Neural Abstractive Text Summarizer for Telugu Language

GADI ANUSHA
PG scholar, Department of MCA, DNR college, Bhimavaram, Andhra Pradesh.
B.S.MURTHY
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: Abstractive Text Summarization is the process of constructing semantically relevant shorter sentences which captures the essence of the overall meaning of the source text. It is actually difficult and very time consuming for humans to summarize manually large documents of text. Much of work in abstractive text summarization is being done in English and almost no significant work has been reported in Telugu abstractive text summarization. So, we would like to propose an abstractive text summarization approach for Telugu language using Deep learning. In this paper we are proposing an abstractive text summarization Deep learning model for Telugu language. The proposed architecture is based on encoder-decoder sequential models with attention mechanism. We have applied this model on manually created dataset to generate a one sentence summary of the source text and have got good results measured qualitatively. Keywords: Deep Learning, LSTM, Telugu, Neural Networks, NLP, Summarization

Keywords: LSTM,NN,NLP,DL.

I. INTRODUCTION

Textual data is ever increasing in the current Internet age. We need some process to condense the text and simultaneously preserving the meaning of the source text. Text summarization is creating a short, accurate and semantically relevant summary of a given text. It would help in easy and fast retrieval of information. Text summarization can be classified into two categories.

- Extractive Text Summarization methods form summaries by copying from the parts of the source text by taking some measure of importance on the words of the source text and then joining those sentences together to form a summary of the source text.
- Abstractive text summarization methods create new semantically relevant phrases, it can also form summaries by rephrasing or by using the words that were not in the source text. Abstractive methods are actually harder .

For an accurate and semantically relevant Msummaries, the model is expected to comprehend the meaning of the text and then try to express that understanding using the relevant words and phrases. So, abstractive models can have capabilities like generalization, paraphrasing. Significant work is being focused on extractive text summarization methods and especially with English as the source language.

There is no reported work for Telugu abstractive Text summarization using Deep learning models and also there are no available datasets for Telugu text summarization. Our goal is to build a model such that when given the telugu news article it should output semantically relevant sentence as the summary/title sentence for the corresponding telugu article. We have proposed a Deep learning model using encoder-decoder architecture and we have achieved good results measured qualitatively. We have manually created the Dataset because of the fact there are no available datasets.

Training Dataset has been created from the Telugu News websites by taking the headline as the summary and the main content as the source text and we have created a dataset with 2000 telugu news articles with their corresponding summaries which are taken as the headline of the respective article. We have created the dataset in such a way that the articles belonging to the different domains i.e, Politics, Entertainment, Sports, Business, National are more or less equally distributed to maintain a balance to the dataset. To create word embeddings for the telugu words, we have made use of wordembeddings by FastText, which has created word embeddings for nearly 157 languages with each word-embedding of 300 dimensions.[1]

With the growing amount of data in the world, interest in the field of automatic summarization generation has been widely increasing. Text summarization involves reducing a text file into a passage or paragraph that conveys the main meaning of the text. Searching for important information from a large text file is a very difficult job for the users thus automatically extracting the important information or summary of the text file. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs, etc. from the original document and concatenating them into a shorter form.

The importance of sentences is decided based on the statistical and linguistic features of sentences. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. The extractive summarization systems are typically based on techniques for sentence extraction and aim to cover the set of sentences that are most important for the overall understanding of a given document. Abstractive methods create an internal semantic representation to create a summary that is closer to what a human might generate.

Such a summary might contain words not explicitly present in the original. With the rapid growth of the World Wide Web (internet), information overload is becoming a problem for an increasingly large number of people. Automatic summarization can be an indispensable solution to reduce the information overload problem on the web.

In the era of information overload, the ability to distill large volumes of text into concise and informative summaries is becoming increasingly important. Text summarization, the process of condensing lengthy documents into shorter versions while retaining key information, has garnered significant attention in natural language processing (NLP). Abstractive summarization, a technique that involves generating summaries using natural language generation (NLG) techniques, offers a more flexible and human-like approach compared to extractive methods that simply select and rearrange existing sentences. In this context, the development of neural abstractive text summarizers has emerged as a cutting-edge solution, particularly for languages with rich morphology and syntax like Telugu.

Telugu, one of the Dravidian languages predominantly spoken in the Indian states of Andhra Pradesh and Telangana, presents unique challenges for text summarization due to its complex grammar, extensive vocabulary, and diverse writing styles. Traditional approaches to Telugu text summarization have often relied on rule-based or statistical methods, which struggle to capture the nuances and semantic intricacies of the language. However, recent advancements in deep learning and neural network architectures have paved the way for more effective and context-aware summarization models tailored to Telugu.

II. LITERATURE SURVEY

[1] Edouard Grave et al. (2018) Grave et al. proposed a large-scale multilingual word embedding framework by training high-quality word vectors for 157 languages using Wikipedia and Common Crawl data. Their approach is based on an extended version of the fastText model, which incorporates subword information to generate robust word representations, especially for morphologically rich and low-resource languages. To evaluate their work, they introduced new word analogy datasets for French, Hindi, and Polish and assessed performance across 10 languages, achieving results superior to previously released models. This work is significant for expanding NLP capabilities beyond English, providing useful pretrained embeddings for global language applications.

[2] Rush, A. M., Chopra, S., & Weston, J. (2015) In their work on abstractive sentence summarization, Rush et al. introduced a neural attention-based model capable of generating summaries word-by-word based on input sentences. Unlike extractive approaches that copy sections of text, this model is trained end-to-end to generate condensed, paraphrased summaries using a soft alignment attention mechanism. Evaluated on the DUC-2004 dataset, the model demonstrated considerable performance improvements over traditional summarization methods. This work helped shift the focus in summarization research toward data-driven, generative techniques using neural networks, laying the foundation for future abstractive summarization models.

[3] Konstantin Lopyrev (2015) Lopyrev explored headline generation using an encoder-decoder framework with Long Short-Term Memory (LSTM) units and attention mechanisms. Applied to news articles, the model was effective in producing concise and contextually accurate headlines. The study also analyzed how the model learned attention weights, revealing insights into which words influenced the output most. Interestingly, a simplified attention mechanism was found to outperform more complex versions in certain cases. This research reinforced the potential of recurrent neural networks in text generation tasks, particularly in capturing semantic relevance for summarization.

[4] Bahdanau, D., Cho, K., & Bengio, Y. (2014) This influential paper introduced a major advancement in neural machine translation (NMT) by proposing an attention mechanism to replace the limitations of fixed-length encoding in the encoder-decoder architecture. Instead of compressing the entire input into a single vector, the model learns to (soft-)align with relevant parts of the source sentence dynamically during decoding. The proposed method achieved translation performance on par with state-of-the-art systems and offered better handling of longer sentences. The attention-based alignment approach pioneered in this paper has since become a core component in various NLP tasks beyond translation.

[5] Allahyari et al. (2017) Allahyari and colleagues conducted a comprehensive survey on text summarization techniques, categorizing them into extractive and abstractive approaches. The study reviewed classical and modern methods including statistical, graph-based, and deep learning-based summarization techniques. It also highlighted the growing importance of neural models in handling the increasing volume of textual data. The survey provides valuable insights into the strengths and limitations of different summarization algorithms, serving as a foundational reference for researchers and developers working in the domain of automatic text summarization.

[6] Alquliti & Abdul Ghani (2019) Alquliti and Abdul Ghani proposed CNN-ATS, a convolutional neural network-based system for automatic text summarization. By representing sentences as text matrices, the model assesses and extracts the most informative sentences from input documents. The authors experimented with 26 CNN configurations and validated the performance on the DUC 2002 dataset using ROUGE metrics. Their findings demonstrated that deeper CNN layers improved summary quality, making this approach competitive with existing summarization systems. The study highlighted how deep learning, particularly CNNs, can be adapted for non-visual tasks like text summarization with effective results.

[7] Bahdanau, D., Cho, K., & Bengio, Y. (2014) In a separate work on neural machine translation, Bahdanau et al. redefined the encoder-decoder framework by introducing a joint learning mechanism for alignment and translation. By enabling the model to attend selectively to parts of the input sentence during translation, the system overcame the bottleneck caused by fixed-length representations. This soft attention mechanism improved translation quality, especially for longer sequences, and inspired widespread adoption in other tasks like summarization and question answering. The paper's core contribution lies in its elegant solution to dynamic alignment, which became a key innovation in the evolution of attention-based models.

III. PROPOSED METHOD

In propose work we have used Telugu Summary dataset to train different deep learning algorithms such as LSTM and BI-LSTM to predict summary. LSTM algorithm is suitable for textual analysis and for summary generation.

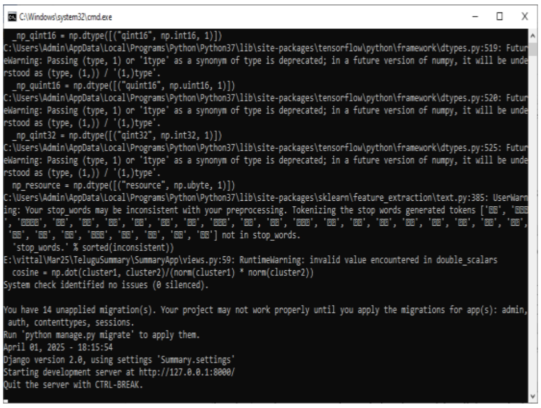
In summary generation, LSTMs (Long Short-Term Memory) are a type of recurrent neural network (RNN) that excel at processing sequential data, like text, and capturing long-term dependencies, making them suitable for tasks like understanding context and generating coherent summaries.

To implement this project we have designed following modules

1) User Login: user can login to system using username and password as ‘admin and admin’.
2) Load & Process Dataset: after login user will run this module to load and process dataset and then generate TFIDF vector which will be input to LSTM algorithm to process sequential data and tp describe summary
3) Train LSTM & BILSTM Encoder: processed data will be input to LSTM and BILSTM to algorithm to train encoder and decoder where LSTM encoder will get trained on Telugu data and then Decoder will get trained on Telugu Summary and both this model will be utilize to generate summary from new test data and then calculate accuracy on test data
4) Text Summarization: using this module user can upload test document and then best performing encoder and decoder model will be applied to generate summary

IV. Results Analysis

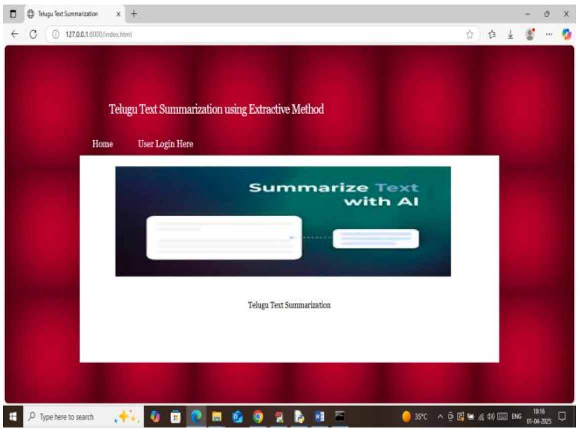
To run project install Python 3.7.2 and then install all packages given in requirements.txt file. Now double click on ‘run.bat’ file to start python web server and then will get below page



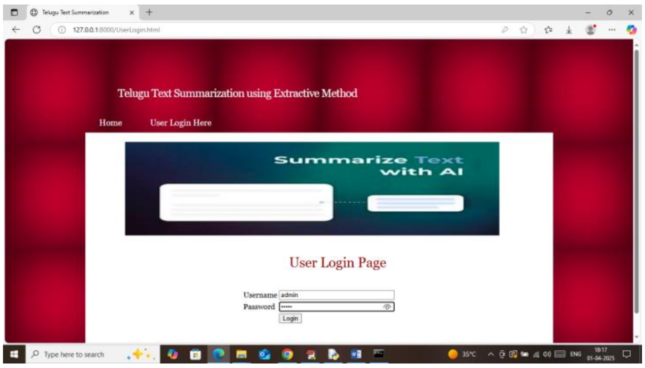
```
  _np_qint16 = np.dtype([("qint16", np.int16, 1)])
C:\Users\Admin\AppData\Local\Programs\Python\Python37\lib\site-packages\tensorflow\python\framework\dtypes.py:519: Futur
eWarning: Passing (type, 1) or '1type' as a synonym of type is deprecated; in a future version of numpy, it will be unde
rstood as (type, (1,)) / '(1,)type'.
  _np_quint16 = np.dtype([("quint16", np.uint16, 1)])
C:\Users\Admin\AppData\Local\Programs\Python\Python37\lib\site-packages\tensorflow\python\framework\dtypes.py:520: Futur
eWarning: Passing (type, 1) or '1type' as a synonym of type is deprecated; in a future version of numpy, it will be unde
rstood as (type, (1,)) / '(1,)type'.
  _np_qint32 = np.dtype([("qint32", np.int32, 1)])
C:\Users\Admin\AppData\Local\Programs\Python\Python37\lib\site-packages\tensorflow\python\framework\dtypes.py:525: Futur
eWarning: Passing (type, 1) or '1type' as a synonym of type is deprecated; in a future version of numpy, it will be unde
rstood as (type, (1,)) / '(1,)type'.
  np_resource = np.dtype([("resource", np.ubyte, 1)])
C:\Users\Admin\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\feature_extraction\text.py:385: UserWarn
ing: Your stop_words may be inconsistent with your preprocessing. Tokenizing the stop words generated tokens ['□□', '□□□
', '□□□□', '□□', '□□', '□□', '□□', '□□', '□□', '□□□', '□□', '□□', '□□□', '□□', '□□', '□□', '□□', '□□', '□□', '□□', '□□',
 '□□', '□□', '□□', '□□□', '□□', '□□', '□□', '□□'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
E:\vittal\Mar25\TeluguSummary\SummaryApp\views.py:59: RuntimeWarning: invalid value encountered in double_scalars
  cosine = np.dot(cluster1, cluster2)/(norm(cluster1) * norm(cluster2))
System check identified no issues (0 silenced).

You have 14 unapplied migration(s). Your project may not work properly until you apply the migrations for app(s): admin,
 auth, contenttypes, sessions.
Run 'python manage.py migrate' to apply them.
April 01, 2025 - 18:15:54
Django version 2.0, using settings 'Summary.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

In above screen python web server started and now open browser and enter URL as http://127.0.0.1:8000/index.htmla and then press enter key to get below page

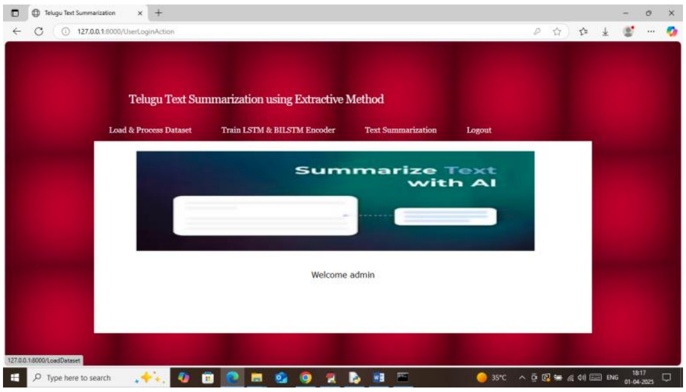

Telugu Text Summarization
127.0.0.1:8000/index.html
Telugu Text Summarization using Extractive Method
Home
User Login Here
Summarize Text with AI
Telugu Text Summarization
Type here to search

In above screen click on ‘User Login Here’ link to get below page


Telugu Text Summarization
127.0.0.1:8000/UserLogin.html
Telugu Text Summarization using Extractive Method
Home
User Login Here
Summarize Text with AI
User Login Page
Login
Type here to search

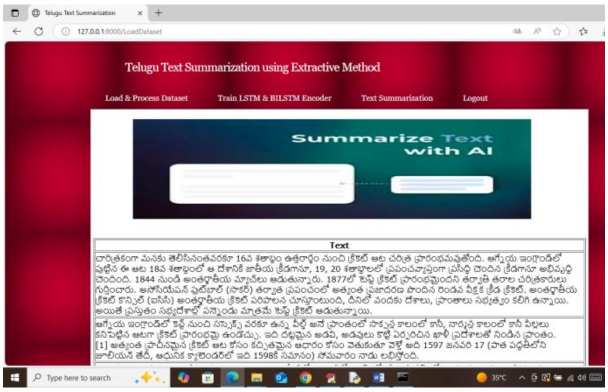
Username admin
Password •••••

In above screen user is login and after login will get below page


Telugu Text Summarization
127.0.0.1:8000/UserLoginAction
Telugu Text Summarization using Extractive Method
Load & Process Dataset
Train LSTM & BILSTM Encoder
Text Summarization
Logout
Summarize Text with AI
127.0.0.1:8000/LoadDataset
Type here to search

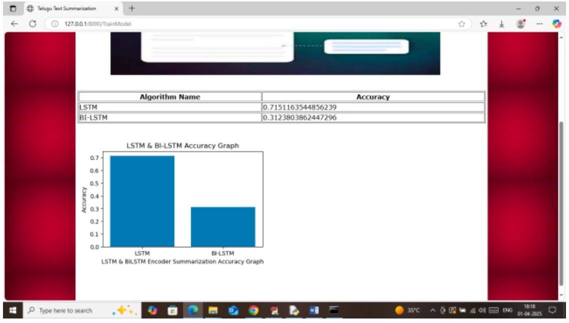
Welcome admin

In above screen click on ‘Load & Process Dataset’ link to load dataset and then will get below page


Telugu Text Summarization
127.0.0.1:8000/LoadDataset
Telugu Text Summarization using Extractive Method
Load & Process Dataset
Train LSTM & BILSTM Encoder
Text Summarization
Logout
Summarize Text with AI
Type here to search
35°C

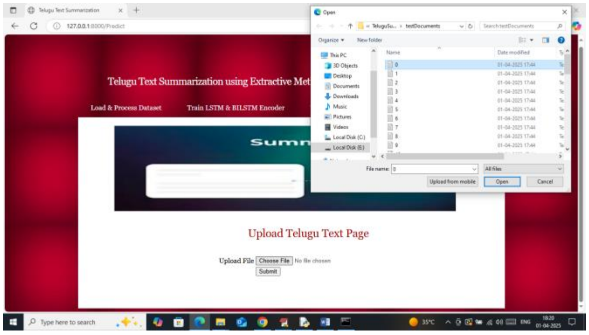
Text
చారిత్రకంగా మనకు తెలిసినంతవరకూ 16వ శతాబ్దం ఉత్తరార్థం నుంచి క్రికెట్ ఆట చరిత్ర ప్రారంభమవుతోంది. ఆగ్నేయ ఇంగ్లాండ్‌లో పుట్టిన ఈ ఆట 18వ శతాబ్దంలో ఆ దేశానికి జాతీయ క్రీడగానూ, 19, 20 శతాబ్దాలలో ప్రపంచవ్యాప్తంగా ప్రసిద్ధి చెందిన క్రీడగానూ అభివృద్ధి చెందింది. 1844 నుండి అంతర్జాతీయ మ్యాచ్‌లు ఆడుతున్నారు. 1877లో టెస్ట్ క్రికెట్ ప్రారంభమైందని తర్వాతి తరాల చరిత్రకారులు గుర్తించారు. అసోసియేషన్ ఫుట్‌బాల్ (సాకర్) తర్వాత ప్రపంచంలో అత్యంత ప్రజాదరణ పొందిన రెండవ వీక్షక క్రీడ క్రికెట్. అంతర్జాతీయ క్రికెట్ కౌన్సిల్ (ఐసీసీ) అంతర్జాతీయ క్రికెట్ పరిపాలన చూస్తూంటుంది, దీనిలో వందకు దేశాలు, ప్రాంతాలు సభ్యత్వం కలిగి ఉన్నాయి. అయితే ప్రస్తుతం సభ్యదేశాల్లో పన్నెండు మాత్రమే టెస్ట్ క్రికెట్ ఆడుతున్నాయి.
ఆగ్నేయ ఇంగ్లాండ్‌లో కెన్ట్ నుంచి ససెక్స్ వరకూ ఉన్న వీల్డ్ అనే ప్రాంతంలో సాక్సన్ కాలంలో కానీ, నార్మన్ కాలంలో కానీ పిల్లలు కనిపెట్టిన ఆటగా క్రికెట్ ప్రారంభమై ఉండొచ్చు. ఇది దట్టమైన అడవి, అడవులు కొట్టి ఏర్పరిచిన ఖాళీ ప్రదేశాలతో నిండిన ప్రాంతం. [1] అత్యంత ప్రాచీనమైన క్రికెట్ ఆట కోసం కచ్చితమైన ఆధారం కోసం వెతుకుతూ వెళ్తే అది 1597 జనవరి 17 (పాత పద్ధతిలోని జూలియన్ తేదీ, ఆధునిక క్యాలెండర్‌లో ఇది 1598కి సమానం) సోమవారం నాడు లభిస్తోంది.

In above screen dataset loaded and processed and now click on 'Train LSTM & BILSTM Encoder' link to train algorithms and then will get below page

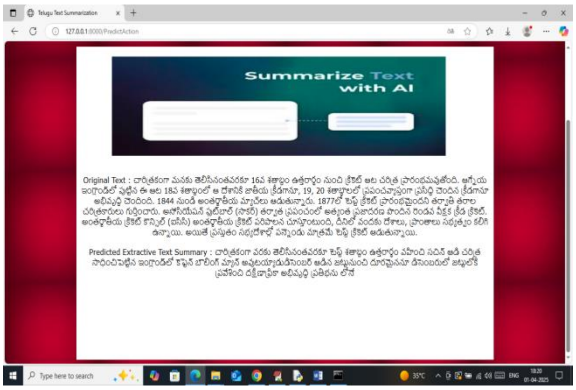

Telugu Text Summarization
127.0.0.1:8000/TrainModel
LSTM & BI-LSTM Accuracy Graph
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
LSTM
BI-LSTM
LSTM & BILSTM Encoder Summarization Accuracy Graph
Type here to search
35°C
ENG

Algorithm Name	Accuracy
LSTM	0.7151163544856239
BI-LSTM	0.3123803862447296

In above screen in table format can see LSTM got 71% accuracy and BILSTM got 31% accuracy and can see performance of both algorithms in graph format where x-axis represents algorithm names and y-axis represents accuracy. Now click on 'Text Summarization' link to get below page



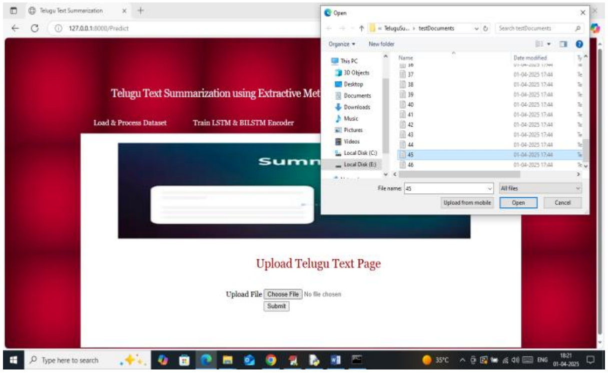
In above screen selecting and uploading 'Telugu Text Document' and then click on 'Submit' button to load document and then will get below summary output



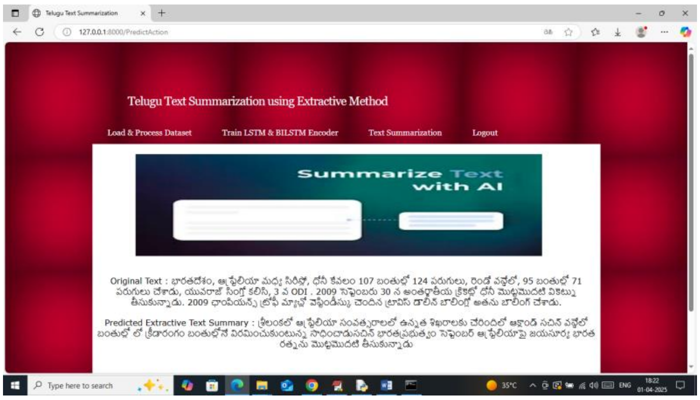
Original Text : చారిత్రకంగా మనకు తెలిసినంతవరకూ 16వ శతాబ్దం ఉత్తరార్థం నుంచి క్రికెట్ ఆట చరిత్ర ప్రారంభమవుతోంది. ఆగ్నేయ ఇంగ్లాండ్‌లో పుట్టిన ఈ ఆట 18వ శతాబ్దంలో ఆ దేశానికి జాతీయ క్రీడగానూ, 19, 20 శతాబ్దాలలో ప్రపంచవ్యాప్తంగా ప్రసిద్ధి చెందిన క్రీడగానూ అభివృద్ధి చెందింది. 1844 నుండి అంతర్జాతీయ మ్యాచ్‌లు ఆడుతున్నారు. 1877లో టెస్ట్ క్రికెట్ ప్రారంభమైందని తర్వాతి తరాల చరిత్రకారులు గుర్తించారు. అసోసియేషన్ ఫుట్‌బాల్ (సాకర్) తర్వాత ప్రపంచంలో అత్యంత ప్రజాదరణ పొందిన రెండవ వీక్షక క్రీడ క్రికెట్. అంతర్జాతీయ క్రికెట్ కౌన్సిల్ (ఐసీసీ) అంతర్జాతీయ క్రికెట్ పరిపాలన చూస్తూంటుంది, దీనిలో వందకు దేశాలు, ప్రాంతాలు సభ్యత్వం కలిగి ఉన్నాయి. అయితే ప్రస్తుతం సభ్యదేశాల్లో పన్నెండు మాత్రమే టెస్ట్ క్రికెట్ ఆడుతున్నాయి.

Predicted Extractive Text Summary : చారిత్రకంగా వరకు తెలిసినంతవరకూ టెస్ట్ శతాబ్దం ఉత్తరార్థం వహించి సచిన్ ఆడి చరిత్ర సాధించిపెట్టిన ఇంగ్లాండ్‌లో కెప్టెన్ బౌలింగ్ మ్యాన్ అవుటయ్యాడుడిసెంబర్ ఆడిన జట్టునుంచి దూరమైననూ డిసెంబరులో జట్టులోకి ప్రవేశించి దక్షిణాఫ్రికా అభివృద్ధి ప్రతిభను లోనే

In above screen in first paragraph can see uploaded original TEXT and then in second paragraph can see generated summary. Similarly you can upload any document and generate summary. Below is another example



In above screen uploading another document and below is the summary output


Telugu Text Summarization
127.0.0.1:8000/PredictAction
Telugu Text Summarization using Extractive Method
Load & Process Dataset
Train LSTM & BILSTM Encoder
Text Summarization
Logout
Summarize Text with AI
Type here to search
35°C
ENG

Original Text : భారతదేశం, ఆస్ట్రేలియా మధ్య సిరీస్లో, ధోనీ కేవలం 107 బంతుల్లో 124 పరుగులు, రెండో వన్డేలో, 95 బంతుల్లో 71 పరుగులు చేశాడు, యువరాజ్ సింగ్తో కలిసి, 3 వ ODI . 2009 సెప్టెంబరు 30 న అంతర్జాతీయ క్రికెట్లో ధోనీ మొట్టమొదటి వికెట్ను తీసుకున్నాడు. 2009 ఛాంపియన్స్ ట్రోఫీ మ్యాచ్లో వెస్టిండీస్కు చెందిన ట్రావిస్ డౌలిన్ బౌలింగ్లో అతను బౌలింగ్ చేశాడు.

Predicted Extractive Text Summary : శ్రీలంకలో ఆస్ట్రేలియా సంవత్సరాలలో ఉన్నత శిఖరాలకు చేరిందిలో ఆక్లాండ్ సచిన్ వన్డేలో బంతుల్లో లో క్రీడారంగం బంతుల్లోనే విరమించుకుంటున్న సాధించాడుసచిన్ భారతప్రభుత్వం సెప్టెంబర్ ఆస్ట్రేలియాపై జయసూర్య భారత రత్నను మొట్టమొదటి తీసుకున్నాడు

In above screen can see original uploaded text and predicted summary.

V. CONCLUSION

The proposed deep learning-based summarization system successfully generates concise Telugu text summaries using LSTM and Bi-LSTM models. The results indicate that LSTM performs better, achieving a higher accuracy rate. The system is user-friendly, allowing seamless text input and summary generation through a web-based interface. Future improvements may involve experimenting with transformer-based models like BERT or GPT for enhanced summarization quality.

REFERENCES

1. Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, Tomas Mikolov. "Learning Word Vectors for 157 Languages."arXiv:1802.06893v2 [cs.CL]
2. Alexander M. Rush, Sumit Chopra, Jason Weston, "A Neural Attention Model for Abstractive Sentence Summarization." arXiv:1509.00685v2 [cs.CL]
3. Konstantin Lopyrev, "Generating News Headlines with Recurrent Neural Networks" arXiv:1512.01712v1[cs.CL]
4. Dzmitry Bahdanau, Kyughyun Cho, Yoshua Bengio Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473v7 [cs.CL]
5. Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) Text summarization techniques: a brief survey. arXiv. https://doi.org/10.1177/1010428317692226
6. Alquliti WH, Abdul Ghani NB (2019) Convolutional neural network based for automatic text summarization. Int J Advan Comput Sci Applic (IJACSA) 10(4). https://doi.org/10.14569/IJACSA.2019.0100424
7. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. ArXiv:1409–0473
8. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166. https://doi.org/10.1109/72.279181