https://doi.org/10.17576/apjitm-2020-0902-02
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 9 No. 2 December 2020: 11 - 24
e-ISSN: 2289-2192

Received: 9 June 2020
Accepted: 12 October2020
Published: 1 December 2020

TEXT CLUSTERING FOR REDUCING SEMANTIC INFORMATION IN MALAY SEMANTIC REPRESENTATION

TUAN NORHAFIZAH TUAN ZAKARIA
MOHD JUZAIDDIN AB AZIZ
MOHD ROSMADI MOKHTAR
SAADIYAH DARUS

ABSTRACT

The generation of texts are dramatically increased in this era. A text basically consists of structured and unstructured texts. The enormous amount of unstructured texts can be easily perceived by humans, unfortunately cannot be simply processed by computer. It needs efficient techniques to reduce the information into more valuable vectors. In this article, we introduce text clustering method using Malay linguistic information to reduce the unstructured semantic information derived from Wikipedia Bahasa Melayu's articles. The proposed method uses the linguistic features in Malay language to cater the morphological issues of Malay words. We have incorporated semantic information from semantic lexical resource for Malay, which called Wikipedia Bahasa Melayu (WikiBM). Then, an experiment was conducted to evaluate the effects of text clustering to the semantic similarity value using gloss definition of WikiBM's article. We used Jaccard similarity to calculate the overlaps vectors from the text of WikiBM. Then, the correlation was computed using Pearson's correlation. The score between original text definition was compared to the new text definition using text clustering method. From the experiment, we can conclude that the correlation value was increased after the semantic information was reduced to more valuable vectors using text clustering method (from 0.39 to 0.43).

Keywords**:** text clustering, Malay, semantic representation, Wikipedia Bahasa Melayu, semantic similarity measurement.

INTRODUCTION

Texts are high-dimensional objects. Every word could be considered as an independent attribute. The huge size of attribute-value representation of texts make it a problem that need to be solved in most of natural language processing (NLP) applications. Reducing the dimensionality of texts can be undertaken based on the knowledge specified manually by experts, derived from corpus statistics, or computed from linguistic resources (Awajan, 2015; Alghamdi and Selamat, 2019). Text clustering is an application of cluster analysis to text-based documents. It uses natural language processing and machine learning to understand and categorize unstructured textual data. Text clustering can be accomplished based on two methods which are entity recognition and part-of- speech (POS) tags. Traditionally, text clustering is only based on keywords occurring in texts. It includes named entities (NE), which referred to names such as people, organizations and locations (Sekine, 2004).

The main objective of this work is to demonstrate that using the linguistic features of natural languages may yield an improved representation of texts. In our research, we proposed text clustering based on part-of-speech methods. Part-of-speech is a category to which type of word is assigned in accordance to its syntactical functions. In English, the main parts of speech are noun, verb, pronoun, adjective, determiner, adverb, preposition, conjunction and interjection.

Knowledge based lexical source such as WordNet and Wikipedia are very useful in variety of language processing tasks such as semantic searches (Ma et al., 2016; Liu et al., 2017; Zainodin et al., 2017; Kwan et al., 2015, Liu et al., 2019; Zhang et al., 2019) and word sense disambiguation (Chen et al., 2015; Moro et al., 2014; Scarlini et al., 2020; Bevilacqua and Navigli, 2020). Wikipedia has a tremendous information that is frequently updated by millions of voluntary donors. In semantic similarity application using Wikipedia, gloss definition is the important feature to be used. However, this text consists of various words with different types of words. The gloss definition is written by enormous number of volunteers which make the texts vary and unstructured. There are many unimportant words in the gloss definition. Sometimes, the words are redundant and repeated. Therefore, a sufficient method needs to be taken to reduce the gloss definition in Wikipedia's article to form a more concise and valuable vectors. Various studies have been conducted and several techniques have been proposed such as grouping of texts, bag of words and so on (Zhang et al., 2014; Saif et al., 2016; Wu et al., 2017).

Wikipedia Bahasa Melayu (WikiBM) (https://ms.wikipedia.org/wiki/Laman_Utama) is an online Malay encyclopaedia. WikiBM was developed inspired by English Wikipedia. WikiBM has 326,477 articles (as in 2019). WikiBM has gloss definition just as English Wikipedia and has the problem of unstructured texts as English Wikipedia. However, Malay language has different structures compared with English language. The previous techniques were efficient to English words but may not suitable for Malay word structure. Hence, a study needs to be done to evaluate and adapt the clustering technique towards Malay word.

RELATED WORKS

The semantic similarity can be measured using three basic methods, namely: path-based (Wu and Palmer, 1994; Rada et al., 1989; Leacock et al., 1998), information content (Seco et al., 2004; Zhou et al., 2008; Sanchez et al., 2012) and gloss-based (Lesk, 1987; Banarjee and Pedersen, 2003; Ponzetto and Navigli, 2010). The path-based and information content are measured using semantic taxonomy. The gloss-based method can be measured using the definition of the concept. Lexical sources such as WordNet and Wikipedia are very useful to measure the semantic similarity between words. Wikipedia contains a feature named gloss definition. The text similarity between the Wikipedia's definition of two concept can be used to measure their similarity. However, the texts are unstructured and long. Sometimes it has redundant and repeated words. The unstructured of Wikipedia's text may affects the similarity value of words.

Text clustering is the unsupervised technique of partitioning a set of texts into different groups that has similar (or related) to the other. This method aims at separating a set of texts into several groups with reasonable content without knowledge of any predefined categories (Manning et al., 2008). Text clustering is an efficient technique, largely used in pattern recognition, text mining and machine learning (Abasi et al., 2019). The similarity of text clustering usually defined through a representation of texts using some (or all) words or tokens that appear in them. Two texts are considered similar if they share the same words.

In this article, we use k-means clustering algorithm by Jain et al. (1999). K-means clustering is one of the partitioning algorithms widely used in the data mining. This algorithm cluster and partitioning n documents in the text data into k clusters representative around which cluster are built. The basic form of k-means algorithm is as follows:

ALGORITHM 1: k-means clustering algorithm

Input: Document set *D*, similarity measure *S*, number *k* of cluster
Output: Set of *k cluster*

initialization
Select randomly *k* data points as starting centroids
while *not converged do*
Assign documents to the centroids based on the closest similarity
Calculate the cluster centroids for all the clusters.
end
return *k clusters*

Stemming can increase similarity between texts, as it allows us to treat several related tokens as the same although the texts consist of different forms of a word. Part-of-speech (POS) tagging can be used to cluster the set of words into the same group and improve the similarity value.

Jaccard similarity is a measure often used to compare the similarity, dissimilarity and distance of the data set. Jaccard similarity coefficient is measured by the division between the number of features that are common to all divided by the number of properties, as shown below (Niwattanakul et al., 2013)

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

where,

A = set of words of text A,
B = set of words of text B,
$\cap$ = intersection of set A and text B,
$\cup$ = union of set A and B.

MALAY LINGUISTIC FEATURES

Malay is a language rich in morphology (word formation processes). A root word could be transformed to several new words through various morphological processes such as affixation, compounding and duplication. For example, the word '*kerja*' (work) can be transformed to other words like '*pekerja*' (worker) via affixation, '*kerjasama*' (cooperation) via compounding and '*kerja-kerja*' (works) via duplication.

These transformations may lead to the change of part-of-speech of the word. For example, the word '*sepak*' (kick) originally is a verb, but the word '*sepakan*' (kick) and '*bola sepak*' (football) become nouns. Moreover, a word may have two different part-of-speeches depending on the context of the sentence. For example, a word '*semak*' (check) has two different part-of-speeches; a noun in this sentence example: "*Bola itu ditendang oleh Ali ke dalam semak*" (The ball is kicked by Ali into the bushes) and a verb in this sentence: "*Sila semak jawapan anda sebelum menghantarnya*" (Please check your answer before submitting).

The change of part-of-speech could affect the similarity value between texts because the computer recognizes these words as different words although it has similar or related semantic meaning. Below is the example of the effects of morphological process in Malay language:

Text 1: "*Perkuburan merupakan tempat untuk pengebumian jenazah*"

Text 2: "*Kubur merupakan satu saluran tanah untuk pengkebumian*"

The word '*perkuburan*' (cemetery) and '*kubur*' (grave) are similar because they shared the same root word '*kubur*'. '*Pengebumian*' (funeral) and '*pengkebumian*' (burial) are also similar word from the root word '*bumi*' (earth). Unfortunately, computer cannot recognize these similarities and counting these words as different words. It affects the similarity value between these two texts.

The Malay sentence construction were based on the combination of several types of words. '*Kata nama*' (noun) and '*kata kerja*' (verb) are two main part of speeches in Malay language and often used to construct a sentence. The subject – verb – object rule was used in the construction of Malay sentence. The example below showed the Malay sentence construction using S-V-O rule:

Sentence: "Ali menendang bola".

'Ali' is a noun subject, menendang' is a verb while 'bola' is a noun predicate.

Moreover, noun and verb are two main components to construct a gloss in WordNet (Rubenstein and Goodenough, 1965). Therefore, in this article, we adapt the text clustering method based on POS tag (KN and KK). It means that our text will be clustered into KN and KK only to reduce the semantic information from definition text of WikiBM.

APPROACH

Text clustering is the application of clustering analysis to text-based documents. In our research, we used text clustering method based on Malay POS tags to reduce the unstructured texts into valuable vectors. Figure 1 shows the steps involved in this text clustering method.

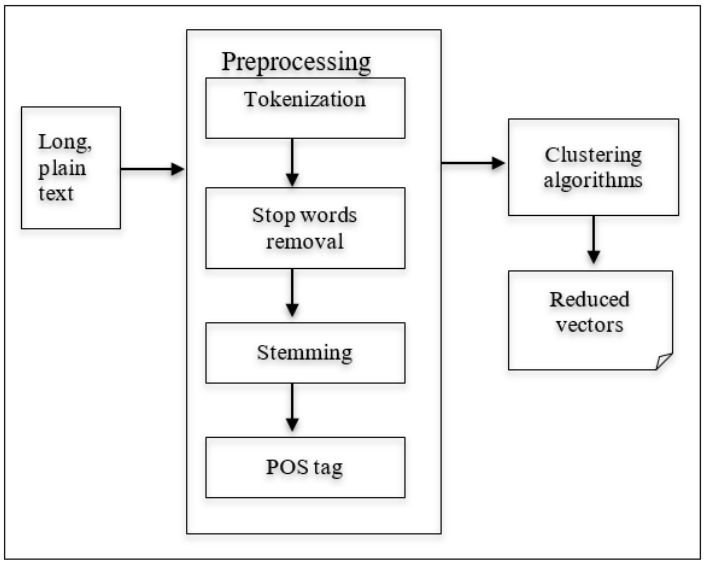

FIGURE 1. Steps in Text Clustering Methods

Based on Figure 1, the input of this text clustering method is long and plain text of gloss definition from WikiBM's article. These texts are long, plain and unstructured where they contain repeated words and less valuable in terms of semantic information. In this article, we chose first paragraph of WikiBM's article as our input. Then, the long text will be preprocessed. The preprocessing tasks include tokenization, Malay stop words removal, stemming and POS tagging.

Pre-processing task plays an important phase in this method. We need to pre-process the text and transforms it into a sequence of tokens where each token is labelled to identify its category. The pre-processing phase starts with tokenization where the tokenizer detects and separates the individual word. Then, the Malay stop words such as 'ialah', 'ini', 'hanya' and so on (Muhamad Taufik et al., 2005) were removed to clean and reducing the vectors of the text. Later, the words were stemmed to their root word. For example, the word 'perkuburan' (cemetery) were stemmed into root word 'kubur' to reduce the vector and improve the similarity value. The last step for pre-processing task is tagging the word into its part of speech such as 'kata nama' (noun), 'kata kerja' (verb), 'kata adjektif' (adjectives) and so on.

The next phase in clustering method is the clustering algorithms. The clustering algorithms used in this article is K-means algorithm using Malay Toolkit (https://malaya.readthedocs.io/en/latest/index.html). K-means algorithm can be classified as the simplest clustering algorithm and it was implemented in various and different methods (Wagstaff et al., 2001; Kodinariya and Makwana, 2013). Moreover, this algorithm works computationally faster than the hierarchical clustering and can works for a large number of variables (Dhanachandra et al., 2015; Ashour et al., 2018). The clustering algorithm for this article are including the choice of POS tag that will be used in this study.

'*Kata Nama* -KN' (noun) and '*Kata Kerja* - KK' (verb) are two important types of words in Malay language. Hence, these types of words were chosen as our choice of POS tag. The word with KN and KK tag will be collected and clustered using k-means algorithm from Jain et al. (1999). The output of this method is the clustered of reduced vectors based on KN and KK. The K-means algorithm used in this paper is as follows:

ALGORITHM 1: k-means clustering algorithm

Input: Document set D, similarity measure S, number k of cluster

Output: set of k clusters, based on verbs and nouns only

initialization

Select randomly k data as starting centroids

while *not converged* **do**

Assign documents to the centroids based on closest similarity

Calculate the cluster centroids for all the clusters

end

return *k clusters* based on verbs and nouns only

The algorithm selects randomly k data as starting centroids, and assign documents based on the closest similarity. The text was clustered based on 'kata nama' (noun) and 'kata kerja' (verb) only.

EXPERIMENT

Several steps have been taken for this experiment. The steps include data set collections, translating the data into Malay, collecting the gloss definition from WikiBM, clustering the text into reduced vectors, calculating the text similarity of each word-pair and calculating the correlation. Figure 2 shows the framework of text clustering used in this experiment.

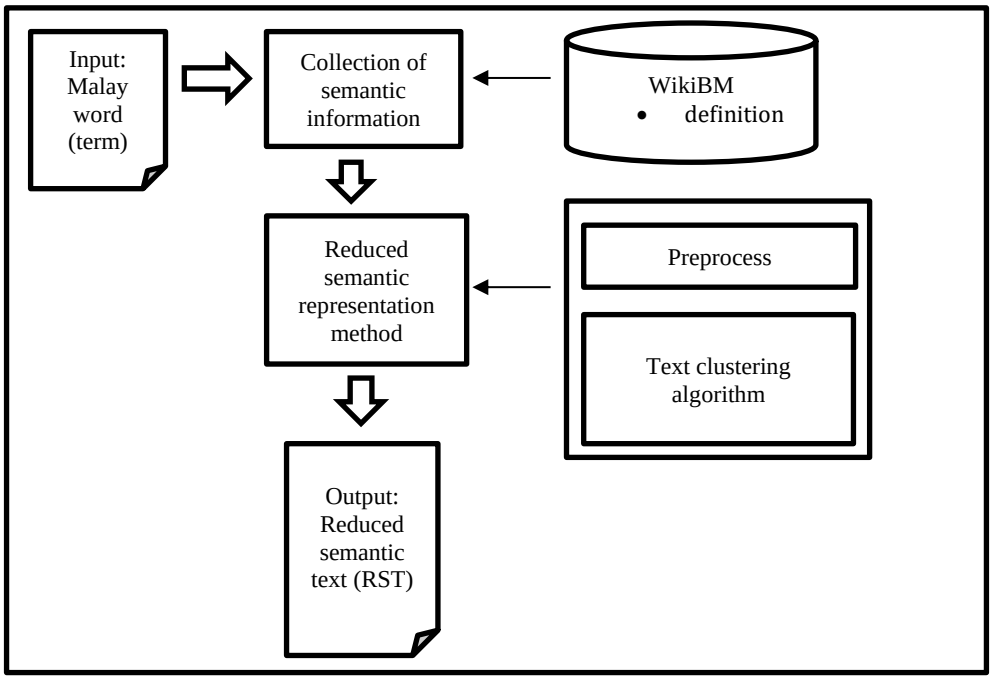

FIGURE 2. Framework of text clustering for experiment

Referring to figure 2, the input used in this experiment is the Malay word pair or we called them as terms, for example '*kereta*' (car) and '*automobil*' (automobile). The semantic information of these terms was collected from WikiBM's articles. The first paragraph of definition from WikiBM's article were used. Then, reduced semantic representation method

were applied including the pre-processing tasks such as tokenization, stop words removal, stemming and part of speech tagging. The algorithm of text clustering was used to cluster the words based on Malay nouns and verbs. The output of this experiment is the reduced semantic text (RST). The steps of this experiment were discussed later.

DATASET COLLECTION

Currently, there is no existing baseline dataset of Malay words containing semantic similarity value and semantic senses which available online and can be downloaded freely from internet that can be used in our experiment. Hence, we used the dataset from other language. In this article, we collected three English datasets named RG-65 (Rubenstein and Goodenough, 1965), MC-28 (Miller and Charles, 1991) and FL-353 (Finkelstein et al., 2002). RG-65 contains 65 word-pairs. The similarity between two words is given a score between 0-4 by two people, based on assessment of 51 subjects. MC-28 contains 28 word-pairs and is a subset of RG-65 data sets. Meanwhile, the FL-353 contains two sets of English words together with an assessment of the similarity by human assessors. The first set contains 153 word-pairs based on 13 subjects, while the second set contains 200 word-pairs, based on 16 subjects. This data set is rated on scale of 0-10. A total of 60 word-pairs (120 words) has been selected for our experiment. Table 1 shows the example of word-pairs in the dataset.

TABLE 1. The Example of Word-Pairs in The Dataset

Word 1	**Word 2**
car	automobile
tiger	fauna
monk	slave
beach	forest
forest	woodland
food	fruit
tiger	mammal
hill	forest
grave	forest
glass	jewel
bird	forest
glass	alloy
planet	star
professor	doctor
bread	butter
rain	lightning
king	queen
lobster	food
cup	object
century	year

Table 1 shows some word-pairs in English dataset. This dataset was very popular in semantic similarity researches not only for English research but also in other language researches such as Arabic (Saif et al., 2016).

MALAY DATA SET

The selected words from three English datasets were translated into Malay. Table 2 shows the translation of dataset from English to Malay.

TABLE 2. The Translation of English Dataset to Malay

English Word-Pair	Malay Word-Pair
car - automobile	kereta - automobile
tiger - fauna	harimau - fauna
monk - slave	sami - hamba abdi
beach - forest	pantai - hutan
forest - woodland	hutan - rimba
food - fruit	makanan - buah
Tiger - mammal	harimau - mamalia
hill - forest	bukit - rimba
grave - forest	makam - rimba
glass - jewel	kaca - permata
bird - forest	burung - rimba
glass - alloy	kaca - aloi
planet - star	planet - bintang
professor - doctor	profesor - doktor
bread - butter	roti - mentega
rain - lightning	hujan - petir
king - queen	raja - permaisuri
lobster - food	udang karang - makanan
cup - object	cawan - objek
century - year	abad - tahun

As seen in Table 2, the English dataset was translated to Malay using Online Dictionary: Cambridge Dictionary (https://dictionary.cambridge.org/) dan Kamus Oxford Fajar (Hawkins, 2006).

COLLECTING GLOSS DEFINITION

After the dataset were translated to Malay, we collected the gloss definition of each word from WikiBM. The first paragraph of the WikiBM's gloss definition were used as input texts. Table 3 shows the gloss definition collected from WikiBM.

TABLE 3. The gloss definition collected from WikiBM

Word 1	Gloss Definition	Word 2	Gloss Definition
kereta	Kereta ataupun automobil ialah kenderaan bertayar empat (biasanya) yang mempunyai enjin. Pada masa dahulunya ia dikenali sebagai kenderaan bermotor kerana 'motor' merujuk kepada enjin. Kereta mempunyai tempat duduk untuk pemandu dan sekurang kurangnya satu tempat duduk untuk penumpang dan selebih-lebihnya tujuh orang.	automobil	Kereta ataupun automobil ialah kenderaan bertayar empat (biasanya) yang mempunyai enjin. Pada masa dahulunya ia dikenali sebagai kenderaan bermotor kerana 'motor' merujuk kepada enjin. Kereta mempunyai tempat duduk untuk pemandu dan sekurang kurangnya satu tempat duduk untuk penumpang dan selebih-lebihnya tujuh orang.
sami	Rahib atau sami ialah seorang yang bersifat zahid dari segi keagamaan, seseorang yang menjaga minda dan badannya demi rohnya, dan sering bertapa sama ada sendirian atau dengan rahib-rahib lain di biara, terasing dari orang lain. Konsep ini berasal di zaman purba dan boleh dilihat dalam banyak agama dan falsafah. Istilah biarawan khusus untuk rahib lelaki, manakala rahib wanita digelar biarawati.	hamba abdi	Perhambaan adalah sistem sosial-ekonomi di mana manusia - hamba atau abdi - dirampas atau dinafikan kebebasan dan dipaksa untuk bekerja.
pantai	Pantai ialah bentuk bumi geologi sepanjang tebing atau tepi lautan, laut atau tasik yang mempunyai zarah longgar. Zarah-zarah yang membentuk pantai biasanya terdiri daripada batu seperti pasir, batu kerikil, batu bujur, atau batu bulat.	hutan	Hutan merujuk kepada kawasan yang ditumbuhi Zharif Ikhwan Parut al Sinan secara meliar dan bercampur-campur. Hutan berbeza dengan ladang yang pipinya ditanam oleh manusia dan pada kebiasaannya dipenuhi hanya satu atau dua spesies pokok sahaja.

The gloss definition from WikiBM is a very long text. Some words are redundant and repeated. It has root words such as '*zarah*' (particles) and '*bentuk*' (shape) and morphological words such as '*meliar*' (wildly) and '*ditanam*' (planted). These words can be reduced to a more valuable word using clustering method.

CLUSTERING TEXT INTO REDUCED VECTORS

In this experiment, we used text clustering method based on Malay POS tags to reduce the unstructured texts into valuable vectors. '*Kata Nama* -KN' (noun) and '*Kata Kerja* - KK' (verb) were chosen as our choice of POS tag. The word with KN and KK tag will be collected and clustered using k-means algorithm from Dhanachandra et al. (2015). The output of this method is the clustered of reduced text (RST) based on KN and KK. Table 4 shows the original text, clustered text by KN and KK, and clustered text by KN only.

TABLE 4. The Original and Clustered Text

Word	Original Text	RST (KN & KK)	RST (KN)

kereta	Kereta ataupun automobil i alah kenderaan bertayar em pat (biasanya) yang mempunyai enjin. Pada masa dahulunya ia dikenali sebagai kenderaan bermotor kerana 'motor' merujuk kepada enjin. Kereta mempunyai tempat duduk untuk pemandu dan sekurang kurangnya satu tempat duduk untuk penumpang dan selebih-lebihnya 7 orang.	kenderaan motor enjin orang duduk masa bagai tempat pandu rujuk kereta bertayar automobil punya tumpang	kenderaan motor enjin orang masa bagai tempat kereta automobil
hamba abdi	Perhambaan adalah sistem sosial-ekonomi di mana manusia - hamba atau abdi - dirampas atau dinafikan kebebasan dan dipaksa untuk bekerja.	nafi sistem manusia rampas kerja abdi sosial-ekonomi hamba	sistem manusia kerja abdi sosial-ekonomi hamba
pantai	Pantai ialah bentuk bumi geologi sepanjang tebing atau tepi lautan, laut atau tasik y ang mempunyai zarah longgar. Zarah-zarah yang membentuk pantai biasanya terdiri daripada batu seperti pasir, batu kerikil, batu bujur, atau batu bulat.	punya tepi diri geologi laut zarah pantai bujur pasir bentuk bulat batu bumi tebing kerikil	diri geologi laut zarah pantai bujur pasir bentuk bulat batu bumi tebing kerikil

From Table 4, we can see that the texts were reduced into shorter text. The morphological words were transformed into root words only. The redundant and repeated words were clustered. Then, we will calculate the gloss similarity of each word-pairs based on the original and clustered texts.

CALCULATING THE GLOSS SIMILARITY

After we reduced the definition texts using preprocessing tasks and clustering algorithms, we computed the text similarity of each word-pair using Jaccard similarity to calculate the overlap of vectors from definition of WikiBM's article. Jaccard Similarity is a statistic measurement used for measuring the similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets, as shown below:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (2)$$

The overlap vectors of word 1 and word 2 for each word-pair were computed using Jaccard Similarity. The similarity value of original text definition (long and plain text) and the

clustered text (based on KN and KK) will be compared with human score. We collected the similarity score from 12 respondents of Malay native speakers. Table 5 shows the similarity value of the dataset.

TABLE 5. The Similarity Value of The Dataset

Word 1	Word 2	Gloss Similarity (Ori)	Gloss Similarity (RST - Noun + Verb)	Human Score
kereta	automobil	1.00	1.00	0.98
rahib	sami	1.00	1.00	1.00
sami	hamba abdi	0.06	0.00	0.14
pantai	hutan	0.04	0.00	0.21
hutan	rimba	0.07	0.02	0.91
makanan	buah	0.08	0.07	0.67
hamba	hamba abdi	1.00	0.56	0.86
bukit	rimba	0.05	0.03	0.37
makam	rimba	0.07	0.03	0.30
kaca	permata	0.07	0.14	0.45
burung	rimba	0.05	0.00	0.31
kaca	aloi	0.03	0.07	0.56
planet	bintang	0.02	0.00	0.81
profesor	doktor	0.11	0.13	0.66
roti	mentega	0.06	0.05	0.62
hujan	petir	0.07	0.02	0.63
raja	permaisuri	0.09	0.07	0.86

CALCULATING THE CORRELATION

The correlation of original definition text and clustered text were calculated using Pearson correlation. Table 6 shows the correlation value of original definition text and clustered text based on Malay POS tag (KN and KK).

TABLE 6. Correlation Value of Original and Clustered Text

Gloss Similarity (Ori)	Gloss Similarity (RST - KN + KK)	Gloss Similarity (RST - KN)
0.39	0.43	0.43

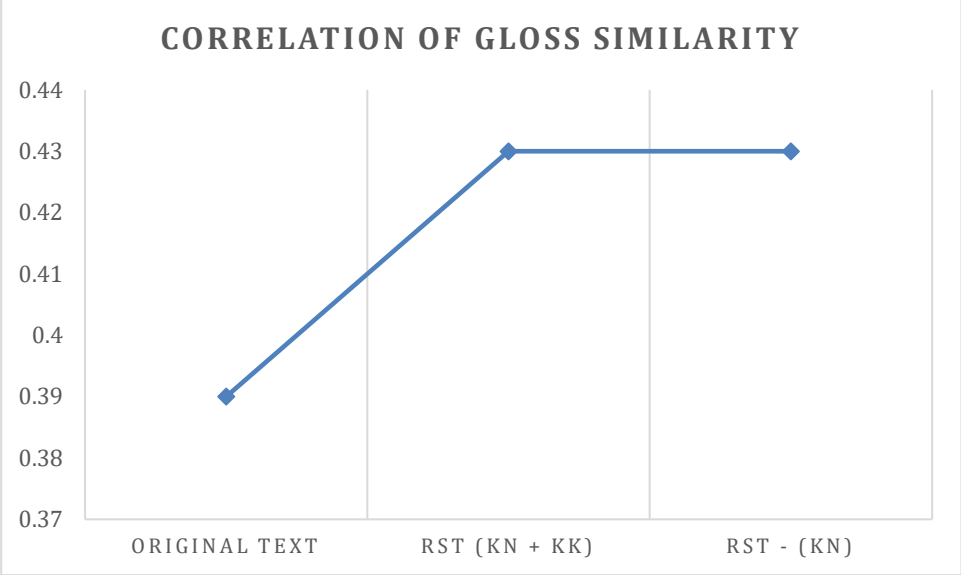

FIGURE 3. The Correlation of Gloss Similarity Between Original Text and Clustered Text

Correlation value increased when the semantic information was reduced using text clustering method based on Malay POS tag (from 0.39 to 0.43). The value of clustered text using KN + KK is similar to the value of clustered text using KN only (0.43). It means that both POS tag (KN and KK) can be used to reduce the semantic information. Hence, for our further research, we will be using text clustering based on KN and KK to reduce our data. This reduced data is important in our next study, which is the mapping of WordNet synset to Wikipedia article for Malay words.

CONCLUSION AND FUTURE WORK

This article describes a text clustering method using Malay linguistic information to reduce unstructured text of articles derived from Wikipedia Bahasa Melayu. The linguistic features of natural language improve the representation of the texts. In our research, we proposed text clustering based on part-of-speech methods. Two part-of-speech were chosen which are '*Kata Nama* -KN' (noun) and '*Kata Kerja* - KK' (verb). The word with KN and KK tag were collected and clustered using k-means algorithm. Then, an experiment was conducted to evaluate the effects of the text clustering method to the semantic similarity value using gloss definition from WikiBM's article. The Jaccard similarity was used to calculate the overlap vectors from the text of WikiBM. The correlation was computed using Pearson's correlation.

The clustered text in this article will be used in the next phase of our research which is the mapping of WordNet Bahasa's synset to WikiBM's article using a combination of Explicit Semantic Analysis (ESA) and similarity of definition (SD) technique.

REFERENCES

Alghamdi, Hanan M. and Selamat, Ali. 2019. Arabic Web Page Clustering: A Review. Journal of King Saud University – Computer and Information Sciences (2019), Volume 31, Issue 1, Pages 1-14.

Abasi, A. K., Khader, A. T., Al-Betar, M. A., Naim, S., Makhadmeh, S. N. and Alyasseri, Z. A. A. 2019. A text feature selection technique based on binary multi-verse optimizer for text clustering. In *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. pp. 1–6.

Ashour, A.S., Hawas, A.R., Guo, Y., and Wahba, M.A. 2018. A novel optimized neutrosophic k-means using genetic algorithm for skin lesion detection in dermoscopy images. Signal Image Video Process. Vol 12, 1311–1318.

Awajan, A. 2015. Semantic Vector Space Model for Reducing Arabic Text Dimensionality. *Digital Information and Communication Technology and Its Applications, IEEE*, pp. 129-135.

Banarjee, S, and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In IJCAI International Joint Conference on Artificial Intelligence, pp. 805-810.

Bevilacqua, M., and Navigli R. 2020. Breaking through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In Proc. of ACL.

Chen, X., Liu, Z., and Sun, M. 2015. A unified model for word sense representation and disambiguation. In *EMNLP* 2015, pp. 1025-1035.

Chenxi, L., Liang-Chieh, C., Florian, S., Hartwig, A., Wei, H., Alan, Y., and Li F. 2019. Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation. arXiv preprint arXiv:1901.02985.

Dhanachandra, N., Manglem, K., and Chanu, Y. J. 2015. Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm. Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), pp. 764-771.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman G., and Ruppin E. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1): 116-131.

Hawkins, J. M. 2006. Kamus Dwibahasa Oxford Fajar. Oxford Fajar Sdn. Bhd.

Jain, A. K., Murty, M. N., and Flynn, P. J. 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3): 264-323.

Kodinariya, T. M., and Makwana, P. R. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, vol 1(6): 90-95.

Kwan, B. R., Nebot, V., and Perez, M. 2015. Tailored semantic annotation for semantic search. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30: 69-81.

Leacock, C., Chodorow, M., and Miller, G. A. 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistic*, 24(1): 147-165.

Lesk, M. 1987. Automatic sense disambiguation using machine readable dictionaries. In *Association for Computing Machinery (ACM)*, pp. 24-26.

Liu, Z., Zheng, V. W., Zhao Z., Zhu F., Chang K. C., Wu M., and Ying J. 2017. Semantic proximity search on heterogeneous graph by proximity embedding. In Proc. 26th AAAI Conf. Artificial Intelligence 2017, pp. 154-160.

Liu, C., Chen, L. C., Schroff, F., Adam, H., Hua, W., Yuille, A. and Fei-Fei, L. 2019. Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. arXiv preprint arXiv:1901.02985.

Ma, B., Zhang, N., Liu, L., and Yuan, H. 2016. Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing and Management*, 52: 1-17.

Manning, C. D., Raghavan, P., and Schutze, H. 2008. Introduction to Information Retrieval. Cambridge University Press.

Miller, G., and Charles, W. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1–28.

Moro, A., Raganato, A., and Navigli, R. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transaction of the Association for Computational Linguistics*, 2014, 2: 231-244.

Muhamad Taufik, A., Fatimah, A., Ramlan, M., and Tengku Mohd, T S. 2005. Improvement of Malay Information Retrieval Using Local Stop Word. International Advanced Technology Congress.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. 2013. Using of Jaccard Coefficient for Keywords Similarity. In Proceedings of the International MultiConference of Engineers and Computer Scientists.

Ponzetto, S. P., and Navigli, R. 2010. Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems. In Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics, pp. 1522-1531.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1): 17–30.

Rubenstein, H., and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10): 627–633.

Saif, A., Ab Aziz, M. J., and Omar, N. 2016. Reducing explicit semantic representation vectors using Latent Dirichlet Allocation. *Knowledge-Based Systems*, 100: 145-159.

Sanchez, D., Batet, M., Isern, D., and Valls, A. 2012. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9): 7718–7728.

Scarlini, B., Pasini, T. and Navigli, R. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proc. of AAAI*, pp. 699– 709.

Sekine, S. 2004. Named Entity: History and future. Proteus Project Report.

Seco, N., Veale, T. and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in word net. In Proceedings of the 16th European Conference on Artificial Intelligence.

Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. 2001. Constrained K-means Clustering with Background Knowledge. Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 577–584.

Wu, Z. and Palmer, M. 1994. Verbs semantics and lexical selection. In *Association for Computational Linguistics (ACL)*, pp. 133-138.

Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J., and Xu, G. 2017. An efficient Wikipedia semantic matching approach to text document classification. *Information Sciences*, 393: 15-28.

Yiheng, Z., Zhaofan, Q., Jingen, L., Ting, Y., Dong, L., and Tao, M. 2019.Customizable architecture search for semantic segmentation. In CVPR, 2019.

Zainodin, U. Z., Omar, N., and Saif, A. 2017. Semantic Based on Features in Lexical Knowledge Sources. *Asia-Pacific Journal of Information Technology and Multimedia*. Vol 6 (1), pp. 39-55.

Zhang, C., Zhang, L., Wang, C. J., and Xie, J. Y. 2014. Text summarization based on sentence selection with semantic representation. In Proceedings – International Conference on Tools with Artificial Intelligence, ICTAI 2014, pp. 584-590.

Zhang, Y., Qiu, Z., Liu, J., Yao, T., Liu, D. and Mei, T. 2019. Customizable architecture search for semantic segmentation. In *CVPR*, 2019.

Zhou, L., Zhang, L., Feng, C., and Huang, H. 2008. Extracting Chinese multi-word terms from small corpus. In 3rd International Conference on Intelligent System and Knowledge Engineering, pp. 813–818.

Tuan Norhafizah Tuan Zakaria
Mohd Juzaiddin Ab Aziz
Mohd Rosmadi Mokhtar
Faculty of Information Science & Technology,
Universiti Kebangsaan Malaysia
tn_hafizah@yahoo.com, juzaiddin@ukm.edu.my, mrm@ukm.edu.my

Saadiyah Darus
Faculty of Social Science and Humanities,
Universiti Kebangsaan Malaysia.
adi@ukm.edu.my