National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Srishti Subhash Chandra Prasad
Student ID: x20142218

School of Computing
National College of Ireland

Supervisor: Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Srishti Subhash Chandra Prasad
Student ID:	x20142218
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Majid Latifi
Submission Due Date:	31/01/2021
Project Title:	Configuration Manual
Word Count:	1154
Page Count:	8

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Srishti Subhash Chandra Prasad
x20142218

1 Introduction

The configuration manual outlines the several stages and steps required in putting the research idea into action. It includes information on the environmental configurations, collection of dataset, preparation of input data, building model using Extractive and Abstractive approach, results obtained from the experiments and evaluation of summaries, as well as references, piece of codes, and screenshots of obtained summaries or results.

2 System Configuration

2.1 Hardware Requirement

Minimum hardware requirement for running the code

- CPU with operating frequency of minimum 1 GHz
- Disk space: 10 GB minimum
- RAM: 4 GB minimum
- 64-bit operating system

2.2 Software Requirement

Software requirements or pre-requisite for running the code

- Microsoft Edge
- Google Chrome
- IDE: Google Colab, Visual Studio
- Programming language: Python 3.7

3 Data Collection and Preparation

3.1 Data Collection

• Open the URL `https://groups.inf.ed.ac.uk/ami/icsi/download/` to reach the ICSI Corpus download page.

• Download the ICSI original MRT format transcripts with documentation. A zip file will be downloaded having "ICSI_original_transcripts" folder indide it. Files are in the .mrt format.

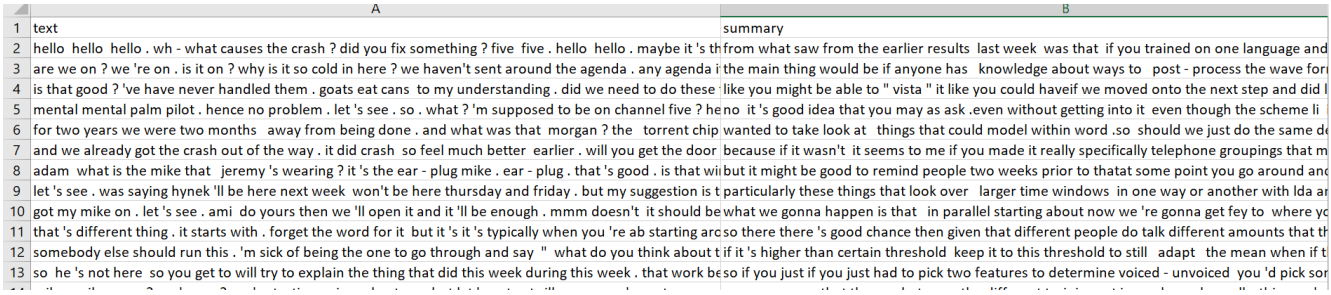
• ICSI_dataset folder is kept inside the folder "FinalCode" also.

3.2 Data Preparation

• Run ICSI_preprocess.py for the conversion of .mrt to .txt and clean the file.

• The cleaned and preprocessed transcripts are kept inside a separate folder "Preprocess_cleaned_Transcripts".

• After generating the summaries from extractive summarization the text and the generated summaries from extractive are combined into single CSV with column name "text" and "summary" as shown in Figure 1



	A	B
1	text	summary
2	hello hello hello . wh - what causes the crash ? did you fix something ? five five . hello hello . maybe it 's th	from what saw from the earlier results last week was that if you trained on one language and
3	are we on ? we 're on . is it on ? why is it so cold in here ? we haven't sent around the agenda . any agenda i	the main thing would be if anyone has knowledge about ways to post - process the wave for
4	is that good ? 've have never handled them . goats eat cans to my understanding . did we need to do these	like you might be able to " vista " it like you could haveif we moved onto the next step and did
5	mental mental palm pilot . hence no problem . let 's see . so . what ? 'm supposed to be on channel five ? he	no it 's good idea that you may as ask .even without getting into it even though the scheme li
6	for two years we were two months away from being done . and what was that morgan ? the torrent chip	wanted to take look at things that could model within word .so should we just do the same d
7	and we already got the crash out of the way . it did crash so feel much better earlier . will you get the door	because if it wasn't it seems to me if you made it really specifically telephone groupings that m
8	adam what is the mike that jeremy 's wearing ? it 's the ear - plug mike . ear - plug . that 's good . is that wi	but it might be good to remind people two weeks prior to thatat some point you go around an
9	let 's see . was saying hynek 'll be here next week won't be here thursday and friday . but my suggestion is t	particularly these things that look over larger time windows in one way or another with lda a
10	got my mike on . let 's see . ami do yours then we 'll open it and it 'll be enough . mmm doesn't it should be	what we gonna happen is that in parallel starting about now we 're gonna get fey to where y
11	that 's different thing . it starts with . forget the word for it but it 's it 's typically when you 're ab starting arc	so there there 's good chance then given that different people do talk different amounts that t
12	somebody else should run this . 'm sick of being the one to go through and say " what do you think about t	if it 's higher than certain threshold keep it to this threshold to still adapt the mean when if t
13	so he 's not here so you get to will try to explain the thing that did this week during this week . that work be	so if you just if you just had to pick two features to determine voiced - unvoiced you 'd pick so

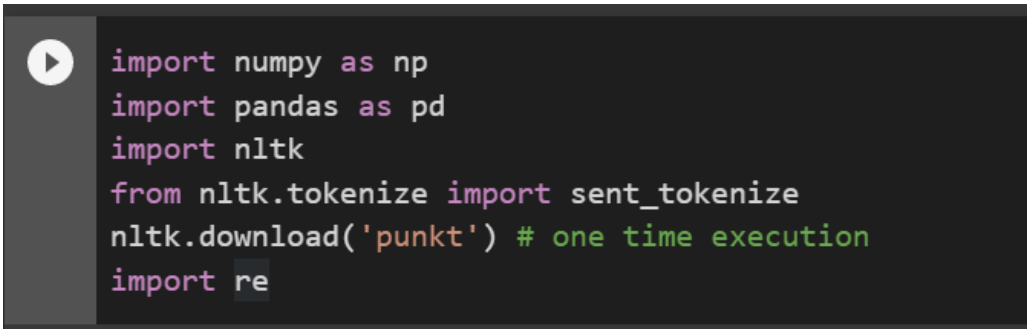
Figure 1: CSV File

4 Model Development

4.1 Importing Important Libraries

The libraries which are essential to run abstractive and extractive model:

Figure 2 shows the important libraries for extractive text summarization.



```
import numpy as np
import pandas as pd
import nltk
from nltk.tokenize import sent_tokenize
nltk.download('punkt') # one time execution
import re
```

Figure 2: Libraries for Extractive approach

Figure 3 shows the important libraries for abstractive text summarization.

```
%tensorflow_version 1.x
import numpy as np   #Package for scientific computing and dealing with arrays
import pandas as pd  #Package providing fast, flexible and expressive data structures
import re            #re stands for RegularExpression providing full support for Perl-like Regular Expressions in Python
from bs4 import BeautifulSoup   #Package for pulling data out of HTML and XML files
from keras.preprocessing.text import Tokenizer  #For tokenizing the input sequences
from keras.preprocessing.sequence import pad_sequences  #For Padding the seqences to same length
from nltk.corpus import stopwords   #For removing filler words
from tensorflow.keras.layers import Input, LSTM, Attention, Embedding, Dense, Concatenate, TimeDistributed   #Layers required to implement the model
from tensorflow.keras.models import Model  #Helps in grouping the layers into an object with training and inference features
from tensorflow.keras.callbacks import EarlyStopping  #Allows training the model on large no. of training epochs & stop once the performance stops im
import warnings  #shows warning message that may arise

pd.set_option("display.max_colwidth", 200) #Setting the data sructure display length
warnings.filterwarnings("ignore")
```

Figure 3: Libraries for Abstractive approach

4.2 Extractive Text Summarization

The Figure 4 shows the table of text and summary as 2 columns with multiple files. In Figure 5 shows the output summary generated from extractive model before feeding to abstractive summarization. The preprocessed data given to extractive approach and after the output it is combined and stored in a CSV to use as a input to abstractive. It consist of 2 columns and 60 rows as tere are 60 transcript given as input.

```
data.head()
```

	text	summary
0	hello hello hello . wh - what causes the crash ? did you fix something ? five five . hello hello . maybe it 's the turning off and turning on of the mike you think that 's you ? aaa - aaa	from what saw from the earlier results last week was that if you trained on one language and tested on another say that the results were relatively poor .example when we go from ti - digits t...
1	are we on ? we 're on . is it on ? why is it so cold in here ? we haven't sent around the agenda . any agenda items anybody has wants to talk about what 's going on ? could talk about the meetin...	the main thing would be if anyone has knowledge about ways to post - process the wave forms that would give us better recognition but just about the wish list item of getting good quality clos...
2	is that good ? 've have never handled them . goats eat cans to my understanding . did we need to do these things ? could hit - seven to do that ? the remote will do it cuz 'm already up there ? i...	like you might be able to " vista " it like you could haveif we moved onto the next step and did learning of some sort according bhaskara we 'd be handicapped .it 's talks about it just refers to...
3	mental mental palm pilot . hence no problem . let 's see . so . what ? 'm supposed to be on channel five ? her . nope . doesn't seem to be hello 'm channel one . what does your thing say on the b...	no it 's good idea that you may as ask .even without getting into it even though the scheme li is really documented in the festival .and see if he has any something already .what does your thin...
4	for two years we were two months away from being done . and what was that morgan ? the torrent chip . we went through it jim and went through old emails at one point and for two years there ...	wanted to take look at things that could model within word .so should we just do the same deal where we go around and do status report things ?wh why would that be considering that we actua...

Figure 4: Stage1 'text' and 'summary' CSV

```
actually , even though liz was kind enough to offer to be the first subject , felt that she knew too much ,
seems like you could put magic special ingredient in , so that everyone know which one was yours .
we need to so that 's one thing .
that probably is why of it that way .
just on the spur of the moment , and she was kind enough to serve as the first subject .

however there is always more people in in facul in department than are just taking his class or anybody else 's
how however suggest that if you if you look at your email carefully you may think you may find that you already
and it proved finally fruitful in the sense that we came up with new scenario for how to get the subject to rea
and how we can make it work for us .
that 's generally the way it 's done .

because , it might be easy to figure out that this person is going to need more film eventually from their utte
and so 've tried to come up with some initial things one could observe
so go - there in the first place or not is definitely one of the basic ones .
let 's first of all let 's see if it does influence anything .
one could go there 's

the we got to the point where we can now speak into the smartkom system , and it 'll go all the way through and then say s
so , the reason 'd like you to understand what 's going on in this demo system is not because it 's important to the resea
right now it 's brittle and you need to ch start it up and then make ts twenty changes on on seventeen modules before they
actually want , at least , maybe , you should be able to start it on your own .
and , , it 's not going to be problem because we decided to stick to the so - called concept to speech approach .
```

Figure 5: Summary generated from Extractive model

4.3 Abstractive Text Summarization

In this stage the CSV file generated after phase 1 is supplied to phase 2. After that cleaning is done for the text and summary. The Figure 6 shows the cleaning process.

```
[ ]  from nltk.corpus import stopwords
     stop_words = stopwords.words('english')

[ ]  stop_words = set(stopwords.words('english'))

     def text_cleaner(text,num):
         newString = text.lower()
         newString = BeautifulSoup(newString, "lxml").text
         newString = re.sub(r'\([^)]*\)', '', newString)
         newString = re.sub('"','', newString)
         newString = ' '.join([contraction_mapping[t] if t in contraction_mapping else t for t in newString.split(" ")])
         newString = re.sub(r"'s\b","",newString)
         newString = re.sub("[^a-zA-Z]", " ", newString)
         newString = re.sub('[m]{2,}', 'mm', newString)
         if(num==0):
             tokens = [w for w in newString.split() if not w in stop_words]
         else:
             tokens=newString.split()
         long_words=[]
         for i in tokens:
             if len(i)>1:                                 #removing short word
                 long_words.append(i)
         return (" ".join(long_words)).strip()
```

Figure 6: Cleaning of data in phase2

The Figure 7 shows output after the cleaning process.

```
[ ]  for i in range(10):
         print("Meetings:",data['cleaned_text'][i])
         print("summary:",data['cleaned_summary'][i])
         print("\n")

Meetings: hello hello hello wh causes crash fix something five five hello hello maybe turning turning mike think aaa aaa mine working gonna digits end
summary: _START_ from what saw from the earlier results last week was that if you trained on one language and tested on another say that the results we


Meetings: cold sent around agenda agenda items anybody wants talk going could talk meeting everyone everyone met agenda item one quick question know d
summary: _START_ the main thing would be if anyone has knowledge about ways to post process the wave forms that would give us better recognition but ju


Meetings: good never handled goats eat cans understanding need things could hit seven remote cuz already control control high tech yet another powerpo
summary: _START_ like you might be able to vista it like you could haveif we moved onto the next step and did learning of some sort according bhaskara


Meetings: mental mental palm pilot hence problem let see supposed channel five nope seem hello channel one thing say back nnn five alright five sibilar
summary: _START_ no it good idea that you may as ask even without getting into it even though the scheme li is really documented in the festival and se


Meetings: two years two months away done morgan torrent chip went jim went old emails one point two years thing saying two months away done believable
summary: _START_ wanted to take look at things that could model within word so should we just do the same deal where we go around and do status report
```

Figure 7: Cleaned text and summary code

The Figure 8 shows the distribution of text and summary through the histogram in which it helps to get maximum text length and maximum summary length. The text and summary are preprocessed and cleaned in the abstractive summarization (phase 2).

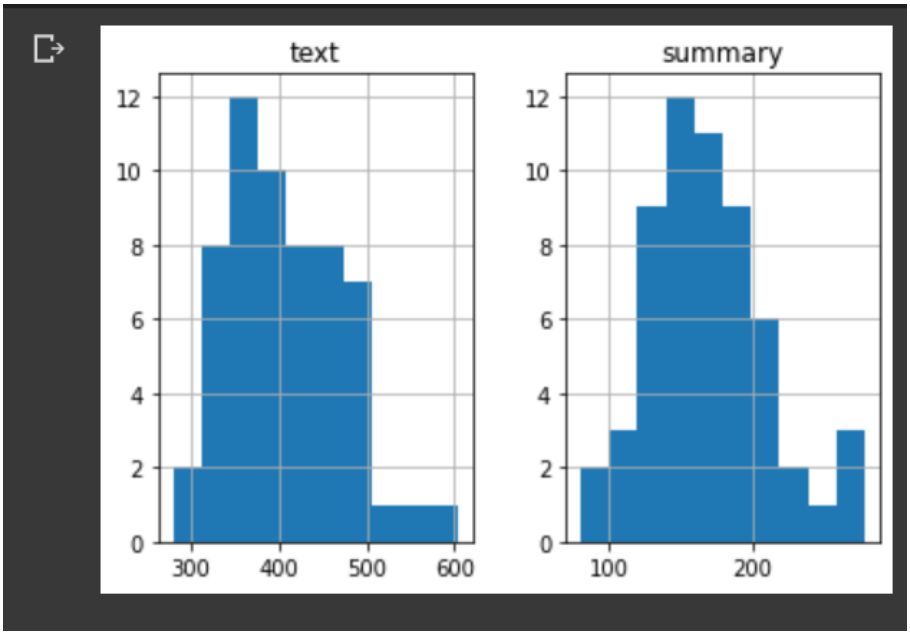


Figure 8: Distribution of text and summary through histogram

The Figure 9 shows the addition of tokens to the START and END by which it is easy to understand the starting and ending point of a sentence. This is done before feeding the data to the phase 2 model.

```
df.head()
```

	text	summary
0	hello hello hello wh causes crash fix something five five hello hello maybe turning turning mike think aaa aaa mine working gonna digits end channel channel three mm channel five mike number mike ...	sostok _START_ from what saw from the earlier results last week was that if you trained on one language and tested on another say that the results were relatively poor example when we go from ti d...
1	cold sent around agenda agenda items anybody wants talk going could talk meeting everyone everyone met agenda item one quick question know discussion previous meeting missed wish list item getting...	sostok _START_ the main thing would be if anyone has knowledge about ways to post process the wave forms that would give us better recognition but just about the wish list item of getting good qua...
2	good never handled goats eat cans understanding need things could hit seven remote cuz already control control high tech yet another powerpoint presentation makes easier johno let see one buttons ...	sostok _START_ like you might be able to vista it like you could haveif we moved onto the next step and did learning of some sort according bhaskara we be handicapped it talks about it just refers...
3	mental mental palm pilot hence problem let see supposed channel five nope seem hello channel one thing say back nnn five alright five sibilance three three see matches seat coming cuz starts count...	sostok _START_ no it good idea that you may as ask even without getting into it even though the scheme li is really documented in the festival and see if he has any something already what does you...
4	already got crash way crash feel much better earlier get door collected agenda collect agenda gonna go first mwa ha take long digits gone set thing ten right go pick ones problems either correct r...	sostok _START_ because if it was not it seems to me if you made it really specifically telephone groupings that maybe people would not go and do numbers so much would like to get away from having ...

Figure 9: Addition of START and END token

The splitting of the dataset is in the 70:30 ratio as shown in Figure 10.

```
[ ]  from sklearn.model_selection import train_test_split
     x_tr,x_val,y_tr,y_val=train_test_split(np.array(df['text']),np.array(df['summary']),test_size=0.3,random_state=12,shuffle=True)
```

Figure 10: 70:30 Dataset Split

The Figure 11 shows the Recurrent neural network in which an embedding layer for decoder and encoder networks, as well as an attention layer to memorize extended sequences, make up the model, which is a three-layer LSTM encoder and a one-layer LSTM decoder, and a function of SoftMax activation to the output layer. The embedding

```
from keras import backend as K
K.clear_session()

latent_dim = 300
embedding_dim=200

# Encoder
encoder_inputs = Input(shape=(max_text_len,))

#embedding layer
enc_emb =  Embedding(x_voc, embedding_dim,trainable=True)(encoder_inputs)

#encoder lstm 1
encoder_lstm1 = LSTM(latent_dim,return_sequences=True,return_state=True,dropout=0.4,recurrent_dropout=0.4)
encoder_output1, state_h1, state_c1 = encoder_lstm1(enc_emb)

#encoder lstm 2
encoder_lstm2 = LSTM(latent_dim,return_sequences=True,return_state=True,dropout=0.4,recurrent_dropout=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_dropout=0.4)
encoder_outputs, state_h, state_c= encoder_lstm3(encoder_output2)
```

Figure 11: RNN model

layers are 200 units while hidden layers are 300 units in size and the hidden layer has 0.4 value as a dropout to minimize overfitting and increase performance of the model.

The Figure 12 shows the epocs with 50 and batch size of 32.

```
[ ]  history = model.fit([x_tr,y_tr[:,:-1]], y_tr.reshape(y_tr.shape[0],y_tr.shape[1], 1)[:,1:] ,epochs=50,callbacks=[checkpoint], batch_size=32, validati
```

Figure 12: Epochs and Batch size

The Figure 13 shows the learning curves of the accuracy and loss of train and test data after running the number of epochs and batch size to identify how the model are trained for both train sample and validation data.

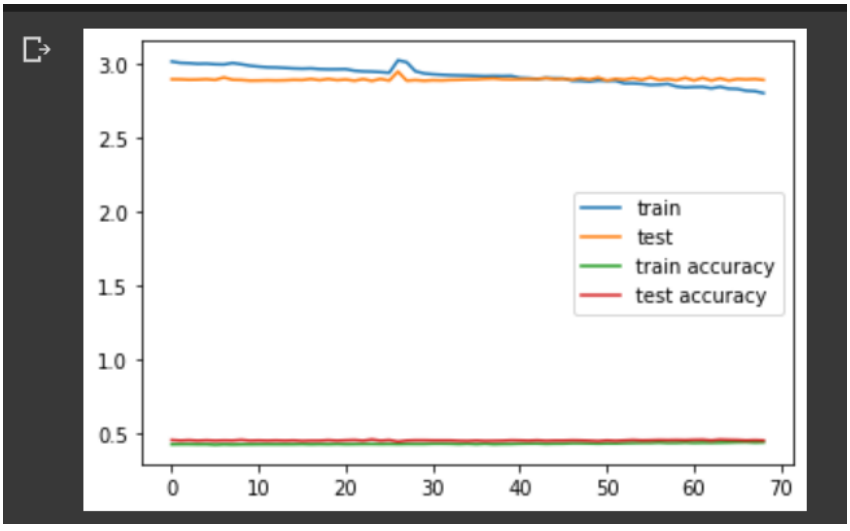


Figure 13: Accuracy and loss graph

The Figure 14 shows the final output summary which is generated from the hybrid model.

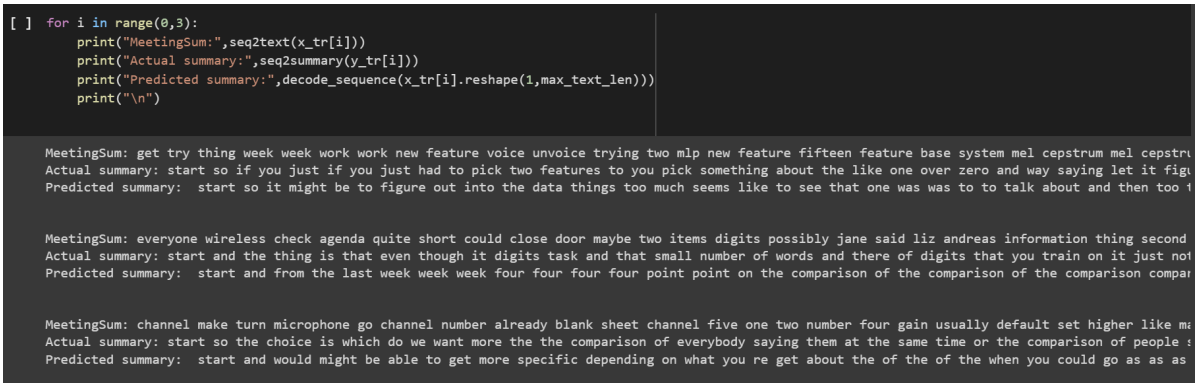


Figure 14: Final summary generated from Abstractive model

5 Evaluation

5.1 ROUGE Metrics

There are various metrics to evaluate based on content based, co-selection based, text quality based etc. ROUGE score for the text summarization is used to evaluate the reference summary with the generated summary. ROUGE scores are of different types like ROUGE N (ROUGE1, ROUGE2), ROUGE L, ROUGE S and ROUGE W. It states how much reference summary and actual summaries have similarity between them. Figure 15,16 shows ROUGE score calculation

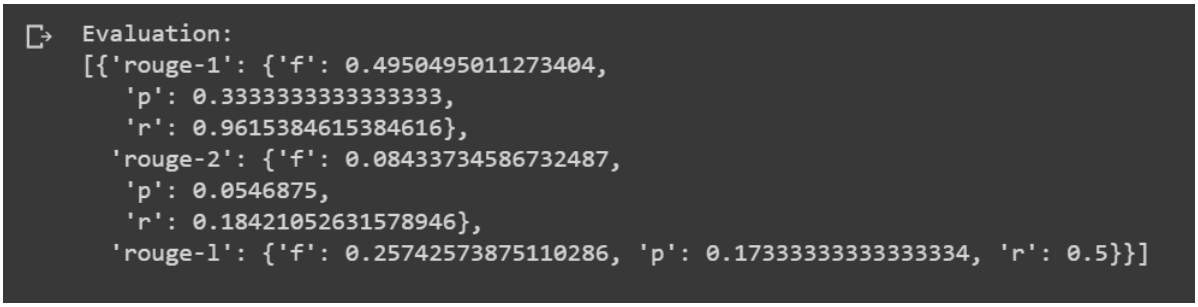


Figure 15: ROUGE Evaluation

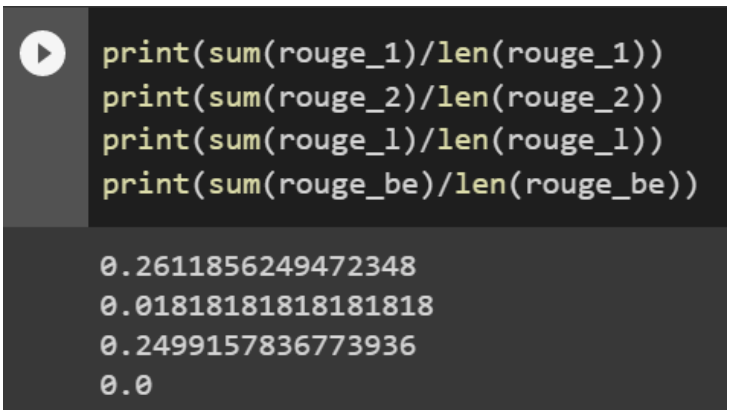


Figure 16: ROUGE 1,ROUGE 2,ROUGE L scores

5.2 Human Evaluation

There was also human evaluation done in which 5 people evaluated the summaries. As it was little difficult to read the long input text to know if the summary is generated correct or not. So, they reviewed reference summaries (extractive summary), actual summaries (abstractive summary), according to the ROUGE scores and mainly according to the human readability and understandability the output summaries was evaluated.

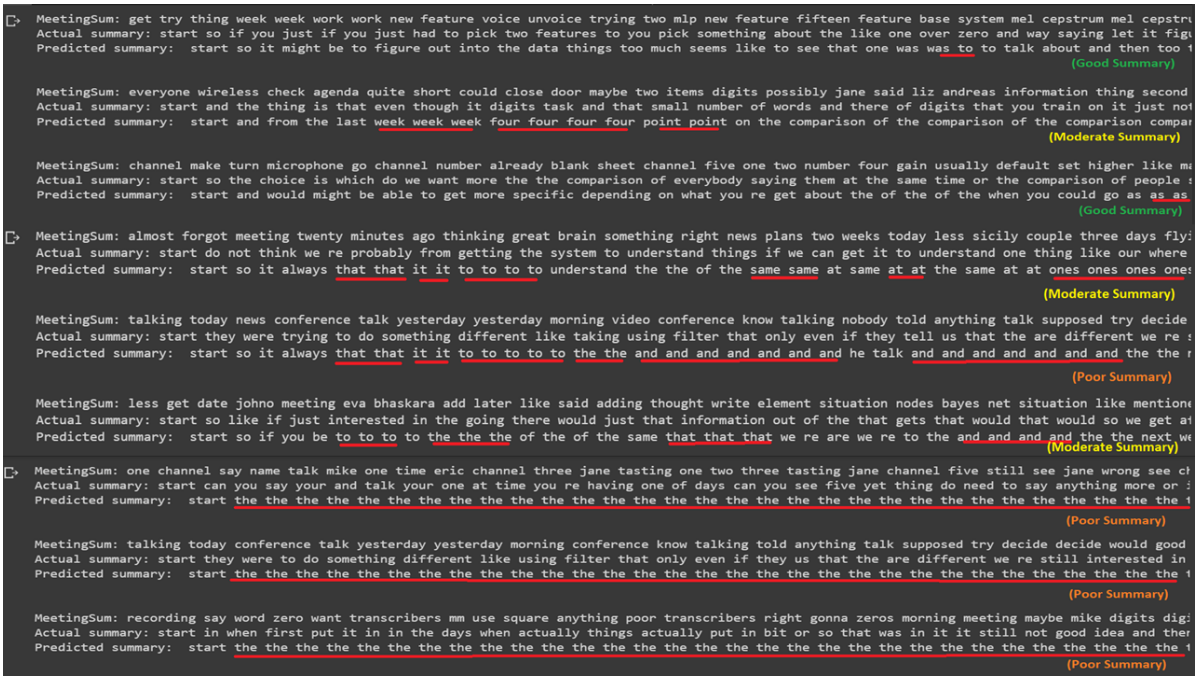


Figure 17: Human evaluation on final summary