An Enhanced Malay Named Entity Recognition using Combination Approach for Crime Textual Data Analysis

Siti Azirah Asmai[1], Muhammad Sharilazlan Salleh[2], Halizah Basiron[3], Sabrina Ahmad[4]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya
76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—Named Entity Recognition (NER) is one of the tasks in the information extraction. NER is used for extracting and classifying words or entities that belong to the proper noun category in text data such as person's name, location, organization, date and others. As seen in today's generation, social media such as web pages, blogs, Facebook, Twitter, Instagram and online newspapers are among the major contributors to the generation of information. This paper presents an enhanced Malay Named Entity Recognition model using combination fuzzy c-means and K-Nearest Neighbours Algorithm method for crime analysis. The results showed that this combination method could improve the accuracy performance on entity recognition of crime data in Malay. The model is expected to provide a better method in the process of recognizing named entities for text analysis particularly in Malay.**

***Keywords*—Named entity recognition; information extraction; fuzzy c-means; k-nearest neighbors; malay language; crime data**

I. Introduction

Information is one of the important sources in human life that is increasingly rising and technologically. At all times, various types of information have been generated on the internet and the amount of information is constantly increasing from time to time. Information consisting of various types such as text, images, audio, video, data, and so on are increasingly being generated on the internet which are largely unstructured. This growing number of information affects the daily lives of people in work, learning and lifestyle. Effective management and organization of information is a key strategy for addressing the problem of finding useful information. The appropriate techniques and methods are very necessary to process and extract the essential knowledge contained in this information.

Therefore, this paper presented the Malay named entity recognition using clustering and classification method. The rest of this paper is organized as follows. In Section 2, it discusses the related work for the named entity recognition task. Section 3 presents techniques and machine learning algorithms for NER. Then, Section 4 discusses the Malay NER and follow by its approach in Section 5. Next, the experiment result and discussion are elaborated in Section 6. Finally, Section 5 covers the conclusion.

II. Related Work

Named Entity Recognition (NER) is important in analyzing the crime report to address the problem of crime due to the use of different languages in writing crime reports for each country. When a lot of information relates to crime occurrences are available on the web with many specific entities, many techniques can be used in NER for extracting useful information for better crime analysis and execution actions that explain by Hosseinkhani, Koochakzaei, and Keikhaee [1].

Shabat and Omar [2] have implemented NER tasks using an ensemble framework that focuses on designing models to extract specific criminal information from the Web. Their main goal is to integrate the set of features and classification algorithms in an orderly way to synthesize more precise classification procedures. Three base-classifiers specifically Naïve Bayes, Support Vector Machine and K-Nearest Neighbor classifiers are used for each of the feature sets and these three classifiers are combined using a weighted voting ensemble method.

Alkaff and Mohd [3] have analyzed online news, blogs and social networking sites on the internet using gazetteers and rule-based extraction for named entity recognition in identifying crime hot spots. Therefore, an accurate natural language processing technique is needed to be explored to capture and recognize named entity within open domain textual data effectively.

Execution processing recognition named entity analysis requires several steps to achieve the objectives of the research. The steps including the pre-processing stage, the annotation stage and evaluation or developing system stage. Based on Jurafsky and Martin [4] there some basic steps in the statistical sequence labelling approach to creating a named entity recognition system. The following Fig. 1 shows the steps illustration.

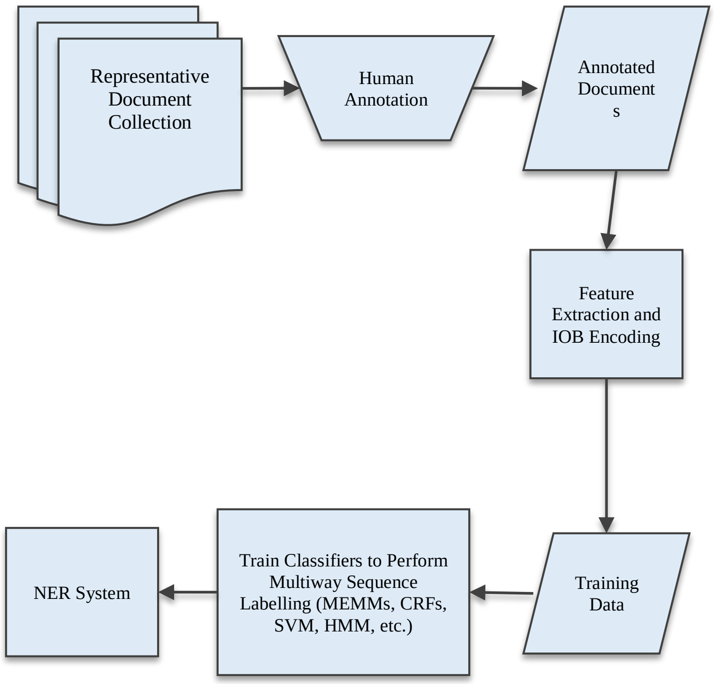

Fig. 1. Basic steps approach for NER

III. Techniques and Machine Learning Algorithms for NER

Many method and techniques are being continuously developed which is it more focus on managing of information and knowledge. Previous knowledge management a strongly focuses on just keeping large amounts of data for data mining. Now the growing use of the Internet and the information burden placed a huge demand for managing intelligent information efficiently and effectively. This application of artificial intelligence methods and research in the growing area of human-machine interaction is ahead grounds for more investigations.

A. Rule-based Approaches

In computer science, rule-based systems are used as a means of storing and manipulating knowledge to interpret information in a useful way. They are often used in artificial intelligence applications and research. Normally, the term Rule-Based System ('rules-based system') is used for systems involving a set of man-made rules or rules outlined. Today, these rules-based systems are widely being used and implemented for many kinds of problem and tasks. As developing the text analysis that focuses on NER task, the rule-based approach is used for the recognition of named entities by defining rules regarding the status of entity members' position in the phrase or sentence. The constraints in the implementation of this method lie in the capability of a pattern definition that is usually done by a linguist. Rule-based NER is also too dependent on the language used.

In general, the NER system using a rule-based approach has Part-of-Speech (POS) tagger, sentence or phrase syntax and orthographic, such as word capitalization pattern combined with the data dictionary. Eftimov et al. [4] state that the NER method using a rule-based approach uses a regular expression that combines information from the source terminology and interests of the feature entity. The main drawback of this method is the construction of manual rules, which are time-consuming and dependent on the domain. Eftimov et al. [4] combined the terminological-driven NER with rules-based NER as their proposed rule-based method called as DrNER extracting knowledge for evidence-based dietary recommendations. The basic structure of the rule-based expert system is shown in Fig. 2.

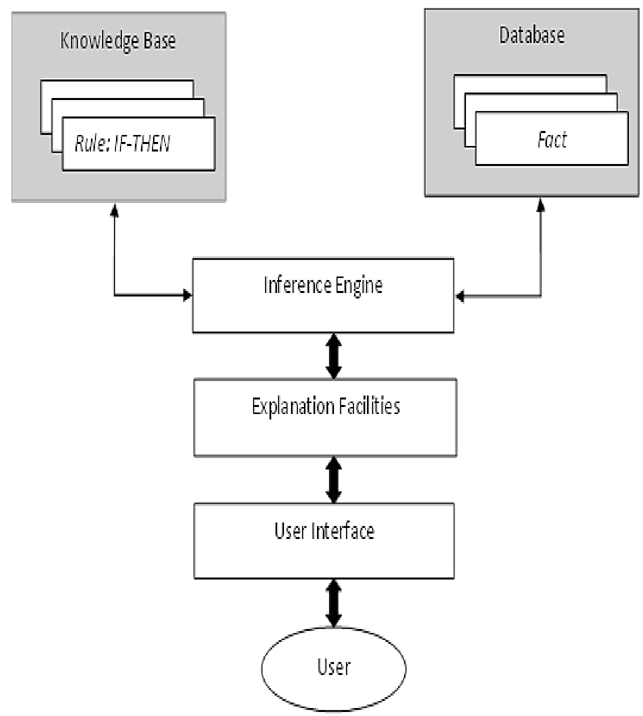

Fig. 2. Basic structure of rule-based expert system (Abraham, 2005)

B. Learning-based Approaches

- Supervised Learning

The ability to learn unnamed entities is an essential part of the NER solution. Early studies were mostly based on the supervised learning (SL). The supervised learning algorithm is the process of forming a relationship model and dependence between predictive output and input characteristics so prediction of output values for new data can be predicted based on the relationships studied from previous datasets. Kotsiantis [5] stated that supervised machine learning is an algorithm that generates the general hypothesis based on externally supplied examples and hence is used in making predictions about future instances. In other meaning, the purpose of this learning is to build a brief model that distribute class labels based on predictor features.

Morwal [6], Chopra and Morwal [7] use Hidden Markov in named entity recognition. While Ahmed and Sathyaraj [8] applied maximum entropy to recognize entity sets from a given text such as name, location and organization. With the different variant of SL techniques, it offers tagging words of the test corpus from the define corpus that require a large set of heuristic rules and clusters.

- Unsupervised Learning

One of the learning based approaches for pattern recognition is unsupervised learning (USL). Unsupervised learning is an artificial intelligence algorithm (AI) that performs data isolation in a dataset using unlabelled or classified information where the isolation is based on the hidden features contained in the data. This algorithm acts on this information or data without guidance. The AI system used can arrange information based on similarities and differences in information although no category is provided among the data. The AI system algorithm also acts on data without prior training. Sathya and Abraham [9] stated that unsupervised learning model recognises information based on heuristic patterns and Reinforcement learning learns through trial and error interactions with their surroundings (rewards / penalties).

Unsupervised learning is also used in named entity recognition tasks. This learning-based is one of the approaches in solving the problems encountered in the task of named entity recognition. Li et al. [10] presented the unsupervised NER system without explicit human label efforts named TwiNER for targeted tweet streams in the Twitter application. The system not dependent on unreliable local linguistic features. Furthermore, S. Zhang and Elhadad [11] also proposed an unsupervised approach in the biomedical field for NER task by extracting named entities from biomedical text. This unsupervised approach for NER was conducted using three main step which are seed term collection, boundary detection and entity classification.

- Semi-supervised Learning

Semi-supervised learning is a technique that is a combination of supervised learning and unsupervised learning. A variety of semi-supervised learning method tries to generate high-quality training data automatically from the unlabelled corpus. By using the semi-supervised learning technique, it can produce considerable improvement in learning accuracy. This improvement in learning accuracy can help in the structured process of extracting named entities such as location, person, type of crime and other entities involved in the crime situation more accurately from any unstructured data like email messages, word processing documents and web blogs.

However, traditional semi-supervised learning methods remain to rely on the high quality of the labelled entity to learn the context of unlabelled data in textual data. Fuzzy semi-supervised clustering it offers a new opportunity to overcome classical methods and crisp semi-supervised hierarchical clustering. However, fuzzy semi-supervised clustering is still a new subject and not many studies have been done with fuzzy semi-supervised cluster related on named entity recognition in the literature. Diaz-Valenzuela, Vila, and Martin-Bautista [12] use fuzzy semi-supervised clustering approach to classifying scientific publications in digital web libraries. They use the concepts of fuzzy must-link and fuzzy cannot-link constraints for identifying optimum α-cut of a dendrogram.

Castellano, Fanelli, and Torsello [13] use a semi-supervised fuzzy clustering algorithm to group shapes into some clusters. Each cluster is represented by a prototype that is manually labelled and used to annotate shapes belonging to that cluster. To capture the evolution of the image set over time, the previously discovered prototypes are added as pre-labelled objects to the current shape set and semi-supervised clustering is applied again. Both of these recent studies improve the accuracy of the group clusters under the supervision of a limited number of labelled data.

IV. Malay NER

This research discusses the overview of Malay language based on some aspects related to this scope. The Malay language is also one of the language fields that get researchers interest to implement the named entity recognition task. It focuses on the identification of proper nouns in Malay. Like other languages, the Malay language also has its own characteristics in the presentation of information based on the order of sentences and the form of words that have certain meanings. The Discussions on the execution of named entity recognition in the Malay language include orthography, morphology, structure, and so on.

Alfred, Chin Leong, Kim On, and Anthony [14] explains that as one of the processes in Text Mining, a named entity recognition is very useful for information extraction by helping user for entities identification and detection like the person, location and organization. They also argue that different NER processes need to be applied to different languages due to morphological differences. So, a Rule-Based Named-Entity Recognition algorithm for Malay articles has been proposed based on a Malay part-of-speech (POS) tagging features and contextual features in dealing with Malay language articles. The use of a set of rules and manually-specified dictionary lists by the human is a method used in the Rule-Based NER algorithm in identifying named entities. Due to the lack of annotated corpus sources for the Malay language which can be used as training data, they have used rule-based methods rather than using machine learning method to identify person, organization and location as three named entities major types. The rule has been made based on the POS-tagging contexts. The F-Measure result's value during conducted the NER experimental was 89.47%.

Furthermore, another experiment was conducted by Sulaiman et al [15] to detect Malay named entity recognition. Stanford NER and Illinois NER tools are used to identify the Malay named entity using online news articles as a process of measuring the capabilities of this tool in the identification of Malay entities. Experimental comparisons have found that Stanford NER tends to yield higher results on F1 and Precision than Illinois NER. These two tools, Illinois NER and Stanford NER are developing based on machine learning method. They conclude that, for improvements in the named entity task in Malay, most NER Malays are used rule-based methods. After conducting experiments, they found that both NERs tools showed a low detection result for the Malay corpus because there were many errors when identifying entities. This is because of the morphological differences between Malay and English.

Besides that, Salleh, Asmai, Basiron, and Ahmad [16] was applied conditional random fields method in developed an automated Malay Named Entity Recognition (AMNER) conceptual model to recognize entities for the Malay language. Current approaches for Malay NER are more using a set of rules and list of dictionaries set by the human to identify entities. These rules work to extract the pattern of an entity such as location, organization and other entities based on their basic pattern. Due to limitation, the libraries or dictionaries used should always be updated for recognizing named entities. The Malay language features as the main factor on their development model as the guidance for the named entity recognition process. There are several structures in Malay language writing as follows.

A. *Orthography*

In the execution of named entity recognition tasks, one of the things involved is the conventional spelling system of a language called orthography. The Malay language also has its own orthography in the spelling structure. Based on Cho [17], they explain that in the present time, the Latin alphabet has been used for orthography and spelling system for the Malay and Indonesian languages that have been made by Western linguists. Besides that, Zaidi, Rozan, and Mikami [18] stated that with the use of Malay language standard words using 26 letter alphabets known as Rumi in Malay, it is compatible with communication technology and has the potential to use only the text-based features for communicating in Malay. Orthography used in Malay includes spelling norms, hypotheses, emphasis, punctuation, capitalization, fractions of words.

B. *Morphology*

Furthermore, morphology is also used in the research of named entity recognition. Morphology in linguistics is the study of the words inner structure and word formation that forms the essential part of today's linguistic study. It describes how the words are formed and their relationship to other word focus on the same language. By breaking the words down into smaller, meaningful part, this smallest meaningful part of a word is called a morpheme. Word structure and part of words analyzed by morphology include stems, prefixes, suffixes and root words. In addition, it also sees the part of speech, the way the context can change the word's pronunciation and meaning, as well as the intonation and pressure in one word.

V. A Malay Named Entity Recognition Approach

The research is conducted through five phases represented in the form of research design. Each phase in the research design is intensively investigated and then used to facilitate the next phase of the research. The Phase One begins with data acquisition, data obtained in the form of web pages and unstructured. The Phase Two is pre-processing data and is followed by a Phase Three that focused on features extraction. Then, the development of the NER Malay model was carried out in Phase Four. Finally, an accuracy of the entity recognition is evaluated in Phase Five. Fig. 3 illustrates the design of the proposed Malay Named Entity Recognition (MNER) approach.

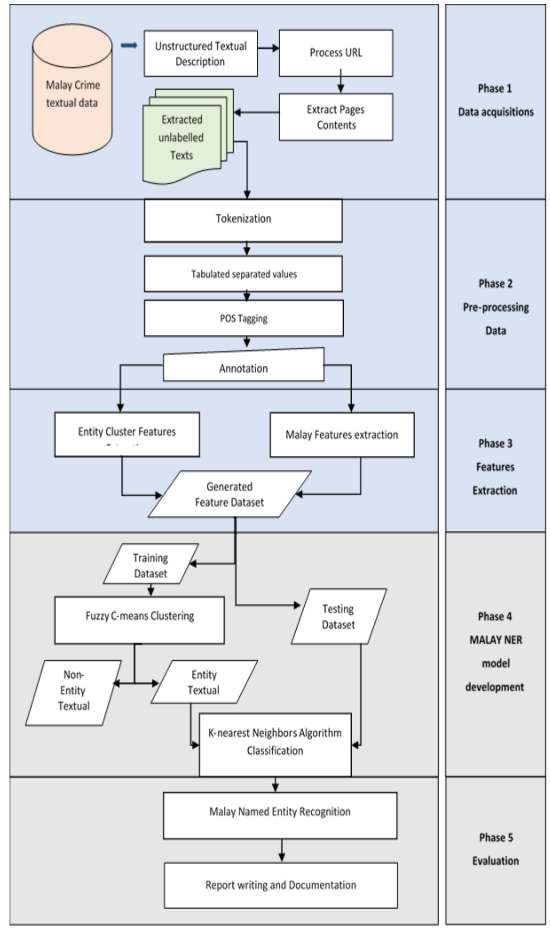

Fig. 3. The Proposed Malay Named Entity Recognition Design

A. *Data Acquisition*

Based on research design in Fig. 3, data acquisition is conducted in Phase One. Data is obtained from the Malay Crime News PDRM Website in the form of web pages. These web pages contain some elements such as URL links, images, and texts that need to be processed as they are in unstructured form. The page contents are extracts to obtain the required information which as extracted unlabeled PDRM News Texts.

B. *Pre-processing Data*

Pre-processing involved four tasks towards the data. As the process in Phase Two, the documents that contain many unstructured data need to delimit into meaningful units by performing tasks like tokenization, tabulation values, POS tagging and annotation. Then, after the annotation process was done, the data were divided into two parts: training data and testing data. The following Fig. 4 shows the process for pre-processing data.

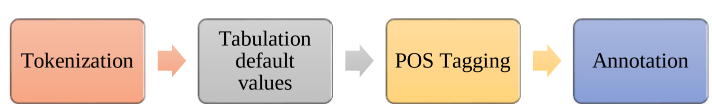

Fig. 4. Pre-processing Data

TABLE I. THE PENN TREEBANK PART-OF-SPEECH TAG SET

Tag	Details
CC	conjunction, coordinating
CD	cardinal number
DT	determiner
EX	existential there
FW	foreign word
IN	conjunction, subordinating or preposition
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
LS	list item marker
MD	verb, modal auxillary
NN	noun, singular or mass
NNS	noun, plural
NNP	noun, proper singular
NNPS	noun, proper plural
PDT	predeterminer
POS	possessive ending
PRP	pronoun, personal
PRP$	pronoun, possessive
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	adverb, particle
SYM	symbol
TO	infinitival to
UH	interjection
VB	verb, base form
VBZ	verb, 3rd person singular present
VBP	verb, non-3rd person singular present
VBD	verb, past tense
VBN	verb, past participle
VBG	verb, gerund or present participle
WDT	*wh*-determiner
WP	*wh*-pronoun, personal
WP$	*wh*-pronoun, possessive
WRB	*wh*-adverb
.	punctuation mark, sentence closer
	punctuation mark, comma
:	punctuation mark, colon
(	contextual separator, left paren
)	contextual separator, right paren

- Tokenization

The text data file (.txt) that were presented in unstructured data consisted of sentences and paragraph which were tokenized as the process of separating a text into valuable elements, words, phrases, symbols or digits called tokens. The tokens were presented in a list as the input for further processing.

- Tabulation Values

Next, the token text file was processed to store data in a tabulator structure like spreadsheet data. The file was divided into three rows namely token data, part of speech tag (POS) and named entity tag. Before continuing to the annotation stage, entity tag column was set as default value "O" as outside or other.

- POS Tagging

Every token in the file was also annotated with POS tagging bands such as CC, CD, NN, VB and others. The description of The Penn Treebank POS tagset is based on Table 1.

- Annotation

Then, the file was annotated with entities types. There are five types of entities that are being worked out in this research. Those entities are person name, location, organization, date, and types of crime labelled as PERSON, LOCATION, ORGANIZATION, DATE and CRIME TYPE. For non-entity types, they are labelled as OTHER. The final preprocessing dataset produced is shown in both Fig. 6 and Fig. 7 respectively with their features extraction.

C. Features Extraction

In Phase Three, the process of extracting features for the named entity recognition task has been performed. Feature extraction is divided into two parts. The first part, some features have been extracted for use in clustering process and in the second part; some other features have been extracted for use in the process of classification. The generated feature dataset is produced in this phase for further analysis. The features selected for both parts are as appropriate to carry out the task of recognizing named entities in the Malay language. The details process of extracting these features are discussed as shown in Fig. 5.

D. Malay NER Model Development

Furthermore, in Phase Four, there are two types of learning used, namely clustering and classification. Fuzzy C-Means as clustering method is used to cluster the data either entity or non-entity. After that, the correct entities that have been clustered are labelled based on more detailed entity types which are person, location, organization, date, and type of crime. Then, these entities through the classification process by using K-nearest Neighbors Algorithm Classification.

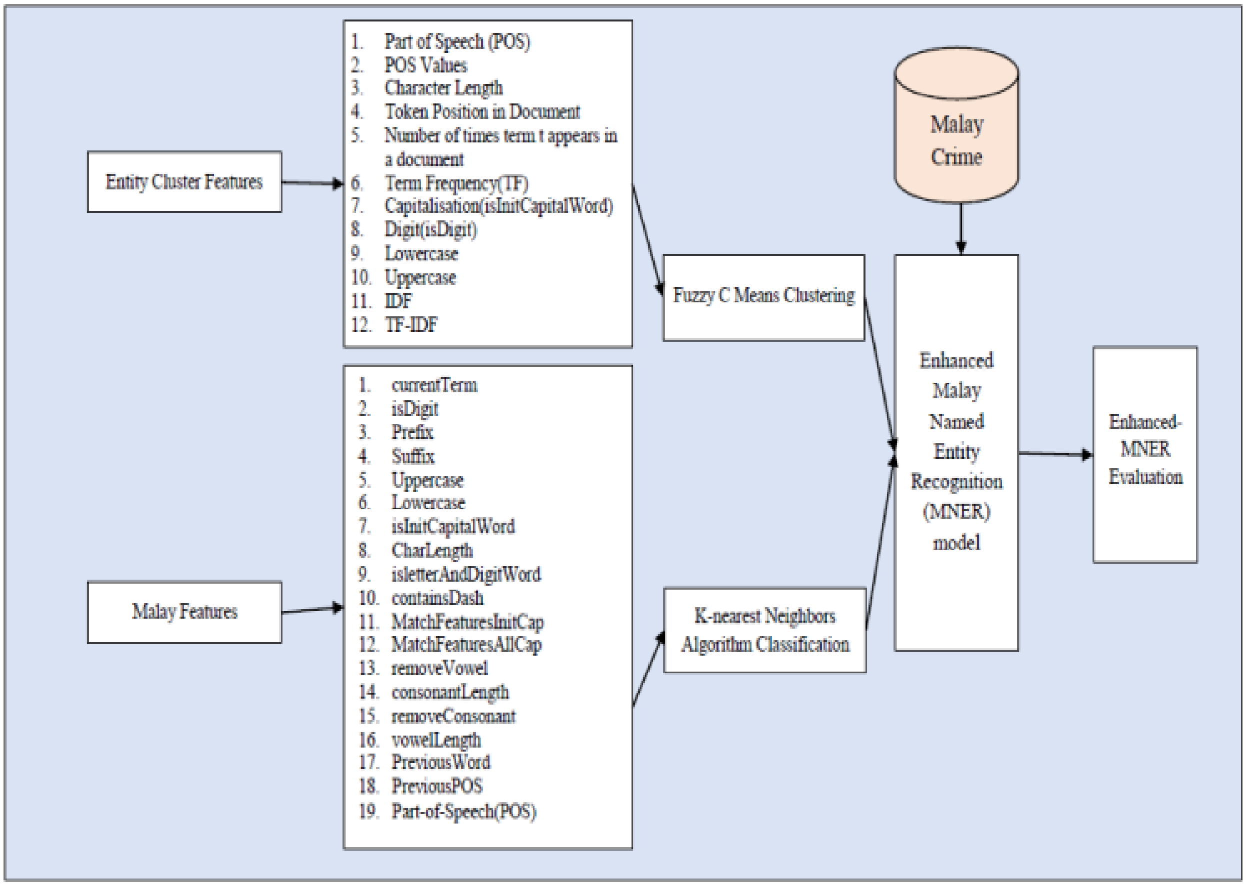

Fig. 5. MNER Features Extraction

Row No.	Term	POS	POS_ValueNormalize	Character Length_normalize	Token Position in Document	noAppearNorm	Term Frequency(TF)	Lowercase	Uppercase	TFIDF	CLASS
1	BELUKAR	NN	0.0666666666666667	0.2692307692307692	1.0	0.07272727272727272	0.016597510373443983	0.0	1.0	0.028885687792435583	NON ENTITY
2	JADI	VB	0.4666666666666667	0.15384615384615385	0.995850622406639	0.03636363636363636	0.008298755186721992	0.0	1.0	0.014442843896217791	NON_ENTITY
3	TEMPAT	NN	0.0666666666666667	0.23076923076923078	0.991701244813278	0.03636363636363636	0.008298755186721992	0.0	1.0	0.009446495420467065	NON ENTITY
4	JUAL	VB	0.4666666666666667	0.15384615384615385	0.9875518672199171	0.05454545454545454	0.012448132780082987	0.0	1.0	0.01791700448751364	NON_ENTITY
5	HEROIN	NN	0.0666666666666667	0.23076923076923078	0.983402489626556	0.05454545454545454	0.012448132780082987	0.0	1.0	0.01791700448751364	NON ENTITY
6	Jabatan	NNP	0.13333333333333333	0.2692307692307692	0.979253112033195	0.07272727272727272	0.016597510373443983	0.0	0.0	0.0	ENTITY
7	Sumber	NNP	0.13333333333333333	0.23076923076923078	0.975103734439834	0.01818181818181818	0.004149377593360996	0.0	0.0	0.0	ENTITY
8	Strategik	NNP	0.13333333333333333	0.34615384615384615	0.970954356846473	0.01818181818181818	0.004149377593360996	0.0	0.0	0.0	ENTITY
9	Dan	NNP	0.13333333333333333	0.11538461538461539	0.966804979253112	0.12727272727272726	0.029045643153526972	0.0	0.0	0.0	ENTITY
10	Teknologi	NNP	0.13333333333333333	0.34615384615384615	0.9626556016597511	0.01818181818181818	0.004149377593360996	0.0	0.0	0.0	ENTITY
11	Jabatan	NNP	0.13333333333333333	0.2692307692307692	0.9585062240663901	0.07272727272727272	0.016597510373443983	0.0	0.0	0.0	ENTITY
12	Integriti	NNP	0.13333333333333333	0.34615384615384615	0.9543568464730291	0.01818181818181818	0.004149377593360996	0.0	0.0	0.0	ENTITY
13	Dan	NNP	0.13333333333333333	0.11538461538461539	0.950207468879668	0.12727272727272726	0.029045643153526972	0.0	0.0	0.0	ENTITY
14	Pematuhan	NNP	0.13333333333333333	0.34615384615384615	0.946058091286307	0.01818181818181818	0.004149377593360996	0.0	0.0	0.0	ENTITY
15	Standard	NNP	0.13333333333333333	0.3076923076923077	0.941908713692946	0.01818181818181818	0.004149377593360996	0.0	0.0	0.0	ENTITY
16	-LRB-	(	0.8	0.19230769230769232	0.9377593360995851	0.09090909090909091	0.02074688796680498	0.0	1.0	0.0	NON ENTITY

Fig. 6. Sample of Feature Extraction for FCM

Uppercase	Lowercase	isInitCapit	isDigit	isletterAn	MatchFea	isAllCapita	containsD	MatchFea	Term	POS	Class	id	Term-1	Term-2	Term-3	POS-1	POS-2	POS-3	CharLengt	Prefix	Suffix	removeVo	consonant	removeCo	vowelLen	currentTe	AverageAt
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	PREMIS	NN	O	1.0							6.0	PREMI	REMIS	PRMS	4.0	EI	2.0	premis	1.2
1.0	.0	.0	.0	.0	.0	.0	.0	.0	VCD/DVD	NN	O	2.0	PREMIS			NN			7.0	VCD/D	D/DVD	VCD/DVD	7.0	/	1.0	vcd/dvd	1.3
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	HARAM	JJ	O	3.0	VCD/DVD	PREMIS		NN	NN		5.0	HARAM	HARAM	HRM	3.0	AA	2.0	haram	1.0
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	DISERBU	VB	O	4.0	HARAM	VCD/DVD	PREMIS	JJ	NN	NN	7.0	DISER	SERBU	DSRB	4.0	IEU	3.0	diserbu	1.3
1.0	1.0	.0	.0	.0	.0	.0	.0	.0	,	,	O	5.0	DISERBU	HARAM	VCD/DVD	VB	JJ	NN	1.0	,	,	,	1.0	,	1.0	,	.4
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	INDIVIDU	NN	O	7.0	5	,	DISERBU	CD	,	VB	8.0	INDIV	IVIDU	NDVD	4.0	IIIU	4.0	individu	1.5
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	DITAHAN	VB	O	8.0	INDIVIDU	5	,	NN	CD	,	7.0	DITAH	TAHAN	DTHN	4.0	IAA	3.0	ditahan	1.3
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Jabatan	NNP	ORGANIZA	9.0	DITAHAN	INDIVIDU	5	VB	NN	CD	7.0	Jabat	batan	Jbtn	4.0	aaa	3.0	jabatan	1.3
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Sumber	NNP	ORGANIZA	10.0	Jabatan	DITAHAN	INDIVIDU	NNP	VB	NN	6.0	Sumbe	umber	Smbr	4.0	ue	2.0	sumber	1.2
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Strategik	NNP	ORGANIZA	11.0	Sumber	Jabatan	DITAHAN	NNP	NNP	VB	9.0	Strat	tegik	Strtgk	6.0	aei	3.0	strategik	1.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Dan	NNP	ORGANIZA	12.0	Strategik	Sumber	Jabatan	NNP	NNP	NNP	3.0	Dan	Dan	Dn	2.0	a	1.0	dan	.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Teknologi	NNP	ORGANIZA	13.0	Dan	Strategik	Sumber	NNP	NNP	NNP	9.0	Tekno	ologi	Tknlg	5.0	eooi	4.0	teknologi	1.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Jabatan	NNP	ORGANIZA	14.0	Teknologi	Dan	Strategik	NNP	NNP	NNP	7.0	Jabat	batan	Jbtn	4.0	aaa	3.0	jabatan	1.3
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Integriti	NNP	ORGANIZA	15.0	Jabatan	Teknologi	Dan	NNP	NNP	NNP	9.0	Integ	griti	ntgrt	5.0	Ieii	4.0	integriti	1.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Dan	NNP	ORGANIZA	16.0	Integriti	Jabatan	Teknologi	NNP	NNP	NNP	3.0	Dan	Dan	Dn	2.0	a	1.0	dan	.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Pematuha	NNP	ORGANIZA	17.0	Dan	Integriti	Jabatan	NNP	NNP	NNP	9.0	Pemat	tuhan	Pmthn	5.0	eaua	4.0	pematuha	1.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Standard	NNP	ORGANIZA	18.0	Pematuha	Dan	Integriti	NNP	NNP	NNP	8.0	Stand	ndard	Stndrd	6.0	aa	2.0	standard	1.5
1.0	.0	.0	.0	.0	.0	.0	1.0	.0	-LRB-	(	O	19.0	Standard	Pematuha	Dan	NNP	NNP	NNP	5.0	-LRB-	-LRB-	-LRB-	5.0	--	2.0	-lrb-	1.2
1.0	.0	.0	.0	.0	.0	1.0	.0	1.0	JIPS	NNP	ORGANIZA	20.0	-LRB-	Standard	Pematuha	(	NNP	NNP	4.0	JIPS	JIPS	JPS	3.0	I	1.0	jips	.9
1.0	.0	.0	.0	.0	.0	.0	1.0	.0	-RRB-	)	O	21.0	JIPS	-LRB-	Standard	NNP	(	NNP	5.0	-RRB-	-RRB-	-RRB-	5.0	--	2.0	-rrb-	1.2
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Jabatan	NNP	ORGANIZA	22.0	-RRB-	JIPS	-LRB-	)	NNP	(	7.0	Jabat	batan	Jbtn	4.0	aaa	3.0	jabatan	1.3
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Pencegaha	NNP	ORGANIZA	23.0	Jabatan	-RRB-	JIPS	NNP	)	NNP	10.0	Pence	gahan	Pncghn	6.0	eeaa	4.0	pencegaha	1.8
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Jenayah	NNP	ORGANIZA	24.0	Pencegaha	Jabatan	-RRB-	NNP	NNP	)	7.0	Jenay	nayah	Jnyh	4.0	eaa	3.0	jenayah	1.3
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Dan	NNP	ORGANIZA	25.0	Jenayah	Pencegaha	Jabatan	NNP	NNP	NNP	3.0	Dan	Dan	Dn	2.0	a	1.0	dan	.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Keselamat	NNP	ORGANIZA	26.0	Dan	Jenayah	Pencegaha	NNP	NNP	NNP	11.0	Kesel	matan	Kslmtn	6.0	eeaaa	5.0	keselamat	2.0
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Komuniti	NNP	ORGANIZA	27.0	Keselamat	Dan	Jenayah	NNP	NNP	NNP	8.0	Komun	uniti	Kmnt	4.0	ouii	4.0	komuniti	1.5
1.0	.0	.0	.0	.0	.0	.0	1.0	.0	-LRB-	(	O	28.0	Komuniti	Keselamat	Dan	NNP	NNP	NNP	5.0	-LRB-	-LRB-	-LRB-	5.0	--	2.0	-lrb-	1.2
1.0	.0	.0	.0	.0	.0	1.0	.0	1.0	JPJKK	NNP	ORGANIZA	29.0	-LRB-	Komuniti	Keselamat	(	NNP	NNP	5.0	JPJKK	JPJKK	JPJKK	5.0		.0	jpjkk	1.1
1.0	.0	.0	.0	.0	.0	.0	1.0	.0	-RRB-	)	O	30.0	JPJKK	-LRB-	Komuniti	NNP	(	NNP	5.0	-RRB-	-RRB-	-RRB-	5.0	--	2.0	-rrb-	1.2
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Jabatan	NNP	ORGANIZA	31.0	-RRB-	JPJKK	-LRB-	)	NNP	(	7.0	Jabat	batan	Jbtn	4.0	aaa	3.0	jabatan	1.3
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Siasatan	NNP	ORGANIZA	32.0	Jabatan	-RRB-	JPJKK	NNP	)	NNP	8.0	Siasa	satan	Sstn	4.0	iaaa	4.0	siasatan	1.5
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Dan	NNP	ORGANIZA	33.0	Siasatan	Jabatan	-RRB-	NNP	NNP	)	3.0	Dan	Dan	Dn	2.0	a	1.0	dan	.7
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Penguatku	NNP	ORGANIZA	34.0	Dan	Siasatan	Jabatan	NNP	NNP	NNP	14.0	Pengu	asaan	Pngtksn	7.0	euauaaa	7.0	penguatku	2.5
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Trafik	NNP	ORGANIZA	35.0	Penguatku	Dan	Siasatan	NNP	NNP	NNP	6.0	Trafi	rafik	Trfk	4.0	ai	2.0	trafik	1.2
1.0	.0	.0	.0	.0	.0	.0	1.0	.0	-LRB-	(	O	36.0	Trafik	Penguatku	Dan	NNP	NNP	NNP	5.0	-LRB-	-LRB-	-LRB-	5.0	--	2.0	-lrb-	1.2
1.0	.0	.0	.0	.0	.0	1.0	.0	1.0	JSPT	NNP	ORGANIZA	37.0	-LRB-	Trafik	Penguatku	(	NNP	NNP	4.0	JSPT	JSPT	JSPT	4.0		.0	jspt	.9
1.0	.0	.0	.0	.0	.0	.0	1.0	.0	-RRB-	)	O	38.0	JSPT	-LRB-	Trafik	NNP	(	NNP	5.0	-RRB-	-RRB-	-RRB-	5.0	--	2.0	-rrb-	1.2
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	PREMIS	NN	O	39.0	-RRB-	JSPT	-LRB-	)	NNP	(	6.0	PREMI	REMIS	PRMS	4.0	EI	2.0	premis	1.2
1.0	.0	.0	.0	.0	.0	.0	.0	.0	VCD/DVD	NN	O	40.0	PREMIS	-RRB-	JSPT	NN	)	NNP	7.0	VCD/D	D/DVD	VCD/DVD	7.0	/	1.0	vcd/dvd	1.3
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	HARAM	JJ	O	41.0	VCD/DVD	PREMIS	-RRB-	NN	NN	)	5.0	HARAM	HARAM	HRM	3.0	AA	2.0	haram	1.0
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	DISERBU	VB	O	42.0	HARAM	VCD/DVD	PREMIS	JJ	NN	NN	7.0	DISER	SERBU	DSRB	4.0	IEU	3.0	diserbu	1.3
1.0	1.0	.0	.0	.0	.0	.0	.0	.0	,	,	O	43.0	DISERBU	HARAM	VCD/DVD	VB	JJ	NN	1.0	,	,	,	1.0	,	1.0	,	.4
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	INDIVIDU	NN	O	45.0	5	,	DISERBU	CD	,	VB	8.0	INDIV	IVIDU	NDVD	4.0	IIIU	4.0	individu	1.5
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	DITAHAN	VB	O	46.0	INDIVIDU	5	,	NN	CD	,	7.0	DITAH	TAHAN	DTHN	4.0	IAA	3.0	ditahan	1.3
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	PREMIS	NN	O	47.0	DITAHAN	INDIVIDU	5	VB	NN	CD	6.0	PREMI	REMIS	PRMS	4.0	EI	2.0	premis	1.2
1.0	.0	.0	.0	.0	.0	.0	.0	.0	VCD/DVD	NN	O	48.0	PREMIS	DITAHAN	INDIVIDU	NN	VB	NN	7.0	VCD/D	D/DVD	VCD/DVD	7.0	/	1.0	vcd/dvd	1.3
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	HARAM	JJ	O	49.0	VCD/DVD	PREMIS	DITAHAN	NN	NN	VB	5.0	HARAM	HARAM	HRM	3.0	AA	2.0	haram	1.0
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	DISERBU	VB	O	50.0	HARAM	VCD/DVD	PREMIS	JJ	NN	NN	7.0	DISER	SERBU	DSRB	4.0	IEU	3.0	diserbu	1.3
1.0	1.0	.0	.0	.0	.0	.0	.0	.0	,	,	O	51.0	DISERBU	HARAM	VCD/DVD	VB	JJ	NN	1.0	,	,	,	1.0	,	1.0	,	.4
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	INDIVIDU	NN	O	53.0	5	,	DISERBU	CD	,	VB	8.0	INDIV	IVIDU	NDVD	4.0	IIIU	4.0	individu	1.5
1.0	.0	.0	.0	.0	.0	1.0	.0	.0	DITAHAN	VB	O	54.0	INDIVIDU	5	,	NN	CD	,	7.0	DITAH	TAHAN	DTHN	4.0	IAA	3.0	ditahan	1.3
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Kuala	NNP	LOCATION	55.0	DITAHAN	INDIVIDU	5	VB	NN	CD	5.0	Kuala	Kuala	Kl	2.0	uaa	3.0	kuala	1.0
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Lumpur	NNP	LOCATION	56.0	Kuala	DITAHAN	INDIVIDU	NNP	VB	NN	6.0	Lumpu	umpur	Lmpr	4.0	uu	2.0	lumpur	1.2
1.0	1.0	.0	.0	.0	.0	.0	.0	.0	,	,	O	57.0	Lumpur	Kuala	DITAHAN	NNP	NNP	VB	1.0	,	,	,	1.0	,	1.0	,	.4
1.0	1.0	.0	1.0	1.0	.0	.0	.0	.0	6	CD	DATE	58.0	,	Lumpur	Kuala	,	NNP	NNP	1.0	6	6	6	1.0	6	1.0	6	.6
.0	.0	1.0	.0	.0	1.0	.0	.0	.0	Oktober	NNP	DATE	59.0	6	,	Lumpur	CD	,	NNP	7.0	Oktob	tober	ktbr	4.0	Ooe	3.0	oktober	1.3
1.0	1.0	.0	1.0	1.0	.0	.0	.0	.0	2017	CD	DATE	60.0	Oktober	6	,	NNP	CD	,	4.0	2017	2017	2017	4.0	2017	4.0	2017	1.3

Fig. 7. Sample of Feature Extraction k-NN Classification

- Fuzzy C-means Clustering Method

The research proposed the fuzzy c-means method that applies to Malay named entity recognition task. The experiment is conducted by analyses the data that have done the pre-processing stage. The data that consists with features set is processed by using clustering method called as fuzzy c-Means algorithm. Fuzzy clustering is categorized as an unsupervised learning method that influential for data analysis and model's construction. Sakinah [19] stated that the desired number of clusters and preliminary predictions for each grade of membership is the beginning of the FCM algorithm. Therefore, for each cluster, all data points have their respective membership grades. The goal algorithm is to guide the central cluster to the optimum location in the data space by gradually updating the membership grade along with prototype (cluster centers) of the data point.

Suganya and Shanthi [20] stated that fuzzy c-means use fuzzy division to allow the sharing of data by all groups with different grades of membership between 0 and 1. They explain that the fuzzy c-means algorithm works by providing membership to each data point equivalent to each cluster center. Membership value given was calculated based on the distance between the center of the cluster and data points. The membership value of each data increases according to the closeness of data to the specified cluster center. This fuzzy C-means clustering makes a performance to cluster data by iteratively searching for a set of fuzzy clusters and the associated cluster centers which represent the data structure. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. The following Fig. 8 is the algorithm for fuzzy C-Means clustering.

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

4. *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2.*

Fig. 8. Fuzzy C-Means Clustering Algorithm

- K-Nearest Neighbors Algorithm

Classification is a machine learning technique in a supervised learning category that can be used to develop a model that describes the classification of important data. The development of the classifier is based on the class attributes involvement. Another method used in this experiment for classification is by using the K nearest neighbors algorithm. In pattern recognition, k-Nearest neighbors (k-NN) is one of the algorithms that are very simple, efficient, effective and most widely used classification methods. KNN classifier is a straightforward classifier in classifying data where sample data is classified according to the nearest neighbor class.

The K number of the nearest neighbors used has been given first in achieving high precision in the classification and relies heavily on the data set used. As the most basic instance-based method, the data used in the KNN algorithm are represented in vector space. There are two steps that are used in simple K nearest neighbor algorithm, firstly is finding the K training example that is closest to the unknown example and the second step is to pick the most classify occur for these K examples. The following Fig. 9 is the pseudo code of k nearest neighbors algorithm.

***k*-Nearest Neighbor**

1. Classify (X,Y,x) // *X:training data, Y:class labels of X, x:unknown* sample
2. Calculate "d (x, xi)" i =1, 2, ….., n; where d denotes the Euclidean distance between the points.
3. Arrange the calculated n Euclidean distances in non-decreasing order.
4. Let k be a +ve integer, take the first k distances from this sorted list.
5. Find those k-points corresponding to these k-distances.
6. Let ki denotes the number of points belonging to the ith class among k points i.e. $k \geq 0$
7. If ki >kj $\forall$ i $\neq$ j then put x in class i.

Note:
where x_i is the training data point

Fig. 9. Pseudo code of k Nearest Neighbors algorithm

VI. Result & Discussion

The collection of data is produced from PDRM news web pages in Malay languages cover on a few categories such as general topics, sports, crimes and others. Examples of the dataset before pre-processing are shown in Fig. 10 and after pre-processing in both Fig.6 and Fig. 7 respectively.

Fig. 10. Example of the dataset before pre-processing phase

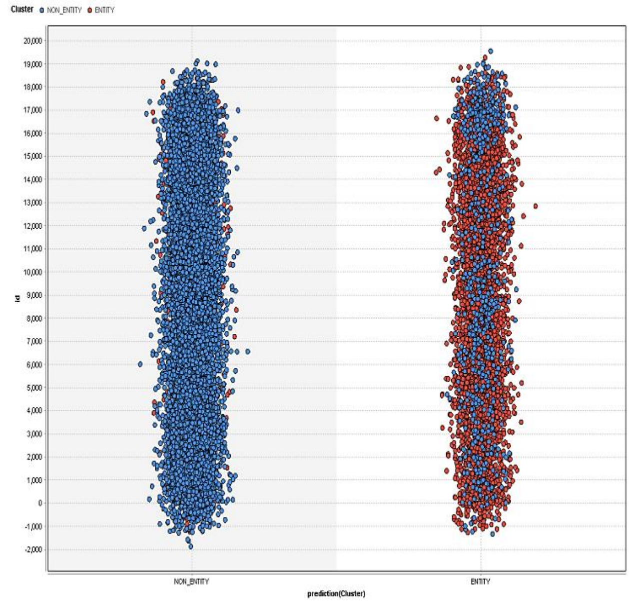

Fig. 11. Prediction FCM Clustering Chart

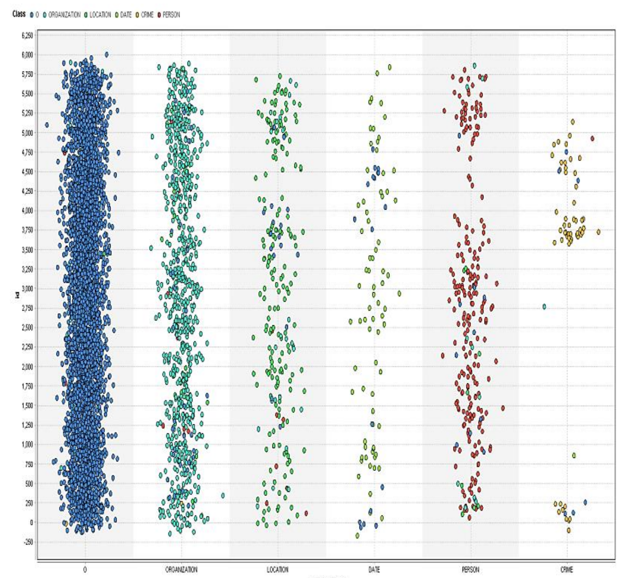

Fig. 12. k-NN Classification Chart

		True		class precision	Accuracy
		NON_ENTITY	ENTITY		
Predicted	NON_ENTITY	10451	1367	88.43%	88.51%
	ENTITY	646	5062	88.68%	
class recall		94.18%	78.74%		

Fig. 13. FCM Clustering Result

		True						
		OTHER	ORGANIZATION	LOCATION	DATE	CRIME	PERSON	class precision
Predicted	OTHER	4204	39	28	10	6	8	97.88%
	ORGANIZATION	33	737	38	0	3	10	89.77%
	LOCATION	16	23	194	0	0	7	80.83%
	DATE	17	0	0	76	0	0	81.72%
	CRIME	6	1	0	1	53	1	85.48%
	PERSON	10	10	8	0	0	233	89.27%
class recall		98.09%	90.99%	72.39%	87.36%	85.48%	89.96%	Accuracy 95.24%

Fig. 14. Result for Malay Named Entity Recognition

Based on prediction clustering chart of Fig. 11 and the cluster result in Fig. 13, the overall percentage accuracy had gave markedly good results based on clustering matching with 88.51% due to the calculation from all recall and precision results from all class entities. This accuracy was evaluated according to 17527 data samples, which have been pre-processed and undergone feature extraction. The precision result for NON_ENTITY class is 88.43% with 94.18% recall, whereas the precision for ENTITY class is 88.68% with 78.74% recall. Based on the analysis with other languages including English, NER has been implemented in the Malay language, which has the same characteristics as English in named entity recognitions such as capitalisation feature.

Then, for k-NN classification chart and result in the Fig.12 and Fig. 14 respectively, the prediction of classified entities consists of ORGANIZATION, LOCATION, DATE, CRIME, PERSON and OTHER is evaluated according to precision and recall. For ORGANIZATION entity, the precision is 89.77% and recall is 90.99%. For LOCATION entity, its precision is 80.83% and 72.39% recall. Next, the DATE entity produces 81.72% and 87.36% for both precision and recall respectively. For CRIME type entity, it produces both precision and recall as many as 85.48%. Then, for PERSON entity, it produces 89.27% for precision and 89.96% for recall. Lastly, for OTHER entity, the result for both precision and recall are 97.88% and 98.09% respectively.

VII. Conclusions

As conclude, the overall accuracy produced for Malay NER analysis is 95.24% during k-NN classification. This accuracy that can be an overall perspective of the evaluation process can be improved by undergoing another experiment by increasing the training dataset for a better result. This is because the percentage of accuracy increment for recognizing Malay entities liable on the model trained and suitable features sets used. The generated model from the small amount of dataset during the training process affected the assessment of the test's results. Therefore, the bigger dataset is needed to develop the Malay model to increase the results. As significant, the produced NER model can help to extract text data by determining exact text or term in the Malay language as named entity for the further police investigation.

In addition, the selection of appropriate features need to be continuously focused as these features can affect the performance of the NER model especially for Malay language because the language has complex structure in sentences.

The proposed Malay NER model can be further improved by increasing the corpus references in Malay for solving the problem of ambiguities for recognizing named entity types in Malay texts.

Acknowledgment

The publication is sponsored by Center of Advanced Computing Technology(C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka.

References

[1] Hosseinkhani, M. Koochakzaei, S. Keikhaee, and J. H. Naniz, “Detecting Suspicion Information on the Web Using Crime Data Mining Techniques,” Int. J. Adv. Comput. Sci. Inf. Technol., vol. 3, no. 1, pp. 32–41, 2014.

[2] H. Shabat and N. Omar, “Named Entity Recognition in Crime News Documents Using Classifiers Combination,” vol. 23, no. 6, pp. 1215–1222, 2015.

[3] A. Alkaff and M. Mohd, “Extraction of nationality from crime news,” J. Theor. Appl. Inf. Technol., vol. 54, no. 2, pp. 304–312, 2013.

[4] D. Jurafsky and J. H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,” Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. Speech Recognit., vol. 21, pp. 0–934, 2009.

[5] S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” Informatica, vol. 31, pp. 249–268, 2007.

[6] S. Morwal, “Named Entity Recognition Using Hidden Markov Model (HMM): An Experimental Result on Hindi , Urdu and Marathi Languages,” vol. 3, no. 4, pp. 671–675, 2013.

[7] D. Chopra and S. Morwal, “Named entity recognition in english language using Hidden Markov Model,” Int. J. Comput. Sci. Appications, vol. 3, no. 1, pp. 1–6, 2013.

[8] I. Ahmed and R. Sathyaraj, “Named Entity Recognition by Using Maximum Entropy,” Int. J. Database Theory Appl., vol. 8, no. 2, pp. 43–50, 2015.

[9] R. Sathya and A. Abraham, “Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification,” Int. J. Adv. Res. Artif. Intell., vol. 2, no. 2, 2013.

[10] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, and A. Sun, “Twiner: named entity recognition in targeted twitter stream,” in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’12, 2012, p. 721.

[11] S. Zhang and N. Elhadad, “Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts,” J. Biomed. Inform., vol. 46, no. 6, pp. 1088–1098, 2013.

[12] I. Diaz-Valenzuela, M. A. Vila, and M. J. Martin-Bautista, “On the Use of Fuzzy Constraints in Semisupervised Clustering,” IEEE Trans. Fuzzy Syst., vol. 24, no. 4, pp. 992–999, 2016.

[13] G. Castellano, A. M. Fanelli, and M. A. Torsello, “Shape annotation by incremental semi-supervised fuzzy clustering,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial

Intelligence and Lecture Notes in Bioinformatics), 2013, vol. 8256 LNAI, pp. 193–200.

[14] R. Alfred, L. Chin Leong, C. Kim On, and P. Anthony, “Malay Named Entity Recognition Based on Rule-Based Approach,” Int. J. Mach. Learn. Comput., vol. 4, pp. 300–306, 2014.

[15] S. Sulaiman, R. A. Wahid, S. Sarkawi, N. Omar, A. N. E. R. English, and G. Languages, “Using Stanford NER and Illinois NER to detect Malay Named Entity Recognition,” vol. 9, no. 2, pp. 2–5, 2017.

[16] S. Salleh, A. Asmai, H. Basiron, and S. Ahmad, “A Malay Named Entity Recognition Using Conditional Random Fields,” in 5th International Conference on Information and Communication Technology (ICoIC7), 2017, 2017, vol. 0, no. c, pp. 44–49.

[17] T. Cho, “Differences in the Romanized Spelling of Arabic Loanwords in Bahasa Melayu in Malaysia , and Bahasa Indonesia,” MELAYU J. ANTARABANGSA DUNIA MELAYU JILID 9 BIL. 2 2016, 2016.

[18] M. Zaidi, A. Rozan, and Y. Mikami, “Orthographic Reforms of Standard Malay Online : Towards Better Pronunciation and Construction of a Cross-language Environment *,” Education, no. March, pp. 129–159, 2007.

[19] S. Sakinah and S. Ahmad, “Fuzzy modeling through granular computing,” Thesis, 2012.

[20] R. Suganya and R. Shanthi, “Fuzzy C-Means Algorithm-A Review,” Int. J. Sci. Res. Publ., vol. 2, no. 11, pp. 2250–3153, 2012.