Utilizing Noun-Verb Extraction in Enhancing Information Retrieval

Ahmad Zaki Yousef Al Abdala[1], Qusay Bsoul, Noor Hafizah Hassan[2], Mohd Shahidan Abdullah[3], Mohd Shahidan Abdullah[4]

[1]Advanced Informatics Department, Razak Faculty of Technology and Informatics, Kuala Lumpur, Malaysia
[2]Computer science Department, Faculty of Science and Technology, Universiti Sains Islam Malaysia
[3]Advanced Informatics Department, Razak Faculty of Technology and Informatics, Kuala Lumpur, Malaysia
[4]Advanced Informatics Department, Razak Faculty of Technology and Informatics, Kuala Lumpur, Malaysia
ahmadzaki22@yahoo.com[1], qusay068@usim.edu.my[2], noorhafizah.kl@utm.my[3], mshahidan@utm.my[4]

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: The increasing growth of the news and a large number of users on information retrieval (IR) has resulted in making the retrieval of documents complex and more difficult. The IR process consists of pre-processing, extraction and representation, feature selection and indexing, querying, and retrieving results. The weakness of IR is concerned with the process of extraction, where most important words focused on verbs or nouns. Unsupervised Feature selection is an important task in content classification for being among the most popular and effective methods for retrieval reduction. The use of the verbs and nouns as extraction was recently introduced in the IR technique to avoid irrelevant and redundant unsupervised features. This paper aims to enhance IR using noun-verb extraction using Word-Net and Krill Herd Algorithm as Unsupervised Feature Selection (KHUFS) combine with Simulated Annealing as Unsupervised Feature Selection to find the suitable retrieval of ranking. The external of Mean Average Precision (MAP) and Mean Average Recall (MAR) internal of Mean Average Distance (MAD) as measurements were used to verify the proposed retrieval of ranking. The results demonstrate that the proposed nouns verbs method extraction outperformed other extraction methods, in which the proposed extraction was 26.15 % using MAP measure and using MAR was 45.41%. In comparison with other unsupervised feature selection algorithms such as Harmony Search, Simulated Annealing, Particle Swarm, and Genetic Algorithm, the combined combination outperformed other unsupervised feature selection algorithms with an accuracy 26.163 % MAP, and 11.653% MAR. On the other side, the effect of use proposed extraction on the proposed unsupervised feature selection was 39.563% MAP, and 8.96% MAR. The other evaluation using the number of features, the using combined Krill Herd with Simulated Annealing number of features has decreased to more than 50 %, which the total feature in the dataset was 10682 features and after used the proposed was 4723 features.

Keywords: Information Retrieval, Unsupervised Feature Selection, Noun-Verb extraction

1. Introduction

The creation of machine learning is behind establishing algorithms of general-purpose that are characterized by practical values limited to a limited amount [1]. There are tight and direct relationships between machine learning and Artificial intelligence (AI), which refers to simulating human intelligence mimic their actions, or, AI can be applied to any machine that shows traits associated with a human mind such as learning and problem-solving [2].

The general purpose of AI is to distinguish and utilize data patterns as an important step towards multi-application purposes that can be applied in many fields including weather forecast, fraud detection, and medical diagnosis [3]. Machine learning, then, serves in at least two key domains as supervised or unsupervised learning. The supervised generates a mapping of the supervised (called sometimes as labeled training) data to an output of classes or predictions. Unsupervised learning performs the tasks without the need for supervision which, in this case, helps in finding the unknown patterns, clustering, and associating the data [4]. For the supervised is mainly used to classification and to produce a label or class which transfers (maps) the input objects to output values via a function called a classifier or a model. This model comprises items for classification while these items are termed as tuples [5] as they were reported to having a considerable amount of classification determined by the quality of extraction aiming at employing feature selection. The feature selection could take time tailed by heavy cost originated from the colossal amount of extraction terms and overflood the collections of the selection of a feature. The current study helps in providing accurate examples from the pool to be labeled. The idea of this approach lies in the possibility of establishing a training set that uses enhanced classifiers where the focus is to avoid considering all terms needed, which, means a severe reduction in the number of labeled feature selections needed to be employed [6].

Text mining can be classified in two ways, one includes a text or a document, while the second is text clustering. The document classification relies on a strategy in which the supervised learning task compromises pre-defining categories and labeling documents [7]. Further, the document classification can detect new events under certain criteria. Hence, the approach of the document classification is characterized by two phases -training and testing. The training phase uses the incorporated training set called corpus which can generate classifiers by assigning a subset for each category in the training phase. The mechanism, then, depends on using techniques called Information

Retrieval (IR) [8]. The goal of this mechanism is to search for the main features altering it as a classifier for each category. The testing phase, on the other side, is to test and evaluate the performance of searching for a test set to enabling categorizing classes known as unseen documents. The classification requires a comparison between the estimated categories and the pre-defined categories to measure the performance of the classification.

The real challenge to the classification task can be seen in the presence of the large volumes of data combined with a huge number of related documents [9]. The other challenge, yet less important, is the insufficient handling of the data and identification of new events [10]. Accordingly, the challenge can be categorized into two main aspects of the complete scenario [11]. The first aspect is about the weakness in the extraction by using words or terms from texts which require discriminating topics, sections, or document types. Traditionally, several successful techniques have been employed to solve the problem for the term extraction using a very well-known entity mechanism [12]. Adversely, the proposed technique was found unbeneficial for the extraction process for a huge feature amount. For this reason, another technique has been proposed in which the noun-verb (NV) extraction was used in document classification. Consequently, NV extraction was proven as an appropriate technique for the most crucial words or ‘terms’ found in the documents datasets. For the syntactic extracted terms, the central meaning is normally concealed to ignore the main meaning or sentence sensing [13].

In text mining, language ambiguity is widely considered as the highest-ranked problem in restricting the capability of understanding two or more possible senses. The problem is stemmed from being some words and/or phrases have multiple meanings resulting in ambiguity and, hence, classification of text or documents remained unimproved [9, 10]. Based on this argument, several authors have been trying to solve the above-mentioned problem with several techniques by decreasing the number of features using Chi-square and Information gain [5-8]. The new techniques consider any word replacement with its prospective concepts could expand the feature space; however, no enhancement in the performance [11, 12]. As a possible solution, ignoring the feature dependencies may provide the methodology of the feature selection (FS) [14, 15]. Yet, optimization still represents the most wanted goal to discriminate between the multi-solution involved in a set by employing meta-heuristic optimization. Optimization is principally to improve the performance of the classifier algorithm that is responsible to differentiate between hundreds or thousands of solutions. Optimization is a very advancing and progressing technique that can be achieved by several methods such as ant colony feature selection [16], Genetic Algorithm feature selection [17], Harmony Search (HS) feature selection [15], and Practical Swarm feature selection [14].

Briefly, there is a need to tackle these problems by minimizing or reducing solutions to, if possible, a single technique that is strong enough to overcome the problem of document classifiers. This study continues the effort made by several authors to employ the noun-verb extraction, which has been showing strong benefits as far as the construction and the extraction of the terms were concerned. The paper confines available various class labels in the document classifier to make it very helpful and easy for the term extraction technique.

This paper is divided into five section, firstly, the introduction of overall research is described. Secondly, a review of related work of machine learning approach is presented. Third section discussed on the development of noun-verb extraction.

2. Literature Review

Documentation is a very old technique that was used at different levels amongst people since the beginning of the writing era [18]. When computer technology has begun, paper writing has begun to disappear little by little, replacing it with the digital system, where information, documents, and literature were stored. Thus, a new era was born, in which the recovery of this information was of great importance [19]. Hence, the documentation in terms of storing and retrieval has been changed by exploring the digital and online text information extraction [20]. To achieve this technology, machine learning has been employed under two categories: supervised and unsupervised [21]. In this study, assigning an approach that is relevant to every document according to similarities of training and labeling in an unsupervised approach that is purely independent of any human intervention or interaction, at any time of execution, with the labeled documents. The focus, then, is on the classifiers which are using supervised techniques called detect classifier that involves the news, events, or topics. The unsupervised technique, on the other hand, adopts clustering, organizing, and categorizing documents that include news, events, or topics. The purpose of the classification of data and the documents is to show the connection between machine learning and natural language processing (NLP). The classification of the documents can be performed on a wide variety of texts such as the email systems in which discriminating the spasm or non-spasm email based on the binary system. For news and articles, for instance, different benchmark datasets are considered based on a particular class of the area under the process.

In the past two decades, various algorithms associated with the machine learning approach such as k-nearest neighbour (KNN) [22], Support Vector Machines (SVM) [23], Neural Networks (N-Net) [24], Linear Least Squares Fit (LLSF) [25] and Naïve Bayes (NB) [26]. The common factor between these algorithms is that all can be derived from natural occurrences. However, these algorithms are different in the capability of extraction documents where, for example, SVM, KNN, and LLSF, significantly outperform N-Net and NB.

It seems that finding out certain categories relevant to a specific category of searching or browsing operations are organized hierarchically. The new technique replaces the mere query posting to a specified text categorization system or search engine into a posting enabling finding out the categories in the documents of interest. Another issue is about the documents hierarchy where the classification is sub-grouped into smaller subdivisions [27].

Recently, as another line of research, there is a movement towards finding out areas of the extraction domain. The first stage of the extraction begins with the text which is represented by features according to an approach explained by [20]. Besides, a heuristic-based ontology was developed based on weighting ontology [28] such as the lexical ontology which can be utilized in various words available in the text [21]. Also, [29] has developed semantic features that are available for natural acceptance and, more specifically, [30] has established a general common type of semantic relationship. Meanwhile, WordNet equipped with capabilities for finding a solution was demonstrated by [31]. Besides, another sophisticated statistical semantic method was prescribed by [32].

The reference [33] has elaborated using various techniques to preserve the semantic procedure. As prescribed by [34] for WordNet and [35] for verbs with nouns approach, the semantic procedure was highly developed.

The most relevant proposal is the one that was revealed by [36] in which the NB classifier was adopted for verb-centric relationship extraction. The main application of the NB classifier was in the biomedical text by proposing an algorithm able to identify the relationship between sentences and phrases derived from a sentence. The central theme of this approach was to find out the entities which are classified by keys that are depicting the phrase. The missing or incomplete entities can be located using an algorithm specified in extracting phrases for the participating entities. The experimental results have shown that the precision of this approach ranges between 0.86 to 0.95, and, more recently, was estimated between 0.88 to 0.92 [21].

3. Development of noun-verb extraction

The search engines are characterized by the presence of a huge database connected to the Internet (known as WWW). The users can search for information using keywords or phrases (queries) looking for relevant information according to an architecture search engine compromises the contents and refinement, core search, and application interfaces s shown in Fig. 1 [37].

Reference [38] has stated that extraction of words or a phrase can be performed via algorithms that were built on the textual representation of web pages musing natural language processing (NLP). It is known that these algorithms are limited in interpreting sentences and extracting valued information [39]. Algorithm extraction is very effective in retrieving texts and split these texts in spelling and counting words; however, it requires high-level symbolic abilities [40]. The high-level symbolic appears in creation and propagation, manipulation of recursive, constituent structures; acquisition and access of lexical, semantic, and episodic memories, control of multiple learning/processing modules, routing of information, and grounding of basic-level language constructs [41].

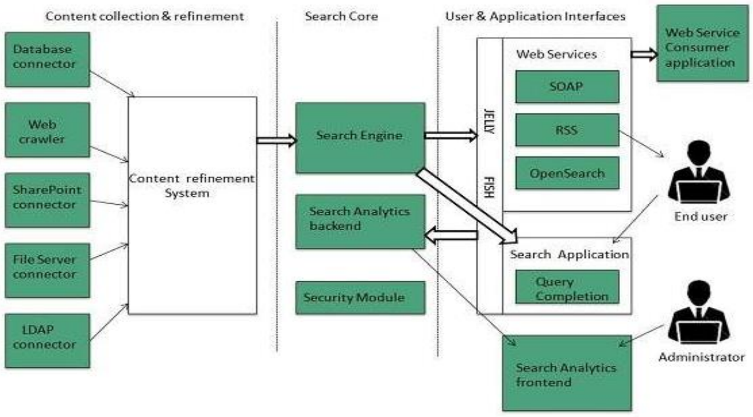


Fig:1Search engine architecture [37]

Combined detection of noun phrases with the use of the Word-Net as background knowledge. This was to explore better ways of representing documents semantically for clustering [42]. Here, three divers' clusters were used: k-means, bisecting k-means, and Hierarchical Agglomerative Clustering. Then, based on noun phrases as well as single-term analysis, they exploited different document representation methods. This was to analyze the effectiveness of hypernymy, hyponymy, holonymy, and meronymy. In this study, Reuters-21578 was used as the dataset.

The results showed that the best method is hypernymy. Further, [43] used hypernyms of the Word-Net. The purpose was to enhance document clustering using the Fuzzy-based Multi-label Document Clustering. In their experiment, they used many datasets benchmark. Their findings showed that using the hypernyms seemed better than not using the hypernyms. Also, the researchers were able to increase the accuracy and effectiveness of text mining. However, there was one weakness of their work, particularly, with reducing the dimensionality of terms.

In this work, new patterns of retrieval documents followed by detecting the relevant words on basis of noun-verb combination has been employed for feature selection. The novelty of this approach resulted in reducing the number of non-important features by ignoring the irrelevant and redundant words. The other significance of this study is the analytical procedure that relies on analysis in terms of documents. Having done so, taking the advantageous procedure, the information retrieval will be more accurate than standard information retrieval. The domain of the retrieved information by search engines for the websites is very important. This retrieval domain includes combining local search with global unsupervised feature selection can simplify the detection of the information retrieval [44]. The huge number of users which could hit a billion at the same time makes severe pressure on both the website and the search engines. For this reason, optimizing performance is indeed the most relevant part of this study.

4. Methodology

In the computer science field, developing methodologies requires testing, evaluating, and conducting experiments [8]. Regarding the experimental part, both computer science and natural science share the same concept; however, for different reasons.

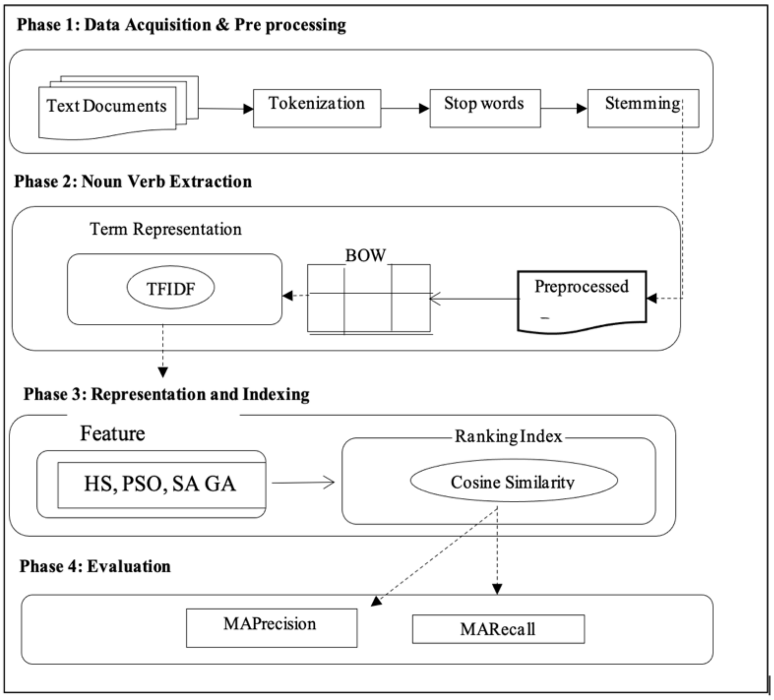
For instance, in natural science, scientists and scholars carry out experiments as the only way to obtain a better understanding of problems or natural phenomena being investigated. However, in computer science, scholars conduct such experiments because they find it hard to investigate the behavior or characteristic of a certain solution method analytically [45]. Research methodology in computer science has been utilized under either descriptive, correlational, explanatory, or exploratory [46]. However, complementary work requires proof by demonstration, empiricism, mathematical proof, and hermeneutics. There are some cases in which it is possible to classify one research method into more than one of these four research methodologies. For instance, if a particular study aims at developing or enhancing a given solution method for solving a particular problem, that, the researcher can mathematically prove such proposed method and demonstrate evaluated using computer programs as a proof such as used by [47]. Usually, such an implementation is regarded as an iterative process in which in each trial, a researcher can introduce or even modify some factors. However, proving a given proposed method mathematically in many real-world applications tends to be complex or challenging, and therefore, researchers resort to using empiricism proof [48].

For many years, researchers have described the most challenging real-world problems involved in computer science such as the problem being investigated in the present study as noun-phrase (NP) since no specific exact methods can handle such problems accurately and more effectively within tractable computational time [49]. Therefore, optimization approaches are used alternatively to solve unsupervised feature selection problems as they return good quality solutions within a reasonable amount of time. Fig. 2 shows the phases of groundwork, induction, and improvement, evaluation, and Comparison the quality.

The groundwork phase begins with identifying the most relevant works by comprehensively understanding the challenges in developing effective information retrieval. This was achieved by reviewing the state of the art of quality information retrieval in addition to the details of each process to identify the strengths and limitations of the current approaches. The induction phase represents identifying and testing the information retrieval provided demonstrating generality, consistency, and the performance of the proposed information retrieval frameworks. The identification phase is adopted similarly to simulating human thinking. The first problem in the feature selection is the dependent which ignores the feature dependencies. The meta-heuristic solves this problem using optimization unsupervised feature selection such as Harmony search, Genetic Algorithm, practical swarm optimization, and ant colony unsupervised feature selection.

The extraction of useful information from texts was proposed by [50] and then was evolved by [51]. Information retrieval searches and manipulates a large database collection via obtaining information relevant to a request from the collection of information resources via communication networks (e.g., the Internet, intranets, and extranets) [52]. Different types of applications that involve IR are available. Currently, information retrieval systems have been used worldwide by millions of people who are promoting business, education, and entertainment [53]. Fig. 2 shows the processes of information retrieval.

Fig. 1. Framework of the unsupervised feature selection experiments for information retrieval



Each web was represented by a vector t_i with term as the attributes and the attribute value being its Term Frequency × Inverse Document Frequency TFIDF weight (Salton & Buckley, 1988). This weight is a statistical measure used to evaluate how important a word is to a web in a collection or corpus. (TF×IDF) weighting is seen as the most popular method used for term weighting since it considers this property. By using this approach, assigning the weight of term i (w_i) to the number of times the term appears in the web is proportional (tf_i). tf_i weighting approach gives weight to the frequency of a term in a web N is the number of documents in the corpus with a factor discounting its importance in the case when it appears in most of the webs as described in Equation (1).

$$w_i = tf_i \log(\frac{N}{n}) \tag{1}$$

Evaluating information retrieval quality depends on considering the Mean Average Precision (MAP), Mean Average Recall (MVR), and number of features [54]. Meanwhile, the external quality depends on the labelled ranking of the information retrieval corpora. The corresponding methodology is to make a comparison between the resulting information retrieval and labelled ranking and to measure the extent to which queries from the same rank are assigned to the same retrieved. In the current study, MAP is used as an external quality measure, which is the most commonly used measures in information retrieval, and the internal evaluation used to evaluate each solutions in unsupervised feature selection called Mean Absolute Difference, The internal quality measure mainly depends on maximize the similarity of the value of the objective function.

The retrieval documents are viewed as either relevant or non-relevant depending on the user search knowledge. Based on an article of [55], the following explanations of the factors used in this study to assess the quality and the performance of the information retrieval.

The precision represents the proportion of relevant documents to all documents retrieved according to Equation (2).

$$Precision = \frac{number\ of\ relevant\ documents\ retrieved}{no\ of\ relevant\ documents} \tag{2}$$

In addition, the recall refers to the proportion of relevant documents retrieved out of all relevant documents available as shown in Equation (3),

$$\mathrm{Recall} = \frac{\text{number of relevant documents retrieved}}{\text{Total number of relevant documents in the collection}} \quad (3)$$

Besides the precision and recall, there are two means: mean average precision (MAP) and mean average recall (MAR). MAP is defined in Equation (4) in terms of the rank (*r*), the number of documents retrieved (*N*), the binary function (*rel*), and the precision (*p*),

$$\mathrm{MAP} = \frac{\sum_{r=1}^{n}(p(r)\ x\ \mathrm{rel}(r))}{\text{no of relevant documents}} \quad (4)$$

The other average, MAR, is normally considered after each relevant document is retrieved in which the MAP and executed.

The aim of this paper is to better understand the process of classification and to support the performance. Consequently, it is very difficult to generate good performance without detecting the weakness.

5. Results and Discussion

The datasets, for instance, showed improvement in retrieving when stemmed nouns are used as presented in MAPrecision measurement. The results have shown for Word-Net achieved 0.255136% compared to 0.230463% achieved by BOW, while the last extraction has achieved 0.120494%. Consequently, the datasets have been improved when stemmed verbs and nouns are used as shown using MARecall measurement achieving 0.266246% and 0.447438%, respectively, while the worst related to BOW achieved 0.512492%. The results in Table I suggest that using Word-Net to identify nouns-verbs is not only reduced the terms set dimensionality but also improved the retrieving for the datasets. The optimal performance of information retrieval when the score of MAPrecision is high and MARecall is low. related to the number of terms, it can be found that the verb as extraction reduce the number of terms more than 80%, then noun as extraction may reduce the number of terms up to 50%, which the BOW takes the biggest number of terms 10682.

TABLE I. COMPARISON BETWEEN THE THREE EXTRACTIONS METHOD ON BENCHMARK DATASETS

Extraction method	#of terms	MAPrecision	MARecall
BOW	10682	0.230463	0.512492
Noun	7103	**<u>0.255136</u>**	0.447438
Verb	**<u>1875</u>**	0.120494	**<u>0.266246</u>**

<u>best result: underline and bold</u>**

BOW has proven its superiority in verbs as extraction as shown using MAPrecision measurement; however, using MARecall measurement, the verbs extraction is better than BOW and noun as extraction. One of the purposes of this chapter is to compare the nouns as features with BOW. Table I has shown that the extraction of verbs as terms have been conclusive for the importance of those extracted for retrieving despite that the MARecall measurement was the lowest. The reducing number of terms pointed out that the verb, as extraction, has shown the lowest number of terms extracted. The information retrieval performance increased in the MAPrecision when the number of terms low with the noun as extraction. Therefore, it is suggested that the number of terms plays important role in the performance of information retrieval.

Based on the results presented in Table II, a technique that was recently used by Bounhas et al. (2020) in which a combination of nouns and verbs in the field of information retrieval as terms extraction. To evaluate the effectiveness of the combined (noun, verbs) identification method, this research used the same datasets that were used in previous sections by using cosine similarity as indexing ranking of documents and takes the average of 20 times independent runs. Table II compares retrieving performance between using a combination between nouns and verbs as terms extraction and comparing with the previous section. The proposed noun-verb as extraction got a score 0.26% in the MAPrecision measure than a noun, BOW, and verb as extraction. On the other hand, the proposed increases the number of features comparing with noun 97103) terms and verb (1875) terms compare to BOW which involves the highest number of terms with (10682) comparing with proposed extraction 8324 terms. Comparing between the extraction methods using MARecall, the best extraction which has the lowest score as shown in verb with 0.266246%, then noun (0.447438%), and the proposed noun-verb (0.454098922%0 while the worst is related to BOW (0.512492%). Still, the performance of information retrieval is inconclusive, where some of the numbers feature high and the MAPrecision high in noun-verb or low in BOW as extraction. On the contrary, the MARecall, which the number of features when low the MARecall is low but if the number of features is high the MARecall will be high.

TABLE II. THE RESULT OF THREE EXTRACTIONS METHOD ON BENCHMARK DATASETS

Extraction method	#of terms	MAPrecision	MARecall
BOW	10682	0.230463	0.512492
Noun	**7103**	0.255136	0.447438
Verb	**1875**	0.120494	**0.266246**
Noun with Verb	8324	**0.261475078**	0.454098922

****best result: underline and bold**

Tables III showed the results of the mean average precision and mean an average recall for five unsupervised feature selection comparing with proposed krill heard us unsupervised feature selection with their hybridizations. The highest MAP is the best in information retrieval for contrasting the lowest mean average recall. The Combined hybridized krill heard with SA outperform others hybridizations KH with SA as unsupervised feature selection with MAP (26.1%), KH (24.9%), then high-level, middle, low-levels of hybridizations and GA achieved the same results (24.8%), and the last in HS-1 (14.4%), which is the lowest. On another side, the lowest MAR got with Combined hybridized KHSAUFS (11.6%), middle hybridization (12.7%), low-level (13.1%), high- level, and GA (14.4%0, SA (48.2%) and without feature selection (51.2%), respectively.

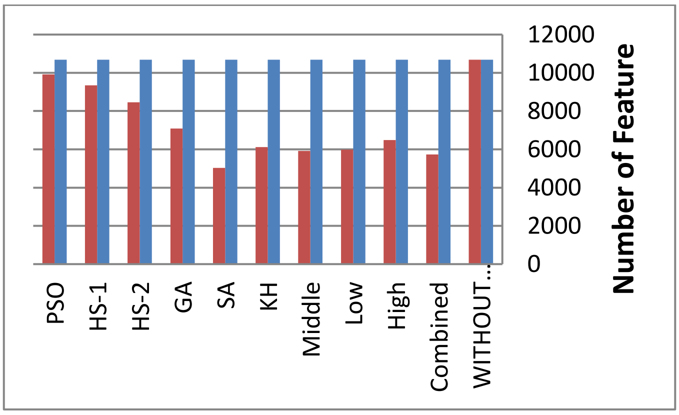
TABLE III. THE RESULT OF BOW AS EXTRACTIONS METHOD USING PROPOSED FEATURE SELECTION

Extraction method	MAPrecision	MARecall
combined of hybridization	**0.261632653**	**0.116530612**
high	0.249991837	0.144897959
low	0.249959184	0.131489796
middle	0.248367347	**0.127755102**
KH	**0.248500021**	0.13940243
HS-2	0.144477077	0.297724077
HS-1	0.154231531	0.400133939
PSO	0.156347804	0.473732922
SA	0.164326794	0.482066549
GA	0.248571429	0.144081633
without feature selection	0.23046298	0.512492143

****best result: underline and bold; **Second best result: bold**

The combined hybridized KHSAUFS achieves the best result in MAP and lowest in MAR. However, KH achieves the second rank in MAP, and middle hybridization got the second rank in MAR. Conversely, the worst algorithm got with SA in MAP, and the worst algorithm got in MAR was with/without used feature selection. This leads to study of the effect of the number of features on the MAP and MAR as shown in Fig. 3.

Fig. 2. Comparison between 6 algorithms and proposed unsupervised feature selection based on number of features and BOW extraction



The first finding was about the verbs, as important terms, in extraction for retrieving, as presented in the results. However, noun-verb was found better than others as shown in MARecall measurement. Identifying the number of features that affect the performance of information retrieval. Introducing the nouns as a term having the same importance as the verb as the term. Hence, the new extraction for extracting nouns with verbs, which have proven that, in most results, it is better than others. The improvement of proposed extraction and proposed feature selection in information retrieval was 33%. The number of terms using proposed noun-verb extraction and number of features using proposed combined hybridized KHSAUFS reduce to around 50%.

6. Acknowledgment

This work is supported by UNIVERSITI TEKNOLOGI MALAYSIA with cost center number Q.K130000.2656.17J22.

References

A. Holzinger, A. (2019). Introduction to MAchine Learning & Knowledge Extraction (MAKE). *Machine learning and knowledge extraction*, 1(1), 1-20.

B. Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., ... & Lee, S. (2019). *Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence*. USDOE Office of Science (SC), Washington, DC (United States).

C. Corea, F. (2019). *Applied artificial intelligence: Where AI can be used in business* (Vol. 1). Springer International Publishing.

D. Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE Access*, 7, 65579-65615.

E. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2), 261-318.

F. Wang, Y., Liu, Y., Feng, L., & Zhu, X. (2015). Novel feature selection method based on harmony search for email classification. *Knowledge-Based Systems*, 73, 311-323.

G. Wang, B., Xue, B., & Zhang, M. (2020). *Particle swarm optimization for evolving deep convolutional neural networks for image classification: Single-and multi-objective approaches*. In Deep Neural Evolution (pp. 155-184). Springer, Singapore.

H. Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., ... & Cheng, X. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6), 102067.

I. Mohammadi, M., Al-Fuqaha, A., Sorour, S., & Guizani, M. (2018). Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials*, 20(4), 2923-2960.

J. Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps*, 323-350.

K. Faris, H., Mafarja, M. M., Heidari, A. A., Aljarah, I., Ala'M, A. Z., Mirjalili, S., & Fujita, H. (2018). An efficient binary salp swarm algorithm with crossover scheme for feature selection problems. *Knowledge-Based Systems*, 154, 43-67.

L. Masnizah, M., Bsoul, B. W., Ali, N. M., Mohd Noah, S. A., Saad, S., Omar, N., and Aziz Abd, M. J. (2012). Optimal Initial Centroid In K-Means for Crime Topic. *Journal of Theoretical and Applied Information Technology*, 45(1), 19-26.

M. Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2020, March). *Fake news types and detection models on social media a state-of-the-art survey*. In Asian Conference on Intelligent Information and Database Systems (pp. 562-573). Springer, Singapore.

N. Aghdam, M. H., & Heidari, S. (2015). Feature selection using particle swarm optimization in text categorization. *Journal of Artificial Intelligence and Soft Computing Research*, 5(4), 231-238.

O. Wang, Y., Liu, Y., Feng, L., & Zhu, X. (2015). *Novel feature selection method based on harmony search for email classification*. Knowledge-Based Systems, 73, 311-323.

P. Guru, D. S., Suhil, M., Raju, L. N., & Kumar, N. V. (2018). An alternative framework for univariate filter based feature selection for text categorization. *Pattern Recognition Letters*, 103, 23-31.

Q. Ghareb, A. S., Bakar, A. A., & Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications*, 49, 31-47.

R. Oppenheim, A. L. (2013). *Ancient Mesopotamia: portrait of a dead civilization*. University of Chicago Press.

S. Bell, W. (2011). *Foundations of futures studies: human science for a new era: values, objectivity, and the good society* (Vol. 2). Transaction Publishers.

T. Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). *Using text mining techniques for extracting information from research articles.* In Intelligent natural language processing: Trends and Applications (pp. 373-397). Springer, Cham.

U. Al-Omari, O. & Omari, N. (2019). Enhanced Document Classification Using Noun Verb (NV) Terms Extraction Approach. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(1), 85-92.

V. Li, W., Chen, Y., & Song, Y. (2020). *Boosted K-nearest neighbor classifiers based on fuzzy granules.* Knowledge-Based Systems, 195, 105606.

W. Yu, D., Xu, Z., & Wang, X. (2020). Bibliometric analysis of support vector machines research trend: a case study in China. *International Journal of Machine Learning and Cybernetics*, 11(3), 715-728.

X. Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2020). *Machine learning for neural decoding*. Eneuro, 7(4).

Y. Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. P*olitical Analysis*, 143-168.

Z. Shafiq, M., Tian, Z., Sun, Y., Du, X., & Guizani, M. (2020). *Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city*. Future Generation Computer Systems, 107, 433-442.

AA. Misra, S., & Laskar, R. H. (2019). Development of a hierarchical dynamic keyboard character recognition system using trajectory features and scale-invariant holistic modeling of characters. *Journal of Ambient Intelligence and Humanized Computing*, 10(12), 4901-4923.

BB. Wang, R., Nellippallil, A. B., Wang, G., Yan, Y., Allen, J. K., & Mistree, F. (2018). Systematic design space exploration using a template-based ontological method. *Advanced Engineering Informatics*, 36, 163-177.

CC. Fodeh, S., Punch, B., & Tan, P. N. (2011). On ontology-driven document clustering using core semantic features. *Knowledge and information systems*, 28(2), 395-421.

DD. Zheng, H. T., Kang, B. Y., & Kim, H. G. (2009). Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences*, 179(13), 2249-2262.

EE. Dumais, S., & Chen, H. (2000, July). *Hierarchical classification of web content.* In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 256-263).

FF. Daghir, K. (2011). *Semantic Document Clustering for Crime Investigation* (Doctoral dissertation, Concordia University).

GG. Howard, M. (2012). *Semantic preserving text te presentation and its applications in text clustering*.

HH. Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, 42(4), 2264-2275.

II. Bsoul, Q., Salim, J., & Zakaria, L. Q. (2016). *Effect verb extraction on crime traditional cluster*. World Appl. Sci. J, 34(9), 1183-1189.

JJ. Yuan, S., & Yu, B. (2019). HClaimE: A tool for identifying health claims in health news headlines. *Information Processing & Management*, 56(4), 1220-1233.

KK. Lu, X., Chen, Y., & Li, X. (2019). Discrete deep hashing with ranking optimization for image retrieval. *IEEE transactions on neural networks and learning systems*, 31(6), 2052-2063

LL. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational intelligen Ce magazine,* 13(3), 55-75.

MM. Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine,* 9(2), 48-57.

NN. Dyer, M. G. (1995). *Connectionist natural language processing: A status report.* In Computational architectures integrating neural and symbolic processes (pp. 389-429). Springer, Boston, MA.

OO. Chowdhary, K. R. (2020). *Natural language processing. In Fundamentals of Artificial Intelligence* (pp. 603-649). Springer, New Delhi.

PP. Zheng, Z., Lan, Z., Park, B. H., & Geist, A. (2009, June). *System log pre-processing to improve failure prediction*. In 2009 IEEE/IFIP International Conference on Dependable Systems & Networks (pp. 572-577). IEEE.

QQ. Liu, Q., Jiang, H., Wei, S., Ling, Z. H., & Hu, Y. (2015, July). *Learning semantic word embeddings based on ordinal knowledge constraints*. In Proceedings of the 53rd Annual Meeting of the

Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1501-1511).

RR. Ziewitz, M. (2019). Rethinking gaming: The ethical work of optimization in web search engines. *Social studies of science*, 49(5), 707-731.

SS. Moret, B. M., & Shapiro, H. D. (2001). Algorithms and experiments: The new (and old) methodology. *Journal of Universal Computer Science*, 7(5), 434-446.

TT. Abutabenjeh, S., & Jaradat, R. (2018). Clarification of research design, research methods, and research methodology: A guide for public administration researchers and practitioners. *Teaching Public Administration*, 36(3), 237-258.

UU. Sabar, N. R., Ayob, M., Qu, R., & Kendall, G. (2012). A graph coloring constructive hyper-heuristic for examination timetabling problems. *Applied Intelligence*, 37(1), 1-11.

VV. Bartz-Beielstein, T., Chiarandini, M., Paquete, L., & Preuss, M. (Eds.). (2010). *Experimental methods for the analysis of optimization algorithms* (pp. 311-336). Berlin: Springer.

WW. Talbi, E. G. (2009). *Metaheuristics: from design to implementation* (Vol. 74). John Wiley & Sons.

XX. Subbaiah, S. (2013, February). Extracting knowledge using probabilistic classifier for text mining. In 2013 International Conference on Pattern Recognition, *Informatics and Mobile Engineering* (pp. 440-442). IEEE.

YY. Wang, S., Hsu, C. J., Trent, L., Ryan, T., Kearns, N. T., Civillico, E. F., & Kontson, K. L. (2018). Evaluation of performance-based outcome measures for the upper limb: a comprehensive narrative review. *PM&R*, 10(9), 951-962.

ZZ. Kraenzel, C. J. (2016). *U.S. Patent No. 9,288,000. Washington, DC: U.S. Patent and Trademark Office*

AAA. Amarasinghe, M., Kottegoda, S., Arachchi, A. L., Muramudalige, S., Bandara, H. D., & Azeez, A. (2015, August). *Cloud-based driver monitoring and vehicle diagnostic with OBD2 telematics*. In 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer) (pp. 243-249). IEEE.

BBB. Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.

CCC. Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5), 217-240.