# Concise Semantic Analysis based Text Categorization using Modified Hybrid Union Feature Selection Approach

Amol P Bhopale
Dept. of Computer Science and Engineering
Visvesvaraya National Institute of Tech.
Nagpur, India
amolpbhopale@gmail.com

Sowmya Kamath S
Dept. of Information Technology
National Institute of Tech. Karnataka
Surathkal, India
sowmyakamath@nitk.edu.in

Ashish Tiwari
Dept. of Computer Science and Engineering
Visvesvaraya National Institute of Tech.
Nagpur, India
at@cse.vnit.ac.in

***Abstract*—Text categorization mainly comprises of deriving a representation of the corpus in a standard bag-of-words format. The merit of bag-of-word representations is that they considering every term as a feature, while the downside of this is that the computation cost increases with the number of features and the representation of relations between documents and features. Semantic analysis can help in gaining an edge through document and term correlation in a concept space. However, most semantic analysis techniques have their own limitations when used for text categorization. In this work, a Concise Semantic Analysis (CSA) technique that extracts concepts from corpus and then interpret the document & word relationship in a given concept space is proposed. To improve the performance of CSA, a novel feature selection technique called the Modified hybrid union (MHU) was designed, which considerably reduced computation time and cost. To experimentally validate the proposed approach, MHU based CSA was applied to the problem of text categorization. Experiments performed on standard data sets like Reuters-21578 and WSDL-TC, show that the proposed CSA with MHU approach significantly improved performance in terms of execution time and categorization accuracy.**

***Keywords*—*Feature Selection, Text Representation, Concise Semantic Analysis, Document Classification, Supervised Learning***

## I. Introduction

In a digital era, large amount of information is constantly generated, due to which text categorization often plays an important role during information retrieval. Automating document categorization is one of the most important tasks in document management, especially for large corpora. Text categorization is the task of assigning documents to a predefined set of labels, for improving document management. The task is usually a supervised learning technique which is used to automatically identify and annotate the predefined set of class labels [1]. Conventionally, information retrieval by means of text categorization can be achieved by using standard models such as Vector Space Model (VSM) proposed by Salton et al [2] also known as Feature Vector Model (FVM) that helps represent text documents in a 2D space modeled using term frequency. Normally, text representation comprises of two steps - term extraction and term weighting. Term extraction identifies important terms that potentially capture the context of a document. Term weighting is the numeric weight-age of a particular term in a document, indicating its importance in both that particular document, as well as in the entire corpus. The Bag-Of-words (BOW) [2] technique also is popular for generating such representations of documents in term space. The BOW model represents documents in terms of words i.e. the size of a document vector increases with document size. Hence, in case of large corpora, document representation in 2D space is very expensive, due to increase in both processing cost and computation time.

Semantic analysis based techniques can be used to boost up the performance and accuracy during automatic text categorization. They focus on capturing the inherent relationship between words and documents over the predefined labels using fundamental semantic elements called concepts, and represent these words and documents as a combination of concepts. Most existing text categorization methods use statistical weighting methods for calculating term weightage. Deerwester et al [3] proposed a purely statistical technique called Latent Semantic Analysis (LSA), that uses the linear algebraic method of Singular Value Decomposition (SVD) for eliminating the least significant singular values from a document's vector to identify 'latent semantic concepts'. The document by word matrix is then decomposed into word-Vs-concept and concept-Vs-document matrix to represent it in the latent concept space.

Gabrilovich and Markovitch [4] proposed a Explicit Semantic Analysis (ESA) technique to infer the knowledge of word-document relationship using easily readable concepts maintained by human resources. Wikipedia is a good example of this set of concepts where every concept has a huge set of documents. Similar approaches have been proposed in [5] which an objective to develop an concept ontology, where background knowledge extracted from Wikipedia is used as a semantic kernel to improve document representation. Jiang et al [6] used an improved K Nearest Neighbors (KNN) algorithm for text classification. They used constrained one pass-clustering technique to construct a model that uses the least distance principle to define constraints for splitting text documents into hyper spheres with same radius, over which KNN is applied. The model can dynamically update using efficient clustering algorithms. An efficient approach, Fuzzy Similarity and KNN have been proposed by Jiang et al [7] for multi-label text classification. First, fuzzy similarity measure was used to form the cluster from training samples. Then the training documents whose fuzzy similarity score is more than

978-1-5386-3039-6/18/$31.00©2018 IEEE

the threshold value was used in KNN for further classification. Silva and Ribeiro [8] proposed a hybrid hierarchical classification model using Support Vector Machines (SVM) and Relevance Vector Machines (RVM). In the first level, they used RVM to determine those classified samples with low confidence value and in second level, SVM is used to rectify samples for classifying text documents.

In this work, we propose a novel Concise Semantic Analysis (CSA) technique that can handle complex relationships between words and concepts, and is based on two premises. Firstly, the category labels which are derived using human interference can provide additional information that can be used for document classification. Secondly, the categorization process automatically labels of documents using pre-defined labels without looking at the whole content of the document. Hence, it is suitable to have a *word-document* interpretation in the space of concepts, as CSA uses only category labels as a source of information content. A method for extracting concepts from category labels called the Modified Hybrid Union (MHU) is also presented, which is a weighting method used for calculating the degree of relationship between words and concepts. The MHU approach is a feature selection technique that helped in optimizing the space and time taken in representing the document corpora in term space. Experimental validation of proposed approach was done by applying it for the task of text categorization using various supervised learning algorithms, on two standard datasets, Reuters-21578 and WSDL-TC.

The rest of this paper is organized as follows - Section II presents the details of the proposed feature selection technique that is used for enhancing document categorization and the implementation specific details. Section III details the experimental results and performance evaluation of the proposed approach, followed by conclusion and future work in section IV.

## II. Proposed System

Fig. 1 depicts the major processes in the proposed methodology. These processes are discussed in detail in this section.

### A. Document Collection

The Reuters-21578 corpus[9], available from the standard UCI repository, was used for experimental evaluation. It has multi-labeled 3049 documents from top 8 categories such as - acq (1000 documents), crude(253 documents), earn (1000), grain (41), interest (190), money-fx (206), ship (108), trade (251). Another standard dataset which we used is WSDL-TC[10], with a collection of 1090 web service documents from various domains such as - Communication (58), Economy (359), Education (285), Food (34), Geography (60), Medical (73), Simulation (16), Travel (165) and Weapon (40).

### B. Data Preprocessing

Several natural language processing (NLP) techniques were applied to the document corpus during preprocessing, each of which have been discussed in detail next.

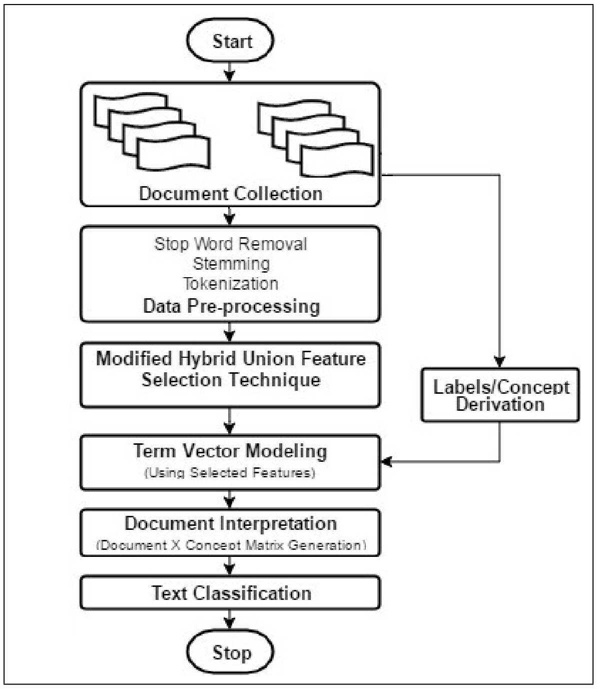

Fig. 1: Workflow of the proposed methodology

*1) Element name extraction :* Since WSDL-TC dataset contain XML files and the service-specific information present as element names, hence we used a XML parser to extract only element name-phrases from the XML DOM tree. This process is not required for Reuters documents.

*2) Tokenization: :* a process of splitting any document into words or tokens. A token is the instance of stream of characters on which pre-processing techniques are applied. The element name-phrases obtained after WSDL-TC parsing are processed to obtain tokens which forms the initial feature space for each document. Same process is applied for Reuter-21578 corpus.

*3) Stopword Removal: :* In the initial feature space, low-value words often occur (for example, *is, am, we, thus, where, a, the, who, be, also, on* etc), which contribute very little towards the domain of each category. Hence, these stopwords are removed by using a standard English language stopword list, thus reducing the computational complexity.

*4) Stemming: :* A process of stemming is used to identify root word from the derivationally related formatted term. For example, consider words like *nationalist, nationalism, national* etc, which are derived from the original root word ‘nation’. Removing these multiple terms which have the same stem can further reduce the original term space. We used the Porter Stemmer [11] for performing stemming of terms obtained from element name-phrases.

*5) Document Vector formation::* After generating a global dictionary of features obtained from all documents in the corpus, a $Document \times Feature$ matrix is created, which contains feature frequency w.r.t to each document. This matrix representation of documents in terms of features is also known as a bag-of-word (BOW) representation.

978-1-5386-3039-6/18/$31.00©2018 IEEE

*6) Similarity calculation: :* To differentiate between documents based on their content and context, a similarity based measure was applied to the document vectors. It generates a score which is used to decide the clusters. We used cosine correlation measure which is given by Eq. (1), where, $x_p$ and $x_j$ are the two document vectors, between which the similarity is to be measured.

$$cos(x_p, x_j) = \frac{x_p.x_j}{|x_p||x_j|} \tag{1}$$

### C. Modified Hybrid Union Feature Selection Technique

To optimize categorization performance and improve accuracy, the proposed CSA technique uses a combination of feature selection techniques for achieving an optimal feature space. Modified hybrid union approach of feature selection [12] helps to extract top scored and most common features from the global dictionary. We used the following feature selection methods in the proposed work.

*1) Mean Absolute Difference (MAD):* This method is used to assign a relevance score based on the difference of sample weight and the mean value of a feature w.r.t all other documents. For experimental purpose, we considered a threshold relevance score value as 0.90. Eq. (2) gives the MAD value for $i^th$ feature, where, $X_{ij}$ is the value of $i^{th}$ feature with $j^{th}$ document and mean value can be calculated as per Eq. (3).

$$MAD_i = \frac{1}{n}\sum_{j=1}^{n}(X_{ij} - \bar{X_i}) \tag{2}$$

$$\bar{X_i} = \frac{1}{n}\sum_{j=1}^{n}(X_{ij}) \tag{3}$$

*2) Absolute Cosine (AC):* AC helps in pruning out redundant features based on their similarity score. The number of similar characters present two different features will decide the similarity score and if it is greater than a threshold value then, that term will be discarded. A threshold value of 0.40 was considered for similarity in this case. Absolute Cosine values for a term can be calculated using Eq. (4), where, $W_i$ and $W_t$ are derived features. The denominator denotes the euclidean norm which measures the angle between two vectors and has value between 0 and 1, where 0 means features are orthogonal and 1 means features are co-linear.

$$cos(\theta_{w_i}, w_t) = \left|\frac{(W_i.W_t)}{||W_i|||W_t||}\right| \tag{4}$$

Among MHU techniques, a combination of MAD and AC achieved better performance over other feature selection techniques [13]. We considered the union of 20% features selected from the union of MAD and AC models and 80% features selected from intersection of MAD and AC models as shown in Eq. (5).

$$MHU_{FS} = 20\%(MAD \cup AC) \cup 80\%(MAD \cap AC) \tag{5}$$

### D. Concept Derivation

Categorizing text documents is not just a statistical or a mathematical problem like image categorization and credit scoring. It is a language modeling problem where the meaning of document content should be identified by a classifier. However, numerical calculation oriented classifiers do not treat text classification as a language problem. Semantic analysis technique helps to increase the capability of a system to understand the meaning of a word by its occurrence in the specific document. *Concepts* cannot be practically defined without external knowledge, but it is also impracticable to define a global set of concepts for all classifiers. As information can be obtained only from the corpus, it is not easy to understand the relationship between words and concepts which are irrelevant to files in same corpus. Concepts can be obtained from any word or any phrase present in the corpus. Depending on the type of corpora used in existing literature, three methods have been proposed for concept derivation [14].

*1) Direct Derivation:* Direct derivation is the most simple way of getting concepts, here category labels are considered as concepts. This technique is applicable for all types of corpora.

*2) Split Derivation:* The number of high level labels in hierarchical corpora are considered as generalized labels, which may not express complete information as low-level labels. Hence, it is more helpful to use low-level labels than high-level labels. Each high-level label is further divided into low-level labels so that more information can be extracted.

*3) Combined Derivation:* In most corpora, labels or categories are formed as per requirement, but it is not always good to have specialized labels for all sets of corpora. Reuters-21578 corpora has labels like 'wheat', 'corn', 'grain' etc which are related to food, hence the high-level label '*Food*' can be used instead of 3 different labels. This technique is known as combined derivation, which is used for concept derivation in multi-label corpora.

### E. Modeling the Term Vector

For deriving the relationship between terms and concepts, the process of modeling the term vector plays an important role in CSA. Terms can be anything amongst any index, phrase or word which is used to define the contents of the document. Wang and Domeniconi [15] found that phrases do not contribute more towards accuracy, while words when considered as a term achieve comparatively better performance. We too have considered a word as a term in our work.

In the process of modeling term vector, only positive samples, i.e., the terms which belong to a particular concept are only considered. We also considered negative samples, as increasing the membership of a term with one concept implies a decreasing membership with other concept, hence all training samples must belong to any of the derived concept. The size of the document and the frequency of a particular term in the document is a deciding factor in the calculation of membership of any term w.r.t concepts. In this paper, the term vector was modeled as per Eq. (6).

$$W(c_i, t_j) = \sum_k H(c_i, d_k)\frac{log(1 + tf(d_k, t_j))}{log(1 + length(d_k))} \tag{6}$$

978-1-5386-3039-6/18/$31.00©2018 IEEE

Here, if $d_k$ belongs to $c_i$, then, H($c_i$,$d_k$) = 1, otherwise H($c_i$,$d_k$) = 0.

### *F. Document Interpretation*

After the CSA term modeling process, every document is now represented as a vector in a concept space. As the concepts are defined under human observation, i.e. we directly considered the category labels as concepts in the given corpus, CSA helps provide word interpretation in the form of a concept. Once the word vectors is formed, they can directly be used as samples for a classification model.

As each word contributes differently to the context of different documents, various weighting methods were employed to measure their relative importance. A well-known technique called Term Frequency Relevance Frequency (tfrf)[16] measures the relevance of a document in the concept space. It is calculated using the term vector and term weighting vector i.e., the vector is formed as per Eq. (7).

$$tfrf(d_k,t_j) = t_f(d_k,t_j) * log(2 + \frac{a}{c}) \qquad (7)$$

where, $a$ is a count of positive documents that contains term $t_j$ and $c$ is a total count of documents containing term $t_j$.

$$Wd(c_i,d_k) = \sum_{t_j \in d_k} W(c_i,t_j) * tfrf(d_k,t_j) \qquad (8)$$

where, $d_k$ is calculated as per Eq. (9).

$$\vec{d_k} = \sum_{t_k \in d_k} tfrf(d_k,t_j) * \vec{t_j} \qquad (9)$$

### *G. Text Classification Learning Algorithms*

To evaluate the performance of the proposed CSA technique for text categorization applications, we applied various learning algorithms to the final document vectors generated using it.

*1) Support Vector Machines (SVM):* SVM is the best known learning classification algorithm used to classify both liner and non-linear data. If a data contains linear values then it can be separated using linearly drawn optimal separable hyperplane i.e. the linear kernel; this is deciding factor between data of two classes. If the data content is not separable by linear hyperplane, then SVM uses nonlinear mapping for transforming data into higher dimensions. We used both linear and nonlinear SVM in our work and found linear SVM performed well on Reuters dataset whereas non-linear SVM performed well with WSDL-TC. Mathematically, the separating hyperplane can be written as per Eq. (10).

$$W.X + b = 0 \qquad (10)$$

where, $W$ is a weight vector and $w = w_1, w_2, ..., w_n$. $X$ is a training tuple whereas $b$ is a scalar. The parameters $b$ and $W$ decide the offset of the hyperplane from the origin along the normal vector $W$.

*2) Naïve Bayes Classifier:* Naïve Bayes (NB) is a probabilistic model based on Bayesian Theroused for text classification which is based on a Bayesian theorem. We have used the multinomial NB model since it performed better compare to other variations of NB and it gave good results in text categorization. Consider a training data D where every tuple $X$ is represented by $n$-dimensional feature vector such as $X = x_1, x_2, .., x_n$ where X is a tuple with n attributes or features. Assuming that corpus has $m$ different classes, $C_1, C_2, ..., C_m$, the Naïve Bayes classifier can predict that the class $X$ belongs to $C_i$ iff: $P(C_i \mid X)\, P(C_j \mid X)$ where $i$, $j$ in the range of 1 to $m$. $P(C_i \mid X)$ is computed as per Eq. (11),

$$P(C_i \mid X) = \prod_{k=1}^{n} P(X_k \mid Ci) \qquad (11)$$

*3) Random Forests: :* an ensemble learning method based on the concept of bagging which is used for classification and regression. The classification decision is based on the voting by multiple decision trees on training samples and prediction w.r.t the class for individual tree. Studies have shown that Random Decision Forest method produces good results for high dimensional data with high dimensions, and repetitive random sampling on training data helps in producing robust and quicker results.

Let D be the total number of documents i.e. data points and T be the number of terms extracted for experiment and C be the total number of concepts i.e. categories. Therefore, classifier will create $k$ bootstrap samples of $D$. Every sample is then denoted by $D_i$, and $D_i$ has the same number of tuples as $D$ which mean some tuples are not included in $D_i$ whereas some are included more than once. The classifier then will create k different decision trees for each $D_i$. To classify any unknown sample $X$ prediction counting by vote is used and $X$ then will be assigned to the class with the most votes. We used WEKA for implementing decision tree algorithm on document vector, by considering maximum 10 trees in the model.

*4) Adaboost M1:* Adaboost i.e. adaptive boosting is the first successful boosting machine learning meta algorithm designed to boost up the performance of decision tree on binary classification problems. For given training samples, the algorithm calculates the uniformly distributed weights $W$. After every time span $t$, Adaboost recalculates the distribution with factor $W_t$ as per the predicted values on training samples. In every cycle, a higher weightage is given to the misplaced samples where as correctly placed samples get a lower weightage. The process continues till the $T$ cycles and finally all training data points are linearly combined into a single hypothesis. Data points with greater weights can be also assigned to component classifiers with lower training errors.

## III. Experimental Evaluation and Results

To evaluate the performance of the proposed technique, we used well-known evaluation measures like - Purity, Receiver operating characteristic (ROC) curves, Rand Index and F1 Measure which are discussed in detail here.

### *A. Purity Measure*

Results of classification models should be evaluated depending on the correctly placed data points. We used a well-known purity measure to assess the performance of the models. Purity can be calculated as ratio of the number of correctly placed objects in each class to the total number of objects considered for study. Fig. 2 & 3 gives purity percent values for each classification technique using conventional and MHU model.

$$Purity(\%) = \frac{\sum_0^k \max_0^j (Samples\ obtained\ in\ each\ class)}{Total\ number\ of\ Documents} \tag{12}$$

### B. ROC and Area under Curve region (AUC)

ROC curve is graphically represents the performance of binary classifiers and AUC is a way of summarizing classifier performance in a single value. ROC curves are drawn as a plot of False Positive Rate (FPR) values on X-axis against the True Positive Rate(TPR) on Y-axis. Here, the FPR and TPR are obtained from the classifier's confusion matrix. More the area under ROC curve, better is the classifier's performance. Conventionally, a classifier with an AUC value of 0.9 to 1 is considered to be the best classifier. In Fig. 5 to 8 graphs of area under ROC curves are shown.

### C. Rand Index (RI)

Rand Index is a measure of similarity between two data clusters which considers all samples that are correctly or incorrectly assigned between different or same clusters. For model evaluation, we considered Rand Index between different category samples. RI varies between 0 and 1, 0 indicates samples identified between two categories are totally different, whereas 1 indicates samples are exactly same between two categories (given by Eq. (13)).

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

where TP is True positive, FP is False positive, TN is True Negative and FN is False Negative.

### D. F1 Measure

The F1-measure or F-score is a measure of accuracy, which can be calculated as a weighted average of precision and recall. It varies between 1 and 0 where 1 is considered to be best and 0 as worst and is calculated as per Eq. (14).

$$F - score = \frac{2P * R}{P + R} \tag{14}$$

where,

$$Precision\ P = \frac{TP}{TP + FP} \tag{15}$$

$$Recall\ R = \frac{TP}{TP + FN} \tag{16}$$

The observed results from the experimental analysis are tabulated in Table I, II, III and IV. A reduction of 1/4th in execution time was observed when the proposed CSA with MHU was used for classification, as can be observed from Table III and Fig. 4. The reason for this is evident from Table IV, as there is significant decrease in number of features finally selected for representing documents in the concept space when the proposed method is used.

From Table I, it can be seen that different classification models performed differently for the two datasets, with and without CSA with MHU. For Reuters dataset, the linear SVM model and Random Forests achieved the best results, while for WSDL-TC data set, SVM with linear & polynomial kernels and Random Forest techniques produced good results. Fig. 2 depicts the comparative results of purity of each classification model using conventional feature selection techniques, while purity values for classification models using modified hybrid union approach is highlighted in Fig. 3. From Table I and II, it is evident that the proposed MHU approach performed very well in comparison to conventional feature selection techniques, with an improvement of over 5% in accuracy and significantly computationally faster that the traditional approach. In Fig. 5 to 8, the plots of area under ROC curves are shown. The curves are drawn considering False Positive Rate on X-axis against True Positive Rate on y-axis.

TABLE I: Experimental statistics using traditional feature selection techniques (without MHU)

| **Classification Models** | **Purity** | | **ROC/AUC** | | **Rand Index** | | **F1 Measure** | |
|---|---|---|---|---|---|---|---|---|
| | **Reuters** | **WSDL-TC** | **Reuters** | **WSDL-TC** | **Reuters** | **WSDL-TC** | **Reuters** | **WSDL-TC** |
| SVM Linear | 83.56% | 92.01% | 0.884 | 0.963 | 0.861 | 0.9374 | 0.829 | 0.92 |
| SVM Polynomial | 53.36% | 93.11% | 0.655 | 0.961 | 0.4381 | 0.9386 | 0.49 | 0.931 |
| Naive Bayes | 60.67% | 37.98% | 0.902 | 0.901 | 0.6 | 0.8075 | 0.505 | 0.674 |
| Random Forest | 88.88% | 94.03% | 0.98 | 0.991 | 0.9179 | 0.954 | 0.888 | 0.94 |
| Adaboost M1 | 47.19% | 49.63% | 0.67 | 0.688 | 0.5607 | 0.5937 | 0.366 | 0.375 |

TABLE II: Experimental statistics using Modified Hybrid Union feature selection techniques (With MHU)

| **Classification Models** | **Purity** | | **ROC/AUC** | | **Rand Index** | | **F1 Measure** | |
|---|---|---|---|---|---|---|---|---|
| | **Reuters** | **WSDL-TC** | **Reuters** | **WSDL-TC** | **Reuters** | **WSDL-TC** | **Reuters** | **WSDL-TC** |
| SVM Linear | 87.77% | 89.26% | 0.912 | 0.928 | 0.8711 | 0.918 | 0.877 | 0.893 |
| SVM Polynomial | 76.24% | 91.28% | 0.808 | 0.942 | 0.7271 | 0.9309 | 0.735 | 0.913 |
| Naive Bayes | 61.03% | 51.55% | 0.845 | 0.879 | 0.5744 | 0.7856 | 0.559 | 0.632 |
| Random Forest | 90.17% | 91.74% | 0.974 | 0.987 | 0.8992 | 0.9408 | 0.901 | 0.917 |
| Adaboost M1 | 62.73% | 49.08% | 0.7 | 0.671 | 0.6156 | 0.5829 | 0.556 | 0.371 |

TABLE III: Total Execution Time

| Technique | Reuters-21578 | WSDL-TC |
|---|---|---|
| Without MHU | 274.4 min | 8.49 min |
| With MHU | 48.2 min | 2.39 min |

TABLE IV: Selected Feature Count

| Technique | Reuters-21578 | WSDL-TC |
|---|---|---|
| Without MHU | 11181 | 1370 |
| With MHU | 686 | 141 ‘ |

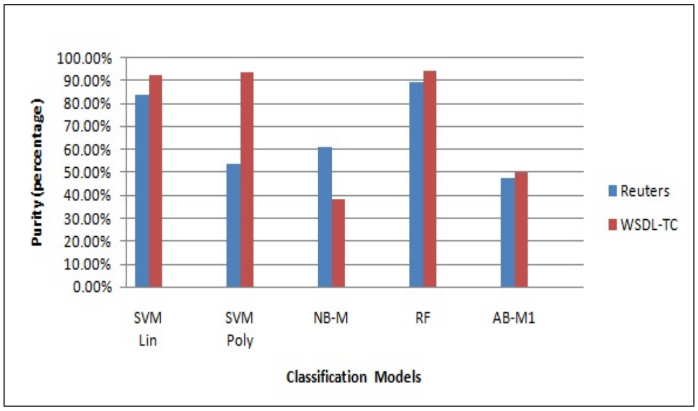

Fig. 2: Observed purity values using traditional feature selection techniques

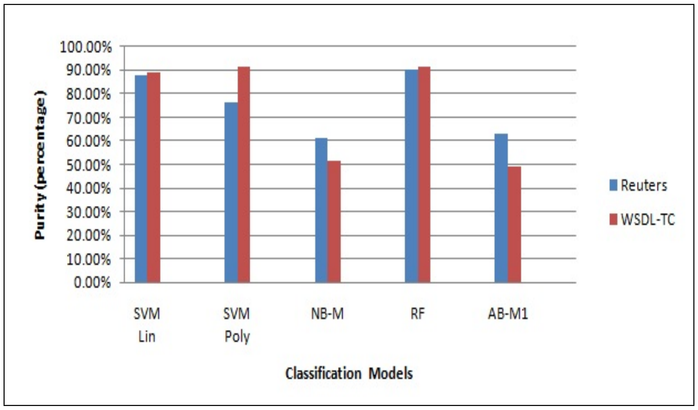

Fig. 3: Observed values of purity using proposed MHU

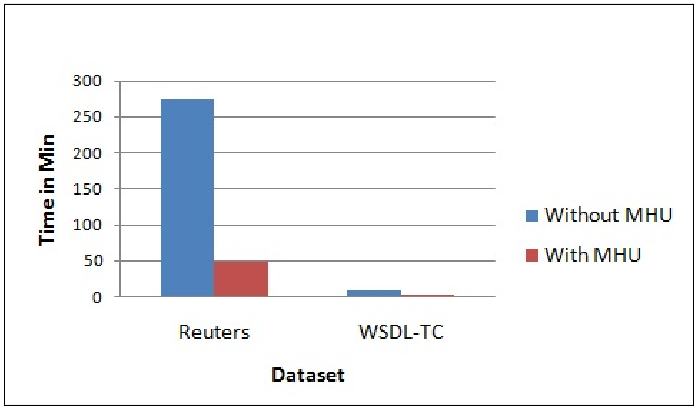

Fig. 4: Total Execution Time comparison

## IV. Conclusion and Future Work

In this paper, a concise semantic analysis (CSA) technique based on a novel feature selection technique called Modified Hybrid Union was presented. The proposed CSA technique incorporate effective tfrf feature weighting method which considers the length of document for finding the relationship between documents and words in a concept space. Further, the use of newly adopted feature selection method i.e. modified hybrid union (MHU) approach helps in reducing computational time by a quarter over traditional techniques in addition to 5-10% increase in classification accuracy. It was also found that the concepts derived by the proposed derivation method were suitable for all types of corpora. Text categorization experiments performed on two different corpora viz. Reuters-21578 and WSDL-TC, showed significant improvements achieved using MHU and CSA, in comparison to traditional feature selection methods. As part of future work, we intend to study the effect of the proposed technique on large, diverse corpora, and also optimize it through parallelization.

## Acknowledgment

This work was carried out when the first author was a Masters student at the Department of Information Technology, National Institute of Technology Karnataka (NITK) Surathkal. The authors gratefully acknowledge the use of NITK-IT facilities and services during the course of this research work.

## References

[1] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 42–49. [Online]. Available: http://doi.acm.org/10.1145/312624.312647

[2] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975. [Online]. Available: http://doi.acm.org/10.1145/361219.361220

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, vol. 41, no. 6, pp. 391–407, 1990.

[4] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *J. Artif. Int. Res.*, vol. 34, no. 1, pp. 443–498, Mar. 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622716.1622728

[5] P. Wang and C. Domeniconi, "Building semantic kernels for text classification using wikipedia," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 713–721. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401976

[6] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved k-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503 – 1509, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411011511

[7] J.-Y. Jiang, S.-C. Tsai, and S.-J. Lee, "Fsknn," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2813–2821, Feb. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.08.141

[8] C. Silva and B. Ribeiro, "Two-level hierarchical hybrid svm-rvm classification model," in *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*, Dec 2006, pp. 89–94.

[9] D. Lewis, "Reuters-21578," 1997.

[10] M. Klusch, B. Fries, and K. Sycara, "Automated semantic web service discovery with owls-mx," in *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM, 2006, pp. 915–922.

978-1-5386-3039-6/18/$31.00©2018 IEEE

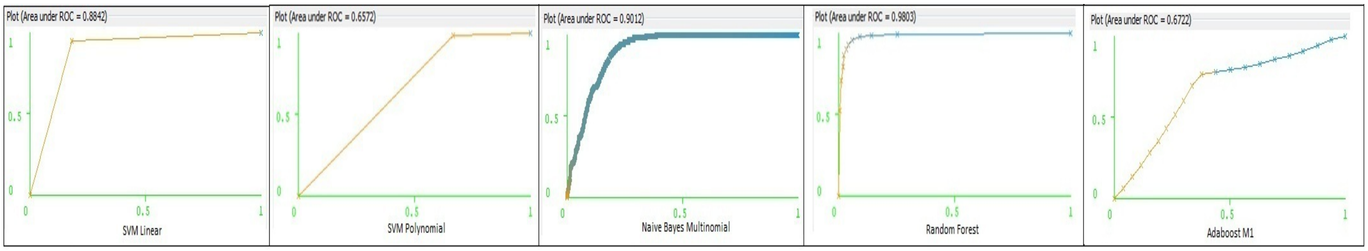

Fig. 5: Area under ROC curve for Reuters-21578 dataset (without MHU)

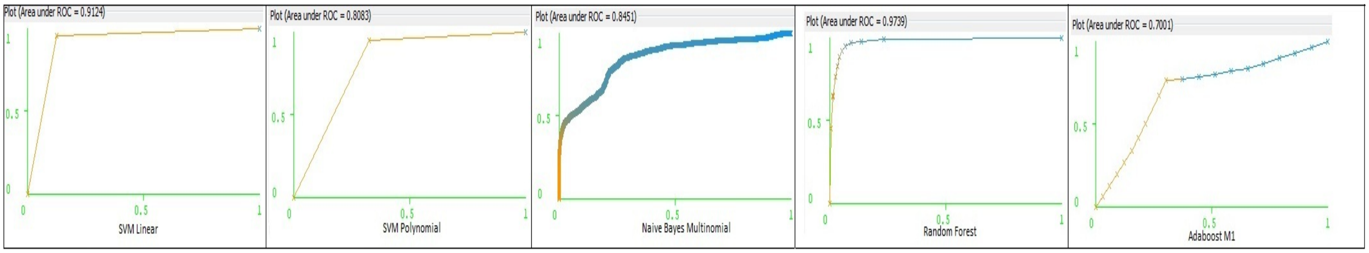

Fig. 6: Area Under ROC curve for Reuters-21578 dataset (with MHU)

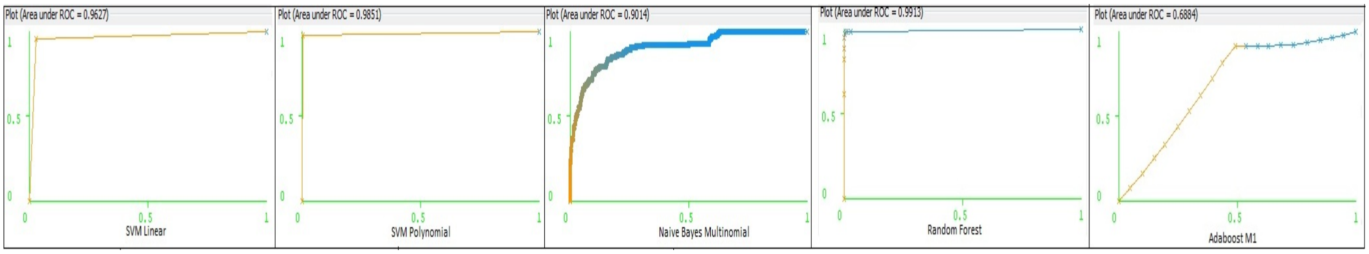

Fig. 7: Area under ROC curve for WSDL-TC dataset (without MHU)

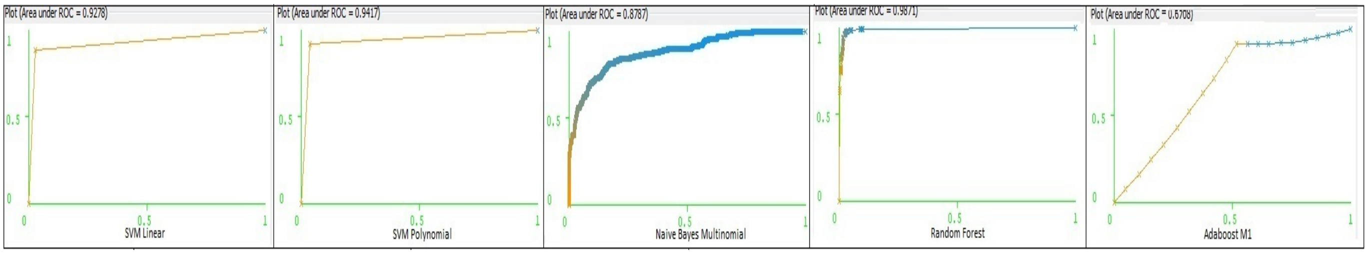

Fig. 8: Area under ROC curve for WSDL-TC dataset (with MHU)

[11] M. F. Porter, “Readings in information retrieval,” K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. An Algorithm for Suffix Stripping, pp. 313–316. [Online]. Available: http://dl.acm.org/citation.cfm?id=275537.275705

[12] K. K. Bharti and P. K. Singh, “Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering,” *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3105–3114, Apr. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2014.11.038

[13] A. P. Bhopale and S. S. Kamath, “Novel hybrid feature selection models for unsupervised document categorization,” in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sept 2017, pp. 1471–1477.

[14] Z. Li, Z. Xiong, Y. Zhang, C. Liu, and K. Li, “Fast text categorization using concise semantic analysis,” *Pattern Recogn. Lett.*, vol. 32, no. 3, pp. 441–448, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2010.11.001

[15] X. B. Xue and Z. H. Zhou, “Distributional features for text categorization,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 3, pp. 428–442, March 2009.

[16] M. Lan, C. L. Tan, J. Su, and Y. Lu, “Supervised and traditional term weighting methods for automatic text categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721–735, April 2009.