ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 12(33), DOI: 10.17485/ijst/2019/v12i33/146130, September 2019* ISSN (Online) : 0974-5645

Classification of Sindhi Headline News Documents based on TF-IDF Text Analysis Scheme

Irfan Ali Kandhro[1*], Sahar Zafar Jumani[1], Ajab Ali Lashari[2], Saima Sipy Nangraj[1], Qurban Ali Lakhan[1], Mirza Taimoor Baig[3] and Subhash Guriro[4]

[1]Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan;
irfan@smiu.edu.pk, sahar@smiu.edu.pk, snangraj@smiu.edu.pk, qurban@smiu.edu.pk
[2]Department of Education, Sindh Madressatul Islam University, Karachi, Pakistan;
ajablashari@smiu.edu.pk
[3]Department of Computer Science, University of Karachi, Pakistan;
Taimoor.baig@iunc.edu.pk
[4]Faculty of Media and Communication Studies, Sindh Madressatul Islam
University, Karachi, Pakistan; subhash@smiu.edu.pk

Abstract

Objectives: Sindhi language, historically rich belongs to Indo-Aryan language with diverse background and diverse dialects. Recent drive in globalization, e-commerce and e-literacy have influenced on languages as well. There are lots of magazines, Sindhi books, newspapers and web material available online, but unluckily still proper dataset is not designed for Sindhi information processing. This research study focuses on the Sindhi language news headline texts dataset and automated tool for the online texts' classification based on the predefined label. **Methods/Statistical Analysis:** For the collection of datasets, the scraping tool is designed for extraction of the headline news from most popular newspapers: Awami Awaz and Daily Jhoongar. The dataset contains 2800 Sindhi headline news with five categories: 0. Entertainment, 1. Sports, 2. Science and Technology, 3. International, 4. National, 5. Sindhi news. The dataset is normalized by removing stop words and cleaning the spaces, punctuations and other unnecessary texts. Furthermore, the language feature is analyzed using TF-IDF and vector model. This paper presents Sindhi headline news classification model with implementation of the machine learning classification algorithms, namely. Multinomial NB, Linear SVC, Logistic Regression, MLP classifier, SGD Classifier, Random Forest Classifier, Ridge Classifier. **Findings**: The results show that the performance of the Linear SVC and MLP Classifier indicate better results on Sindhi headlines news categorization as compared to other classification techniques. This research study helps in improving the automatic classification of Sindhi text headline news.

Application/Improvements: It is recommended that LSVC and MLP Classifiers should be used in Sindhi language news headline classification.

Keywords: IR Models, Machine Learning, News Classification, SHN, Sindhi News, Text Classification, TF-IDF

1. Introduction

With the rapid growth of language dominance and online information on the web, the survival of different human languages has become a major issue due to the increasing day by day use of information and internet-based technologies[9]. The Natural Languages (NL) are the best methods for the communication between humans but its critical task for the machines to decipher the sense like human[4]. The text categorization has become one of

**Author for correspondence*

the key techniques of text mining to manage and organize the text information more efficiently by classifying the documents into classes by using classification methods. Text classification techniques are used to classify the news information, stories, contents that refer to identify the problems and get them to resolve by documents based on contextual information of the text and with respect to predefined labels. The purpose of the text classification is to assign a category to a new data/document[1]. Recently, the text categorization has been used for classification of the headline news in various languages, but work related to Sindhi text or Sindhi news headlines classification has not been carried out. The main focus of this research study is on Sindhi news headlines classification, and it has been proposed that Sindhi Headline News Model (SHN) with implementation of various machine learning approaches (namely: Multinomial NB, Linear SVC, Logistic Regression, MLP classifier, SGD Classifier, Random Forest Classifier, Ridge Classifier) are effective to classify Sindhi news, predefined classes. First, the SHN will extract important features from the Sindhi text by using Vector and TF-IDF and then model implements classification algorithms to classify the selected features to determine the respective news categories of Sindhi news headlines[1]. The significant contribution of this research study is to highlight the importance of Sindhi language with respective of the text classification/headline news classification and to discuss the significant classification algorithms for Sindhi language processing. Figure 1 shows categorize sample size of Sindhi news corpus.

This paper has been organized in the following sequence: Section II shows related work of news classification techniques; Section III represents methodology of the research SNH model, Section V compares results and discussion of the model and Section VI represents conclusion and future research contribution of this study.

Recent work shows the significant contribution of researchers in the field of text mining and data classification. The term text mining includes text cleaning and extraction of meaningful information from raw and unstructured textual data. The procedure of text processing comprises of pre-processing text i.e. (tokenization, stop words removal, POS tagging, common word stemming and selection of features and weights of words)[4]. Lots of preprocessing steps have been done in various languages. Once tokenization is performed, then tokenized data generated which is also stated in the dictionary list. Secondly, the various methods have been designed for removing stop words from the text documents. Plenty of stemming methods have been designed for the different languages i.e. (English, Urdu, Persian, Arabic and Turkish)[5,6]. There are few generic stemming methods of the language Urdu and selection of the text features. Many researchers have designed approaches which combine Document Frequency, Boolean Weighting scheme, TF-IDF, mutual and gain information[3]. In[11] designed corpus for organizing the grammatical and morphological structure of Sindhi language.

The text classification model is proposed for separating the predefined labels from the text documents by using supervised machine learning techniques. It uses different practical applications like sentiment analysis, e-mail spam detection and natural language processing. Further, this model applied five well-known classification approaches on Urdu corpus and labeled with the class to the document by the majority voting scheme.

They[8] carried out the classification of text which comprises seven different categories (Culture, Health, Sports, Business, Entertainment and Weird) on Urdu corpus based on 21769 documents regarding the news. Various Machine learning algorithms have been applied to predict classes on 93400 features taken out from multiple data sets which ensure 94% precision and recall using classify class. NLP is a diverse field regarding the complex nature of language and ambiguity in language and speech. For NLP tasks Machine Learning and statistical tools are the best to analyze the data.

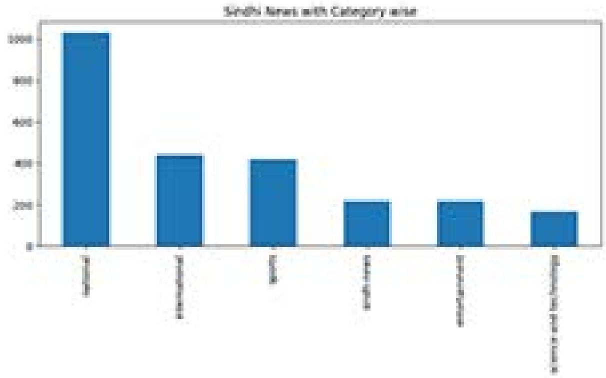


Figure 1. Sindhi headline news corpus label wise.

2. Proposed Methodology

2.1 Stop Words

Information retrieval models, the stop words are not providing important and significant information from the text. The main aim of these words is just complete the sentences and give a proper understanding of the sentence. Therefore, the stop words construct the sentences correctly. In the field of computer science does not give significantly to stop words for any kind of text processing application, for that reason, it filters these words during searching and text process. To analysis the dataset of Sindhi headline news text, Sindhi stop words are identified from the headline news text dataset. The selected stops from the corpus are articles, prepositions, conjunctions, determiners, verbs, interjections. In Table 1 described some random stop words of Sindhi text and show the structure and style of Sindhi stop words in the text corpus. [8]In Sindhi text corpus, there are some words have high ranking, but the contribution of these words meaningless, the word (کي) has highest ranked in the corpus and second-highest ranked stop words is (۾ or م) which means (in) in English in the text corpus and third highest ranked word is (in) which is determiner (آهي) and means that/this English. Therefore, these words have the highest frequency of stop words described the importance of prepositions and verbs[9].

Table 1. Random samples of stop words of Sindhi text

STOP WORDS		
هي	ڪان	جو
اهو	م	کي
هن	اسان	ته
مون	آهن	تي
آهن	۾	۽
هلي	جتي	آهي
چيو	هر	چو
..	..	..
ويو	ٿي	جتي

3. TF-IDF

Term Weighting schemes are the new and noteworthy approach for information retrieving system. The core functionality of TF weighting scheme is to identify the significant features of the term of document. Therefore, the document is ranked based on the term weighting. The term frequency and inverse document frequency are called TF-IDF and it extracts the important feature and useful statistical model for information retravel systems and text mining. Furthermore, it is to find out the most important and key terms from the text corpus and make them useful for further processing. Addition to this, TF-IDF labels the importance of the word and uses them as a term to the text document. TF counts the number of words exists in the text document. The frequency of words is divided by a complete list of terms in the documents. Table 2 shows the computed term frequency and inverse documents frequency of all documents[7].

Table 2. Random samples term frequency and inverse document frequency

DOC NO	FEATURE NAME	FEATURE	TF-IDF
1	لاس	2	0.021739
1	نيوز	4	0.021738
1	هالي	6	0.043478
1	اندسٿري	9	0.021749
1	ڳالهائيندي	12	0.021738
2	لاهور	14	0.060000
2	صحافين	19	0.021739
….	….	….	….
2494	پاڪستان	55252	0.075223
2494	واپار	55553	0.023122
2494	ايشيائي	55654	0.032242
2494	تنظيم	55752	0.054212

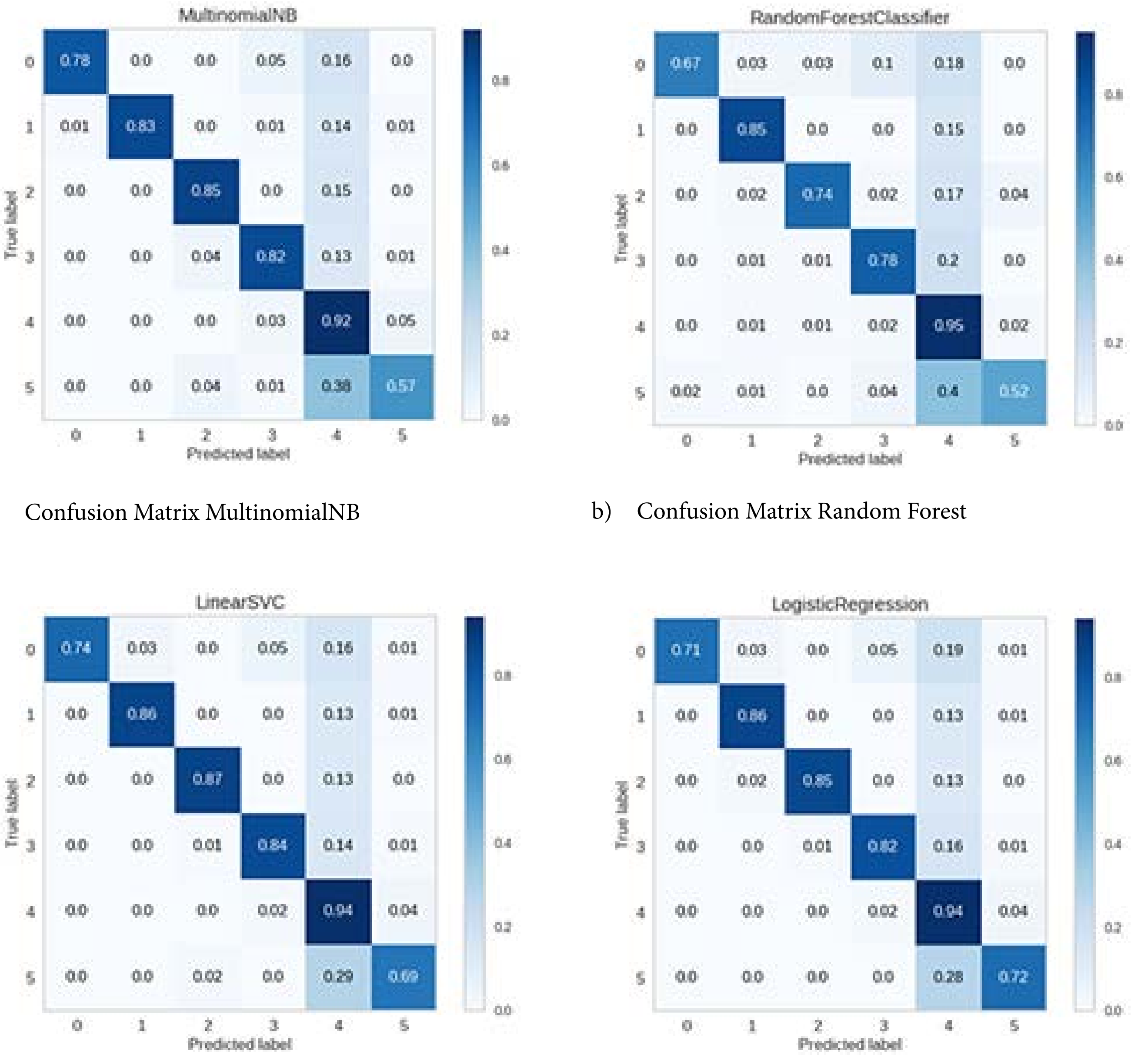
4. Classification Algorithms

The Multinomial Naïve Bayes (MNB) Classifier is used for the text classification for extracting distinct features. MNB separations generally need integer features, but in practical implementation, for counting features used TF-IDF. MNB is an updated version of simple naïve Bayes synchronized classification. It upgraded the conventional bag of words. Linear SVC is a modified version of support vector classifier with different hyper parameter. In addition to this, Linear SVC computes loss function, penalties and scaling quite better on a large number of samples. The running time and implementation are better than Linear SVM[13]. The Logistic Regression is used for the analysis of high dimensional data, such as text, images, videos, NLP computational text. The Bayesian logistic regression method is used with Laplace before cover over fitting the results and besides produce sparse representation for the text[14]. The MLP Classifier is a modified model of the Artificial Neural Network. ANN reproduces the learning in steps and it adopts the behaviors just like humans with an attempt to model the structure of biological neural network[15]. The Stochastic Gradient Descent (SGD) used with the regularized linear network for the classification of text. In each sample gradient loss is estimated along the way of reducing asset and learning rate. Further, fixed SGD optimization approach improves automatic classification in various fields[16]. The Random forest used estimator somehow called a meta estimator and adjust different samples and sub-samples of the dataset, and computes averaging for improving accuracy and over fitting. The size of the sample is always sample as like original, but samples are drawn with parameter bootstrap true[17]. The ridge Classifier also called cyclic coordinate descent. It works in steps, where each step minimizes the coordinate before moving to the next. The fashion gives overall optimum solution[18].

5. Results and Discussion

A good care has been done in order to validate the results and use reliable tools for data analysis. Keeping in view the results, it is reported that for the text classification F1 Score (F-measure), Precision and Recall has been used as evaluation metrics. Precision has been used as a number which is the sum of true positive predictions divided by the whole sum of true and false-positive predictions. The Recall is the ratio of true positive predictions divided by the whole sum of true positive and false negative prediction in the set. Precision metric helps to assess the correctness of the classifier. If there is Precision in a greater amount which means less false positive, while if there is Precision is in lower amount means false positive is more. In recall scenario a simple method is to improve precision is to decrease recall. Recall metrics is used to find out the completeness of the classifier or its sensitivity. If there is less recall means higher false, negative while if there is a higher amount of recall means lower false negatives. Strengthening recall mostly reduces precision because it will be difficult to be precise as the sample size increase. F-score is the combination of Precision and Recall considered as harmonic mean. Precision and Recall metrics have been widely used as valid metrics to evaluate the effectiveness of classifier while Accuracy metrics fail to achieve the desired results. The Recall and Precision metrics are significant metrics which can help to study in-depth about the performance attributes of the multiclass classifier.

Figure 2 shows the Confusion Matrix for the headline text news features classifier obtained from the corpus by dividing into testing and training parts. The actual categories are listed vertically and the predicted categories are listed horizontally. According to all matrixes' results; it's not necessary that all models should perform well on the predefined labels. The news classification model is dividing the dataset into training and testing which leads to the analysis of main sources of misclassification on the test set. Major source to identify error is confusion matrix based on predicted and actual labels discrepancies. The matrix summarizes the performance of model news headline text classification and evaluates the classification report in five categories. It computes the true positive and negative, false positive and negative. 1. Multinomial NB classifier highest correct predication is (Label 4, 92%) and the lowest of (label 5, 57%), 2. The random forest classifier shows the high correct prediction of (label 5, 95%) and low predication of (label 5, 52%). 3. Linear SVC classifier, correct predication (Label 4, 94%) and (label 5, 69%), shows a correct prediction on the diagonal side, where shows correct label of entertainment (74%), sports (86%), science-and-technology (87%), International (84%), National (94%) and Sindh-news (94%). 4. The logistic regression classification report of label 4, 94% and Label 5, 72%. The correctness of label 4 is significantly greater than the Label 5. Comparatively, the classification report of Linear SVC model is better than 1, 2 and 3. Furthermore, it shows the high predication of four labels out of six. In Figure 3, 1. SGD classifier highest correct predication is (Label 4, 92%) and the lowest of (Label 5, 57%) 2. In MLP Classifier shows the high correct prediction of (label 5, 95%) and low predication of (label 5, 52%). 3. Ridge classifier, correct predication (Label 4, 94%) and (label 5, 70%), its represents correct predication of the model on diagonal side, where label entertainment (74%) and sports (86%), science and technology (87%), International (83%), National (94%) and Sindh-news (70%). Figure 4 shows the results of Precision-Recall Curve, where X and Y direction indicate the Recall and Precision and the zigzag curves line frequently moves up and down which represents the changes in Precision and Recall with respect to class, time and moreover, both evaluation parameters cross each other more quickly than normal ROC Curve on certain time frame as graphical representation shows the Precision of 1. MNB Classifier with respect to class 0, 1, 2, 3, 4, 5 are (90%, 93%, 90%, 88%, 94%, 74%) and 2. RF classifier the precision results are (90%, 93%, 90%, 88%, 94%, 74%) and similarity 3. LSVC (90%, 93%, 90%, 88%, 94%, 74%) and 4. LG (90%, 93%, 90%, 88%, 94%, 74%). Furthermore, analyzing the



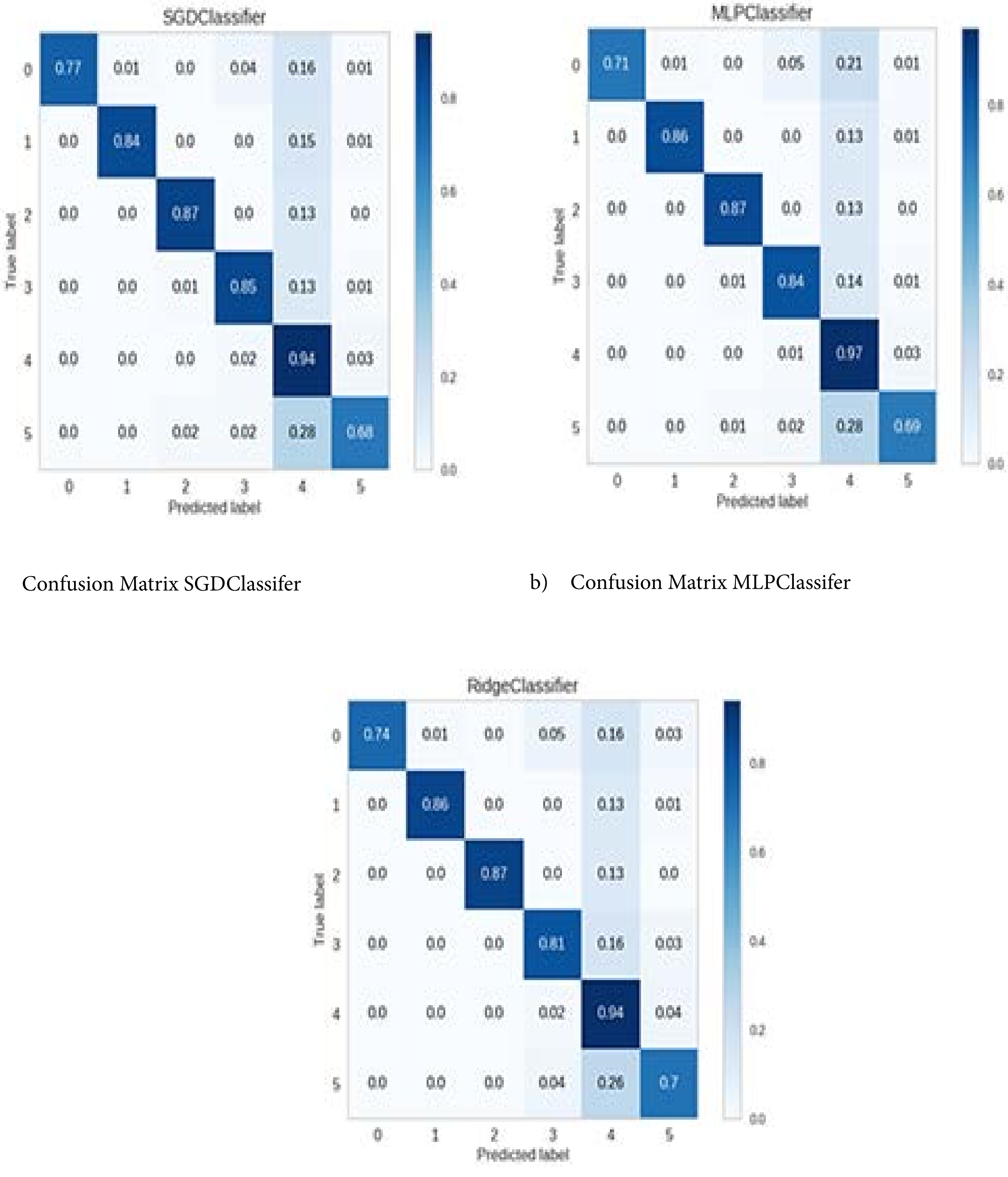
a) Confusion Matrix MultinomialNB

b) Confusion Matrix Random Forest

c) Confusion Matrix Linear SVC

d) Confusion Matrix Logistic Regression

Figure 2. Confusion Matrix (a) Multinomial NB Classifier, (b) Random Forest Classifier, (c) Linear SVC Classifier and (d) Logistic Regression

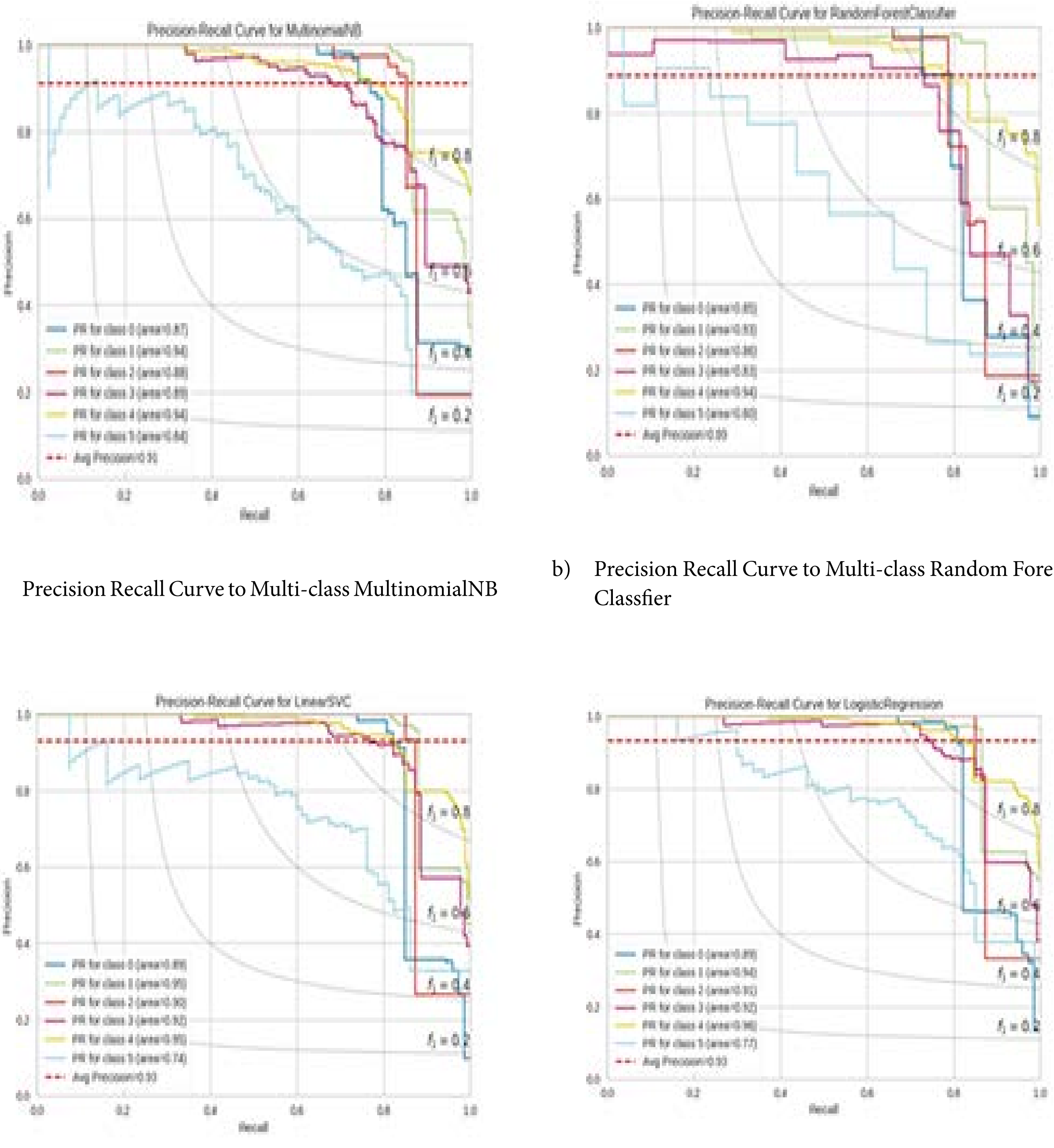


a) Confusion Matrix SGDClassifer

b) Confusion Matrix MLPClassifer

c) Confusion Matrix RidgeClassifer

Figure 3. Confusion Matrix (a) SGD classifier, (b) MLP Classifier and (c) Ridge Classifier.



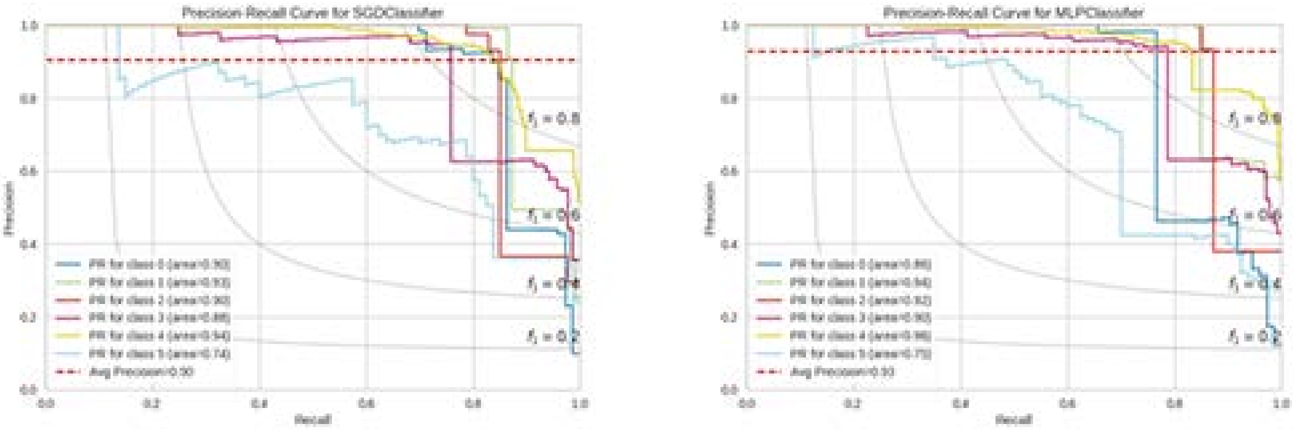
a) Precision Recall Curve to Multi-class MultinomialNB

b) Precision Recall Curve to Multi-class Random Forest Classfier

c) Precision Recall Curve to Multi-class LinearSVC

d) Precision Recall Curve to Multi-class Logistic Regression

Figure 4. Precision-Recall Curve (a) Multinomial NB Classifier, (b) Random forest Classifier, (c) Linear SVC Classifier and (d) Logistic Regression.



a) Precision Recall Curve to Multi-class SGDClassfier

b)Precision Recall Curve to Multi-class MLPClassifier

c) Precision Recall Curve to Multi-class Ridge Classifer

Figure 5. Precision-Recall Curve (a) SGD classifier, (b) MLP Classifier and (c) Ridge Classifier.

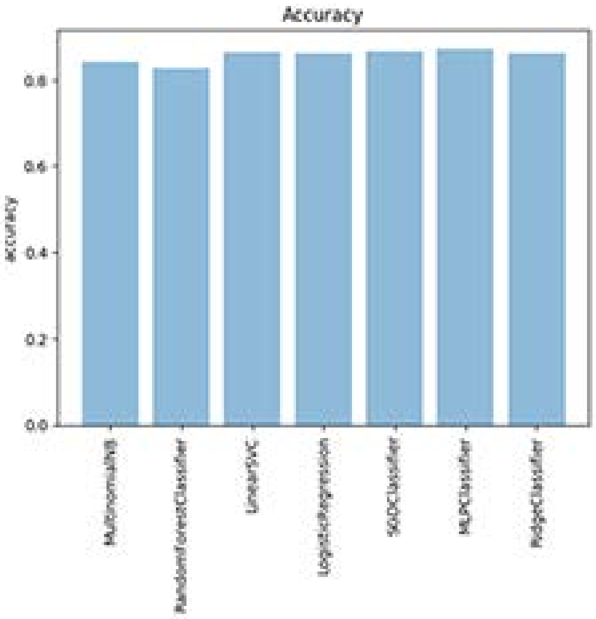


Figure 6. Results comparison of various Machine Learning Techniques.

overall performance system, the average Precision also compute the average precision of the (RF) 90%, (RF) 90%, (RF) 90% and (RF) 92%. Figure 5 depicts the Precision and Recall Curve of 1. SGD Classifier with respect to class 0, 1, 2, 3, 4, 5 are (90%, 93%, 90%, 88%, 94%, 74%) and 2. MLP classifier the Precision results are (90%, 93%, 90%, 88%, 94%, 74%) and similarity 3. Ridge classifier (90%, 93%, 90%, 88%, 94%, 74%). The average precision is (RF) 90%, (RF) 90%, (RF) 90% and (RF) 92%. Figure 6 shows the accuracy of different machine learning classification algorithms, such as Multinomial NB, Linear SVC, Logistic Regression, MLP classifier, SGD Classifier, Random Forest Classifier, Ridge Classifier) to classify Sindhi text headline news. The accuracies are 92%, 92%, 92%, 92%, 92%, 92%, and 92%. Its shows the accuracy RF classification algorithm better than others, it has been reported that RF classifier extremely good for identifying the features from the Sindhi text.

6. Conclusion

This paper presents a new model for automated online news text classification for the Sindhil anguage. The study has been carried out the online web-based Sindhi headline news text classification by incorporating the information retrieval models and machine learning classification algorithms. The corpus is collected from Awami Awaz and Daily Jhoongar using self-designed scrapper tool. Furthermore, the corpus split into two parts testing and training, 30% for testing and the remaining 70% for training. In the study model, at the first stage the term weighting method to assign the term weights and computes the relevant documents based on the user queries. Moreover, to analyze the most important features from the documents TF-IDF and count vectorization has been computed also. And then machine approaches have been implemented namely: Multinomial NB, Linear SVC, Logistic Regression, MLP classifier, SGD Classifier, Random Forest Classifier, Ridge Classifier approaches. The performance of model evaluated through, Confusion Matrix, Precision and Recall Curve, Average Precision and Accuracy metrics. The results show the accuracy of MNB (82%), LSVC, (84%), LR (83%), MLPC (84%), SGDC (82%), RFC (83%) and Ridge Classifier (83%). The representation of graphs shows that the performance of LSVC and MLP classifier is better than other classification algorithms.

7. Future Work

In future, this work will be extended, to increasing the more samples in the dataset and order to implement the Long Short Term Memory (LSTM), deep learning model.

8. References

1. Rajan K, et al. Automatic classification of Tamil documents using vector space model and Artificial Neural Network. Expert Systems with Applications. 2009: 36(8):10914–8. https://doi.org/10.1016/j.eswa.2009.02.010
2. Ahmed K, et al. Framework for Urdu News Headline Classification. 2016; 10(21):17. https://doi.org/10.4316/JACSM.201601002
3. Ali M, Khalid S, Saleemi MH. A novel stemming approach for Urdu language. Journal of Applied Environmental and Biological Sciences. 2014; 4(7S):436–43.
4. Lovins JB. Development of a stemming algorithm. Mech Translat and Comp Linguistics.1968: 11(1-2):22–31.
5. Porter MF. An algorithm for suffix stripping. Program. 2006; 40(3):211–8. https://doi.org/10.1108/00330330610681286
6. Ali M, et al. A rule based stemming method for multilingual Urdu text. International Journal of Computer Applications. 2016: 134(8):10–8. https://doi.org/10.5120/ijca2016907784
7. Ali M, WaganA. An analysis of Sindhi annotated corpus using supervised machine learning methods. Mehran

University Research Journal of Engineering and Technology. 2019: 38(1):185–96.

8. Ali M, Wagan A. Sentiment summerization and analysis of Sindhi text. Int J Adv Comput Sci Appl. 2017: 8(10):296–300. https://doi.org/10.14569/IJACSA.2017.081038
9. Jamro WA. Sindhi language processing: A survey. 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT); 2017.
10. Khan W, Daud A, Nasir JA, Amjad T. A survey on the state-of-the-art machine learning models in the context of NLP. Kuwait Journal of Science. 2016; 43(4):95–113.
11. Usman M, Ayub S, Shafique Z, Malik K. Urdu text classification using majority voting. International Journal of Advanced Computer Science and Applications. 2016; 7(8):265–73. https://doi.org/10.14569/IJACSA.2016.070836
12. Sharma NS, Singh M. Modifying Naive Bayes classifier for multinomial text classification. 2016 IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE); 2016. https://doi.org/10.1109/ICRAIE.2016.7939519
13. Kaur G, Kaur P. Novel approach to text classification by SVM-RBF kernel and linear SVC. International Journal of Advance Research, Ideas and Innovation in Technology. 2017; 3(3):1014–7.
14. Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics. (2007): 49(3):291–304. https://doi.org/10.1198/004017007000000245
15. Basu S, et al. Handwritten Bangla alphabet recognition using an MLP based classifier. arXiv preprint arXiv: 1203.0882; 2012.
16. Diab S. Optimizing stochastic gradient descent in text classification based on fine-tuning hyper-parameters approach. A Case Study on Automatic Classification of Global Terrorist Attacks. arXiv preprint arXiv: 1902.06542. 2019.
17. Xu B, et al. An improved random forest classifier for text categorization. JCP. 2012; 7(12):2913–20. https://doi.org/10.4304/jcp.7.12.2913-2920
18. Genkin A, Lewis DD, Madigan D. Sparse logistic regression for text categorization. DIMACS Working Group on Monitoring Message Streams Project Report. 2005.