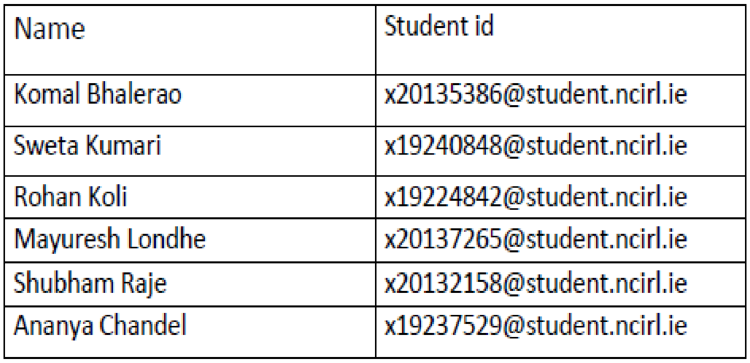
ENFUSE: Historical Text Summarization Using NLP and Pre-Trained Models

Project Report Document
Data Mining and Machine Learning 2
MSc Data Analytics
Prof. Michael Bradford



Name	Student id
Komal Bhalerao	x20135386@student.ncirl.ie
Sweta Kumari	x19240848@student.ncirl.ie
Rohan Koli	x19224842@student.ncirl.ie
Mayuresh Londhe	x20137265@student.ncirl.ie
Shubham Raje	x20132158@student.ncirl.ie
Ananya Chandel	x19237529@student.ncirl.ie

ENFUSE: Historical Text Summarization Using NLP and Pre-Trained Models

Komal Bhalerao
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x20135386@student.ncirl.ie

Sweta Kumari
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x19240848@student.ncirl.ie

Rohan Koli
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x19224842@student.ncirl.ie

Shubham Raje
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x20132158@student.ncirl.ie

Ananya Singh Chandel
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x19237529@student.ncirl.ie

Mayuresh Londhe
MSc Data Analytics
National College of Ireland)
Dublin, Ireland
x20137265@student.ncirl.ie

***Abstract*—Human language interpretation and generation is one of the most complex problems faced and it remains an unsolved mystery. To address this issue, Natural Language processing (NLP) which is a subset of AI has been developed and used in various applications such as email filtering, sentiment analysis, language translation. Another important usage of NLP is text summarization which breaks down a large corpus of sentences into few meaningful sentence. This process largely reduces the read-times and important information is filtered and extracted in the shortest amount of time. In this research paper, we implement a NLP model for the company DMC Tours to help them with their tourism business. Large relevant historical texts such as books, articles which are relevant to the history of two countries India and Ireland are collected as a part of input for the summarization model. Next, we implement Sequence to Sequence summarization mode with attention mechanism which is trained using the new summary dataset to get relevant summary. We compare the model with three pretrained models (BERT, GPT2 and XLNet) and find that BERT model performed with superior results (ROUGE score of 0.5).**

***Index Terms*—NLP, LSTM, Attention Model, BERT, XLNet, GPT2, Tokenization, Encoder-Decoder**

I. Introduction

The tourism industry has an important role in developing a country's cultural values, traditions and showcases them to the world. There is no final product as such, still changes in this industry directly affect the country's GDP. Transport, Accommodation, Tourist Attractions, Food are the industries which are an essential part of Tourism. In the year 2019, the industry has contributed a total of 9.25 Trillion US Dollars in a global economy with 1.46 billion tourist arrivals worldwide. This results in creating employment opportunities, revenue and Infrastructure development of a country.

This industry has seen significant evolution from ancient times due to advances in technology. Mainly due to the invent of faster transportation means, the use of the internet to minimize bureaucracy and making global awareness. Due to this, new stakeholders emerged, which also altered the role of existing stakeholders. The Use of Data-driven methods also made its impact by improving the decision-making process such as management of logistics, pricing, comparison tools for pricing. The industry is comprised of both B2C and B2B relations. Now once the customer completes the payment, all the subsequent steps are completed automatically which includes booking of Transportation services, hotel reservations, food preferences. Thus, providing focused, pleasant and value-added services to a consumer as all the B2B relations are automated by integration portals.

The culture and history of the country play an important role in attracting tourist's generation after generation. United Nation World Tourism Organization (UNWTO) study revealed that 40% of tourist worldwide inclined towards visiting sites which has cultural and historical aspect attached to them. This is a part of the tourism industry that is more focused on exhibiting archeological monuments, cult and civil architecture, monuments of landscape, Rural Settlements, socio-cultural infrastructure, religious structures. Cultural similarities and historical events between the two countries also play an important role. The majority of sources of ancient history were primarily in the form of books. Tourist guides spend more of their time reading these books and accumulating historical knowledge, the impact of historical events on country and world, chronological order of events. Widespread use of the Internet has made these books available in digital readable formats.

DMC tours is a business that focuses on managing tours for international tourists arriving in Ireland and Immigrants who are settled in Ireland but want to get more connected and blended to the culture of Ireland by exploring the heritage and historical places. The tagline of the business itself depicts the kind of tours are arranged for the tourist. "Listen Deeply:

Create Stories”, Thus the focus is more given on providing information which will make tourist more aware of Ireland culture, heritage and history. This makes it essential for a guide to be conversant with the history, culture and heritage of Ireland by reading the relevant literature. Below Figure 1, represents the website of DMC Tour. Tourists can make the selection of tours according to their proclivity. This study focuses on finding cross-cultural relationships, events, places between two countries using available literature which is in the form of PDF. Initially, this study will be implemented for India and Ireland, the structure and flexibility of the study allows us to extend the study for different pairs of countries. The structure of the study is divided into two stages, In the first part of the study, all the related literature and keywords were gathered. The list of keywords will be used to find the presence of the keyword in the literature. After identifying the presence of the keyword, the text before and after the point of the match is extracted, i.e., about 25 lines before and after. The second part of the study focuses on summarizing the extracted text using natural language processing. This extracted text can also be used in adding content to the website. The NLP model is based on Encoder-Decoder with Bidirectional LSTM. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Bilingual Evaluation Understudy (bleu) and are used as quantitate ve evaluation metrics of the NLP model and Human-based evaluation is used as a qualitative method of evaluation.

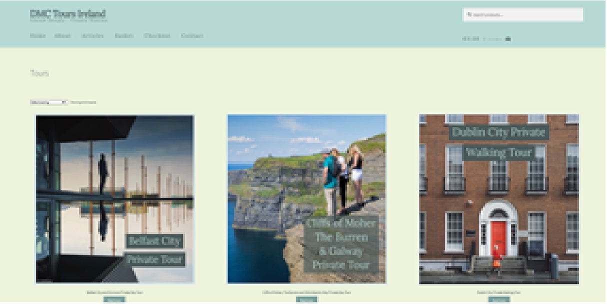


Figure 1. DMC Tours Website

The primary goal of the research is accurately summarizing the extracted text, which will provide a gist of the topic around the matched keyword. Therefore, there will not be a need for reading the entire literature. The indexing will help present the data in structured format thus searching information in extensive sources of literature will become easy.

The paper is structured in the following sections. Section I provides an introduction to the topic. Section II contains related work implemented in this domain. The methodology used for implementing the study and detailed work performed within the steps of the methodology is explained in section III. Evaluation of implemented model is described in Section IV. Section V contains the Conclusion and Future work of the paper.

II. Literature Review

Automatic summarization of documents available on web has become the most critical task ever since the generation of data available on web has increased exponentially. [4] proposed a novel approach to reduce human effort in summarizing documents by extracting the sentences represented by input document to process the generation of the summary. The process is divided into four phases distinctly including, Generating list of frequently used words, generating sentences, updating database, setting up web services. A extraction based approach is followed by implementing text summarization model. The conducted experiment showed the results of compression ratio to be 33% lesser compared to the input text.

In the age of blowing up information, to save people’s energy and time information available in documents should be summarized using the Automatic Summarization technology. [24] implemented a study of abstractive automatic summarization. Abstractive summarization is capable and consistent of human-like abstract writing providing simple, flexible and diverse use. In this reseacrh, implementation of Attention and coverage mechanism and pointer network along with comparative experiment of neural network to datasets of Chinese short text summary and english automatic summary. The results proved to be more efficient after adding location features and Batch Normalization with an apparent improvement to rogue evaluation index to the Seq2Seq mode.

In order to achieve better results in the domain of Automatic document extraction [22] proposed a Seq2Seq (Sequence-to-Sequence) architecture. To provide the system with additional information about the sequence of the input, bi-directional encoder is used to add extra advantage along with combination of two LSTM neural networks. The two types of decoder used in the implemented project are vanilla seq2seq one-way decoder for learning or initial phase and beam search decoder for interfacing. Additionally, relevant embedding layer and attention mechanism is used to support with text generation. The outputs of the research project is evaluated using different metrics such as Recall, precision, F1 score of ROUGE-1 with results as 5.5% Recall, 19.3% precision and 8.5% F1-Score. The overall results indicated that only a small portion of words are generated in the content to that of original text document concluding the abstractive feature to be strong.

[21] reviews the recent approaches for abstractive text summarization using deep learning algorithms. Additionally, validating and training existing datasets are reviewed along with limitations and features are presented. The challenges along with solutions are discussed. Some of the challenges that are addresses with probable solutions such as Out of Vocabulary(OOV) words that can be solved by employing pointers to point to actual place in the original document, Fake facts, Inaccurate information and Summary sentence

repetition, Unavailability of reference summary tokens that occurs while using small datasets can be addressed using DaD (Data-as-Demonstrator and other challenges that has not been addressed are also discussed.

RNNs and other Seq2Seq models are robust and one of the best approaches for a e-Commerce setting proved by [11] by proposing a novel implementation of Seq2Seq abstractive and extractive summarization model using Document-context. The high-level human understanding way of understanding document by reading title, abstract and other information instead of reading entire document is incorporated using Seq2Seq model generating rich level summaries specific to documents. Training is implemented using both supervised and semi-supervised setting to extract summary. In conclusion, the results derived from the implementation suggests that seq2seq based RNN technique of summarization out-performs other techniques where Extractive-context RNN depicts better performance on small data whereas, Abstractive context RNN depicts significant difference of improvement using large scale data.

A review conducted by [14] discusses and summarizes the currently applied and potential future application of text mining of big data techniques focusing on the tourism industry. The main take-aways included in the paper suggested that deep learning can be used for enabling text and sentiment features of words for exploiting and understanding short texts. Topic model is one of the basic models used in many topic extraction models. Co-training, meta-learning, transfer-learning and training samples are some of the strategies proposed particularly focusing on abundant labeled data as a requirement for supervised learning.

A review conducted by[5] depicts abstractive text summarization techniques using Natural Language Processing. As increasing amount of articles, papers and documents these days it is very difficult to extract important things from the huge amount of text data. This issue can be handled using Text Summarization technique. Basically Text Summarization is the technique which converts large amount of text into short and meaningful sentences. Using various Machine Learning approaches, text summarization technique can be accomplished to understand the large amount of text document at first and then further to create a summary from it. In this paper Abstractive text summarization is done using Long Short Term Memory network and Recurrent Neural Networks to extract prime words from the text document or creating human-like sentences to form a summary from it. Text summarization technique generates comprehensive summary which saves time and effort. So, in this paper another two methods are used to accomplished the task such as encoder-decoder model and pointer generator mechanism. For results ROGUE i.e. Recall-Oriented Understudy for Gisting Evaluation scores were compared between this methods based on thier outputs. ROGUE 1 and ROGUE 2 reffered as overlap of bigram and unigram between the system. Whereas, ROGUE L is calculated for longest common subsequence statistics.

Research conducted by[13] have been studied an unsupervised approach for extractive multi-document summarization based on sentence embedding and centroid approach. Extractive multi-document summarization is a technique of automatic text summarization from a collection of documents to extract important information from it. Sentence embeddings is an effective technique for several natural language processing tasks for Automatic text summarization. This paper proposes an unsupervised method for generic extractive document summarization using sentence embedding representations and centroid approach. The proposed research choose relevant sentences according to the score obtained by three score such as sentence novelty, sentence content relevance and sentence position scores. This paper provides comparative analysis of nine different sentence embedding methods for extractive multi-document text summarization. Analysis was done on two datasets such as DUC'2002–2004 benchmark and Multi-News dataset. The results of this analysis showed that the use of sentence embedding method is effective in extractive document summarization. Also models of centriod approach achieved comparable results in this proposed research.

In research[19] Automatic text summarization model is performed using Sequence 2 Sequence technique for summarizing text from the document. From past few years text summarization has became topic of research. Various methods of Natural language processing enables researchers to generate effective results with the large amount of documents. In this research Seq2Seq technique is used with Recurrent Neural Networks to perform abstractive text summarization. The data was collected from harvard NLP project, comprises two datasets such as Gigaword dataset and CNN/DM dataset, Where CNN/DM dataset contains news articles and hand written summary. Data was trained using TensorFlow and Seq2Seq architecture. The model performed have showed effective results on summarization of text on larger and legal documents. The results generated from summary generation and ROUGE scores ranges from 0.6-0.7, which is very good. Along with the abstraction technique, extractive method was also implemented which can be used to do comparison between two methods.

Conducted research by[9] have implemented automatic text summarization using Gensim Word2Vec and K-Means clustering algorithm. Due to increasing virtual textual materials, significance of text summarization in the field of Natural Language Processing has expanded. Text summary is a operation which is generated from one or more text documents, which gives insights in a small form from the main text. This reduces time and efforts which is required to read a whole document. Extractive and abstractive are two main techniques for summarizing the text. In this research sentence based clustering algorithm i.e K-Means is used for a single document. Whereas, Genism Word2Vec algorithm is used for the feature extraction. Which automatically extracts important topics from the large document in the efficient way. K-Means clustering is used for clustering the sentences. This methods are implemented on BBC news articles database. Sentence scoring algorithm is performed on this dataset. The results were more effective on business related articles because the business articles have more numbers in it and sentence scoring algorithm gives greater importance to numeric values. BLEU score have been seen in this model and the result are between 0 and 1, where 1 is best similarity and 0 is the lowest similarity.

A survey conducted by[7] presenting Automatic Text summarization for extracting useful information from huge documents. Automatic Text Summarization (ATS) is on fleek these days because of large amount of textual content is growing on internet in the form of articles, scientific papers,legal documents and many more. Manually text summarising consumes lot of efforts, time and cost and even becomes unfeasible sometimes because of large amount of textual content. Automatic Text Summarization have extractive, abstractive or hybrid approaches. The hybrid approach combines both abstractive and extractive approaches. Various evaluation techniques used to test the performance of model such as Precision score metric, Recall score metric, F-Measure score metric and ROUGE metric. Results generated by the model showed that the model performed well and the results are shown by all this evaluation techniques.

Research conducted by [23] is a comparative study on abstractive text summarization. Text summarization have became most interesting and vital research point in the area of Natural language Processing. Text summarization is divided into two groups such as Extractive Text Summarization and Abstractive Text Summarization. ETS is more easier than ATS. ETS is based on algorithms, which extracts important words or sentences from the large text document. Whereas, ATS generates summary by itself. Three kind of techniques are applied in this research such as Word Graph Methodology, Semantic Graph Reduction Algorithm and Markov Clustering Principle. Results generated by all algorithms are very efficient. Accuracy rate is above 70% for all the three techniques which is very good.

In past few years it has been seen that a lot of information is being dumped online and through various sources. Hence, we have seen surge in information [6]. Almost every organization is having their own websites to contain essential information about the organization. Similarly for educational institutions the websites are containing loads of information and processing that information all at once can be tough for students and their parents. To get the query answered at times it takes loads of time and efforts by students, academicians, and parents. In order to cut down on time and efforts, for searching relevant information on the websites without reading through several sub-pages. A paper presented a model to retrieve accurate and relevant information about the query without going through all the information. The model used Natural language processing (NLP) to perform text summarization for giving relevant results. Clustering algorithm and Hybrid similarity measure was used for extracting relevant data according to the query and removing the redundancy. The results achieved by the model were accurate for 86% of the times.

Similarly, humans are accustomed to remembering only the important point in an article which they find relevant. Like a synopsis of a book or the gist of the book can help buyer in understanding and making the decision. Likewise, a product review can help a buyer in making the buying decision. In a paper text summarization is used to help the buyer decide by providing a concise version of reviews to the buyer. [20] uses an abstractive text summarization in which the extraction picks uo the words and later abstracts and summarizes it. This involves Natural Language processing and Machine learning to replace the traditional reviewing system without hurting the sentiments of the reviewer. The model could be further improved by taking entire paragraph at a time by using window model.

To answer questions at court it is of utmost importance for the ordinary citizens and lawyers to do an exhaustive research on their related case and be prepared. Since the judgements are long and exhaustive at times, they have to hire legal editors so that they can summarize the long judgements for their understanding. In a research an automated text summarization is created for performing a similar task by generating short summaries from the long judgements. To achieve the goal Natural language processing technique is used, it is called latent semantic analysis (LSA) and it captures key concepts in a single document. Two approaches have been used by [16] one is using a single document untrained approach and another one is using a multi-document trained approach. The data for the analysis was collected from the official government sites. Latent Semantic Analysis (LSA), also known as Latent Semantic Indexing (LSI), is an unsupervised statistical-algebraic summarization technique that is entirely automated. that uses an extractive approach to document analysis to uncover secret semantic relationships between words and sentences in the text. The model achieved an average ROGUE-1 score of 0.58.

A pre trained language model can be used to express text span or semantics of a word, it has wide applications in the field of NLP tasks and text summarization. Since bidirectional encoder from transformers lot of work and many models for text summarization have been based on BERT and fine-tuning parameters end-to-end. Multiple researches have been carried out to enhance or created different versions of BERT for achieving state of the art performance to handle Natural language processing tasks. Study by [8] explores the different versions of BERT for handling text-summarization. The paper presented a two-stage encoder model for summarization and extraction. First stage introduces a Lite BERT model for securing embedding at sentence level and identifying the content that may be valuable based on a Lite BERT. While the second stage involves extraction of meaningful document embeddings using a fine tune BERT strategy. Lastly, it involved selecting the combination of best matched combination of sentences along with the document source for composing summarization.

With the ever-increasing data and the data available in the form of text is in abundance. Extracting useful insights from the data is extremely important similarly extracting useful insights from large texts can be very useful. We come across large number of websites, customer reviews and blogs with long articles and huge texts extracting important information to gain insights is a necessity. Several papers have been published so far using different techniques for performing text-summarization. [1] researched several methods for text summarizations have been mentioned like abstractive and extractive. Query based summarizations have also been discussed. The paper discusses about structure based and semantic based approach in text documents for summarization. Several datasets were used for the study including DUC2000, CNN corpus and other multiple and single text documents.

Due to COVID-19 pandemic the medical communities have been extremely busy.[12] has done research on where Medical communities have to be updated with corona virus updates and the new literature that is generated every day. Due to this problem corpus scholarly articles released a competition Open research dataset challenge to get help in getting machine learning approaches to get solution. In similar regard a research focused on using pre-trained NLP model, OpenAI GPT-2 and BERT were used for performing text summarization on covid related documents to prevent users from reading the whole covid literature. The model was evaluated using visual inspection and ROGUE score. The model provided comprehensive information and abstractive summary; the summary was based on keywords that were extracted from the original articles. These kind of text summarizations can help the community in this difficult time. This work motivates our research for using GPT-2 method over BERT for text summarization of our project.

As per author [17] from last two decades using text summarization on lectures has been used. Using text summarization on lectures can be useful as reading through all the lecture notes can be tedious using extractive text summarization on lecture notes can help in finding key phrases and important sentences. Clustering output embeddings via deep learning and new machine learning approaches have provided us with extractive text summarizations. In a research a python based RESTful service was used for lecture summarization service. The model utilized BERT model for using K-Means clustering and text embeddings for extracting sentences that are closest to the summary selection. BERT resulted in extractive text summarization with promising results. This research showed potential for using BERT method as one of implementations for our project.

The research implemented by author [3] performs classification of twitter tweets into optimistic and pessimistic. The deep learning architecture used to perform the research is XLNet. The research has outperformed state of the art model by 6% increase in accuracy from 90.32% to 96.45%. The initial step of the model building was Language modeling for that pre trained networks GPT and ELMo were implemented. XLNet model was fine-tuned using CLS. Data is divided into 80% training and 20% testing. The data is collected from 1000 different users but count of tweets in each class are imbalanced. There is no information regarding users consent before utilizing the tweets for research.

Research [2] has performed comparative study using 4 different models for accurate detection of emotions using text data. Models used in this study are XLNet, BERT, DISTILBERT, RoBERTa. Dataset ISEAR is used to fine tune all the models. Each model detects 7 different emotions. Dataset contains equal number of sentences for each emotion thus bias due to data imbalance is avoided. The RoBERTa model achieved the highest accuracy among the implemented models that is 74%. There is more Scope to generalize the model to improve the accuracy. If we look at accuracy of individual emotion the accuracies are varied that needs to similar in all emotions.

III. METHODOLOGY

This section throws light on the adopted methodology for this project on implementation for solution provided to business.Here for our business problem where we suggested a customized model which later bifurcated into two parts one focuses on Text Search Using a dictionary and another one is abstract text summarization which summarized the text obtained from former part with machine/deep learning with usage of NLP-Natural Language Processing,RNN - Recurrent Neural Network,LSTM etc, hence building the efficient model is an imperative step.

As [15] authors discussed on evolution and how using CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology is significant in building data mining project and most used and widely accepted analytical methodology and how still various data mining paradigm still applicable for many range of data science research projects and as this project is based on business requirement to help the business in achieving goal using deep learning hence considering CRISP-DM over KDD was important.

It is comprises of six steps:

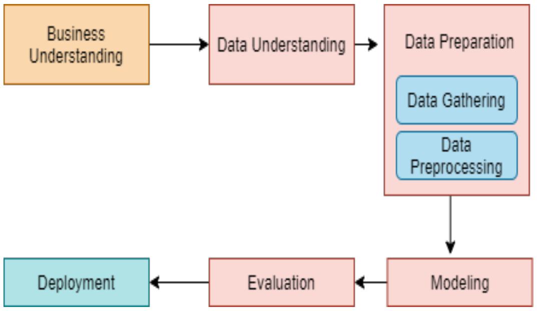


Figure 2. CRISP-DM Methodology

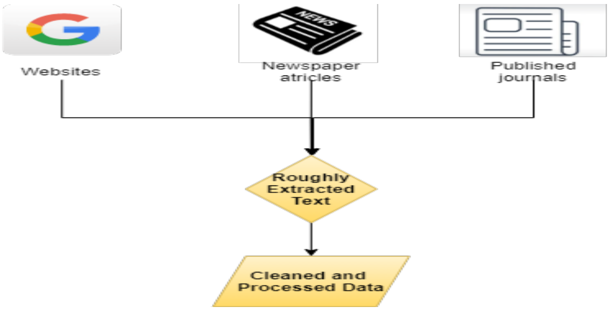


Figure 3. Data Gathering Process

A. Business understanding – Focused on What does the business/client need?

B. Data understanding – What data we have what it does? Is it clean for next step?

C. Data preparation – How does we organize/scrub/pre-process the data for modeling? The Data-Preparation is further divided in two categories mentioned below:

1) Data Gathering: The initial task for this assignment was to gather historical data that may have some co-relation between two countries India and Ireland and as the future scope of this project similar co-related data can be found between different countries for performing abstractive text summarization. Historical data has been gathered in the form of texts from journals, articles, newspapers, websites, and books. To begin with we have used different search strings for finding the relevant texts that may be related towards the goal of our work. All the data collected in the form of text is made sure to be collected from publicly available sources hence not violating any ethical concern. The data collected from different sources was cleaned and filtered to remove the un-necessary texts. Few data sources contained images and captions, all the images and captions were cleaned manually. After all the data was cleaned to obtain simple text, the data was collected and collated to a single pdf document. This pdf document now contains clean historical text data for any co relations that could be found between India and Ireland.
The model build for this paper will be using this collated historical text document to parse search strings or keyworks in order to pick relevant co-related connection between India and Ireland and will help the reader to pinpoint the query instead of going through long historical texts or at times even book. Further this data can be extended by using other historical sources like online libraries, books etc.

2) Data Pre-processing: In the above section, we looked at how and which type of data was gathered to perform the research. Now we will look at data pre-processing methods used to make data ready for further analysis. Keyword datasets were gathered from varied sources therefore structure of each keyword document is different. Four datasets contain keywords dictionary of Lakes, Mountains, Heritage and Towns land.
Here we need to first remove the irrelevant columns from each CSV file. Instead of pre-processing each file at a time. We created a dictionary that will store the Data Frame name and columns to be removed from that Data Frame. Data Frame names and Columns to be removed will act as key-value pairs of the dictionary. Now we created the function columnremoval() which takes three parameters as an argument and returns the Data Frame which contains a single column of keywords. These three parameters are Data Frame from which columns need to be removed, a dictionary that contains key-value pairs and Data Frame name which is key in the dictionary. The returned Data Frame is overwritten on the original Data Frame. This creates a well-structured and scalable method to remove the irrelevant columns. If the tourist guide wants to increase the dictionary of keywords it will be easy to just specify the key-value pairs in the dictionary and make a function call.
All four dictionaries contain different names of the columns in which keywords are present. For better understanding, we renamed those with one single name. To make this change, rename() function of Pandas Data Frame is used. Now we can merge all the keywords from four different files into one single CSV file. To perform this step, append() function is used. The merged pdf contains duplicates keywords, if these keywords are not removed the processing time for searching the keywords in pdf literature will increase as well as there will duplicate entries for extracted surrounding text. This is done using drop_duplicates() function of pandas. Now last step of dictionary pre processing is converting all the keywords to Lowercase format for that str.lower() is used.

Below is the data distribution of Text_Title and Text_Summary for News Summary dataset after data is clean.Below Figure 5 for reference based on length of content and title was plotted using 'Seaborn' library 'distplot' function.Hence below analysis gives us firm decision to chose optimal fixed length of News summary and New title.

For tokenizing and vector distribution we have add 'sos' termed as start of string and suffix with 'eos' termed as end of the string before feeding the data in to the model as it does helps the decoder to understand when to start generating output.

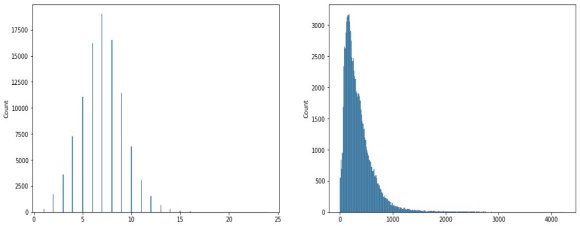
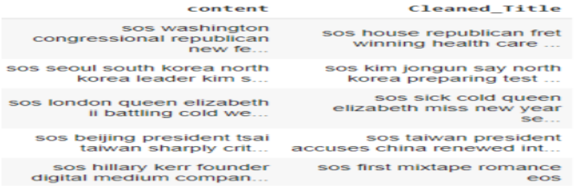


Figure 4. Data Distribution plot based on Length



content	Cleaned_Title
sos washington congressional republican new fe...	sos house republican fret winning health care ...
sos seoul south korea north korea leader kim s...	sos kim jongun say north korea preparing test ...
sos london queen elizabeth ii battling cold we...	sos sick cold queen elizabeth miss new year se...
sos beijing president tsai taiwan sharply crit...	sos taiwan president accuses china renewed int...
sos hillary kerr founder digital medium compan...	sos first mixtape romance eos

Figure 5. SOS and EOS addition tokens

D. Modeling – Focuses on most efficient modeling techniques to apply to achieve the requirement?

E. Evaluation – Evaluation to reaffirm Which model best meets the business objectives and with quantitative and qualitative method of evaluation.

F. Deployment – How to deploy the code and provide to business and access the results?

IV. Modeling

Post Data pre-processing modeling the next step as per CRISP-DM Methodology is Modeling before which environment to run the model was setup. For this project we have done implementation using Google Colab's which Cloud based Jupyter notebook - pro version (Limit of 25 GB RAM and 125 GB Disk Space) runs on Google Chrome where as Dataset was stored in Google Drive as training model as with deep learning techniques required here high computational power. After the Google drive is mounted we have installed various to be used libraries and packages to implement in our project analysis (Packages like Tensorflow,Keras,NLTK,Seaborn,Matplotlib,Numpy,Pandas etc)
Below is the Figure:6 for the reference of Model Architecture used for this project.

As two data source one for testing and other training prepared differently the training dataset is the news summary article data set has been sourced from Kaggle[1] with three available different csv files hence we merged them using program into one single file and used further for analysis.These three files were initially uploaded into Google drive taken from there for using 'pandas' and concatenation lead to convert them into one single source for training the model where as for testing or Business requirement data was manually exported and later converted into paragraph format using python.

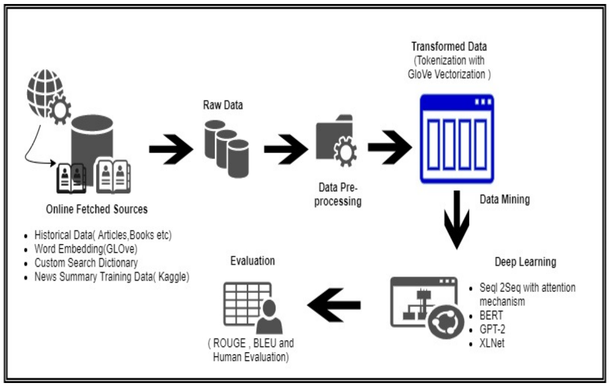


Figure 6. Model Architecture

Training dataset was divided into trained and test dataset initially and later used on business dataset prepared separately.The dataset used for training model consists of news article from 15 different publications and has different article according to event and publication's before feeding data into the model before and after of text string was added with 'sos' and 'eos' at the end respectively which helps the decoder to identify start and end of the sequences.

Training and test data is divided in the ration of 80:20 then data taken to splitting and tokenizing the data which later passed through Glove embedding matrix of almost 100 dimensions to get the vector representation of the words from the used data.As seen above in the 6 image we have made an optimization block for saving the words using GloVe embedding matrix and discarded missing words to compress the size of vocabulary for smooth computation with better results. Below is our project code generated architecture Figure:12 for detailed architecture view.

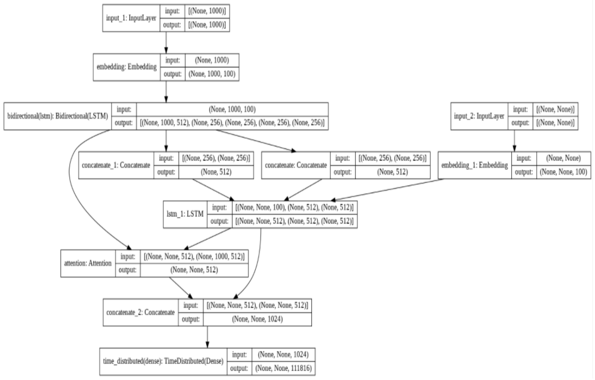


Figure 7. System Generated Model Architecture

For this first experiment modeling we have used two different architecture with different mechanism under encoder-decoder framework and also attention mechanism can be termed as Seq2Seq model along with separately added attention mechanism. Here 'Adam', 'RMSProp' optimizer are preferred under optimization block.

[1] https://www.kaggle.com/snapcrack/all-the-news

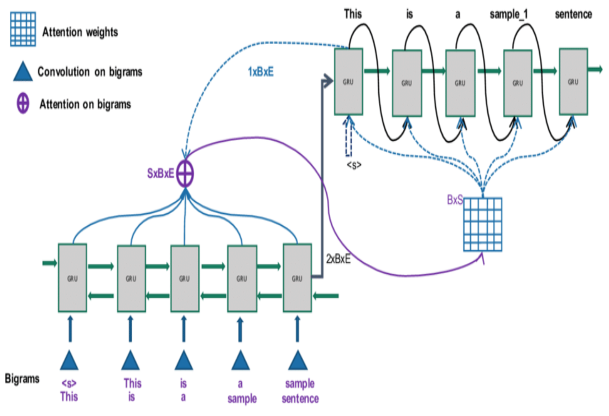


Figure 8. Encoder and Decoder Architecture for Seq2Seq WSD with attention on bigrams.(S represents Sentence length, B represents Batch size and E represents Embedding dimension)

Post compilation we found in the summary that same hyperparameters used in the stacked LSTM. Under encoder we kept three LSTM layer and one LSTM at the decoder block side.Post compilation we trained model using model.fit function and also early stop function used to prevent over fitting issues and led to save the best model.10 epochs are set and below is the diagnostic diagram of loss plot of both stacked as well as Bidirectional LSTM.

Below Figure represents the loss over the number of epochs which is shown from loss plots of both Bidirectional and Stacked LSTM architectures.

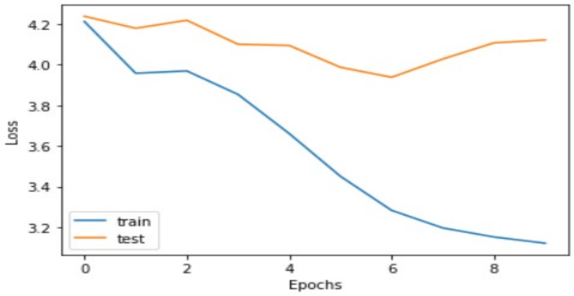


Figure 9. LOSS Plot of Stacked LSTM

Finally decoder block generates outputs each time and step and provide the headline/summarized output and this decoder architecture has a feature of generating words openly from the created vocabulary earlier where as hidden states both encoder and decoder goes into attention layer which combines the states and with the help of softmax layer generates attention scores.Highest score words will be taken into next output and that's the mechanism how summary created.

In the next experiment we have considered State-of-art language pre-trained model under NLP specially for text summarization such as BERT,GPT2 and XLNet.

BERT - Bidirectional Encoder Representations from Transformer's is transformer-based machine learning technique under NLP pre-trained developed by Google in 2019.[2] was published by Jacob and his colleague modelled with corpus of data.BERT is comprises of two models (i) the BERTBASE has 12 Encoders and inclusion of 12 bidirectional self-attention heads. (ii) the BERTLARGE has 24 Encoders with the 24 bidirectional self attention head. Above models are pre-trained using unlabeled extracted data with corpus of data from varied domains and dictionary with 800M words and English wikipedia words upto 2500M words.One of the research [10] where the experiment was performed to execute abstractive text summarization on Japanese text using BERT encoder and Transformer-based decoder using livedoor corpus data approx 130,000 datapoints.Here BERT has performed well and able to learn efficiently.However there were places where repetition of summary sentence and unknown words were not able to handle could be due to word mistakes this led us to give a try on our client formed dataset which has performed decent and tested.

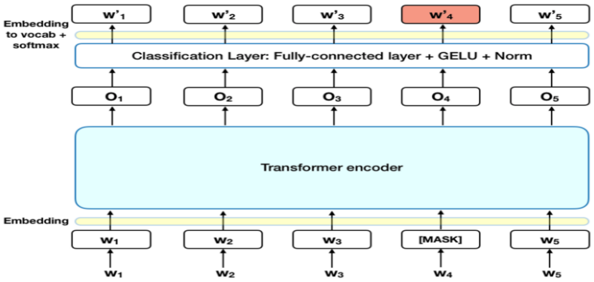

W'1
W'2
W'3
W'4
W'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
W1
W2
W3
[MASK]
W5
W1
W2
W3
W4
W5

Figure 10. BERT Architecture

XLnet[3] is Denoising autoencoding has shown the potential to model bidirectional contexts. Related pretraining, such as the BERT, outperforms pretraining techniques based on autoregressive modeling for language. Although, when model relies on inputs with mask, BERT ignores the relation between position that are masked hence, pretrain divergence is suffered because of it. Looking at the pros and cons a XLNet is proposed.XLNet is an autoregressive pretraining methodology, it overcomes the limitations posed by BERT because of its auto-regressive design. Bidirectional context is enabled by XLNet XLNet makes it possible by optimizing the expected probability across all contexts over a variety of factorization order permutations. Transformer XL which is a state-of-art autoregressive model, XLNet adopts and integrates its ideas for pre training.[**?**] In some experiments conducted in a research XLNet outperformed BERT. Under certain experimental settings that were comparable XLNet was able to outperform BERT for 20tasks, with a large margin for tasks including, Natural language processing, question answering and sentiment analysis. Hence, XLNet seemed to be an optimal choice for the implementation of our model.

[2] https://en.wikipedia.org/wiki/BERT$_{(}language_{m}odel)$
[3] https://cloud.google.com/tpu/docs/tutorials/xlnet-2.x

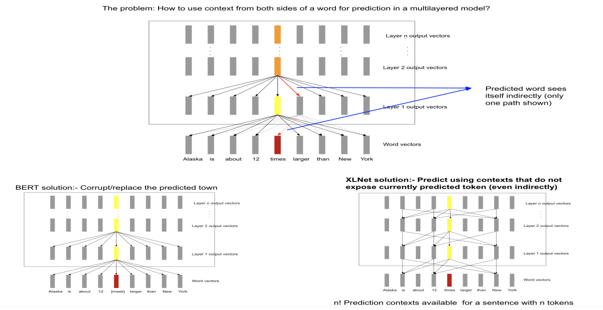

The problem: How to use context from both sides of a word for prediction in a multilayered model?
Layer n output vectors
Layer 2 output vectors
Layer 1 output vectors
Word vectors
Predicted word sees itself indirectly (only one path shown)
Alaska
is
about
12
times
larger
than
New
York
BERT solution:- Corrupt/replace the predicted town
[mask]
XLNet solution:- Predict using contexts that do not expose currently predicted token (even indirectly)
n! Prediction contexts available for a sentence with n tokens

Figure 11. XLNet Architecture

GTP: stands for Generative Pre-trained Transformer.GTP-2 is a large model contains huge amount of compressed knowledge. Which can be used for predicting the probability of sentences and for text auto-correction.GTP-2 is pre-trained model which is used to generate coherent text which is very efficient. This is architecture which is very identical to decoder only transformer. GTP-2 which is very huge and transformer based language model was trained on huge datasets in past. The smallest version of trained GPT-2 takes upto 500 MBs of storage for storing all of its parameters. Whereas, largest variant can take up more than 6.5GBs of storage. BERT architecture is built using transformer encoder blocks. Whereas, GTP-2 is built with transformer decoder blocks. GTP2 is model which is auto-regressive in nature. Whereas, BERT is not. We are using this architecture in our research project and it has implemented

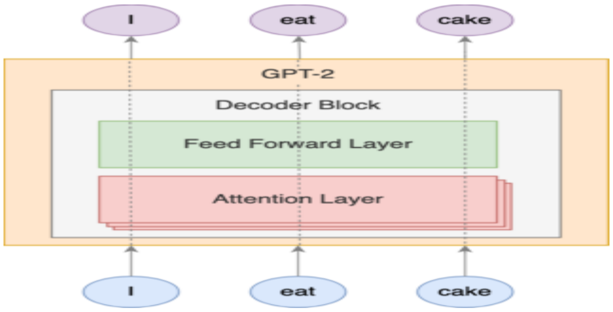


Figure 12. GTP-2 Architecture

V. EVALUATION

Our motive is not only to achieve the abstract output summary but also to evaluate the result is crux of business requirement and significant step before deploying it to business environment.Evaluation gives us the detailed performance over modelling performance in terms of prediction.As our project is based on Abstract Text Summarization after implementation of the models on dataset, its indispensable to measure how well the model has performed.Here we have divided thee evaluation into Quantitative and Qualitative analysis accordingly.

A. Quantitative Analysis

1) ROUGE Metrics: For text summarization related evaluation specially on NLP models ROUGE metrics is considered as state-of-art.ROUGE(Recall-Oriented Understudy for Gisting Evaluation metrics) is based on recall measure which calculate the n-gram overlaps among the output generated summary and the existed reference summary.As we know, there are several types of rouge metrics such as ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-N, ROUGE-L with their own speciality.Author [18] have implemented ROUGE in evaluation and found RMSProp optimizer among other two outperforms and has the highest R-1,R-2 and R-L scores hence higher the score better the performance model is. and able to .

For this project work ROUGE-1, ROUGE-2 and ROUGE-L are taken and considered further.ROUGE-1 is the based on overlapping of unigrams and when it bigrams overlapping then we reference it as ROUGE-2,lastly ROUGE-L spotting the longest most common sequence (LCS) occurring in n-grams.Below R_n is referred as ROUGE-N, S is Sentence,RT_{set} are basically the summarise reference m-grams and $count_{match}$ counts the maximum numbers of n-grams, N is n-gram's length.

$$R_N = \frac{\sum_{S \in RS_{set}} \sum_{gram_n \in S} Count_{match}\,(\ gram_n)}{\sum_{S \in RS_{set}} \sum_{gram_{n \in S}} Count\,(gram_n)} \tag{1}$$

The generated title and the original title are compared using ROUGE evaluation metrics. The ROUGE scores for Bi-directional LSTM model are as shown below and calculated using Sumeval package in python.

2) BLEU: BLEU - Bilingual Evaluation Understudy metrics is another renowned and latest method used in text summarization.It usually helps to tell how much the words (or n-grams) in the model generated output summaries appeared in the provided human references summaries.As [10]here author has implemented and able to evaluate the extractive summarization well hence we will be using this method for evaluating the final generated summarized data.

$$BP = \exp\left\{\min\left(0, \frac{h-R}{h}\right)\right\} \tag{2}$$

BP-Brevity Penalty:Existed to penalize short hypothesis given as

$$BP = \exp\left\{\min\left(0, \frac{h-R}{h}\right)\right\} \tag{3}$$

where h is the length of hypothesis and R is termed as referenced length.

Hence overall BLEU score is n-gram precision times BP (Brevity Penalty) which is given as below

$$BLEU = BP \times P \tag{4}$$

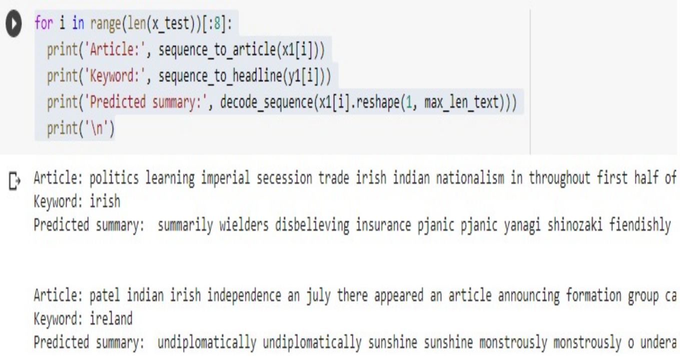
The BLEU score below shown figure is calculated from NLTK library.

Below is the Evaluation Matrix for reference :

Model	BLEU	Rouge -1	Rouge -2	Rouge -l
BERT	0.5470	0.5038	0.4821	0.5038
GPT-2	0.5511	0.4682	0.4443	0.4682
XLNet	0.5598	0.4546	0.4319	0.4546
Seq2Seq	0.6156	0.1928	0.0448	0.1821

B. Qualitative Analysis

Finally human evaluation performed on the output summaries of test dataset to confirm if the prediction summary with highest ROUGE Score able to produce high quality and more sensible summaries.Here our team evaluated randomly selected samples from validation dataset.We tried to analyzed the original text with actual summary and system predicted summary and then implemented same on test dataset to retrieve the appropriate the summarized textual data.Later we have divided the section into good and Moderate based on rating.The human evaluation results showed that nearly 20% of summaries were rated as 'Good' whereas about 80% into 'Moderate'. Below is generated summary for client tested output where we can see generated summarised data is not proper can observed many repetitive word and sentence is not proper hence we need to optimize and train model with other datasets and also customise the existed model with advance feature into it.



```
for i in range(len(x_test))[:8]:
  print('Article:', sequence_to_article(x1[i]))
  print('Keyword:', sequence_to_headline(y1[i]))
  print('Predicted summary:', decode_sequence(x1[i].reshape(1, max_len_text)))
  print('\n')
```

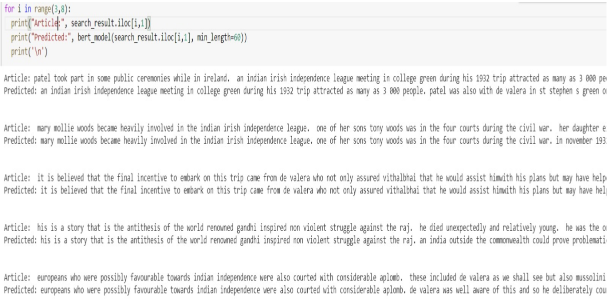
```
Article: politics learning imperial secession trade irish indian nationalism in throughout first half of
Keyword: irish
Predicted summary:  summarily wielders disbelieving insurance pjanic pjanic yanagi shinozaki fiendishly


Article: patel indian irish independence an july there appeared an article announcing formation group ca
Keyword: ireland
Predicted summary:  undiplomatically undiplomatically sunshine sunshine monstrously monstrously o undera
```

Figure 13. Text Summarization on Our Model

BERT Model Text Summarized Output : As shown below in the figure and as per BLEU and ROUGE measures BERT has outperformed our first model based on seq2seq model and as seen the summary it's almost closely relevant to the context and performed best as it bagged highest ROUGE value.

GPT-2 Model Text Summarized Output : The generated output is better than BERT but it took repetition of many words in the summarised input but still we can say it's much better than Seq2Seq model but could not score higher than BERT.

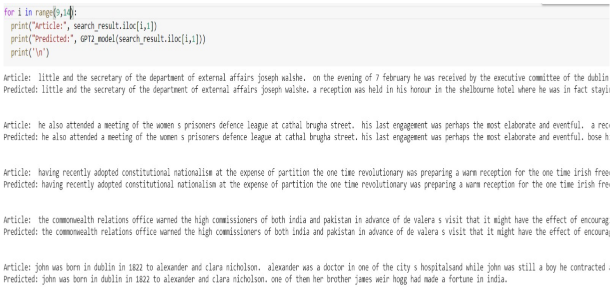
XLNet Model Text Summarized Output : Text summarization of the output looks much better grammatically correct and based on both ROUGE and BLEU we can say it has outperformed GPT-2 but not better than BERT.Hence considering the BERT pre-trained model is our go to preference.



```
for i in range(3,8):
  print("Article:", search_result.iloc[i,1])
  print("Predicted:", bert_model(search_result.iloc[i,1], min_length=60))
  print('\n')
```

```
Article: patel took part in some public ceremonies while in ireland.  an indian irish independence league meeting in college green during his 1932 trip attracted as many as 3 000 pe
Predicted: an indian irish independence league meeting in college green during his 1932 trip attracted as many as 3 000 people. patel was also with de valera in st stephen s green o


Article:  mary mollie woods became heavily involved in the indian irish independence league.  one of her sons tony woods was in the four courts during the civil war.  her daughter e
Predicted: mary mollie woods became heavily involved in the indian irish independence league. one of her sons tony woods was in the four courts during the civil war. in november 193


Article:  it is believed that the final incentive to embark on this trip came from de valera who not only assured vithalbhai that he would assist himwith his plans but may have help
Predicted: it is believed that the final incentive to embark on this trip came from de valera who not only assured vithalbhai that he would assist himwith his plans but may have hel


Article:  his is a story that is the antithesis of the world renowned gandhi inspired non violent struggle against the raj.  he died unexpectedly and relatively young.  he was the o
Predicted: his is a story that is the antithesis of the world renowned gandhi inspired non violent struggle against the raj. an india outside the commonwealth could prove problemati


Article:  europeans who were possibly favourable towards indian independence were also courted with considerable aplomb.  these included de valera as we shall see but also mussolini
Predicted: europeans who were possibly favourable towards indian independence were also courted with considerable aplomb. de valera was well aware of this and so he deliberately cou
```

Figure 14. Text Summarization on Our Model



```
for i in range(9,14):
  print("Article:", search_result.iloc[i,1])
  print("Predicted:", GPT2_model(search_result.iloc[i,1]))
  print('\n')
```

```
Article:  little and the secretary of the department of external affairs joseph walshe.  on the evening of 7 february he was received by the executive committee of the dublin
Predicted: little and the secretary of the department of external affairs joseph walshe. a reception was held in his honour in the shelbourne hotel where he was in fact stayi


Article:  he also attended a meeting of the women s prisoners defence league at cathal brugha street.  his last engagement was perhaps the most elaborate and eventful.  a rec
Predicted: he also attended a meeting of the women s prisoners defence league at cathal brugha street. his last engagement was perhaps the most elaborate and eventful. bose h


Article:  having recently adopted constitutional nationalism at the expense of partition the one time revolutionary was preparing a warm reception for the one time irish free
Predicted: having recently adopted constitutional nationalism at the expense of partition the one time revolutionary was preparing a warm reception for the one time irish fre


Article:  the commonwealth relations office warned the high commissioners of both india and pakistan in advance of de valera s visit that it might have the effect of encourag
Predicted: the commonwealth relations office warned the high commissioners of both india and pakistan in advance of de valera s visit that it might have the effect of encoura


Article: john was born in dublin in 1822 to alexander and clara nicholson.  alexander was a doctor in one of the city s hospitalsand while john was still a boy he contracted
Predicted: john was born in dublin in 1822 to alexander and clara nicholson. one of them her brother james weir hogg had made a fortune in india.
```

Figure 15. Text Summarization on Our Model

VI. Deployment

Post testing the model and evaluation of results over generated summarized texts next step is to deploy the code into Business environment as our client is "DMC Tours Ireland". Code shall be given to Client with real-time test data.

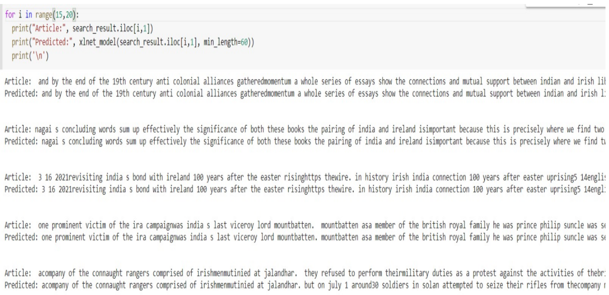
VII. CONCLUSION AND FUTURE WORK

We have tested our client dataset on both new model created by us using sequence to sequence model with attention mechanism and also on existed model like BERT,GPT2 and XLNet which has outperformed traditional models.Out of the four models implemented, BERT outperformed other models in terms of evaluation metrics with the overall score of all four evaluation techniques in the range of 45% to 51%.

For future work we can see scope of building an IDE or API based frame specially user friendly where our client can upload the available historical data source and update keywords based on tourists regional culture easily and able to run model together and get the generated abstract summarized output in less span of time also can see more scope on training model with more variety of relevant data-set and also we must develop more efficient technique provided or cross-lingual based structure and can try to generate better concise and few redundant summarized by amalgamation of new and available techniques.

References

[1] S. Adhikar et al. Nlp based machine learning approaches for text summarization. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 535–538. IEEE, 2020.



```
for i in range(15,20):
  print("Article:", search_result.iloc[i,1])
  print("Predicted:", xlnet_model(search_result.iloc[i,1], min_length=60))
  print('\n')
```

```
Article:  and by the end of the 19th century anti colonial alliances gatheredmomentum a whole series of essays show the connections and mutual support between indian and irish li
Predicted: and by the end of the 19th century anti colonial alliances gatheredmomentum a whole series of essays show the connections and mutual support between indian and irish l


Article: nagai s concluding words sum up effectively the significance of both these books the pairing of india and ireland isimportant because this is precisely where we find two
Predicted: nagai s concluding words sum up effectively the significance of both these books the pairing of india and ireland isimportant because this is precisely where we find t


Article:  3 16 2021revisiting india s bond with ireland 100 years after the easter risinghttps thewire. in history irish india connection 100 years after easter uprising5 14engli
Predicted: 3 16 2021revisiting india s bond with ireland 100 years after the easter risinghttps thewire. in history irish india connection 100 years after easter uprising5 14engl


Article:  one prominent victim of the ira campaignwas india s last viceroy lord mountbatten.  mountbatten asa member of the british royal family he was prince philip suncle was s
Predicted: one prominent victim of the ira campaignwas india s last viceroy lord mountbatten. mountbatten asa member of the british royal family he was prince philip suncle was s


Article:  acompany of the connaught rangers comprised of irishmenmutinied at jalandhar.  they refused to perform theirmilitary duties as a protest against the activities of thebr
Predicted: acompany of the connaught rangers comprised of irishmenmutinied at jalandhar. but on july 1 around30 soldiers in solan attempted to seize their rifles from thecompany
```

Figure 16. Text Summarization on Our Model

[2] A. F. Adoma, N.-M. Henry, and W. Chen. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121, 2020. doi: 10.1109/ICCWAMTIP51612.2020.9317379.

[3] A. Alshahrani, M. Ghaffari, K. Amirizirtol, and X. Liu. Identifying optimism and pessimism in twitter messages using xlnet and deep consensus. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[4] A. Bagalkotkar, A. Kandelwal, S. Pandey, and S. S. Kamath. A novel technique for efficient text document summarization as a service. In *2013 third international conference on advances in computing and communications*, pages 50–53. IEEE, 2013.

[5] P. Batra, S. Chaudhary, K. Bhatt, S. Varshney, and S. Verma. A review: Abstractive text summarization techniques using nlp. In *2020 International Conference on Advances in Computing, Communication & Materials (ICACCM)*, pages 23–28. IEEE, 2020.

[6] G. V. M. Chandu, A. Premkumar, N. Sampath, et al. Extractive approach for query based text summarization. In *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, volume 1, pages 1–5. IEEE, 2019.

[7] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, page 113679, 2020.

[8] W. Guo, B. Wu, B. Wang, and Y. Yang. Two-stage encoding extractive summarization. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*, pages 346–350. IEEE, 2020.

[9] M. M. Haider, M. A. Hossin, H. R. Mahi, and H. Arif. Automatic text summarization using gensim word2vec and k-means clustering algorithm. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 283–286. IEEE, 2020.

[10] D. Jain, M. D. Borah, and A. Biswas. Automatic summarization of legal bills: A comparative analysis of classical extractive approaches. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 394–400, 2021. doi: 10.1109/ICCCIS51004.2021.9397119.

[11] C. Khatri, G. Singh, and N. Parikh. Abstractive and extractive text summarization using document context vector and recurrent neural networks. *arXiv preprint arXiv:1807.08000*, 2018.

[12] V. Kieuvongngam, B. Tan, and Y. Niu. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *arXiv preprint arXiv:2006.01997*, 2020.

[13] S. Lamsiyah, A. El Mahdaouy, B. Espinasse, and S. E. A. Ouatik. An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems with Applications*, 167:114152, 2021.

[14] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu. A review of text corpus-based tourism big data mining. *Applied Sciences*, 9(16):3300, 2019.

[15] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández Orallo, M. Kull, N. Lachiche, M. J. Ramírez Quintana, and P. A. Flach. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[16] K. Merchant and Y. Pande. Nlp based latent semantic analysis for legal text summarization. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1803–1807. IEEE, 2018.

[17] D. Miller. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*, 2019.

[18] K. Muthiah. Automatic Coherent and Concise Text Summarization using Natural Language Processing. Master's thesis, Dublin, National College of Ireland, Jan. 2020. URL http://norma.ncirl.ie/4133/.

[19] C. Prasad, J. S. Kallimani, D. Harekal, and N. Sharma. Automatic text summarization model using seq2seq technique. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 599–604. IEEE, 2020.

[20] J. Shah, M. Sagathiya, K. Redij, and V. Hole. Natural language processing based abstractive text summarization of reviews. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 461–466. IEEE, 2020.

[21] D. Suleiman and A. Awajan. Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges. *Mathematical Problems in Engineering*, 2020, 2020.

[22] G. Szűcs and D. Huszti. Seq2seq deep learning method for summary generation by lstm with two-way encoder and beam search decoder. In *2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 221–226. IEEE, 2019.

[23] M. A. I. Talukder, S. Abujar, A. K. M. Masum, S. Akter,

and S. A. Hossain. Comparative study on abstractive text summarization. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–4. IEEE, 2020.

[24] J. Wang, H. Su, H. Zheng, B. Yan, S. Xu, and W. Tang. Research on abstractive automatic summarization technology based on deep learning. In *2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, pages 433–438. IEEE, 2019.