NEWS ARTICLE SUMMARIZATION USING NLP

B. Shreya[1*], B. Archana[2], K. Raju[3]

[1]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.
[2]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.
[3]Department of Electrical and Electronics Engineering SR University, Warangal, Telangana, India.

Email: rajurgd1277@gmail.com

Abstract: Text Summarization is the process of creating a condensed form of text document which maintains significant information and general meaning of source text. Automatic text summarization becomes an important way of finding relevant information precisely in large text in a short time with little efforts. Text summarization approaches are classified into two categories: extractive and abstractive. This paper focuses on approaches to build a text automatic summarization model for news articles, generating a one-sentence summarization that mimics the style of a news title given some paragraphs. We explored Recurrent Neural Network models with encoder-decoder using LSTM and Seq2Seq. We obtained 87% accuracy using the LSTM for summarzing an news article.

Keywords: LSTM , Seq2Seq , RNN.

1. INTRODUCTION

Text summarization is a way of creating a accurate and fluent summary of a long data. The origins of it start from 1940 at the time of world war 2 to convert russsian language into english. Google uses featured snippets to show the summary of an article or the answer for a user's query. Text Summarizer is used to convert huge amount data into small lines of data the summarizer data. It focuses on the vital information of the original group of sentences and generates a new set of

sentences for the summary. This developed text or sentence may not be present in the original sentence.

As the amount of information on the web is increasing rapidly day by day in different formats such as text, video, images. It has become difficult for individual to find relevant information of his/her interest. Suppose user queries for information on the internet he may get thousands of result documents which may not necessarily relevant to his concern. To find appropriate information, a user needs to search through the entire documents this causes information overload problem which leads to wastage of time and efforts. To deal with this dilemma, automatic text summarization plays a vital role. Automatic summarization condenses a source document into meaningful content which reflects main thought in the document without altering information. Thus it helps user to grab the main notion within short time span. If the user gets effective summary it helps to understand document at a glance without checking it entirely, so time and efforts could be saved.

Text summarization process works in three steps : analysis, transformation and synthesis. Analysis step analyzes source text and select attributes. Transformation step transforms the result of analysis and finally representation of summary is done in synthesis step. There are two types in summarization of text - abstractive and exctractive text summarization. In extractive text smmarization , it only tries to highlight the sentences without noticing the meaning between the sentences. In abstractive text smmarization ,it extract the sentences an embed the words and form meaningful sentences , then creates an overall summary. Our project mianly focuses on developing abstractive summary using LSTM.

The main objective of text summarization is to understand the key concepts easily and get to know the whole concept in shortened version which is fast and time saving.

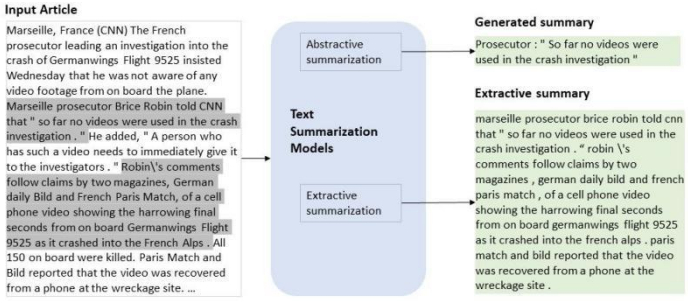


Figure 1: Text Summarization

2. PROBLEM DEFINATION :

Summarizing helps you understand and learn important information by reducing information to its key ideas.

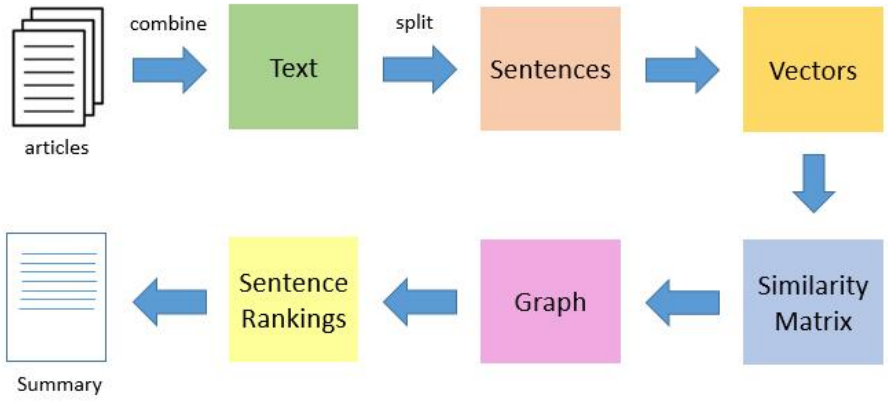


Figure 2. Processing Steps for Text Summarization using NLP.

3. DATASET AND ATTRIBUTES :

A total of 98,400 rcords are used to represent the original summary in the dataset.

.**Input feature:**

- ✧ Author

  Value : Text

- ✧ Date

  Value :0-31(Numeric)

- ✧ Headlines

  Value : Text

- ✧ Original text

  Value : Text

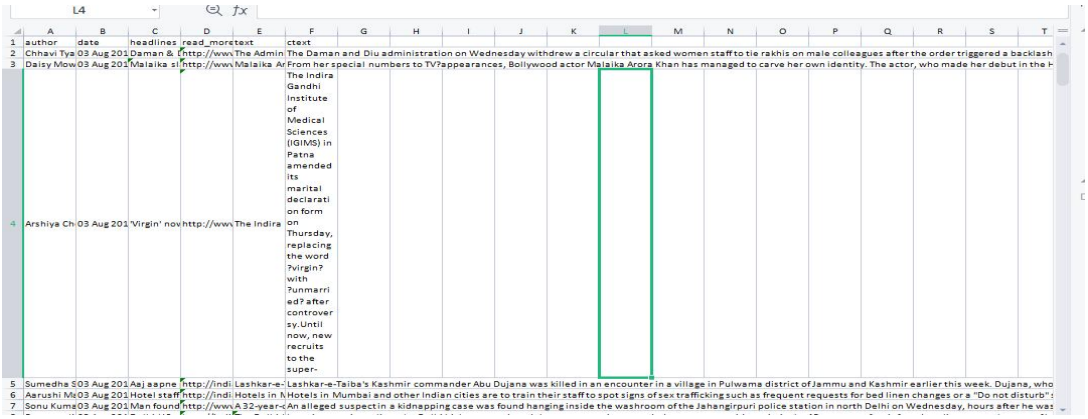
Output feature:

- ✧ Summary Text

  Value : Text

- ➢ Data sets are two types :

1. Train data
2. Test data



	author	date	headlines	read_more	text	ctext
2	Chhavi Tya	03 Aug 201	Daman & D	http://www	The Admin	The Daman and Diu administration on Wednesday withdrew a circular that asked women staff to tie rakhis on male colleagues after the order triggered a backlash
3	Daisy Mow	03 Aug 201	Malaika sl	http://www	Malaika Ar	From her special numbers to TV?appearances, Bollywood actor Malaika Arora Khan has managed to carve her own identity. The actor, who made her debut in the H
4	Arshiya Ch	03 Aug 201	'Virgin' nov	http://www	The Indira	The Indira Gandhi Institute of Medical Sciences (IGIMS) in Patna amended its marital declaration form on Thursday, replacing the word ?virgin? with ?unmarried? after controversy.Until now, new recruits to the super-
5	Sumedha S	03 Aug 201	Aaj aapne	http://indi	Lashkar-e-	Lashkar-e-Taiba's Kashmir commander Abu Dujana was killed in an encounter in a village in Pulwama district of Jammu and Kashmir earlier this week. Dujana, who
6	Aarushi Ma	03 Aug 201	Hotel staff	http://indi	Hotels in M	Hotels in Mumbai and other Indian cities are to train their staff to spot signs of sex trafficking such as frequent requests for bed linen changes or a "Do not disturb"
7	Sonu Kuma	03 Aug 201	Man found	http://www	A 32-year-c	An alleged suspect in a kidnapping case was found hanging inside the washroom of the Jahangirpuri police station in north Delhi on Wednesday, hours after he was

Fig 3 : Data-set.

3.1 DATA PRE-PROCESSING :

- Real-world data collection has its own set of problems. It is often very messy which includes missing data, presence of outliers, unstructured manner, etc. Before looking for any insights from the data, we have to first perform preprocessing tasks which then only allow us to use that data for further observation and train our machine learning model. We use missing values treatment, outliers detection, normalization and data split to process our data before feeding it to the machine learning model.

Data info:

```
In [2]:
summary = pd.read_csv('/kaggle/input/news-summary/news_summary.csv', encoding='iso-8859-1')
raw = pd.read_csv('/kaggle/input/news-summary/news_summary_more.csv', encoding='iso-8859-1')

In [3]:
pre1 =  raw.iloc[:,0:2].copy()
# pre1['head + text'] = pre1['headlines'].str.cat(pre1['text'], sep =" ")

pre2 = summary.iloc[:,0:6].copy()
pre2['text'] = pre2['author'].str.cat(pre2['date'].str.cat(pre2['read_more'].str.cat(pre2['text'].str.cat(pre2['ctext'], sep = " "), sep =" "),sep= "

In [4]:
pre = pd.DataFrame()
pre['text'] = pd.concat([pre1['text'], pre2['text']], ignore_index=True)
pre['summary'] = pd.concat([pre1['headlines'],pre2['headlines']],ignore_index = True)

In [5]:
pre.head(2)
```

Out[5]:

	text	summary
0	Saurav Kant, an alumnus of upGrad and IIIT-B's...	upGrad learner switches to career in ML & AI w...
1	Kunal Shah's credit card bill payment platform...	Delhi techie wins free food from Swiggy for on...

Fig 4 : Data preprocessing

Missing values treatment:

The real world's dataset often has many missing values which can be treated by using certain methods. But in our data, there are no missing values, because we collected the data manually through google forms survey and made sure not to miss out any data. To treat the missing values, we generally use the following strategies:

- Remove the entire row (If missing values are less in number)
- Replace the missing value with either mean or median
- Replace the missing value with most frequent value in the column (This is generally used only for large dataset)

Normalization:

Normalization is a technique for organizing data in a database. Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. It is important that a database is normalized to ensure only related data is stored in each table and to avoid biasing towards huge values. When we normalize the data while feeding it to the model, we also have to de-normalize it. This process can be done using the formulas below:

- $x_{nor} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$
- $y_i = y_{nor}(y_{max} - y_{min}) + y_{min}$

Data split:

To train any machine learning model irrespective what type of dataset is being used, we have to split the dataset into training and testing data. The reason to split the data is to give the machine learning model an effective mapping of input to outputs and to evaluate the model performance. We pass the training data to train our machine learning model and then test the model on testing data.We can do the data split using train_test_split module in python.

5 . ALGORITHMS:

This section talks about the algorithms used for the project. We used algorithms like LSTM and Seq2Seq and achieved an accuracy of 87% .

LSTM :

LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images. This finds application in speech recognition, machine translation, etc. LSTM is a special kind of RNN, which shows outstanding performance on a large variety of problems.

Sequence prediction problems have been around for a long time. They are considered as one of the hardest problems to solve in the data science industry. These include a wide range of problems; from predicting sales to finding patterns in stock markets' data, from understanding movie plots to recognizing your way of speech, from language translations to predicting your next word on keyboard.LSTMs have an edge over conventional feed-forward neural networks and RNN in many ways. This is because of their property of selectively rembeThe central role of an LSTM model is held by a memory cell known as a 'cell state' that maintains its state over time. The cell state is the horizontal line that runs through the top of the below diagram. It can be visualized as a conveyor belt through which information just flows, unchanged. LSTM Applications are - Language modeling , Machine translation, Handwriting recognition.

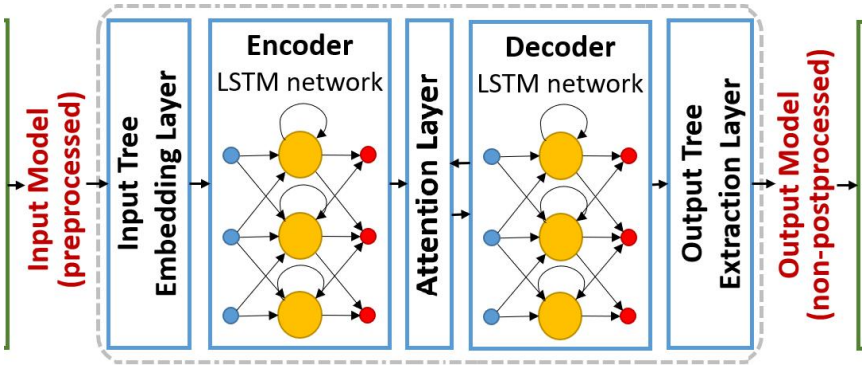


Fig 5 : LSTM architecture

A Long Short Term Memory Network consists of four different gates for different purposes as described below:-

1. **Forget Gate(f):** It determines to what extent to forget the previous data.
2. **Input Gate(i):** It determines the extent of information be written onto the Internal Cell State.
3. **Input Modulation Gate(g):** It is often considered as a sub-part of the input gate and much literature on LSTM's does not even mention it and assume it is inside the Input gate. It is used to modulate the information that the Input gate will write onto the Internal State Cell by adding non-linearity to the information and making the information **Zero-mean**.
4. **Output Gate(o):** It determines what output(next Hidden State) to generate from the current Internal Cell State.

SEQ2SEQ :

Seq2Seq model is a model that takes a stream of sentences as an input and outputs another stream of sentences. This can be seen in neural machine Translation where input sentences is one language and output sentences are translated versions of that language. Encoder and Decoder are the two main techniques used in seq2seq modeling.

Encoder Model :

Encoder model is used to encode or transdorm the input sentences and genreate feedback after every step. This feedback can be an internal state I.e, hidden state or cell state if we are using the LSTM layer.Encoder models capture the vital information from the input sentences while maintaining the context throughout. In Neural Machine translation, out input language will be passed into the encoder model where it will capture the contexual information without modiying the meaning of the input sequence. Outputs from the encoder model are then passed into the decoder model to get the output sequences.

Decoder Model :

The decoder model is used to decode or predict the target sentences word by word. Decoder input data takes the input of target sentences and predicts the next word which is then fed into the next layer for the prediction.<start> to start the sentence and <End> to end the target sentence are the two word that help the model to know what will be the initial variable to predict the next word and the ending variable to know the ending of the sentence.

While training the model, we first provide the word<start>, the model then predicts the next word that is the decoder data target. This word is then fed as input data for the next timestep to get the next word prediction.

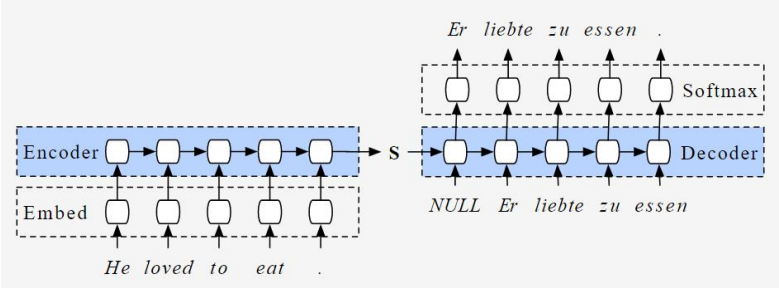


Fig 6 : Seq2Seq architecture

6. RESULTS:

We used an machine learning algorithms in this project, and the algorithm LSTM have high level accuracies for our dataset. We consider testing accuracy for final deployment of the project. Here, the testing accuracy of LSTM is highest with 87%. As this model has highest accuracy , we use that model to develop the web application for the project. Accuracies with training and testing data of the model is represented in graphical and tabular form below.

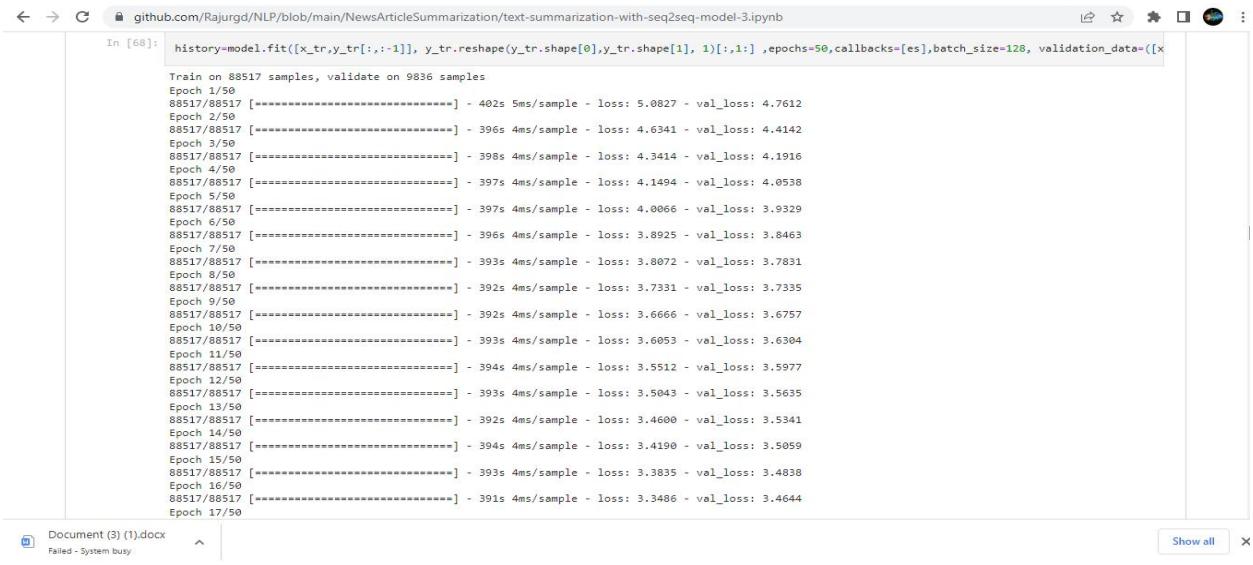
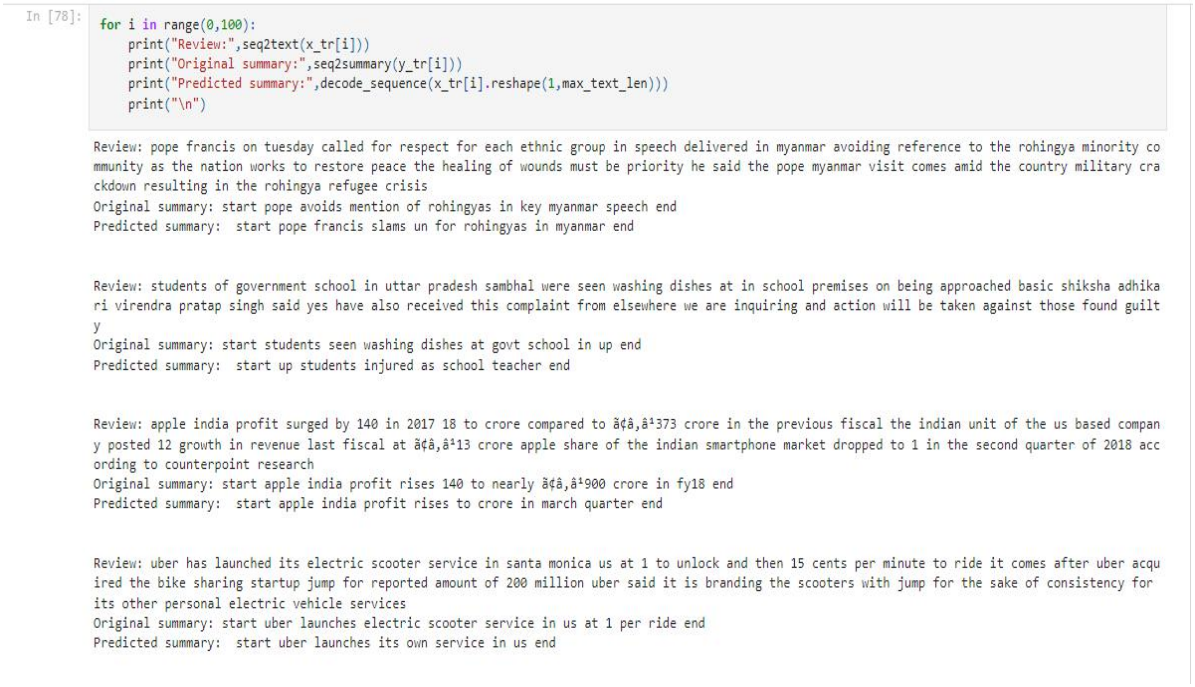


Fig 7 : Accuracy

7. OUTPUT PREDICTED:

We have developed application through Kaggle Notebook by implementing LSTM (as it has the highest testing accuracy)



```
In [78]: for i in range(0,100):
             print("Review:",seq2text(x_tr[i]))
             print("Original summary:",seq2summary(y_tr[i]))
             print("Predicted summary:",decode_sequence(x_tr[i].reshape(1,max_text_len)))
             print("\n")

Review: pope francis on tuesday called for respect for each ethnic group in speech delivered in myanmar avoiding reference to the rohingya minority co
mmunity as the nation works to restore peace the healing of wounds must be priority he said the pope myanmar visit comes amid the country military cra
ckdown resulting in the rohingya refugee crisis
Original summary: start pope avoids mention of rohingyas in key myanmar speech end
Predicted summary:  start pope francis slams un for rohingyas in myanmar end


Review: students of government school in uttar pradesh sambhal were seen washing dishes at in school premises on being approached basic shiksha adhika
ri virendra pratap singh said yes have also received this complaint from elsewhere we are inquiring and action will be taken against those found guilt
y
Original summary: start students seen washing dishes at govt school in up end
Predicted summary:  start up students injured as school teacher end


Review: apple india profit surged by 140 in 2017 18 to crore compared to ã¢â‚â¹373 crore in the previous fiscal the indian unit of the us based compan
y posted 12 growth in revenue last fiscal at ã¢â‚â¹13 crore apple share of the indian smartphone market dropped to 1 in the second quarter of 2018 acc
ording to counterpoint research
Original summary: start apple india profit rises 140 to nearly ã¢â‚â¹900 crore in fy18 end
Predicted summary:  start apple india profit rises to crore in march quarter end


Review: uber has launched its electric scooter service in santa monica us at 1 to unlock and then 15 cents per minute to ride it comes after uber acqu
ired the bike sharing startup jump for reported amount of 200 million uber said it is branding the scooters with jump for the sake of consistency for
its other personal electric vehicle services
Original summary: start uber launches electric scooter service in us at 1 per ride end
Predicted summary:  start uber launches its own service in us end
```

Fig 10 : Application

8. CONCLUSION :

The Seq2Seq encoder-decoder model has been successfully applied along with LSTM to obtain good results on summarizing the text.The model is manipulated with an accuracy of 87% accordingly to produce human-written like summaries. The future work is to improve the scalability and generalize large paragraphs to obtain summaries.

9.REFERENCES

[1] Ramesh Nallapati, Bowen Zhou et al., "Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond", The SIGNLL Conference on Computational Natural Language Learning (CoNLL), Aug 2016.

[2] Sutskever et al., "Sequence to Sequence Learning with Neural Networks", *Conference on Neural Information Processing Systems (NIPS 2014)*.

[3] Cho, B. van Merrienboer, D. Bahdanau and Y. Bengio, "On the Properties of Neural Machine translation: Encoder Decoder Approaches", *Workshop on Syntax Semantics and Structure in Statistical Translation (SSST-8)*, Oct 2014.

[4] Peter J. Liu et al., "Generating Wikipedia by Summarizing Long Sequences", International Conference on Learning Renresentation (ICLR), 2018.

[5] K. Chen et al., "Extractive Broadcast News Summarization Leveraging Recurrent Neural Network Language Modeling Techniques," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 8, pp. 1322-1334, Aug. 2015.

[6] M. Steinert, R. Granada, J. P. Aires and F. Meneguzzi, "Automating News Summarization with Sentence Vectors Offset," 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), 20.

[7] S. Huang, R. Wang, Q. Xie, L. Li and Y. Liu, "An ExtractionAbstraction Hybrid Approach for Long Document Summarization," 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), 2019.

[8] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 535-538,