Journal of Computer and Electrical Sciences (JCES)- Volume-1 Issue-4, April 2026
ISSN: 3049-2602

A HYBRID MULTILINGUAL NEXT WORD PREDICTION SYSTEM USING STATISTICAL AND LSTM MODELS WITH NLP-BASED EXPLAINABILITY

Roshini Baruva[1] Mohammad Baig Mohammad[2]
[1]Student, Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning)
[2]Assistant Professor, Department of Computer Science and Engineering (Artificial Intelligence & Machine Learning)
Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh, India
Email id: kavithakumari*baruva676@gmail.com*

Abstract*: Next word prediction is a key task in Natural Language Processing (NLP) that enhances user interaction by providing intelligent text suggestions. This paper presents a hybrid multilingual next word prediction system that integrates both statistical and deep learning approaches to improve prediction accuracy and usability. The proposed system combines a frequency-based statistical model with a Long Short-Term Memory (LSTM) model to capture both short-term and long-term dependencies in text. It accepts a single-word input and generates top predicted words along with their probability scores, supporting both English and Telugu for broader usability. A language validation mechanism ensures correct prediction based on the selected language. The system also emphasizes explainability by displaying NLP processing steps such as tokenization, normalization, encoding, and prediction generation, along with model comparison and performance metrics including accuracy, loss, and prediction status. Graphical visualization is used to represent prediction confidence and model behavior. The system is implemented using Python, Flask, and TensorFlow with a web-based interface, and experimental results demonstrate effective handling of various prediction scenarios.*

Keywords — *Next Word Prediction, NLP, LSTM, Statistical Model, Multilingual Prediction, Explainable AI, Text Prediction, Deep Learning*

1. INTRODUCTION

Next word prediction is an important application of Natural Language Processing (NLP) that enhances typing efficiency and improves user experience. It is widely used in applications such as chat systems, search engines, and smart keyboards. Traditional approaches rely on statistical models such as N-grams, which predict the next word based on frequency. However, these models are limited in capturing deeper contextual relationships within text. In contrast, deep learning models such as Long Short-Term Memory (LSTM) networks provide better contextual understanding by learning sequential dependencies, but they are computationally expensive and often lack interpretability.

To address these limitations, this work proposes a hybrid approach that combines both statistical and LSTM models. The system also incorporates multilingual support and explainability features, making it more practical, efficient, and user-friendly.

2. LITERATURE SURVEY

Existing research in next word prediction highlights the use of statistical models such as N-grams, which are simple and computationally efficient but limited in capturing contextual relationships within text. Neural network-based approaches, particularly Long Short-Term Memory (LSTM) models, have been

widely adopted to improve prediction accuracy by learning sequential dependencies and contextual information. Furthermore, advanced deep learning models enhance performance but introduce increased computational complexity and resource requirements.

Multilingual prediction systems extend functionality across different languages; however, they require large-scale datasets and involve complex processing mechanisms. Additionally, many existing approaches lack explainability, providing limited insight into how predictions are generated.

To address these challenges, the proposed system adopts a hybrid approach that combines statistical and LSTM models, while also incorporating multilingual support and explainability features to improve both performance and usability.

3. PROPOSED SYSTEM

The proposed system is a Hybrid Multilingual Next Word Prediction System designed to overcome the limitations of traditional prediction models by combining statistical and deep learning approaches. The system integrates a frequency-based statistical model with a Long Short-Term Memory (LSTM) neural network to achieve both efficiency and contextual understanding in prediction.

Unlike conventional systems that rely solely on either statistical or neural methods, the proposed hybrid architecture leverages the strengths of both approaches. The statistical model ensures fast and interpretable predictions based on word frequency, while the LSTM model captures long-term dependencies and semantic relationships within text.

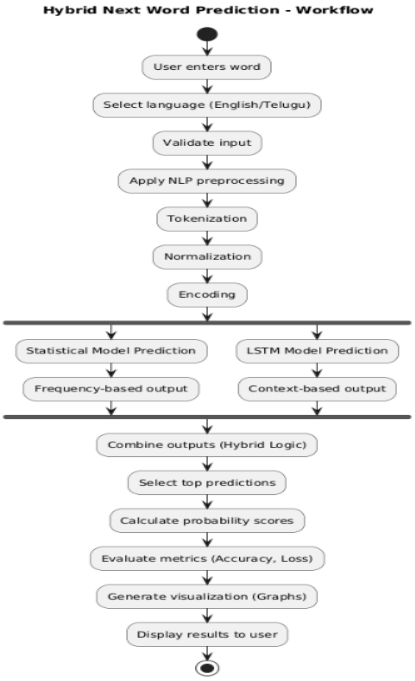


Fig. 1: Proposed System

As shown in Fig. 1, the system follows a structured hybrid workflow that combines preprocessing, model prediction, and result generation stages.

Key Features

- **Single-word input prediction:** The system accepts a single word as input to simplify user interaction and reduce computational complexity.

- **Top 4 predictions with probabilities:** The system provides multiple predicted words along with probability scores, enabling better decision-making.
- **Multilingual support:** Supports both English and Telugu, making the system suitable for diverse users.
- **Language validation:** Ensures that the input matches the selected language to prevent incorrect predictions.
- **NLP step visualization:** Displays preprocessing steps such as tokenization, normalization, and encoding for better understanding**.**
- **Model comparison:** Compares outputs from both statistical and LSTM models, highlighting their contributions.
- **Performance metrics**: Provides accuracy, loss, and prediction status (Correct, Partial, Neutral, Wrong, Error).
- **Graph visualization:** Dynamically represents prediction confidence and model performance.

Working of the System

The system follows a structured pipeline:

1. The user enters a single word and selects the desired language (English or Telugu).
2. The input is validated to ensure correctness and language compatibility.
3. NLP preprocessing is applied, including tokenization, normalization, and encoding.
4. The processed input is passed to both the statistical and LSTM models.
5. The statistical model generates predictions based on frequency patterns in the dataset.
6. The LSTM model generates predictions based on learned contextual relationships.
7. The outputs from both models are combined using hybrid decision logic.
8. The system selects the top predicted words along with probability scores.
9. Performance metrics such as accuracy and loss are calculated.
10. Graphical visualization is generated to represent prediction confidence.
11. The final results, including predictions, model comparison, NLP steps, metrics, and graphs, are displayed to the user.

This workflow ensures that the system is both efficient and intelligent, providing accurate predictions along with clear explanations.

4. Methodology

The methodology of the proposed system is based on a hybrid approach that combines statistical and deep learning techniques to improve prediction performance and efficiency.

4.1 Statistical Model

The statistical model is based on frequency analysis of word sequences. It learns patterns from the dataset by analyzing how frequently a word follows another word.

During training:

- The dataset is processed to generate word sequences.
- Frequency counts of word pairs are stored.
- Probabilities are calculated based on occurrence.

During prediction:

- The model retrieves the most frequent next words.
- Words are ranked based on probability.

Advantages:

- Fast and computationally efficient.
- Easy to interpret.
- Performs well for common word patterns.

Limitation:

- Unable to capture long-term context or semantic meaning.

4.2 LSTM Model

The LSTM model is a type of recurrent neural network designed to handle sequential data. It is capable of learning long-term dependencies and contextual relationships between words.

Working:

- Text is converted into numerical sequences using tokenization.
- These sequences are fed into the LSTM network.
- The model learns patterns during training.
- It predicts the probability of the next word.

LSTM Components:

- **Input gate:** Controls new information entering the network.
- **Forget gate:** Removes irrelevant information.
- **Output gate:** Generates the final prediction.

Advantages:

- Effectively captures contextual information.
- Provides higher prediction accuracy.
- Handles complex sentence structures.

Limitations:

- Requires higher computational resources.
- Less interpretable compared to statistical models.

4.3 Hybrid Decision Logic

The hybrid model combines the outputs of both statistical and LSTM models to generate the final prediction.

- If strong frequency patterns exist, the statistical model is prioritized.
- If contextual understanding is required, the LSTM model is preferred.
- Final predictions are selected based on combined probability scores.

This hybrid approach:

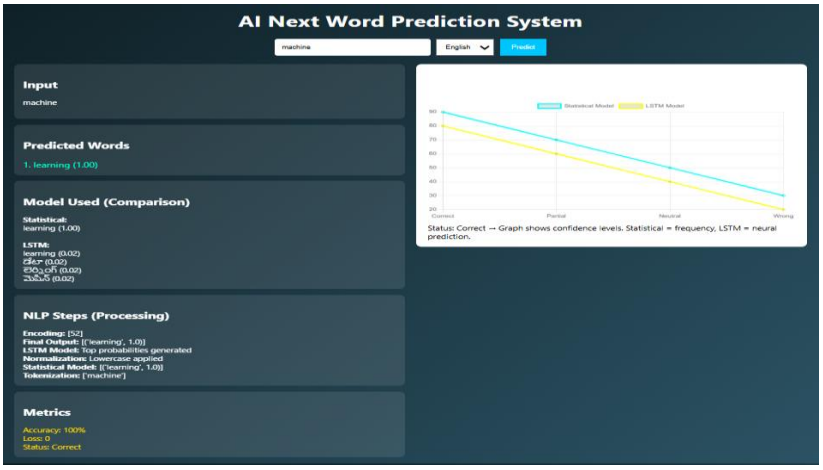
- Improves overall prediction accuracy.
- Reduces the limitations of individual models.
- Balances computational efficiency and contextual intelligence.

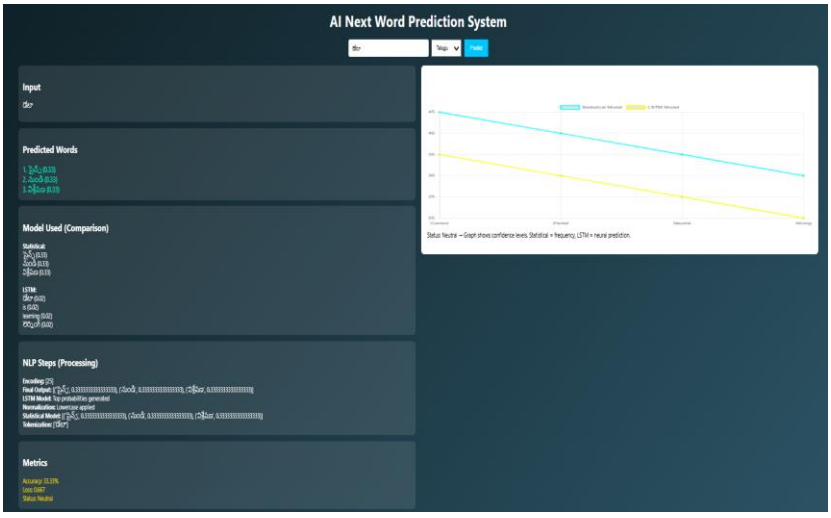
5. RESULTS AND DISCUSSION

The proposed system was evaluated using multiple test cases representing different input scenarios. The results demonstrate the effectiveness of the hybrid approach in handling diverse prediction conditions.

Prediction Scenarios

- **Strong words:**
  Frequently occurring words result in high-confidence predictions. The accuracy is high, and the prediction status is classified as *Correct*.
- **Medium words:**
  Words with moderate frequency produce multiple possible outputs. The accuracy is moderate, leading to a *Partial* prediction status.
- **Weak words:**
  Rare or less frequent words result in low-confidence predictions. The prediction status is typically *Neutral* or *Wrong*.
- **Unknown words:**
  Words not present in the dataset result in prediction failure. The accuracy is zero, and the loss is maximum.





System Performance

The system successfully:

- Generates meaningful predictions for valid inputs.
- Supports both English and Telugu languages.
- Displays detailed NLP preprocessing steps.

- Compares outputs from statistical and LSTM models.
- Calculates dynamic performance metrics.
- Updates graphical visualizations based on prediction status.

Discussion

The results clearly indicate that:

- The statistical model performs well for frequently occurring patterns.
- The LSTM model performs better for context-based predictions.
- The hybrid approach provides balanced and improved performance.

Graphical visualization further enhances understanding of model confidence and behavior across different scenarios. Overall, the system effectively handles all prediction cases, including correct, partial, neutral, wrong, and error conditions.

6. CONCLUSION

The proposed Hybrid Multilingual Next Word Prediction System effectively addresses the limitations of existing prediction models by integrating statistical and LSTM approaches. The system provides accurate, efficient, and explainable predictions, making it suitable for real-world applications.

The inclusion of multilingual support enhances accessibility by enabling users to interact with the system in both English and Telugu, while the language validation mechanism ensures correctness and prevents invalid predictions. A key contribution of this work is its emphasis on explainability, achieved through the visualization of NLP processing steps, model outputs, performance metrics, and graphical representations.

Furthermore, the hybrid model improves overall performance by combining the speed of statistical methods with the contextual intelligence of deep learning models, ensuring reliable predictions across diverse input scenarios.

Overall, the system demonstrates strong performance, usability, and scalability, and can be extended to support additional languages, larger datasets, and real-time applications such as smart keyboards, chatbots, and assistive writing tools.

REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[2] C. Aliprandi, "Advances in NLP Applied to Word Prediction," 2008.
[3] P. F. Brown et al., "Predicting Sentences Using N-gram Language Models," Proc. Empirical Methods in Natural Language Processing (EMNLP), pp. 141–149, 2009.
[4] T. Mikolov et al., "Recurrent Neural Network Based Language Model," Proc. Interspeech, 2010.
[5] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.
[6] D. Nagalavi and M. Hanumanthappa, "N-gram Word Prediction Language Models to Identify the Sequence of Article Blocks in English E-Newspapers," Proc. CSITSS, 2016.
[7] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.
[8] R. V. Regalado, "Use of Word and Character N-grams for Low-Resourced Local Languages," Proc. IALP, 2018.
[9] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.
[10] A. Tiwari, "Next Word Prediction Using Deep Learning," IEEE GlobConPT, 2022.

[11] S. Singh, "On-Device User-Adaptive Next Word Prediction System," Proc. ICCSEA, 2022.
[12] S. Singh, "NLP-Based Next Word Prediction Model," Proc. CSITSS, 2023.
[13] A. Radford et al., "Language Models are Few-Shot Learners," OpenAI, 2020.
[14] T. Brown et al., "GPT-3: Language Models are Few-Shot Learners," NeurIPS, 2020.
[15] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," 2022.
[16] Z. Zhang et al., "Recent Advances in Neural Language Models for Text Prediction," IEEE Access, 2023.
[17] K. Lee et al., "Efficient Multilingual Language Models for NLP Applications," IEEE Transactions on AI, 2024.