# ABSTRACT

The growth of Internet has led to a great deal of interest in developing useful and efficient tools to support users in searching the Web. Exploitation of the Internet usage has grown to multifold, which has become an annoyance for the Internet users. So, the necessity of retrieval of relevant information has become an essential requirement. The ability to automatically classify documents into a fixed set of categories is highly desirable to improve the relevancy of the search results. Document classification aims at learning the properties of each category from a set of training documents and classifies unknown documents to an appropriate class.

The classification task has become more challenging due to the unstructured nature and high dimensionality of text documents. A review of literature reveals the importance of feature selection in document classification. The main goal of this research work is to improve the accuracy and relevancy of classification by using suitable feature selection approaches to remove redundant and irrelevant features. Performance of the classifier is also improved by building labeled documents from unlabeled web documents. Experiments were conducted with two benchmark datasets, namely, Reuter's 21578 and 20Newsgroup.

The following are the major conclusions arrived from the research work:

i. Experimental analysis depict that Latent Dirichlet Allocation (LDA) provides better result with an average F1 measure of 86% for Reuter's dataset and 88% for 20Newsgroup dataset when compared to Term Frequency-Inverse Document Frequency (TF-IDF), CHI Square, Rough set approach, Concept based approach and Latent Semantic Indexing (LSI).

ii. New words are constantly being created and existing words are assigned with new senses in the Web. Latent Semantic Indexing (LSI) is a process that uses Singular Value Decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of web snippets. The research work trains SVM and ABC algorithm with the keywords and concepts extracted from web snippets using LSI. For Reuter's dataset, the proposed feature selection approach recorded an increase in F1 measure of 5% with SVM classifier and 3% with ABC algorithm when compared to features obtained from Reuter's dataset alone. The F1 measure of classification with the proposed feature selection algorithm for 20Newsgroup dataset is improved by 3% with SVM classifier and 2% with ABC algorithm when compared to features obtained from 20Newsgroup dataset alone.

iii. Combination of multiple feature selection methods outperformed a single method. Features extracted using LDA are weighted using CHI-square, Normalized Google Distance (NGD) and TF-IDF approaches. The features are scored and ranked based on the weights assigned. Combinatorial Fusion Analysis (CFA) is performed to combine multiple feature selection methods to obtain discriminate features and to reduce the features in the feature set. SVM classifier for the individual methods like LDA with TF-IDF, LDA with CHI and LDA with NGD recorded F1 measure of 92%, 92% and 89% with top 550 features for Reuter's dataset. The F1 measure of the SVM classifier for the fusion scheme is 95% with the same number of features for Reuter's dataset. SVM

classifier for the individual methods like LDA with TF-IDF, LDA with CHI and LDA with NGD recorded F1 measure of 92%, 90% and 91% with top 550 features for 20Newgroup dataset. The F1 measure of the SVM classifier for the fusion scheme is 94% with the same number of features for 20Newsgroup dataset.

iv. Performance of feature weighted LDA and fusion of multiple feature selection methods are verified for online news document dataset from the BBC news and the Guardian websites. In order to improve time efficiency, preprocessing and LDA topic modeling has been parallelized. The parallel approach shows 59% improvement in speed and 39% improvement in CPU utilization when compared to serial approach for the news document dataset prepared using the BBC news and the Guardian websites. Feature selection using fusion method improves the F1 measure by 6% for 'Business' category, 2% for 'Health' category, 3% for 'Sports' category and 2% for 'Technology' category when compared to LDA with TF-IDF, LDA with CHI and LDA with NGD. Further, 25% reduction in features is obtained when compared to term weighted LDA feature selection approach.

The experimental results indicate that fusion of multiple feature selection methods outperformed when compared to individual feature selection methods for the classes in Reuter's and 20Newsgroup data sets. The classification accuracy recorded approximately 94%, which is comparable to the performance of other feature selection and feature weighting approaches like CHI square, TF-IDF, Concept based, Rough set, LSI, LDA, LDA with CHI, LDA with TF-IDF and LDA with NGD.