# A FUSION BASED APPROACH FOR EXTRACTIVE TEXT SUMMARIZATION OF TELUGU DOCUMENTS

**ABSTRACT**

Nowadays, the data is becoming voluminous both online and offline. Understanding and analysing the vital information from available information is a painstaking and time-consuming task. Automated text summarization systems address the challenges mentioned above by helping the readers to understand the content quickly without reading the entire text. Text Summarization performs extraction/abstraction of the original text's condensed form by retaining its full meaning. Text Summarizations are used in various applications, such as making the headlines for news documents, outline notes for the students, generating minutes of the meeting, reports of weather forecasting-stock exchange constructing the movie previews, and preparing a summary of histories by arranging the events in chronological order. Automatic Text summarization has become a thirst challenging topic of research in natural language processing.

A vast majority literature focused on summarization methods for English. Several advancements were made for building English text summarization systems but not for Indian languages. Especially for Telugu, only a few methods are proposed because of its agglutinative and lexical complexity nature of language. Telugu is the 2nd famous language in India and the 15th most popularly speaking world language. Text summarizations developed for other Indian languages like Hindi, Bengali does not support Telugu due to its complex linguistic characteristics. Text summarization for Telugu obtained little attention due to the non-availability of Telugu resources like data sets, dictionaries, wordnet, etc. This thesis focuses on building a fusion-based approach that combines both the Feature-Extraction and Long short term memory (LSTM) based methods for effectively scoring the sentences in text summarization for Telugu documents.

The research work commences from developing an improved sentence ranking algorithm with the event and named entity scores to extract the informative sentences into the summary. In the feature-extraction approach, only the linguistic and syntactic features of the text are extracted. In long-short term memory (LSTM) with rectified Adam Optimizer ranks the sentences based on the semantic features of words. In this, the 300-dimension word vector of Telugu words is used to train the model. The sentences are scored based on the significance of words context. The proposed architecture jointly captures the syntactic and semantic features to identify worthy sentences, which eventually generates a better summary.

The highest-ranked sentences above the threshold are selected to form the summary. The uniqueness detection step performs the task of choosing the non-redundant sentences into the generated summary. We build a new corpus of 360 Telugu text documents collected from online news sources (i.e., Eenadu, Sakshi, Andhra Jyothi, Namaste Telangana). Evaluation metrics like precision, recall, and F-score is used to measure the proposed method's performance. The results showed improved performance compared with existing Telugu text summarization methods mentioned in the state-of-the-art.