CHAPTER 6

EMPLOYING INTELLIGENT TOPIC CLASSIFICATION MODEL IN PATENT DOMAIN

6.1 INTRODUCTION

Knowledge documents are growing remarkably to serve the organization for information processing and various management tasks. Patents are rich knowledge source needed to be organized efficiently and conveniently thereby providing a way for gathering business intelligence and identifying key trends in technology development. The main focus of this work is to propose a patent classification methodology based on Semantic Deep Learner (SDL). In this approach, the key terms are extracted by two proposed feature selection methods. The first method is achieved through the weighted clustering and the second method of selection is through correlation coefficient in which the highly correlated terms in a document corpus are projected as a reduced set of terms. The correspondence between the patent documents and key features generated through both the methods in the form of matrix are used for training SDL and accordingly patents are classified. The trained SDL is further utilized for testing the patents without the topic or class label. The target output identifies the category of patent information with respect to hierarchical classification scheme of the International Patent Classification (IPC) standard. The accuracy of the proposed system is considerably improved when compared to other state of the art classifier and moreover the current approach is new to the Patent domain.

6.2 SEMI-STRUCTURED PATENT DOCUMENT

Patents are semi-structured documents that focuses on the explicit knowledge management (Turban & Aronson 2001). The International Patent Classification (IPC) is a standard taxonomy developed and administered by the World Intellectual Property Organization (WIPO) for classifying patents and patent applications(Tikk et al 2005). The IPC covers all areas of technology and is currently used by the industrial property offices of more than 90 countries. The use of patent documents and the IPC for research is interesting for several reasons. The IPC covers a range of topics that spans all human inventions and uses a diverse technical and scientific vocabulary. A large part of it is concerned with Chemistry, Mechanics, Computers, Electrical and Electronics. The organization has got several reasons to analyze the collection of the patents that are applied or granted. For example, the Research lab run by Government requires to access the patent details for knowing about the field where more number of patents are registered and where the patents are lagging. Another example is where a Research and Development unit of industry needs to find its close competitor with respect to the number of paper publications spanning different research areas. Looking from economical side, Intellectual Patent Rights (IPR) are becoming one of the most important mechanisms for business in extracting economic value from creativity and encouraging greater investment in innovation.

A Patent document has got a prescribed format which includes structured and unstructured data. Structured data are located on the front page and provide bibliographic information of the granted patent or patent application, such as document number, filing and publication dates, name of the inventors, assignees, addresses. The rest of the document comprises unstructured text given in predefined parts. These parts are: title, abstract that gives a summary of the technology of the invention, detailed description that

discloses the technical details of the invention normally illustrated by working examples showing how to practically realize the invention, claims that define the scope of protection for the invention and represent the legal aspect of the patent document. Text classification approaches for patent classification problems have to manage simultaneously very large size of hierarchy, large documents, huge features set and multi-labeled documents (Karki 1997). The clustering algorithms for analyzing the various parts of the patents were investigated by Miroslava Drazic et al (2013), the accuracy values are estimated with only few testing patents. The k-Means and Neural gas algorithms provide limited accuracy.

6.3 PATENT CLASSIFICATION METHODOLOGY

In this work, a novel approach for applying Semantic Deep Learner for classifying the patents related with the electrical field is proposed. The abstract and claim information are considered for classification tasks. Initially, the patent documents are preprocessed using NLP tasks and bifurcated into training and testing documents. The proposed SDL is trained with two sets of key features generated from patent documents. The one set of features are generated through centroids of clusters and another set is generated through correlation measure. The first set of the key features and corresponding documents are represented as document-term matrix $DTM_{cent.}$ The feature set generated through second method is represented as DTM_{corr} . These two matrices are supplied as an input to SDL.

The performance of SDL using DTM_{cent} and DTM_{corr} are analyzed separately. As deep learner is better than neural networks and other learners, the accuracy of the classification is improved to some extent especially for patent document. Applying deep learning technique for classification of the documents is a new approach in patent domain.

6.3.1 Patent Extraction and Representation

The input patent documents are preprocessed by removing the insignificant terms using stemming and stop words removal techniques . The resultant documents are bifurcated into training and testing sets. These training and testing documents are further analysed to extract the additional semantic features with the help of WordNet. The extracted training features(terms) are futher processed to generate the shortlisted key terms that is made ready for training.

6.3.2 Selection of Key Terms Using Cluster Centroids

Intially, preprocessing is carried out on set of patent documents to generate training and testing set. Next, the semantic features are extracted for each term in a training document through WordNet and incorporated as additional terms. For selecting the significant features for training, each one-length terms are identified and weighted using TF-IDF measure (as discussed in chapter 4).Similarly, two-length unique terms (phrases) are identified and weighted using the same measure. Note that the current weighting method identifies infrequent terms in the patent document. The calculated weights are represented as a VSM. Next, the Clustering process is applied to group the weighted vectors based upon their similarity score. Here, the similarity score is calculated using Cosine Measure (as discussed in Chapter 4). The vectors in each clusters are ranked based on the TF-IDF weighting. The vector with maximum weight is designated as a centroid of the cluster. These centroids are selected as a relevant key terms to form a document-term matrix DTM_{cent} for training SDL.

6.3.3 Relevant Term Selection Through Correlation Coefficient

This is an alternative method where the key terms are selected from the preprocessed training documents using the correlation measure given in

the Equation(6.1). Here, the dominant terms those occur frequently in a document are treated as significant key features. A correlation matrix is constructed by extracting all the corresponding significant terms and their frequency of occurrences in the same document.

Let CT_i and CT_j are the two significant features residing in set of patent documents, the correlation among these terms are represented as

$$X_{ij}=\frac{\sum_{i=1}^{C_d} X_{i,I}X_{j,I}-C_d\overline{X_iX_j}}{\sqrt{\sum_{i=1}^{C_d} X^2{}_{i,I}-C_d\overline{X^2}_i}\sqrt{\sum_{i=1}^{C_d} X^2{}_{j,I}-C_d\overline{X^2}_j}} \quad (6.1)$$

where $X_{i,j}$ denotes the correlation value of CT_i and CT_j in a set of patent documents; $X_{i,I}$is the number of times CT_i occurs in a document d_i ; $X_{j,I}$ denotes the number of times CT_j occurs in a document d_j; $\overline{X_i}$ denotes the mean frequency of CT_i occurring in all training documents; $\overline{X_j}$ is the mean frequency of CT_j occurring in all training documents;$C_d$is document count. The significant features that have high correlations are identified based on $X_{i,j}$ value and hence represented as related feature list. Finally, all the highly correlated features in the list are merged to form a training feature and make it easier to train SDL with reduced key features. Whenever a new patent document is loaded into a patent knowledge management system, the significant key features and their frequency of occurrences in a document are calculated. Then $XTf_{ij}$is used to represent the number of occurrences of related-feature XT_{ij}. The correlation of related feature $XT_{ij}$and $CT_i$is listed as $X_{ij}$and hence the final frequency of $CT_i$is given the Equation (6.2)

$$CTF_i=CTf_i+\sum_{j=1}^{n} XTf_{ij}.X_{ij} \quad (6.2)$$

Once CTF of all key features are selected , it is represented in a form of vectors as $\{CTF_1, CTF_2, CTF_3, \ldots\ldots\ldots CTF_n\}$. These generated key

features contribute to generate the correlation matrix DTM_{corr} used for training SDL.

6.3.4 Semantic Deep Learner for Patent Classification

The matrix DTM_{cent} generated through clustering process and DTM_{corr} generated through correlation method are provided as an input to SDL separately. The topic of the patent documents are treated as target and provided at the output layer. SDL is trained using the target given. The training is done by the hidden layers of SDL that exploits the target and the input. After training the system based on SDL, the testing is done by giving the testing patent documents for validation. During the testing process, the testing patents are given as input to the system, the DTM_{cent} and DTM_{corr} matrices are fabricated accordingly with the respective key terms that were already generated in the training phase. The semantic deep learner will give a score for the given input document and based on the score a patent document is classified to which topic it belongs.

- **Pre-training Phase with DTM_{corr}**

In the pre-training phase, the key feature vectors in the matrix are supplied as input to SDL for training. All key features are normalized between 0 and 1 using the normalization function given in the Equation (6.3) and supplied as input vector to the visible layer of SDL.

$$CTF'_i = \frac{CTF_{i,}}{\max(CTF_{1,}CTF_{2,}CTF_{3,}\ldots\ldots\ldots CTF_n)} \tag{6.3}$$

Every node in the input layer of SDL is thereby set with vectors of DTM_{corr}. The result generated through the activation function of the each node is received and propagated to the next layer. Likewise each node in the hidden layer also process the vector until vector reaches the final output layer. Hence

the input feature vectors from the input layer of node m to the node n of hidden layer is given as

$$T_n^h = \sum_{li \in prevlayer} w_{mn}^h v_m + bias_n \quad (6.4)$$

where w_{mn}^h is the weight of connection between input layer node m and hidden layer node n; The input value pertaining to node m is given as v_m, the bias value of the hidden layer n is denoted as $bias_n$. Then the output of hidden layer node n is determined using the Equation (6.5).

$$H_n = f(T_n^h) \quad (6.5)$$

The sigmoid activation function f(x) is given as $\frac{1}{1+e^{-x}}$ and the values from hidden layer to output layer of node p is computed using Equation (6.6)

$$T_p^0 = \sum_n w_{np}^0 H_n \quad (6.6)$$

where w_{np}^0 denotes the weight value in the link between node n of hidden layer and node k of output layer. Finally, the output of the SDL is determined by the Equation (6.7)

$$O_r = g(T_r^0) = g(\sum_n w_{nr}^0 H_n) \quad (6.7)$$

where g(x) denotes the activation function of output node r.

The error of the network is expressed as

$$\text{error} = \frac{1}{2} \sum_r (A_r - O_r)^2 \quad (6.8)$$

where A_r is the actual output of the given training data.

- **Fine-Tuning Phase with Target**

In this phase, the output values obtained in the output layer is propagated backward to the hidden layers. The main role of this phase is to carry out the weight adjustment by determining the error between target value and obtained output. As error is defined as the function of $O_r$and intern O_r is the function of w_{nr}^0 , therefore the weight is revised between output layer and the corresponding hidden layer and it is expressed in Equation (6.9)

$$\Delta w_{nr}^0 = Lr * ((A_r - O_r)\, g'(T_r^0) H_n = Lr * \partial_r^0 H_n \quad (6.9)$$

where Lr is the learning rate and $\partial_r^0 = (A_r - O_r) g'(T_r^0)$.

In the same way, the weight $w_{mn}^h$is adjusted between hidden layer and output layer by using the Equation (6.10)

$$\Delta w_{mn}^h = Lr * \partial_r^0 w_{nr}^0 f'((T_n^h) * H_n = Lr * \partial_r^h * H_n \quad (6.10)$$

therefore, the revised weight is given as

$$\Delta w_{mn} = Lr * \partial_n * f(T_m) \quad (6.11)$$

The resultant values demonstrates the goodness-of-fit between a given test document and all potential patent topics. In this approach, IPC provides the target topics for the patent documents.

When new patents are given as input to the model, the model automatically generates the testing matrix DTM_{corr} and it is fed as input to SDL to predict the category of the patent. In the same way the behavior of the model is studied using DTM_{cent} . The predicted result of SDL shows that features represented with DTM_{cent} if found to be better. To demonstrate the patent classification task, the sequential step carried out by the model is depicted in the Figure (6.1)

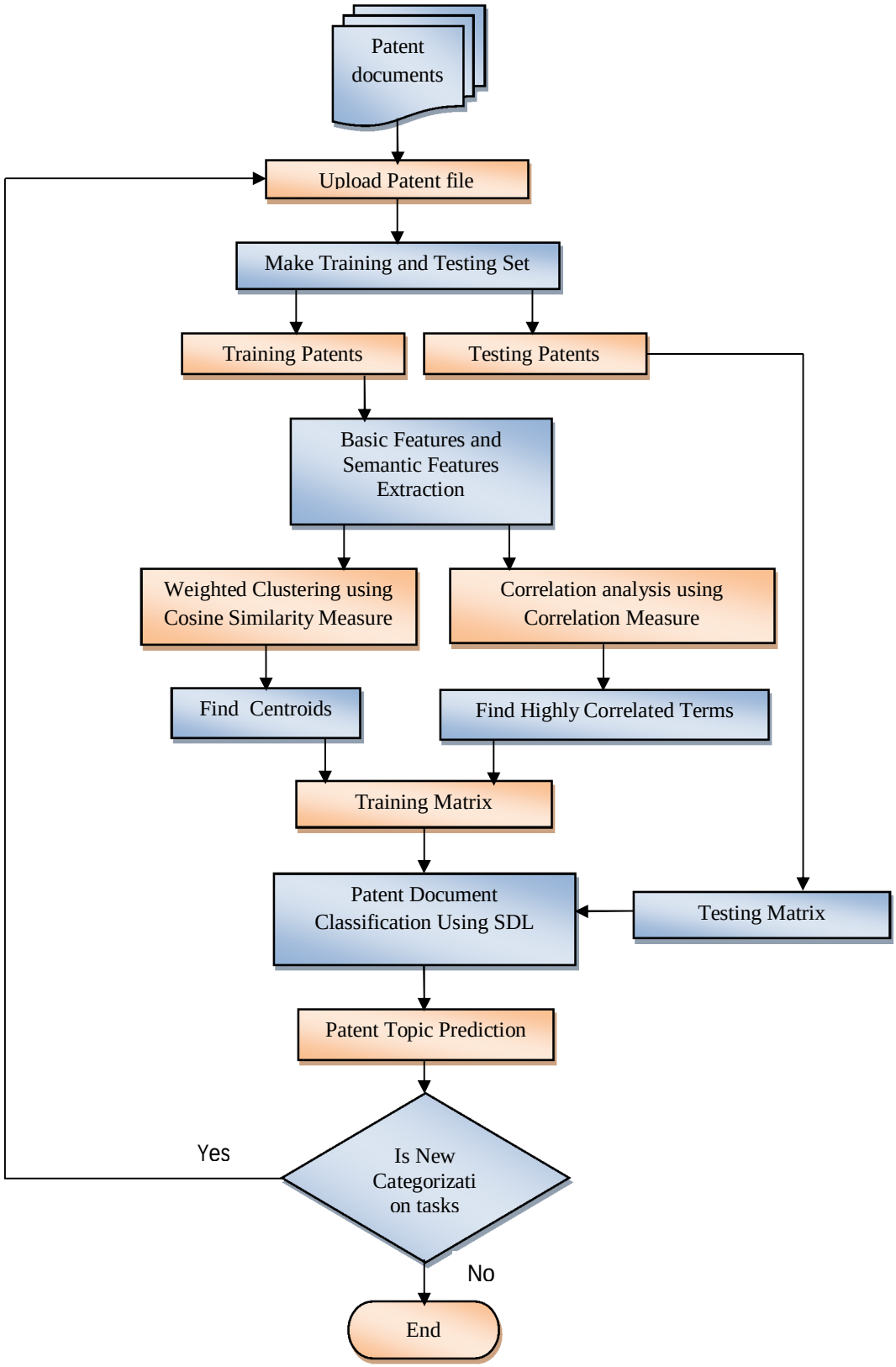

Figure 6.1 Sequential Steps in the Proposed Patent Classification

6.4 RESULTS AND DISCUSSIONS

The proposed method of patent classification is experimented with the 120 US patent documents that are extracted from Free Patent Online(FPO)- IP Research and Communities. The proposed system is executed in Java platform JDK 1.6 with Personal Computer configured with an i5 processor and 4GB main memory. All the patents collected for analysis pertains to electrical field under the topics Conductor, Connector, Devices and Outlets. The unstructured portion of patents includes abstract, description and claim. Only abstract and claim information of the patents are investigated and accordingly classified using the proposed patent classification model.

The performance of the system is measured in terms of accuracy by setting a learning rate as 0.3, epoch as 400 and number of hidden layers of SDL as 3 for better learning. The output layer corresponds to patent topic and the number of input neurons depends on the significant terms selected for training. The proposed system provides a good accuracy with correct prediction.

Table 6.1 shows the sample patent document with abstract and claim under the topic "device". From each topic, 25 patent documents are extracted for training and 5 documents for testing. Totally 100 documents are used for training the model and 20 for validation. The top ten key terms selected by the model, the training patents and correlation matrix generated are depicted in the Table 6.2.

Table 6.1 Sample Patent Extracted Under the Topic "Device"

Device : Interface and fabrication method for lighting and other electrical devices
Abstract : Interfaces for electrical (e.g., lighting) devices involve use of electrically conductive edge contacts arranged on or protruding from edges of printed circuit boards (PCBs) that provide or facilitate electrical connections to first and second externally accessible electrical contacts, such as may include threaded and foot contacts of a lighting device including a screw-shaped male base. First and/or second edge contacts of a PCB may protrude through first and second openings in a housing to form first and second externally accessible contact, or directly engage first and second externally accessible contact elements associated with (e.g., retained by) the housing. A contact element retained by a housing may define a slot in the interior of the housing to directly engage an edge contact of the PCB. Electric power is supplied to the PCB via edge contacts without need for intervening wires or soldered connections. What is claimed is: 1. A light bulb comprising: a housing defining at least one opening; at least one light emitting element associated with the housing; and a printed circuit board configured for electrical connection to the at least one light emitting element, the printed circuit board including a first face, a second face, at least one edge, a first electrically conductive edge contact arranged on an edge of the at least one edge or protruding from the printed circuit board, and a second electrically conductive edge contact arranged on an edge of the at least one edge or protruding from the printed circuit board; wherein the housing comprises a first portion proximate to the at least one light emitting element, a second portion proximate to the first and second electrically conductive edge contacts, and a sidewall extending between the first portion and the second portion; wherein an opening of the at least one opening defined in the housing extends through the sidewall; and wherein at least a portion of the printed circuit board is arranged within the housing, the first electrically conductive edge contact extends through the opening defined in the sidewall and protrudes laterally beyond the sidewall to form a first externally accessible electrically conductive contact, and the second electrically conductive edge contact extends through the at least one opening to form a second externally accessible electrically conductive contact.

Table 6.2 Selected Key terms, Testing Patent Documents, Correlation Matrix of the Proposed Model

Selected Key terms : fit, plate, engage, circuit, sides, mounting, adjacent, member, rear, front

Topics : Conductors, Connectors, Devices, Outlets

Testing Documents	Correlation Matrix - Training	Correlation Matrix - Testing
Testing Documents t_doc testing\conductors _45877.html t_doc testing\conductors _47585.html t_doc testing\conductors _47754.html t_doc testing\conductors _47778.html t_doc testing\conductors _54878.html t_doc testing\connectors_58421.html t_doc testing\connectors_58448.html t_doc testing\connectors_58741.html t_doc testing\connectors_58854.html t_doc testing\connectors_67567.html t_doc testing\devices_19.html t_doc testing\devices_2.html t_doc testing\devices_20.html t_doc testing\devices_21.html t_doc testing\devices_22.html t_doc testing\outlets_19.html t_doc testing\outlets_2.html t_doc testing\outlets_20.html t_doc testing\outlets_21.html t_doc testing\outlets_22.html	1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 0 0 0 0 1 1 1 0 0 1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 1 1 1 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 0 0 1 1 1 1 0 0 1 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0	0 6 0 0 0 20 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 2 0 2 1 1 0 0 10 0 0 0 22 4 2 0 25 0 0 0 16 0 41 0 0 0 0 2 0 0 4 1 0 0 0 6 0 2 0 0 0 0 8 0 0 0 0 0 0 8 8 0 0 0 1 0 2 0 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 14 25 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 26 0 2 0 0 0 0 0 0 0 0 0 12 4 6 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 8 1 0 0 0 0 0 0 9 0 0 0 0 0

The document vectors of training matrix are supplied as an input to SDL in the visible layer. The targets value are set at the output layer. The output of the trained SDL is validated with test patents. The table 6.3 shows the sample test results obtained by the system. In table 6.3, the first row represents the topic name with index values, the second row shows the test results of SDL and the third row shows target value and retrieved value obtained. The highest score value of SDL is highlighted in the table 6.3. The investigation is made by extracting 25 training patents and 5 testing patents from each topic. Hence, the proposed SDL is trained with 100 training patents and 20 testing patents with accuracy of 90 %.

Table 6.3 Sample Test Results and Evaluated Accuracy

The Output Values generated by SDL for four topics : [conductors-0, connectors-1,devices-2, outlets-3]

Test Results

0.6801270780182215	0.6021192812488169	0.4923371518235496	0.6302152748993258
0.5288571151837651	0.6885287598003434	0.46712678729751855	0.5224810866608722
3.4237146657699826	0.33451200111618956	0.2955934500389267	0.2584073861471042
5.027042520385333	0.5259642266932262	0.5267498314243713	0.57837976457204
5.485716366992804	0.6762808775142823	0.5964215622440646	0.6513424633544365
0.7278378254438732	5.5551749442474785	0.7234616343475166	0.6502415654795011
0.3633663783130884	4.3009620301909015	0.39271785126635933	0.3069654772717713
0.553510872965372	4.950539583496848	0.6340252566335857	0.6332146229313704
0.33856125053381686	0.4743296415680428	0.39771279328524284	0.30666968062674227
0.21228188510209853	3.692893143612802	0.17073007655758995	0.22572147962663375
0.14304468160846404	0.3313946830293363	0.3458727941173493	0.23474271210572595
0.24764885041787768	0.21159214623830724	3.806253109762009	0.12539049439691297
0.1913507666151021	0.32470458034071503	0.3189536433659855	0.22498509025823307
0.11013505243552454	0.3388315819986284	4.080103664935729	0.17523303723567454
0.11114077359773189	0.19516237064499534	3.889689521999923	0.19771752328103104
0.15996127090628418	0.18702327152325124	0.19094103836981352	0.21877362702405645
0.11451962502611317	0.20861412022186449	0.3237672150201221	0.3664007691722125
0.4981460119006549	0.36686433412535924	0.5161778800378503	4.331779345328009
0.15156231805237771	0.1651960725335811	0.21187202951835876	0.22538565306204328
0.1848949754520957	0.2571055813585946	0.25085574585813025	0.27782257485390952

DName	Target	Retrieved	
conductors and insulators_45877.html	0	0	The document in conductors
conductors and insulators_47585.html	0	1	The document in connectors
conductors and insulators_47754.html	0	0	The document in conductors
conductors and insulators_47778.html	0	0	The document in conductors
conductors and insulators_54878.html	0	0	The document in conductors
connectors_58421.html	1	1	The document in connectors
connectors_58448.html	1	1	The document in connectors
connectors_58741.html	1	1	The document in connectors
connectors_58854.html	1	1	The document in connectors
connectors_67567.html	1	1	The document in connectors
devices_19.html	2	2	The document in devices
devices_2.html	2	2	The document in devices
devices_20.html	2	1	The document in connectors
devices_21.html	2	2	The document in devices
devices_22.html	2	2	The document in devices
outlets_19.html	3	3	The document in outlets
outlets_2.html	3	3	The document in outlets
outlets_20.html	3	3	The document in outlets
outlets_21.html	3	3	The document in outlets
outlets_22.html	3	3	The document in outlets

Accuracy: 0.9

BUILD SUCCESSFUL (total time: 12 minutes 25 seconds)

Figure 6.2 shows the performance of SDL using document-term matrix with correlated features DTM_{corr} and SDL using document-term matrix with cluster centroids DTM_{cent}. Investigations are made with different units of patent document such as abstract, claim and both. From the figure 6.3, it is evidenced that the classification made with respect to claim information of the patent shows good accuracy. The accuracy values are estimated for two different SDL flavors with three units of patents. The accuracy of SDL with DTM_{cent} projects a reasonable improvement over SDL with DTM_{corr} . Hence, the accuracy with respect to the centroid based features is found to be good.

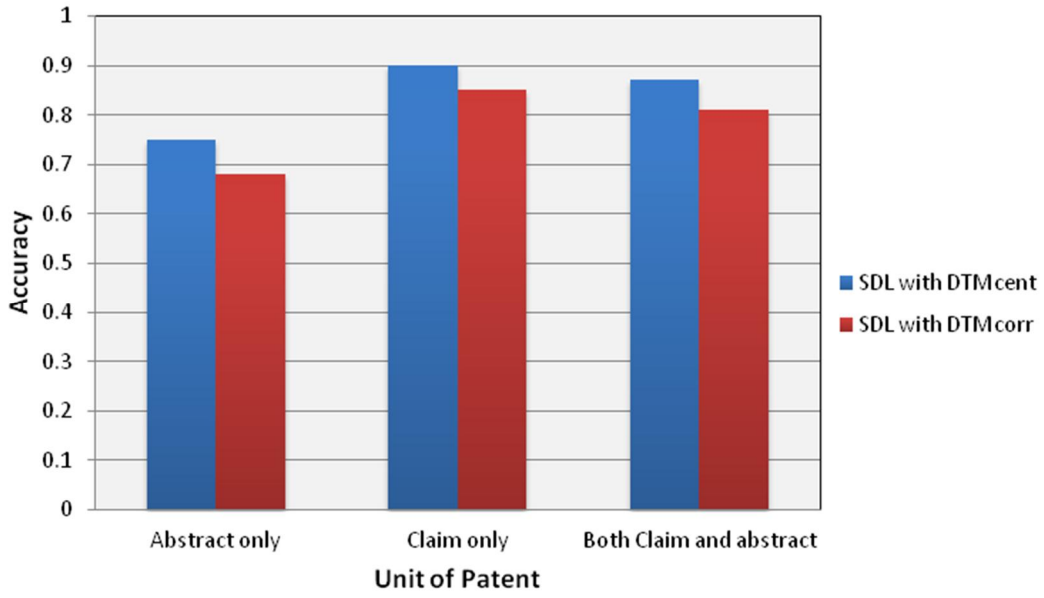

Figure 6.2 Classification Accuracy of SDL for the Predicted Topics

The accuracy of each patent topic is examined through the Neural Network based classifier and the proposed SDL classifier with two set of features. The table 6.5 shows the performance comparison of SDL and Neural Network classifier with respect to correlated feature matrix and centroid feature matrix. The investigation made with top ten key terms based on the claim information of the patents extracted from 4 domain topics.

Table 6.4 Classification Accuracy of Various Learning Techniques for Patent Classification Based on Claims

	Classification Accuracy			
	Conductors	Connectors	Devices	Outlets
NN based classifier	0.6	0.7	0. 5	0.6
SDL with DTM_{cent}	0.8	1.0	0.8	1.0
SDL with DTM_{corr}	0.6	0.8	0.8	0.8

From Table 6.4, it is clearly evidenced that 100 % accuracy is achieved under the topics "Connector" and "Outlets".

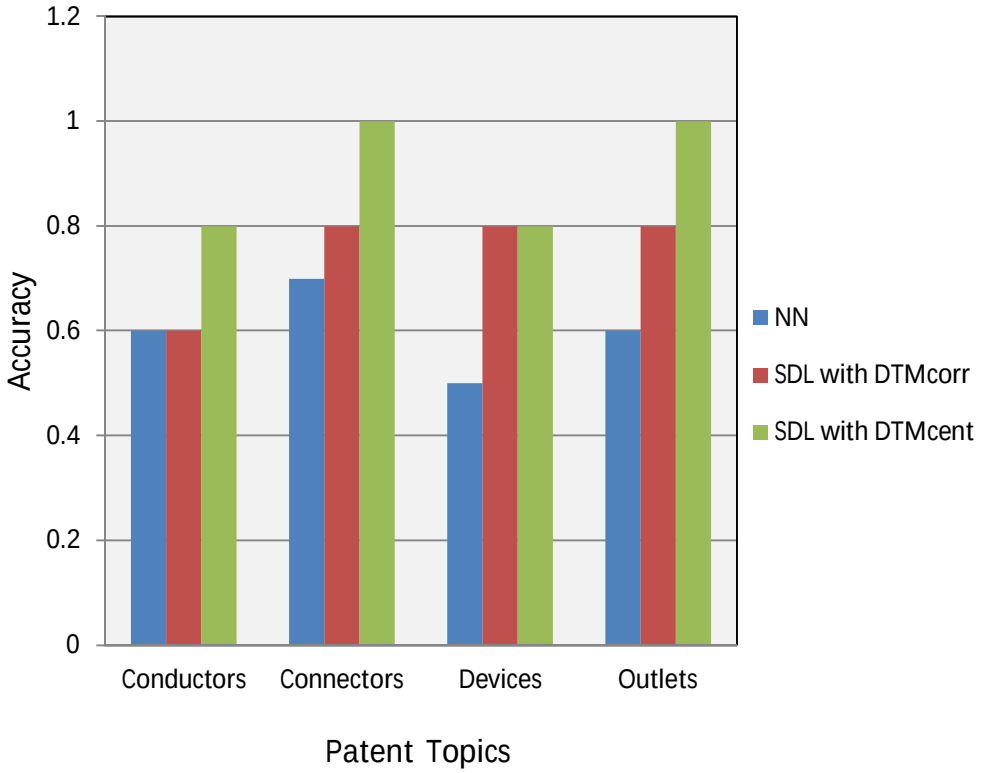

Figure 6.3 Accuracy comparison of SDL with Neural Network based Classifier on different Topics of Patent Dataset

6.5 CONCLUSION

The approach proposed in this work suggests that efficient solution for patent document classification using SDL. There are certain shortcomings in applying SDL because insufficient training data may lead to the unreliable model and at the same time the training procedure involves more computing resources. A well-trained semantic model will help the companies to organize and process the documents in better way. The ability of deep learning approach to discover the hidden structures and features at different levels of abstraction is useful for efficiently classifying patents. The accuracy of the trained model is found to be better than other classifier especially when Electrical patents are considered. This work can be extended to support more informative semantic ontology and smoothing techniques in order to make better prediction with improved flexibility and accuracy. Besides the classification of patent documents, this work presents a framework which can be used to extract more meaningful data representation for analysis of other type of un-structured text documents.