Innovative Project Report

PROJCS601

B. Tech

Department of Computer Science & Engineering

3rd Year, 6th Semester

(Batch 2021-25)

Group Mentor: Dr. Anupam Mondal

Group Members:

Name	Section	Roll No.	Enrolment No.	Year
Aritra Ghosh	A	74	12021002002137	3rd
Subhojit Ghosh	B	97	12021002002160	3rd

Table of Contents:

Acknowledgement

We extend our heartfelt gratitude to Dr. Anupam Mondal, our mentor, whose unwavering guidance and encouragement propelled us forward in this project. Without his invaluable support, our journey wouldn't have been as successful.

We also express our sincere appreciation to the Institute of Engineering and Management for granting us the opportunity to embark on this remarkable project. It has been an enriching experience that has significantly contributed to our personal and professional growth.

Additionally, we extend our deepest thanks to our parents for their boundless love and unwavering support throughout this endeavor. Their encouragement has been a constant source of strength, driving us to reach new heights.

Overview:

Our goal is to develop a highly effective summarization model that addresses the limitations and challenges encountered by existing models in this domain. To achieve this, we have outlined a comprehensive five-step approach.

In the first step, we will conduct an exhaustive analysis of all existing research efforts to identify models that show promise for our task. This involves scrutinizing various methodologies, techniques, and architectures to pinpoint areas where improvements are necessary.

Moving on to step two, we will delve deeper into the top 10 models identified in the initial analysis. Our focus will be on evaluating these models against key metrics such as accuracy, coherence, and efficiency. This comparative analysis will provide valuable insights into the strengths and weaknesses of each model.

Step three entails the meticulous curation and creation of a custom dataset tailored to the specific requirements of our project. This dataset will serve as the foundation for training and evaluating our summarization models.

In step four, we will embark on the training phase, where we will deploy the top three models identified from our analysis onto the custom dataset. Through rigorous training iterations, we aim to optimize the performance of these models to better suit our summarization objectives.

Finally, in step five, we will consolidate the results obtained from the trained models and fine-tune them further to develop an optimized version. This iterative refinement process will involve tweaking parameters, adjusting algorithms, and incorporating feedback to enhance the overall effectiveness and robustness of our summarization model.

Introduction

Summarization plays a crucial role in distilling the essence of extensive textual content across a spectrum of contexts. In journalism, it enables the rapid dissemination of key information, allowing readers to stay informed amidst their busy schedules. Similarly, in academic research, summarization facilitates the efficient comprehension of complex studies, accelerating the exploration of current literature. Moreover, within corporate environments, summarization streamlines decision-making processes by synthesizing lengthy business reports into concise summaries, providing executives with quick insights.

A multitude of techniques are employed in text summarization, each leveraging distinct methodologies to extract essential information. Fuzzy logic approaches, grounded in fuzzy set theory, adeptly handle imprecise or ambiguous information present in texts.

Machine learning algorithms play a significant role in text summarization, encompassing various models such as Naive Bayes, Decision Trees, Support Vector Machines, and neural network architectures like Recurrent Neural Networks (RNNs) and Transformers. These algorithms are trained on labeled datasets to discern patterns conducive to summarization, enabling the generation of informative summaries.

Additionally, summarization techniques include conditional random fields (CRFs), which treat summarization as a sequence labeling task, identifying and extracting salient sentences or phrases. Tree-based approaches, such as Tree Summarization or Rhetorical Structure Theory, scrutinize the discourse structure of the text to pinpoint pertinent information. Template-driven methods rely on predefined templates or schemas to extract relevant details from the text and generate summaries based on populated templates.

Rule-based techniques apply manually crafted rules or heuristics to identify and extract crucial information for summarization. The ontology method capitalizes on domain-specific ontologies or knowledge bases to prioritize essential concepts and information, while multimodal semantic approaches synthesize multiple semantic representations to generate comprehensive summaries.

Semantic graph-based techniques construct a semantic graph depicting relationships between entities, concepts, and events in the text, which is then utilized for summarization purposes. Despite the efficacy of these methods, challenges persist, including oversimplification, dataset biases, reliance on static knowledge bases, and a predominant focus on extractive rather than abstractive summarization.

Work Methodology

Our methodology is structured into five distinct stages:

1. **Literature Review and Model Analysis:**

    - Conduct an exhaustive review of existing research to identify prevalent summarization models and ascertain areas in need of improvement.
    - Evaluate the effectiveness of each model against predefined criteria, laying the groundwork for subsequent analysis.

2. **Comparative Analysis of Top Models:**

    - Identify and select the top ten summarization models based on their prominence and relevance in the field.
    - Perform a comprehensive comparative analysis, scrutinizing these models across key metrics such as coherence, informativeness, and fluency.

3. **Dataset Curation and Development:**

    - Curate a bespoke dataset tailored to the nuances and intricacies of the summarization task, encompassing diverse sources and topics.
    - Ensure the integrity and representativeness of the dataset to facilitate robust model training and evaluation.

4. **Model Training and Evaluation:**

    - Utilize the top three selected models to train on the custom dataset, leveraging state-of-the-art techniques and algorithms.
    - Employ rigorous evaluation methodologies to assess the performance and efficacy of each model in generating accurate and coherent summaries.

5. **Fine-Tuning and Optimization:**

    - Aggregate the results obtained from model evaluations, identifying strengths, weaknesses, and areas for improvement.
    - Employ iterative refinement techniques to fine-tune the selected models, optimizing their performance for enhanced summarization outcomes.

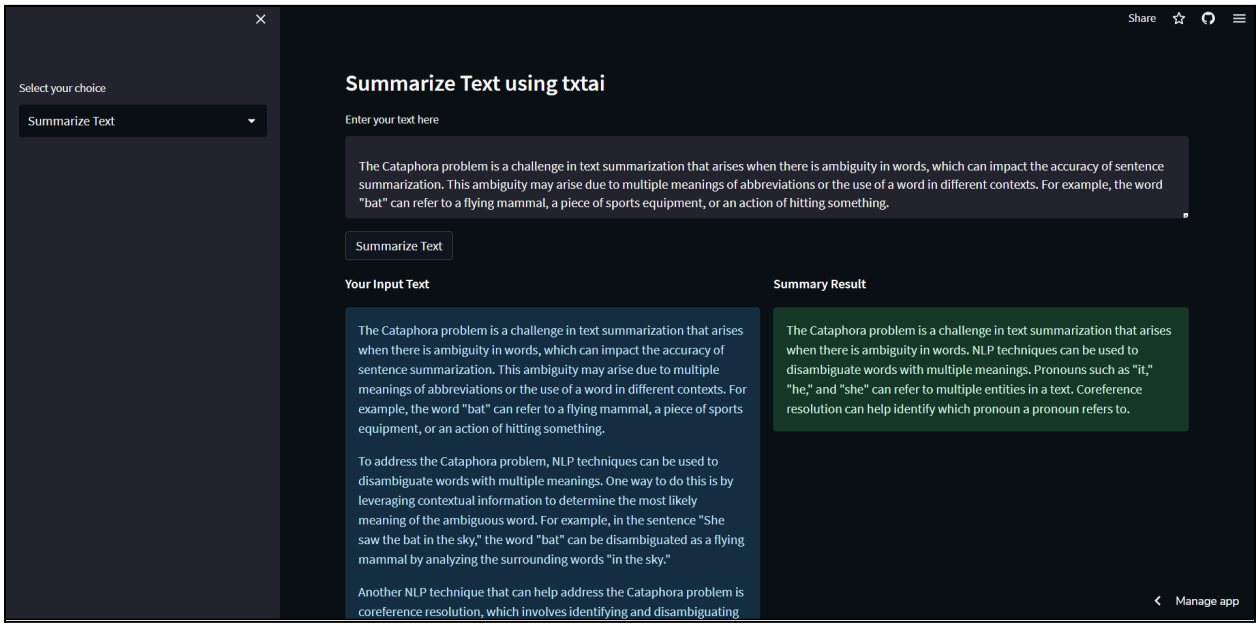
Timeline Chart

Step No.	Description	Status	Result
1.	**Literature Review and Model Analysis** **a.** Generation of Report of Survey, Research Paper Review **b.** Develop a basic Model Using inbuilt library functions in Python to gain a better understanding of the problem and ask in hand	Completed	**a.** https://docs.google.com/document/d/1sNxGtEG7QVPPM6zPaM07_Tsb2P3EmxzEy48yHFyzX80/edit?usp=sharing **b.** https://github.com/TheCleverIdiott/summarizer
2.	**Comparative Analysis of Top Models** **a.** Comparative Analysis of Top Models **b.** Research Paper Publication - I	Completed	a. https://iemtronics.org/wp-content/uploads/2024/04/IEMTRONICS-2024-Conference-Proceedings-1.pdf b. https://iemtronics.org/wp-content/uploads/2024/04/IEMTRONICS-2024-Conference-Proceedings-1.pdf
3.	**Dataset Curation and Development** a. Creating a Custom Dataset from several Genres that target humane text summary to train Further Models	Completed	a. https://mega.nz/file/sdBwHQRI#yykHpuwcHldDz-K_EG-qhXGcgvvJyO4eNydVMEkKWhs
4.	**Model Training and Evaluation**	Ongoing	a. https://huggingface.co/spaces/TheCleverIdiot/Text_Summarization/
5.	**Fine-Tuning and Optimization**	Scheduled	

Table 1: Timeline

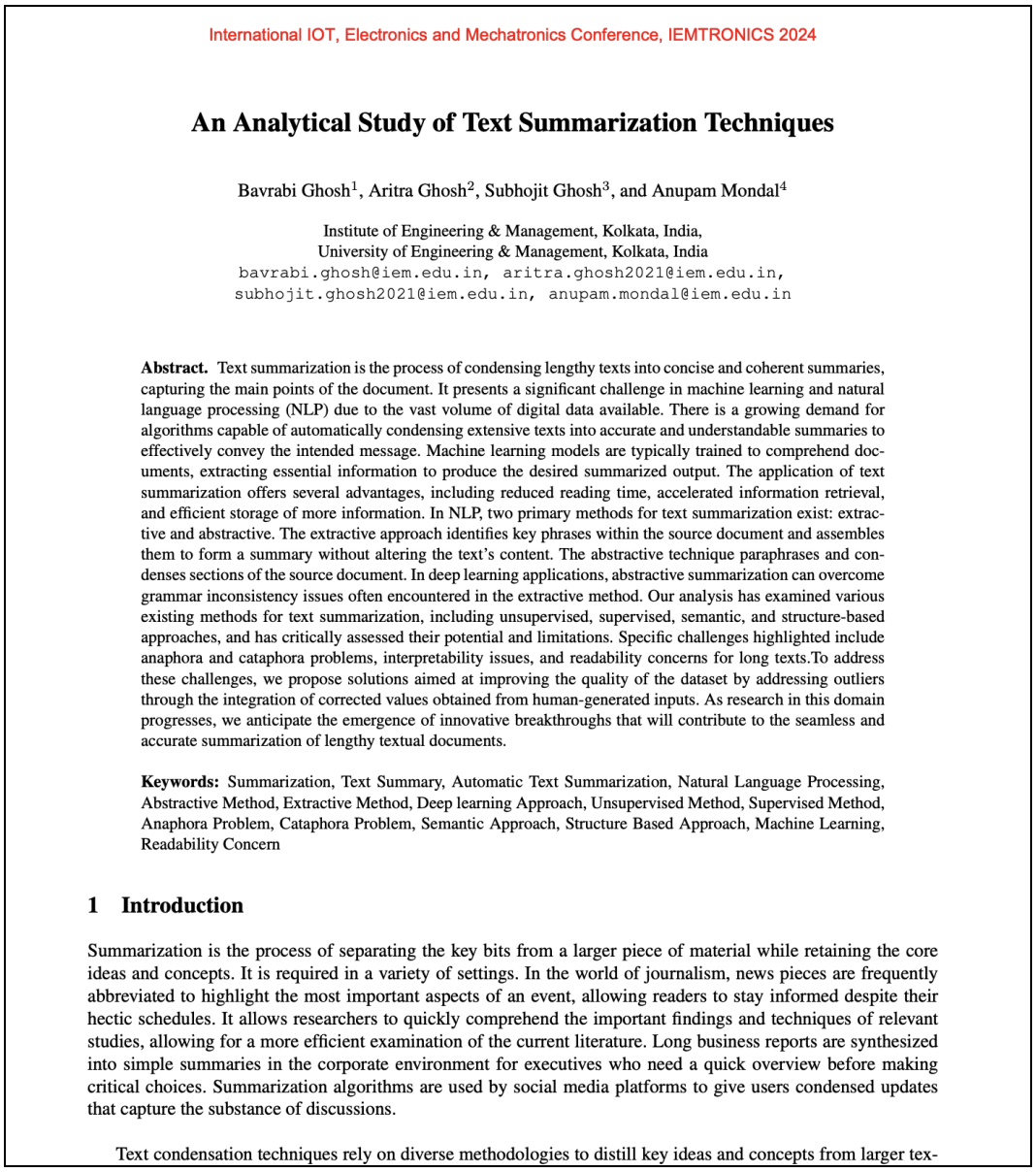
Completed Work

Developed an library bsed summarization model using txtai and python. (https://github.com/TheCleverIdiott/summarizer)



We embarked on an exhaustive assessment of various text summarization techniques, employing the CNN Daily dataset as our benchmark. Through meticulous analysis, we systematically ranked the performance of each method to discern their efficacy.

Informed by an in-depth exploration of existing scholarly works, we contributed to the field by publishing our research paper titled "An Analytical Survey of Text Summarization Techniques." This comprehensive study was presented at the prestigious IEMTronics, International IOT, Electronics and Mechatronics Conference, hosted from April 3rd to 5th, 2024, at Imperial College London, UK. (https://iemtronics.org/wp-content/uploads/2024/04/IEMTRONICS-2024-Conference-Proceedings-1.pdf)



International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2024

An Analytical Study of Text Summarization Techniques

Bavrabi Ghosh[1], Aritra Ghosh[2], Subhojit Ghosh[3], and Anupam Mondal[4]

Institute of Engineering & Management, Kolkata, India,
University of Engineering & Management, Kolkata, India
bavrabi.ghosh@iem.edu.in, aritra.ghosh2021@iem.edu.in,
subhojit.ghosh2021@iem.edu.in, anupam.mondal@iem.edu.in

Abstract. Text summarization is the process of condensing lengthy texts into concise and coherent summaries, capturing the main points of the document. It presents a significant challenge in machine learning and natural language processing (NLP) due to the vast volume of digital data available. There is a growing demand for algorithms capable of automatically condensing extensive texts into accurate and understandable summaries to effectively convey the intended message. Machine learning models are typically trained to comprehend documents, extracting essential information to produce the desired summarized output. The application of text summarization offers several advantages, including reduced reading time, accelerated information retrieval, and efficient storage of more information. In NLP, two primary methods for text summarization exist: extractive and abstractive. The extractive approach identifies key phrases within the source document and assembles them to form a summary without altering the text's content. The abstractive technique paraphrases and condenses sections of the source document. In deep learning applications, abstractive summarization can overcome grammar inconsistency issues often encountered in the extractive method. Our analysis has examined various existing methods for text summarization, including unsupervised, supervised, semantic, and structure-based approaches, and has critically assessed their potential and limitations. Specific challenges highlighted include anaphora and cataphora problems, interpretability issues, and readability concerns for long texts.To address these challenges, we propose solutions aimed at improving the quality of the dataset by addressing outliers through the integration of corrected values obtained from human-generated inputs. As research in this domain progresses, we anticipate the emergence of innovative breakthroughs that will contribute to the seamless and accurate summarization of lengthy textual documents.

Keywords: Summarization, Text Summary, Automatic Text Summarization, Natural Language Processing, Abstractive Method, Extractive Method, Deep learning Approach, Unsupervised Method, Supervised Method, Anaphora Problem, Cataphora Problem, Semantic Approach, Structure Based Approach, Machine Learning, Readability Concern

1 Introduction

Summarization is the process of separating the key bits from a larger piece of material while retaining the core ideas and concepts. It is required in a variety of settings. In the world of journalism, news pieces are frequently abbreviated to highlight the most important aspects of an event, allowing readers to stay informed despite their hectic schedules. It allows researchers to quickly comprehend the important findings and techniques of relevant studies, allowing for a more efficient examination of the current literature. Long business reports are synthesized into simple summaries in the corporate environment for executives who need a quick overview before making critical choices. Summarization algorithms are used by social media platforms to give users condensed updates that capture the substance of discussions.

Text condensation techniques rely on diverse methodologies to distill key ideas and concepts from larger tex-

To facilitate our experimentation, we meticulously curated a custom dataset comprising 200,000 data points, each consisting of ['article', 'summary'] pairs. Recognizing the importance of training context, we ensured the inclusion of diverse data from various regions and fields of topics to address any potential biases.
(https://mega.nz/file/sdBwHQRI#yykHpuwcHldDz-K_EG-qhXGcgvvJyO4eNydVMEkKWhs)

	article	highlights
afc246a7964130f43ae940af6bc6c57f01	By . Associated Press . PUBLISHED: . 14:11 EST, 25 October 2013 . | . UPDATED: . 15:36 EST, 25 October 2013 . The bishop of the Fargo Catholic Diocese in North Dakota has exposed potentially hundreds of church memb	Bishop John Folda, of North Dakota, is taking time off after being diagnosed He contracted the infection through contaminated food in Italy . Church members in Fargo, Grand Forks and Jamestown could have been ex
5e55fcbd3a2f366d9bf92a95433dc305ef	(CNN) -- Ralph Mata was an internal affairs lieutenant for the Miami-Dade Police Department, working in the division that investigates allegations of wrongdoing by cops. Outside the office, authorities allege that the 45-year-(	Criminal complaint: Cop used his role to help cocaine traffickers . Ralph Mata, an internal affairs lieutenant, allegedly helped group get guns . He also arranged to pay two assassins in a murder plot, a complaint alleges.
965c8264c35cc1bc55556db388da82b07f	A drunk driver who killed a young woman in a head-on crash while checking his mobile phone has been jailed for six years. Craig Eccleston-Todd, 27, was driving home from a night at a pub when he received a text message	Craig Eccleston-Todd, 27, had drunk at least three pints before driving car . Was using phone when he veered across road in Yarmouth, Isle of Wight . Crashed head-on into 28-year-old Rachel Titley's car, who died in hospital . Police say he would have been over legal drink-drive limit at time of crash . He was found guilty at Portsmouth Crown Court of causing death by danger
7436637c4fe1837c935c04de47adb18e9a	(CNN) -- With a breezy sweep of his pen President Vladimir Putin wrote a new chapter into Crimea's turbulent history, committing the region to a future returned to Russian domain. Sixty years prior, Ukraine's breakaway peni	Nina dos Santos says Europe must be ready to accept sanctions will hurt bot Targeting Russia's business community would be one way of sapping their s But she says Europe would have a hard time keeping its factories going with
6ef0c37534f80b55b4235108024b407f0b	Fleetwood are the only team still to have a 100% record in Sky Bet League One as a 2-0 win over Scunthorpe sent Graham Alexander's men top of the table. The Cod Army are playing in the third tier for the first time in their	Fleetwood top of League One after 2-0 win at Scunthorpe . Peterborough, Bristol City, Chesterfield and Crawley all drop first points of th Stand-in striker Matt Done scores a hat-trick as Rochdale thrash Crewe 5-2 . Wins for Notts County and Yeovil . Coventry/Bradford and Oldham/Port Vale both end in draws . A late Stephen Bywater own goal denies Gillingham three points against Milt
6354494f090ee2d7bc5ddbf80b63e80de6	He's been accused of making many a fashion faux pas while on holiday. But the Prime Minister seems to be deaf to his critics. Yesterday David Cameron was seen in the same pair of beige loafers he wore on holiday last yea	Prime Minister and his family are enjoying an Easter break in Lanzarote . Sported the same £20.99 beige loafers as he wore in Portugal last year . PM sat and had a drink at a beach-side cafe on the Spanish Island .
1497d21ff37a17751829bd7e3b6e4a7c5c	By . Daily Mail Reporter . PUBLISHED: . 01:15 EST, 30 November 2013 . | . UPDATED: . 01:23 EST, 30 November 2013 . More than two decades after Magic Johnson announced that he had HIV, the basketball player says he	NBA star calls for black and Hispanic communities to get tested . Former Lakers player dedicated life to raising awareness about disease .
1f772fad0aa78a977ce4a31b3faa6e6fe5	By . Daily Mail Reporter . This is the moment a train announcer stunned passengers by announcing over a tannoy as they pulled into a station to beware of pickpockets and gipsies. The London Midland service had been pull	London Midland service had been pulling into Telford Station in Shropshire . Passenger Chris Downes, 46, was recording on his mobile at the time . Announcer can clearly be heard saying: "Telford Central - please be aware of London Midland said it is now launching an investigation into the incident .
97263e215e5e7eda753070f08aa374dd45	There are a number of job descriptions waiting for Darren Fletcher when he settles in at West Brom but the one he might not have expected is Saido Berahino's nanny. Fletcher's unveiling as the deadline day signing from Ma	Tony Pulis believes Saido Berahino should look up to Darren Fletcher . Pulis insists Berahino has been listened to the wrong advice . Berahino said he wants to move on to bigger things earlier in the week . READ: Berahino available for £20m after Liverpool target angers club . CLICK HERE for all the latest West Brom news .
f2bb357ac04a236a232156d8b9b18d1667	Canberra, Australia (CNN) -- At first glance, it doesn't look like much. Hidden behind an unmarked door, in a nondescript government office building in the Australian capital, it could be mistaken for a high school science clas	Black box data from Flight 370 could be analyzed at a laboratory in Australia Even if the flight data recorder is damaged, information is retrievable . About 2,000 parameters are decoded, like altitude, engine performance and The data is used to create a visual representation, helping the public underst
0b1967511741629926ef9f5faea2bb6be24	By . Ellie Zolfagharifard . Take a look at a map today, and you're likely to see that North America is larger than Africa, Alaska is larger than Mexico and China is smaller than Greenland. But in reality China is four times bigger t He argues that all maps are of their time, their place and serve certain purposes. 'No world map is, or can be, a definitive, transparent depiction of its subject that offers a disembodied eye onto the world,' he writes. "Each on	The distortion is the result of the Mercator map which was created in 1596 to It gives the right shapes of countries but at the cost of distorting sizes in favo For instance, north America looks larger, or at least as big, as Africa, and Gre In reality, you can fit north America into Africa and still have space for India, A Map suggests Scandinavian countries are larger than India, whereas in reality The biggest challenge for cartographers is that it is impossible to portray real
5555db62e319854d9f8912061cdca1893e	Two lawyers representing a woman who . claims to have had sex as a minor with prominent U.S. criminal defense lawyer Alan Dershowitz have filed a counter-defamation . lawsuit against him. Former federal judge Paul Cass	Alan Dershowitz has filed defamation suits against two other U.S lawyers . He is accused of having sex with Virginia Roberts when she was a minor . Dershowitz says that the two lawyers representing her have defamed him . Those two lawyers are now counter-suing him for defamation . Paul Cassell and Bradley Edwards say their character has been attacked .
1fc9d877f8d4bb2ebd1d6858c69be571fd8	It's the moment every pet owner dreads - when the time comes when they have to say a final goodbye to a faithful friend. These heart-breaking end-of-life snaps are meant to highlight the special relationship between an own	Sarah Ernhart, the owner of Sarah Beth Photography in Minneapolis, created She dubbed the shoot a 'Joy Session' in which she records owners' last mor Her service has been so popular Mrs Ernhart has had more than 100 shoots
ee0098c4d510a03ddc97d11764448ebac2	Louis van Gaal said he had no option but to substitute Paddy McNair in the first half against Southampton because the defender's 'confidence' was shot - but believes that it will benefit the youngster in the long run. The 19-	Manchester United beat Southampton 2-1 at St Mary's on Monday night . Paddy McNair was substituted by Louis van Gaal after only 39 minutes . Van Gaal admitted he 'had to' replace the 19-year-old against Saints . United boss said McNair 'had no confidence' after struggling early on . But Van Gaal is adamant substitution was "in best interests' of McNair .
9f6b1d954d827c9a550f3f24a5474ee82b	(CNN) -- One can hardly read the news these days without learning that yet another American corporation has announced plans to invert, which is corporate-speak for restructuring as a foreign company to avoid U.S. taxes. I	U.S. corporations merge with foreign companies, move their headquarters . McIntyre: Such "inversions" enable firms to greatly lower their U.S. corporate He says government can lose billions of tax revenue from such maneuvers . McIntyre: Congress should pass administration proposal to bar inversions .
a19e2c96de11276b3cce11566ccfed0030	For most people, it has become a travel essential. Taking your smartphone or tablet away on holiday keep you in touch with what's going on back home, as well as offering a chance to monitor 'work emails.' But a 'digital det	Half of Brits admit to checking work e-mails while on holiday, while a third re Rural getaways are becoming more popular in 'digital detox' revolution, many Offers a chance to leave smartphones and tablets firmly switched off and enj
70d3fc53ea7f2a4a50f6496dc2df17e02f	By . Margot Peppers . Nigerian and Cameroonian pop star Dencia has hit out at Lupita Nyong'o for her new contract with Lancome, accusing her of bowing to 'white people companies'. In an angry tweet directed at the 12 Y	Dencia's comment is hypocritical considering she recently courted controver
91cfcab17bac165f7a4719019a628a9db3	Britain and the West must brace themselves for more bloody atrocities before Islamist jihadists in Iraq are defeated, former top brass said last night. Retired commanders issued the chilling warning as they urged David Cam have been more careful and shown restraint but enough is enough. 'I'm prepared to take more risks to defy these evil people. I utterly condemn IS even if it puts me at personal risk, at danger of people coming to my home. I	Ex-military chiefs suggested . stepping up Special Forces operations to 'spoil the day' of fanatics, including Iraq . Air Chief Marshal Sir Michael Graydon, and Air Commodore Andrew Lambert up military options in Iraq . They spoke out after gruesome murderer of US journalist James Foley – appa air strikes .

In our approach, we leveraged a pre-trained BERT model to analyze initial performance metrics and subsequently fine-tuned a custom BERT model on the curated dataset. By integrating the strengths of these state-of-the-art models with our tailored dataset, our aim was to elevate the quality and generalizability of text summarization techniques. This strategic amalgamation allowed us to explore the full potential of advanced summarization methodologies and pave the way for more effective and adaptable solutions in this domain.

Ongoing Work

Currently, our focus is on refining a customized BERT model using our meticulously curated dataset comprising 200,000 ['article', 'summary'] pairs sourced from diverse regions and topics. Our objective in this fine-tuning phase is to tailor the language understanding capabilities of the pre-trained BERT model to better suit the intricacies of text summarization.

To achieve this goal, we are experimenting with a range of strategies including hyperparameter tuning, gradient clipping, warm-up techniques, and attention visualization methods. These endeavors aim to optimize both the performance and interpretability of our model, ensuring it can effectively distill key information from the input articles to generate concise summaries.

Moreover, we are exploring advanced techniques such as the integration of graph neural networks and attention mechanisms. By incorporating these methodologies, we aim to enhance our model's ability to capture semantic relationships and contextual nuances present within the text, thereby improving the quality and coherence of the generated summaries.

Through this comprehensive approach of fine-tuning on our tailored dataset, we seek to harness the inherent strengths of BERT while adapting it to the specific requirements of text summarization. Our ultimate aim is to develop a robust and versatile model capable of generating high-quality summaries across a wide range of domains and topics, thereby advancing the state-of-the-art in text summarization technology.

Breaking the 200,000 dataset into segments of 4 x 50,000 for easier training

```
In [2]: import pandas as pd

In [3]: df = pd.read_csv('train.csv', nrows=50000)

In [4]: df = df.rename(columns={'highlights': 'summary'})
        df = df.drop(columns=['id'])

In [5]: df.columns
Out[5]: Index(['article', 'summary'], dtype='object')

In [6]: num_rows = df.shape[0]

In [7]: df.to_csv("f50k.csv", index=False)

In [ ]:
```

Training and performing BERT on the first 50k sets of data

```
In [1]: import pandas as pd

In [2]: df = pd.read_csv('train50k.csv')

In [3]: df.columns
Out[3]: Index(['article', 'summary'], dtype='object')

In [4]: null_values = df.isnull().sum()
        null_values
Out[4]: article    0
        summary    0
        dtype: int64

In [5]: from sklearn.model_selection import train_test_split

In [6]: # Split the DataFrame into features (X) and target (y)
        X = df['article']
        y = df['summary']

        # Split the data into training (80%) and the rest (20%)
        X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.2, random_state=42)

        # Further split the rest into validation (50%) and testing (50%)
        X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)

        # Now we have X_train, y_train for training
        # X_val, y_val for validation
        # X_test, y_test for final testing

In [7]: from transformers import BartTokenizer, BartForConditionalGeneration, AdamW
        from tqdm import tqdm
        import torch

In [8]: tokenizer = BartTokenizer.from_pretrained("facebook/bart-large-cnn")

In [9]: train_encodings = tokenizer(list(X_train), truncation=True, padding=True, max_length=512, return_tensors="pt"

In [10]: model = BartForConditionalGeneration.from_pretrained("facebook/bart-large-cnn")
```

```
In [9]: rain_encodings = tokenizer(list(X_train), truncation=True, padding=True, max_length=512, return_tensors="pt", return_

In [10]: model = BartForConditionalGeneration.from_pretrained("facebook/bart-large-cnn")

In [11]: model.train()

Out[11]: BartForConditionalGeneration(
           (model): BartModel(
             (shared): Embedding(50264, 1024, padding_idx=1)
             (encoder): BartEncoder(
               (embed_tokens): Embedding(50264, 1024, padding_idx=1)
               (embed_positions): BartLearnedPositionalEmbedding(1026, 1024)
               (layers): ModuleList(
                 (0): BartEncoderLayer(
                   (self_attn): BartAttention(
                     (k_proj): Linear(in_features=1024, out_features=1024, bias=True)
                     (v_proj): Linear(in_features=1024, out_features=1024, bias=True)
                     (q_proj): Linear(in_features=1024, out_features=1024, bias=True)
                     (out_proj): Linear(in_features=1024, out_features=1024, bias=True)
                   )
                   (self_attn_layer_norm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
                   (activation_fn): GELUActivation()
                   (fc1): Linear(in_features=1024, out_features=4096, bias=True)
                   (fc2): Linear(in_features=4096, out_features=1024, bias=True)
                   (final_layer_norm): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
                 )

In [12]: optimizer = AdamW(model.parameters(), lr=5e-5)

         /Users/aritra/anaconda3/lib/python3.10/site-packages/transformers/optimization.py:306: FutureWarning: This implementation of AdamW is deprecated and will be removed in a future version. Use the PyTorch implementation torch.optim.AdamW instead, or set `no_deprecation_warning=True` to disable this warning
           warnings.warn(

In [13]: y_train_tensors = tokenizer(list(y_train), truncation=True, padding=True, max_length=128, return_tensors="pt", retur

In [*]: # Training loop
        batch_size = 4
        num_epochs = 3
        for epoch in range(num_epochs):
            for i in tqdm(range(0, len(train_encodings["input_ids"]), batch_size)):
                # Clear gradients
                optimizer.zero_grad()
```

```
y_train_tensors = tokenizer(list(y_train), truncation=True, padding=True, max_length=128, return_tensors="pt", retur
```

Since BART doesn't have an in-built train function we push the model into a training loop of epoch = 3 and batch size = 4

```
# Training loop
batch_size = 4
num_epochs = 3
for epoch in range(num_epochs):
    for i in tqdm(range(0, len(train_encodings["input_ids"]), batch_size)):
        # Clear gradients
        optimizer.zero_grad()

        # Get batch inputs
        batch_inputs = {key: val[i:i+batch_size] for key, val in train_encodings.items()}
        batch_labels = {key: val[i:i+batch_size] for key, val in y_train_tensors.items()}

        # Shift labels to the right
        decoder_input_ids = batch_labels["input_ids"].clone()
        decoder_input_ids[:, :-1] = decoder_input_ids[:, 1:]
        decoder_input_ids[:, -1] = tokenizer.pad_token_id

        # Forward pass
        outputs = model(**batch_inputs, decoder_input_ids=decoder_input_ids)

        # Compute loss
        loss = outputs.loss

        # Check if loss is None
        if loss is None:
            print("Warning: Loss is None!")
            continue

        # Backward pass
        loss.backward()

        # Update parameters
        optimizer.step()
```

Output from the trained Model

```
In [1]: from transformers import pipeline

        summarizer = pipeline("summarization", model="facebook/bart-large-cnn")
```

Downloading: 100% 1.36M/1.36M [00:01<00:00, 1.39MB/s]

```
In [2]: ARTICLE = """ New York (CNN)When Liana Barrientos was 23 years old, she got married in Westchester County, New York.
        A year later, she got married again in Westchester County, but to a different man and without divorcing her first hu
        Only 18 days after that marriage, she got hitched yet again. Then, Barrientos declared "I do" five more times, somet
        In 2010, she married once more, this time in the Bronx. In an application for a marriage license, she stated it was
        Barrientos, now 39, is facing two criminal counts of "offering a false instrument for filing in the first degree," r
        2010 marriage license application, according to court documents.
        Prosecutors said the marriages were part of an immigration scam.
        On Friday, she pleaded not guilty at State Supreme Court in the Bronx, according to her attorney, Christopher Wright
        After leaving court, Barrientos was arrested and charged with theft of service and criminal trespass for allegedly s
        Annette Markowski, a police spokeswoman. In total, Barrientos has been married 10 times, with nine of her marriages
        All occurred either in Westchester County, Long Island, New Jersey or the Bronx. She is believed to still be married
        Prosecutors said the immigration scam involved some of her husbands, who filed for permanent residence status shortl
        Any divorces happened only after such filings were approved. It was unclear whether any of the men will be prosecute
        The case was referred to the Bronx District Attorney\'s Office by Immigration and Customs Enforcement and the Depart
        Investigation Division. Seven of the men are from so-called "red-flagged" countries, including Egypt, Turkey, Georgi
        Her eighth husband, Rashid Rajput, was deported in 2006 to his native Pakistan after an investigation by the Joint T
        If convicted, Barrientos faces up to four years in prison.  Her next court appearance is scheduled for May 18.
        """

In [3]: print(summarizer(ARTICLE, max_length=130, min_length=30, do_sample=False))

        [{'summary_text': 'Liana Barrientos, 39, is charged with two counts of "offering a false instrument for filing in t
        he first degree" In total, she has been married 10 times, with nine of her marriages occurring between 1999 and 200
        2. She is believed to still be married to four men.'}]
```

Rogue Metrics:

```
1  rouge_names = ["rouge1", "rouge2", "rougeL", "rougeLsum"]
2  rouge_dict = dict((rn, score[rn].mid.fmeasure ) for rn in rouge_names )
3
4  pd.DataFrame(rouge_dict, index = ['pegasus'])
```

	rouge1	rouge2	rougeL	rougeLsum
pegasus	0.015596	0.000296	0.015525	0.015562

Links

- **Github:** https://github.com/TheCleverIdiott/Innovative_Project
- **Dataset:** https://mega.nz/file/sdBwHQRI#yykHpuwcHldDz-K_EG-qhXGcgvvJyO4eNydVMEkKWhs
- **Implementation**: https://huggingface.co/spaces/TheCleverIdiot/Text_Summarization/

Future Scope

Looking ahead, our future endeavors involve delving into the utilization of pre-trained LSTM and Pegasus models for text summarization tasks. To realize this vision, we plan to embark on the training of custom LSTM and Pegasus models using our meticulously curated dataset.

Furthermore, we aspire to explore the potential of an ensemble approach by combining the strengths of BERT, Pegasus, and LSTM architectures. This ambitious strategy entails accumulating and fine-tuning these distinct models to create a unified ensemble model. By integrating multiple architectures, we aim to capitalize on their respective strengths and potentially achieve superior performance in text summarization.

The envisioned ensemble model has the capability to capture various facets of the data, including semantic relationships, contextual nuances, and long-range dependencies. This holistic approach is anticipated to yield more coherent and informative summaries, thereby enhancing the overall quality of summarization outputs.

We believe that this multi-model strategy, coupled with our tailored dataset, will serve as a catalyst for advancements in robust and generalizable text summarization techniques. By pushing the boundaries of current methodologies and embracing innovative approaches, we are poised to make significant strides towards the realization of more efficient and effective text summarization solutions.

Conclusion

In conclusion, this project represents a significant stride towards advancing the field of summarization through the development of an optimized model. By leveraging a systematic approach encompassing thorough analysis, comparative evaluation, dataset curation, and iterative refinement, we lay the groundwork for future innovations in this domain. The insights garnered from this endeavor pave the way for the creation of more efficient and effective summarization solutions capable of meeting the evolving needs of information processing in the digital age.

References

1. Aker, A., Cohn, T., Gaizauskas, R.: Multi-document summarization using a* search and discriminative learning. In: Pro- ceedings of the 2010 conference on empirical methods in natural language processing. pp. 482–491 (2010)
2. AL-Banna, A.A., AL-Mashhadany, A.K.: Natural language processing for automatic text summarization [datasets]-survey. Wasit Journal of Computer and Mathematics Science 1(4), 156–170 (2022)
3. Bala, A., Mitra, R., Mondal, A.: Recommendation system to predict best academic program. In: 2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech). pp. 1–6. IEEE (2023)
4. Dey, M., Mondal, A., Das, D.: Ntcir-12 mobileclick: Sense-based ranking and summarization of english queries. In: Ntcir (2016)
5. Edmundson, H.P.: New methods in automatic extracting. Journal of the ACM (JACM) 16(2), 264–285 (1969)
6. Ferreira, R., de Souza Cabral, L., Lins, R.D., e Silva, G.P., Freitas, F., Cavalcanti, G.D., Lima, R., Simske, S.J., Favaro, L.: Assessing sentence scoring techniques for extractive text summarization. Expert systems with applications 40(14), 5755–5764 (2013)
7. Garc´ıa-Herna´ndez, R.A., Ledeneva, Y.: Word sequence models for single text summarization. In: 2009 Second Interna- tional Conferences on Advances in Computer-Human Interactions. pp. 44–48. IEEE (2009)
8. Kupiec, J., Pedersen, J., Chen, F.: A trainable document summarizer. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 68–73 (1995)
9. Mahata, S.K., Mondal, A., Dey, M., Sarkar, D.: Sentiment analysis using machine translation. In: Applications of Machine intelligence in Engineering, pp. 371–377. CRC Press (2022)
10. Mondal, A., Cambria, E., Dey, M.: An annotation system of a medical corpus using sentiment-based models for summa- rization applications. In: Computational Intelligence Applications for Text and Sentiment Data Analysis, pp. 163–178. Elsevier (2023)
11. Mondal, A., Dey, M., Das, D., Nagpal, S., Garda, K.: Chatbot: An automated conversation system for the educational domain. In: 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP).
12. pp. 1–5. IEEE (2018)
13. Mondal, A., Dey, M., Mahata, S.K., Sarkar, D.: An automatic summarization system to understand the impact of covid-19 on education. In: Applications of Machine intelligence in Engineering, pp. 379–386. CRC Press (2022)
14. Mridha, M.F., Lima, A.A., Nur, K., Das, S.C., Hasan, M., Kabir, M.M.: A survey of automatic text summarization: Progress, process and challenges. IEEE Access 9, 156043–156070 (2021)
15. Sinha, S., Mandal, S., Mondal, A.: Question answering system-based chatbot for health care. In: Proceedings of the Global AI Congress 2019. pp. 71–80. Springer (2020)

16. Syed, A.A., Gaol, F.L., Matsuo, T.: A survey of the state-of-the-art models in neural abstractive text summarization. IEEE Access 9, 13248–13265 (2021)
17. Tas, O., Kiyani, F.: A survey automatic text summarization. PressAcademia Procedia 5(1), 205–213 (2007)
18. Zhang, Y., Zincir-Heywood, N., Milios, E.: Narrative text classification for automatic key phrase extraction in web docu- ment corpora. In: Proceedings of the 7th annual ACM international workshop on Web information and data management.
19. pp. 51–58 (2005)