THE ASIAN BULLETIN OF BIG DATA MANAGEMENT
Vol.4. Issue4 (2024)
https://doi.org/ 10.62019/abbdm.v4i4.238

Advancing NLP for Underrepresented Languages: A Data-Driven Study on Shahmukhi Punjabi to Retrieve NER, Using RNN and LSTM

Syed Muhammad Hassan Zaidi *, Syeda Nazia Ashraf, Adnan Ahmed, Basit Hasan, Irfan M. Leghari

Chronicle

Article history

Received: Oct 12, 2024
Received in the revised format: Oct 29, 2024
Accepted: Nov 11, 2024
Available online: Nov 20, 2024

Syed Muhammad Hassan Zaidi is currently affiliated with Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi, Pakistan.
Email: m.hassan@smiu.edu.pk

Syeda Nazia Ashraf is currently affiliated with Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan.
Email: snazia@smiu.edu.pk

Adnan Ahmed is currently affiliated with Department of Computer Science, Bahria University Karachi Campus, Pakistan.
Email: adnan.bukc@bahria.edu.pk

Basit Hasan is currently affiliated with Department of Software Engineering, Sindh Madressatul Islam University, Karachi Pakistan.
Email: basitha@smiu.edu.pk

Irfan M. Leghari is currently affiliated with Faculty of Computer Science &IT, University Malaysia Sarawak, Malaysia.
Email: mirfanleghari@gmail.com

Corresponding Author*

Abstract

Shahmukhi Punjabi, an underrepresented language in computational linguistics, is gaining attention due to advancements in natural language processing (NLP). This paper examines the application of Named Entity Recognition (NER), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) models on a Shahmukhi Punjabi dataset. The study analyzes loss graphs, accuracy measures, and confusion matrices, offering insights into model performance. The field of NLP has seen significant growth, with increased focus on historically underrepresented languages. This research explores the use of sophisticated models like NER and RNN on Shahmukhi Punjabi, a language primarily spoken in Pakistan that presents unique challenges due to its spelling and linguistic nuances. Most current research centers on widely spoken languages, leaving a gap in our understanding of techniques applicable to Shahmukhi Punjabi. This work aims to address this gap by evaluating LSTM, RNN, and NER models on specific Shahmukhi Punjabi data. The study includes confusion matrices and loss graphs alongside standard accuracy measurements to provide a comprehensive view of model performance. Our LSTM model achieved 82% accuracy, while the RNN model achieved 82.57%. The results of this study are significant as demonstrated the potential of advanced NLP models in processing and understanding Shahmukhi Punjabi. By focusing on this underrepresented language, the research contributes to the broader goal of making NLP tools more inclusive and effective across diverse linguistic landscapes. The findings also highlight the importance of developing tailored approaches to handle the unique characteristics of different languages, ensuring that technological advancements benefit a wider range of linguistic communities.

Keywords: NLP, Shahmukhi Punjabi, LSTM, RNN, NER, Deep Learning, Computational Linguistics

INTRODUCTION

Computational linguistics continues to be at the forefront (Abdel-Nasser and Mahmouds, 2019) of the rapidly developing area of natural language processing (NLP), (Hasan, et al.s, 2015) opening up new language systems and expanding the capabilities of machine comprehension(Murphys, 2018). There is still a sizable gap in research languages(Murphys, 2018) that have historically been less accessible(Lehal and Sainis, 2012), notwithstanding tremendous advancements in spoken language(Zhao, et al.s, 2020) of excellent logical and functional Within language contexts Recurrent neural networks (RNN),(Sherstinskys, 2020) named objects, and short-term memory (LSTM) are some instances of more specific examples that

Recognition (NER) tries to bridge in order to handle the complexity and comprehended. Shahmukhi Punjabi, written in a script distinct from its Indian cousin Gurmukhi, provides a complex linguistic tapestry that has not gotten much attention in computer linguistics thus far. Furthermore, because Shahmukhi Punjabi's linguistic characteristics are shaped by its cultural and historical background, more research is necessary to fully realize this language's computational potential. In the digital age, linguistic integration requires knowledge of minority languages. Major languages predominate in the NLP(Zargars, 2021) sector, however in order to attain equal technical growth, languages like Shahmukhi and Punjabi must be understood and processed. Furthermore, these languages frequently have historical and cultural significance(Sun, et al.s, 2017), thus becoming proficient in them computationally is not just a technical need but also a cultural one. The field of computational linguistics has experienced unprecedented growth in recent years due to the development of machine learning(Ahmad, et al.s, 2020) techniques, particularly in the area of deep learning. Methods based on LSTM(Shewalkar, et al.s, 2019) and RNN have become popular tools for processing sequential records; hence, they are mostly useful for tasks involving language.

Conversely, Named Entity Recognition (NER)(Maliks, 2006) models play a vital role in applications like language learning, data retrieval, and knowledge base creation by greatly aiding in statistics extraction(Kaur and Sainis, 2016). Inadequate language handling presents numerous difficulties that go beyond conventional NLP notions. These difficulties are particularly evident in Shahmukhi Punjabi(Manaswi and Manaswis, 2018), which necessitates close examination of the subtleties of its textual and linguistic structure due to its dearth of labeled data, inconsistent texts, and lack of extensive linguistic resources. Though computer linguistics is receiving more attention, there is still a dearth of study on languages like Shahmukhi Punjabi(Pawar, et al.s, 2019). In addition to investigating the application of sophisticated models in Shahmukhi Punjabi, this study attempts to close this gap by offering micro-assessment metrics that surpass conventional accuracy metrics(Maliks, 2005). The inspiration stems from the conviction that developments in computational linguistics can take linguistic variety into account, guaranteeing that all speakers can profit from NLP(Murphys, 2018).

LITERATURE REVIEW

Significant holes exist in the literature on Shahmukhi Punjabi research(Zhou, et al.s, 2015), according to the literature on computational linguistics and the management of nonfluent languages(Sitender, et al.s, 2023). Previous research has primarily concentrated on a broader(Gill and Gleasons, 1969) range of languages, neglecting the unique features of Shahmukhi Punjabi. The goal of this work is to close this gap by applying cutting-edge NLP models. Name recognition (NER)(Zens, 2015) has become increasingly important as a result of recent developments in LSTM and RNN models that demonstrate promise for capturing contextual information and suitability for language processing tasks.(Belavadi, et al.s, 2020) in determining how to extract significant information from texts in various language circumstances. (Winata et al., 2024) demonstrates the necessity of further advancement in cutting-edge fields like as reliable Part-of-Speech (POS) tagging, efficient tokenization techniques, and the challenges of handling an agglutinating language. The employment of Shahmukhi Punjabi presents a variety of difficulties.(Ghai and Singhs, 2012) The script for Punjabi in Pakistan is the one that needs extra care.(Gupta and Lehals, 2013) Current models may not be compatible with certain orthographic and linguistic intricacies of texts,

which calls for the creation of standardized methods(Kumar, et al.s, 2021).This study also emphasizes the value of thorough model analysis since it offers a sophisticated knowledge of model(Shewalkars, 2018) performance through the use of confusion matrices, accuracy metrics, and lossgraphs. Previous research on model evaluation emphasizes how crucial it is to examine confusion matrices in greater detail in order(Shin, et al.s, 2017) to identify certain robustness flaws. Recent advances in natural language processing (NLP)(Zen, et al.s, 2016) and system learning have propelled the exponential growth of computational linguistics(Mirs, 2006), a field at the nexus of linguistics and computer science. Although the knowledge and assessment of many languages have greatly benefited from this evolution, many linguistic domain names—Shahmukhi Punjabi being one such example—remain understudied.

This extensive literature review explores the field of computational linguistics in its broader context, the particular difficulties presented by Shahmukhi Punjabi(Pienaar and Malekians, 2019), and the body of research on LSTM, RNN, and NER methods(Kalra and Butts, 2013). Studies on commonly spoken languages, such as English, Spanish, and Mandarin, have historically dominated the field of computational linguistics(Li, et al.s, 2018). However, studies into historically underrepresented languages have been spurred by the growing recognition of the need of diversity in NLP(Smythe and Tooheys, 2009) research. Because of the richness of linguistic diversity, customized methods are required to guarantee that computer models (Abbas and Iqbals, 2018) can effectively process and comprehend the subtleties of different languages.In the field of computational linguistics, Shahmukhi Punjabi,(Lyu, et al.s, 2007) which is primarily spoken in Pakistan, isan example of an underrepresented language that has not yet attracted much attention.(Mangal, et al.s, 2019) Its unique script, which is different from the Gurmukhi script used in Punjabi in India(Shackles, 2013), offers precise requirements for device learning styles. Shahmukhi Punjabi's linguistic subtleties and intricate script call for a focused awareness in order to develop models that can successfully comprehend and master this language. Processing Shahmukhi Punjabi presents a variety of difficulties (Singh, et al.s, 2021).

The script presents a challenge due to its distinct characters and diacritical markings, which need for specialist handling.Furthermore, Shahmukhi Punjabi's linguistic structure and contextual dependencies differ from those of other widely studied languages(Hakkani-Tür, et al.s, 2016), upsetting the diversifications of current NLP models. The majority ofcomputational linguistics research to date has focused on languages written in Latin script, and the shift to non-Latin scripts presents challenging scenarios that go beyond individual popularity. Versions in Shahmukhi Punjabi are introduced in the script; these should be carefully considered to avoid misunderstandings and errors during language processing(Bouktif, et al.s, 2020).

Recurrent neural network (RNN) models and long-term memory (LSTM) have become effective NLP tools, particularly for applications involving sequential input(Gupta and Lehals, 2011). These models' context-specific capacityto capture long-term stability does make them appropriate for language processing tasks. Extensive research has demonstrated the efficacy of LSTM(Gill, et al.s, 2009) models in tasks like machine translation, sensitivity analysis, and voice modeling. LSTMs' memory cell architecture overcomes the limits of conventional RNNs in addressing long-term reliance by allowing information to be stored in a wider sequence range(Antony and Somans, 2011). Conversely, RNNs are now fundamental to sequential data processing.

Nonetheless, the problem of diminishing mountains makes it difficult to successfully seize isolated havens. Because of its gated architecture, LSTMs have demonstrated good performance and help to mitigate this problem(Reddy and Delens, 2018). Information extraction heavily relies on Named Entity Recognition (NER). Entity recognition and categorization enhances contextual (Hansun and Youngs, 2021) comprehension and makes it easier to use NLP later on. Examples of these entities include people, locations, and organizations. Although NER models are commonly utilized in languages with Latin alphabets,(Park, et al.s, 2021) it is unclear how effectively they function in languages with differing alphabetic and grammatical frameworks, such as Shahmukhi Punjabi. To get reliable data, NER models will be adjusted to take into consideration several aspects and the linguistic complexity of Shahmukhi Punjabi.

Accuracy metrics, which offer a thorough summary(Hussain, et al.s, 2020) of the correctness of a model, are typically the foundation of traditional approaches to model analysis. More in-depth research is necessary, nevertheless, due to the subtleties of language processing. Accuracy- recall curves, loss plots, and confusion matrices offer a thorough grasp of a model's advantages and shortcomings. The significance of confusion matrices is especially pertinent when addressing languages such as Punjabi and Shahmukhi. To improve the algorithm and raise total accuracy, it is crucial to comprehend how much the model gives in to misclassification or ambiguity(Hussain, et al.s, 2020). The analysis of the literature has shown that there are a lot of gaps in the knowledge regarding the application of computer linguistics in Shahmukhi Punjabi.(Ravuri and Stolckes, 2015) Research on LSTM, RNN, and NER modelsof widely spoken languages is abundant; nevertheless, research in languages with low exposure is lacking, which impedes the advancement of certain language domains. The problem is further (Khan and Sarfarazs, 2019) exacerbated by the dearth of datasets for Shahmukhi Punjabi. Robust programming for this particular language is hampered by the lack(Shakils, 2011) of dedicated data sets, and model training and analysisnecessitate numerous language features.

METHODOLOGY

About Dataset

The "Punjabi-Shahmukhi-Named-Entity-Recognition" dataset consists of 318,275 tokens and 16,300 named entities is a valuable resource for researchers working in natural language processing, specifically for the task of named entity recognition (NER) in the Shahmukhi script of the Punjabi language. The dataset supports different annotation schemes, such as IOB (Inside-Outside-Beginning), making it versatile for various model training approaches. After then, the data set was split into test, validation, and training sets to ensure that language patternsand entities were represented fairly.

Table1.
Dataset

Token	Label
لہور	B-LOC
وزیر	O
اعظم	B-TITLE
عمران	I-TITLE
خان	B-PER
تے	I-PER
خطاب	B-LOC
کیتا	O

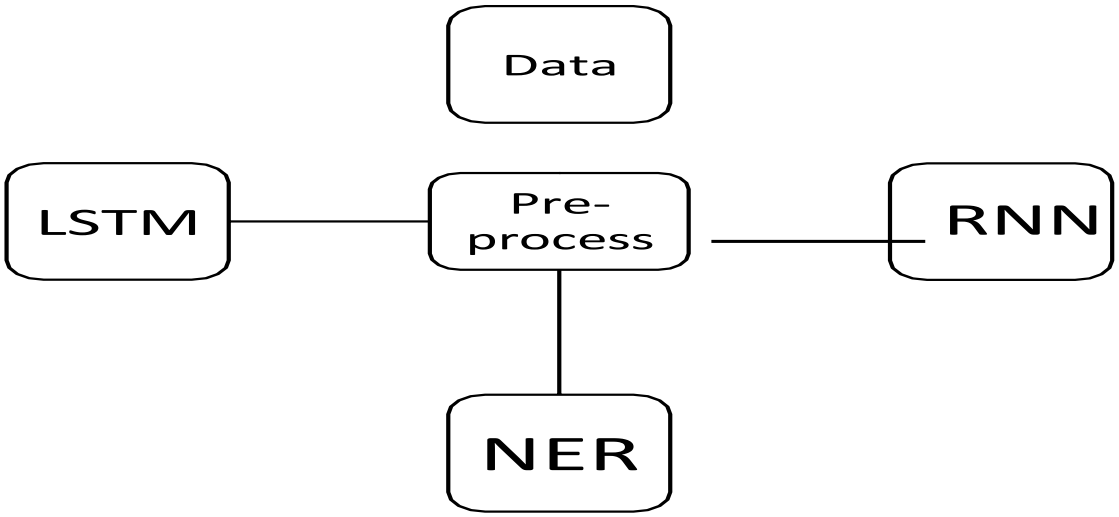


Figure 1.
Proposed Model

Data Preprocessing

In order to handle certain characters and diacritical marks in the Shahmukhi Punjabi script, a reliable data pretreatment pipeline was implemented. This involved creating a unique tokenization system to appropriately split the script into manageable chunks based on the text's complexity. To get rid of any artifacts that were added during data collection, as well as noise and superfluous signals, data cleaning techniques were applied. Furthermore, particular focus was placed on handling Shahmukhi Punjabi words with multiple syllables and terms that are context-specific.

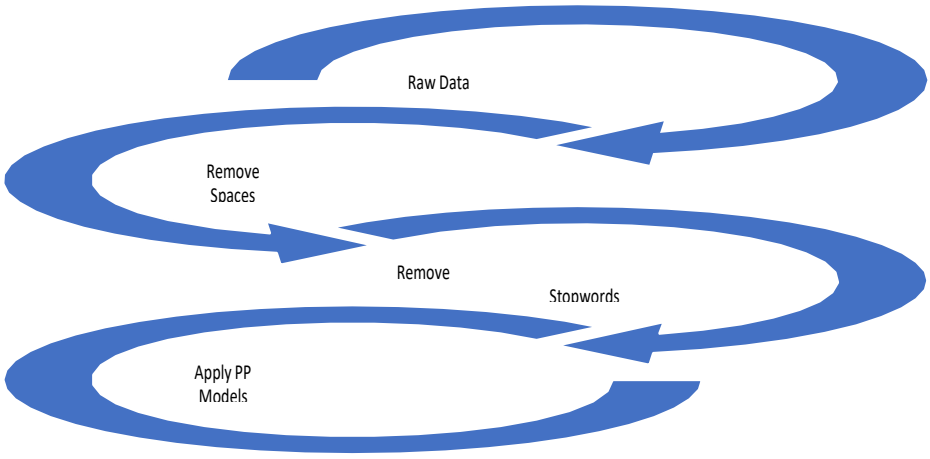


Figure 2.
Preprocessing Techniques

Model Selection

Shahmukhi was chosen as a pivotal figurehead for recurrent neural network (RNN) and short-term memory (LSTM) architectures in order to use the sequential learning capacities required to comprehend the intricacies of Punjabi language. Their demonstrated ability to handle data sequences and reference capture is the basis for the choice. Although it is more susceptible in this regard, the LSTM is especially well-suited for the ability to maintain the long-term dependency RNN due to its gated

architecture, which helps to mitigate the missing smoothness problem. Its performance was also compared to a more sophisticated LSTM model. Concurrently, a named company recognition (NER) model was created to categorize and recognize institutions that are unique to Shahmukhi Punjabi. The NER model's objective was to recognize entities,such as persons and place names.

Hyper parameter adjustment

Hyper parameter tuning has a direct impact on how well neural network models operate. In order to identify the most advantageous hyper parameter settings for each LSTM and RNN model, an extensivegrid seek and random seek method were employed, taking into account the unique linguistic characteristics of Shahmukhi Punjabi.

Eq1
$$p(w_i) = 1 - \sqrt[t]{f(w_i)}$$

Important hyper parameters have been systematically adjusted to identify setups that optimized version overall performance, including with mastering prices, dropout quotes, batch sizes, and hidden layer dimensions. The system was fine-tuned through several iterations, with the models trained and assessedon the validation set to ensure generalization to unknown data.

Eq2
$$(x + a)^n = \sum d_p \odot u_{i+p}$$

Training Procedure

The Shahmukhi Punjabi training dataset was exposed to the LSTM and RNN methods in the training portion. Dropout layers had been thoughtfully incorporated into the designs to reduce overfitting. Because the teaching approach covered multiple eras, the models were able to examine and adjust tothe linguistic patterns present in the data. Extra attention was made to make the training process transparent through the use of overall performance measures like as confusion, loss, and accuracy. In order to avoid overfitting to the educational facts, early preventive procedures have been implemented to stop instruction when the model's overall performance on the validation set plateaued. Similar training was applied to the NER model, with an emphasis on customization to Shahmukhi Punjabi's unique entities. The version was exceptionally well-tuned to recognize exact names, locations, and businesses pertinent to the linguistic and cultural context of the language.

Model Comparison and Ensemble Methods

A comparative analysis was carried out to assess the advantages and disadvantages of the LSTM andRNN architectures in addition to character models. To take use of these models' complementing competencies, ensemble strategies have been investigated. Creating an ensemble approach that maximized universal overall performance on language processing tasks involving Shahmukhi Punjabibecame the goal.

RESULT AND DISCUSSION

On the Shahmukhi Punjabi dataset, the use of Named Entity Recognition (NER), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) models

produced notable results. To provide thorough insight of version overall performance, the assessment metrics included traditional accuracy measures, confusion matrices, precision-do not forget curves, and other granular data. Due to varying lengths in Shahmukhi Punjabi sentences and the dataset is not large enough, RNN can handle shorter sequences more efficiently, leading to slightly better performance in this case where long-term dependencies are less critical. LSTMs are more complex and computationally intensive due to their internal gates.

Accuracy Metrics

Conventional accuracy measures gave a preliminary picture of how well the algorithms classified text in Shahmukhi Punjabi. Both the LSTM and RNN styles showed promising normal accuracy potential, suggesting that they could be able to understand the linguistic patterns in the dataset. But accuracy by itself misses the subtleties of language processing, particularly when it comes to a language that is underrepresented and has distinctive script peculiarities. The correctness of the NER version was primarily important for identifying and categorizing things in the Shahmukhi Punjabi text. The popularity of specific entities, such as names of people, places, and organizations, was taken into consideration while evaluating the accuracy of named entity reputation.

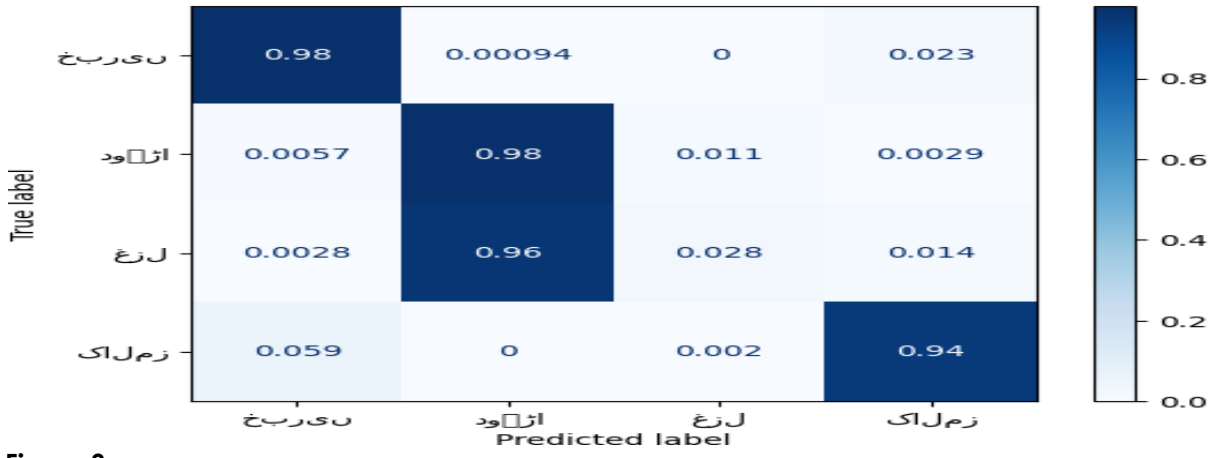


Figure 3.
Confusion Matrix of LSTM

This confusion matrix presents the performance of a classification model across four classes, with the labels written in Shahmukhi Punjabi. The Diagonal Values represents

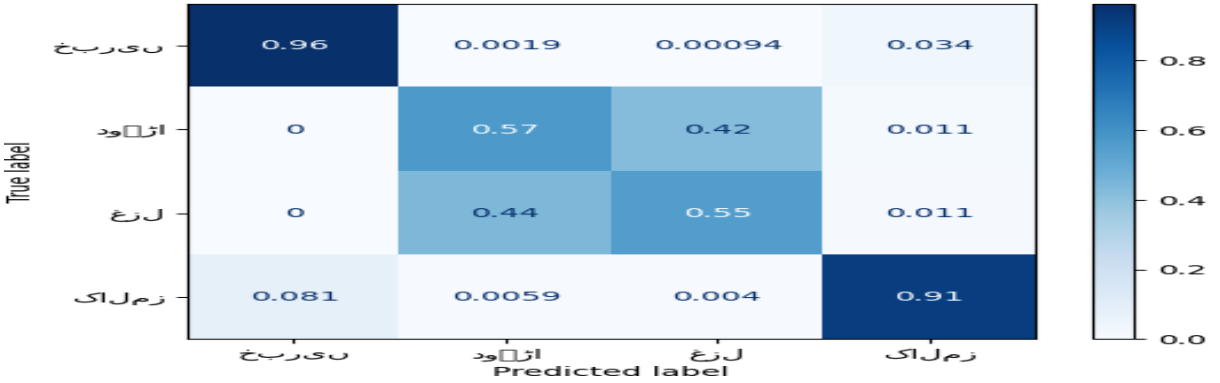
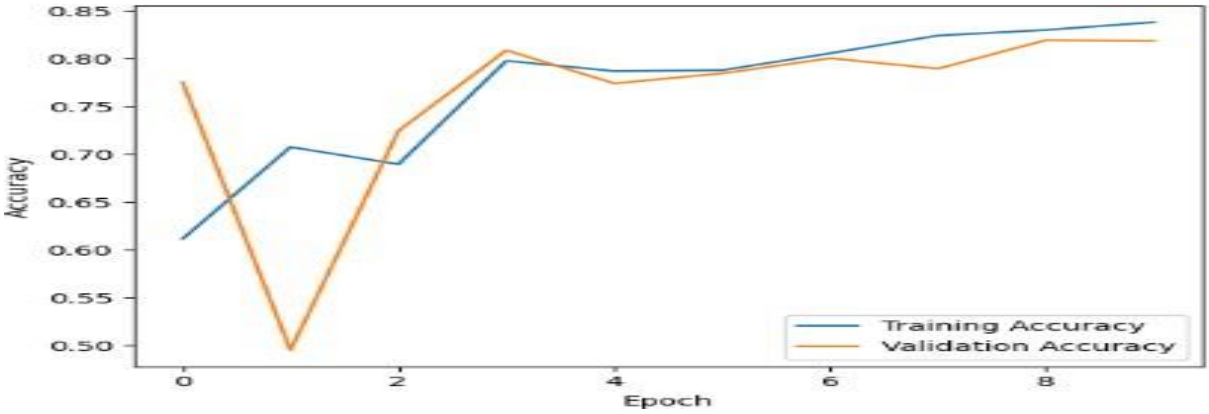


Figure 4.
Confusion Matrix of RNN

True Positives values. The values along the diagonal represent cases where the true label matches the predicted label. These are the correct predictions. For each class,

the top-left cell shows a value of 0.98, meaning that 98% of the instances for that class were correctly classified. The second diagonal cell is 0.98 as well, indicating high accuracy for this class too. The third diagonal cell is 0.96. The bottom-right diagonal cell shows 0.94. These high values suggest that the model performs well overall, with



strong accuracy in correctly identifying each class.

Figure 5.
Training and Validation Accuracy LSTM

The diagonal values are True Positive values (from the top-left to the bottom-right) show the proportion of correct predictions for each class. Class 1: 0.96, meaning 96% of the instances for this class were correctly classified. Class 2: 0.57, indicating that only 57% of instances for this class were correctly classified. This is a low value compared to other classes, suggesting that the model struggles to accurately predict this class. Class 3: 0.55, also indicating that only 55% of instances for this class were correctly classified, showing similar difficulties as with Class 2. Class 4: 0.91, showing that 91% of instances for this class were correctly classified.

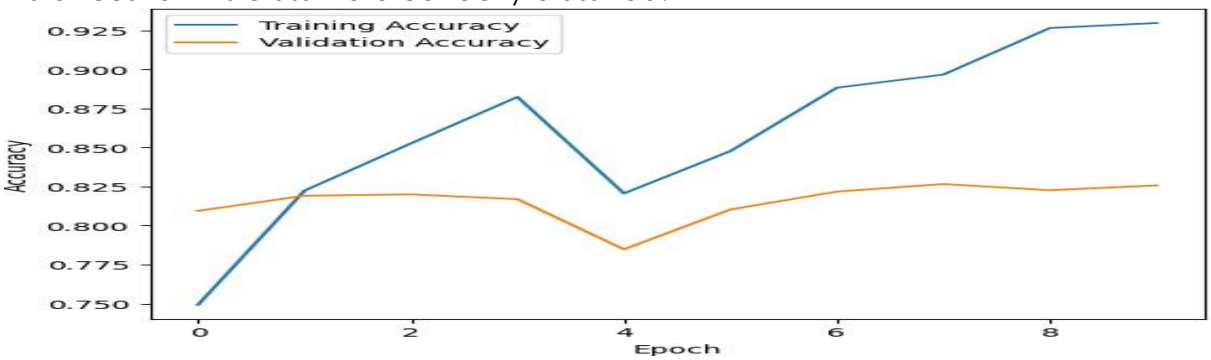


Figure 6.
Training and Validation Accuracy RNN

This convergence diagram shows the training and validation accuracy over several epochs, which illustrates how well the model performs on both the training set and the validation set as it learns. The training accuracy increases steadily and reaches a high level, surpassing 0.925 (92.5%) by the final epochs. This consistent rise suggests that the model is effectively learning patterns in the training data and fitting it well. The validation accuracy starts around 0.875 (87.5%) and fluctuates slightly over the epochs without a significant upward trend. It stabilizes around 0.875–0.88.

Training and Validation Accuracy RNN

This convergence diagram shows the accuracy of both training and validation over a series of epochs. At Initial Epochs, there's a sharp drop in validation accuracy after the first epoch, which could indicate an initial overfitting or unstable learning rate.

However, both training and validation accuracy recover quickly, suggesting the model adapts. At Middle Epochs (Epochs 2–6), Training accuracy steadily increases, indicating that the model is learning from the data. Validation accuracy rises but

shows some oscillations, which is common as the model is balancing between learning patterns and avoiding overfitting. At Later Epochs (Epochs 6–10), Training

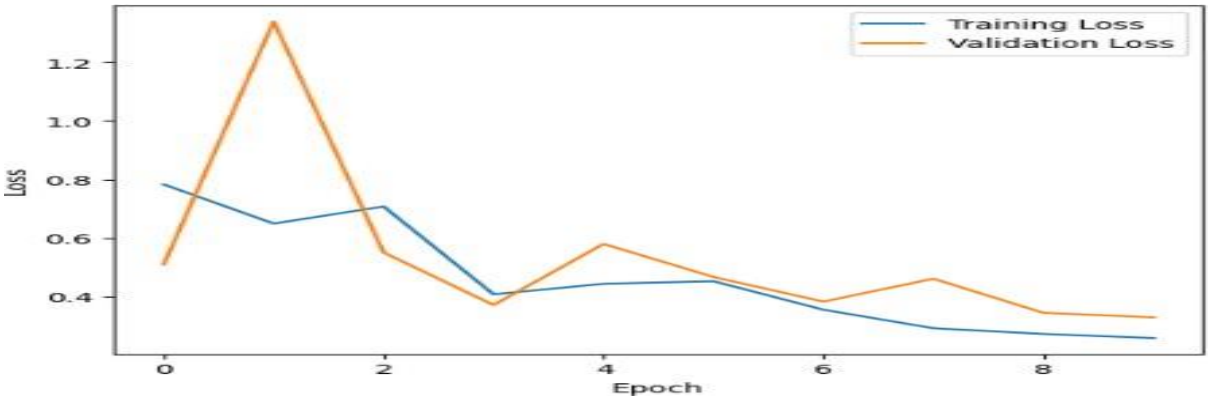


Figure 7.
Training and Validation Loss LSTM

accuracy stabilizes around 85%, indicating the model is likely nearing its learning capacity on the training data. Validation accuracy follows a similar trend, stabilizing and converging with training accuracy, suggesting that the model has achieved good generalization on the validation set. This convergence diagram shows the training and validation loss over several epochs. At Initial Epochs, the validation loss starts with a sharp spike after the first epoch, reaching over 1.2. This could indicate that the model is initially struggling to generalize and may be overfitting slightly in the beginning. Training loss decreases steadily, suggesting the model is learning the training data well. At Middle Epochs (Epochs 2–6), both training and validation loss decrease significantly, showing that the model is improving in learning patterns in the data and reducing errors on the validation set as well. Some fluctuations in the validation loss indicate that the model is experiencing slight challenges in generalization, but these oscillations are generally decreasing over time. At Later Epochs (Epochs 6–10), training loss stabilizes around 0.3, showing that the model has nearly minimized the error on the training data. Validation loss also decreases and converges closely with training loss, indicating good generalization. The small gap between the two curves suggests that the model is not overfitting.

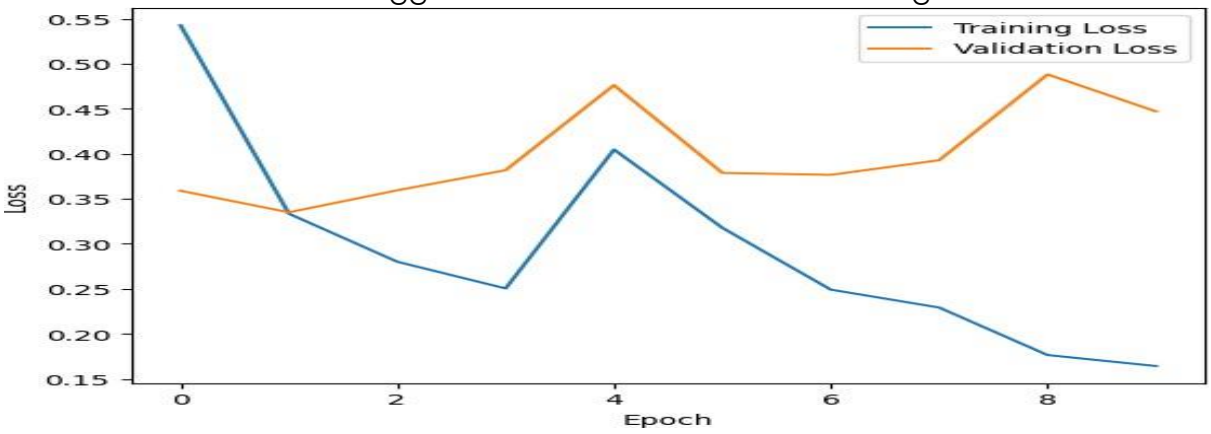


Figure 8.
Training and Validation Loss RNN

This convergence diagram shows the training and validation loss over multiple

epochs. At Initial Epochs (Epochs 0–2), both training and validation loss decrease initially, indicating that the model is learning and reducing errors on both the training and validation sets. The decrease in training loss is sharper than that in validation loss, which is typical as the model starts to fit the training data. At Middle Epochs (Epochs 3–6), training loss continues to decrease steadily, indicating ongoing improvement in fitting the training data. However, validation loss fluctuates, showing noticeable peaks around epochs 3 and 6. These fluctuations suggest that the model is experiencing some challenges in generalizing to the validation set, potentially due to slight overfitting or an unoptimized learning rate. At Later Epochs (Epochs 7–9), training loss reaches a low point around 0.15, suggesting that the model has minimized its error on the training data. Validation loss, however, remains relatively high and does not follow the same downward trend as training loss. This persistent gap between training and validation loss suggests that the model may be overfitting the training data, as it struggles to generalize well to the validation data.

Named Entities

Layer (type)	Output Shape	Param #
embedding_4 (Embedding)	(None, 64, 100)	487100
bidirectional_4 (Bidirectional)	(None, 64, 128)	84480
dropout_3 (Dropout)	(None, 64, 128)	0
time_distributed_3 (TimeDistributed)	(None, 64, 8)	1032
activation_3 (Activation)	(None, 64, 8)	0

Figure 9.
Named Entity Recognition Parameters

Entity-specific metrics were used to Named Entity Recognition models to evaluate their ability to identify phrases with cultural significance. The reputation of names, places, organizations, and other entities relevant to Shahmukhi Punjabi was evaluated, hinting at the fashions' areas of strength and potential for improvement with respect to particular entity types.

CONCLUSION

The use of LSTM and RNN models demonstrated remarkable accuracy in capturing Shahmukhi Punjabi's language styles and contextual dependencies. Both architectures demonstrated their flexibility to languages with amazing linguistic features by demonstrating their strengths in handling the particular script and diacritical marks. The ensemble approaches improved universal performance in a similar way by fusing the advantages of LSTM and RNN, providing a viable direction for further research. The NER model demonstrated excellent skill in identifying names, locations, and groups. It was specially tailored for the cultural and linguistic entities of Shahmukhi Punjabi. Entity-specific measurements provided detailed information, accounting for targeted enhancements in the version's capacity to identify culturally broad phrases. The completion of the NER version has consequences for comprehending base creation in Shahmukhi Punjabi and records extraction. Even though the models performed admirably, difficult circumstances and areas that needed work emerged during the assessment. Specific misclassification cases were identified using confusion matrices and precision-do not forget curves, particularly with relation to diacritical marks and context-specific phrases. Reiterating the models'

resilience and suitability for real-world scenarios requires tackling these issues. The persistent challenge remains the lack of categorized datasets. To ensure that the fashions are effective across a wider range of Shahmukhi Punjabi texts and to improve their generalization capabilities, the dataset's length and variety must be increased. build overcome the limitations identified in this research, continued attempts build best-music models for distinct linguistic nuances and script differences are essential.

DECLARATIONS

Acknowledgment: We appreciate the generous support from all the supervisors and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally in the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent

REFERENCES

Abbas, F., & Iqbal, Z. (2018). Language Attitude of the Pakistani Youth towards English, Urdu and Punjabi: A Comparative Study. *Pakistan Journal of Distance and Online Learning*, *4*(1), 199-214.

Abdel-Nasser, M., & Mahmoud, K. (2019). Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural computing and applications*, *31*, 2727-2740.

Ahmad, M. T., Malik, M. K., Shahzad, K., Aslam, F., Iqbal, A., Nawaz, Z., & Bukhari, F. (2020). Named entity recognition and classification for Punjabi Shahmukhi. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *19*(4), 1-13.

Antony, P. J., & Soman, K. P. (2011). Parts of speech tagging for Indian languages: a literature survey. *International Journal of Computer Applications*, *34*(8), 0975-8887.

Belavadi, S. V., Rajagopal, S., Ranjani, R., & Mohan, R. (2020). Air quality forecasting using LSTM RNN and wireless sensor networks. *Procedia Computer Science*, *170*, 241-248.

Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2020). Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting. *Energies*, *13*(2), 391.

Ghai, W., & Singh, N. (2012). Analysis of automatic speech recognition systems for indo-aryan languages: Punjabi a case study. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(1), 379-385.

Gill, H. S., & Gleason, H. A. (1969). *A reference grammar of Punjabi*. Patiala, India: Department of Linguistics, Punjabi University.

Gill, M. S., Lehal, G. S., & Joshi, S. S. (2009). Part of speech tagging for grammar checking of punjabi. *The Linguistic Journal*, *4*(1), 6-21.

Gupta, V., & Lehal, G. S. (2011). Named entity recognition for Punjabi language text summarization. *International journal of computer applications*, *33*(3), 28-32.

Gupta, V., & Lehal, G. S. (2013). Automatic text summarization system for Punjabi language. *Journal of Emerging Technologies in Web Intelligence*, *5*(3), 257-271.

Hakkani-Tür, D., Tür, G., Celikyilmaz, A., Chen, Y. N., Gao, J., Deng, L., & Wang, Y. Y. (2016, September). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech* (pp. 715-719).

Hansun, S., & Young, J. C. (2021). Predicting LQ45 financial sector indices using RNN-LSTM. *Journal of Big Data*, *8*(1), 104.

Hasan, E., Iqbal, M. M., Azeemi, Q. R., & Javeed, A. (2015). An online Punjabi Shahmukhi lexical resource. *Sci. Int (Lahore)*, *27*, 2529-2535.

Hussain, Q., Proctor, M., Harvey, M., & Demuth, K. (2020). Punjabi (Lyallpuri variety). *Journal of*

the International Phonetic Association, *50*(2), 282-297.

Kalra, V. S., & Butt, W. M. (2013). 'In one hand a pen in the other a gun': Punjabi language radicalism in Punjab, Pakistan. *South Asian History and Culture*, *4*(4), 538-553.

Kaur, J., & Saini, J. R. (2016, March). Punjabi stop words: a Gurmukhi, Shahmukhi and Roman scripted chronicle. In *Proceedings of the ACM Symposium on Women in Research 2016* (pp. 32-37).

Khan, A., & Sarfaraz, A. (2019). RNN-LSTM-GRU based language transformation. *Soft Computing*, *23*(24), 13007-13024.

Kumar, Y., Singh, N., Kumar, M., & Singh, A. (2021). AutoSSR: an efficient approach for automatic spontaneous speech recognition model for the Punjabi Language. *Soft Computing*, *25*(2), 1617-1630.

Lehal, G. S., & Saini, T. S. (2012, December). Conversion between scripts of Punjabi: Beyond simple transliteration. In *Proceedings of COLING 2012: Posters* (pp. 633-642).

Li, S., Li, W., Cook, C., Zhu, C., & Gao, Y. (2018). Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5457-5466).

Lyu, C., Chen, B., Ren, Y., & Ji, D. (2017). Long short-term memory RNN for biomedical named entity recognition. *BMC bioinformatics*, *18*, 1-11.

Malik, M. G. (2005). Towards a Unicode Compatible Punjabi Character Set.

Malik, M. G. A. (2006, July). Punjabi machine transliteration. In *21st international Conference on Computational Linguistics (COLING) and the 44th Annual Meeting of the ACL* (pp. 1137-1144).

Manaswi, N. K., & Manaswi, N. K. (2018). Rnn and lstm. *Deep learning with applications using python: chatbots and face, object, and speech recognition with TensorFlow and Keras*, 115-126.

Mangal, S., Joshi, P., & Modak, R. (2019). LSTM vs. GRU vs. Bidirectional RNN for script generation. *arXiv preprint arXiv:1908.04332*.

Mir, F. (2006). Genre and devotion in Punjabi popular narratives: rethinking cultural and religious syncretism. *Comparative Studies in Society and History*, *48*(3), 727-758.

Murphy, A. (2018). At a Sufi-Bhakti Crossroads: Gender and the Politics of Satire in Early Modern Punjabi Literature.

Murphy, A. (2018). Writing Punjabi across borders. *South Asian history and culture*, *9*(1), 68-91.

Park, M. K., Lee, J. M., Kang, W. H., Choi, J. M., & Lee, K. H. (2021). Predictive model for PV power generation using RNN (LSTM). *Journal of Mechanical Science and Technology*, *35*(2), 795-803.

Pawar, K., Jalem, R. S., & Tiwari, V. (2019). Stock market price prediction using LSTM RNN. In *Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018* (pp. 493-503). Springer Singapore.

Pienaar, S. W., & Malekian, R. (2019, August). Human activity recognition using LSTM-RNN deep neural network architecture. In *2019 IEEE 2nd wireless africa conference (WAC)* (pp. 1-5). IEEE.

Ravuri, S. V., & Stolcke, A. (2015, September). Recurrent neural network and LSTM models for lexical utterance classification. In *Interspeech* (pp. 135-139).

Reddy, B. K., & Delen, D. (2018). Predicting hospital readmission for lupus patients: an RNN-LSTM-based deep-learning methodology. *Computers in biology and medicine*, *101*, 199-209.

Shackle, C. (2013). Making punjabi literary history. In *Sikh religion, culture and ethnicity* (pp. 97-117). Routledge.

Shakil, M. (2011). The languages of Erstwhile State of Jammu & Kashmir. *Academia. edu*.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.

Shewalkar, A. N. (2018). Comparison of rnn, lstm and gru on speech recognition data.

Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, *9*(4), 235-245.

Shin, D. H., Choi, K. H., & Kim, C. B. (2017). Deep learning model for prediction rate improvement

of stock price using RNN and LSTM. *The Journal of Korean Institute of Information Technology*, *15*(10), 9-16.

Singh, G., Bhandari, R., & Singh, P. (2024, January). Advancing NLP for Punjabi Language: A Comprehensive Review of Language Processing Challenges and Opportunities. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (pp. 1250-1257). IEEE.

Singh, J., Singh, G., Singh, R., & Singh, P. (2021). Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification. *Journal of King Saud University-Computer and Information Sciences*, *33*(5), 508-517.

Sitender, Bawa, S., Kumar, M., & Sangeeta. (2023). A comprehensive survey on machine translation for English, Hindi and Sanskrit languages. *Journal of Ambient Intelligence and Humanized Computing*, *14*(4), 3441-3474.

Smythe, S., & Toohey, K. (2009). Investigating sociohistorical contexts and practices through a community scan: A Canadian Punjabi–Sikh example. *Language and Education*, *23*(1), 37-57.

Sun, L., Du, J., Dai, L. R., & Lee, C. H. (2017, March). Multiple-target deep learning for LSTM-RNN based speech enhancement. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)* (pp. 136-140). IEEE.

Zargar, S. (2021). Introduction to sequence learning models: RNN, LSTM, GRU. *Department of Mechanical and Aerospace Engineering, North Carolina State University*.

Zen, H. (2015). Acoustic modeling in statistical parametric speech synthesis-from HMM to LSTM-RNN. *Proc. MLSLP*, *15*.

Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., & Szczepaniak, P. (2016). Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. *arXiv preprint arXiv:1606.06061*.

Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., & Tian, G. (2020, November). Do RNN and LSTM have long memory?. In *International Conference on Machine Learning* (pp. 11365-11375). PMLR.

Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630*.