Sindhi Poet Classification Using Deep Learning

Ronit Kumar Kataria
Computer Science
Habib University
Karachi, Pakistan

Faraz Ali
Computer Science
Habib University
Karachi, Pakistan

Muhammad
Computer Science
Habib University
Karachi, Pakistan

***Abstract*—Sindhi poetry, an integral part of the rich cultural heritage of Sindh, has a profound and diverse tradition that spans centuries. Sindhi poetry has many forms, one of which is ghazals. Sindhi poetry is primarily composed in the Sindhi language and is written in the Arabic script. In this paper, with the help of machine and deep learning models, we will explore the poetry classification on corpus of Sindhi Ghazals. Our primary objective is to establish a suitable and precise representation of ghazals that comprehensively captures the unique writing style of the poet. We have curated our dataset of over 3000 couplets from four notable poets. This paper not only helps identify poets but also highlights the lasting importance of Sindhi poetry, honoring the voices that have remained significant over time.**

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

I. Introduction

Poetry is a form of literary expression that uses language and its words to create a unique and artistic structure. It is placement of words in a unique and expressive style that grabs a reader attention, and share ideas, express emotions, and create imagery. The poetry is characterized by the use of the rhythmic and metaphorical language, employing various literary devices like rhyme, meter and symbolism. The words and the sentences in the poem are not be perceived literally, but there is a deeper meaning imposed within it. This is what attracts computational linguistics to poetry.

Sindhi heritage and culture hold a special place in the hearts of its people. They are the threads that connect generations, carrying forward stories, traditions, and values. When we delve into Sindhi poetry using deep learning, we're not just studying words; we're exploring the essence of a community's identity. Sindhi poetry is a vibrant and culturally significant form of literature that originates from the Sindh region, situated in present-day Pakistan, as well as parts of India. It has a rich history and continues to be a valuable part of the Sindhi cultural heritage. Shah Abdul Latif Bhitai's "Shah Jo Risalo" is a cornerstone of Sindhi poetry and literature. This epic work is celebrated for its lyrical beauty and profound spiritual themes.

Poet attribution is the task of determining the poet who wrote a given poem. It is a subfield of authorship attribution, which is the more general task of determining the author of a given text. This task is carried out by identifying the unique writing styles of different authors. One common approach is to use stylometric features. Stylometric features are statistical measures of a text's language use, such as the frequency of certain words or phrases, the use of certain grammatical structures, and the overall length of sentences. Poet attribution has a few distinguishing features. While authorship attribution models are trained on articles, the poet attribution models are trained and tested on couplets that are composed of two lines. In addition to stylometric features, poet attribution systems may also use data such as the poem's rhyme scheme, meter, and use of figurative language. This is because poetic language can be more nuanced and expressive than other types of language, and these additional features can help to better distinguish between different poets.

Our research on Sindhi poetry classification using deep 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. Research Question

The research question we are addressing are "Poet classification of Sindhi Ghazals". In this research we propose a deep learning model in which given a couplet from a ghazal, it will identify the poet. The problem statement can be formulated as
"Given a couplet from a Sindhi Poetry/Ghazal, identify who the poet is'

In this study, we will undertake an examination and implementation of various deep learning models. This will involve utilising pre-existing models as well as developing our own neural networks in order to effectively tackle the categorization challenge at hand. 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 process of selecting the poets is conducted from a compilation of 4 renowned Sindhi poets, as enumerated below.

1) Shah Abdul Latif Bhittai
2) Sheikh Ayaz
3) Rukhsana Preet
4) Ahmed khan madhosh

Our motivation for this research stems from a noticeable gap in preserving Sindhi literature. Many Sindhi Ghazals are still in hardcopy books, and many of the traditional work exists but their poet are not known. In today's era of social media many of couplets are shared under the name of Sindhi poets for likes and shares but there is no way to verify them. Apart from it very few work has been done in the area of Sindhi Language, hence this research will allow us to delve more deeper into Sindhi Literature. Using deep learning models devised in our research, it will allow us to to determine whether a couplet attributed to a poet, such as Ustad Bukhari, is actually one of Ustad Bukhari's. Classification problems in other languages make use of machine learning methods such as Support Vector Machine, K-Nearest Neighbor, Naive Bayes and Decision Tree Classifiers will allow us to train those models on Sindhi Ghazals dataset.

III. Literature Review

Poetry classification tasks has been done in different languages such as Urdu [3] , Persian [1], Malay [2], and other languages. Coming towards the novelity of Sindhi, there has not been single prior work done on Sindhi or specifically Sindhi Poetry. Different techniques have been used to extract the features from the poetry and classify them according to poets. For longer texts such as ghazals Support Vector Machine results in a better accuracy. [3].

In this study [3], they have trained several models which are SVM, Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbor (KNN). For feature selections, they have made use of chi square, and L1 based feature selection. They have used around 4000 ghazals to train the model. In the dataset, complete ghazals are used for a particular poet. 80 % data is used for training while 20 % data is for testing. Out of all classification models mentioned, SVM outperforms them all with and without feature selection. Overall, the SVM performed the best and achieved 72 % F1-measure score.

In addition, with reference to Poet Classification, we came across a study [6]of poet classification of Turkish poems using Artificial Neural Network (ANN) and Deep Neural Network (DNN) achitecture. The data is collected from 5 poets and their subsequent 314 Turkish poetries. The pre-processing starts by converting the text to lowercase, eliminating out-of-vocabulary (OOV) words, removing stop-words, tokenizing the poems, and applying stemming using the Snowball stemmer. They utilized two classification models: Multilayer Perceptron (MLP) for ANN and Convolutional Neural Network (CNN) for DNN. For MLP, they set parameters like regularization parameter alpha, maximum iteration count, and solver [6]. In CNN, they defined word embeddings dimension, neuron counts in dense layers, activation functions, optimization solver, and batch size. The authors evaluated the classification models based on precision, recall, F-score, and accuracy metrics. n. The results indicated that the Multilayer Perceptron (MLP) demonstrated higher accuracy compared to the Convolutional Neural Network (CNN). Specifically, MLP achieved an accuracy score of approximately 81%, signifying that it correctly classified poems with an accuracy of 81%. On the other hand, the CNN achieved a lower accuracy score of about 61%, indicating that it correctly classified poems with an accuracy of 61%. This outcome highlights the superior performance of the MLP model in accurately attributing poems to their respective poets based on writing style, underscoring its effectiveness in this specific authorship attribution task.

Additionally, we conducted a thorough examination of a scholarly article [4] that centred on the classification of poet attribution in the Urdu language. The dataset consisted of literary works authored by 15 highly esteemed and recognised poets, amounting to a total of 18,472 pairs of lines. This article critically assesses the utilisation of deep learning models for the purpose of attributing Urdu poetry. The four models, namely Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN), are subjected to a series of preprocessing processes. Transformer-based models, such as BERT and roBERTa, have also been investigated in the academic literature. The key parameters for each model are explicitly defined, so offering a full assessment of their efficacy in the analysis of Urdu literature. The findings demonstrate that the Support Vector Machine (SVM) model attained an accuracy rate of 64%, whilst the Random Forest model obtained a score of 25%. Long Short-Term Memory (LSTM) models shown superior performance compared to alternative deep learning models. Transformer-based models, particularly BERT, demonstrated exceptional performance, with an accuracy rate of 80%, so highlighting their supremacy. Transformers, specifically BERT, demonstrated precise outcomes for each class owing to their attention mechanism and encoder-based architecture.

Another study [1], we came across was the classification of Persian *Ghazals* using sequential learning. The era of poet was the key highlight of the paper where the developed model was able to classify the chronological order of each poem with respect to author's life span. The corpus consisted of multiple *Ghazals* of renowned poet Hafez. The data-set consisted of 496 samples, where each instance was presented in Persian, translated into English, Six labels of chronological labels, and Raad Label. The pre-processed data was sent for Word-Embedding process to vectorize the couplet and bringing in the form which can be sent to Classification models. The important features of data were extracted using Bag of Words(BoW), and Latent Dirchlet Allocation(LDA). The output of the model were class labels, Chronological and Raad which were applied on both Persian and English. For the model, both Machine Learning(ML) and Deep Learning (DL) models were used for Classification.

With context of Machine Learning, we came across study [7] that employs the XGBoost algorithm to create an automatic classification model (XGBoost-MCP) for modern Chinese poetry styles. The data collected was 836 samples which were preprocessed by the application of labelling, word segmentation using Jieba, and the removal of stopwords. The process of feature extraction involves the utilisation of the Doc2Vec method in order to decrease the dimensionality of the data. The study conducted a comparison of four classification methods, namely XGBoost-MCP, SVM, DNN, and DT. The XGBoost-MCP model demonstrated exceptional performance across many evaluation criteria, encompassing accuracy (93.62%), precision (94.09%), recall (93.69%), and F1-score (93.65%). The performance of the model surpassed that of Support Vector Machines (SVM), Deep Neural Networks (DNN), and Decision Trees (DT) by substantial margins. The findings of this study confirm that XGBoost remains the favoured classification model, consistent with prior research. Although DT had impressive performance, especially in specific datasets, XGBoost-MCP frequently emerged as the highest-performing model. The graphical representation in the study served to underscore the superior performance of XGBoost-MCP in comparison to the other models.

Similary, we reviewed another paper [5] which focused on Gujarati poery classification based on emotions. The paper used sentiment classification of the poetry and attributed them in different classes. The authors of this paper were able to collect the dataset of more than 300 poetries and divided them into 9 emotions based on **"Navarasa"** theory. The input of their model is the Gujrati poetry which is classified as one of the emotions from **"Navarasa"**. The methodology started by processing the **Kavan** dataset and making embeddings using NLP and vectorizing the important featured-tokens. The use of Zipf's law enables the determination of the likelihood associated with a specific rasa or emotion, as derived from the provided poetry. The presence of the extracted word is determined by referencing a pre-existing dataset called 'Rasa'. This dataset contains words that have been identified as capable of expressing emotions in poetry, in accordance with the Navrasa Concept. The employed model used for the classifications were Machine Learning (ML) which gave the accuracy of 87% in identigying the 9 emotions of **"Navarasa"**.

The literature review elucidates the predominance of numerous significant classification models including, Support Vector Machine (SVM), Multilayer Perceptron (MLP), Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbour (KNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Transformer-based models (such as BERT and roBERTa), and XGBoost. These models have been widely employed for the purpose of classifying poetry in several languages, demonstrating their adaptability and efficacy in the realm of literary analysis.

Moreover, we have conducted an exhaustive review of pertinent literature, exploring additional studies that delve into similar classification endeavors. The ensuing section encapsulates the comprehensive details of the dataset used and provides a succinct summary of the model accuracies, as depicted in the table I

TABLE I
Summary of Models, Dataset, and Accuracy

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

TABLE II
Corpus Breakdown

No	Poet Name	No of Couplets
1	Shah Abdul Latif Bhittai	1173
2	Sheikh Ayaz	232
3	Rukhsana Preet	117
4	Ahmed Khan Madhosh	72

IV. Dataset

In this section we have discussed about the information related to the samples used for poem classification for the model training. We have discussed of how we have obtained the samples, including the corpus size, and its acquisition and characterization.

We have collected corpus of around 1500 couplets for four Sindhi poets. 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 searched for different sites related to Sindhi literature and looked for Ghazals of known Sindhi poets. In the links mentioned above, we web scraped the websites using *beautiful soup* and *requests* library, while we manually copied the couplets from some. The websites have substantial data related to Sindhi Ghazals for multiple poets. They were suggested by Sindhi Sikhiya course professor at Habib University, Prof. S. Seelro.
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.

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