National College of Ireland

Abstractive-Based Multilingual Text Summarization and Sentimental Analysis using NLP Techniques

MSc Research Project
Data Analysis

Rutuja Anil Pande
Student ID: x21239444

School of Computing
National College of Ireland

Supervisor: Dr.Ahmed Makki

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rutuja Anil Pande
Student ID:	x21239444
Programme:	Data Analysis
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr.Ahmed Makki
Submission Due Date:	14/08/2023
Project Title:	Abstractive-Based Multilingual Text Summarization and Sentimental Analysis using NLP Techniques
Word Count:	7794
Page Count:	21

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:	17th September 2023

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):	

Abstractive-Based Multilingual Text Summarization and Sentimental Analysis using NLP Techniques

Rutuja Anil Pande
x21239444

Abstract

The study proposes generating text summarization from the news articles and performing sentimental analysis on those generated summaries to identify the semantics behind the sentences. These summaries are then translated to Hindi language to attract the non-english speakers from India.To implement the objectives proper research methodology is followed. Dataset used in this study are Indian News summary and BBC News Summary. Both the datasets are pre-processed using function to remove stopwords, punctuation, empty spaces from the text for better analysis. Basic Exploratory Data Analysis (EDA) is performed to understand the structure of data in detail. Models like Bidirectional and Auto-Regressive Transformers (BART), Bidirectional Encoder Representations from Transformers (BERT) and Google Translator are used for the objectives of this research. The summaries generated are of short lengths between 50 to 150 range. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score for both the datasets. The ROUGE-L F1 score for first dataset is approximately 10%. ROUGE-1, ROUGE-2, ROUGE-L score are 40%, 20%, 40% respectively for second dataset. The scores are comparatively low due to the eliminations of outliers and filtered data. The translation is measured using Bilingual Evaluation Understudy (BLEU) score which falls close to 0. which indicates that the translation are not very good in quality. This can be due to the translation performed on the summaries generated from filtered data. The sentimental analysis performed on summaries produced output close to the sentiments of the text.

1 Introduction

1.1 Background and Motivation

Text summarization has been in genuine help for long articles, movie summaries, video , and audio summaries. Nowadays, these mediums are given their expected importance but news articles are not recognized that much. Basically, text summarization is a short story of the long documents. People are less interested nowadays for reading the whole article as it may sometimes seem boring or may be not understandable. In this case, a short summary for the long document can be efficient for reading. This study targets the purpose of reducing the time for reading the long articles by short articles. It can save people's time and they even can get an idea of the topic that the article is about. There are various of blogs, websites, news stories on the internet which contains unstructured data. So, these data can be used for gaining in proper format which is readable by humans

and they can explore the information more easily in the manner of summaries. Generating the text summaries are of two types extractive and abstractive text summaries. In this study, abstractive text summarization is employed using pre-trained model BART. It is an sequence-to-sequence transformer model trained on large documents which are corrupted with denoising for training purposes. It has an architecture of encoder-decoder where encoder takes the input data and decoder produces the output summaries. BART is selected for generating text summaries as it has 10% more features than BERT model. Therefore, "facebook/bart-large-cnn" model is used here and after this BART tokenizer helps in tokenizing or encoding words differently to form an abstractive output summary.Gupta et al. (2022) This pre-training model called as language model are said to very effective in terms of improving the natural language processing tasks. Devlin et al. (2018) BERT is used for performing sentimental analysis on generated tasks as it can extract more features and semantics from the sentences and present it in a contextual format with its relationships.Hoang et al. (2019) A package googletrans==4.0.0-rc1 is installed to access the translation services of Google API's for English to Hindi translation which can help attracting many audience in India for reading news articles. Fig 1. shows the basic flow of generated abstractive text summaries.

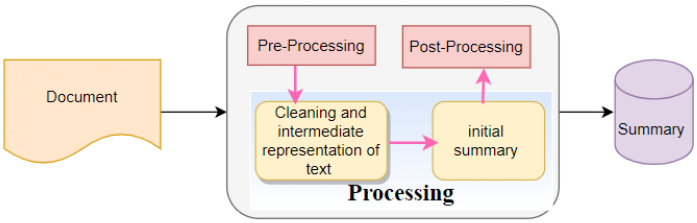

Figure 1: Process of Abstractive Text Summarization ResearchGate (n.d.)

The modelling pipeline to combine 3 different NLP tasks in this study is based on Hybrid Pipeline Modelling where the output of task 1 is used as an input for task 2 and 3. Further, the tasks 2 and 3 generates separate outputs. The study is divided in three parts:

- Generating text summaries from the news articles using BART model
- Performing sentimental analysis on the generated text summaries using DistilBERT model
- Translation of English text summaries to Hindi text summaries using Google Translator

In Fig 1. below The diagram of hybrid modelling pipeline is depicted. The news articles which is the source data are passed to BART model as task 1 and output of text summaries are generated. These generated text summaries are then passed on as an input to the next step where two tasks are preformed. Sentimental Analysis is performed on the generated summaries using DistilBERT as task 2 and the translation of English text summaries into Hindi text summaries is achieved in the 3rd task using Google Translator. Both tasks outputs are generated separately. The output for task 2 is produces as labels associated with each summaries indicating the sentiments as positive or negative. The output of task 3 is the new generated summaries but in Hindi language.

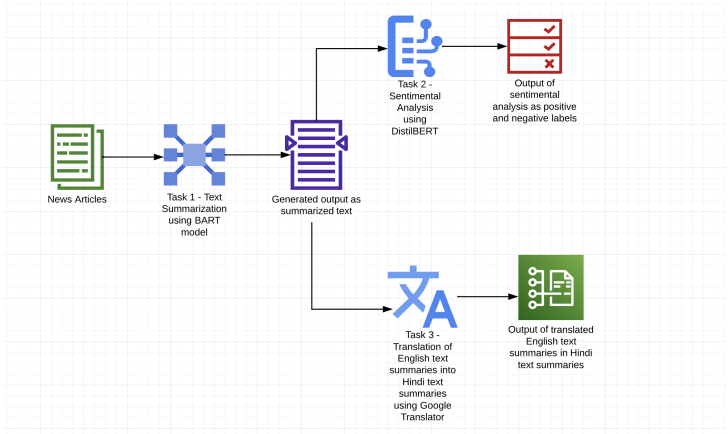

Figure 2: Hybrid Modelling Process

1.2 Research Question and the Objectives

How to generate High quality Abstractive Text Summaries using pre-trained models and identify semantics behind it? What strategies to apply for attracting Non-English speaking audience in India?

The objective of this research is to improve the quality of text summarization of long news articles. For this, best pre-trained models are selected to achieve the objective. Then performing sentimental analysis on the text summaries to know whether the summary is based on positive or negative events which can help the audience to decide if they want to continue reading the whole article or not in real time. This study also targets the audience in India who does not read or speak in English. Therefore, by converting the summaries in Hindi Language. For this, a simple package called googletrans==4.0.0-rc1 is used which has access to google translator services and API's.

2 Related Work

In this section, the previous work done on the similar topic is being discussed. The machine learning or deep learning techniques used and the quality results achieved for text summarization in the previous works are going to be discussed in detail while analysing it critically.

2.1 Text Summarization using DL, ML and NLP Techniques

The studyKahla et al. (2021) discusses about some challenges faced by the abstractive text summarization for Arabic language. Arabic being a complex language it is hard to generate summaries from it. The study highlighted some of the challenges like absence of short vowels and presence of multiple form of words. The quality of results are based on various evaluation metrics the scores are very similar and the performance and the outcome generated is very promising in both automatic and manual evaluations. The results of these evaluation methods are not compared in details and only the overview is given in the paper. In the current study, BART model is being used for text summarization in English language and the results quality is measured using ROUGE score which given

an idea about the accuracy of the output.

To achieve an automatic text summarization from BBC news articles. The author Haider et al. (2020) has used k-means clustering algorithm, Word2Vec and Gensim library. This study aimed to generate summaries from only a single document by extracting features andb by k-means clusters. The feature extraction is done using Gensim and Word2Vec libraries. The business category as this category had more numeric values and the library used here focuses mainly on numeric values. As this study was based on unsupervised learning, the authors decided to use k-means clustering which resulted best for their study. The evaluation of the text summaries are done using BLEU score and highest BLEU score was 0.894 for business category. In the current study, same dataset is used with only three categories as entertainment, tech, and politics and is using supervised learning which can help in developing more quality text summaries.

The paper La Quatra and Cagliero (2022) discusses about developing a tool that generates abstractive text summarization in Italian language using a transformers models. These models generates typically very fluent, concise and coherent summaries. Many models are applied and the importance of each models are highlighted like BART, mBART, mT5 transformers. The experiments results are evaluated on the basis of performance and quality of summaries. The paper has evaluated the model on two model ROUGE and BERT score for both abstractive ans well as extractive summaries on three different datasets Fanpage, WITS, and IlPost out of which the BART-IT model performed best on WITS dataset achieving the highest scores for all the evaluation metrics being used. In the current study, ROUGE score and BLEU scores are used to evaluate the models performance and BART model is used for English text later this text is converted to Hindi Language.

The author Krishnan et al. (2019) has an objective of generating extractive but supervised learning based text summaries using BBC News Dataset. This started by robust feature extraction from the news articles after the pre-processing of the data. The model chosen for this purpose are the Machine Learning algorithms like Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF). The quality of generated text summaries are evaluated based on ROUGE score in this study and the score for ROUGE-1 is of 0.51 approximately for all the categories. Firstly, experimenting with ML for text summarization can be very helpful in understanding the concept and to have a firm hand on this field. Using ML approach turned to be good for this research. But, generating extractive summaries and that for long articles requires a large number of resources. In the current study, abstractive text summarization is produced by using pre-trained natural language processing (nlp) models to handle the complexity of the dataset and resource allocations as they are said to be best performing for textual data.

In this studyAlshibly et al. (2023) author used Named Entity Recognition (NER) model for generating text summaries. The python library called SpaCy is imported initially to use NER model. NER identifies the important entities from the text and return it. Using NER the results are more accurate in terms of F-scores as compared to word frequency method. After applying both models to datasets the outputs are evaluated based on ROUGE score and NER model performed better according to the score. Two data-

sets are used here and the dataset with shorter text performed well with NER model in terms of F-score, ROUGE-1, ROUGE-2 and ROUGE-3 as compared to word frequency method. AS the current research is using pre-trained model the results of NER model did not affect much to the text summaries. Later on NER model was removed from the current study as the dataset in this study is quite large and NER did not performed well. The text summarization is measured using ROUGE scores by defining the quality of the summaries generated using BART model.

In this paper Gupta et al. (2022) author discusses about the comparison between different training objectives on language models and evaluation of performance is done on basis of different tasks like Xsum, ConvAI2, SQuad, MNLI, ELI5, and CNN/DM. Various models are explored in this research mainly based on BART , BERT that is seq2seq models. F1-score, accuracy, precision and perplexity are used as evaluation metrics for this study. The performance of these pre-trained models varied significantly among all the tasks. Left-to-Right pre-training models are decided to be very important for tasks generation and token masking is crucial in terms of performance. BART model performed well for the tasks with text infilling but ELI5 dataset worked good with pure language models. The models used in this research are pre-trained model and gave tremendous amount of knowledge for all the language models from which BART was suitable for the current study and thus was selected. BERT is used in the current study for sentimental analysis as it identifies semantics from the texts.

The paperManakul and Gales (2020) is based on the discussion of Podcast Summarization challenge selected by the authors. The challenge was to generate the best summaries with the most important information from the podcast episodes and their transcripts. The proposed system is in the two steps where the author first uses a step to reduce or remove the sentences which are duplicate or means the same from the transcription. This improves the length of the text and then, BERT is applied for generating the text. this study also includes ensemble approach with nine models. The evaluation is done by both manual and automatic evaluation methods such as ROUGE-L and human evaluation by NIST. Sometimes, including so many models for the implementation can result in not identifying the best model for the tasks and the resources and computation power required for such approaches is very expensive.

2.2 Sentimental Analysis using DL and pre-trained Models

In this study the author Devlin et al. (2018) discusses BERT implementation and its performance on various nlp tasks. BERT is a pre-trained model that is bidirectional which means it remember previous and future memories therefore, it is very good at extracting the semantics. It can always be fine-tuned with just an extra output layer for specific tasks. There are some tasks performed and BERT performed as state-of-the-art results on eleven tasks like question-answering, sentence-level tasks, language inference. The Evaluation is done based on GLUE benchmark including BERTBASE and BERTLARGE that outperformed the other systems. BERTLARGE performed best on dataset with limited training data. On SQuAD BERT performed better by achieving top leaderboard scores. The BERT performed best in this study, thus including this model in the current study for quality results and better implementation.

The authors ResearchGate (n.d.) in this paper presented the concept of sentimental analysis to understand the semantics behind the text to be judged by humans. Lexicon-based approach is proposed for sentimental analysis on BBC news articles. The business and sports categories got the highest positive reviews then the other categories. The study is performed on unsupervised learning data and it mostly focused on two things sentence polarity and word polarity. The sentimental analysis performed in this work by lexicon-based approach which a machine learning method gave very good results in identifying the sentiments. In the current study, BERT is used for this purpose to basically implement the model easily and get the results as positive or negative. A variant of bert known as Distilbert is used which is popular to identify the semantics and give contextual representation alongside labels.

This research discusses the challenges faced during sentimental analysis for Urdu Language. The scarcity of the linguistic resources are highlights for Urdu Sentimental Analysis. Various ML and DL models are included in this study for Urdu text classification like GRU, LSTM, mBERT, CNN-1D. The author Khan et al. (2022) a manual dataset in Urdu which is also manually annoted. UCSA-21 dataset is used for sentimental analysis which has comments of users in multiple genre. The study found that, mBERT performed best compared to other models. mBERT achieved an accuracy of 82.5%. Similarly, precision, recall, and F1-score are 81.35%, 81.65%, and 81.4% respectively. The comparison in this study between deep learning models and mBERT proves that the pre-trained model produces a quality of performance and gives accurate semantics from the texts. In the current research sentimental analysis is measured using BLEU score. The BLEU evalution metric will given an perspective of the quality of the sentiments extracted and labeled out of the summaries.

The paperNaik et al. (n.d.) works on Amazon URLs for improving the efficiency. The review which is given under the amazon product and the review which are given for customer services are together. To distinguish between this a ranking method is used by which it makes easier for the customer who are surfing on amazon website to the know the review is for what thing. The analysis is performed to employ if the reviews are positive, negative or neutral using VADER library. In the current research study, for sentimental analysis distilbert is used for extracting the semantics and relations of the sentences associated with the labels which is easy implementing and provide results with positive and negative labels.

The paperHoang et al. (2019) is based on the discussion on aspect-Based sentimental Analysis (ABSA) which uses BERT model. between ABSA and traditional tasks ABSA is more complex because it requires to identify the attributes or aspects present in the text. The paper describes the ABSA tasks like aspect classification, sentiment polarity classification, and target expression. ABSA models are compared with and without BERT and then the advantages are marked down for ABSA using BERT. BERT uses semantic similarities in word embeddings for finding sentiments for an aspect in the text. The Evaluation of result is done by SemEval-2016 Task 5. Aspect-based sentimental analysis outperforms the results of previous state-of-the-art. The uniqueness in this paper makes the research interesting and the results shown in the research is well performed by BERT.

This paperLibrary (n.d.) has used unsupervised approach on dataset if clinical discharge summaries where two libraries were introduced Word2Vec and Doc2Vec. the aim of the research is to detect the underlying bias for a particular diseases. SentiWordNet is used for sentimental analysis.

In this studySanh et al. (2019) the proposed model are KNN, naive bayes, SVM, Logistic Regression, CNN and BERT for sentimental analysis. Naive Bayes outperformed in finding the emotion in the tweets than KNN. Logistic regression got the highest accuracy count in analyzing product reviews tweets. The research uses a hybrid approaches of these models. The classification of sentiments is done in three categories as neutral, negative and positive. For Bert and DL models Bert achieved an 92% of accuracy for twitter sentiments recognition. CNN and BERT in combination on the other hand scored 92.6% and 91.89% on different datasets. The best performance of this models RNN, LSTM, and CNN with BERT was achieved. In the current study, the attempt to combine CNN with NER and CNN with BERT was done but due to longer computation only BERT was used.

2.3 Translation of English Summaries to Hindi Summaries

The authors Gupta et al. (2023) in this research ATS system known as Auto text summarization is used for Indian languages. The authors have discussed about the less attention given to Hindi language. In this study the datasets BBC News, CNN daily news are translated using machine translation engine. Microsoft bing , Google translator are used here. The ROUGE-3 score for this model is around 56%, 46%, and 44% for BBC news, DUC 2004, and CNN daily news respectively. The Google Translator is used in the current study. The implementation is not very hard. The translation also depends upon the data in the dataset. ROUGE scores for the previous work are average and therefore, can be considered as average accuracy.

In this paper the authorPrates et al. (2020) addresses the issue regarding the bias AI where, trained model societal biases are reflected like racial or gender bias. The proposed system highlights the harm caused by the biased AI tools which classifies on race and gender. This problem is addressed using automatic translation tool through gender - neural languages. They used and trained some of the sentences like HE/SHE and translated the sentences into English language using Google translator. After the translation the frequency of male and female in the sentences is analyzed. After the analysis the results showed that the Google translator mostly implies on male pronounce than female. Finally, the study suggests that the results are not up to the mark and it does not address the machine biases using machine translation technique. As the current study does not involve any detection of the sarcasm and biasness in the text. Therefore, Google Translator is used here for translation of English to Hindi language.

The studyLarassati et al. (2019) focuses on the translation errors during the process of translation of Instagram post in Indonesian language using Google Translator API. The errors are addressed using an ATA system which is American Translator Association as the analytical framework. After the translation it is observed that the original post carries more errors categorized as Literalness (L), Terminology (T), and Syntax (SYN) which are interconnected to each other. The research finally suggested that the Google

Translation has an potential in future to be the better and reliable machine translation tool. In the current research, Google translator gives BLEU score as low as sometimes it does not define the quality appropriately but eventhough, the results or the translation generated are not very bad.

The studyMoslem et al. (2023) focuses on achieving the real time consistency and domain adaption using machine translation (MT). The Google translator is explored on the basis of large-scale where MT learns input-output text, or any pattern generation without any fine-tuning. The experiments showed the the MT is reliable to enhance in aspect of in-domain pairs, terminology when translation the news sentences, and only with short context learning it has high chance of surpassing the encoder decoder system. The experiments was performed on various languages pairs like. From English to Arabic, Kinyarwanda, and Spanish. The Google Translation is properly used and tested in this study which makes the translation reliable for translation purpose. In the current study, English text in translated to Hindi which also performed very well. The implementation and processing was very smooth using Google Translator.

3 Methodology

This research has followed the CRISP-DM methodology throughout the study. The Fig 2. depicts the flow of CRISP-DM.

Figure 3: Flow of Research Methodology ItsaLocke (n.d.)

3.1 Data Acquisition

.

In this research 2 datasets are used based on news articles and the summaries. Theses datasets are obtianed from an open- source platform known as Kaggle that has a tremendous collection of datasets for analysis and data processing purpose from the data science community. The datasets have around 4000 plus records which contains the long articles and respective summaries as well. One dataset is CSV file of Indian News Summary and another one is read only text files of BBC News Summary. Datasets available on Kaggle are undergone some data cleaning processes. The datasets are also available publicly and can be downloaded on big scale.

3.2 Data Pre-processing

The datasets are gone under various text pre-processing techniques such as removing special characters, removing punctuation, stopwords, converting to lower case and tokenization. The CSV file has the input data and target data as news articles and headlines for the analysis purpose. The second dataset has two directories as News Articles and Sumamries. These directories are further classified into 5 categories as business, sports, tech, politics, and entertainment. The CSV file has 6 columns but only two were selected for the study as an input data and target data. the text file has only text and summaries into the folder. As the texts in these files are very long and are of unequal length which takes a lot of time for executing any operations, the text with same length were extracted and the further operation are performed to only those sample records.

3.3 Data Transformation

The data transformation steps taken for this research involves some of the steps in general which converts the raw data into a data which can be useful for operations. Such as data augmentation which paraphrased the text using extra words. Outliers are removed from the dataset, Data filtration using sampling techniques is done for improving the computation.

3.4 Model building and evaluation

Model building and Evaluation stage discusses about the models selected for the study and the evaluation model used for testing the generated results. There are three parts in this research text summarization, sentimental analysis, and language translation. Therefore, three models are selected for this.

- BART - Generating text summaries is done by applying pre-trained models on the dataset. BART is a pre-trained model which is specially designed for various natural language processing task. It is sequence-to-sequence model, trained on multiple long corpus of various different languages. Hence, BART is a suitable model for the purpose of generating text summaries. BART is better than BERT in terms performance and computational power Lewis et al. (2019).

- BERT - Sentimental Analysis performed on generated text summaries was employed by DistilBERT. It is a variant of BERT model. It is a transformer-based model with the capability of identifying sentiments for the text. The results are produced in two categories as LABEL-1 and LABEL-0 which positive and negative. DistilBERT is

designed for such tasks which inherits some of the performances from BERT model. Sanh et al. (2019)

- Google Translator - It is an online language translation service which Google developed for translation purpose for different languages. The users are able to translate any text document or any other content from language to language. Installing googletrans==4.0.0-rc1 package gives access to the google services and can easily translate the text in different language by calling translator() function. Moslem et al. (2023)

The quality of text summaries generated, the analysis of sentiments extracted from the summaries and the translation of the English to Hindi is calculated or evaluated based on some scores or model. Hence, ROUGE-1, ROUGE-2, ROUGE-L scores and BLEU scores.

3.4.1 ROUGE Score

ROUGE short form of (Recall-Oriented Understudy for Gisting Evaluation) is an evaluation metrics used to evaluate the quality of text summaries which are generated by machines. ROUGE score gives a score by comparing the human generated summaries to machine generated summaries. It is known for measuring the overlap between sequence of words which n-grams between the reference and machine generated summaries. Kahla et al. (2021)

In this study, ROUGE-1 , ROUGE-2, and ROUGE-L is used for measuring the unigrams, bigrams, and trigrams between the reference and generated summaries.

Sometimes, the accuracy or score given by these evaluation metrics does not defines the exact quality. Depending upon some factors, it can give very low score eventhough, the summaries generated are not that bad

3.4.2 BLEU Score

The third part of this research is translation of English to Hindi language. BLEU score is the best evaluation metric to measure the quality of translation. (Bilingual Evaluation Understudy). It compares more than one reference human generated translation and gives a score that defines the quality of the translation. The BLEU score also gives less score depending upon some factors. Therefore, this score gives an idea of the quality of translations comparing the human reference and machine translations. Haider et al. (2020)

4 Design Specification

This section discusses about the models architecture or detailed description used for the implementation of the research.

4.1 BART - Bidirectional and Auto-Regressive Transformers for Text Summarization

BART is a transformer-based model which is very powerful model specially designed by training long documents on the model. Its application is for various natural languages

processing tasks. One of the task is text summarization. The architecture of BART is a encoder-decoder architecture and has some unique features of denoising training and auto-regressive objectives. The explanation of the architecture in details is discussed below:-

- Encoder-Decoder architecture: As BART is sequence-to-sequence model, it has encoder-decoder architecture similar to other models. The steps are that the encoder processes the input data from the dataset and the decoder generates the results which is the output data such as summaries. Therefore, the structure of this model is build for handling the tasks like text summarization.
- Denoising Autoencoder (DAE) Architecture: DAE is also a part which is employed during the pre-training process. The flow is to corrupt the input data or text file by randomly selecting the parts or sentences from the paragraphs and then training the model by reconstructing the original sequence. Lewis et al. (2019)
- Genenration and Summarization of Sentence: As the BART decoder is responsible for producing generated output summaries. During fine tuning this decoder generates a concise summary. Its training is based on predicting the next token based on the previous one and this step helps the model to generate the text as informative and coherent.

Fig 3. shows the architecture of BART model

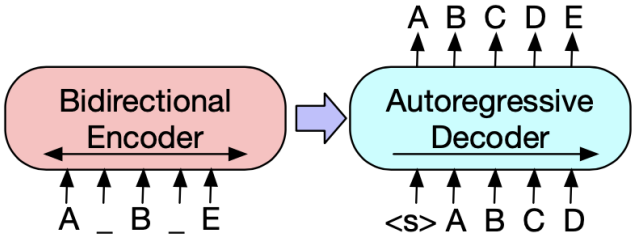

Figure 4: BART Architecture Lewis et al. (2019)

4.2 BERT - Bidirectional Encoder Representations from Transformers

BERT - known as a transformer-based model used for sentimental analysis in this study. The variant of BERT called as distilbert is used for the purpose of identifying the sentiment behind the sentences. The concept behind the BERT model is that is is pre-trained to produce the contextual sentences along with its relationship. It also associate the labels predicting the sentiments of the sentences as label-0 and label-1.

- Pre-trained BERT Model: The model or the variants of model are already trained on a big amount of data in textual formats which helps in producing the contextual words, meaning and relationships.
- Fine-Tuning for Sentiment Analysis: fine tuning of model is specifically done for extracting the sentiments from the text.

- CLS Token for Classification: This CLS tokens are used in this study for representing the sentiment classification.
- Binary Classification Task: For predicting whether the sentence is positive or negative, binary classification is used. A classification layer is added on top of the CLS tokens. Bello et al. (2023)

Fig 4. shows the design of distilbert model.

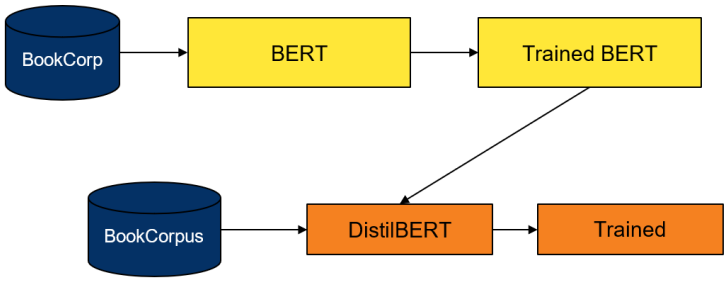

Figure 5: Distilbert Model Sanh et al. (2019)

4.3 Google Translator

The translation part performed on generated text summaries is achieved through installing googletrans==4.0.0-rc1 package which can get access to Google services and used as a python package. This package communicates with various services of Google. The library is very lightweight and does not have very complex architecture. The translator() function effortlessly translates the text. The translate engine used is neural machine translation model (NMT). These translation engine is trained on big amount on data based on different languages parallelly. Larassati et al. (2019)

5 Implementation

The implementation section discusses the description of the output produces in this research with the software, hardware, tools, and languages used. The transformation of data and nlp techniques selected for the study is also explained.

5.1 Configuration of Software and Hardware Components

As this research is based on abstractive Text Summarization using natural language processing techniques. Therefore, the models are selected for generating text summaries are suitable for this task. BART, BERT are pre-trained models for generating text summaries and performing sentimental analysis, they have a very large number of parameters which makes the model heavy for computing accessing a lot of time and memory resources. Hence, the project was developed on Mac Book Air with M2 processing chip with 8GB RAM and 512GB storage. Coding is done on Google Colab pro platform throughout the development using Python Language.

5.2 Data Transformation and Pre-processing Step

The datasets used in this study are in CSV and text read only format. The csv file already has news articles as text and its generated summaries. Therefore, there is not much cleaning of data involved like converting numeric to text. There is a lot of inequality and imbalance with the lengths of the text. For Indian news dataset the lengths varies 3 to 76045. But only few articles and are that long in length. The other dataset has different categories out of which only three categories are selected for the study. The average range of lengths in these categories falls between 1000 to 2000. The categories politics, entertainment, and tech are selected as they have approx similar number of articles around 300 to 400. To avoid this imbalance of data, outliers were removed by calculating the z-scores of the data points. Also after this samples of the filtered data was taken for easy computation of the model. removal of some special words, stop words, punctuation were done for cleaning the data. As some of the length were very less in the samples, data augmentation was done to increase the size of generated summaries. Tokenization of sentences was done before applying the input data to the model. Same steps were performed for the other dataset.

5.3 Model Implementation

BART model used for generating text summaries which a pre-trained model generated the summaries very easily with its capability of producing high quality summaries. Initially, the BART pipeline was initialized from the Hugging face transformers library. This pipeline gives access to use the BART model without worrying about the complex architecture of the model. In this study, pre-trained variant of BART is used which is "facebook/bart-large-cnn" is specially designed for text summaries tasks.

The min and max length is defined in the code between the range of 100 to 500 for both the datasets. This lengths helps in generating concise summaries and within this range only. Initially, with this lengths the model didn't performed well with the imbalance data. But after removing the outliers this model worked perfectly with the filtered data. There was a consequence as well with this filtration. As the ouliers were extreme records with min lengths as 3 and max lengths as 70000, by removing this records the accuracy was affected a bit. The BART model generates the abstractive summaries out of input data with batch size of 8.

BERT is used here for sentimental analysis of summaries generated which is also from Hugging Face Transformers library. Distilbert a variant of BERT called as sentiment analyzer pipeline is a pre-trained model uses a function that iterates through all the generated summaries and uses this sentiment analyzer model for getting the sentiments out of it and than labeling it as positive or negative for each summary.

The translation of English text to Hindi language is achieved using Google Translator and python library called translate. Both the function gave similar output and implementation is also the same. Here, a function is defined which takes the parameter as source and target language as English and Hindi and the function is applied on the list of English text summaries. It translate each summaries in the loop and stores the translation inside new list. As some of the component in the list were of type 'None' thus, it was giving an error. To handle this, a check for 'None' is also included in the function to generate the

output appropriately. This translations are done only for short or smaller tasks and gave robust solutions but not suitable for longer tasks.

6 Evaluation

The experiments performed in this research is discussed here in detail. The results and their evaluation will be also discussed in this section. There are different experiments conducted for this study with different size of samples and lengths defined
There are two datasets selected for this study on which models are applied. It will be observed here on which experiment the accuracy or the scores are better.

6.1 Indian News Summary Dataset

Experiment 1: The Experiment is performed on the data filtered with the articles having length equal to 2000, tolerance = 100.
The parameters passed to the BART model are min = 100, max = 500. With this definition, the time taken by the model to run was very consuming and involving a lot of resources. Fig 5. shows the outliers before and after.

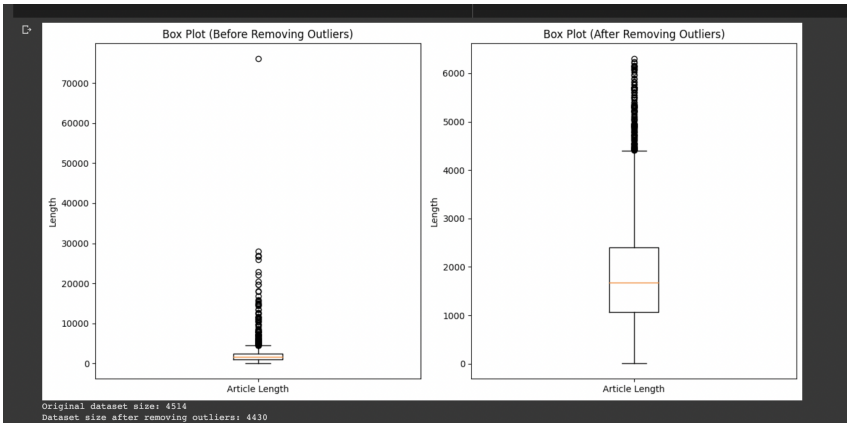

Figure 6: Before and After Outliers

The original dataset of Indian News Summary is of 4514 rows and 6 columns. From this, outliers are identified and removed and the models are applied on final filtered data.

There are some imbalance seen in the dataset. The below Fig 6. depicts the unequal lengths of the dataset.

6.1.1 Results and Evaluation

The Bart model is applied on the sentences which are pre-processed and are filtered out with the outliers. The Fig 7. shows the output of generated summaries.

The quality of the generated summaries are measured using ROUGE score. The average ROUGE score of the generated summaries is around 8%. The lower score achieved can be because of the sampling techniques used for improving the performance of the model and for efficient computation. Keeping this in mind the accuracy on original data

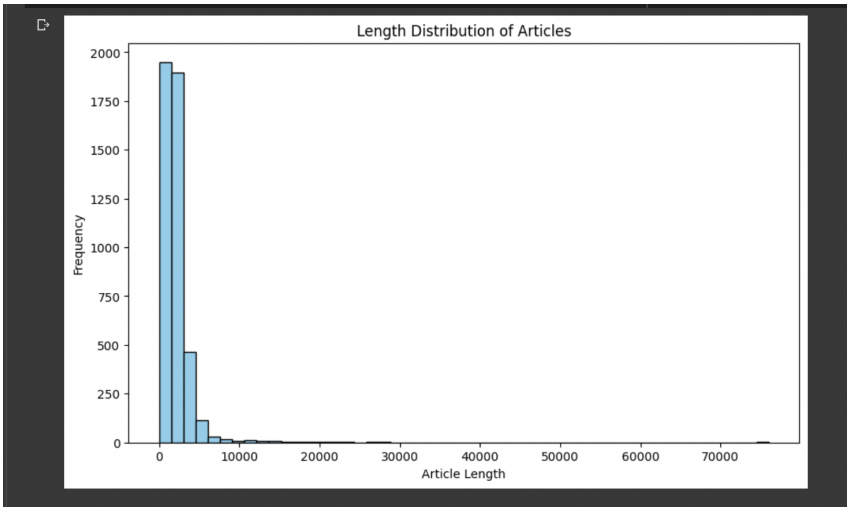

Figure 7: Distribution of Article Lengths

```
indian_news_sampled_data['summary']

22      Farmers including women took to the streets in...
25      The saudiled coalition fighting in yemen is ob...
47       naseeruddin shah has watched the film lipstic...
51      A teacher of a private school in haridwar dist...
60      Comedian sunil grover has stayed away from the...
                              ...
4425    Scientists from hokkaido university in japan h...
4446    Super cars were seized from near akkarai on ea...
4460    Gurmehar kaur is the daughter of a kargil mart...
4471    India is the worlds second largest mobile serv...
4499     demonetisation led to huge cash shortages tha...
Name: summary, Length: 307, dtype: object
```

Figure 8: Generated Summaries

can be achieved as a decent score. Due to less resources, running model on whole dataset was very time consuming and allocated a huge amount of resources.

The BERT model is used for sentimental analysis and the below Fig 8. depicts the output of the model.

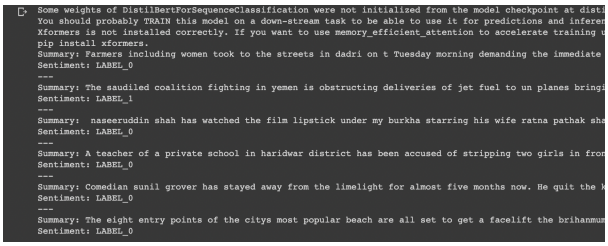

Figure 9: Result of Sentimental Analysis

The generated summaries are then translated using Hindi language using Google Translator. The score for the translation is again not up to the mark. As the summaries in English are measured as low accuracy. Below Fig 9. shows the translated summaries.

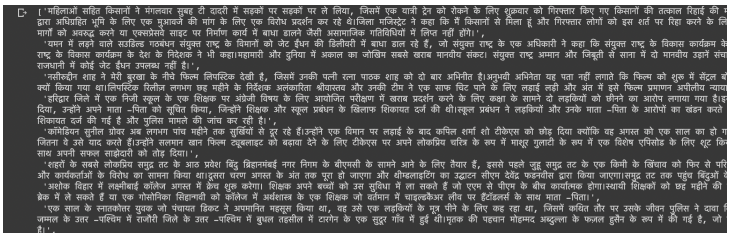

Figure 10: The Hindi Translation of English Summaries

Experiment 2: The next experiment, The parameters passed to the BART model are min = 50, max = 100. The time consumed with this parameters was very less as compared to 1st experiment.But the generated summaries are of the same lengths. Different sample are taken to check the execution time. In this experiment the sample size and the size of lengths passed are compatible with BART model thus, generate summaries in less time.

6.2 BBC News Summary Dataset

Experiment 3: The Experiment performed on BBC News Dataset is similar as with the first dataset. Same models and techniques are used here. Some changes are done as per requirements. Data augmentation is applied here on the pre-processed data to increase the accuracy. The outliers are removed and the operations are performed on final dataset. The parameters of BART model for text summarization are min = 50 , max = 150, batch size = 8. The time to execute the model was not very engaging as the data in this dataset is quite similar in lengths.

Below Fig 10. shows the dataframe created by extracting the data from the categories.

There are three categories and each categories have some number of news articles. The count of news articles is shown in the below Fig 11. for all three categories.

bbc_news_dataset

	news articles	summaries	categories
0	UK's 'useless' quangos under fire\n\nThe UK ha...	Mr Johnson told Age Concern's Age Agenda in Lo...	politics
1	Royal couple watch nation's mood\n\nPrince Cha...	But Mr Leigh said the department was "dragging...	politics
2	Straw to attend Auschwitz service\n\nForeign S...	Speaking outside the Olympic Primary School, M...	politics
3	Brown ally rejects Budget spree\n\nChancellor ...	Among the over-65s, 70% said they would be cer...	politics
4	McConnell details Scots wave toll\n\nAt least ...	Tony Blair has said he does not want higher ta...	politics
...	...	...	...
1199	Warnings on woeful wi-fi security\n\nCompanies...	But she did not, because having fast, always o...	tech
1200	Mobile games come of age\n\nThe BBC News websi...	An attachment in the e-mail contains the virus...	tech
1201	Humanoid robot learns how to run\n\nCar-maker ...	"Consumer electronics companies want UWB to re...	tech
1202	Open source leaders slam patents\n\nThe war of...	Security firm iDefence, which notified users o...	tech
1203	Slim PlayStation triples sales\n\nSony PlaySta...	But they have slowly realised that P2P is a go...	tech

1204 rows × 3 columns

Figure 11: Dataframe of BBC News Summary

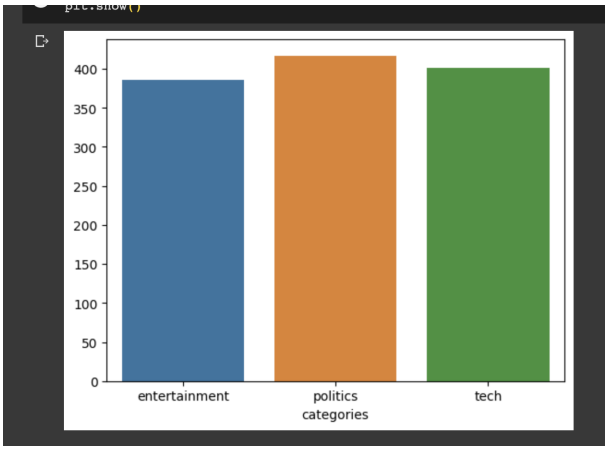

Figure 12: Count of News Articles for Categories

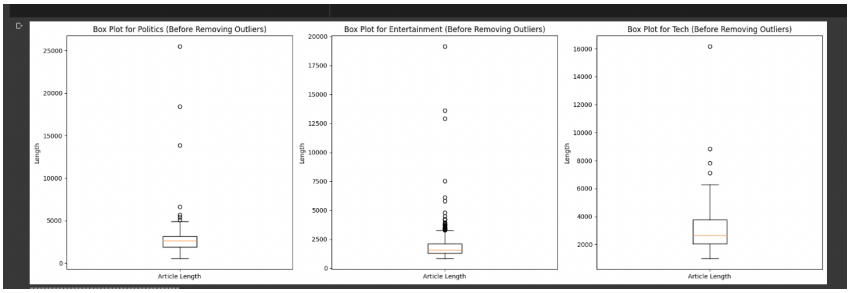

Figure 13: Outliers in each categories

The outliers observed in the dataset is fixed by removing the outliers. Fig 12. shows the outliers seen in the dataset.

The data augmentation is applied on the sentences and the result of augmentation and summaries generated by BART model are compared in the below Figures 13, 14.

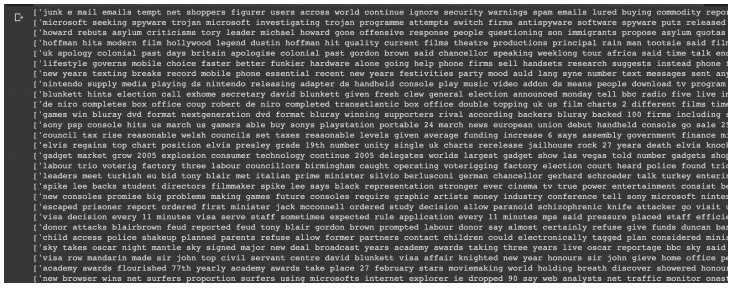

Figure 14: Data Augmentation

Figure 15: Generated Summaries on BBC News Summary

The sentimental analysis performed on the dataset produced output with labels indicating positive and negative text.

Translation of English summaries to Hindi summaries show in Fig 15.

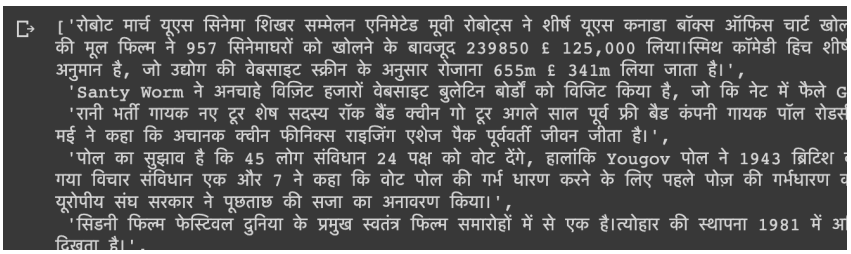

Figure 16: Translation of the English Summaries to Hindi

ROUGE score for the generated summaries are 40%, 20%, and 40% respectively for ROUGE-1, ROUGE-2, ROUGE-L score.

6.3 Discussion

The Experiments performed on this study is based on two datasets Indian News Summary and BBC News Summary. There are total three experiments altogether combining the

datasets. The first two experiments shows different min, max values with filtered data. The experiment justifies that even when different sizes of lengths are passed in parameters it generates the same lengths of summaries. BART model was not able to generate very long summaries. The summaries are generated in a decent length. The accuracy of the generated summaries is a bit lower around 8.2% for first experiment and 9.71% for other experiment is due to the sampling techniques used for filtering the data for better performance of the model. After eliminating the outliers the size of dataset was not sufficient for scoring better accuracies. The original data passed to the model might generate a good score for the summaries. The BERT model performed sentimental analysis which gave an average level of results. The result is labeled as 0,1. The sentences are labeled approximately close to the semantics in the sentences. The translation of the English summaries to Hindi summaries are pretty good. The summaries translation is measured using BLEU score which is very less eventhough, the translated texts are readable and has formed proper meaning and sentences. The less score might be because of the original summaries are generated on filtered data. The improvement for the accuracy can be done by improving the quality of the summaries and translation by running the model on whole set of data and by using different techniques such as BART plus CNN for text summaries and sentimental analysis. Another reason can be because Google translator was not able to manage the numbers, signs, and few words are still in English. Google Translator is a popular model but it has few limitation. The translation can be improved by using more complex translation models or API's such as Google Cloud translation or Microsoft Bing Translator.

7 Conclusion and Future Work

In Conclusion, the objective of the study is divided in three parts: Generating Abstractive Text Summary, Sentimental Analysis, English to Hindi Translation on two datsets. One of the dataset is related to Indian News Summary and other one is BBC News Summary. Both the datasets are based on News articles and have news around the globe. To achieve this objectives, models like BART, BERT, and Google Translator are selected for respective parts. All the three parts are achieved using these pre-trained models. The generated summaries are not of very long length due to the imbalanced data. The maximum length of summaries is defined to 50 to 500 in the model. The dataset contains the column with generated summaries and these are then compared with the machine generated summaries using evaluation metrics and the generated scores are a bit lower. The lower score in the result can be due to the sampling techniques used in the study and the elimination of the outliers. Without this filtration BART model was not able to process the execution. The resources available for implementing this research on large scale were lacking. Due to which, implementation was don on small scale. The samples taken from the dataset offered some of the benefits in terms of efficient execution and computation of the model, reducing the time and memory resources engaged in the execution. As consequence, the performance of models affected the results and scores. Therefore, the interpretation of accuracy is considered by keeping this in mind. Further, running the whole dataset may provide more coherent outputs and higher scores and performance from the models.

Future Work: The proposed model can be enhanced further by using techniques

and methods which can improve the accuracy of the summaries. Modifying the usage of models which can handle large data and computational power. For the purpose of sentimental analysis, the combination of NLP techniques with deep learning can be used to gain more emotions out of the sentences. The models can be run on better environment with sufficient resources available. The proposed model can be implemented as a tool in real-time which generates short summaries from the news articles associated with the semantic comments or labels and with the a translation in different language if the readers wish to. This will help more people in India to read the summaries and decide on the basis of semantic labels whether to read the whole article or not.

References

Alshibly, I., Al-Shorfat, S., Otair, M., Shehab, M., Tarawneh, O. and Daoud, M. S. (2023). Text summarization of news articles based on named entity recognition using spacy library, *arXiv preprint arXiv:2303.12345* .

Bello, A., Ng, S. C. and Leung, M.-F. (2023). A bert framework to sentiment analysis of tweets, *Sensors* **23**(1): 506.
URL: *https://www.mdpi.com/1424-8220/23/1/506*

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .
URL: *https://arxiv.org/pdf/1810.04805.pdf*

Gupta, A., Chugh, D., Anjum and Katarya, R. (2022). Automated news summarization using transformers, *Sustainable Advanced Computing: Select Proceedings of ICSAC 2021*, Springer Singapore, Singapore, pp. 249–259.
URL: *https://link.springer.com/chapter/10.1007/978-981-16-9012-9*$_2$1

Gupta, P., Nigam, S. and Singh, R. (2023). A statistical approach for extractive hindi text summarization using machine translation, *Proceedings of Fourth International Conference on Computer and Communication Technologies*, Springer, Singapore, pp. 275–282.

Haider, M. M., Hossin, M. A., Mahi, H. R. and Arif, H. (2020). Automatic text summarization using gensim word2vec and k-means clustering algorithm, pp. 283–286.

Hoang, M., Bihorac, O. and Rouces, J. (2019). Aspect-based sentiment analysis using bert, *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 187–196.
URL: *https://aclanthology.org/W19-6120.pdf*

ItsaLocke (n.d.). Crisp-dm and why you should know about it, `https://itsalocke.com/blog/crisp-dm-and-why-you-should-know-about-it/`. ItsaLocke Blog.

Kahla, M., Yang, Z. and Novák, A. (2021). Cross-lingual fine-tuning for abstractive arabic text summarization, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 655–663.
URL: *https://aclanthology.org/2021.ranlp-1.74.pdf*

Khan, L., Amjad, A., Ashraf, N. and Chang, H.-T. (2022). Multi-class sentiment analysis of urdu text using multilingual bert, *Scientific Reports* **12**(1): 5436.
URL: *https://www.nature.com/articles/s41598-022-09381-9*

Krishnan, D., Bharathy, P., Anagha and Venugopalan, M. (2019). A supervised approach for extractive text summarization using minimal robust features, *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, pp. 521–527.

La Quatra, M. and Cagliero, L. (2022). Bart-it: An efficient sequence-to-sequence model for italian text summarization, *Future Internet* **15**(1): 15.
URL: *https://www.mdpi.com/1999-5903/15/1/15*

Larassati, A., Setyaningsih, N., Nugroho, R. A., Suryaningtyas, V. W., Cahyono, S. P. and Pamelasari, S. D. (2019). Google vs. instagram machine translation: multilingual application program interface errors in translating procedure text genre, *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, IEEE, pp. 554–558.
URL: *https://ieeexplore.ieee.org/abstract/document/8884334*

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, *arXiv preprint arXiv:1910.13461* .
URL: *https://arxiv.org/pdf/1910.13461.pdf*

Library, C. U. (n.d.). Arxiv. Accessed: April 16, 2023.

Manakul, P. and Gales, M. (2020). Cued_speech at trec 2020 podcast summarisation track, *arXiv preprint arXiv:2012.02535* .
URL: *https://arxiv.org/abs/2012.02535*

Moslem, Y., Haque, R., Kelleher, J. D. and Way, A. (2023). Adaptive machine translation with large language models, *ADAPT Centre School of Computing Dublin City University* . Contact Emails: yasmin.moslem@adaptcentre.ie, rejwanul.haque@adaptcentre.ie, john.kelleher@adaptcentre.ie, andy.way@adaptcentre.ie.

Naik, N. V., Prathusha, K. A., Nagari, P., Binjadagi, V. and Shaikh, R. N. (n.d.). Text based sentiment analysis.

Prates, M. O., Avelar, P. H. and Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate, *Neural Computing and Applications* **32**: 6363–6381.
URL: *https://link.springer.com/article/10.1007/s00521-019-04144-6*

ResearchGate (n.d.). Abstractive summarization process, `https://www.researchgate.net/figure/Abstractive-Summarization-Process_fig4_354291874`. Retrieved from ResearchGate.

ResearchGate (n.d.). News sentiment analysis. Accessed: April 16, 2023.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* .
URL: *https://arxiv.org/pdf/1910.01108.pdf*