Text Document Classification by using WordNet Ontology and Neural Network

Manisha Gawade
Computer Department of Engineering, Sinhgad Academy of Engineering Kondhwa (BK)
Pune – 411043

Tejashree Mane
Computer Department of Engineering, Sinhgad Academy of Engineering Kondhwa (BK)
Pune – 411043

Dhanashree Ghone
Computer Department of Engineering, Sinhgad Academy of Engineering Kondhwa (BK)
Pune - 411043

Prasad Khade
Computer Department of Engineering, Sinhgad Academy of Engineering Kondhwa (BK)
Pune – 411043

Nihar Ranjan, PhD
Computer Department of Engineering, Sinhgad Academy of Engineering Kondhwa (BK)
Pune 411043

ABSTRACT
Every day the mass of information available, merely finding the relevant information is not the only task of automatic text classification systems. The main problem is to classify which documents are relevant and which are irrelevant. The Automated text classification consists of automatically organizing clustered data. We propose a method of automatic text classification using Convolutional Neural Network based on the disambiguation of the meaning of the word we use the WordNet ontology and word embedding algorithm to eliminate the ambiguity of words so that each word is replaced by its meaning in suitable context. The closest ancestors of the senses of all the words in a given document are selected as folders for the specified document.

Keywords
Neural network, classification, wordsense, feature selection, model selection, WordNet

1. INTRODUCTION
Every day the mass of information available to us increases. This information would be irrelevant if our ability to productively get to did not increment too. For most extreme advantage, there is need of devices that permit look, sort, list, store and investigate the accessible information. One of the promising region is the automatic text categorization. With the quick increase of information and knowledge, automatically classifying text documents is becoming a hotspot of knowledge management. A critical capability of knowledge management systems is to classify the text documents into different categories, which are meaningful to users. Feature selection is an important step in any method of text classification. in this work, we have used an efficient method of feature selection. It uses a threshold to include only those words in the given set of text documents that are useful for the purpose of classification. At present, many scholars at home and abroad have studied the text classification technology by using main methods, including the traditional machine learning and the deep learning which is popular currently. The deep learning (dl) is a new field of the machine learning research, which aims to establish a neural network to simulate human brain for analysis and learning. As a superior model of the deep learning technology, the convolutional neural network (cnn) has become one of the research focuses in many fields such as image recognition speech analysis and natural language processing [3]. Experiments have showed that the text document classification algorithm proposed in this paper can greatly improve the classification accuracy compared with the traditional methods [6]. However, word ambiguity is a severe problem in the keywords-based methods. For example, if 'bat' occurs several times in a document, should the file be classified to "sport" or "mammal"? A number of computer engineers tried to retrieve articles about "board", but a large number of web pages about "board game" or "message board" were retrieved. Each word may have multiple senses (meanings), and multiple words may have the same sense. it is not trivial for a computer to know which sense the keyword is using in a given context. Extensive research has been done in word sense disambiguation. However, to the best of our knowledge, disambiguation research is focused in retrieval or in query, not for text classification [1][2]. In this paper, we propose a text classification method based on sense disambiguation. In order to define an appropriate mid-level category for each sense is implemented on wordnet [7].

Fig 1: Text document classification

2. MOTIVATION
In today's world of internet the data is the main concern which performs most important part in different operation. In this most of data is available in text/pictorial/numeric format which use to perform the operations. In this process document having big shares which may be relevant or irrelevant. The main problem with text classification is word sense. One word having multiple meaning available. With the current size of documents, it has become hard to try and manually index and categorize all of its content. Evidently, there is need for automatic text document classification [8][9].

3. RELATED WORK

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers.

In this section, we briefly review the related work on Text classification and their different techniques.
Wang et al.have worked on a convolution NN for achieving text classification. The feature selection was modified through integrating similar words into groups and thereby, the NN was assisted in learning to achieve better classification performance, in terms of accuracy. The various limitations were the supervised feature down-sampling, the task-specific embedding learning and the embedding affinity measurement in the vector spaces [1].

J.-T. Chien, describe the "Hierarchical theme and topic modeling," in that Taking into account hierarchical data sets in the body of text, such as words, phrases and documents, we perform structural learning and we deduce latent themes and themes for sentences and words from a collection of documents, respectively. The relationship between arguments and arguments in different data groupings is explored through an unsupervised procedure without limiting the number of clusters. A tree branching process is presented to draw the proportions of the topic for different phrases. They build a hierarchical theme and a thematic model, which flexibly represents heterogeneous documents using non-parametric Bayesian parameters. The thematic phrases and the thematic words are extracted. In the experiments, the proposed method is evaluated as effective for the construction of a semantic tree structure for the corresponding sentences and words. The superiority of the use of the tree model for the selection of expressive phrases for the summary of documents is illustrated [2].

Bernardini, C. Carpineto, and M. D'Amico, describe the "Full-subtopic retrieval with keyphrase-based search results clustering," in that Consider the problem of restoring multiple documents that are relevant to the individual sub-topics of a given Web query, called "full child retrieval". To solve this problem, they present a new algorithm for grouping search results that generates clusters labeled with key phrases. The key phrases are extracted generalized suffix tree created by the search results and merge through a hierarchical agglomeration procedure improved grouping. They also introduce a new measure to evaluate the performance of full recovery sub-themes, namely "look for secondary arguments length under the sufficiency of k documents". they have used a test collection specifically designed to evaluate the recovery of the sub-themes, they have found that our algorithm has passed both other clustering algorithms of existing research results as a method of redirecting search results underline the diversity of results (at least for k> 1), that is when they are interested in recovering more than one relevant document by sub-theme) [3].

Lu Pan, Haibo Tang and Lei Zhou, Liuyang Wang, Quanyin Zhu describes the "An Identification Method of News Scientific Intelligence Based on TF-IDF" in that they propose an identification method With the development of Internet, the amount of Information has been rapidly growing which is spread widely. In order to improve the value and accuracy of science information that is pushed in this paper, an intelligence dichotomous method for science information categorization to identify science information from massive Web news is presents. During the experiment, 85.3% recognition rate of the recognition non-tech news are realized and 82.9% accuracy rate, the results show that the method can effectively identify Web science information news and reduce the amount of independent news[4].

Ning Li, Hui Zhang, Yong Chen describes the "Convolutional Neural Network with SDP-based Attention for Relation Classification" in that Relation classification plays an important role in the field of natural language processing (NLP). The state-of-theart methods for this task use prior knowledge as features such as WordNet, Part-of-Speech(POS), shortest dependency path (SDP), which is helpful but brings error propagation. In this paper, we propose a convolutional neural network architecture, which builds word-level attention mechanism based on SDP to capture task-oriented patterns in sentences. We explore the way of combining prior knowledge and deep models properly to ease errors in prior knowledge. Additionally, a new objective function is designed to reduce the impact of artificial class which is seldom touched in previous works [5].

S. Dumais, J. Platt, D. Heckerman, and M. Sahami, describe the "Inductive learning algorithms and representations for text categorization," in that Text categorization the assignment of natural language texts to one or more predefined categories based on their content is an important component in many information organization and management tasks. They compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real-time classification speed and classification accuracy. They also examine training set size, and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. Linear Support Vector Machines (SVM) are particularly promising because they are very accurate, quick to train and quick to evaluate [6].

Ying Liu1, Peter Scheuermann2, Xingsen Li1, and Xingquan Zhu, describe the "Using WordNet to Disambiguate Word Senses for Text Classification," in that they propose an automatic method of text classification. Based on the disambiguation of the meaning of words. We use the "bell" algorithm to eliminate the word ambiguity so that every word is replaced by its meaning in context. The closest ancestors of the senses of all words without stopping in a given document Selected as classes for the specified document [7].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero,and A. Saarela, describe the "Self-organization of a massive document collection," This paper describes the implementation of a system that is able to organize vast document collections according to textual similarities. It is based on the self-organizing map (SOM) algorithm. As the feature vectors for the documents statistical representations of their vocabularies are used. The main goal in our work has been to scale up the SOM algorithm to be able to deal with large amounts of high-dimensional data. In a practical experiment we mapped 6 840 568 patent abstracts onto a 1 002 240-node SOM. As the feature vectors we used 500-dimensional vectors of stochastic figures obtained as random projections of weighted word histograms [8].

Q. Mei, X. Shen, and C. Zhai, describe the "Automatic labeling of multinomial topic models,"In this paper, they propose probabilistic approaches to automatically labeling multinomial topic models in an objective way. They cast this labeling problem as an optimization problem involving minimizing Kullback-Leibler divergence between word distributions and maximizing mutual information between a label and a topic model. Experiments with user study have been done on two text data sets with different genres. The results show that the

proposed labeling methods are quite effective to generate labels that are meaningful and useful for interpreting the discovered topic models. Our methods are general and can be applied to labeling topics learned through all kinds of topic models such as PLSA, LDA, and their variations [9].

K. Lagus and S. Kaski, describe the "Keyword selection method for characterizing text document maps," in that Characterization of subsets of data is a recurring problem in data mining. They propose a keyword selection method that can be used for obtaining characterizations of clusters of data whenever textual descriptions can be associated, with the data. Several methods that cluster data sets or form projections of data provide an order or distance measure of the clusters. If such an ordering of the clusters exists or can be deduced, the method utilizes the order to improve the characterizations. The proposed method may be applied, for example, to characterizing graphical displays of collections of data ordered e.g. with the SOM algorithm. The method is validated using a collection of 10,000 scientific abstracts from the INSPEC database organized on a WEBSOM document map [10].

4. OPEN ISSUES

Lot of work has been done in this field because of its extensive usage and applications. In this section, some of the approaches which have been implemented to achieve the same purpose are mentioned. These works are majorly differentiated by the algorithm for Text Classification.

In another research, to access the relevant information from mass of data is very difficult and time consuming task as every day mass of information increases because of digital world. Every day, the mass of information available to us increases. This information would be irrelevant if our ability to efficiently access did not increase as well. Automated text classification provides us with maximum benefit that allows us to search, sort, index, store, and analyze the available data. It also allows us to find in desired information in a reasonable time.

The internal structure of deep learning method tries to find the data, found the real relationship between variables. A large number of studies have shown that basic data representation approach to training a large impact on the success of the study, said good to eliminate the change of the input data that is unrelated to the learning task factors influence on the learning performance, keep on learning tasks at the same time useful information.

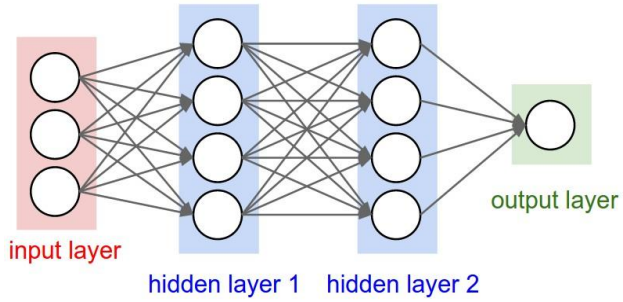

Fig 2: Typical CNN Model Structure

As my point of view when I studied the papers the issues are related to Text Classification. The challenge is to addressing automatic text classification problem convolutional neural network[5].

5. PROPOSED ARCHITECTURE

In Proposed System training is creation of train data set using which classification of unknown data in predefined categories is done. Here a learning system is created using neural network approach. It is a supervised learning where unlabeled data (test data) is classified using labeled data (training dataset). Training data is always a labeled dataset based on its features.

Project had considered no of scientific papers form different publication of different domains for creating training dataset. These papers are input for creating training dataset. This input is first preprocessed and most informative features are extracted using TF/IDF and Word embedding word sense algorithm[10].

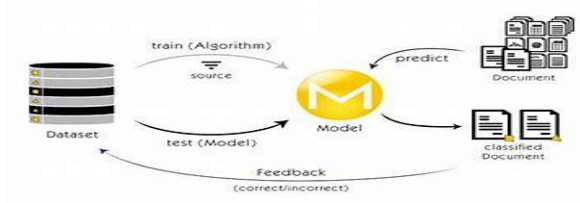

Fig 3: Proposed architecture

Step1: Remove all stop words such as 'the', 'a', 'an' etc., and also all functional words such as adverbs, preposition, conjunction etc, from the text of all the documents in the given set. **Step 2:** Necessary morphological analysis to extract the root words from the given set of texts is carried and remove all the repetition of the root words. **Step 3:** We define an elimination factor TF/IDF for each word as = Number of occurrence in its own context / Total number of occurrences in all contexts [4].Step **4:** Select the features to be all the remaining words of all the documents in the given set of text documents. **Step 5:** Create one pattern vector for each document with the features (i.e. words) selected in step 3. The numeric value for each component of such vector would be the number of occurrence for the particular word corresponding to that component in the given document. Note: Step 3 eliminates all those words that are used almost to the similar extent in all the given classes of documents. Thus these words have almost no discriminatory significance so far classification is concerned.

Convolutional neural network (CNN) is a kind of typical artificial neural network. In this kind of network, the output of each layer is used as the input of the next layer of neuron. Multi-layer convolution operation is used to transform the results of each layer by nonlinear until the output layer. In general, the convolution neural network model used in text analysis.which includes four parts: embedding layer, convolutional layer, pooling layer and fully connected layer. Compared with the traditional models for image analysis, the difference is that the input layer of the CNN model used in text analysis is the word vector.

WordNet is a manually-constructed lexical system developed by George Miller at the Cognitive Science Laboratory at Princeton University.It reflects how human beings organize their lexical memories. The basic building block of WordNet is synset consisting of all the words that express a given concept. Synsets, which senses are manually classified into, denote synonym sets. Within each synset, the senses, although from different keywords, denote the same meaning.

5.1 Feature Selection using Entropy Model

Once the features are extracted, feature selection, which is an important step in classification, is done to reduce the training time by reducing the dimensionality of search space. In text categorization, feature selection is essential which not reduce

the index sizeand improve the performance of the classifier. The feature selection approach adopted for the proposed classification method is entropy. The entropy model is based on the distribution of the documents containing the term in the categories, and it considers its entropy. The features are selected based on a criterion that determines the quality of the feature. Entropy can be defined as the measure of uncertainty of a random outcome. Let A × Y be the dimension of the feature database. The selected keywords are then organized in a class of dimension H. By matching each keyword with that in the class, a new database is created [11].

5.2 Feature Extraction using Semantic Word Processing

The key words from pre-processing are subjected to the semantic processing, which helps to identify the important words of the text documents. Features of text classification perspective are the significant words or multi-words or frequently occurring phrases indicative of the text category. The semantic processing technique uses a bag of word approach for text classification to overcome two basic problems, namely, 'polysemy' and 'synonymy'. polysemy/hyponymy resembles a word that has distinctive meaning, whereas synonymy resembles different words that have the same meaning. In semantic processing, the keywords that are selected from the preprocessing techniques are applied to the word net ontology to extract the synonyms and hyponyms of every keyword. A synonym is a word, which can be used to substitute another word without a change in the meaning of the words. Hyponymy is the lexical relationship between the meanings of the words. Based on the semantic processing, unique keywords are selected in association with the extracted synonym and hyponymy words. The semantic processing is the effective way of text classification with robustness, reliability, and effectiveness. The organizational diagram of the semantic keyword processing [12].

6. CONCLUSION

In this paper, an automatic text classification system has been proposed for classifying the bulk documents in the database based on the topic label. The proposed classification methodology has used semantic processing in feature extraction to reduce the dimensionality problem by avoiding the repetition of words as well as the occurrence of words with same meaning. The growing use of textual data need's text mining, machine learning and methodologies to organize and extract pattern and knowledge from the document. In addition, the general step of text classification algorithm using wordnet ontology and cnn which is most representative model in neural networks.

7. REFERENCES

[1] WangP,XuB,XuJ,etal. Semantic expansionusingword embedding clustering and convolutional neural network for improving short text classification. Neurocomputing. 2016;174:806–814.

[2] J.-T. Chien, "Hierarchical theme and topic modeling," IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 565–578, 2016.

[3] Bernardini, C. Carpineto, and M. D'Amico, "Full-subtopic retrieval with keyphrase-based search results clustering," in IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intelligent Agent Technol., 2009, pp. 206–213

[4] Lu Pan, Haibo Tang and Lei Zhou, Liuyang Wang, Quanyin Zhu, "An Identification Method of News Scientific Intelligence Based on TF-IDF" 2015 IEEE DOI 10.1109/DCABES.2015.131

[5] Ning Li, Hui Zhang, Yong Chen, "Convolutional Neural Network with SDP-based Attention for Relation Classification" 2018 IEEE DOI 10.1109/BigComp.2018.00108

[6] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in Proc. Int. Conf. Inform. Knowl. Manag., 1998, pp. 148–155

[7] Ying Liu1, Peter Scheuermann2, Xingsen Li1, and Xingquan Zhu1Using WordNet to Disambiguate Word Senses for TextClassification.

[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero,and A. Saarela, "Self-organization of a massive document collection,"IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.

[9] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2007, pp. 490–499

[10] K. Lagus and S. Kaski, "Keyword selection method for characterizing text document maps," in Int. Conf. Artificial Neural Networks (ICANN), 1999, pp. 371–376

[11] Nihalr M. Ranjan [a,b,*] Rajesh S. Prasad [b] "LFNN: Lion fuzzy neural network-based evolutionary model for text classification using context and sense based features "https://doi.org/10.1016/j.asoc.2018.07.016 1568-4946/©2018 Published by Elsevier B.V

[12] Nihar M. Ranjan [a] and Rajesh S. Prasad [b] "Automatic text classification using BPLion-neural network and semantic word processing" THE IMAGING SCIENCE JOURNAL, 2017https://doi.org/10.1080/13682199.2017.1376781