Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

A Comparative Analysis for Effective Text Document Classification Using Machine Learning Algorithms and Deep Convolution Neural Network

P.Ramya[1],B.Karthik[2]

[1] Assistant Professor, Dept. of CSE, Sona College of Technology, Salem
[2] Associate Professor, Dept. of EEE, Sona College of Technology, Salem
[1]shriramyabe@gmail.com,[2]karthik_pse@yahoo.co.in

ABSTRACT

The enormous amount of text documents keeps on increasing day by day to a greater extent on the web. Almost 80% of the data are available in the form of text on the web. The voluminous of text documents in this digital era requires organizing them consistently which facilitates the information retrieval process. Hence text mining plays a vital role in the process of information retrieval. This paper is focusing on text document classification that has its wider applications in information retrieval, document indexing based on controlled vocabulary, word sense disambiguation, generating hierarchical categorization of web pages, spam detection, email categorization, sentiment analysis, named entity recognition(NER), topic labeling, web search and ranking, document summarization etc. Text document classification belongs to the category of Natural Language Processing tasks where the machine itself automatically categorizes the text documents based on the content to its classes. A lot of manual effort and time is saved by using automatic text document classification. Text document consists of a huge, sparse, non-uniform distribution of features. Mining informative features and performing text classification still exist as a challenging task. This paper contributes techniques involved in text document classification and performs comparative analysis by using machine learning algorithms and deep learning algorithms. The proposed model is experimented with 20-Newsgroups dataset and also evaluated using different performance measures. It has been proven that the proposed model using deep convolution neural network gives superior performance when compared to machine learning algorithms. It gives accuracy 96.3% precision 100%, recall 100% and f1-score99.8%.

Keywords : Machine learning, Deep Convolution Neural Network, Natural Language Processing, Information Retrieval

Introduction

Nowadays the incremental growth of text documents on the web emphasizes the importance of text document classification. These text documents provide valuable information to the user during the searching process. The search engine; tool retrieves precise and reliable information to the user once these text documents are properly categorized. The user may be naïve to the subject of data whilesearching[10][15]. Hence organizing the text documents to their classes helps the user to obtain relevant results. Text classification is widely used in applications such as

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

spam detection, sentiment analysis, fake news detection, etc. Previously the task of text document classification is performed manually. It requires human expertise of the particular domain to classify those documents. Moreover, it is a very tedious and time-consuming process. It highlights the importance of involving the machine in the text classification process.

Regarding the structure of the text document, it is available in the form of semi-structured or in an unstructured format. The sheer size of the text documents in a corpus shows different relationships of the data across the corpus and within that corpus. A text document consists of a huge amount of non-uniform distribution of features. Mostly a document contains nearly 60% of irrelevant and redundant features[13]. The curse of dimensionality is an issue in text document classification. It can be resolved by using preprocessing techniques such as stop word removal and stemming.

The evolution of artificial intelligence particularly its subsetsmachine learning and deep learning algorithms have a greater impact in various applications such as pattern recognition, image processing, language translation, speech recognition, natural language processing, etc. The popularity behind these techniques is that it replicates the functions of the human brain. Human learns things by experience. Likewise, the machine learns things from the labeled data through machine learning algorithm. Deep learning is the subset of machine learning by which it goes a step ahead in constructing the neural network similar to the human brain. Our proposed model utilizes both these techniques and evaluates the performance in terms of accuracy, precision, recall,f1-score, and also loss in text document classification.

Literature Review

Most of the existing system uses feature extraction techniques or feature selection techniques to obtain the top informative features from the documents in a corpus.Feature extraction is a technique that creates new dimensions from the text documents.PCA is a widely used Feature extraction technique that creates principal components that are orthogonal to each other.By this way, it reduces the number of features from text documents by identifying principal components.Feature selection extracts the subsets of features depending on certain weighting mechanism[13].Feature selection is classified into the filter, wrapper, and hybrid method [18][19].The filter method uses certain measures such as information gain, entropy, gain ratio,chi-square function, etc. to select the optimal feature set. Then it performs classification using machine learning algorithms.There are different types of representation models for text documents.Bag of Words(BoW) is the widely used model.It ignores syntactic and semantic representation of the text. These limitations are resolved by using word embedding model.It preserves context of the words in the text document. It representseach word as one-hot code encoding for the entire vocabulary in the text corpus[8][12].The evolution of deep learning algorithms uses the word embedding representational model for extracting the features in the text documents.Deep learning algorithms consist of convolution neural network and also sequence

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

models such as Recurrent Neural Network are used for classification. Almost all the existing system uses the combination of word embedding representational model and different classification models.

Methodology

The text documents in a corpus are preprocessed by using tokenization, stop word removal, and lemmatization. After preprocessing, the traditional BoW model is used as a representational model for the text documents in a corpus. This data representation of the term-document matrix is known as Vector Space Model. Feature weighting is important to identify the significance of the feature to the text document on a corpus.TF-IDF is the weighting measure that widely used to perform the task. After computing the feature weight for each document, average feature weight across the documents in a corpus is determined. Data partitioning is performed to split the dataset into training data and test data. The proposed model uses two different techniques to compare the performance of the model; they are machine learning classifier model and deep convolution neural network. The first technique explores and learns the characteristics of the data and classifies the data by using different classification algorithms such as linear support vector machine, naïve Bayes, and logistic regression. Secondly the technique; Deep Convolution Neural Network (DCNN) is used in which term weighting is given as an input. It uses the filter to slide over the distribution of the word vector[18]. The word vector and the distribution of words in a corpusbased on tfidf values are preserved as horizontal and vertical spatial information respectively.The performance of two techniques machine learning and deep convolution neural network is validated by using the given dataset. On experimental analysis,the Deep Convolution Neural Network outperforms the traditional machine learning algorithms in terms of performance measures such as accuracy, precision, recall, f1-score, and loss.

Proposed Model

The proposed model consists of modules such as Data preprocessing, Bag of Word(BoW) representation model,Term Weighting, Data Partitioning, and also classification models. It covers steps involved in performing text document classification and used two different techniques to evaluate their results based on performance measures.

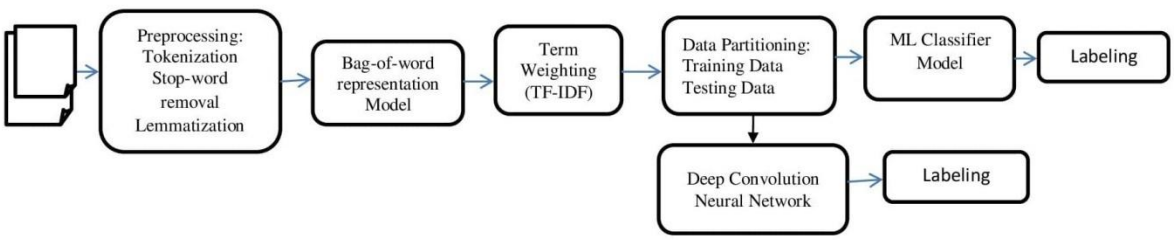

Figure.1. Block diagram of the proposed model

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

A. Preprocessing Techniques

Tokenization

Tokenization is the process that converts the document into tokens. It ignores punctuation, white spaces, and numerals in the text documents. The token is a character sequence that represents the semantic unit for text processing. It is transformed into a term that acts as a semantic identifier that is uniquely identified from IR's system dictionary.

Stop word removal

A typical text document consists of words such as conjunction and verbs which are less informative. Since it is not informative, stop word removal helps to eliminate these insignificant words or features from the text documents. Stop words can be removed from all the documents in a text corpus either by using the NLTK library or a manually prepared list of stop words.

Lemmatization

Almost 60% of the terms or features are removed after performing stop word removal. The remaining features are getting reduced to their base word present in the dictionary by the process called Lemmatization.

B. Traditional BOW representational model

The traditional BOW model is popularly used for texts. The features and the documents in a text corpus are represented in the term-document matrix as rows and columns respectively. In this model, features are considered independent by disregarding the syntactical and semantically relationship that exists between words in a document[2]. As the size of the text corpus grows, the number of features in the text documents also increases. The curse of dimensionality is a very big issue related to the bag of word model.

C. Term weighting

TF-IDF is the weighting scheme widely used for texts which represents the importance of the features in the text corpus. In the term-document matrix, each cell represents TF; term frequency of a document. This data representation is also known as the Vector Space Model in which documents and features of a text corpus are represented as vectors.IDF represents the number of documents that contain the feature.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

	T1	T2	T3 …. .Ti		
D1	w11	w12	w13	w1i	c1
D2	w21	w22	w23	w2i	c2
Dj	wj1	wj2	wj3	wji	ck

$$w_{j,i}=tf(i,j)*idf(j,i)=tfij*\log(N/dfi) \quad (1)$$

Ti is the number of features in a corpus. Nis the total number of documents in a corpus. Ck is the number of classes in a corpus. Dj is the number of documents in a corpus[20]. The features with high TF-IDF values are considered more informative features to the text corpus [4][5]. The features with low TF-IDF values are considered least informative to the text corpus.

D. Data Partitioning

Data are partitioned into three datasets. They are training dataset, validation dataset, and test dataset. The classifier model learns a characteristic of data from the training dataset. The validated dataset evaluates the classifier model how well it generalizes the training dataset. The classifier predicts the class of unseen data from the test dataset. We used the stratified k fold cross-validation method to avoid over-fitting, selection biasing, and improvising accuracy[1]. The stratified k-fold cross-validation method picks data evenly from all classes[7]. To make it robust, data is shuffled each time before splitting into batches. It is tested and validatedK times uniquely, on a different part of the data each time. Though it takes more time for generalization, itshows performance improvement when compared with the traditional method of data partitioning that splits the data into 80% for training data and 20% for testing data.

E. Text Document Classification

Two different techniques are used in our proposed model. They are,

I. Machine learning classifier model
II. Deep Convolution Neural Network classifier model

I. Machine Learning Classifier Model

The proposed model is experimented with different machine learning (ML) algorithms for text document classification. Machine learning algorithm is broadly classified into three types. They are supervised learning, unsupervised learning, and reinforcement learning. The proposed work belongs to the type of supervised learning where the class label is known. As the name implies, ML algorithms learn the characteristics of data from the corpus. It categorizes the text document appropriately to its classes based on the features present in it. More the labeled data is provided as an input to the system, improvises the results in performance.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

i) Naïve Bayes Algorithm

It is the probabilistic model that classifies the text documents into its classes based on the frequency of words in the document.Bayes theory assumes that each feature is independent and has equal contribution to the outcome.According to the Naïve Bayes theorem, the posterior probability tells the class of document given the features.Itisdefinedas the product of apriori probability which isbased on the frequency of the class during experimentation and the class conditional probability factor based on the probability a document belongs to a class depending on its features.It classifies a document to class for which it has maximum aposterior probability (MAP)[6]. It is given by,

$$P(C|d)=P(C)*P(d|C)/P(d) \qquad (1)$$

P(C/d) is the posterior probability, P(C) is the apriori probability, and P(d|C) is the class conditional probability. Denominator is constant for all classes.

ii) Linear Support Vector Machine

It is a supervised technique which is widely used for classification and regression.It is linear model which draws hyper-plane to separate the text documents of its classes. We could draw as many hyperplanes as possible but finding the optimal hyperplane is important.We find the points close to hyperplane from the different classes.These data points are called support vectors.The distance between the support vectors and hyperplane is called margin.The optimal hyperplane is drawn based on maximizing the marginof the data points from its classes. The linear decision boundary is given by,

$$yi=w^{T}xi+b>=1 \text{ for } yi=1 \qquad (2)$$
$$yi=w^{T}xi+b<=-1 \text{ for } yi=-1$$

Where yi is the class label, wi weight function and xi={x1,x2…xn} is the dataset[6]. The decision boundary can be found by solving minimizing $\frac{1}{2}||w||^{2}$ subjectedto yi $(w^{T}xi+b)>=1$ for all i.This classification algorithm is applicable for both linear non-linear data.

iii) Logistic Regression

It is a linear statistical model widely used for classification. It is used to describe the data and also explains the relationship between the dependent variable and one or more independent variables. It uses the sigmoid function to compute the maximum likelihood of the text documents that belong to the class. The output of the target function Z (dependent variable) is given by,

$$Z = WX + B \qquad (3)$$

Where X is independent variables also known as input variables, B is the bias,and the underlying hypothesis function is sigmoid function and it is given by,

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

$$h\Theta(x) = \text{sigmoid}(Z) \quad (4)$$

$$sig(t)=1/1+e^{-t}$$

If Z goes infinity, then, Y (Predicted) will become 0. If Z goes negative infinity,then Y (Predicted) will become 1.

II. Convolution Neural Network

The success of Deep Convolution Neural Network in Computer Vision and Image analysis initiate us to use it in our proposed work for text classification.It is the subset of machine learning. It simulates the human brain and its functionalities. It is a feed-forward neural network in which information flow is in one direction. The advantage of CNN is that learns the hierarchical structure of data and also handles data of variable length [3]. It consists of layers such as convolution, pooling, and a fully connected layer. The convolution layer is used as a feature extractor. Our proposed model uses a one-dimensional convolution layer. The tfidf vector is given as an input to the convolution layer. The rows of the input matrix represent the distribution of words. The column of the input matrix represents the word vector i.e. number of words in the document.The width of the convolution filter is same as that of input width size. Hence it is identical with convolution filter,only vertical striding is necessary. A fixed size sub-matrix known as receptive field from input matrix produces scalar valuesby adding element-wise product between receptive fields and the convolution filter;. Since the input is of variable size, zero padding is used [3]. It uses a filter of size 64 that captures feature maps ranging from high level to lower level. It uses ReLU as its activation function to speed up the training process. For example, the kth output featuremap is given by,

$$Yk = f(Wk * x) \quad (5)$$

Where convolution filter related to kth feature map is Wk.* represents the inner product of the filter model over the feature maps. f(.) represents non-linear activation function. Drop-out layers are used to avoid selection biasing and overfitting the data.It is also used to regularize the model complexity [1].Then it is followed by the pooling layer. The purpose of the pooling layer is to reduce the computation complexity of the neural network. It reduces the number of features by eliminating irrelevant and redundant features. We used max-pooling layers to maximize the feature map. Finally, the output of the pooling layer is directed to the fully connected dense layer which is used to perform classification. It uses the soft-max function as its activation function to compute the class of the document based on the maximum likelihood function[1].The loss function used in the model is sparse categorical cross entropy .

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

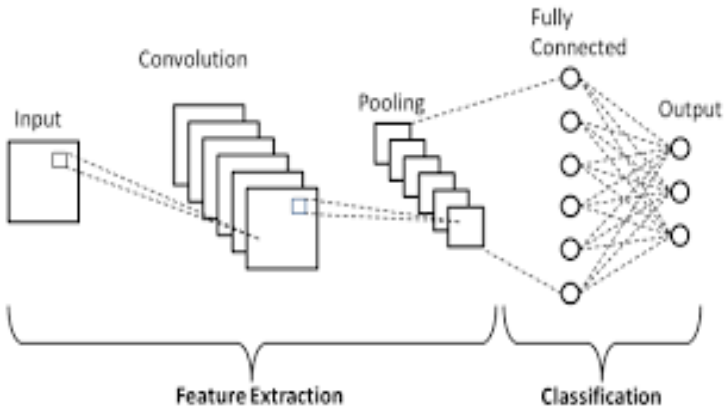

Figure.2.Architecture of Convolution Neural Network

The following is the model structure of our proposed work. The model structure consists of single convolution layer followed by max-pooling layer and dense layer.

Model: "sequential_5"

Layer (type)	Output Shape	Param #
reshape_5 (Reshape)	(None, 16931, 1)	0
conv1d_5 (Conv1D)	(None, 16931, 64)	128
dropout_5 (Dropout)	(None, 16931, 64)	0
max_pooling1d_5 (MaxPooling1	(None, 8465, 64)	0
flatten_5 (Flatten)	(None, 541760)	0
dense_10 (Dense)	(None, 100)	54176100
dense_11 (Dense)	(None, 4)	404

Total params: 54,176,632
Trainable params: 54,176,632
Non-trainable params: 0

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

Experiments

A. DataSet Collection

Our proposed model uses two datasets 20-Newsgroups dataset. Four classes that include electronics, hockey, sales, and politics which consists of 1000 text documents per class from 20-Newsgroup are utilized for experimental analysis.

B. Experimental setup

All the computation work is performed on Intel® core™ i-5-8250 CPU @ GPU 1.60GHz-1.80GHZ 8GB RAM with 64-bit Windows OS. The proposed model is implemented using python programming in the Google colab environment as we used a voluminous dataset.

C. Evaluation measures

Accuracy is defined as the ratio between the number of documents that are correctly classified to their classes and the total number of documents in a corpus

Accuracy=TP+TN/(TP+TN+FP+FN)

Precision is defined as the ratio between no of documents retrieved that are relevant and the total number of documents that are retrieved.

Precision =TP/(TP+FP)

Recall is defined as the ratio between the total number of relevant documents that are retrieved and the total number of documents in the text corpus.

Recall=TP/(TP+FN)

F1score is the harmonic weighted average of precision and recall.

F1-score=2*precision*recall/ (precision+ recall)

These performance measures can be clearly defined by following the contingency table.

	Relevant	Not Relevant
Retrieved	True Positive(TP)	False Positive(FP)
Not retrieved	False Negative(FN)	True Negative(TN)

Results and Discussions

Our proposed model has experimented with machine learning algorithms and the results are shown below.

```
NB Result stratified
precision    recall  f1-score   support
```

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

hockey	0.99	0.98	0.98	100
politics	0.92	0.99	0.95	100
sales	0.96	0.85	0.90	100
electronics	0.91	0.95	0.93	100
accuracy			0.94	400
macroavg	0.94	0.94	0.94	400
weightedavg	0.94	0.94	0.94	400

Linear SVC Result stratified

precision	recall	f1-score	support
hockey	1.00	0.98	0.99	100
politics	0.99	0.99	0.99	100
sales	0.91	0.94	0.93	100
electronics	0.93	0.92	0.92	100
accuracy			0.96	400
macroavg	0.96	0.96	0.96	400
weightedavg	0.96	0.96	0.96	400

Logistic RegressionResult stratified

precision	recall	f1-score	support
hockey	1.00	0.97	0.98	100
politics	0.96	0.98	0.97	100
sales	0.88	0.92	0.90	100
electronics	0.93	0.90	0.91	100
accuracy			0.94	400
macroavg	0.94	0.94	0.94	400
weightedavg	0.94	0.94	0.94	400

It is revealed that the three classifier models by machine learning algorithms provide with minimal variation ranges 1% to 2% in terms of accuracy,precision,recall, and F1-score .

We further want to experiment the proposed model by implementing the deep convolution neural network to perform comparative analysis in text document classification. The results of the proposed model by using a deep convolution neural network and its corresponding plots are

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

shown below. From the experimental analysis, it is observed that the deep convolution neural network gives superior results in terms of accuracy, precision, recall, and F1-score

```
Epoch 1/20
113/113 [==============================] - 12s 96ms/step - loss:
0.8890 - accuracy: 0.7047 - precision: 0.6444 - recall: 0.3555 -
f1_score: 0.4109 - val_loss: 0.1874 - val_accuracy: 0.9425 -
val_precision: 1.0000 - val_recall: 0.9615 - val_f1_score:
0.9800
Epoch 2/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0815 - accuracy: 0.9854 - precision: 1.0000 - recall: 0.9888 -
f1_score: 0.9942 - val_loss: 0.1484 - val_accuracy: 0.9525 -
val_precision: 1.0000 - val_recall: 0.9784 - val_f1_score:
0.9889
Epoch 3/20
113/113 [==============================] - 10s 91ms/step - loss:
0.0275 - accuracy: 0.9942 - precision: 1.0000 - recall: 0.9943 -
f1_score: 0.9971 - val_loss: 0.1327 - val_accuracy: 0.9550 -
val_precision: 1.0000 - val_recall: 0.9880 - val_f1_score:
0.9939
Epoch 4/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0159 - accuracy: 0.9961 - precision: 1.0000 - recall: 0.9948 -
f1_score: 0.9974 - val_loss: 0.1372 - val_accuracy: 0.9600 -
val_precision: 1.0000 - val_recall: 0.9904 - val_f1_score:
0.9951
Epoch 5/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0142 - accuracy: 0.9968 - precision: 1.0000 - recall: 0.9982 -
f1_score: 0.9991 - val_loss: 0.1480 - val_accuracy: 0.9550 -
val_precision: 1.0000 - val_recall: 0.9880 - val_f1_score:
0.9939
Epoch 6/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0090 - accuracy: 0.9984 - precision: 1.0000 - recall: 0.9997 -
f1_score: 0.9999 - val_loss: 0.1374 - val_accuracy: 0.9650 -
val_precision: 1.0000 - val_recall: 0.9904 - val_f1_score:
0.9951
Epoch 7/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0107 - accuracy: 0.9974 - precision: 1.0000 - recall: 0.9975 -
f1_score: 0.9987 - val_loss: 0.1504 - val_accuracy: 0.9625 -
val_precision: 1.0000 - val_recall: 0.9880 - val_f1_score:
0.9939
Epoch 8/20
```

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

```
113/113 [==============================] - 10s 91ms/step - loss:
0.0106 - accuracy: 0.9989 - precision: 1.0000 - recall: 0.9962 -
f1_score: 0.9981 - val_loss: 0.1436 - val_accuracy: 0.9575 -
val_precision: 1.0000 - val_recall: 0.9880 - val_f1_score:
0.9939
Epoch 9/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0095 - accuracy: 0.9976 - precision: 1.0000 - recall: 0.9992 -
f1_score: 0.9996 - val_loss: 0.1636 - val_accuracy: 0.9475 -
val_precision: 1.0000 - val_recall: 0.9976 - val_f1_score:
0.9988
Epoch 10/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0073 - accuracy: 0.9979 - precision: 1.0000 - recall: 0.9998 -
f1_score: 0.9999 - val_loss: 0.1394 - val_accuracy: 0.9600 -
val_precision: 1.0000 - val_recall: 0.9976 - val_f1_score:
0.9988
Epoch 11/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0176 - accuracy: 0.9932 - precision: 1.0000 - recall: 0.9973 -
f1_score: 0.9986 - val_loss: 0.1918 - val_accuracy: 0.9500 -
val_precision: 1.0000 - val_recall: 0.9952 - val_f1_score:
0.9976
Epoch 12/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0062 - accuracy: 0.9986 - precision: 1.0000 - recall: 0.9996 -
f1_score: 0.9998 - val_loss: 0.1752 - val_accuracy: 0.9600 -
val_precision: 1.0000 - val_recall: 0.9928 - val_f1_score:
0.9963
Epoch 13/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0115 - accuracy: 0.9960 - precision: 1.0000 - recall: 0.9991 -
f1_score: 0.9995 - val_loss: 0.1476 - val_accuracy: 0.9600 -
val_precision: 1.0000 - val_recall: 0.9928 - val_f1_score:
0.9963
Epoch 14/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0093 - accuracy: 0.9967 - precision: 1.0000 - recall: 0.9981 -
f1_score: 0.9991 - val_loss: 0.1600 - val_accuracy: 0.9550 -
val_precision: 1.0000 - val_recall: 0.9904 - val_f1_score:
0.9951
Epoch 15/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0070 - accuracy: 0.9984 - precision: 1.0000 - recall: 0.9971 -
f1_score: 0.9985 - val_loss: 0.1698 - val_accuracy: 0.9550 -
val_precision: 1.0000 - val_recall: 0.9952 - val_f1_score:
0.9976
```

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

```
Epoch 16/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0034 - accuracy: 0.9994 - precision: 1.0000 - recall: 0.9979 -
f1_score: 0.9989 - val_loss: 0.1610 - val_accuracy: 0.9575 -
val_precision: 1.0000 - val_recall: 0.9928 - val_f1_score:
0.9963
Epoch 17/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0070 - accuracy: 0.9964 - precision: 1.0000 - recall: 0.9981 -
f1_score: 0.9991 - val_loss: 0.1654 - val_accuracy: 0.9525 -
val_precision: 1.0000 - val_recall: 0.9952 - val_f1_score:
0.9976
Epoch 18/20
113/113 [==============================] - 10s 91ms/step - loss:
0.0074 - accuracy: 0.9975 - precision: 1.0000 - recall: 0.9961 -
f1_score: 0.9980 - val_loss: 0.1671 - val_accuracy: 0.9525 -
val_precision: 1.0000 - val_recall: 0.9952 - val_f1_score:
0.9976
Epoch 19/20
113/113 [==============================] - 10s 92ms/step - loss:
0.0064 - accuracy: 0.9970 - precision: 1.0000 - recall: 0.9980 -
f1_score: 0.9990 - val_loss: 0.1664 - val_accuracy: 0.9550 -
val_precision: 1.0000 - val_recall: 0.9952 - val_f1_score:
0.9976
Epoch 20/20
113/113 [==============================] - 10s 91ms/step - loss:
0.0040 - accuracy: 0.9987 - precision: 1.0000 - recall: 0.9989 -
f1_score: 0.9994 - val_loss: 0.1745 - val_accuracy: 0.9525 -
val_precision: 1.0000 - val_recall: 1.0000 - val_f1_score:
1.0000
```

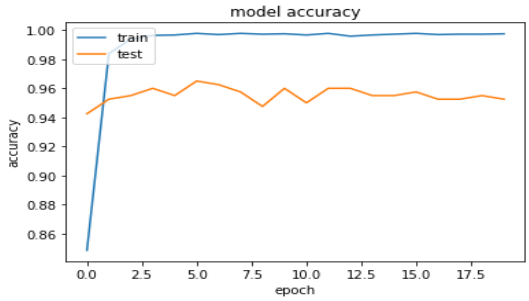

Figure.3. Plot Accuracy Vs. no of Epochs by CNN model

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

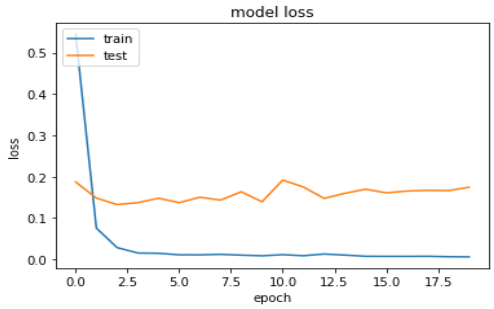

Figure.4. Plot loss Vs. no of Epochs by CNN model

Table 1:Performance measures of the proposed model

Performance measures	NB classifier	Linear SVC Classifier	LR Classifier	LR Classifier
Accuracy	94%	96%	94%	96.3%
Precision	94.5%	95.8%	94%	100%
Recall	94.3%	95.8%	94%	100%
F1-Score	94%	95.8%	94%	99.8%

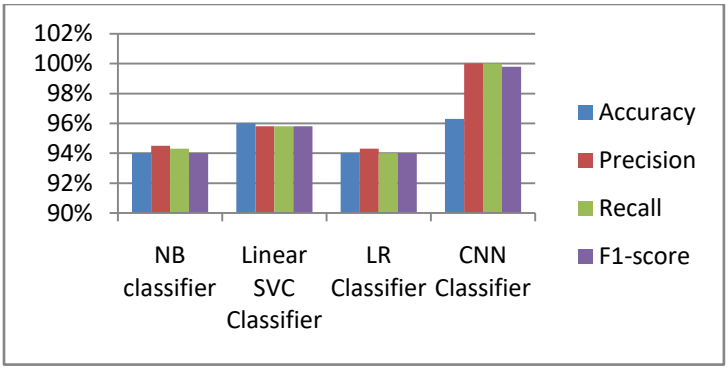

Figure.5. Performance measures of our proposed work

Conclusion

In this digital era,the large volume of text documents isgenerated on the web.Two million tweets are generated per second on social media and other online discussion and community forums[4].The text data generated are in a semi-structured or unstructured format.Mining the knowledge out of it is a critical task.Hence classification plays a vital role in organizing the text

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

data. The features in the text documents are reduced by using preprocessing techniques and only the informative features are identified using tfidf weighting mechanism. In our proposed model, we perform comparative analysis for classifying text documents into their categories by utilizing two techniques;they are machine learning and deep learning.Among them, convolution neural network give superior results in terms of precision 100%,recall 100%,f1-score 99.8% and accuracy 96.3%.

Future Studies

In the future, we could extend our work by using the data at a greater scale and want to explore different representation models of the text documents. Also plan to use sequential model in text classification task.

References

[1] M. Alhawarat , (Member, IEEE), And Ahmad O. Aseeri "A Superior Arabic Text Categorization Deep Model (SATCDM)" IEEE Access Vol.8 Pg.No24653-24661 2020
[2] Christoph Tauchert ,Marco Bender,Neda Mesbah, "Towards an Integrative Approach for Automated Literature Reviews Using Machine Learning" Proceedings of the 53rd Hawaii International Conference on System Sciences | 2020
[3] Seungwan Seo1, Czangyeob Kim1, Haedong Kim2, Kyounghyun Mo3, and Pilsung Kang 1 ,"Comparative Study of Deep Learning-Based Sentiment Classification" IEEE Access Vol.8,Pg.No6861-6875
[4] Jiun-Yu Wu , Yi-Cheng Hsiao and Mei-Wen Nian," Using supervised machine learning on large-scale online forums to classify course-related Facebook messages in predicting learning achievement within the personal learning environment" Taylor& Franchis Group
[5] Alejandro Moreo , Andrea Esuli, and Fabrizio Sebastiani" Learning to Weight for Text Classification"IEEE Transaction on Knowledge and Data Engineering Vol.32.Pg.No302-316
[6] Qi-na Li1, Ting-hui Li1," Research on the application of Naive Bayes and Support Vector Machine algorithm on exercises Classification" International symposium on Big data and Applied Analytics
[7] Rungroj Maipradit, Hideki HAta, Kenichi Matsumota, "Sentiment Classification using N-gram IDF and Automated Machine Learning"Pg.No1-4
[8] Roger Allen Stein,Patricia, A.Jaques,Joya Franchisco Valiati," An Analysis of Hierarchical Text Classification Using Word Embeddings" 2018
[9] Syed Muzamil Basha1, K. Bagyalakshmi2, C. Ramesh3, Robbi Rahim4, R. Manikandan5 and Ambeshwar Kumar5,"Comparative Study on Performance of Document Classification Using Supervised Machine Learning Algorithms: KNIME" 2019
[10] Christopher D.Manning ,PrabhakarRagavan,HinrichSchutze,"An Introduction to Information Retrieval" Cambridge University Press England

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 16071 - 16086
Received 05 March 2021; Accepted 01 April 2021.

[11] MowafyM,Rezk A and El-bakryHM,"An Efficient Classification Model for Unstructured Text Document"American Journal of Computer Science and Information Technology ISSN 2349-3917,Feb 20 ,2018

[12] EunjeongL.Park, Sungzooncho and Pilsungkang,"Supervised Paragraph Vector:Distributed Representations of Words ,Documents and Class Labels"IEEE Transaction Vol 7 Feb 27,2019

[13] Laith Mohammad Qasim Abualigah,"Feature Selection And Enhanced Krill Herd Algorithm For Text Document Clustering"

[14] Mahdi Abdollahi1 , Xiaoying Gao1 , Yi Mei1 , Shameek Ghosh2 , Jinyan Li3,"An Ontology-based Two-Stage Approach to Medical Text Classification with Feature Selection by Particle Swarm Optimisation"

[15] Muhammad Pervez Akhter 1 , Zheng Jiangbin 1 , Irfan Raza Naqvimohammed Abdelmajeed 2 , Atif Mehmood 3 , And Muhammad Tariq Sadiq," Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network" IEEE Access 2020

[16] Laith Mohammad Qasim Abualigah Al-abayt University, Mafraq, Jordan Essam S. Hanandeh Zarqa University,Zarqa, Jordan," Applying Genetic Algorithms To Information Retrieval Using Vector Space Model",IJCSEA Feb2015

[17] Said Bahassine a , Abdellah Madani b , Mohammed Al-Sarem c , Mohamed Kissi," Feature selection using an improved Chi-square for Arabic text classification"

[18] Rajeswari C., Sathiyabhama B**.,** Devendiran S., Manivannan K. Bearing fault diagnosis using wavelet packet transform, hybrid PSO and support vector machine,Procedia Engineering, ,Vol.97 (1) PP:1772-1783,2014

[19] Rajeswari C., Sathiyabhama B., Devendiran S., Manivannan K. A Gear fault identification using wavelet transform, rough set based GA, ANN and C4.5 algorithm. Procedia Engineering, Vol – 2 PP: 338-344 , DOI : 10.1016/j.procs.2010.11.044,2014

[20] Laith Mohammad Abualigah1 · AhamadTajudin Khader1 · Essam Said Hanandeh2," Hybrid clustering analysis using improved krill herd algorithm"Springer 2018

[21] A.M. Barani, R.Latha, R.Manikandan, "Implementation of Artificial Fish Swarm Optimization for Cardiovascular Heart Disease" International Journal of Recent Technology and Engineering (IJRTE), Vol. 08, No. 4S5, 134-136, 2019.

[22] Asraf Yasmin, B., Latha, R., & Manikandan, R. (2019). Implementation of Affective Knowledge for any Geo Location Based on Emotional Intelligence using GPS. International Journal of Innovative Technology and Exploring Engineering, 8(11S), 764–769. https://doi.org/10.35940/ijitee.k1134.09811s19