iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60113

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Deep Learning Based Hybrid Technique for Improved Summarization of Abstractive Text

Lanka Shriya[1], Mekala Kamala[2]
[1]*UG Student*, [2]*Assistant Professor, Department of Computer Science & Engineering, CMR College of Engineering & Technology, Hyderabad, India*

Abstract: ***From various source sentences, construct summary sentences by merging facts. The process of preserving information material and overall meaning while condensing it into a shorter representation is known as abstractive text summarization, or ATS. It takes a lot of effort and time for humans to manually summarise large textual publications. In this paper, we present an ATS framework (ATSDL) based on (Long Short-Term Memory-Convolutional Neural Network) LSTM- CNN that can generate new sentences by investigating semantic phrases, which are finer- grained fragments than sentences. ATSDL, in contrast to current abstraction- based methods, consists of two primary phases: the first extracts phrases from source sentences, and the second uses deep learning to produce text summaries. Experimental results on CNN and Daily Mail datasets show that our ATSDL framework outperforms the state-of-the-art models in terms of both syntactic structure and semantics, and achieves competitive results on manual linguistic quality evaluation. In this application hybrid model is giving better performance over other state of art techniques.***

Keywords: ***Deep learning Relation extraction, Abstractive text summarization, Text mining.***

I. INTRODUCTION

A task generating a short summary the Text summarization (TS) consisting of a some sentences that capture salient ideas of a text [16]. The objectives of the paper are,

1) Develop a deep learning framework for abstractive text summarization that leverages the strengths of LSTM and CNN architectures.
2) Train the model on large-scale datasets of text documents and human-generated summaries to learn to generate informative and fluent summaries.
3) Conduct qualitative analysis to assess the linguistic quality, coherence, and in formativeness of the generated summaries.

In Natural language understanding overcomingthis task is an important step. In a very short time also good summary and concise can help bitterly to humans comprehend the text content. Into two various categories the text summarization can be broadly classified Based on previous studies. [6] One uses traditional methods and is called Extractive Text Summarization (ETS).

Its methods for creating summaries involve selecting significant sections from the original text and putting them together to create a logical summary. The alternative is Abstractive TextSummarization (ATS), which creates summaries from scratch without limiting itself to specific phrases from the source text by producing more qualitatively human-written sentences. Initially,ETS models were suggested as a way to extract andcondense the essential semantic data included inthe source text. Sequence-to-sequence models, or ATS models, have been proposed more recently asa result of advances in computer performance and the growth of deep learning theory. The long short-term memory (LSTM) encoder-decoder model, which was proposed in [11], is a particularly important model within the sequence- to-sequence model framework for our purpose. It has achieved state-of-the-art performance inmachine translation, a natural language problem. Though TS and machine translation share manysimilarities, they are completely different problems. The target output sequence in machine translation is roughly the same length as the source text. Nonetheless, the output sequence in text summaryis usually quite brief and is not highly dependent onthe length of the source text.

A major difficulty in text summarization is to best compress the source material in a lossy way while maintaining the main ideas, whereas machine translation aims for a lossless translation. While near-word-for-word alignment between source and target is widely recognised in machine translation, it is less evident in text summarization. Even though the current TS models have produced excellent results in numerous well-known datasets, not all issues have been overcome by these models. While syntactic structure and semantics are twocrucial aspects to consider when assessing TS models, each type of model concentrates on just one of them.

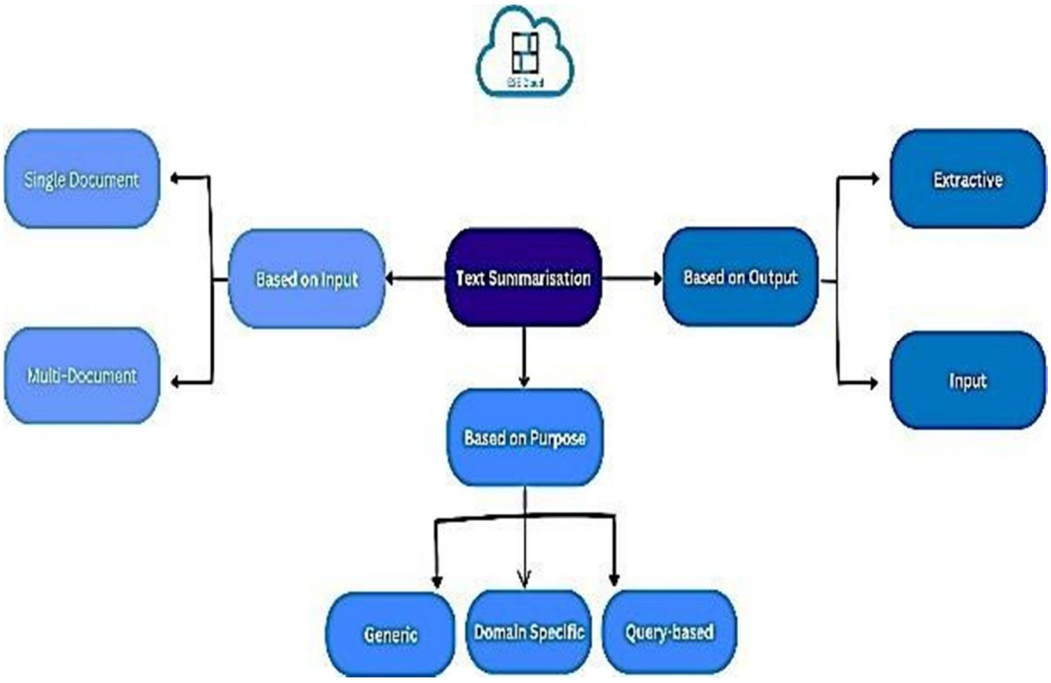


Figure 1: Different types of summarizations of text[16]

Text summarization has basic three types asmentioned below,

1) Type of Input data
2) Type of Output data
3) Based on the Purpose

Above three are the main categories and subtopicsare shown in figure.

The present ETS models are summaries with a few isolated sentences that are coarsely textured. One benefit of using ETS models is that the summary' sentences have to follow the syntacticstructure's rules. The potential for syntactic disarray in the summary is an inherent drawback of ETS models. This drawback arises from the fact that the summaries' neighbouring sentences aren't always adjacent in the source material. The current iterations of ATS models are detailed and incorporate semantic objects (words, sentences, etc.) in their summaries. ATS models have the benefit of inclusive semantics as, during training, they identify word collocations and produce a listof keywords based on those collocations.

We suggest an ATS framework, called ATSDL, based on LSTM-CNN to address the issues. The primary contributions of this work areas follows:

a) To boost TS performance, we integrate CNN and LSTM, an LSTM model that was first created for machine translation, with summarization. With existing ATS and ETS models, ATSDL takes syntactic structure and semantics into account. The new model will produce a series of words after training. This series is a natural sentence-based summarization of the text.
b) Using phrase location information, we address the primary issue of rare words and produce more naturally occurring sentences.
c) The experiment's findings show that, on two distinct datasets, ATSDL performs better than modern abstractive and extractive summarization algorithms.

II. LITERATURE SURVEY

Distant supervision naturally aims to significantly reduce the high cost of data annotation, it is still highly constrained by the quality of training data. Based on the well-known MIML framework, we create architecture in this paper called MIML-sort (Multi-instance Multi- label Learning with Sorting Strategies). We provide three ranking-based sample selection techniques based on MIML-sort that we use to find relation extractors from a portion of the training data. The KBP (Knowledge Base Propagation) corpus, a sizable and noisy benchmark dataset for remote supervision, is used for experiment setup. Incomparison to earlier research, the suggested techniques yield significantly superior outcomes. Moreover, the combination of the three approaches produces the best F1 on the official testing set,maximising F1 from 27.3% to 29.98%. [1]

We suggest an abstraction-based multi- document summarising system that can create new sentences by examining noun/verb phrases, which are more fine-grained syntactic units than sentences. Unlike other abstraction-based methods, ours builds a pool of facts and concepts first, represented by phrases taken from the input documents. Subsequently, in order to maximisephrase salience and meet sentence construction limitations, informative phrases are chosen and combined to create new sentences. To find the global optimal solution for a summary, we simultaneously perform phrase selection and merger using integer linear optimisation. [2]

In extractive query-focused summarization, the two primary goals are query relevance rating and sentence saliency ranking. Historically, supervised summarization systems have frequently completed the two tasks independently. Nevertheless, neither of the two rankers could be trained properly since reference summaries represent the trade-off between saliency and relevance when used as supervision.

In this paper, a novel summarising system named AttSum is proposed that addresses both goals simultaneously. Together with the document cluster, it automatically learns distributed representations for sentences. When a question is posed, it uses the attention mechanism to mimic thecareful observation of human behaviour. Benchmark datasets for DUC query-focused summarization are utilised for extensive experiments. In the absence of any custom features,AttSum reaches competitive results. Additionally, we note that the words identified as query-focused do address the requirement for the question. [3]

Text document representation in information retrieval, machine learning, and text mining typically uses variations of sparse Bag of Words (sBoW) vectors, such as TF-IDF [1]. Word-level synonymy and polysemy cannot be captured by sBoW style representations, despite their simplicity and intuitiveness, due to their intrinsic over- sparsity. Over fitting and decreased generalisation accuracy result from this. In this research, we present an unsupervised approach to discover enhanced sBoW document features: Dense Cohort of Terms (dCoT). By deleting and recreating random subsets of words from the unlabelled corpus, dCoT directly models absent words. In this way, the high dimensional sparse sBoW vectors are mapped into a low dimensional dense representation, and dCoT learns to rebuild frequent words from co-occurring infrequent words. We demonstrate that the reconstruction can be solved for in closed form and that the feature removal can be marginalised out. Using a number of benchmark datasets, we provide empirical evidence that dCoT features greatly increase classification accuracy for a variety of document classification tasks. [4]

Conventional methods of extractive summarization mostly depend on attributes that have been designed by humans. In this work,we offer a data-driven method based on continuous sentence characteristics and neural networks. We create a general framework consisting of an attention-based extractor and a hierarchical document encoder for single-document summarization. We are able to create many classes of word or sentence extraction summarization models thanks to this design. Using extensive datasets with hundreds of thousands of document- summary pairings, we train our models1. Tested on two summarization datasets, our models achieve performance on par with the state of the art even in the absence of language annotation. [5]

A conditional recurrent neural network (RNN) is presented here, capable of producing an overview of an input text. A unique convolutional attention-based encoder provides the conditioning, making ensuring the decoder concentrates on the right input words at every stage of generation. Our model is simple to train end-to-end on big data sets and only depends on learned features. Our tests demonstrate that the model performs competitively on the DUC-2004 shared task and achieves significant advantages over the newly proposed state-of-the-art technique on the Gigaword corpus. [6]

One of document summarization's numerous documented uses is automatic headline generation. We provide a sequence-prediction method in this work to understand how news editors title theirpieces. The approach that has been presented framesthe issue as a discrete optimisation effort within a feature-ich domain. Using dynamic programming, the global optimum in this space canbe obtained in polynomial time. We use a large corpus of financial news to train and test our model, and we assess its performance against several baselines using common measures from the document summarization area and some new metrics suggested in this study. We also use human evaluation to determine how are readable and formative the generated titles.[7]

III. PROPOSED METHOD

Text summaries can help in understanding any topic easily without reading whole content as it will give short description of complete articles. It’s verydifficult to build summaries for each book manually and there are many deep learning algorithms are available while will read whole content and give summary but this summaries are not accurate as they work directly on words. Inpropose paper author using phrases (combination ofwords) from sentences to train LSTM-CNN algorithm and this phrases can help deep learning algorithm in obtaining accurate summary.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

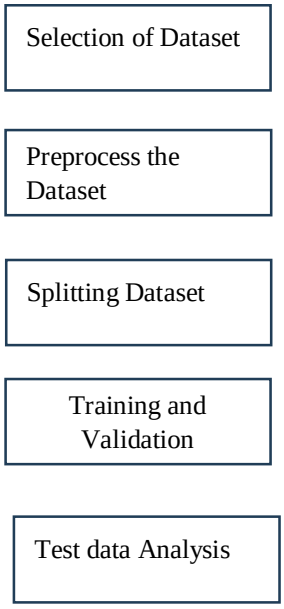


Figure 2: Proposed Block diagram of Proposed Model

There are above mentioned steps are discussed as below,

1) Selection of dataset : select proper dataset is key for getting improved performance
2) Preprocessing Dataset : to bring dataset in standard format
3) Splitting dataset : dataset is splitted in 80% and 20% (80% training and 20% testing)
4) Training the model with 80% train data from dataset and 20% data is used for validation
5) Test data Analysis : we can enter any text data and generate its abstractive summarization.

Abstractive Text Summary using deep learning(ATSDL) consist of two parts where first part concentrate on extracting phrases from dataset and then extracted phrases will be converted to vector and this vector will be input to LSTM-CNN algorithm to build text summary model.

User can input any text and then ATSDL will predict summary from it and this predicted summary and test data summary will be used to calculate ROUGE 1 and 2 scores. The higher the score the accurate is the summary.

In propose paper author using CNN-Daily Mail dataset which contains articles and summaries and by using this data author training LSTM-CNN (ATSDL) algorithm.

We have coded this project using JUPYTER notebook and below are the code and output screens with blue colour comments Data Collection and Pre-processing: Gather a diverse dataset of text documents and their corresponding human-written summaries from various domains such as news articles, scientific papers, and online forums. Preprocess the data by tokenization, stemming, and removing stop words.

Design a hybrid deep learning architecturethat combines LSTM and CNN components. The LSTM module captures long-term dependencies and sequential patterns in the input text, while the CNN module extracts hierarchical features and local structures.

Train the LSTM-CNN model on the pre-processed dataset using techniques such as mini-batch gradient descent and back propagation through time. Fine-tune hyper parameters, including learning rate, dropout rate, and embedding dimensions, to optimize performance.

Evaluate the performance of the trained model using standard evaluation metrics, including ROUGE-1, ROUGE-2, and ROUGE-L scores.

Compare the results against baseline models, including traditional LSTM, CNN, and seq2seq models, to assess improvements in summarization quality.

Conduct human evaluation and qualitative analysis to assess the linguistic quality, coherence, and informativeness of the generated summaries. Solicitfeedback from domain experts to validate the effectiveness of the LSTM-CNN model incapturing key information and preserving the original meaning of the text.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

IV. RESULT ANALYSIS

Performance analysis of proposed model is analysed using LSTM -CNN combination with python software. Python 3.7 is used for design and analysis with installation of relevant libraries. Many libraries are used which are opensource and make programming simple like NLTK used for text data processing, Keras used for deep learning , sk learn used for performance analysis.

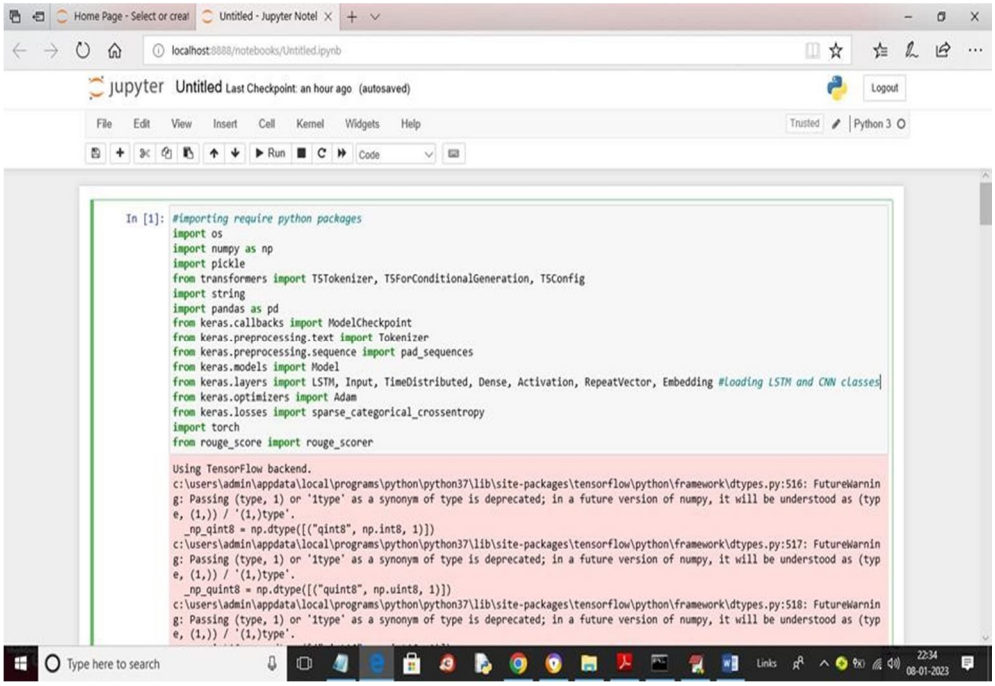


Figure 3: importing python packagesIn above screen we are importing require python packages

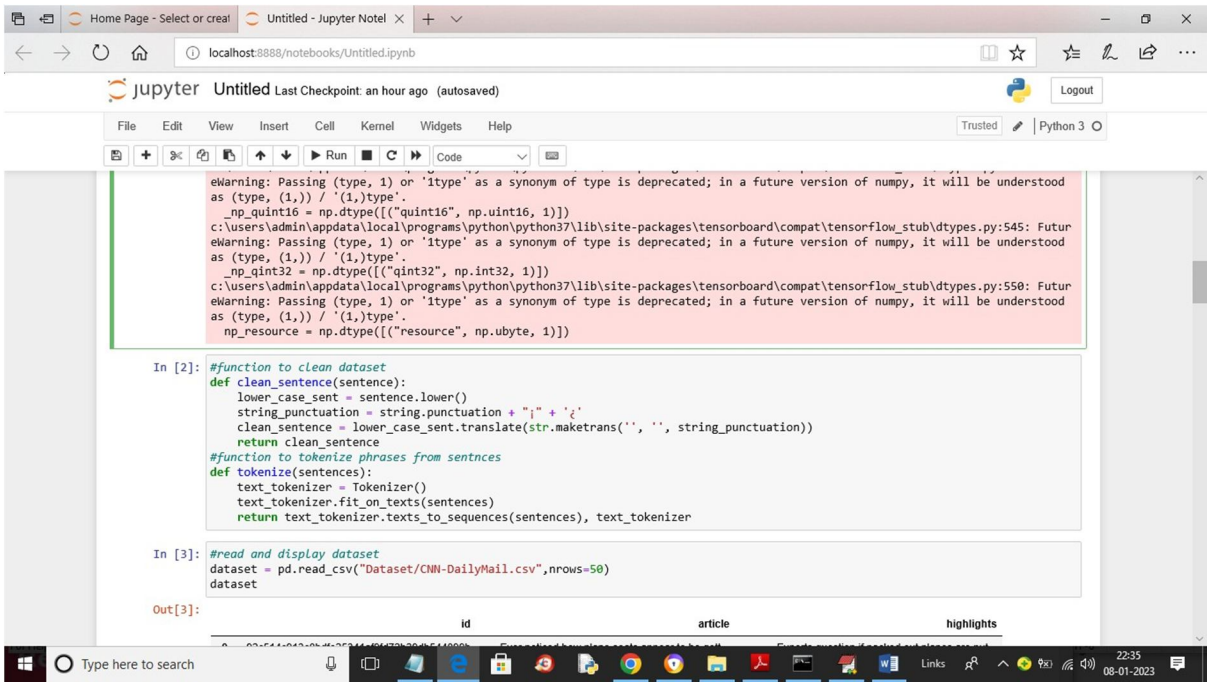


Figure 4: Defining Function

In above screen we are defining functions to clean text and the extract phrases from the sentences

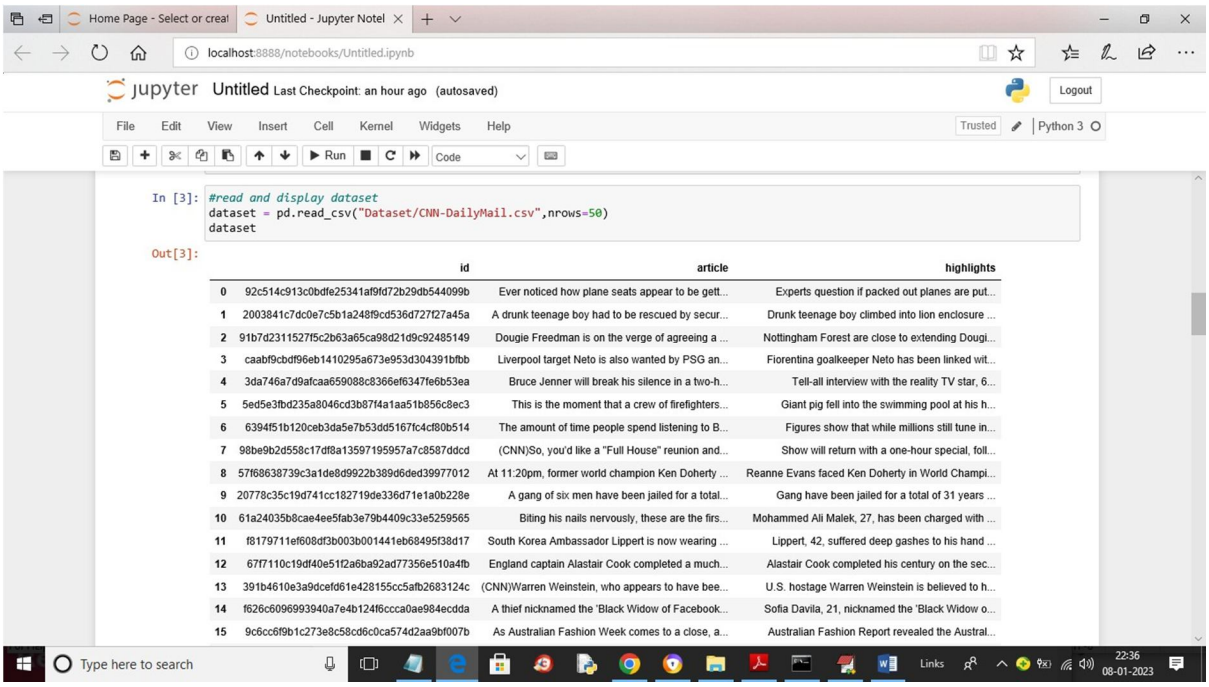


Figure 5: Reading and Displaying Dataset

In above screen reading and displaying dataset values. Dataset with multiple rows and columns is displayed in above tabular format.

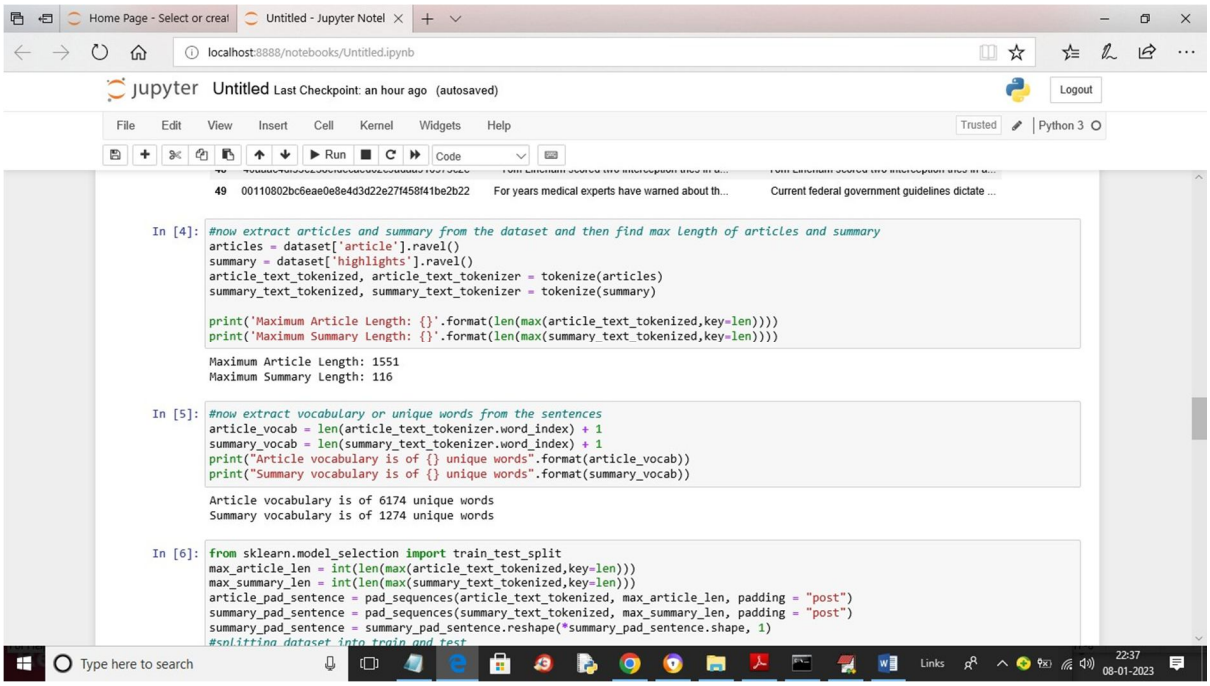


Figure 6: extracting article and summary size

In above screen extracting article and summary sizeand then extracting total articles and summary unique phrases words.

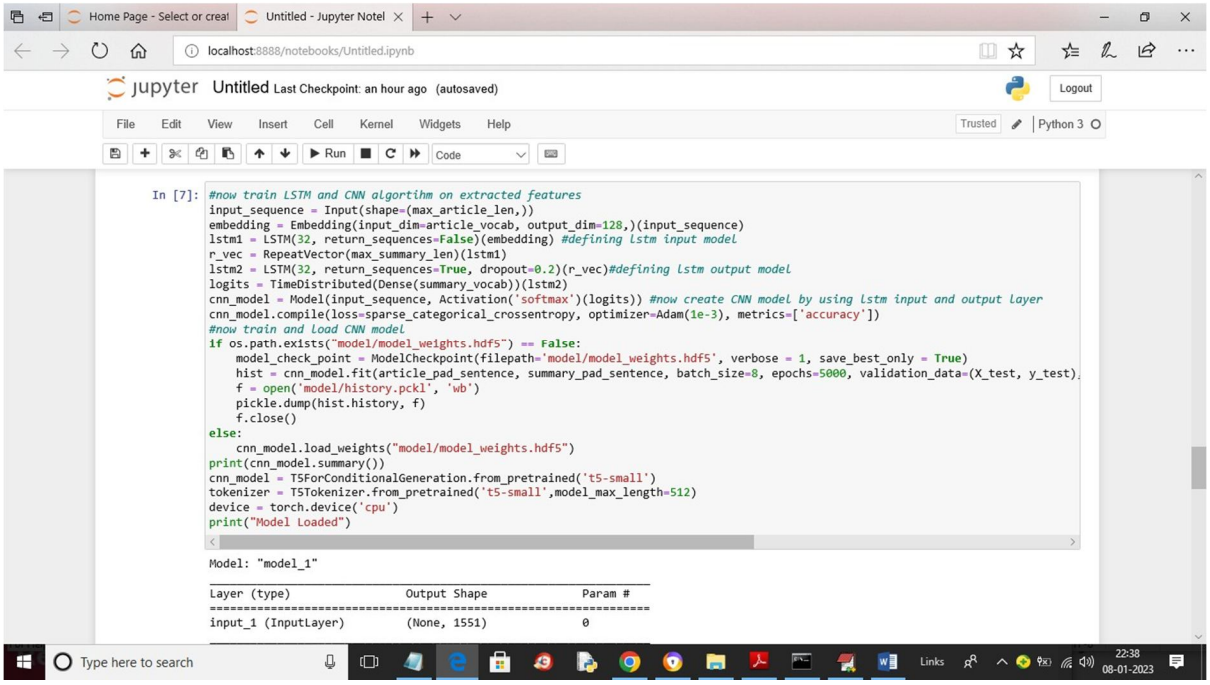


Figure 7: Defining LSTM-CNN model

In above screen defining LSTM-CNN model and then training this model to get below output

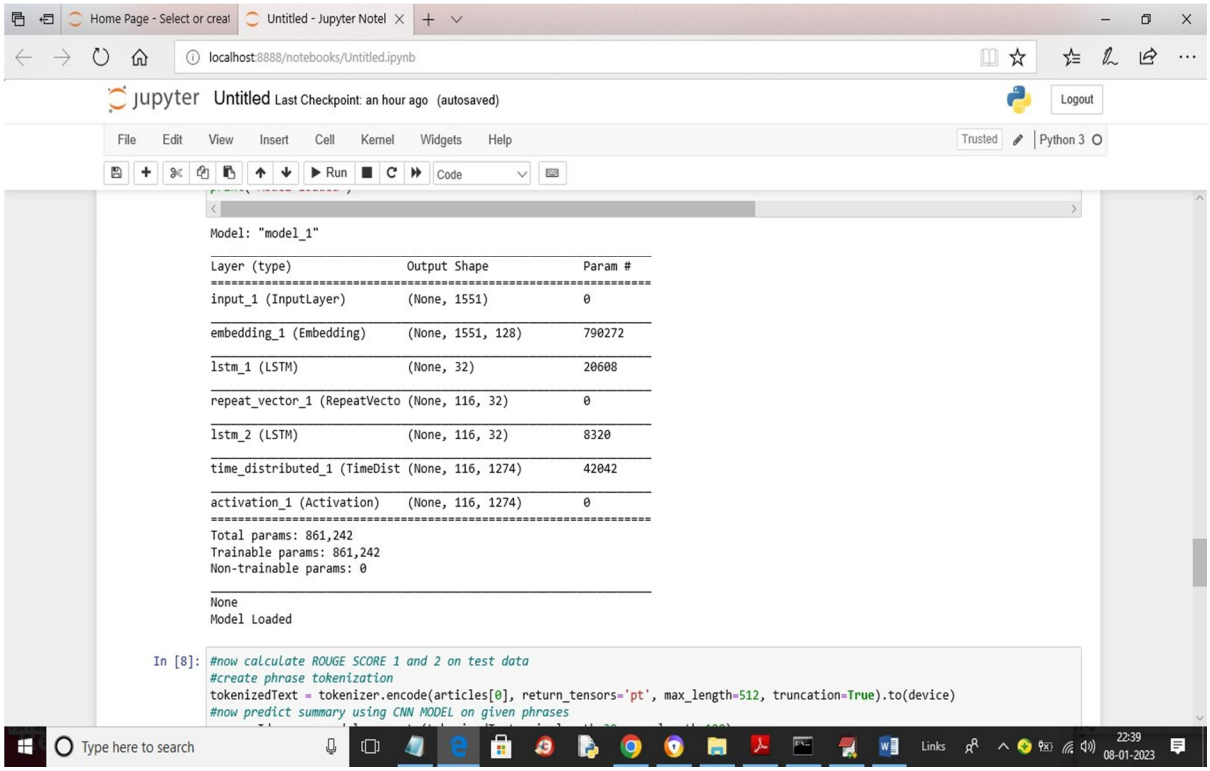


Figure 8: LSTM-CNN model details with differentlayers

In above screen we can see LSTM-CNN modeldetails with different layers and then model is loaded

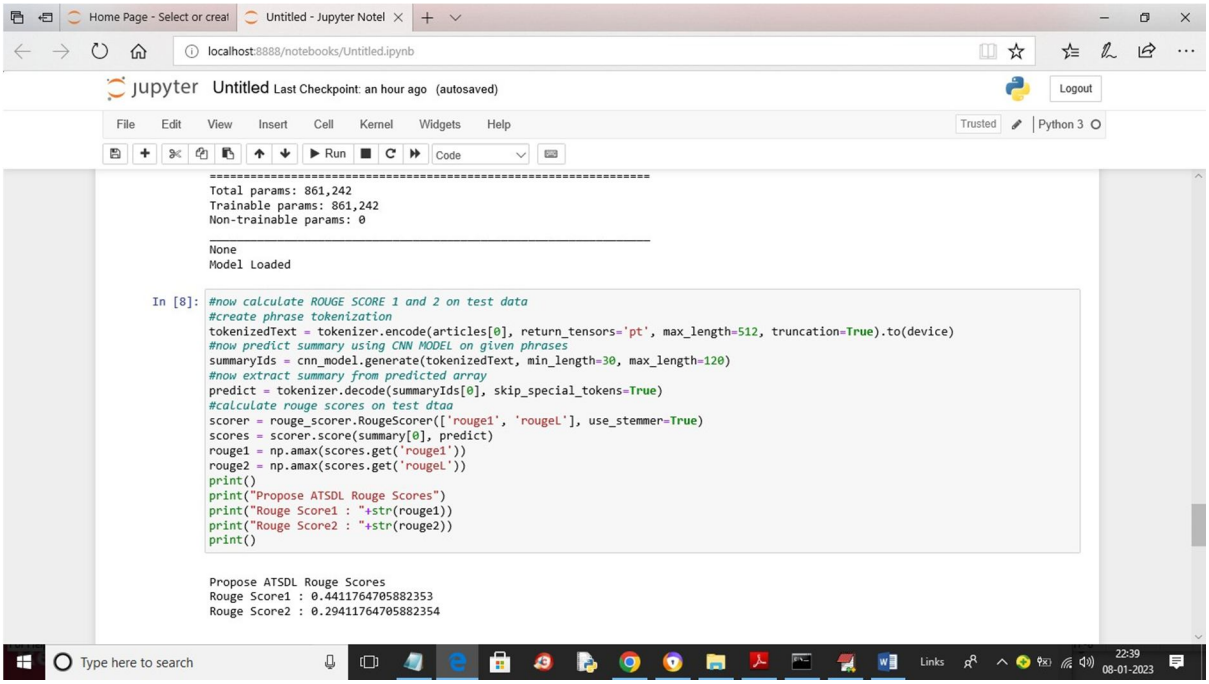


Figure 9: Applying test article on LSTM-CNNtrained model

In above screen we are applying test article on LSTM-CNN trained model and then calculate Rouge 1 and 2 score on predicted summary and we got Rouge 1 score as 0.44 and Rouge Score as 0.29 which is closer and little higher which is given in base paper.

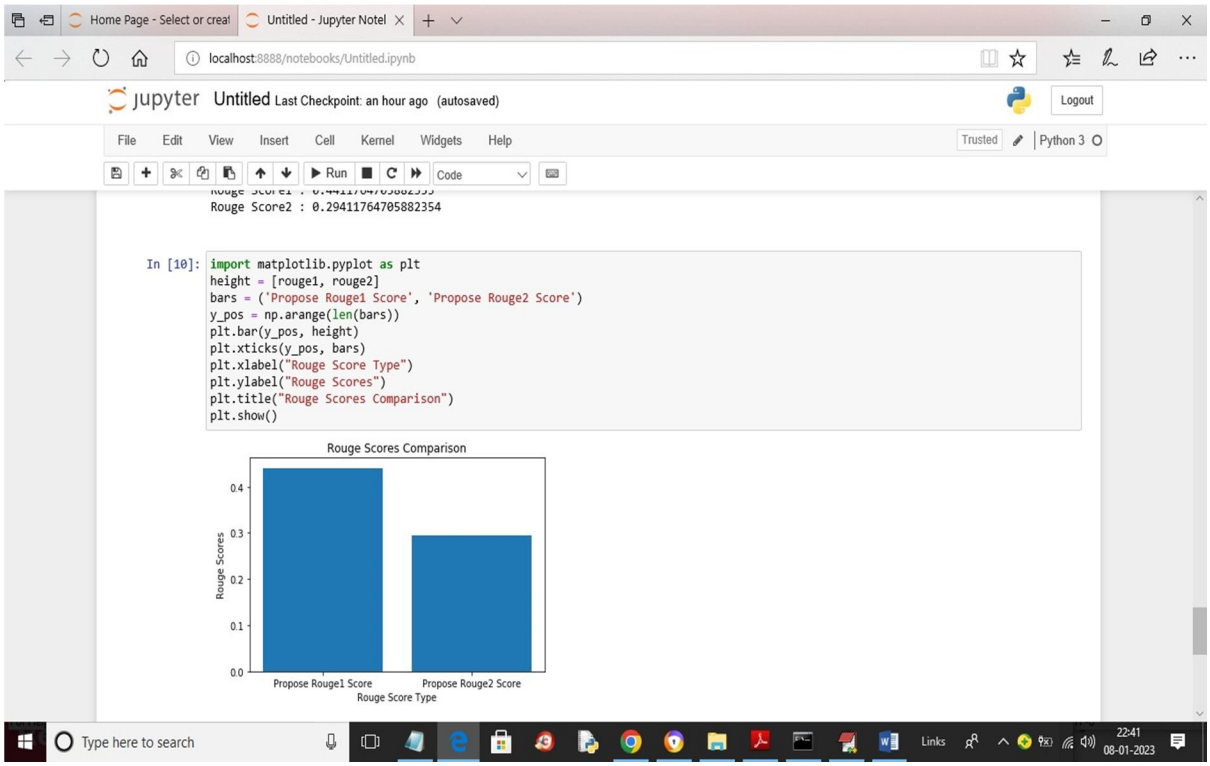


Figure 10: Plotting graph for LSTM-CNN predictedsummary

In above screen we are plotting graph for LSTM- CNN predicted summary rouge scores where x-axisrepresents rouge 1 and 2 and y-axis represents scores

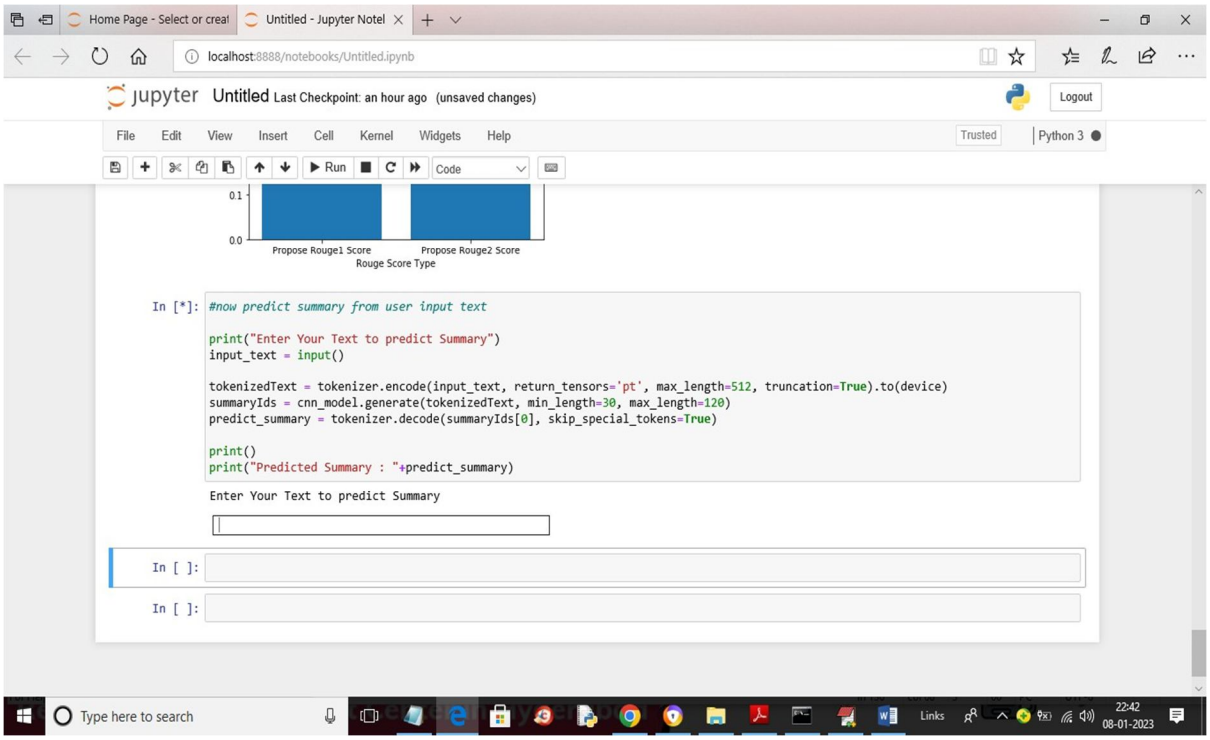


Figure 11: Input Text Field for Abstractive Summarization

In above screen after executing block you will get text field and then enter some TEXT in that text field and press enter key toget summary output

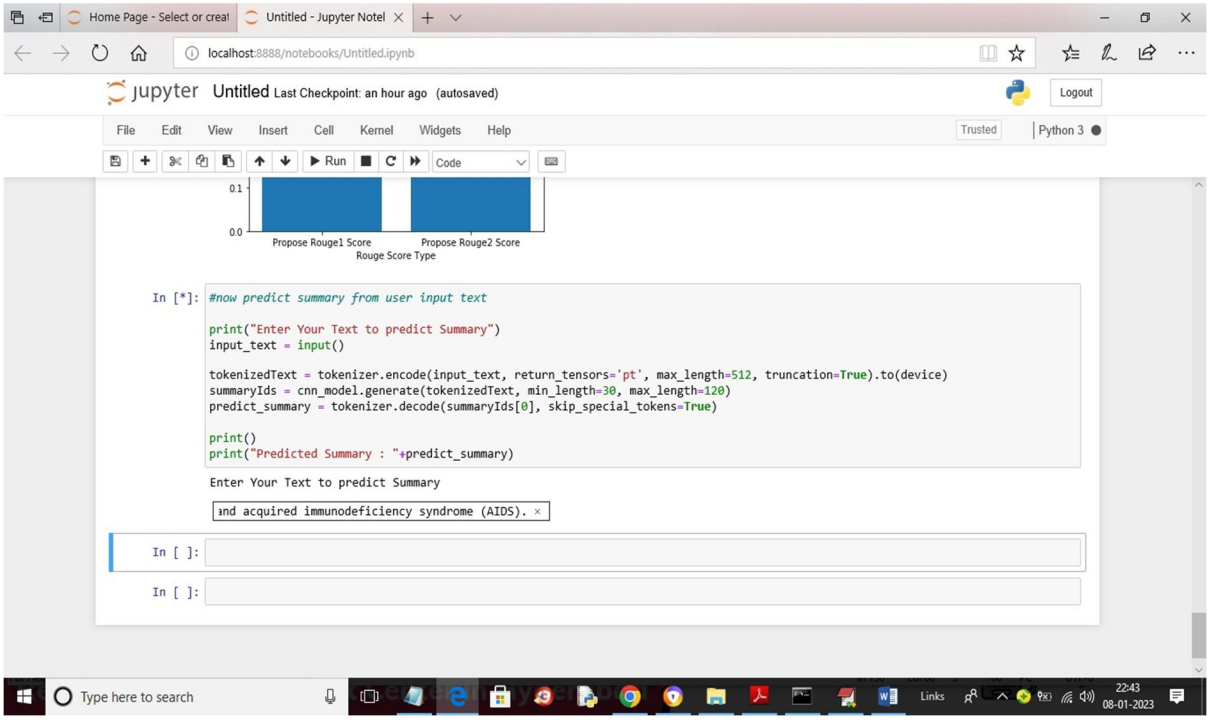


Figure 12: Entered some TEXT in text field

In above screen in text field I entered some TEXT and press enter key to get below output

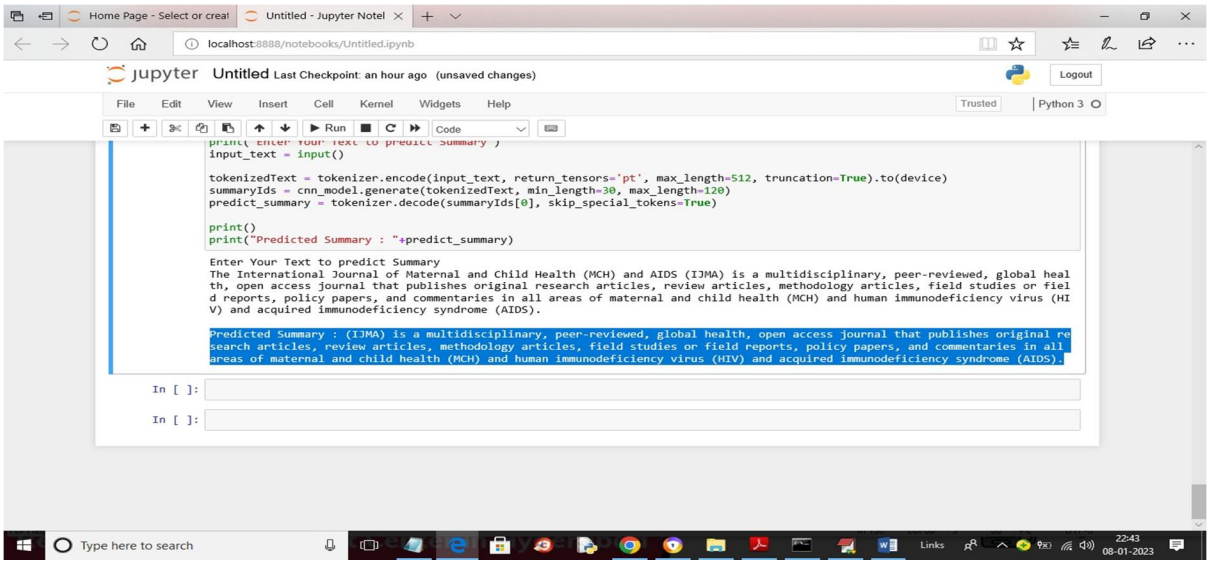


Figure 13: Displaying entered TEXT

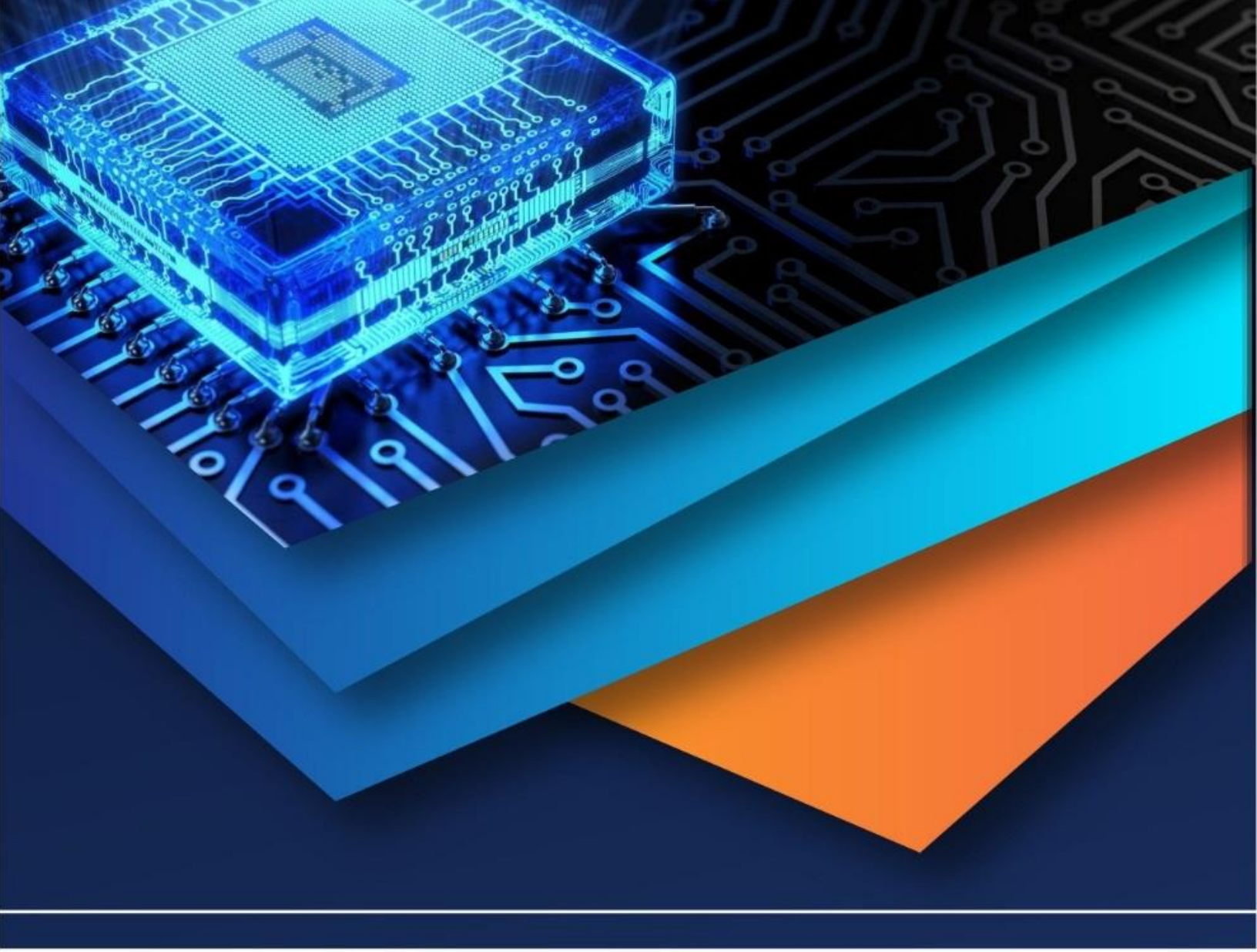
In above screen first we are displaying entered TEXT and in blue colour displaying predicted shortsummary from given TEXT. Similarly you can input any text and get predicted summary

V. CONCLUSION

In this study, we construct a unique LSTM- CNN-based ATSDL model in the field of TS, which addresses several importantissues. Recent ETS models are more concerned with syntactic structure, whereas recent ATS models concentrate more on semantics. The advantages of both summarization approaches are combined in our model. The new ATSDL model learns the collocation of phrases after first extractingimportant phrases from the source text using a phrase extraction technique called MOSP. Lastly, we carry out thorough tests on two distinct datasets, and the outcome demonstrates that our model performs better than the most advanced methods in terms of both syntactic structure and semantics. So LSTM-CNN is used for performance analysis and to give abstractive text summarization.

REFERENCES

[1] Angeli G, Tibshirani J, Wu J et al (2014) Combining distant and partial supervision for relation extraction[C]. EMNLP, pp 1556–1567
[2] Bing L, Li P, Liao Y et al (2015) Abstractive multi- document summarization via phrase selection and merging[J]. arXiv preprint arXiv:1506.01597Buss D.M., Shakelford T.K. Human aggression in evolutionarypsychological perspective. Clinical Psychology Review, 17, 1997. P. 605–619. DICTIONARIES:
[3] Cao Z, Li W, Li S et al (2016) Attsum: joint learning of focusing and summarization withneural attention[J]. arXivpreprint arXiv:1604.00125LITERATURE**:**
[4] Chen M, Weinberger KQ, Sha F (2013) An alternative text representation to TF-IDF and Bag-of-Words[J]. arXiv preprint arXiv:1301.6770
[5] Cheng J, Lapata M (2016) Neural summarization by extracting sentences and words[J]. arXiv preprint arXiv:1603.07252
[6] Chopra S, Auli M, Rush AM (2016) Abstractive sentence summarization with attentive recurrent neural networks[C]. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human LanguageTechnologies, pp 93–98
[7] Colmenares CA, Litvak M, Mantrach A et al (2015) HEADS: headline generation as sequence prediction using an abstract feature-rich space[C]. HLT-NAACL,pp 133– 142
[8] rkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization[J]. J Artif Intell Res 22:457–479
[9] Filippova K, Altun Y (2013) Overcoming the lack of parallel data in sentence compression[C]. EMNLP, pp 1481–1491
[10] Gu J, Lu Z, Li H et al (2016) Incorporating copying mechanism in sequence-to-sequence learning[J]. arXiv preprint arXiv:1603.06393
[11] Hu B, Chen Q, Zhu F (2015) Lcsts: a large scale chinese short text summarization dataset[J]. arXiv preprint arXiv:1506.05865
[12] Li J, Luong MT, Jurafsky D (2015) A hierarchical neural autoencoder for paragraphs and documents[J]. arXiv preprint arXiv:1506.01057
[13] Lin CY, Hovy E (2003) Automatic evaluation of summaries using n-gram co-occurrence statistics[C]. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, pp 71–78 1
[14] Litvak M, Last M (2008) Graph-based keyword extraction for single-document summarization[C]. Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization. Association for Computational Linguistics, pp 17–24
[15] Lopyrev K (2015) Generating news headlines with recurrent neural networks[J]. arXiv preprint arXiv: 1512.01712
[16] https://www.e2enetworks.com/blog/an- introduction-to-text-summarization-with-nlp



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY