Text Summarization Using Natural Language Processing for Books

A PROJECT REPORT

Submitted by:

Akhil Sanker- [RA1811026020035]

Melvin Abraham- [RA1811026020029]

K Prasath -[RA1811026020061]

D Raj Praneeth-[RA1811026020058]

Under the guidance of

MS.GOUTHAMI

(M.E., Department of Computer Science and Engineering)

in fulfillment for the award of the degree

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY RAMAPURAM CAMPUS, CHENNAI -600089

MAY 2020

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY (Deemed to be University U/S 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this project report titled -
"Text Summarization Using Natural Language Processing for Books"
is the bonafide work of

Akhil Sanker (RA1811026020035), Melvin Abraham (RA1811026020029), K. Prasath (RA1811026020061), D Raj Praneeth (RA1811026020058)

who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an occasion on this or any other candidate.

SIGNATURE

MS.GOUTHAMI,
Assistant Professor,
Computer Science and Engineering,
SRM Institute of Science and Technology,
Ramapuram Campus, Chennai.

SIGNATURE

Dr. N. KANNAN., Ph.D.,
Professor and Head
Computer Science and Engineering,
SRM Institute of Science and Technology
Ramapuram Campus, Chennai.

Submitted for the project viva-voce held on 10-11-2020 at SRM Institute of Science and Technology , Ramapuram Campus, Chennai -600089.

INTERNAL EXAMINER **EXTERNAL EXAMINER**

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
RAMAPURAM, CHENNAI - 89

DECLARATION

We hereby declare that the entire work contained in this project report titled "**Text Summarization Using Natural Language Processing for Books**"

has been carried out by **:**
Akhil Sanker (RA1811026020035), Melvin Abraham (RA1811026020029),
K. Prasath (RA1811026020061), D Raj Praneeth (RA1811026020058) at
SRM Institute of Science and Technology, Ramapuram Campus, Chennai 600089,

under the guidance of **M.S.GOUTHAMI , Assistant Professor**, Department of Computer Science and Engineering.

Place: Chennai

Date: 10-11-2020

Department of Computer Science and Engineering
SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course : B.Tech / Computer Science Engineering(AI & ML)

Student Name : Akhil Sanker , Melvin Abraham , K Prasath , D Raj Praneeth

Registration Number: RA1811026020035, RA1811026020029, RA1811026020061, RA1811026020058

Title of Work : TEXT SUMMARIZATION USING NATURAL LANGUAGE PROCESSING FOR BOOKS

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

· Clearly references / listed all sources as appropriate

· Referenced and put in inverted commas all quoted text (from books, web, etc)

· Given the sources of all pictures, data etc. that are not my own

· Not made any use of the report(s) or essay(s) of any other student(s) either past or present · Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
· Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalised in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

ACKNOWLEDGEMENT

We place on regard of our deep sense of gratitude to our lionized Chairman **Dr.R.SHIVAKUMAR** for providing us with the requisite infrastructure throughout the course.

We take the opportunity to extend our hearty and sincere thanks to our Dean, **Dr.G.SELVAKUMAR, M.Tech, PhD.,** for maneuvering us into accomplishing the project.

We take the privilege to extend our hearty and sincere guidance to the Professor and Head of the Department, **Dr.N.KANNAN, M.Tech, PhD.,** for his suggestions, support and encouragement towards the completion of the project with perfection.

We convey our hearty and sincere thanks to our Project Coordinator **Dr.V.SELLAM, M.Tech, PhD.,** for her fortification. We express our hearty and sincere thanks to our guide **M.S.MINU, M.E.,** Computer Science and Engineering Department for her sustained encouragement, consecutive criticism and constant guidance throughout this project work.

Our thanks to the teaching and non-teaching staff of the Computer Science and Engineering Department of SRM Institute of Science and Technology, Ramapuram Campus, who provided necessary resources for our project.

Akhil Sanker (RA1811026020035)
Melvin Abraham (RA1811026020029)
K Prasath (RA1811026020061)
D Raj Praneeth(RA1811026020058)

ABSTRACT

Over the Past Years, many books have been written, many documents have been published online and it still goes on as we speak.

There is an enormous amount of textual material, and it is only growing every single day.

Think of the internet, comprised of web pages, news articles, status updates, blogs and so much more. The data is unstructured and the best that we can do to navigate it is to use search and skim the results.

There is a great need to reduce much of this text data to shorter, focused summaries that capture the salient details, both so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for.

We cannot possibly create summaries of all of the text manually; there is a great need for automatic methods.

1. Summaries reduce reading time.
2. When researching documents, summaries make the selection process easier.
3. Automatic summarization improves the effectiveness of indexing.
4. Automatic summarization algorithms are less biased than human summarizers.
5. Personalized summaries are useful in question-answering systems as they provide personalized information.
6. Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process.

we have focused on analyzing and pre- processing data sets as well as the deployment of the Stacked LSTM models with attention for the purpose of Getting an Apt Summary for our documents.

Table Of Contents:

Problem Statement:

As of late, there has been a blast in the measure of text data from an assortment of sources.

This volume of text is a priceless source of information and knowledge, which should be effectively summarized to be useful.

In this problem, the main objective is to automatic text summarization are described below for lighting more about processes. With the dramatic growth of the Internet, people are overwhelmed by the tremendous amount of online information and documents.

This expanding availability of documents has demanded exhaustive research in automatic text summarization. Now days many research is going on for text summarization. Because of increasing information in the internet, these kinds of research are gaining more and more attention among the researchers.

Extractive text summarization generates a summary by extracting proper set of sentences from a document or multiple documents by deep learning. The whole concept is to reduce or minimize the valuable information present in the documents.

The procedure can be manipulated Neural Networks with the help of Specific model namely Stacked LSTM , it also uses a mechanism of Attention algorithm for better efficiency by removing redundant sentences.

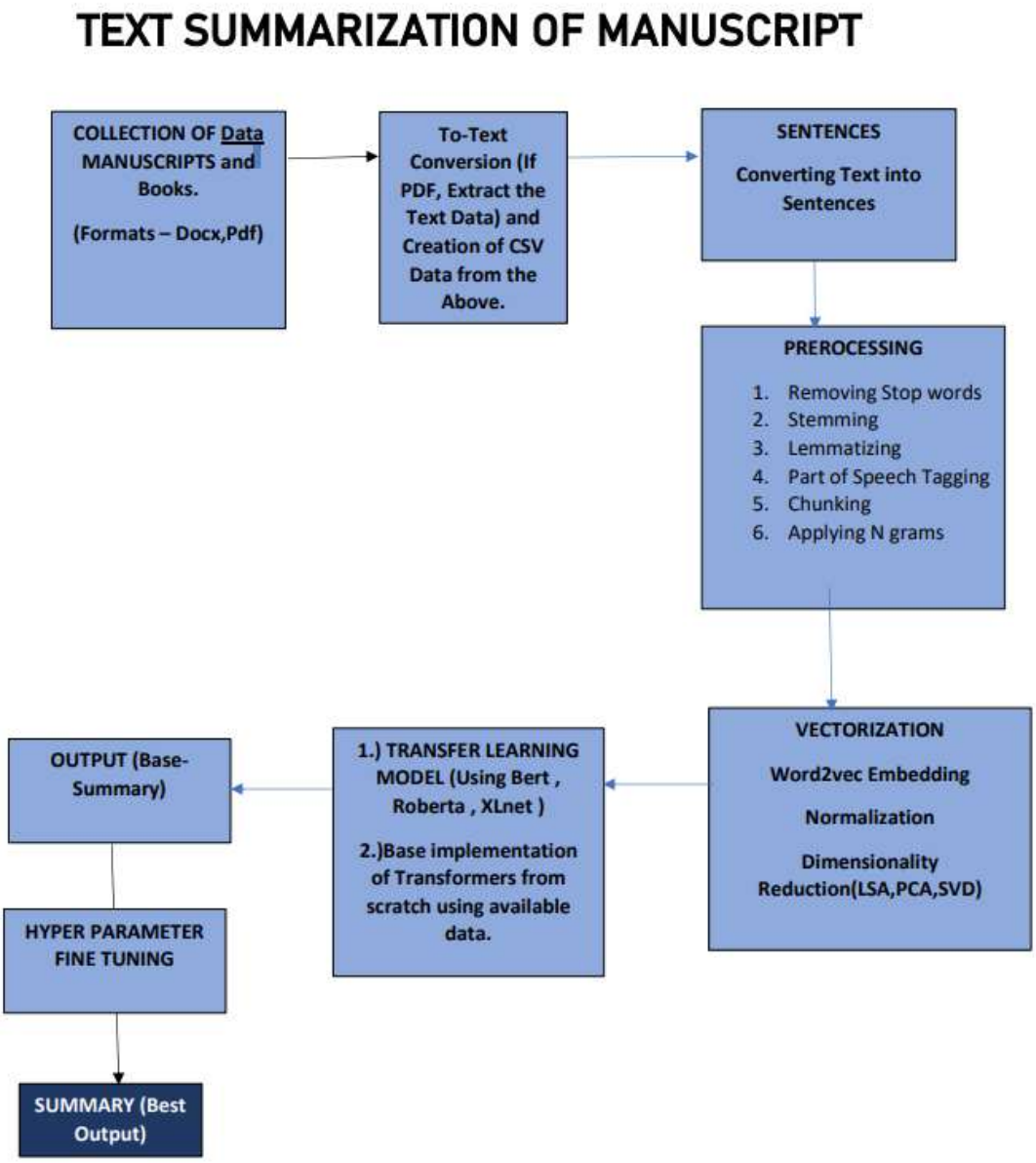
Literature Survey:

Reference Paper TITLE and AUTHOR	Concept and Algorithms	Advantages	Drawbacks
Text Summarization Techniques: SVM versus Neural Networks Shang Gaoa, Jawad Attari, Ken Barker	-Text Summarization using two models and comparing their performances ! -SVM vs Neural networks , because both are good at non-linear data and at large scale. -Find applicable corpus =>Extract features=>assign predictor class=>train and validate =>compare	-SVM is faster than Neural network -Neural network is a bit efficient compared to SVM. -When moved from normal MLP to Transformer models such as Bert , significant improvements were found -Helps to find the best approach	-Time consuming and Computational -Lack of Good Corpuses. -Might not be the perfect one (real time) , Example as when taken into something more complex such as medicine , we might never know .
Extractive Summarization using Continuous Vector Space Models Olof Mogren, Devdatt Dubhashi, Mikael Kageb ˚	-Focuses on notion of similarity of sentences . -Embeddings are created and mapped into latent space -tf-idf , vector distance , cosine-similarity are used	-Improve the existing models by applying better techniques. -Compare DeepNN, RNN,Autoencoders . -State of the Art Performance achieved	-Highly Time consuming & Computationally costly -Word embedding Approach is comparatively slower and high dimensional as embeddings itself will take a huge time.
Sequence GAN for long Text Summarization Hao Xu, Yanan Cao, Yanbing Liu	Triple RNN as Discriminator and Encoder Decoder as Generative. Attention mechanism	(seq2seq model achieves sota it uses Most Likekihood Estimation (MLE) principle for training,	the generated summaries consist of repeating phrases. our model is still a supervised learning

	RNN Encoder Decoder Architecture. Positional Encoding	which suffers from exposure bias. GAN in machine translation using ANMT. Double Attention machanism. Compare LexRank,abs-baseline atttention-based seq2seq model,abs+INRNN by introducing attention in encoder,abs+ enahnced version of abs,DeepRL,ANMT. ATRNN - 41.565 IN DAILY MAIL Corpus and 31.40% in NLPCC corpus	one relying on high-quality training datasets which is scarce. we will study an unsupervised or semi-supervised framework which can be applied to the text summarization task.
PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization Jingqing Zhang * 1 Yao Zhao * 2 Mohammad Saleh 2 Peter J. Liu 2	Sparse Attention mechanism. MLM (Masked Language Model) Transformer	evaluated PEGASUS model on 12 downstream./test data summarization tasks spanning news, science, stories, instructions, emails, patents, and legislative bills. Experiments demonstrate it achieves state-of-the-art performance on all 12 downstream datasets measured by ROUGE scores.. model was able to adapt to unseen summarization datasets very quickly	Need Large Corpus of text data for pre-training objective. Other than that no any drawbacks because it latest paper released in july 2020 overcomes all the drawback of previous bird-pagasus. Need indepth understanding of transformer and its architecture to implemention.. Of course really high computation speed .

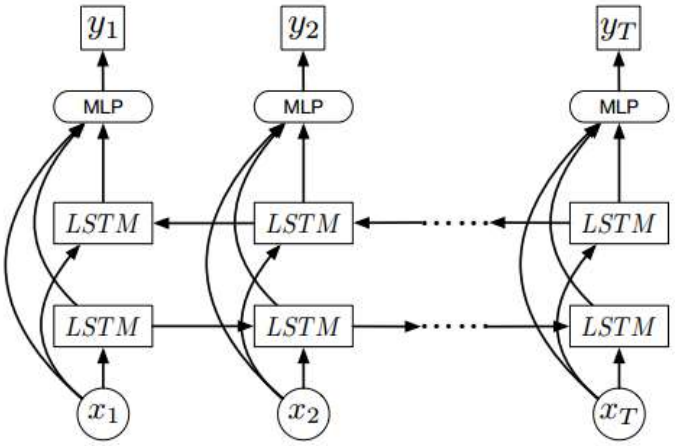
Multi documents on text summarization techniques **Author:** Chintan Shah and Anjali G Jivani	Graph LSA Term frequency Cluster	New approaches can be made and developed with the help of NLP and Linguistic apperances which can help us to get better summary	By using exiting techniques approaches there will be more time consuming and effort towards will be more
A Survey on Automatic Text Summarization **Authors:-** D.Das , AFT MArtins	-Uses Naïve bayes -Decision Trees -Hidden Markov Models -Non Linear Models	-Compares Performance and decides which one is the best among these.	-it is diffcult to replicate or extend the broader domains in abstractive summarization
Improving performance of Text Summarization **Authors:** S.A.Babara Pallavi D.Patil	- This model makes use of fuzzy logic extraction approach for text summarization. - Performs Latent Semantic Analysis (LSA) as opposed to performing direct word matching. - Has high recall and precision significance test with manual evaluation results	- Can extract hidden semantic relations between concepts in a text unlike the traditional methods. - Accurately captures semantic contents in sentences with the help of latent semantic analysis.	
Automatic summarising: the state of the art **Author:** Karen Spark Jones	- content is extracted from the original data, but the extracted content is not modified in any way. - Abstraction transforms the extracted content by paraphrasing sections of the source document, to condense a text much more strongly.	- Works Instantly. Reading the entire article, dissecting it and separating the important ideas from the raw text takes time and effort - Can work with any languages without the need for manual intervention	- Automatic summarization is a complex task that consists of several sub-tasks. - Each of the sub-tasks directly affects the ability to generate high quality summaries.

Architecture Diagram:



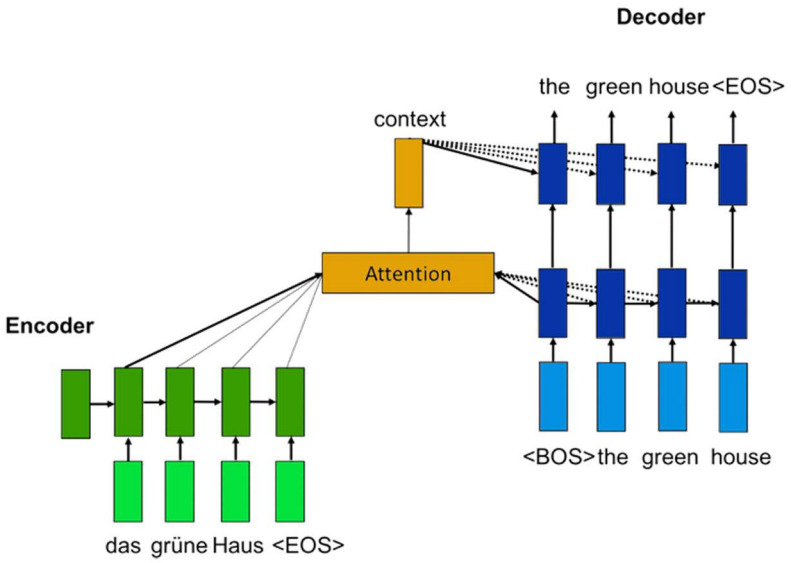
METHODOLOGY :

The approach that this document proposes, uses a Specific neural network architecture of stacked LSTM , along with attention model implementation. It ensures that the results we get are the best ones so far.



The above diagram Represents Stacked LSTM , along with this , Attention Algorithm has been implemented.
The data was obtained from various different sources and the model was trained.

There was a shuffling of data from different sources , hence huge variance is shown. Imbalance issues have been solved using techniques such as SMOTE.



Application Programming:

So we made use Jupyter Notebook Environment to implement our sought out code and program, it took some real time and we can hence show the coding details as follows , the following runs on kaggle kernel as we lack the resources :

In [10]:
```
df.head()
```

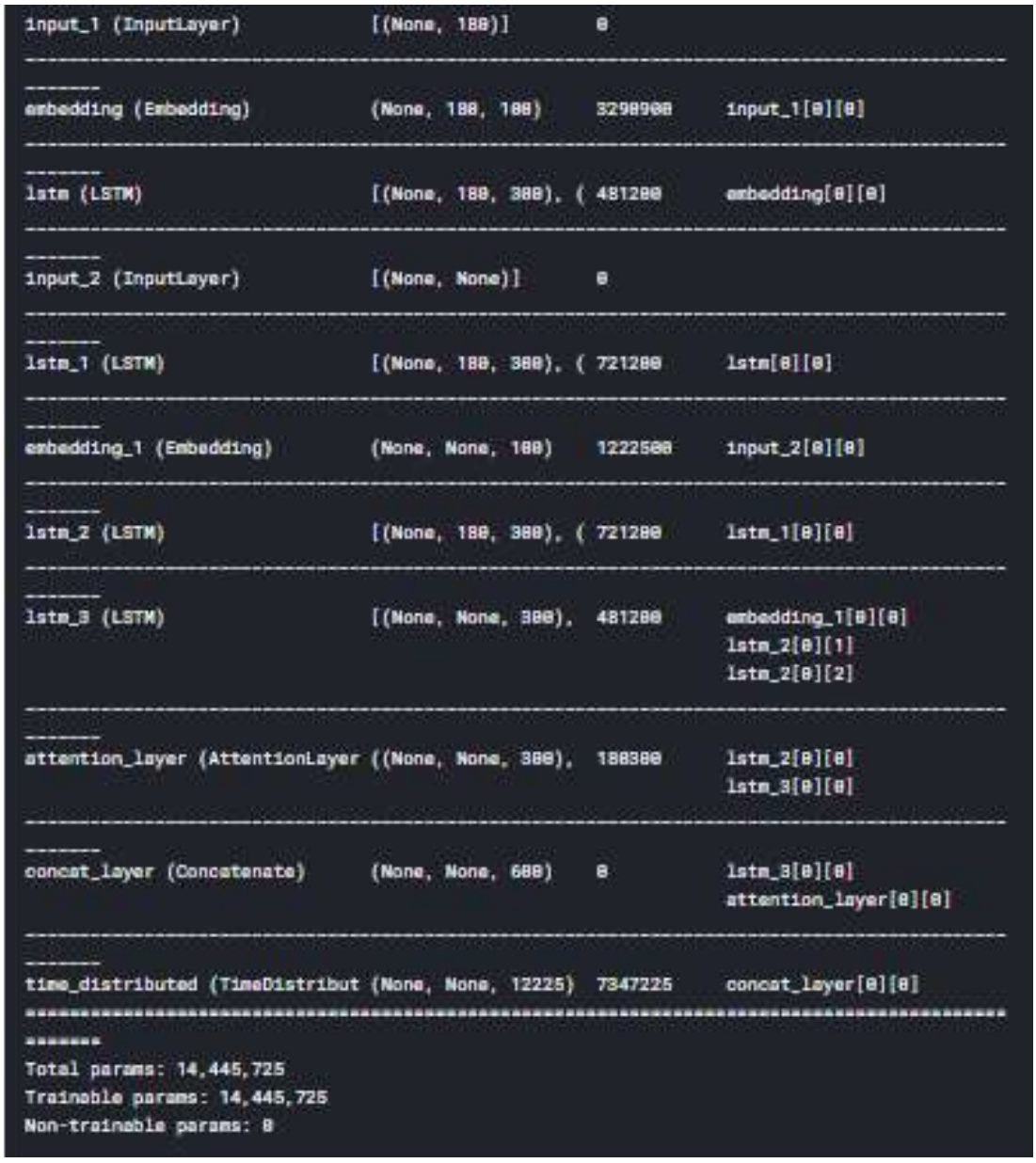
Out[10]:

	Text	Summary	cleaned_text	cleaned_summary
0	ad sales boost time warner profit quarterly profits at us media giant timewarner jumped 76% to 1.13*bn*(â 639m £600*m*)*forthethreemonthstodecember*, *from* year-earlier. the firm, which is ...	timewarner said fourth quarter sales rose 2% to 11.1*bn from* 10.9bn.for the full-year, timewarner posted a profit of 3.36*bn*, *up*27 42.09b...	sales boost time warner profit quarterly profits media giant timewarner jumped three months december year earlier firm one biggest investors google benefited sales high speed internet connections ...	timewarner said fourth quarter sales rose to bn from bn for the full year timewarner posted profit of bn up from its performance while revenues grew to bn quarterly profits at us media giant timew...
1	dollar gains on greenspan speech the dollar has hit its highest level against the euro in almost three months after the federal reserve head said the us trade deficit is set to stabilise. an...	the dollar has hit its highest level against the euro in almost three months after the federal reserve head said the us trade deficit is set to stabilise.china currency remains pegged to the doll...	dollar gains greenspan speech dollar hit highest level euro almost three months federal reserve head said trade deficit set stabilise alan greenspan highlighted government willingness curb spendin...	the dollar has hit its highest level against the euro in almost three months after the federal reserve head said the us trade deficit is set to stabilise china currency remains pegged to the dolla...

The model is represented by the following code:

```
#encoder lstm 2
encoder_lstm2 = LSTM(latent_dim,return_sequences=True,return_state=True,dropout=0.4,recurrent_d
ropout=0.4)
encoder_output2, state_h2, state_c2 = encoder_lstm2(encoder_output1)

#encoder lstm 3
encoder_lstm3=LSTM(latent_dim, return_state=True, return_sequences=True,dropout=0.4,recurrent_d
ropout=0.4)
encoder_outputs, state_h, state_c= encoder_lstm3(encoder_output2)

# Set up the decoder, using `encoder_states` as initial state.
decoder_inputs = Input(shape=(None,))

#embedding layer
dec_emb_layer = Embedding(y_voc, embedding_dim,trainable=True)
dec_emb = dec_emb_layer(decoder_inputs)

decoder_lstm = LSTM(latent_dim, return_sequences=True, return_state=True,dropout=0.4,recurrent_
dropout=0.2)
decoder_outputs,decoder_fwd_state, decoder_back_state = decoder_lstm(dec_emb,initial_state=[sta
te_h, state_c])

# Attention layer
attn_layer = AttentionLayer(name='attention_layer')
attn_out, attn_states = attn_layer([encoder_outputs, decoder_outputs])

# Concat attention input and decoder LSTM output
decoder_concat_input = Concatenate(axis=-1, name='concat_layer')([decoder_outputs, attn_out])

#dense layer
decoder_dense =  TimeDistributed(Dense(y_voc, activation='softmax'))
decoder_outputs = decoder_dense(decoder_concat_input)

# Define the model
model = Model([encoder_inputs, decoder_inputs], decoder_outputs)
```

The final Model is:



```
input_1 (InputLayer)            [(None, 180)]        0
_________________________________________________________________
embedding (Embedding)           (None, 180, 100)     3290900     input_1[0][0]
_________________________________________________________________
lstm (LSTM)                     [(None, 180, 300), ( 481200      embedding[0][0]
_________________________________________________________________
input_2 (InputLayer)            [(None, None)]       0
_________________________________________________________________
lstm_1 (LSTM)                   [(None, 180, 300), ( 721200      lstm[0][0]
_________________________________________________________________
embedding_1 (Embedding)         (None, None, 100)    1222500     input_2[0][0]
_________________________________________________________________
lstm_2 (LSTM)                   [(None, 180, 300), ( 721200      lstm_1[0][0]
_________________________________________________________________
lstm_3 (LSTM)                   [(None, None, 300),  481200      embedding_1[0][0]
                                                                 lstm_2[0][1]
                                                                 lstm_2[0][2]
_________________________________________________________________
attention_layer (AttentionLayer ((None, None, 300),  180300      lstm_2[0][0]
                                                                 lstm_3[0][0]
_________________________________________________________________
concat_layer (Concatenate)      (None, None, 600)    0           lstm_3[0][0]
                                                                 attention_layer[0][0]
_________________________________________________________________
time_distributed (TimeDistribut (None, None, 12225)  7347225     concat_layer[0][0]
=================================================================
Total params: 14,445,725
Trainable params: 14,445,725
Non-trainable params: 0
```

Application Output Screenshots:

Head of the Input:

```
In [10]: df.head()
```

Out[10]:

	Text	Summary	cleaned_text	cleaned_summary
0	ad sales boost time warner profit quarterly profits at us media giant timewarner jumped 76% to 1.13*bn*(â 639m £600*m*)*forthethreemonthstodecember*, *from* year-earlier. the firm, which is ...	timewarner said fourth quarter sales rose 2% to 11.1*bn from* 10.9bn.for the full-year, timewarner posted a profit of 3.36*bn*, *up*27 42.09b...	sales boost time warner profit quarterly profits media giant timewarner jumped three months december year earlier firm one biggest investors google benefited sales high speed internet connections ...	timewarner said fourth quarter sales rose to bn from bn for the full year timewarner posted profit of bn up from its performance while revenues grew to bn quarterly profits at us media giant timew...
1	dollar gains on greenspan speech the dollar has hit its highest level against the euro in almost three months after the federal reserve head said the us trade deficit is set to stabilise. an...	the dollar has hit its highest level against the euro in almost three months after the federal reserve head said the us trade deficit is set to stabilise.china currency remains pegged to the doll...	dollar gains greenspan speech dollar hit highest level euro almost three months federal reserve head said trade deficit set stabilise alan greenspan highlighted government willingness curb spendin...	the dollar has hit its highest level against the euro in almost three months after the federal reserve head said the us trade deficit is set to stabilise china currency remains pegged to the dolla...

Vocabulary Size:

```
In [21]: x_voc
```

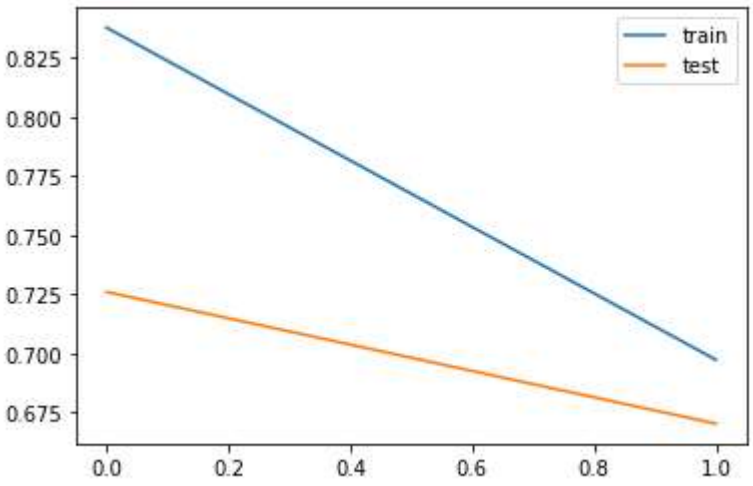
```
Out[21]: 32909
```

The Evaluation of Accuracies :

```
history=model.fit([x_tr,y_tr[:,:-1]], y_tr.reshape(y_tr.shape[0],y_tr.shape[1], 1)[:,1:] ,epochs=2,callbacks=[es],batch_size=128, validation_data=([x_val,y_val[:,:-1]], y_val.reshape(y_val.shape[0],y_val.shape[1], 1)[:,1:]))

model.save("new_model_A.h5")
```

```
Epoch 1/2
713/713 [==============================] - 2381s 3s/step - loss: 0.8376 - accuracy: 0.8954 - val_loss: 0.7260 - val_accuracy: 0.9010
Epoch 2/2
713/713 [==============================] - 2464s 3s/step - loss: 0.6974 - accuracy: 0.9032 - val_loss: 0.6704 - val_accuracy: 0.9047
```

Validation:



Output:

In [37]:

```
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: swedish company sued country discrimination watchdog call
ing woman job interview emerged would shake hands religious reaso
ns watchdog took case swedish labour court demanding company pay
kronor damages woman
Original summary: start firm sued for calling off woman interview
over handshake end
Predicted summary:  start us founder ceo quits to be in us end
```

References:

Paper 1
https://www.researchgate.net/publication/221237689_Text_summarization_techniques_SVM_versus_neural_networks

Paper 2
https://www.aclweb.org/anthology/W14-1504/

Paper 3
https://www.researchgate.net/publication/329740404_Sequence_Generative_Adversarial_Network_for_Long_Text_Summarization

Paper 4
https://arxiv.org/abs/1912.08777

Paper 5
https://www.researchgate.net/publication/310596578_Literature_Study_on_Multi-document_Text_Summarization_Techniques

Paper 6
https://www.researchgate.net/publication/228989228_A_survey_on_automatic_text_summarization

Conclusion:

We have described the a study that explores extractive summarization using Neural Networks and abstractive summarization using seq-seq LSTM stacked model.

We used multiple Stacked - LSTM . Representation capacity is increased by using this model. Experiments on the given data is by using these two techniques which outperforms several previously proposed models.

In addition to it , we've used attention mechanism to improve the outputs of our model.