ISSN No:-2456-2165

A Learning Based Approach for Automatic Text Document Classification

[1]Ravi Prasad Ravuri
Application Developer, Sriven Technologies, Ashburn, VA, USA

Abstract:-Text documents over Internet, social media and in internal applications of various organizations such as judiciary are increasing exponentially. Manual observation of such documents and classifying them for further processing is tedious task. There is need for automatic text document classification. Traditional heuristics based approaches have limitations to scale up to the demand in terms of volumes of input documents. To overcome this problem, machine learning (ML) techniques are used as they can learn from the training data and perform classification. They can also deal with large corpus. However, existing ML models when used directly their performance gets deteriorated due to lack of training quality. In this paper we proposed a framework that has a hybrid approach including feature selection and also ML models towards leveraging prediction performance. Our framework is named as Learning based Text Document Classification Framework (LbTDCF). We also proposed an algorithm known as Intelligent Document Classification Algorithm (IDCA) to realize our framework. Experimental results have revealed that our framework improves classification results with 99.30% accuracy.

Keywords:- *Machine Learning, Text Document Classification, Supervised Learning, Intelligent Document Classification*

I. INTRODUCTION

Text documents over Internet, social media and in internal applications of various organizations such as judiciary are increasing exponentially. Manual observation of such documents and classifying them for further processing is tedious task. There is need for automatic text document classification. Traditional heuristics based approaches have limitations to scale up to the demand in terms of volumes of input documents. To overcome this problem, machine learning (ML) techniques are used as they can learn from the training data and perform classification. They can also deal with large corpus [1], [2]. It is observed that ML models provide a scalable solution to the problem in hand.

There are many contributions found the literation based on ML methods. In [2] there is deep learning process explored for automatic classification of documents associated with requests to government from citizens. Toxicological abstracts are considered in [5] for automatic classification based on the similarity of the content. They used deep learning models to achieve this. Intellectual property related documents are considered in [8] for analysis. They used ML models, deep learning models along with AI approaches to gain knowledge from analysis of IP records. Fuzzy inference rule based methodology is used in [10] for automatic categorization of documents. Their research has focused on the web text documents by modelling ML and fuzzy logic. Accident prone areas are discovered in [13] using text mining process made up of NLP and deep learning. An improved CNN model known as faster RCNN is used in [15] for identification of complex documents and classify them automatically. Other important contribution. From the literature, it is observed that there are ML and deep learning methods used for text document classification. However, NLP and strong feature engineering cloud improve classification performance further. Our contributions in this paper are as given below.

- We proposed a framework, named Learning based Text Document Classification Framework (LbTDCF), that has a hybrid approach including feature selection and also ML models towards leveraging prediction performance.
- We also proposed an algorithm known as Intelligent Document Classification Algorithm (IDCA) to realize our framework.
- We built an application to evaluate LbTDCF and its underlying algorithm.

The rest of the paper is divided into multiple sections. Section 2 reviews literature on different existing methods used for classification of text documents. Section 3 presents our proposed framework that reflects a hybrid approach with feature selection and multiple ML models. Section 4 provides our results on document classification. Section 5 draws conclusions on our work and provides possible scope of future work.

II. RELATED WORK

This section reviews literature on methods of text document classification. Scanned health records are subjected to classification in [1] for helping healthcare unit. It could achieve automatic classification of documents with supervised learning process. In [2] there is deep learning process explored for automatic classification of documents associated with requests to government from citizens. In [3], bio-medical documents are considered for processing and classification. They exploited different advanced neural network based models for document classification and found that deep learning showed higher level of accuracy in classification. An improved approach based on deep learning is made in [4] for classification of safety reports. Their work considered text classification focusing on near-misses in the text corpora. Toxicological abstracts are considered in [5] for automatic classification based on the similarity of the content. They used deep learning models to achieve this. Text classification methods found in [6] are with diversified approaches. They exploited different kinds of methods for text classification. Different supervised learning approaches are used in [7] for automatic text classification. Their study has provided a comparative study to ascertain the strength of each ML model.

Intellectual property related documents are considered in [8] for analysis. They used ML models, deep learning models along with AI approaches to gain knowledge from analysis of IP records. In [9] explored different techniques in ML and deep learning to solve problems in different applications. They found that deep learning techniques provide better learning approaches. Fuzzy inference rule based methodology is used in [10] for automatic categorization of documents. Their research has focused on the web text documents by modelling ML and fuzzy logic. Clinical notes are considered for their research in [11]. Their methodology is to learn from such documents using ML and deep learning models along with the process known as ensemble for better performance. In [12] considered the accident narratives in construction industry to perform automatic classification and visualization u sing deep learning techniques. Accident prone areas are discovered in [13] using text mining process made up of NLP and deep learning. In [14] considered open source java applications as the text corpora and proposed a methodology for automatic categorization of the projects. An improved CNN model known as faster RCNN is used in [15] for identification of complex documents and classify them automatically. Other important contributions include Arabic text classification [16], novel feature selection method [17], schizophrenic discourse classification [18] and ML optimization method [19]. From the literature, it is observed that there are ML and deep learning methods used for text document classification. However, NLP and strong feature engineering cloud improve classification performance further.

III. PROPOSED FRAMEWORK

We proposed a framework that has a hybrid approach including feature selection and also ML models towards leveraging prediction performance. Our framework is named as Learning based Text Document Classification Framework (LbTDCF). We also proposed an algorithm known as Intelligent Document Classification Algorithm (IDCA) to realize our framework. Figure 1 shows the proposed framework that has both feature selection process and also document classification. The framework takes text corpora, pre-processes it and divides it into training and test documents.

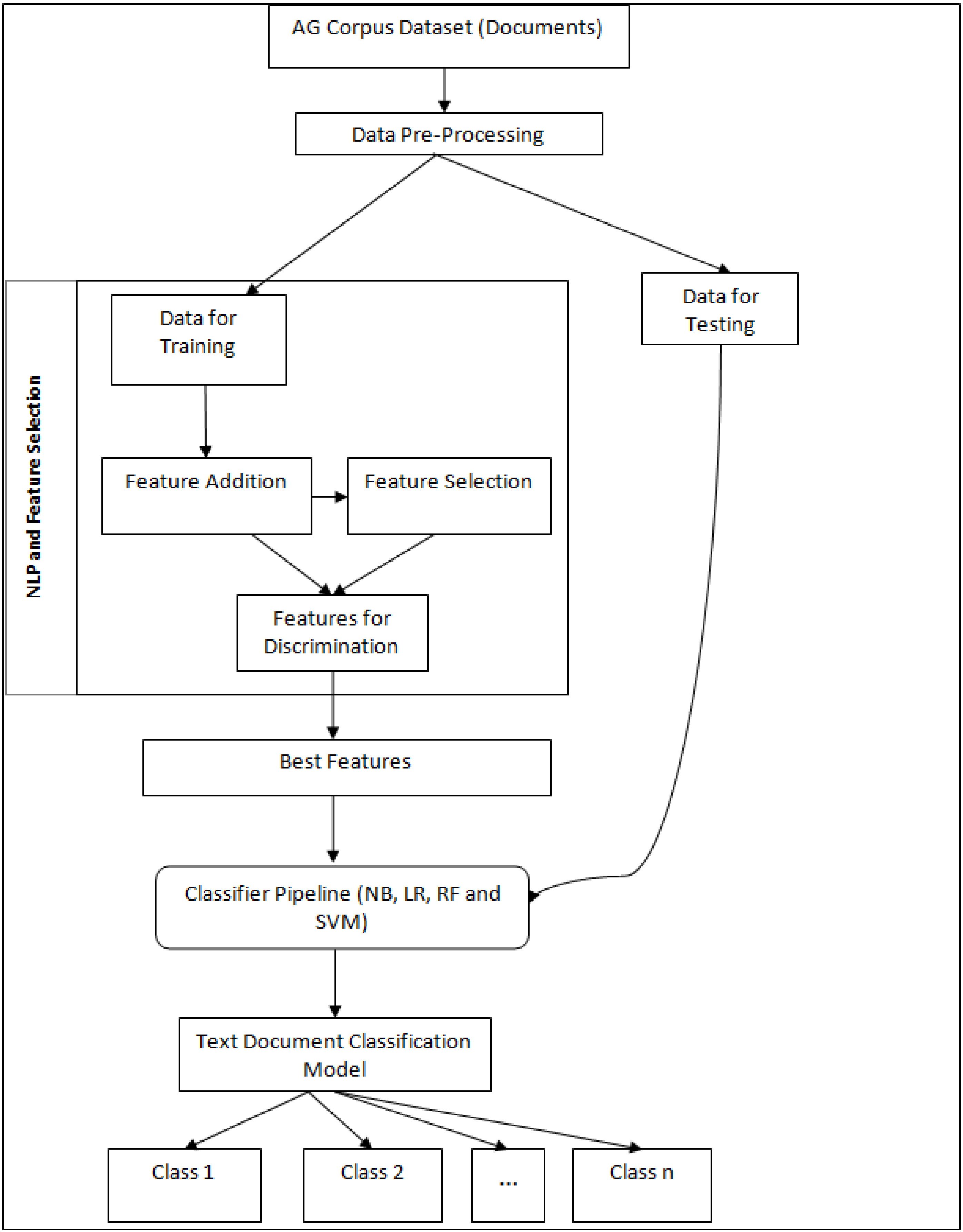

AG Corpus Dataset (Documents)
Data Pre-Processing
NLP and Feature Selection
Data for Training
Data for Testing
Feature Addition
Feature Selection
Features for Discrimination
Best Features
Classifier Pipeline (NB, LR, RF and SVM)
Text Document Classification Model
Class 1
Class 2
...
Class n

Fig 1 Learning Basedtext Document Classification Framework (Lbtcf)

- The framework has support for NLP where it makes use of techniques like removal of stop words and lemmatization to improve quality of training data. Feature selection is carried out to improve predication performance of classifiers. TF-IDF is a vectorization method that helps in data representation which is crucial for feature selection. Word embedding is also used to select features to make the pre-processing must stronger. TF-IDF computation is as expressed in Eq. 1.

$$W_{i,j} = tf_{i,j} \times \text{Log}\left(\frac{N}{df_i}\right) \qquad (1)$$

- Where $tf_{i,j}$ denotes number of occurrences of i in j, df_i denotes number of documents containing i and N indicates total number of documents. After feature engineering ML models are trained and used for document classification. Naïve Bayes technique is one of the ML models used. It has based on Bayes theorem.

$$\text{P}(\text{c}/\text{x}) = \frac{\text{P}(\text{x}/\text{c})\text{P}(\text{C})}{\text{P}(\text{x})} \qquad (2)$$

- As in Eq. 2, it involves different aspects like likelihood, class prior probability, predictor prior probability and posterior probability. This can be further modelled as in Eq. 3.

$$P(c/x) = P(x_1/c) \times P(x_2/c) \times \cdots P(x_n/c) \times P(c) \qquad (3)$$

- Logistic Regression is a statistical model which models a binary dependent variable by using a logistic function. It is also known as sigmoid function which is as given in Eq. 4.

$$\text{F(x)} = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \qquad (4)$$

- This function helps the model to obtain values required by binary classification. If p(x), an unbounded linear function, is assumed as linear function, probability is denoted by p which ranges from 0 to 1. To solve the problem, let log p(x) is a linear function and log p(x)/(1-p(x)) is expressed as in Eq. 5.

$$\text{Log}\frac{p(x)}{1-p(x)} = \alpha_0 + \alpha.x \qquad (5)$$

- Once the problem of p(x) is solved, it can be expressed as in Eq. 6.

$$\text{P(x)} = \frac{e^{\alpha_0+\alpha}}{e^{\alpha_0+\alpha}+1} \qquad (6)$$

- In order to make logistic regression as a linear function there is need for a threshold which is set to 0.5 and rate of misclassification is minimized.

- RF is the model based on the generation of number of trees from the data and have an explicit ensemble. SVM is another model which has discriminative power in classification tasks based on its hyperplane phenomeno

```
Algorithm:Intelligent Document Classification Algorithm (IDCA)
Inputs:
Dataset D
Pipeline of ML models M (include NB, LR, RF and SVM)
Output:
Classification results R

Begin
D'←DatePreProcess(D)
(T1, T2)←DataSplit(D')
F←ExtractFeatures(T1)
Initialize knowledge models map K
For each model m in pipeline M
model←TrainTheCurrentModel(m, F)
Add m and model to K
End For
F←Extract Features(T2)
For each entry k in K
  For each test instance t in T2
r←ClassifyDocument(k.m, F, t)
  Update R with r
End For
End For
Return R
End
```

- *Algorithm 1: Intelligent Document Classification Algorithm (IDCA):-*

Algorithm 1 presents the proposed algorithm. It takes AG Corpus dataset that contains text documents as input. It performs pre-processing using NLP techniques and then split the data into training and test documents. It also takes a pipeline of ML models that exploit feature selection in order to improve classification performance. There is an iterative process to train all the models. Then the testing phase involves classification of each document is classified by each ML model and the classification results are provided. Accuracy is the metric used as expressed in Eq. 7 for performance evaluation.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Accuracy metric helps in computing the percentage of correct predictions in terms of documents belonging to specific classes.

IV. EXPERIMENTAL RESULTS

An application is built to evaluate our framework and algorithm. The dataset is taken from [20]. It has different categories of documents associated with news. It has 7600 test samples and 120000 training samples. This section presents our experimental results.

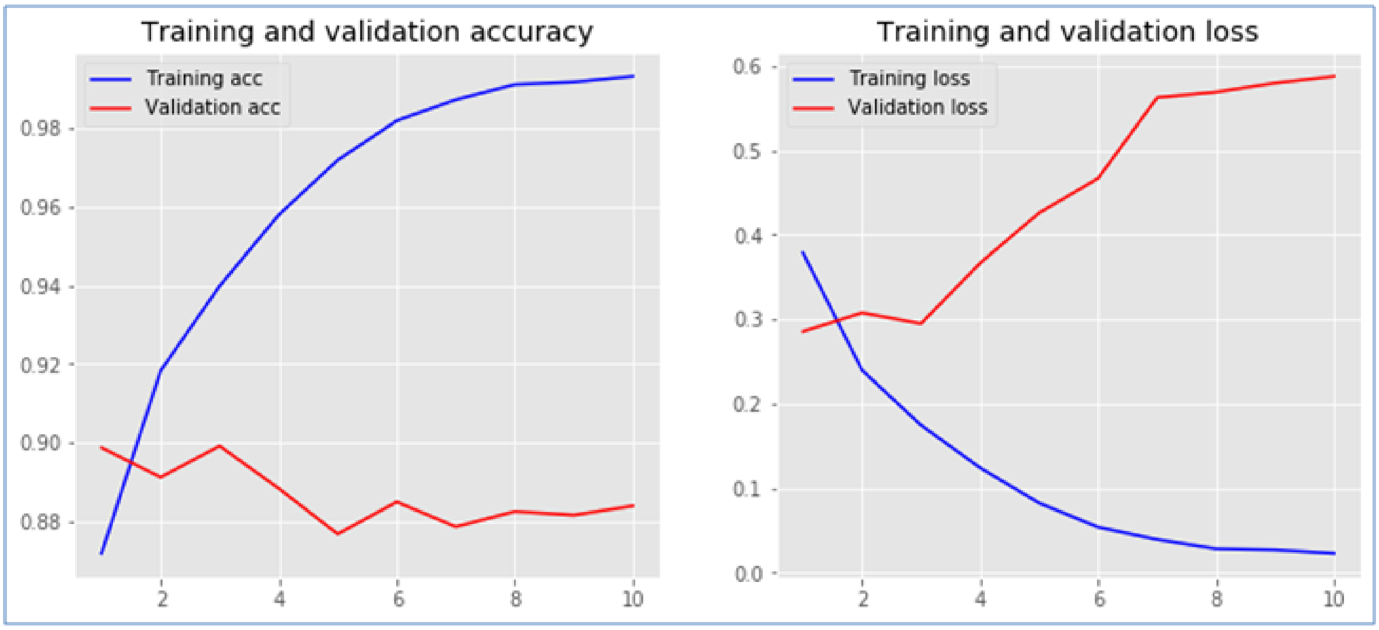


Fig 6 Document Classification Performance of RF

As presented in Figure 6, our methodology when used with RF has given its accuracy and also loss function values.

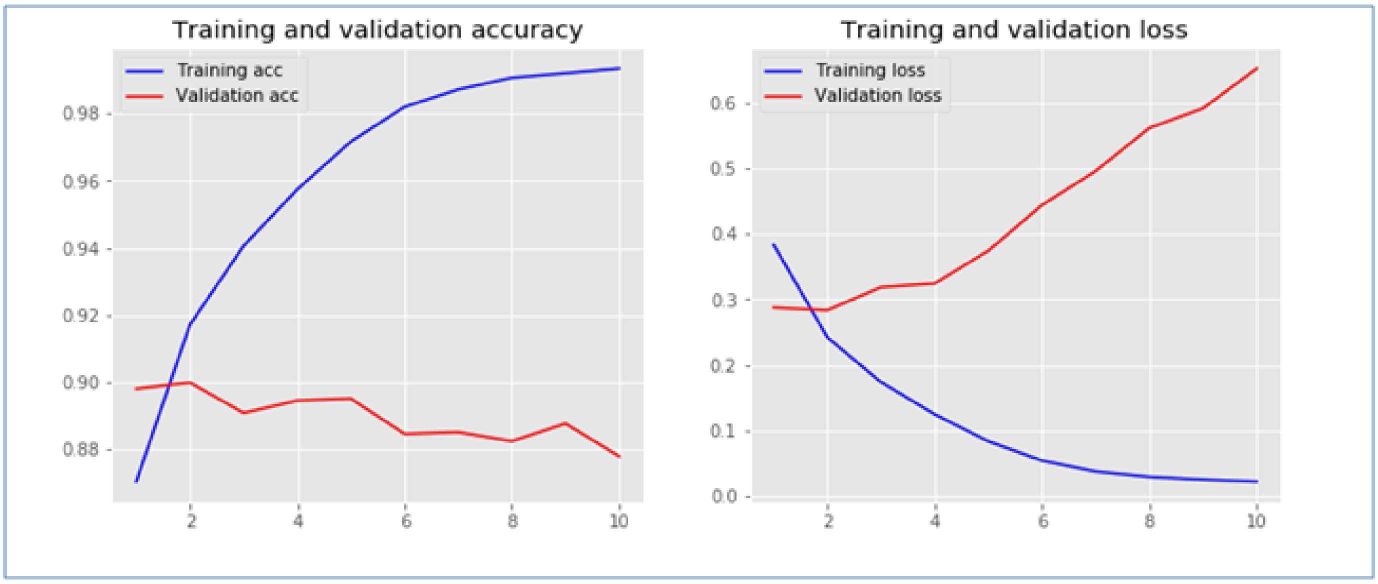
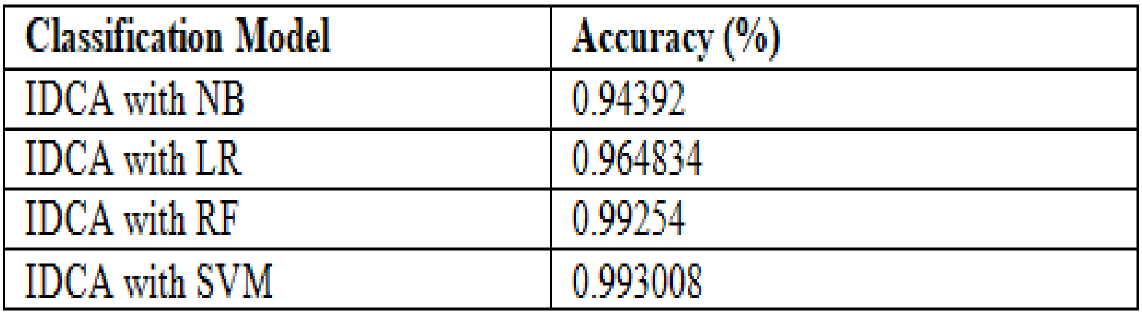


Fig 7 Document classification performance of SVM

As presented in Figure 7, our methodology when used with SVM has given its accuracy and also loss function values.

Table 1 Performance Comparison
As presented in Table 1, the IDCA algorithm when used with different classification models showed its accuracy.



Classification Model	Accuracy (%)
IDCA with NB	0.94392
IDCA with LR	0.964834
IDCA with RF	0.99254
IDCA with SVM	0.993008

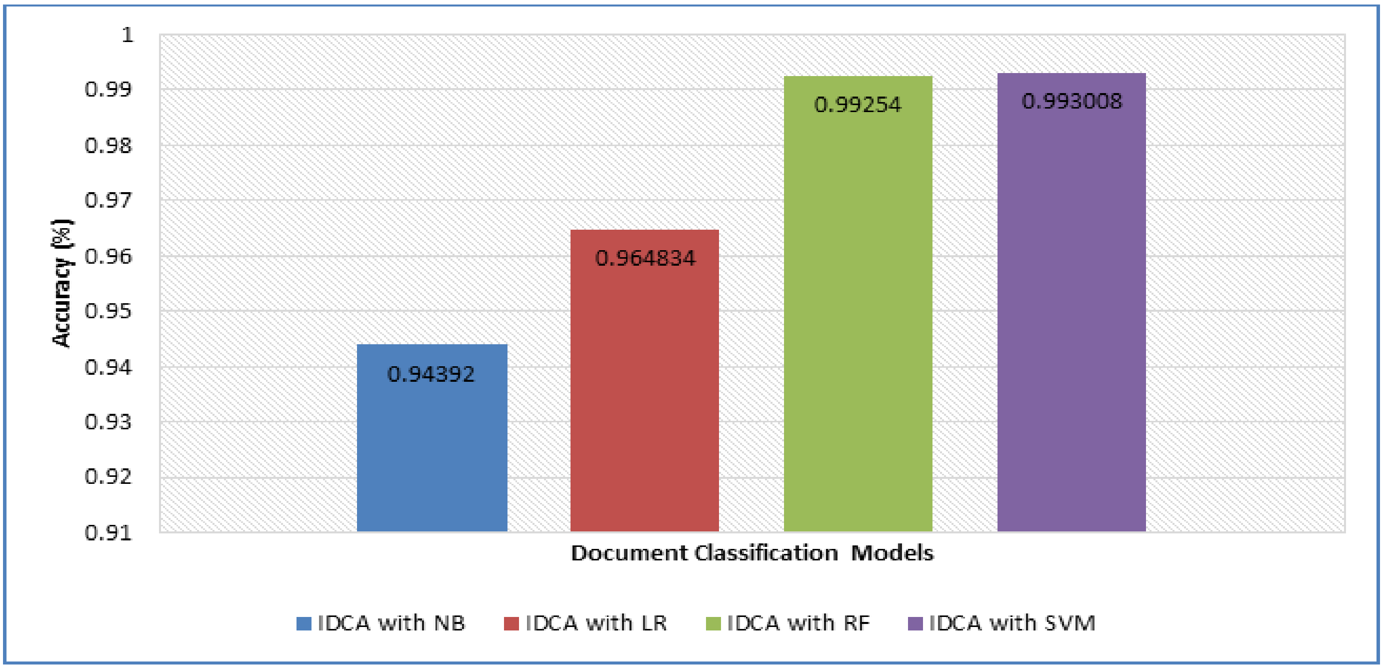


Fig 8 IDCA Performance With Different ML Models

IDCA is an efficient algorithm that has feature selectin and also exploits multiple ML models. When this algorithm is used, instead of using an ML model directly, there is performance improvement in the text document classification. When IDCA is used with NB model, the accuracy is recorded at 94.39%. IDCA with LR showed better performance over IDCA with NB with 96.48% accuracy. IDCA with RF and IDCA with SVM showed better performance other the preceding two models. Accuracy of RF with IDCA algorithm is 99.25% while IDCA with SVM outperformed all the models with highest accuracy 99.30%.

V. CONCLUSION AND FUTURE SCOPE

In this paper we proposed a framework that has a hybrid approach including feature selection and also ML models towards leveraging prediction performance. Our framework is named as Learning based Text Document Classification Framework (LbTDCF). We also proposed an algorithm known as Intelligent Document Classification Algorithm (IDCA) to realize our framework. It performs pre-processing using NLP techniques and then split the data into training and test documents. It also takes a pipeline of ML models that exploit feature selection in order to improve classification performance. There is an iterative process to train all the models. Then the testing phase involves classification of each document is classified by each ML model and the classification results are provided. Experimental results have revealed that our framework improves classification results with 99.30% accuracy. In future, we intend to improve our framework with deep learning models.

REFERENCES

[1]. Goodrum, Heath; Roberts, Kirk and Bernstam, Elmer V. (2020). Automatic classification of scanned electronic health record documents. International Journal of Medical Informatics, 144, 104302–.http://doi:10.1016/j.ijmedinf.2020.104302.

[2]. Kim, Narang and Hong, Soongoo (2021). Automatic classification of citizen requests for transportation using deep learning: Case study from Boston city. Information Processing & Management, 58(1), 102410–.http://doi:10.1016/j.ipm. 2020. 102410.

[3]. Behera, B., Kumaravelan, G., & Kumar.B, P. (2019). Performance Evaluation of Deep Learning Algorithms in Biomedical Document Classification. 2019 11th International Conference on Advanced Computing(ICoAC). http://doi:10.1109/icoac48765.2019.246843.

[4]. Fang, Weili; Luo, Hanbin; Xu, Shuangjie; Love, Peter E.D.; Lu, Zhenchuan and Ye, Cheng (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach. Advanced Engineering Informatics, 44, 101060 http://doi:10.1016/j.aei.2020.101060.

[5]. Varghese, Arun; Agyeman-Badu, George and Cawley, Michelle (2020). Deep learning in automated text classification: a case study using toxicological abstracts. Environment Systems and Decisions. http://doi:10.1007/s10669-020-09763-2.

[6]. Hartmann, Jochen; Huppertz, Juliana; Schamp, Christina; Heitmann, Mark (2018). Comparing automated text classification methods. International Journal of Research in Marketing, S0167811618300545http://doi:10.1016/j.ijresmar.2018.09.009.

[7]. Kadhim and Ammar Ismael (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review. http://doi:10.1007/s10462-018-09677-1.

[8]. Aristodemou, Leonidas and Tietze, Frank (2018). The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. World Patent Information, 55, 37–51. http://doi:10.1016/j.wpi.2018.07.002.

[9]. Neha Sharma;Reecha Sharma and Neeru Jindal; (2021). Machine Learning and Deep Learning Applications-A Vision . Global Transitions Proceedings. http://doi:10.1016/j.gltp.2021.01.004.

[10]. Dhar, Ankita; Mukherjee, Himadri; Dash, Niladri Sekhar and Roy, Kaushik (2020). Automatic categorization of web text documents using fuzzy inference rule. SÄ•dhanÄ•, 45(1), 168–. http://doi:10.1007/s12046-020-01401-6.

[11]. Kumar, Vivek; Recupero, Diego Reforgiato; Riboni, Daniele and Helaoui, Rim (2020). Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes. IEEE Access, 1–1. http://doi:10.1109/access.2020.3043221.

[12]. Zhong, Botao; Pan, Xing; Love, Peter E.D.; Ding, Lieyun and Fang, Weili (2020). Deep learning and network analysis: Classifying and visualizing accident narratives in construction. Automation in Construction, 113, 103089–. http://doi:10.1016/j.autcon.2020.103089.

[13]. Zhong, B., Pan, X., Love, P. E. D., Sun, J., & Tao, C. (2020). Hazard analysis: A deep learning and text mining framework for accident prevention. Advanced Engineering Informatics, 46, 101152. http://doi:10.1016/j.aei.2020.101152.

[14]. Nguyen, Anh Tuan and Nguyen, Tien N. (2017). IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C) - Automatic Categorization with Deep Neural Network for Open-Source Java Projects, 164–166. http://doi:10.1109/ICSE-C.2017.118.

[15]. Jun, Chen; Suhua, Yang; Shaofeng, Jiang (2019). 14th IEEE International Conference on Electronic Measurement & Instruments (ICEMI) - Automatic classification and recognition of complex documents based on Faster RCNN, 573–577. http://doi:10.1109/ICEMI46757.2019.9101847.

[16]. Sundus, Katrina; Al-Haj, Fatima and Hammo, Bassam (2019). 2nd International Conference on new Trends in Computing Sciences (ICTCS) - A Deep Learning Approach for Arabic Text Classification, 1–7. http://doi:10.1109/ICTCS.2019.8923083.

[17]. Agnihotri, Deepak; Verma, Kesari and Tripathi, Priyanka (2017). Variable Global Feature Selection Scheme for automatic classification of text documents. Expert Systems with Applications, 81, 268–281. http://doi:10.1016/j.eswa.2017.03.057.

[18]. Allen-dencid, Hector; Zamora, Juan; Alfaron-faccio, Pedro and Alonso-Sanchez, Marma Francisca (2019). A Machine Learning Approach for the Automatic Classification of Schizophrenic Discourse. IEEE Access, 1–1. http://doi:10.1109/ACCESS.2019.2908620.

[19]. Janani, R. and Vijayarani, S. (2020). Automatic text classification using machine learning and optimization algorithms. Soft Computing. http://doi:10.1007/s00500-020-05209-8.

[20]. AG Corpus of News Articles Dataset. Retrieved from http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html