A Survey on Semantic Document Object Classification

PROF.S.M.PRASAD
Department of Computer Engineering
JSPM's
Rajarshi Shahu College of Engineering Tathawade,
Pune, India.
tghumade210@gmail.com

.Dr.ARVIND PRASAD
Department of Computer Engineering
JSPM's
Rajarshi Shahu College of Engineering Tathawade,
Pune, India.
radesh19@gmail.com

ABSTRACT

Now a day the text mining studies are obtaining more importance because the availability of number of the documents from a different variety of sources, which includes unstructured and semi structured information. The main purpose of text mining is to enable users to extract information from resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization. Natural Language Processing (NLP), Data Mining, and Machine Learning work together to automatically classify and discover patterns from the different types of the documents. This paper proposed an overview and details study of document classification system using various NLP and Machine learning approaches. Many aspects have been proposed before for classification but none shows solution for multi label classification redundancy, such issue can be handled by introducing Recurrent Neural network (RNN) for classification.

Keywords— ***Document classification, RNN, Deep Learning, Multi label classification.***

I. INTRODUCTION

Text classification (TC) is very important part of text mining, looked to be that of manually building automatic TC systems by means that of knowledge-engineering techniques, i.e. manually process a group of logical rules that convert skilled data on a way to classify documents below the given set of categories.

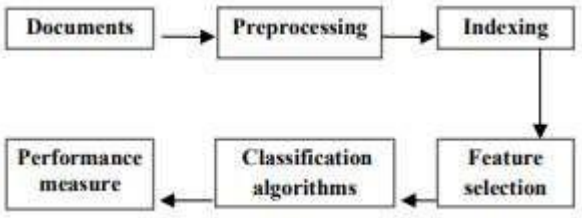

Figure 1: Document classification traditional approach

Figure 1 show the traditional approach of document classification, below are the phases which are basically used for classification.

Pre-Processing: The first step of pre-processing that is employed to presents the text documents into clear word format. The documents ready for next step in text classification as diagrammatic by a good quantity of options. Usually the steps taken are: Tokenization: A document is treated as a string, and so partitioned off into a listing of tokens. Removing stop words: Stop words like "the", "a", "and", etc sometimes occurring, therefore the insignificant words got to be removed. Stemming word: Applying the stemmer that converts totally different kind into similar canonical form. This step conflating tokens to their root kind, e.g. connection to connect, computing to compute.

Indexing: The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector.

Feature Selection: After pre-processing and indexing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. Most of algorithms used TF-IDF or density based approach for select the features.

Classification: The automatic classification of documents into predefined classes has established as a strong attention, the documents may be classified by many ways, unattended, supervised and semi supervised strategies. From previous couple of years, the task of automatic text classification is extensively studied and fast progress appears during this space, together with the machine learning approaches.

Several supervised learning techniques are exists for the classification of text documents namely Decision trees, Support Vector machine (SVM), Neural Network, Ada Boost and Naïve Bayes etc. Several clustering techniques are also available for text categorization namely K-means, Suffix Tree Clustering (STC), Semantic Online Hierarchical Clustering (SHOC), Label Induction Grouping Algorithm (LINGO) etc. [16]

Here, a survey of various document classification and clustering techniques are presented. Section II highlights the literature survey. In this section we cover the work that has already been done previously by researchers. Section III covers proposed system overview with various modules. Section IV presents the summary of survey in the form of conclusion and scope for future enhancement.

II. LITERATURE SURVEY

Sr. No	Paper Title	Techniques	Dataset	Result	Remark
1	A Text Classification Model Using Convolution Neural Network and Recurrent Neural Network.[2]	CNN and RNN Algorithms	Used Essay dataset 1000 for training and 500 for testing	Test Accuracy obtained by CNN is 50% and by RNN is 55%.	Here, the Evaluation performance is less.
2	Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification?[3]	CNN and RNN	They used unstructured data in English, French and Greek Restaurants and Hotels of 92 k reviews	Result shows we can avoid extra complex module when performing Multilingual classification.	Bigger dataset is not used for classification.
3	An Ontology-based and Domain Specific Clustering Methodology for Financial Document.[6]	They Proposed Words sense disambiguation method and evaluate two algorithms sequential Information Bottleneck and K-means algorithm	They used 446 document downloaded from EMMA repository. Document Categorized class label as state and purpose.	Obtain generalized and unbiased clustering result. Proposed methodology shown better cluster purity result compared to other disambiguated dataset.	Lack of availability of free financial dataset, evaluated experiment on limited dataset.
4	Text Dissimilarities Predictions using Convolutional Neural Networks and Clustering. [4]	Clustering Combining with CNN methods	Used 10 Arabic and 10 English long texts.	Result classify the text into two classes as reliable or suspicious texts and in many cases it can confirm the previous evaluation	Not using different encoding of words.

5	Activities of Daily Living Classification using Recurrent Neural Networks. [5]	RNN	Used DaLiAc(Daily Life Activities) Dataset,13 monitored daily life activity	The result shows accuracy of 82.5% using basic cross validation .	As future development create real time monitoring for data acquisition.
6	Automatic modulation classification using recurrent neural networks. [7]	RNN with LSTM and GRU	GNU Radio as benchmark dataset	Compared with CNN based method the proposed method obtains distinct advantage and high SNR regime and accuracy improved from 80% to 91%	Improve the performance for AMC problem at lower SNR Region
7	Convolutional and Recurrent Neural Networks for Real-time Data Classification. [8]	CNN and RNN with LSTM	German Credit Data split with 70:30 ratio	Experimental result of F1 score in case of CNN 0.8F1 and by LSTMs0.92F1. Precision and recall can adjusted by changing decision threshold.	Convolutional Neural Network's feature sharing methodology does not really work for Real time data
8	Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. [10]	CNN and RNN algorithm	Used 3 datasets 1. DSTC 4: Dialog State 2. MRDA: ICSI Meeting Recorder Dialog Act Corpus 3. SwDA: Switchboard Dialog Act Corpus	It achieves state-of-the-art results on Three different datasets for dialog act prediction. Accuracy by using datasets as DSTC 4, MRDA, SwDA by CNN 65.5 %,84.6%, 73.1%and LSTM 66.2 %,84.3%, 69.6% 65.5% resp.	Accuracy obtained is less in percentage
9	Text Categorization using Rocchio Algorithm and Random Forest Algorithm. [11]	Combines Rocchio algorithm and Random Forest	Used 20 newsgroup dataset in which each category contains 50 document for testing, RCV1 dataset consist of 804,414 new stories having 109 feature	Hybrid model of Rocchio and Random Forest for text categorization provide more accuracy and overcome disadvantages of Rocchio and Random forest Algorithm.	Marginally accurate than the existing.
10	Multi-label Text Categorization Based On Feature Optimization using Ant Colony Optimization and Relevance Clustering Technique. [15]	Ant Colony optimization algorithm	Used 3 datasets WebKB,Yahoo,RCV1	Performance with all tree dataset along F1,BEP,and HLOSS,Result of classification by ACO is better as compared with RSVM AND ML-FRC algorithm.	The optimization process reduce problem of data dimension and loss of data.

Table 1: Literature survey

III. PROPOSED SYSTEM

There are many algorithm and method has proposed for data classification in existing system we analysis those approaches in our literature survey. Figure 2 shows the design approach for document based text classification using supervised learning technique.

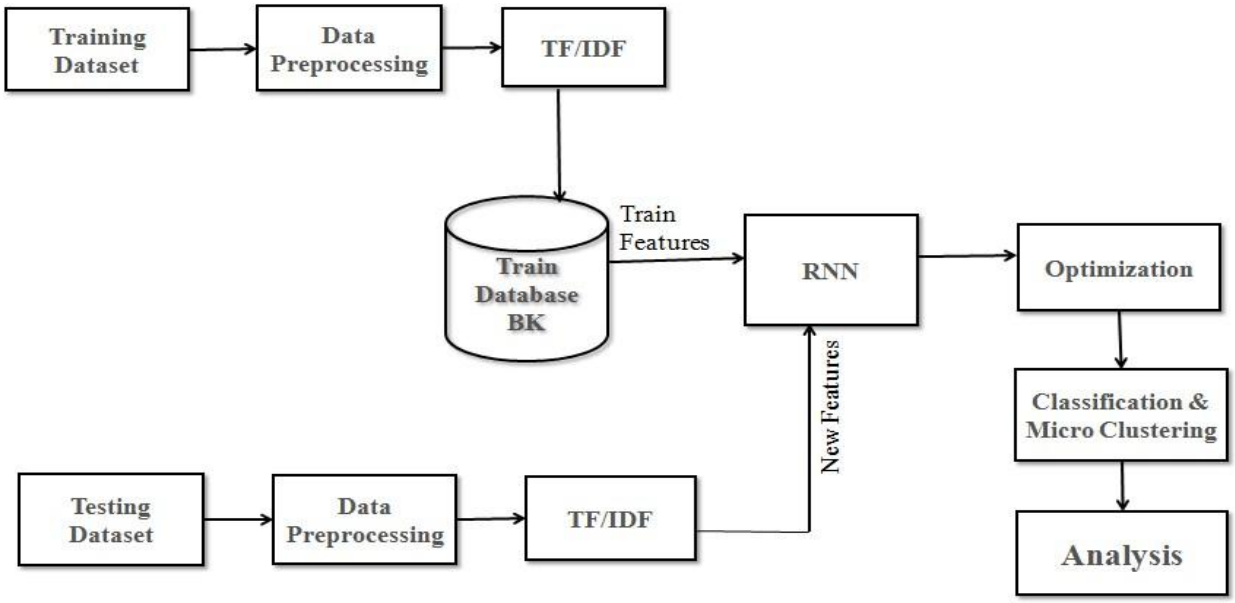

Figure 2: Classification System Overview

Basically NLP and ML has used for actual text classification for high dimensional as well as structured data. This text classification that classifies the documents with predefined classes. In this we have to provide the introduction of text classification, as well as the overview of the classifiers and tried to compare the some existing classifier on basis of few criteria like time complexity, performance evaluation.

A. System Modules

Module1: Data Training phase with pre-processing.
This module system creates the Background Knowledge (BK) according to given input dataset. Once we upload dataset system will read the abstract section from PDF using PDFBOX API.T hen tokenization, stop word removal and porter's stemmer will execute. Finally, TF-IDF will provide the availability of current vector and store into feature database.
When training phase has completed we have complete BK for all domain like cloud computing, data mining etc.
Module 2: Testing phase with preprocessing and TF-IDF.
First upload the testing dataset which does not have labels. The initial phase of testing is same like training phase till TF-IDF score calculation, it has use to identify the density of current test object. Then features will extract using RNN can calculate the similarity vector with all train features.
Module 3: Clustering Phase.
The similarity vector will return the current weight of test object with all training instances. Classification has to be done with respective weight factor. It will assign the label according to maximum weight generated by algorithm.
Module 4: Micro-clustering phase.
Final phase works for micro clustering base classification. It provides sub class categorization. Each cluster has categorized into multiple similar clusters, under the one master cluster. Finally, similarity score will classify each bucket into the respective domain.

IV. CONCLUSION

In this paper several existing classification ways has compared. It includes survey of traditional and various document classification approaches. Various steps such as stemming, stop word removal, tokenization and document classification

are followed in both techniques. From the higher than discussion it's understood machine learning base methods generate high time complexity as well as low accuracy like ANN and RF algorithm.

For further future enhancement we can introduce deep learning base approach to achieve the high text classification accuracy and will perform an effective solution to handle issue of multi label classification approach using Recurrent Neural Network.

REFERENCES

[1] Salem A, Almarimi A, Andrejková G. Text Dissimilarities Predictions Using Convolutional Neural Networks and Clustering. In2018 World Symposium on Digital Intelligence for Systems and Machines (DISA) 2018 Aug 23 (pp. 343-347). IEEE.

[2] Radhika K.1, Bindu K.R.2*, Latha Parameswaran" A Text Classification Model Using Convolution Neural Network and Recurrent Neural Network" International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018, 1549-1554.

[3] Medrouk L, Pappa A. Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification?. In2018 International Joint Conference on Neural Networks (IJCNN) 2018 Jul 8 (pp. 1-6). IEEE.

[4] Asmaa Salem, Abdulwahed Almarimi, Gabriela Andrejkov´a" Text Dissimilarities Predictions using Convolutional Neural Networks and Clustering" The research is supported by the Slovak Scientific Grant Agency VEGA,Grant No. 1/0056/18.

[5] Jurca R, Cioara T, Anghel I, Antal M, Pop C, Moldovan D. Activities of Daily Living Classification using Recurrent Neural Networks. In2018 17th RoEduNet Conference: Networking in Education and Research (RoEduNet) 2018 Sep 6 (pp. 1-4). IEEE.

[6] Kulathunga, Chalitha, and D. D. Karunaratne. "An ontology-based and domain specific clustering methodology for financial documents."Advances in ICT for Emerging Regions (ICTer), 2017 Seventeenth International Conference on. IEEE, 2017.

[7] Hong D, Zhang Z, Xu X. Automatic modulation classification using recurrent neural networks. In Computer and Communications (ICCC), 2017 3rd IEEE International Conference on 2017 Dec 13 (pp. 695-700). IEEE.

[8] Abroyan N. Convolutional and recurrent neural networks for real-time data classification. In Innovative Computing Technology (INTECH), 2017 Seventh International Conference on 2017 Aug 16 (pp. 42-45). IEEE.

[9] Zhang Y, Er MJ, Venkatesan R, Wang N, Pratama M. Sentiment classification using comprehensive attention recurrent models. In Neural Networks (IJCNN), 2016 International Joint Conference on 2016 Jul 24 (pp. 1562-1569). IEEE.

[10] ji Young Lee, Franck Dernoncourt "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks" Proceedings of NAACL-HLT 2016, pages 515–520, San Diego, California, June 12-17, 2016

[11] Thamarai Selvi. S, Karthikeyan. P, Vincent. A, Abinaya.VNeeraja. G, Deepika. "Text Categorization using Rocchio Algorithm and Random Forest Algorithm" IEEE Eighth International Conference on Advanced Computing (ICoAC) 2016.

[12] Gupta, Aditi, Jyoti Gautam, and Ajay Kumar. "A survey on methodologies used for semantic document clustering." 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). IEEE, 2017.

[13] Pengfei Liu,Xipeng Qiu,Xuanjing Huang et.Al."Recurrent Neural Network for Text Classification with Multi-Task Learning" (IJCAI-16).

[14] Huo, Zhuo-Liang, et al. "A topic-based cross-language retrieval model with PLSA and TF-IDF." Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on. IEEE, 2018.

[15] Nema, Puneet, and Vivek Sharma. "Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique." Computers, Communications, and Systems (ICCCS), International Conference on. IEEE, 2015.

[16]Sushma R. Vispute ; M. A. Potey"Automatic text categorization of marathi documents using clustering technique" 15th International conference on advanced computing technology(ICACT)2013.