Telugu News Classification

Aditya Rajesh Sakri, Vaka Satwik Reddy, Vamsi Krishna Vunnam, Ms. Priyanka Vivek
Department of Computer Science and Engineering, Amrita School of Engineering Bengaluru, Amrita Vishwa Vidyapeetham, India

Introduction

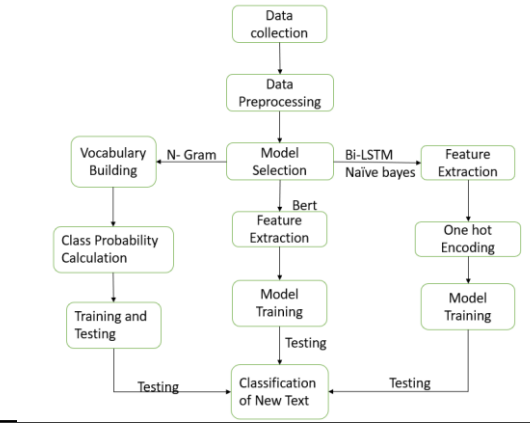
In our project on Telugu news classification, we explored the application of both Bi-LSTM (Bidirectional Long Short-Term Memory) and BERT (Bidirectional Encoder Representations from Transformers) models. These models were employed to address the challenges posed by the linguistic characteristics and complexity of the Telugu language.

We first utilized the Bi-LSTM model, which is a type of recurrent neural network capable of capturing sequential information effectively we also incorporated the BERT model, a powerful transformer-based language model known for its ability to capture contextual information. Leveraging BERT's pre-trained model, we fine-tuned it on our Telugu news dataset to specialize its understanding of Telugu language nuances and news topics. By combining the deep contextual understanding of BERT with the sequential information processing of Bi-LSTM, we aimed to enhance the accuracy and performance of our Telugu news classification system

Methodology

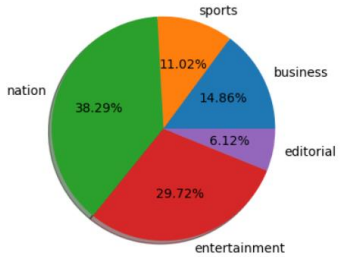
The methodology for text classification involving Naive Bayes, LSTM, and BERT models with Count Vectorizer, word embeddings, and BERT embeddings includes the following steps. Firstly, the data is split into training and testing sets. For Naive Bayes, the text is transformed into numerical features using Count Vectorizer. The Naive Bayes model is then trained and evaluated based on accuracy.

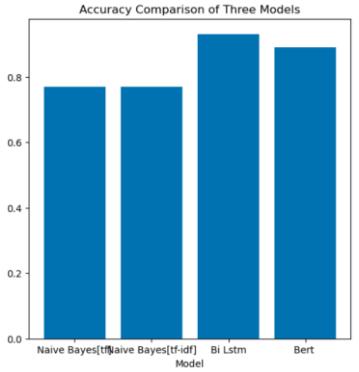
.



Experiments and Results

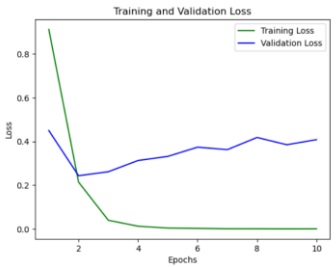
The feature extraction process in text classification involves preprocessing text, generating n-grams, and representing them numerically. Naive Bayes uses TF and TF-IDF, while Bidirectional LSTM utilizes tokenization and padding. In addition, BERT, a contextual language model, can be employed for text classification by obtaining BERT embeddings. The models used include N-Grams, Naive Bayes, Bidirectional LSTM, and BERT, enabling effective classification of text data into predefined categories

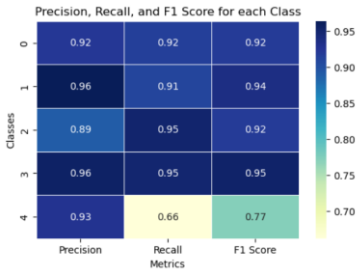




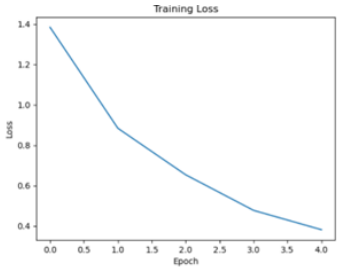
Models and Outputs

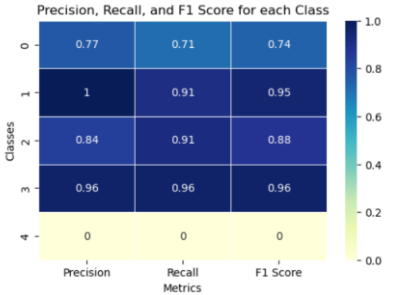
Bi - LSTM





Bert





Conclusion

In this study, we compared Naive Bayes with n-gram features and a Bi-LSTM model for Telugu text classification. The Bi-LSTM model achieved remarkable accuracy of 93%, surpassing the Naive Bayes model's limited accuracy of 38.5% for unigrams. Additionally, when incorporating BERT and training it on the full sequence of data, an accuracy of 87% was attained. These findings highlight the superiority of deep learning models for accurate Telugu text classification.