Poet Attribution of Urdu Ghazals using Deep Learning

Haania Siddiqui
Computer Science
Habib University
Karachi, Pakistan
hs06188@st.habib.edu.pk

Fizza Rubab
Computer Science
Habib University
Karachi, Pakistan
fr06161@st.habib.edu.pk

Iqra Siddiqui
Computer Science
Habib University
Karachi, Pakistan
is06176@st.habib.edu.pk

***Abstract*—Poet attribution focuses on determining ownership of a piece of poetry by insights obtained from analyzing his existing poetry. Its significance is immense including in detection of plagiarism and characterization of poetry of a poet. Urdu, Pakistan's lingua franca with the richest poetic tradition, has been a subject of misinformation and misattribution. In this paper, we are exploring poet attribution on a corpus of Urdu Ghazals using machine and deep learning models. The main objective of the paper is to define an appropriate and accurate characterization of ghazals that holistically capture the writing style of the poet. As a result, Machine Learning, Deep Learning and transformer based classification models are trained and tested on a dataset of 18,609 couplets of 15 notable ghazal poets. Transformer based models, namely, Bert and Roberta resulted in highest accuracy of about 80 percent. This paper takes an analytical approach to experiment and analyze each model and explores their capabilities of capturing writing style of Urdu Ghazal poets.**

***Index Terms*—poet attribution, authorship attribution, urdu poetry, classification, computational linguistics, deep learning, transformers**

I. Introduction

One of the most popular means of generating a form of literature with metrical attributes of linguistics including sound and symbolism, is called poetry and this generated literary form is called poem. Poems in general are a combination of words organized in a repeated rhythm with metaphors and similes. Poetry is a figurative language that exists in all humane languages including Urdu.

Urdu is a language with rich asset and heritage of poetry since as early as the 13th century. The language is widely used in South Asian countries including Pakistan and India. Urdu has a scripture drawn from other languages including Persian, Arabic, Hindi, Sanskrit and Turkish. It has different poetic forms such as Nazm, Marsiya, Qaseeda, etc. with Ghazal being the most popular among poets and the readers. Ghazal is a short poem on themes of pain/love/separation, consisting of five to fifteen rhyming couplets, called *shair*, each following certain refrain rules such as *qafiya* or *radif*. Out of all poetic forms in Urdu, Ghazal is the most simple and well-structured one. Even though all ghazal poets follow these refrain and thematic rules however they differ in lexicons and semantics that assists in poet attribution thus making each poet's piece stand out from the rest.

Poet attribution refers to the identification of the owner of a piece of poetry from a list of candidate poets. It is a variant of author attribution problem that is solved by studying different writing styles and characteristics of authors, extracting features from the author's texts(documents and prose) and training AI model to perform classification based on these features. Poet Attribution has a few distinguishing features. In contrast to the articles used in training of authorship attribution models, the couplets used for training and testing in poetry attribution are just composed of two lines(*misra*) and at most 30 words. Out of those words some are repetitive because of the ghazal radif constraints.

There has been extensive research in the domain of authorship attribution as well as poet attribution using deep learning but in different languages and is relatively less in Urdu. A reason behind is the complex nature of Urdu Poetry in general due to elements like meter, weight, radif and qafiya. The potential of identifying poets to a degree of accuracy by just a ghazal couplet is not only interesting but can be used to prevent misinformation and misattribution.

II. Research Question

The research question we are addressing is 'Poet Attribution of Urdu Ghazals using Deep Learning". Our problem statement can be formulated as

'Given a couplet from a Ghazal, identify who the poet is'

We will be exploring and training different deep learning models by using existing models and designing our own neural networks to address this classification problem. The input to the model will be a couplet from any Ghazal and its output would be class label of the poet. The poets selection is carried out from list of 15 notable Urdu Ghazal poets listed below. More information is provided in the data set section.

1) Mir Taqi Mir
2) Mirza Ghalib
3) Haider Ali Atish
4) Nazeer Akbar Abadi
5) Faiz Ahmed Faiz
6) Ahmed Faraz
7) Parveen Shakir
8) Jaun Elia

9) Allama Iqbal
10) Muneer Niazi
11) Qateel Shifai
12) Siraj Aurangabadi
13) Riyaz Khairabadi
14) Nida Fazli
15) Zafar Iqbal

The primary motivation behind this research is that there is a lot of misinformation about poets and their poetry pieces especially in South Asian languages. With the increasing use of the social media and communication platforms, people are observed sharing couplets with incorrect poet names and readers believe them unquestioningly. The offence of plagiarism and wrong attribution pertaining to Urdu ghazals and other poetic forms has become very common. The deep learning model devised as a part of this research can be used for validation and verification of such poetry. It can be gauged whether a couplet claimed to be written by a poet, for example Mirza Ghalib, is indeed of Ghalib's.

Furthermore, there is little research available in the domain of poet attribution particularly in Urdu language however there is available work on sibling languages of Urdu including Arabic and Hindi. This compels us to explore this in Urdu as well. The existing researches in Urdu use classical machine learning methods such as Naive Bayes, Support Vector Machines and Decision tree Classifiers. Moreover, they cater to very few poets so there is room for catering to more by acquiring a more extensive data set. Therefore, the selection of feature set and techniques for better model training on Urdu Ghazal Corpus is at the core of our research.

III. Literature Review

The papers that we reviewed can be classified into two categories that is, a category of research in poet attribution and other of authorship attribution. A brief review of major concerned papers of the two categories is mentioned in this section.

A study [1] was conducted for the classification of Persian *Ghazals* according to era of the poet using sequential learning. The method of classification developed was able to detect the chronological order of each poem with respect to the author's lifespan. The dataset collected was of Hafez's *ghazals*. It consisted of 496 samples, each with a Persian version, English translation, chronological label (six classes) and Raad label[1]. Word embedding, which is an important technique in Natural Language Processing, and deep learning based classification models were used for the purpose of classification. Comparison and evaluation of the their approaches with other classification models was also done. Features were extracted using Bag of Words(BoW), and Latent Dirichlet Allocation(LDA). Machine Learning (ML) and Deep Learning(DL) models were analyzed for the purpose of classification. For ML classifiers, Random Forest and Logistic Regression using k-fold validation. For *ghazal* classification, Long Short Term Memory (LSTM), Gated Recurrent Units (GRU) and Bidirectional-LSTM (Bi-LSTM) were used. The two types of class labels, chronological and Raad were applied to both Persian and English data. For evaluation metrics, accuracy, F1-score, precision and recall were used to analyze the performance. The ML classifiers were compared with Support Vector Machine (SVM). The classifiers outperformed SVM in terms of F1-score but not in terms of accuracy on all the datasets. Among DL classifiers, LSTM gave highest precision and recall on Persian dataset.

[1]Dataset was labelled according to two politicians and was categorized as 'before Amir Mobarezeddin', 'Amir Mobarezeddin', 'Shah Shoja' and 'after Shah Shoja', consisting of four classes.

Another research [6] was done on the poet identification in Urdu Language. The dataset consisted of 11406 couplets from poetries of three different Urdu Poets, *Faiz*, *Iqbal*, *Ghalib*. For machine learning classification, Support Vector Machine, Multilayer Perceptron, Word2vec (pretrained) with sentence embedding, Multinomial Naive Bayes was used. Evaluation metrics used were f1-score, precision, recall and accuracy. SVM got the highest accuracy of 82.85 percent and F1-score of 82.67 percent. The study suggested to use other deep learning models, CNNs and LSTM for the similar purpose to improve the performance.

In addition, we came across a study [5] that has performed comparative analysis to analyze different identification models to find the one which performs the best for poet attribution in Urdu poetry. The dataset consisted of 1563 samples of top 5 poets with highest count of poems. The poets were Mir Taqi Mir, Mirza Ghalib, Nazeer Akbar Abadi, Faiz Ahmed Faiz and Allama Muhammad Iqbal. During the pre-processing of data, stop words were removed in other studies but the authors of this research didn't remove it because the stop words are able to capture the writing style of the author. 4690 experiments were conducted based on the classification algorithms, number of lines in a poetry, and use of n-gram features to evaluate the optimal approach. Naive Bayes, Support Vector Machines, and LSTM were used as classification algorithms. To evaluate if proper identification can be done using less number of lines, experiments were performed on different lengths of poetry where number of lines ranged from 2 to 10. For n-gram features, unigram, bigram and combination of unigram and bigrams were used. The experiment where SVM and combination of unigram and bigram was used outperformed all other experiments with an accuracy of 88.7 percent. They also concluded that minimum 100 samples per poet, more than two lines per sample, can be used to get good results in Urdu poet identification.

A similar research [8] worked on comparative analysis of different machine learning classification models for the purpose of identification of Urdu *ghazals*. The dataset consisted of 4000 samples of four different poets: Ahmed Faraz, Meer Taqi Meer, Zafar Iqbal, and Mirza Ghalib. The machine learning classifiers used were Support Vector Machine, Naive Bayes, K-Nearest Neighbor, Random Forest

and Decision Trees. Three different experiments were conducted, one without feature selection and other two with Chi-2 and L1 feature selection. F1-score was used as evaluation metric. SVM outperformed all other classifiers in all three experiments and performance of SVM was improved using feature selection achieving a score of 74 percent.

In addition to Urdu, research in other language poetry such as Arabic were also found such as a study [12] that was conducted for poet identification in Arabic poetry. Naive Bayes and Support Vector Machines were used as classification algorithms. The dataset comprised 18646 *Qasidahs* of 73 poets. Lexical and character features were extracted and used for determination of style. Feature selection, using Chi-Squared and Information Gain, was applied to remove the unimportant features and use the most relevant features for the classification. The six features used were character, word length, sentences length, first word length, meter, rhyme and average. Recall and accuracy were used as evaluation metrics. The best performance, 98.86 percent, in terms of accuracy was of SVM when all six features were used. However, Naive Bayes gave the similar accuracy using two features only. Moreover, average recall of Naive Bayes, 91 percent was higher than that of SVM.

We came across other approaches such as using emotion classification for the poet attribution as presented in one of the study [4] that uses sentiment classification in the Gujarati Poems. The dataset was a collection of 300 poems with nine classes of emotions. For feature extraction, Natural Language Processing Token Extraction is used where extracted tokens are then processed for emotion classification. Using Zipf's law, probability of one *rasa*, emotion, is found from the given poem. The occurence of the extracted word is found from a predefined dataset, 'Rasa', consisting of words that can be used to identify the emotion in a poetry according to the 'Navrasa Concept'[2]. The word with maximum frequency is calculated and poetry is classified accordingly. The approach was successful for emotion classification in Gujarati Poems.

Another research [7], worked on the identification of poets in Hindi poetry. The dataset consisted of 100 poems from three different authors. Features, including lexical, statistical and syntactic, were extracted and calculated which were then given as input to the machine learning classifiers. The classification was performed using decision tree using seventeen machine learning algorithms where Logitbost gave the best performance in terms of accuracy, 75.67 percent.

As we expanded our review to the second category that is the authorship attribution, we came across a study [3] that focused on the identification of authors given Urdu text samples. A stylometric feature space was developed using richness of characteristic and vocabulary, n-gram and character n-gram features. Using these features, "structural, syntactic and lexical" information about author's style was captured. Probabilistic k nearest neighbors classifier was used for the purpose of classification because it can be trained with less data. Each text sample was represented as a set of vectors for the identification of variations in writing style, hence, to calculate similarity between two text samples(or vectors), partial Hausdorff distance was used. The dataset consisted of 985 Urdu text samples of 90 authors. The suggested method outperformed all the methods implemented in previous studies,and classical machine learning methods which included "decision trees, naive bayes, support vector machines and random forests" with an accuracy of around 94 percent.

Other than the aforementioned important researches relevant to our problem statement, other papers were also reviewed to get a better understanding of the challenges and approach to solve the problem of poet attribution and its distinction from the authorship attribution problem. The details are summarized in Table I.
According to the literature review, machine learning algorithms including SVM, Naive Bayes, LSTMs were used the most for the identification task. Moreover, the task of poet identification in Urdu was performed on maximum 5 poets and comparatively less data. Our goal is to use a larger dataset. Feature selection was also an integral part in all the studies because it captures the style of the writing of the poet, according to which classification was performed.

IV. Material and Methodology

In this section, major steps of our proposed framework for poet attribution are discussed, with emphasis on the dataset, models along with their specific parameter settings and experimentation. The discussion on corpus and dataset is in Section IV-A, then we present three broad experimentation's for poet attribution, namely using traditional Machine learning techniques in Section IV-B, Deep Learning models in Section IV-C and state of art natural language processing models in Section IV-D.

A. Data Set

In this section, we have discussed necessary information related to the samples that were used in Poet attribution model training and testing, including details regarding corpus, data set, and its acquisition and characterization.

We have collected a corpus of 18,609 couplets from 15 different notable Urdu poets. These couplets(consisting of 2 misras concatenated together) will be passed to our model for training and testing. Our sole source of data is Rekhta. Rekhta is an Indian literary website owned by the Rekhta Foundation, a nonprofit and non-governmental organization which is devoted to promoting and encouraging Urdu poetry and prose in the subcontinent. The data set is publically available and can be accessed from here.

[2]According to this concept, nine emotions are the core of any art and literature

TABLE I
LITERATURE REVIEW SUMMARY

Year	Model	Dataset	Accuracy
[1] Sep 2022	DMM and LSTM	495 ghazals	85%
[2] May 2022	CNN	94 authors	98.9%
[3] Mar 2022	NN Classifier	N/A	94.03%
[4] May 2021	Classifier	'Kavan'	85.62%
[5] 2020	SVM+uni&biagram	5 Poets	88.7%
[5] 2020	NB Classifier	5 Poets	77%
[6] June 2020	Multinomial NB	3 Poets	78.81%
[6] June 2020	RBF Kernel SVM	3 Poets	82.85%
[6] June 2020	MLP	3 Poets	80.67%
[6] June 2020	MultinomialMLP	3 Poets	78.67%
[7] Mar 2020	Logit-bost algorithm	3 authors	75.67%
[7] Mar 2020	Decision tree algorithm	3 authors	70%
[7] Mar 2020	Iterative Classifier Optimizer	3 authors	75%
[7] Mar 2020	Classification via Regression	3 authors	71%
[7] Mar 2020	KStar and BayesNet	3 authors	60%
[8] Oct 2019	KNN[1]	4 Poets	45%
[8] Oct 2019	Decision Tree[1]	4 Poets	46%
[8] Oct 2019	Random Forest[1]	4 Poets	50%
[9] 2019	LDA[2]	Pan12	84.52%
[9] 2019	LDA[2]	12 authors	93%
[10] Dec 2018	LDA	21,938 news	92.89%
[11] July 2018	CNN	5.9M Words	57.3%
[12] 2017	NB	73 Poets	97%
[12] 2017	SVM	73 Poets	98%

[1]without feature selection and also with Chi-2 and L1 Based
[2]with cosine metric for KNN classifier

1) Acquisition: The primary source of data is ghazals and couplets from Rekhta. Its acquisition was a challenge that required both manual and programmatic effort. Each poet has a webpage on rekhta which contains links of his/her ghazals. The page on loading shows 50 ghazal links. If there are more, then the user has to navigate to the bottom and wait for them to load. We visited each poets page manually and waited for all of the links to load, then we extracted the element containing the links and extracted the links. Then using an automated program, each link was visited, and finally Urdu couplets were extracted using regular expressions and string processing. We have chosen Rekhta as the only source of data because it is reliable unlike the other sources used by literature. The data at this stage seems sufficient. If more is required at any later stage, we will scrape other sources like Urdu Web, Iqbal, and Sukhansara etc.

2) Characterization: There are total of 18,609 couplets from 15 poets. Classwise count can be seen in the following table

The choice of the above poets can be justified by the following factors:

1) The chosen poets are an almost equal blend of notable contemporary and historical poets of Urdu Ghazal. Poets from [1-8] wrote in the 19th-20th century while the rest [9-15] belong to the older 18th-19th era. Since the styles are considerably different in terms of word usage and urdu dialects, it is beneficial for the model as the patterns will be more evident and our model will be able to cater to both.
2) Another criteria for selection of poets is their abundance of ghazals. We handpicked the poets that had most ghazals so our dataset could be large enough, for better training our deep learning model. There were poets that were more popular but had only 200 couplets and hence were discarded. In the current data set, every poet has at least 450 couplets which the model could be trained on with the upper bound being 2986 for Mir Taqi Mir.
3) All poets chosen are popular and cherished in Urdu literary circles so the analysis and results of the research can be appreciated by the non computer science readers alike.

TABLE II
CORPUS BREAKDOWN

No	Poet/Shayar	Couplet Count
1	Ahmed Faraz	926
2	Zafar Iqbal	1104
3	Qateel Shifai	780
4	Parveen Shakir	593
5	Nida Fazli	474
6	Faiz Ahmad Faiz	504
7	Jaun Elia	1470
8	Muneer Niyazi	523
9	Allama Iqbal	799
10	Riyaz Khairabadi	1700
11	Haider Ali Atish	1330
12	Siraj Aurangabadi	860
13	Mir Taqi Mir	2971
14	Nazeer Akbar Abadi	1643
15	Mirza Ghalib	1934

B. Traditional Machine Learning Models

We found from our literature review that traditional machine learning models are quite popular in Poet attribution as discussed in Section III. Therefore we experimented with four machine learning classification techniques on the extracted features, that have showed better performance. This section presents discussion on each experiment. The following diagram presents an overview of our experimentation with each machine learning models.

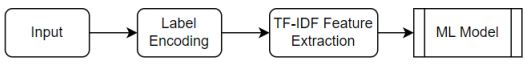

Fig. 1. Experimentation Overview for Machine Learning

Pre-Processing: Even though a deep pre-processing pipeline was not required as observed from Section III however, the data was encoded and features were extracted before experimenting the machine learning models on it.

a) Label Encoding: Label encoding is a process of converting labels into numeric values that can be fed into the

model. Label encoding was implemented using preprocessing module of sklearn library that encodes the labels from 0 to n-1 where n is the number of classes. In our case, data was encoded from 0 to 14 as dataset has 15 poets as label classes.

b) Feature Extraction: The process of extracting numerical information from the raw data is called feature extraction. In this procedure, only those features are selected that are best for the training model. Term Frequency-Inverse Document Frequency (TD-IDF) feature extraction for this purpose which has also been used in cited literature. It comprises of two parts i.e. the term frequency metric (tf) and inverse document frequency metric (idf), where
$tf(t)$ is the number of times word t appeared
and

$$idf(t) = log(\frac{1+n}{1+df(t)}) + 1 \text{ [13]}$$

Hence, tf-idf is given by:

$$tfidf(t) = tf(t) * idf(t) \text{ [13]}$$

Sklearn library's built in TfidfVectorizer was used to obtain the features automatically.
The dataset was then divided into 80% training and 20% testing data. Finally, following machine learning classification techniques were implemented on the pre-processed data.

1) SVM: Section III presented Support Vector Machine (SVM) as the commonly used ML technique for poet attribution, hence we experimented SVM on our extracted features. SVM finds a hyperplane/support vector in the multi-dimensional space where number of dimensions is the number of features that classifies the data points. We have trained Linear SVC for a maximum of 100000 iterations on our training dataset. The implementation was done in python using svm module of sklearn library.

2) Logistic Regression: The Logistic regression model measures the relationship between the categorical dependent variable with independent variables by approximating the probability of occurrence using logistics function and classifies based on these probabilities. We have trained the logistic regression linear model for a maximum of 100000 iterations on our training dataset. The implementation was done in python using logistic regression module of sklearn library.

3) Random Forest: Random forest creates set of decision trees of randomly selected subsets of training data and classifies based on the score obtained from each decision tree. We have trained the Random forest Classifier over train dataset, with 450 random decision trees, arbitrary random state and height of each tree set to 9. This combination of parameters resulted in relatively better results and were set based on trial and error. The testing results will be discussed in Section V. The implementation was done in python using RandomForestClassifier module of sklearn library.

4) Naive Bayes: Naive Bayes algorithm classifies by applying Bayes theorem with the conditional independence assumption between every pair of class features and then uses maximum A Posteriori estimation to conclude the relative frequency of a certain class in the training set, i.e.,

$$y_{predict} = argmax P(y) \prod_{i=1}^{n} P(x_i|y) \text{ [14]}$$

The Naive Bayes implemented for multinomially distributed data is called multinomial Naive Bayes that parameterize the distribution for each class. We have trained the multinomial naive bayes over train dataset and have used the trained model to predict test data. The implementation was done in python using MultinomialNB module of sklearn library.

We have evaluated above models based on the accuracy and validation loss. Results will be discussed in Section V.

C. Deep Learning Models

Deep learning model works based on the structure and operation of human brain. As a consequence of which, they are elegant and powerful models that are used for myriad applications in diverse areas. Here, we present four common deep learning models trained for poet attribution of urdu ghazals.The following diagram presents an overview of our experimentation with each deep learning models.

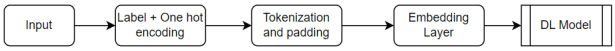

Fig. 2. Experimentation Overview for Deep Learning

Pre-processing: It was deduced from the Section III that robust pre-processing and cleaning was not required for Poet Attribution as each poet has one's own syntactical/grammatical mistakes and preferences of word styles. For distinctive features, it is important that grammatical mistakes and terminals are not corrected, as all of these adds to the poet specific feature list. Moreover the problem has been simplified into categorical classification task by encoding the data set. The encoded data is then tokenized into padded sequences for input to the embedding layer.

a) Encoding: The dataset has been encoded using one-hot encoding scheme to convert categorical data variables into a numerical value. The encoding was done using python numpy library. As we have 15 poets in the dataset, all couplets are mapped to an integer from the range 0-14.

b) Tokenization:: After encoding the train and test datasets, vocabulary of words was built to generate sequences using tokenization. The word-index dictionary was created where each word in the vocabulary was assigned a particular index. The training sentences were then used to generate sequences and pad them. The procedure was repeated for validation and test examples. The implementation was done using Tokenizer module of tensorflow.keras python library.

c) Embedding:: With vocabulary generated in the tokenization phase, next step was defining an embedding. This was done by adding an embedding layer of a standard dimension of 300 neurons as the first layer of each DL model using the keras.Layers package.

The dataset was then divided into 80% training and 20% testing data. Finally, The padded sequences obtained from the pre-processing stage were used for training following four traditional deep learning models.

1) Flatten: We trained a model with flattened embedding layer of dimension 300. It is a 6 layer model with two dense layers of 256 and 64 dimensions respectively with $relu$ activation, a dropout layer of 0.3, followed by another dense layer with 15 dimension and $sigmoid$ activation.

2) LSTM: Long-Short Term Memory (LSTM) is an efficient RNN model with multiple hidden layers. We have used a four layer model for training with LSTM. The first layer is the embedding layer of dimension 300 on our vocabulary. We have used bi-directional LSTM with dimension 64 and $tanh$ activation. Finally we have added two dense layers, each of dimension 128, one for the $relu$ activation and the $sigmoid$ activation. The final layer contains 15 neurons to output the probabilities of each class.

3) GRU: Gated Recurrent Unit (GRU) is similar to LSTM with fewer parameters and just two gates: update and reset. We have used a four layer model for training with GRU. The first layer is the embedding layer of dimension 300 on our vocabulary. We have used bi-directional GRU with dimension 128. Finally we added two dense layers, first of dimension 10, one for the $relu$ activation and the $sigmoid$ activation.

4) CNN: Convolution Neural Networks are a recent type of deep learning models that is currently famous for image classifications however here we have attempted to use it for text. We have used 1d convolution with 128 filters and kernal size of 5 along with 1D average pooling. The input to convolution was provided by the embedding layer of dimension 16. Finally we have applied two dense layers of dimension 6 and 15 respectively or $relu$ and $sigmoid$ activations.

All aforementioned models were optimized using Adam optimization and learning rate of 0.001. It was trained for 30 epochs, each on a batch size of 64. The implementation was done using Tensorflow and its high level API, Keras. We have evaluated above models based on the categorical cross entropy loss and accuracy. Results will be discussed in Section V.

D. State of art NLP Models

The power of attentions and encoder decoder architecture of transformers was explored in class. Transformers are artificial neural network trained to solve natural language processing, input to output sequence predication and generation problems. Although they have not been used in any cited literature, they harness the power of self-attentions to achieve an unmatched performance. This paper experiments with two transformer based models.

1) roBERTa-Urdu Model: roberta-urdu-small is a published urdu transformer based model on hugging face which was trained on urdu news corpus. This was used through an interface provided by the simpletransformers library in python. Input was loaded in a pandas dataframe through the csv file. Poet labels were converted to numbers using sklearn library's label encoder and the dataframe was directly fed to the train method of ClassificationModel wrapper of the roberta-urdu-small model. The model was run over 15 epochs. Batch size was limited to 8 couplets and learning rate was set to 0.0004 which is the recommended value by the library. Due to its ability of capturing context which is an important aspect in urdu poetry, it is able to capture more poetry patterns than the other methods previously discussed. Results will be further explored in section V.

2) Bert-base-Multilingual-uncased Model: Bert multilingual model is one of the most popular models used in NLP applications and has been trained over 102 languages using wikipedia pages. This model, published on hugging face was used in our experiments through the wrapper simple transformers library. Model training and testing pipeline is identical to that of roberta's. Input was loaded in a pandas dataframe through the csv file. Poet labels were converted to numbers using sklearn library's label encoder and the dataframe was directly fed to the train method of Classification Model wrapper of the bert-base-multilingual-uncased model. The model was run over 10 epochs. Batch size was limited to 8 couplets and learning rate was set to 0.0004 which is the recommended value by the library. Results will be further explored in section V.

V. Results

This section explores the outcomes of aforementioned experiments and compares and contrasts the results with similar ones from related work. It has been subdivided into three sections pertaining to the broad categories of experimentation.

A. Machine Learning Models

The first set of experiments were conducted on four different Machine Learning models with TF-IDF vectorization for feature extraction.

1) SVM: Support Vector Machine model demonstrated the highest accuracy of 64%.

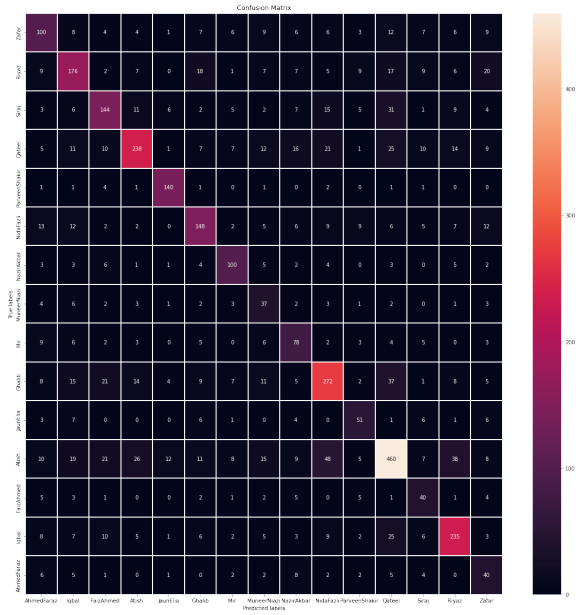

Fig. 3. Confusion Matrix of SVM Model

2) Logistic Regression: Logistic Regression classifier showed an accuracy of 60%.

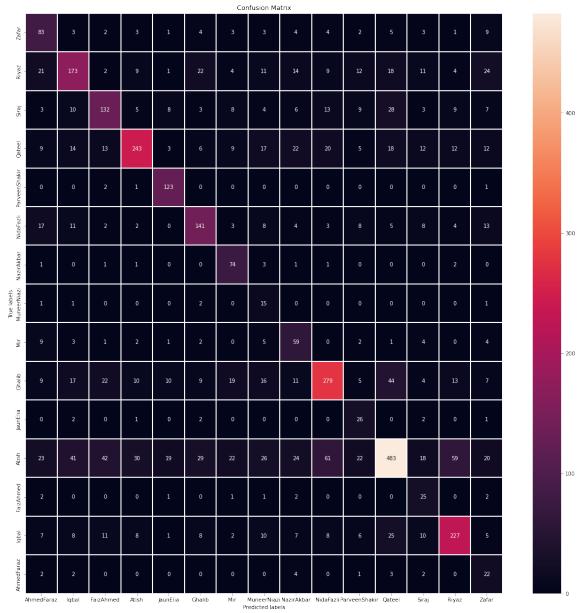

Fig. 4. Confusion Matrix of Logistic Regression Model

3) Random Forest: Random Forest Classifier demonstrated an accuracy of 22% on the testing data.

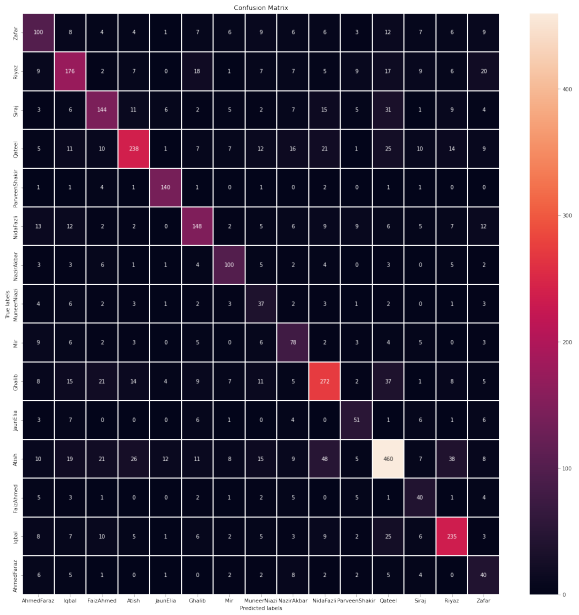

Fig. 5. Confusion Matrix of Random Forest Classifier Model

4) Naive Bayes: Multinomial Naive Bayes classifier showed an accuracy of 25%.

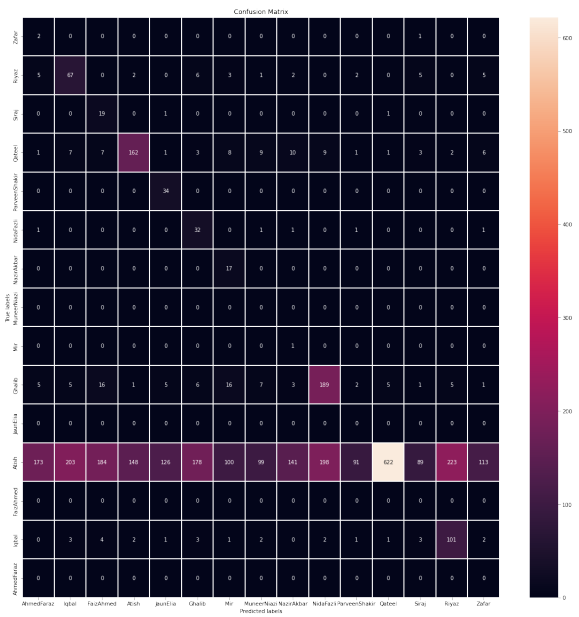

Fig. 6. Confusion Matrix of Logistic Regression Model

B. Deep Learning Models

The second set of experiments were conducted on four Deep Learning models of different nature as an attempt to find out if poetry patterns could be better captured by DL models instead of ML ones.

1) Flatten Model: Flatten model resulted in an accuracy of 55.45%. Receiver Operating Characteristic (ROC) curve shows class wise performance and loss curve indicate the progress of the model over 30 epochs.

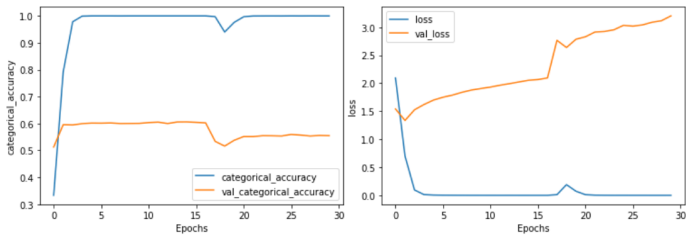

Fig. 7. Accuracy and Loss curves of Flatten Model

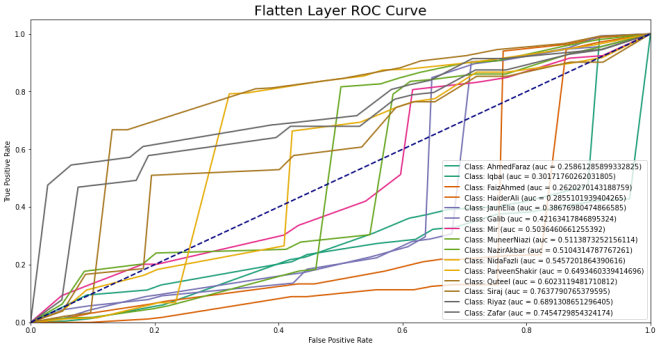

Fig. 8. ROC of Flatten Layer

2) LSTM: Long Term Short Term (LSTM) model resulted in an accuracy of 59.96%. Training was time intensive and accuracy resembled that of SVM and Logistic Regression classifier. Figures below indicate classwise performance and loss statistics over 30 epochs.

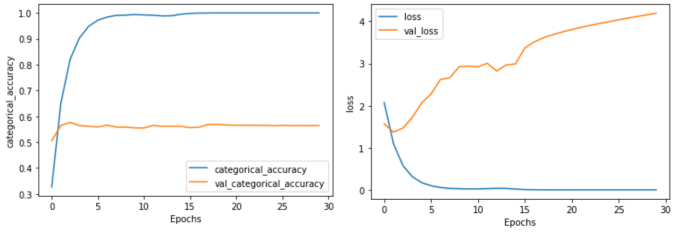

Fig. 9. Accuracy and Loss curves of LSTM

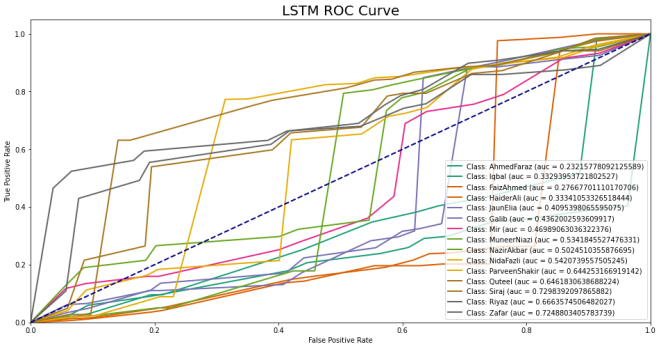

Fig. 10. ROC of LSTM

3) GRU: Gated Recurrent Unit (GRU) model resulted in an accuracy of 55.27%. This model was relatively faster to train. Figures below indicate classwise performance and loss statistics over 30 epochs.

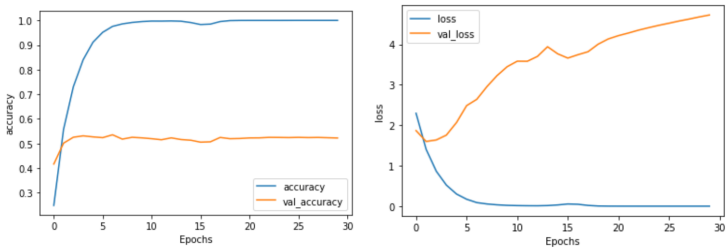

Fig. 11. Accuracy and loss curves of GRU

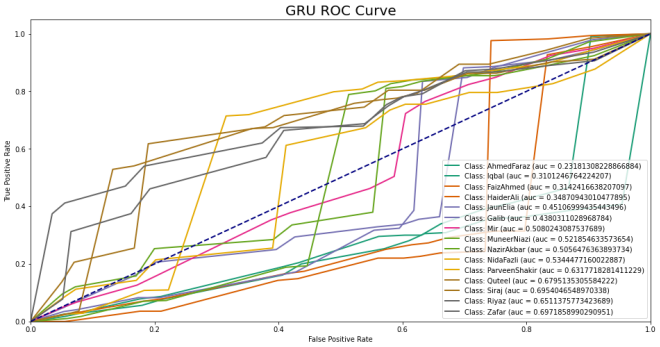

Fig. 12. ROC of GRU

4) CNN: Convolutional Neural Network (CNN) model resulted in a poor accuracy of 37.5%. Figures below indicate classwise performance and loss statistics over 30 epochs.

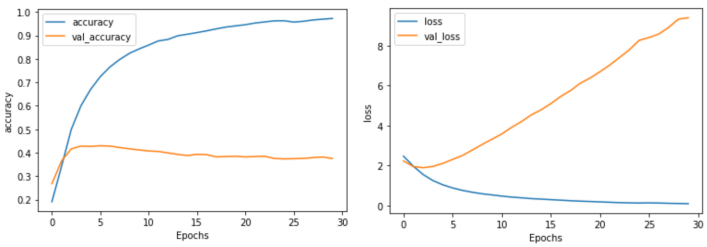

Fig. 13. Accuracy and Loss curves of Convolution

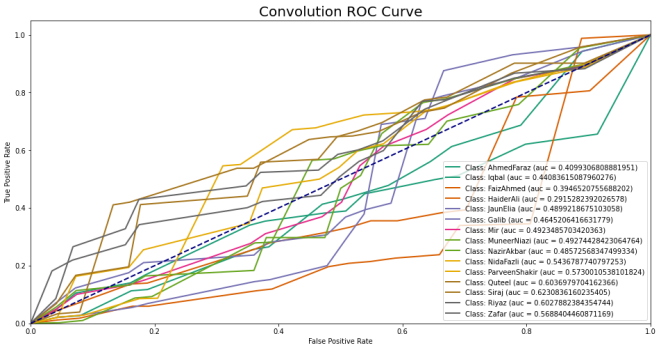

Fig. 14. ROC of Convolution

C. Transformer based Models:

The last set of experiments were conducted on two different state of art transformer based models. These deeply bidirectional pretrained models were used as an attempt to capture writing styles and vocabulary patterns more extensively.

1) roBERTa-Urdu Model: Bert-base-multilingual-uncased model demonstrated the highest accuracy out of all of 79.62%. The Matthew correlation constant indicate 77.09%. The evaluation loss at the end of 15 epochs was 2.03.

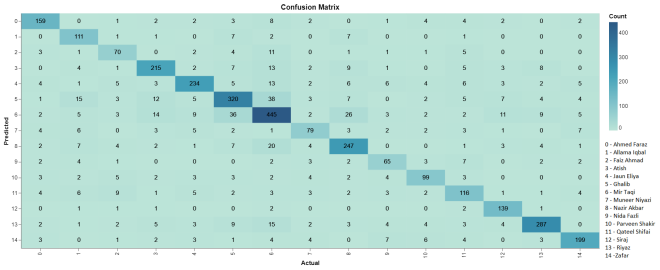

Fig. 15. Confusion Matrix for roBERTa

The following figure shows the gradual decrease in training loss over 10 epochs and 2000 steps.

Fig. 16. Loss vs Steps

2) Bert-base-Multilingual Model: Bert-base-multilingual-uncased model demonstrated the highest accuracy out of all of 80.38%. The Matthew correlation constant indicate 78.52%. The evaluation loss at the end of 15 epochs was 1.556.

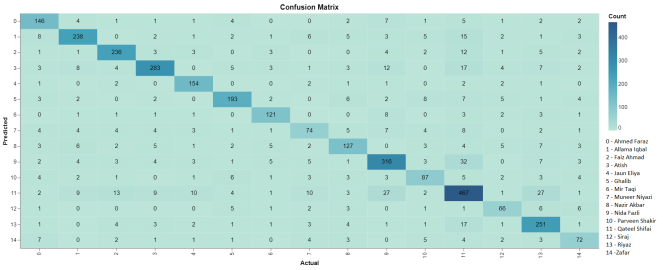

Fig. 17. Confusion Matrix for BERT

The following figure shows the gradual decrease in training loss over 10 epochs and 2000 steps.

Fig. 18. Loss vs Steps

D. Summary of Experiments

TABLE III
RESULTS SUMMARY

Model	Accuracy(%)
Decision Tree Classifier	25
Random Forest Classifier	35
Multinomial Naive Bayes	35
Conv1D	37.5
GRU	55.27
Flatten MLP Model	55.45
LSTM	59.96
Logistic Regression	60
SVM	64
Roberta	79.52
Bert	80.32

VI. DISCUSSION

The first set of experiments were performed on traditional machine learning models. As shown by Figures 3, *SVM* performed the best with 64 percent accuracy whereas *Random Forest Classifier*'s performance was worst with 25 percent accuracy. However, the results given by these classifiers are not according to the state-of-the-art results and improvement is needed. We also observed that using pre-trained embedding for the purpose of finetuning and improving the accuracy provided no additional advantage and the results remained the same.

Even though ML classifiers did not perform well, their accuracy was still better than traditional deep learning models. Amongst the DL models, LSTMs performed the best(see

Figure 9), mainly because they are good at capturing long term dependencies. The *Conv1D* performed the worst (see Figure 13) which shows that those are not suitable for multi-class text classification. We did not experiment with other convolution models, hence there is still a possibility that they perform better. As far as, our results are concerned, convolutions were worst. However, in comparisons to the work shown in Section III, our DL models performed much better for the purpose of poet identification in Urdu poetry. As shown in Table I, [11] achieved 57 percent accuracy whereas *Conv1D* gave almost 60 percent. This shows there was a significant improvement in accuracy in our work.

The performance of transformer based models was relatively similar with *Bert* and *Roberta* having 80.4 percent and 79.52 percent accuracy respectively. The former achieved the state-of-the-art accuracy and outperformed all the models we had experimented with. This shows that transformers are the best model that must be used for the purpose of poet attribution. Transformer based models are an entirely novel approach for poet attribution application and has not been experimented by anyone in the cited literature.

Although, similar to the state-of-the-art results, the highest accuracy was still less than the results achieved by previous works (see table I). We noticed that this maybe due to the imbalance of data in between classes. Every class should have an equal size of data in order for the the model to generalize well. Second clear reason is that our model caters to 15 poets as compared to 3-5 as in literature.

Additionally, results from our experimentation show that the collected data and it's overall size was enough to achieve good accuracy. Our work also showed that efficient identification of poets in Urdu poetry can be done using couplets only. Previous works showed that than four to six *misras* were needed to achieve this accuracy but our work disproves that.

All in all, transformers proved to be the best choice with considerably well overall accuracy. As shown by Figure 18 and 17, *Bert* offers good results of accuracy per class unlike the traditional ML and DL models due to their attention and encoder based framework.

VII. Future Work

Many different modifications, tests and experiments have been left for future work due to time constraints as model training is time-intensive.

1) *Incorporation of more poets* - Current dataset consists of 15 notable poets of Urdu Ghazals. Decent performance of experimented models gives us confidence to expand dataset by scraping couplets of more poets such as Nasir Kazmi, Jigar Moradabadi and Shakeel Badayuni, etc.
2) *Imposing Class Balance* - Training data for models is class imbalanced. The range of couplet count is 2500. Abundance of data for a few classes and lack of data in others is one of the factors affecting accuracy of the models. Future work will explore undersampling and oversampling. Synthetic sampling is not a feasible option due to nature of data.
3) *Data-driven investigation of poorly attributed Poets* Some poet classes are showing unusually poor results as shown in the ROC curves. In future, we will explore the causes behind this observation. The causes will be dissected based on the writing styles of the poorly attributed poets, variations in their styles and vocabulary. Domain experts will be consulted for this analysis.
4) *Training over larger Epochs* - In this paper, time constraints limited training to at most 20 epochs. Models like Bert and Roberta are very time consuming to train but are expected to show better results over larger epochs. Extensive long duration training will be carried out later.
5) *Web-based sandbox for testing Urdu ghazals* Another objective of this project is to deploy a website based sandbox for users to experiment and try model results by predicting the closest poet match to the provided couplet provided on the web-interface. This tool will be implemented and made publicly accessible as a part of future work.

VIII. Conclusion

Urdu Poetry attribution has been largely neglected despite being a challenging yet thoroughly interesting area of research. This paper takes an experimental approach to perform the arduous task of identifying the poet given an arbitrary couplet from a ghazal using deep learning techniques. A corpus of 17,608 couplets belonging to ghazals of 15 notable urdu poets were gathered from Rekhta. Bert multilingual uncased model performed best and achieved an MCC accuracy of 79%. Machine learning methods cited in the literature were explored and SVM model gave the highest f1 accuracy score of 62%. Clearly, deep learning techniques performed better than traditional machine learning approaches.

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] Z. Nazir, K. Shahzad, M. K. Malik, W. Anwar, I. S. Bajwa, and K. Mehmood, "Authorship attribution for a resource poor language—urdu," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 3, pp. 1–23, May 2022.

[3] R. Sarwar and S.-U. Hassan, "Urduai: Writeprints for urdu authorship identification," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 21, no. 2, pp. 1–18, Mar. 2022.

[4] 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.

[5] 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.

[6] "(PDF) authorship attribution in urdu poetry-researchgate," Jun-2020. [Online]. Available: https://www.researchgate.net/publication/344561377 Authorship Attribution in Urdu Poetry. [Accessed: 16-Oct-2022].

[7] D. A. Pandian, P. Maurya, and N. Jaiswal, "[PDF] author identification of Hindi poetry: Semantic scholar," [PDF] AUTHOR IDENTIFICATION OF HINDI POETRY — Semantic Scholar, 01-Mar-2020. [Online]. Available:https://www.semanticscholar.org/paper/AUTHOR-IDENTIFICATION-OF-HINDI-POETRY-Pandian-Maurya/7d14ac5b51edb43577e03e8cbd173f91db8d93ef. [Accessed: 16-Oct-2022].

[8] N. Tariq, I. Ijaz, M. K. Malik, Z. Malik, and F. Bukhari, “Identification of urdu ghazal poets using SVM,” Mehran University Research Journal of Engineering and Technology, vol. 38, no. 4, pp. 935–944, Oct. 2020.
[9] W. Anwar, I. S. Bajwa, and S. Ramzan, “Design and implementation of a machine learning-based authorship identification model,” Scientific Programming, 16-Jan-2019. [Online]. Available: https://www.hindawi.com/journals/sp/2019/9431073/. [Accessed: 16-Oct-2022].
[10] W. Anwar, I. S. Bajwa, M. A. Choudhary, and S. Ramzan, “An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution,” IEEE Access, vol. 7, pp. 3224–3234, Dec. 2018.
[11] S. Waijanya and N. Promrit, “The poet identification using convolutional neural networks,” SpringerLink, 01-Jul-2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-60663-717. [Accessed: 16-Oct-2022].
[12] A.-F. Ahmed, R. Mohamed, and B. Mostafa, “Machine learning for authorship attribution in Arabic poetry,” International Journal of Future Computer and Communication, vol. 6, no. 2, pp. 42–46, 2017.
[13] “Understanding TF-IDF for Machine Learning,” Capital One. [Online]. Available: https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/. [Accessed: 03-Dec-2022].
[14] D. D. J. D. Scientist and K. B. S. Engineer, “Machine learning mastery,” MachineLearningMastery.com, 25-Oct-2021. [Online]. Available: https://machinelearningmastery.com/. [Accessed: 03-Dec-2022].