Deep Learning based Poet Attribution model for Punjabi Poetry

Fatima Tariq
School of Science and Engineering
Habib University
Karachi, Pakistan
ft07200@st.habib.edu.pk

Ragini Gopchandani
School of Science and Engineering
Habib University
Karachi, Pakistan
rg05951@st.habib.edu.pk

Raza Hashim Nizamani
School of Science and Engineering
Habib University
Karachi, Pakistan
rn07380@st.habib.edu.pk

Abdul Samad
School of Science and Engineering
Habib University
Karachi, Pakistan
abdul.samad@sse.habib.edu.pk

Muhammad Munawwar Anwar*
Department of Computer Science
Emory University
Atlanta, Georgia
muhammad.munawwar.anwar@emory.edu

***Abstract*—Poetry is a profound medium for expressing emotions and beliefs, with South Asian languages like Urdu, Hindi, and Punjabi boasting rich poetic traditions. The digital age and the rise of social media have amplified the visibility of these poetic works but also heightened instances of plagiarism and misattribution. In this paper, we present a Deep Learning model for Poet Attribution for Punjabi poetry using Shahmukhi, Gurmukhi, and Roman scripts. Our dataset consists of 830 poems from 11 different poets. We utilize Multilingual DistilBERT to generate the embedding of each poem in the 768-dimensional vector space. By conducting extensive experiments and utilizing advanced approaches, including Bi-LSTM and Bi-GRU we obtained a test accuracy of 91.57% on the Roman script. Additionally, we achieved accuracies of 90.36% and 87.95% on the Gurmukhi and Shahmukhi scripts, respectively.**

I. Introduction

Poetry is one of the most popular ways to express emotions and beliefs. South Asian languages such as Urdu, Hindi, Farsi, Sindhi, and Punjabi have rich poetic traditions due to the significant contributions of poets like Rumi, Allama Iqbal, Bulleh Shah, and Shah Abdul Latif Bhittai. Despite their passing, their poetic works and philosophical thoughts remain relevant today.

With the advent and rise of social media platforms, these works of art have gained prominence in the public sphere, with many being widely shared. However, this rise has also led to increased instances of plagiarism and misattribution. One potential solution to counter this trend is the development of a Poet Attribution model.

In this paper, we present a Deep Learning Model-based Poet Attribution approach for the Punjabi language. We have chosen Punjabi because, to the best of our knowledge, little work has been done in this area for Punjabi. Additionally, Punjabi is spoken by approximately 100 million people in the Punjab region of Pakistan and India, making it the eighth most spoken native language in the world [15]. The 2023 census of Pakistan lists the population of Punjab at 127 million, approximately 52% of Pakistan's total population [8].

Punjabi is a digraphic language, using more than one writing system. In Pakistan, it is written in the Shahmukhi alphabet, based on the Perso-Arabic script, while in India, it uses the Gurmukhi alphabet, based on Indic scripts. A significant number of social media users also use the Roman script. Therefore, we have developed separate Poet Attribution models for Roman, Shahmukhi, and Gurmukhi scripts. [15]

Since no existing dataset for Punjabi Poet Attribution in Shahmukhi, Gurmukhi, and Roman scripts was available, we curated our dataset from scratch. Our dataset includes a total of 830 poems from 11 different poets listed below:

1) Ali Haider
2) Baba Farid
3) Fazal Shah Sayyad
4) Guru Nanak
5) Kareem Bakhsh
6) Khawaja Ghulam Farid
7) Kush Taba
8) Shah Muhammad
9) Shiv Kumar Batalvi
10) Sultan Bahu
11) Waris Shah

The remainder of this paper is structured as follows: Section II provides an overview of existing approaches to Poet Attribution in various languages. Section III details our proposed methodology, including data preparation and the Machine Learning and Deep Learning

* This research was performed while the author was at Habib University.

models chosen. Section IV presents the experimental setup, including evaluation metrics used to assess model performance. Section V discusses the results obtained and analyzes the effectiveness of our approach. Finally, Section VI highlights the limitations of our study.

II. Literature Review

Pandian et al. [10] developed an author identification model for Tamil classical poems using the *Mukkoodar Pallu* dataset with 800 anonymous poems. They extracted lexical, syntactic, and semantic features, performed feature selection with ID3, and used the C4.5 decision tree algorithm, achieving an accuracy of 88.23% with specific parameter settings.

Pandian et al. [9] proposed machine learning models for Punjabi poetry with a dataset of 400 poems from five poets. They extracted various features and used J48 for feature selection, applying 20 different models. The J48 algorithm achieved the highest accuracy of 86.66%.

Tariq et al. [17] focused on poet identification for Urdu couplets using a dataset of 3,967 couplets from four poets. They employed Term Document Frequency (TDF) for feature extraction, and feature selection was done with Chi-Square and L1-Norm. The SVM model with L1-Norm performed best, achieving an F1-Score of 72%.

Ahmed et al. [1] worked on Arabic poetry with a dataset of 21,929 poems from 114 poets. They extracted seven types of features and tested Naive Bayes, SVM, and LDA classifiers. LDA with Specific Word features achieved the highest accuracy of 99.12%.

Pandian et al. [11] developed models for Hindi poetry using 100 poems from three poets. They extracted lexical, statistical, and semantic features, performed feature selection with J48, and tested 17 models. Logitboost performed best with an accuracy of 75.67%.

Dar [5] compared machine learning and deep learning models for Urdu couplet attribution with a dataset of 11,406 couplets. After data cleaning and tokenization, models such as SVM, Naive Bayes, and MLP were evaluated. The RBF Kernel SVM achieved the best accuracy of 82.85%, with F1-measure, recall, and precision scores of 82.67%, 82.67%, and 83.00%, respectively.

Ahmed and Rao [12] compared SVM, LSTM, Naive Bayes, and DNN models for Urdu poetry with a dataset of 1,563 poems. They varied the number of lines and n-grams. The SVM model with 10 lines per poem and a combination of unigram and bigram achieved the highest accuracy of 88.77%.

Ekinci et al. [6] assessed ANNs and DNNs for Turkish poetry identification using 314 poems. They used TF-IDF features for MLP and GloVe embeddings for CNN. The MLP outperformed the CNN with an accuracy of 81%.

Siddiqui et al. [16] compared various models for Urdu couplet identification using a dataset of 18,472 couplets from 15 poets. They used TF-IDF and Word Embedding for feature extraction and tested models including SVM, Logistic Regression, Naive Bayes, Random Forest, MLP, LSTM, GRU, CNN, BERT, and RoBERTa. BERT and RoBERTa achieved the highest accuracy of 80%.

III. Methodology

A. Methodology Overview

Our methodology includes the following key steps:

1) **Data Splitting:** The dataset, with poems in Roman, Gurmukhi, and Shahmukhi scripts, is divided into training (80%), validation (10%), and test sets (10%) using Scikit-learn.
2) **Tokenization:** Poems are tokenized into subwords with padding and truncation, setting the maximum length to 80 tokens.
3) **Embedding Generation:** Tokenized poems are converted into 768-dimensional embeddings using the DistilBERT multilingual model.
4) **Model Training:** Machine learning models (RandomForest [2], SoftMax [3], SVM [18]) and deep learning models (Bi-GRU, Bi-LSTM [4], [7], [14], DistilBERT [13]) are trained on these embeddings.
5) **Model Evaluation:** Performance is evaluated on the test set using accuracy.

B. Dataset

Our Dataset contains 830 poems from 11 distinct poets in three different scripts: Gurmukhi, Shahmukhi, and Roman. Punjabi text in different scripts can be seen in Table I.

Script	Text
Gurmukhi	ਲਹੌਰ ਪਾਕਸਿਤਾਨੀ ਪੰਜਾਬ ਦੀ ਰਾਜਧਾਨੀ ਹੈ। ਲੋਕ ਗਣਿਤੀ ਦੇ ਨਾਲ਼ ਕਰਾਚੀ ਤੋਂ ਬਾਅਦ ਲਹੌਰ ਦੂਜਾ ਸਭ ਤੋਂ ਵੱਡਾ ਸ਼ਹਿਰ ਹੈ। ਲਹੌਰ ਪਾਕਸਿਤਾਨ ਦਾ ਸਆਸੀ, ਕਾਰੋਬਾਰੀ ਅਤੇ ਪੜ੍ਹਾਈ ਦਾ ਗੜ੍ਹ ਹੈ ਅਤੇ ਇਸੇ ਲਈ ਇਹਨੂੰ ਪਾਕਸਿਤਾਨ ਦਾ ਦਲਿ ਵੀ ਕਹਿਾ ਜਾਂਦਾ ਹੈ। ਲਹੌਰ ਰਾਵੀ ਦਰਆਿ ਦੇ ਕੰਢੇ 'ਤੇ ਵੱਸਦਾ ਹੈ। ਇਸਦੀ ਲੋਕ ਗਣਿਤੀ ਇੱਕ ਕਰੋੜ ਦੇ ਨੇੜੇ ਹੈ।
Shahmukhi	لہور پاکستانی پنجاب دی راجدھانی اے۔ لوک گنتی دے نال کراچی توں بعد لہور دوجا سبھ توں وڈا شہر اے۔ لہور پاکستان دا سیاسی، رہتلی کاروباری اتے پڑھائی دا گڑھ اے اتے، ایسے لئی ایہہ نوں پاکستان دا دل وی کہا جاندا اے۔ لہور راوی دریا دے کنڈھے تے وسدا اے۔ ایہدی لوک گنتی اک کروڑ دے نیڑے اے۔
Roman	Lahore Pakistani Punjab di rajdhani hai. Lok ginti de naal Karachi ton baad Lahore dooja sab ton vadda shehar hai. Lahore Pakistan da siyasī, kārobārī ate paṛhāī da gaṛh hai ate, ise leī ihnū Pakistan da dil vī keha janda hai. Lahore Ravi dariya de kanḍhe te vasda hai. Isdī lok ginti ikk karoṛ de neṛe hai.

Table I: Punjabi text in Gurmukhi, Shahmukhi, and Roman scripts

Our dataset was sourced from two main platforms: GitHub and Folk Punjab. The GitHub repository provided poems by Baba Bulle Shah, Bhai Veer Singh, Fazal Shah, Sultan Bahu, and Ustad Daman in Gurmukhi script. These poems were converted to Shahmukhi and Roman scripts using ChatGPT, and each conversion was verified by experts. The Folk Punjab website supplied poems in Gurmukhi, Shahmukhi, and Roman scripts. To ensure a fair comparison, we only included poets with data available in all three scripts. Details about the dataset,

including poets and poem counts, are summarized in Table II.

Poet	Roman	Gurmukhi	Shahmukhi
Ali Haider	29	29	29
Baba Farid	131	131	131
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
Waris Shah	90	90	90

Table II: Number of Poems per poet

IV. Experiments

A. Experimental Setup

We evaluated model performance using accuracy. The dataset was divided into 664 samples for training, 83 for testing, and 83 for validation. Model training was performed on cloud platforms, specifically Google Colab with 12GB RAM and an NVIDIA Tesla T4 GPU, and Kaggle's Kernel with 29GB RAM and two NVIDIA Tesla T4 GPUs.

B. Machine Learning Models

The hyperparameters used for the Machine Learning Models are summarized in Table III.

Model	Hyper Parameter	Value
Random Forest	Number of Trees	150
Softmax Regression	Max Iterations	1000
SVM	Kernel	Linear

Table III: Hyperparameters for Machine Learning Models

C. Deep Learning Models

The hyperparameters used for the Deep Learning Models are summarized in Table IV.

Hyper Parameter	Value
Epochs	100
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
Loss function	Cross Entropy Loss
Early Stopping Patience	10

Table IV: Hyperparameters for Deep Learning Models

V. Results

We tried multiple Machine Learning and Deep Learning models for Poet Attribution on three different Punjabi scripts and then evaluated these models using accuracy. The results that we obtained on the test dataset are summarised in Table V

Model	Gurmukhi (%)	Roman (%)	Shahmukhi (%)
Bi-LSTM	87.95	91.57	81.93
Bi-GRU	89.16	87.95	80.72
DistilBERT	90.36	87.95	87.95
SoftMax	86.75	84.34	79.52
SVM	85.54	89.16	80.72
RandomForest	63.86	71.08	68.67

Table V: Performance comparison of models on Gurmukhi, Roman, and Shahmukhi scripts

A. Machine Learning Models

Since most machine learning models displayed similar trends, we have included only the confusion matrices for the best-performing model, the SVM Linear Classifier. This model achieved the highest performance on the Gurmukhi script, with an accuracy of 89.16%. It performed next best on the Roman script with 85.54% accuracy, and the Shahmukhi script with 80.72% accuracy. The confusion matrices for these three scripts, generated by the SVM model, are presented in Figures 1, 2, and 3. Each confusion matrix shows a consistent pattern across the scripts.

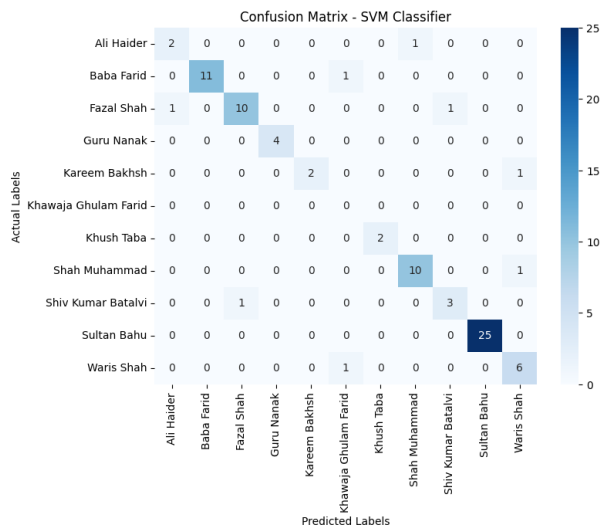

Figure 1: Confusion Matrix for the Gurmukhi script using SVM

B. Deep Learning Models

Since most deep learning models exhibited similar trends, we have included only the plots for the best-performing model, the Bi-LSTM classifier. This model achieved varied performance across different scripts: 91.47% accuracy for the Roman script, 87.95% for the Gurmukhi script, and 81.93% for the Shahmukhi script. The plots for accuracy vs. epochs, loss vs. epochs, and the confusion matrix for the Gurmukhi script are shown in Figures 4, 5, and 6, respectively. The plots for the other two scripts displayed similar trends and are therefore not included.

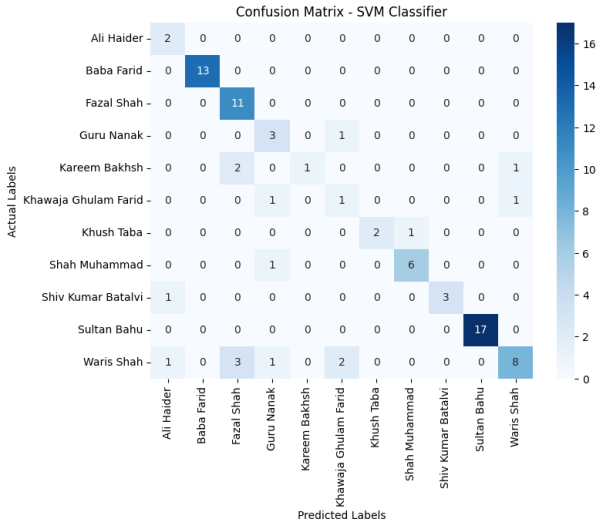

Figure 2: Confusion Matrix for the Shahmukhi script using SVM

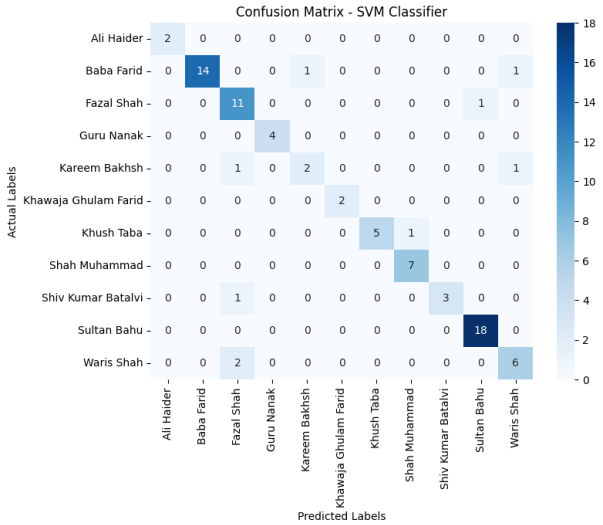

Figure 3: Confusion Matrix for the Roman script using SVM

C. Analysis

In this section, we highlight the notable trends identified after analyzing the results reported in the previous sections. The notable trends are listed below.

1) Overall Trends: Overall, the results on the test dataset indicate that deep learning techniques such as Bi-LSTM, Bi-GRU, and DistilBERT are more effective for poet attribution tasks across various scripts. Traditional machine learning techniques like SVM, SoftMax, and Random Forest, while still performing adequately, generally lag behind their deep learning counterparts. This highlights the advantage of using neural network-based architectures for tasks involving textual data in different scripts.

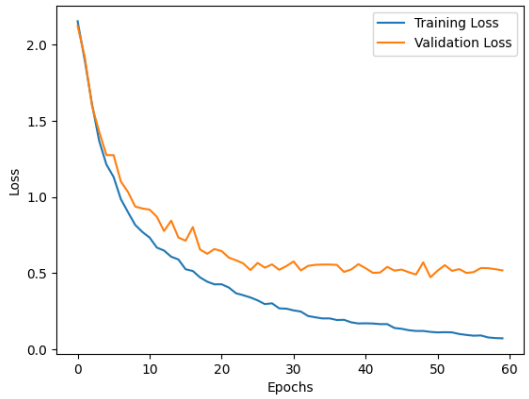

Figure 4: Plot of Loss vs. Epochs of the Bi-LSTM Model on the Gurmukhi Script

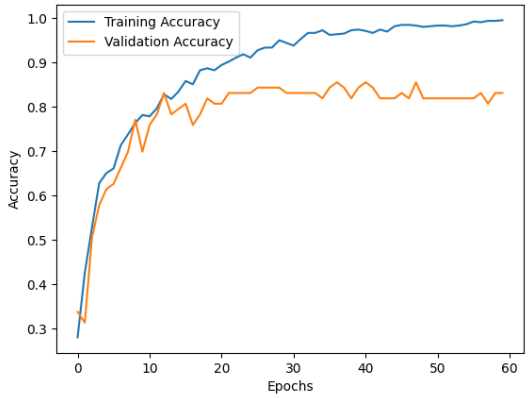

Figure 5: Plot of Accuracy vs. Epochs of the Bi-LSTM Model on the Gurmukhi Script

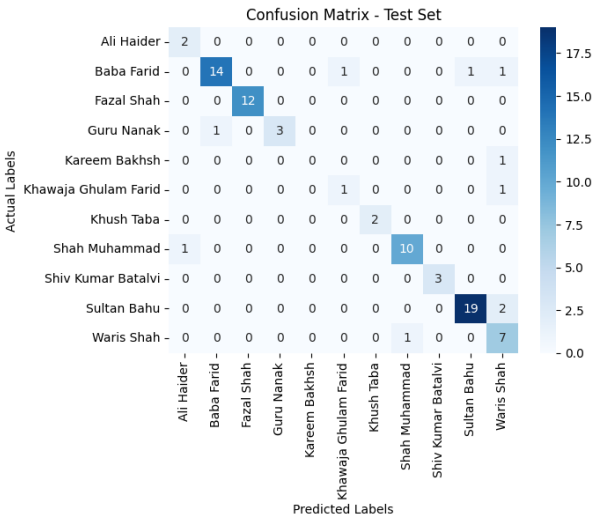

Figure 6: Confusion Matrix for the Gurmukhi Script Using Bi-LSTM

2) Script Performance: In comparing the scripts, the Roman script consistently achieves the highest accuracy across all models, with the Bi-LSTM model achieving a notable 91.57%. This suggests that the Roman script's structure and features are more easily captured by the models, leading to better performance. The Gurmukhi script also performs well, especially with the DistilBERT model, which achieves 90.36% accuracy. In contrast, the Shahmukhi script tends to have lower accuracy across most models, highlighting potential challenges in feature extraction and classification for this script. These differences emphasize the importance of script-specific considerations in developing effective poet attribution models.

3) Model Performance: When comparing the performance of different models across the Gurmukhi, Roman, and Shahmukhi scripts, it is evident that deep learning models generally outperform traditional machine learning models. DistilBERT, Bi-LSTM, and Bi-GRU consistently achieve higher accuracy rates compared to SoftMax, SVM, and RandomForest. Specifically, DistilBERT stands out with the highest accuracy for both Gurmukhi (90.36%) and Shahmukhi (87.95%), indicating its robustness and effectiveness in handling script-based text classification. Bi-LSTM, on the other hand, shows the highest accuracy for the Roman script at 91.57%, suggesting its strong performance across different languages and scripts. The traditional machine learning models, particularly RandomForest, perform significantly worse, highlighting the advantages of deep learning models in this domain.

4) Poet Performance: When comparing the confusion matrices of different models obtained across the Gurmukhi, Roman, and Shahmukhi scripts, it is evident that models perform well overall, with most poets such as Baba Farid and Sultan Bahu being correctly identified. However, there are notable misclassifications, such as certain poets being frequently confused with others. For example, Kareem Baksh being confused with Fazal Shah and Waris Shah. This trend indicates that while the models are generally accurate, there are specific poets whose writing styles are similar, leading to some level of confusion.

5) Training Process: The training plots of loss versus epochs show a clear learning trajectory. The loss versus epochs plot indicates a consistent decrease in both training and validation loss, signifying that the models are learning the patterns in the data effectively. The training loss decreases rapidly initially and then gradually stabilizes, while the validation loss exhibits a similar trend with occasional fluctuations. This indicates that the models are fitting well to the training data while also generalizing to unseen data. The plots of accuracy versus epochs complement these trends by showing a steady increase in both training and validation accuracy, eventually becoming stable. This underscores the effectiveness of the models in learning and adapting to the textual data throughout the training process.

VI. Limitations

Although the dataset that we gathered was much larger than any previous work in the field of Poet Attribution for Punjabi and had comparable results, our project faced multiple limitations.

The first major limitation was the limited representation of some poets. For example, our dataset contained only 20 poems of Khawaja Ghulam Farid, even though he had written many more poems. This limited representation led to class imbalance, which impacted the model's performance on unseen data. A balanced dataset could lead to better training results and improved model performance.

The second major limitation was that our dataset can be more diversified. Currently, our dataset contains poems from 11 poets and omitted significant figures such as Faiz Ahmed Faiz and Allama Muhammad Iqbal. Including a more diverse range of poets would not only make the dataset more appealing to a broader audience, but also enable us to cover more unique styles and subjects. This increased diversity would likely enhance the model's ability to generalize across different poetic styles and themes.

References

[1] Al-Falahi Ahmed, Ramdani Mohamed, and Bellafkih Mostafa. Arabic poetry authorship attribution using machine learning techniques. *Journal of Computer Science*, 15(7):1012–1021, Jul 2019.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Françoise Fogelman Soulié and Jeanny Hérault, editors, *Neurocomputing*, pages 227–236, Berlin, Heidelberg, 1990. Springer Berlin Heidelberg.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[5] Momna Dar. Authorship attribution in urdu poetry. 06 2020.

[6] Ekin Ekinci, Hidayet Takcı, and Sultan Alagöz. Poet classification using ann and dnn. *Electronic Letters on Science and Engineering*, 18(1):10–20, 2022.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[8] Pakistan Bureau of Statistics (PBS). Announcement of results of 7th population and housing census-2023. Technical report, Pakistan Bureau of Statistics (PBS), 2023.

[9] A. Pandian, Stephen Wahid, Yash Tokas, and V. V.Ramalingam. Authorship identification of punjabi poetry. *International Journal of Engineering amp; Technology*, 7(4.19):13–16, Nov. 2018.

[10] A. Anbarasa Pandian, V. 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*, 9:1–5, 2016.

[11] Dr. A. Pandian, Paritosh Maurya, and Nitin Jaiswal. Author identification of hindi poetry. 2020.

[12] M. Adil Rao and Tafseer Ahmed. Poet attribution for urdu: Finding optimal configuration for short text. *KIET Journal of Computing and Information Sciences*, 4(2):12, Jul. 2021.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[14] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[15] Christopher Shackle. Punjabi language.

[16] Iqra Siddiqui, Fizza Rubab, Haania Siddiqui, and Abdul Samad. Poet attribution of urdu ghazals using deep learning. In *2023 3rd International Conference on Artificial Intelligence (ICAI)*, pages 196–203, 2023.

[17] Nida Tariq, Iqra Ijaz, Muhammad Kamran Malik, Zubair Malik, and Faisal Bukhari. Identification of urdu ghazal poets using svm. *Mehran University Research Journal of Engineering and Technology*, 38(4):935–944, 2019.

[18] Vladimir Vapnik, Steven E. Golowich, and Alex Smola. Support vector method for function approximation, regression estimation and signal processing. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 281–287, Cambridge, MA, USA, 1996. MIT Press.