International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Enhancing Abstractive Text Summarization using Two-Staged Network for Telugu Language (EATS2N)

**Srisudha Garugu[1*], D. Lalitha Bhaskari[2]**

**Submitted**: 09/01/2024 **Revised**: 15/02/2024 **Accepted**: 24/02/2024

**Abstract:** The process of producing a clear and short synopsis of lengthy texts without sacrificing the overall meaning by concentrating on the passages that provide important information is known as text summarization Extractive summaries that highlight a significant portion of the input texts frequently include crucial keywords. The vast majority of strategies for extractive summarization are based on the idea of locating keywords and extracting sentences that have a disproportionately high number of keywords compared to the others. The process of extracting keywords often involves identifying relevant terms that occur more frequently than other words and putting an emphasis on the most significant of them.Selecting keywords manually is challenging, susceptible to inaccuracies, and demands considerable time and attention.A technique that can automatically extract keywords from Telugu e-newspaper datasets was proposed by using this work. The keywords may then be used for text summarizing. The proposed method compares two different datasets, the telugu newspaper and the biology text book and the performance metrics are compared using the accuracy and ROGUE score values.

***Keywords****: Summarization, Extractive Summary, Keywords, Telugu News Paper, Biology Textbook*

## 1. Introduction

In this day and age of the internet, there is a wealth of knowledge available to users online for free. This material may be found in the form of technical reports, journal articles, e-Newspapers, transcribed dialogues, and so on.

The aforementioned digital media each contain a vast number of documents, making it a challenging task for users to glean only the information that is pertinent from each of these media within the allotted amount of time. There is a requirement for a computerized system that is able to glean only the information that is pertinent from the given sources of data. In order to accomplish the given task, one must do text mining included within the papers. Text mining is the method of extraction of information from massive text quantities using specialized software. Text mining makes use of a number of the natural language processing (NLP) techniques in order to do text analysis. These techniques include tokenization, N-grams, parsing, POS (parts of speech) tagging, and others. It incorporates activities such as the automatic extraction of keywords and the summarizing of text. The practice of selecting phrases and words from a text document which could at best portray the fundamental sentiment of the document without any interaction from a human being is referred to as automatic keyword extraction [1].

*[1]Research Scholar, Department of Computer Science & System Engineering, srisudha.garugu@gmail.com, Andhra University College of Engineering (A), Andhra University, Visakhapatnam-530003, India[1]*
*[2]Professor, Department of Computer Science & System Engineering, Coordinator, IQAC, lalithabhaskari@yahoo.co.in, Andhra University College of Engineering (A), Andhra University, Visakhapatnam-530003, India[2]*

The level of human engagement in automatic keyword extraction depends on the model utilized. The primary objective is to capitalize on current computational strengths to refine information retrieval and structuring. Significantly, this method sidesteps the added costs of involving human annotators.

Summarization is a procedure that involves extracting the highly essential aspects of a piece of writing and compiling them into a condensed version of the original document that you are attempting to summarize [2]. Text summarizing, as defined by Mani and Maybury [3] is the method of extracting the most vital information from a given piece of writing in order to provide a condensed version suitable for a certain purpose and audience. Even though they are typically only around 17% as long as the original text, summaries nonetheless include all of the information that a reader could have gained from studying the primary source [4]. In the aftermath of big data analysis, summarization has emerged as a technique that is both effective and powerful in providing an overview of the entire data set. The text could be summarized in two different methods: extractive and abstractive. Both of these summarizing approaches are possible.

The abstractive summary is now the subject of a significant amount of research; nevertheless, no universal algorithm has been developed as of yet. The summaries are produced by first comprehending what was conveyed in the article and then translating that comprehension into a form that can be comprehended by a machine. These summaries may then be read and utilized by the machine. It is analogous to the manner in which a person might summarize an article after having read it. On the other

hand, an extractive summary will pull information directly from the originating article and provide it to the reader.

## 2. Literature Review

The creation of seq2seq models [5] and the attention mechanism [6] both contributed to the consolidation of neural networks as a main tool for ATS. The attention-based Transformer architecture [7] has served as the foundation for a significant number of large-scale pre-trained language models that have achieved state-of-the-art outcomes in ATS [8,9]. Recent work in the given area have primarily focused on making minor adjustments to the designs that are already in place [10,11].

Convolutional algorithms were utilized for encoding the input text in the current neural network that was suggested by Rush et al. [12]. This network was used for the task of abstractive text summarization. A neural network with attentional feed-forward processing was used so that a summary could be produced. The pointer network was a sequence-to-sequence model that was introduced by Vinyals et al. [13]. This model was on the basis of the soft attention distribution method that was introduced by Bahdanau et al. [14]. Additionally, hybrid techniques to language modeling, neural machine translation [15], and summarization [16], [17] have been devised as a result of the pointer network. This work was extended by Rush et al. [18], which utilized a similar convolutional technique for the encoder, but the RNN was used in place of the convolutional method for the decoder to produce higher performance. By utilizing RNN, Hu et al. [19] were able to demonstrate the effective results of the Chinese dataset through the use of text summarization.

In order to perform extractive text summarization of the source, Cheng and Lapata [20] used an RNN-based encoder-decoder. Nallapati et al. [17], who studied the research utilizing the DailyMail/CNN dataset, made use of a sequence-to-sequence model. Ranzato et al. [21] choose to use an assessment matrix in place of the conventional training matrix. Some examples of evaluation matrices include ROUGE and BLEU. See et al. [22] and Jin et al. [13] incorporated pointer networks into their proposed models in order to better characterize OOV words. See et al. [22] came up with an alternative methodology to help reduce the amount of summary text that contained repeated terms.

The underlying model was constructed using reinforcement learning with an attention layer, as described in Yadav et al. [24]. In order to attain a high score with human review, Li et al. [25] utilized generative adversarial networks. In their study, Bahdanau et al. [26] put out the idea of an attention mechanism. Yang et al. [27] came up with the idea of a hierarchical attention mechanism that may be used for document classification. The researchers Nallapati et al. [17] integrated word-level and sentence-level attention, with the emphasis being placed on the sentence level.

## 3. Methodology

This proposal aims to enhance the abstractive text summarization process in the Telugu language by incorporating linguistic features, utilizing a two-staged network, and using keyword extraction with PSO. The proposed approach leverages the Telugu Books Corpus dataset, available on Kaggle, to train and evaluate the model's performance. Through the use of sophisticated linguistic features, effective feature weighting, a larger and diverse dataset, and keyword extraction with PSO, The goal is to enhance the depth and detail of the produced summaries. To execute the suggested approach, Python frameworks like NLTK, spaCy, and TensorFlow will be employed. Figure 3.1 depicts the structure of the proposed study, and Figure 3.2 showcases its procedural flow.

**1.Dataset Collection and Annotation:**

The Telugu Books Corpus dataset will be collected and annotated with linguistic features, including sentence position, sentiment analysis, Coreference resolution, and semantic role labeling. The dataset will be annotated manually or by using machine learning algorithms for automatic annotation.

**2. Preprocessing Steps:**

The Telugu Books Corpus dataset will undergo several preprocessing steps to prepare it for training and evaluation:

- **Document segmentation:** The Telugu text will be segmented into individual documents for better organization and analysis.
- **Tokenization and normalization:** The text will be tokenized into words and normalized to ensure consistency in representation.
- **Stemming:** Stemming is utilized to condense words to their fundamental or root form, enhancing matching and analytical processes.
- **Paragraph segmentation:** The text will be segmented into paragraphs to capture higher-level semantic units.
- **Stop word filtering:** Common stop words in Telugu will be filtered out to remove noise and improve the focus on important content.

Since the dataset does not have any pre-established summary column we have to add that column by creating some summary of each data by using pre-establish bert model from hugging face API for which does not work on

telugu language, to do so we first have to translate it to English then create summary then convert that summary back to telugu and save in summary column in the dataset. (this process is not a text summarization model training process) which further will be used for Keyword Extraction.

**3. Keyword Extraction with PSO:**

The PSO optimizer will be used to identify the most important words in the preprocessed text. The PSO optimizer will be used to generate a summary of the text document that focuses on the most important words.

**4. Effective Feature Weighting:**

Statistical methods will be used to assign weights to linguistic features effectively. By analyzing the frequency and importance of each feature in the training data, we can ensure that the most relevant features are given appropriate emphasis during the summarization process. Python libraries with statistical capabilities, such as SciPy or scikit-learn, will be utilized for this purpose.

**5. Training a Two-Staged Network**:

A two-staged network will be trained on the annotated Telugu Books Corpus dataset. The first stage of the network will focus on extracting important sentences from the text, while the second stage will abstract and generate concise summaries using the linguistic features and keywords extracted by the PSO optimizer. The network will be implemented using Python libraries such as Tensor Flow or PyTorch, allowing for efficient training and inference.

**6. Utilizing a Larger and Diverse Dataset:**

To enhance the performance of the two-staged network, the training dataset will be supplemented with additional Telugu text sources, augmenting the existing Telugu Books Corpus. By incorporating a larger and more diverse dataset, the model will be exposed to a wider variety of text, improving its ability to generalize and generate high-quality summaries. Python scripting and data manipulation libraries, such as pandas, will facilitate the integration and preprocessing of the extended dataset.

**7. Testing and Evaluation:**

The proposed method will be thoroughly tested and assessed.It will measure the effectiveness of the summarization model using the well-regardedROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric. Additionally, we will conduct a human evaluation to assess the informativeness,conciseness,and comprehensiveness of the generated summaries. To compare the performance of our model, we will benchmark it against state-of-the-art abstractive summarization approaches in the Telugu language, including Nallapati et al.'s approach and the pointer generator model.

**3.1 Two stage network**

The learning process is proposed to be more effective by adopting simultaneously learning of the segmentation and classification layers in an end-to-end aspect. This would involve learning the entire text from beginning to end. The process of learning would be facilitated by this. The new architecture not only makes it easier and faster for the network to train, but it also results in an increased defect detection rate. This is one of the many benefits of the new architecture.

**Pseudocode of the proposed research is as follows**

**Start**

**# Stage 1:** Training the 2SAutoencoder

**Input:** Training Data

**Process:** Initialize and train the 2Sautoencoder with the training data

**Output:** Trained 2SAutoencoder Model

**# Stage 2**: Using the 2SAutoencoder for Feature Extraction and Training Summarization Model

**Input:** New Data, Trained 2SAutoencoder Model

**Process:** Use the encoder part of the 2SAutoencoder to extract features from new data. Train a summarization model (e.g., Seq2Seq, Transformer) with the extracted features

**Output:** Trained Summarization Model

**# Making Predictions on Unseen Data**

**Input:**UnseenData,Trained 2SAutoencoderModel, Trained Summarization Model

**Process:**Use the encoder part of the 2SAutoencoder to extract features from unseen data

Use the trained summarization model to generate summaries on the extracted features

**Output:** Summaries

**End**

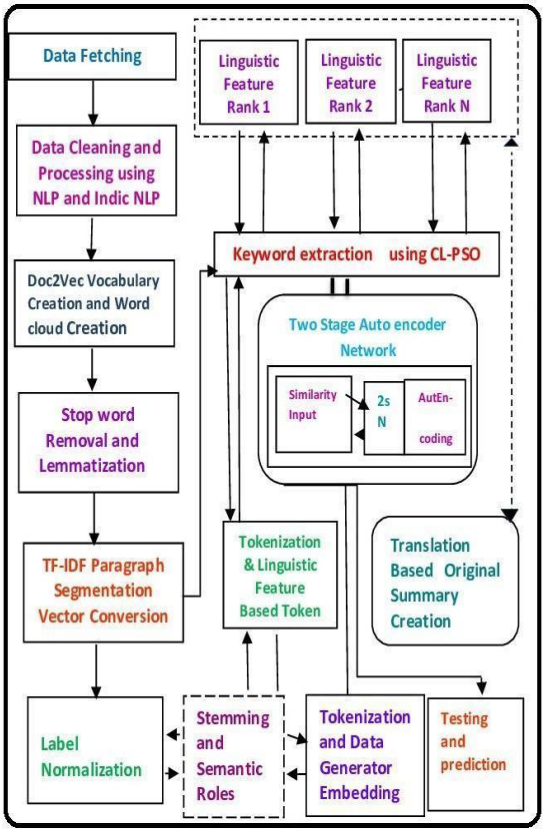


**Fig 3.1:** Block diagram representation of the proposed research

Making a modification is recommended to the gradient flow so that it takes into consideration the annotations that are made at the word level in order to achieve this goal. However, because the accurate annotations at the pixel level are difficult to get, our research is to make use of the annotations that are less precise but easy to acquire. Because the loss function is enhanced to take into consideration the ambiguities of the region-based annotations, substantially coarser annotations can be obtained while maintaining a level of simplicity that is quite easy to achieve. Additionally, a sampling method based on the frequency of usage is used for the non-defective samples, which leads to an even higher improvement in the performance of defect identification [28].

The Encoder-decoder architecture converts a variable length input sequence to a compressed representation vector, which the decoder uses to construct the output sequence. The encoder and decoder in this study are developed using Recurrent Neural Networks (RNN).To transform the input sequence X = (x1, x2,...xm) into a set of hidden representations h = (h1, h2,..., hm), the encoder employs a 3-layer stacked Long ShortTerm Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network, where each hidden state is obtained repeatedly as follows:

$$h_t=LSTM(X_t,h_{t-1}) \quad \text{Eq(1)}$$

The decoder likewise employs a three-layer LSTM network to accept encoder output and construct a variable-length sequence Y, as seen below:

$$S_t=LSTM(y_{t-1},s_{t-1},c) \quad \text{Eq(2)}$$

$$p(y_t)=softmax(y_{t-1},s_t,c) \quad \text{Eq(3)}$$

where $s_t$ is the hidden state of the LSTM decoder at time t and c is the context vector introduced afterwards. We utilize the softmax layer to compute $y_t$'s output probability and choose the word with the highest probability.

### 3.2 Particle Swarm Optimization

Particle swarm optimization (PSO), which is used in the system that has been developed, makes it simple to extract keywords and search for them according to user queries. When using the PSO algorithm, searching for keywords is easy, and the results deliver the user-searched terms that are closest. It will display some of the greatest keyword lists from around the world, which it will judge to be the top N keywords.


**Fig 3.2:** Proposed Framework

Particle swarm optimization is going to be utilized after feature extraction. It is going to be utilized for extracting the coverage of the important themes of the conversation, and it is going to maximize that coverage. In addition, in order to cover a wider range of topics, it will select the most significant keywords from each of those themes. The extraction of keywords from the conversation is required here for the keyword list that the system must recommend. These keywords should be able to cover as many of the issues brought up in the conversation as they possibly can.

## 4. Results and Discussion

### 4.1 Dataset used

This dataset signifies a development in Natural Language Processing specific to the Telugu language. The collection includes Telugu news segments primed for multi-class classification tasks. It's segmented into two key files: train

and test. Topics within the news encompass areas like business, editorial, entertainment, nation, and sport. Figure 4.1 offers a visual insight into this dataset.

Link:https://www.kaggle.com/datasets/sudalairajkumar/telugu-nlp

### 4.2 Pre-processing steps

#### 1. Data Cleaning

The process of rectifying or eliminating inaccurate, corrupted, duplicate, or improperly formatted data within a dataset is known as "data cleaning." Combining data from multiple sources increases the chances of encountering duplicates or mislabeled entries. For this dataset, the employed data cleaning techniques include checking for null values and addressing indefinite values.

| S.NO | | DATE | HEADING | BODY | TOPIC |
|---|---|---|---|---|---|
| 0 | 7771 | 19-05-2022 13:44:10 | కేశవ | హీరోగా తెలుగు సినిమాల్లోకి ఎంటీ ఇచ్చిన | Entertainment |
| 1 | 9591 | 01-08-2021 | డబ్బుల కోసం ప్రాణాల మీదకు | సినిమాల్లో యాక్షన్ స౦ఘటులు | Entertainment |
| 2 | 12622 | 05-04-2017 | దక్షిణాదిన బీజేపీ పుంజుకోవడం | దక్షిణ భారతాన్ని ఆక్రమించేందు | Nation |
| 3 | 10422 | 10-04-2017 | ఏయ్ జయప్రకాష్ అసే కేక | ఏమి రా... అబ్బి... యాడకి | Entertainment |
| ... | ... | ..... | ..... | ..... | ..... |
| 4324 | 4093 | 18-11-2017 01:17:03 | వర్గ హత్యలు ఆగాలి! | మొన్న త్రిసూరులో పటపగలు | Editorial |
| 4325 | 13874 | 22-03-2017 19:02:03 | 'ఫేస్ బుక్ పై నిషేధం..? | ఇస్లామాబాద్: దైవదూషణకు సంబంధించిన | Nation |
| 4326 | 15364 | 15-06-2017 03:18:08 | అమెరికా చట్టసభ్యుడిపై కాల్పులు | ఒక రిపబ్లికన్ సభ్యుడు, నలుగురు | Nation |
| 4327 | 17166 | 18-05-2017 19:22:27 | కోటి రూపాయల.. రదిన పాత | థానే: కోటి రూపాయల విలువైన రదిన | Nation |
| 4328 | 8138 | 15-11-2017 22:37:02 | ప్రస్తుతం ఆ కేటగిరీలో ఉన్నా | సంగీతం అనేది మహా సముద్రం. | Entertainment |

**Fig 4.1:** Sample Telugu News Dataset

#### 2. Normalization

For normalization, the original labels are considered and after label encoding the specific values are given for the labels. Encode target labels with values that range from 0 to n classes minus one. Instead of encoding the value of the input variable X, you should use this transformer to encode the target values, which are denoted by y. The label Count is represented in figure 4.2. Fetch the Resource and Dictionary of Telugu language from Indic_NLP and configure Environment for Body Sentences from data and apply resource tokens on It. Then remove Special Symbols and punctuations from the dataset.

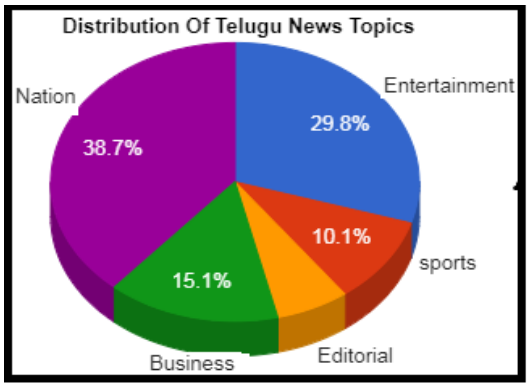


**Figure 4.2:** Representation of Label Count

#### 3.Sentence Split and Tokenization (Single Sentence)

The sentences are split and tokenized. The OG Sentence size is 67231 and the split Sentence Size is 4330. The shape after tokenization is 728061. Heap Law Graph is considered and is represented in figure 4.3. The Heaps Law in linguistics (also called Herdan's law) is an empirical law which describes the number of distinct words in a document (or set of documents) as a function of the document length (so called type-token relation). It can be formulated as

$$V\_R(n) = K*(n^{\beta}) \qquad \text{Eq(4)}$$

where V_R is the number of distinct words in an instance text of size n. K and β are free parameters determined empirically. With English text corpora, typically K is between 10 and 100, and β is between 0.4 and 0.6.

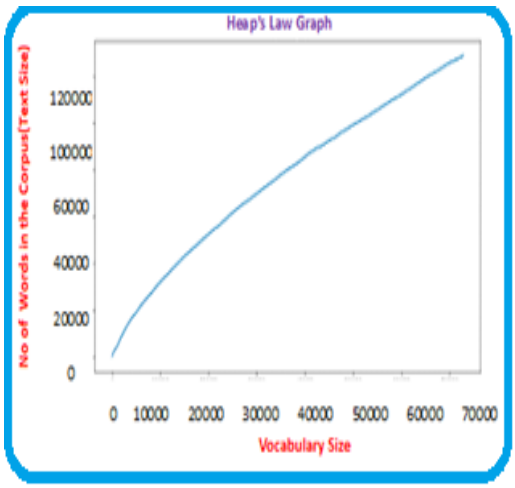


**Fig 4.3:** Heaps Law Graph

#### 4.Vocabulary Distribution Word Frequency Sorting

Insert all of the good words into an unordered set, and then loop through each word of each phrase contained in the data array. While doing so, maintain a count of the good words by determining whether or not each individual word is included in the set of good words. Then, making use of a reliable sorting algorithm, we order the elements of the array according to the number of positive comments contained inside each review that is present in the array.

Figure 4.4 gives the Vocabulary Distribution Word Frequency Sorting.

| | WORD | FREQ |
|---|---|---|
| 0 | ఈ | 9607 |
| 1 | కూడా | 4250 |
| 2 | ఆ | 3855 |
| 3 | నుంచి | 3115 |
| 4 | ఆయన | 2576 |

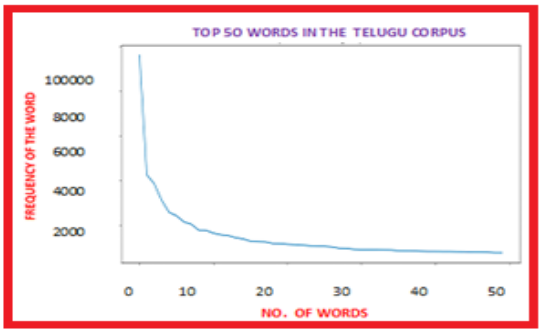


**Fig 4.4:** Vocabulary Distribution Word Frequency Sorting

### 5. Vocabulary sentence provocation

The vocabulary Sentence Provocation is performed next and then N-GRAM Analysis is performed on the dataset. In a text document, an n-gram represents a series of n sequential elements, which could be words, numbers, symbols, or punctuation. An n-gram may also be referred to as a sequence. N-gram models are helpful in many applications of text analytics where sequences of words are relevant, such as in sentiment analysis, text categorization, and text production.

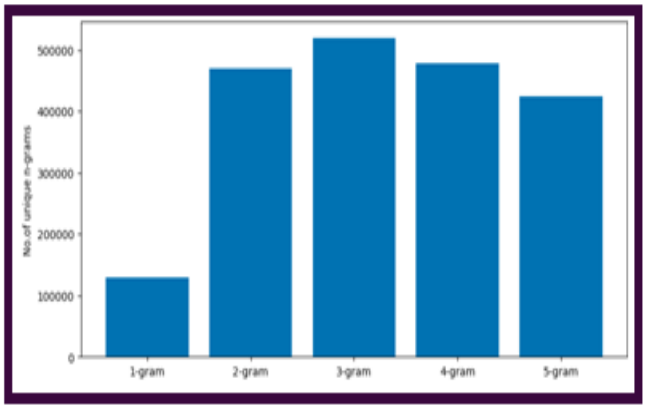


**Fig 4.5:** Graphical Representation of Unique ngrams_count

N-gram models have the capability to produce fresh text. Specifically, there are 470,062 unique bi-grams, 519,886 unique tri-grams, 477,725 unique 4-grams, and 423,869 unique 5-grams. These counts are represented as ngrams_count: [128,859, 470,062, 519,886, 477,725, 423,869]. A visual depiction of these unique ngram counts can be found in Figure 4.5.

The dataset is classified as uni Gram Classify, Bi Gram Classify, Tri Gram Classify, 4 Gram Classify and 5 Gram Classify. The data after analysis and classification is given in figure 4.6. Next the Sentence Positions are calculated and the graphical representation of the calculation of the sentence positions is given in figure 4.7. The document segmentation is 728061.

| Topic | Body_Processed |
|---|---|
| 0 0 | హీరోగా తెలుగు సినిమాల్లోకి ఎంట్రీ ఇచ్చిన నిఖిల.... |
| 1 0 | సినిమాల్లో యాక్షన్ స్టంట్లు చేసేటప్పుడు ఎక్కు... |
| 2 1 | దక్షిణ భారతాన్ని ఆక్రమించేందుకు బీజేపీ పంచెలు... |
| 3 1 | నేటి మధ్యాహ్నం 1:59కి కొంట్లోస్ షురూ!హారికేట... |
| 4 0 | "ఏమి రా అబ్బి యాడికి పోయినావు' అంటూ రాయలసీమ యా... |

**Figure4,6:** Telugu News Dataset After Analysis and Classification

| Topic | Body_Processed | Sentence_Position |
|---|---|---|
| 0 0 | హీరోగా తెలుగు సినిమాల్లోకి ఎంట్రీ ఇచ్చిన నిఖిల.... | 447 |
| 1 0 | సినిమాల్లో యాక్షన్ స్టంట్లు చేసేటప్పుడు ఎక్కు... | 182 |

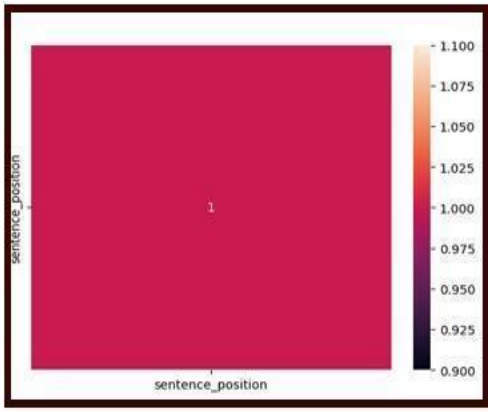


**Fig 4.7:** Calculation of the Sentence Positions

### 6. Annotation based Tokenization

Tokenization is one of the most fundamental forms of Structure Annotation, and it is required for the vast majority of the linguistic annotation kinds that FoLiA provides. Words and tokens are often concealed within other kinds of structural elements, such as sentences and paragraphs, when they are used in a document.

**7.Stemming, Paragraph Segmentation and Stop Word Filtering**

In the fields of linguistic morphology and the domain of information retrieval, stemming is the act of transforming inflected or derivative words into their core root or foundational form, often in written format. Stemming seeks to streamline a word to its primary representation. While this derived stem may not align perfectly with the word's morphological origin, the goal is for related terms to link consistently to that stem, even if it isn't a complete root by itself.

Paragraph segmentation involves splitting continuous text into distinct sections, taking into account textual structure and linguistic considerations. Stop words, listed in a stoplist or negative dictionary, are bypassed during natural language data processing because of their negligible importance..

**8. Word Cloud**

The word cloud of news article text is given in figure 4.8. These word clouds, also termed as tag clouds, visually emphasize words based on their frequency within the text. The frequency with which a word appeared in the document(s) was proportional to its size in the visual.


**Figure 4.8:** Word Cloud of Telugu News

**4.3 Keyword identification using CL-PSO Optimizer**

CL-PSO Optimization Configuration is initially done and CL-PSO Execution is completed. Optimum Keyword Index Selection using CLPSO (Single Sentence) is completed. The keywords are: ['వ్యవస్తాపకుడు', 'మిలాప్', 'కొడ్వండింగ్కు', 'హైదరాబాద్', 'గత']

Executing for All Data and then the Keywords are saved.

Font Characterization is later executed. It is the font identification that is coded into a program in order to correlate the qualities of a font with a print item such as a field or a literal.

These print items can be any type of text. The linguistic features are also calculated. The Semantic Roles are calculated as 4329.

Target Summarize data for testing using English translation-based summarization is next calculated.

The transformers are fetched and Feature Weights are calculated. The final Dataset is given in figure 4.9. After the final dataset is created after all of the preprocessing steps, the tokenization and Argument Setup is done.

**4.4 Testing and Training using two stage network**

Two stage network is used for evaluating the performances of the dataset. The new architecture not only makes it easier and faster for the network to train, but it also results in an increased defect detection rate. The shape of the dataset is (4329, 10).

The train Test Split (X-Y Split) is given as (4329, 9), (4329, 1). The Target keyword Tokenization is represented in figure 4.10.

| TOPIC | Body_Processed | Sentence_Pos | Document | Tokens | Stemmed Tokens | Keywords | Semantic_Roles |
|---|---|---|---|---|---|---|---|
| 00 | హీరోగా తెలుగు సినిమాల్లో కి ఎంట్రీ ఇచ్చిన | 447 | హీరోగా తెలుగు సినిమాల్లో కి ఎంట్రీ ఇచ్చిన | హీరోగా తెలుగు సినిమా ల్లోకి ఎంట్రీ | హీరోగా, తెలుగు, సినిమాల్లో కి,ఎంట్రీ,ఇ చ్చిన | [సుధీర్, చిన్నవా డా,స్వా మిరారా, ఎలాంటి] | Positive |
| 10 | సిని మాల్లో యాక్ష న్ స్టంట్లు చేసేట | 182 | సినిమాల్లో యాక్షన్ స్టంట్లు చేసేటప్పు డు ఎక్కు... | సినిమా ల్లో యాక్షన్ స్టంట్లు చేసేట ప్పుడు | సినిమాల్లో, యాక్షన్, స్టంట్లు ,చేసేటప్పు డు, ఎక్కు... | [చేసేట ప్పుడు,లే కుండా,, డూప్లీ,జా గ్రత్తలు,ఐ | Negative |

**Fig 4.9:** Final Dataset of TeluguNews

| | |
|---|---|
| 0 | చిన్నవాడా, ఎక్కడికి, స్నేహితుడు, హీరోగా, చిత్రాన్ని |
| 1 | లేకుండా, కనిపించేందుకు, తీసుకుంటుంటారు, సినిమాల్ల... |
| 2 | వైఖరికి, పార్టీ, సహకరిస్తోంది, దక్షిణ, సీఎం |
| 3 | లాంచ్, రాకెట్ ప్రయోగవేదికపై, నేటి, రివ్యూ |
| 4 | మెప్పించారు,పాత్రలను, ముద్రవేసి, 'ఏమి, ప్రజ్ఞాశాలి' |
| ..... | ..... |
| 4324 | ఆణిముత్యాలయిన, దేశానికి, సంఖ్యలో, మొన్న, జనగణమన |
| 4325 | సిద్ధిఖి, న్యాయమూర్తి, అంశం'పై, ఇస్లామాబాద్, సంబంధి... |
| 4326 | సీనియర్ పార్టీకి, వక్ష, ఒక, అధికారులకు |
| 4327 | వాహనంలో, జరిగింది, మహారాష్ట్రలోని, థానే, కారును |
| 4328 | నాన్నగారు ఉన్నారు, పనిచేస్తారు, "సంగీతం, వారి |
| Name: Keywords, Length:4329, dtype:object | |

**Fig 4.10:** Target keyword Tokenization

The model Summary of the two stage network is as given

**<boundmethodModel.summaryof(_main_.TwoStaged SummarizationModel object at 0x7c54ed954520>>**

And the testing and Prediction Score for accuracy is 0.90. After prediction, the predicted summary of the text is given as follows

| | |
|---|---|
| **Original Passage :** | ఇప్పుడు ఆహారం అనేది కడుపుని వదిలి చిన్న ప్రేగులలోకి ప్రవేశించినప్పుడు మిశ్రమం వంటి మిశ్రమం. ఆహారం ప్రేగులలోకి ప్రవేశించినప్పుడు, చైమ్ యొక్క ఆమ్ల స్వభావం సెక్రెటిన్ మరియు కోలిసిస్టోకినిన్ వంటి హార్మోన్ల ఉత్పత్తిని |
| **English Translation :** | **Now food is a soup like mixture when it leaves the stomach and enters the small intestine. When the food enters the intestine the acidic nature of the chyme initiates the production of hormones like secretin and cholecystokinin.....** |
| **Summary Passage :** | ఆహారం కడుపు ద్వారా చిన్న ప్రేగులలోకి ప్రవేశిస్తుంది మరియు పోషకాలను గ్రహించడానికి క్లోమం, కాలేయం..... |

The original passage and the English translation of the passage is initially given and the summary passage after prediction is given. The elaborated version of the summaries passage is given as follows:

ఆహారం కడుపు ద్వారా చిన్న ప్రేగులలోకి ప్రవేశిస్తుంది మరియు పోషకాలను గ్రహించడానికి క్లోమం, కాలేయం మరియు గోడలను ప్రేరేపించే హార్మోన్లను ఉత్పత్తి చేస్తుంది. వేగు గోడల కారణంగా పోషకాల పోషణ ఎంపిక చేయబడుతుంది.

The ROUGE Score is evaluated for the telugu dataset. The ROUGE score is a set of metrics that is typically utilized for text summarizing projects.

| Sample Output-1 | | No of Words | No of Keywords |
|---|---|---|---|
| Source _Text (Biology _Text Book) Before Prediction | నాలుక అంగిలికి వ్యతిరేకంగా నొక్కినప్పుడు, మనకు తెలిసినట్లుగా, నాలుక పనితీరులో ఇంద్రియ సంబంధమైనది మరియు రుచి మొగ్గలను కలిగి ఉంటుంది. ఈ రుచి మొగ్గలు ప్రతి �పిపింగ్‌తో చిన్న పాపిల్లే. నాలో అనేక రుచి గుచ్చి, తమైన కణాలు ఉన్నాయి. ఏదైనా ఆహార పదార్థాన్ని, నాలుకపై ఉంచినప్పుడు నేలలోని లాలాజల గ్రంథుల ద్వారా స్రవించే లాలాజలంలో కరిగిపోతుంది. నాలుకను అంగిలికి వ్యతిరేకంగా నొక్కినప్పుడు, ఆహార పదార్థం యొక్క మొగ్గ తెరవడానికి వ్యతిరేకంగా నొక్కినప్పుడు అది రుచి కణాలను చేరుకోవడానికి మరియు రుచి సంకేతాలను ప్రేరేపిస్తుంది, చివరకు మెదడులో రుచి గుర్తించబడుతుంది. | 63 | 22 |
| Corresponding English _Text (Biology _Text Book) Before Prediction | when the tongue is pressed against the palate. As we know the tongue is sensory in function and contains taste buds. These taste buds are tiny papillae with an opening on top. Within them there are several taste sensitive cells. Any food substance when placed on the tongue gets dissolved in the saliva secreted by salivary glands in the mouth. When the tongue is pressed against the palate the food substance is pressed against the opening of the taste bud letting it to reach the taste cells and triggering taste signals. Finally the taste is recognized in the brain. | 99 | 17 |
| Telugu Summary After Prediction | నాలుక అంగిలికి వ్యతిరేకంగా నొక్కినప్పుడు, మనకు తెలిసినట్లుగా, నాలుక పనితీరులో జ్ఞానేంద్రియం మరియు రుచి మొగ్గలను కలిగి ఉంటుంది. ఈ రుచి మొగ్గలు ప్రతిగాంలో ఓపెనింగ్ ఉన్న చిన్న పాపిల్లే. | 22 | 11 |
| Corresponding Translated English Text After Prediction | When the tongue presses against the palate. As we know, tongue functions as sense organ and taste buds. These taste buds are small papillae with an opening at the top. | 27 | 9 |

| Sample Output-2 | | No of Words | No of Keywords |
|---|---|---|---|
| Source _Text (Telugu _News) Before Prediction | 1868లో, జర్మనీలోని ఫ్రీబర్గ్ విశ్వవిద్యాలయంలో పాథాలజీ ప్రొఫెసర్ పాల్ లాంగర్‌హాన్స్, ప్యాంక్రియాస్ యొక్క నిర్మాణంపై పని చేస్తూ, అవయవం యొక్క సాధారణ కణజాల కణాల నుండి దాని విస్తృతమైన కణాలను గుర్తించాడు మరియు రక్త నాళాలతో సమృద్ధిగా సరఫరా చేయబడింది. వాటిని లాంగర్‌హాన్స్ ద్వీపాలు (ద్వీపాలు అంటే ద్వీపాలు) అని పిలుస్తారు, కానీ వాటి పనితీరు తెలియదు. ప్యాంక్రియాస్ పనితీరుపై చాలా మంది ఆసక్తి కలిగి ఉన్నారు మరియు ప్రయోగాత్మక జంతువు యొక్క శరీరం నుండి దాని తొలగింపు ప్రసిద్ధ మానవ వ్యాధి 'షుగర్ డయాబెటిస్' వంటి వ్యాధుల అభివృద్ధికి దారితీస్తుందని కనుగొన్నారు. రక్తంలో మరియు మూత్రంలో ఉచిత చక్కెర పరిమాణం అసాధారణంగా ఎక్కువగా ఉండే పరిస్థితి ఇది. మనిషిలో ఇది ఒక కారణం తెలియని కానీ సాక్ష్యం క్లోమం పాథాలజీకి ప్రాథాన్యం సూచించింది. | 88 | 29 |
| Corresponding English _Text (Telugu _News) Before Prediction | In 1868 Paul Langerhans, Professor of Pathology at the University of Freiburg in Germany, working on the structure of the pancreas, noted certain patches of cells quite different in appearance from the normal tissue cells of the organ and richly supplied with blood vessels. They are known Islets of Langerhans (Islets stands for islands), but their function remained unknown. Many others interested in the function of pancreas and found that its removal from the body of an experimental animal would lead to the development of diseases similar to a well-known human ailment 'sugar diabetes'. This is a condition in which the amount of free sugar in the blood and in the urine is abnormally high. It's a cause in man was unknown but evidence pointed to the pancreas as a possible role. | 132 | 35 |
| Telugu Summary After Prediction | 1868లో, పాల్ లాంగర్‌హాన్స్ యొక్క "ద్వీపాలను" కనుగొన్నాడు, అయితే వాటి పనితీరు ఇంకా తెలియలేదు. ప్రయోగాత్మక జంతువు నుండి వాటిని తొలగించడం వల్ల చక్కెర మధుమేహం అభివృద్ధి చెందడానికి దారితీసింది, ఈ పరిస్థితి రక్తం మరియు మూత్రంలో అసాధారణంగా అధిక స్థాయి ఉచిత చక్కెరను కలిగి ఉంటుంది. మానవులలో ఈ పరిస్థితిలో సాధ్యమయ్యే పాత్రను సాక్ష్యం సూచిస్తుంది. | 42 | 24 |
| Corresponding Translated English Text After Prediction | In 1868, Paul Langerhans discovered "Islets" of the pancreas, but their function remained unknown. Removing them from an experimental animal led to the development of sugar diabetes, a condition with abnormally high levels of free sugar in the blood and urine. Evidence suggests a possible role in this condition in humans. | 51 | 20 |

**Table 4.1:** Results

The purpose of these activities is to automatically construct a condensed summary of a lengthier piece of writing. ROUGE was developed to evaluate the quality of summaries that are produced by machines by comparing them to reference summaries that are produced by people. The ROUGE scores of the first dataset are given in table 4.2.

### 4.5 Comparison with dataset2 Results

The first dataset is compared with the Biology Book Dataset which is created from a biology book. The dataset Sample is represented in figure 4.11. The label Distribution of the biology book dataset are divided into 3 clusters as represented in figure 4.12.

| | Body | English_Sentence | Topic |
|---|---|---|---|
| 0 | జీవులన్నింటికీ పెరుగుదల వంటి | All living things need food to | Cluster 2 |
| 1 | జీవులు శారీరక ఉష్ణోగ్రతను | Organisms also need food to | Cluster 1 |
| 2 | మానవ శరీరంలో వివిధ కణాలకు తమ | Different cells in the human body | Cluster 2 |
| 3 | స్వయంపోషకాలు అనగానేమి? అవి | What are autonutrients? | Cluster 3 |
| 4 | అవి నేలలోని నీటిని మరియు ఖనిజ | They consume water and | Cluster 2 |
| ..... | ..... | ..... | .... |
| 2067 | భారతదేశంలో అతివేగంగా | Fastest Depleting Energy Resources | Cluster 1 |
| 2068 | పరిశ్రమలు గనుల కీమిసంహారకాలు . | Industries Mines cesticides | Cluster3 |
| 2069 | వృధాను అరికట్టడం స్థిరమైన పెరుగుదల... | curb of wastage sustainable | Cluster 3 |
| 2070 | నష్టం వాటిల్లకుండా అభివృద్ధి చేయడం | To develop without loss ,To | Cluster 3 |

**Figure 4.11:** Sample Dataset of Biology Book

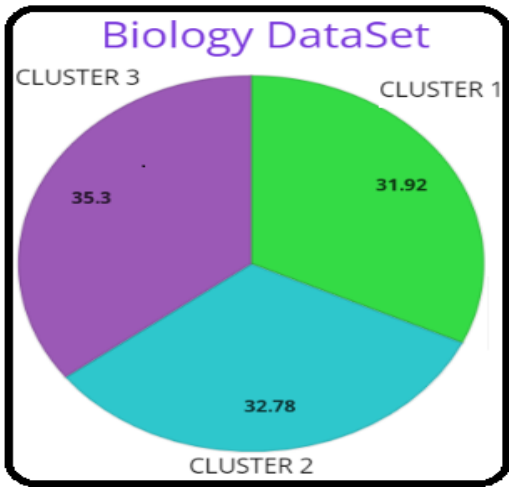


**Fig 4.12:** Biology Dataset Label Distribution

The testing and Prediction Score for accuracy for the second dataset is 0.7339 and the ROUGE scores are given in table 4.2. The comparison of the ROUGE scores with the scores obtained from the state-of-the-art methods are given in table 4.3.

| Data sets | Accuracy | ROUGE 1 | ROUGE 2 | ROUGE L |
|---|---|---|---|---|
| Telugu Newspaper | 0.9035 | 0.7975 | 0.5978 | 0.7973 |
| Biology Book | 0.7339 | 0.6172 | 0.5314 | 0.6314 |

**Table 4.2:** Comparison of Telugu News and Biology Book Datasets

| Algorithm | ROUGE 2 | ROUGE L |
|---|---|---|
| Proposed method | 0.5978 | 0.7973 |
| Lex-Rank [29] | 0.0489 | 0.1525 |
| Sum DSDR (SM) [30] | 0.0985 | 0.2602 |

**Table 4.3:** Comparison of ROUGE scores with State of the Art Method

## 5. Conclusion

In this work, the study that was offered focused on interdependent algorithms for the extraction of keywords and the summarizing of text. The efficiency with which the top-scoring keywords were discovered by the keyword extraction algorithm was comparable to that of a human. A notion with the primary purpose of "not all keywords are equal" was introduced with the assistance of a summarization algorithm that was proposed with the help of a keyword extraction algorithm. Further refinement of the Hybrid Optimization Approach may involve parameter tuning and the exploration of alternative optimization algorithms to achieve optimal performance. This study lays the groundwork for future endeavours that seek to advance language technologies, particularly in under-resourced languages like Telugu.

## References

[1] C. Zhang, “Automatic keyword extraction from documents using conditional random fields,” Journal of Computational Information Systems, vol. 4 (3), 2008, pp. 1169-1180.

[2] E. Hovy, C.-Y. Lin, “Automated text summarization and the summarize system,” in: Proceedings of a workshop held at Baltimore, ACL, 1998, pp. 197-214.

[3] Mani, M. T. Maybury, “Advances in automatic text summarization,” Vol. 293, MIT Press, 1999.

[4] G. Erkan, D. R. Radev, “Lexrank: graph-based lexical centrality as salience in text summarization,” Journal of Artificial Intelligence Research, 2004, pp. 457-479.

[5] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

[6] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in neural information processing systems*, *28*.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[8] Guo, Y., Rennard, V., Xypolopoulos, C., & Vazirgiannis, M. (2021). BERTweetFR: Domain adaptation of pre-trained language models for French tweets. *arXiv preprint arXiv:2109.10234*.

[9] Pàmies, M., Öhman, E., Kajava, K., & Tiedemann, J. (2020). LT@ Helsinki at SemEval-2020 Task 12: Multilingual or language-specific BERT?. *arXiv preprint arXiv:2008.00805*.

[10] Kawintiranon, K., & Singh, L. (2022, June). PoliBERTweet: a pre-trained language model for analyzing political content on twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7360-7367).

[11] Akomeah, K. O., Kruschwitz, U., & Ludwig, B. (2021). University of Regensburg@ PAN: Profiling Hate Speech Spreaders on Twitter. In *CLEF (Working Notes)* (pp. 2083-2089).

[12] M. Rush, S. Chopra, and J. Weston, ‘‘A neural attention model for abstractive sentence summarization,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 1–11.

[13] O. Vinyals, M. Fortunato, and N. Jaitly, ‘‘Pointer networks,’’ in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2015, pp. 2692–2700.

[14] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, ‘‘End-to-end attention-based large vocabulary speech recognition,’’ in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 4945–4949.

[15] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, ‘‘On using monolingual corpora in neural machine translation,’’ 2015, *arXiv:1503.03535*.

[16] J. Jin, P. Ji, and R. Gu, ‘‘Identifying comparative customer requirements from product online reviews for competitor analysis,’’ *Eng. Appl. Artif. Intell.*, vol. 49, pp. 61–73, Mar. 2016.

[17] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, ‘‘Abstractive text summarization using sequence-to-sequence RNNs and beyond,’’ in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, 2016, pp. 280–290.

[18] M. Rush, S. Chopra, and J. Weston, ‘‘A neural attention model for abstractive sentence summarization,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 1–11.

[19] Hu, Q. Chen, and F. Zhu, ‘‘LCSTS: A large scale Chinese short text summarization dataset,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., 2015, pp. 1967–1972.

[20] J. Cheng and M. Lapata, ‘‘Neural summarization by extracting sentences and words,’’ in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 484–494.

[21] M. A. Ranzato, S. Chopra, M. Auli, and W. Zaremba, ‘‘Sequence level training with recurrent neural networks,’’ in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–16.

[22] See, P. J. Liu, and C. D. Manning, ‘‘Get to the point: Summarization with pointer-generator networks,’’ in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, vol. 1, 2017, pp. 1073–1083.

[23] J. Jin, P. Ji, and R. Gu, ‘‘Identifying comparative customer requirements from product online reviews for competitor analysis,’’ *Eng. Appl. Artif. Intell.*, vol. 49, pp. 61–73, Mar. 2016.

[24] K. Yadav, A. Singh, M. Dhiman, R. Kaundal, A. Verma, and D. Yadav, ‘‘Extractive text summarization using deep learning approach,’’ Int. J. Inf. Technol., vol. 14, no. 5, pp. 2407–2415, 2022.

[25] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, ‘‘Adversarial learning for neural dialogue generation,’’ 2017, *arXiv:1701.06547*.

[26] D. Bahdanau, K. Cho, and Y. Bengio, ‘‘Neural machine translation by jointly learning to align and translate,’’ 2014, *arXiv:1409.0473*.

[27] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, ‘‘Hierarchical attention networks for document classification,’’ in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[28] L. Xu, S. Lv, Y. Deng, and X. Li. A Weakly Supervised Surface Defect Detection Based on Convolutional Neural Network. IEEE Access, 8:42285–42296, 2020.

[29] Erkan G, Radev DR (2004) LexRank: graph-based lexical centrality as salience in text summarization. J Artif Intell 22:457–479.

[30] He Z et al. (2012) Document summarization based on data reconstruction. Twenty-Sixth AAAI Conference on Artificial Intelligence: 620–626.