ANALYTICAL ASSESSMENT OF NOUN VERB TERM EXTRACTION FOR DOCUMENT CLASSIFICATION USING T-TEST

Omaia Mohammad Al-Omari¹, Nazlia Omar²

¹Center for Artificial Intelligence Technology
Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia,

²Center for Artificial Intelligence Technology
Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia,

¹Omaiaomari@yahoo.com, ²Nazlia@ukm.edu.my

Corresponding Author: Omaia Mohammad Al-Omari

Abstract

There has been a significant growth in the digital word as per the documents are concerned. The classification of digital document is a big trend in the market as a revolution. However the classification of the document is a big task for the modern applications. There are various terms that are used for the extraction of information from the documents. The main concerned areas for the document classification are the noun and the verbs that broadly signify the topics and events. The use of NV (Noun Verb) techniques is a common and powerful practice for the words to be classified. The performance of the document depends on the NV technique due to the classification of the document. The main aim of the work shown in this study is to enhance the capability of the NV extraction methodology to classify the documents. Three classifiers namely, K-Nearest Neighbor (KNN), Naive Bayes (NB), and Support Vector Machine (SVM) are used for the comparison of the results. Various benchmark set are used in this study for the evaluation of the accuracy of the data sets. The data sets were taken from Reuters 8 and WebKb for this purpose. Other extraction methods were also enhanced and incorporated with the NV method extraction e.g., Nouns, Bag of Word (BOW), and Verbs. The results are studied and the conclusion follows them.

Keywords: BOW extraction, Document classification, NV extraction, KNN classifier, NB classifier, SVM classifier.
I. Introduction

The method of natural language processing (NLP) extraction has been a popular method for the representation of text of web pages. These algorithms are useful in retrieving texts and dividing them into parts, checking the spelling as well as counting how many words there are. Somehow, as indicated by [III], the capacities of these algorithms are highly restricted, particularly in sentence interpretation and in the extraction of meaningful information. [V] Mentioned the requirement of high-level symbolic capacities in the use of NLP. This includes the attainment and access of lexical, episodic, and semantic memories, generation and proliferation that include dynamic binding; handling of reoccurring and component based structures, management of enormous processing / learning modules and sending of the data among these modules, and representation of abstract conceptions. The classification algorithms generally depend on BOW (Bag of Words) as syntactic extraction. However, due to the fast progression of text documents, the textual data comprise the diverse vocabulary, aside from being greatly dimensional, because of large feature extracted from documents. Hundreds or thousands of terms are present. Therefore, improving the extraction yields to increase the performance classifying the document to the correct class.

As demonstrated by experiments carried out in several studies, the use of the concepts yields superior outcomes [VIII]. However, as mentioned in the works of [VIII, VI] the improvement is not always significant. Rather, it has been documented by other researchers that the ontological concepts add no value, and they may harm document classification in terms of performance [VIII, VI, and II]. Nonetheless, in this work, the major limitation with regard to the assessment on the past work, particularly on the comparison between the extraction methods of document classification, will be highlighted. Accordingly, two benchmark datasets are used in the assessment of the extraction methods [X]. 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. In addition, this work will attempt to provide an answer to the question, namely: Finding out the most important terms that can be taken out from documents in order to enhance the performance of the document classifiers. In this regard, the current work attempts to achieve the starting objective.

II. Related Works and Background

A large number of English Lexical item from the WordNet database are available online. The four major parts of the speech, i.e. noun, verb, adjective and adverb are the major contributors because of which the WordNet was created. Synset is responsible for creation of a specific meaning of a word. Sense is the availability of the specific meaning for a word of a fixed kind called as POS. Synset contains all the items that are having synonyms meaning. Thus Synset will have a glossary that contains the concept it shows[X]. For example, the words night, night time as well as dark represents the same Synset. These Synset have some common relations within the set. Some of these sets follow the relation of being kind – of or a part – of type relationship. This research uses WordNet to extract the verbs and nouns from the
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, this work will combine the powerful of extractions noun terms and verb terms as the extraction. The following figure 1 illustrate the extraction process of verbs and nouns from the datasets.

**Fig.1:** Extraction process of verbs and nouns from the datasets documents

**III. Research Methodology and Data Sets**

In the experiments executed in the research, two standard datasets with differing properties are employed. The datasets are used in order that the performance of the algorithms can be thoughtfully analyzed and evaluated. Accordingly, the first dataset contains the group of document classification of Reuters-21578 [XI]. It is a set of financial dispatches emitted during the year 1987 by the Reuters agency in the English language and available free on the Internet. The documents were assembled and indexed with categories by personnel from Reuters Ltd. Also. It contains a number of limitations. Firstly, more than half of the documents have no class labels while some are allotted to numerous classes. This dataset. Furthermore, certain classes, for instance, earn and acquisition has a large amount of documents whereas other classes for instance reserve and veg-oil contain only a handful of documents; this is as well a limitation to this dataset. These limitations were addressed by selecting a dataset containing the eight largest classes which called by literature R8. This dataset is a collection of web pages from four different college websites, contains 8282 web page assigned to 7 classes. In this work, only four classes are used as in the literature, which contains 2803 documents. Experiments were performed on two standard corpora of English text datasets namely Reuters R8 and WebKB dataset.

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. Additionally, numerous morphological procedures can be added to the dialects containing its exact sense. In order to find out the similar nature, these forms are standardized in a general root-form, which is called as word stemming. In this study, word stemming is executed based on Porter's stemmer [XVI]. Ultimately, in this study, the nouns and verbs within the benchmark’s documents are recognized and employed as features [XV]. A word could be in a form of verb or noun. In this study, WordNet is applied in order to ascertain whether the word under the consideration belongs to the dataset of nouns or the dataset of verbs. 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. 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. The effectiveness of the WordNet based (noun, verb) method of identification was evaluated in this study. For this purpose, the samples produced by the benchmarks datasets were used. In this regard, [XI] gathered the Reuters-21578 and 20-newsgroups. In this work, KNN, NB, and SVM are employed as the classification algorithms, while cross-validation is used in dividing the data into different sets of training the data and then finally testing on them.

Cross validation is a method that uses statistical modelling in the evaluation and comparison that uses the division rules on the learning algorithm. This rule divides the data into two segments, where one segment contains data for finding out the result as derivations and the other one uses that training on the dataset that is obtained from the model. The other contains data for model validation. The standard cross validation requires the sets of training and validation to crossover in successive rounds. This is in order that each data point can have an opportunity that has to be validated. Accordingly, a sub category of cross-validate is k-fold data cross validation [XII]. Using this cross-validation, at the beginning, the data is apportioned into k segments, which are equal or supposedly nearly equal in size or folds. Consequently, k iterations of exercising the results and training them along with the validation are executed in a manner that a different data folds is held-out inside each iteration for validation, while the (k-1)th iterations that remains are utilized for the basic learning purpose. Let us consider that with k = 3 is shown in Figure 4.8. As shown, in the figure, the shadowed section is employed for the purpose the training whereas the light shaded section is for the purpose of validation. In the context of data extraction, text mining and learning by machines, 10-fold cross-validate (k = 10) appears to be the widely used and spread value for the data set. The process of cross validation is utilized in the assessment or comparison of learning algorithms. Here, one or the other learning algorithm is applied in each iteration that utilizes k-1 group of data from the whole dataset for learning respectively one or the other models [IX]. In the same way, the upcoming learned models are instructed for forecasting the whole data within the validation iteration and get results. Then, with the use of certain predetermined performance metric (e.g., accuracy), tracking can be performed for the dataset in each iteration and corresponding to the learning algorithm [I]. The samples
of k-performance metrics will be accessible for every algorithm, on completion. Variable methodologies such as averaging are employable in attaining a compiled measure from the supplied samples. Alternatively, statistical hypothesis test can make use of these sample sets for the demonstration of the superiority of one algorithm to another [VII]. Cross-validate process has two important goals namely, correlate the performance of the data from the learning model, which exists in the data set using of one algorithm, and to make comparison of the evaluations on the existing algorithms to determine the suitable algorithm for the existing set of information.

Fig.2: Procedure of three-fold cross-validation.

IV. Results

This section shows the results of the experiments conducted for BOW, Noun, Verb, and Noun Verb extraction methods with the three classifiers. Finally, the statistical tests are done for two objectives. The first is to evaluate the classifier-dependency of with different extraction methods, and the second is to statistically analyze the difference between the results of extraction methods.

IV.i. Results of the Proposed NV Feature Extraction Method

The proposed method is evaluated by conducting the document classification using the generated feature sets from the proposed method. The same setting used for BOW, Noun, and Verb is also used for NV (i.e., datasets, classification methods and evaluation metrics). The results achieved by NV compared with those achieved by the other extraction methods are reported in Table 4.1
Table 4.1: Classification results achieved by extraction methods on benchmarks

| Dataset | Classifier | BOW | Noun | Verb | NV |
|---------|------------|-----|------|------|----|
|         |            |     |      |      |    |
| R8      |            |     |      |      |    |
|         | Micro average F1 |     |      |      |    |
|         | NB          | 75.23 | 77.95 | 76.6 | 77 |
|         | SVM         | 80.86 | 69.82 | 73.5 | 81.3 |
|         | KNN         | 66.84 | 82.67 | 83.6 | 84 |
|         | Macro average F1 |     |      |      |    |
|         | NB          | 73.71 | 76.05 | 74.3 | 74.1 |
|         | SVM         | 79.96 | 68.79 | 72.1 | 75.6 |
|         | KNN         | 65.64 | 80.7  | 82   | 82.6 |
|         |             |     |      |      |    |
| WebKB   |            |     |      |      |    |
|         | Micro average F1 |     |      |      |    |
|         | NB          | 76.65 | 80.38 | 73.4 | 83.7 |
|         | SVM         | 73.15 | 78.11 | 73.9 | 81.8 |
|         | KNN         | 79.38 | 82.76 | 72.6 | 83.4 |
|         | Macro average F1 |     |      |      |    |
|         | NB          | 74.5  | 77.87 | 72.2 | 79.9 |
|         | SVM         | 70.35 | 76.22 | 71.5 | 80.4 |
|         | KNN         | 78.2  | 79.67 | 71   | 81.8 |

Table 4.2: Precision (P), Recall (R) and F-measure (F) of each class in R8 dataset using BOW extraction method with three classifiers

|        | NB | KNN | SVM |
|--------|----|-----|-----|
|        | P  | R   | F   | P  | R   | F   | P  | R   | F   |
| Interest       | 72.1 | 63.4 | 67.5 | 67.1 | 60.4 | 63.6 | 80.8 | 69.2 | 74.5 |
| Money-fx       | 76  | 70.3 | 73  | 70.7 | 63.2 | 66.8 | 84.1 | 77.1 | 80.5 |
| Grain          | 85.1 | 79.7 | 82.3 | 79.1 | 54.5 | 64.5 | 89.3 | 85.1 | 87.2 |
| Ship           | 81.7 | 73.5 | 77.4 | 74.7 | 53.7 | 62.5 | 75.9 | 80.2 | 78 |
| Trade          | 75.7 | 68.6 | 71.9 | 71.1 | 74.2 | 72.6 | 82.7 | 76.7 | 79.5 |
| Crude          | 67.3 | 62.5 | 64.8 | 62.1 | 68.5 | 65.2 | 75.3 | 68.2 | 71.6 |
| Earn           | 66.1 | 72.6 | 69.2 | 65.3 | 62.4 | 63.8 | 74.1 | 79.3 | 76.6 |
| Acquisition    | 81.3 | 86.3 | 83.7 | 75.6 | 59  | 66.2 | 90.3 | 93.6 | 91.9 |
| Average        | 75.6 | 72.1 | 73.7 | 70.7 | 62  | 65.6 | 81.5 | 78.7 | 80  |
Table 4.3: Precision (P), Recall (R) and F-measure (F) of each class in WebKB dataset using BOW extraction method with three classifiers

| Class     | NB   | KNN  | SVM  |
|-----------|------|------|------|
|           | P    | R    | F    | P    | R    | F    | P    | R    | F    |
| Course    | 78.3 | 77   | 77.6 | 67.9 | 74.2 | 70.9 | 81.3 | 85.6 | 83.4 |
| Project   | 78.4 | 70   | 73.9 | 75.3 | 71.1 | 73.2 | 70.4 | 74.3 | 72.3 |
| Student   | 75.4 | 68.1 | 71.5 | 61.5 | 66.1 | 63.7 | 78.8 | 73.8 | 76.3 |
| Faculty   | 72.7 | 77.3 | 74.9 | 73.3 | 74.1 | 73.7 | 78.3 | 83.7 | 80.9 |
| Average   | 76.2 | 73.1 | 74.5 | 69.5 | 71.4 | 70.4 | 77.2 | 79.4 | 78.2 |

Table 4.4: Precision (P), Recall (R) and F-measure (F) of each class in R8 dataset using Noun extraction method with three classifiers

| Class   | NB   | KNN  | SVM  |
|---------|------|------|------|
|         | P    | R    | F    | P    | R    | F    | P    | R    | F    |
| Interest| 74.6 | 66.4 | 70.3 | 70.3 | 63.3 | 66.6 | 84.6 | 69   | 76   |
| Money-fx| 78.7 | 70.3 | 74.2 | 74.1 | 66.3 | 70   | 78.4 | 88.2 | 83   |
| Grain   | 88.1 | 76.1 | 81.6 | 82.9 | 57.1 | 67.6 | 93.2 | 78.4 | 85.2 |
| Ship    | 84.5 | 85.7 | 85.1 | 78.2 | 56.3 | 65.5 | 90.2 | 75.9 | 82.4 |
| Trade   | 78.3 | 80.6 | 79.5 | 74.5 | 77.7 | 76.1 | 86.6 | 81   | 83.7 |
| Crude   | 69.6 | 75   | 72.2 | 65.1 | 71.8 | 68.3 | 78.9 | 73.2 | 75.9 |
| Earn    | 68.4 | 67.5 | 67.9 | 68.4 | 65.4 | 66.9 | 77.6 | 71.8 | 74.6 |
| Acquisiti on | 84.1 | 72 | 77.6 | 79.2 | 61.8 | 69.4 | 91.5 | 78.9 | 84.8 |
| Average | 78.3 | 74.2 | 76.1 | 74.1 | 65   | 68.8 | 85.1 | 77   | 80.7 |

Table 4.5: Precision (P), Recall (R) and F-measure (F) of each class in WebKB dataset using Noun extraction method with three classifiers

| Class     | NB   | KNN  | SVM  |
|-----------|------|------|------|
|           | P    | R    | F    | P    | R    | F    | P    | R    | F    |
| Course    | 84.8 | 75.5 | 79.9 | 80.2 | 78   | 79.1 | 82.9 | 79.9 | 81.4 |
| Project   | 77.4 | 81.6 | 79.5 | 76.7 | 79.5 | 78.1 | 81.4 | 83.6 | 82.4 |
| Student   | 84.8 | 78.7 | 81.6 | 68.2 | 71.6 | 69.9 | 77.7 | 81.6 | 79.6 |
| Faculty   | 71.7 | 69.4 | 70.5 | 79.3 | 76.6 | 77.9 | 81.7 | 69.9 | 75.3 |
| Average   | 79.7 | 76.3 | 77.9 | 76.1 | 76.4 | 76.2 | 80.9 | 78.7 | 79.7 |
### Table 4.6: Precision (P), Recall (R) and F-measure (F) of each class in R8 dataset using Verb extraction method with three classifiers

| Class    | NB | KNN | SVM |
|----------|----|-----|-----|
|          | P  | R   | F   | P  | R   | F   | P  | R   | F   |
| Interest | 71 | 62.4 | 66.5 | 73.7 | 66.4 | 69.8 | 88.7 | 73.3 | 80.3 |
| Money-fx | 74.9 | 69.2 | 71.9 | 77.6 | 69.5 | 73.3 | 82.1 | 76.5 | 79.2 |
| Grain    | 83.8 | 78.5 | 81 | 86.9 | 59.8 | 70.9 | 90.7 | 80.9 | 85.5 |
| Ship     | 80.4 | 72.4 | 76.2 | 82 | 59 | 68.6 | 86 | 76.3 | 80.8 |
| Trade    | 74.6 | 67.6 | 70.9 | 78.1 | 81.5 | 79.7 | 84.9 | 90.8 | 87.7 |
| Crude    | 76.3 | 77.9 | 77 | 68.2 | 75.3 | 71.6 | 76.7 | 82.7 | 79.6 |
| Earn     | 65.2 | 71.5 | 68.2 | 71.7 | 68.5 | 70.1 | 75.3 | 81.3 | 78.2 |
| Acquisition | 84.9 | 80.5 | 82.7 | 83 | 64.8 | 72.8 | 89.4 | 80.2 | 84.5 |
| Average  | 76.4 | 72.5 | 74.3 | 77.6 | 68.1 | 72.1 | 84.2 | 80.2 | 82 |

### Table 4.7: Precision (P), Recall (R) and F-measure (F) of each class in WebKB dataset using Verb extraction method with three classifiers

| Class    | NB | KNN | SVM |
|----------|----|-----|-----|
|          | P  | R   | F   | P  | R   | F   | P  | R   | F   |
| Course   | 70.7 | 79.5 | 74.8 | 68.5 | 69.8 | 69.2 | 69.8 | 68.3 | 69.1 |
| Project  | 67.2 | 72.2 | 69.6 | 69.9 | 73.3 | 71.6 | 72.3 | 80.1 | 76 |
| Student  | 70.7 | 77.4 | 73.9 | 79.7 | 70.9 | 75 | 76.6 | 69.2 | 72.7 |
| Faculty  | 64.4 | 77.6 | 70.4 | 71.6 | 68.6 | 70.1 | 69.2 | 63.5 | 66.3 |
| Average  | 68.3 | 76.7 | 72.2 | 72.4 | 70.7 | 71.5 | 72 | 70.3 | 71 |

### Table 4.8: Precision (P), Recall (R) and F-measure (F) of each class in R8 dataset using NV extraction method with three classifiers

| Class    | NB | KNN | SVM |
|----------|----|-----|-----|
|          | P  | R   | F   | P  | R   | F   | P  | R   | F   |
| Interest | 67.6 | 65.4 | 66.5 | 77.2 | 69.6 | 73.2 | 93 | 73.3 | 82 |
| Money-fx | 71.3 | 69.2 | 70.2 | 81.3 | 72.8 | 76.8 | 86.1 | 87.9 | 87 |
| Grain    | 79.8 | 74.9 | 77.3 | 91.1 | 62.7 | 74.3 | 90.7 | 81.4 | 85.8 |
| Ship     | 76.5 | 84.4 | 80.3 | 85.9 | 61.8 | 71.9 | 86 | 78.9 | 82.3 |
| Trade    | 71 | 79.4 | 74.9 | 81.8 | 85.4 | 83.6 | 85.7 | 76.5 | 80.8 |
| Crude    | 76.3 | 77.9 | 77 | 71.5 | 78.9 | 75 | 78.1 | 80.9 | 79.5 |
| Earn     | 62 | 66.5 | 64.2 | 75.1 | 71.8 | 73.4 | 76.1 | 79.3 | 77.7 |
| Acquisition | 84.9 | 80.5 | 82.7 | 87 | 67.9 | 76.2 | 89.4 | 82.3 | 85.7 |
| Average  | 73.7 | 74.8 | 74.1 | 81.4 | 71.4 | 75.6 | 85.6 | 80.1 | 82.6 |
Table 4.9: Precision (P), Recall (R) and F-measure (F) of each class in WebKB dataset using NV extraction method with three classifiers

|        | NB       |        | KNN     |        | SVM     |        |
|--------|----------|--------|---------|--------|---------|--------|
|        | P        | R      | F       | P      | R      | F      |
| Course | 86.5     | 80.2   | 83.2    | 88.7   | 82.9   | 85.7   |
| Project| 79.5     | 81.7   | 80.6    | 78.1   | 81.4   | 79.7   |
| Student| 81.2     | 76     | 78.5    | 70.7   | 76.8   | 73.6   |
| Faculty| 75.4     | 79.3   | 77.3    | 83.1   | 81.7   | 82.4   |
| Average| 80.6     | 79.3   | 79.9    | 80.1   | 80.7   | 80.4   |

Table 4.10: Average F-measure (F) Results of each class in R8 dataset using three classifiers

|        | NB |        | KNN |        | SVM |        |
|--------|----|--------|-----|--------|-----|--------|
|        | BOW| N      | V   | NV     | BOW| N      | V   | NV     | BOW| N      | V   | NV     |
| Interest|67.45|70.3|66.5|66.5|63.5|7|66.6|69.8|73.2|74.54|76|80.3|82|
| Money-fx|73.01|74.2|71.9|70.2|66.7|5|70|73.3|76.8|80.45|83|79.2|87|
| Grain |82.3|81.6|81|77.3|64.5|2|67.6|70.9|74.3|87.15|85.2|85.5|85.8|
| Ship |77.38|85.1|76.2|80.3|62.4|7|65.5|68.6|71.9|77.96|82.4|80.8|82.3|
| Trade |71.94|79.5|70.9|74.9|72.6|76.1|79.7|83.6|79.54|83.7|87.7|80.8|
| Crude |64.76|72.2|77|77|65.1|5|68.3|71.6|75|71.57|75.9|79.6|79.5|
| Earn |69.19|67.9|68.2|64.2|63.7|9|66.9|70.1|73.4|76.57|74.6|78.2|77.7|
| Acquisition |83.68|77.6|82.7|82.7|66.2|4|69.4|72.8|76.2|91.91|84.8|84.5|85.7|
Table 4.11: F-measure (F) Results of each class in WebKB dataset using three classifiers

| Class   | NB     | KNN    | SVM    |
|---------|--------|--------|--------|
|         | BOW N V NV | BOW N V NV | BOW N V NV |
| Course  | 77.64  79.9  74.8  83.2  70.9  79.1  69.2  81.4  83.35  81.4  69.1  80.5 |
| Project | 73.92  79.5  69.6  80.6  73.15  78.1  71.6  82.4  72.3  82.4  76  85.1 |
| Student | 71.53  81.6  73.9  78.5  63.69  69.9  75  79.6  76.25  79.6  72.7  81.5 |
| Faculty | 74.9  70.5  70.4  77.3  73.66  77.9  70.1  75.3  80.9  75.3  66.3  80.2 |

IV.ii. Statistical Tests

For the statistical test, a two-sample T-test on data sets from two independent populations with unequal variances was executed to test if the difference between two groups of results is significant or not. The T-test between the results obtained by NB classifier and other classifiers is conducted to test the classifier dependency of extraction methods BOW, Noun, Verb, and the proposed NV. The statistical analysis is also performed between the classification results of Verb and those of NV. However, before conducting the T-test, the normality test is needed to ensure that the data samples are normally distributed. The results of the normality test are reported in Table 4.12.

Table 4.12: P-values of the groups of results, where each group contains 40 samples and α=0.05

| Classifier | Method | P-value R8 | P-value WebKB |
|------------|--------|------------|--------------|
| NB         | BOW    | 0.92       | 0.48         |
|            | Noun   | 0.33       | 0.68         |
|            | Verb   | 0.45       | 0.55         |
|            | NV     | 0.18       | 0.73         |
| KNN        | BOW    | 0.92       | 0.68         |
|            | Noun   | 0.64       | 0.74         |
|            | Verb   | 0.16       | 0.25         |
|            | NV     | 0.62       | 0.92         |
| SVM        | BOW    | 0.68       | 0.84         |
|            | Noun   | 0.71       | 0.36         |
|            | Verb   | 0.63       | 0.45         |
|            | NV     | 0.35       | 0.67         |

Table 4.12 show the result of normality test for each results group, where each group contains forty samples. The table show that all results groups are normally distributed.
distributed, as p-value is greater than α (i.e., 0.05) in all cases. Therefore, T-test is conducted as described in the following subsections.

Analyzing classifier-dependency

Tables 4.13 and 4.14 show the statistical analysis for the Micro average F1 results of NB vs. KNN and NB vs. SVM for R8 and WebKB datasets, respectively. After the T-test is conducted, the values of t State and t Critical two-tail are observed. If t Stat < -t Critical two-tail or t Stat > t Critical two-tail, the null hypothesis (i.e., the two groups have equal means) is rejected. This means that the difference between the two groups is significant. This is the case in the conducted tests for Verbs. As shown in Table 4.13 of Reuters dataset, the t State is greater than t Critical two-tail (13.47 > 2.00) and (9.63 > 2.00) for NB vs. KNN and NB vs. SVM, respectively. The same case in Table 4.14 of WebKB dataset, where t State is greater than t Critical two-tail (17.06 > 2.00) and (18.98 > 2.00) for NB vs. KNN and NB vs. SVM, respectively. The results of the statistical test show that the feature extraction in Verb method is classifier dependent.

Table 4.13 T-test for Verb results with different classifiers for R8 dataset

|          | NB     | KNN     |
|----------|--------|---------|
| Mean     | 78.3   | 81.36   |
| t State  | 13.47  |         |
| t Critical two-tail | 2       |         |

Table 4.14: T-test for Verb results with different classifiers for WebKB dataset

|          | NB     | KNN     |
|----------|--------|---------|
| Mean     | 77.2   | 79.25   |
| t State  | 17.06  |         |
| t Critical two-tail | 2       |         |

Table 4.15: T-test for Verb results with different classifiers for R8 dataset

|          | NB     | KNN     |
|----------|--------|---------|
| Mean     | 77.2   | 79.41   |
| t State  | 18.98  |         |
| t Critical two-tail | 2       |         |

The same test is done for the results of NV. Tables 4.15 and 4.16 show the results of T-test for NB vs. KNN and NB vs. SVM for R8 and WebKB datasets, respectively.
Unlike the results of Verb method, the difference between the various groups of results of NV is not significant. As shown in Table 4.15, t State is greater than -t Critical two-tail (-1.09 > -2.00 and -1.31 > -2.00) for Reuters results of NV. In Table 4.15, the first case (i.e., NB vs. KNN) shows that t State is less than t Critical two-tail (-1.09 < 2.00), while in the second case (i.e., NB vs. SVM), t State is greater than -t Critical two-tail (-1.62 > -2.00). All these cases mean that the null hypothesis (i.e., the two groups have equal means) is accepted (i.e., the null hypothesis is accepted if t State is inside the interval [-t Critical two-tail, t Critical two-tail] which is in our case [-2, 2]). These results demonstrate that the feature set generated by the proposed NV method is independent from the classification method. In other words, the difference between the two groups of results (NB vs. KNN and NB vs. SVM) is not significant.

Table 4.15: T-test for NV results with different classifiers for R8 dataset

|        | NB  | KNN |
|--------|-----|-----|
| Mean   | 81.35 | 81.64 |
| t State| -1.09 |     |
| t Critical two-tail | 2 |     |

|        | NB  | SVM |
|--------|-----|-----|
| Mean   | 81.35 | 82.38 |
| t State| -1.31 |     |
| t Critical two-tail | 2 |     |

Table 4.16: T-test for NV results with different classifiers for WebKB dataset

|        | NB  | KNN |
|--------|-----|-----|
| Mean   | 81.15 | 80.78 |
| t State| -1.36 |     |
| t Critical two-tail | 2 |     |

|        | NB  | SVM |
|--------|-----|-----|
| Mean   | 81.15 | 79.48 |
| t State| 1.78  |     |
| t Critical two-tail | 2 |     |

The statistical analysis of classification results obtained by Verb and NV for R8 and WebKB datasets is conducted and the results are reported in Tables 4.17 and 4.18, respectively. The tables reported the statistical test results between the two groups of classification results in terms of Micro average F1 measures with the utilized classifiers.
### Table 4.17: T-test of R8 dataset Micro average F1 results of NV vs. Verb

|       | NV          | V          |
|-------|-------------|------------|
| **NB** |             |            |
| Mean  | 80.05       | 87.51      |
| t State | -24.83     |            |
| t Critical two-tail | 2          |            |
| **KNN** |             |            |
| Mean  | 81.25       | 88.39      |
| t State | -33.78     |            |
| t Critical two-tail | 2          |            |
| **SVM** |             |            |
| Mean  | 82.75       | 88.39      |
| t State | -11.68     |            |
| t Critical two-tail | 2          |            |

### Table 4.18: T-test of WebKB dataset Micro average F1 results of NV vs. V

|       | NV          | V          |
|-------|-------------|------------|
| **NB** |             |            |
| Mean  | 77.13       | 74.85      |
| t State | -15.54     |            |
| t Critical two-tail | 2          |            |
| **KNN** |             |            |
| Mean  | 77.68       | 75.86      |
| t State | -24.28     |            |
| t Critical two-tail | 2          |            |
| **SVM** |             |            |
| Mean  | 78.12       | 76.68      |
| t State | -8.47      |            |
| t Critical two-tail | 2          |            |
As shown in table 4.17, the t State is less than -t Critical two-tail with all classifiers for R8 dataset. The same case in Table 4.16 for WebKB dataset as the t State is less than -t Critical two-tail with all classifiers. These results indicate that the classification performance in terms of Micro-average-F1 is significantly higher with NV than V. This is attributed to the ability of NV to extract more representative feature set by considering the features’ dependencies.

V. Conclusion

In this paper 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. Somehow, the result is inconclusive owing to the impact of classification or features extraction. 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.

References

I. 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.

II. Bsoul, Q., &Salim, Z. 2016. Effect Verb Extraction on Crime Traditional Cluster, world applied science journal.

III. Cambria, E., & White, b. 2014. Jumping NLP Curves: A Review of Natural Language Processing Research. IEEE Computational Intelligence Magazine, 9(1): 48-57.

IV. Ding, X. & Tang, Y. 2013. Improved Mutual Information Method For Text Feature Selection. The 8th International Conference on Computer Science & Education. IEEE, pp: 163-166.

V. Dyer, M. 1995. Connectionist natural language processing: a status report. in Computational Architectures Integrating Neural and Symbolic Processes, Sun and L. Bookman, Eds. Dordrecht. The Netherlands: Kluwer Academic, 292(1):389–429.
VI. Fodeh, S., Punch, W. & Tan, P. 2011. On ontology-driven document clustering using core semantic features. On ontology-driven document clustering using core semantic features, Journal of KnowlInfSyst, Springer-Verlag London. 28(2): 395-421.

VII. 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

VIII. Hotho, A., Staab, S., &Stumme, G. 2003. WordNet improves text document clustering. In Proc. of the SIGIR 2003 Semantic Web Workshop, pp: 541-544.

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

X. Kummer, O., Savoy, J., &Argand, E. 2012. Feature selection in sentiment analysis.

XI. Lewis, D. 1997. Reuters-21578 text categorization test collection. AT&T Labs Research.

XII. Liu, Xin&Beyrend-Dur, Delphine&Dur, Gael & Ban, Syuhei. (2014).

XIII. OumaymaOueslati, 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

XIV. Porter, F. 1997. An algorithm for suffix stripping in K. Sparck Jones, P. Willett (1st Eds) Readings in Information Retrieval, Morgan Kaufmann Multimedia Information and Systems Series, pp: 313–316.

XV. Rogati, Monica & Yang, Yiming. 2002. High-performing feature selection for text classification. 659. 10.1145/584902.584911.

XVI. Yao, H., Liu, C., Zhang, P., & Wang, L. 2017. A feature selection method based on synonym merging in text classification system. Journal on Wireless Communications and Networking. Springer. pp: 1-8.