places and unique destinations. These words are expected to contain nouns with capitalization, for example **Sarawak** and **Malaysia**.

- **PERSON*:*** Retrieval of any words that share the most similarities with names of persons, ignoring ranks, jobs, and prepositions. For example, **Farid**, **Arima**.

- **ORGANIZATION*:*** Extracting all form of words that is synonymous with organizational names including abbreviations & word characters. Among the examples of organization names are **UMNO**, and **Perbadanan Tabung Pendidikan Nasional**.

- **MISC:** Miscellaneous characters are nouns that may appear in a similar manner with proper noun, however is not considered a proper noun as they cannot stand alone without the presence of certain word features to support. Examples of miscellaneous words are **Perdana Menteri** and **Naib Canselor**.

Words that adhere to premediated criteria are deemed as correct and positive, while those that fail to associate with any of the text pattern rules is referred to as not the target and a negative output. A manually created regex is currently referred to as the commonly improvised solution in order to retrieve certain desired text information (Li et al., 2008). The following example demonstrates how regex attempts to detect certain word features from a resource.

To extract the name of any location in a paragraph:

The simple regular expression of the task is formed by detecting the feature that the names of locations would have as compared to the other word in a sentence structure. Most location names come in capital letters to distinguish their importance; however, there are instances in which location names are not capitalized depending on importance in society. Therefore, the most common assumption to form a regex pattern is "1 or more capitalized word, followed by small capitalized letters in an order of alphabetical sequence".

The regex for location is **R = / ([A-Z])\w+/g**

To extract only the phone number that appears within a sentence, without discrimination:

As phone numbers are usually formed by a combination of numbers and occasional hyphenation to distinguish their unique structure as compared to other numerical features, the forming of regular expression pattern for phone number must take into account the detection of number before any other string characters. Phone number also contains distinguishing area codes and number sequence depending on the location in which they are used. However, the most common of regex pattern for phone number detection should detect a sequence of number alignment together with special characters, such as whitespaces, inward dashes, and hyphenations. The assumption to form a regex pattern of phone number is "non-alphanumerical characters".

The regex of phone number is **R =/ \b\d{3}[-.]?\d{3}[-.]?\d{4}\b /g**

Figure 4.3. Examples on How Regular Expression Feature is Derived, based on Word Characteristics and Definitions.

This research effort is motivated to investigate the regex pattern behind catching all word structures that fit the defined criteria, before merging it with more intricate morphological features mentioned in Section 4.2. In order to discuss the characteristics of regex pattern devised to detect Malay proper nouns that bears similarities with English, the following excerpt further elaborates the features that is attempted to be imparted within the process.

1. Begins by searching for sets of capitalized words. This is basically 2 halves separated by an "OR" (the pipe "|")

```
(([A-Z]([a-z]+|\.+))+(\s[A-Z][a-z]+)+) | ([A-Z]{2,})
```

2. Searches for one or more strings that begin in capitalized form, followed by either a string of lowercase letters, or a full stop.

```
(\s[A-Z][a-z]+)+
```

3. Word structure followed by a space, another capital letter, accompanied with 1 or more lower case letters.

```
|([a-z][A-Z])[a-z]*[A-Z][a-z]*
```

The full regex pattern when morphed into 1 statement would be:

```
(([A-Z]([a-z]+|\.+))+(\s[A-Z][a-z]+)+) | ([A-Z]{2,})
(\s[A-Z][a-z]+)+ |([a-z][A-Z])[a-z]*[A-Z][a-z]*
```

The next listing shows the miscellaneous features applied to extract or eliminate certain word features from articles.

Date in format dd/mm/yyyy

Regex:

```
/^(0?[1-9]|[12][0-9]|3[01])([/\-])(0?[1-
9]|1[012])\2([0-9][0-9][0-9][0-9])(([-])([0-1]?[0-
9]|2[0-3]):[0-5]?[0-9]:[0-5]?[0-9])?$/
```

Time in 24-hour format

Regex:

```
/^([01]?[0-9]|2[0-3]):[0-5][0-9]$/
```

Date and time in ISO-8601 format

Regex:

```
/^(?![+-]?\d{4,5}-?(?:\d{2}|W\d{2})T)(?:|(\d{4}|[+-
]\d{5})-?(?:|(0\d|1[0-2])(?:|-?([0-2]\d|3[0-1]))|([0-
2]\d{2}|3[0-5]\d|36[0-6])|W([0-]\d|5[0-])(?:|-?([1-
]))))(?:(?!\d)|T(?=\d)))(?:|([01]\d|2[0-4])(?:|:?([0-
5]\d)(?:|:?([0-5]\d)(?:|\.(\d{3})))(?:|[zZ]|([+-
](?:[01]\d|2[0-4]))(?:|:?([0-5]\d)))))$/
```

HTML tags

Regex:

```
/^<([a-z1-6]+)([^<]+)*(?:>(.*)<\/\1>| *\/>)$/
```

Email

Regex:

```
/^.+@.+$/
```

URL

Regex:

```
/^((https?|ftp|file):\/\/)?([\da-z\.-]+)\.([a-
z\.]{2,6})([\/\w \.-]*)*\/?$/
```

Phone number

Regex:

```
/^\+?(\d.*){3,}$/
```

4.3.2 Difference in Text Data Detection (Before & After Word Processing)

Various researchers have described the advantage of pre-processing any type of dataset before further analysis (Al-Moslmi et al., 2015; Trstenjak et al., 2014). This factor is influenced by the motivation of improving the feature detection within the cluster itself, besides preventing unwanted outliers influencing the overall rating of system performance. There is also evidence of a slightly improved system performance evaluation after the domain dataset undergo processing steps (also known as normalization), such as the elimination of punctuation marks and lowercasing all present string characters (Le Nguyen et al., 2014).

In order to isolate the newly located proper noun when the article is feed to the regex detection system schematic, pre-processing is done between intervals of article retrieval and data cleaning. The article collection from the scheduled time range is processed by few pre-requisites, such as extra whitespace elimination, and punctuation marks removal. The product is then feed to the regex detection system to categorise the present proper noun, as assumed during the annotation of pattern detection schemes above.

4.4 Cluster Analysis

For this phase involving the detection of text pattern using regular expression, the target which is emphasized in the process is the identification of noun structure from a given word paragraph. Proper nouns consist of various nomenclatures and motives within the definition depending on its contextual usage affecting the entire sentence structure that contains the information. Usually proper noun begins with a capitalized form, lining up the frontal part of each sentence construction (Nadeau, 2007). There are fewer possibilities of proper noun to appear at the back of a sentence capitalized, which if it does the noun would be appearing in small capital letters. An example is shown as follows:

Abang saya meminati kumpulan Red Velvet dari Korea. (*My brother's favourite group is Red Velvet from Korea*).

In this sentence, "Red Velvet" and "Korea" are proper nouns. Although Red Velvet is an English term that remains in its native form, a Malay sentence does not convert the second proper noun into lower capitalization. "Red Velvet" and "Korea" also remains capitalized. Proper nouns could exist to present adjectives, prepositions, or a representative for important real-world objects, as per the major named entity classes such as Person or Location. Languages such as Arab do not place much emphasis on the order of affixes and suffixes (Althobaiti et al., 2014), while word segmentations in Chinese do not properly represent definitions of word meanings

(Luo et al., 2016). The alignment of proper noun in a sentence structure varies in terms of linguistic context, and the chronological order in which the word is arranged to form a sentence to express something.

To investigate the relationship between states of text data before and after they undergo cleaning and processing from outliers such as unwanted whitespace and punctuation marks, the research projects one experimental classifier derived from the regular expression application before and after the text data is processed. The text data used in this experiment utilizes 2 forms: raw form, where all text elements are retained as its native form retrieved from the webpage, and the processed collection in which punctuation marks are eliminated and texts are normalized. As the study indicates, there is a slight variation detected in the efficiency of proper noun detection after the text data is modified.

The collecting process of news content from the 3 major news corpuses is delegated into stages for a range of 1 month, beginning on 24th March 2016 and ends on 29th April 2016. This stage further divides the news content collection into 10 days between intervals, beginning on 24, 28 (March 2016) and 1, 5, 9, 13, 17, 21, 25, 29 (April 2016) consecutively. The web scraping process only views the news content with the most quality content, along with the range of words. The scraping process also takes into consideration any most highly acquirable genre with appropriate information relay, however each scrapped article would undergo a brief review to determine its compatibility with the research scope. For example, gossips and content-

sensitive articles are omitted from extraction, even though they exist in abundance as compared to sports news.

Figure 4.4 visualizes the supposed process on how to evaluate the obtained proper noun for the rate of Precision and Recall. The initial procedure starts with the retrieval of appropriate news articles from the online news website. Labelling in this sense means to identify the presence of proper noun, based on how regex rule recognises the capitalization of wordings and punctuation marks in which would determine where the proper noun is positioned in a sentence. Input refers to the retrieved news articles. The extraction of features in this sense means to excerpt necessary traits from a collection of words, carefully isolate them into the wanted words (in this scope, Malay proper nouns). Feature extraction includes identification of prominent word traits that could indicate a proper noun, such as common names, locations, and uppercase consisting of initial letters.

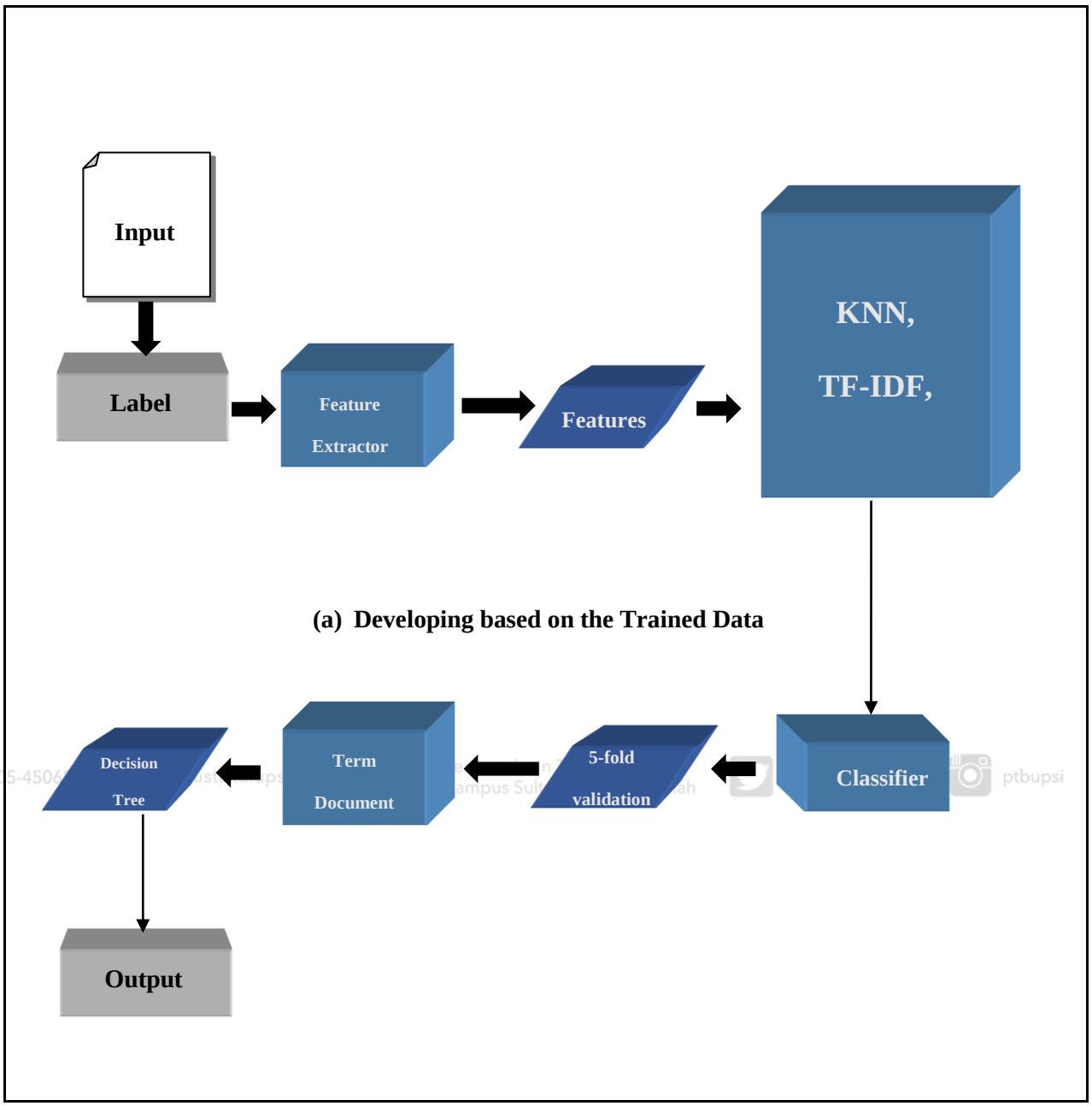
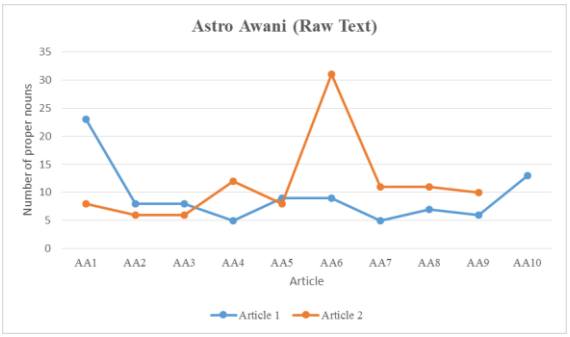
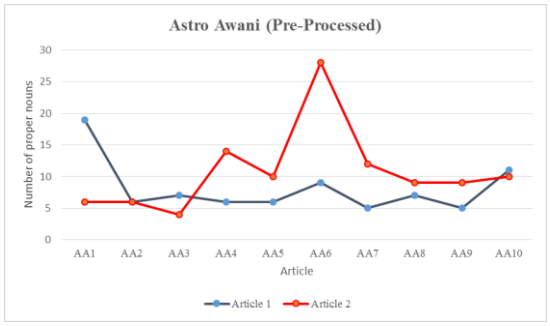
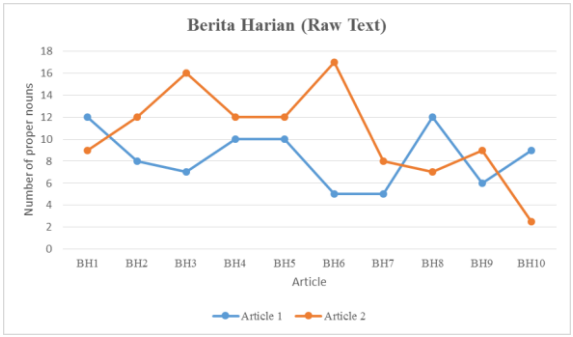
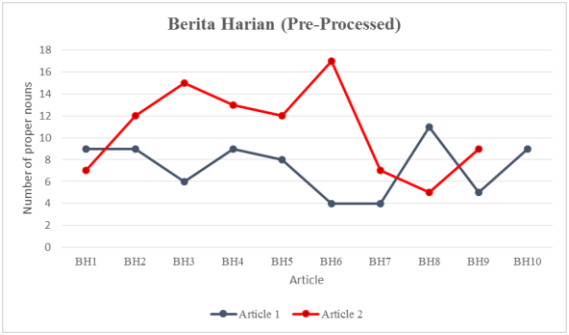


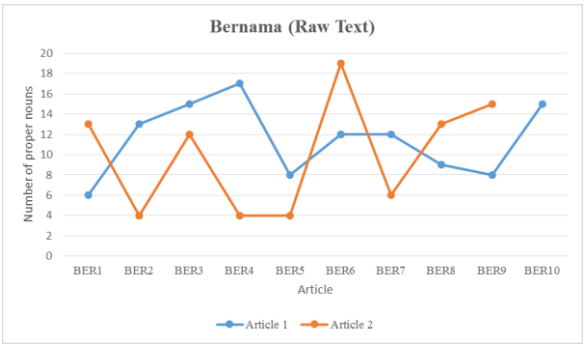
Figure 4.4. Fundamental Framework of a Supervised Classification.











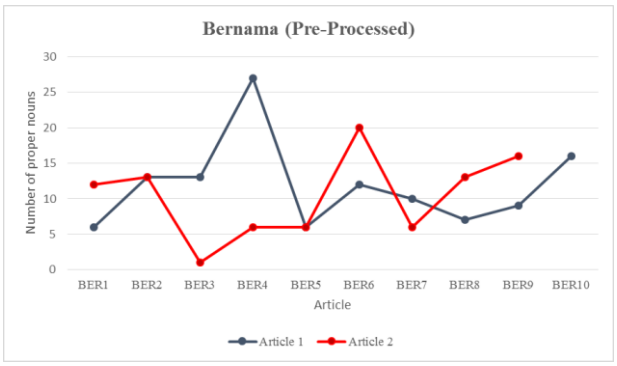


Figure 4.5. Detection Rate Proper Noun from Regex Rule Before Processing

Figure 4.6. Detection Rate Proper Noun from Regex Rule after Processing

Figures 4.5 and 4.6 is the generated graph for frequency of Malay proper noun appearing in news articles. As seen from the graph, there is a distinct difference of the number of proper nouns that is managed to be detected between the raw collection and the pre-processed collection from the targeted news article in terms of slight percentage increment of the latter over the former. The average percentage of increments is seen in between 5-10%. However, with the majority results obtained shows a slight increase in values, there are also random instances in which the value dropped slightly as well. This value also ranges between 5-10%. All three news corpuses practice an ethic of uniformity in its effort of delivering information to the crowd, according to the motive of the news articles. Most of the accumulated text data is collected from politics, sports, technology, and current affairs genre. This news is constantly updated between intervals over the period of 24 hours, therefore there is almost no instances where new content is not added.

Each of these corpuses review content based on its conformity in the latest content obtained from various media, where Astro Awani and Bernama delivers its content asynchronously with its broadcasting news station. The contents that both of them publishes were derived directly after hour-to-hour basis news show is broadcasted on television. However, Astro Awani news content is seen to detect more noun usage in its online articles with the number of occurrences that is managed to be detected occasionally rallies in between 5 to 35 per articles. Bernama news articles are also seen to practice high integrity in the sense of persistent article span and coverage over its news content by hour of day due to the multilingual domain.

Each article was translated after another version of its content is published online, be it English over Malay, Malay over Chinese, and so on. However, upon close review, these articles still possess content which is relevant and persistent with each other. Bernama has a lower proper noun discovery per article, with a range between 2 to 28. Berita Harian possesses the lowest value of discovery among those two, with a value ranging in between 3 to 17. The reason behind the low value could be the inconsistency of article length and content.

The major difference that emerge before and after the unannotated text is collected is in terms of how much after processing influence the detection of any noun presence, possibly improving them better. In this instance, it could be assumed that there are inconsistencies in the detection percentage after the cleaning process is performed. The most noticeable difference is shown by Bernama text data, where the output results greatly varies compared to before. For Astro Awani and Berita Harian, both of these text data collections illustrate an elevation of detection percentage of the latter over the former, where the most noticeable change is the fluctuation of the latter over the former. After the text data is processed, it is shown that the rate of detection is higher.

However, all 3 of these news corpuses share a common similarity, which is the detection of articles that is done latter indicates a higher usage of noun in context, as compared to when the data is obtained at an older time. The scraping process is done based on the comparison between the old and newer news article emergence. The older news article is labelled as Article 1, and the latest is deemed as Article 2. The

reason that the research approach only targets two news articles per day is to determine the difference in between the older and newer article entry; indication of the urgency of the news content is seen in how much proper noun is utilized in its content.

In this scope, the targeted articles are those that is brief however carries substantial information. In all the three news corpuses, it is observable that there has been a rising trend of publishing short but summative content, ranging between 200 to 500 words. Only on comprehensive broadcasts, articles that contain more than 600 words could be seen. All the three corpuses exhibit the similar behaviour in journalism. The only difference that emerges among the three is based on the word structure that the writer stylises so that their articles is simple and expresses direct point conjecture. Except the case with Bernama that consists of multilingual translated content, both Astro Awani and Berita Harian publishes articles in a similar range of word count that is stated earlier. Chapter 5 briefly analyses the word count frequency for each included article.

Before regex rule is applied further, the research attempts to review the benefits of clearing the text materials from outliers that may dampen its final performance. In this stage, pre-processing takes into consideration the magnitude of importance for the information that is tried to be conveyed, therefore all unrelated text heading and footer is removed alongside punctuation marks in order to normalize its content. Table 4.2 shows the difference in word count, where T1 = total number of words before cleaning, T2 = total number of words after cleaning, O1 = total number

of word collection before cleaning, and O2 = total number of words after cleaning. The total number of words before and after this alteration is performed varies, where before (Astro Awani=2729, Berita Harian=2434, Bernama=4418) and after (Astro Awani=2529, Berita Harian=2233, Bernama=4254) bears a slight decline in word count. The total number of characters eliminated in the process for Astro Awani, Berita Harian, and Bernama stands at 7.33%, 8.24%, and 3.71% respectively. These results are gathered for further investigation of overall system performance after applying clustering and rule-based approach.

Table 4.2

Difference of Word Count between the 3 News Corpus Before and After Pre-Processing

Article	Total Number of Words					
	Before Processing	After Processing	T_1	T_2	O_1	O_2
AA01	351	333	**2729**	**2529**	**9581**	**9016**
AA02	329	312				
AA03	153	132				
AA04	206	186				
AA05	206	185				
AA06	168	151				
AA07	524	505				
AA08	196	176				
AA09	277	253				
AA10	319	296				
BH01	258	233	**2434**	**2233**		
BH02	181	158				
BH03	209	189				
BH04	232	212				
BH05	288	269				
BH06	303	287				
BH07	210	191				
BH08	266	247				
BH09	323	303				

(Continued in next page)

Table 4.2 *(Continued)*

Article	Total Number of Words					
	Before Processing	After Processing	T_1	T_2	O_1	O_2
BH10	164	144				
BER01	313	299				
BER02	1086	1068				
BER03	452	436				
BER04	478	456				
BER05	241	226	**4418**	**4254**		
BER06	592	574				
BER07	400	384				
BER08	195	180				
BER09	362	346				
BER10	299	285				

Where:

T1 = total number of words before cleaning,

T2 = total number of words after cleaning,

O1 = total number of word collection before cleaning, and

O2 = total number of words after cleaning.

4.5 Implementation of Regex Rule Pattern Extraction to Detect Proper Nouns

In this research, a framework to develop Malay proper noun detection system based on regex algorithm is introduced. In the base of detecting named entity from a collection of unknown data group, regular expression approach were initiated earlier to identify these named entity existence to the maximum (Li et al., 2008). Although regular expression techniques are seen to lean more towards locating certain text pattern that is predetermined, the concept of regex is used as an alternative to replace the conventional POS tagging that is usually applied for further identification of unannotated text documents (Chapman et al., 2016).

In terms of developing training and testing classifier to emulate the results of a system's performance in terms of real-world practice, via its precision result and rate of retention, a classifier is produced from the final testing of both the training and testing classifier models combined. This works to initiate a primary model as a guidance for future attempts in evaluating new data input into the existing collection, and determining how much new input is suitable to optimize the overall performance of the entire learning system. During the process of classification, the development process might slant towards whichever type that is suitable for the current data collection, also influenced by the factor of data availability. Two types of classification systems may emerge, namely the type that is devised to work automatically, and those that exist to assist user in decision making (Král, 2014).

In order to properly augment rule pattern identification theory to the conventional NER classifier methods, several aspects need to be taken into

consideration, such as robustness and level of comprehension for users. Malay orthography contains an obvious difference to English due to the nature of Malay words that consists of more syllables but fewer letters, based on the variance in term of the syllabic structure of root and bound morphemes (Isa et al., 2013). The constituent of Malay is more agglutinative among its used words. This in turn makes the nature of Malay morphology contribute to the lower number of letters against the number of syllables.

Compared to English, Malay has shallow orthography-phonology mappings, general morphology, along with brief syllable structures. The current form dictates more towards using Baku. Malay is an immersive language, where its trait possesses a lot of values derived from neighbouring linguistics or the ruler before Independence's language loanwords. In morphological processes for Malay there lies 3 main elements, which are affixation, reduplication, and compounding (Yap et al., 2010). Prefixes and suffixes are both identifiable as prone to rule based; however, both are distinctively transparent and rule-based. Both of them functions as appendages to the morphemic stem. These rules are tested by moulding into the proper noun detection system via various simulations.

Malay is written in alphabetical form called Rumi, an alphabet rule originated from Latin. Transliteration of other languages into Malay also usually undergo transformations in terms of sentence structure and translations before being conveyed into a meaningful form. Rumi consists of 5 vowels and 20 consonants, and there are also features that were not fully utilized in the formation of a sentence (Yap et al.,

2010). Some instances of characters not normally used are consonant *x*, and several consonants such a *v* and *q* that is only discovered in foreign loanwords. Among the pronunciation system widely inducted in the Malay Archipelago, there are three that are used synchronously: Non-standard Malay, Standard Baku, and Singapore Baku (Yap et al., 2010). In the context of public usage such as those by official government documents and online articles, Standard Baku is used.

According to Karim, a renowned Malay linguist, there are three main known morphological processes in Malay when a proper noun is formed with its sequence and contextual usages in a sentence: affixation, reduplication, and compounding (Yap et al., 2010). Prefixes and suffixes are incurred extensively to relay grammatical relationships and to form new words. Rule-based approach came into emphasis during the determining of prefixes and suffixes.

To identify the presence of proper noun, features such as affixes in front and end of a word is identified prior, some of these such as *ke-*, *ke-an*, *peN-an*, and *–kan*. Singular text chunk is improvised with prefixes and suffixes in order to relay its state of usage, be it from first, second, or third person view. Stems also appeared in front of proper noun to reform monosyllabic word, for example *meN-* and *meng-*. The Malay language has very persistent mappings for orthography-phonology aspect; however, the syllable structure may reduce accessibility more than the orthography seems to indicate. The following explains the formation of Malay word under stemming and morphology mappings, and the structure that could be broken down into singular proper nouns.

1. Affixes

i. Retrieving nouns

Example:

peN-, pe-, -an, ke-an, per-an, peN-an

ii. Verb inflections

Example:

meN-, beR-, ter-, peR-, -I, -kan

iii. Some word changes its structure according to the stem that follows.

Example:

meN- to *menge-* (monosyllabic word)

meN- to *meng-*, *mem-*, *meny-*, and *me-*

iv. Pushes the first/last letter to the stem word

Example:

makanan (food)

makan (to eat) + *-an*

2. Prefix and suffix

i. Appended to morphemic stem

Prefix pe- is appended to singular adjective nouns to form affixed word.

Example:

pe- + *sara* (to provide support financially) = **pesara** (retiree)

pe- + *nyata* (state a situation/condition in the environment that is visible)

= **penyata** (financial statement)

ii. All stems changed when prefix is added to the noun (*p*, *k*, *t*, *s*)

Example:

peN- is added in front of words and the state remains unchanged. However, this suffix changes into peM- when added to verb to form the new affixed word.

pem- + *beri* (to give/giver) = **pemberi** (to provide something/provider)

pem- + *bayang* (shadow) = **pembayang** (clue)

The instance does not change with the affix men-, and is usually applied to compliment another word's definition or amplifying the action of a verb.

men- + *kenal (acknowledge/recognise)* = **mengenal** (recognize)

men- + *sapu (sweep)* = **menyapu** (to sweep)

3. Reduplication

Used to mark plurals, repetition of words that is considered as 1 existence. These words are calculated as 1 quantity.

Example:

i. *Rama-rama (butterfly)*

ii. *Kanak-kanak (children)*

4. Compounding

Words consisted of more than one stem. These words are generated from two (or more) words with different meaning to form 1 new noun.

Example:

i. *tanggungjawab* (responsibility)

tanggung (bear the burden of) + *jawab* (to answer)

i. *keretapi* (train)

kereta (car) + *api* (fire)

The rule pattern identification for this research context is a combination of small fraction of data clustering approach augmented with local search algorithms encroached with regular expression detection in text pattern. Previous works addresses the efficiency of word structure detection in terms of its distance and similarity frequency when clustering algorithm is applied, as compared to the standard NER identification tools and methods. In the upcoming chapter, a more intrinsic explanation will be placed on the research's implementation of text pattern identification process.

As proposed in this research, Malay text data that is collected in phased time is experimented in their pre and post processing nature. The alteration is omitted on the entire phrases in the text collection during the pre-processing phase, as per the pre-

requisite of eliminating any outliers that might influence the optimal time of word structure detection. The transformation is done as the rules defined below:

- The removal of all stop words that had been predetermined as may possess the highest frequency of occurrence, such as *adakah*, *atau*, and *dari.*

- Removal of all punctuation marks, such as fullstops (.), comma (,), and apostrophe (‘). However open and close brackets are retained, as some words tend to borrow foreign words to justify its importance therefore need to be encased with the origin word. For example, *UMNO (United Malays National Organization).*

- Collapse all repetitive characters into a single character in order to remove duplicates. Malay language tend to use several repetition words as conjunctions to enforce the degree of importance for a state of condition or to express doubt. For example, *Kadang-kadang*, *Sama-sama*.

4.6 Structure Flow

The following section would elaborate on the structure hierarchy of the program codes used during the detection of noun words from the collected text resource. From here on out, the data clustering process is performed simultaneously to be divided into two phases to fulfil the purpose of developing a training and testing model. The implementation of regular expression is executed in the 1st phase: detection of raw proper noun structure. For the 2nd phase, further classification process involving

clustering and evaluation methods is initiated to document the method used, before the final proper noun listing is formed.

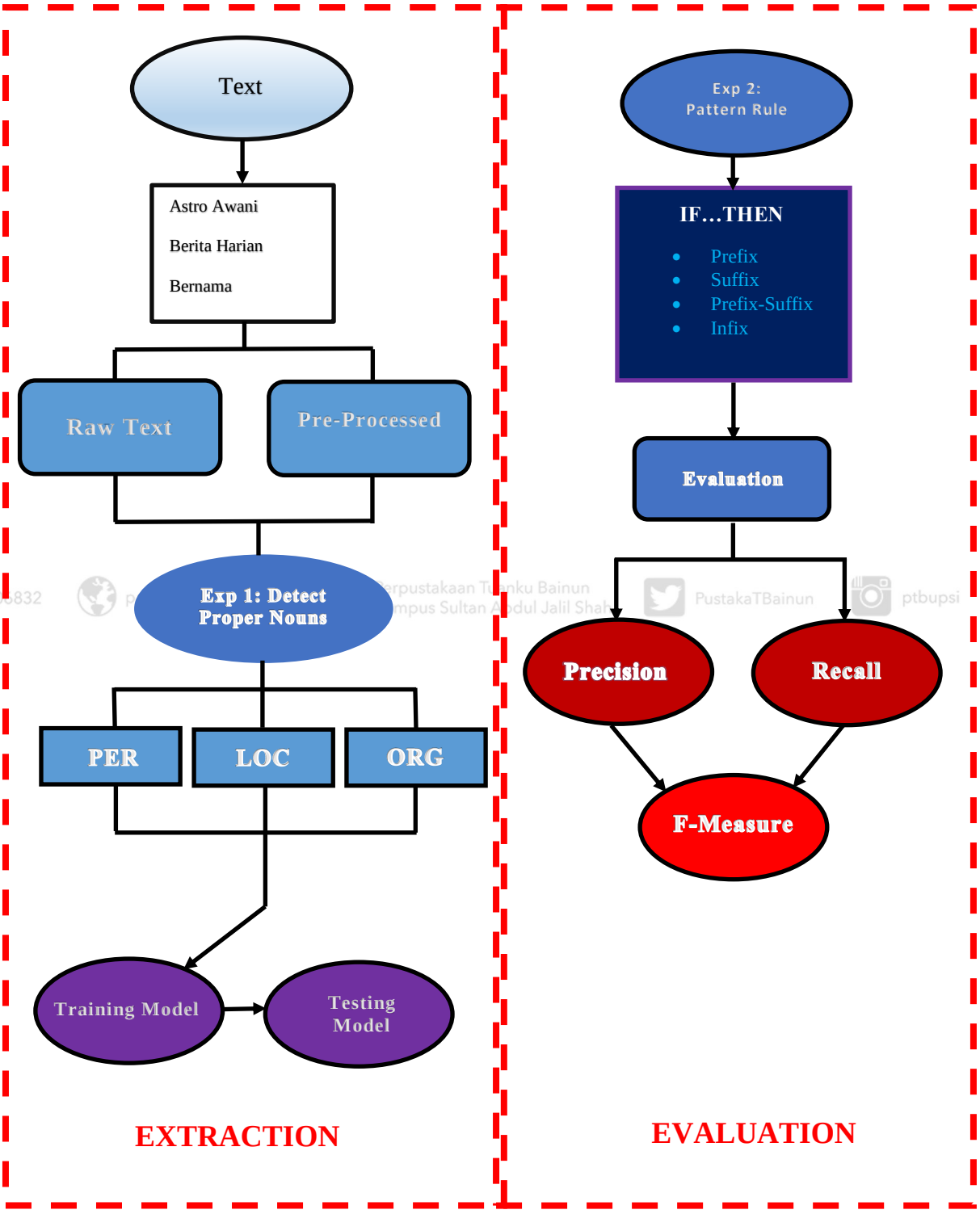


Figure 4.7. Workflow on Extraction of Proper Noun Structure using Regex Rule.

Table 4.3 (*a*)

Clustering Process to Identify and Classify Unlabelled Data from Labelled Data during the Training Phase

Algorithm 1 Training Phase

Input: *L*, *U*, *R*

Output: *C*, δ

1 Tag **POS**, lemmatization of *L*, *U*

2 Collecting **PN** of Malay NE

3 $C_s \rightarrow$ cluster *(U)*

4 **repeat**

5 form model **KNN** for *C* based on *L*

6 **for** C_{Si} in C_S **do**

7 Classify C_{Si} by *C*

8 Label C_{Si} by *R*

9 Calculate δ

10 Label *PN*

11 Assign similar category of **NE**

12 Add C_{Si} to *L*

13 **end for**

14 **until** no new NE is detected

15 **return** *C*, δ

Table 4.3 (*b*)

Clustering Process to Identify and Classify Unlabelled Data from Labelled Data during the Testing Phase.

Algorithm 2 Testing Phase

Input: *U, R, C, δ*

Output: *L*

for Malay NE ∈ *U* **do**
POS of Malay NE
Classify **PN**
    Classify NE by *C*
    Label NE by *R*
    Label PN based on average *δ*
end for
return *L*

Where:

L=Labelled Data, *U*=Unlabelled Data, *R*=Rule, *C*=Classifier, *Δ*=Co-Occurrence Coefficient, And **PN**=Proper Noun.

In the proposed approach, evaluation on rule output is performed after all text data undergoes prior identification of class membership in the NER roster, for this initial concept only the main classes are selected: Person, Location, Organization, and Miscellaneous. To give a brief overview, the reason behind the selection of KNN over other clustering algorithm is in term of its ability to learn from small set of examples (Liu et al., 2011). The approach itself stems from the NN rule schematics, the oldest and most simple method of non-parametric pattern classification. KNN carries the

assumption that any patterns in the vicinity of a feature space are most likely to be classed in the similar class. As in the case of semi-supervised learning method, new data categories are produced from the input of a minute set of examples. In semi-supervised learning concept, new data are assimilated with the older ones (David Nadeau, 2007a).

To elaborate on the benefits of KNN in terms of clustering, among the main potential seen in its implementation is its non-bias ability to treat all training instances equally during the generalization phase, by assignation of equal weight to all training instances available (Liu et al., 2011). The algorithm approach also minimizes the size of the training set. For unannotated text data instances, such as the one used in the research scope, the utilization of KNN is still seen as viable and possible with the execution of semi-supervised learning. The main metric used in the KNN algorithm execution is to measure the distance metric among individual cluster elements.

The entity classification process is known as a regular unsupervised learning problem, as its goal is to attempt to classify unlabelled data into appropriate clusters to obtain equality within the similar clusters and differentiate data from other clusters. As supervised learning is the collaboration between unlabelled and labelled corpus, classifications could also be reviewed via a probabilistic point of view. In the scope of this research, combinations for both training and testing phase algorithms is used as the determining factor for precision and recall rate. As KNN classifier algorithm is chosen as the foundation of this scope's semi-supervised data training approach, the

process of obtaining clusters and data aggregation is divided as follows, due to the fact that NLP research tend to focus on the solution to sentence sequence problems.

Table 4.4 reveals the final adaptation incurred for the KNN algorithm after regex pattern identification of proper nouns had taken place. During this phase, several subcomponents such as Conditional Random Fields and feature vector had been implemented as well. This is to obtain the word to number vectorization for further analysis and other subsequent processes.

Table 4.4

Modified Version of KNN Algorithm Used in Classification

Proposed KNN Classifier Model	
1	Initialize l_s, CRF labeller: $l_s = \text{train}_s\,(\text{ts})$
2	Initialize l_k, KNN classifier: $l_k = \text{train}_k\,(\text{ts})$
3	Initialize n, number of named entities where $n = 0$.
4	**While** n is collected from C_s, $C_s \neq$ null, **do**
5	**for** word (w) $\in n$ **do**
6	Get feature vector $\vec{w}$: $\vec{w} = \text{repr}_{\,w}\,(\text{w, t}) \rightarrow$
7	Classify w with KNN: (c, cf) = knn (l_k, $\vec{w}$)
8	**if** (f > T) **then**
9	Pre-label: t = update (t, w, c)
10	**end if**
11	**end for**

(continued in next page)

Table 4.4 *(Continued)*

Proposed KNN Classifier Model

12 Get feature vector *t*: $t = \text{repr}_t(t, ga)$

13 Label $\vec{t}$ with Conditional Random Field: $(t, cf) = \text{crf}(l_s, \vec{t})$

14 Put labelled result (t, cf) into 0

15 **if** $(f > \vec{y})$ **then**

16 add labelled result *t* to *ts*, $n = n+1$

17 **end if**

18 **if** n > N **then**

19 retrain l_s; $l_s = \text{train}_s(ts)$

20 retrain l_k; $l_k = \text{train}_k(ts)$

21 n = 0

22 **end if**

23 **end while**

24 return 0;

Where:

n = number of proper nouns,

C_s = initial cluster,

w = word,

f = word feature extracted (proper nouns), and

T = current total number of words accumulated.

4.7 Learning Method

This section is going to further discuss the process that had been performed during the data classification progress using regular expression in Chapter 4 (section 4.3.1), and augment it with some initializations of rule-based systems. The end results are further omitted in the subsequent components (Experiment 1, 2, and 3), as the system performance is discussed in detail.

The regular expression text pattern detection in the previous chapter serves as a foundation to test the theory on how a pre-determined text pattern could locate the desired output from a given text structure as planned. The approach itself could be developed to pinpoint certain features that the end system desires to be achieved, such as linguistic feature detection (capitalization) and outlier character analysis (whitespace, punctuation mark elimination). A further testament to text character allocation is on the implementation of learning model and pre-processing methods to augment the discovery of proper nouns aside from visualizing the relationship between word levels in a specific domain.

Rule-based system is seen applicable for corpus that still sees a lower abundance of resource due to the nature of entities that could be correlated more viable among each other. In this process the usage of Decision Tree prunes the word entities into its corresponding group. The side objective in this chapter is to predict the frequencies of the top word stems extracted from the accumulated Malay news

articles. The standard baseline obtained in recent studies sees an average of 50% in the Recall, Precision, and F-Measure values.

To support the research's initial purpose of combining data clustering techniques with regex rule-based approach, experimentations had been done on modification for various parameters to acknowledge the best model with a relatively acceptable tolerance of evaluation values for Precision and Recall after cross-validation on the training dataset. Comparison of different learning models had been done using KNN and Decision Trees to replicate the high average precision and recall rate to improve the accuracy of the learning models. Several pre-processing methods have also been improvised with the custom Malay dataset to further testament the correlation between word similarity and frequency in which they appear. Among the methods applied includes TF-IDF, feature selection, clustering, and term document matrix.

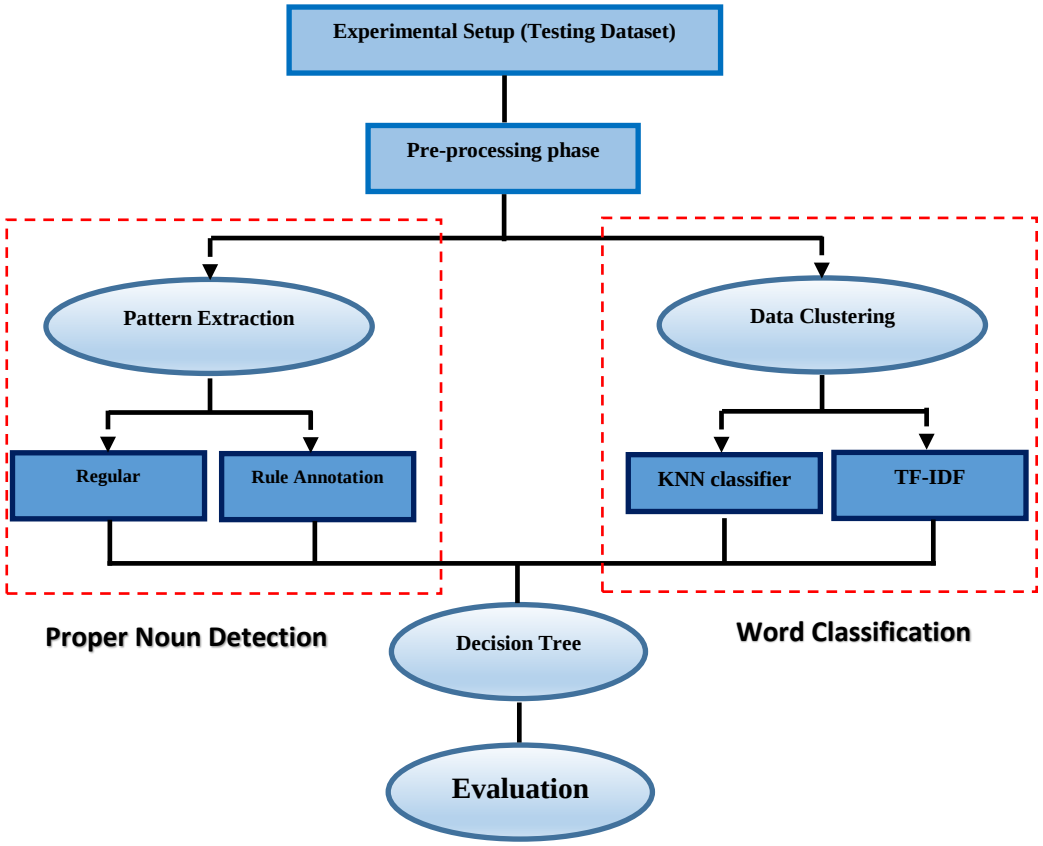


Figure 4.8. An Overview of the Proposed Proper Noun Detection System Process Flow for Malay.

Figure 4.8 above summarizes the entire process conducted for this research into 1 flow, culminating from Chapter 4. As in previous explanations, the research has been carried out by accumulating Malay web articles from the chosen newswires, feed the articles through a regex proper noun detection system, data clustering, and eventually evaluation measures and data visualization. During the course of this process, some approaches may be carried out simultaneously as a proof-of-concept, such as the comparison between the performances of word identification before and after the articles undergo pre-processing. However, all of the reunites at the end during data clustering, where they undergo similar data evaluation & visualization.

The progress of this later research does not always come with the desired end result where there are some unexpected occurrences in the process: most of the techniques suggested to assist the data manipulation fails to support the initial theory. This happened due to the nature of dataset that is highly ambiguous and lack in feature selection after the initial data gleaning is performed. The approaches implemented so far only involves frameworks from Information Retrieval, where future recommendation to overcome this problem is to include more approaches from Information Extraction as the raw Malay dataset is too vague and requires further validation.

4.8 Experiment Results

This stage mainly serves to evaluate the results obtained from the processes conducted earlier, including identification of Malay proper nouns present in the articles using regex pattern, comparisons with the gold data, determining positive and negative values, and initial clustering with KNN algorithm. For this particular stage, the objective emphasizes the results of precision and recall in order to be summed up for F-score.

4.8.1 Extraction of News Articles

The output of all the collected Malay articles is generated and collected in files called **document9** and **document9-propernoun** respectively. These files each undergone pre-processing accordingly.

- **document9.txt**: Contains the 30 final selected news articles that contain the appropriate traits of proper noun categorisation. These traits depend on article's number of words, concise content, formality, and context of topic. This 30-article collection is picked from 60 articles after going through evaluation in terms of content appropriateness, formality of information, timestamps, content location, concise of materials, and redundancy of information transfer.
- **document9-propernoun.txt**: Consisted of all the detected proper nouns from document9. The proper nouns are obtained from the generated process from the developed Python & PHP script. However, the detected items have not undergone any expert validation on its authenticity. This process will be performed later.

4.8.2 Experiment 1: Regex Detection of Proper Nouns

This phase is a continuation from the extraction of news articles, where the text articles collected in document9.txt is passed for proper noun detection in a web app developed in PHP, JavaScript, & jQuery. Mentioned in Chapter 4 section 4.3, the process in which proper noun is detected is from several predefined lists of text

pattern algorithm based on regular expression. Each of the pattern devised would detect certain features present in text character from the news articles, in this scope such as word capitalization, stop word scanning, and punctuation marks elimination. The following tables represent the total number of proper nouns classified after undergoing regex process. Both Table 4.5 (a) and 4.5 (b) contain the results from the experiments done in Chapter 4.

***Note: Gray-out contents is chosen to form training model.**

Table 4.5 (a)

Output of Annotated Named Entity (NE) For the Collected Malay Proper Nouns

Article	Date Collected	Person	Location	Organization	Total
AA01	24/03/2016	2	10	11	**23**
		1	6	1	**8**
AA02	28/03/2016	3	1	4	**8**
		2	2	2	**6**
AA03	01/04/2016	4	1	3	**8**
		2	2	2	**6**
AA04	05/04/2016	3	2	0	**5**
		4	4	4	**12**
AA05	09/04/2016	1	3	5	**9**
		4	1	3	**8**
AA06	13/04/2016	5	3	1	**9**
		9	20	2	**31**
AA07	17/04/2016	2	1	2	**5**
		4	6	1	**11**
AA08	21/04/2016	2	2	3	**7**
		5	3	3	**11**
AA09	25/04/2016	2	4	0	**6**
		2	4	4	**10**
AA10	29/04/2016	5	4	4	**13**
		2	6	2	**10**
BH01	24/03/2016	3	5	4	**12**
		3	4	2	**9**
BH02	28/03/2016	3	3	2	**8**
		8	4	0	**12**

(continued in next page)

Table 4.5 (a) *(Continue)*

Article	Date Collected	Person	Location	Organization	Total
BH03	01/04/2016	3	3	1	**7**
		2	6	8	**16**
BH04	05/04/2016	4	3	3	**10**
		2	3	7	**12**
BH05	09/04/2016	4	4	2	**10**
		9	3	0	**12**
BH06	13/04/2016	3	1	1	**5**
		6	8	3	**17**
BH07	17/04/2016	2	3	0	**5**
		2	5	1	**8**
BH08	21/04/2016	2	8	2	**12**
		4	2	1	**7**
BH09	25/04/2016	2	1	3	**6**
		5	2	2	**9**
BH10	29/04/2016	6	2	1	**9**
		2	3	3	**8**
BER01	24/03/2016	1	3	2	**6**
		5	2	6	**13**
BER02	28/03/2016	3	5	5	**13**
		1	1	2	**4**
BER03	01/04/2016	5	8	2	**15**
		2	4	6	**12**
BER04	05/04/2016	10	6	1	**17**
		3	1	0	**4**
BER05	09/04/2016	3	2	3	**8**
		1	2	1	**4**
BER06	13/04/2016	4	4	4	**12**
		4	13	2	**19**
BER07	17/04/2016	2	10	0	**12**
		2	1	3	**6**
BER08	21/04/2016	5	1	3	**9**
		2	8	3	**13**
BER09	25/04/2016	2	3	3	**8**
		2	9	4	**15**
BER10	29/04/2016	4	10	1	**15**
		5	12	4	**21**

***Note: Gray-out contents is the best-chosen output to form testing model.**

Table 4.5 (b)

Output of annotated Named Entity (NE) for the collected Malay proper nouns

Article	Date Collected	Person	Location	Organization	Total
AA01	**24/03/2016**	2	8	9	**19**
		1	4	1	**6**
AA02	**28/03/2016**	2	1	3	**6**
		1	3	2	**6**
AA03	**01/04/2016**	3	1	3	**7**
		1	2	1	**4**
AA04	**05/04/2016**	2	3	1	**6**
		4	4	6	**14**
AA05	**09/04/2016**	1	3	2	**6**
		6	2	2	**10**
AA06	**13/04/2016**	3	4	2	**9**
		7	19	2	**28**
AA07	**17/04/2016**	1	2	2	**5**
		4	6	2	**12**
AA08	**21/04/2016**	1	1	5	**7**
		3	4	2	**9**
AA09	**25/04/2016**	1	4	0	**5**
		1	4	4	**9**
AA10	**29/04/2016**	4	3	4	**11**
		1	6	3	**10**
BH01	**24/03/2016**	3	3	3	**9**
		2	3	2	**7**
BH02	**28/03/2016**	2	4	3	**9**
		7	4	1	**12**
BH03	**01/04/2016**	2	3	1	**6**
		1	5	9	**15**
BH04	**05/04/2016**	3	3	3	**9**
		2	3	8	**13**
BH05	**09/04/2016**	3	3	2	**8**
		8	4	0	**12**
BH06	**13/04/2016**	2	1	1	**4**
		4	9	4	**17**
BH07	**17/04/2016**	1	3	0	**4**
		1	5	1	**7**
BH08	**21/04/2016**	1	8	2	**11**
		1	2	2	**5**
BH09	**25/04/2016**	1	1	3	**5**
		4	2	3	**9**
BH10	**29/04/2016**	5	3	1	**9**
		1	3	4	**8**
BER01	**24/03/2016**	1	2	3	**6**
		5	2	5	**12**

(continued in next page)

Table 4.5 (b)

Article	Date Collected	Person	Location	Organization	Total
BER02	**28/03/2016**	2	5	6	**13**
		1	4	8	**13**
BER03	**01/04/2016**	4	6	3	**13**
		1	4	6	**11**
BER04	**05/04/2016**	15	11	1	**27**
		3	1	2	**6**
BER05	**09/04/2016**	3	2	1	**6**
		1	3	2	**6**
BER06	**13/04/2016**	4	4	4	**12**
		5	12	3	**20**
BER07	**17/04/2016**	2	8	0	**10**
		2	0	4	**6**
BER08	**21/04/2016**	4	0	3	**7**
		2	8	3	**13**
BER09	**25/04/2016**	2	3	4	**9**
		2	9	5	**16**
BER10	**29/04/2016**	4	11	1	**16**
		3	13	5	**21**

Training model and testing model in this context consists of text data with varied composition ratio, and both consisted of 60 and 30 articles respectively. The difference between both is in the sense of text articles in training model is unprocessed, while testing model had undergo pre-processing steps as stated. The motive to divide this dataset is to review the proficiency in which regex rule could sort out proper nouns appropriately according to the predefined criteria.

4.9 Learning Models

This section highlights the learning models that is executed to meet the requirements of the research scope. The learning model for this is obtained from the testing model from the annotated Malay proper noun obtained using regex detection. Each of the consecutive methods is implemented during clustering process.

4.9.1 Conditional Random Field (CRF)

CRF is considered a statistical modelling approach that is normally utilized in machine learning, basically performing structural assumptions. As contrary to ordinary classifier's approach of assuming labels for an unannotated term ignoring neighbouring values, CRF instead takes this value into account. In Named Entity Recognition, CRF perform a vital role in including both internal and external features from the context, such as prefix and suffix. However, CRF have its negative effects to take a longer time span in vectorising data without annotation. Therefore, in this context CRF is not been implemented in entirety but during the tokenization of unannotated text and attempts to label sequential data.

4.9.2 KNN (K-Nearest Neighbours)

The next list highlights the procedure included for obtaining the performance of clustering.

- KNN algorithm is run on the targeted dataset. This practice is to locate the optimal point for *k*, which would be utilized in the next testing process. This process assumes the 3 sub clusters to be merged into 1 component for processing.
- KNN algorithm is run twice, once during the classification of unannotated proper nouns and once during the generation of decision tree. The first execution is done to produce results for further comparison with the ones produced from after the proper noun list is annotated. The difference between the generated graphs of both data after the KNN algorithm is obtained.
- The processed clusters are validated for its classification via graph representations. In this step, the number of *k* is also obtained to observe the optimal group of clusters in which the text data could possibly be classified.

4.9.3 Experiment 2: Data Clustering

The dataset is constructed in 2 base forms: raw form where no changes are commit as per retrieved from web articles, and the processed form that had been gleaned. From the regex program developed in PHP, an array of proper noun is gathered into a list. For this practice, the test is conducted with the overall 60 articles in order to provide

the general overview of the cluster's condition. Figure 5.3 shows the outcome from the clustering process of the tested dataset.

The process of clustering is done simultaneously in Python's Scipy package, where the text data is accumulated into a list before it is processed through random clustering. The usage of clustering is to enable processing of the data properly into its groupings, regardless of labelling. This practice assumes that the text dataset is clustered into 3 main clustering, indicated by the 3 colours in the graph (blue=AA, yellow=BH, green=BH). The numbering on both x and y-axis represents sigma in which the cluster centre is determined to belong. From this cluster centre, the data expands its functions according to TF-IDF and vector similarities aforementioned.

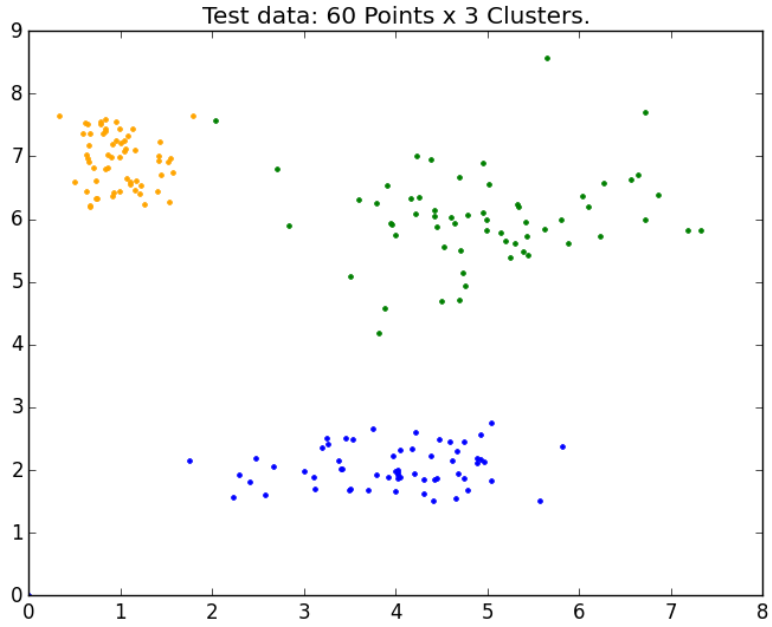


Figure 4.9. Clustering of the Accumulated Text Data of the Three News Datasets.

Figure 4.9 represents the visualization of the data clusters based on the most optimal number of clusters. This process targets the least number of centres as an optimal point of expansion. The red dots in between the 3 clusters indicates the similar points that each of the data constitutes. As the number of centres increases, the tendency of red dots to remain at the centre seems to dissipate away from the centre as well. As seen with AA and BH cluster, the data demonstrates the clustered red dots converged at the centre. With the yellow cluster, red dots seem to move away from gathering at the centre. The value of each data point is represented later in the term document matrix obtained from word to number vectorization done in order to obtain the distance between data points in a single cluster.

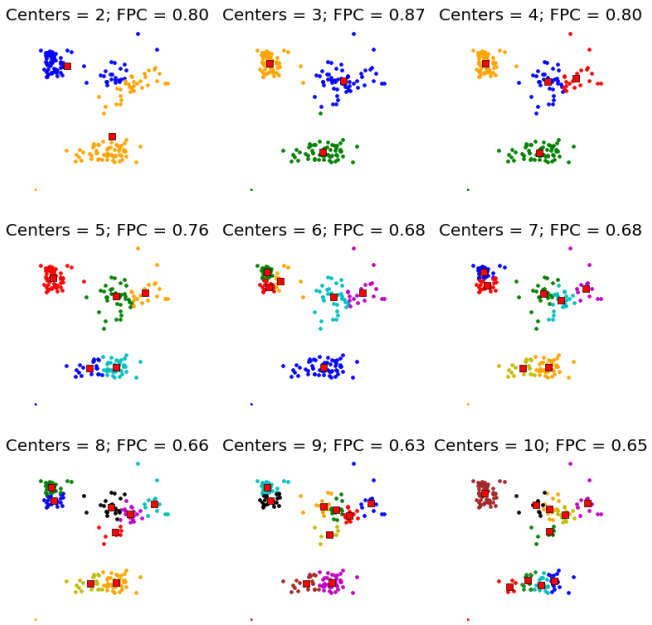


Figure 4.10. KNN Algorithm Defining the Optimal Number of Clusters from the Combined Dataset.

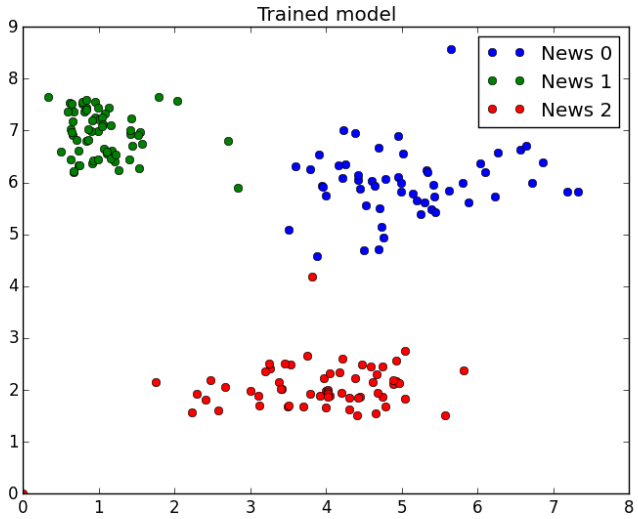


Figure 4.11. Scattered Pattern of the Dataset Constitution Based on The Distance between Objects (TF-IDF)

Figure 4.10 presents the number of clusters obtained from the execution of KNN algorithm. The more the points accumulates at the centre, the less optimal the cluster is. Figure 4.11 further illustrates the correlation between distance of data points and similarity between clusters. Each of the colours represents the respective newswires (News 0=BER, News 1=BH, and News 2=AA). The plot colours represent each of the 3 groupings (red=AA, green=BH, blue=BER, not in order) processed in a separate Python package. Based on the graph above, it is most obvious that BH group contains data with most similarity points, followed by AA and BER consecutively. BH group contains the most clumped points in the group.

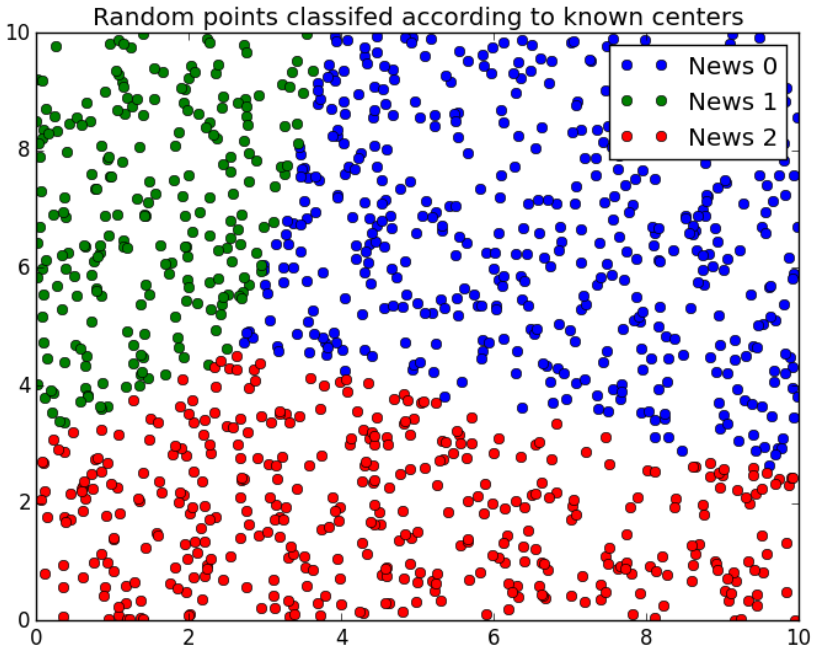


Figure 4.12. Simulation of the Tendency of the Scattered Pattern for the 3 Clusters When the Number of Data is Increased over Time.

Figure 4.12 represents a simulation done by Python library when the number of data points is incremented by option in further processing. This practice is to demonstrate the pattern growth of data points over the course of increase in data quantity.

4.9.4 Decision Tree

In the case of ambiguous dataset such as Malay text that still needs further verification and authentication, producing a high level of precision is desirable should it could be enhanced via the combination with suitable learning model. Considering the pattern rule approach implemented in this project, Decision Tree modelling is selected as the

learning model that could assist in binary classification. Decision Tree determine the rules based on an attribute value and relational operator due to the nature of data row that possess a more generalized concept than a collection of objects that would meet qualification of rule annotation Unsupervised Decision Tree had been discovered to serve well for various forms of data such as text (Balachandran et al., 2012).

Decision Tree depicts text clusters which consisted by rules on word frequencies, by associating the dataset's antecedent and consequent with the root node before dividing them into child nodes based on word frequency conditions recursively. The division attribute is looped as long as the child node is greater than the threshold. The only problem persists when using Decision Tree is the necessity to select appropriate stopping criteria. Due to the small feature space in this experimentation, parameter selection carried over from using KNN algorithm could not help much in improving the results obtained. However, the research decides to use Decision Tree as a preferable method to visualize the correlation between raw, unannotated data.

The objective of using decision tree is to produce a model that predicts the target value based on several other parameters. Each component in a domain is identified as class; where the tree structure is produced from each input feature taken from the available nodes. Each node from the tree is flagged with a class or a probability distribution from the classes. The tree is improvised from dividing the source into subsets based on the attribute obtained from test set. This procedure is replicated on each obtained subset in a recursive order known as recursive partitioning. Within this process, the steps implemented involves as follows.

- The dataset representing the detection of proper noun is annotated in a table form, comma separated values (*.csv*) format where each obtained proper noun is considered as an entity
- The dataset is spread into 2 columns; predictors (type of proper noun) and target (article type)
- Each column header is denoted as predictors
- Entire column is saved as target that would be processed into nodes and leaf of the tree
- Cross validation is implemented

4.10 Data Classification & Further Analysis

The following section explains on the classification approach that has been implemented via learning models and algorithms. Results is also displayed together with brief discussion related to the methods of further data analysis.

4.10.1 TF-IDF (Term Frequency-Inverse Document Frequency)

As briefed in the previous chapter, TF-IDF is applied in word categorisation to perform comparison of word importance in a corpus domain. The technique incorporates Boolean frequency, logarithmic scaled frequency, and augmented frequency to annotate the number of times certain word appears in a document. The

approach that had been implemented in this research includes performing regex to isolate proper noun structure from other linguistic categories, aside from frequency analysis to count number of occurrence for a distinctive word for that article alone.

TF-IDF eliminates the need to count the number of times any particular term appears for a single document, known as *d*. However, this formula takes into consideration the appearance of that particular word across all documents. The idea of TF-IDF execution is: a high score for the word that always occur in a phrase or entire paragraph, and lower score is marked for the word that occurs in many other documents. The formula for TF-IDF is annotated as:

$$\text{tf-idf}\ (t, d, D) = tf(t, d) \cdot idf\ (t, D)$$

Which consists of $t\ f\ (t,\ d)$ that stands for term frequency, and d as the frequency of the term occurring in a particular document. $idf\ (t,\ D)$ represents the inverse document frequency that dictates the amount of information the term represents in the domain. The calculation of *idf* is stated by:

$$idf\ (t, D) = \log \frac{N}{1\ +\ |\{d\ \in D: t\ \in d\}|'}$$

Where N is the total frequency of term in the domain, and $|\{d\ \in D: t\ \in d\}|$ is the number of documents where the term, *t* resides. In the research context, the purpose of selecting TF-IDF is due to its nature of evaluate word distinctiveness. This means that the measure takes into account the presence of common or rare words for

classification. Moreover, this technique is deemed selective towards corpus with low term frequency such as Malay. In this sense, the approach attempts to extract the most relevant percentage of recall and precision by applying TF-IDF than word frequency analysis for the suggested model afterwards.

4.10.2 Feature Selection

In the research work, it had been processed that there are too much features cohabiting an appearance in the word paragraph. These features exist in the form of varied linguistic syntax & morphology orders that indirectly affects the pattern recognizing and data clustering phase. As currently there is no pre-existing methods that could perform assortment on these word features, cross validation had been selected as the evaluation measure to vindicate the dataset training. This is identified as a good probability to make training ambiguous data faster and to prevent overfitting on the output. In the case of a relatively small dataset such as the one devised, cross validation is deemed to provide a good balance for feature selection.

After running the TF-IDF using KNN and K-Means, output for the 3 mini clusters is produced in the form of data matrix. This data matrix is in unilateral form; binary classified integers that generates the value of the similarity and distance between individual terms. Feature selection is applied for further extraction in order to maintain meaningful attributes and eliminate the least useful ones. After performing this action, it could be seen that the rate of precision and recall improved

slightly. The comparison results obtained from this practice is the average values for both precision and recall after applying KNN classifier. Among the features that is highlighted in the course of this investigation is listed in Table 4.6 below.

Table 4.6

Feature Selection in Eliminating Unwanted Character Parentheses

Feature	Example
Part of Speech tags	lengan, telinga
Common words	atau, dan, jadi
Location Suffix	situ, sana, di, ke
Digit	777
Denominator	,.;-=[]””
Common person’s names	Siti, Mohammad, Mohd
Location	Putrajaya, Mozambique
Date	21 Mac
Times	1 petang, 9.30 malam
Telephone numbers	016-1238908
Letters	FACEBOOK
Digits	777, 30000
Uppercase	PUTRAJAYA
Uppercase consisting of initial letters	Menteri Pengangkutan
Lowercase consisting of initial letters	analisis, serpihan
Preposition	dari, daripada, ke, kepada
Combination of preposition + end of sentence	kepada ASTRO AWANI
Verb	selamat, loya
Noun	bangun, rehat
Characters in sentence endings	.:?!]

In order to achieve this task, a few separate programs had been scripted in Python and PHP to obtain the results.

- **SciPy, NumPy:** Mathematical computation and learning module in Python
- **Matplotlib, Seaborn:** Data visualization in Python

- **PyPDF2:** Generating the encoded graphs and decision tree, Python library extension
- **PHP:** Web scraping, regular expression pattern detection, frequency analysis

In the accomplished task, feature selection is applied to the training data and then trained on several models according to the following order:

- A variate feature selection algorithm is implemented in the program, together with the manually skimmed through feature dataset that would generate an average accuracy score for each data point, where the data is stored in the final array of verified values. The data is stored as DecisionTree_Accuracy.csv
- The selected learning models goes through word vectorization, where several parameter values are decided and collected in the process, such as TF-IDF, number of optimal *k*, and similarity distance between terms in the dataset
- Accuracy value is directly generated in this process, however it is not placed into consideration for the generation of the ROC curve
- KNN algorithm is used to validate the distance value obtained prior via data visualization
- Decision Tree list is generated, where final evaluation measures is applied to indicate the final performance values for precision and recall

4.10.3 Term Document Matrix

After undergoing model & feature selection, clustering is also improvised on the accumulate dataset. In the research scope, Malay words are identified by pattern recognition. From the tests conducted, it could be deduced that each term (data points) have instinctively a close correlation between each other depending on the article subject. For example, newsfeeds with political headlines tend to contain more usage of proper nouns that describes high formality in contexture. Proper noun is detected highly in these articles.

However, articles that emphasis on leisure content may persist of least proper nouns that conjectures the information in a sense that is easy to be understood by readers. Hence, proper nouns appear in moderation. These findings had been highlighted in the following result table. For term with identical feature vectors, clustering need to be performed beforehand where different learning model is implemented with each consecutive attempt. Due to the nature of the experimental dataset cluster that is small, each group is processed with KNN algorithm as discussed before.

4.11 Experiment 3: Evaluation Measure and Decision Tree

This phase involves analysis of word frequency, calculation of evaluation measures, and generation of respective decision trees. These processes were performed separately corresponding to each other. The following Table 4.6 shows the proper nouns collected from the three news sites, in the period of 1 month.

Table 4.7

Actual Detection of Proper Nouns and Segregation into its Respective Categories

Article	PER	LOC	ORG	MISC	PREF	SUF	PRE-SUF	INFIX
Current	4	4	2	3	4	0	2	0
National	2	3	0	0	2	0	2	0
World	0	3	1	0	4	2	3	0
World	2	2	0	1	1	0	0	0
World	8	20	3	6	3	0	10	0
National	2	4	1	1	5	0	2	0
Politics	5	5	1	2	2	0	5	0
Politics	1	2	7	2	5	1	5	0
Politics	3	6	2	5	12	4	12	0
Current	1	2	2	0	9	2	15	0
Current	3	4	5	3	11	2	12	1
Current	2	5	1	3	3	0	9	1
Sports	8	4	0	2	5	3	9	0
Finance	1	6	2	3	7	2	16	0
Current	4	9	4	0	13	4	12	0
Sports	8	4	0	1	10	4	9	1
Current	1	8	1	1	6	0	8	0
Politics	2	4	6	2	14	2	13	0
Politics	4	3	2	0	7	1	9	0
Current	5	2	1	2	4	0	7	0
Current	2	6	3	0	10	4	20	0
Current	1	5	6	1	5	3	9	0
World	4	7	2	5	14	3	16	0
Current	12	3	0	4	10	4	17	0
World	4	2	3	6	10	3	13	0
Current	4	11	3	2	9	5	14	0
Finance	2	9	3	1	8	3	10	0
Current	2	11	4	2	4	0	6	0
Politics	4	8	1	4	7	4	9	0
Politics	4	14	3	2	6	3	13	0

As stated in Chapter 3, the clustering stages is done before and after pre-processing in order to compare the convergence pattern between sub-clusters with different frequency of proper nouns. After the Malay proper nouns is classified from news articles, these data are isolated and compared with the gold standard data for this research. The data is identified via 8 major traits of rule annotation stated earlier, mainly PER, LOC, ORG, MISC, PREF, SUF, PRE-SUF, and INFIX based on the major categories for NER. Inference on which category appears the most frequent is performed during the process. PRE-SUF constitutes the highest appearance in the dataset, where INFIX seems to be the least used morphological feature for information relaying. It can be observed that organization names and location titles serves as a vital aspect to deliver information chunks in simple but concise manner among news articles.

4.12 Evaluation Results

This section unveils the result of the experiments carried out in the research, in terms of reviewing the evaluation metrics for the system's performance.

4.12.1 TF-IDF

In order to compare the correlation between applying naïve word count or TF-IDF in selecting features from the dataset, KNN and K-Means had been applied during the clustering process. Each data cluster contains 10 estimators for each method respectively for the 3 clusters (AA, BH, and BERNAMA). The training involves word to number vectorization with Euclidean distance to produce a matrix array that would calculate the distance between individual terms.

4.12.2 Precision, Recall

Table 4.8

Output of Recall and Precision Rates

Evaluation		
Article	Precision	Recall
AA01	0.548	0.676
AA02	0.461	0.75
AA03	0.6	0.191
AA04	0.5	0.473
AA05	0.525	0.815
AA06	0.478	0.393
AA07	0.55	0.647
AA08	0.526	0.5
AA09	0.542	0.394
AA10	0.5	0.438
BH01	0.571	0.6
BH02	0.563	0.643
BH03	0.5	0.571
BH04	0.516	0.8
BH05	0.526	0.434
BH06	0.48	0.667
BH07	0.556	0.333
BH08	0.923	0.522
BH09	0.5	0.63

(Continue)

Table 4.7 *(Continued)*

Evaluation		
Article	Precision	Recall
BH10	0.521	0.857
BER01	0.52	0.867
BER02	0.5	0.178
BER03	0.536	0.455
BER04	0.386	0.361
BER05	0.5	0.444
BER06	0.31	0.214
BER07	0.5	0.406
BER08	0.484	0.6
BER09	0.484	0.441
BER10	0.5	0.618

The overall results are averaged over 3 datasets, as the size is considerably small. It is concluded that domain adaptation is important to bring out a good performance of NER detection. However, the research scope has not delved deep into detailed NER classifications, as the technique tends more to identify the presence of proper noun from an organized article. Consequently, from further analysis done there is a significant misses of proper noun detection and false positives. It could be seen that the value is inconsistent over time.

There are several factors that lead to the low output of the detection system. The annotation scheme tends to be inconsistent; as different news portal relays their contents in mannerism that adapts to the formality of the content itself. The mapping is not absolute, as pattern detection that is implemented adheres to certain rule induction. Terms that is considered identical to the rule induction would be assumed as the correct proper noun. For instance in Astro Awani, ORGANIZATION could exist as facility or company name, while Berita Harian only takes into account abbreviation of group as classification for ORGANIZATION. Different newspaper

corpus exhibits varied styles in composing news articles. Therefore, detections could be accurate at times, or bias depending on the term structure.

Reviewed from other works, the standard baseline for Precision and Recall is always classed over half of the total percentage. As this dataset is designed for use as a frequency analysis of proper noun appearance, the experiment goes through every term that is pre-processed in a uniform manner. Therefore, there might be discrepancies over the actual performance value. Another problem that had been detected is the classification of MISC categories. The dynamic context of this category made actual classification difficult, due to the state of terms that MISC could be associated with. Across the three sub clusters in the dataset, it is seen that MISC sometimes is detected in ORGANIZATION or LOCATION. Only via further validation of gold data could these terms be categorised properly.

Proper nouns appearing in articles are considered distinctive even though the domain is in Malay, making it difficult for systems to tag them correctly. As this research neglects the reliance on conventional tools such as gazette and POS tagging, unless the regex pattern is annotated very specifically to cater the detection of only certain terms, it is difficult to locate proper nouns for the main specified categories.

Noisy data such as unannotated documents need to be pre-processed in order to improve the performance rate so as to avoid omitting important features. For example, in news articles, person names are usually annotated via full name, preceded by a respective title, starts at the sentence beginning, and often capitalized. This

linguistic context is persistent across all news articles. Unprocessed data would suffer in performance if they are excluded from further linguistic analysis, where even cases with annotated data fails. In this case, the dataset is only utilized as data evaluation of proper noun and not for further modelling neither training.

4.12.3 F-Score

To further process the binary classifier values from the document matrix, F-Score is applied to annotate the correlation between precision and recall values for a dataset performance. Mentioned in the previous chapters, F-Score is the harmonic convergence value from the precision, P and the recall, R duo. In order to obtain this value, the calculation takes into account the number of correct positive results that is obtained and the number of positive results that should have been returned. F-Score is the weighted average of the two, and range in between 0 to 1.

$$F - Score = 2 \cdot \frac{P \cdot R}{P + R} \qquad (4.3)$$

4.12.4 Term-Document Matrix

In experimentation, obtaining the most accurate grouping data is considered as the subsequent important task to achieve after evaluation measure implementations. Data clustering assist in providing the user an overview of the actual semantic values encompassed within a certain domain. However in this study, unsupervised method

(KNN algorithm) contains a probability of failing to improvise the data content into its relative partitions. For example, articles that discusses on compounding topics however consists of many similar words may be categorised into the similar domain. This research attempts to incorporate word similarities aside including clustering processes in order to produce the dataset with the best achievable results even if the data is unannotated such as applied.

The approach that had been divided is as follows: a set of unlabelled data is obtained, where the clustering algorithm is executed with the major categories (in this case, the defined categories of NER such as PERSON, ORGANIZATION, LOCATION, and MISC). The matrix is defined from the word to number (Google's Word2vec algorithm) vectorization results obtained earlier. To reiterate, the purpose to vectorise word to numbers is to obtain a binary representation for the text in order to calculate the ratio of word similarity using TF-IDF. The term document matrix results obtained for each cluster structures are illustrated in Figure 4.13 and the result for the true positive, false negative, and F-score for gold data is stated in Table 4.14.

i. Bernama

0	8	2	6	5	1	3	4	7	9	**10**
1	0	8	2	3	5	6	4	7	9	**10**
2	0	4	6	8	5	7	9	1	3	**10**
3	0	6	4	7	2	8	1	5	9	**10**
4	2	8	0	7	3	6	9	5	1	**10**
5	0	8	2	6	1	3	4	9	7	**10**
6	0	2	3	5	8	4	7	9	1	**10**
7	0	4	2	3	8	6	9	5	1	**10**
8	0	4	1	5	2	6	3	7	9	**10**
9	0	2	4	7	6	5	8	3	1	**10**
10	7	6	4	8	1	3	2	9	5	**0**

Figure 4.13. Term document matrix of the Bernama data

Figure 4.13 represents an array of numerical data representations gathered from the vectorization process for the 3 clusters in the dataset, based on Google's Word2Vec Algorithm. The text data is converted into numerical form via CRF before been feed to a decision tree generator to simulate the visualized comparison between the clusters. In order to evaluate the word similarities between entities in a single cluster, this matrix representation calculates the occurrences of 0-10 for a document. Term document matrix takes into consideration the equilibrium of data rows and columns, where certain fluctuations in row values indicate that data possesses high closeness in proximity with each other. However this rule only comes in hand with data clusters that have been structured and carefully tokenized (Ma et al., 2013). From this cluster there are a lot of instances where high fluctuation occur, this indicating the data's correlation with each other. However there are also occurrences in documents that shows low values, which is to be expected from unstructured data.

Table 4.9

Results for True Positive (TP), False Positive (FP), False Negative (FN), and F-Score based on Gold Data Comparison

Article	Evaluation			
	TP	FP	FN	F1
AA01	23	19	11	**0.6052**
AA02	12	14	4	**0.571**
AA03	9	6	38	**0.2898**
AA04	9	9	10	**0.4861**
AA05	31	28	7	**0.6386**
AA06	11	12	17	**0.4313**
AA07	11	9	6	**0.5946**
AA08	10	9	10	**0.5127**
AA09	13	11	20	**0.4411**
AA10	10	10	13	**0.4669**
BH01	12	9	8	**0.5851**
BH02	9	7	5	**0.6003**
BH03	12	12	9	**0.5331**
BH04	16	15	4	**0.6274**
BH05	10	9	13	**0.4756**
BH06	12	13	6	**0.5583**
BH07	10	8	20	**0.4165**
BH08	12	1	11	**0.6669**
BH09	17	17	10	**0.5575**
BH10	12	11	2	**0.648**
BER01	13	12	2	**0.6501**
BER02	13	13	47	**0.2625**
BER03	15	13	18	**0.4922**
BER04	17	27	30	**0.3731**
BER05	12	12	15	**0.4703**
BER06	9	20	33	**0.2532**
BER07	13	13	19	**0.4481**
BER08	15	16	10	**0.5358**
BER09	15	16	19	**0.4615**
BER10	21	21	13	**0.5528**

After data is performed matrix comparisons, the data is further vectorised in order to gain an intrinsic numeric representation over the Decision Tree. During the vectorization process, TF-IDF is implemented as well to produce word similarity index. This is in order to enable data to be correctly represented during the final clustering simultaneously. From the derivation of term document matrix, vectorization process further indicates the degree of similarity between data in that particular domain. Evaluation is done based on the tendency of value increment, or vice versa. From the aforementioned experiment, it could be concluded that different newswire cluster would indicate a varied level of word correlativity, in other words how far the document contains information chunks that is actually persistent with each other. This is shown in the vectorised output from the 3 clusters: AA, BH, and BER.

Table 4.10

Term Document Vectorization for Astro Awani Documents

File	ASTRO AWANI										
AA 01	**1.00000000 e+00**	6.93007858 e-02	1.27260181 e-01	1.13681369 e-01	1.12500050 e-01	1.67314487 e-01	8.89864634 e-02	1.46790878 e-01	1.40791224 e-01	1.39195376 e-01	0.0000000 0e+00
AA 02	6.93007858 e-02	**1.00000000 e+00**	8.30192936 e-02	7.59459264 e-02	6.40467220 e-02	6.74983567 e-02	5.74456751 e-02	7.29880796 e-02	1.09232331 e-01	8.56697376 e-02	0.0000000 0e+00
AA 03	1.27260181 e-01	8.30192936 e-02	**1.00000000 e+00**	9.80074039 e-02	8.00634168 e-02	9.03343073 e-02	5.52673865 e-02	1.02427294 e-01	9.38946784 e-02	8.66262382 e-02	1.0861920 3e-03
AA 04	1.13681369 e-01	7.59459264 e-02	9.80074039 e-02	**1.00000000 e+00**	9.57388054 e-02	1.12163230 e-01	8.69737924 e-02	1.14807135 e-01	1.19347788 e-01	1.14975122 e-01	2.2903624 3e-03
AA 05	1.12500050 e-01	6.40467220 e-02	8.00634168 e-02	9.57388054 e-02	**1.00000000 e+00**	9.59291339 e-02	1.13682713 e-01	1.49586853 e-01	2.01517433 e-01	1.01862759 e-01	9.3300652 1e-04
AA 06	1.67314487 e-01	6.74983567 e-02	9.03343073 e-02	1.12163230 e-01	9.59291339 e-02	**1.00000000 e+00**	7.26131552 e-02	1.25406828 e-01	1.47763716 e-01	1.00007797 e-01	1.3064528 4e-03
AA 07	8.89864634 e-02	5.74456751 e-02	5.52673865 e-02	8.69737924 e-02	1.13682713 e-01	7.26131552 e-02	**1.00000000 e+00**	3.07301513 e-01	1.83332261 e-01	1.65977671 e-01	0.0000000 0e+00
AA 08	1.46790878 e-01	7.29880796 e-02	1.02427294 e-01	1.14807135 e-01	1.49586853 e-01	1.25406828 e-01	3.07301513 e-01	**1.00000000 e+00**	3.22380573 e-01	1.44059436 e-01	0.0000000 0e+00
AA 09	1.40791224 e-01	1.09232331 e-01	9.38946784 e-02	1.19347788 e-01	2.01517433 e-01	1.47763716 e-01	1.83332261 e-01	3.22380573 e-01	**1.00000000 e+00**	1.74228003 e-01	0.0000000 0e+00
AA 10	1.39195376 e-01	8.56697376 e-02	8.66262382 e-02	1.14975122 e-01	1.01862759 e-01	1.00007797 e-01	1.65977671 e-01	1.44059436 e-01	1.74228003 e-01	**1.00000000 e+00**	0.0000000 0e+00

Table 4.11

Term Document Vectorization for Berita Harian Documents

File	BERITA HARIAN										
BH 01	**1.00000000 e+00**	5.84388712 e-02	1.03607140 e-01	7.58051439 e-02	1.38450237 e-01	9.29990610 e-02	6.68218531 e-02	1.12551358 e-01	1.04120160 e-01	7.12036668 e-02	9.81133827e-04
BH 02	5.84388712 e-02	**1.00000000 e+00**	8.48284517 e-02	1.18674621 e-01	8.09746421 e-02	1.34795971 e-01	4.67614956 e-02	7.84232952 e-02	9.69879334 e-02	9.87185178 e-02	8.36966655e-04
BH 03	1.03607140 e-01	8.48284517 e-02	**1.00000000 e+00**	8.99779312 e-02	1.11454627 e-01	6.57616120 e-02	6.96377644 e-02	1.09381690 e-01	1.19479534 e-01	7.35907558 e-02	0.00000000e+00
BH 04	7.58051439 e-02	1.18674621 e-01	8.99779312 e-02	**1.00000000 e+00**	9.77849251 e-02	9.28127979 e-02	1.00152237 e-01	6.41961311 e-02	1.59014677 e-01	5.05430617 e-02	0.00000000e+00
BH 05	1.38450237 e-01	8.09746421 e-02	1.11454627 e-01	9.77849251 e-02	**1.00000000 e+00**	7.42548946 e-02	9.79288105 e-02	1.26502166 e-01	1.23659088 e-01	8.07611917 e-02	0.00000000e+00
BH 06	9.29990610 e-02	1.34795971 e-01	6.57616120 e-02	9.28127979 e-02	7.42548946 e-02	**1.00000000 e+00**	3.84204587 e-02	9.50531227 e-02	9.23333432 e-02	8.31407615 e-02	7.38356310e-04
BH 07	6.68218531 e-02	4.67614956 e-02	6.96377644 e-02	1.00152237 e-01	9.79288105 e-02	3.84204587 e-02	**1.00000000 e+00**	3.75060025 e-02	9.77071092 e-02	8.42640849 e-02	0.00000000e+00
BH 08	1.12551358 e-01	7.84232952 e-02	1.09381690 e-01	6.41961311 e-02	1.26502166 e-01	9.50531227 e-02	3.75060025 e-02	**1.00000000 e+00**	1.31656254 e-01	6.84743080 e-02	0.00000000e+00
BH 09	1.04120160 e-01	9.69879334 e-02	1.19479534 e-01	1.59014677 e-01	1.23659088 e-01	9.23333432 e-02	9.77071092 e-02	1.31656254 e-01	**1.00000000 e+00**	8.01003056 e-02	0.00000000e+00
BH 10	7.12036668 e-02	9.87185178 e-02	7.35907558 e-02	5.05430617 e-02	8.07611917 e-02	8.31407615 e-02	8.42640849 e-02	6.84743080 e-02	8.01003056 e-02	**1.00000000 e+00**	0.00000000e+00

Table 4.12

Term Document Vectorization for Bernama Documents

File	BERNAMA										
BER 01	**1.00000000e+00**	1.80147344e-01	2.24419379e-01	1.77904447e-01	1.68828272e-01	1.91184860e-01	2.15909738e-01	1.60633997e-01	1.60633997e-01	1.46408371e-01	0.00000000e+00
BER 02	1.80147344e-01	**1.00000000e+00**	1.14148657e-01	1.07796273e-01	7.41001119e-02	1.01885705e-01	8.91899159e-02	6.41817213e-02	1.53588578e-01	5.42110307e-02	5.86542653e-04
BER 03	2.24419379e-01	1.14148657e-01	**1.00000000e+00**	1.13822543e-01	2.03610385e-01	1.39123653e-01	1.53609368e-01	1.38229378e-01	1.40377991e-01	1.30854847e-01	4.52398158e-04
BER 04	1.77904447e-01	1.07796273e-01	1.13822543e-01	**1.00000000e+00**	1.30789142e-01	1.01676244e-01	1.31878011e-01	1.22102361e-01	1.09315649e-01	6.52095536e-02	4.90570565e-04
BER 05	1.68828272e-01	7.41001119e-02	7.41001119e-02	1.30789142e-01	**1.00000000e+00**	9.98906835e-02	1.14530386e-01	1.40030240e-01	1.90555956e-01	1.07618144e-01	2.17760741e-03
BER 06	1.91184860e-01	1.01885705e-01	1.39123653e-01	1.01676244e-01	9.98906835e-02	**1.00000000e+00**	1.30740910e-01	9.32865566e-02	1.53000955e-01	9.55273479e-02	0.00000000e+00
BER 07	2.15909738e-01	8.91899159e-02	1.53609368e-01	1.31878011e-01	1.14530386e-01	1.30740910e-01	**1.00000000e+00**	9.73493321e-02	1.79083255e-01	9.57488694e-02	2.75214443e-03
BER 08	1.60633997e-01	6.41817213e-02	1.38229378e-01	1.22102361e-01	1.40030240e-01	9.32865566e-02	9.73493321e-02	**1.00000000e+00**	1.06673463e-01	9.72998371e-02	7.42697044e-03
BER 09	2.45405340e-01	1.53588578e-01	1.40377991e-01	1.09315649e-01	1.90555956e-01	1.53000955e-01	1.79083255e-01	1.06673463e-01	**1.00000000e+00**	7.36912632e-02	5.93301996e-04
BER 10	1.46408371e-01	5.42110307e-02	1.30854847e-01	6.52095536e-02	1.07618144e-01	9.55273479e-02	9.57488694e-02	9.72998371e-02	7.36912632e-02	**1.00000000e+00**	0.00000000e+00

It could be implicated from the output that the three clusters replicate a similar pattern in ascending matrix. Although the huge variance of matrix indices produced shows the terms are sparsely unrelated with each other, upon closer inspection of the article content, this problem only occurs when the term used in the article is used in least appropriate with the information context. Since the desired research objective is to inquire learning models that could provide more stability, TF-IDF method is still considered preferable to provide the most optimal value of matrix contingency.

4.12.5 Decision Tree

In the research process, decision tree structure is generated from continuous numeric variables, as the value in the dataset all consisted of number forms obtained from frequency analysis. This instance is also known as multiple binary features. The reason that this method is implemented lies in the fact of the character handling when they are interpreted into binary variables. During the testing phase, problems existed in the form of value error where within the data table there persisted character values that cannot be instantiated into vectorization. Replacing the strings with a hash code is tested to overcome the problem, however this causes the data to be less credible. Therefore, the creation of data table is done in numerical form to remove subtle inconsistencies for the rule-based method to take effect.

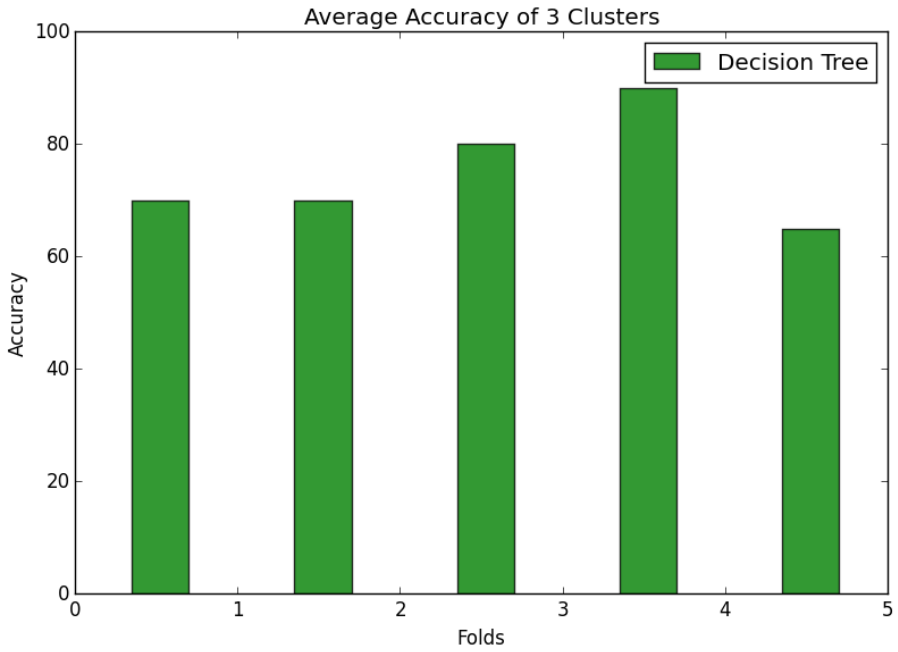


Figure 4.14. Accuracy Value of the Testing Dataset by Implementing 5-Fold Validation

```
Iteration #1

Accuracy Score is 70.0%
[[5 0 0 0]
 [2 2 0 0]
 [0 3 3 0]
 [0 1 0 4]]
             precision    recall  f1-score   support

     health       0.71      1.00      0.83         5
   politics       0.33      0.50      0.40         4
 technology       1.00      0.50      0.67         6
      world       1.00      0.80      0.89         5

avg / total       0.80      0.70      0.71        20
```

```
Iteration #2

Accuracy Score is 70.0%
[[4 0 0 0]
 [1 5 3 1]
 [0 0 2 1]
 [0 0 0 3]]
             precision    recall  f1-score   support

     health       0.80      1.00      0.89         4
   politics       1.00      0.50      0.67        10
 technology       0.40      0.67      0.50         3
      world       0.60      1.00      0.75         3

avg / total       0.81      0.70      0.70        20
```

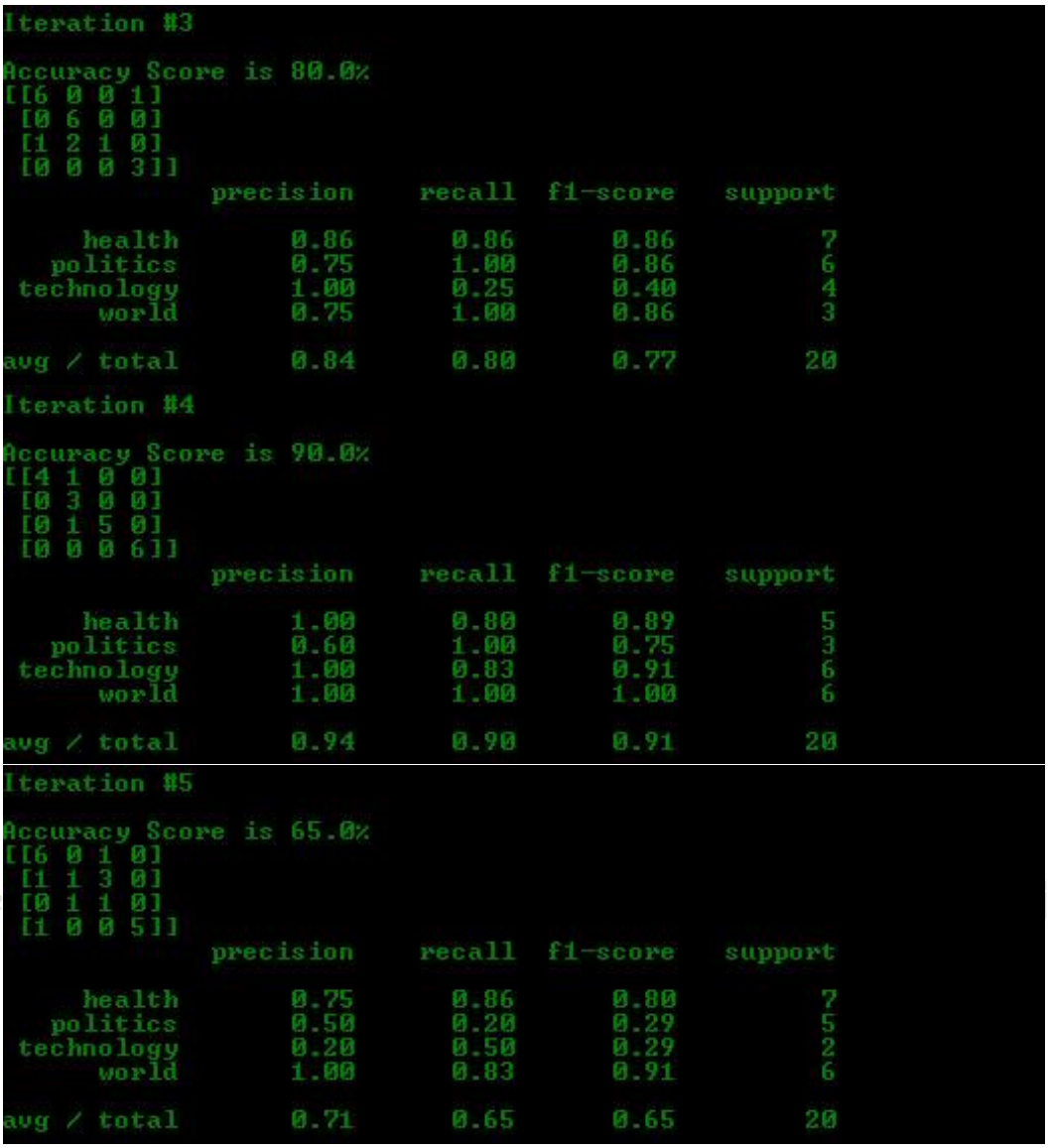


Figure 4.15. Generated Report on the Performance for the 3 Clusters

As seen from the output, the decision tree produced an inconsistent result over several testing. Both the initial and final testing conducted on the character dataset produced irregular results, depending on the pre-processing methods improvised. The initial result expected an organised generated structure of tree traversals after the data organised into tabular form, however it is observed that the data needs to be tagged

and tokenised accordingly to be generated appropriately according to the ruleset. As the dataset had been arranged based on frequency analysis of proper nouns, only numerical forms are emulated. Therefore, the structure of the entire decision tree is represented via number form.

Figure 4.16 illustrates the value of the generated decision tree together with its traversals and nodes. From the implemented table data consisting of predictors and targets, there are 30 articles that is examined and verified with the gold data to be used as the final input of the modelling target.

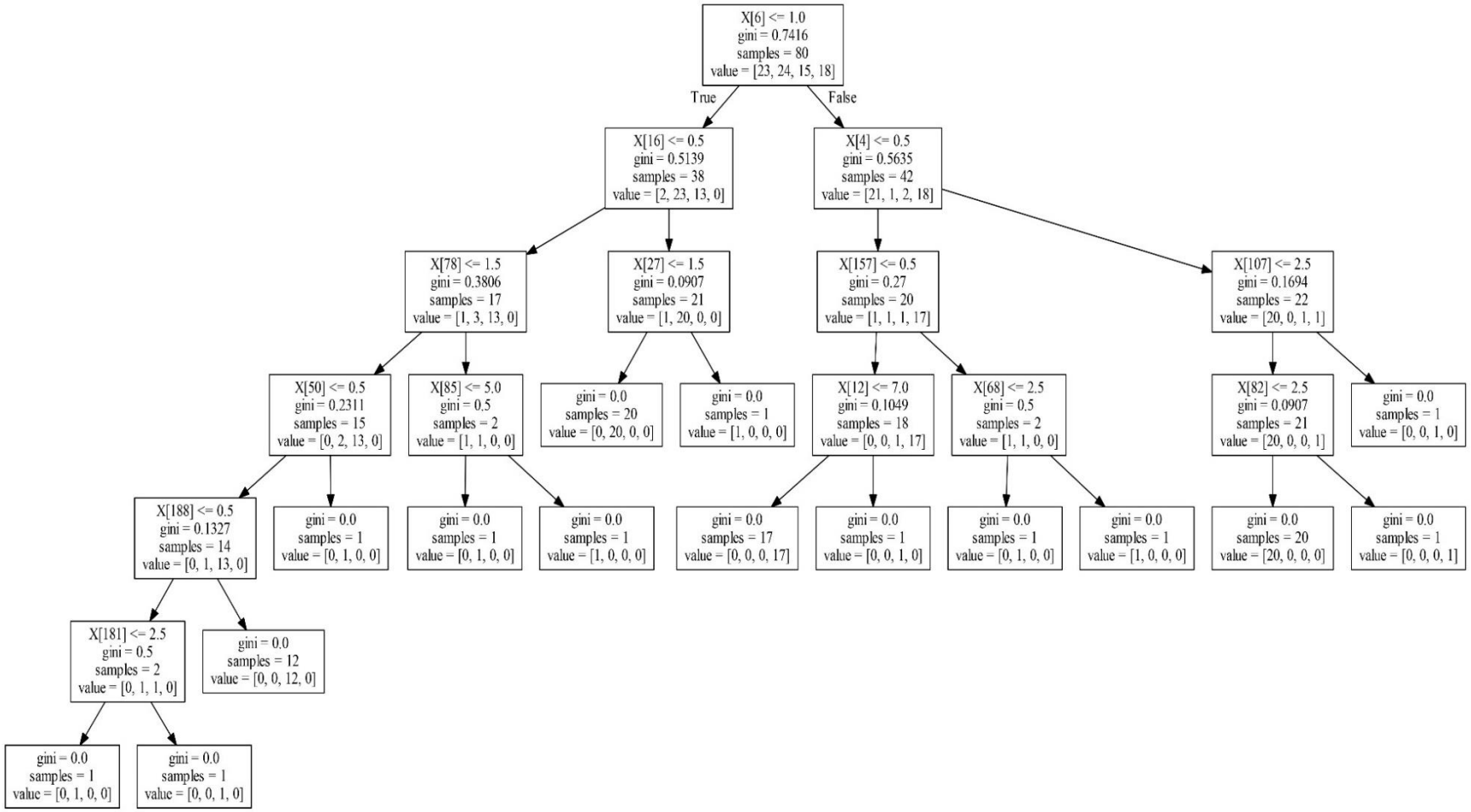


Figure 4.16. Tree Structure for the Dataset

Figure 4.16 serves as a visualization mapping to correspond each of the data elements that resides within a certain domain. The higher the number of branching nodes produced from a tree structure, the more elements is proven related with the main hierarchy. The balance of the entire cluster domain is also reflected on the equality of left and right nodes. The experiments performed from feeding the vectorization seed data to the Decision Tree structure shows that there exist imbalance within the network. Left nodes primarily grow inconsistent with the development of the middle & right nodes, which shows that unstructured data is significant within components that share a bearing resemblance in terms of word similarity, and eventually the node growth would wilt when there is no more similarities to further encapsulate on.

4.12.6 Number of Words

In order to determine the pros and cons of applying word count or TF-IDF for the feature extraction, the identification of total word character present is performed with frequency analysis. A program in PHP is scripted to detect the presence of term occurrence from each article. Afterwards, simple precision and recall calculation is compared between total words that has not been gleaned with those that had been pre-processed. The results are omitted, as the experimental measure produced output that is too vast in comparison. Due to the nature of this experiment that is inferring for learning models with a more preferable stability, the best output value represented

from TF-IDF is chosen. All of the results from this research stems from a total usage of 54000 tokens.

Table 4.13

Word Frequency Analysis on the Total Sentence Structure

Document	Lowercase	Sentence Case	Total Word
AA01	1938	138	2076
AA02	878	69	947
AA03	1063	50	1113
AA04	1130	46	1176
AA05	1457	123	1580
AA06	2013	91	2104
AA07	884	85	969
AA08	1356	88	1444
AA09	1755	126	1881
AA10	2312	49	2361
BH01	1420	105	1525
BH02	809	64	873
BH03	1037	61	1098
BH04	1222	108	1330
BH05	1763	92	1855
BH06	1460	53	1513
BH07	842	55	897
BH08	1853	104	1957
BH09	1091	78	1169
BH10	711	97	808
BER01	6504	212	6716
BER02	1409	105	1514
BER03	2755	133	2888
BER04	2431	165	2596
BER05	1320	111	1431
BER06	3197	132	3329
BER07	2271	123	2394
BER08	1029	96	1125
BER09	1853	161	2014
BER10	1639	161	1800

4.12.7 ROC (Receiver Operating Characteristic) Curve

Construction of a matrix representations produces in turn 4 distinctive variables for each elements constituting a certain dataset, where this elements is assimilated by its actual (positive) or false (negative) value. The harmonic convergence obtained in the form of F-measure is done as to alleviate any possibilities of system performing under optimal states. Bias impacts both Precision and Recall itself, where a system might get a bad F-measure or vice versa when the system's directly affected by how the data is pre-processed. ROC curve is applied to review the trade-offs between True Positive Rate and False Positive Rate obtained from Precision and Recall. Area under ROC curve is desired to represent a better classifier (Masud et al., 2012), where the degree is represented under bell shaped.

True negatives were not being taken into account during these experimental setups due to the criterion of measuring precision and recall that does not require the stating of true negative. Therefore, a complete harmonised value for both positive and negatives could not be relayed in a confusion matrix table which is a prominent feature in IE in the effort to rationalize the relationship between the main variables that is compared in a particular system. The X-axis is dedicated for the false positive rate, whilst Y-axis is constituted by true positive rate. Occasionally, ROC curve graph attempts to correlate the true and false positive values by 0 to 1 point, however with the research scope the actual obtained result was not adjusted into the ratio rate to demonstrate the tendency in value fluctuation or decrease to affect the development of the curve, which would prove the classifier model by the bell shape of the graph.

The detected proper nouns from the experiments performed are still unlabelled. Dispersion and impurity measures are important in building clusters (Masud et al., 2012), however this research implements clustering for proper noun data that had only been classified under each of respective categories (PER, LOC, ORG, MISC, PRE, SUF, PRE-SUF, INFIX). With the case of smaller clusters such as the ones produced, evaluating system performance under a limited amount of labelled data is more realistic in order to establish a baseline record of what degree of extent would the system be able to perform should the data is optimized appropriately; in this context annotated via gold standard and gradual increment of data entry.

ROC curves would give an insight into examining the trade-off for the model to identify correctly positive cases and those of negative cases that were not classified correctly. Figure 4.17-4.19 illustrates the ROC obtained from the experimental setups, where the contingency values are derived from False Positive and True Positive of the testing model.

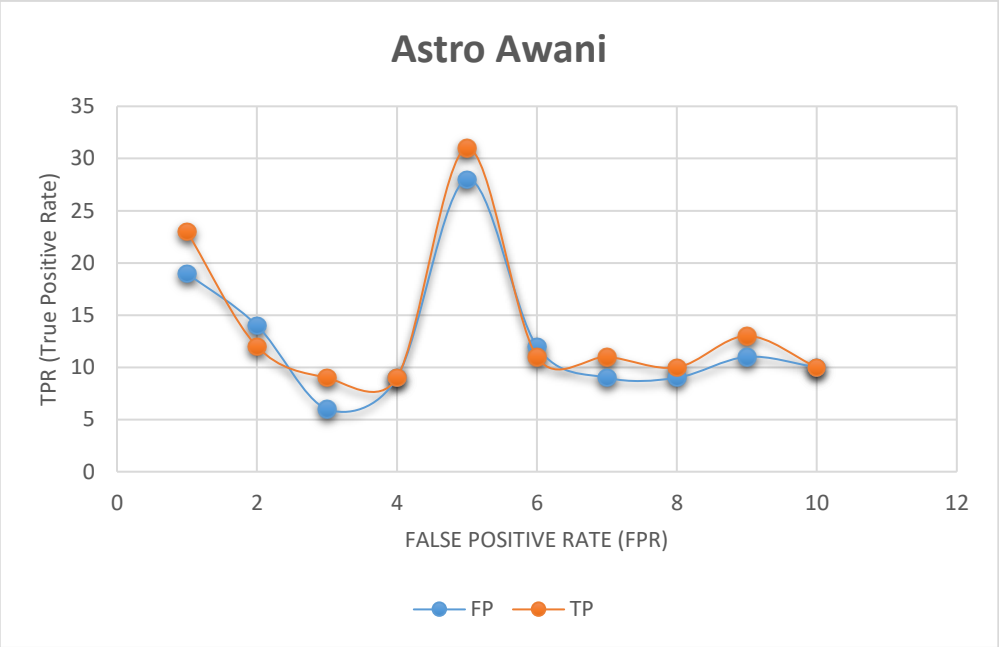


Figure 4.17. Plotted ROC Graph for Astro Awani Data

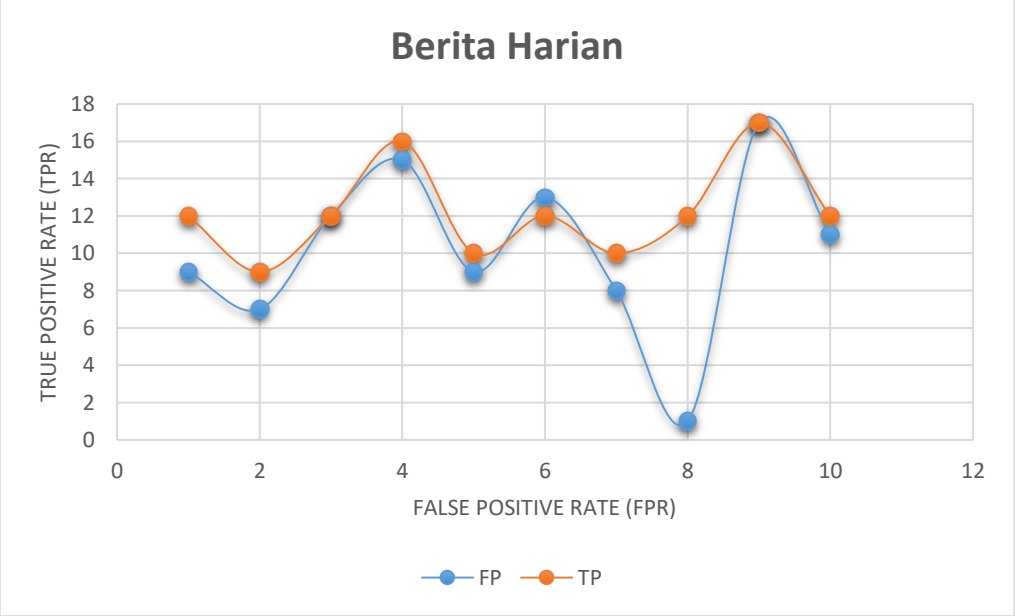


Figure 4.18. Plotted ROC Graph for Berita Harian Data

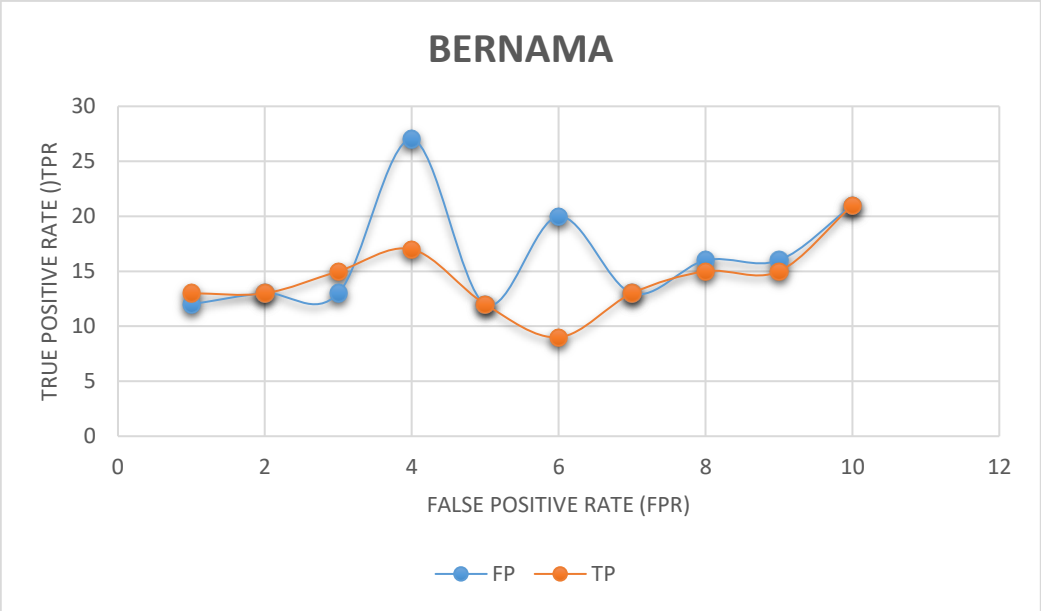


Figure 4.19. Plotted ROC Graph for Bernama Data

The ROC graph does not achieve the optimized bell curve shape as occasionally would be demonstrated in normal classifier models that have its value ratio being converted to percentage decimals. Graphs that applies the converted ratio would display a more consistent shape in the graph generation. However, in the three graph sets for the categories, all results demonstrate a correlation between true and false positives that always be consistent in parallel with each other's position.

This result is obtained even when the named entities are not being clustered to the most optimized position using KNN algorithms as mentioned in the previous chapters. Therefore, it could be hypothesized that data in monocular corpus exist in parallel with accordance with the corpus' growth in size. However, it is safe to assume that the data could not achieve a convergence point when the data cluster is

not being optimized greatly. The graph shape does not achieve bell-shaped as a normal ROC curve does, where it tends to develop in a mountain-like gradient curve.

Figure 4.16 clearly illustrates the growth pattern of named entity identification from the initial experimentation setup after applying regular expression and data clustering, derived from the 3 previous figures above. As shown, the graph demonstrates an inconsistency in the classifier's parameter growth rate over a progression of time. This factor could be related to the classification of the named entity itself, as a more progressive relationship between False Positive and True Positive values would produce a curve that grows steadily upwards until it reaches a climatic point where the growth rate is slower eventually and ends at a certain point in time. All the 3 data clusters from Astro Awani, Berita Harian, and Bernama had illustrated a strikingly similar fluctuation and reduction rate in its True Positive Rate, demonstrated on the y-axis.

4.13 Conclusion

From the techniques applied in gathering evidence of correlation between term frequency occurrence and level of importance a word exudes in an article, it is concluded that the methods listed serves its purpose in determining the existence of proper noun within a domain. The approach introduced induces pattern recognition, data clustering, and term frequency analysis into observing the performance of unannotated documents classification, such as news articles in Malay. Feature

selection and clustering alone is seen as difficult for improving the learning model's evaluation values alone. Further pre-processing methods is required to achieve this task.

In order to produce a more credible learning model, this research implemented ensemble method in the form of Decision Tree to infer the relationship between the predicator (row) and column (target) of the data table. Graphical representations of the data category had been drafted in data table so as to illustrate the output obtained from pre-processing and pattern identification phase. Although the practice had produced a moderate result, it is identified that the performance is widely affected by the factors of term classifications that had been done prior with learning algorithms. Thus, a more credible and basic structure of the data term might replicate a more encouraging result.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

CHAPTER 5

CONCLUSION & FUTURE RECOMMENDATION

5.1 Introduction

This chapter will conclude all the investigation findings and discussions based on the development for proper noun detection system to classify the frequency of Malay proper noun appearance from myriad of techniques. In addition, this chapter contains, suggestions for further improving the research topic, as well as the pros and cons from the proposed theories derived from prior researches.

The research is contrived from the proprietary research for Malay Named Entity Recognition in order to investigate the most beneficial approach that could be performed to detect Malay proper noun using the theories construed in Information Retrieval. Several prominent research is referred as the pioneer guideline to carry out this research more proactively, among those that had been mentioned includes NER-based Malay study done by Rayner. Studies conducted in Data Mining and Text Mining regarding the creation of annotated Malay resources for the purpose of reference of impending information systems had been extensively carried out since the current years. This is to address the lacklustre presence of existing Malay text corpuses. As reviewed in Chapter 2, the investigation of Malay named entity recognition aspect adopts morphological analysis measures from the most closely related linguistic clusters. Examples include English with its morphological structure and Indonesian for its synonymous definition in context.

Research done to overcome lack in availability of Malay data is mostly done in the context of specific specialization fields, or leaning to text pattern identification methods to classify the available text to its most appropriate categories. Embedding into the approach used, the newly acquired data is subjugated into pre-existing ones via combination of training and testing models that would be further evaluated as a measure to validation of the adaptability for the chosen techniques with that particular linguistic structure. During this process, learning techniques are applied altogether to improvise the compatibility of the newly obtained data to constitute the existing collection.

Most of named entity identification for linguistic purpose adopts context-based approach, as distinctive words represent the identity of the linguistic group itself to a certain degree of extent. Context-based approach assimilates new, properly categorised data into its respective groups that are deemed to be most appropriate in terms of precision and accuracy. However, rate of recall is also considered a side-lining element that could assist in determining the extent of quality in which a system could blend in new entries and to function at the best, optimal rate in terms of information retrieval. In particular, the research effort pinpoints the adoptable methods that have utilized named entity detection in a collection of unannotated text via extracting important information from a context portal.

5.2 Further Discussion

From the research flow and methodologies suggested to detect Malay proper nouns in order to be classified into named entities, there are some progressive outputs and undeniably several faults had been detected. Guided by the baseline objective, the end results managed to bring forth several deductions. Below are the main objectives been broken down with analysis on their relevance in the study, and how they affected the research method applied in details.

5.2.1 Objective 1: To Propose Annotated Unstructured Malay Proper Noun Newspaper Corpus That Can Be Used To Classify Named Entity For Malay Text.

The main motivation of the research target is to define and implement a suitable technique to extract proper nouns in order to be associated with the major categories in Named Entity Recognition, among these include Person, Location, Organization, and Miscellaneous. Among the few methods available are tokenization, sentence segmentation, part-of-speech extraction, parsing, Text-to-Speech recognition, and co-reference resolution. These techniques have been implemented in the major linguistic resources such as English and Arabic where there are already supportive repositories and extraction tools made available to comply with specific purposes. However, Malay language had seen a relatively slower growth over the years, with the mentioned techniques had been extensively explored by neighbouring languages, the closest with Malay is the Indonesian. This lack of supportive methodologies made it difficult to improve the lack of abundance for Malay linguistic resource.

Researchers tend to focus on entity extraction where extraction is performed on articles and the pieces is augmented together in order to locate important words that represents major categories such as person, location, and organization names. For the research scope, attempts have been made to perform frequency analysis on the presence of proper nouns in a paragraph via identifying the expression pattern in which they are mostly composed of, and how they constitute a sentence structure. Output from the conducted investigation includes a list of annotated Malay proper noun that is derived from the collected newspaper articles. The research effort omits the major requirement of tokenisation, where regex identification is done by defining

sentence boundary and punctuation character marks at the end of sentence. Initialization of Experiment 1 detects capitalization and counts total number of proper noun frequency serves to prove the theory of every proper noun exists in upper capitalization. From the obtained analysis on word count processing portrayed in Chapter 4, a total of 9581 and 9016 words were obtained before and after processing.

5.2.2 Objective 2: To Develop Proper Noun Detection Method Using Regex Algorithm to Cluster and Classify Word Categories as Named Entity For Malay Text

The utilization of NER techniques supports the word extraction methods as syntactic parsers. However, expensive costs and extensive human efforts are still required in order to properly assimilate the new data and categorise them accordingly. Research work referred in this study emphasized on the importance of proper noun as an agglutination factor that constitutes most of a normal sentence structure. From the conducted research, there had been 4 main regular expression extraction pattern that had been assimilated into the final product. Extraction technique is seen effective should the size of the corpus is in a considerable ratio. Within this research effort, several techniques had been combined to improvise the currently available methodologies in terms of proper noun detection. The emphasized technique in context is regex pattern detection algorithm, along with the inclusion on pre-processing data for outlier elimination, clustering algorithms of text groups into each of respective categories, classification algorithms to isolate the majority groups, and Decision Tree for illustrating data correspondence within a single domain. Results obtained shows regex algorithm could indicate the presence of Malay proper noun

with a high credibility, thus making it an appropriate candidate for extraction tool to detect Malay named entities.

These techniques are selected based on the magnitude of data quantity, along with the purpose of careful annotation and identification of proper nouns into their respective settings. Contrary to major language classes such as English and Chinese who had already nurtured their research techniques and effective tools, this research effort attempts to reduce the reliability on the creation of dictionary and gazetteers, which previous works had indicated vital for language with less availability to have its linguistic elements properly categorised as. Within this scope itself, a supervised algorithm is introduced as a catalyst for data clustering and classification.

Results obtained from this practice shows that for a fixed setting where data size is small and unannotated, the methods implemented can achieve a moderate performance and fulfils its functionalities. A decision analysis tool popular in evaluating newswire, Decision Tree is also introduced in the process. Contrary to the normal types of Decision Tree which reproduces the sentence structure, this experiment converts the text into numerical form first using Google-based Word2Vec style vectorization algorithm so that a number matrix that represents the similarity distance of each cluster items could be visualized.

5.2.3 Objective 3: To Evaluate the Performance of the Proposed Regex Algorithm in Detecting Proper Nouns for Categorisation into Respective Named Entity Classes

All evaluation methods devised to measure the level of system performance serves as a guideline on how well could that system functions under a fixed setting, and how external influence could affect the outputs obtained. In terms of performance indicators, majority of computational linguistics field invoke evaluation measure via observing the value of several key indicators, with this including Precision, Recall, and Accuracy. Each of these parameters represent the elements in which the creation or improvisation of the system attempts to resolve, for example review on Precision values would observe the frequency of matched occurrences, and Recall attempts to examine the degree of extent in which occurrences of new variables could be replicated as compared with the original.

Learning system tend to rely on the value of positive and negatives from structural features annotated on the individual data in which the domain resides. In cases such as the accomplished regex algorithm proper noun detection, determinant factor to evaluate the system performance is to apply the similar measure intended to provide an overview of the system's performance, in terms of precision and recall.

Although accuracy value of 74% is obtained during the generation of Decision Tree, this value serves to give a brief analogy on how well could the system performs when the data size is incremented, or so otherwise. Generation of Decision Tree is done to assign training examples to the same category. Explained in Section 5.2.2,

Word2Vec algorithm is applied to convert and synthesize text structures into numerical form. This process is known as vectorization. As the nature of each obtained articles may seem like they possess lack of similarity with each other, vectorization is done prior the generation of Decision Tree in order to reproduce the tree structure of the antecedents and consequents from the data table into hierarchies of tree nodes that is easier to understand. From the experimented tree structure, the best is chosen to represent the entire dataset.

All of these methods are chosen to suit with the nature of the research undertaken. Although there are several suited methods that could be improvised as well, these methods were chosen based on adaptability and the state of the current research scope. The main motive to be achieved is to carry out an evaluation measure package on the proper noun data to issue a generalized, non-bias assessment on the system's performance on a smaller scale.

5.3 Further Review

Early research in the field of NER is mostly done to discover text patterns that could identify the structure of separate, distinctive word entities in a single domain according to their own usage and definition in context. Due to the absence of a dependable corpus as a gold standard data for further evaluation of obtained data, efforts to prove the presence of a pattern to detect each word entity had been proven futile and not without any difficulties. Past research tend to lean on to concurrent

database to produce an evaluation metric that could differentiate the presence of a match should a named entity is detected from an organized word structure.

As reviewed in Chapter 2, prior research in detecting named entity from certain word cluster relies on approaches conducted by the English language in order to determine the data compatibility. These obtained data chunks were tagged and annotated before its existence as a named entity is further justified. In short, co-occurrence information represents the relation between the test word and its gold standard data (also known as ground truth). Context-based approach had been used to tackle the problem of similarities among text patterns. However, with the linguistic type such as Malay that is still in its premature establishment the methodologies implemented across corpus that is wider in availability is not robust enough to be incorporated entirely in the named entity detection phase. In the research scope, the detection of proper noun is prioritized as the trait to be detected and classified before further categorisation into named entity is performed.

The presence of proper noun in text articles are normally co-referenced as the discovery of a named entity itself. Due to the characteristic of proper noun annotation that usually contains uniform features such as front word capitalization and white space allocation, they are closely associated with named entity. However as proven in prior English NER researches, not all proper nouns are named entities. Although there are aforementioned research experiments that stems toward the classification of named entity by pinpointing towards specific traits within the linguistic domain, such as only acknowledging category-related words or indication of crime proper nouns

from the paragraph, these research only studies a portion of which Malay language has to offer.

In Chapter 3, further elaborations on the proposed concept is discussed further, where several respective theories are reviewed accordingly such as the inception of Information Retrieval techniques and noun detection patterns. These concepts are what that is proposed into the suggested research framework to locate Malay noun structure from the testing domain, in this purpose for news domain. The discussion briefs on the critical aspects that should be included during the classification process. The minute incorporation of human participation and machine learning technique into tracking the noun structure pattern are occasionally composed into a normal article content is discussed accordingly to cater the need towards data categorisation. Based on the popular research pattern currently emerging dedicated to Malay noun architecture, it is seen that most objectives initiated to recognise Malay named entity debates on which technique should be induced to represent the best approach to produce a Malay NER system with high precision and recall rate.

Chapter 4 reviews the framework deposition itself, along with results obtained from the investigation. From the results emerged, there are a few conclusions that could be deduced not only leaning towards the research scope alone, but also to the entire Malay corpus. Many research had reached a similar verdict that the total number of words should not be used as evaluation tool alone to determine the success rate of the assimilation of new data into its older collection, however these presumptions should be judged from the evaluation performance when the data group

had undergone training and classifier in order to obtain the most credible result. The involvement of positive and negative values from the accumulated data should not be underestimated either, as either dominance of true or false values of both positive and negative values of the obtained data would bring imbalance to the final result as shown in the ROC curve on Chapter 4

This study had investigated the detection of proper nouns from 60 Malay news articles containing different composition of words, types, and degree of importance. Subsequent research methods had implemented progressive methods to identify the presence of proper noun from a given text article, among these include regular expression and decision tree approach that is only incorporated in accordance to the research need and not its entirety. From all these approaches, it may be concluded that the application of various information retrieval methods would indeed fluctuate or reduce the efficiency of the system performance, depending on the feature exploited and the motive of data extraction.

5.4 Research Contributions

From the beginning until the conclusion of research, various traits had been highlighted on the importance of overcoming the problems of ambiguity and absence in current Malay text data. Due to its nature, a lot of ongoing research seems to be focused more on categorising intrinsic traits of Malay noun before the result is subjugated into a larger corpus towards the incorporation of a unified corpus. The

final result from the research that is considered to have a major contribution to Malay language in its entirety is listed as follows.

i. To Produce An Annotated Malay Proper Noun Newspaper Corpus

The primary motives for conjuring the detection of proper nouns from online news resources stems from the lack of availability of the said resource in mainstream media, even in the case of national resource itself. The ever improvement of linguistic information everyday requires a constant update of respective materials to be utilized in many aspects of applications, some of these include information systems and archival references. The literature review interpreted in Chapter 2 highlights several efforts done by researchers to categorise the presence of certain word categories in relative domains so as to complement the availability of their current material abundance. Detection of proper nouns in Malay grants a significant contribution in the formation of Malay corpus annotation as each identification already proves the associating feature that each element in a Malay sentence constitutes in. This research effort could be seen as a starting point for endowment of Malay linguistic materials apart from other prominent techniques for Information Retrieval such as post-of-speech tagging and Text-to-Speech. From the researches carried out, a list of Malay proper noun according to the major categories in NER is produced based on the detection using regex algorithm.

ii. Proper Noun Detection Method That Could Be Innovated To Improve The Detection Of Malay Named Entity Recognition

As suggested in Chapter 3, there has been several methods that have been widely acclaimed in order to detect word categories into each of its respective classes, among these involving classification algorithms, learning methods, and word tagging. As Malay language is still at its infancy in terms of providing supportive repositories for further linguistic analysis, other effective methods should be improvised into the identification of noun features. The approach that had been implemented in this research is the regular expression algorithm, an intrinsic way to detect word features based on the breaking down of word via customized pattern identification expressions, along with efficient classification of the potential similar word candidate without discrimination. The introduction of this method could be further encouraged to augment the current system's ability in analysing word features and carefully classifying them accordingly. This in turn could assist in improving and updating the current Malay linguistic resources.

iii. To Evaluate The Performance Of The Proposed Regex Algorithm In Detecting Proper Nouns for Categorisation into Respective Named Entity Classes

In the near future to come, it could be predicted that the establishment of unified Malay corpus data could be realised based on the budding research that have been currently ongoing. During the progression of these researches, it is inevitable that there would be workflows and frameworks which dictates the identification of traits available in the Malay language structure to emerge. In this research alone, it had

been proven that the application of 2 extraction approaches such as pre-processing of unannotated articles and detection of proper nouns using pattern detection analysis could replicate results that corresponds with each other. However, it is proposed that it's better should there is similar tool to extrapolate Malay proper noun into named entities, as seen by work done by Purwarianti for Indonesian NER.

5.5 Recommendations for Future Work

As with the purpose of fulfilling the absence of available Malay text corpus for modern system reference, the research side lines several suggestions that had been considered as a key element for the improvement of the current state after a review of the research's final result. Among these are the most ones that is discussed in detail. This session concludes the research work.

i. Accreditation Of The Real-Time Characteristics For Malay Text Resources

From the experimentation efforts already performed during the research course, it could be deduced that named entity recognizing process needs a credible resource as a gold standard data for guidelines in performance. Learning methods would almost be certain to require a specified amount of seed data in order to help the system digest and regurgitate its new data entry be it supervised, unsupervised, or semi-supervised learning. This would need the actual characteristic of the target resource to be

carefully devised and defined in its entirety. Only after doing this process, the components that is emphasized for identification could be properly detected. Although this measure would increase the rate of detection for the targeted data, to evaluate a corpus's characteristic in detail and particular is not an easy feat to accomplish. To improve the result of the current research, it is preferable that the process of defining a Malay corpus be done simultaneously with the development of the detection tools in the future to improve the current result.

ii. Relationship Between Malay Word Structure, Usage In Context, And Arrangement In Actual Practice To How They Are Classified

Most stages of identifying named entity from raw data involve techniques that have their own significant roles in the whole extraction task. The process in identifying proper noun from random text requires a thorough understanding in how each of these aspects function, be it from retrieving information chunks, pre-processing approaches on the obtained data, or clustering methods according to the data groups in that particular domain. Via comprehension of the appropriate extraction plans applicable on the process, the impact on the system's performance could be decided initially.

This is due to the fact that the selection of techniques deemed appropriate to the detection process could attenuate or improve each of the task's individual results, therefore creating the most optimal result. This research suggests the identification of the inflicting variables and parameters of each applied concepts, whether it is derived from concurrent study fields such as Information Retrieval and Extraction, or by the

deduction of the initial research hypothesis. This identification will enable the proposal of suitable techniques for data extraction and categorisation.

iii. Development of a gold data corpus for reference

As per the main objective of this research to produce seed data that could amplify the existence of current Malay corpus, it is suggested that corresponding future research attempts should be focused on adding the repository of current Malay text data. A unified, credible gold standard corpus could be put into fruition that would provide any corresponding effort with a similarity of data structure in terms of form, density, and structure. This will streamline the data categorisation process into a synchronous setting.

iv. Further Investigation Into The Optimization For A Malay Corpus In Terms Of Precision And Rate Of Recall

One major limitation of the proposed methods is that the rate of detection for proper nouns from newly acquired data is inconsistent with any attempt to introduce new methods. The current method would only produce data closest to consistency when the collected new data undergoes proper pre-processing techniques. As seen during the research course, the results from overall precision and recall directly influence the final harmonic convergence of the F-Measure, thus defining the proficiency of the entire system. More effort is needed to introduce and improvise upon effective techniques already used in Malay text collection that is still under levels of vague and

ambiguity and produce a framework that would detect the most optimal precision and recall rate.

REFERENCE

Abdallah, S., Shaalan, K., & Shoaib, M. (2012). Integrating rule-based system with classification for arabic named entity recognition. In *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 7181 LNCS, pp. 311–322). http://doi.org/10.1007/978-3-642-28604-9_26

AbdelRahman, S., Elarnaoty, M., & Magdy, M. (2010). Integrated Machine Learning Techniques for Arabic Named Entity Recognition. *International Journal of Computer Science*, 7(4), 27–36. Retrieved from http://ijcsi.org/papers/IJCSI-Vol-7-Issue-4-No-3.pdf#page=41

Abdul-hamid, A., & Darwish, K. (2010). Simplified Feature Set for Arabic Named Entity Recognition. *Proceedings of the 2010 Named Entities Workshop, (July)*, 110–115. Retrieved from http://www.aclweb.org/anthology/W10-2417

Abdullah, M., & Ahmad, F. (2009). Rules frequency order stemmer for malay language. … International Journal of …, 9(2), 433–438. Retrieved from http://paper.ijcsns.org/07_book/200902/20090258.pdf

Abedinpourshotorban, H., Hasan, S., Shamsuddin, S. M., & As'Sahra, N. F. (2016). A differential-based harmony search algorithm for the optimization of continuous problems. *Expert Systems with Applications*, 62, 317–332. http://doi.org/10.1016/j.eswa.2016.05.013

Aboaoga, M., & Aziz, M. J. A. (2013). Arabic person names recognition by using a rule based approach. *Journal of Computer Science*, 9(7), 922–927. http://doi.org/10.3844/jcssp.2013.922.927

Abu Bakar, J., Omar, K., Nasrudin, M. F., & Murah, M. Z. (2013). Part-of-Speech for Old Malay Manuscript Corpus: A Review. In *Communications in Computer and Information Science* (Vol. 378 CCIS, pp. 53–66). http://doi.org/10.1007/978-3-642-40567-9_5

Abu Bakar, J., Omar, K., Nasrudin, M. F., Murah, M. Z., Al-shoukry, S., Omar, N., … Klose, A. (2013). Processing natural malay texts: A data-driven approach. *Neurocomputing*, 79(3), 2670–2676. http://doi.org/10.3176/tr.2010.1.06

Agarwal, S. K., Shah, S., & Kumar, R. (2015). Classification of mental tasks from EEG data using backtracking search optimization based neural classifier. *Neurocomputing*, 166, 397–403. http://doi.org/10.1016/j.neucom.2015.03.041

Aggarwal, C., & Zhao, P. (2013). Towards graphical models for text processing. *Knowledge and Information Systems*, 36(1), 1–21. http://doi.org/10.1007/s10115-012-0552-3

Ahmad, Z. H., & Khalifa, O. (2008). Towards designing a high intelligibility rule based standard Malay text-to-speech synthesis system. *Proceedings of the International Conference on Computer and Communication Engineering 2008*, ICCCE08: Global Links for Human Development, 89–94. http://doi.org/10.1109/ICCCE.2008.4580574

Ahmed, Z. (2013). Named Entity Recognition and Question Answering Using Word Vectors and Clustering.

Akbari, R., Hedayatzadeh, R., Ziarati, K., & Hassanizadeh, B. (2012). A multi-objective artificial bee colony algorithm. *Swarm and Evolutionary Computation*, 2, 39–52. http://doi.org/10.1016/j.swevo.2011.08.001

Alfred, R. (2016). Intelligent Information and Database Systems. In ACIIDS 2016, Part II (pp. 447–457). http://doi.org/10.1007/978-3-642-12145-6

Alfred, R., Leong, L. C., On, C. K., & Anthony, P. (2014). Malay Named Entity Recognition Based on Rule-Based Approach. *International Journal of Machine Learning and Computing*, 4(3), 300–306. http://doi.org/10.7763/IJMLC.2014.V4.428

Aljoumaa, H. (2012). Development of a Self-Learning Approach Applied to Pattern Recognition and Fuzzy Control, (September 2012), 127.

Al-Moslmi, T., Gaber, S., Al-Shabi, A., Albared, M., & Omar, N. (2015). Feature Selection Methods Effects on Machine Learning Approaches in Malay Sentiment Analysis, (October), 2–5.

Alshalabi, H., Tiun, S., Omar, N., & Albared, M. (2013). Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization. *International Conference on Electrical Engineering and Informatics (ICEEI 2013)*, 11(Iceei), 748–754. http://doi.org/10.1016/j.protcy.2013.12.254

Al-shammaa, M., & Abbod, M. F. (2015). Automatic Generation of Fuzzy Classification Rules from Data.

Al-shoukry, S., & Omar, N. (2015). Proper Nouns Recognition in Arabic Crime Text Using Machine Learning Approach, 79(3), 506–513.

Althobaiti, M., Kruschwitz, U., & Poesio, M. (2015). Combining Minimally-supervised Methods for Arabic Named Entity Recognition. *Transactions of the Association for Computational Linguistics*, 3, 243–255. Retrieved from https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/564

Althobaiti, M., Kruschwitz, U., & Poesio, M. (2013). A Semi-supervised Learning Approach to Arabic Named Entity Recognition, (September), 32–40. http://doi.org/10.1177/0165551513502417

Althobaiti, M., Kruschwitz, U., & Poesio, M. (2014). Automatic Creation of Arabic Named Entity Annotated Corpus Using Wikipedia. *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 106–115. Retrieved from http://www.aclweb.org/anthology/E14-3012

Ananiadou, S., & McNaught, J. (2006). Text Mining for Biology and Biomedicine. Boston: Artech House.

Ananiadou, S., Pyysalo, S., Tsujii, J., & Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*. http://doi.org/10.1016/j.tibtech.2010.04.005

Ando, R. R. K., & Zhang, T. (2005). A high-performance semi-supervised learning method for text chunking. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (June), 1–9. http://doi.org/10.3115/1219840.1219841

Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1), 4–20. http://doi.org/10.4304/jait.1.1.4-20

Bali, R.-M., Chua, C. C., & Ng, P. K. (2007). Identifying and Classifying Unknown Words In Malay Texts. The Seventh International Symposium on Natural Language Processing

(SNLP2007), 493–498. Retrieved from http://eprints.usm.my/9442/1/Identifying_and_classifying_unknown_words_in_Malay_texts.pdf%5Cnhttp://eprints.usm.my/9442/

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open Information Extraction from the Web. *Proceedings of IJCAI-07, the International Joint Conference on Artificial Intelligence*, 2670–2676. http://doi.org/10.1145/1409360.1409378

Bawane, M. S., & Gadicha, P. V. B. (n.d.). Analysing the result of GRIAS framework by using Precision , Recall and F-measure, 24–30.

Benajiba, Y., Diab, M., & Rosso, P. (2008). Arabic named entity recognition using optimized feature sets. *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (October), 284–293. Retrieved from http://dl.acm.org/citation.cfm?id=1613715.1613755

Benajiba, Y., & Rosso, P. (2008). Arabic Named Entity Recognition using Conditional Random Fields. Proc. of Workshop on HLT & NLP within the Arabic World, LREC. Vol. 8., 143–153. Retrieved from http://www.dsic.upv.es/~prosso/resources/BenajibaRosso_LREC08.pdf

Benajiba, Y., Rosso, P., & BenedíRuiz, J. (2007). ANERsys: an Arabic named entity recognition system based on maximum entropy. Gelbukh, A. (Ed.) CICLing 2007. LNCS, 143–153. Retrieved from http://www.springerlink.com/index/5g6n298843878701.pdf

Bezdek, J. C. (1993). A Physical Interpretation of Fuzzy ISODATA. Readings in Fuzzy Sets for Intelligent Systems, (November), 615–616. http://doi.org/10.1109/TSMC.1976.4309506

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. a, Maynard, D., & Aswani, N. (2013). TwitIE : An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 83–90). Retrieved from https://www.aclweb.org/anthology/R/R13/R13-1011.pdf

Brief, T. (2005). Agreement , the F-Measure , and Reliability in Information Retrieval, 296–298. http://doi.org/10.1197/jamia.M1733.Informatics

Brill, E. (2000). Pattern-based disambiguation for natural language processing. Annual Meeting of the ACL, 1. Retrieved from http://portal.acm.org/citation.cfm?id=1117795

Bsoul, Q., Salim, J., & Zakaria, L. Q. (2013). An Intelligent Document Clustering Approach to Detect Crime Patterns. Procedia Technology, 11(Iceei), 1181–1187. http://doi.org/10.1016/j.protcy.2013.12.311

Cao, T. H., Tang, T. M., & Chau, C. K. (2012). Text Clustering with Named Entities: A Model, Experimentation and Realization. Intelligent Systems Reference Library, 23, 267–287. http://doi.org/10.1007/978-3-642-23166-7_10

Carlson, A., & Betteridge, J. (2010). Coupled semi-supervised learning for information extraction. *Proceedings of the Third ACM International Conference on Web Search and Data Mining (2010)*, 101–110. http://doi.org/10.1145/1718487.1718501

Chapman, C. A. (2016). Usage and refactoring studies of python regular expressions by. Graduate Theses and Dissertations. This, Paper 1513.

Chapman, C., & Stolee, K. T. (2016). Exploring regular expression usage and context in Python. In *Proceedings of the 25th International Symposium on Software Testing and Analysis - ISSTA 2016* (pp. 282–293). http://doi.org/10.1145/2931037.2931073

Chart, G., Algorithm, G., Tun, U., & Onn, H. (2012). Single Disciplinary Project Application Form Fundamental Research Grant Scheme (FRGS), (i), 1–16. http://doi.org/10.1155/2013/782519.(ISI-Q2).

Che, W., Wang, M., Manning, C. D., & Liu, T. (2013). Named Entity Recognition with Bilingual Constraints. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (June), 52–62. Retrieved from http://www.aclweb.org/anthology/N13-1006

Chen, K., Dong, X., Zhu, J., & Shen, B. (2016). Building a domain knowledge base from wikipedia: A semi-supervised approach. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, 2016–Janua. http://doi.org/10.18293/SEKE2016-051

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010). Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, (October)*, 1002–1012. Retrieved from http://portal.acm.org/citation.cfm?id=1870756

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12(Aug), 2493–2537. http://doi.org/10.1145/2347736.2347755

Derczynski, L., Maynard, D., Rizzo, G., & Erp, M. Van. (n.d.). Analysis of Named Entity Recognition and Linking for Tweets, 1–35.

Diab, M. (2009). Second Generation AMIRA Tools for Arabic Processing : Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, 285–288. Retrieved from http://www.elda.org/medar-conference/pdf/56.pdf

Duan, H., Zheng, Y., & Random, C. (2011). A Study on Features of the CRFs-based Chinese. *International Journal of Advanced Intelligence*, 3(2), 287–294.

Dumais, S., & Chen, H. (2000). Hierarchical classification of Web content. *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 256–263. http://doi.org/10.1145/345508.345593

Ek, T., Kirkegaard, C., Jonsson, H., & Nugues, P. (2011). Named entity recognition for short text messages. *Procedia - Social and Behavioral Sciences*, 27(September), 178–187. http://doi.org/10.1016/j.sbspro.2011.10.596

Ekbal, A., & Saha, S. (2011). A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies. *Expert Systems with Applications*, 38(12), 14760–14772. http://doi.org/10.1016/j.eswa.2011.05.004

Ekbal, A., Saha, S., & Sikdar, U. K. (2012). Multiobjective Optimization for Biomedical Named Entity Recognition and Classification. *Procedia Technology*, 6(0), 206–213. http://doi.org/http://dx.doi.org/10.1016/j.protcy.2012.10.025

Elsayed, H., & Elghazaly, T. (2015). A Named Entities Recognition System for Modern Standard Arabic using Rule-Based Approach. *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, 12(1), 51–54. http://doi.org/10.1109/ACLing.2015.14

Elsebai, a, Meziane, F., & Belkredim, F. (2009). A Rule Based Persons Names Arabic Extraction System. *Communications of the IBIMA*, 11(August), 53–59. Retrieved from http://usir.salford.ac.uk/2206/

Elyasir, A. M. H., Sonai, K., & Anbananthen, M. (2013). Comparison between Bag of Words and Word Sense Disambiguation, (Icacsei), 413–417.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S.,… Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1), 91–134. http://doi.org/10.1016/j.artint.2005.03.001

Fadzli, S. A., Norsalehen, A. K., Syarilla, I. A., Hasni, H., & Dhalila, M. S. S. (2012). Simple rules malay stemmer. *The International Conference on Informatics and Applications (ICIA2012)*, 28–35. Retrieved from http://sdiwc.net/digital-library/download.php?id=00000187.pdf

Fuchs, G., Stange, H., Samiei, A., Andrienko, G., & Andrienko, N. (2015). A semi-supervised method for topic extraction from micro postings. *Information Technology*, 57(1), 49–56. http://doi.org/10.1515/itit-2014-1078

Fung, P., Fung, P., Cheung, P., & Cheung, P. (2004). Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. *EMNLP 2004 - Conference on Empirical Methods in Natural Language Processing*, 57–63. Retrieved from http://www.aclweb.org/anthology-new/W/W04/W04-3208.pdf

Gosselin, L., Tye-Gingras, M., & Mathieu-Potvin, F. (2009). Review of utilization of genetic algorithms in heat transfer problems. *International Journal of Heat and Mass Transfer*. Elsevier Ltd. http://doi.org/10.1016/j.ijheatmasstransfer.2008.11.015

Goyvaerts, J., & Levithan, S. (2012). Regular Expressions Cookbook, 612. http://doi.org/9780596802837

Gunawan, Purnama, I. K. E., & Hariadi, M. (2015). Supervised learning Indonesian gloss acquisition. *IAENG International Journal of Computer Science*, 42(4), 337–346.

Hassan, M., Nazlia, O., & Mohd Juzaiddin, A. A. (2015). Malay Part of Speech Tagger : A Comparative Study on Tagging Tools. *Asia-Pacific Journal of Information Technology and Multimedia*, 4(1), 11–23. http://doi.org/10.17576/apjitm-2015-0401-02

Hemmati, M., Amjady, N., & Ehsan, M. (2014). System modeling and optimization for islanded micro-grid using multi-cross learning-based chaotic differential evolution algorithm. *International Journal of Electrical Power and Energy Systems*, 56, 349–360. http://doi.org/10.1016/j.ijepes.2013.11.015

Heydt, M. (2015). Learning pandas: Get to grips with pandas - a versatile and high-performance Python library for data manipulation, analysis, and discovery. Retrieved from http://gen.lib.rus.ec/book/index.php?md5=75566423DC8A5A9411165F24EF9DD886

Hu, B., Tang, B., Chen, Q., & Kang, L. (2016). A novel word embedding learning model using the dissociation between nouns and verbs. *Neurocomputing*, 171, 1108–1117. http://doi.org/10.1016/j.neucom.2015.07.046

Isa, N., Puteh, M., & Kamarudin, R. M. H. R. (2013). Sentiment classification of malay newspaper using immune network (SCIN). *Lecture Notes in Engineering and Computer Science*, 3 LNECS, 1543–1548. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84887882006&partnerID=40&md5=652fdc713458c4dfedcbc4e3a0b736b6

J.M., M. M. U. J. S.-C. S. M. J. G.-B. (2013). Named Entity Recognition: Fallacies challenges and opportunities. Computer Standards and Interfaces, 3554824891(http://www.scopus.com/inward/record.url?eid=2-s2.0-84878302542&partnerID=40&md5=fa0cc4fcfad6db514533c129e08333d6).

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. http://doi.org/10.1016/j.patrec.2009.09.011

Kanagavalli, R. V, & K, R. (2013). Detecting and resolving spatial ambiguity in text using named entity extraction and Self-Learning fuzzy logic techniques. Retrieved from http://arxiv.org/abs/1303.0445

Kantardzic, M. (2011). Data Mining: Concepts, Models, Method, and Algorithms (2nd Edition) (2nd ed.). New Jersey: John Wiley & Sons, Inc.

Khalaf, Z. (2015). MAHIR System: Unsupervised Segmentation for Malay Spoken Broadcast News Stories. *International Journal of Information and Electronics Engineering*, 5(3). http://doi.org/10.7763/IJIEE.2015.V5.532

Kondrak, S. B. and G. (2007). Alignment-Based Discriminative String Similarity. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 656–663.

Kraft, D. H., Martin-Bautista, M. J., Chen, J., & Sanchez, D. (2003). Rules and fuzzy rules in text: Concept, extraction and usage. *International Journal of Approximate Reasoning*, 34(2–3), 145–161. http://doi.org/10.1016/j.ijar.2003.07.005

Král, P. (2014). Named entities as new features for Czech document classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8404 LNCS (PART 2), 417–427. http://doi.org/10.1007/978-3-642-54903-8_35

Kummerfeld, J., & Curran, J. (2008). Classification of Verb-Particle Constructions with the Google Web1T Corpus. *Australasian Language Technology Association Workshop 2008*, 6 (December), 55–63. Retrieved from http://aclweb.org/anthology-new/U/U08/U08-1.pdf#page=114

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 8(June), 282–289. http://doi.org/10.1038/nprot.2006.61

Larasati, S. (2012). Towards an Indonesian-English {SMT} System: A Case Study of an Under-Studied and Under-Resourced Language, Indonesian. {WDS}*'12 Proceedings of Contributed Papers*, 123–129.

Le Nguyen, M., & Shimazu, A. (2014). A semi supervised learning model for mapping sentences to logical forms with ambiguous supervision. In *Data and Knowledge Engineering* (Vol. 90, pp. 1–12). Elsevier B.V. http://doi.org/10.1016/j.datak.2013.12.001

Le, T., Nguyen, K., Nguyen, V., Nguyen, V., & Phung, D. (2016). Scalable Support Vector Machine for Semi-supervised Learning, 1–18. Retrieved from http://arxiv.org/abs/1606.06793

Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., Arbor, A., & Jagadish, H. V. (2008). Regular Expression Learning for Information Extraction. *Conference on Empirical Methods in Natural Language Processing*, (October), 21–30. Retrieved from http://portal.acm.org/citation.cfm?id=1613719

Liao, W., & Veeramachaneni, S. (2009). A simple semi-supervised algorithm for named entity recognition. *Workshop on Semi-Supervised Learning for Natural Language Processing*, (June), 58–65. http://doi.org/10.3115/1621829.1621837

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing Named Entities in Tweets. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1(2008), 359–367. Retrieved from http://acl.eldoc.ub.rug.nl/mirror/P/P11/P11-1037.pdf

Lu, Y., Ji, D., Yao, X., Wei, X., & Liang, X. (2015). CHEMDNER system with mixed conditional random fields and multi-scale word clustering. *Journal of Cheminformatics*, 7(Suppl 1), S4. http://doi.org/10.1186/1758-2946-7-S1-S4

Luis Eduardo, P., Iacobelli, F., & Su, S. (2015). Semi-Supervised Approach to Named Entity Recognition in Spanish Applied to a Real-World Conversational System, 224–235. http://doi.org/10.1007/978-3-319-19264-2

Luo, W., & Yang, F. (2016). An Empirical Study of Automatic Chinese Word Segmentation for Spoken Language Understanding and Named Entity Recognition, 238–248.

Malanyon, D. (2009). Malay Lexical Analysis through Corpus-Based Approach. Eprints.Usm.My. Retrieved from http://eprints.usm.my/10608/

Mangasi, T., Erwin, A., & Ipung, H. P. (2014). Defined entity extraction based on Indonesian text document. In *Proceedings - 2014 International Conference on ICT for Smart Society: "Smart System Platform Development for City and Society, GoeSmart 2014", ICISS 2014* (pp. 61–65). http://doi.org/10.1109/ICTSS.2014.7013152

Manning, C. D., & Raghavan, P. (2009). An Introduction to Information Retrieval. Online, 1, 1. http://doi.org/10.1109/LPT.2009.2020494

Markov, Z., & Larose, D. T. (2007). Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. John Wiley & Sons, Inc.

Mikolov, T., Le, Q. V, & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. arXiv Preprint arXiv:1309.4168v1, 1–10. Retrieved from http://arxiv.org/abs/1309.4168v1%5Cnhttp://arxiv.org/abs/1309.4168

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 1st ed. Elsevier. Oklahoma: Academic Press. http://doi.org/10.1016/B978-0-12-386979-1.00009-8

Mohamed, H., Omar, N., & Ab. Aziz, M. J. (2015). Malay Part of Speech Tagger: A Comparative Study on Tagging Tools. *Asia-Pacific Journal of Information Technology and Multimedia*, 4(1), 11–23. http://doi.org/10.17576/apjitm-2015-0401-02

Mohd Don, Z. (2010). Processing natural malay texts: A data-driven approach. Trames, 14(1), 90–103. http://doi.org/10.3176/tr.2010.1.06

Mohit, B., Schneider, N., Bhowmick, R., Oflazer, K., & Smith, N. a. (2012). Recall-oriented learning of named entities in Arabic Wikipedia. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 162–173. Retrieved from http://dl.acm.org/citation.cfm?id=2380816.2380839

Nadeau, D. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 8(30), 3–26. http://doi.org/10.1075/li.30.1.03nad

Nogueira, T. M., Rezende, S. O., & Camargo, H. a. (2010). On the use of fuzzy rules to text document classification. *Hybrid Intelligent Systems (HIS), 2010 10th International Conference on*, 19–24. http://doi.org/10.1109/HIS.2010.5600076

Noh, N., Rusydi, M., Talib, A., Ahmad, A., Halim, S. A., & Mohamed, A. (2009). Malay Language Document Identification Using BPNN. In *Proceedings of the 10th WSEAS international conference on Neural networks* (pp. 163–168).

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from Wikipedia. Sydney: Elsevier Science. http://doi.org/10.1016/j.artint.2012.03.006

Ojo, A., & Adeyemo, A. B. (2012). Framework for Knowledge Discovery from Journal Articles Using Text Mining Techniques. *African Journal of Computing & ICT*, 5(2), 35–44. Retrieved from http://www.ajocict.net/uploads/Pre-print_-_O__Ojo___A_B__Adeyemo__2012___Framework_for_Knowledge_Discovery_from_Journal_Articles_Using_Text_Mining_Techniques.pdf

Oudah, M., & Shaalan, K. (2012). A Pipeline Arabic Named Entity Recognition using a Hybrid Approach. *COLING* (December 2012), 2159–2176. Retrieved from http://www.newdesign.aclweb.org/anthology/C/C12/C12-1132.pdf

Oudah, M., & Shaalan, K. (2016). Studying the impact of language-independent and language-specific features on hybrid Arabic Person name recognition. *Language Resources and Evaluation*, 1–28. http://doi.org/10.1007/s10579-016-9376-1

Petrov, S., Das, D., & McDonald, R. (2011). A Universal Part-of-Speech Tagset. Retrieved from http://arxiv.org/abs/1104.2086

Pham, Q. H., Nguyen, M.-L., Nguyen, B. T., & Cuong, N. V. (2015). Semi-supervised Learning for Vietnamese Named Entity Recognition using Online Conditional Random Fields. In *Proceedings of the Fifth Named Entity Workshop* (pp. 50–55). Retrieved from http://www.aclweb.org/anthology/W15-3907

POWERS, D.M.W. (AILab, School of Computer Science, Engineering and Mathematics, Flinders University, South Australia, A. (2011). Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63. http://doi.org/10.1.1.214.9232

Powers, D. M. W. (2015). What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes, 19. http://doi.org/KIT-14-001

Prasad, G., Fousiya, K. K., Kumar, M. A., & Soman, K. P. (2015). Named Entity Recognition for Malayalam Language : A CRF based Approach, (May), 16–19.

Ramli, I., Jamil, N., Seman, N., & Ardi, N. (2015). An Improved Syllabification for a Better Malay Language Text-to-Speech Synthesis (TTS). *2015 IEEE International Symposium On*

Robotics and Intelligent Sensors, 76 (Iris), 417–424. http://doi.org/10.1016/j.procs.2015.12.280

Rao, R. V., & Saroj, A. (2017). A self-adaptive multi-population based Jaya algorithm for engineering optimization. Swarm and Evolutionary Computation, (October 2016), 1–26. http://doi.org/10.1016/j.swevo.2017.04.008

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named Entity Recognition in Tweets: An Experimental Study. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1524–1534. Retrieved from http://dl.acm.org/citation.cfm?id=2145595

Rosso, P., Benajiba, Y., & Lyhyaoui, A. (2006, December). Towards an Arabic question answering system. In *Proc. 4th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROIV, Damascus, Syria* (pp. 11-14).

Rozenfeld, B., & Feldman, R. (2008). Self-supervised relation extraction from the Web. *Knowledge and Information Systems*, 17(1), 17–33. http://doi.org/10.1007/s10115-007-0110-6

Sam, R. C., Le, H. T., Nguyen, T. T., & Nguyen, T. H. (2011). Combining proper name-coreference with conditional random fields for semi-supervised named entity recognition in Vietnamese text. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6634 LNAI (PART 1), 512–524. http://doi.org/10.1007/978-3-642-20841-6-42

Samat, N. A., Murad, M. A. A., Abdullah, M. T., & Atan, R. (2005). Malay Documents Clustering Algorithm Based on Singular Value Decomposition. *Journal of Theoretical and Applied Information Technology*, 180–186.

Sari, Y., Hassan, M. F., & Zamin, N. (2009). A Hybrid Approach to Semi-supervised Named Entity Recognition in Health, Safety and Environment Reports. *2009 International Conference on Future Computer and Communication*, 599–602. http://doi.org/10.1109/ICFCC.2009.52

Sari, Y., Hassan, M. F., & Zamin, N. (2010). Rule-based pattern extractor and Named Entity Recognition: A hybrid approach. In *Proceedings 2010 International Symposium on Information Technology - Engineering Technology, ITSim'10* (Vol. 2, pp. 563–568). http://doi.org/10.1109/ITSIM.2010.5561392

Satoshi Sekine, K. S., & Nobata, C. (2002). Extended named entity hierarchy. *Third International Conference on Language Resources and Evaluation (LREC 2002)*, 1818–1824.

Sazali, S. S., Rahman, N. A., & Bakar, Z. A. (2017). Information extraction: Evaluating named entity recognition from classical Malay documents. In *2016 3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016 - Conference Proceedings* (pp. 48–53). http://doi.org/10.1109/INFRKM.2016.7806333

Seeger, M., & King, I. (2002). Learning from labeled and unlabeled data. Learning, (January), 1–62. http://doi.org/10.1109/IJCNN.2002.1007592

Sekine, S., Sudo, K., & Nobata, C. (2002, May). Extended Named Entity Hierarchy. In *LREC*.

Selvaperumal, P., & Suruliandi, A. (2016). Semi-Supervised Personal Name Disambiguation Technique for the Web. *International Journal of Modern Education and Computer Science*, 8(3), 28–36. http://doi.org/10.5815/ijmecs.2016.03.04

Servan, C., Berard, A., Elloumi, Z., Blanchon, H., & Besacier, L. (2016). Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources? Retrieved from http://arxiv.org/abs/1610.01291

Shaalan, K., & Oudah, M. (2013). A hybrid approach to Arabic named entity recognition. *Journal of Information Science*, 40(1), 67–87. http://doi.org/10.1177/0165551513502417

Shaalan, K., & Raza, H. (2007). Person Name Entity Recognition for Arabic. *Computational Linguistics*, (June), 17–24. http://doi.org/10.3115/1654576.1654581

Shabat, H. (2015). Named Entity Recognition in Crime News Documents Using Classifiers Combination, 23(6), 1215–1222. http://doi.org/10.5829/idosi.mejsr.2015.23.06.22271

Sharma, D., Devale, P. R., & Khare, A. K. (2011). Approach for Multiword Expression Identification in Natural Language Processing, 2 (August 2011), 663–666.

Sidi. (2011). Malay Interrogative Knowledge Corpus. American Journal of Economics and Business Administration, 3, 171–176. http://doi.org/10.3844/ajebasp.2011.171.176

Sinoara, R. A., Sundermann, C. V., Marcacini, R. M., Domingues, M. A., & Rezende, S. O. (2014). Named entities as privileged information for hierarchical text clustering. *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS '14*, 57–66. http://doi.org/10.1145/2628194.2628225

Srivastava, A. N., & Sahami, M. (2009). Text Mining: Classification, Clustering, and Applications. Boca Raton: Chapman and Hall/CRC.

Suakkaphong, N., Zhang, Z., & Chen, H. (2013). Disease Named Entity Recognition Using Semisupervised Learning and Conditional Random Fields. *Journal of the American Society for Information Science and Technology*, 14(4), 90–103. http://doi.org/10.1002/asi

Sun, a, Grishman, R., & Sekine, S. (2011). Semi-supervised relation extraction with large-scale word clustering. *Proceedings of the 49th Annual Meeting …*, 521–529. Retrieved from http://www.aaai.org/Papers/AAAI/2007/AAAI07-224.pdf%5Cnhttp://dl.acm.org/citation.cfm?id=2002539

Suwarningsih, W., Supriana, I., & Purwarianti, A. (2015). ImNER Indonesian medical named entity recognition. In *Proceedings of 2014 2nd International Conference on Technology, Informatics, Management, Engineering and Environment, TIME-E 2014* (pp. 184–188). http://doi.org/10.1109/TIME-E.2014.7011615

Tabuchi, N., Sumii, E., & Yonezawa, A. (2003). Regular expression types for strings in a text processing language. *Electronic Notes in Theoretical Computer Science*, 75, 97–115. http://doi.org/10.1016/S1571-0661 (04)80781-3

Tan, T. P., Xiao, X., Tang, E. K., Chng, E. S., & Li, H. (2009). MASS: A Malay language LVCSR corpus resource. *2009 Oriental COCOSDA International Conference on Speech Database and Assessments, ICSDA 2009*, 25–30. http://doi.org/10.1109/ICSDA.2009.5278382

Tran, V. C., Hwang, D., & Jung, J. J. (2015). Semi-supervised Approach Based on Co-occurrence Coefficient for Named Entity Recognition on Twitter, 141–146.

Triguero, I., García, S., & Herrera, F. (2013). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, pp. 1–40. http://doi.org/10.1007/s10115-013-0706-y

Triguero, I., Sáez, J. A., Luengo, J., García, S., & Herrera, F. (2014). On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. *Neurocomputing*, 132, 30–41. http://doi.org/10.1016/j.neucom.2013.05.055

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. In *Procedia Engineering* (Vol. 69, pp. 1356–1364). Elsevier B.V. http://doi.org/10.1016/j.proeng.2014.03.129

Tuffery, S. (2011). Data Mining and Statistics for Decision Making. Wiley.

Turian, J., Ratinov, L., Bengio, Y., & Turian, J. (2010). Word Representations: A Simple and General Method for Semi-supervised Learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, *(July)*, 384–394. http://doi.org/10.1.1.301.5840

Wibawa, A. S., & Purwarianti, A. (2016). Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning. *Procedia Computer Science*, 81(May), 221–228. http://doi.org/10.1016/j.procs.2016.04.053

Witten, I. H., Frank, E., & Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques (2nd ed.). http://doi.org/citeulike-article-id:8827086

Worden, K., Staszewski, W. J., & Hensman, J. J. (2011). Natural computing for mechanical systems research: A tutorial overview. *Mechanical Systems and Signal Processing*. Elsevier. http://doi.org/10.1016/j.ymssp.2010.07.013

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H.,Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems* (Vol. 14). http://doi.org/10.1007/s10115-007-0114-2

Xian, B. C. M., Lubani, M., Ping, L. K., Bouzekri, K., Mahmud, R., & Lukose, D. (2016). Benchmarking Mi-POS: Malay Part-of-Speech Tagger. *International Journal of Knowledge Engineering*, 2(3), 115–121. http://doi.org/10.18178/ijke.2016.2.3.064

Yang, F., & Vozila, P. (2014). Semi-Supervised Chinese Word Segmentation Using Partial-Label Learning With Conditional Random Fields. Emnlp, 90–98. Retrieved from http://emnlp2014.org/papers/pdf/EMNLP2014010.pdf

Yesilbudak, M., Sagiroglu, S., & Colak, I. (2017). A novel implementation of kNN classifier based on multi-tupled meteorological input data for wind power prediction. *Energy Conversion and Management*, 135, 434–444. http://doi.org/10.1016/j.enconman.2016.12.094

Yong, S.-F., Ranaivo-Malançon, B., & Wee, A. Y. (2011). NERSIL : the named-entity recognition system for Iban language. *25th Pacific Asia Conference on Language, Information and Computation*, 549–558.

Yong, Z., Youwen, L., & Shixiong, X. (2009). An Improved KNN Text Classification Algorithm Based on Clustering. Journal of Computers, 4(3), 230–237. http://doi.org/10.4304/jcp.4.3.230-237

Zamin, N., & Oxley, A. (2011). Building a Corpus-Derived Gazetteer for Named Entity Recognition, 73–80.

Zamin, N., Oxley, A., Abu Bakar, Z., & Farhan, S. A. (2012). A statistical dictionary-based word alignment algorithm: An unsupervised approach. In *2012 International Conference on Computer and Information Science, ICCIS 2012 - A Conference of World Engineering,*

Science and Technology Congress, ESTCON 2012 - Conference Proceedings (Vol. 1, pp. 396–402). http://doi.org/10.1109/ICCISci.2012.6297278

Zatarain Salazar, J., Reed, P. M., Herman, J. D., Giuliani, M., & Castelletti, A. (2016). A diagnostic assessment of evolutionary algorithms for multi-objective surface water reservoir control. *Advances in Water Resources*, 92, 172–185. http://doi.org/10.1016/j.advwatres.2016.04.006

Zeng, H., Song, A., & Cheung, Y. M. (2013). Improving clustering with pairwise constraints: A discriminative approach. *Knowledge and Information Systems*, 36(2), 489–515. http://doi.org/10.1007/s10115-012-0592-8

Zhan, Q. (2017). An Improved K-means Algorithm Based on Structure Features, 12(1), 62–80. http://doi.org/10.17706/jsw.12.1.62-81

Zhang, C., Hong, X., & Peng, Z. (2012). An automatic approach to harvesting temporal knowledge of entity relationships. In *Procedia Engineering* (Vol. 29, pp. 1399–1409). http://doi.org/10.1016/j.proeng.2012.01.147

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6), 1088–1098. http://doi.org/10.1016/j.jbi.2013.08.004

Zhou, D., & Zhong, D. (2015). A semi-supervised learning framework for biomedical event extraction based on hidden topics. *Artificial Intelligence in Medicine*, 64(1), 51–58. http://doi.org/10.1016/j.artmed.2015.03.004

Zirikly, A., & Diab, M. (2015). Named Entity Recognition for Arabic Social Media. *Proceedings of NAACL-HLT 2015*, 176–185. Retrieved from http://www.aclweb.org/anthology/W15-1524.pdf