###### 

**Sindhi Headline News Text Classification**

|  |  |
| --- | --- |
| Mr A  Department of Computer Science  Sindh Madressatul Islam University  Karachi, Pakistan  a@smiu.edu.pk | Mr B  Department of Computer Science  Sindh Madressatul Islam University  Karachi, Pakistan  b@jinnah.edu |

***Abstract* —** Sindhi language is Indo-Aryan language and historically rich with diverse background and diverse dialects. Recent drive in globalization, e-commerce and e-literacy have influence that language as well. There are lots of magazines, sindhi books, newspapers and online 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 predefined label. In this paper, for the collection of datasets, the scraping tool is designed for extraction of the headline news from mostly 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: Sindh-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 TFIDF and vector model. This paper presents Sindhi headline news classification model with implementation of the machine learning classification algorithms, namely. MultinomialNB, LinearSVC, Logistic Regression, MLP classifier, SGD Classifier, Random Forest Classifier, Ridge Classifier. 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. It is recommended that LSVC and MLP Classifiers should be used in Sindhi language news headline classification.

Keywords — Sindhi news, Text classfication, Machine learning, News classfication, TFIDF, IR models, SHN.

# INTRODUCTION

With the rapid growth of languages dominance and online information on web, the survival of different human languages has become 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 the human to human but its critical task for the machines to decipher the sense like human.[4] The text categorization has become one of the key techniques of the 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 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 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: (MultinomialNB, LinearSVC, 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 TFIDF and then model implements classification algorithms to classify the selected features to determine the respective news categories of Sindhi news headlines [1].

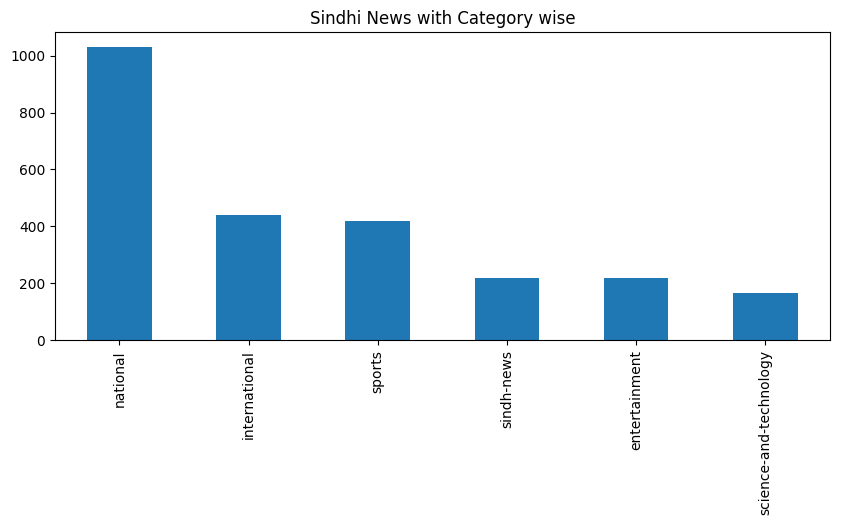


Figure 01: Sindhi Headline News corpus Label wise

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.

This paper has been organized in the following sequence: the 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 model and section VI represents conclusion and future research contribution of this study.

# RELATED WORKS

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) etc. [4]. Lots of of preprocessing steps have been done in various languages. Once tokenization is performed, then tokenized data generated which is also stated in dictionary list. For elimination of the stops words the numerous methods have been designed in the literature that combine the usage of statistical approach and static words list, past, the plenty of stemming methods have 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 of researchers have designed approaches which combine Document Frequency, Boolean Weighting scheme, TFIDF, mutual and gain information etc. [3]. (Mazhar et al) designed corpus for organizing 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, email spam detection and natural language processing. Further, this model applied five well known classification approaches on Urdu corpus and labeled with class to the document by majority voting scheme.

Muhammad Usman et al. carried out a research study on the classification of text which comprises seven different categories (Culture, Health, Sports, Business, Entertainment and Weired) on Urdu corpus based on 21769 documents regarding news. Various Machine learning algorithms have been applied to predict classes on 93400 features taken out from multiple data sets which ensures 94% precision and recall using classify class. [10] Wahab et al, NLP is diverse field regarding 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.

# METHODOLOGY

1. **Stopwords**

An information retrieval models, the stop words are not providing the important and significant information from the text. the main aim of these words is just complete the sentences and give the proper understanding of sentence. therefore, the stop words construct the sentences correctly. In the field of the computer science does not give significant 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 corpus are articles, prepositions, conjunctions, determiners, verbs, interjections. In figure 1 described some random stop words of sindhi text and show the structure and style of sindhi stop words in the text corpus.[8]

**TABLE I. Random Samples of Stopwords of Sindhi text**

|  |  |  |
| --- | --- | --- |
| STOPWORDS | | |
| هي | کان | جو |
| اهو | م | کي |
| هن | اسان | ته |
| مون | آهن | تي |
| آهن | ۾ | ۽ |
| هلي | جتي | آهي |
| چيو | هر | ڇو |
| .. | .. | .. |
| ويو | ٿي | جتي |

In Table I shows random Stopwords of 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 highest frequency of stop words described the importance of prepositions and verbs.

1. **TFIDF**

Term Weighting schemes are 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 (Paik, 2013). Therefore, the document is ranked based on the term weighting. The Term frequency and Inverse document frequency are called TFIDF and it extract the important feature and useful statistical model for information retravel systems and text mining. furthermore, the 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 use them as term to the text document.TF counts the number of words exists in the text document. that the reason, the frequency of words is divided by complete list of terms in the documents. In Table II computed the term frequency and inverse documents frequency of all documents

**TABLE II. Random Samples Term Frequency and inverse Document Frequency**

|  |  |  |  |
| --- | --- | --- | --- |
| **DOC N0** | **FEATURE NAME** | **FEATURE** | **TFIDF** |
| 1 | لاس | 2 | 0.021739 |
| 1 | نيوز | 4 | 0.021738 |
| 1 | هالي | 6 | 0.043478 |
| 1 | انڊسٽري | 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 |

1. **Classification Algorithms**

The Multinomial Naïve Bayes Classifier used for the categorization or classification distinct of features (eg count the word type features in the text classification). MNB separations generaly need integer features. But in practical implementation, for counting features used TFIDF. Its extension of simple naïve Bayes approach by updating conventional Bag of words.[12] **LinearSVC** same as support vector classifier with kernel type linear parameter, but LinearSVC implements liblinear parameter rather than libsvm. Furthermore, it provides more easiness in used of loss function and penalties and scaling quite better on large number of samples. the running time and implementation is 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 overfitting the results and besides produce sparse representation for the text [14]. The MLP Classifier is modified model of the artificial neural network. ANN reproduce the learning in steps, and it adopts the behaviors just like humans with attempt to model the structure of biological neural network [15]. The Stochastic gradient descent (SGD) used with 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 improving 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 the accuracy and overfitting. The size of sample is always sample as like original, but samples are drawn with parameter bootstrap true [17]. The ridge Classifier also called cyclic coordinate descent. Its works in steps, where each step minimizes the coordinate before moving to the next. The fashion gives overall optimum solution [18].

# RRESULT AND DISUCSSION

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 to as evaluation metrics.

Precision has been used as a number which is the sum of true positive predictions divided by 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 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 higher amount of recall means lower false negatives. Strengthening recall mostly reduces precision because it will be difficult to be precise as sample size increase. F-score is the combination of Precision and Recall considered as harmonic mean (C. J. van Rijsbergen, 1979): Precision and Recall metrics has 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 01 shows the Confusion Matrix for the headline text news features classifier obtained from the corpus by dividing in testing and training parts. The actual categories are listed vertically, and the predicted categories are listed horizontally. According to all matrix’s results; it’s not necessary that all models should perform well on the predefined labels. The news classification model is dividing 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 evaluate the classification report in five categories. It computes the true positive and negative, false positive and negative. (a) MultinomialNB classifier highest correct predication is (Label 4, 92%) and the lowest of (label 5, 57%) (b) The random forest classifier shows the high correct prediction of (label 5, 95%) and low predication of (label 5, 52%). (c)Linear SVC classifier, correct predication (Label 4, 94%) and (label 5, 69%), shows a correct prediction on diagonal side,

Where shows correct label of entertainment (74%), sports (86%), science-and-technology (87%), International (84%), National (94%) and Sindh-news (94%). (d) 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 (a), (b) and (c), and furthermore, it shows the high predication of four labels out of six.

In figure 02 (a) SGD classifier highest correct predication is (Label 4, 92%) and the lowest of (Label 5, 57%) (b)In MLP classifier shows the high correct prediction of (label 5, 95%) and low predication of (label 5, 52%). (c) 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 03 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 (a) MNB Classifier with respect to class 0,1,2,3,4,5 are (90%, 93%, 90%, 88%, 94%, 74%) and (b) RF classifier the precision results are (90%, 93%, 90%, 88%, 94%, 74%), and similarity (c) LSVC (90%, 93%, 90%, 88%, 94%, 74%) and (d) LG (90%, 93%, 90%, 88%, 94%, 74%). Furthermore, analyzing the overall performance system, the average Precision also computed, the average precision of the (RF) 90%, (RF) 90%, (RF) 90% and (RF) 92%.

|  |  |
| --- | --- |
|  |  |
| a) Confusion Matrix MultinomialNB | b) Confusion Matrix Random Forest |
|  |  |
| c) Confusion Matrix Linear SVC | d) Confusion Matrix Logistic Regression |

Figure 01: Confusion Matrix (a) MultinomialNB Classifier, (b) Random Forest Classifier, (c) Linear SVC Classifier, (d) Logistic Regression.

|  |  |
| --- | --- |
|  |  |
| a) Confusion Matrix SGDClassifer | b) Confusion Matrix MLPClassifer |
|  | |
| c) Confusion Matrix RidgeClassifer | |

Figure 02: Confusion Matrix (a) SGD classifier, (b) MLP Classifier, (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 03: Precision-Recall Curve (a) MultinomialNB Classifier, (b) Random forest Classifier, (c) Linear SVC Classifier, (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 RidgeClassifer | |

Figure 04: Precision-Recall Curve (a) SGD classifier, (b) MLP Classifier, (c) Ridge Classifier

Figure 04 depicts the Precision and Recall Curve of (a) SGD Classifier with respect to class 0,1,2,3,4,5 are (90%, 93%, 90%, 88%, 94%, 74%) and (b) MLP classifier the Precision results are (90%, 93%, 90%, 88%, 94%, 74%), and similarity (c) Ridge classifier (90%, 93%, 90%, 88%, 94%, 74%) .The average precision of the (RF) 90%, (RF) 90%, (RF) 90% and (RF) 92%.

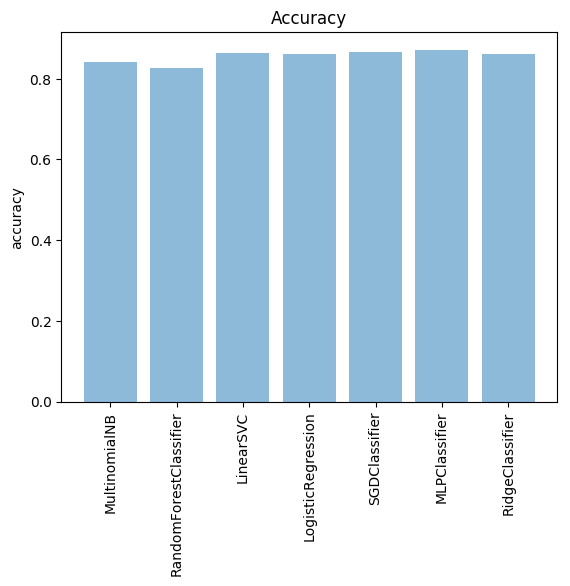


Figure 05: Results comparison of various Machine Learning Techniques.

Figure 05 shows the accuracy of different machine learning classification algorithms, such as MultinomialNB, LinearSVC, 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.

# CONCLUSION

This paper presents new model for automated online news text classification for Sindhi Language. 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 remaining 70% for training. In the study model, at 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 have been computed also. And then machine approaches have been implemented namely: MultinomialNB, LinearSVC, 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.

##### **References**

1. Rajan, K., et al. "Automatic classification of Tamil documents using vector space model and artificial neural network." Expert Systems with Applications 36.8 (2009): 10914-10918.
2. Ahmed, Kashif, et al. "Framework for Urdu News Headline Classification." (2016).
3. Ali, Mubashir, Shehzad Khalid, and Muhammad Haneef Saleemi. "A novel stemming approach for Urdu language." Journal of Applied Environmental and Biological Sciences 4.7S (2014): 436-443.
4. Lovins, Julie Beth. "Development of a stemming algorithm." Mech. Translat. & Comp. Linguistics 11.1-2 (1968): 22-31.
5. Porter, Martin F. "An algorithm for suffix stripping." Program (2006).
6. Ali, Mubashir, et al. "A rule based stemming method for multilingual Urdu text." International Journal of Computer Applications 134.8 (2016): 10-8.
7. Ali, Mazhar, and Asim Imdad Wagan. "An Analysis of Sindhi Annotated Corpus using Supervised Machine Learning Methods." Mehran University Research Journal of Engineering and Technology 38.1 (2019): 185-196.
8. Ali, Mazhar, and Asim Imdad Wagan. "Sentiment summerization and analysis of Sindhi text." Int. J. Adv. Comput. Sci. Appl 8.10 (2017): 296-300.
9. Wazir Ali J amra. " Sindhi Language Processing: A Survey" 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)
10. Wahab Khan, Ali Daud, Jamal A. Nasir, Tehmina Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP", Kuwait journal of Science, Vol. 43. No. 4, pp. 95-113, 2016.
11. Muhammad Usman, Saba Ayub, Zunaira Shafique, Kamran Malik, "Urdu Text Classification using Majority Voting", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 8, pp. 265-273, 2016.
12. Sharma, Neha, and Manoj Singh. "Modifying Naive Bayes classifier for multinomial text classification." 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE). IEEE, 2016.
13. Kaur, Gurvir, and Er Parvinder Kaur. "Novel approach to text classification by SVM-RBF kernel and linear SVC." International Journal of Advance Research, Ideas and Innovation in Technology 3.3 (2017).
14. Genkin, Alexander, David D. Lewis, and David Madigan. "Large-scale Bayesian logistic regression for text categorization." Technometrics 49.3 (2007): 291-304.
15. Basu, Subhadip, et al. "Handwritten Bangla alphabet recognition using an MLP based classifier." arXiv preprint arXiv:1203.0882 (2012).
16. Diab, Shadi. "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, Baoxun, et al. "An Improved Random Forest Classifier for Text Categorization." JCP 7.12 (2012): 2913-2920.
18. Genkin, Alexander, David D. Lewis, and David Madigan. "Sparse logistic regression for text categorization." DIMACS Working Group on Monitoring Message Streams Project Report (2005).