Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HYBRID APPROACH FOR TEXT CLASSIFICATION

M.KAVITHA[1] , Dr.P.PRABHAVATHY[2]

[1]Research Scholar, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai-26, India.
[2]Assistant Professor (S.G), Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai-26, India.
E-mail: [1]km2693@srmist.edu.in, [2]prabhavp1@srmist.edu.in

ABSTRACT

On the internet, a huge amount of text data is collected, and segregating it based on a particular category is a crucial task. The data collected can be structured or unstructured. In the proposed method, machine learning algorithms and ensemble technique is used to handle the unstructured text data for classifying the text. The paper aims to evaluate the performance of the machine and deep ensemble classifiers. Ensemble classifiers provide solutions to numerous problems. There are various methods for the ensemble. General ensemble techniques are bagging, boosting, and stacking. In this paper, the bagging and boosting techniques are used to evaluate the performance of the models. Different voting schemes are available in the bagging method. The ensemble learner used in this paper makes predictions based on voting techniques.

Keywords: *Bagging, Boosting, Ensemble, Machine Learning, Deep Learning*

1. INTRODUCTION

One of the common unstructured forms of data is text. Classifying the text helps to quickly retrieve the data from the internet. It can be done easily with the help of natural language processing tools. Text categorization [1] is used in several applications. In an organization, text categorization [2] plays a significant role to resolve end-user issues. A manual approach to categorizing customer support tickets is a time-consuming and error-prone process. To overcome this problem, ML and DL techniques are used to automatically categorize the text. Text tagging is used to understand customer insights. It is used for both binary classification problems like spam mail classification and multi-class classification problems. Understanding the subtle meaning of text and tagging is a complicated task. Ensemble methods are widely used in classical machine learning to improve the robustness of the model. Basic ensemble methods are averaging and voting. Advanced ensemble techniques are bagging, boosting, and stacking. An algorithm that uses the bagging technique is random forest and the algorithms using boosting technique are Adaboost, XGBoost, Gradient Boosting Machine (GBM). Ensemble techniques are extensively used in many research works. It constructs multiple models and combines their outputs to predict the results. The ensemble is used to obtain a collective output from diverse models. It is used to attain a reliable model. Deep learning has shown good results when compared to traditional machine learning models. To improve further we can ensemble deep learning models.

The scope of the paper is to analyze the performance of ensemble machine learning and deep learning models for binary and multi-class text classification problems. Two datasets are used for muti-class text classification and one dataset is applied for binary classification problems. This study uses bagging and boosting techniques to ensemble the models. Accuracy and loss metrics are used to evaluate the performance of the models.

In this paper ensemble method using ML and DL techniques is applied for the text classification. This paper is organized as follows: Section 2 deals with different literature surveys on this problem. Section 3 gives a brief overview of the datasets used for text classification and the methodology implemented. Section 4 describes the machine learning classifiers used in this study and section 5 explains various ensemble methods. All the obtained results are compared and discussed in section 6. Finally, section 7 concludes the whole work.

2. BACKGROUND

Zhang, Yuebing, et al. [3] suggest that traditional ensemble methods used weak base classifiers and

not strong base classifiers. The author aims to use strong base classifiers. The proposed technique outperforms the accuracy and cost of other ensemble approaches such as weighted voting, majority voting, and meta-learning.

Liang, Decui, and Bochun Yi [4] propose a technique to avoid misclassification in classifying policy text through two-stage and (3WD) 3-way decision ensemble technique. Classifying policy text is helpful for medium-sized enterprises. CNN is used as a base classifier and later AdaBoost and bagging methods are employed to classify policy text.

Khai Tran, Thien, and Tuoi Thi Phan [5] handle various feature types such as sentiment shifting, language features, and statistical techniques using the classifier ensemble method. The paper uses real-time datasets from social media, blog posts, and product reviews.

Anand, Manish, et al [6] used fuzzy-based CNN for feature selection. HASOC 2020, offensive language identification dataset (OLID), and CAALDYC dataset are used for implementation. The online dataset is taken from Twitter, Youtube, and Facebook. Different metrics are used in this paper to assess the performance of classifiers. Ensemble architecture with Bi-LSTM, SVM, and naïve Bayes algorithms are used to handle Multilingual Text Classification (MTC) in this paper.

Roy et al [7] focussed on text posted on social media to identify hate speech and aggressive language. The performance of ML models, DL models, transformer models, and ensembled transformers with DL models are discussed in this research. It uses an ensemble technique to handle a code-mixed dataset that contains text from two or more languages.

Lin et al [8] identify harmful news and fake news in their research. The correlation between harmful news and text sentiment is analyzed. Transformer based ensemble technique is used in this study.

Sharif et al [9] used ML models, DL models, transformers models, and ensemble methods for implementation. The paper uses weighted ensemble techniques to identify aggressive text in social media and categorize it.

J Briskilal and C.N. Subalalitha [10] proposed a hybrid model using BERT and RoBERTa to classify idioms and literal texts. TroFi dataset is used for this purpose. The Fscore and accuracy of the ensemble model are greater than BERT model and RoBERTa model.

Many researchers are working on text classification problems using different ensemble methods for improving the performance of the models. Table 1 shows the summary of previous works using ensemble techniques in machine learning, deep learning, and transformer-based approaches. Most of the previous research works used datasets with binary labels. This research paper uses both binary and multiple-label datasets and the work focuses on ensemble methods using machine and deep learning techniques.

3. PROPOSED METHODOLOGY

Combining several different models to create an ensemble classifier. The output from each model is combined to produce ensemble classifier output. For classification problems, the document is labeled using voting methods. In the case of regression, the ensemble learner finds the mean value to predict the result. Then we can test the model with the test data that is kept aside. As each kind of learner has a sort of bias, they are put together to reduce the overall bias. Thus, ensemble learner [11] reduces error and leads to less overfitting.

In this article, ML algorithms are fused to handle multi-class classification problems. Later the performance of the individual classifier and hybrid classifiers are evaluated to classify the news articles [12]. Ensemble learning takes opinions from multiple classifiers to make predictions. It is used to solve myriad problems. This study also analyses the performance of the deep ensemble model using the IMDB dataset. It comes under two categories.

3.1 Dataset collection

In this study, two datasets are used for the machine learning ensemble process. BBC news dataset and AGNews dataset are taken for ML ensemble implementation. BBC dataset contains 2225 samples with five categories. The first 5 articles of the BBC dataset are shown in figure 1. The dataset is partitioned as training and testing sets. Figure 2 shows the counts of each category and the train-test split of the BBC dataset. The ensemble is made on this dataset using hard voting, soft voting, and voting with weights. The second dataset is AGNEWS containing one million news articles with train and test sets. Figure 3 presents the number of features extracted using TFIDF [9] for training and test samples. It also gives the shape of training and test data. Hard voting is tried for this dataset. The third dataset is the IMDB dataset which contains 50000 positive and negative reviews on movies. The first five rows of the IMDB dataset are shown in figure 4. The paper assesses the performance of traditional machine learning algorithms, ensemble classifiers using bagging and boosting techniques on ML, deep learning models, and ensemble deep learning models for the IMDB dataset.

TABLE 1: Summary Of The Previous Works Using Ensemble Techniques

Reference	Ensemble technique	Metric used	Dataset	Number of classes
Zhang, Yuebing, et al. [3]	Deep Learning ensemble	Accuracy	IMDB, MR, Customer Review, SUBJ MPQA	Binary labels
Liang, Decui, and Bochun Yi [4]	Ensemble CNN model	Accuracy	Policy text	Binary labels
Khai Tran, Thien, and Tuoi Thi Phan [5]	Machine learning ensemble, Deep learning ensemble	Accuracy	Social media, Blog posts, Product reviews	Binary labels
Anand, Manish, et al [6]	BiLSTM+NB+SVM	Accuracy, Precision, Recall, F1-score, RMSE	HASOC 2020, OLID, CAALDYC	Hierarchical labels
Roy et al [7]	Ensemble model using deep learning and transformers	Precision, Recall, F1-score	Hate speech and offensive language from Social media	Binary labels
Lin et al [8]	Transformer based ensemble technique	Precision, Recall, F1-score, Accuracy	Fake and harmful news	Binary labels
Sharif et al [9]	Machine learning ensemble, Deep learning ensemble, Transformer based ensemble	Precision, Recall, F1-score	Aggressive and non-aggressive text from Social media	Binary labels, Multiple labels
J Briskilal and C.N. Subalalitha [10]	Transformer based ensemble technique	F1-Score, Accuracy	Trofi dataset	Binary labels

	category	filename	title	content
1	business	001.txt	Ad sales boost Time Warner profit	Quarterly profits at US media giant TimeWarne...
2	business	002.txt	Dollar gains on Greenspan speech	The dollar has hit its highest level against ...
3	business	003.txt	Yukos unit buyer faces loan claim	The owners of embattled Russian oil giant Yuk...
4	business	004.txt	High fuel prices hit BA's profits	British Airways has blamed high fuel prices f...
5	business	005.txt	Pernod takeover talk lifts Domecq	Shares in UK drinks and food firm Allied Dome...

Figure 1. Description Of BBC Dataset

```
(2225, 4)
sport            511
business         510
politics         417
tech             401
entertainment    386
Name: category, dtype: int64
(1557,)
(668,)
```

Figure 2: Counts of each category and train-test split in BBC dataset

```
X_train.shape :  (120000, 43679)
X_valid.shape :  (6619, 43679)
y_train.shape :  (120000,)
y_valid.shape :  (6619,)
```

Figure. 3: Train And Test Split Of Agnews Dataset

	review	sentiment
1	One of the other reviewers has mentioned that ...	positive
2	A wonderful little production. The...	positive
3	I thought this was a wonderful way to spend ti...	positive
4	Basically there's a family where a little boy ...	negative
5	Petter Mattei's "Love in the Time of Money" is...	positive

Figure 4: First 5 Rows Of Imdb Dataset

3.2 Preprocessing

The pre-processing [9] comprises scaling, normalization, and binarization methods. After data gathering, it is necessary to pre-process the text content in order to retrieve the useful features. Pre-process consists of various subprocesses like tokenization, stop word removal, stemming, lemmatization, named entity recognition, and POS tagging. Tokenization is to extract tokens from the text. Stop words are common words from the text which should be removed from the text as it increases the sparsity in the matrix. Stemming is to find out the root word of the tokens without dictionary meaning. Lemmatization is to identify the base form of the word with dictionary meaning. Named entity recognition is recognizing the fundamental entities in the text such as person name, organization name, location, date, time, etc. POS tagging is used to describe the tokens with parts of speech tags like nouns, verbs, adverbs, adjectives, etc. The token count before removing unwanted characters is computed for the IMDB dataset and it is shown in figure 5. The token count is reduced after removing unwanted characters and is portrayed in figure 6. The sentiment labels are replaced with numeric values using a label encoder and are depicted in figure 7. Figure 8 represents the shape of the IMDB dataset, the number of samples in each category, and the training and test data split. Once pre-processing is completed, features are extracted using feature extraction techniques [13].

	review	sentiment	token_count
1	One of the other reviewers has mentioned that ...	positive	307
2	A wonderful little production. The...	positive	162
3	I thought this was a wonderful way to spend ti...	positive	166
4	Basically there's a family where a little boy ...	negative	138
5	Petter Mattei's "Love in the Time of Money" is...	positive	230
...	...	...	...
49996	I thought this movie did a down right good job...	positive	194
49997	Bad plot, bad dialogue, bad acting, idiotic di...	negative	112
49998	I am a Catholic taught in parochial elementary...	negative	230
49999	I'm going to have to disagree with the previou...	negative	212
50000	No one expects the Star Trek movies to be high...	negative	129

50000 rows × 3 columns

Figure 5: Token Count Before Removing Unwanted Characters

	review	sentiment	token_count
1	one of the other reviewers has mentioned that ...	positive	300
2	a wonderful little production the filming tech...	positive	156
3	i thought this was a wonderful way to spend ti...	positive	161
4	basically theres a family where a little boy j...	negative	127
5	petter matteis love in the time of money is a ...	positive	222
...	...	...	...
49996	i thought this movie did a down right good job...	positive	188
49997	bad plot bad dialogue bad acting idiotic direc...	negative	108
49998	i am a catholic taught in parochial elementary...	negative	225
49999	im going to have to disagree with the previous...	negative	211
50000	no one expects the star trek movies to be high...	negative	124

50000 rows × 3 columns

Figure 6: Token Count After Removing Unwanted Characters

	review	token_count	sentiment
1	one of the other reviewers has mentioned that ...	300	1
2	a wonderful little production the filming tech...	156	1
3	i thought this was a wonderful way to spend ti...	161	1
4	basically theres a family where a little boy j...	127	0
5	petter matteis love in the time of money is a ...	222	1
...	...	...	...
49996	i thought this movie did a down right good job...	188	1
49997	bad plot bad dialogue bad acting idiotic direc...	108	0
49998	i am a catholic taught in parochial elementary...	225	0
49999	im going to have to disagree with the previous...	211	0
50000	no one expects the star trek movies to be high...	124	0

50000 rows × 3 columns

Figure 7: Sentiment Labels Are Replaced With Numeric Values Using Label Encoder

```
(50000, 3)
0    25000
1    25000
Name: sentiment, dtype: int64
(35000,)
(15000,)
```

Figure 8: IMDB Shape, Count Of Samples In Each Category, Training And Test Data Split

4. CLASSIFIERS

4.1 Logistic Regression

The logistic regression algorithm uses a sigmoid function to predict the dependent variable from the independent variables. The logistic function shown in equation (1) takes the input 't' as a real value and outputs the value 0 or 1. The standard logistic function is given as

$$\sigma(t) = \frac{1}{1+e^{-t}} \qquad (1)$$

It comes in various types. Linear regression estimates the value of the dependent variable when there is a change in the independent variable whereas logistic regression finds the probability of an event. Hence linear regression algorithm is used for regression problems and logistic regression for classification problems. It is efficient to implement and train logistic regression algorithms. When the samples are more than the features, this technique is recommended to prevent overfitting. Regarding the distribution of classes in feature space, it makes no assumptions.

4.2 Random Forest

It is an ensemble technique with a collection of decision trees. The samples are divided into 'n' subsamples and each subsample is given to a decision tree. The decision trees predict the output of all subsamples. The final output is predicted from the output of all the subsamples. It is a meta-estimator that uses various decision tree classifiers on subsamples to improve accuracy and overfitting. It applies the bagging technique. It is quite similar to k-fold cross-validation. When compared to logistic regression or support vector machines, random forest performs worse.

4.3 Multinominal Naïve Bayes

The multinominal Naïve Bayes (MNB) technique depends on Baye's statement of predictor independence. To simply express, an NB classifier assumes the presence of a feature in a class has no relation with any other features. It is simple to build and specially applied for very large data sets. It is well known that Naive Bayes outperforms even the most sophisticated classification methods.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \qquad (2)$$

Equation (2) represents Bayes theorem which computes P(c|x), the posterior probability of class 'c' given predictor 'x' from the prior likelihood of class P(c), the prior likelihood of predictor P(x), and the probability of predictor given the class P(x|c). The naive Bayes method predicts the probability of different classes based on various attributes. NB method is mainly useful for text classification [14] when a problem contains several classes. The class of the test dataset might be correctly predicted quickly by the Naive Bayes approach. To solve multi-class prediction issues, it is used. The Naive Bayes algorithm outperforms other classifiers with fewer training samples when the independence of the features is assumed. Compared to numerical variables, it does remarkably well with categorical input variables. If a categorical variable is present in the test dataset but not in the training dataset, the NB model will give zero probability. Making predictions is quite difficult in such cases. Zero frequency is a phenomenon that is solved by using a smoothing technique and the model assumes every feature is independent of other features. While it seems fantastic in theory, it is difficult to find a set of independent properties in real-time applications.

This method can be used to produce real-time predictions because it is quick and effective. It is used to quickly determine the probability of several target classes. This algorithm is used to determine whether or not an email is spam. For spam filtering, this algorithm works incredibly well. It is easy to implement sentiment analysis with the assumption of feature independence. To determine whether a target client group has positive or negative feelings sentiment analysis is used.

There are various types of this algorithm. The predictors in Bernoulli Naive Bayes are Boolean variables with the values True and False ('Yes' or 'No'). When samples are drawn from a multivariate Bernoulli distribution, it can be employed. Problems with document classification are resolved with multinomial naive Bayes. This algorithm sorts documents according to the category and it makes use of word frequency as an attribute. If the predictors take continuous value or if the samples are from gaussian distribution, Gaussian Naive Bayes is utilized.

4.4 Support Vector Machine

A hyperplane is the decision boundary to separate the data points. The data points that are near the hyperplane are said to be support vectors. If data points are linearly separable then linear SVM is used otherwise non-linear SVM is used. The hyperparameters C and gamma are used to fine-tune

the SVM. If the value of the C-hyperparameter increases, misclassification will be less as it acts as a penalty. If the gamma value increases, it leads to overfitting. Kernels in SVM can be used to transform high-dimensional data into low-dimensional data. SVM executes admirably when there is a significant gap between classes. It achieves good performance when there are more dimensions than samples, as well as in high-dimensional spaces. The SVM algorithm does not work well for large and noisy data sets.

4.5 MultiLayer Perceptron

If the data points are linearly separable then a single-layer perceptron can be used otherwise it is preferable to use a multilayer perceptron. MLP consists of input and hidden and output layers. To bring nonlinearity into the model activation functions are used. Adam is the default optimizer and relu is the default activation function used in MLP.

5. ENSEMBLE METHODS

The ensemble [15,16] combines the output of various classifiers to produce a final output. It can be done using bagging, boosting, and stacking techniques. Figure 9 depicts the ensemble approach using the voting concept. The original dataset is split into sample datasets. The output of each sample dataset is predicted using base classifiers. The final output of the ensemble model [17] is predicted by the voting technique. The ensemble algorithm is given in Table 2. The bagging algorithm is described in Table 3. Table 4 explains the boosting algorithm.

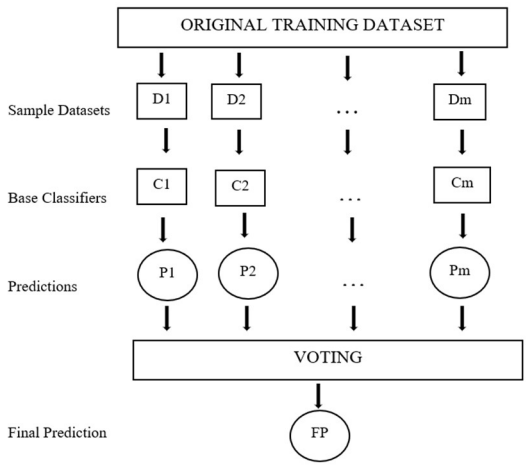

Figure 9. An ensemble approach using voting

Table 2: Ensemble Algorithm

Ensemble algorithm
Let N be the size of the training dataset
Make M samples with the replacement of size N from the original training set
for each sample do
Apply the learning algorithm to each sample.
Predict the class of each classifier
Find the class with the maximum value
return class

5.1. Bagging

Bootstrap aggregating or Bagging classifier [18] is an ensemble classifier constructed using multiple estimators and aggregating the results either by voting or averaging method. When a random dataset is used with replacement, it is known as bagging else it is pasting. An Independent dataset called bootstrap is given for each classifier in the bagging ensemble method. Each classifier predicts the output of the bootstrap sample. Bagging employs an aggregation method to compute the final predictions. For classification problems [19], the voting technique is employed and for regression problems, the average technique is adopted. Hard and soft are the two types of voting. Hard voting [20] predicts the class with the max rule method. The majority voting of multiple models is taken as ensemble classifier output. Soft voting uses a probability averaging technique to compute the probability scores of several models to determine the overall score of the ensemble learner. In the case of the weighted average method, the weighted average of probability is computed. Ensemble learner results are calculated by allotting different weightage to each classifier based on the significance of the classifier.

Table 3: Bagging Algorithm

Bagging algorithm
Input the samples D
Divide the samples into subsamples Dm
For each subsample do
Build the model Gm
Predict the output of Gm
End for
For the regression problem, averaging the output from all the models
For the classification problem, voting is taken from all the models

5.2 Boosting

The bagging algorithm adopts a parallel processing method whereas boosting algorithm works sequentially. In boosting ensemble technique, misclassifications from the previous classifier are fed as input for the subsequent classifier. Then, the final result is taken from all the weak or base classifiers. All the weak classifier combines together to form a strong learner.

Table 4: Boosting Algorithm

Boosting algorithm
Input the samples D Initially, equal weights are assigned to the samples. The entire sample is given to the first classifier Repeat Assigning low/high weights to wrongly classified points. Wrongly classified samples are sent to the next model Reinitialize the weights of the samples. Training the model Computing the error Predicting the output Until all the samples are trained

5.2.1. Adaboost classifier

Adaboost uses decision stumps to predict the original dataset by assigning equal weights to each observation. From the first classifier, it identifies the misclassified observation and gives higher weightage to them. It continues the process in an iterative manner with new classifiers until no more misclassified observations are seen or a particular limit is reached. Adaboost algorithm can handle both regression and classification problems. The performance of the adaboost algorithm can be tuned (optimized) using parameters such as estimators, learning rate, and base estimators. The weak classifiers are controlled by estimators. The contribution of each learner is determined by the learning rate parameter. The various machine algorithms employed are specified in the base estimators.

5.2.2. Gradient Boosting Machine

It sequentially trains many classifiers by minimizing the loss function using the gradient descent method. It constructs a new classifier to provide an accurate estimate. It uses a decision tree as a weak classifier. It optimizes the performance of the model using estimators, maximum depth, and learning rate parameters. There is always a trade-off between estimators and learning rates. The number of node counts in a tree can be restricted using the maximum depth parameter.

5.2.3. XGboost classifier

It stands for extreme gradient boosting. It also uses decision trees as weak learners. It is the extension of the gradient boosting technique. It performs processing at the node level and hence it is quite faster than the gradient boosting method. By using various regularization techniques and hyperparameters, the XGboost method reduces overfitting.

5.3 Stacking

The stacking method works on bootstrapped data similar to the bagging technique. The stacking ensemble technique adopts two levels of training. The first level classifier represents the individual estimator and the second level classifier is said to be meta learner. The output of all classifiers is given to the meta-classifier [21] to predict the final output. The idea behind the meta classifier is to determine if the training data is correctly learned.

6. RESULTS AND DISCUSSION

6.1. Ensemble results using machine learning

The performance of state of art classifiers and ensemble models [22,23] is assessed in this paper. When compared to individual learners, the performance of the ensemble classifier [24] is superior. The training and testing accuracy score is computed for both datasets. The precision, recall, f1score, and support for all the labels are generated using a classification report. It also computes the accuracy, macro average, and weighted average for the metrics. Figure 10 to 14 depicts the performance metrics value of MNB, LR, SVM, MLP, and RF for AGnews dataset. The precision, recall, f1score, and support for all the labels are generated using a classification report. It also computes the accuracy [25], macro average, and weighted average for the metrics.

```
training accuracy Score    :  0.9114166666666667
Validation accuracy Score :  0.8948481643752833
              precision    recall  f1-score   support

           1       0.90      0.91      0.90      1639
           2       0.97      0.95      0.96      1672
           3       0.85      0.86      0.85      1642
           4       0.87      0.86      0.86      1666

    accuracy                           0.89      6619
   macro avg       0.90      0.89      0.89      6619
weighted avg       0.90      0.89      0.90      6619
```

Figure 10. Performance Metrics Of MNB Classifier For Agnews Dataset

```
Training accuracy Score    :  0.93565
Validation accuracy Score :  0.9082943042755703
              precision    recall  f1-score   support

           1       0.90      0.93      0.91      1614
           2       0.98      0.95      0.96      1685
           3       0.87      0.89      0.88      1632
           4       0.89      0.87      0.88      1688

    accuracy                           0.91      6619
   macro avg       0.91      0.91      0.91      6619
weighted avg       0.91      0.91      0.91      6619
```

Figure 11. LR Classifier Performance Results For Agnews Dataset

```
Training accuracy Score    :  0.9636166666666667
Validation accuracy Score :  0.9104094274059525
              precision    recall  f1-score   support

           1       0.90      0.93      0.92      1610
           2       0.97      0.95      0.96      1669
           3       0.87      0.88      0.88      1656
           4       0.90      0.88      0.89      1684

    accuracy                           0.91      6619
   macro avg       0.91      0.91      0.91      6619
weighted avg       0.91      0.91      0.91      6619
```

Figure 12. SVM Classifier Implementation Results On Agnews Dataset

The results of individual classifiers for BBC dataset are shown in figure 15. Logistic regression produces higher accuracy when compared to other classifiers. When standard SVM is used instead of linear SVM, SVM shows better accuracy than other algorithms which is depicted in figure 16. Table 5 represents the test accuracy of all classifiers for BBC and AGNews datasets. Figure 17 shows the bagging output of IMDB dataset and figure 18 represents IMDB boosting output.

```
Training accuracy Score    :  0.962675
Validation accuracy Score :  0.8945460039280858
              precision    recall  f1-score   support

           1       0.89      0.90      0.90      1650
           2       0.95      0.95      0.95      1637
           3       0.85      0.86      0.86      1651
           4       0.88      0.86      0.87      1681

    accuracy                           0.89      6619
   macro avg       0.89      0.89      0.89      6619
weighted avg       0.89      0.89      0.89      6619
```

Figure 13. MLP Classifier Results On Agnews Dataset

```
RF Training accuracy Score    :  0.9985
RF Validation accuracy Score :  0.882610666263786
              precision    recall  f1-score   support

           1       0.88      0.90      0.89      1632
           2       0.97      0.91      0.94      1740
           3       0.84      0.86      0.85      1629
           4       0.85      0.86      0.86      1618

    accuracy                           0.88      6619
   macro avg       0.88      0.88      0.88      6619
weighted avg       0.88      0.88      0.88      6619
```

Figure 14. RF Classifier Performance For Agnews Dataset

	Model	Test accuracy
1	Logistic Regression	0.971557
2	MLP	0.968563
0	SVM	0.967066
3	Naive Bayes	0.962575
4	Random Forest	0.953593

Figure 15. Performance of classifiers for BBC dataset

	Model	Test accuracy
0	SVM	0.974551
1	Logistic Regression	0.971557
2	MLP	0.968563
3	Naive Bayes	0.962575
4	Random Forest	0.950599

Figure 16. Test Accuracy Of Various Classifiers For BBC Dataset

Table 5: Test Accuracy Of Classifiers For BBC And Agnews Dataset

Classifier	BBC News	AGNews
SVM	97.45	91.04
Logistic Regression	97.15	90.82
MLP	96.85	89.45
Naïve Bayes	96.25	89.48
Random Forest	95.05	88.26
Ensemble Classifier	97.90	91.11

```
LogisticRegression
training accuracy Score    :  0.8664
Validation accuracy Score :  0.8606666666666667
DecisionTreeClassifier
training accuracy Score    :  1.0
Validation accuracy Score :  0.7086666666666667
RandomForestClassifier
training accuracy Score    :  1.0
Validation accuracy Score :  0.8242666666666667
ExtraTreesClassifier
training accuracy Score    :  1.0
Validation accuracy Score :  0.8374
KNeighborsClassifier
training accuracy Score    :  0.8237714285714286
Validation accuracy Score :  0.7061333333333333
BAGGING
Training accuracy Score    :  1.0
Validation accuracy Score :  0.8458
```

Figure 17. Bagging Output Of IMDB Dataset

```
AdaBoostClassifier
training accuracy Score    :  0.801
Validation accuracy Score :  0.7908
GradientBoostingClassifier
training accuracy Score    :  0.8150857142857143
Validation accuracy Score :  0.7942666666666667
XGBClassifier
training accuracy Score    :  0.9236857142857143
Validation accuracy Score :  0.8348666666666666
BOOSTING VotingClassifier
Training accuracy Score    :  0.8388285714285715
Validation accuracy Score :  0.8066
```

Figure 18. Boosting Output Of IMDB Dataset

6.2. Ensemble results using Deep Learning

Convolutional neural network (CNN) [7] is excellent at extracting local and position-invariant features. Long short-term memory (LSTM) performs classification with the help of long-gap semantic dependency. LSTM [5] is a type of recurrent neural network which is used for many applications. It is efficient in learning the order dependence in sequence prediction problems. Hence it can be applied for machine translation and speech recognition tasks. Bidirectional LSTM uses information from both ends of the sequence to estimate the output. The current output depends on future and past observations. Gated Recurrent Unit (GRU) [25] provides a gated approach to efficiently capture dependencies on different time scales. Conventional RNNs face the problems of vanishing and exploding gradients. To overcome this problem GRU is used. CNN, LSTM, GRU, and bidirectional long short-term memory (Bi-LSTM) [6] are used for deep learning ensembles. The loss and accuracy are calculated for deep ensemble models. 5 epochs are taken for computing loss and accuracy of the models. Figure 19 displays the accuracy of classifiers for the IMDB dataset.

Table 6: Training And Testing Accuracy For The IMDB Dataset

Classifiers	Training Accuracy	Testing Accuracy
CNN	100	88
GRU	95	87
Bi-LSTM	95	87
LSTM	93	86
CNN+GRU	97	88
CNN+Bi-LSTM	97	88
CNN+LSTM	97	87

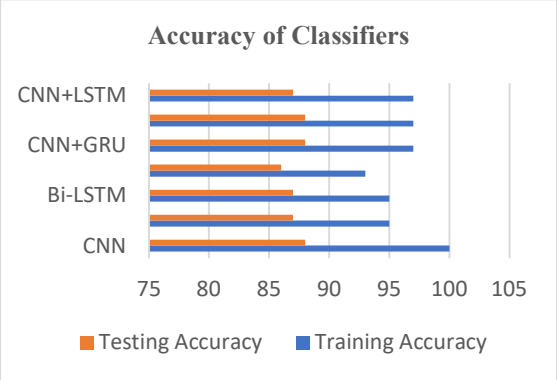

Figure 19. Accuracy For IMDB Dataset

Table 7: Training And Testing Loss For The IMDB Dataset

Classifiers	Training Loss	Testing Loss
CNN	0.17	54
GRU	13	39
Bi-LSTM	13	37
LSTM	16	37
CNN+GRU	7	39
CNN+Bi-LSTM	7	36
CNN+LSTM	7	40

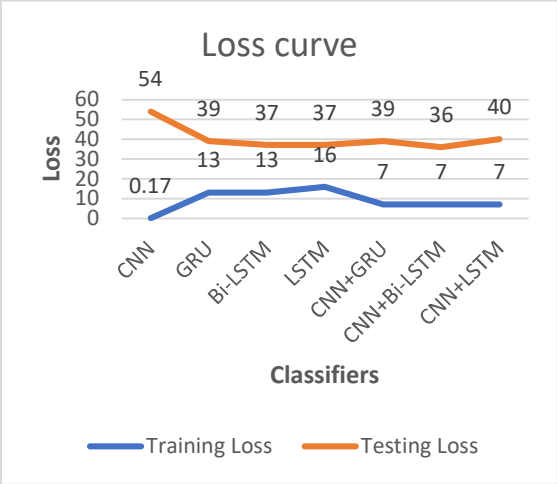

Figure 20. Training And Testing Loss For IMDB Dataset

Figure 20 illustrates the loss of classifiers for the IMDB dataset. Table 6 represents the accuracy of DL and ensemble DL classifiers for the IMDB dataset. The accuracy of DL classifiers is higher than ML classifiers. Table 7 describes the loss values during training and testing for the IMDB dataset. The loss value of ensemble CNN with BiLSTM is less than other classifiers.

7. CONCLUSION

The Ensemble technique is the hybrid model of various classifiers. A hybrid model is used to improve the predicted accuracy and to avoid overfitting problems. In this paper, various ensemble techniques like bagging, boosting, and stacking are discussed. This study is performed to implement an ensemble technique for text classification problems. Three datasets are used for this study. The performance of individual classifiers and hybrid models are compared. The experiment makes use of bagging and boosting techniques. According to the experimental findings, hard voting performs better than weighted and soft voting. Ensemble classifier using ML algorithms gives slightly better results for multi-class text classification datasets when compared to ML algorithms. IMDB dataset is utilized for the deep ensemble model. The accuracy and loss of deep ensemble classifiers are more or less the same as DL models for this dataset. Ensemble learner using DL algorithms gives better results for binary text classification dataset when compared to traditional machine learning algorithms.

REFERENCES:

[1] Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms. In Mining Text Data; Springer:Berlin/Heidelberg, Germany, 2012; pp. 163–222.

[2] Al-Garadi, Mohammed & Yang, Yuan-Chi & Cai, Haitao & Ruan, Yucheng & Oconnor, Karen & Graciela, Gonzalez-Hernandez & Perrone, Jeanmarie & Sarker, Abeed. (2020). Text Classification Models for the Automatic Detection of Nonmedical Prescription Medication Use from Social Media. 10.21203/rs.3.rs-58679/v2.

[3] Zhang, Yuebing, et al. "A cost-sensitive three-way combination technique for ensemble learning in sentiment classification." International Journal of Approximate Reasoning 105 (2019): 85-97.

[4] Liang, Decui, and Bochun Yi. "Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification." Information Sciences 547 (2021): 271-288.

[5] Khai Tran, Thien, and Tuoi Thi Phan. "Deep learning application to ensemble learning—the simple, but effective, approach to sentiment classifying." Applied Sciences 9.13 (2019): 2760.

[6] Anand, Manish, et al. "Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques." Theoretical Computer Science (2022).

[7] Roy, Pradeep Kumar, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. "Hate speech and offensive language detection in Dravidian languages using deep ensemble framework." Computer Speech & Language 75 (2022): 101386.

[8] Lin, Szu-Yin, Yun-Ching Kung, and Fang-Yie Leu. "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis." Information Processing & Management 59.2 (2022): 102872.

[9] Sharif, Omar, and Mohammed Moshiul Hoque. "Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers." Neurocomputing 490 (2022): 462-481.

[10] Briskilal, J., and C. N. Subalalitha. "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa." Information Processing & Management 59.1 (2022): 102756.

[11] Sagi, Omer, and Lior Rokach. "Ensemble learning: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4 (2018): e1249.

[12] J. E. Sembodo, E. B. Setiawan and M. A. Bijaksana, "Automatic Tweet Classification Based on News Category in Indonesian Language," 2018 6th International Conference on Information and Communication Technology (ICoICT), 2018, pp. 389-393, doi: 10.1109/ICoICT.2018.8528788.

[13] Conference on Information and Communication Technology (ICoICT). IEEE, 2018. Kowsari, Kamran, et al. "Text classification algorithms: A survey." Information 10.4 (2019): 150

[14] Thangaraj, M. & Sivakami, M. (2018). Text Classification Techniques: A Literature Review. Interdisciplinary Journal of Information, Knowledge, and Management. 13. 117-135. 10.28945/4066.

[15] Dong, Xibin, et al. "A survey on ensemble learning." Frontiers of Computer Science 14.2 (2020): 241-258.

[16] Zhou, Zhi-Hua. "Ensemble learning." Machine learning. Springer, Singapore, 2021. 181-210.

[17] Dwivedi, S.K.; Arya, C. Automatic Text Classification in Information retrieval: A Survey. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, Udaipur, India, March 2016; p. 131.

[18] Al-Azani, Sadam, and El-Sayed M. El-Alfy. "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text." Procedia Computer Science 109 (2017): 359-366.

[19] Akanksha Patro, Mahima Patel, Richa Shukla and Dr. Jagurti Save, "Real Time News Classification Using Machine Learning", IJAST, vol. 29, no. 9s, pp. 620 - 630, May 2020.

[20] Mohammed, Ammar, and Rania Kora. "An effective ensemble deep learning framework for text classification." Journal of King Saud University-Computer and Information Sciences (2021).

[21] Sultana, Naznin, and Mohammad Mohaiminul Islam. "Meta classifier-based ensemble learning for sentiment classification." Proceedings of International Joint Conference on Computational Intelligence. Springer, Singapore, 2020.

[22] Fattahi, Jaouhar, and Mohamed Mejri. "Spaml: a bimodal ensemble learning spam detector based on NLP techniques." 2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP). IEEE, 2021.

[23] Rezaeinia, S.M.; Ghodsi, A.; Rahmani, R. Improving the Accuracy of Pre-trained Word Embeddings for Sentiment Analysis. arXiv 2017, arXiv:1711.08609.

[24] Nigam, K.; McCallum, A.; Mitchell, T. Semi-supervised text classification using EM. In Semi-Supervised Learning; MIT Press: Cambridge, MA, USA, 2006; pp. 33–56.

[25] Mohammadi, Azadeh, and Anis Shaverizade. "Ensemble deep learning for aspect-based sentiment analysis." International Journal of Nonlinear Analysis and Applications 12. Special Issue (2021): 29-38.