*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

# A Robust Deep Model for Improved Categorization of Legal Documents for Predictive Analytics

**Divya Mohan[1], Latha Ravindran Nair[2]**
[1]School of Engineering
Cochin University of Science and Technology
Kerala, India
div.mohan01@gmail.com
[2]School of Engineering
Cochin University of Science and Technology
Kerala, India

**Abstract**— Predictive legal analytics is a technology used to predict the chances of successful and unsuccessful outcomes in a particular case. Predictive legal analytics is performed through automated document classification for facilitating legal experts in their classification of court documents to retrieve and understand the details of specific legal factors from legal judgments for accurate document analysis. However, extracting these factors from legal texts document is a time-consuming process. In order to facilitate the task of classifying documents, a robust method namely Distributed Stochastic Keyword Extraction based Ensemble Theil-Sen Regressive Deep Belief Reweight Boost Classification (DSKE-TRDBRBC) is proposed. The DSKE-TRDBRBC technique consists of two major processes namely Keyword Extraction and Classification. At first, the t-distributed stochastic neighbor embedding technique is applied to DSKE-TRDBRBC for keyword extraction. This in turn minimizes the time consumption for document classification. After that, the Ensemble Theil-Sen Regressive Deep Belief Reweight Boosting technique is applied for document classification. The Ensemble boosting algorithm initially constructs' set of Theil-Sen Regressive Deep Belief neural networks to classify the input legal documents. Then the results of the Deep Belief neural network are combined to built a strong classifier by reducing the error. This aids in improving the classification accuracy. The proposed method is experimentally evaluated with various metrics such as F-measure , recall, accuracy, precision, , and computational time. The experimental results quantitatively confirm that the proposed DSKE-TRDBRBC technique achieves better accuracy with lowest computation time as compared to the conventional approaches.

**Keywords**- legal document processing, t-distributed stochastic neighbor embedding Keyword Extraction, Ensemble Theil-Sen Regressive Deep Belief Reweight Boosting technique

## I. INTRODUCTION

The categorization of documents has emerged as a potentially fruitful endeavor to support intelligent information services Of the information contained inside them, the papers are sorted into the appropriate categories. The similar use in the legal realm has recently acquired a great deal of relevance. There are several chances for information extraction and application made available as a result of the availability of judicial decision papers in digital form. Due to the unusual architecture of these papers and their high level of complexity, automatic categorization of these legal texts is a task that is both essential and difficult to do.

In [1], a model called Joint Bidirectional Label Attention Conditional Network (JBLACN) was developed to categorize court record documents. This model was constructed using information that was retrieved from the evidence.The designed model performed court record documents classification but with less accuracy and required more computation time. Another method employing deep learning architecture namely Label-attended Multi-task Multi-label Classification (LAMT_MLC) model was proposed in [2]. The model demonstrates better precision and recall but classification accuracy reduced with insufficient training records and also with long texts.

A general framework was designed in [3] for the development of a system to categorize the inherent inter-correlations between parts of a legal text. The designed framework was not efficient to improve the results of classification algorithms. A semi structured document classification framework was developed in [4] using the semantic hierarchical attention method. While the time consumption during the classification task was not analyzed.

For document classification, artificial neural network model was introduced in [5].This method worked considerably well for long documents but accuracy was not improved. In [6], a graph neural network model was created for structure information categorization that was motivated by quantum probability. With more sophisticated document

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

categorization, however, the planned approach proved ineffective. A transfer learning system for document classification based on feature selection using Genetic Programming was introduced in [7]. The proposed method could not provide promising result by effectively combining Genetic programming with deep learning model architecture. For domain based Chinese legal document classification with knowledge extraction, [8] marked the beginning of the use of graph LSTM, which stands for long short-term memory. Even though the model improves classification accuracy, neither the performance nor the time consumption of legal document categorization was investigated. Convolutional Neural Network using Single-layer Multisize Filters (SMFCNN) was proposed in [9] for text document classification.

Three deep neural network architectures were designed in [10] based on a wide variety of network topologies and configurations. The designed architecture failed to consider the more real-case financial documents to maximize the effectiveness of deep learning to improve the functioning of the network.

Major Contributions

To address the issues mentioned in the literature review, a DSKE-TRDBRBC technique is proposed. This robust method provides contributions as given below,

- To enhance the document classification accuracy, a DSKE-TRDBRBC technique is introduced based on keyword extraction and classification.
- A t-distributed stochastic neighbor embedding method is applied to DSKE-TRDBRBC for keyword extraction based on frequency occurrence score. Based on score value, significant keywords are extracted. This helps to minimize the time consumption of document classification.
- Ensemble Theil-Sen Regressive Deep Belief Reweight Boosting technique is applied for document classification. The Ensemble boosting algorithm initially uses Theil-Sen Regressive Deep Belief neural network to categorize the input legal documents based on the extracted keywords using Tucker's congruence correlation coefficient. The Reweight Boosting technique combines the classification performance of Theil-Sen Regressive Deep Belief in order to increase the categorization performance neural network and minimize the error.
- The performance of the proposed method namely DSKE-TRDBRBC is estimated quantitatively and compared with other related works. The results thus obtained demonstrates better performance of our method in terms of recall, F-measure, computational time, precision and accuracy.

## II. Literature Review

Document classification using method namely hierarchical multi-attention network (HMAN) was proposed in [11] but performance in terms of accuracy was not improved. Semantic document classification strategies were developed in [12] to improve the accuracy by extracting the valuable features. However, it failed to analyze a multiple deep learning models for improving the document classification accuracy.

A new document representation method called Bag-of-Concepts (BoC) was developed in [13] for document classification. But it failed to integrate conceptual information into deep neural networks for better document classification. A novel Document Classification and Analysis (DoCA) framework was introduced in [14] for document analysis. However, the outlined framework was not more robust, flexible, and a feature-rich solution.

The document vector extension model was introduced in [15] for document representation learning. However, the feature matching was not applied for improving the better performance. Indian legal documents classification with a neural network which is simple and generic was depicted in [16]. The performance in terms of minimizing the time complexity of legal documents classification was not achieved.

Legal knowledge extraction from a assortment of legal documents using a knowledge-based framework was outlined in [17]. Though the method supported extraction of documents, the deep learning model failed to enhance the classification accuracy. A hybrid method for improving classification performance was proposed in [18] by the utilization of the Term Frequency-Inverse Document Frequency (TF-IDF) in conjunction with Supports Machine. The performance was not appreciable when applied to the larger dataset where the feature selection technique failed in terms of efficiency and effectiveness.

Document Clustering by Consensus and Classification (DCCC) was developed in [19]. However, the designed technique was not efficient to perform keyword extraction. A new taxonomy that is aware of hierarchy and attention graph capsules that are repeated For large-scale, multi-label text categorization, CNNs framework was created in [20]. Yet, the models that were developed failed to apply complex text classification datasets.

## III. Proposed Methodology

Massive legal documents consist of the knowledge that is both rich and valuable. The information included in these papers is mined so that attorneys general and judges can receive intelligent assistant case handling services. Hence it is vital to device an effectual method for legal document

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
*Article Received: 20 December 2022 Revised: 19 January 2023 Accepted: 26 January 2023*

classification and conventional manual methods are inefficient due to the enormous number of similar documents. A novel and robust method namely, DSKE-TRDBRBC technique is introduced for document classification.

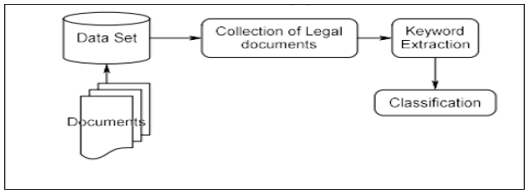


Figure 1.Proposed Methodology

The suggested DSKE-TRDBRBC approach may be broken down into two stages, which are referred to as the keyword extraction and classification stages, respectively, as illustrated in Figure 1 above. In the beginning, a collection of legal papers taken from the dataset are gathered. After that, the t-distributed stochastic neighbor embedding method, which is based on the frequency score, is utilized in the process of extracting keywords. This is followed by a classification step, which is performed using the boosting technique namely ensemble Theil-Sen Regressive Deep Belief Reweight.

### *A. t-distributed stochastic neighbor embedding technique based keyword extraction*

The first step of the proposed DSKE-TRDBRBC technique is to perform the keyword extraction refers to extracting the impotent keywords from documents. The main aim of the keyword extraction is to reduce the complexity of the process namely document classification. Conventional keyword extraction algorithms may lead to poor performance. Therefore the proposed DSKE-TRDBRBC technique uses the machine learning technique called t-distributed stochastic neighbor embedding is a dimensionality reduction approach that works well for embedding high-dimensional data. This technique is well-suited for embedding high-dimensional data. This technique is well-suited for embedding high-dimensional data. into low-dimensional space.

Let us consider the number of legal documents $DL_i \in DL_1, DL_2, DL_3 \ldots DL_m$ collected from the dataset. The document is made up of a certain amount of words.

Each piece of writing is made up of a certain amount of words. '$v_t'$. Therefore the word frequency score is calculated as follows,

$$S_{Freq} = \left(\frac{v_t(DL)}{v_k}\right) \qquad (1)$$

'$S_{Freq}$'- the frequent occurrence score of the word and '$v_n(DL)$' - the total number of occurrences of words inside the document. and '$v_k$'- the total amount of words contained within the report in question as shown in equation (1).

Based on the frequency score, significant keywords are extracted by applying the t-distributed stochastic neighbor embedding technique.

$$S_{ij} = \frac{(1+\|S_{Freq}-\vartheta\|^2)^{-1}}{\sum \quad \sum \quad (1+\|S_{Freq}-\vartheta\|^2)} \qquad (2)$$

From equation (2), $S_{ij}$ denotes a similarity function to measure the similarity between the number of times that the term appears often and the score '$S_{Freq}$' and threshold '$\vartheta$'. Significant keywords are selected for classification based on the proximity of words to the set threshold value.In this way, keywords are extracted from the documents. This helps to minimize time consumption.

**// Algorithm 1: t-distributed stochastic neighbor embedding technique based keyword extraction**
**Input**: Dataset, number of legal documents $DL_i \in DL_1, DL_2, DL_3 \ldots DL_m$
**Output**: Select the keywords
**Begin**
**Step 1:** Collect the number of legal documents '$DL_i = DL_1, DL_2, DL_3 \ldots DL_m$' from dataset
**Step 2:** **For** each legal document '$DL_i$'
**Step 3:** **For each** word in the document
**Step 4:** Measure the frequency score '$S_{Freq}$'.
**Step 5:** Find the significant keywords using (2)
**Step 6:** Select the keywords
**Step 7:** **end for**
**End**

The above algorithm namely, Algorithm 1 shows the step-by-step process for extraction of keywords which enhances the performance of classification task. For each word in the document, the frequent occurrence score value is measured. Then the significant keywords based on the threshold value are obtained by applying t-distributed stochastic neighbor embedding technique. This process helps to minimize the time complexity of document classification and hence enhance its performance.

### *B. Ensemble Theil-Sen Regressive Deep beliefs reweight boost classifier based document classification*

The proposed DSKE-TRDBRBC technique uses Ensemble TheilSen Regressive Deep beliefs reweight boost classifier for classifying the legal documents. A machine learning paradigm namely ensemble learning is employed which trains multiple models often referred as weak learners to solve a problem and results are combined to get a better output. Ensemble Deep belief neural network reweight boost

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

is boosting is the name of a strategy for supervised ensemble classification that is part of the family of meta-algorithms that seeks to combine learners with poor performance. The term "boosting" refers to an ensemble technique used in machine learning. This approach aggregates classification results to provide accurate and reliable classification outputs.

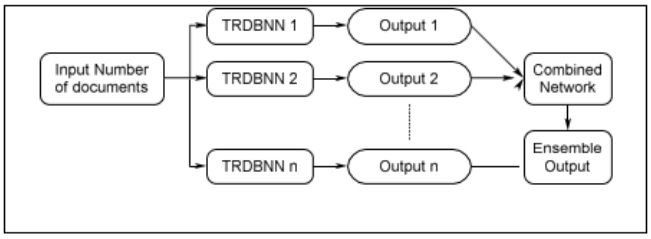


Figure 2. Ensemble Theil-Sen Regressive Deep beliefs reweight boost classifier

The above Figure 2 depicts the process of the Ensemble Theil- Sen Regressive Deep beliefs reweights boost classifier to receive the final results of the categorization. The ensemble classifier takes into account the training sets as (Di, Z), where D stands for the total number of documents and "Z" shows the results of the ensemble classification for the inputs that were provided. The ensemble method creates '$k$' number of Theil- Sen Regressive Deep belief neural network to classify the given input document.

The Theil-Sen Regressive Deep belief neural network is a type of deep artificial neural network consisting of many layers, which are highly effective in document classification. The proposed architecture consists of two units namely input unit and output unit and three hidden units. The model of the Theil-Sen Regressive Deep belief neural network is shown in figure 3.

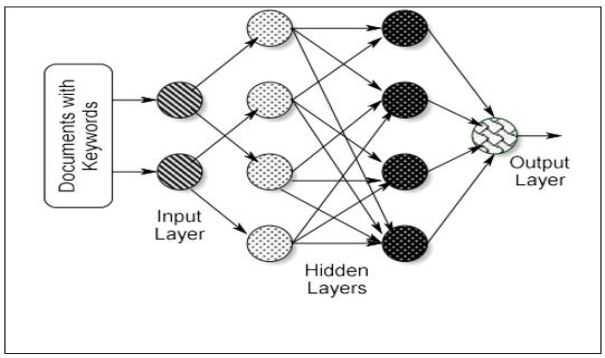


Figure 3. Schematic diagram of the Theil-Sen Regressive Deep belief neural network

Figure 3 depicts schematic illustration of the proposed TheilSen Regressive Deep belief neural network (TRDBNN) for document classification. Initially, set of legal documents are fed as input of initial layer. The input layer learns the given input and transforms it into the hidden layers. The network is modelled with two hidden layers. The nodes in different layers are connected in a feed- forward pattern to form the entire network. A machine learning technique namely regression function is used to identify keywords in the documents. In the first hidden layer, regression function initializes the mean value, number of classes . The correlation between the classes and documents are measured in second hidden layer using Tucker's congruence correlation coefficient.

$$\beta = \frac{(D_i)\,(m_j)}{\sqrt{\sum \quad ({D_i}^2 \sum \quad {m_j}^2)}} \qquad (3)$$

Where $\beta$ with value ranging from -1 to +1 indicates Tucker's congruence correlation coefficient, $D$i denotes documents, $m$j denotes a mean of particular class. Then the correlated values are passed to the second hidden layer.

The congruence correlation coefficient value ranges from -1 to +1. A negative correlation is indicated by the value '-1' while '+1' indicates a positive correlation. Based on the correlation value, the documents are suitably classified into their respective class with minimum time. The ensemble method boosts the classification performance by combining several deep learning classifiers.

$$Z = \sum_{i=1}^{k} \quad W_i \qquad (4)$$

Where $Z$ indicates an output of the ensemble classifier,$_i$ denotes the deep learning classifier results.

$$Z = \sum_{i=1}^{k} \quad W_i * \varphi \qquad (5)$$

Where '$\varphi$' denotes the weight initialized. Random numerical values are assigned as weights. After that, the squared error is computed. This is determined by taking the difference between the predicted classification results and the actual classification results and squaring it. Therefore, the error is mathematically represented using the given formula.

$$\vartheta_E = \{exp\,(W_i) - act(W_i)\} \qquad (6)$$

Where, $\vartheta_E$ shows the error of weak classifiers, $exp\,(W_i)$ represents the expected results, $act(W_i)$ indicates the actual results. The weight of each deep classifier is reassigned by checking the error rate and hence it is called re-weighting. The misclassification results attain higher weight and lesser weight for correctly classified results. The ensemble classifier computes the deep classifier results with minimum error. The algorithmic process of Ensemble Theil-Sen Regressive Deep belief reweight boost classifier based document classification is described as follows,

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

**// Algorithm 2: Ensemble Theil-Sen Regressive Deep belief reweight boost classifier based document classification**
**Input**: Dataset, number of legal documents $DL_i \in DL_1, DL_2, DL_3 \ldots DL_m$, Extracted keywords
**Output**: Increase classification Accuracy
**Begin**
**Step 1: For each** legal document '$DL_i$'
**Step 2:** Construct 'k' set of Deep belief neural network
**Step 3:** **Apply** Generalized Theil-Sen Regression
**Step 4:** **I**nitialize the number of classes $c_1, c_2, c_3 \ldots . c_k$ and mean '$m$' **---hidden layer 1**
**Step 5:** **Apply** Tucker's congruence correlation coefficient- **--hidden layer 2**
**Step 6:** **If the** document is closer to mean '$m_j$' **then**
**Step 7:** Classify the legal documents into a particular class
**Step 8:** Combine all the weak learner's results $Z = \sum_{i=1}^{k} W_i$
Step 9: **For all** $W_i$
Step 10: Assigns the similar weight '$\varphi$'
Step 11 **for each result of** $W_i$
Step 12: Calculate the error $\vartheta_E$
Step 13: Update the weight based on the error
Step 14: Find a weak learner with minimum error
Step 15: **Return** (classified legal documents) at the **output layer**
Step 16: **end for**
**End**

The above algorithm namely Algorithm 2 illustrates the document classification process. The weak classifiers – Theil-Sen Regressive Deep belief neural models are ensemble to built a strong classifier. Weights are initialized with random values for each deep learning classifier. The error rates are calculated and weights are reassigned. This is done for various iterations. Thus the ensemble process evaluates and comes up with the best classifier with minimum error. This in turn results in increasing the classification accuracy of the proposed method with minimum time.

## IV. Experimental Scenario

The proposed DSKE-TRDBRBC technique and existing works specifically JBLACN [1] and LAMT_MLC [2] are tested. The work is simulated using Java Programming language with the aid of Cloud Simulator. The dataset used for the work is fetched from the UCI repository. This is a popular repository with datasets for machine learning applications.
https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports.
A collection of 4000 legal case documents from the Federal Court of Australia (FCA) are used for the work. The case documents through the years 2006 to 2009 are included in the dataset. Classification task is performed on the text dataset. Various classes collected are 1.catchphrases 2. citations sentences 3. citation catchphrases 4. citation classes.

## V. Results And Discussions

The experimental evaluation of the proposed method namely, DSKE- TRDBRBC and existing works specifically JBLACN [1] and LAMT_MLC [2] are compared using metrics such as precision, recall, accuracy, computational time and F-measure. The performance of all the above mentioned three methods is visualized with the help of tables and graphs.

**Accuracy**: The accuracy is measured as given below,

$$Acc = \left[\frac{T_p+T_n}{T_p+T_n+F_p+F_n}\right] * 100 \qquad (7)$$

From (7), accuracy '$Acc$' is measured in percentage (%)
'$T_p$ – number of true positive
'$T_n$'- number of true negative
'$F_p$'- number of false positive
'$F_n$'- number of false-negative

TABLE I. Comparison of Accuracy

| Number of legal documents | Accuracy (%) | | |
|---|---|---|---|
| | DSKE-TRDBRBC | JBLACN | LAMT_MLC |
| 350 | 91.42 | 84.28 | 87.14 |
| 700 | 93.57 | 85.71 | 88.57 |
| 1050 | 92.38 | 87.61 | 89.52 |
| 1400 | 93.57 | 86.42 | 87.85 |
| 1750 | 92 | 85.14 | 87.42 |
| 2100 | 91.42 | 87.14 | 89.52 |
| 2450 | 91.02 | 82.44 | 85.71 |
| 2800 | 92.5 | 86.78 | 88.21 |
| 3150 | 93.01 | 86.66 | 87.61 |
| 3500 | 94 | 87.14 | 88.57 |

The experimental results of accuracy for the proposed method and its comparison with other two methods are depicted in Table 1. Various iterations are carried out with documents in the range of 350 to 3500. The accuracy values of all iterations are observed and on analysis it can be shown that the proposed method performs well compared to [1] [2]. For example, with the number of 350 documents the accuracy observed when applied DSKE-TRDBRBC technique is 91.42% and the accuracy was found to be 84.28%, 87.14% when applied with the existing [1] [2]. Ten iterations are run with varied number of documents and the analysis performed on the result of iterations shows that the accuracy of the proposed method is increased by 8% and 5% with respect to methods [1] and [2] respectively.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

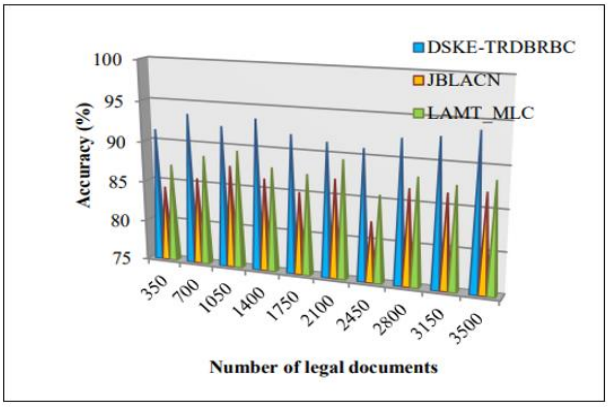


Figure 4. Numbers of documents versus accuracy

Figure 4 illustrates the graphical representation of accuracy for 10 different numbers of documents. As shown in figure 4, $x$ axis represents the number of documents and $y$ axis represents accuracy. The figure illustrates that for different numbers of documents, the accuracy is also different. The accuracy of the proposed method is shown to be better compared to the other two methods [1] [2].This is because the Ensemble Theil-Sen Regressive Deep Belief Reweight Boost Classification is performed. The Ensemble algorithm initially constructs' set of Theil-Sen Regressive Deep Belief neural networks to classify the given input legal documents. Then the results of the Deep Belief neural network are combined into a strong classifier by minimizing the error. This helps to enhance the accuracy.

**Precision**: The precision is calculated as follows

$$PR = \left[\frac{T_p}{T_p+F_p}\right] * 100 \qquad (8)$$

'PR' is measured in terms of percentage (%).

True positive -'$Tp$'
False positive - '$Fp$'

TABLE II. COMPARISON OF PRECISION

| Number of legal documents | Precision (%) | | |
|---|---|---|---|
| | DSKE-TRDBRBC | JBLACN | LAMT_MLC |
| 350 | 93.54 | 89.47 | 91.52 |
| 700 | 95.34 | 90 | 91.93 |
| 1050 | 94.84 | 91.20 | 92.47 |
| 1400 | 96.15 | 91.05 | 92 |
| 1750 | 95 | 90.72 | 92.25 |
| 2100 | 94.73 | 91.71 | 93.01 |
| 2450 | 94.57 | 86.66 | 88.53 |
| 2800 | 95.68 | 91.02 | 91.96 |
| 3150 | 95.84 | 91.27 | 91.75 |
| 3500 | 96.59 | 91.83 | 92.88 |

The performance of the proposed method in terms of precison and its comparison with other two methods [1] [2] is depicted in Table 2. The observed results illustrate that the precision founds to be higher using the DSKE-TRDBRBC when compared to other methods. As shown in table 2, '350' numbers of documents, the observed results of precision using DSKE-TRDBRBC technique is 93.54%. The precision of JBLACN [1], LAMT_MLC [2] are 89.47% and 91.52 % respectively. Ten iterations are run with varied number of documents and the analysis performed on the result of iterations shows that the precision of the planned method is increased by 5% and 4% with respect to methods [1] and [2] respectively.

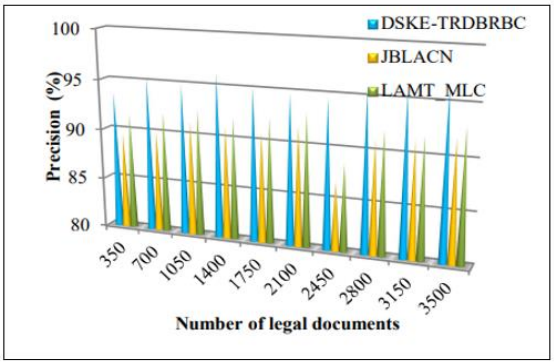


Figure 5. Numbers of documents versus precision

Figure 5 given above plots the convergence graph of precision with respect to 10 different runs observed. The plot indicates that precision using the DSKETRDBRBC technique is found to be increased when compared to the other two existing works.

**Recall**: is calculated as below

$$RR = \left[\frac{Tp}{Tp+Fn}\right] * 100 \qquad (9)$$

Recall is measured in terms of (%).

True positive - '$Tp$' , False-negative -'$Fn$'.

TABLE III. COMPARISON OF RECALL

| Number of legal documents | Recall (%) | | |
|---|---|---|---|
| | DSKE-TRDBRBC | JBLACN | LAMT_MLC |
| 350 | 96.66 | 91.07 | 93.10 |
| 700 | 97.61 | 93.10 | 95 |
| 1050 | 96.84 | 94.31 | 95.55 |
| 1400 | 96.89 | 93.33 | 94.26 |
| 1750 | 96.20 | 91.94 | 93.46 |
| 2100 | 95.74 | 93.25 | 95.05 |
| 2450 | 95.43 | 92.38 | 95.07 |
| 2800 | 96.06 | 93.69 | 94.62 |
| 3150 | 96.51 | 93.30 | 94.11 |
| 3500 | 96.89 | 93.35 | 94.09 |

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

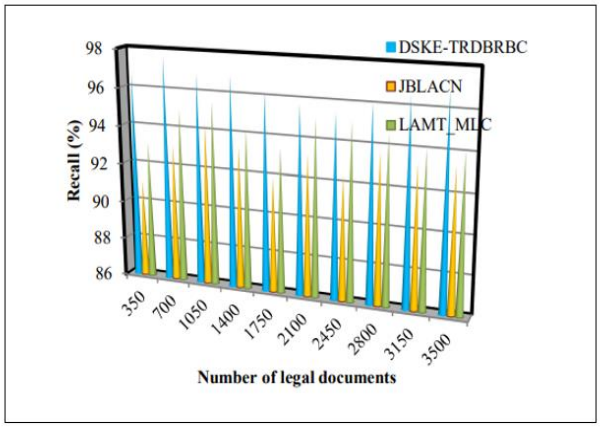


Figure 6. Numbers of documents versus recall

The performance of the proposed method in terms of recall and its comparison with other two methods [1] [2] is depicted in Figure 6 and Table 3. For a fair comparison, ten different runs are noted and the corresponding performance results of a recall are observed and plotted as a graph. The results shows that the recall rate of the DSKETRDBRBC technique is better than the other two existing methods. For example, with 350 documents are considered for experimentation, the recall rate using the proposed DSKE-TRDBRBC technique was observed to be 96.66%, when applied with JBLACN [1] and LAMT_MLC [2], 91.07% and 93.10% of accuracy was observed. Avenge of ten results indicates that the performance analysis of recall is found to be increased by 4% and 2% using DSKE-TRDBRBC technique when compared to existing methods. This is due to the proposed Ensemble Theil-Sen Regressive Deep Belief Reweight Boost Classification technique minimizes the false-negative and increases the true positives.

**F-measure**: is a measure of the mean of precision and recall. The formula for calculating F-measure is given below,

$$Fmeasure = 2 * [(PR * RR / (PR + RR)] * 100 \quad (10)$$

'$PR$' indicates precision, '$RR$' indicates recall.

TABLE IV. COMPARISON OF F-MEASURE

| Number of legal documents | F-measure (%) | | |
|---|---|---|---|
| | DSKE-TRDBRBC | JBLACN | LAMT_MLC |
| 350 | 95.07 | 90.26 | 92.30 |
| 700 | 96.46 | 91.52 | 93.43 |
| 1050 | 95.82 | 92.72 | 93.98 |
| 1400 | 96.51 | 92.17 | 93.11 |
| 1750 | 95.59 | 91.32 | 92.85 |
| 2100 | 95.23 | 92.47 | 94.01 |
| 2450 | 94.99 | 89.42 | 91.68 |
| 2800 | 95.86 | 92.33 | 93.27 |
| 3150 | 96.17 | 92.27 | 92.91 |
| 3500 | 96.73 | 92.58 | 93.48 |

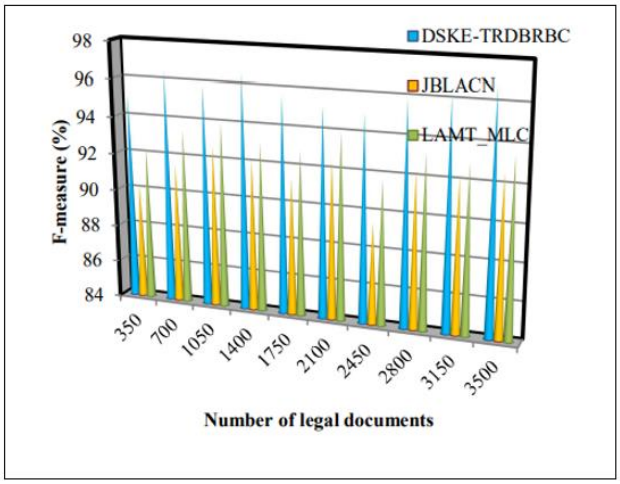


Figure 7. Numbers of documents versus F-measure

The results of F-measure are shown in Table 4 and Figure 7. The obtained result indicates that the performance of F-measure using the proposed DSKE-TRDBRBC technique is higher when compared to other two methods. The performance of the proposed method in terms of F-measure is higher compared to the other methods. Statistical estimation shows that for a collection of 350 documents, F-measure gives a value of 95.07% while the value for other two methods [1] [2] is 89% and 85% respectively. Different iterations are carried out with different number of documents. All the observed values are compared to the other two methods [1] [2].The final results are obtained by taking the average value and are found that the proposed method gives an F-measure value with an increase of 5% and 3% when compared to existing methods.

**Computational time**: gives the amount of time required for legal document classification. The overall time

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

consumption of document classification is formulated as given below,

$$CT = [n] * Time\ [CSD] \qquad (11)$$

Where $CT$ indicates a computational time, $n$ denotes the actual count of documents, $CSD$ indicates classification time of a single document.

TABLE V. COMPARISON OF COMPUTATIONAL TIME

| Number of legal documents | Computational time (ms) | | |
|---|---|---|---|
| | DSKE-TRDBRBC | JBLACN | LAMT_MLC |
| 350 | 21 | 28 | 25 |
| 700 | 26 | 32 | 29 |
| 1050 | 29 | 36 | 34 |
| 1400 | 35 | 43 | 39 |
| 1750 | 39 | 46 | 42 |
| 2100 | 44 | 50 | 48 |
| 2450 | 47 | 54 | 51 |
| 2800 | 50 | 59 | 56 |
| 3150 | 54 | 63 | 57 |
| 3500 | 56 | 67 | 60 |

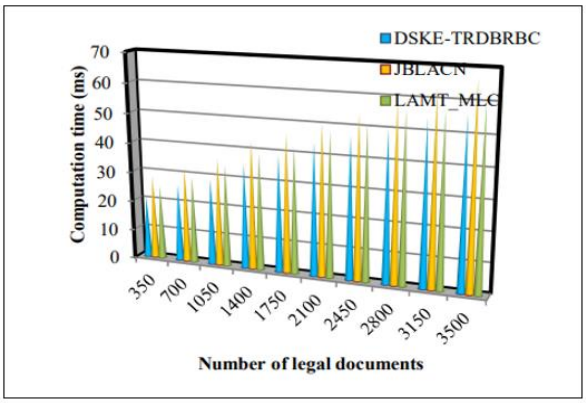


Figure 8. Numbers of documents versus computatuinal time

Table 5 and Figure 8 illustrate the computational time using three methods namely the DSKE-TRDBRBC technique and existing JBLACN [1] LAMT_MLC [2]. Here, computational time refers to the time consumed for legal document classification. For a fair comparison, ten different runs are observed and the corresponding computational time is also observed and plotted as a graph. Using the proposed method, when 350 documents are considered for experimental evaluation the computational time involved is 21ms. However it takes 28$ms$ and 25$ms$ while employing methods [1] and [2] respectively. Hence, from the convergence plot, the computational time of all the methods gets increased with the increase in the number of documents. But the computational time using the DSKE-TRDBRBC technique is found to be minimized. This is because the technique of t distributed stochastic neighbor embedding is used for keyword extraction, and the explanation for that is as follows: A score based on the frequency with which a word appears in the document is determined for each word. The relevant keywords are retrieved by referring to the threshold value, which is based on the value that has been computed. This in turn highly aids in minimizing the time complexity of classifying documents.

## VI. CONCLUSION

In this paper, a robust method namely DSKE-TRDBRBC is presented which aids in predictive analytics of legal documents. The ensemble deep learning models showed superior performance than the machine learning. Hence the proposed DSKE-TRDBRBC technique performs the document classification using ensemble deep learning models. The DSKETRDBRBC first performs the keyword extraction using the t-distributed stochastic neighbor embedding technique. Followed by, Ensemble Theil-Sen Regressive Deep Belief Reweight Boosting technique is applied for document classification with the extracted keywords. The Ensemble boosting algorithm uses the Theil-Sen Regressive Deep Belief neural network to classify the legal documents. A strong classifier with minimized error is then built with the results of the Deep Belief Neural networks combined together. Thus the accuracy of the classifier is improved. To analyze the performance of proposed method significant evaluation metrics such as precision, accuracy, recall, computational time and F-measure in terms of the number of different legal papers that are employed. According to the findings of the comparison, the suggested approach achieves higher levels of accuracy than the other standard methods with minimum time.

## REFERENCES

[1] Donghong Ji, Peng Tao, Hao Fei, Yafeng Ren, “An end-to-end joint model for evidence information extraction from court record document”, Information Processing & Management, Elsevier,Volume 57, Issue 6, 2020, Pages 1-14

[2] Dezhao Song, Andrew Vold, Kanika Madan, Frank Schilder, “Multi- label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training”, Information Systems, Elsevier, 2021, Pages 1-12

[3] Emilio Sulis, Llio Humphreys, Fabiana Vernero, Ilaria Angela Amantea, Davide Audrito, Luigi Di Caro, “Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts”, Information Systems, Elsevier, 2021, Pages 1-12.

[4] Weizhong Zhao, Dandan Fang, Jinyong Zhang, Yao Zhao, Xiaowei Xu, Xingpeng Jiang, Xiaohua Hu, Tingting He, “An

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 3s*
*DOI: https://doi.org/10.17762/ijritcc.v11i3s.6179*
***Article Received**: 20 December 2022 **Revised**: 19 January 2023 **Accepted**: 26 January 2023*

effective framework for semistructured document classification via hierarchical attention model", International Journal of intelligent system, Wiley, Volume 36, Issue 9, 2021, Pages 5161-5183

[5] Kshitij Tripathi, Rajendra G. Vyas, and Anil K. Gupta, "Document Classification Using Artificial Neural Network", Asian Journal of Computer Science and Technology, Volume 8 Issue 2, 2019, Pages 55-58

[6] Peng Yan, Linjing Li, Miaotianzi Jin, Daniel Zeng, "Quantum probability-inspired graph neural network for document representation and classification", Neurocomputing, Elsevier,Volume 445, 2021, Pages 276-286

[7] Wenlong Fu, Bing Xue, Xiaoying Gao, Mengjie Zhang, "Output- based transfer learning in genetic programming for document classification", Knowledge-Based Systems, Elsevier, Volume 212, 2021, Pages 1-11

[8] Guodong Li, Zhe Wang, Yinglong Ma, "Combining Domain Knowledge Extraction With Graph Long Short-Term Memory for Learning Classification of Chinese Legal Documents", IEEE Access,Volume 7, 2019, Pages 139616 – 139627

[9] Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, Atif Mehmood, and Muhammad Tariq Sadiq, "Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network", IEEE Access,Volume 8, 2020, Pages 42689 – 42707

[10] Zenun Kastrati, Ali Shariq Imran, Sule Yildirim Yayilgan, "The impact of deep learning on document classification using semantically rich representations", Information Processing & Management, Elsevier, Volume 56, Issue 5, 2019, Pages 1618-1632

[11] Yingren Huang, Jiaojiao Chen, Shaomin Zheng, Yun Xue, Xiaohui Hu, "Hierarchical multi‐attention networks for document classification", International Journal of Machine Learning and Cybernetics, Springer, Volume 12, Issue 3, 2021

[12] Shuo Yang, Ran Wei, Jingzhi Guo, Hengliang Tan, "Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis", Journal of Web Semantics, Elsevier, Volume 63, 2020, Pages 1-15

[13] Pengfei Li, Kezhi Maoa, Yuecong Xu, Qi Li, Jiaheng Zhang, "Bag- of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base", Knowledge-Based Systems, Elsevier, Volume 193, 2020, Pages 1-14

[14] Süleyman Eken, Houssem Menhour, Kübra Köksal, "DoCA: A Content-Based Automatic Classification System Over Digital Documents", IEEE Access, Volume 7, 2019, Pages 97996 – 98004

[15] Veena Hosamani,H S Vimala, "Data Science: Prediction and Analysis of Data using Multiple Classifier System", International Journal of Computer Engineering in Research Trends, Volume 5, Issue 12, 2019, Page(s): 216- 222.

[16] Deepa Anand, Rupali Wagh, "Effective deep learning approaches for summarization of legal texts", Journal of King Saud University -Computer and Information Sciences, Elsevier, 2019, Pages 1-18

[17] Silvana Castano, Mattia Falduti, Alfio Ferrara, Stefano Montanelli, A knowledge-centered framework for exploration and retrieval of legal documents", Information Systems, Elsevier, 2021, Pages 1-14

[18] Y.Yashasree, K.Venkatesh Sharma "Creditcard Fraud Detection and Classification Using Machine Learning Based Classifiers", International Journal of Computer Engineering in Research Trends, Volume 7, Issue 9, 2020, Page(s): 1- 8.

[19] Ahmad Muqeem Sheri, Muhammad Aasim Rafique, Malik Tahir Hassan, Khurum Nazir Junejo, Moongu Jeon Gwangju, South Korea, "Boosting Discrimination Information Based Document Clustering Using Consensus and Classification", IEEE Access, Volume 7, 2019,Pages 78954 – 78962

[20] Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, Lifang He, and Philip S. Yu, "Hierarchical Taxonomy-Aware and Attentional Graph Capsule RCNNs for Large- Scale Multi-Label Text Classification", IEEE Transactions on Knowledge and Data Engineering, Volume 33, Issue 6, 2021, Pages 2505 – 2519