A Document Classification using NLP and Recurrent Neural Network

Trupti G. Ghumade, R.A. Deshmukh

Abstract: ***The classification technique is most important for supervised and semi supervised base machine learning task. Many classification algorithms has introduced already for existing systems. Class-label classification is an important machine learning task wherein one assigns a subset of candidate without label to an object. Classification of various document models based on short text, metadata, heading levels these are the existing techniques which are introduced in literature survey. Sometime whole data reading and processing might be take a much time for classification, so it increase the time complexity for entire system. We proposed a new document classification method based on deep learning using NLP and machine learning approach. In this work system has several attractive properties: it captures some metadata from entire abstract section and built the training set first. Once complete all document process, it deals with optimization algorithm. Recurrent Neural Network has used to categories the individual object according to their weights. And it provides final class label for entire test dataset. Based on the various experimental analysis system provides data classification accuracy as well as minimum time complexity than classical machine learning algorithms.***

Keywords ***: Document Classification, Deep Learning, NPL, RNN.***

I. INTRODUCTION

A level of quality classification compilation is established and the combine evaluation method is adopted to classify the English text based on machine learning, which has made large progress now. Categorizing text is the activity of labeling natural language texts with one or more categories from a predefined set, which is one such task. According to machine learning (ML) methodology, we can build an automatic text classifier by learning, from a set of previously classified text documents based on the categories of interest.RNN is used for sequential information. Dissimilar to the traditional neural networks, in RNN all inputs neurons are independent. The recurrent concept applies to perform the same task upon each instance of succession, as they produce output which depends on previous estimation and outcomes. A fixed-size vector is produced to represent a sequence by feeding index one by one to a recurrent unit. RNNs have memory over the previous estimation and this information is used in current processing. RNN also provides network support to perform time distributed joint processing. RNN can be used for applications such as document classification, multi-label text categorization, multimodal sentiment analysis, and subjectivity detection [15].Many aspects have been proposed before for classification but none of these shows solution for multi-label classification redundancy, such an issue can be handled by introducing Recurrent Neural network (RNN) for classification.The proposed work focuses on document classification using the RNN algorithm, the system work with a supervised learning approach with 70-30% documents for training as well as testing respectively. In training and testing, common features extraction NLP based method and RNN weight classification are used for object classification.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Ms. Trupti G. Ghumade*, Student, Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering. Tathawade (Maharashtra), India.

Mrs. Rushali Deshmukh, Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering. Tathawade (Maharashtra), India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

II. LITERATURE REVIEW

In this section, we demonstrate the complete literature review of many classification and clustering techniques for various existing systems has done previously. We had also found some gaps in all those given surveys and our contribution to eliminating such problems in automatic text classification.

In [1] to sort the documents using a similarity metric that is based on keyword matching, they propose the k-means clustering algorithm and use information from WordNet and DBPedia as a probabilistic graph that can be used to determine the similarity between two terms. It gives higher precision and recall which correspond to the classification.

In [2] they build a text classification model using Convolution Neural Network and Recurrent Neural Network for essay dataset. From the train and test accuracy obtained they concluded that RNN performs better than CNN for essay dataset.In [3] they used two different networks one is convolution neural networks and the other is recurrent neural networks (RNNS). For this work, they used unstructured text data, which is gathered from the web, and written in French, English, and Greek and the result found that there is no need for complex modules during classification. In [4] they assert improved TF-IDF algorithm for calculating the weight of feature word and used deep learning tool for representing feature word into a vector. Then multiplying the feature word and word vectors weights, the vector representation of a word is perceived. At last each text is categorized according to all word vectors. Since the text classification is carried out. In [5] systems examine the text segments of a few long texts and find different stylometry of segment. They developed the two-steps methods one as clustering of segments; other as classification of segments using CNN. This technique has experimented with ten Arabic and English long texts.

The result classifies the text into two classes as reliable or suspicious text.In [6] they use an AMC method based on the (RNN), which has the potential to moderately utilize the timely corresponding sequence characteristic of received communication signals. This method retreats to raw signals directly with limited data length and ignores extracting signal features manually. As compared to the convolution neural network (CNN) based method and the output shown by the RNN method is advanced, in the signal-to-noise ratio (SNR. The accuracy is improved from 80% to 91%.In [7] they used methods for classification of real-time data using convolution and recurrent neural networks. They explored, experimenting and providing new approaches of classification non-stationary data using the RNN neural network. Experimental result of F1 score in the case of CNN 0.8 and by LSTMs 0.92.In [8] the bidirectional recurrent neural networks (BRNN) are used to retrieve the past and future data while a convolutional layer is used to encapsulate local data. Here the standard RNN is replaced by two recently appeared RNN modifications, called long short-term memory (LSTM) and gated recurrent unit (GRU), to increase the effectiveness of the new architecture for real-time classification. The basic advantage is that the experimental model is trained end-to-end without human involvement and it is easily implemented.In [9] system based on the ANTS algorithm and cluster mapping technique. The ant colony optimization algorithm plays a task of feature optimization of the text data mapping for classification. Performance with all tree dataset webKB, Yahoo, Rcv1, along F1, BEP, and HLOSS, Result of classification by ACO is better as compared with RSVM AND ML-FRC algorithm.In [10] they used Words sense disambiguation method and evaluate two algorithms Sequential Information Bottleneck and K means algorithm. They used 446 documents downloaded from the EMMA repository and the Document Categorized with a class label as state and purpose. The proposed methodology has shown better in cluster purity results.[16] Presents a survey of different techniques for clustering that have been studied and reflect the ascendancy and down sides of each algorithm. Semantic relationship of the term considered for clustering and did not depend on its only meaning.In [19] system detects the complaint tweets automatically using supervised and unsupervised learning. Here they applied the supervised learning to classify complaint tweets topic, whereas the unsupervised learning is applied to cluster tweets data based on the equivalence information of the complaints. Using the evaluation algorithms as Sequential Minimal Optimization (SMO), Naive Bayes, Multinomial, and Random Forests they evaluate the accuracy, precision, recall, and F1 score. SMO provides an accuracy average of 95% for single-label classification and Random Forests for multi-label topic classification with a 97.92F1 score. Unsupervised learning is used to evaluate the topic clusters detected by a clustering index value.In [17] neural network-based methods have obtained large progress on a variety of natural language processing tasks. Here using recurrent neural network framework to jointly learn across multiple related tasks. Text classification tasks show that the given model exceeds the execution of a task with the help of other related tasks.

In [18] they used deep learning to Arabic keyphrases extraction introduced Bi-LSTM neural network model, used to extract keyphrases from Arabic text. They have an insufficient large-scale dataset; therefore, they build a new dataset consisting of 6,000 abstracts of scientific Arabic documents. The attributes of this are equivalent to the English datasets. The evolutionary result of this method is significantly better as compared with the existing system as morphke, KP-minor for wiki all datasets.

In [12] system used a hybrid text categorization model that combines both the Rocchio algorithm and the Random Forest algorithm to perform Multi-label text categorization. These methods find a correlation between training Data sets and only related specifications filter required Classification. The accuracy obtained by hybrid text categorization is marginally more than individuals.

III. RESEARCH METHODOLOGY

This section represents the research work, where the system uses standard IEEE transaction PDF data set for training and testing. The architecture of the proposed system is shown in Fig.1. In the first phase of a system, Machine Learning (ML) based NLP features have been used for extract the feature set and RNN has used for generate the weight vector and define the similarity with a respective domain.

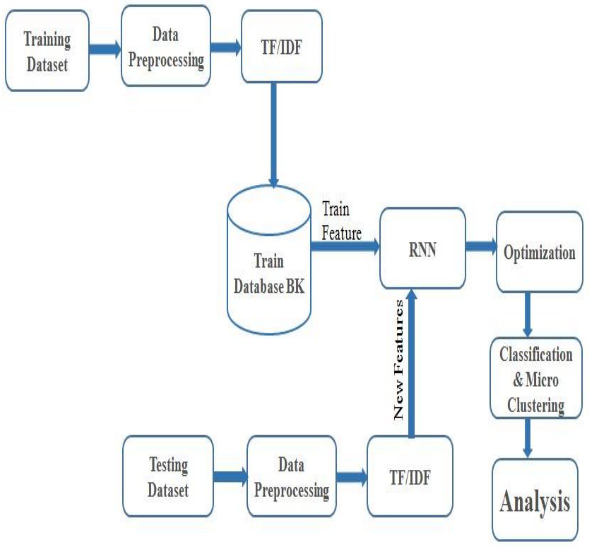

Fig. 1: System Architecture

The following modules show the linear execution of the system.

1. Data Training phase with pre-processing.
2. Testing phase with preprocessing and TF-IDF.
3. Clustering Phase.
4. Micro-clustering phase.

A. Data training phase with preprocessing

In this module, the system creates Background Knowledge (BK) according to the given input dataset.

Step 1: Once we upload the dataset system reads the abstract section from PDF using PDFBOX API.

Step 2: Then tokenization, stop word removal and porter's stemmer execute.

Step 3: Finally, TF-IDF (Term frequency-inverse Document frequency) provides the availability of the current vector and store into the feature database.

Step 4: When the training phase is completed we have complete BK for all domains like cloud computing, data mining, etc.

B. Testing phase with preprocessing and TF-IDF

Step1: First upload the testing dataset which does not have labels.

Step2: The initial phase of testing is the same as the training phase until the TF-IDF score calculation; it is used to identify the density of the current test object.

Step3: Then features are extracted using RNN can calculate the similarity vector with all train features.

C. Clustering phase with RNN

Step 1: The similarity vector returns the current weight of the test object with all training instances.

Step 2: Classification has done with respective weight factors.

Step 3: It assigns the label according to the maximum weight generated by the algorithm.

D. Micro clustering phase

Step 1: Final phase works for micro clustering-based classification.

Step 2: It provides sub class categorization.

Step3: Each cluster is categorized into multiple similar clusters, under the one master cluster.

Step 4: Finally, the similarity score will classify each bucket into the respective domain.The proposed system describes deep learning-based Recurrent Neural Network (RNN), basically, the system contains two different phases like training as well as testing. The training phase consists of the initial process; system used Natural Language Processing (NLP) to extract the best features. During the training phase features are extracted by system. Once training has done the system stores relevant features into the behalf of the respective domain, actually system works like a supervised learning and this extracted features known as background knowledge of the desired domain. After the completion of the training phase, it moves to evaluate the specific test object with the help of the proposed classification algorithm. The system uses text feature evaluation technique based on the similarity index, and RNN is used to classify the respective test object in testing. When multiple objects have given as testing module, the system first extracts the background Knowledge (BK) using NLP and selects top features according to their weights. TF-IDF has used to generate the respective terms of weight using NLP. When systems execute in the RNN phase, it needs input neuron as a feature vector which is extracted from a test object using the NLP process. Each term of an extracted feature called as neuron and such neurons should communicate with the hidden layer during the execution. When all input neurons communicate with the hidden layer, using cross-layer validation, it must need to pass training data to a hidden layer. Each hidden layer is associated with respective domain background information. In single-loop system creates any (T) instances from the n training set, and each input neuron communicates with respective Tth instance distance. Once the evaluation has done it generates a similarity score for specific text input with the respective hidden neuron. The feedback count will be increased whenever the scenario has generated like a similarity weight is higher than the given quality threshold, moreover, the feedback set is the final list of the systems output set. To reduce such weak instances from the given output set which will be generated by RNN, according to their weights, here the system uses the Hash map based optimization technique and find the superior based best instance for predict the result. Once the whole execution has done system creates automated confusion matrix evaluation for various experimental analyze.

IV. SYSTEM ANALYSIS

1. Word Removal Approach

Input: Stop words list L[], String Data D to remove the stop words

Output: Verified data D with removal all stop words

```
Step 1: Initialize the data string S[].
Step 2: initialize a=0,k=0
Step 3: for each(read a to L)
        If(a. equals(L[i]))
        Then Remove S[k]
        End for
Step 4: add S to D.
Step 5: End Procedure.
```

2. Stemming Algorithm

Input: Word w

Output: w with removing past participles as well

```
Step 1: Initialize w
Step 2: Initialize all steps of Porter stemmer
Step 3: for each (Char ch from w)
        If(ch.count==w.length()) && (ch.equals(e))
        Remove ch from(w)
Step 4: if(ch.endswith(ed))
        Remove 'ed' from(w)
Step 5: k=w.length()
        If(k (char) to k-3 .equals(tion))
        Replace w with te.
Step 6: end procedure.
```

3. TF-IDF

Input: Each word from vector as Term T, All vectors V[i…n]

Output: TF-IDF weight for each T

```
Step 1: Vector = {c1, c2, c3….cn}
Step 2: Aspects available in each comment
Step 3: D = {cmt1, cmt2, cmt3, cmtn}
        and comments available in each document
        Calculate the Tf score as
Step 4: tf (t,d) = (t,d)
        t=specific term
        d= specific document in a term is to be found.
Step 5: idf = t →   sum(d)
Step 6: Return tf *idf
```

4. Recurrent Neural Network

Input: Test Dataset which contains various test instances TestDBLits [], Train dataset which is built by training phase TrainDBLits[], Threshold Th.

Output: HashMap <class_label, SimilarityWeight> all instances in which weight violates the threshold score.

Step 1: For each read each test instances using the below equation

$$testFeature(m) = \sum_{m=1}^{n} (.\ featureSet[A[i] \ldots\ldots A[n] \leftarrow TestDBLits) \quad (1)$$

Step 2: extract each feature as a hot vector or input neuron from $testFeature\,(m)$using the below equation.

$$Extracted\_FeatureSetx\ [t\ldots\ldots n] = \textstyle\sum_{x=1}^{n}(t) \leftarrow testFeature\ (m) \quad (2)$$

Extracted_FeatureSetx[t] contains the feature vector of the respective domain

Step 3: For each read each train instances using the below equation

$$trainFeature(m) = \sum_{m=1}^{n} (.\ featureSet[A[i] \ldots\ldots A[n] \leftarrow TrainDBList) \quad (3)$$

Step 4: extract each feature as a hot vector or input neuron from $testFeature\,(m)$using the below equation.

$$Extracted\_FeatureSetx[t\ldots\ldots n] = \textstyle\sum_{x=1}^{n}(t) \leftarrow testFeature\ (m) \quad (4)$$

Extracted_FeatureSetx[t] contains the feature vector of the respective domain.

Step 5: Now map each test feature set to all respective training feature set

$$weight = calcSim\ (FeatureSetx \mid\mid \sum_{i=1}^{n} FeatureSety\,[y]) \quad (5)$$

V. RESULT AND DISCUSSION

For the system execution analysis, calculate the matrices for accuracy. The system is executed on a java 3-tier architecture framework with INTEL 2.7 GHz i3 processor and 4 GB RAM with a deep learning approach. Fig. 2 shows the classification results for different test data size. The 500 objects have used to validate the proposed system, which is categorized as 70-30% documents for training as well as testing respectively.

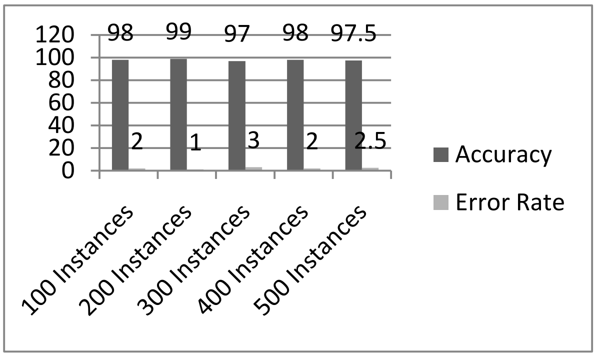

Fig. 2: Accuracy of the proposed system with different test instances

The second experiment evaluates the proposed system accuracy with some existing machine learning algorithms like NB [13], ANN [12], RF [11], etc. The propose RNN provides better accuracy for structured data classification. The proposed system is completely supervised learning which also shows a low error rate during the classification for the entire data set. The Fig. 3 shows the execution evaluation of the proposed system with existing classification algorithms.

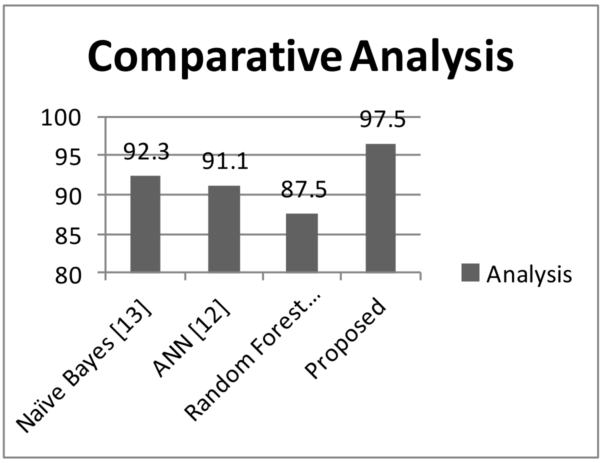

Fig.3: Comparative analysis of proposed system vs. existing classification algorithms

VI. CONCLUSION

Experimental results have demonstrated the effectiveness of our approach over several benchmark datasets; it produces better results than some existing approaches. The proposed system provides around 97.5% average accuracy after the multiple experiments. The system was evaluated with three different classification algorithms like [11, 12, 13] respectively, and shows the document object classification system perform an effective solution to handle redundancy generated during multi-label classification using Recurrent Neural Network then existing classification algorithms.

To deals with high dimensional unstructured as well as semi-structured data in a distributed environment will be the future work for this system.

REFERENCES

1. Lubomir Stanchev," Semantic Document Clustering Using Information from WordNet and DBPedia" 2018 12th IEEE International Conference on Semantic Computing, 0-7695-6360-0/18/$31.00 ©2018 IEEE DOI 10.1109/ICSC.2018.00023
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 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. Cai-Zhi Liu1, Yan-Xiu Sheng1, Zhi-Qiang Wei1 And Yong-Quan Yang" Research of Text Classification Based on Improved TF-IDF Algorithm".The International Conference of Intelligent Robotic and Control Engineering, 978-1-5386-7416-1/18/$31.00 ©2018 IEEE
5. 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.

6. 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.

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

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

9. 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.

10.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.

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

12.Pacifico LD, Macario V, Oliveira JF. Plant Classification Using Artificial Neural Networks. In2018 International Joint Conference on Neural Networks (IJCNN) 2018 Jul 8 (pp. 1-6). IEEE.

13.Singh T, Singla V, Bhatia P. Score and winning prediction in cricket through data mining. In2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI) 2015 Oct 8 (pp. 60-66). IEEE 2015.

14.Van Der Walt E, Eloff J. Using machine learning to detect fake identities: bots vs humans. IEEE Access. 2018;6:6540-9.

15.Tom Younga, Devamanyu Hazarikab, Soujanya Poriac, Erik Cambriad "Recent trends in deep learning-based natural language processing " arXiv:170 8.02709v4 [cs.CL] 1 6 Aug 20 17.

16.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.

17.Pengfei Liu, Xipeng Qiu, Xuanjing Huang .“Recurrent Neural Network for Text Classification with Multi-Task Learning” (IJCAI-16).

18.Muhmmad Hemly, R. M. Vigneshwaram, Giuseppe Serra, Carlo Tasso" Applying Deep Learning for Abarbic Key Phrases Extraction "2018, The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates. Published by Elsevier B. V. 1877-0509 © 2018.

19.Pratama, Timothy, and Ayu Purwarianti. ”Topic classification and clustering on Indonesian complaint tweets for Bandung government using supervised and unsupervised learning.” International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017. IEEE, 2017.

AUTHORS PROFILE

Ms. Trupti G. Ghumade is a research student currently pursuing M.E. Computer Engineering from JSPM's Rajarshi Shahu College of Engineering. The author has published papers in international and national conferences.

Mrs. Rushali Deshmukh pursued a Bachelor of Computer Engineering from Pune University, India in 1999 and Master of Computer Engineering from Pune University, India, in the year 2007. The author is having 19 years of teaching experience. She has published 24 papers in reputed Journals/Conferences. She has registered copyright on “Marathi Sentiment Dataset” and also published a patent “Improved smart shopping cart using client-server”. She is currently pursuing a Ph.D. in the area of Semantic Analysis of Natural Languages. Her main research work focuses on Natural Language Processing using Machine learning, and Data Mining.