Supervised machine learning techniques for automated classification of Sindhi poets.

Ronit Kumar Kataria
Habib University
Email: rk06451@st.habib.edu.pk

Ragini Gopchandani
Habib University
Email: rg05951@st.habib.edu.pk

Shah Jamal Alam
Habib University
Email: sj.alam@sse.habib.edu.pk

***Abstract*—Poet classification is a crucial task that involves identifying the authorship of a specific poem by analyzing its distinct features. This task carries significant importance for establishing a connection between poetry and its original author, as well as for addressing cases of plagiarism. The main objective of our research study is to construct an accurate representation of Sindhi ghazals to capture and differentiate the unique writing styles of different poets by categorizing poetry couplets. This study explores the classification of Sindhi poets, with a specific focus on a collection of Sindhi couplets. Using supervised machine learning models, our objective is to create a strong framework for accurately determining the authorship of Sindhi ghazals. Our research aims to enhance the understanding and conservation of the extensive Sindhi poetry tradition through the implementation of a methodical approach to classifying poets.**

***Index Terms*—classification, deep learning, sindhi poetry, attribution**

I. Introduction

Sindhi poetry is a constantly evolving and culturally significant genre of literature that derives from the Sindh region, located in present-day Pakistan, as well as certain areas of India. Sindhi poetry is a significant component of Sindh's cultural history, characterized by a deep and varied tradition that has endured for centuries. When we engage with Sindhi poetry, we are not only analyzing words; rather, we are delving into the very core of a community's collective identity.

Poet attribution is the act of determining the poet who is accountable for composing a certain poem. It is a specific field within the study of authorship attribution, which is the broader topic of determining the author of a certain document. This process is accomplished by discerning the distinct writing styles exhibited by various authors. A frequently employed method involves the utilisation of stylometric traits. Stylometric aspects refer to statistical metrics that quantify the language usage in a text, including the occurrence rate of specific words or phrases, the utilisation of particular grammatical structures, and the overall sentence length. In th Poet Attribution classification, we have used couplet as a single data stream with a maximum length of 27 words for training and testing. Due to the restrictions of ghazal rhymes, some words in these couplets may be repeated.

Our motivation for this research stems from a noticeable gap in preserving Sindhi literature. Numerous Sindhi Ghazals are currently preserved in physical books, and a significant portion of the classic compositions remain, but with unknown authors. In the contemporary age of social media, numerous couplets attributed to Sindhi poets are widely circulated for the purpose of garnering likes and shares. However, the authenticity of these couplets cannot be verified. Only a limited amount of work has been conducted in the field of Sindhi Language. Therefore, this research will provide us with the opportunity to explore Sindhi Literature in more depth. we aim to employ machine learning techniques such as Support Vector Machine, K-Nearest Neighbour, Naive Bayes, and Decision Tree Classifiers to train these models on a dataset of Sindhi Ghazal and accurately determine the true author of a couplet. Moreover, this study aims to elucidate the distinctions between conventional poetic structures and contemporary expressions in Sindhi poetry, shedding light on the evolution of writing methodologies and theme elements over various periods.

Our research on Sindhi poetry classification using supervised machine learning represents a pioneering endeavor in a relatively unexplored domain. Surprisingly, there has been a notable absence of prior work in the identification of Sindhi poets through computational techniques. Therefore, our study serves as foundational groundwork, laying the initial framework for future investigations in this field. By employing cutting-edge technologies, we aim to not only shed light on the distinctive features of Sindhi poetry but also open avenues for further exploration, providing a critical starting point for subsequent research endeavors in Sindhi literary analysis.

II. Literature Review

Poetry classification task has been carried out in multiple languages such as Persian [1], Malay [2], Urdu [3], and others. However, when it comes to Sindhi, there has been no prior work specifically focused on Sindhi poetry. To extract the features from the poetry and classify them based on poets, various techniques have been utilized. For lengthy texts like ghazals, Support Vector Machine yields higher accuracy in classification [3].

The study [3] trained various models—SVM, Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbor. They employed chi-square and L1-based feature selection methods. About 4000 ghazals were used for training the model. Specifically, they used complete ghazals attributed to each poet within the dataset. 80 % of the data was used for training, while the remaining 20% was for testing. SVM outperformed all other

models, achieving the highest F1-measure score of 72% with or without feature selection.

Additionally, a scholarly article [4] focused on attributing poets in the Urdu language. The dataset comprised 18,472 pairs of lines from 15 renowned poets. This article takes a close look at how deep learning models are being used in Urdu poetry attribution. Four models namely, Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN) underwent preprocessing. Additionally, Transformer-based models like BERT and roBERTa were explored. Each model's parameters were clearly defined, giving a detailed assessment of their effectiveness in analyzing Urdu literature. Results showed the Support Vector Machine (SVM) achieving 64% accuracy, while the Random Forest model scored 25%. LSTM models outperformed other deep learning methods, and Transformer-based models, especially BERT, stood out with exceptional 80% accuracy. Notably, BERT's attention mechanism and encoder-based architecture produced precise outcomes for each category.

We encountered another study [1] focusing on classifying Persian Ghazals using sequential learning. The paper emphasized the poets' eras, by developing a model that classified the chronological order of each poem based on the author's lifespan, particularly featuring the renowned poet Hafez. The dataset comprised 496 samples, each including Persian and English versions, six chronological labels, and a Raad Label. They pre-processed the data and applied Word Embedding to vectorize the couplets for classification models. Essential data features were extracted using Bag of Words (BoW) and Latent Dirichlet Allocation (LDA). The model's output involved class labels (Chronological and Raad) applied to both Persian and English texts. Both Machine Learning and Deep Learning models were employed for the classification task.

With the context of Machine Learning, we came across study [7]. They used the XGBoost algorithm to create a modern Chinese poetry style classification model (XGBoost-MCP). They worked with 836 samples, preprocessing them by labeling, word segmentation using Jieba, and removing stopwords. They applied Doc2Vec for feature extraction to reduce data dimensionality. The study compared four classification methods namely XGBoost-MCP, SVM, DNN, and DT. The XGBoost-MCP model showed exceptional performance across various evaluation criteria: accuracy (93.62%), precision (94.09%), recall (93.69%), and F1-score (93.65%). XGBoost-MCP outperformed SVM, DNN, and DT by significant margins, confirming its preference as a classification model. While Decision Trees (DT) had an impressive performance in specific datasets, XGBoost-MCP consistently emerged as the top-performing model.

The literature review highlights various significant classification models, including Support Vector Machine, Multilayer Perceptron, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbour, Convolutional Neural Network, Long Short-Term Memory, Gated Recurrent Unit, Transformer-based models like BERT and roBERTa, and XGBoost. These models have been widely utilized to classify poetry in multiple languages, showcasing their versatility and effectiveness in literary analysis.

Additionally, we thoroughly examined related literature that delves into similar classification tasks. The following section outlines the dataset used and summarizes model accuracies, displayed in a table. I

TABLE I
SUMMARY OF MODELS, DATASET, AND ACCURACY

Paper	Model	Dataset	Language	Accuracy
[1]	DMM & LSTM	495	Persian	85%
[2]	SVM + Uni&Biagram	32667	Urdu	88.70%
[2]	Naïve Bayes Classifier	32667	Urdu	77%
[3]	SVM	4000	Urdu	72.00%
[4]	MLP(ANN)	18472	Urdu	64%
[4]	Random Forest	18472	Urdu	25%
[4]	BERT	18472	Urdu	80%
[5]	Classifier	Kavan'	Gujrati	87%
[6]	MLP(ANN)	314	Arabic	81%
[6]	CNN (DNN)	314	Arabic	61%
[7]	XGBoost-MCP	836	Chinese	93.62%
[7]	Support Vector Machine	836	Chinese	87.29%
[7]	Decision Tree	836	Chinese	90%
[7]	Deep Neural Network	836	Chinese	86.85%

III. METHODOLOGIES

In this section, we present key stages of our proposed framework of Sindhi poet classification, where we have focused on important aspects such as datasets, model preparation and their respective experiments of Fine tuning.

A. Dataset Curation

We have collected corpus of around 2500 couplets for six Sindhi poets as as enumerated in the table II. To avoid class imbalance, we strictly followed the principle of equal data theory to gain more accurate classification results. A couplet has two lines which may be of equal or variable length. A Sindhi Ghazal consists of multiple couplets. We have scraped the poet data from different Sindhi sites. Some of the major links we got our data from are

1) https://poetofsindh.blogspot.com/,
2) https://sindhishayari.blogspot.com/,
3) http://www.sindhiadabiboard.org/.

1) *Acquisition:* We conducted a thorough search for several websites pertaining to Sindhi literature, specifically focusing on locating Ghazals composed by renowned Sindhi poets. In the aforementioned links, we utilised the *Beautiful Soup* and *Requests* library to perform web scraping on the websites. However, we manually duplicated the couplets from certain sources.
2) *Characteristics:* We have selected the poets based on following two factors:
   a) How known a particular poet is in the Sindhi Literature and how significant their contribution is.
   b) The amount of data available online for that particular poet from reliable sources.

TABLE II
CORPUS BREAKDOWN

No	Poet Name	No of Couplets
1	Shah Abdul Latif Bhittai	500
2	Sheikh Ayaz	500
3	Ustad Bukhari	488
4	Masroor Pirzado	500
5	Adal Soomro	500

B. Machine Learning Models

In this study, we will undertake an examination and implementation of various classification models. The model will receive a couplet extracted from any poetries as its input, and its output will consist of the class label corresponding to the poet. The training involved the following process in a sequential Order:

1) Data Prepossessing:

a) Data-Cleaning: The scrapped data included various impurities, including stop words and extraneous symbols. The data was cleaned with removal of linguistic and non essential artifacts to enhance the quality and integrity of underlying Sindhi Poetry corpus.

b) Label Encoding: We utilized the label encoding technique to assign numerical labels to classes, starting from 0 to $n-1$ where n represents the number of classes we have in the dataset. In our case, we are predicting 6 unique poets so class labels will range from 0 to 5.

c) Feature Extraction: To train machine learning model. we need numerical feature information from raw data. For that we used Term Frequency-Inverse Document Frequency (TF-IDF) to convert a single couplet into a vector of word embedding.

d) Testing and Training Split: we divided our dataset into three separate subsets.The training set, consists 70% data, helped the learning of the model, while the testing set, containing 20% instances, was used to evaluate the model's ability to generalise. The validation set comprised of 10%, which was not used during training, played a vital role in preventing over fitting and optimising the model for optimal performance.

2) Hyper-Parameters: The experimented models were fine tuned using **GridSearchCV** which is a technique to try differnet set of parameters and find the optimal.

a) Stochastic Gradient Descent: We Utilized SGDClassifier with hinge loss, max iteration as 1000, and a random seed of 42. We Trained the model in batches of 64, achieving optimal performance on sequence classification tasks

b) Random Forest: We also applied a Random Forest Classifier with 100 estimators and a random seed of 42 with each tree of height 7.

c) XGBoost: We also experiments XGBoost classifier, with 100 estimators and a random seed of 42 that combines the predictions of multiple decision tree models to enhance predictive accuracy and generalization performance.

d) Logistic Regression: A logistic Regression model was also employed with regularization parameter of c as 1 with 100000 iterations.

e) Naive Bayes: In addition, a multinomial naive Bayes algorithm with alpha as *0.5* and fit prior as *False*, was employed for this objective. The Naive Bayes algorithm classifies data by applying Bayes' theorem under the premise of conditional independence between each pair of class attributes. Afterward, it employs maximum likelihood estimation to generate predictions.

f) CatBoost: To improve predictive accuracy and generalization, we utilized the CatBoost Classifier with 100 iterations, a depth of 5, a learning rate of 0.1, and a loss function of 'MultiClass'.

g) LightGBM: Finally, we utilised the LightGBM Classifier with the 'multiclass' objective and a random state of 42.

IV. RESULTS

Here, we will analyze the results of our investigation into different machine-learning models to identify the authors of Sindhi poetry. Various models were evaluated, including Stochastic Gradient Descent (SGD), Random Forest, XGBoost, Logistic Regression, Naive Bayes, CatBoost, and LightGBM. We will look at their accuracy scores to evaluate their ability to identify the poets by distinguishing their distinctive writing styles.

A. Model Performance Metrics

1) Stochastic Gradient Descent: The SGD model achieved the highest accuracy rate of 88.1%.

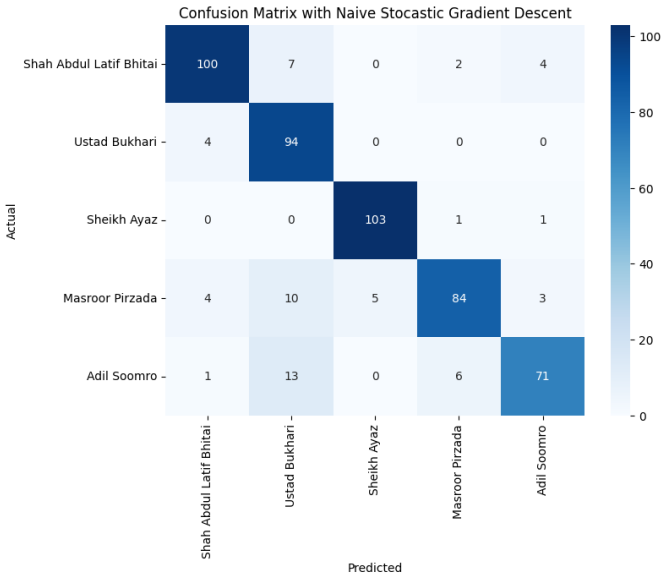

Fig. 1. Confusion Matrix for Stochastic Gradient Descent

2) Naive Bayes: The Naive Bayes classifier attained an accuracy of roughly 86.74%.

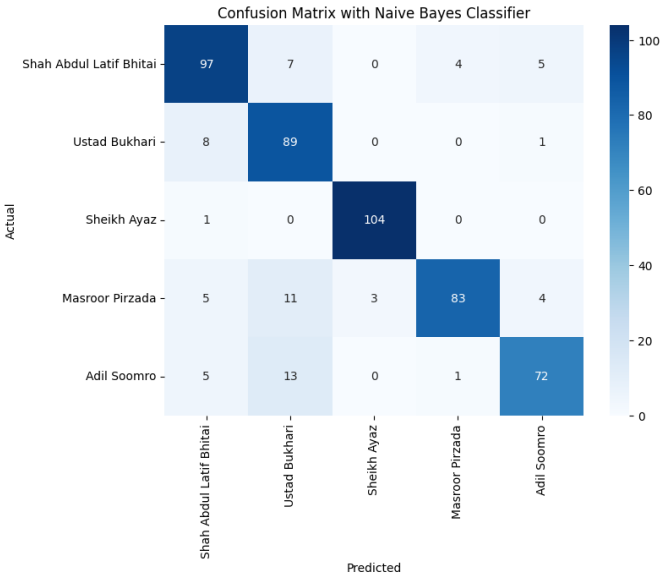

Fig. 2. Confusion Matrix for Naive Bayes Classifier

3) Random Forest: Random Forest showed an accuracy of 71.2%.

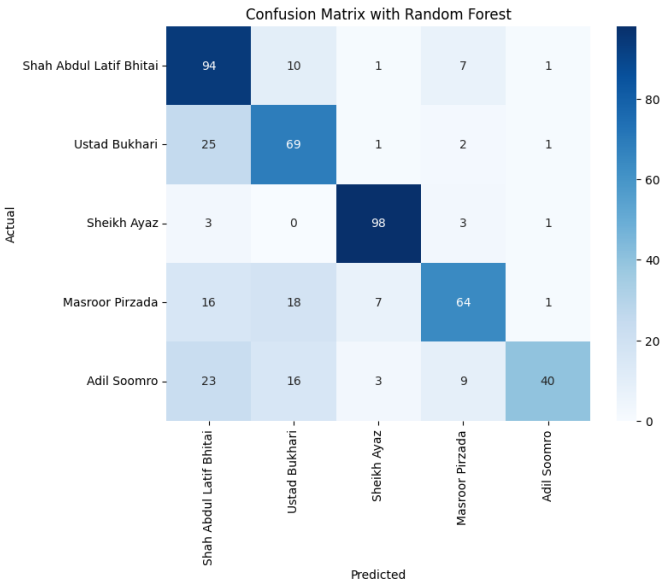

Fig. 3. Confusion Matrix for Random Forest Classifier

4) XGBoost: The XGBoost Classifier showed a 68% accuracy rate.

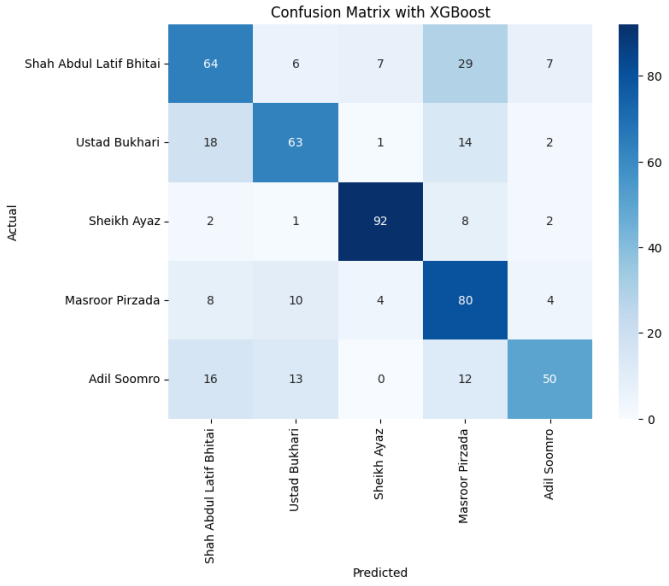

Fig. 4. Confusion Matrix for XGBoost Classifier

5) CatBoost: CatBoost Classifier showed an accuracy of 61.8%.

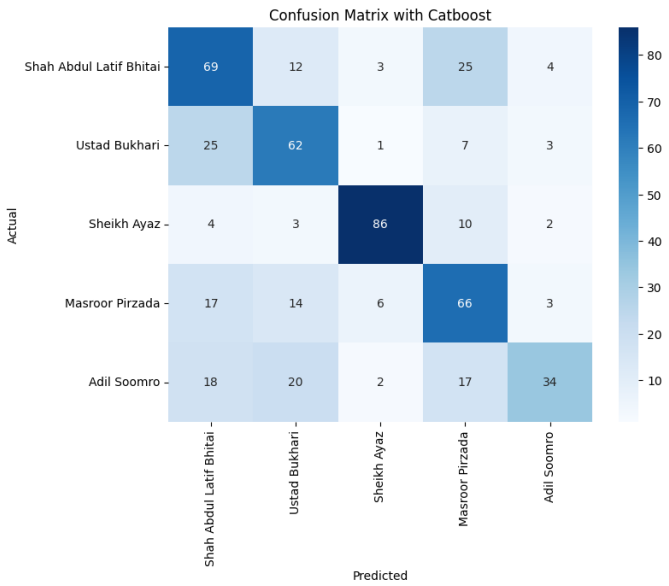

Fig. 5. Confusion Matrix for CatBosst Classifier

6) LightGBM: LightBGM classifier showed an accuracy of 60%.

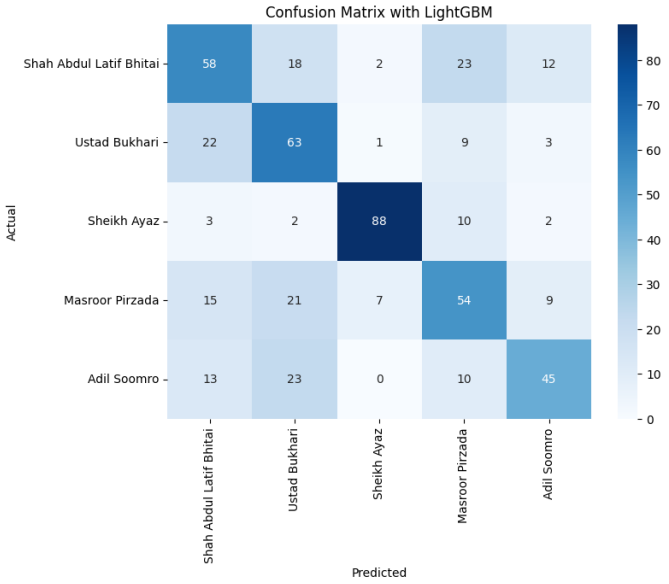

Fig. 6. Confusion Matrix for LightGBM Classifier

7) Logistic Regression: The Logistic Regression model had an accuracy rate of 84.6%.

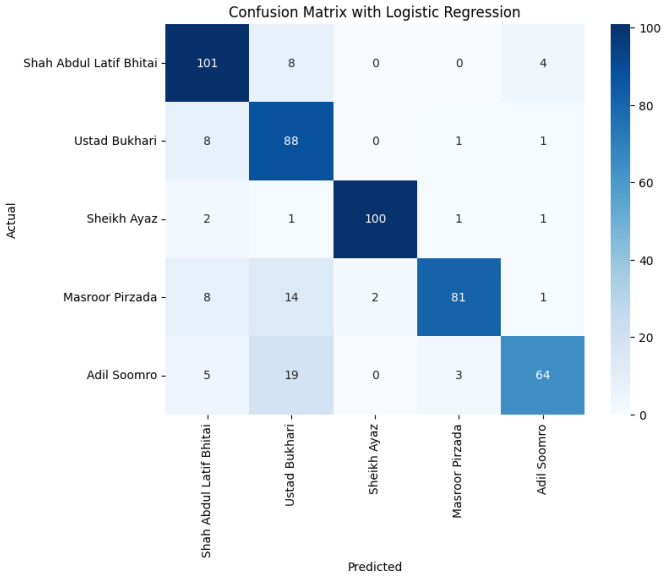

Fig. 7. Confusion Matrix for Logistic Regression Classifier

B. Summary of Results: Assessment of various classifiers in the task of assigning authorship to Sindhi poetry

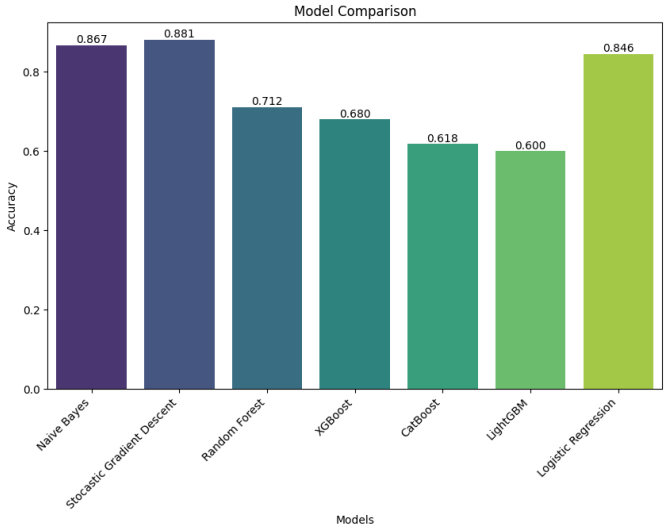

Fig. 8. Comparison of Model Accuracies

V. Discussion

We have conducted training and evaluation on a range of machine-learning models to classify Sindhi poetry couplets based on the poet. The range of accuracies observed in these models, which varied from 60.0% to 88.1%, emphasizes the influence of algorithm selection on the classification of Sindhi poetry.

The Stochastic Gradient Descent (SGD) and Logistic Regression models exhibited higher accuracies of approximately 88.1% and 84.6% respectively, possibly due to their efficacy for text sequence categorization and adeptness in handling intricate interactions within the data. On the other hand, Random Forest, XGBoost, CatBoost, and LightGBM models demonstrated less accurate results, ranging from 60.0% to 71.1%. The reason for this limitation could be that ensemble-based models struggle to capture the intricate linguistic patterns found in the Sindhi poetry corpus, as they primarily emphasize variables specific to the text. Furthermore, the Naive Bayes model achieved a commendable accuracy rate of around 86.7%, utilising the concept of conditional attribute independence.

Furthermore, our dataset, consisting of 2500 Sindhi poetry couplets authored by 6 poets, probably had an impact on the observed accuracies. SGD and Logistic Regression models, which can handle larger datasets, exhibited superior accuracies due to the utilization of a vast data corpus. On the other hand, ensemble models may have encountered restrictions because of the comparatively smaller group of data inside the 2500 couplets. This could have affected their capacity to identify subtle linguistic characteristics, leading to lower levels of

accuracy. Therefore, it is evident that the selection of the algorithm, the size of the dataset, and the performance of models are interrelated.

VI. Conclusion

We considered classifying couplets of Sindhi poetry based on respective poets using diverse machine-learning models. The accuracies ranged from 60.0% to 88.1%. Models such as stochastic gradient descent and logistic regression were good at processing text sequences, but ensemble models faced challenges in capturing the intricate linguistic patterns within the Sindhi poetry. With our dataset of 2500 couplets, some models worked very well, highlighting that the choice of model, the dataset size, and model performance, are all interconnected. Thus, by looking at our findings, we can see there is a need for personalized strategies in future investigations within this area.

References

[1] J. F. Ruma, S. Akter, J. J. Laboni, and R. M. Rahman, "A deep learning classification model for Persian hafez poetry based on the poet's era," Decision Analytics Journal, vol. 4, p. 100111, September 2022.

[2] M. A. Rao and T. Ahmed, "Poet attribution for urdu: Finding optimal configuration for short text," KIET Journal of Computing and Information Sciences, vol. 4, no. 2, p. 12, 2021

[3] N. Tariq, I. Ejaz, M. K. Malik, Z. Nawaz, and F. Bukhari, "Identification of Urdu Ghazal Poets using SVM," Mehran University Research Journal of Engineering & Technology, vol. 38, no. 4, pp. 935-944, October 2019. p-ISSN: 0254-7821, e-ISSN: 2413-7219. DOI: 10.22581/muet1982.1904.07.

[4] I. Siddiqui, F. Rubab, H. Siddiqui and A. Samad, "Poet Attribution of Urdu Ghazals using Deep Learning," 2023 3rd International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2023, pp. 196-203, doi: 10.1109/ICAI58407.2023.10136675.

[5] B. Mehta and B. Rajyagor, "Gujarati poetry classification based on emotions using Deep Learning," International Journal of Engineering Applied Sciences and Technology, vol. 6, no. 1, May 2021.

[6] Ekin Ekinci, Hidayet Takcı, Sultan Alagöz 'Poet Classification Using ANN and DNN' , Elec Lett Sci Eng, vol. 18(1), (2022), 10-20

[7] Zhu, M.; Wang, G.; Li, C.; Wang, H.; Zhang, B. Artificial Intelligence Classification Model for Modern Chinese Poetry in Education. Sustainability 2023, 15, 5265. https://doi.org/10.3390/su15065265