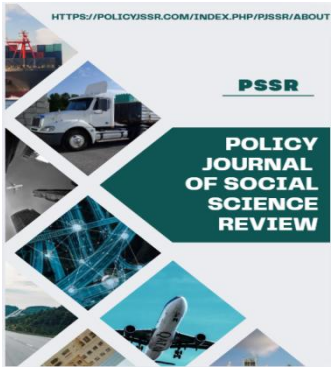
DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR SINDHI NEWS CLASSIFICATION

[*1]Muhammad Ameen Chhajro, [2]Abdul Qayoom, [3]Najma Imtiaz Ali, [4]Seema Sultana Bhurgri, [5]Zubair Uddin, [6]Aadil Jamali

[*1]Department of Software Engineering, Sindh Madressatul Islam University, Karachi
[2]Department of Computer Science, Sindh Madressatul Islam University, Karachi
[3]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan. Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal, Malaysia Melaka, Melaka, Malaysia.
[4]Assistant Professor, Department of Computer Science, Government Nazareth Girls Degree College, Hyderabad
[5]Department of Software Engineering, Sindh Madressatul Islam University, Karachi
[6]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Sindh, Pakistan

[*1]ameen.chhajro@smiu.edu.pk, [2]engr.qaziqayoom@gmail.com, [3]najma.channa@usindh.edu.pk, najma@utem.edu.my, [4]seema.bhurgri@gmail.com, [5]zubair@smiu.edu.pk, [6]aadil.jamali@usindh.edu.pk

Article Details

Received on 14 April, 2026

Accepted on 01 April, 2026

Published on 02 May, 2026

Corresponding Author: *

Muhammad Ameen Chhajro*

Page No: 27-38

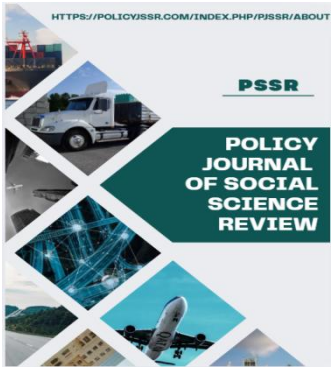
ABSTRACT

Sindhi language is one of the oldest languages of the Asia region, which still has insufficient computational advancement and availability of structured datasets for natural language processing tasks. This study performs a comparative analysis of various machine learning models for Sindhi news headline classification. A Sindhi news dataset of 87,553 headlines was collected from different online newspaper websites and blogs, which consists of two categories: technology and scientific discoveries news data. This research uses the text pre-processing techniques on Sindhi news data to clean the less meaningful data from the dataset, followed by feature extraction techniques to convert the text into a number representation. This research utilizes the machine learning models for classification tasks, including Logistic Regression, Linear SVM, Multinomial Naive Bayes, Radial Basis Function Support Vector Machine (RBF SVM), Random Forest, and Bernoulli Naïve Bayes, which were trained on the Sindhi news dataset, followed by evaluation considering the accuracy, precision, recall, and F1-score metrics. This proposed research study's experimental results reveal that the RBF SVM achieved an excellent accuracy of 98.91% along with precision (0.99) and F1-score (0.98) for the text classification task in Sindhi language news classification.

Keywords: *NLP, Text Classification, Machine Learning Algorithms, Low-resource language, Random Forest, Linear SVM, RBF SVM, Logistic Regression, Bernoulli Naïve Bayes, Sindhi language.*

DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

1. Introduction

The Sindhi language is thousands of years old and has millions of speakers worldwide **(Dootio & Wagan, 2019)**. Research still lacks in terms of benchmarks and a detailed comparative analysis between different machine learning models. However, some research studies have been conducted on classification models and their comparative studies using traditional machine learning approaches for classification tasks **(Soomro et al., 2025)**. This research study focuses on the text classification of Sindhi news headlines using machine learning techniques. Text classification is an important task in Natural Language Processing (NLP) that involves assigning text documents to predefined categories. One of the research studies explores the benchmark performance standards for Sindhi NLP tasks by comparing traditional machine learning approaches with transformer-based models **(Prakash et al., n.d.);** it still lacks some machine model inclusion. In this research, various machine learning models are trained to automatically classify Sindhi news headlines by their content and perform a detailed comparative analysis. The study explores how preprocessing and feature extraction contribute to the performance of classification algorithms. Natural Language Processing is the field of computer science and artificial intelligence that enables computers to understand, interpret, and process human language. NLP bridges the gap between human communication and machine understanding, allowing computers to analyze text or speech in a meaningful way. The success of a text classification system depends heavily on how well the text is prepared and represented.

1.1 Research Objectives

The main objectives of this study are as follows:

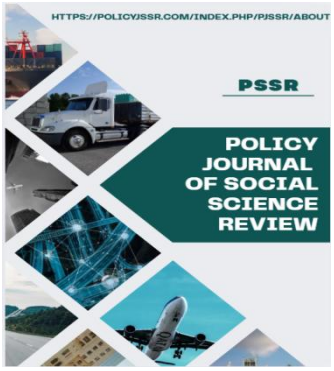
- To apply comprehensive text preprocessing techniques to clean and standardize Sindhi text data.
- To implement and analyze multiple feature extraction techniques for converting text into numerical representations.
- To perform the data analysis on the Sindhi text corpus.
- To train and evaluate different machine learning classification algorithms on the Sindhi language dataset.
- To compare the performance of all machine learning models and identify the best-performing approach for the Sindhi news text classification task.

1.2 Problem Statement

The Sindhi language, despite being spoken by over 30 million people globally, lacks robust computational resources and empirically validated machine learning models for automated text classification. The absence of benchmark performance standards, Linguistic complexity, its 52-character extended Perso-Arabic script, rich morphology with inflection patterns

DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

for Sindhi NLP tasks creates three critical problems. including its unique syntactic structures **(Petrov et al., 2013; Nivre, 2015)** has not been quantitatively assessed for its impact on the text classification performance of machine learning models. The research studies conducted previously do not clearly define the reference point or baseline models for evaluations of the performance accuracy that machine learning models can achieve on Sindhi text data in classification tasks, hindering progress in Sindhi language technology development.

2. Literature Review

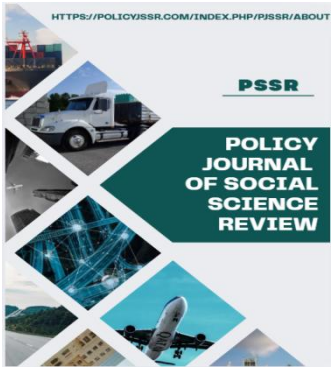
A recent literature review conducted on the Sindhi language classification task by **Soomro et al. (2025)** demonstrates that from 2017 to 2025, no comparative analysis has been performed on text classification based on the category of Sindhi news headlines using machine learning approaches. In another study conducted by **(Dootio & Wagan, 2019; Kandhro et al., 2019)** reveals that present research on Sindhi language in the domain of natural language processing mainly targets the preprocessing pipeline, and limited research work has been conducted in the domain of sentiment analysis in Sindhi language. However, few available studies on related low-resource languages, such as Urdu, have achieved classification accuracies ranging from 80% to 94% using traditional machine learning methods. However, different language has their own linguistic structure, syntax and language morphology, therefor they cannot be directly transmitted into Sindhi language classification task. Keeping in view such gap and limitation demands that there is a need of developing the new classification models implementation and evaluation particularly for Sindhi language classification. One of the research study previously conducted by **Ali et al. (2021) and Hammad & Anwar (2019)** on Sindhi dataset development via different online resources, and applied the text sentiment analysis on Sindhi language text data. Exploring the further literature review in (Ali & Xu, 2021; Arshad et al., 2022) revealed and evaluated the different machine learning models, including Support Vector Machine, Naive Bayes, and Logistic Regression for sentiment classification purposes in different languages, including Sindhi language. This research study further demonstrates that machine learning models perform well in combination with preprocessing and feature extraction techniques.

Another research conducted in Ali et al. (2019) on Sindhi text sentiment analysis using CNN models highlighted that deep learning methods can get semantic relationships from textual data and improve classification performance compared to traditional machine learning techniques.

In recent years, deep learning methods, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and Transformers, have gained popularity for text

DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

classification because they can automatically learn complex patterns from raw text data **Devlin et al., 2019**). The key idea behind the transformer's logic is to capture the context of the words based on their surrounding text data. The main base of Transformer models relies on the attention mechanism demonstrated by Vaswani **et al. (2017)**. **Prakash et al. (n.d.)**, who also performed the comparative study and concluded that transformers perform better than common machine learning models in the text sentiment classification tasks for low-resource languages, and achieved higher accuracy due to their ability to understand the context from the text data. The conclusion from the literature regarding the machine learning tasks indicates that while traditional machine learning techniques provide a basis for sentiment classification tasks **(Arshad et al., 2022; Kandhro et al., 2019)**, in contrast, the transformer-based models enhanced the contextual understanding and performs outstanding for low-resource languages **(Prakash et al., n.d.; Soomro et al., 2025)**. However, research on Sindhi news sentiment classification is still limited, highlighting the need for comparative studies that evaluate both machine learning and transformer models for better sentiment analysis in Sindhi textual data **(Dootio & Wagan, 2019)**. From the literature, it has been concluded that there is still limited work that has been conducted on the text classification task on Sindhi Language corpora.

3. Research Methodology

This proposed research follows a specific methodology pipeline, which includes data cleaning, data analysis, feature extraction, machine learning models training, evaluation, and comparative results analysis as illustrated in Figure 1 below. Each phase of this research is wisely designed to achieve excellent performance on the text classification task in the Sindhi language dataset.

DOI:

Volume. 4, issue No. 5 (2026)

Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

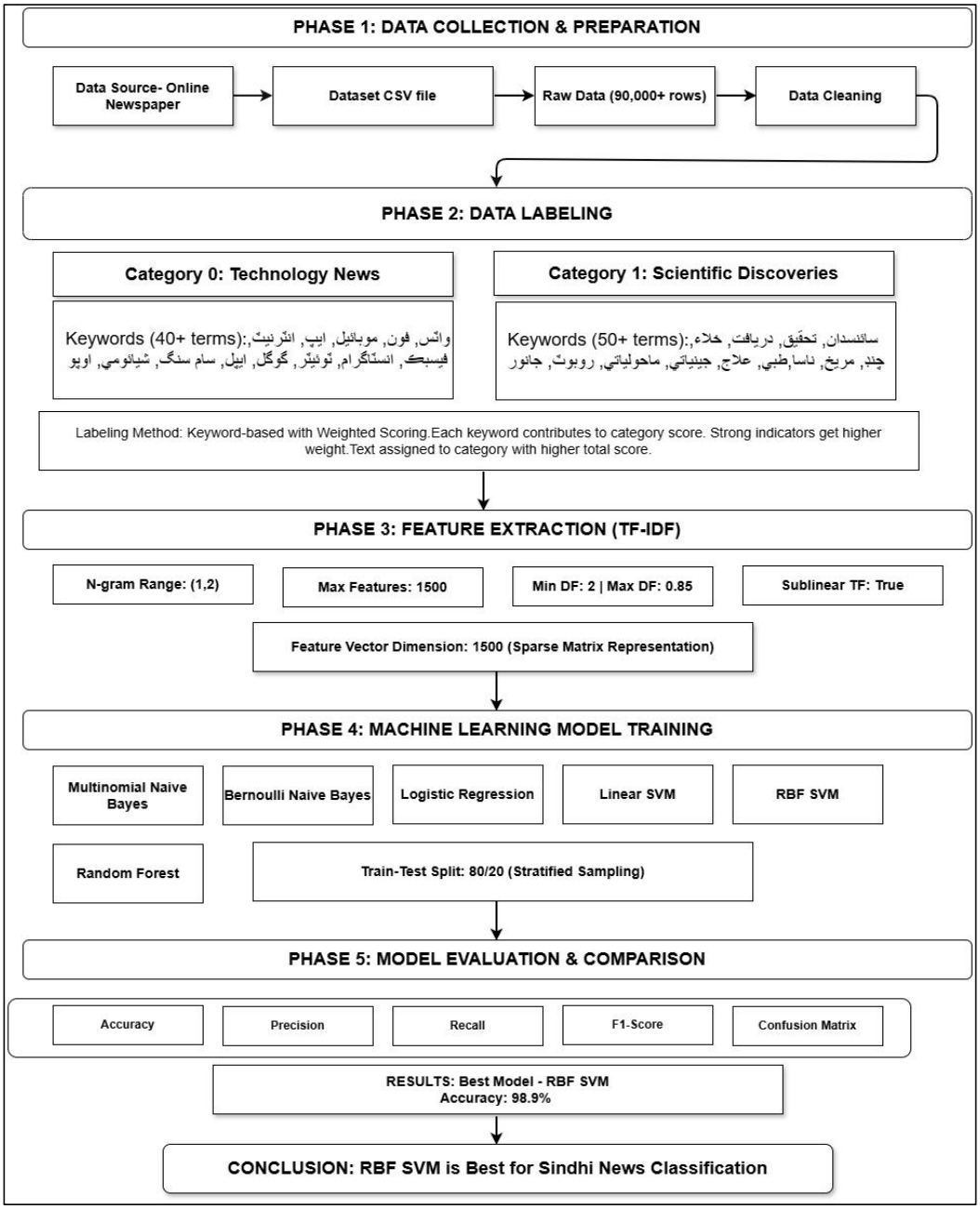


Figure 1. A Systematic Research Methodology Approach for Sindhi News Classification

3.1 Dataset Description

The scraped dataset of the Sindhi language utilized in this research consists of 87553 Sindhi headline news. The different sources used for data collection include the Sindhi newspaper websites and blogs, followed by two main data categories, including the scientific news

DOI:

Volume. 4, issue No. 5 (2026)

Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

and technology news. The structured form of the language dataset makes it suitable for preprocessing, feature extraction, data analysis, and model training.

3.2 Pre-Processing Techniques

The preprocessing techniques are a very crucial step for data cleaning and contribute more to the model performance and evaluation. Utilization and adoption of preprocessing techniques effectively results the more accurate models prediction in the classification task. Some of the most common pre-processing techniques used in this research study are: Tokenization: Breaking text into individual words or tokens. Stop word Removal: Removing common words. Feature Extraction: Representing text as numbers using techniques like Bag of Words, TF-IDF, or Word Embeddings. In this research, text preprocessing was performed using a step-by-step pipeline to clean and standardize the Urdu news headlines for machine learning.

3.2 Exploratory Data Analysis

After applying preprocessing techniques, the exploratory data analysis was performed on the Sindhi dataset, and the data visualization indicated the data misbalancing issue within the dataset categories. Considering the problem, the dataset was balanced as illustrated in Figure 2. Figure 2 shows the dataset category distribution, Word count, and text data length through box plot representation.

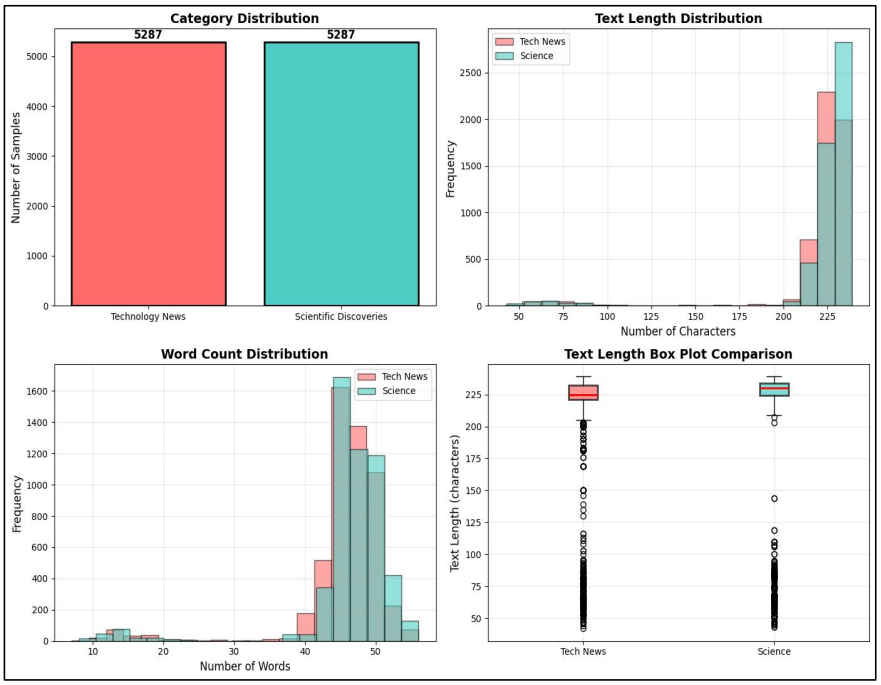
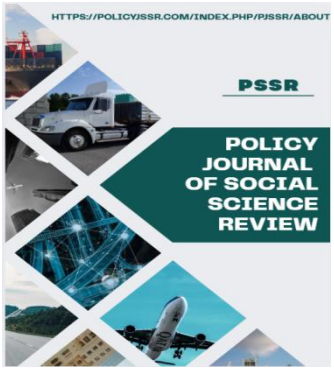


Figure 2. Exploratory Data Analysis of Sindhi News Dataset

DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

3.4 Feature Extraction Techniques

After preprocessing, feature extraction techniques were applied to convert text data into a numerical representation so that the text data can be made understandable for training the machine learning models. The various feature extraction techniques were implemented in this research study, including the One-Hot Encoding, Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), and Word Embeddings.

3.5 Machine Learning Models

The adoption of machine learning classification models in this research study was revealed by some previous research studies, which were conducted on low-resource languages, and the comparative analysis was performed, including the Urdu text classification **Arshad et al. (2022)** and Sindhi syntactic parsing techniques **Dootio & Wagan (2019)**.Considering the research methodology, after feature extraction techniques, different machine learning models were trained on text data features for the classification tasks on Sindhi news. The machine learning models that were trained for the classification task in this research are Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Linear SVM, RBF SVM, and Random Forest. The performance of each machine learning model on Sindhi news text headlines classification has been discussed in the following sections of this research study.

4. Results and Discussion

This section of the research demonstrates the experimental results, comparative analysis, and discusses the performance of machine learning techniques for the Sindhi news headlines text classification task. In order to monitor the performance of different machine learning classifiers, we use the precision, recall, and F1-score.

Precision (Positive predictive value), it is the performance measuring indicator of machine learning classifiers, which evaluates the correctly predicted data samples from all predicted positives.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (1)$$

Where TP is (True Positive), FP is (False Positive). A high precision value indicates a low false positive rate.

Recall (Sensitivity/True positive rate): Recall measures a classifier's performance by the proportion of actual positive samples it correctly identifies. Meaning how many data samples the model correctly identified.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (2)$$

Where TP is (True Positive), FN is (False Negative). If we get high recall values meaning that low false negatives.

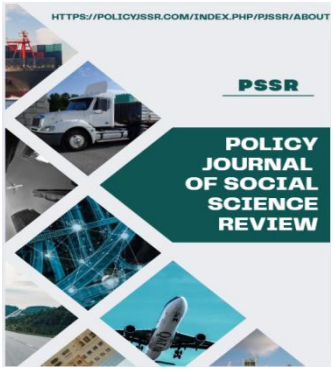
F1-score: In terms of performance evaluation of classifiers, it is called a harmonic mean of precision and recall.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (3)$$

The F1-score is very important when we need a single metric that maintains the balance of both; it is a good choice, particularly when we have imbalanced

DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

datasets. Here is the performance Table 1, which shows the results of the accuracy, Precision, Recall, and F1-score, and a comparative study of different machine learning models on the Sindhi news headlines dataset.

Table 1

Performance Metrics of Machine Learning Models on Sindhi News Headlines

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	92.96%	0.946	0.911	0.928
Bernoulli Naive Bayes	90.87%	0.948	0.864	0.904
Logistic Regression	97.68%	0.966	0.987	0.977
Linear SVM	98.82%	0.987	0.988	0.988
RBF SVM	98.91%	0.995	0.983	0.989
Random Forest	94.66%	0.912	0.987	0.948

The experimental results mentioned in Table 1 show that the RBF SVM has achieved the highest accuracy and performance overall in comparison to other machine learning classification models by achieving the 98.91% accuracy and exceptional outcome of precision (0.995) on the Sindhi news headlines text data. One important aspect coming from this model's outcome is that it has maintained the perfect balance between precision and recall on both classes, which makes it a more reliable classifier for the Sindhi news text classification task. In follow-up, the Linear SVM is at 2nd highest accuracy value achiever (98. 82%). Logistic Regression has also performed well, and along with strong recall results (0.9877), which shows that less likely to miss the positive sample cases. However, the Logistic Regression has achieved a slightly lower precision than the SVM models. The Random Forest classifier also obtained good accuracy (94.66), excellent recall (0.9877), and lower precision performance (0.913), which shows that the falser positive rate. The experimental results show that among all machine learning classifiers, the Multinomial Naïve Bayes and Bernoulli performances are moderate in terms of accuracy and precision values, as illustrated in Table 1 above.

4.1 Confusion Matrix Analysis

The experimental results and analysis of confusion matrices have been illustrated in Figures 3, 4, and 5 below. Looking closely at the confusion matrices outcome, it provides the overall performance of classification tasks of

DOI:

Volume. 4, issue No. 5 (2026)

Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

each machine learning model by showing the correct and incorrect predictions output through distribution across both categories, including the technology news and scientific discoveries. Figure 3 shows the analysis of Multinomial Naïve Bayes and Bernoulli Naïve Bayes machine learning classifiers. The confusion matrix illustrated that the Multinomial NB classified 1,005 correctly (True Positive) from technology news headlines text data, while achieving 960 scientific discoveries (True Positive) samples classified. However, the classification results from confusion matrices show that 97 text data headings from the technology news were misclassified and the model considered them as science news (False Negatives), similarly, 53 science news headlines were misclassified as technology news. Bernoulli Naïve Bayes predicts a more number of false negatives (143), revealing that the binary presence or absence feature representation fails to capture the frequency importance of distinguishing terms in Sindhi text. This indicates that Multinomial NB struggles more with identifying Science category instances, likely due to overlapping vocabulary between scientific terminology and technology terms. Figure 4 shows the confusion matrix analysis of Linear Support Vector Machine and Logistic Regression. Logistic Regression performance is remarkably good, with only 13 false negatives and 35 false positives from 2,115 test cases. The diagonal results of the confusion matrix show that two categories are being effectively separated by the linear decision boundary.

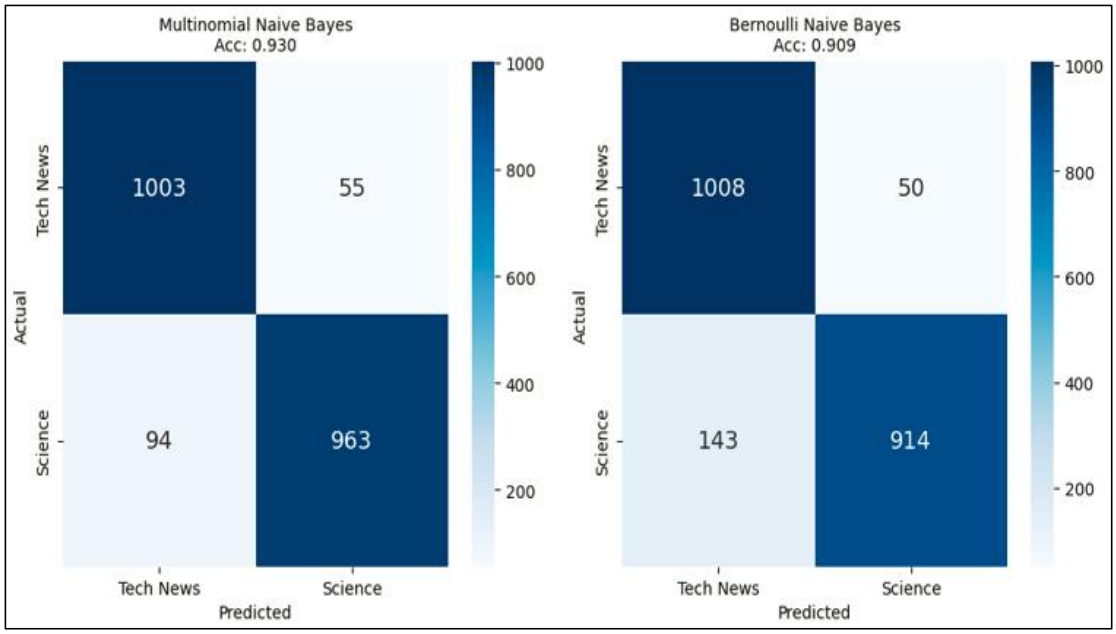


Figure 3. Confusion Matrix of Multinomial Naive Bayes and Bernoulli Naive Bayes

DOI:

Volume. 4, issue No. 5 (2026)

Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

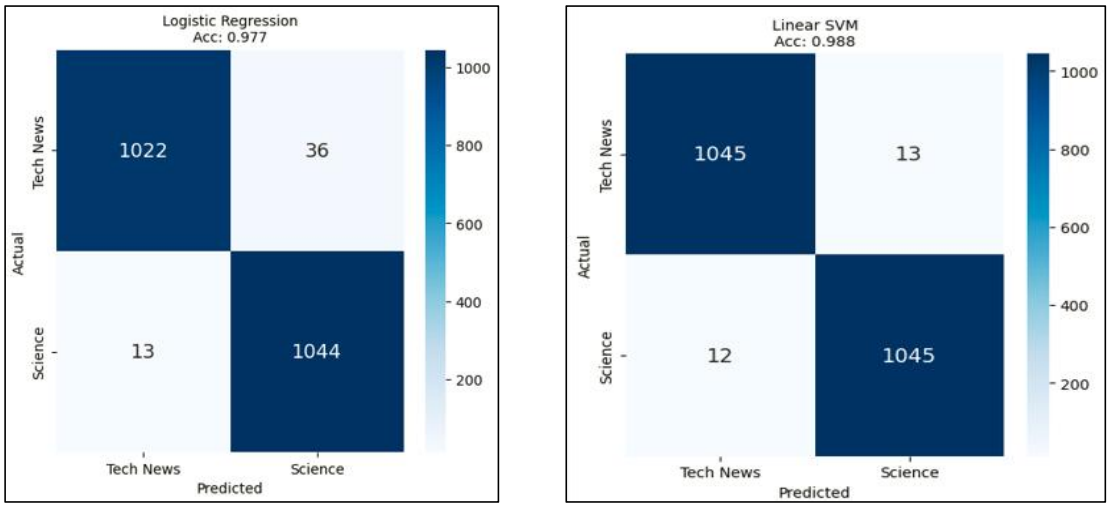


Figure 4. Confusion Matrix of Logistic Regression and Linear SVM

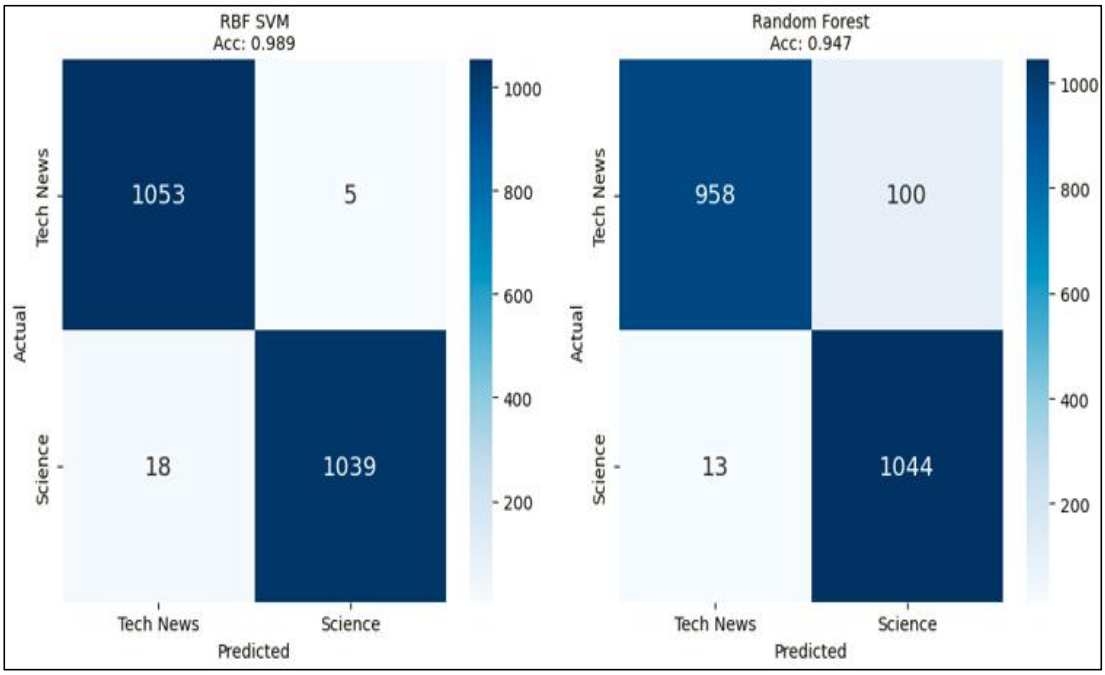
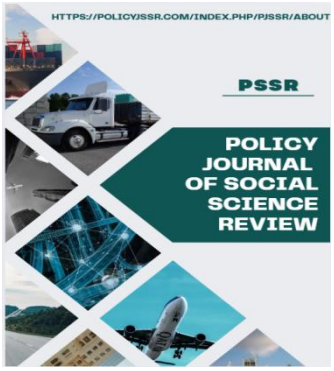


Figure 5. Confusion Matrix of RBF SVM and Random Forest

DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

Table 2
Summary of Confusion Matrix Findings

Model	True Positives (Tech→Tech)	True Negatives (Sci→Sci)	False Positives	False Negatives
Multinomial NB	1,005	960	53	97
Bernoulli NB	1,005	914	53	143
Logistic Regression	1,026	1,041	17	31
Linear SVM	1,047	1,046	11	12
RBF SVM	1,058	1,038	19	18
Random Forest	962	1,046	97	11

5. CONCLUSION

This research study demonstrates that the Radial Basis Function (RBF SVM) achieved the highest accuracy of 98.91%, and with a precision of 0.995 among all machine learning models evaluated in this research. These experimental results show that traditional machine learning models proved highly effective for the Sindh news headlines classification task. This research also frames a benchmark for new research in the field of natural language processing (NLP). The research findings conclude that proper preprocessing, feature extraction, data analysis, and an appropriate choice of machine learning classifier can achieve exceptional results despite having limited digital resources in low-resource languages.

6. Future Work

After conducting this research, multiple directions have evolved for future research and advancing the NLP applications in the Sindhi language, including the other low-resource languages. The current language dataset can be extended to multiple classes, making it suitable for broader applications in the NLP domain. The future research should explore the different deep learning models, including LSTM, BiLSTM, and CNN models, for better sequence capturing from Sindhi language text data. For the text classification task, a pre-trained multilingual model can be fine-tuned on Sindhi text, like mBERT, XLM-RoBERTa, and DistilRoBERTa can further improve the accuracy of the classification task in Sindhi language. This research study directs there is gap for advancement in computational languages, including Sindhi and other udder-developed languages in the region.

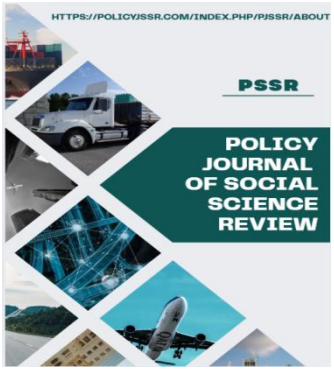
References

[1] M. A. Dootio and A. I. Wagan, "Syntactic parsing and supervised analysis of Sindhi text," *Journal of King Saud University – Computer and Information Sciences*, vol. 31, no. 1, pp. 105-112, 2019.

[2] S. A. Soomro, S. S. Yuhaniz, M. A. Dootio, G. Mujtaba, and J. A.

DOI:

Volume. 4, issue No. 5 (2026)



Policy Journal of Social Science Review

ISSN Online:3006-4635

ISSN Print: 3006-4627

Siddiqui, "Category-Based Sentiment Analysis of Sindhi News Headlines Using Machine Learning, Deep Learning, and Transformer Models," *IEEE Access*, vol. 13, 2025.

[3] K. J. Prakash, C. Anjali, R. Indira, M. K. Kumar, and K. Pushkar, "A Comparative Analysis of Machine Learning and Transformer Models for Sindhi News Sentiment Classification," *International Journal of Data Science and IoT Management System*, vol. 5, no. 2.

[4] W. Ali, N. Ali, Y. Dai, J. Kumar, S. Tumrani, and Z. Xu, "Creating and evaluating resources for sentiment analysis in the low-resource language: Sindhi," in *Proc. 11th Workshop on Computational Approaches to Subjectivity*, 2021, pp. 188-194.

[5] I. A. Kandhro, S. Z. Jumani, A. A. Lashari, S. S. Nangraj, Q. A. Lakhan, M. T. Baig, and S. Guriro, "Classification of Sindhi headline news documents based on TF-IDF text analysis scheme," *Indian Journal of Science and Technology*, vol. 12, no. 33, pp. 1-10, 2019.

[6] M. Hammad and H. Anwar, "Sentiment analysis of Sindhi tweets dataset using supervised machine learning techniques," in *Proc. 22nd International Multitopic Conference (INMIC)*, 2019, pp. 1-6.

[7] W. Ali and Z. Xu, "SiPOS: A benchmark dataset for Sindhi part-of-speech tagging," in *Proc. Student Research Workshop Associated with RANLP*, 2021, pp. 22-30.

[8] W. Ali, J. Kumar, J. Lu, and Z. Xu, "Word embedding based new corpus for low-resourced language: Sindhi," *arXiv preprint arXiv:1911.12579*, 2019.

[9] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," *arXiv preprint arXiv:1104.2086*, 2013.

[10] J. Nivre, "Towards a universal grammar for natural language processing," in *Computational Linguistics and Intelligent Text Processing*, Springer, 2015, pp. 3-16.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[12] A. Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017.

[13] U. Arshad, K. I. Malik, and H. Arooj, "Urdu news content classification using machine learning algorithms," *LGU Research Journal of Computer Science & IT*, vol. 6, no. 1, 2022.