SEMANTIC BASED TEXT CLASSIFICATION OF EMERGENT BRAIN TUMOUR DISEASE REPORTS USING MACHINE LEARNING TECHNIQUES

[1]M.Vengateshwaran, [2]M.A.Mohamed Aslam, [3]K.Ajithkumar, [4]S.Vishnu Varthanan

[1]Assistant Professor in CSE, [2,3,4]UG Scholars
Department of Computer Science and Engineering
Arasu Engineering College, Kumbakonam

Abstract: With the rapid growth of stored information has been enormously increasing day by day which is generally in the unstructured form and cannot be used for any processing to extract useful information. Text classification is a technique to find meaningful patterns from the available text documents. Medical domains are rich knowledge source needed to be organized efficiently and conveniently. Disease reports documents are used for gathering business intelligence and identifying key trends in technology development. The main focus of this paper is to propose a Disease reports document classification framework based on Semantic Deep Learner (SDL). In this framework, initially key terms of the Disease reports documents are extracted and represented using Vector Space Model(VSM), the importance of the key terms are weighted based up on their frequencies using TF-IDF. The semantic similarity between the key features are computed using cosine measure. Terms with higher correlations are synthesized into a smaller set of features. Finally the Semantic Deep Learner is trained using the correlated features and accordingly Disease reports are classified. The target output identifies the category of a medical brain disease document based on a hierarchical classification scheme of the International Patent Classification (IPC) standard. Our approach is new to the medical domain and shows some improvement in the classification accuracy when compared to the other state of art classifier.

Keywords: Text classification, Medical Domain, Disease reports, SDL, VSM, TF-IDF

1. INTRODUCTION

In recent years, we can find many applications related for text classification using text mining. Text mining is the automatically find out the text from a large set of document.. It is a one of the research field. Each class or category represents a particular topic, e.g., sports, sciences, politics or Research. There are many real-world problems require more refined classification based on some semantic perspectives. For example, collection of documents about a disease, some documents may report outbreaks of the disease, some may describe how to cure the disease, some may discuss how to prevent the disease, etc. To classify text at this semantic level, the traditional bag-of-words model is no longer sufficient.

In this paper, we propose to integrate the bag-of-words scheme and semantic features extracted from texts for classification to various kinds of documents. As a case study, we investigate the disease reporting domain. We want to classify sentences that report disease outbreaks, and sentences that do not. For example, the following sentence reports a possible disease outbreak "*the district hosipital reported today that 10 people were diagnosed with brain pathological this morning*". However, the following sentence does not report an outbreak, "*the district hospital reported today that they have successfully tested a new brain pathological treatment procedure*". Both sentences are on the topic of *brain pathological*. However, they are entirely different semantically. The problem is how to separate sentences based on the required semantic categories, i.e., reporting a possible outbreak or not in this case. This classification task is an important application in its own right. It shows that both the words used in sentences and the sentence semantic characteristics are important.

Our novel approach explores the document disease report classification framework using SDL. The automatic document classification methodology is described in the following steps. Initially, the important terms are extracted from Disease report documents and represented as feature vectors using VSM and features vectors are weighted using TF-IDF based on the frequency of terms in a Disease report document. Next, Cosine measure is applied to find the similarities between the features and depicted in a correlation matrix in order to synthesize features into a smaller set representing key features within the medical domain. Finally SDL is trained using the consolidated set of key features available in a correlation matrix. The output of SDL gives the category of the corresponding disease report document. As deep learner is better than neural networks and other learners, the accuracy of the classification will be improved to some extent especially for disease report document. Applying deep learning technique for classification of the documents is a new approach in medical domain. Our experimental results confirm that this integrated approach produces much more accurate classifiers than each of them alone.

II. SEMANTIC TEXT CLASSIFICATION

The semantic text classification is the same as traditional topic-based text classification. We have a collection of documents D and each document $d_i \in D$ is labeled with a class $c_j \in C$, where C is a set of known classes. A supervised learning algorithm is applied to build a classification model. However, semantic text classification usually has more refined categories or classes. Different classes are hard to be separated based on bag-of-words or n-grams alone. Semantic information is required. For example, the sentence, "*the district hospital reported that 10 people were diagnosed with brain pathological early today*", reports a possible *brain pathological* outbreak. It is easy to observe that the words "reported" and "diagnosed" are indicative of an outbreak in multiple document in various kind of medical domain disease reports.

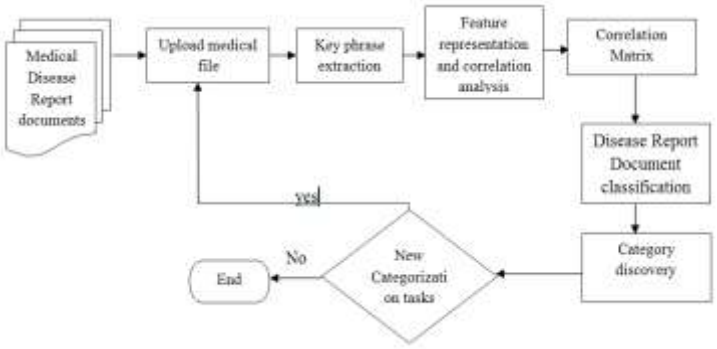

Fig.1. Illustrates the difference between traditional text classification and semantic text classification as described in this work.

III. IMPORTANT SEMANTIC FEATURES

Semantic Features

Our aim is to classify sentences that report possible disease outbreaks and those that do not, which is a classification problem. We will use a Semantic Deep Learner approach, e.g., naïve Bayesian (NB) or Vector Space Model(VSM). Thus, we only need to design and construct a features. As we mentioned above, we use both keywords and semantic features. Keyword features are obtained in the same way as in traditional text classification. Here, we only focus on semantic features.

3.1 Vector Space Model (VSM)

It is proposed by Salton (Salton, 1968) is a general technique for document representation in classification. In this model each document is represented as a vector of features. Each and every feature is associated with a weight. Usually these features are simple words. The feature weight can be simply a Boolean indicating the presence or absence of the word in document, its occurrence number in document or it can be calculated by a formula like the well known tf*idf method. VSM has been widely used in traditional information retrieval and for automatic document categorization. There are three key steps where terms are first extracted from the document text, then the weights of the indexed terms are derived to improve the document retrieval accuracy, and then the documents are ranked with respect to a similarity measure.VSM is a multi-dimensional vector where each feature of a document is a dimension. For instance, term frequency (TF) and inverted document frequency (IDF) are two features of a text document. After the vector of a text document is derived, a cosine function is applied to measure the similarity between two documents.

$$\text{Cos}(X,Y)=\frac{\sum_{i=1}^{n} xiyi}{\sqrt{\sum_{i=1}^{n} xi2 * \sum_{i=1}^{n} yi2}}$$

Where $x=\{x_1,x_2,\ldots..x_n\}$,$x_i$ represents i_{th} feature of document X.
$Y=\{y_1,y_2,\ldots.y_n\}$,y_i with similarity between X and Y calculated by using cosine function.

3.2 Naive Bayes Approach

Naive Bayes classification should does not require more observations for all possible combinations of the variables.other than, each and every variables are assumed to be independent to each other. In Another words, naive Bayes classifiers assume that the influence of a variable is independent of other variables for a given class, an assumption called class conditional independence. This algorithm uses the joint probability of document features to calculate the probability that a new document belongs to a specific class.

$$P(Ci|d')=\frac{P(d'|Ci).P(Ci)}{\Sigma_{Cj\in C} P(d'|Ci).P(Cj)}$$

P(ci) is the probability of a given document belonging to class ci and P(d'|ci) is the conditional probability of document d'belonging to a specific class ci. The probability of document d' belonging to class ci can be derived using the following equation:

$$P(d'|Ci)=\prod_{j=1}^{|d'|} P(Wj|Ci)$$

Naive Bayes are use to handle labeled and unlabeled documents. It compares two Naive Bayes models are Multivariate Bernoulli model and Multinomial model to provide a different vocabulary sizes.

IV. SYSTEM METHODOLOGY

This section depicts the detailed methodologies for document classification. First, a document content extraction model is built to represent the medical domain disease report document content with a vector consisting of key phrase frequencies. Next a document

classification model SDL is developed based on the training set to classify the disease report document. Finally, SDL is validated by testing disease report documents.

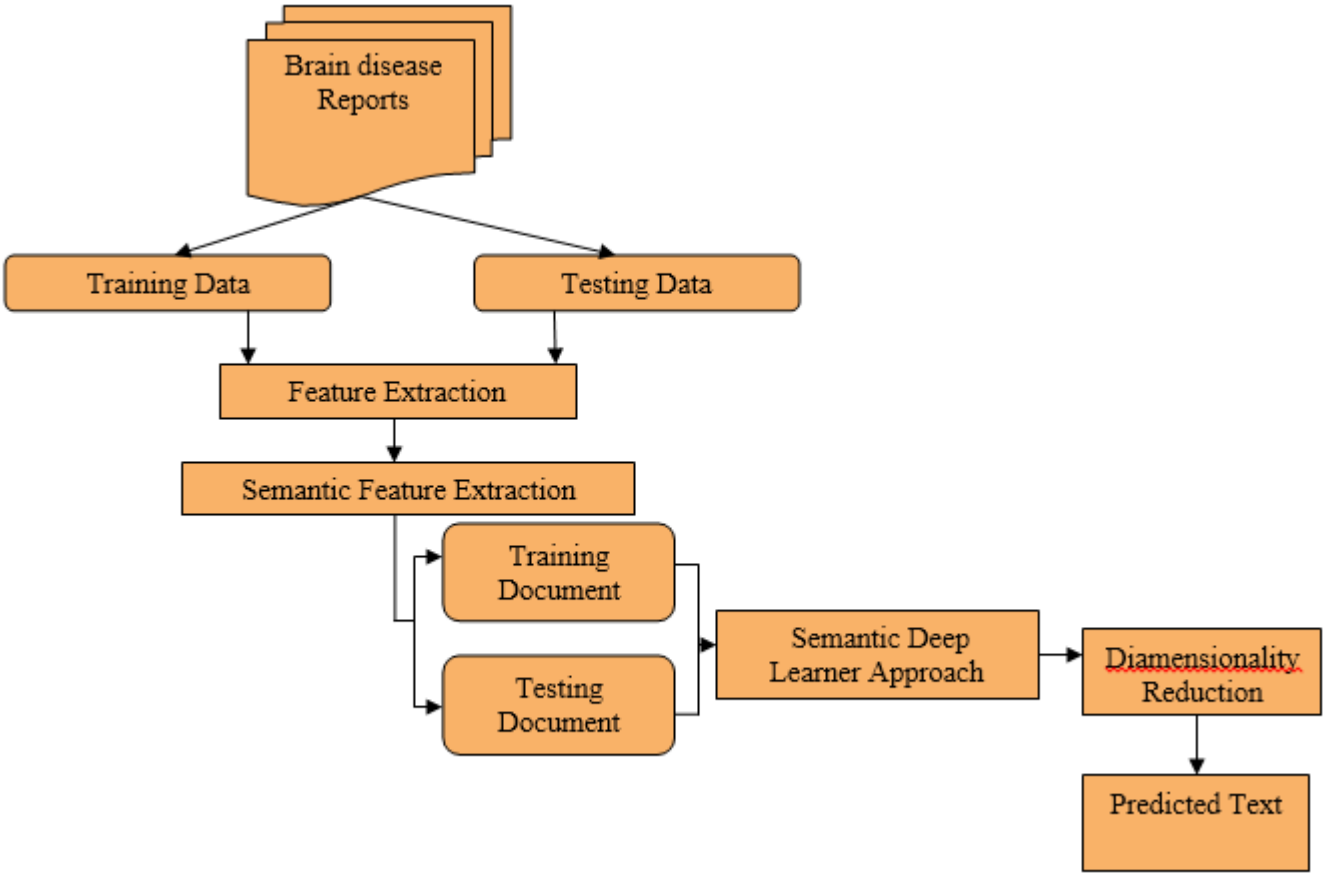

Fig.2 System Architecture

4.1 Key Features Extraction and Representation

The input disease report documents are preprocessed by using a stemming and stop words removal. In this preprocessing step,the phrases are tokenised to get a cleaned or processed data. The resulting key phrases are further processed as follows : Initially, the key phrases of the document are split into terms and these terms are then used to construct a key features base. The correlations among the terms are identified and two or more highly co-related terms are mapped to small set of significant terms. After the document features extraction, the document is represented by a vector containing the frequencies of all key features by using of a VSM.

The key features are weighted and ranked using TF-IDF. A domain term that appears in a document with high frequency indicates that the term is a significant keyword or key features. After extracting all high frequency terms, a correlation matrix of terms is created by calculating their frequency of occurrence within same documents. The correlation of two key features(KT_i, KT_j)appeared in set of patent documents are determined using equation.

$$C_{ij}=\frac{\sum_{i=1}^{N_D} X_{i,I}X_{j,I} - N_D\ \overline{X_i}\,\overline{X_j}}{\sqrt{\sum_{i=1}^{N_D} X^2{}_{i,I} - N_D\overline{X^2}_i}\cdot\sqrt{\sum_{i=1}^{N_D} X^2{}_{j,I} - N_D\overline{X^2}_j}}$$

Where C_{ij} is the correlation of KT_i *and* KT_j that appear in a set of patent documents; $X_{i,I}$ is the frequency that KT_i appeared in document D_i, $X_{j,I}$ is the frequency that KT_j appeared in document D_j. $\overline{X_i}$ is the average frequency that KT_i appeared in all documents(all D_s); $\overline{X_j}$ is the average frequency that KT_j appeared in all documents(all D_s); N_D is the total number of documents.

The highly correlated key features are selected and stored as the related features list. Finally, the key features list is completed when highly correlated features are merged, a necessary step since it is easier to train SDL models using fewer variables. When a new document is uploaded into a patent knowledge management system, the key features and their frequencies are extracted from the document. The frequencies of all terms are derived. Then KTf_i is used to present the frequency of key terms KT_i in the document, and CPf_{ij} to represent the frequency of related-term CT_{ij} .The correlation of CT_{ij} and KT_i are listed as C_{ij} and the final frequency of KT_i is

$$KTF_i = KTf_i + \sum_{j=1}^{k} CTf_{ij} \,.\, C_{ij}$$

After calculating the KTF of all key features, a vector of key features frequencies is listed as

$$[KTF_1, KTF_2, KTF_3, \ldots\ldots\ldots KTF_n]$$

This vector serves as the input to SDL.

4.2 Document Categorization Using SDL

In this section, we describe the document categorization methodology based on the Semantic Deep Learner (SDL). It is a multi-layered network model also called as a supervised learning model that can be used to solve non-linear problems. An advantage of SDL is that it does not need to change the network structure and achieve the target output. Another advantage of using SDL is its rapid execution when a trained network is applied. The learning stage of SDL involves a forwardpass and a backward pass. The structure of the simple learning model is depicted in figure 3. The two passes of SDL are described in the following section.

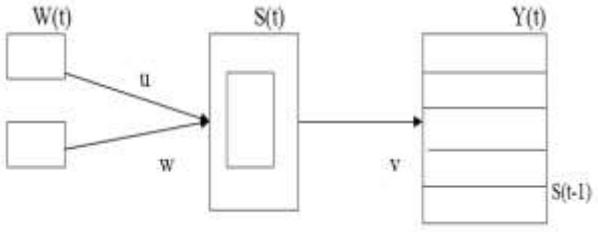

Fig.3 Structure of the simple learning model

Where W(t)-Input Layer, S(t)-Hidden Layer, Y(t)-Output Layer and u,v,w -weight matrices.
The hidden layer provide recurrent connection to S(t-1) and thus provide short term memory that models context of a word. The generated feature frequency matrix is given as input to SDL. The neocortex, which is associated with many cognitive abilities, has a complex multilayer hierarchy. The development of intelligence follows with the multi-layer structure. From an evolutionary viewpoint, the phylogenetically most recent part of the brain is the neocortex. In humans and other primates, starting from catarhinians, the multi-layers structure began to appear in the neocortex. It thus provides a possible way to achieve the ultimate target of natural language processing, which is to enable the computer to understand the human (natural) languages. The structure of SDL is depicted in Fig. 3
Here, the correlation matrix formed using the whole document taken for training is given as input to the semantic deep learner. The deep learner will train the system based on the target given. The training is done by the hidden layer of the deep learner that exploits the target and the input. The target would be based on the documents taken for training belongs to the topic, because we know the topic (domain) for the documents taken for training. After training the system based on the semantic deep learner, the testing is done by giving the testing document. When the testing document is given as input to the system, the correlation matrix is formed for the input document using the keywords that formed the correlation in the training process. The semantic deep learner will give a score for the given input document and based on the score the document will be classified to which category it belongs.

V. PERFORMANCE EVALUATION

There are three fundamental measures for assessing the quality of information retrieval:

- Precision
- Recall
- F-score

1. Precision

Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as:

Precision= |{Relevant} ∩ {Retrieved}| / |{Retrieved}|

2. Recall

Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as:

Recall = |{Relevant} ∩ {Retrieved}| / |{Relevant}|

3. F-score

F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa. F-score is defined as harmonic mean of recall or precision as follows:

F-score = recall x precision / (recall + precision) / 2

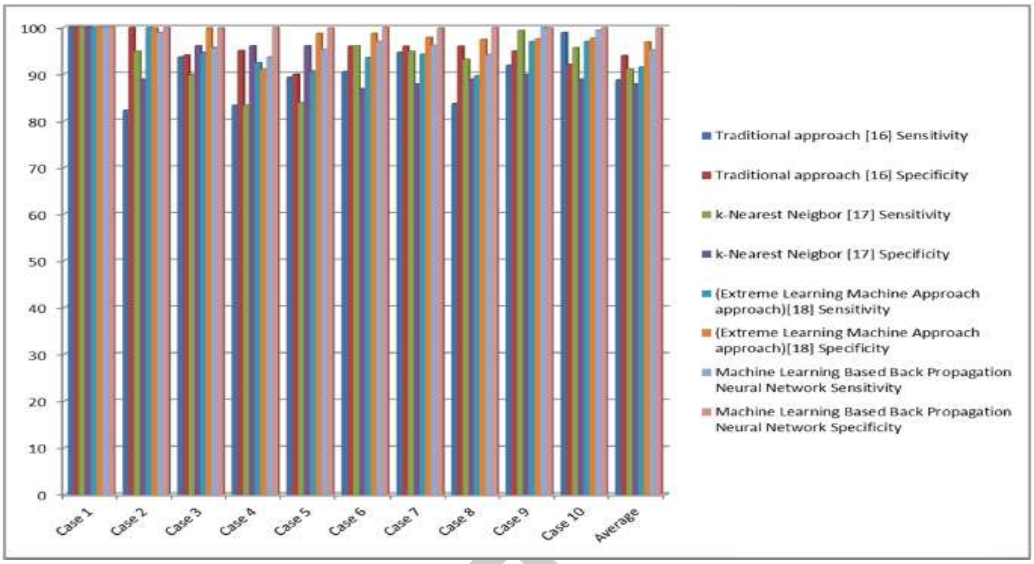

Fig.4 Representation of tumor prediction methods with modular machine learning based neural networks.

VI.CONCLUSION

In this paper, we proposed a topic classification algorithm based on semantic deep learner (SDL) using semantic smoothing model. Initially, different documents from different domains are taken and it is separated for training and testing. The training documents are applied for four different phases to extract the keywords to form frequency matrix. The top keywords from each phase are used to form the frequency matrix with respect to the documents taken for training. Finally the Semantic Deep Learner is trained using the correlated features and accordingly Disease reports are classified. The target output identifies the category of a medical brain disease document based on a hierarchical classification scheme of the International Patent Classification (IPC) standard. Our approach is new to the medical domain and shows some improvement in the classification accuracy when compared to the other state of art classifier. More experiments on SDL using different semantic extraction models will be conducted and compared with other state-of-art models in future.

REFERENCES

[1] Bashar Tahayna, Ramesh Kumar Ayyasamy, Saadat Alhashmi, "A Novel Weighting Scheme for Efficient Document Indexing and Classification",Information Technology (ITSim), Vol. 2, PP. 783-788, 2010.
[2] Nikos Tsimboukakis and George Tambouratzis, "Word-Map Systems for Content-Based Document Classification", IEEE Transactions on Systems, man, and Cybernetics, Vol. 41, No. 5, PP. 662-673, 2011.
[3] Jemma Wu, "A Framework for Learning Comprehensible Theories in XML Document Classification", IEEE Transactions On Knowledge and Data Engineering, Vol. 24, No. 1, PP. 1-14, 2012.
[4] Mohammad Khabbaz, Keivan Kianmehr, and Reda Alhajj, "Employing Structural and Textual Feature Extraction for Semistructured Document Classification", IEEE Transactions on Systems, man, and Cybernetics, Vol. 42, No. 6, PP. 1566-1578, 2012.
[5] G. W. Jiji, L. Ganesan, and S. S. Ganesh, ``Unsupervised texture classification," J. Theor. Appl. Inf. Technol., vol. 5, no. 4, pp. 373_381, Apr. 2009.
[6] K. V. Ramana and B. R. Korrapati, ``A neural network based classification and diagnosis of brain hemorrhages," Int. J. Artificial Intelligence. Expert System., vol. 1, no. 2, 2009.
[7] P. M. Shakeel, S. Baskar, V. R. S. Dhulipala, S. Mishra, and M. M. Jaber, ``Maintaining security and privacy in health care system using learning based deep-Q-networks," J. Med. Syst., vol. 42, p. 186, Oct. 2018.
[8] S. Lu, Z. Lu, J. Yang, M. Yang, and S. Wang, ``A pathological brain detection system based on kernel based ELM," Multimedia Tools Appl., vol. 77, no. 3, pp. 3715_3728, 2018.

AUTHOR'S PROFILE

Name : Mr.M.Vengateshwaran **B.E., M.E.,**

Designation : Assistant Professor , Department of CSE
Arasu Engineering College, Kumbakonam

Specialization: BigData, Data mining, IR, Machine Learning etc.,

Name : M.A.Mohamed Aslam **(B.E.-CSE)**

College : Arasu Engineering College, Kumbakonam

Specialization : Machine Learning

Name : K.Ajithkumar **(B.E.-CSE)**

College : Arasu Engineering College, Kumbakonam

Specialization: Machine Learning

Name : S. Vishnu Varthanan **(B.E.-CSE)**

College : Arasu Engineering College, Kumbakonam

Specialization: Machine Learning