Implementing Machine Learning Algorithms to Classify Paper Peer Reviews

1st Rabin Chhetri
Applied Machine Learning
001185145

***Abstract*—This paper presents the implementation of two machine learning(ML) algorithms to classify whether the ML paper is accepted or not and the review score based on the ML paper peer reviews.Linear Support Vector Machine (SVM) and Random Forest Classifier (RFC) are implemented on ML paper reviews dataset. for the effective implementation of ML algorithms, the dataset is partitioned as 80 percentage to training set and 20 percentage to test set.Validation set is also created from the test set to validate the trained model.Models are evaluated using confusion matrix and classification accuracy, precision, recall, f1-score and support value. Based on the evaluation metrics, Random Forest Classifier shows higher accuracy as compared to Support Vector Machine for the given peer reviews dataset.**

I. Introduction and related work

In todays world, Text classification is mostly used in real time applications. Majority of the researchers and data scientists are showing high interest in developing applications that leverage text classification methods.text classification happens mainly in four stages namely feature extraction, dimension reductions, model selection and evaluation.there are multiple number of algorithms designed for classifying the text and among them this paper highlights on the implementation of Linear Support vector machine and Random Forest Classifier techniques. the dataset used to train the models is of ML paper peer reviews from International Conference of Learning Representation in the years between 2017 and 2020 [1], [2].this is the very huge dataset having random data so various data pre- processing techniques are used to remove unnecessary words, spaces, special characters, punctuations, duplicate records and records with invalid data. It has four attributes and each attribute has 6119 records for it. after cleaning and pre-processing the data, it is further into training and test data. The two models are trained using training data and are evaluated with the help of testing data. Validation set is also created to validate the accuracy of the trained model. Each model is trained to predict the attributes acceptance status and review score separately as acceptance status has categorical value whereas review score has a numerical value. Once the model is trained and tested then their efficiecny is tested using various evaluation metrices like confusion matrix, precision, recall and accuracy.

II. Ethical discussion

This report is based on the classification task by implementing algortihms over the dataset of ML paper peer reviews from International conference of Learning Representation which is freely available in the website and has no copyright issues.this dataset doesn't contain any personal data so modification of the dataset for the local use is not a matter of legal issues as per IEEE policy of Ethical and legal use of data.I don't consider that this task violates any sort of social, ethical and legal aspects as this is just to classify the published paper whether it is accepted or rejected on the basis of review comments given during the peer reviews.while performing every steps of this task from dataset preparation to training and evaluating the models, every aspects of IEEE code of conduct is taken into account so that none of the policies are violated.

III. Dataset preparation

As this is a large dataset with thousands of rows in it, there were many data quality issues including invalid data for attributes, missing values, duplicate records, unnecessary features and many more inconsistencies which may lead to failure of the model prediction. In order to prepare proper dataset, various steps of Exploratory data analysis and data cleaning were performed like describing the structure of the dataset, exploring the data types of all the attributes, identifying the missing values and removing the records associated with it, dropping all the duplicate records. after successfully cleaning the data, the attributes are classified into feature variable and target variable. feature variable is then taken through various text pre-processing techinques including lemmatization, tokenization and many more to remove unnecessary spaces, punctuations, stopwords and special characters. the text data is now converted into numerical features and finally splitted into training set and test set to train and test the models.

IV. Methods

This section of the report contains the brief description of the text classification algorithms that has been implemented on the ML paper peer reviews dataset to classify the paper as accepted or rejected and to predict the review score based on the review comments. The implemented algorithms are Support Vector Machine (SVM) and Random Forest Classifier (RFC).

A. Support Vector Machine (SVM)

Support vector Machine (SVM) are a set of supervised machine learning algorithm used for outliers detection, regression and classification tasks.This algorithm is beneficial

in a high dimensional spaces [3]. It transforms the data in high dimensional space to change a linearly inseparable dataset to a linearly separable dataset due to which it is considered memory efficient.This technique is more accurate in small dataset and is prone to over fitting issue as well.
Support Vector Machine works by identifying a boundary which separates values from each other. In 2-Dimensional space, the boundary is called line, In 3-Dimensional space it is called plane and in higher dimensions it is called hyperplane.

$$h(x_i) = 1 if w.x + b >= 0 \tag{1}$$

$$h(x_i) = -1 if w.x + b < 0 \tag{2}$$

here, wx + b =0 is called as the hyperplane.
equation (1) says it is on or above the hyperplane so the class is +1 whereas equation (2) says it is below the hyperplane so the class is -1.
It uses the below equation to transform the data into the higher dimensional space.

$$\phi(x) = [x, x^2] \tag{3}$$

where x, x^2 represents the data in the lower dimensions and $\phi(x)$ represents data in higher dimensional space.

B. Random Forest Classifier (RFC)

Random Forest Classifier (RFC) is considered as the most chosen and powerful machine learning algortihm widely used for pattern recognition and classification of high dimensional and skewed data [4].The hierarchial nature of the tree classifier makes is uncommon for small dataset. with an aim to make this tree classification algortihm more stable and effective, a decision forest methodology was invented initially by Ho [5],Amit and Geman [6]and later by Breiman [7] as Random Forest.In this method the accuracy is fully dependent on the set of classifiers used to classify the attributes.
For N number of instances in a given dataset, approximately 2/3 of the initial size of dataset is picked up in a random manner through bootstrapping way for N times. the rest of the instances is used as a out-of-bag to be evaluated.
For M as a number of features in a given dataset, the quantity of features that is selected as each split is equal to $\sqrt{M} or \sqrt{M}/21001$[8]

V. Experiments and evaluation

This section of the report talks about the structure of the ML paper peer reviews from International Conference of Learning Representation dataset, the details about the feature variables and the target variable of the dataset and different criteria used to evaluate the performance of implemented model like precision, accuracy, f1-score, support and the comparative results between two different models for two different attributes.

A. Experimental settings

Text Classification Methods described in section IV are implemented on the ML paper peer reviews dataset. this dataset is taken from International Conference of Learning Representation. The experiment is carried out using Jupyter notebook. Python libraries like numby, sklearn, pandas, matplotlib, nltk, re and pickle are used to import and manipulate the dataset to properly fit in the above models. The ML paper peer review Dataset contains 2 feature variables namely text and confidence score. Out of these two feature variables I have reduced the feature variables to just text so that it will be easier to fit into the model. this dataset originally contains 6119 sample records. there are some noisy data, some attributes have invalid values, some have missing values so different stages of exploratory data analysis and data cleaning is done over the dataset to make the dataset proper for training into the model.after successfully cleaning the data, there are 2992 sample records out of which 80 percentage of samples are used as a training data set and rest 20 percentage of samples are used as a testing data set.After this setup, different models are created using Support vector machine algorithm and Random Forest Classifier algorithm. The prediction of attributes acceptance status and review score are done separately and their performance metrices are also evaluated separately as acceptance status has categorical value whereas review score has a numerical value.

B. Evaluation criteria

After implementing the above text classification algorithms on the ML paper peer reviews dataset, different evaluation metrices are observed on the predicted data to determine on how accurately the model can classify the text data. The below sub section decribes the different evaluation metrices used in the experiment to test the correctness of the trained model using text classification algorithms namely Support Vector Machine and Random Forest Classifiers.

1) Accuracy: Accuracy is defined as the fraction of number of total predictions which are correctly predicted by the model. It can have value from minimum 0 to maximum 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Here,
TP= True Positive
TN= True Negative
FP= False Positive
FN= False Negative

2) Precision: Precision is defined as the total number of fraction of correctly predicted positive values out of all predicted positive values.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

3) Recall: Recall is defined as the total number of fraction of the correctly predicted positive values out of all positive values.Sensitivity of the model is also determined using Recall.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

4) F1-Score: F1 Score is defined as the harmonic mean of precision and recall from the model.It is considered as a better option to evaluate classification performance on imbalanced data.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

C. Results

This section contains some experimental values obtained after implementing the above classification algorithms over the ML paper peer review dataset.

TABLE I
ACCURACY OF SVM AND RFC METHODS FOR ATTRIBUTE ACCEPTANCE STATUS USING TEST SET

	SVM	RFC
Accuracy	0.6844	0.6978

TABLE II
ACCURACY OF SVM AND RFC METHODS FOR ATTRIBUTE REVIEW SCORE USING TEST SET

	SVM	RFC
Accuracy	0.2404	0.2704

TABLE III
ACCURACY OF SVM AND RFC METHODS FOR ATTRIBUTE ACCEPTANCE STATUS USING VALIDATION SET

	SVM	RFC
Accuracy	0.8033	0.9333

TABLE IV
ACCURACY OF SVM AND RFC METHODS FOR ATTRIBUTE REVIEW SCORE USING VALIDATION SET

	SVM	RFC
Accuracy	0.5266	0.8133

TABLE V
CLASSIFICATION REPORT FOR ACCEPTANCE STATUS USING SVM

	Precision	Recall	F1 Score	Support
Accept	0.64	0.46	0.54	237
Reject	0.70	0.83	0.76	362
Accuracy			0.68	599
Macro Avg	0.67	0.65	0.65	599
Weigh Avg	0.68	0.68	0.67	599

TABLE VI
CLASSIFICATION REPORT FOR ACCEPTANCE STATUS USING RFC

	Precision	Recall	F1 Score	Support
Accept	0.88	0.27	0.42	237
Reject	0.67	0.98	0.80	362
Accuracy			0.70	599
Macro Avg	0.78	0.62	0.61	599
Weigh Avg	0.75	0.70	0.65	599

TABLE VII
CLASSIFICATION REPORT FOR REVIEW SCORE USING SVM

	Precision	Recall	F1 Score	Support
2	0.00	0.00	0.00	14
3	0.22	0.04	0.07	48
4	0.21	0.22	0.21	117
5	0.19	0.23	0.21	121
6	0.26	0.41	0.32	139
7	0.32	0.26	0.29	117
8	0.00	0.00	0.00	36
9	0.00	0.00	0.00	7
Accuracy			0.24	599
Macro Avg	0.15	0.15	0.14	599
Weigh Avg	0.22	0.24	0.22	599

TABLE VIII
CLASSIFICATION REPORT FOR REVIEW SCORE USING RFC

	Precision	Recall	F1 Score	Support
2	0.00	0.00	0.00	14
3	0.00	0.00	0.00	48
4	0.29	0.39	0.33	117
5	0.21	0.16	0.18	121
6	0.26	0.55	0.35	139
7	0.38	0.18	0.24	117
8	0.00	0.00	0.00	36
9	0.00	0.00	0.00	7
Accuracy			0.27	599
Macro Avg	0.14	0.16	0.14	599
Weigh Avg	0.23	0.27	0.23	599

TABLE IX
CONFUSION MATRIX OF ACCEPTANCE STATUS FOR SVM

	Accept	Reject
Accept	109	128
Reject	61	301

TABLE X
CONFUSION MATRIX OF ACCEPTANCE STATUS FOR RFC

	Accept	Reject
Accept	65	172
Reject	9	353

TABLE XI
CLASSIFICATION REPORT OF VALIDATION SET FOR ACCEPTANCE STATUS USING SVM

	Precision	Recall	F1 Score	Support
Accept	0.80	0.68	0.74	121
Reject	0.80	0.89	0.84	179
Accuracy			0.80	300
Macro Avg	0.80	0.78	0.79	300
Weigh Avg	0.80	0.80	0.80	300

VI. DISCUSSION AND FUTURE WORK

These statistical data visualised in the form of above twelve tables in section V.C shows the comparative evaluation metrices between two implemented methods Support vector Machine and Random Forest Classifier. Data from table I to table IV clearly demonstrates that Random Forest Classifier is found to be more accurate than Support Vector Machine in terms of classifying the large data like ML paper peer review

TABLE XII
CLASSIFICATION REPORT OF VALIDATION SET FOR ACCEPTANCE STATUS USING RFC

	Precision	Recall	F1 Score	Support
Accept	1.00	0.83	0.91	121
Reject	0.90	1.00	0.95	179
Accuracy			0.93	300
Macro Avg	0.95	0.92	0.93	300
Weigh Avg	0.94	0.93	0.93	300

dataset.

Though Random Forest Classifier is found to be comparatively more accurate but still the accuracy of both the models is not as expected. this low level of accuracy is obtained because of the improper tuning of the parameters while cleaning, tokenizing the text data and converting text into corresponding numerical features.these two models are found to be less accurate for numerical target attributes and good for categorical target attributes. Deeper level of Data cleaning, feature scaling and proper tuning of the parameters for constructors while training the model can be done in the future implementation interms of improving the performance of the model.

VII. CONCLUSIONS

Hence, after successfull implementation of Linear Support Vector Machine and Random Forest Classifier algorithm to classify whether the paper is accepted or rejected and the review score based on the review comments on ML paper peer review dataset, It is concluded that RFC can classify more accurately based on the information available from the ML paper peer review dataset. Also, other Machine learning algorithms can be implemented over this dataset to classify the text more accurately and precisely. Further, It is also concluded that proper cleaning of the text data, proper lemmatization and tokenization, proper distribution of data into training and test set, precise tuning of different paramters for classifiers and reducing the number of feature varibales in the dataset helps in improving the overall performance of the models.

REFERENCES

[1] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (PeerRead): Collection, insights and nlp applications,"2018. [Online]. Available: https://arxiv.org/abs/1804.09635.

[2] Openreview, github. Accessed: 11/01/2022. [Online]. Available: https://github.com/Seafoodair/Openreview.

[3] Paul Pavlidis Ilan Wapinski, William Stafford Noble, "Support vector machine classification on the web", In: Bioinformatics 20.4,March 2004, pp. 586–587

[4] Ahmad Taher Azar, Hanaa Ismail Elshazly, Aboul Ella Hassanien, Abeer Mohamed Elkorany, "A random forest classifier for lymph diseases,"Computer Methods and Programs in Biomedicine, vol. 113, pp. 465–473, 2014.

[5] T. Ho, Random decision forest,in:3rd International Conf. on Document Analysis and Recognition,August 14-18, Montreal, Canada, 1995, pp. 278-282.

[6] Y. Amit, D. Geman, Shape quantization and recognition with randomized trees, Neural comput, 9(1997), 1545-1588.

[7] L. Breiman, Random Forests, Mach. Learn. 45 (2001), 5-32

[8] O.Okun, H. Priisalu, Random Forest for Gene Expression Based Cancer Classification:overlooked issues, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 483-490.