Named Entity Recognition using Fuzzy C-Means Clustering Method for Malay Textual Data Analysis

M.S.Salleh, S.A.Asmai, H.Basiron and S.Ahmad
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya,
76100 Durian Tunggal, Melaka, Malaysia
azlanrs87@gmail.com

***Abstract*— The Named Entity Recognition (NER) task is among the important tasks in analysing unstructured textual data as a solution to gain important and valuable information from the text document. This task is very useful in Natural Language Processing (NLP) to analyse various languages with distinctive styles of writing, characteristics and word structures. The social media act as the primary source where most information and unstructured textual data are obtained through these sources. In this paper, unstructured textual data were analysed through NER task focusing on the Malay language. The analysis was implemented to investigate the impact of text features transformation set used for recognising entities from unstructured Malay textual data using fuzzy c-means method. It focuses on using Bernama Malay news as a dataset through several steps for the experiment namely pre-processing, text features transformation, experimental and evaluation steps. As a conclusion, the overall percentage accuracy gave markedly good results based on clustering matching with 98.57%. This accuracy was derived from the precision and recall evaluation of both classes. The precision result for NON_ENTITY class is 98.39% with 100.00% recall, whereas for an ENTITY class, the precision and recall are 100.00% and 88.97%, respectively.**

***Index Terms*— Fuzzy C-Means; Malay language; Named Entity Recognition; Unstructured Textual Data.**

I. INTRODUCTION

Nowadays, studies analysing unstructured textual data have been increasing as these data provide a lot of valuable and useful information in various fields such as education, medicine, health care, political, military and security. With the advancement in the Internet of Things (IoT) that becomes increasingly pervasive, masses of unstructured textual data are accessible on the wide web world from various sources such as the online document and newspapers, web journals, Facebook, as well as Twitter or Instagram. However, without any suitable approaches, unstructured textual data might not reveal useful information. Some studies have shown an increased need for analysing unstructured text data. There are various textual data analysis tasks covered; for instance, text summarisation, part of speech tagging and named entity recognition. These include the study conducted by Castellucci *et al.* presenting the sentiment analysis by modelling Aspect Term Extraction as a sequential tagging task for classification problem to generalise several linguistic information [1] .

Many studies in NLP have been conducted for text analysis including Named Entity Recognition (NER). NER is one of the textual analysis approaches used to recognise entities in the open-domain text documents such as person, facility or organisation entities. The current NER process is commonly conducted based on an expert labelled document that consumes a lot of time resources. Most of these NER studies were conducted in processing English using various methods that include artificial intelligence and ruled-based methods. However, the NER study is rarely implemented in the Malay language to obtain valuable information from Malay language documents. One of the studies conducted in the Malay language is by using a ruled-based method. Alfred *et al.* have proposed a rule based method applied to named entity recognition task for Malay articles using Malay part-of-speech (POS) tagging features and contextual features to handle Malay articles [10]. Therefore, an improved strategy for automated NER process needs to be carried out. This paper aims at presenting an improved automated NER for Malay text analysis. Its objective is to save time resources to recognise the Malay entities according to the transformation features set used for Malay NER tasks. The dataset used in this study is Bernama Malay.

II. RELATED WORK

Many applications require text analysis for decision making process in any suitable situations that involve important documents. The application of artificial intelligence methods and research has become popular with the growing area of human-machine interaction that is ahead grounds for more investigations. These techniques covered three types of learning namely unsupervised learning (USL) technique, semi-supervised learning (SSL) technique and supervised learning (SL) technique.

The unsupervised learning (USL) technique is a kind of machine learning task used to make a conclusion from datasets comprising information of input data without labelled reactions. The goal of unsupervised learning is to train the machine to learn how to do something without expert intervene. Unsupervised learning can be used to bridge the gap between cause and effect observation of input and output. Clustering analysis is the most common method of unsupervised learning, which is used for exploratory data analysis to discover hidden patterns or groups of data. These unsupervised learning (USL) techniques have been also used in many text analysis study areas. Habib and Keulen proposed disambiguation clues to investigate unsupervised Semantic

Web-driven to improve named entity recognition [2]. On the other hand, the semi-supervised learning (SSL) technique is used in NER where the model is trained by using unlabelled data together with labelled data (human supervision). It is a technique that is combined with supervised learning and unsupervised learning. The goal is to utilise the unlabelled data during the training on labelled ones. This system uses training data to create a "model" of what is to be learned. The idea is for the system to general from a small set of examples to handle the new text of impunity. The training data consist of human-annotated tags for the named entities to be learned. There are various semi-supervised learning techniques that basically tried to automatically generate high-quality training data from an unlabelled corpus with bootstrapping being considered as the most common semi-supervised learning technique used. By using the semi-supervised learning technique, considerable improvement in learning accuracy can be achieved. Liao and Veeramachaneni used conditional random fields (CRFs) as a simple semi supervised algorithm for named entity recognition process [3].

Meanwhile, the supervised learning (SL) is a technique using labelled training data with the correct result to conclude the feedback on how the learning process is progressing in machine learning tasks. This supervised learning is employed to get the machine to learn the classification system created by the experts. This type of learning consists of a set of training examples where each example has an input object and the desired output value. For example, in text analysis, each text training data is labelled with desired output by human-annotated tags for machine learning. The ability to learn unnamed entities is an essential part of NER solution. Ahmed and Sathyaraj applied maximum entropy to recognise entity sets from a given text such as name, location and organisation[4].

III. Named Entity Recognition (NER)

Named Entity Recognition (NER) is a significant tool applied in NLP areas like information extraction or text retrieval, text summarisation and entity resolution [5]. Named Entities (NEs) refer to the proper nouns present in documents. In the NLP areas, NER task is an important task to be carried out especially for extracting information where the NER system serves to identify and classify vocabularies or words indicating the type of entity or proper nouns as the locations' name, organisations' name, persons' name, time, dates and facilities. Studies on the NER task have attracted many researchers to investigate the problems within the NER by utilising various languages and domains [6]. Others also defined that NER as one of the tasks in NLP that extracts information intended to classify text documents or corpus into several categories, which are defined as a person's name, locations' name, organisations' name, months, dates and time [7].

NER has been also applied in investigating various languages and field areas to obtain useful information in the unstructured text [6] such as that carried out by Asharef *et al.* extracting the information in the Arabic language for the criminal field [6]. This is due to an increase in crime rate within the Arab countries and various criminal information are available on the web, making this NER task important to be carried out to obtain information such as named entities from documents crimes. This was also due to the lack of studies in the field of criminal texts. Thus, the rule-based method has been used for the Arabic NER system associated with the domain of crime to identify and classify named entities in Arabic crime texts [6].

Other languages are also provided in NER including Bengali, which is the seventh most popular verbal language in the world and second in India [5]. This language is used as the national language of Bangladesh [5]. Ekbal and Bandyopadhyay explained that Support Vector Machines (SVMs) could be applied to develop the NER system for Bengali using different contextual information with a lot of features for predicting named entity classes. This SVMs method has been used for a variety of pattern recognition problems as it is known for its good generalisation performance [5].

NER tasks have been also undertaken in the field of medical and biomedical. In the biomedical field, Yao *et al.*, have applied named entity recognition using many machine learning approaches and produced good results on GENIA corpus [8]. They presented the Biomedical Named Entity Recognition method using deep neural network architecture with the multi-layer concept [8]. Besides, Bodnari *et al.* utilised NER to identify disorder named entities in electronic medical records by developed CRF supervised model based on a rich set of features of learning to predict interference named entities [9]. Moreover, NER in the Malay language has encouraged researchers to carry out studies in various fields like crime, education and economy with less studies conducted using artificial intelligence techniques. This is because the Malay language uses a different morphology from other languages including English. One of the studies conducting Malay NER analysis had used rule-based method for Malay articles. The ruled-based NER for Malay is basically executed based on POS tagging rules-based process for the Malay language and characteristics of contextual rules [10].

IV. A Malay Named Entity Recognition Approach

The Malay language is a language that appeals to the community in the field of NLP for ongoing studies to process information in many fields such as economics, business, and medical. This language is widely used in the document texts and named entity recognition task is one of the solutions to extract textual data from documents in terms of proper nouns or named entities. These proper nouns or named entities can be extracted from the feature set used that will be elaborated in sub section below. The purpose of this study is to produce useful information from Malay language documents using NER tasks. Using NER, various named entities can be extracted for analysing Malay language documents that are not widely used in various fields. Figure 1 shows the phases for Malay named entity clustering that begin with pre-processing data followed by the clustering of data using fuzzy c-means and lastly evaluation of accuracy for a correct match.

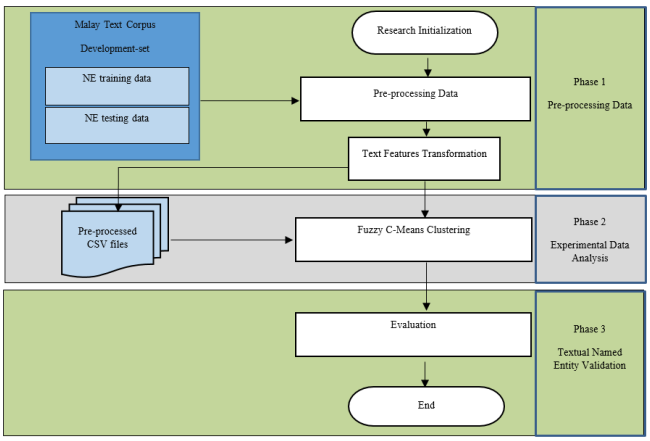

Figure 1: Process of Malay Named Entity Recognition Analysis

A. Pre-processing Data

The dataset used for this study is the Bernama Malay news in the format of plain text. As an initial process, pre-processing of data was performed where the text documents with a lot of unstructured data were separated into useful and meaningful units by doing tokenisation task, features extraction and annotation process. The estimating features are very important at this stage for analysis and evaluation process. This Malay NER analysis was conducted by deciding several features sets used. The following Figure 2 below displays the pre-processing step for NER task analysis.

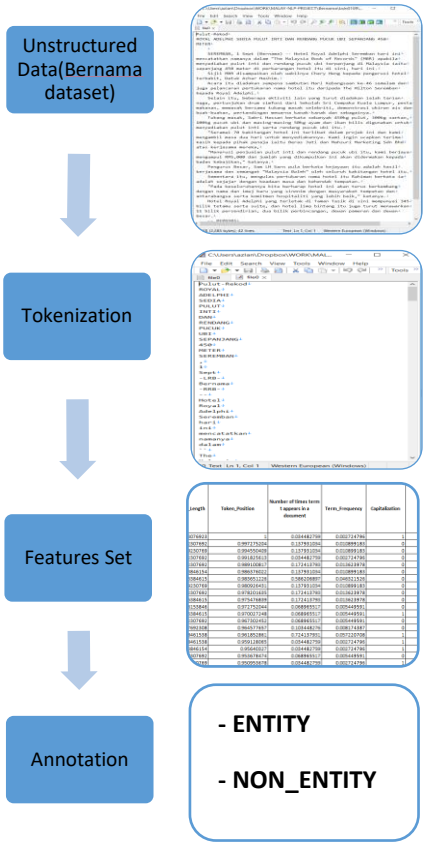

Figure 2: Pre-processing Data

a. Tokenisation

The text data file (.txt) presented in unstructured data from Bernama dataset comprised the sentences and paragraphs, which were tokenised as the process of separating text into valuable elements, words, phrases, symbols or digits called tokens. Tokens were presented in the list as inputs to simplify the next process.

b. Annotation

Then, the file was annotated by labelling with two categories namely ENTITY and NON_ENTITY. ENTITY category was labelled according to person, location, organisation and facility term entities. Meanwhile, NOT_ENTITY was labelled based on other terms. Every token in the file was also annotated with POS tagging bands such as CC, CD, NN, VB and others. The description on The Penn Treebank POS tag set is illustrated based on Table 1 [12].

Table 1
The Penn Treebank POS 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

B. Text Features Transformation

The features were determined based on compatibility with the named entity classes. Features such as part of speech value, character length, token position in the document, number of times term t appears in a document, Term Frequency (TF), capitalisation (proper noun), Uppercase, inverse document frequency (IDF) and TFIDF were selected as the features extraction for named entity recognition. These features were extracted in numeric type values for the analysis process.

a. Part of Speech (POS)

A part of speech (POS) is a category of words having similar grammatical properties. Words that were assigned to the same word part of speech generally displayed similar behaviour in terms of syntax playing similar roles within the grammatical structure of sentences. The Malay language also contains POS in their sentence structures. POS in the Malay language was identified as verbs, nouns and adjectives. Then, this POS feature was assigned with numeric type values for analysis process as outlined in Table 2.

b. Character Length

This feature is useful to filter and classify the word with similar length and helps other features to recognise token entities. In this study, the length of characters for each token was recognised by calculating the number of letters in each token word.

c. Token Position in Document

This feature determines the position of the token in every document by calculating each token word position in a document over with all token words in that document.

***No.TokInDoc*:**

$no.of\ token\ in\ document$

***No.NTokPos*:**

$$token\ position = \frac{no.of\ n\ token\ position}{(No.TokInDoc - No.NTokPos) + 1}, n = 1,2,3\ .. \quad (1)$$

d. Number of times term t appears in a document

This feature is used to count the appearances of term or token in a document divided with the total number of terms in the document to calculate term frequency (TF).

No.Term_t:

$Number\ of\ times\ term\ t\ appears\ in\ a\ document$

e. Term Frequency(TF)

This feature measures the frequency of a term that occurs in a document. Since every document is different in length, it is possible that a term would appear more frequently in long documents than shorter ones. Thus, the term frequency is often divided by the document length (the total number of terms in the document) as a way of normalisation.

TotalNo.TermInDoc:

$Total\ number\ of\ terms\ in\ the\ document$

$$TF(t) = \frac{No.Term\_t}{TotalNo.TermInDoc} \quad (2)$$

f. Capitalisation (PROPER NOUN)

This feature is very useful to recognise entities in a document. This is because named entities in the Malay language often start with capital letter. In the sample dataset as depicted in Figure 3, if the token word starts with a capital letter, the word is declared as 1 or true, otherwise it is declared as false or 0.

g. Uppercase

This feature is used to distinguish the entity structure with the structure of the common words in a document. Similar to capitalisation feature, this feature also applies the true or false (0 or 1) declaration.

h. IDF

Inverse Document Frequency measures the importance of a term. While computing TF, all terms are considered equally important.

TotalNo.Docs:

$Total\ number\ of\ documents$

N or No.DocsWithTerm_t:

$Number\ of\ documents\ with\ term\ t\ in\ it$

$$IDF(T) = log\_e\frac{(TotalNo.Docs)}{(No.DocsWithTerm\_t)} \quad (3)$$

$$idf_t = log\frac{N}{df_t} \quad (4)$$

TF-IDF

TF-IDF weight is the product of term frequency and inverse document frequency.

$$TF - IDF = TF \times IDF \quad (5)$$

The combination of term frequency and inverse document frequency is to produce a composite of every term weight in each document. The idf and TF-IDF are illustrated as in (4) and (5) above.

Table 2
POS Assign Value

POS	Assign Value
NN	1
NNP	2
RB	4
CC	3
CD	4
JJ	6
VB	7
IN	8
DT	9
.	10
,	11
(	12
)	13
O	14

After completing the transformation process of the text features, the dataset was stored in the csv file format for the experimental process. Figure 3 presents the sample of dataset after pre-processing step. The variables or features parameter are shown in the first column containing 10 features known as part of speech (POS), POS value, character length, token position in document, number of times term t appears in a document, term frequency(TF), capitalisation (PROPER NOUN), uppercase, IDF and TF-IDF.

Term	POS	POS_Value	Character_Length	Token_Position	Number of times term t appears in a	Term_Frequency	Capitalization	Uppercase	IDF	TFIDF	Class
Pulut-Reb	NN	0.071428571	0.423076923	1	0.034482759	0.002724796	1	0	0.95424	0.0026	NON_ENTITY
ROYAL	NNP	0.142857143	0.192307692	0.997275204	0.137931034	0.010899183	0	1	0.35218	0.00384	ENTITY
ADELPHI	NNP	0.142857143	0.269230769	0.994550409	0.137931034	0.010899183	0	1	0.35218	0.00384	ENTITY
SEDIA	RB	0.214285714	0.192307692	0.991825613	0.034482759	0.002724796	0	1	0.65321	0.00178	NON_ENTITY
PULUT	NN	0.071428571	0.192307692	0.989100817	0.172413793	0.013623978	0	1	0.25527	0.00348	NON_ENTITY
INTI	NN	0.071428571	0.153846154	0.986376022	0.137931034	0.010899183	0	1	0.35218	0.00384	NON_ENTITY
DAN	CC	0.285714286	0.115384615	0.983651226	0.586206897	0.046321526	0	1	-0.8166	-0.0378	NON_ENTITY
RENDANG	NN	0.071428571	0.269230769	0.980926431	0.137931034	0.010899183	0	1	0.35218	0.00384	NON_ENTITY
PUCUK	NN	0.071428571	0.192307692	0.978201635	0.172413793	0.013623978	0	1	0.25527	0.00348	NON_ENTITY
UBI	NN	0.071428571	0.115384615	0.975476839	0.172413793	0.013623978	0	1	0.25527	0.00348	NON_ENTITY
SEPANJANG	RB	0.214285714	0.346153846	0.972752044	0.068965517	0.005449591	0	1	0.65321	0.00356	NON_ENTITY
450	CD	0.357142857	0.115384615	0.970027248	0.068965517	0.005449591	1	1	0.65321	0.00356	NON_ENTITY
METER	JJ	0.428571429	0.192307692	0.967302452	0.068965517	0.005449591	0	1	0.65321	0.00356	NON_ENTITY
SEREMBAN	NNP	0.142857143	0.307692308	0.964577657	0.103448276	0.008174387	0	1	0.47712	0.0039	ENTITY
-	,	0.785714286	0.038461538	0.961852861	0.724137931	0.057220708	1	1	-1.2011	-0.0687	NON_ENTITY
1	CD	0.357142857	0.038461538	0.959128065	0.034482759	0.002724796	1	1	-0.0872	-0.0002	NON_ENTITY
Sept	NN	0.071428571	0.153846154	0.95640327	0.034482759	0.002724796	1	0	-0.0458	-0.0001	NON_ENTITY
-KRB-	-	0.857142857	0.192307692	0.953678474	0.068965517	0.005449591	0	1	-0.426	-0.0023	NON_ENTITY
Bernama	NNP	0.142857143	0.269230769	0.950953678	0.034482759	0.002724796	1	0	-0.4075	-0.0011	ENTITY
-RRB-	)	0.928571429	0.192307692	0.948228883	0.068965517	0.005449591	0	1	-0.426	-0.0023	NON_ENTITY
--	O	1	0.076923077	0.945504087	0.034482759	0.002724796	1	1	-0.2762	-0.0008	NON_ENTITY
Hotel	NNP	0.142857143	0.192307692	0.942779292	0.344827586	0.027247956	1	0	-0.0458	-0.0012	ENTITY
Royal	NNP	0.142857143	0.192307692	0.940054496	0.137931034	0.010899183	1	0	0.35218	0.00384	ENTITY
Adelphi	NNP	0.142857143	0.269230769	0.9373297	0.137931034	0.010899183	1	0	0.35218	0.00384	ENTITY
Seremban	NNP	0.142857143	0.307692308	0.934604905	0.103448276	0.008174387	1	0	0.47712	0.0039	ENTITY
hari	NN	0.071428571	0.153846154	0.931880109	0.137931034	0.010899183	0	0	-0.3468	-0.0038	NON_ENTITY
ini	DT	0.642857143	0.115384615	0.929155313	0.206896552	0.016348774	0	0	-0.6478	-0.0106	NON_ENTITY
mencatat	VB	0.5	0.423076923	0.926430518	0.034482759	0.002724796	0	0	0.95424	0.0026	NON_ENTITY
namanya	NN	0.071428571	0.269230769	0.923705722	0.034482759	0.002724796	0	0	0.95424	0.0026	NON_ENTITY
dalam	IN	0.571428571	0.192307692	0.920980926	0.068965517	0.005449591	0	0	-0.1919	-0.001	NON_ENTITY
:	O	1	0.076923077	0.918256131	0.172413793	0.013623978	1	1	-0.5082	-0.0069	NON_ENTITY
The	NNP	0.142857143	0.115384615	0.915531335	0.068965517	0.005449591	1	0	0.65321	0.00356	NON_ENTITY
Malaysia	NNP	0.142857143	0.307692308	0.91280654	0.103448276	0.008174387	1	0	-0.1597	-0.0013	NON_ENTITY
Book	NNP	0.142857143	0.153846154	0.910081744	0.034482759	0.002724796	1	0	0.47712	0.0013	NON_ENTITY
of	NNP	0.142857143	0.076923077	0.907356948	0.034482759	0.002724796	0	0	0.47712	0.0013	NON_ENTITY
Records	NNP	0.142857143	0.269230769	0.904632153	0.034482759	0.002724796	1	0	0.47712	0.0013	NON_ENTITY
.	O	1	0.076923077	0.901907357	0.137931034	0.010899183	1	1	-0.1249	-0.0014	NON_ENTITY

Figure 3: Sample of data after pre-processing

V. EXPERIMENT AND RESULT

For fuzzy c-means analysis, the dataset as shown in Figure 3 was processed based on class types of entities. The dataset was imported as a list of file name passed as an argument for the analysis process. The file should only contain the data separated by a space (or change the delimiter as required in split). Then, this data was analysed using a fuzzy c-means method developed by Dunn in 1973, which was then improved by Bezdek in 1981. Fuzzy c-means uses fuzzy division to allow the sharing of data by all groups with different grades of membership between 0 and 1[11].

The fuzzy c-means algorithm works by providing membership to each data point equivalent to each cluster centre. Membership value given was calculated based on the distance between the centre of the cluster and data points. The membership value of each data increases according to the closeness of data to specified cluster centre [11]. The analysis using fuzzy c-means method was conducted through Rapid Miner software with the structure design evaluating the result shown in Figure 4.

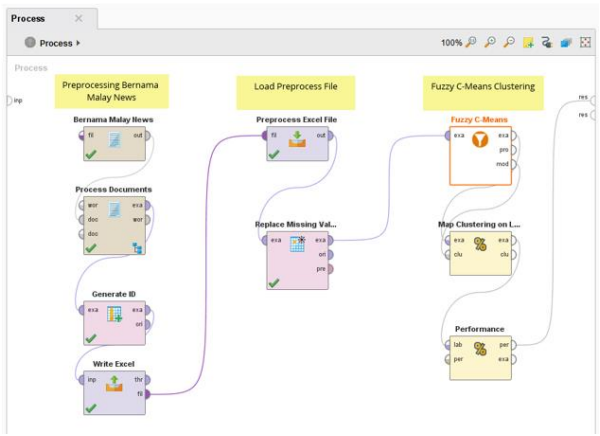

Figure 4: Data analysis using Fuzzy C-Means Method

VI. DISCUSSION

Figure 5 displays the scatter chart for fuzzy c-means analysis for the data. The data were plotted according to the cluster group and entities class. Based on the data plot above, the terms were clustered with two cluster values according to a class assigned, which are ENTITY class and NON_ENTITY class. Then, Figure 6 presents the scatter chart for predicting entity classes with the original entity class. The data plot shows that the result evaluation of prediction class was done based on every term from the dataset. Most data were clustered with a correct group based on original class and predicted class.

Based on Figure 7, the overall percentage accuracy had given markedly good results based on clustering matching with 98.57% due to the calculation from all recall and precision results from all class entities. This accuracy was evaluated according to 3084 data samples, which have been pre-processed and undergone feature transformation phase as shown in Figure 3. The precision result for NON_ENTITY class is 98.39% with 100.00% recall, whereas the precision for ENTITY class is 100.00% with 88.97% 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.

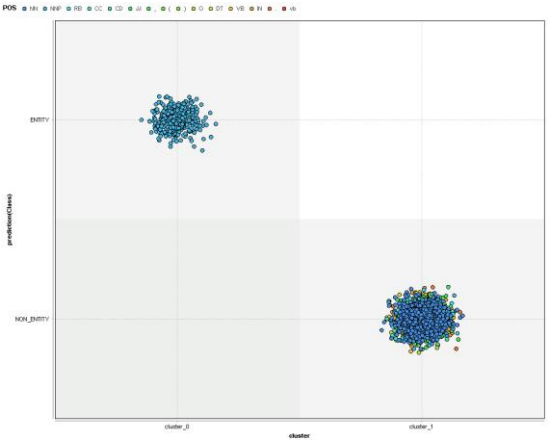

Figure 5: Scatter Chart for Fuzzy C-Means Analysis

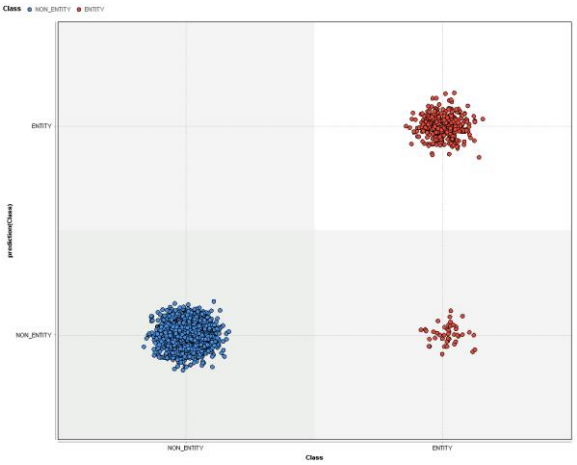

Figure 6: Prediction Chart

accuracy: 98.57%

	true NON_ENTITY	true ENTITY	class precision
pred. NON_ENTITY	2685	44	98.39%
pred. ENTITY	0	355	100.00%
class recall	100.00%	88.97%	

Figure 7: Result Evaluation

Based on the result evaluated, the capitalisation feature was considered one of the main features used to extract the named entities since every named entity or proper noun in POS category for the Malay language starts with capitalisation. For example, the capital letter can be found in a person entity like Muhammad Shahmi as well as location entities like Perlis and

Kuala Lumpur. Other than that, assigned value for POS tagging also helps this study to distinguish between proper nouns as keywords for an entity with other POS tag values.

VII. Conclusion

In conclusion, named entity recognition is one of the tasks undertaken in natural language processing where the NER process is implemented based on features relevant to the types of language examined such including the Malay language.

In this study, the fuzzy c-means method has been implemented on the proposed Malay NER using Rapid Miner software and dataset from Bernama Malay news. There are three phases involved for Malay NER analysis process starting from pre-processing data, fuzzy c-means analysis and lastly the evaluation of results. Using several features such as POS tag, POS assign value, characters length, token position in the document, number of times term t appears in a document, capitalisation and TF-IDF, an improvement has been observed in the accuracy of results for Malay named entity recognition. For subsequent experiments, the entities will be analysed based on their specific types such as a person, location, organisation and facility entities to find more classification differences between those entities while seeking the ambiguity between them.

Acknowledgment

This research work was funded by Fundamental Research Grant Scheme numbered FRGS/1/2015/ICT02/FTMK/02/F00288. Also, we thank the Ministry of High Education, Malaysia and Universiti Teknikal Malaysia Melaka for granting this research.

References

[1] G. Castellucci, S. Filice, D. Croce, and R. Basili, "UNITOR: Aspect Based Sentiment Analysis with Structured Learning," Proc. 8th Int. Work. Semant. Eval. (SemEval 2014), no. SemEval, pp. 761–767, 2014.

[2] M. B. Habib and M. Van Keulen, "Unsupervised improvement of named entity extraction in short informal context using disambiguation clues," CEUR Workshop Proc., vol. 925, pp. 1–9, 2012.

[3] W. Liao and S. Veeramachaneni, "A simple semi-supervised algorithm for named entity recognition," Work. Semi-Supervised Learn. Nat. Lang. Process., no. June, pp. 58–65, 2009.

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

[5] A. Ekbal and S. Bandyopadhyay, "Bengali Named Entity Recognition using Support Vector Machine," no. January, pp. 51–58, 2008.

[6] M. Asharef, N. Omar, and M. Albared, "ARABIC NAMED ENTITY RECOGNITION IN CRIME," J. Theor., vol. 44, no. 1, pp. 1–6, 2012.

[7] S. Morwal, "Named Entity Recognition using Hidden Markov Model (HMM)," Int. J. Nat. Lang. Comput., vol. 1, no. 4, pp. 15–23, 2012.

[8] L. Yao, H. Liu, Y. Liu, X. Li, and M. W. Anwar, "Biomedical Named Entity Recognition based on Deep Neutral Network," Int. J. Hybrid Inf. Technol., vol. 8, no. 8, pp. 279–288, 2015.

[9] A. Bodnari, L. Deleger, and T. Lavergne, "A Supervised Named-Entity Extraction System for Medical Text.," CLEF (Working, 2013.

[10] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach," Int. J. Mach. Learn. Comput., vol. 4, no. 3, pp. 300–306, 2014.

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

[12] A. Taylor, M. Marcus, and B. Santorini, *The Penn treebank: an overview*, (2003) *Treebanks*, 2003 pp. 5-22.