e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 11, November 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711098

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Telugu Document Classification Using Deep Recurrent Neural Network with Bi LSTM

M. Navadeep*[1], A. Sneha*[2], A. Navaneeth Rao*[3], K. Navya Sri*[4], B. Navya*[5],

Prof. Dr.Sujit Das[6]

Department of CSE (AIML), Malla Reddy University, Hyderabad, India[1,2,3,4,5,6]

ABSTRACT: Telugu Document Classification is an essential task for organizing, retrieving, and analyzing large amounts of Telugu textual data. This paper presents a deep learning-based approach for classifying Telugu documents using a Bidirectional Long Short-Term Memory (BiLSTM) model. Due to the agglutinative nature and complex structure of the Telugu language, traditional classification algorithms often struggle to capture the contextual meaning and nuances. The proposed method leverages Bi-LSTM, a type of Recurrent Neural Network (RNN), which can capture long-term dependencies by considering both past and future contexts in the document. The model is trained on a dataset of Telugu documents from various categories, such as politics, sports, and entertainment. The text is preprocessed using tokenization and padding techniques, and the labels are encoded using supervised learning. Results demonstrate that the Bi-LSTM model outperforms traditional machine learning algorithms in capturing the linguistic complexity of Telugu, providing better classification accuracy. Our project can be applied to real-world scenarios such as news categorization, sentiment analysis, and content moderation, enabling better management of Telugu language data. The model's flexibility and performance make it a strong candidate for other low-resource languages as well.

KEYWORDS: Telugu Document Classification ,Deep Learning ,Bidirectional Long Short-Term Memory (BiLSTM) ,Recurrent Neural Networks (RNN) ,Agglutinative Language ,Text Preprocessin ,Tokenization ,Padding Techniques ,Supervised Learning ,Classification Accuracy ,Linguistic Complexity ,News Categorization ,Sentiment Analysis ,Content Moderation ,Low-Resource Languages.

I. INTRODUCTION

The exponential growth of digital content in Telugu has created a pressing need for efficient methods to organize, retrieve, and analyze large volumes of textual data. Telugu, a Dravidian language spoken by millions, presents unique challenges for document classification due to its agglutinative nature, complex sentence structures, and rich vocabulary. Traditional text classification techniques often fail to fully capture the linguistic intricacies of the language, leading to suboptimal performance in real-world applications such as news categorization, sentiment analysis, and content moderation. In response to these challenges, this project explores the use of Deep Recurrent Neural Networks (RNN), with Bidirectional Long Short-Term Memory (Bi-LSTM) for Telugu document classification. Bi-LSTM is particularly well-suited for handling sequential data, as it captures longterm dependencies by processing textual information in both forward and backward directions. This allows for a more comprehensive understanding of context, which is critical in accurately classifying documents with complex sentence patterns and nuances, as seen in Telugu.

II. LITERATURE REVIEW

Recent advancements in natural language processing (NLP) have improved text categorization for Indian languages, including Telugu. Traditional machine learning models like Support Vector Machines (SVM) and Naïve Bayes have limitations in handling Telugu's rich morphology due to their reliance on manual feature engineering. In contrast, deep learning techniques—especially Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Bidirectional LSTM (Bi-LSTM) models—capture contextual information more effectively. Studies have shown Bi-LSTM and hybrid models like LSTM-CNN outperform traditional models in sentiment analysis and document classification tasks by managing sequential data and avoiding overfitting. Building on these findings, this project seeks to develop a deep learning-based classification system tailored to Telugu's unique linguistic features, contributing to

NLP advancements for regional languages

Several methods for domain identification and text categorization have been done on Indian languages, and few of the works have been reported on the Telugu language. In this section, we survey some of the methodologies and approaches used to address domain identification and text categorization. Traditional machine learning algorithms, such as Support Vector Machines (SVM) and Naïve Bayes, have been commonly used for document classification. However, their reliance on feature engineering limits their effectiveness, particularly for languages with rich morphology like Telugu. Recent advancements in deep learning, especially with the introduction of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, have shown significant improvements in text classification tasks. A study by Zhang et al. (2018) demonstrated that LSTM networks outperformed traditional classifiers in various NLP tasks [2]. The Bidirectional LSTM (Bi-LSTM) architecture has gained popularity due to its ability to capture contextual information from both past and future sequences. Research by Huang et al. (2019) highlighted the effectiveness of Bi-LSTMs in sentiment analysis and text classification [3].

Natural Language Processing (NLP) is a systematic [4] and significant approach to produce meaningful text that is understandable by humans. For generating the text, the data is collected from different sources or taken as input from the users. There has been a drastic change in the field of NLP over the past few years[5]. RNN [6] was popularly used for text generation because of its ability to process sequential data. But due to its limitation of vanishing gradients, it is being replaced by other neural networks. LSTM, and other versions of LSTM, i.e., Bi-LSTM , GRU are being popularly used nowadays for generating text in most of the Indian languages. These models are also being used for other NLP related tasks like Query auto-completion, story generation. The morphologically rich Dravidian classical language Telugu was the focus of a text categorization presentation by K. Rajan et al. [7], which utilised a Vector Space Model and an Artificial Neural Network. From our experiments, we know that an A.N.N. model can correctly classify 93.33 percent of Tamil documents. Nguyen and Nguyen [8] proposed a combination of LSTM and CNN in the sentiment analysis. CNN was employed in the filters to capture local dependencies and the LSTM was used to store the information for the long term. The freezing technique was used in this method to avoid the overfitting problem in deep learning. The experimental results showed that the developed model achieved higher performance in the sentiment analysis. The Naïve Bayes with SVM (NBSVM) model was also added in a neural voting ensemble to boost the performance. However, the similarity values were not calculated to form the clusters of NBSVM. Chen, Xu, He and Wang [9] proposed a divide and conquer method to classify the sentence into various types and sentiment analysis was performed on each type. According to the number of targets, the sentences were classified using the Neural Network based sequence model. One dimensional CNN obtained each class of sentence as input for final classification.

III. PROBLEM STATEMENT

The project is focused on developing a deep learning model for classifying Telugu documents by leveraging advanced deep recurrent neural networks, specifically a Bidirectional Long Short-Term Memory (Bi-LSTM) architecture. Given the exponential rise in digital content in regional languages like Telugu, there’s a growing demand for systems that can accurately classify and organize this content to improve information accessibility and retrieval for users. By creating a model that automatically categorizes Telugu documents into predefined classes, the project aims to enhance user experience by making content more organized and accessible.

A critical first step in this project is “data preprocessing” to prepare the text for modeling. Preprocessing includes various tasks such as text cleaning, tokenization, embedding, and handling class imbalance. “Text cleaning” involves removing unwanted characters, special symbols, and stop words from the data, as well as performing any necessary tokenization to convert raw text into manageable components. This step is particularly important for Telugu, as the language has many unique characters and diacritics that need to be processed appropriately to ensure model accuracy. “Tokenization and embedding” follow text cleaning, where Telugu text is converted into sequences of tokens and mapped to numerical embeddings. Given that Telugu is a morphologically rich language, generic word embeddings may not be sufficient; specialized embedding models, or pretrained embeddings compatible with Telugu, are likely required to effectively capture word relationships and meanings. Additionally, “handling class imbalance” is necessary if some categories are over- or under-represented in the dataset, which can impact model performance. Techniques such as oversampling or undersampling may be employed to address this imbalance and ensure that the model learns to classify all categories accurately.

The "Deep Recurrent Neural Network (RNN) architecture" selected for this project is centered around a "Bi-Directional Long Short-Term Memory (Bi-LSTM)" layer. The Bi-LSTM model improves the RNN's ability to handle long sequences by reading the input sequence in both directions—forward and backward. This bidirectional processing allows the model to capture context from both preceding and succeeding words in a sentence, making it highly effective in understanding dependencies and relationships between words. This is especially useful for Telugu, where sentence structure and word morphology can be complex and context-dependent. By leveraging Bi-LSTM, the model will be better equipped to understand the intricacies of Telugu syntax and semantics, leading to more accurate document classification.

IV. SYSTEM DESIGN

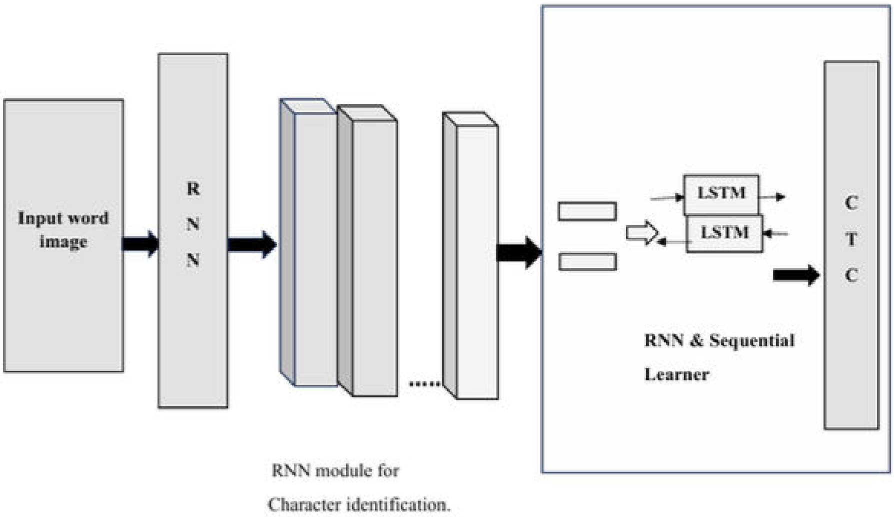


Fig. 1. Architecure

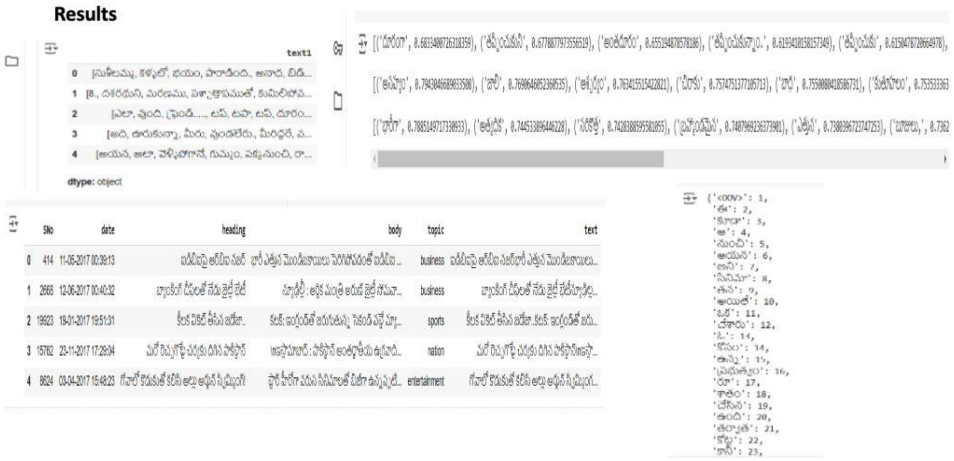
V. METHODOLOGY

TELUGU DOCUMENT CLASSIFICATION USING DEEP RECURRENT NEURAL NETWORK WITH BI LSTM methods and algorithms:

1. **Text Preprocessing**: Techniques like tokenization, stopword removal, and normalization clean Telugu text, reduce noise, and ensure consistency, making it easier for the model to process.
2. **Word Embeddings**: Algorithms like FastText and Word2Vec represent Telugu words as vectors, capturing semantic relationships and context, which helps the model understand meaning and nuances.
3. **Bi-LSTM**: This bidirectional network processes text both forward and backward, capturing long-range dependencies and contextual details, essential for understanding complex Telugu syntax.
4. **Attention Mechanism**: Focuses on important parts of the text, prioritizing key words and phrases, which improves classification accuracy by highlighting contextually relevant information.
5. **Softmax Classifier**: In the final layer, Softmax generates probabilities for each class, enabling clear and confident multi-class categorization.

These methods together create a powerful pipeline for accurate Telugu text classification, handling language-specific complexities effectively.

VI. EXPERIMENTAL RESULTS



Results

```
text1
0    [సుశీలమ్మ, కళ్ళలో, భయం, పారాడింది., అనాథ, బిడ...
1    [8., దశరథుని, మరణము, పశ్చాత్తాపముతో, కుమిలిపోవ...
2        [ఎలా, వుంది, ఫ్రెండ్....., టప్, టపా, టప్, దూరం...
3      [అది, ఊరుకున్నా, మీరు, వుండలేరు., మీరిద్దరే, వ...
4    [ఆయన, అలా, వెళ్ళిపోగానే, గుమ్మం, పక్కనుంచి, రా...

dtype: object
```

```
[('దూరంగా', 0.6833400726318359), ('తప్పించుకుని', 0.6778877973556519), ('అంతదూరం', 0.655194878578186), ('తప్పించుకున్నాం.', 0.6193418158157349), ('తప్పించుకు', 0.6150478720664978),

[('అసహ్యం', 0.7943046689033508), ('జాలి', 0.7690646052360535), ('ఆశ్చర్యం', 0.763415515422821), ('చిరాకు', 0.7574751377105713), ('బాధ', 0.7550080418586731), ('సంతోషం', 0.753533363

[('భారీగా', 0.7885149717330933), ('అత్యధిక', 0.7445338964462228), ('సరికొత్త', 0.7428388595581055), ('ప్రత్యేకమైన', 0.7407969236373901), ('ఎత్తున', 0.7380396723747253), ('బజారు,', 0.7362
```

	SNo	date	heading	body	topic	text
0	414	11-06-2017 00:39:13	ఐడీబీఐపై ఆర్‌బీఐ నజర్	భారీ ఎత్తున మొండిబకాయిలు పెరిగిపోవడంతో ఐడీబీఐ ...	business	ఐడీబీఐపై ఆర్‌బీఐ నజర్భారీ ఎత్తున మొండిబకాయిలు...
1	2668	12-06-2017 00:40:32	బ్యాంకింగ్ చీఫ్‌లతో నేడు జైట్లీ భేటీ	న్యూఢిల్లీ : ఆర్థిక మంత్రి అరుణ్ జైట్లీ సోమవా...	business	బ్యాంకింగ్ చీఫ్‌లతో నేడు జైట్లీ భేటీన్యూఢిల్లి...
2	19923	19-01-2017 19:51:31	కీలక వికెట్ తీసిన జడేజా..	కటక్: ఇంగ్లండ్‌తో జరుగుతున్న సెకండ్ వన్డే మ్యా...	sports	కీలక వికెట్ తీసిన జడేజా..కటక్: ఇంగ్లండ్‌తో జరు...
3	15762	23-11-2017 17:29:04	మరో రెచ్చగొట్టే చర్యకు దిగిన పాకిస్తాన్	\nఇస్లామాబాద్ : పాకిస్తాన్ అంతర్జాతీయ ఉగ్రవాది...	nation	మరో రెచ్చగొట్టే చర్యకు దిగిన పాకిస్తాన్\nఇస్లా...
4	8624	03-04-2017 15:48:23	గోవాలో కొడుకుతో కలిసి అల్లు అర్జున్ స్విమ్మింగ్!	స్టార్ హీరోగా వరుస సినిమాలతో బిజీగా ఉన్నప్పటి...	entertainment	గోవాలో కొడుకుతో కలిసి అల్లు అర్జున్ స్విమ్మింగ...

```
{'<OOV>': 1,
 'ఈ': 2,
 'కూడా': 3,
 'ఆ': 4,
 'నుంచి': 5,
 'ఆయన': 6,
 'అని': 7,
 'సినిమా': 8,
 'తన': 9,
 'అయితే': 10,
 'ఒక': 11,
 'చేశారు': 12,
 'ఓ': 13,
 'కోసం': 14,
 'ఉన్న': 15,
 'ప్రభుత్వం': 16,
 'రూ': 17,
 'శాతం': 18,
 'చేసిన': 19,
 'ఉంది': 20,
 'తర్వాత': 21,
 'కోట్ల': 22,
 'కానీ': 23,
```

VII. CONCLUSION

The "Telugu Document Classification Using Deep Recurrent Neural Network with Bi-LSTM" project effectively showcased the application of deep learning in natural language processing for the Telugu language. The Bi-LSTM model demonstrated strong performance by capturing contextual information and long-range dependencies in text data, resulting in improved classification accuracy. Comprehensive evaluation metrics, including accuracy, precision, recall, and F1 score, validated the model's effectiveness. The project emphasizes the importance of continuous monitoring and updates in deployment to maintain performance. Overall, it contributes to the advancement of multilingual document classification and paves the way for future enhancements in NLP applications.

VIII. FUTURE ENHANCEMENT

future enhancement for the "Telugu Document Classification Using Deep Recurrent Neural Network with Bi-LSTM" project, presented in five points:

a. Exploration of Advanced Architectures :

- Adopting architectures like Transformers or hybrid models that integrate CNNs and RNNs can enhance the model's ability to capture complex contextual relationships within Telugu text. These architectures excel at processing sequential data and can significantly boost classification performance by leveraging self-attention and parallel processing capabilities.

b. Incorporation of Attention Mechanisms :

- Adding attention mechanisms will allow the model to selectively focus on the most relevant parts of longer documents. This capability can improve the extraction of significant information, ensuring that important words or phrases receive more weight during classification, thereby enhancing overall accuracy and relevance in outputs.

c. Domain-Specific Customization :

- Tailoring the model for specific domains such as legal, medical, or technical fields can increase its accuracy and effectiveness. By fine-tuning the model with domain-specific data, it can learn unique terminologies and contexts, resulting in improved classification performance for specialized texts.

D. Utilization of Transfer Learning :

- Implementing transfer learning techniques with pre-trained models can enhance the robustness of the classification system, especially in scenarios with limited training data. This approach allows the model to leverage knowledge from previously trained models, improving performance and generalization across various document types.

e. Expansion for Multi-Language Support :

Broadening the model's capabilities to support multiple languages can increase its applicability and user base. A multilingual framework would enable the system to cater to diverse linguistic needs, making it a versatile tool for

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

document classification across different languages and cultural contexts, thus improving user experience and accessibility.

These enhancements collectively aim to strengthen the model's performance, adaptability, and user satisfaction in practical applications.

REFERENCES

[1] M Abdul Rahiman and M S Rajasree. Printed Malayalam Character Recognition Using Back-propagation Neural Networks, IEEE, IACC 2009.

[2] Abhijit dutta and santanu chaudhury. Bengali alpha-numeric character recognition using curvature features, Pattern Recognition, 26(12): pp. 1757-1770, 1993.

[3]Nidhi Kalidas Sawant and Prof. Sangam Borkar. Devanagari Printed Text to Speech Conversion using OCR.IEEE ISBN:978-1-5386-1442-6.

[4] Alex Graves, Santiago Fernandez, Faustino Gomez and Jurgen Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks International Conference on Machine Learning, 2009.

[5] Baoguang Shi, Xiang Bai and Cong Yao. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition.arxiv.org/abs/1507.05717

[6] Yann LeCun, Patrick Haffner, Leon Bottou and Yoshua Bengio. Object recognition with Gradient-Based Learnin. IEEE, 2278 – 2324, 1998.

[7] Y. Bengio, P. Y. Simard, and P. Frasconi. Learning longterm dependencies with gradient descent is difficult. NN, 5(2):157–166, 1999

[8] Sepp Hochreiter and Jurgen Schmidhuber. Long Short Term Memory, Neural Computation :1735-1780,1997

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com