Pashto Poetry Attribution using Deep Learning Techniques

Dr. Abdul Samad
Associate Professor, CS
Habib University
Karachi, Pakistan
abdul.samad@sse.habib.edu.pk

Lyeba Abid
Computer Engineering
Habib University
Karachi, Pakistan
la07309@st.habib.edu.pk

Ali Muhammad Asad
Computer Science
Habib University
Karachi, Pakistan
aa07190@st.habib.edu.pk

Sadiqah Mushtaq
Computer Engineering
Habib University
Karachi, Pakistan
sm07152@st.habib.edu.pk

Shayan Shoaib Patel
Computer Science
Habib University
Karachi, Pakistan
sp07101@st.habib.edu.pk

I. Introduction

Poetry has been a cornerstone of cultural and literary heritage across the globe, as it is one of the most popular ways to express emotions, thoughts, and experiences. Pashto is one of the major languages spoken in Afghanistan and Pakistan, with 50% of the Afghan population speaking Pashto, and 15% of the Pakistani population speaking Pashto [1] [2]. As such, it also holds a significant place in the rich tapestry of South Asian literature with its first recorded poetic works believed to be dating back to the 8th century with the works of Amir Kror Suri (a warrior poet) [3]. There have also been other notable poets such as Khushal Khan Khattak, Rahman Baba, Ghani Khan, and Hamza Baba who have contributed to the Pashto literary tradition, and are revered for their works. However, despite its rich history and cultural significance, Pashto remains a low-resource language in the context of computational linguistics, with limited availability of annotated data and tools for natural language processing (NLP) apart from translated works. This gap poses challenges for preserving and promoting Pashto literature on a global scale. Thus, the attribution of poetic works to specific poets can sometimes be ambiguous, due to oral traditions, and lack of proper documentation.

In recent years, deep learning techniques have gained substantial traction in the field of language processing, text classification, and poet attribution. Poet attribution refers to the process of identifying the poet who authored a specific poem. This process is crucial for preserving the cultural heritage of a language, as well as for academic research and literary studies. While researchers have explored poetry attribution in various languages such as English, Hindi, Urdu, Arabic, Persian, and Gujrati, Pashto poetry has not been extensively studied in this context. This gap could also be attributed to the lack of resources, and documentation in Pashto. This paper aims to bridge this gap by introducing deep learning based approaches to classify Pashto poetry and attribute it to specific poets, which can also prevent misinformation and mis-attribution due to the lack of proper documentation available. This paper also aims to curate the first Pashto poetry dataset of 13 poets (old and relatively newer poets) with about 12000 couplets in total, which will be made publicly available to the research community.

II. Related Work

Poet attribution using deep learning and machine learning has been explored in various languages previously as well.

A study on poet attribution for Urdu Ghazals [4] employed machine learning, deep learning, and transformer-based techniques on a dataset of 18,609 couplets from 15 notable Urdu poets, covering both the 18th-19th and 19th-20th centuries. The distinct word usage and styles across these periods benefited the model in identifying patterns. The study tested four machine learning models — Logistic Regression, Random Forest, Naive Bayes, and SVM — using label encodings and TF-IDF for feature extraction. They also experimented with four deep learning models: Flatten, LSTM, GRU, and 1D-CNN, utilizing one-hot encoding and tokenization with a 300-dimensional embedding layer. The best performance came from transformer models: Bert and roBerta achieved 80.32% and 79.52% accuracy, respectively, outperforming deep learning methods. Among traditional machine learning models, SVM and Logistic Regression achieved 64% and 60% accuracy, surpassing deep learning approaches, where LSTM performed best at 59.96%. The study notes class imbalance as a challenge but emphasizes the superior performance of transformer models in poet attribution tasks.

Another study was conducted on the classification of Persian poetry based on the poet's era [5]. The authors used the Persian Hafez's Ghazal dataset which contained 496 Persian Ghazals. The authors categorized Persian Ghazals based on the poetic era in which they were written, with the classification focusing on distinguishiing Hafez's ghazals by the time period or era associated with the poetry using sequential learning architectures. Word embeddings were

used, an important technique in NLP, and deep learning classification models. In addition, they also used Bag Of Words (BOW) - a baseline statistical model for textual feature extraction, and Latent Dirichlet Allocation (LDA) for feature extraction. They also used Distributed Memory (DM) for vectorization, and concatenated Distributed BOW with DM. Then they trained Machine Learning (ML) and Deep Learning (DL) models including Random Forest, Logistic Regression, LSTMs, GRU, and Bi-LSTM. For evaluation, they used accuracy, F1 score, precision, and recall to evaluate the performance of their models. The ML models were compared to SVM, which performed better in terms of accuracy but not F1 score. In the DL models, LSTM gave highest accuracy of 77.8% and the highest F1-score of 76.6% on the Persian dataset.

We also came across a study that classified Gujrati Poetry based on emotions using deep learning techniques [6]. The author collected the first Gujrati poetry dataset of more than 300 different Gujrati poems, and called it the "Kavan" dataset that represented the different "Rasa's" emotions. They further used the NLTK library in Python for tokenization and labelling of the data. Poems were used as input one by one, with each word compared to a metadata based on the "Navarasa" concept, returning values between 0 and 8 for the emotions. They implemented a Deep Learning classifier model, and achieved an 87.62% accuracy.

In addition to classification and attribution, another study aimed to perform a comparative analysis on various machine learning models for Urdu poet attribution [7]. They collected a total of 1563 poems from 5 famous Urdu poets. They also tokenized the data using Python's NLTK library, however, they did not remove stop words as they believed that such functions are important in modeling the style of the author. They repeated various experiments on various models repeatedly with different parameters, to conclude that SVM performed the best over most configurations with the highest accuracy of 88.7% and an average accuracy of 81%, with Naive Bayes at a close second with the highest accuracy of 84% and an average accuracy of 77%. They also implemented LSTM models, however, the highest accuracy they received was 45.78%. They concluded that with over 100 samples per poet and more than 2 lines per sample, one could achieve good results in poet attribution.

Regarding Urdu language we found another research on the poet attribution on urdu authorship of poets [8]. They collected a total of 11406 couplets from nonsocial Urdu websites of mainly 3 different Urdu poets. They developed various models for classification including SVM, Multilayer Perceptron (MLP), Multinomial Naive Bayes, and Multinomial MLP Pre-Trained Word2Vec model. SVMs achieved the highest accuracy and F1-score of 82.85% and 82.67%.

Another similar work was done on identification of Urdu Ghazals using SVM [9]. The authors collected a total of 3967 couplets from 4 poets, and generated a total of one million tokens from the dataset with a vocabulary size of 6427. They also created a Term-Document Frequency (TDF) matrix with rows representing couplets and columns representing unique terms. For feature selection, they used Chi-Square and L1-based techniques to select the best features. The Chi-Square evaluates feature importance based on statistical independence, while L1-based selection focuses on non-zero regression coefficients. They selected a total of 5 models; Naive Bayes, Decision Trees, SVMs, KNNs, and Random Forest. Overall, SVMs performed the best out of all the models with an accuracy of 72%, closely followed by Naive Bayes with an accuracy of 70%.

We also found a study that used ML approaches for authorship attribution in Arabic poetry [10]. They established various characteristic of Arabic poetry such as *Meter (Wazn)* and *Rhyme (qafiya)*, and curated a corpus of Arabic poetry of 73 poets with 18646 Qasidah's for training that amount upto 1856436 words. They used SVMs and Naive Bayes for classification, using Chi-Square and Information Gain for feature selection. They came up with 6 features corresponding to character, word length, sentences length, first word length, meter, and rhyme. They achieved the highest accuracy of 98.86% on SVMs when all features were used, while on average SVM had an accuracy of 87.4%. They also implemented a Naive Bayes model which had the highest accuracy of 98.86% as well with only two features, and an average accuracy of 90.68%.

Apart from the aforementioned literature, various other works were also consulted that were relevant to our research. One study we came across was the first to work on Pashto text classification using language processing techniques for single and multi-label analysis [11]. They constructed a dataset of Pashto documents including Sports, History, Health, Scientific, Cultural, Economic, Political, and Technology based - although poetry wasn't included. They used DistilBERT-base-multilingual-cased model, Multilayer Perceptron, SVMs, KNNs, Decision Trees, Gaussian Naive Bayes, Multinomial Naive Bayes, Random Forest, and Logistic Regression models. They got a 94% accuracy using MLP classification and TFIDF feature extraction. DistilBERT - a multilingual model which was not pretrained on Pashto still was able to achieve 66.31% accuracy, thus, the authors conclude that this still gave promising results, and that the model could be further improved by developing a tokenizer specifically tailored for Pashto.

III. Dataset

Since no existing dataset for Pashto Poetry existed, we collected our dataset from scratch. A major challenge in the data collection aspect was the lack of digitalization and proper documentation on Pashto poets. While we were able to find some poetry in blogs and social sites, most of the poetry had been translated, pdf forms, or in apps for more famous poets such as Rahman Baba, Hamza Baba and Khushal Khan Khattak. In addition, poetry was also found as just a couplet or two in the form of images on social media. Thus, our data collection included searching for poetry online through blogs, articles, and sites including Hamari Web, Wordpress, and Rekhta. We also collected pdf

books of poets that we passed through an OCR and then cleaned out. Since there were also mobile apps available for certain Pashto poets, we extracted from mobile apps and converted it into textual form, or took screenshots of the poetry and passed it through the OCR to get the text. All the data was then cleaned and sorted to get just the Pashto poetry of each poet. All in all, we were able to collect 12,124 couplets from a total of 13 poets. The chosen poets also show a good distribution in the sense that some poets are of historic ages such as Rahman Baba, Khushal Khan Khattak, and Hamza Baba, while there are also recent poets such as Ghani Khan, and Khatir Afridi. Thus, the poetry would have varying styles, themes, word usage, dialects, and vocabulary which would help in training a robust model as the patterns would be more diverse.

The table below shows the number of couplets collected for each poet.

No.	Poet/Shayari	Couplets Count
1.	Hamza Baba	1196
2.	Rahman Baba	2276
3.	Ghani Khan	294
4.	Khushal Khan Khattak	2013
5.	Mumtaz Orakazi	1748
6.	Khatir Afridi	710
7.	Dr. Khaliq Ziar	42
8.	Rehmat Shah Sail	486
9.	Munir Jan Buneri	454
10.	Sahib Shah Sabir	661
11.	Aziz Mazerwal	31
12.	Abbasin Yousufzai	117
13.	Salim Riaz	248
Total		**12124**

Table I: Poets and their couplets count

IV. Future Work

We will continue to find more data on Pashto poetry to increase the size of the dataset to gain better results from our models. But mainly our future work will focus on starting with the devleopment of machine learning and deep learning models. If time and resources permit, we are also looking forward to training transformer based models as they've shown promising results as well in contexts of attribution and text classification [4], [11]. But before that, we will be focusing handling the class imbalance within the dataset so that each poet has roughly the same number of couplets to better train our model. We will also perform label encodings, and feature extractions for our machine learning models, and also label and one-hot encodings, tokenization, and embeddings for the deep learning model. Two of our teammates will focus on developing the models while two members will work on processing the data for the models. Once the models have been trained, we will use existing evaluation metrics to evaluate the performance of our models.

References

[1] C. W. Factbook, "Afghanistan." https://www.cia.gov/the-world-factbook/countries/afghanistan/, 2024.

[2] P. B. of Statistics, "Population by mother tongue." https://web.archive.org/web/20220409115251/https://www.pbs.gov.pk/sites/default/files//population_census/census_2017_tables/pakistan/Table11n.pdf, 2021.

[3] A. Aziz, "A brief history of pashto literature," *Journal of Emerging Technologies and Innovative Research*, vol. 6, no. 6, 2019.

[4] I. Siddiqui, F. Rubab, H. Siddiqui, and A. Samad, "Poet attribution of urdu ghazals using deep learning," in *2023 3rd International Conference on Artificial Intelligence (ICAI)*, pp. 196–203, 2023.

[5] 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, 2022.

[6] B. Mehta and B. Rajyagor, "Gujarati poetry classification based on emotions using deep learning," *International Journal of Engineering Applied Sciences and Technology*, vol. 6, 05 2021.

[7] T. Ahmed and A. Rao, "Poet attribution for urdu: Finding optimal configuration for short text," *Poet attribution for urdu: Finding optimal configuration for short text*, vol. 4, 2021.

[8] M. Dar, "Authorship attribution in urdu poetry," 06 2020.

[9] N. Tariq, I. Ejaz, M. K. Malik, Z. Nawaz, and F. Bukhari, "Idenitification of urdu ghazal poets using svm," *Mehran University Research Journal of Engineering and Technology*, vol. 38, no. 4, 2019.

[10] A. Al-falahi, M. Ramdani, and M. Bellafkih, "Machine learning for authorship attribution in arabic poetry," *International Journal of Future Computer and Communication*, vol. 6, pp. 42–46, 01 2017.

[11] J. Baktash and M. Dawodi, "Enhancing pashto text classification using language processing techniques for single and multi-label analysis," 05 2023.