Diachronic Text Classification: A Deep Learning Approach to Identifying Historical Eras in Urdu Literature

Aiza Imran
Computer Science
Habib Univrsity
Karachi, Pakistan
ai07597@st.habib.edu.pk

Syed Hamza
Computer Science
Habib Univrsity
Karachi, Pakistan
hs07141@st.habib.edu.pk

Sandesh Kumar
Computer Science
Habib Univrsity
Karachi, Pakistan
sandesh.kumar@sse.habib.edu.pk

Abdul Samad
Computer Science
Habib Univrsity
Karachi, Pakistan
abdul.samad@sse.habib.edu.pk

***Abstract*—Urdu language has a rich literary heritage, yet the domain of diachronic classification and analysis of literary texts across historical periods has received little attention and remains mostly unexplored. Urdu literature has evolved over centuries and demonstrates a rich linguistic change which presents hurdles for automated text classification . To address this gap in research, our study aims to perform the temporal classification of Urdu literature (nasr) by recognizing shifts in language across history. We use deep learning techniques to analyze a curated corpus of Urdu literature accross centuries which capture the diverse historical and cultural contexts of different eras. This study aims to improve temporal classification approaches, provide new insights for digital humanities, and contribute to the development of resources for Urdu language processing.**

***Index Terms*—Diachronic Text Classification, Urdu Prose, Historical Linguistics, Deep Learning, Computational Linguistics**

I. Introduction

The classification of literary texts according to their historical context, known as diachronic or temporal text classification, is a significant research area in digital humanities and computational linguistics. It involves categorizing documents based on the period in which they were created, enabling the analysis of language evolution, cultural trends, and stylistic changes over time. Temporal text classification is vital for understanding how language and literary styles transform, providing insights into the socio-political and cultural contexts of different eras. While this field has seen considerable progress in well-studied languages like English and Hebrew [1]–[4], research on languages with intricate morphology and deep literary traditions, such as Urdu, remains scarce.

The literary history of Urdu is extensive, spanning several centuries. Its use of language has changed over time, going through several stylistic periods. Since the eleventh century, Urdu literature has flourished throughout South Asia, especially in India and Pakistan [5]. Many languages such as Persian, Arabic, Turkish, Sanskrit, and other regional languages have all had an impact on it. Poetry, prose, and philosophical works are only a few of the many genres that are part of Its rich literary legacy. Every era has distinct literary styles, subjects, and linguistic characteristics. Urdu literature is a fascinating topic for diachronic study since its evolution is influenced by historical events and exchanges between cultures [5].

However, classifying historical Urdu texts using automated methods is not an easy task. It is challenging to use typical methods for natural language processing (NLP) due to the language's intricate structure [6], [7], which includes changes in script and writing styles over time. Additionally, there is a lack of annotated historical texts and resources for Urdu, which limits research opportunities. Urdu text classification has previously been done extensively using traditional machine learning techniques like Naïve Bayes, Support Vector Machines (SVMs), Decision Tree (J48), and K-Nearest Neighbors (KNN) [6], [7], but these approaches often involve extensive manual feature engineering and may not effectively capture temporal changes in language use.

This paper addresses the challenge of diachronic text classification for Urdu prose, a task that has not been explored before. We aim to develop new methods to study the language's evolution over time. Our approach focuses on capturing the unique features of Urdu texts from different time periods to detect changes in language, style, and themes. Through this research, we aim to provide valuable resources for researchers in digital humanities and offer fresh insights into the literary history of Urdu and expanding the study of linguistic changes. This research fills a significant gap in Urdu text analysis and linguistic studies and can also inspire future studies on other low-resource languages with rich literary traditions.

II. Research Question

In this study, we will be addressing the following research question:

"How can deep learning models effectively classify Urdu prose into different historical eras?"

This study focuses on analyzing texts over time to determine the specific time period when a piece of writing was created based on its linguistic style. The main question we want to address is:

"Given a prose excerpt from Urdu literature, classify it into one of the predefined historical eras."

We will exploring different deep learning models, training them using existing architectures and, we will develop customized neural networks to solve this classification problem. The input to our model will consist of prose excerpts, while the output will be the class label corresponding to the identified historical era. The time periods we used for this study span from 1800 to 2024 and have been divided into five distinct eras as follows:

- 1800-1850
- 1850-1900
- 1900-1950
- 1950-2000
- 2000-2024

We decided to investigate how to classify Urdu prose into historical periods using deep learning models due to a notable gap in existing research on Urdu language Diachronic text classification has been examined in other languages such as English and Hebrew [1], [3], however this study has not been done on the Urdu language. This gives us a unique opportunity to contribute to digital humanities using modern deep learning techniques to deepen our understanding of Urdu literature's rich cultural background.

This research aims to find trends in themes and linguistic features of Urdu prose that may indicate sociopolitical developments in South Asia throughout centuries. This can also help to preserve and appreciate the rich literary and cultural heritage of Urdu language. With the help of advanced deep learning models and Natural Language Processing (NLP) techniques, we will be able to capture complex patterns in Urdu prose that depict its diachronic diversity that traditional methods might miss and this will allow for a detailed analyses.

Our research also encourages interdisciplinary collaboration between the fields of computer science and linguistics which may lead to the creation of new resources that benefit both fields. The methodologies and insights we develop could have broader applications for other low-resource languages.

As content continues to expand on online platforms, reliable classification techniques are essential in today's digital environment. By developing a deep learning model for classifying Urdu prose, we can provide valuable resources to scholars and literature enthusiasts navigating this complex landscape of information.

III. Literature Review

Temporal text classification is needed for tracking and analysing how a language evolves over time. Therefore, research in temporal text classification has become increasingly relevant. This section of the paper provides a detailed review of the existing approaches in diachronic text classification. It includes multiple papers which highlight the methods used for this classification in various languages, focusing on both traditional machine learning methods and recent deep learning models. The discussion highlights the different methodologies used in various languages, including Urdu.

Urdu text classification has traditionally relied on statistical methods. A study by Ali Abbas and Maliha Ijaz [6], compared Naïve Bayes and Support Vector Machines (SVMs) for Urdu text classification. The researchers found that Urdu presents some unique challenges due to its complex morphological structure and the lack of tools for basic text processing. In their experiments, the SVM model, especially when using the Radial Basis Function (RBF) kernel, performed really well with an accuracy of 93.34%. The Naïve Bayes model, while faster to run, couldn't handle the complex patterns in the text as well and only achieved 76.79% accuracy. These results emphasize the need for advanced techniques in text classification, but the study only looked at fixed datasets, not how language changes over time.

Building on this, Rasheed et al. [7] conducted a comparative study using SVM, Decision Trees, and k-Nearest Neighbors (KNN) for the classification of Urdu news articles. The dataset comprised 16,678 documents categorized into 16 classes, including international news, sports, and politics. The results demonstrated that SVM achieved the highest accuracy at 68.73%, outperforming both Decision Trees (62.37%) and KNN (55.41%). The study utilized TF-IDF for feature extraction and Information Gain for feature selection, further emphasizing the significance of these techniques in enhancing classification performance.

In recent research on Urdu text classification, Bhaumik and Das [8] conducted a study on sentiment analysis of Urdu tweets using transformer-based models. Their work focused on detecting emotions and threats within Urdu social media content, employing models such as MBERT and MuRIL. The study demonstrated that these transformer models can effectively manage the complexities of the Urdu language, especially in tasks requiring a deeper understanding of context and emotion. MBERT achieved 61.2% accuracy in emotion classification, while MuRIL performed better in threat detection, with an accuracy of 71.6%. This work emphasizes the increasing significance of deep learning methods in processing the Urdu language and underscores the effectiveness of transformer-based models in capturing intricate patterns, particularly in low-resource languages.

The three studies on Urdu text classification [6]–[8], as summarized in Table II, provide foundational information that can help our research on diachronic text classification in Urdu. Although Ali and Rasheed's work used traditional classifiers like SVMs, they highlighted key preprocessing steps—such as tokenization, stop word removal, and feature selection—that are essential for effective classification, even with deep learning models. Additionally, Bhaumik and Das demonstrate the use of advanced models, such as transformers, in addressing challenges specific to Urdu's morphology and script. Given our focus on deep learning for identifying historical eras in Urdu literature, these studies underline the necessity of robust preprocessing methods and highlight the potential of transformer-based models to track temporal and semantic shifts in Urdu texts, particularly when identifying historical eras in Urdu literature.

TABLE I
Diachronic Text Classification Research

Year	Language	Models Used	Accuracy
2021 [1]	Hebrew	CNN, RNN (GRU), Paragraph Vectors	RNN (GRU): 85%
2019 [3]	English	Support Vector Machine (SVM)	46.3% (6-year range), 73.3% (50-year range)
2021 [4]	English	RNN	80% (10-year range)

TABLE II
Urdu Text Classification Research

Year	Models Used	Accuracy
2020 [6]	Naïve Bayes, Support Vector Machines (SVM)	SVM: 93.34%, Naïve Bayes: 76.79%
2021 [7]	SVM, Decision Tree (J48), K-Nearest Neighbors (KNN)	SVM: 68.73%, Decision Tree: 62.37%, KNN: 55.41%
2022 [8]	MuRIL (BERT)	71.6%

In terms of diachronic text classification of languages, some studies have explored this area in languages other than Urdu. For example, Szymanski and Lynch [3] addressed the diachronic classification of English-language news articles using n-grams of characters, words, and syntactic features. Their model achieved up to 73.3% accuracy for classifying texts within a 50-year range, while for a 6-year range, their model achieved 46.3% accuracy. They used Support Vector Machines and combined various feature types such as part-of-speech tags and syntactic structures, which allowed it to capture stylistic changes over time. Although the model was effective for English, it relied heavily on external syntactic data (Google Syntactic N-Grams), which may not be available for Urdu, presenting a challenge for extending this approach to other languages.

Recent advances in deep learning have shown significant potential for improving temporal text classification. In the paper by Yu He et al. [2], the authors proposed two neural frameworks—Temporal Smoothness (TS) and Diachronic Propagation (DP)—to handle time-evolving data. These methods addressed the limitations of traditional classifiers by adapting models to gradual shifts in data distributions. The DP framework, which incorporated historical features into the current model, demonstrated an improvement of up to 6% in accuracy for the NYTimes dataset. The DP approach's success lies in its ability to build on previous time steps, allowing for a smoother adaptation to linguistic evolution. This approach, however, has yet to be applied to Urdu literature, where capturing historical linguistic changes may reveal deeper cultural and stylistic transitions.

Similarly, Liebeskind and Liebeskind [1] applied deep learning to classify historical Hebrew texts by period. Their use of Recurrent Neural Networks (RNNs), particularly Gated Recurrent Units (GRUs) helped in achieving an accuracy of 85%. The success of RNNs in diachronic classification is mainly due to their ability to process sequential data. This makes them particularly suited for tracking language shifts over time. The study emphasized the model's capability to handle non-standard grammar and orthography. These are challenges similar to those faced in Urdu language. Deep learning models like RNNs could address these complexities of Urdu's script and morphology.

In the domain of American diplomatic language, Jo and Algee-Hewitt [4] used Recurrent Neural Networks (RNNs) to predict linguistic changes across decades. Their model achieved high accuracy by identifying periods of significant linguistic shifts, such as during the two World Wars, using decade embeddings to track stylistic and syntactic changes. This approach revealed that even non-content words (stopwords) contributed to linguistic changes, a finding that could be significant for analyzing shifts in Urdu literature, where evolving grammar and syntax play a critical role in distinguishing historical eras.

Another potential field in diachronic categorization is the use of diachronic word embeddings. Huang and Paul [9] presented a text classification model that uses diachronic embeddings to track changes in word usage over time. They showed that combining Bidirectional Long Short-Term Memory (Bi-LSTM) networks with these embeddings improved classifier robustness to temporal shifts, especially for datasets where language changes are more apparent. Their model proved that temporal word embeddings enable a more sophisticated representation of language change, outperforming domain adaption approaches. Applying such techniques to Urdu literature could offer insights into how word meanings and usage have changed across different historical periods, providing a deeper understanding of literary trends.

The reviewed studies, with their models and their corresponding accuracies have been summarized in Table I. These studies highlight that while deep learning models have shown success in English and Hebrew, these techniques need adaptation to handle the specificities of Urdu, including its script complexities and historical changes influenced by cultural and political shifts.

IV. Methodology

A. Dataset:

In this section, we provide essential information regarding the dataset utilized for training and testing our deep learning model for period classification in Urdu prose. Our dataset encompasses notable works of Urdu literature, carefully selected to represent various historical eras.

The choice to create a custom dataset was made because existing datasets [9], [10] for Urdu prose classification are either limited in scope or lack the diachronic diversity required for this study. Furthermore, the selected works exhibit clear linguistic and stylistic variations across time periods, making them ideal for investigating diachronic language changes.

For this study, our dataset is structured around five major time periods, with three significant literary works chosen from each era. The PDFs for these works were sourced from Rekhta [11] and Internet Archive [12].

To prepare the dataset for our deep learning models, we converted the books from their original PDF formats to a page-wise structure. Each page was treated as an independent data point, labeled with its corresponding historical era. This step ensured a granular analysis and allowed the model to focus on linguistic and stylistic features specific to each page, enhancing its ability to identify temporal patterns.

To achieve this, the books in PDF format were first split into individual pages using Python-based scripts. The text from each page was extracted using the Google Vision OCR API, which converted scanned Urdu text into machine-readable formats. Post-extraction, we conducted extensive cleaning to ensure high-quality data. This process involved removing noise such as watermarks, headers, footers, and page numbers, which were irrelevant for text analysis. Each cleaned page was then labeled with one of the five predefined historical eras: 1800-1850, 1850-1900, 1900-1950, 1950-2000, and 2000-2024.

The dataset was randomized to avoid any biases that could arise from sequential page order. It was subsequently divided into training, validation, and testing subsets, following an 80:10:10 ratio. This ensured that the model could learn effectively while maintaining distinct validation and testing sets for unbiased performance evaluation. To simplify input handling during model training, the data was organized into folders representing each era, with subfolders for training, validation, and testing data. For instance, the folder labeled ”1800-1850” contained separate directories for the training, validation, and testing datasets, each comprising the respective pages from this time period.

The breakdown of the time periods and their corresponding literature is as follows:

- **2000-2024**
  - *Ghulam Bagh* by Mirza Athar Baig
  - *Namal* by Nimra Ahmed
  - *Thanda Gosht* (Revised Editions) by Saadat Hasan Manto
- **1950-2000**
  - *Aag Ka Darya* by Qurratulain Hyder
  - *Udaas Naslain* by Abdullah Hussain
  - *Chaklawa* by Shaukat Siddiqui
- **1900-1950**
  - *Godaan* by Munshi Premchand
  - *Angaray* by Rashid Jahan and others
  - *Fasana-e-Azad* by Ratan Nath Dhar
- **1850-1900**
  - *Umrao Jan Ada* by Mirza Hadi Ruswa
  - *Fasana-e-Azad* by Ratan Nath Sarshar
  - *Mirat-ul-Uroos* by Nazir Ahmad
- **1800-1850**
  - *Bagh-o-Bahar* by Mir Amman
  - *Dastan-e-Amir Hamza* (Various versions)
  - *Aab-e-Hayaat* by Umera Ahmed

The selection process for these texts was driven by their literary significance and their ability to effectively represent the diverse historical and cultural contexts of Urdu literature. By incorporating a range of works, we aim to develop a dataset that not only captures the evolution of prose styles but also serves as a rich resource for training our deep learning models.

B. Models Used:

Our methodological approach to diachronic text classification involved experimenting with both traditional recurrent neural networks (RNNs) and advanced transformer-based architectures to evaluate their respective strengths and limitations in handling historical Urdu prose.

The first model we implemented was an RNN, adapted from previous research conducted on English text classification by Jo et al. [4]. This architecture was chosen as a baseline for comparison. The RNN model utilized Word2Vec embeddings in its input layer to capture semantic relationships between words. These embeddings were fed into three hidden layers: a vanilla RNN layer with 128 neurons to process sequential dependencies, an LSTM layer with 128 neurons to capture long-term dependencies, and a GRU layer with 128 neurons to optimize computational efficiency. The outputs of these layers were combined and passed through a softmax layer, which predicted probabilities for each of the five historical eras. The loss function used was multi-class cross-entropy, optimized via the Adam optimizer. Despite achieving a modest accuracy of 79% on the test dataset, the RNN faced challenges in capturing the nuanced temporal and stylistic changes inherent in historical Urdu prose.

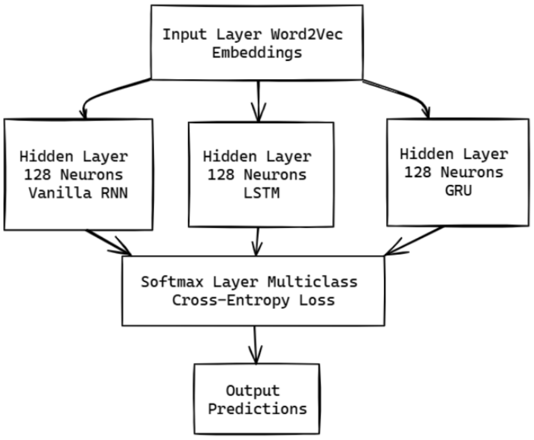
Input Layer Word2Vec Embeddings
Hidden Layer 128 Neurons Vanilla RNN
Hidden Layer 128 Neurons LSTM
Hidden Layer 128 Neurons GRU
Softmax Layer Multiclass Cross-Entropy Loss
Output Predictions

Fig. 1. RNN Implementation [4]

To address the limitations of the RNN, we transitioned to transformer-based architectures, which are renowned for their capability to handle complex textual relationships through self-attention mechanisms. Specifically, we employed XLM-RoBERTa and mBERT, both multilingual models pre-trained on diverse datasets that include Urdu. These transformers excel in processing intricate linguistic structures and capturing temporal shifts in language. XLM-RoBERTa, being trained on over 100 languages, offered superior contextual understanding, while mBERT, though smaller in scale, provided computational efficiency without compromising robustness.

The transformer models tokenized input text into subwords, enabling them to handle rare and historical words effectively. These tokens were processed through multiple layers of self-attention and feed-forward networks to capture semantic and syntactic relationships. The final classification layer mapped the processed features to probabilities for the five historical eras. Both models were trained with a maximum sequence length of 512, a batch size of 16, and a learning rate of 0.0002. Training was conducted for three epochs, utilizing the default loss functions of each respective model. The transformers significantly outperformed the RNN, highlighting their capability to capture the complex interplay of historical, linguistic, and stylistic patterns in the dataset.

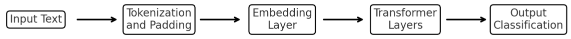

Fig. 2. Transformer Model Workflow Diagram

C. Evaluation Metrics:

To evaluate the performance of our models, we used a combination of quantitative metrics. Accuracy was employed as the primary measure of overall model correctness, providing an aggregate view of predictive performance. To gain deeper insights into class-wise performance, we analyzed confusion matrices, which detailed the model's ability to correctly classify each historical era. Additionally, we monitored loss over epochs to assess model convergence and training stability. These metrics collectively provided a comprehensive evaluation framework for comparing the RNN and transformer-based models.

V. Results and Discussion

A. Model Performance:

The performance of the models was assessed using the evaluation metrics. The confusion matrix for XLM-RoBERTa, shown in Figure 3, demonstrates its superior performance with an accuracy of 98%, the highest among the models. Most samples are correctly classified, as evidenced by the high diagonal values. For instance, 247 out of 254 samples for the era 2000-2024 and all 86 samples for 1800-1850 are correctly predicted. Misclassifications are minimal, such as three samples from 1850-1900 misclassified as 2000-2024. XLM-RoBERTa's high precision and recall highlight its ability to capture complex linguistic and stylistic nuances in the dataset.

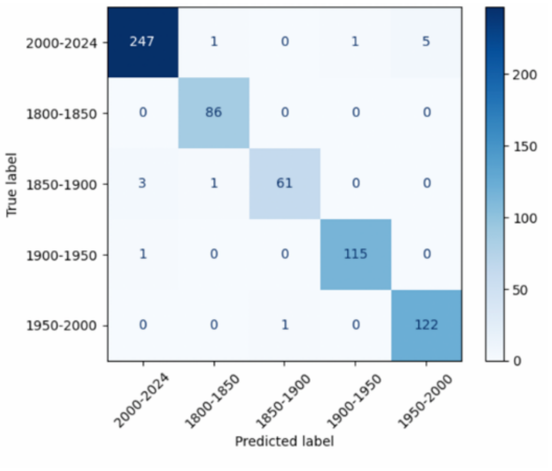

Fig. 3. Confusion Matrix for XLM RoBERTa

The confusion matrix for mBERT, shown in Figure 4, reveals an accuracy of 91%, slightly lower than XLM-RoBERTa. mBERT performs well in recent eras like 2000-2024 (249 out of 254 correctly classified) and 1900-1950 (104 out of 116 correctly classified). However, it struggles with older eras such as 1850-1900, where six samples are misclassified as 2000-2024, and 1950-2000, with ten similar misclassifications. These results suggest that mBERT effectively captures modern linguistic patterns but encounters challenges with the nuanced and overlapping characteristics of historical texts.

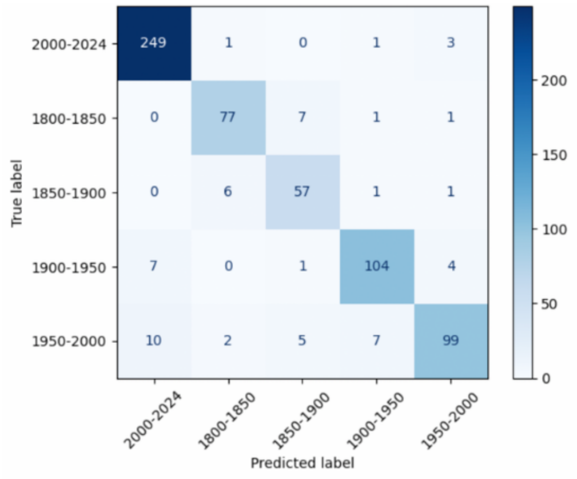

Fig. 4. Confusion Matrix for mBERT

The RNN model, depicted in Figure 5, achieves an accuracy of 79%, the lowest among the models evaluated. It struggles with contextual understanding, particularly in older eras, leading to higher misclassification rates. For instance, seven samples from the 1850-1900 era are misclassified as 2000-2024, and fourteen samples from 1950-2000 are similarly misclassified. However, the RNN performs relatively well in recent eras like 2000-2024, correctly classifying 248 out of 254 samples. These results highlight the limitations of RNNs in handling long-range dependencies and nuanced historical contexts, which are better addressed by transformer-based models.

The loss curve for the RNN, shown in Figure 6, exhibits a steady decline over the five epochs, demonstrating gradual learning. However, the higher initial loss and slower convergence compared to transformer-based models like XLM-RoBERTa and mBERT indicate challenges in optimizing the model effectively for this task.

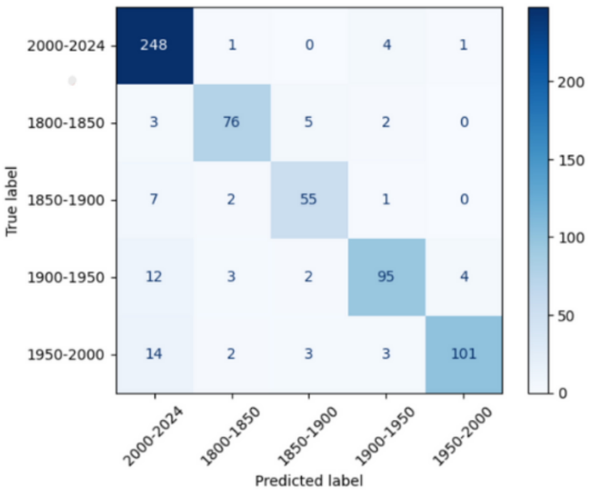

Fig. 5. Confusion Matrix for RNN

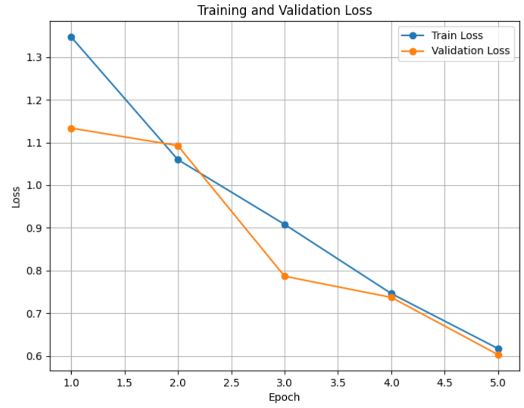

Fig. 6. RNN Training Results

B. Discussion:

The superior performance of XLM-RoBERTa can be attributed to its extensive pretraining on a large multilingual corpus, which endowed it with a robust contextual understanding and the ability to handle complex linguistic relationships. Its self-attention mechanism was instrumental in capturing the intricate interplay of syntax and semantics, particularly in the context of historical Urdu prose, which often features diverse stylistic and thematic elements.

The relatively lower performance of mBERT highlighted the trade-offs associated with smaller transformer models. While computationally efficient, mBERT lacked the extensive pretraining of XLM-RoBERTa, which limited its ability to fully capture the depth and complexity of temporal linguistic shifts. Nonetheless, its strong performance demonstrated its viability as a lightweight alternative for similar tasks.

The RNN, while providing a useful baseline, was inherently limited by its sequential processing nature and lack of pretraining. Its inability to incorporate global contextual information and its reliance on manually derived embeddings further constrained its effectiveness. These limitations highlighted the need for transformer-based models, which leverage advanced mechanisms like self-attention and contextual embeddings to address the challenges of diachronic text classification.

Overall, our study represents a significant advancement in the field of Urdu text classification. By leveraging transformer-based architectures, we demonstrated the feasibility of applying modern deep learning techniques to a low-resource language, setting a new benchmark for future research in this domain.

VI. Conclusion and Future Work

This study tackled the complex task of diachronic text classification in Urdu prose, leveraging deep learning models to capture linguistic and stylistic shifts across five historical eras. The results highlighted the effectiveness of transformer-based models, with XLM-RoBERTa achieving an impressive accuracy of 98%, significantly outperforming mBERT and the RNN baseline. Our findings underscore the potential of multilingual transformers to address the challenges posed by low-resource languages like Urdu, particularly in the context of historical text classification.

One of the key achievements of this research was the creation of a novel page-wise dataset encompassing 15 prominent works of Urdu literature from 1800 to 2024. This dataset serves as a valuable resource for future studies in computational linguistics and digital humanities. Additionally, this study marks the first application of transformer-based models to diachronic classification in Urdu, establishing a strong foundation for further exploration in this area.

However, the study was not without its limitations. The dataset, while diverse, was limited to 15 books, which may not fully capture the breadth of Urdu prose across history. Furthermore, the computational requirements of transformer-based models pose accessibility challenges for researchers with limited resources. Addressing these limitations will be critical for future work.

Looking ahead, there are several avenues for expansion. Increasing the size and diversity of the dataset by incorporating more literary works will enhance the robustness of future models. Experimenting with larger transformer models, such as XLM-RoBERTa Large, may yield even greater performance gains. Additionally, integrating external features, such as historical events and author metadata, could provide richer contextual insights. Finally, this research opens the door for interdisciplinary applications, including sociolinguistic studies and cultural analyses, further bridging the gap between computational linguistics and digital humanities.

By building on the methodologies and findings of this study, future research can continue to advance the field of Urdu language processing, contributing to the preservation and appreciation of its rich literary heritage and paving the way for broader applications in other low-resource languages.

References

[1] C. Liebeskind and S. Liebeskind, “Deep learning for period classification of historical texts,” *Journal for Data Mining and Digital Humanities*, 2019, hAL Id: hal-02324617, version 1 submitted on 22 Oct 2019, last revised 1 Jun 2020 (version 2). [Online]. Available: https://hal.science/hal-02324617v1

[2] Y. He, J. Li, Y. Song, M. He, and H. Peng, “Time-evolving text classification with deep neural networks,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, IJCAI. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 2014–2020, state Key Laboratory of Software Development Environment, Beihang University, China; Department of Computer Science and Engineering, HKUST, Hong Kong. [Online]. Available: https://www.ijcai.org/proceedings/2018/0310.pdf

[3] T. Szymanski and G. Lynch, “Ucd: Diachronic text classification with character, word, and syntactic n-grams,” in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado, USA: Association for Computational Linguistics, June 4-5 2015, pp. 879–883, corpus ID: 6315346. [Online]. Available: https://www.semanticscholar.org/paper/UCD-%3A-Diachronic-Text-Classification-with-Word%2C-and-Szymanski-Lynch/9a286ba134489e0c4173ba2522a1d71cefabb209

[4] E. S. Jo and M. Algee-Hewitt, “The long arc of history: Neural network approaches to diachronic linguistic change,” *Journal of the Japanese Association for Digital Humanities*, vol. 3, no. 1, p. 32, 2021, accessed: 2024-10-15. [Online]. Available: https://www.jstage.jst.go.jp/article/jjadh/3/1/3_1/_pdf

[5] R. Kausar, M. Sarwar, and M. Shabbir, “The history of the urdu language together with its origin and geographic distribution,” *International Journal of Innovation and Research in Educational Sciences*, vol. 2, no. 1, 2015.

[6] A. Ali and M. Ijaz, “Urdu text classification,” 12 2009, p. 21.

[7] I. Rasheed, V. Gupta, H. Banka, and C. Kumar, “Urdu text classification: A comparative study using machine learning techniques,” in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, 2018, pp. 274–278.

[8] A. B. Bhaumik and M. Das, “Emotions & threat detection in urdu using transformer-based models,” in *FIRE’22: Forum for Information Retrieval Evaluation, December 9-13, 2022, India*. CEUR Workshop Proceedings, 2022, pp. 1–8. [Online]. Available: http://ceur-ws.org/

[9] X. Huang and M. J. Paul, “Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 28 - August 2 2019, pp. 4113–4123. [Online]. Available: https://aclanthology.org/P19-1403.pdf

[10] W. Khan, A. Daud, J. Nasir, and T. Amjad, “Named entity dataset for urdu named entity recognition task,” 01 2016.

[11] Rekhta Foundation, “Rekhta,” *Online*, available: https://www.rekhta.org. [Accessed: Oct. 15, 2024].

[12] Internet Archive, “Internet Archive,” *Online*, available: https://archive.org. [Accessed: Oct. 15, 2024].