ISSN 2278-3091
Volume 8, No.1, January – February 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse15812019.pdf
https://doi.org/10.30534/ijatcse/2019/15812019

# Enhanced Document Classification Using Noun Verb (NV) Terms Extraction Approach

**Omaia Al-Omari[1], Nazlia Omari[2]**
[1] Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia, Omaiaomari@yahoo.com
[2] Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia, Nazlia@ukm.edu.my

## ABSTRACT

The exponential growth in digital documents and the constantly increasing online information have called for the necessity and lead to classify the documents. Document classification is increasingly vital and indispensable for modern applications. Generally, documents comprise multiple terms of extraction. Here, the main concentration of the most important words is on verbs and nouns, which signify the topics and events. However, nouns and verbs technique or simply called Noun Verb (NV) as an extraction method will greatly enhance the performance of document classification. The aim and the implication of this research is to improve document classification performance by using and utilizing NV extraction to detect the class of a document. Three classifiers namely, K-Nearest Neighbor (KNN), Naive Bayes (NB), and Support Vector Machine (SVM) are used to compare the results. Nine benchmark datasets were employed in testing the proposed document classification. The anticipated classification was verified by evaluating its accuracy. The results exhibit that the verbs as extraction affect document classification. This encouraged the research work to combine verbs with nouns as extraction. The NV method extraction outperformed other extraction methods (e.g., Nouns, Bag of Word (BOW), and Verbs).

**Key words:** BOW extraction, Document classification, NV extraction, SVM classifier.

## 1. INTRODUCTION

Machine learning is mainly aimed at the creation of general-purpose algorithms of practical value using data of a limited amount. The approaches of machine learning to general AI are distinguished by data utilization and data patterns' discovery, and the application of machine learning can be seen in many areas, including weather forecast, fraud detection and medical diagnosis. There are two key domains of machine learning among which the first one is supervised learning, which comprises the task of machine learning to generate a mapping from supervised data or labelled training data to an output of classes or predictions. The second domain of machine learning is unsupervised learning. Classification task is a central domain of supervised learning, and it is performed to produce a function from input objects to output values called labels or classes. Meanwhile, the mapping or the function is termed a classifier or a model, whereas the input objects comprise items for classification and these items are termed instances, examples or tuples [1] reported the considerable reliance of a system's classification competence on the quality of the extraction and the employed feature selection. Nevertheless, it may take time and costly to collect such as colossal amount of extraction terms and over flood the collections of the selection of a feature. However, with enhanced document classification, this study could meticulously select examples from the pool to be labelled. Hence, when establishing the training set using enhanced classifiers, not all of the terms need to be extracted while all of feature selections to be labelled to do not need to be employed [2].

There are two classifications of text mining namely: Text or Document classification and Text clustering. Document classification refer to a supervised learning task, which requires pre-defined categories and labelled documents. Following certain criteria, document classification detect new events. Accordingly, the document classification approach includes two phases. The first one can be called as the training phase and second one can be termed as the testing phase. The former phase includes the use of the incorporated corpus known as training set. The training set are employed in generating classifier by allotting a subset for each category of the training set applied. Then, using some techniques of Information Retrieval (IR), they are processed in order find out the main features that are to be utilized as characteristics for each of the categories. Meanwhile, the testing phase comprises the testing and evaluation of the performance of the entity involved in employing the rest of the corpus, which is termed as a test set. For this purpose, the documents in each category are classed as unseen documents. Then, to measure the performance of the classification, comparison is made between the estimated categories and the pre-defined categories.

Large volumes of data and their related documents is an issue with the document classification task. The insufficiency of this handling of the data and identification of new events is yet another considerable challenge. Spotting of spam mails from this extraordinary volume of related documents is also a big challenge. The complete challenge can be categorized into two main aspects that relate to the complete scenario as mentioned. The first aspect relates to the weakness in the extraction of key data “as words or as terms” from texts to be discriminate between various topics and sections or documents types. Various successful techniques have been applied in the past in order to solve the problem for the term extraction with the use of the very well-known named entity mechanism [3] [4]. Unfortunately, the technique that was proposed by the authors was not at all beneficial when it comes to the extraction of data on feature or terms of large amount. Furthermore, yet another new technique which makes use of the noun and verb extraction on the document classification was to be established. Nouns and verbs or simply known as noun verb (NV) are the most appropriate and indeed the most crucial words or ‘terms’ that can be used for the extraction from the documents datasets. However, the extracted terms such as the syntactic ones will again conceal the central meaning of the term that is used, which will lead to the ignorance of the main meaning or the sense of the sentence.

In the past, various scholars [5-8] have tried to solve the problem with several techniques that have been employed to decrease the number of features (e.g., Chi-square, and Information gain). The uppermost problem in text mining is the ambiguity of the language, i.e. the capability of being understood in two or more possible sense. Because one word or phrase may have multiple meanings those can lead to ambiguity problem, text or document classification remains unimproved [9, 10]. For example, replacing a word with its prospective concepts opens the possibility of expanding or augmenting the feature space but the performance may remain the same [11, 12]. However, the major issue that is associated with the feature selection (FS) methods that it ignores the feature dependencies “ignore the relationship between the features” [1, 13-18]. Meta-heuristic optimization involves the use of the optimal solution that is available between all possible multi-solutions in the set, and with reference to, the context of the study under consideration. The main target is to improve the performance of the classifier algorithm between all the available hundreds or thousands of solutions. The most important feature that is available for the integral part of document classification is the optimization for the feature, so that, there were several methods that have been given by the past, for instance, Ant colony feature selection [14, 19], Genetic Algorithm feature selection[1, 20], Harmony Search (HS) feature selection[18, 21], and Practical Swarm feature selection[13, 17].

Out of the various problems that were presented in the previous study by various researchers, there was the need for solving these problems with the help of a single technique that was strong enough to overcome the problem of document classifiers. In these regards the technique for the noun verb extraction was used, which proved to be very beneficial as far as the construction and the extraction of the terms were concerned. The classification of the various class labels that are available in the document classifier makes it very helpful and easy for the term extraction technique is proposed in this research. The end result that is needed for the paper is to improve the extraction of important terms in document using Noun Verb (NV) as feature extraction.

## 2. LITERATURE REVIEW

The organization of the documents had become a very important issue in terms of research when the explosion of digital and online text information extraction is concerned. In order to achieve this particular idea mainly two machine learning approaches can be used to find out and to enhance this task: supervised approach, in this approach predefined category is assigned to every document based on the nature of the likeliness for a suggested set of training and labelled; and unsupervised approach, it is an approach in which there is no need for any human intervention or interaction with the labelled documents at any point for the entire process of execution. This thesis will be focused on the classifiers, where supervised techniques which called classifications detect the new news, events, or topics. On the other hand, unsupervised techniques which called clustering of documents is to organize or categorize the news, events, or topics to groups called clusters. The classification of data and the documents clearly illustrates that there is a connection between machine learning and the NLP. However, it is worth notification that the classification of the problems in which the data instances are done is in the form of text. As a very general example for the classification of the documents can find out that the categorization of the email systems can be done on the basis as either spam or non-spam. The task is taken into consideration in this mode can be a clear example for binary classification. Various benchmark datasets are taken into the consideration for the news articles that are belonging to a particular class of the area under the process. This is done in order to label all the documents under several categories which are taken into consideration.

Various algorithms that are based on machine learning approach that can be k-nearest neighbor (KNN)[22-24], Support Vector Machines (SVM)[25], Neural Networks (N-Net)[26], Linear Least Squares Fit (LLSF)[27] and Naive Bayes (NB)[28, 29] have been used for text classification. [27]draw a comparison of these techniques. As a conclusion that can be derived from all the algorithms that are presented by the above references, all the techniques are all comparable in nature. Whenever the category contains more than 300 documents as a whole. That said, when the number of positive training documents per category is less than ten, SVM, KNN, and LLSF outperform N-Net and NB significantly. In a hierarchical way, the process of finding out various categories for any particular document largely depends on the various facts that are searching and browsing operations. Instead of just posting a query to a particular text categorization system or search engine, it is much more convenient to find out the categories in which the documents posted for a particular

system of interest. To classify documents hierarchically, the problem for the classification can be subdivided to various smaller subdivisions. Hierarchical classification for all the documents, which are addressed, and explained by[28, 30].
A huge range of the research was carried out in order to find out the areas that are particularly the extraction domain. The main stage of the extraction starts from the text with is represented by the terms called ‘features’. The technique to find out most frequently occurring words can be used as proposed by (Lia et al. 2008). Yet another heuristic was given for the ontology based on weighting ontology[31]. Lexical ontology can be used for various words that are available inside the text[1]. Some of the semantic features that are available for natural acceptance can be taken into consideration prescribed by[12]. A very common and general type of semantic relationship that was established and it was used by[32]. WordNet and its capabilities for the finding of the solution was typically and productively derived by[30]. A more sophisticated and statistical semantic method, then the word net was also prescribed by[33]. The use of various techniques for preserving the semantic called as semantic preserving was given elevation by[34]. An approach which makes use of the Verb centric approach that was given by[35]. Pronominal anaphoric resolution technique for the semantic analysis was proposed by[36]. Yet another powerful semantic approach using WordNet was discovered by[37] and verbs with nouns method to a group of documents was derived by[38].

[35]proposed for the first time an idea for the extraction of that was based on the verb-centric relationship using the Naïve Bayesian classifier. With the help of checking a particular sentence that can be derived from a biomedical text, the algorithm that was proposed was sufficient enough to identify the relationship between sentences or otherwise. This algorithm had proposed the relationship that extent to present the phrase from the sentence. The group of people were searching for this; try to find out the entities, which were participating presentation and the found out those keys, which were involved around the relationship-depicting phrase. Furthermore, the algorithm that was capable for finding out the missing or incomplete entities that were joining various entity issues that involve using the extraction of phase for the participating entities. The results that were opting from this particular experiment had revealed the fact that amounts the balanced precision from 0.86 to 0.95 and it can be further stated that a recall from 0.88 to 0.92. The basis of this was the assessment on three biomedical datasets. A sentimental analysis can be applied on the datasets for the reviewed predictions [42].

## 3. RESEARCH METHODOLOGY

As mentioned, the purpose of this research is to examine and detect the text with the usage of document classification whereby the most important terms are extracted while preventing large amount of features. Hence, a comprehensive and descriptive overview of the methodology employed in the resolution of issues regarding the performance of document classification is presented in this study. In particular, the construction and performance of document classification are elaborated in this work. This research employs the standard research methodology in the domain of computer science. In view of that, establishing a method of solution, that is, the framework of quality documents classifiers for a given problem, becomes this study’s aim. As also employed by several other researches e.g., [39] the following phases are followed: groundwork phase, induction phase, improvement phase, evaluation and comparison quality phase.

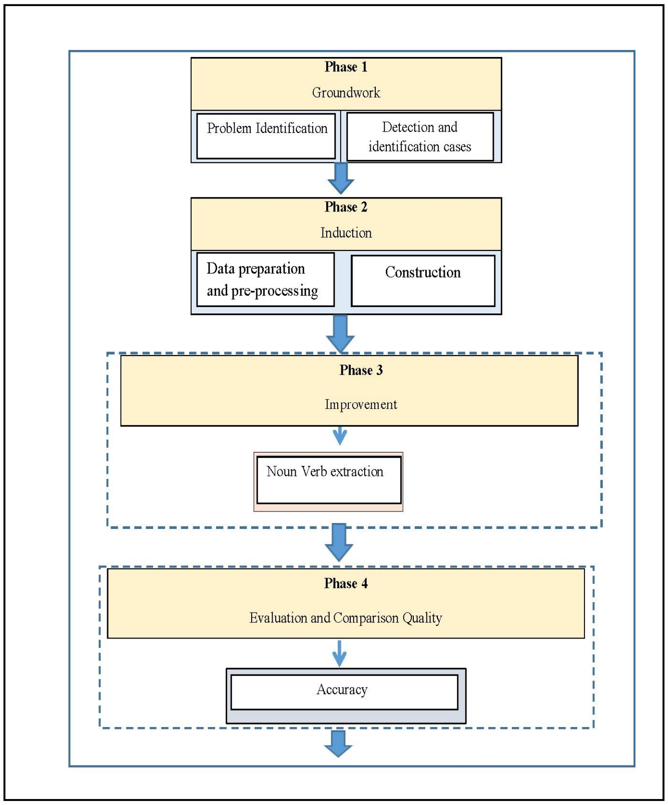

**Figure 1**:Phases involved in the study.

There are four phases included in the research method, which were also highlighted, which are:
1. Ground Work Phase: the groundwork is associated with the document classification problem. Accordingly, the literature on each problem is reviewed in order to detect the methods and algorithms with weaknesses in terms of document classifiers. Identification and detection of the problem as well as the identification case is discussed in this section. This phase involves the identification of two major problems relating to document classification. This becomes a test base in order to show generality, consistency and the performance of the frameworks of document classification proposed. As the Internet is dramatically expanding while online information is increasing, the detection and identification of large amounts of different text information become crucial. In fact, the issue relating to the detection and identification the new events of text is among the biggest challenges that researchers face. It should be noted that the approach of

classifiers could simulate human thinking, and owing to its salient features in mining, the application of this approach has been successful in diverse domains. Still, document classification for detecting and identifying new events, spam, or sentiments is complex due to the issue relating to performance and accuracy associated with the process of feature selection.

As laid down by[40], the assessment of the quality of classifiers employs three types of measures, which are f-measurement, purity, and accuracy. In general, the external measurement of quality relies on the test that are labelled for the document corpora. Accordingly, a comparative study between the resultant classifiers and labelled classes are made. Additionally, the degrees for the similarity of the document from similar class or category are chosen to the similar class is measured as well. Further, as an external quality measure, accuracy is employed in this work. In fact, in text / document classification, accuracy is the most generally employed measure. In problems of classification, measures of evaluation are generally specified using a matrix with the amounts of various examples that are accurately and inaccurately classified for each class. This matrix is called the Confusion Matrix. Accuracy (ACC) is the mostly used measure of evaluation in actual practice. It gives indication about how often the classifier makes the correct prediction, and evaluates the classifier's effectiveness using its percentage of prediction accuracy as expressed below:

$$ACC = ((TP+TN)/ (TP+TN + FP+FN)) *100$$

Where the sum of all correct predictions is divided by the total number of instances in the dataset. The advantage of accuracy this metric is easy to compute with less complexity; applicable for multi-class and multi-label problems; easy-to-use scoring; and easy to understand by human.

2. Induction Phase: The induction phase involves the designing of the initial process for document classification of the frameworks proposed. This comprises the iteration of the phase of improvement, application, and effect of the optimization feature selection phase employing a programming language that is deemed fit. This phase particularly concentrates on the pre-processing and creation of document classification. In the phase of construction, the focus is on the way the initial solution should be fabricated and iteratively enhanced through the framework of document classification proposed. Cross-Validation sometimes called rotation estimation, is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) is retained for subsequent use in confirming and validating the initial analysis.

3. Improvement Phase: The phase of updating is aimed at improving the eminence of the document classification. There are twin parts of this phase. During the first phase, the most important terms in the text is extracted. This is to increase terms that are deemed relevant terms while avoiding those that are regarded as non-relevant. Then, during the second phase, the effectiveness of the reduced amount of feature selection attained during the first phase is improved via the application of the proposed frameworks of document classification on a given area of the problem. This allows quality improvement of the terms while also decreasing the amount of features for classifiers. For document classification improvement, in this proposed research, the emphasis is given on improvement extractions process, the step of extraction comprises the extraction of nouns and verbs in order to increase the related terms while reducing the non-related ones. The terms "words" with the application of Word-Net for nouns and verbs identification are the outputs of this step.

4. Evaluation Phase: a comparative study between the resultant classifiers and labelled classes are made. Additionally, the degrees for the similarity of the document from similar class or category are chosen to the similar class is measured as well. Further, as an external quality measure, accuracy is employed in this work. In fact, in text / document classification, accuracy is the most generally employed measure. In the experiments executed in the research, nine variant datasets with differing properties are employed. The nine datasets are used in order that the performance of the algorithms can be thoughtfully analyzed and evaluated.

**Table 1**: Data Sets used in the Study

| **Document set** | **Source** | **# of documents** | **# of classes** |
|---|---|---|---|
| **DS1** | REUTERS-21578 | 4195 | 8 |
| **DS2** | TDT 2 and TDT3 of TREC 2001 | 1445 | 53 |
| **DS3** | 20-NEWSGROUPS | 3831 | 10 |
| **DS4** | CLASSIC 3 | 3892 | 3 |
| **DS5** | PAIRS 20-NEWSGROUPS | 686 | 2 |
| **DS6** | 6-EVENT CRIME | 247 | 6 |
| **DS7** | PAIRS REUTERS-21578 | 1194 | 2 |
| **DS8** | 5-GROUP 20-NEWSGROUP | 1888 | 5 |
| **DS9** | CLASSIC 4 | 7096 | 4 |

The first dataset contains the group of document classification of Reuters-21578. TDT2 and TDT3, which is the second dataset comprise 1,445 CNN news wire documents. The third dataset employed in this research, it comprises an assortment of 10,000 messages obtained from ten differing Usenet newsgroups. One well-known benchmark dataset used in text mining and classification is the Classic collection, which is the fourth dataset classic 3 that was chosen as the benchmark dataset. It is made up of 3892 documents from three different classes. The fifth dataset comprising of 20-newsgroups is employed in the performance assessment of algorithms on large datasets. The pair classes are obtained from 20 newsgroups consisting of 686 documents. The Crime dataset, which is the sixth dataset, was obtained from Bernama news. The seventh and eighth datasets are pairs reuters-21578, and 5-group 20-newsgroup respectively. The pair classes consisting of 1194 for reuters-21578, and 1888 documents for the five classes in 20-newsgroup.

## 4. NOUN VERB AS EXTRACTION METHOD

The classification algorithms generally depend on BOW (Bag of Words) as syntactic extraction. Semantic extraction is capable of resolving the problem relating to the weakness of syntactic extraction as it extracts the meaning. Accordingly, nine datasets are used in the assessment of the extraction methods. However, there is one main challenge when making comparison between the previously published results, that is, lack of uniformity, particularly with respect for the benchmark data and the baseline algorithms of applied classifiers. The three popular classification algorithms: K-Nearest Neighbor, Naive Bayes and the Support Vector Machine that have been used overall in [5, 8, 27]. Each one of the three classifier algorithms have its own paths, which can be used in order to find out what distinguishes the former algorithm from the other two algorithms. The main highlighting area of focus for this section is actually the addressing the strengths and the weaknesses for all the respective algorithms that can be used for the classification related to the high dimensionality and the features. The bag-of-word contains various numbers of words' representation, which is considered to be one of the simple and preferred models, and so that it can represent a document as a set of distinct words that are all not compatible with each other and by ignoring the order and meaning of words. Considering the general classes for the feature selection algorithms, they can be broadly classified into the basic classes: filter methods, or it can be wrapper methods and embedded methods. Filter feature selection type methods apply a statistical based measure to assign a scoring to the feature. This study is actually discussing the various limitations that can arise in the popular algorithms that are used for the previous researches as the feature selection algorithm. The system of the application need to be very fast and accurate which is free from any kind of intrusion or vulnerable system. A multilevel ensemble classifier system shall be used for sending and receiving data for the same [41].

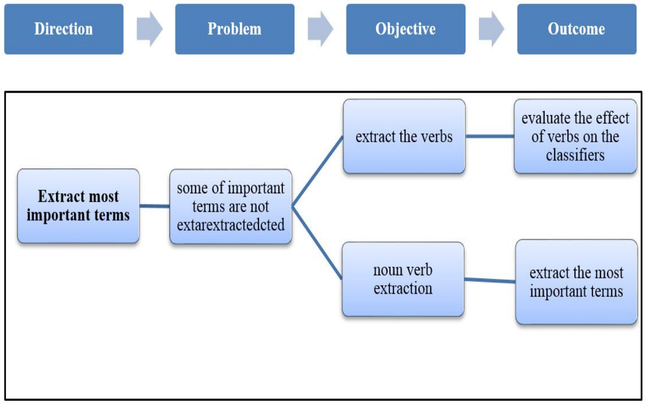
Direction
Problem
Objective
Outcome
Extract most important terms
some of important terms are not extarextractedcted
extract the verbs
evaluate the effect of verbs on the classifiers
noun verb extraction
extract the most important terms

**Figure 2:** Methodology followed for the system

### 4.1 K-Nearest Neighbor (KNN) Classifier Algorithm

The K-nearest neighbor (KNN) is a well-known instance-based classifier. There are two basic steps that are involved in this kind of algorithmic approach to solve the instance-based classifier problems. In step one, the discovery about the training examples in particular is done which, is relative to the closest and the unseen example available. In the second step the algorithm takes, the most frequently taking place about the classification for these various K examples (or they were taking the average of these common K label values based when it comes to regression). In KNN algorithm, a real new input instance may need to be part of the really same class as its K input nearest neighbors are in the training dataset. The weighted sum that can be calculated as in KNN classification is as follows:

$$Scored\ (r,\ t_i) = \sum_{j=KNN(r)} sim(r, r_j)\delta(r_j, c_i)$$

Where KNN(r) shows originally the set of all the K nearest neighbors of that particular instance *r*. If $r_j$ is a part of $c_i$, subsequently $\delta(rj,c_i)$ equals one; or else, it will be zero. For the first test instance *r*, it can be considered that it has to be part of the class and is a member so that has the best resulting weighted sum completely.

### 4.2 Naive Bayes (NB) Classifier Algorithm

The Naive Bayes (NB) type of classifier is a really well known and authentic machine learning technique. It is an uncomplicated and simple probabilistic based classifier determined by the utilization of the Bayes' theorem (from Bayesian Statistics Theory) which is having strong (naive) and cumulative independence assumptions. The more detailed and standardized word for the fundamental theorem of the probability model could have been an independent feature in the model. Simply a proper Naive Bayes classifier could then presume that the actual presence (or absence) can be of a specific feature of a class (that which it is an attribute) is obsolete unrelated to the presence (or could be the absence) of any other salient feature. It can be really formulated as follows:

$$P(C_i|d) = \frac{P(C_i)P(d|C_i)}{P(d)}$$

Where $P(C_i|d)$ is the posterior probability of class $C_i$ given a new instance *d*, $P(C_i)$ is the probability of class $C_i$ which can be calculated by:

$$P(C_i) = \frac{N_i}{N}$$

where we consider that Ni is the number of the proper instance that are associated for being with the class Ci, and N is actually the number of the classes, $P(d|C_i)$ is the probability of an instance *d* given a class $C_i$, and *P(d)* is finally the probability of instance *d*.

### 4.3 Support Vector Machine (SVM) Classifier Algorithm

Support Vector Machine (SVM) is an algorithm that was developed for pattern classification but has recently been adapted for other uses, such as finding regression and distribution estimation and document classification. Since its introduction in 1970s by Vapnik, The complete data is composed of the two groups, and this data is being categorized through the mechanism for dividing space with a proper hyper plane, and in this method. In standard style, an SVM could easily be able to find out what to discover as a linear hyper plane that which can separate the entire negative and positive valued examples with maximal case margin. The data entry points are identified for which they can be as being positive or negative, and the problem also is to be identified as to find a hyper-plane that separates the data points by a maximal margin.

Proposed Extraction:
The methods used for the utilization of Nouns and Verbs as extracted from text documents appropriate for classification of documents is yet to be developed, which is this based on the pervious results appeared that sometimes verb extraction outperformed noun features extraction and BOW extraction, which leads that the verb extraction has powerful to extract important terms. On the other side, noun as extraction outperformed BOW and verb as extraction and that leads this chapter to combine the powerful of extractions noun terms and verb terms as the extraction. Hence, the use of nouns and verbs in combination for feature extraction is proposed for this research. Further, the exact benchmark datasets utilized in past sections are used in assessing the effectiveness of the proposed method. Accordingly, KNN, NB, and SVM are the classifiers used.

**Table 2**: The SVM results on the proposed extraction method on benchmarks

| DS | BOW | | Nouns | | Verbs | | NV | | Improvement |
|---|---|---|---|---|---|---|---|---|---|
| | # Term | Accuracy | # Term | Accuracy | # Term | Accuracy | # Term | Accuracy | % |
| DS1 | 12152 | 69.16 | 5392 | 71.34 | 2931 | **75.33** | 5651 | **72.41** | -2.92 |
| DS2 | 6737 | **68.66** | 4557 | **60.44** | 2580 | 54.3 | 4731 | 59.88 | -8.78 |
| DS3 | 27211 | 49.2 | 9935 | **55.23** | 4398 | 53 | 10346 | **55.6** | **0.37** |
| DS4 | 13310 | 68.64 | 6837 | **73.6** | 2995 | **80** | 7096 | **73.6** | -6.4 |
| DS5 | 9496 | 68 | 5121 | 68.1 | 2772 | **68.8** | 5330 | **70** | **1.2** |
| DS6 | 3863 | 72.45 | 2376 | **74.65** | 1458 | 69.15 | 2375 | **73.45** | -1.2 |
| DS7 | 8306 | 70.25 | 4022 | 79.59 | 2098 | **80.5** | 4636 | **81.67** | **1.17** |
| DS8 | 16383 | 61.5 | 7815 | **69.14** | 3364 | 67.25 | 8350 | **70.35** | **1.21** |
| DS9 | 14938 | 58.35 | 7338 | **69.12** | 3200 | 66.4 | 7621 | **70.65** | **1.53** |

## 5. RESULTS

The benchmark datasets are employed in the study's experiments. As the results demonstrate, the initial finding that regards verbs as terms extraction is essential for classification, and as demonstrated, sometimes it is better than others. Somehow, the result is inconclusive owing to the impact of classification or features selection. Furthermore, nouns and verbs are equally important as term. Hence, a new method for extracting nouns with verbs is proposed in this work, and this method has proven its superiority over other comparable methods. In testing the proposed method, three classifiers techniques were employed. Based on the behavior of KNN, NB and SVM on the effect of stability and the effect of the number of classes on the performance of each one. This part of the study worth to find out all the weaknesses that are associated with the algorithms which is associated with all the classes that are available in accordance with the powerful nature that each one of them has stability on the performance of different independent runs. Considering all the algorithms, it can show very well that SVM is having the best performance, but the weakness of this particular algorithm lies in the fact about the large datasets handling capability like the 20-Newsgroup where NB performed better than KNN. Table below, shows a summary from the results for the three classifiers particularly KNN, NB, and SVM by comparing the accuracies achieved using the BOW extraction method. The results' observation leads to indicate that the SVM classifier achieved best results in five dataset as can be seen by DS1, DS2, DS3, DS6, and DS8 with accuracies of 72.41%, 59.88%, 55.60%, 73.45%, and 70.35% respectively.

The results of classification can be influenced from certain different and unimportant terms that are present in the target dataset document. Therefore, it is necessary that a pre-processing is included in the documents in order that terms that are not informative can be eliminated. The steps of pre-processing include the removal of stop-words. Stop words comprise words, which do not deliver any meaning. These include numbers, determinants and pronouns. Hence, for each document, the feature vector will carry the complete set of noun and verb that are obtained in the pre-processing from the Word Stemming.

**Table 3:** The proposed extraction method results for the three classifiers

| DS | | | KNN | NB | SVM | Best | |
|---|---|---|---|---|---|---|---|
| | # of feature | # of classes | NV extraction | | | Best<br>Classifier | Accuracy |
| | | | KNN | NB | SVM | | |
| DS1 | **5651** | 8 | 65.2 | 61.2 | 72.41 | SVM | 72.41 |
| DS2 | **4731** | 53 | 53.9 | 55.5 | 59.88 | SVM | 59.88 |
| DS3 | **10346** | 10 | 49.7 | 54.7 | 55.6 | SVM | 55.6 |
| DS4 | **7096** | 3 | 66.2 | 91.9 | 73.6 | NB | 91.94 |
| DS5 | **5330** | 2 | 63 | 93 | 70 | NB | 93 |
| DS6 | **2375** | 6 | 65.2 | 65.1 | 73.45 | SVM | 73.45 |
| DS7 | **4636** | 2 | 86.8 | 79.8 | 81.67 | KNN | 86.83 |
| DS8 | **3364** | 5 | 66.9 | 62.1 | 70.35 | SVM | 70.35 |
| DS9 | **7621** | 4 | 66.4 | 81.5 | 70.65 | NB | 81.53 |

Also, in order find out the measurement of terms along with the concepts in terms of weights, the *tfidf* (term frequency-inverted document frequency) must be implemented in the system. Lastly, feature vectors are created from the documents that are utilizing the terms concepts of weights. The vectors are then used for classification. Four of the nine datasets DS1, DS4, DS5, and DS7 are superior in classification. Additionally, the outcomes demonstrate the significance of the application of verbs as extraction method in decreasing the amount of terms of all employed datasets, and in the performance that is close to other extraction methods. It should be noted that there is a problem in terms of the performance. In particular, some of the datasets obtained sound performance with BOW, while others demonstrate good performance using nouns as terms, and others utilizing verbs as terms.

## 6. CONCLUSION

The important terms to be extracted from documents for improving document classification, are highlighted. Accordingly, the benchmark datasets are employed in the study's experiments. As the results demonstrate, the initial finding that regards verbs as terms extraction is essential for classification, and as demonstrated, sometimes it is better than others. Somehow, the result is inconclusive owing to the impact of classification or feature's selection. Furthermore, nouns and verbs are equally important as term. Hence, a new method for extracting nouns with verbs is proposed in this work, and this method has proven its superiority over other comparable methods. In testing the proposed method, three-classifier techniques were employed.

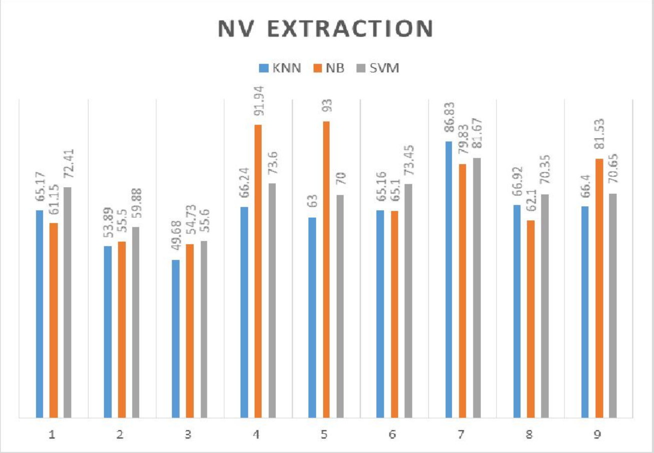

Figure 3. Noun Verb Extraction results

SVM classifier shows better performance with a high number of classes. For instance, DS1, DS2, and DS3 respectively comprising of eight, fifty-three, and ten classes, attained SVM accuracies of 72.41%, 59.88%, and 55.60% correspondingly. For the same datasets, the accuracies with NB and KNN classifiers appear to be higher. The outcomes thus show that based on accuracy, the proposed document classification approach increased the performance compared with baseline approaches. This increases the effectiveness of the document classification domain for modern applications such as sentiment analysis, crime detection, and weather detection.

## REFERENCES

1. Ghareb, S., Bakar, A. & Hamdan, R. , Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications. 2016. 49(1): p. 31-47. https://doi.org/10.1016/j.eswa.2015.12.004
2. Wang, Y., Liu, Y., Feng, L. & Zhu, X. , Novel feature selection method based on harmony search for email classification. Knowledge-Based Systems,. 2015. 73: p. 311-323. https://doi.org/10.1016/j.knosys.2014.10.013
3. Zhiwei, W., Li, M., & Ying, W, A Probabilistic Model for Retrospective News Event Detection. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005: p. 106-113.
4. Mohd, B., Ali, N., & Saad, O. 2012, Optimal Initial Centroid in K-Means For Crime Topic. Journal of Theoretical and Applied Information Technology,. 2012. 45: p. 019-026.
5. Ding, X.T., Improved Mutual Information Method For Text Feature Selection. The 8th International Conference on Computer Science & Education. IEEE,. 2015: p. 163-166.
6. Wang, Y., Liu, Y., Feng, L. & Zhu, X., Novel feature selection method based on harmony search for email classification. Knowledge-Based Systems. 2015. 73: p. 311-323. https://doi.org/10.1016/j.knosys.2014.10.013
7. Sharaff, A., Nagwani, K., & Swami, K, Impact of Feature Selection Technique on Email Classification. 2015.
8. Trivedi, K., & Dey, A Comparative Study of Various Supervised Feature Selection Methods for Spam Classification. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, ACM, . ACM, 2016: p. 61-64.
9. Fodeh, S., Punch, W. & Tan, P., On ontology-driven document clustering using core semantic features. On ontology-driven document clustering using core semantic features,. Journal of KnowlInfSyst, Springer-Verlag London, 2011. 28: p. 395-421.
10. Bhardwaj, A., Text Mining, its Utilities, Challenges and Clustering Techniques. International Journal of Computer Applications. , 2016. 135: p. 22-24.
11. Hotho, A., Staab, S., & Stumme, G. , WordNet improves text document clustering. In Proc. of the SIGIR 2003 Semantic Web Workshop, 2003: p. 541-545.
12. Fodeh, S., Punch, W. & Tan, P., On ontology-driven document clustering using core semantic features. On ontology-driven document clustering using core semantic features. Journal of KnowlInfSyst, Springer-Verlag London, 2011. 28: p. 395-421.
13. Zahran, M., & Kanaan, G., Text feature selection using particle swarm optimization algorithm 1. 2009.
14. Aghdam, H., Ghasem-Aghaee, N., & Basiri, M. E. , Text feature selection using ant colony optimization. Expert systems with applications, . 2009. 36: p. 6843-6853. https://doi.org/10.1016/j.eswa.2008.08.022
15. Uğuz, H., A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems,. 2011. 24: p. 1024-1032. https://doi.org/10.1016/j.knosys.2011.04.014
16. Uysal, K., & Gunal, , Text classification using genetic algorithm oriented latent semantic features. Expert Systems with Applications,. 2014. 41: p. 5943-5947. https://doi.org/10.1016/j.eswa.2014.03.041
17. Aghdam, H., & Heidari, S., Feature selection using particle swarm optimization in text categorization. Journal of Artificial Intelligence and Soft Computing Research,, 2015. 5: p. 231-238. https://doi.org/10.1515/jaiscr-2015-0031
18. Wang, Y., Liu, Y., Feng, L. & Zhu, X., Novel feature selection method based on harmony search for email classification. . Knowledge-Based Systems, 2015. 73: p. 311-323. https://doi.org/10.1016/j.knosys.2014.10.013

19. Guru, S., Suhil, M., Raju, N., & Kumar, V. , An Alternative Framework for Univariate Filter based Feature Selection for Text Categorization. Pattern Recognition Letters. 2018. https://doi.org/10.1016/j.patrec.2017.12.025
20. Hong, S., Lee, W., & Han, M., The feature selection method based on genetic algorithm for efficient of text clustering and text classification. International Journal Advance Soft Computing, 2015. 7: p. 2074-8523.
21. Lu, Y., Liang, M., Ye, Z., & Cao, L., Improved particle swarm optimization algorithm and its application in text feature selection. Applied Soft Computing,, 2015: p. 629-636. https://doi.org/10.1016/j.asoc.2015.07.005
22. Masand, B., Linoff, G., & Waltz, D., Classifying News Stories Using Memory Based Reasoning,. 15th Ann Int ACM SIGIR 1992. Conference on Research and Development in Information Retrieval: p. 59-64.
23. Yang, Y., & Pedersen, O., A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, . Morgan Kaufmann Publishers Inc, 1997. 15: p. 412-420.
24. Yang, Y., An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, 1999. 1: p. 69-90. https://doi.org/10.1023/A:1009982220290
25. Joachims, T., Making Large-Scale SVM Learning Practical, Report LS-8, University Dortmund. 1998.
26. Wiener, E., Pedersen, J., & Weigend, , A Neural Network Approach to Topic Spotting,. Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval., 1995.
27. Yang, Y., & Liu, X., A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,, 1999: p. 42-49. https://doi.org/10.1145/312624.312647
28. Koller, D., & Sahami, M. , Hierarchically Classifying Documents Using Very Few Words. Proceedings of 14th International Conference on Machine Learning,, 1997: p. 170-178.
29. McCallum, A., & Nigam, K, A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization, 1998.
30. Dumais, T., & Chen, H., Hierarchical Classification of Web Content,. Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, 2000: p. 256-263. https://doi.org/10.1145/345508.345593
31. Tar, H., & Nyunt, T., Enhancing Traditional Text Documents Clustering based on Ontology. International Journal of Computer Applications,, 2011. 33: p. 38-42.
32. Zheng, H., Kang, B., & Kim, H, Exploiting noun phrases and semantic relationships for text document clustering. Information Sciences 2009. 179: p. 2249-2262. https://doi.org/10.1016/j.ins.2009.02.019
33. Kabi, D., Semantic Document Clustering For Crime Investigation, . Thesis Of Master Of Applied Science In Information Systems Security, Concordia University, Montréal, Québec, Canada., 2011.
34. Howard, M., Semantic preserving text representation and its applications in text clustering, master of computer science, Missouri University of science and technology, USA. 2012.
35. Sharma, R., Swami, N., & Yang,, A Verb-centric Approach for Relationship Extraction in Biomedical Text,. IEEE Fourth International Conference on Semantic Computing (ICSC),, 2010: p. 377-385.
36. Iqbal, Q., Jin-Woo, J., Jee-Uk, H., & Dong-Ho, L, Concept map construction from text documents using affinity propagation. Journal of Information Science, 2013. 39: p. 719-736. https://doi.org/10.1177/0165551513494645
37. Wei, T., Lu, Y., Chang, H., Xhou, Q., & Bao, X., A semantic approach for text clustering using WordNet and lexical chains. . Journal of Expert Systems with Applications 2015. 42: p. 2264-2275. https://doi.org/10.1016/j.eswa.2014.10.023
38. Bsoul, Q., & Salim, Z. , Effect Verb Extraction on Crime Traditional Cluster. world applied science journal., 2016.
39. Turabieh, H., Population-based algorithms for university timetabling problems. PhD thesis. Faculty of information science and technology. University Kebangsaan Malaysia., 2010.
40. Steinbach, M., & Karypis, G, A comparison of document clustering techniques. KDD Workshop on Text Mining. 2000: p. 24-35.
41. Apoorva Deshpande, Ramnaresh Sharma, Multilevel Ensemble Classifier using Normalized Feature based Intrusion Detection System, International Journal of Advance Trends in Computer Science and Engineering, Vol 7, No.5, September -October 2018.
42. Oumayma Oueslati, Ahmed Ibrahim S. Khalil, Habib Ounelli, Sentiment Analysis for Helpful Reviews Prediction, International Journal of Advance Trends in Computer Science and Engineering, Vol 7, No.3, May - June 2018.