CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENTS

7.1 CONCLUSION

In this research work, a new Intelligent Topic Classification model is proposed which uses semantic information of un-structured text across multiple domains to support IR system. This research work portrays a semantic understanding based approach to tackle the issues in the text mining and provides a better performance than BOW method. Six different research issues have been identified and investigated concerning the topic classification.

Many works have been proposed in the literature over the past few decades. The important works related to the semantic representation and topic classification problems have been surveyed and are presented in the thesis. The literature survey is presented to illustrate that more work on similar problems have been carried out. The literature related to the existing knowledge bases and similarity metrics for semantic relationship of terms in the document are discussed. The important challenges in the existing system are highlighted and motivation for the proposed work is presented in the Chapter 2.

A new topic classification model for web content classification using Hypernym as a semantic feature is proposed and discussed in Chapter 3. It focused a knowledge based feature extraction methods. The semantic

feature extraction through Hypernyms and their importance for supervised classification methods are studied. NLP operations for preprocessing the documents and semantic feature extraction through WordNet knowledgebase are experimented. The Hyponym word tree is created for each document and features are extracted through weighted TF-IDF where word weight is computed based on Hypernym number present in a radix tree. The k-NN and NB are the learning models are used for classifying the input documents using the proposed Hypernym based feature extraction . The performance of the proposed retrieval model is evaluated through the standard dataset in terms of precision, recall and accuracy.

An enhanced Topic Classification model with rich semantics is proposed in the Chapter 4. A new algorithm for Semantic Deep Learner is developed to improve the conventional deep learning algorithm by incorporating document-level semantic information for topic classification. To capture the semantic information from each document, probabilistic based Semantic Smoothing technique is constructed. Thus corpus based feature selection is achieved through semantic smoothing. The training documents are applied to various processes such as un-supervised clustering and semantic smoothing to generate the significant key features to train Semantic Deep Learner. Then, SDL is trained using those features by forming frequency matrix. A separate semantic layer is constructed by exploiting deep architecture based on the sequence of RBM. The results show that, under appropriate parameter settings the proposed topic classification model using SDL significantly outperforms the Neural Network classifier by using two WebKB and 20Newsgroups dataset.

An Intelligent Topic Classification Model is constructed with Decision List based Word Sense Disambiguation, Semantic Smoothing and Semantic Deep Learner(SDL) especially to handle ambiguity problem in

polysemeous word in the training documents. The SDL with respect to DLWSD is termed as Enriched Semantic Deep Learner(ESDL) that encompasses additional feature for training. The deep semantic similarity is achieved to predict the category of documents. The performance is evaluated through 20Newsgroups, WebKB and Electrical datasets. The results obtained using ESDL is presented in Chapter 5. It is evidenced that the accuracy of ESDL is slightly increased in some cases when compared to SDL and it very high when compared to Neural Network classifier.

A new patent classification system is proposed and implemented by adapting SDL. The ability of deep learning approach to discover the hidden structures and features at different levels of abstraction is used for efficiently classifying patents. The correlation matrix constructed through the basic and semantic features are effectively utilized for training the patents. Compared with Neural Network classifier, the proposed patent classification using SDL performs better prediction. The results obtained through various granularity of the patent documents are presented in the chapter 6.

At the nutshell, the overall results suggest that the development of Intelligent Topic Classification model and its application impart a better idea in the way of attaining remarkable classification accuracy through the number of experiments with standard datasets. Empirical evidence from the evaluation of the explored Topic Classification methods indicates that the proposed model provides valuable solutions for reducing the gap between statistics and semantics of text, achieving results superior to traditional supervised machine learning techniques.

7.2 FUTURE ENHANCEMENTS

Future enhancements to this work include the adaptation of the proposed model in the various domains in which the semantics of un-

structured text plays a predominant role. The present work can be evolved by including more semantic resources for feature extraction and semantic similarity metrics. Even though, the model utilizes the semantics of words and phrases, the sentence level semantics are not addressed and hence there is a room for incorporating those features for improving the classification accuracy for better prediction. The training time of the SDL for large corpus can be reduced by providing the distributed way of processing.

Furthermore, the model can be applied in Information Security domain to classify the information contents in the context of the confidentiality. The model can be deployed for sentiment analysis based on the opinion mining. The proposed model can be used for Twitter trending topic prediction. By having a domain specific knowledge bases such as SNOMED-CT and MESH instead of WordNet, the proposed semantic based model can provide a new direction of Research in the Medical domain for emergent disease classification.