Punjabi Poet Attribution

1st Fatima Tariq
Computer Science
Habib University
Karachi, Pakistan

2nd Ragini Gopchandani
Computer Science
Habib University
Karachi, Pakistan

3rd Raza Hashim
Computer Science
Habib University
Karachi, Pakistan

***Abstract*—Poet attribution refers to identifying the authorship of any given piece of poetry based on the key differentiating factors picked up by extensive analysis of works by a given set of poets. The models in question will be given a poem in Romanized script and it will predict the author of that particular poem from the set of given poets. We decided to pursue this cause due to how often we encounter wrongful attribution on social media. Wrongful attribution includes misquotations, claiming a famous poet's words as one's own, or committing plagiarism by presenting a poet's work as their own. We aim to define a mechanism for accurately attributing the given work to its respective poet. The language we have chosen to work on is Punjabi. The reason is that Punjabi is perhaps the language with the richest poetic history in the subcontinent yet at the same time there has not been a lot of work done in it with regard to poet identification. Besides working in Roman script we also conducted experiments with our models using Gurmukhi and Shahmukhi script to see what sort of a difference would that lead to. Overall we ended up having 6 different datasets (3 scripts and Main/Extended divisions) with the largest one containing up to 1125 poems by 14 different poets. We experimented with three distinct deep-learning models to determine the most effective approach for identifying authorship. We found that DistilBERT, a transformer-based model, performed the best among the three models on our primary Roman script, achieving an accuracy of 83.87%. Additionally, it achieved accuracies of 87.50% and 86.67% on the Shahmukhi and Gurmukhi scripts, respectively.**

***Index Terms*—poet attribution, poet identification, Punjabi poetry, Punjabi poet identification, recurrent neural network, deep learning, transformers.**

I. Introduction

Poetry is a staple in cultures of this part of the world, being one the most popular ways of expressing emotions and/or beliefs. Languages like Urdu, Hindi, Farsi, Sindhi, etc have rich traditions when it comes to poetic works. The poets of these languages, such as Maulana Rumi, Allama Iqbal, Shah Abdul Latif Bhittai, etc are globally renowned for their remarkable contributions as great artists and thinkers. Punjabi is no different from these languages in the sense that it has a long and rich poetic tradition that could rival any other. Thinkers like Baba Bulleh Shah have left behind treasures of knowledge and philosophical thoughts in their poetic works which are as relevant today as they were when these men still roamed the world. We believe that these huge treasure troves should be accessible to the common man instead of being restricted to the pages of dusty old books.

Authorship recognition has seen a lot of work but little to none in Punjabi. This is despite the fact that even the most conservative estimates place it as a thousand-year-old language spanning across different cultural eras, leaving it with a diverse abundance of tradition. It has around 113 million speakers worldwide, with around 38% of Pakistan's entire population being native Punjabi speakers. We believe that such a rich language deserves its artworks to be recognized and hence our project came about.

With the advent and rise of social media platforms, these works of art have seen an emergence in the public sphere with works being shared in great numbers on social media platforms. However, that has seen a group of new problems arise, namely plagiarism (individuals passing off another artist's work as their own) and misattribution to an artist, either by quoting one artist's work as another's or by attributing a false work to them. We aim to come up with a model that when given a poem can correctly identify which poet wrote it based on its linguistic style and characteristics. We believe that this can help counter the problems we have mentioned. The fact that we were spurred on by social media is also the reason why our initial and primary focus was on Roman Script over Shahmukhi or Gurmukhi script as many users are far more comfortable with Roman Script over the other two. This also allows us to expand our usefulness to non-native speakers of Punjabi. That being said we did also create and train and test datasets in the other two scripts to expand our scope.

II. Research Question

Our primary research objective is to apply deep learning methods to the field of "Poet Attribution for Punjabi Poetry." The central task of our project is to create a model that can take a Punjabi poem as input and, using deep learning models, produce a class label that identifies the poet of the said poem. Along with that we also wanted to try all three scripts on our chosen model to test how different the results achieved would be, to determine what was the best script for our chosen models.

We are primarily focusing our attention on the literary works of eight different and renowned poets, each of whom has made a significant contribution to the field of Punjabi literature. These poets consist of:

1) Baba Farid
2) Bulleh Shah
3) Fazal Shah Sayyad
4) Shah Muhammad
5) Sultan Bahu
6) Ustad Daman

7) Vir Singh
8) Waris Shah

Besides these main eight we also added a further six to create an additional extended dataset of fourteen total poets. The six additions were:

1) Ali Haider
2) Guru Nanak
3) Kareem Bakhsh
4) Khawaja Ghulam Farid
5) Khush Taba
6) Shiv Kumar Batalvi

The primary use of the Roman script, which hasn't been thoroughly investigated in this context, is one of the distinctive features that makes our research stand out. Despite some scholarly interest in the subject, the work that has been done so far has concentrated on Punjabi poetry written in the Gurmukhi alphabet, which is the language's original script. But even in that regard, our project had a much larger dataset in Gurmukhi while also adding Shahmukhi, meaning we covered the three major scripts for the Punjabi language while also covering many more poets in each script.

It is crucial to realize the profound cultural significance of poet identification in Punjabi poetry. We hope to further contribute to the preservation and dissemination of Punjabi poetic heritage by expanding this analysis to the Roman script and making valuable literary works from Punjabi more available to readers around the world. While also covering the two historical scripts to make it available to the native speakers as well. In the unique context of poet identification, this project emphasises the significance of Romanized Punjabi poetry as a conduit for scholarly investigation and a broader appreciation of Punjabi literary artistry as that has the most presence online and might be the most effective way to preserve the heritage.

III. Literature Review

Literary analysis has been a pivotal area of exploration, particularly in the classification and attribution of poetic works across various languages. In recent studies, classification tasks have been conducted on poems written in languages such as Punjabi, Tamil, Hindi, Arabic, and Urdu. These studies employ diverse methodologies to analyze poetic content and attribute authorship accurately. In this review, we provide a chronological overview of recent studies in this domain, highlighting the methodologies employed and drawing connections with our work.

In [1], Andres Lou et al. conducted a study on multi-label classification of poems based on subjects using a dataset from the Poetry Foundation's archive. The input comprised the text of over 7,000 poems, with the output being the author's name. It was a multi-label classification problem, as each poem could belong to multiple categories and subcategories. The dataset contained over 7,000 samples and was categorized into nine main categories and various subcategories. Feature extraction utilized Term Frequency-Inverse Document Frequency (tf-idf) applied to a Bag-of-Words model and Latent Dirichlet Allocation (LDA), with features filtered using Pearson's Chi-Square test. For classification, a Support Vector Machine (SVM) model was employed. Evaluation criteria included precision, recall, and F-score, resulting in an average accuracy of 84.8% in categorizing poems into main categories and subcategories based on their subjects.

In [2], A. Pandian et al. delved into author classification for unidentified Tamil poems, using features extracted from Mukkoodar Pallu's works. The dataset encompassed around 800 anonymous poem instances, focusing on classification. Feature extraction involved extracting lexical, syntactic, and semantic aspects. The C4.5 algorithm served as the model for classification, achieving an accuracy of 88.23% after parameter adjustments. Evaluation criteria included the confusion matrix. The study varied parameters such as the confidence factor and minimum number of objects to improve classifier accuracy, achieving a maximum accuracy of 88.23% when the confidence factor was set to 0.2 and the minimum number of objects was 4.

A similar study by A. Pandian et al. [3] focused on classifying Punjabi poetry to identify authors. The dataset comprised 400 poems by five poets, treated as a single-label classification problem. Features encompassing lexical, syntactic, and statistical aspects were extracted, enhancing classifier accuracy to 90%. The J48 Decision Tree Algorithm facilitated feature selection. Evaluation included Precision, Recall, F-score, and Accuracy metrics, with J48 achieving the highest accuracy of 86.66%. Other algorithms yielded accuracies ranging from 63.33% to 83.33%.

In [4], Al-Falahi Ahmed et al. focused on authorship identification within Arabic poetry. The dataset comprised 21,929 poems from 114 diverse Arabic poets, divided into a training set of 1,673,465 words from known poets and a test set of 89,456 words from anonymous poets. Features were categorized into seven groups, including Character, Lexical, Structural, Syntactic, Semantic, Poetry, and Specific Words Features. Three classification techniques—Naive Bayes, Support Vector Machine, and Linear Discriminant Analysis (LDA)—were applied. The LDA model notably achieved an impressive accuracy rate of 99.12%, particularly when utilizing specific word attributes. While accuracy rates varied across different feature sets and models (ranging from 71.93% to 98.25%), the research consistently underlines the superiority of LDA, offering valuable insights into authorship attribution in Arabic poetry.

In [5], A. Pandian et al. focused on authorship identification within Hindi literature, analyzing a dataset comprising 100 Hindi poems from three distinguished authors: Ramadhari Singh Dinkar, Maithlisharan Gupt, and Jaisankar Prasad. Extensive feature extraction was conducted to distinguish authors based on their unique writing styles, categorizing features into Lexical, Statistical, and Syntactic groups. The study explored seventeen machine learning algorithms, with the Logitboost algorithm demonstrating an impressive accuracy rate of 75.67%. Additionally, algorithms like J48, Classification via Regression, and Iterative Classifier Optimizer performed well, achieving accuracy scores ranging between 70% and 75%.

Year	Language	Dataset	Problem Type	Models Used	Evaluation Criteria	Results
2015	English	7,000 poems	Multi-label	SVM	Accuracy, Precision, Recall, AUC and F-score	84.8% avg. accuracy
2016	Tamil	800 poems	Classification	C4.5	Accuracy rates	88.23% accuracy
2018	Punjabi	400 poems	Classification	J48, Random Forest, Bayes Net, etc.	Precision, Recall, F-score, Accuracy	86.66% peak accuracy (J48)
2019	Arabic	21,929 poems	Classification	Naive Bayes, SVM, LDA	Accuracy rates, Performance averages	99.12% peak accuracy (LDA)
2020	Hindi	100 poems	Classification	J48, Bayes Net, Naive Bayes, etc.	Accuracy rates	75.67% peak accuracy (Logitboost)
2020	Urdu	11,406 couplets	Classification	SVM (RBF kernel), Naive Bayes, MLP	Precision, Recall, F-score, Accuracy	82.85% accuracy (SVM)

TABLE I: Summary of Literature Review

In [6], Momna Dar focuses on authorship attribution within Urdu literature, specifically analyzing the works of Faiz Ahmad Faiz, Muhammad Allama Iqbal, and Mirza Ghalib. The dataset comprises 11,406 unique couplets. The study utilized various machine-learning libraries for data cleaning and explored multiple classification methods, including Multinomial Naive Bayes, Multi-Layer Perceptron (MLP), and Word2Vec sentence embedding with a pre-trained model. Among these, the Support Vector Machine (SVM) with an RBF kernel emerged as the most effective, achieving an accuracy rate of 82.85%.

We have extensively reviewed various research papers on poet attribution, presenting a broad understanding of the methodologies applied in this domain. Significantly, the aforementioned study [3] focused on Punjabi poet identification, similar to our work. However, it was limited to a set of five poets and relatively smaller datasets. In contrast, our approach stands out for utilizing a more comprehensive dataset encompassing the works of eight prominent poets. Furthermore, while the prior study did not include translations from the Gurmukhi script to the Romanized script, we have undertaken this step to enhance accessibility, intending to reach a broader audience. These adaptations and dataset expansion highlight the uniqueness of our project, detailed further in the subsequent section.

To consolidate the findings from the literature review and provide a comprehensive overview of the studies discussed, Table I summarizes key details of each research paper, including the year of publication, language, dataset details, problem type, models used, evaluation criteria, and results.

IV. Materials and Methodology

In this section, we will go over the dataset(s) that we used for our experimentation and the models that we experimented with.

A. Dataset

This section introduces the critical foundation of our research. Our dataset comprises poems by fourteen different Punjabi poets of various eras and styles in three different scripts; Roman, Shahmukhi, and Gurmukhi. We had initially aimed to use individual couplets rather than whole poems as inputs for our models, but upon further research and consultation with linguistic experts, we decided against that. The reason for that being that the data we were collecting comprised mostly of Nazms as opposed to Ghazals, meaning that they were atomic in nature. So couplets had no standalone meaning hence the decision to work with whole poems instead.

1) Acquisition: The data has primarily been acquired through 2 sources: GitHub and Folk Punjab.

The poetic works of four poets were sourced from a GitHub repository, but they were initially displayed in the Gurmukhi script, which, due to its regional context, posed a challenge in terms of universal comprehension. To bridge this linguistic gap and make these poetic creations accessible to a broader audience, we converted the collection into the Roman script, a more widely understood writing system. To ensure the accuracy, integrity, and fidelity of the Romanized content to the poets' original works, a meticulous verification phase was carried out with the help of an individual proficient in the Gurmukhi script.

We curated the rest of the dataset (eleven poets) with the help of the website Folk Punjab. The poems here were already available in all three scripts. We extracted the available poems of the selected poets and stored them in poet-wise folders. Our datasets were as follows;

Poet	Roman	Gurmukhi	Shahmukhi
Baba Farid	131	131	131
Bulleh Shah	100	100	-
Fazal Shah Sayyad	121	121	121
Shah Muhammad	105	105	105
Sultan Bahu	186	186	186
Ustad Daman	100	100	-
Vir Singh	95	95	-
Waris Shah	90	90	90

TABLE II: Main Poets

Poet	Roman	Gurmukhi	Shahmukhi
Ali Haider	29	29	29
Baba Farid	131	131	131
Bulleh Shah	100	100	-
Fazal Shah Sayyad	121	121	121
Guru Nanak	52	52	52
Kareem Bakhsh	25	25	25
Khawaja Ghulam Farid	20	20	20
Khush Taba	30	30	30
Shah Muhammad	105	105	105
Shiv Kumar Batalvi	41	41	41
Sultan Bahu	186	186	186
Ustad Daman	100	100	-
Vir Singh	95	95	-
Waris Shah	90	90	90

TABLE III: Total Poets

Our dataset is then further divided into three categories: training data (80%), testing data (10%), and validation data (10%) using sklearn's train_test_split library.

2) Reasoning: These poets have been carefully selected for a number of reasons:

- **Availability:** The work of these poets was available online. 4 of the 14 were part of a pre-existing Gurumukhi dataset that we took as is and for Roman conducted transliteration that we verified too. While the rest were also relatively easy to source from a website, Folk Punjab, dedicated specifically to Punjabi poetry.
- **Historical Significance:** Besides availability, another big reason is the fact that these have been some of the most important poets in the Punjabi language. With the likes of Bulleh Shah's work being adapted to this day, nearly 3 centuries after his death. While Heer-Ranjha by Waris Shah being one of 4 most popular Punjabi tragic tales. All in all, these we believe are the most culturally relevant and influential poets in the Punjabi language.
- **Diversity:** These poets also lived and wrote during different times in history. Spanning all the way from Baba Farid in the 12^{th} and 13^{th} centuries to Ustad Daman till the $1980s$. We believe this wide coverage of eras will allow us to capture very different themes, styles and trends in the language.

The reason these eight were chosen as the main dataset is that they had a considerably larger number of available poems and, overall, had a much greater cultural impact. It's important to note that the other six were included much later in our project, while these eight had been selected from the very beginning.

B. Deep Learning Models

Since most of the work from the literature review had been done using traditional machine learning models, we wanted to explore the realm of deep learning models for poet attribution and see how well they would perform. In this section, we will introduce three deep learning models we trained for the poet attribution of Punjabi poems.

1) distilbert-base-multilingual-cased: This model is based on DistilBERT, a smaller version of BERT (Bidirectional Encoder Representations from Transformers). It is pretrained on a large multilingual corpus, covering 104 languages, and has been fine-tuned for various natural language processing tasks. We utilized the DistilBERT tokenizer to convert each poem into tokens, and then used the DistilBERT model to generate embeddings with a maximum length of 50 for these tokens. The maximum length was chosen according to the distribution of all the poem lengths in the dataset as shown in figure 1. To save computational time and increase efficiency, the embeddings were subsequently saved for each set (training, validation, and test), and then loaded when needed. The batch size was kept to 32 and we used a learning rate of 0.001 for the best results (discussed in next section). The model was trained for 100 epochs with the use of early stopping to avoid training overfitting. This comprehensive approach allowed us to capture the intricate linguistic nuances present in Punjabi poetry.

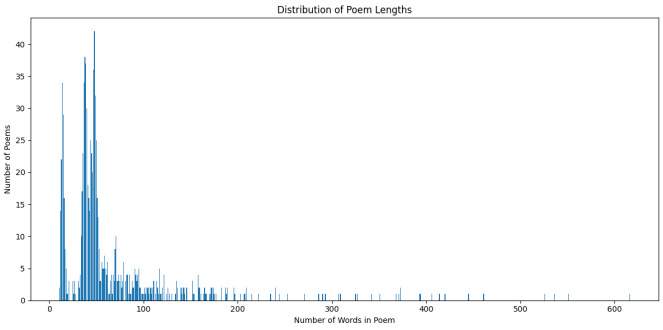

Fig. 1: Distribution of Poem Lengths in the Main Dataset

2) LSTM: We explored the utilization of Long Short-Term Memory (LSTM) networks in our project. LSTMs are a type of recurrent neural network (RNN) designed to capture and learn long-term dependencies in sequential data, making them well-suited for tasks involving sequential information, such as natural language processing. Prior to input, labels are encoded using scikit-learn's LabelEncoder, ensuring effective classification. The poetry texts undergo tokenization and are converted into index sequences and then padded (a dictionary mapping each word in the vocabulary to a specific index was generated). An LSTM-based classifier, named PoetLSTMClassifier, was configured with an embedding layer of dimension 50, a bidirectional LSTM layer with a hidden dimension of 30 to enhance context understanding, and a fully connected layer for classification. During the training process, we employed a learning rate of 0.01, a batch size of 32, and utilized the CrossEntropyLoss criterion along with the Adam optimizer. Early stopping mechanisms were implemented to prevent overfitting and ensure generalization to unseen data.

3) GRU: Next, we employed the Gated Recurrent Unit (GRU) to train and test our data. GRU, a variant of traditional

RNNs, is designed to capture sequential dependencies efficiently while mitigating some of the challenges associated with vanishing gradients. The encoding, tokenizing, and embedding process was carried out similarly to the LSTM model. The model's key hyperparameters, including hidden and embedding dimensions (Hidden Dimension: 30, Embedding Dimension: 50), learning rate (lr=0.01), and batch size (32), were thoughtfully selected to balance complexity. For training, the model underwent 25 epochs, and early stopping with a patience of 10 epochs was employed to prevent overfitting. The best model was saved during training based on validation loss.

V. Results

In this section, we delve into the outcomes of our exploration into deep learning models for the intricate task of poet attribution in Punjabi poetry. The investigation encompasses three distinct models: DistilBERT, LSTM, and GRU. The models were trained and tested separately on each dataset.

A. Main Dataset

Our main dataset consists of 8 poets for Roman and Gurmukhi, and 5 poets for the Shahmukhi script.

1) distilbert-base-multilingual-cased: On our main Roman dataset, distilBERT performed the best with an accuracy of 83.87%. Throughout the training process, the validation loss consistently decreased, culminating in a final value of 0.59. Overall, the time taken for training was much less for distilBERT as compared to the other models. The results are shown below.

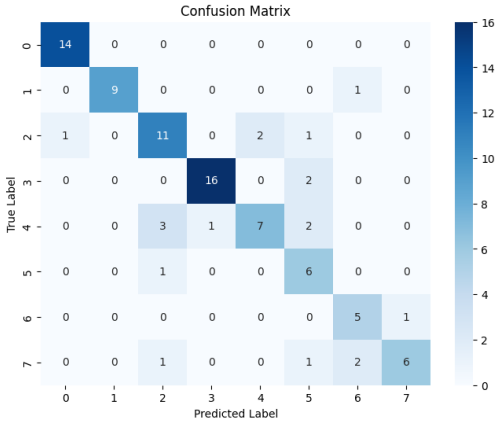

Fig. 2: Confusion Matrix for Roman

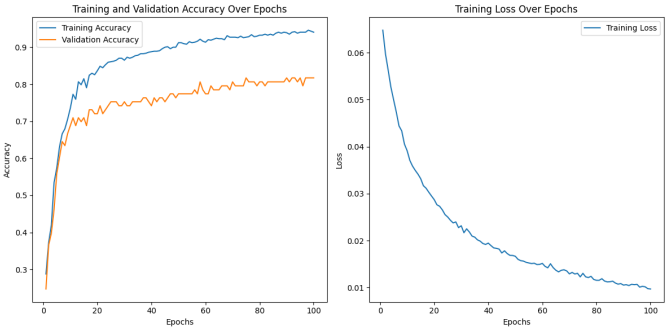

Fig. 3: Accuracy and Training Loss Curve

On the Shahmukhi and Gurmukhi datasets, distilbert gave an accuracy of 87.50% and 86.67& respectively.

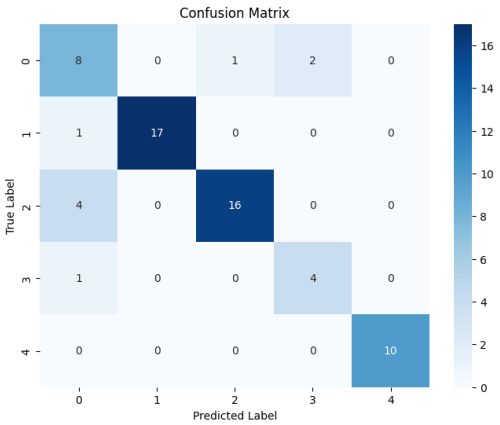

Fig. 4: Confusion Matrix for Shahmukhi

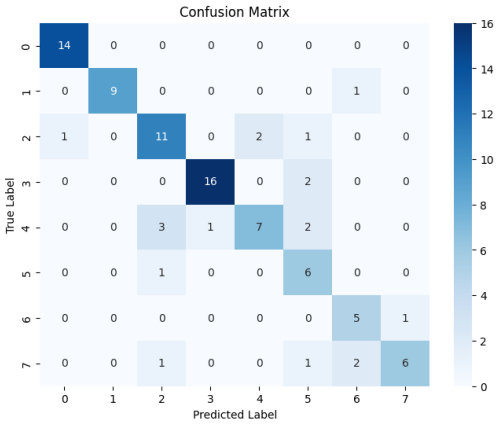

Fig. 5: Confusion Matrix for Gurmukhi

2) LSTM: LSTM gave an accuracy of 73.12% on our main Roman dataset. Despite being outperformed by distilBERT, the LSTM model demonstrated strong classification capabilities. The best validation loss for LSTM was recorded at 1.96.

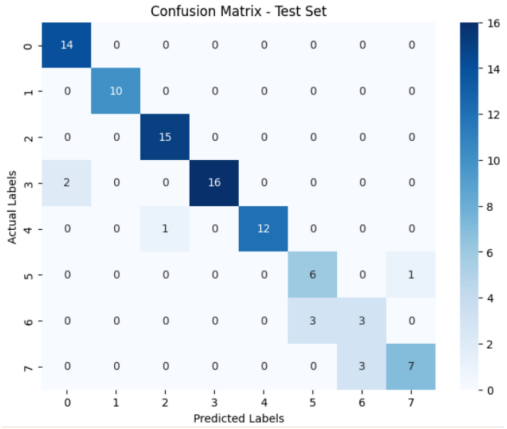

Fig. 6: Confusion Matrix for Roman

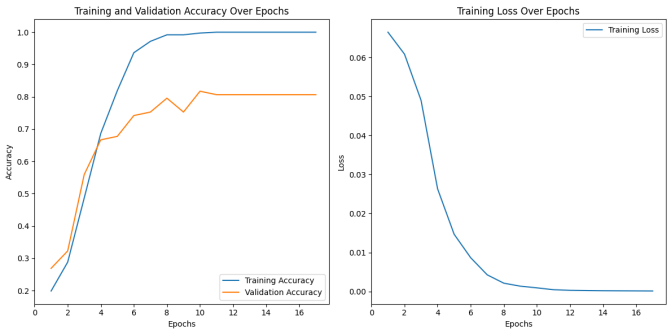

Fig. 7: Accuracy and Training Loss Curve

On the Shahmukhi dataset, LSTM surpassed distilbert, achieving an accuracy of 92.50%. On the Gurmukhi dataset, LSTM demonstrated strong performance with an accuracy of 82.80%.

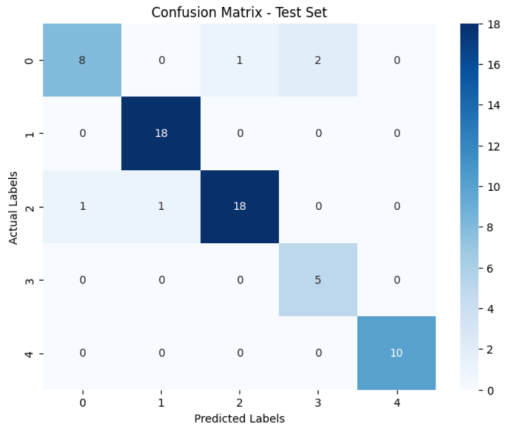

Fig. 8: Confusion Matrix for Shahmukhi

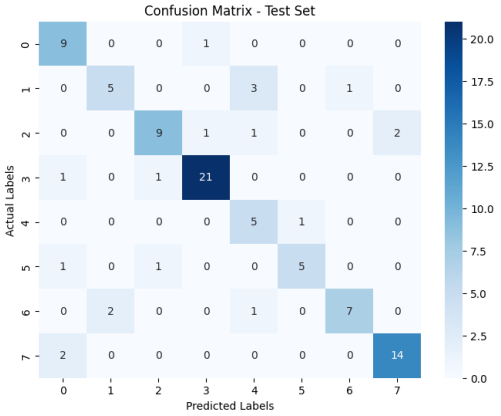

Fig. 9: Confusion Matrix for Gurmukhi

3) GRU: Our GRU model outperformed LSTM on the main Roman dataset with an accuracy of 80.20% and the best validation loss reaching 1.31.

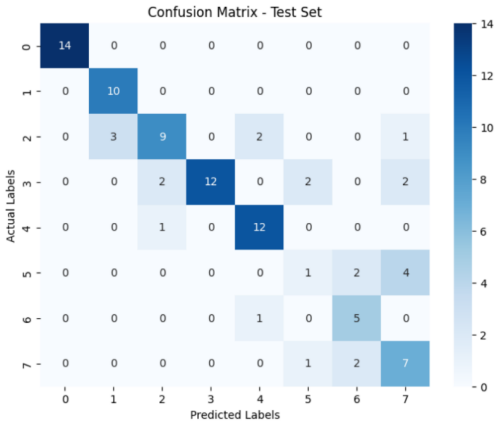

Fig. 10: Confusion Matrix for Roman

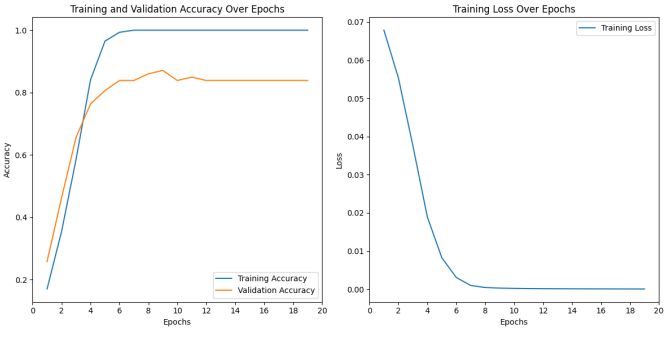

Fig. 11: Accuracy and Training Loss Curve

On the Shahmukhi and Gurmukhi datasets, GRU performed well with an accuracy of 97.50% and 82.56%, respectively.

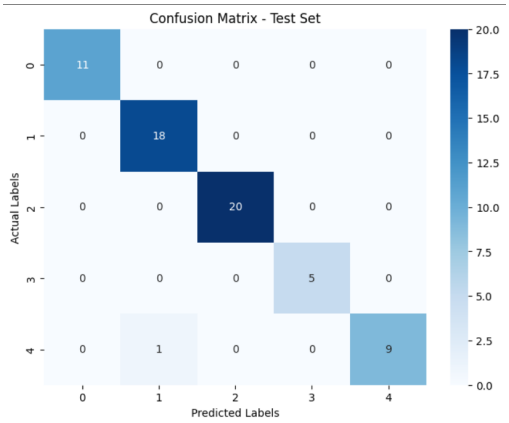

Fig. 12: Confusion Matrix for Shahmukhi

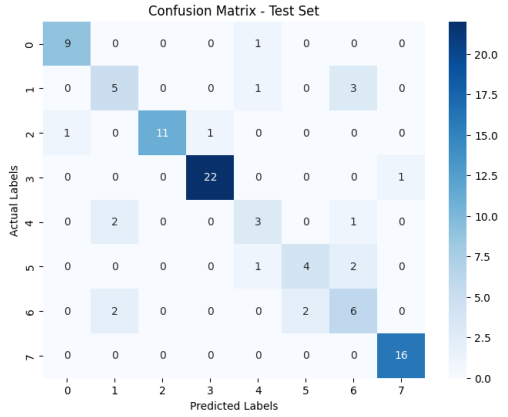

Fig. 13: Confusion Matrix for Gurmukhi

B. Summary of Results for the Main Dataset

	Roman (Main)	Shahmukhi (Main)	Gurmukhi (Main)
DistilBERT	83.87%	87.50%	86.67%
LSTM	73.12%	92.50%	82.80%
GRU	80.20%	97.50%	82.56%

TABLE IV: Accuracy Percentages for the Main Dataset

C. Extended Dataset

In this section, we present the outcomes achieved by our models on the extended dataset which consists of 14 poets for the Roman and Gurmukhi script, and 11 poets for Shahmukhi. Notably, while the distilBERT model completed training within a similar time frame as the main dataset, the LSTM and GRU models required considerably more time on the extended dataset. To streamline the presentation and avoid redundancy, we showcase only the confusion matrices for distilBERT. A comprehensive overview of accuracies for all three models across each script is consolidated in Table VI.

distilbert-base-multilingual-cased:

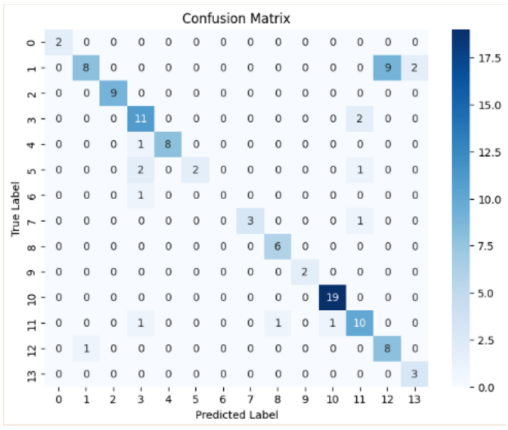

Fig. 14: Confusion Matrix for Roman

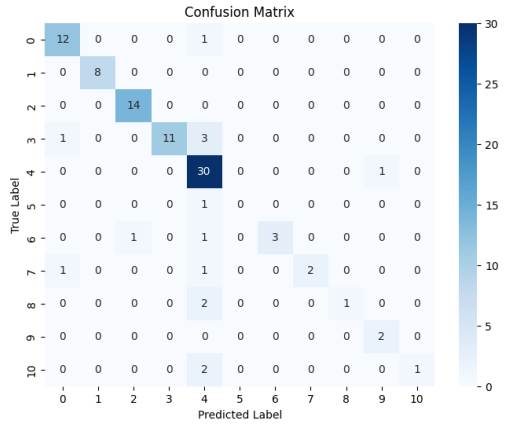

Fig. 15: Confusion Matrix for Shahmukhi

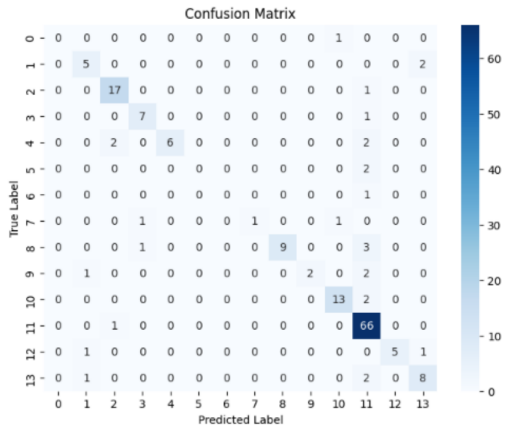

Fig. 16: Confusion Matrix for Gurmukhi

D. Summary of Results for the Extended Dataset

	Roman	Shahmukhi	Gurmukhi
DistilBERT	81.82%	84.85%	78.26%
LSTM	72.81%	83.84%	76.89%
GRU	78.95%	88.89%	81.71%

TABLE V: Accuracy Percentages for the Extended Dataset

VI. Discussion

In examining the performance of DistilBERT, LSTM, and GRU models across the main dataset's different scripts—Roman, Shahmukhi, and Gurmukhi—it's evident that each model exhibited distinct accuracies.

DistilBERT emerged as the top performer across Roman and Gurmukhi scripts due to its multilingual pre-training. Its capability to comprehend diverse linguistic nuances and handle multilingual texts contributed to high accuracies of 83.87% in Roman and 86.67% in Gurmukhi. Notably, Shahmukhi accuracies across all models were higher—DistilBERT reached 87.50% accuracy. This improvement in Shahmukhi accuracies may result due to fewer poets (classes) in this script, potentially enabling better model learning.

On the other hand, GRU demonstrated exceptional accuracy of 97.50% in Shahmukhi, surpassing both DistilBERT and LSTM. Its simpler architecture might have contributed to its effectiveness in discerning patterns in Shahmukhi poetry, especially given the fact that there were comparatively fewer classes than the other two scripts. Conversely, while LSTM demonstrated competitive performance in Roman (73.12%) and Gurmukhi (82.80%), it displayed variability across scripts, possibly influenced by dataset sizes and script-specific intricacies.

Transitioning to the extended dataset encompassing a larger pool of poets in Roman, Gurmukhi, and Shahmukhi scripts, it's evident that the accuracies are lower compared to the main dataset. This decrease might be attributed to the introduction of extra classes not present in the main dataset, with fewer poems available for these additional classes, impacting the models' learning and generalization.

For Shahmukhi, each model outperformed the other scripts, further emphasizing the impact of the dataset size. The lower number of classes in Shahmukhi might have allowed the models to better generalize and attribute poems accurately, compared to the increased complexity and number of classes in Roman and Gurmukhi.

DistilBERT sustained a relatively high accuracy of 81.82% in Roman, maintaining consistency despite the expanded dataset. However, it experienced a slight accuracy decrease compared to its performance on the main dataset (83.87%). GRU, while showcasing resilience with 88.89% in Shahmukhi, exhibited marginally decreased accuracies in Roman (78.95%) and Gurmukhi (81.71%). This could be down to the extreme class imbalance in the extended dataset and might imply challenges in adapting to increased dataset complexities. LSTM, despite its competitive edge, experienced slightly reduced accuracies in Roman (72.81%) and Gurmukhi (76.89%) across the extended dataset. This points toward the model's sensitivity to dataset expansions and nuances present in different scripts, affecting its attribution capabilities.

Moreover, our study exhibits performances akin to a previous paper [1] reviewed in the literature. This prior study specifically focused on poet identification in the Gurmukhi Script using machine learning models and achieved an accuracy of 86.66% using J48, with a dataset comprising around 400 poems from five poets. In contrast, despite handling a more extensive dataset with additional classes and employing deep learning models across multiple scripts, our accuracies are comparable to those achieved by this previous study. This demonstrates the effectiveness and advancement of our approach in poet attribution tasks even with increased complexity and larger datasets.

The variations in accuracies across models within the extended dataset highlight the impact of dataset sizes, script-specific complexities, and model architectures on their attribution capabilities. This underscores the robustness of deep learning models in handling challenges posed by an expanded dataset with diverse scripts and increased classes.

VII. Future work

While our dataset was much larger than any previous work in Punjabi it was still quite limited. That is something that could be improved upon in the future. Some poets had very limited representation (as low as 20) and considering that most of these poets have written a lot more that can be better represented, as that would hopefully result in better training too.

Considering that some of the poets had very few poems, we did have quite the class imbalance in our expanded dataset, ranging from as high as 186 to as low as 20. For any future work that is most certainly something we would like to address.

The dataset can be more diversified too, as it only contained 14 poets and lacked some very big names such as Faiz Ahmed Faiz and Iqbal. More diversity would make it appealing to more people. Not only that but it would also allow us to cover more unique styles and subjects.

Another thing that could most certainly be explored for this purpose is the usage of machine learning models, support vector machines (SVMs) in particular to see how much the results would differ.

VIII. Conclusion

The attribution of Punjabi poetry to respective poets is a complex yet intriguing task that this study aimed to address. By employing DistilBERT, LSTM, and GRU models on a diverse collection of poems in Roman, Gurmukhi, and Shahmukhi scripts, this research explored the intricate nuances in the attribution process. DistilBERT showcased impressive accuracy rates of 83.87% in Roman and 86.67% in Gurmukhi scripts during the evaluation of the primary dataset. Conversely, the GRU model exhibited exceptional performance with an accuracy of 97.50% specifically in the Shahmukhi script. The extended dataset revealed decreased accuracies, highlighting the impact of added classes. Shahmukhi script consistently performed better, likely due to having fewer classes than the rest. These insights emphasize the critical role of dataset composition and script-specific nuances in attributing Punjabi poetry, offering valuable directions for future research in language attribution within poetry analysis.

REFERENCES

[1] A. Lou, D. Inkpen, and C. Tănăsescu, "Multilabel Subject-based Classification of Poetry." Accessed: Mar. 26, 2024. [Online]. Available: https://www.site.uottawa.ca/~diana/publications/flairs_2015_paper.pdf Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) using C4.5 Algorithm

[2] A. Pandian, V. Ramalingam, and R. P. Vishnu Preet, "Authorship Identification for Tamil Classical Poem (Mukkoodar Pallu) using C4.5 Algorithm," Indian Journal of Science and Technology, vol. 9, no. 1, pp. 1–5, Jan. 2016. doi: https://doi.org/10.17485/ijst/2016/v9i47/107944.

[3] A. Pandian, S. Wahid, Y. Tokas, and V. V. Ramalingam, "Authorship Identification of Punjabi Poetry," International Journal of Engineering & Technology, vol. 7, no. 4.19, pp. 13–16, Nov. 2018. Available online: https://www.sciencepubco.com/index.php/ijet/article/view/21987/10663.

[4] A.-F. Ahmed, R. Mohamed, and B. Mostafa, "Arabic Poetry Authorship Attribution using Machine Learning Techniques," Journal of Computer Science, vol. 15, no. 7, pp. 1012–1021, Jul. 2019. Available online: https://thescipub.com/abstract/jcssp.2019.1012.1021.

[5] P. Pandian, N. Maurya, and Jaiswal, "Author Identification of Hindi Poetry." [Online]. Available: https://www.semanticscholar.org/paper/AUTHOR-IDENTIFICATION-OF-HINDI-POETRY-Pandian-Maurya/7d14ac5b51edb43577e03e8cbd173f91db8d93ef.

[6] M. Dar, "Authorship attribution in Urdu poetry," ResearchGate, Jun. 2020, [Online]. Available: https://www.researchgate.net/publication/344561377_Authorship_Attribution_in_Urdu_Poetry