Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

FEATURE EXTRACTION USING REGULAR EXPRESSION IN DETECTING PROPER NOUN FOR MALAY NEWS ARTICLES BASED ON KNN ALGORITHM

S. Sulaiman[2], R. A. Wahid[2] and F. Morsidi[1,*]

[1]Faculty of Art, Computing and Creative Industry, Universiti Pendidikan Sultan Idris, 35900
Tanjong Malim, Perak, Malaysia
[2]Computing Department, Faculty of Art, Computing and Creative Industry, Universiti
Pendidikan Sultan Idris, 35900Tanjong Malim, Perak, Malaysia

Published online: 17 October 2017

ABSTRACT

The identification of proper nouns from text aims to classify named entities according to their respective groupings, an aspect included in Named Entity Recognition (NER). Proper noun disambiguation can adversely affect morphological analysis, a vital trait to improve the corpus availability via classification and new word assimilation. The occurrences of proper nouns can be annotated from the text resources using separate entity mapping from their fragments. The research was carried out to examine the impact of regex on text pattern identification sequence that queried and acquired proper nouns from a collection of unannotated Malay language news articles that envisions several techniques to improve text entities precision and accuracy such as pre-processing and data clustering. The results showed that the F-scores of the output tested on the unannotated news dataset were between 30% and 60%.

Keywords: data mining; named entity recognition; regular expression; natural language processing

Author Correspondence, e-mail: farid.mors90@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v9i5s.16

Journal of Fundamental and Applied Sciences is licensed under a Creative Commons Attribution-NonCommercial 4.0

1. INTRODUCTION

The enormous scope of the field of text mining is due to the unlimited article resources available on the web. After images, textual materials represent the second highest raw information chunk that is open for access on the Internet, which complicates the process of seeking and identifying relevant information. Thus, text mining has been proposed to help manage such tasks. In fact, text mining is closely related with other techniques such as data mining and natural language processing [1].

In essence, all these techniques share some common traits in terms of analyzing the structures of a language. Analyzing the linguistic elements of a written language has been made more efficient by the Information Extraction (IE) technique [2], which extracts the language's beneficial features and then organizes them into proper partitions such that the language can be better understood. Thus, this technique can help improve the number of available resources for other fields of research, notably in information systems and linguistic studies.

In English language, the proper noun represents a unique language feature that is morphologically expressed. Invariably, such proper nouns appear abundantly and frequently in many articles published on the Web. Essentially, proper nouns can exist in the form of names for persons, organizations, and locations [3]. To this end, the identification of proper nouns is enlisted under the classification of Named Entities [4] which can be performed with the use of word processing features.

To distinguish the presence of a specific word structure in text content, various text mining techniques have been introduced within the scope of feature extraction, of which the two most prominent ones are POS (part of speech) tagging and word sense disambiguation [5]. In addition, among various rule schematics, regular expression (regex) is also employed to retrieve specific information chunks. Fundamentally, regex is generalized based on rule patterns to describe data normalization in text [6].

In text mining, a large amount of varied unstructured documents are analyzed to extract relevant information that can be synthesized to generate new knowledge [7]. As such, this process is deemed a branch closely related to information retrieval, the objective of which is to assist users in their effort to retrieve relevant information chunks from huge computerized text data sources. This effort does not rule out the probability of the newly obtained information chunks is rather a coexistence of the other valid information collection.

Research on data mining involving numerical data obtained from text resources is known as computational linguistics [8]. In fact, the field of computational linguistics is the connecting point that binds together computer science and linguistic analysis.In other words, it is the intersection point of linguistics, phonetics, artificial intelligence and formal logic. Additionally, computational linguistics is used to generate the statistics of a huge entry of text collections, with the purpose of locating useful patterns [9]. In this respect, Natural Language Processing (NLP) is closely associated with computational linguistics that deals with linguistic computational model. Several classes of NLP analysis include phonology, morphology, syntactic, semantic and pragmatic [10].

In the linguistic field of the Malay language, there is a lack of research focusing on text mining of the Malay corpus as compared to those of the English or German language. Thus, existing repository analysis techniques, which have been used in these languages can be expanded or adapted to analyzing data written in Malay. For example, regular expression is seen as a potential candidate to help conventional morphological taggers to analyze and contribute to new data entry in such corpus. In essence, regex is used to identify the presence of text from any text array by pattern rather than by exact string [10].To date, regex has been employed in software application for various purposes, ranging from Javascript validation in forms, encoding text in a particular pattern for development motives and decrypting encoded passages in database entries. Despite these obvious benefits, the use of regex entails users to have a comprehensive understanding of its mechanics to ensure it can be executed efficiently.

Thus, designing an appropriate regex pattern involves a complicated design architecture to decide the features that need to included or excluded. In light of this need, this research was undertaken to investigate the efficiency of regex patterns in detecting the desired traits of Malay proper nouns, together with the common features of the Malay language.

1.1.Research Scope

Apparently, the study of information extraction of Malay texts is limited due to a host of problems. The following subsections highlight such limitations.

1.1.1. Lack of Standard Malay Corpus Resource

Machine learning research of Malay language thus far is still in its infancy, thus requiring more efforts to boost the availability of Malay language repository for implementations on information systems. So far, several computational linguistic studies have been carried out that focused on entity recognition and part-of-speech tagging [11]. Although the main aim of such research tends to focus on the analysis of morphological structure of a particular linguistic domain, these techniques are limited to English and German languages, and not to other languages that may have varying morphological and syntactical structures [2]. Hence, existing techniques need to be adapted with some innovative features to help analyze Malay corpus in the public domain. In this respect, regex pattern detection seems an appropriate alternative that can assist in such an effort.

1.1.2. Improve the Prerequisite to Locate Certain Lexical Features of Given Information Chunks

The continual growth of web materials can directly improve the availability of Malay information chunks that can be regulated as a reference source [12]. However, such materials are in their purest form that need to undergo some processing by humans and machines before the technological or linguistic aspects of such information can be shared and accessed by any sort of applications [13]. In computational linguistics, a huge collection of raw information, such as web materials, is required for the annotation of dictionaries and lexicographer with the purpose of composing and constant updating of the Malay dictionary. As such, Natural Language Processing (NLP) and Named Entity Recognition (NER) serve as important tools that incorporate some features of the extraction methods to identify useful text data and to conduct an in-depth analysis of a particular language structure [14]. Although some text mining features have been innovated to further improve the classification of morphological structure of composed words in sentences [15], there is an urgent need to improve current qualifications for linguistic corpuses to indicate specific lexical features of new or previously acquired text data.

1.1.3. Reliance of Machine Learning Approach on the Presence of Seed Data or Annotated Document

Corpus is usually the main reference for any linguistic research as it is the only resource where all text data are clustered in a single domain [9]. As such, lexical databases are important for the advancement of computational linguistic for a particular language. In order to initiate any types of supervised learning techniques for new dataset entries, practices in natural language processing often employ bootstrapping where a minimal amount of seed data is used to regulate new data with old datum collection [16]. Nonetheless, this approach cannot be applied to linguistic studies where resources are still scarce such as Malay language. Furthermore, there is a need to minimize or to remove the reliance of newly developed information systems on scaffolding tools such as dictionaries and wordlists in order to classify new text entries into their respective grouping [17]. To achieve such independence, regular expression methods that are commonly deployed as a problem-solving approach in software development and artificial intelligence can be integrated in existing information extraction methods to further expand annotated dataset repositories [10].

1.2. Related Work

Past studies were mainly concerned with the use of regex techniques in the classification of a minute quantity of character collection. Essentially, regex is a feature for pattern recognition with the attempt to regulate and process equal patterns encompassing object structures [10]. In addition, regex has been developed based on the identification of specific features retrieved from a structure of word sentences that possess a high recurring tendency [6]. Several benefits of applying regex in detecting morphological features include the provision of knowledge pertaining to a particular domain, together with the accountability to restrict the search space [10]

Regex provides users with a clear overview of the extracted named entity structure. In fact, regex represents a specific kind of text patterns that could be applied into information chunks to extract specific features of the text chunks. The output of such extraction mainly consists of an array of normal characters and special meta-characters [10]. The process includes locating a specific text section that is matched with a regex pattern. Interestingly, regex could be developed from a huge plethora of programming languages, namely Java, PHP, Python and VBScript. To date, a number of regex tools are accessible online such as grep, PowerGREP, and RegexMagic. Nonetheless, most developers prefer to develop new tools or innovate existing tools to cater for their software requirements.

Regular expression (regex in short) is a common technique that is usually utilized to learn text patterns for test case generations, in addition to specifying solutions for string constraints [1]. Regex is frequently used in command line tools to investigate the bracket contents and to inquire character delimiters that contains the most common behavioral themes for a particular cluster [10]. Moreover, regex has been widely utilized in query forming in mining frameworks. However, its use in data mining processes has been minimal thus far. In [18] developed a CRF-based semi-supervised learning system to identify named entity categories for diseases via bootstrapping and feature sampling. Specifically, they used this feature to extract sentences that had distinctive characteristics and morphological patterns. Similarly, in [19] deployed a feature extraction technique to detect text patterns such as nouns and character classes, within a supervised learning framework. Additionally, in medical science, text pattern extraction is performed to detect certain terminologies of the medical text that may be related to other common words which share similar context.

According to [20], pattern matching approaches in biologyis diverse ranging from basic processes (e.g., sentence extraction) to complicated processes (e.g., part-of-speech or regex). Another beneficial use of regex is in software engineering, where the method can be used to perform advanced query to inquire a particular code pattern from a repository. For example, Chapman and Steele studied the use of regex in major software repositories such as Githubto observe human collaboration with machine learning. Interestingly, they found that the use of regex in software developments was extremely high in the program compilation, registering 80% (43,525 cases) of the overall use of this technique. The remaining 20% of the overall use of regex was mainly used by in-house parser to scan documents [10]. Additionally, regex is applied in linguistic model making to detect pattern behavior of data chunk. Quite recently, Munkhjargal implemented a statistical classifier by employing regex on a simple pattern- and gazetteer-based recognizer for Mongolian language[21].

2. METHODOLOGY

This section elaborates on the processing of annotating regex patterns in self-crafted web system to assist in detecting the presence of proper nouns in web articles. This section also discusses several traits of linguistic characteristics and information extraction processes.

2.1. Text Corpus

In all studies of information extraction, the most important resource is the collection of meaningful text articles. Most of the formats of text articles, however, do not have the right definitions and contexts of the text corpus. In general, text collections residing in online repositories are maintained under the structural divisions in the form of sentences, paragraphs and chapters [13]. Typically, the contents of these repositories mostly consist of unstructured text sequences that need to undergo further processing before they could be turned into understandable constructs for users.

2.2. Regular Expression Pattern

The approach of applying regex in this context is leaning towards feature extraction rather than sentiment analysis of text characters. The extraction phase would target word appearances that follow certain predetermined criteria, which failing to abide with any of the human-crafted rules will drop such appearances from the output. Currently, manually created regex is being prioritized as the solution to retrieve desired text information [6]. Normally, regex is used to extract unclassified nouns from sentences that are still in their unexplored state-the condition of which the sentences are not annotated or have not been structurally modified [10].

2.3. Pre-Processing Features

Numerous studies of text mining have been carried out in view of its benefit in pre-processing of obtained dataset that can further improve the analysis of morphological structures. In addition, text mining can help enhance feature detection in identifying unwanted outliers in a particular domain that can adversely affect the actual performance of the entire system [10].

In this respect, the extraction of appropriate features based on pre-determined patterns is deemed highly important. Generally, this process ignores all semantic values of the text information [12] and eliminates unwanted characters (e.g., punctuation marks, whitespace and capitalization) to normalize the entire entries. For linguistic analysis, this process acts as an intermediary to sort out desired traits for preliminary analysis. Only after the dataset has been pre-processed, the ensuing categorization process would produce more precise output compared to the output of a raw dataset [22].

2.4. Applying Regex to Detect Sentence Features

This research focused on the identification of a noun structure of a given word paragraph with the use of regex in detecting spatial features of proper nouns. The correct use of proper nouns in a particular language is implied in its native form, thus determining the level of importance and definition of a particular word. In this regard, proper nouns in Malay sentences begin with upper capitalization [23], which is then subsequently followed by lower capitalization of the following letters. Clearly, this rule is similar with other major languages, notably English and Indonesian languages [22].

Nonetheless, there is a probability (thought slight) for Malay proper nouns to appear “back-character capitalized” [17]. In general, proper nouns are utilized to represent important real-word objects, together with morphological categories or annotations such as adjectives and prepositions [24]. For this research, its scope was limited based on the initial assumption that all proper nouns in Malay articles would appear in uppercase form more frequently so as to distinguish their distinctive traits from other morphological forms that could appear as well, ranging from punctuation marks to conjunctions. The following section elaborates on the basic approach in forming a rule pattern for text analysis.

2.5. Extracting Location as Proper Nouns from a Given Paragraph

Important names of locations mostly occur in capitalized forms, whereas general descriptions of locations appear in lowercase forms. As Malay nouns use Roman alphabets, they share similar characteristics with that of the English in compiling proper nouns. For example, the Malay proper nouns such as *AlamandaPutrajaya* and *MenaraBerkembarPetronas* have their first letters written in uppercase for each word and the remaining letters are written in lowercase. Clearly, capitalization signifies the importance of such words.In this respect, the initial assumption to form a regex pattern is “1 capitalized character, followed by smaller capitalized letter”, as denoted in the following expression:

*Hence, the basic regex for finding any Location as a noun is:***R = /[A-Z])\w+/g** (without discrimination of character context)

2.6. Extracting a Phone Number That Appears in a Paragraph, and Ignoring Prerequisites

A phone number is identified by a combination of numbers and uncommon hyphenations to distinguish it from other numerical features. To isolate a normal phone number from other string characters, regex pattern is generated to annotate an alignment of number sequences with several common phone number abbreviation formats such as inward dashes and hyphens. A simple assumption to create a phone number is “non-alphanumerical characters, 000-000-0000” which is the standard for local prefix used by telecom operators.

*The basic regex of detecting phone number structure is:***R** = /^\d{3}-\d{3}-\d{4}(x \d{1,6})?$/

2.7. Framework

In this research, the extraction process was divided into three main phases, namely data collection, pre-processing, and clustering. During data collection, news articles were combined into a repository according to pre-defined parameters. Then, the raw unannotated entries went through the pre-processing phase, in which text data were scanned for outliers. In fact, regex was applied to perform the statistical analysis to produce training and testing models.

2.8. Phase 1: Data Collection

The initial analysis of linguistic traits and the criteria of text data to be extracted from the accumulated source was carried out in this stage. The collection of data began with the extraction of text resource according to predefined traits such as proper nouns, conjunctions, verbs and so on. The extraction process would only initiate after the prerequisite of the text data had been decided.

2.9. Phase 2: Pre-Processing

The collected and accumulated text data were turned into a huge unsorted repository, which would be considered to be in a “human-recognizable” format.However, they needed to be further processed for further analysis. Such a process would require the text to be synthesized according to categories and distinctions, which would help facilitate a proper analysis of such data. In NER, this feature extraction would allow the data to be generalized into major classes such as “person”, “organization” and “location”. Pre-processing techniques for computational linguistic which have been adapted from the Information Extraction can help eliminate outliers that may adversely affect the quality of the obtained data. Such effective pre-processing techniques include whitespace and punctuation marks removal.

2.10. Phase 3: Clustering

Text data is placed into clusters, the objective of which is to mold a testing model for final performance evaluation. Before performing the evaluation, the cluster is analyzed for its distinctive traits, such as noun frequency and verb alignment. In this context, the annotated test dataset would has been improved with the traits of proper nouns. This analysis is vital because the implementation of regex totally ignores semantic analysis as the latter only explores characters based on the defined text characters, thus the analysis helps improve the credibility of the dataset. In addition, this analysis helps minimize the possibility of high false positives. During this process, data-clustering algorithm is applied simultaneously to produce a final output that consists of solely proper nouns for evaluation measures to take place.

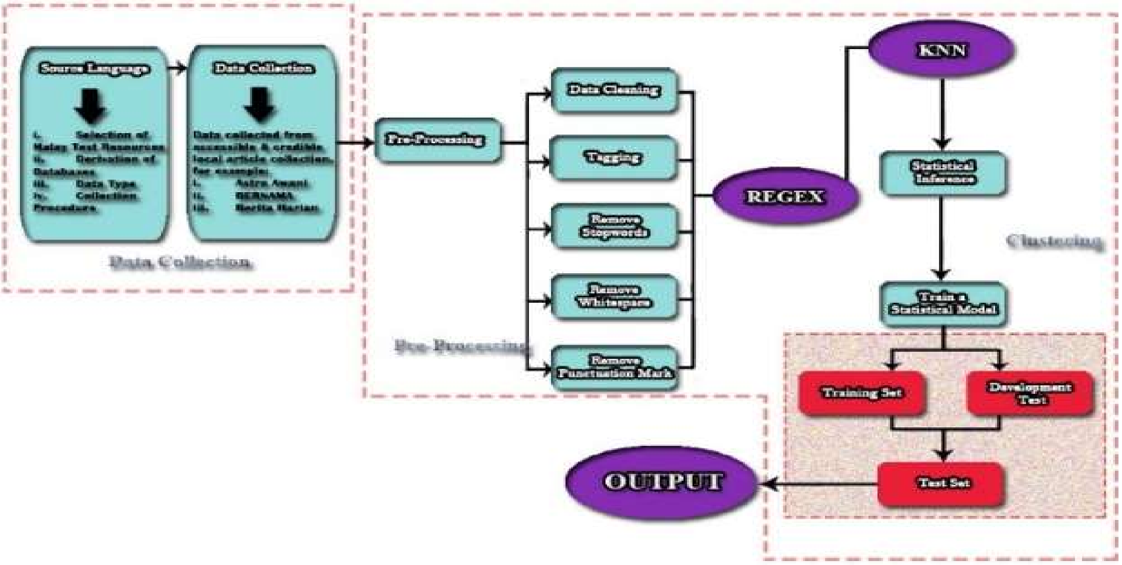

Fig.1.A framework of regex pattern implementation in annotating Malay dataset

3. EXPERIMENTS

For this section, we present an experiment that used the *Malay proper noun regex identifier* which was developed in-house by us to perform extraction tasks on 60 datasets created from the Malay news corpuses. Specifically, the aim of this experiment is to evaluate the benefits of applying regex technique in identifying linguistic traits of unannotated text resource and to investigate its correlation with standard NER techniques used in classifying text entities of various classes.

3.1. Experimental Setup

Malay language contains a shallow orthography and phonology mapping, brief morphology and simple syllable structures [25]. In order to observe the relationship between states of text data before and after cleaning and processing, two experimental classifier models were developed based on the regular expression application used for analyzing unaltered and altered data.

The text data used in this experiment were of two forms: raw form (text elements were retained in their native form, which was retrieved from the webpage) and the pre-processed collection (which had no punctuation marks and normalization of text). The experiment found a slight variation in the efficiency of proper noun detection after the text data had been modified. The selected Malay news corpuses were deemed credible given the formal writing style used to compose such news with high linguistic precision. Before applying regex detection, the scrapped dataset from these news corpuses were examined for their authenticity.The scraping process of the news contents was performed over several weeks, starting from March to April 2016, to collect the news articles in a random order. The scraping process was carried out in a four-day interval and the collected news items were meticulously sorted to select those with high quality contents and a high number of words, roughly 120-500 words.

Before performing the scraping process, an initial survey was carried out to determine the genre with highest frequency and the compatibility of the news articles with the research scope. For example, entertainment articles were found to have credible contents and high frequency compared to general news. The accumulated data were divided into training and testing sets. These sets were then processed by a regex detection program (written in PHP), while the pre-processing was performed using a Python program and manual annotation.

3.2. Data Collection and Evaluation Metrics

The procedure of data collection took into consideration the constitution of linguistic structure based on the morphological, syntactical and lexicalappearance.

3.2.1. Text Article Collection

The scope of this research was limited to analyzing only three prominent news corpuses in Malaysia namely *BeritaHarian, Bernama,* and*AstroAwani.* From the 60 articles accumulated, 10 articles with the highest frequency of collected nouns were selected. These selected articles were then developed into datasets of mixed categories such as general, entertainment, sports, finance and they were selected as the testing dataset. To evaluate the performance of the dataset, the standard evaluation measures of information retrieval were adapted as the evaluation parameters, including Precision, Recall and F-measure. Such evaluation measures were specifically selected because they have the most optimal harmonic convergence value for the cluster measurement.

3.2.2. Extraction Features

To extract relevant features from unaltered or original web articles, several key points had to be carefully considered during the crafting of the appropriate regex pattern to detect the proper nouns. Given the possibility of word structures resembling the syntax of the detected words (which were not under the proper noun category), the annotation of pattern rules and human accreditation of the acquired data were entailed to vouch the integrity of the data.Table 1 summarizes the general distinction of Malay noun taxonomy.

Table 1. Malay noun morphology based on four main language features: Affixes, prefix and suffix, reduplication, and compounding

Language Feature	Purpose
Affixes	Retrieving nouns
	Verb inflections
	Changes structure of word according to stem
	Pushes the first/last letter to the stem word
Prefix and Suffix	Appended to morphemic stem
	All stems changed when prefix is added to the noun (p, k, t, s)
Reduplication	Mark plurals, word repetitions under 1 existence
Compounding	Words consisted of more than 1 stem. These words are generated from more than 2 words with different meaning to form 1 new noun.

4. ALGORITHM IMPLEMENTATION

After the regex process, the revised K-Nearest Neighbor (KNN) algorithm was applied to identify targeted proper noun structure of such Malay articles as shown in Table 2. Proper nouns is extracted from the output of regex clustering. The following table illustrates the pseudocode algorithm applied during the entire process. This experimentation adopts semi-supervised approach by [26], modified according to the feature extraction purpose.

Table 2. The revised K-Nearest Neighbor (KNN) algorithm for unannotated data clustering

Modified KNN Classifier Model

```
1  Initialize $l_s$, CRF labeler: $l_s$= $train_s$(ts)
2  Initialize $l_k$, KNN classifier: $l_k$ = $train_k$(ts)
3  Initialize $n$, number of named entities where $n$ = 0.
4  While$n$ is collected from $C_s$, $C_s \neq$ null, do
   for word ($w$) $\in n$do
   Get feature vector $\vec{w}$: $\vec{w}$= repr $_w$ (w, t)→
   Classify $w$ with KNN: ($c$, $cf$) = knn ($l_k$, $\vec{w}$)
5  if ($f$> T) then
   Pre-label: $t$ = update ($t$, $w$, $c$)
   end if
```

```
    end for
6   Get feature vector $t$: $t$ = repr$_t$($t$, $ga$)
                    Label $\vec{t}$ with Conditional Random Field: ($t$, $cf$) = rf ($l_s$, $\vec{t}$)
    Put labelled result ($t$, $cf$) into 0
7   if ($f > \vec{y}$) then
    add labelled result $t$ to $ts$, $n = n+1$
    end if
8   if n > N then
    retrain $l_s$; $l_s$= train$_s$ ($ts$)
    retrain $l_k$; $l_k$= train$_k$ ($ts$)
    n = 0
    end if
    end while
    return 0;
```

4.1.Performance Evaluation

The process of categorizing named entities from proper noun detection would involve both training and testing models to determine the most optimal setting in which the frequency of discovering proper nouns could be improved or how well does the system fare under the indication of fewer detections. The main criterion to observe this behavior is to isolate the value from the evaluation measures, to which the dataset model would respond after the initial classification of its features. Through this process, the precision and recall rates would improve significantly, depending on the initial parameter of the output.

In fact, the three types of evaluation techniques used are based on the information retrieval technique where each output was calculated for its performance before and after alteration was carried out on the raw text dataset via pre-processing. For this experiment, precision and recall were treated as main parameters to which the performance of the system would be based. The total error rate of the dataset clustering was obtained from the summation of all differences. Labels were assigned to each entity produced from the proper noun recognition procedure. The main classes of the target items used a labelling scheme based on 4 possible values, ranging from positive to negative values as shown in Table 3. Evaluation measures were applied as a tool to further clarify the traits of the obtained nouns. These characteristics were chosen to test the data for the accuracy of information, precision of the obtained item, and retention rate in the real-world context. Table 4 explains the implementation of the evaluation measures.

Table 3. Labelling scheme for classes in clustering/classification

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; predicted as negative
False negative (**FN**)	Actual positive, but predicted as negative

Table 4. General formula to evaluate precision, recall and F-measure

Parameter	Formula
Precision	$p = \frac{tp}{tp + fp}$
Recall	$r = \frac{tp}{tp + fn}$
F-Measure	$2 * \frac{(precision * recall)}{(precision + recall)}$

4.2. RESULTS

The objective to cluster regexes from distinctive values is to detect linguistic features present in articles. Within this context, the limitation of regex pattern in the detection of capitalized characters, whitespaces removal and punctuation marks was emphasized.

Nonetheless, this pattern was still considered primary, as regex could adopt a more complex form to target a more effective detection of proper nouns. In this study, the behaviorof the two experimental datasets were observed namely the training and testing datasets, to test the performance of regex in the detection of proper nouns in the selected Malay corpuses. In fact, 120 datasets were obtained, consisting of raw and pre-processed data (the latter were processed using information extraction methods). These datasets were equally divided into two sets after skimming and accrediting the scraped text contents. From the 30 training and testing datasets, 10 datasets that possessed the most suitable traits in accordance with news content were selected for further analysis. The rates of detection of proper nouns by the regex approach are shown in Fig. 2 and Fig. 3.

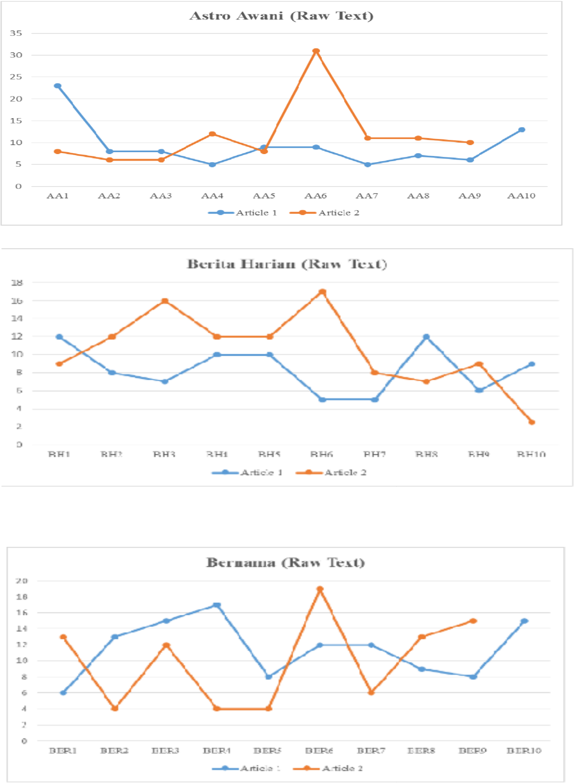

Fig.2. Detection rate of proper nouns before text pre-processing (ground truth)

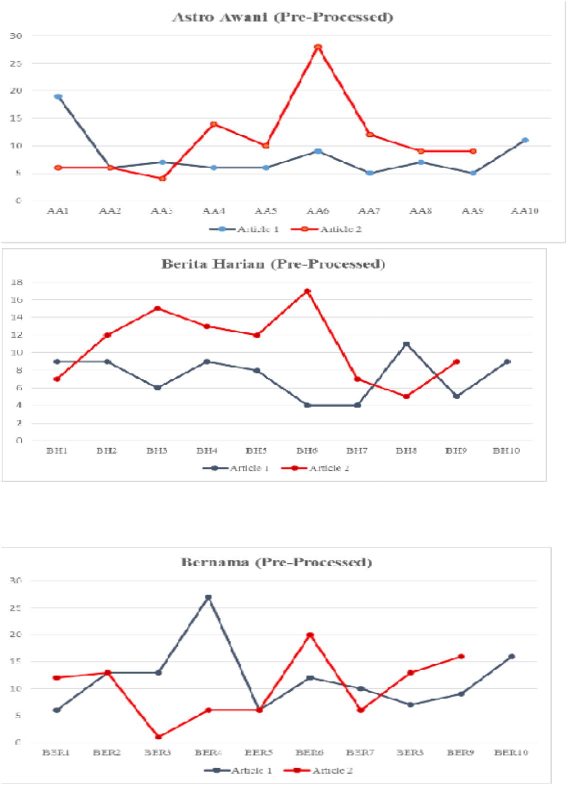

Fig.3. Detection rate of proper nouns after text pre-processing (testing model)

For the three news corpuses, 20 different articles were collected. Evidently, there was a significant difference in the number of detected proper nouns between the raw data and the pre-processed data where the latter was better than the former. A majority of the accumulated text data were of politics, sports, technology and current affairs genres. The output showed that the AstroAwani news contents had more nouns than the other news corpuses, containing between 5 and 35 nouns per article. This particular finding suggests that the occurrence of proper nouns varies significantly among Malay news articles. As anticipated, proper nouns tend to be used more frequently in formal articles and they are used quite extensively in social and political contexts.

Fig. 4 shows the output of KNN clustering of the tested dataset. The result suggests that optimal detection performance can be achieved when data points are uniformly spread, rather than when they are concentrated at a particular point. In this regard, vectorization can be used to help achieve such desired performance.

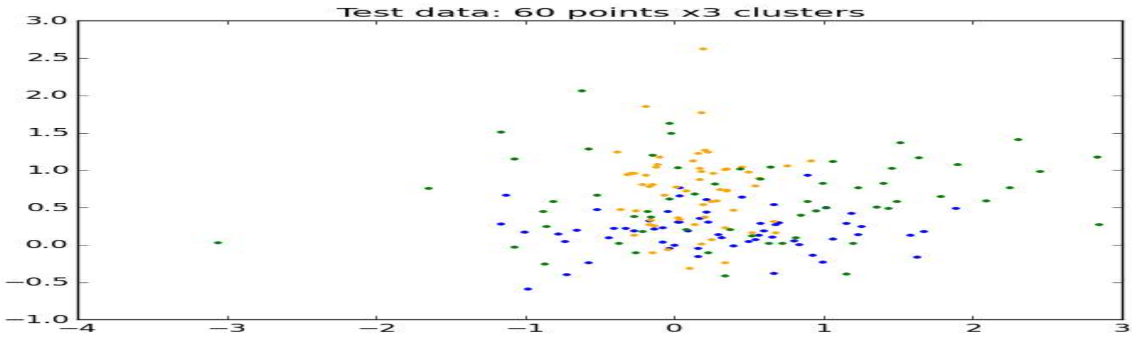

Fig.4. Clustering of the accumulated text data of the three news dataset.

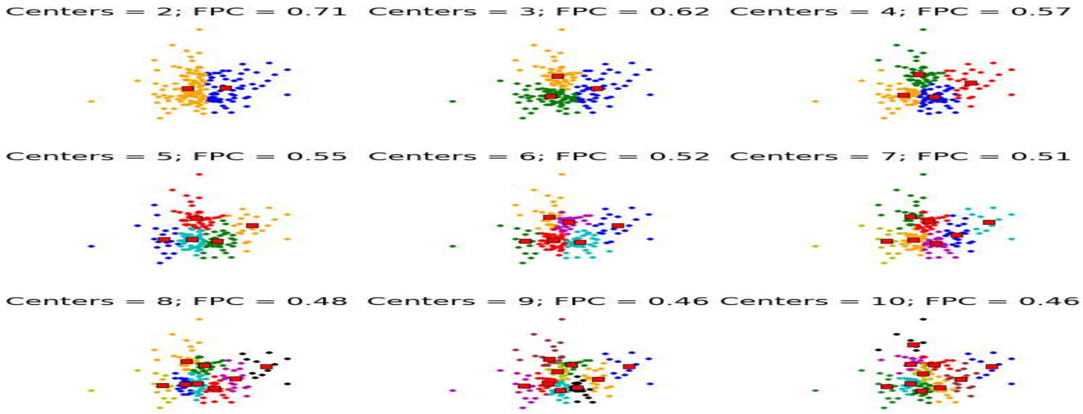

Fig.5. KNN algorithm defining the optimal number of clusters (the more the points at the center, the less optimal the cluster is)

Fig. 6 shows the detailed results of the use of KNN algorithm on the mock dataset of the three news corpuses. This figure shows an ascending scattering pattern of the dataset entities as the number of premediated clusters was increased and the dataset of text entities with least similarity dispersed uniformly.

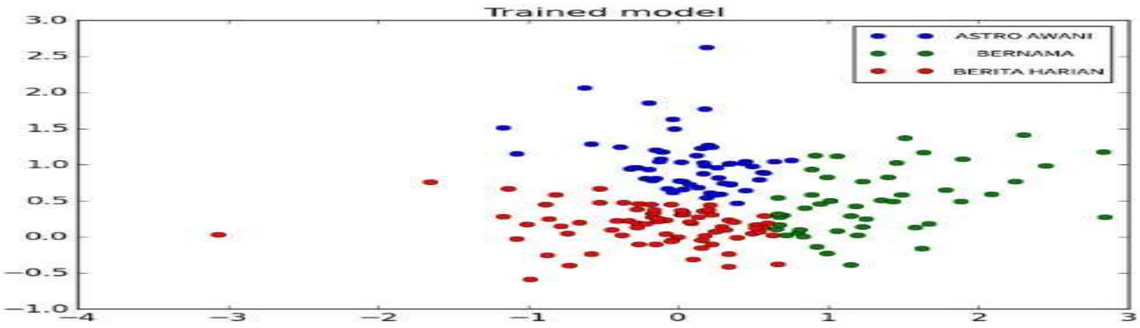

Fig. 6. Scattered pattern of the dataset constitution based on the distance between objects (TF-IDF)

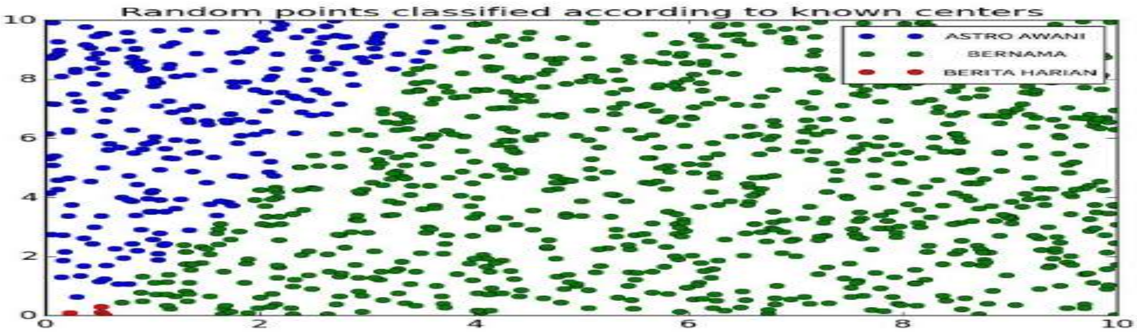

Fig. 7. Illustration of the tendency of the scattered pattern of the three clusters

Table 5. Result of KNN algorithm and regular expression matching on Malay proper noun detection

Dataset	No. of PN in Dataset	No. of PN Identified	Average		
			Precision	Recall	F_1
AA	139	131	0.52	0.53	0.504
BH	122	116	0.57	0.61	0.567
BER	153	164	0.47	0.46	0.450

5. DISCUSSION

For the implementation of regex, several factors need to be accounted for to ensure it can perform the detection of proper nouns effectively. Firstly, the integrity of the detected items need to be checked, because the regex approach is very dependent on the preconfigured rule. In fact, similar word structure detected during the process that exhibits the targeted extraction feature would tamper the accuracy of the obtained output. To help mitigate such problem, Python's pattern language that composes regex can support default character classes.

Secondly, the character class definition can be inquired from every extraction; however, it needs to be validated for its practical use. Hence, regex pattern identification is performed without discrimination, depending on the nature of the environment in which it is implemented. Additionally, other limitations may impede the extraction process of data accumulation, such as endpoint anchors, capture groups, boundaries of word, redundancy and initial and end values. Arguably, regex performs well in software pattern recognition; however, for computational linguistics, it requires fine-tuning to adapt to morphological settings of a linguistic corpus. From the practical standpoint, regex implementation would act as an intermediary to supplement current systems to detect the presence of text entities.

6. CONCLUSION

This paper discusses the findings of the present study in which the extraction of proper noun structures from an unannotated Malay dataset was the focus of the research. Specifically, a proof-of-concept approach was used to simulate regex's functionalities in locating certain features of Malay linguistic structures in the data mining process. As computational linguistics in Malay language is still new, the concept of acquiring certain text features is vital in isolating the traits and classes of a text structure. As demonstrated in this study, regex can efficiently synthesize Malay words with precision. Nonetheless, human intervention is still required, especially in validating the integrity of the processed data.

Clearly, the implementation of regex has improved as more knowledge is acquired from NER studies, particularly from those that focus on part-of-speech tagging and raw corpus annotation. In this study, the occurrence of noun subjects was found to be moderately high thus emphasizing their importance in the Malay corpus for relaying information. Proper nouns have their own nomenclature, which is unique based on the usage context and abbreviations. Essentially, regex can process the sentence structure to detect its class elements which can then be expressed accordingly. In this respect, refactoring of regex pattern would help detect Malay nouns with improved readability thus making them more understandable to both man and machine. As stated earlier, regex libraries could be defined to cater to user objectives thus enabling them to achieve better detection performance compared to other feature extraction techniques.

For future research, efforts to improve the availability of Malay corpus resource should take into consideration the implementation speed of regex patterns in different domains in order to optimize the rate of data detection. Furthermore, the development steps of regex should be configured more efficiently such that it can discover text contents by parsing the string elements with few errors. Future research should also focus on eliminating such errors by using intrinsic regex measures that would help improve the features available in the regex library. Additionally, it would be of great interest and importance to apply critical algorithms, such as Bag of Words and word vectorization in experiments involving Malay text data, specifically that investigate the relationship between word and system performances.

7. REFERENCES

[1] Srivastava A. N., Sahami M.Text mining: Classification, clustering, and applications. Florida: CRC Press, 2009

[2] Etzioni O, Banko M, Soderland S, Weld D S. Open information extraction from the web. Communications of the ACM, 2008, 51(12):68-74

[3] Suwarningsih W,Supriana I, Purwarianti A. ImNER Indonesian medical named entity recognition. In 2014 2nd International Conference on Technology, Informatics, Management, Engineering and Environment, 2015, pp. 184-188

[4] Al-Shoukry S, Omar N. Proper nouns recognition in Arabic crime text using machine learning approach. Journal of Theoretical and Applied Information Technology, 2015, 79(3): 506-513

[5] Elyasir A M H, Anbananthen K S M. Comparison between bag of words and word sense disambiguation. In International Conference on Advanced Computer Science and Electronics Information, 2013, pp. 413-417

[6] Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish H V. Regular expression learning for information extraction. In Conference on Empirical Methods in Natural Language Processing, 2008, pp. 21-30

[7] Sari Y, Hassan M F, Zamin N. Rule-based pattern extractor and named entity recognition: a hybrid approach. In 2010 International Symposium on Information Technology-Engineering Technology, 2010, pp. 563-568

[8] Ojo A, Adeyemo A B. Framework for knowledge discovery from journal articles using text mining techniques. African Journal of Computing and ICT, 2012, 5(2):35-44

[9] Zamin N, Oxley A. Building a corpus-derived gazetteer for named entity recognition. In International Conference on Software Engineering and Computer Systems, 2011, pp. 73-80

[10] Chapman C, Stolee K T.Exploring regular expression usage and context in python. In 25th International Symposium on Software Testing and Analysis, 2016, pp. 282-293

[11] Ritter A, Clark S, Etzioni O. Named entity recognition in Tweets: An experimental study. In 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1524-

1534

[12] Mohd D Z. Processing natural Malay texts: A data-driven approach, Trames, 2010, 14(1):90-103

[13] Ramli I, Jamil N, Seman N, Ardi N. An improved syllabification for a better Malay language text-to-speech synthesis (TTS). Procedia Computer Science, 2015, 76:417-424

[14] Collobert R, Weston J, Bottou L, Karlen M, KavukcuogluK, Kuksa P. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12:2493-2537

[15] Zhang X, LeCun Y. Text understanding from scratch. Learning: Computation and Language. 2015, https://arxiv.org/pdf/1502.01710.pdf.

[16] Althobaiti M, Kruschwitz U, Poesio M. A semi-supervised learning approach to Arabic named entity recognition. In IEEE Proceedings of Recent Advances in Natural Language Processing, 2013, pp. 32-40

[17] Alfred R, Chin LL, Kim O C, Anthony P.Malay named entity recognition based on rule-based approach. International Journal of Machine Learning and Computing, 2014, 4(3):300-306

[18] Suakkaphong N, Zhang Z, Chen H. Disease named entity recognition using semi-supervised learning and conditional random fields. Journal of the American Society for Information Science and Technology, 2013, 14(4):90-103

[19] Nicholson J, Baldwin T. Learning count classifier preferences of Malay nouns. In Australasian Language Technology Workshop, 2008, pp. 115-123

[20] Ananiadou S, Pyysalo S, Tsuji J, Kell D B. Event extraction for systems biology by text mining the literature. Journal of Trends in Biotechnology, 2010, 28(7):381-390

[21] Horak A., Kopecek I., Pala K. Text, speech and dialogue. Heidelberg: Springer-Verlag Berlin, 2010

[22] Nadeau D, Sekine S.A survey of named entity recognition and classification.LingvisticaeInvestigationes, 2007, 30(1):3-26

[23] Wibawa A S, Purwarianti A. Indonesian named-entity recognition for 15 classes using ensemble supervised learning. Procedia Computer Science, 2016, 81:221-228

[24] Abu B J, Omar K, Nasrudin M F, Murah M Z, Al-Shoukry S, Omar N, Klose A. Processing natural Malay texts: A data-driven approach. Journal of Neurocomputing, 2013, 79(3):2670-2676

[25] Fadzli S A, Norsalehen A K, Syarilla I A,Hasni H, Dhalila M S S. Simple rules

Malaystemmer. In International Conference on Informatics and Applications, 2012, pp. 28-35

[26] Tran V C, Hwang D, Jung JJ.Semi-supervised approach based on co-occurrence coefficient for named entity recognition on Twitter.In 2nd IEEENational Foundation for Science and Technology Development Conference on Information and Computer Science, 2015, pp. 141-146

How to cite this article:
Sulaiman S, Wahid R A, Morsidi F. Feature extraction using regular expression in detecting proper noun for malay news articles based on knn algorithm. J. Fundam. Appl. Sci., 2017, *9(5S), 210-231*.