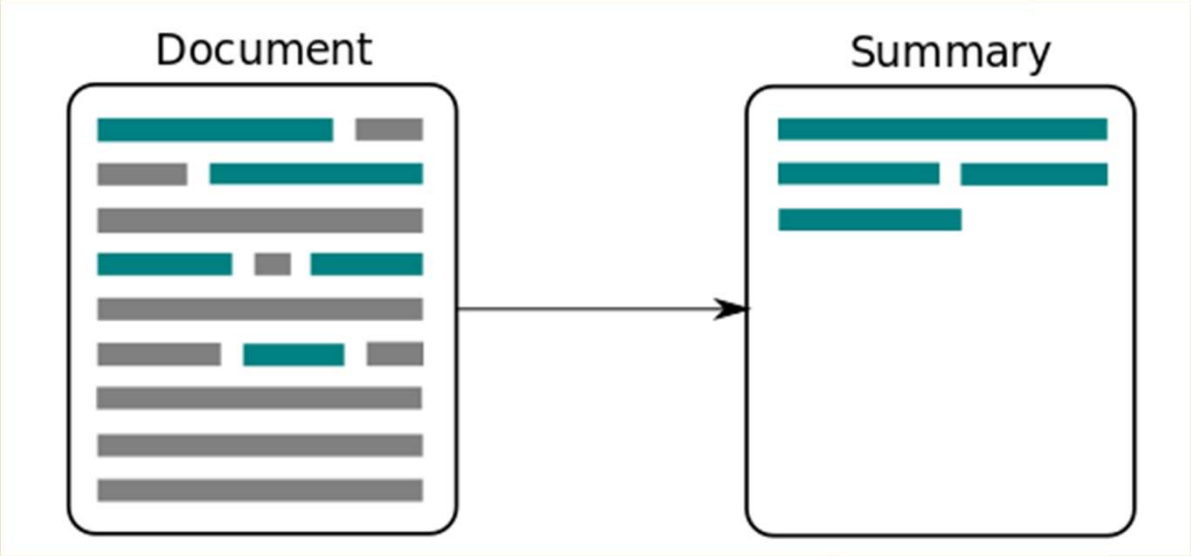
National Institute of Technology Andhra Pradesh

NEWS ARTICLE SUMMARIZATION

by
Ashish-521116
Massevale Karishma Taj-521159

Application of news article summarization

- Research and Data Analysis
- Professional Decision-Making
- Journalism and Media
- Content Aggregation Platforms
- Personalized News Feed
- Education and Learning



Contents

Introduction

There are 2 summarization types:

1.Extractive Summarization: Extractive summarization involves identifying and extracting key phrases, sentences, or segments directly from the original text to form a condensed version. It's akin to highlighting parts of the text that are deemed most informative or relevant.

2.Abtractive Summarization: Abstractive summarization goes beyond mere extraction; it involves understanding the main ideas and then expressing them in new words. It's akin to reading a text and then explaining it in one's own words.

BERT MODEL FOR EXTRACTIVE SUMMARIZATION

1. BERT (Bidirectional Encoder Representations from Transformers) leverages a transformer-based neural network to understand and generate human-like language. BERT employs an encoder-only architecture. In the original Transformer architecture, there are both encoder and decoder modules. The decision to use an encoder-only architecture in BERT suggests a primary emphasis on understanding input sequences rather than generating output sequences.
2. 2.Traditional language models process text sequentially, either from left to right or right to left. This method limits the model's awareness to the immediate context preceding the target word. BERT uses a bi-directional approach considering both the left and right context of words in a sentence, instead of analyzing the text sequentially, BERT looks at all the words in a sentence simultaneously.

3.Extractive summarization aims to select the most relevant sentences from an article to create a summary. BERT, with its powerful embeddings, plays a crucial role in this process.

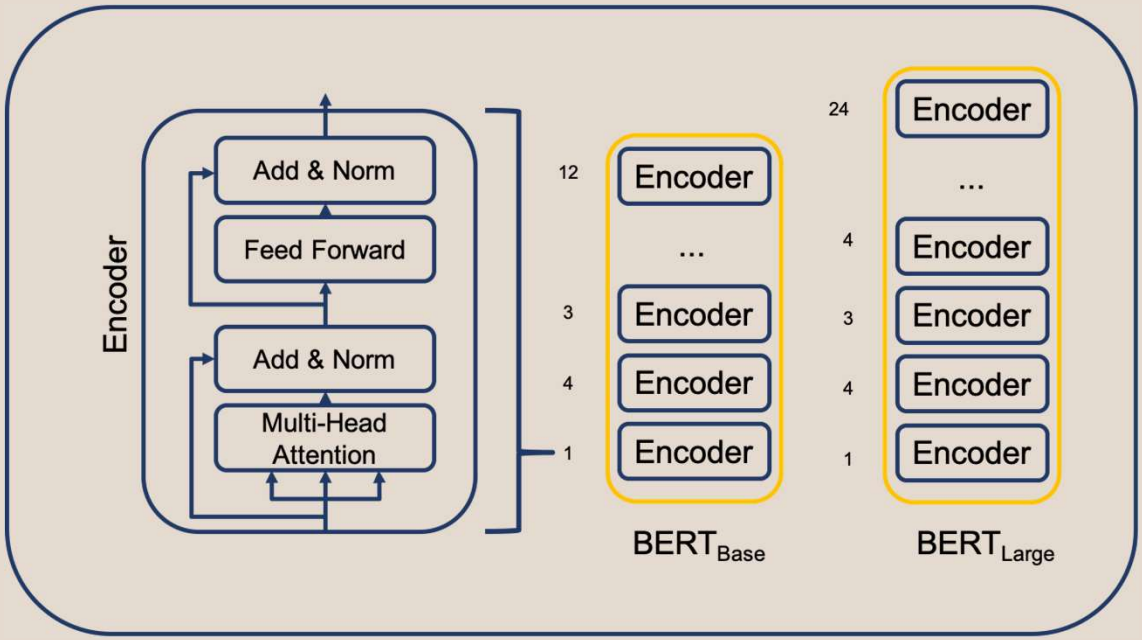
- Here are two common approaches for extractive summarization using BERT:

    a.Embeddings-Based Approach:

    b. Sequential Information Approach.

4.Metrics for Evaluation

BERT MODEL ARCHITECTURE



- encoder-only architecture
- Bidirectional Approach
- Pre-training and Fine-tuning
- BERT's architecture consists of a stack of Transformer's Encoder layers.
- Key parameters:
- L: Number of layers (e.g., 12 for BERT Base, 24 for BERT Large).
- H: Hidden size (size of q, k, and v vectors)
- .A: Number of attention heads.
- BERT Base: L=12, H=768, A=12 (Total Parameters=110M).
- BERT Large: L=24, H=1024, A=16 (Total Parameters=340M)[1]

code:

```
from newspaper import Article
import torch
from models.model_builder import ExtSummarizer
from ext_sum import summarize
import textwrap
import nltk
nltk.download('punkt')


# Crawl URL with `newspaper3k`
url = "https://www.cnn.com/2020/05/29/tech/facebook-violence-trump/index.html" #@param {
article = Article(url)
article.download()
article.parse()
print(wrapper.fill(article.text))

# Save input text into `raw_data/input.txt`
with open('raw_data/input.txt', 'w') as f:
    f.write(article.text)
```

(CNN) Over and over again in 2018, during an apology tour that took him from the halls of the US Congress to an appearance before the European Parliament, Mark Zuckerberg said Facebook had failed to "take a broad enough view of our responsibilities." But two years later, Zuckerberg and Facebook are still struggling with their responsibilities and how to handle one of their most famous users: President Donald Trump. Despite Zuckerberg having previously indicated any post that "incites violence" would be a line in the sand — even if it came from a politician — Facebook remained silent for hours Friday after Trump was accused of glorifying violence in posts that appeared on its platforms. At 12:53am ET on Friday morning, as cable news networks carried images of fires and destructive protests in Minneapolis, the President tweeted : "These THUGS are dishonoring the memory of George Floyd, and I won't let that happen. Just spoke to Governor Tim Walz and told him that the Military is with him all the way. Any difficulty and we will assume control but, when the looting starts, the shooting starts. Thank you!" His phrase "when the looting starts, the shooting starts," mirrors language used by a Miami police chief in the late 1960s in the wake of riots. Its use was immediately condemned by a wide array of individuals, from historians to members of rival political campaigns. Former Vice President and presumptive Democratic nominee Joe Biden said Trump was "calling for violence against American citizens during a moment of pain for so many." Read More

```
# Load model
model_type = 'mobilebert' #@param ['bertbase', 'distilbert', 'mobilebert']
checkpoint = torch.load(f'checkpoints/{model_type}_ext.pt', map_location='cpu')
model = ExtSummarizer(checkpoint=checkpoint, bert_type=model_type, device="cpu")


%%time
# Run summarization
input_fp = 'raw_data/input.txt'
result_fp = 'results/summary.txt'
summary = summarize(input_fp, result_fp, model, max_length=3)

    CPU times: user 473 ms, sys: 4.96 ms, total: 478 ms
    Wall time: 609 ms


# Print summary
wrapper = textwrap.TextWrapper(width=80)
print(wrapper.fill(summary))

    (CNN) Over and over again in 2018, during an apology tour that took him from the
    halls of the US Congress to an appearance before the European Parliament, Mark
    Zuckerberg said Facebook had failed to "take a broad enough view of our
    responsibilities." But two years later, Zuckerberg and Facebook are still
    struggling with their responsibilities and how to handle one of their most
    famous users: President Donald Trump. Despite Zuckerberg having previously
    indicated any post that "incites violence" would be a line in the sand — even if
    it came from a politician — Facebook remained silent for hours Friday after
    Trump was accused of glorifying violence in posts that appeared on its
    platforms.
```

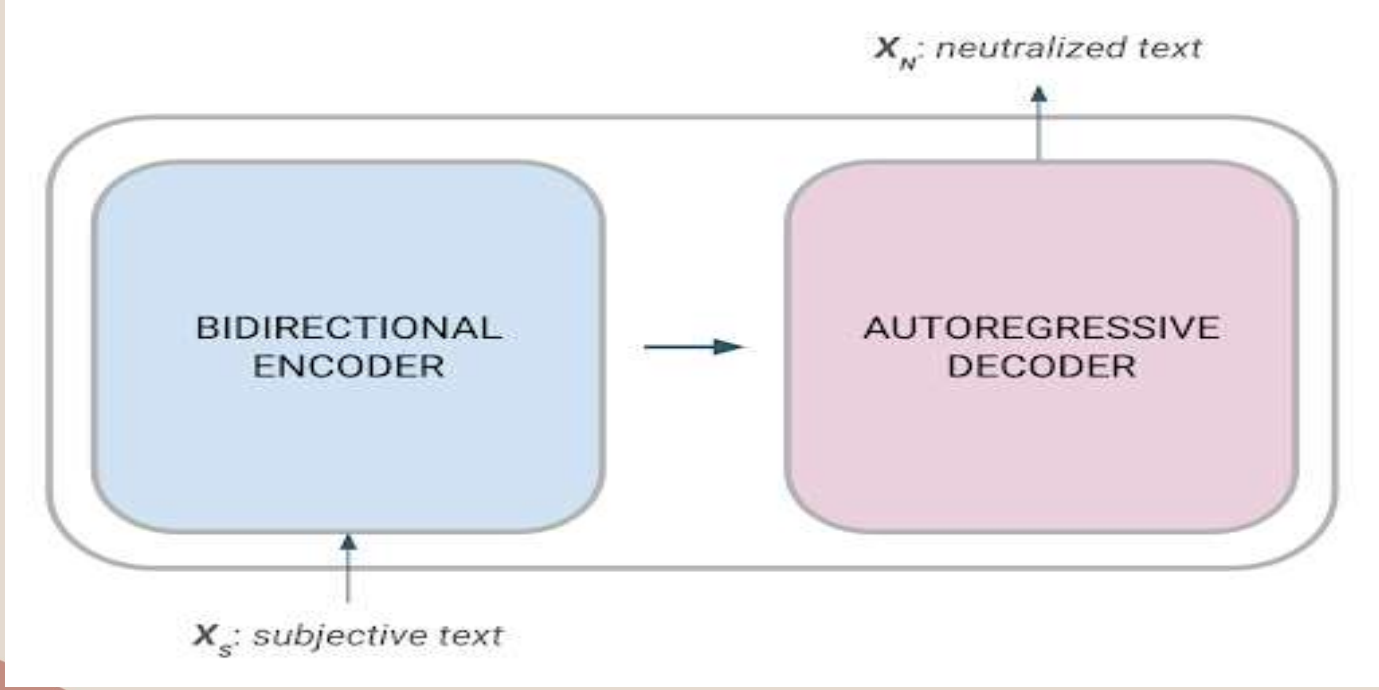
Bart model for Abstractive summarization

- The Bart model was proposed in BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer on 29 Oct, 2019.
- BART is a denoising autoencoder for pretraining sequence-to-sequence models. It is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Transformer-based neural machine translation architecture. It uses a standard seq2seq/NMT architecture with a bidirectional encoder (like BERT) and a left-to-right decoder (like GPT). This means the encoder's attention mask is fully visible, like BERT, and the decoder's attention mask is causal, like GPT2.
- The pretraining task involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where spans of text are replaced with a single mask token. BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks. It matches the performance of RoBERTa with comparable training resources on GLUE and SQuAD, achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks, with gains of up to 6 ROUGE

Applications of BART Summarization:

- BART can create concise summaries that may introduce new phrases not present in the original text.
- Domains where it's useful include:
    1. Science
    2. Literature
    3. Finance
    4. Legal analysis
    5. Meetings
    6. Video conferencing
    7. Programming languages

BART ARCHITECTURE DIAGRAM



METHODOLOGY
1.DATA PROCESSING & EDA

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import torch
import warnings
warnings.filterwarnings("ignore")


from sklearn.model_selection import train_test_split
from torch.utils.data import Dataset, DataLoader
from lightning.pytorch import Trainer
from lightning.pytorch.callbacks import ModelCheckpoint
from lightning.pytorch.loggers import TensorBoardLogger


import transformers
from transformers import T5Tokenizer, T5ForConditionalGeneration, AutoTokenizer, AutoModel
from transformers import PegasusForConditionalGeneration, PegasusTokenizer, BartForConditionalGeneration
from transformers import DataCollatorForSeq2Seq, AdamW, get_linear_schedule_with_warmup
from transformers import AutoModelForSeq2SeqLM, Seq2SeqTrainingArguments, Seq2SeqTrainer
from transformers import create_optimizer, AdamWeightDecay
from transformers import pipeline
import datasets
from datasets import Dataset, DatasetDict
#from torch.utils.data import DataLoader, Dataset
from tqdm.auto import tqdm
from rouge import Rouge


from torch import cuda
device = 'cuda' if cuda.is_available() else 'cpu'
device
```

```
'cuda'
```

```
df = pd.read_csv("news_summary.csv", encoding="latin-1")
df.head()
```

	author	date	headlines	read_more	text	ctext
0	Chhavi Tyagi	03 Aug 2017,Thursday	Daman & Diu revokes mandatory Rakshabandhan in...	http://www.hindustantimes.com/india-news/raksh...	The Administration of Union Territory Daman an...	The Daman and Diu administration on Wednesday ...
1	Daisy Mowke	03 Aug 2017,Thursday	Malaika slams user who trolled her for 'divorc...	http://www.hindustantimes.com/bollywood/malaik...	Malaika Arora slammed an Instagram user who tr...	From her special numbers to TV?appearances, Bo...
2	Arshiya Chopra	03 Aug 2017,Thursday	'Virgin' now corrected to 'Unmarried' in IGIMS...	http://www.hindustantimes.com/patna/bihar-igim...	The Indira Gandhi Institute of Medical Science...	The Indira Gandhi Institute of Medical Science...
3	Sumedha Sehra	03 Aug 2017,Thursday	Aaj aapne pakad liya: LeT man Dujana before be...	http://indiatoday.intoday.in/story/abu-dujana-...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...
4	Aarushi Maheshwari	03 Aug 2017,Thursday	Hotel staff to get training to spot signs of s...	http://indiatoday.intoday.in/story/sex-traffic...	Hotels in Maharashtra will train their staff t...	Hotels in Mumbai and other Indian cities are t...

```
df.describe()
```

	author	date	headlines	read_more	text	ctext
count	4514	4514	4514	4514	4514	4396
unique	45	240	4514	4461	4514	4341
top	Chhavi Tyagi	19 Jul 2017,Wednesday	More than half of India's languages may die in...	http://indiatoday.intoday.in/story/assembly-el...	At least 400 languages or more than half langu...	AAJ TAK LIVE TV WITH LIVE ELECTION RESULTS I c...
freq	559	76	1	13	1	13

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4514 entries, 0 to 4513
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   author     4514 non-null   object
 1   date       4514 non-null   object
 2   headlines  4514 non-null   object
 3   read_more  4514 non-null   object
 4   text       4514 non-null   object
 5   ctext      4396 non-null   object
dtypes: object(6)
memory usage: 211.7+ KB
```

```
df = df[['headlines', 'text', 'ctext']]
df.head()
```

	headlines	text	ctext
0	Daman & Diu revokes mandatory Rakshabandhan in...	The Administration of Union Territory Daman an...	The Daman and Diu administration on Wednesday ...
1	Malaika slams user who trolled her for 'divorc...	Malaika Arora slammed an Instagram user who tr...	From her special numbers to TV?appearances, Bo...
2	'Virgin' now corrected to 'Unmarried' in IGIMS...	The Indira Gandhi Institute of Medical Science...	The Indira Gandhi Institute of Medical Science...
3	Aaj aapne pakad liya: LeT man Dujana before be...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...
4	Hotel staff to get training to spot signs of s...	Hotels in Maharashtra will train their staff t...	Hotels in Mumbai and other Indian cities are t...

```
# drop na
df = df.dropna()
df.describe()
```

	headlines	text	ctext
count	4396	4396	4396
unique	4396	4396	4341
top	More than half of India's languages may die in...	At least 400 languages or more than half langu...	AAJ TAK LIVE TV WITH LIVE ELECTION RESULTS I c...
freq	1	1	13

```
# Converting to lowercase
df['ctext'] = df['ctext'].apply(str.lower)
df['text'] = df['text'].apply(str.lower)
df['headlines'] = df['headlines'].apply(str.lower)
df.head()
```

	headlines	text	ctext
0	daman & diu revokes mandatory rakshabandhan in...	the administration of union territory daman an...	the daman and diu administration on wednesday ...
1	malaika slams user who trolled her for 'divorc...	malaika arora slammed an instagram user who tr...	from her special numbers to tv?appearances, bo...
2	'virgin' now corrected to 'unmarried' in igims...	the indira gandhi institute of medical science...	the indira gandhi institute of medical science...
3	aaj aapne pakad liya: let man dujana before be...	lashkar-e-taiba's kashmir commander abu dujana...	lashkar-e-taiba's kashmir commander abu dujana...
4	hotel staff to get training to spot signs of s...	hotels in maharashtra will train their staff t...	hotels in mumbai and other indian cities are t...

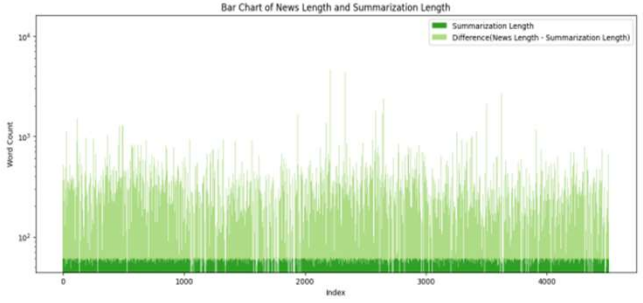
```
df['headlines_length'] = [len(x.split()) for x in df.headlines]
df['text_length'] = [len(x.split()) for x in df.text]
df['ctext_length'] = [len(x.split()) for x in df.ctext]
df.head()
```

	headlines	text	ctext	headlines_length	text_length	ctext_length
0	daman & diu revokes mandatory rakshabandhan in...	the administration of union territory daman an...	the daman and diu administration on wednesday ...	9	60	364
1	malaika slams user who trolled her for 'divorc...	malaika arora slammed an instagram user who tr...	from her special numbers to tv?appearances, bo...	10	60	396
2	'virgin' now corrected to 'unmarried' in igims...	the indira gandhi institute of medical science...	the indira gandhi institute of medical science...	8	60	335
3	aaj aapne pakad liya: let man dujana before be...	lashkar-e-taiba's kashmir commander abu dujana...	lashkar-e-taiba's kashmir commander abu dujana...	10	60	404
4	hotel staff to get training to spot signs of s...	hotels in maharashtra will train their staff t...	hotels in mumbai and other indian cities are t...	11	60	526

```
df = df[df['ctext_length'] >= df['text_length']]
df.describe()
```

	headlines_length	text_length	ctext_length
count	4274.000000	4274.000000	4274.000000
mean	9.300889	58.299719	351.740056
std	1.407168	2.314246	358.884472
min	4.000000	44.000000	50.000000
25%	8.000000	57.000000	193.000000
50%	9.000000	59.000000	288.000000
75%	10.000000	60.000000	416.000000
max	14.000000	62.000000	12202.000000

```
df['diff'] = df['ctext_length'] - df['text_length']
#sns.set(style="whitegrid")
color1 = sns.color_palette("Paired")[3]
color2 = sns.color_palette("Paired")[2]

plt.figure(figsize=(15, 6))
bar1 = plt.bar(df.index, df['text_length'], color=color1, label='Summarization Length', width=1.0)
bar2 = plt.bar(df.index, df['diff'], bottom=df['text_length'], color=color2, label='Difference(News Length - Summarization Length)', width=1.0)

plt.xlabel('Index')
plt.ylabel('Word Count')
plt.title('Bar Chart of News Length and Summarization Length')
plt.legend()
plt.yscale('log')
plt.show()
```




```
df = df.drop(columns=['headlines_length', 'text_length', 'ctext_length', 'diff'])
df.head()
```

DATA PREPROCESSING &EDA

	headlines	text	ctext
0	daman & diu revokes mandatory rakshabandhan in...	the administration of union territory daman an...	the daman and diu administration on wednesday ...
1	malaika slams user who trolled her for 'divorc...	malaika arora slammed an instagram user who tr...	from her special numbers to tv?appearances, bo...
2	'virgin' now corrected to 'unmarried' in igims...	the indira gandhi institute of medical science...	the indira gandhi institute of medical science...
3	aaj aapne pakad liya: let man dujana before be...	lashkar-e-taiba's kashmir commander abu dujana...	lashkar-e-taiba's kashmir commander abu dujana...
4	hotel staff to get training to spot signs of s...	hotels in maharashtra will train their staff t...	hotels in mumbai and other indian cities are t...

```
df['news'] = df['headlines'] + '. ' + df['ctext']
df.rename(columns={"text": "summary"}, inplace = True)
df = df.drop(columns=['headlines', 'ctext'])
df.head()
```

	summary	news
0	the administration of union territory daman an...	daman & diu revokes mandatory rakshabandhan in...
1	malaika arora slammed an instagram user who tr...	malaika slams user who trolled her for 'divorc...
2	the indira gandhi institute of medical science...	'virgin' now corrected to 'unmarried' in igims...
3	lashkar-e-taiba's kashmir commander abu dujana...	aaj aapne pakad liya: let man dujana before be...
4	hotels in maharashtra will train their staff t...	hotel staff to get training to spot signs of s...

```
# Making the dataset
prefix = 'summarize: '
df['news'] = prefix + df['news']
df.head()
```

	summary	news
0	the administration of union territory daman an...	summarize: daman & diu revokes mandatory raksh...
1	malaika arora slammed an instagram user who tr...	summarize: malaika slams user who trolled her ...
2	the indira gandhi institute of medical science...	summarize: 'virgin' now corrected to 'unmarrie...
3	lashkar-e-taiba's kashmir commander abu dujana...	summarize: aaj aapne pakad liya: let man dujan...
4	hotels in maharashtra will train their staff t...	summarize: hotel staff to get training to spot...

```
# Converting the pandas dataset to huggingface dataset
# first split the train and test set
train_df, test_df = train_test_split(df, test_size=0.01, shuffle=True)
print("train and val shape:", train_df.shape, "test shape:",test_df.shape)
# save for every model inference
global_train_df = train_df
global_test_df = test_df
train_df = datasets.Dataset.from_pandas(train_df)
train_df = train_df.remove_columns(["__index_level_0__"])
# split train into train and val
train_df = train_df.train_test_split(test_size=0.2, shuffle=True) # split train and test
train_df["train"][0]
```

```
train and val shape: (4231, 2) test shape: (43, 2)
{'summary': "lenovo phab 2 pro, the first phone with google's augmented reality camera, has been launched in india at ₹29,999. through its depth-sensing google tango camera, the phone can
map out physical spaces, track its own position in a room, and follow objects. powered by a snapdragon 652 processor, the 6.4-inch phone has 4 gb ram and 4,050 mah battery.",
 'news': "summarize: 1st google tango camera phone launches in india at ₹29,999. how crazy are you about the pokã©mon go? if not much, what about lenovo phones? nothing still? what about
a lenovo phone priced cheaper in india than in the us?lenovo's first smartphone touted to have google tango -- lenovo phab2 pro -- has been launched in india and available for rs 29,999
exclusively on flipkart, discounted from its us price of $499 (roughly rs 33,500) where it was first announced in november 2016. the phone was expected to sell at rs 40,000 given the
usual price difference between two markets.also read:  xiaomi mi max v lenovo phab 2 plus: top specs shootout lenovo phab 2 pro is the world's first smartphone to feature google's tango
technology -- a set of software and specially calibrated group of embedded sensors developed by google to bring augmented reality (ar) experiences on relatively affordable phablet form
factor by complimenting them with depth perception and 3d mapping to understand shapes and own location in the immediate surroundings. asus also launched the zenfone ar at ces that has
both google daydream vr and google tango.\xa0 augmented reality means your smartphone that will be aware of it surroundings, and create computer generated images to superimpose them on a
user's view of the real world to provide a merged view that they can interact with. the pokã©mon go and house of the dying sun are some of the top ar games in the world.phab2 pro is a 4g
smartphone powered by the qualcomm snapdragon 652 processor and comes with 4gb ram and 64gb storage expandable up to 128gb. the phone has a 6.4 inch quad hd ips display, 16-megapixel rear
camera with pdaf and 8-megapixel front camera with f/2.2 aperture and 1.4 micro meter pixel size. it features dolby audio capture 5.1 to record surround sound using three microphones
deployed on the phone while there is dolby atmos playback for surround sound output."}
```

```
# Fitting into dataset dict
train_val_test_dataset = DatasetDict({
    'train': train_df['train'],
    'val': train_df['test']})

print(type(train_val_test_dataset))
train_val_test_dataset
```

```
<class 'datasets.dataset_dict.DatasetDict'>
DatasetDict({
    train: Dataset({
        features: ['summary', 'news'],
        num_rows: 3384
    })
    val: Dataset({
        features: ['summary', 'news'],
        num_rows: 847
    })
})
```

```
# Padding
data_collator = DataCollatorForSeq2Seq(tokenizer=tokenizer, model=model_name)

def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    decoded_preds = tokenizer.batch_decode(predictions, skip_special_tokens=True)
    labels = np.where(labels != -100, labels, tokenizer.pad_token_id)
    decoded_labels = tokenizer.batch_decode(labels, skip_special_tokens=True)
    result = Rouge().get_scores(decoded_preds, decoded_labels, avg=True, ignore_empty=True)

    # prediction_lens = [np.count_nonzero(pred != tokenizer.pad_token_id) for pred in predictions]
    # result["gen_len"] = np.mean(prediction_lens)

    return result


# tokenize the data
model_name = "facebook/bart-large-cnn"
tokenizer = AutoTokenizer.from_pretrained(model_name, use_fast = False)
tokenized_data = train_val_test_dataset.map(prepare_dataset, batched=True)
```

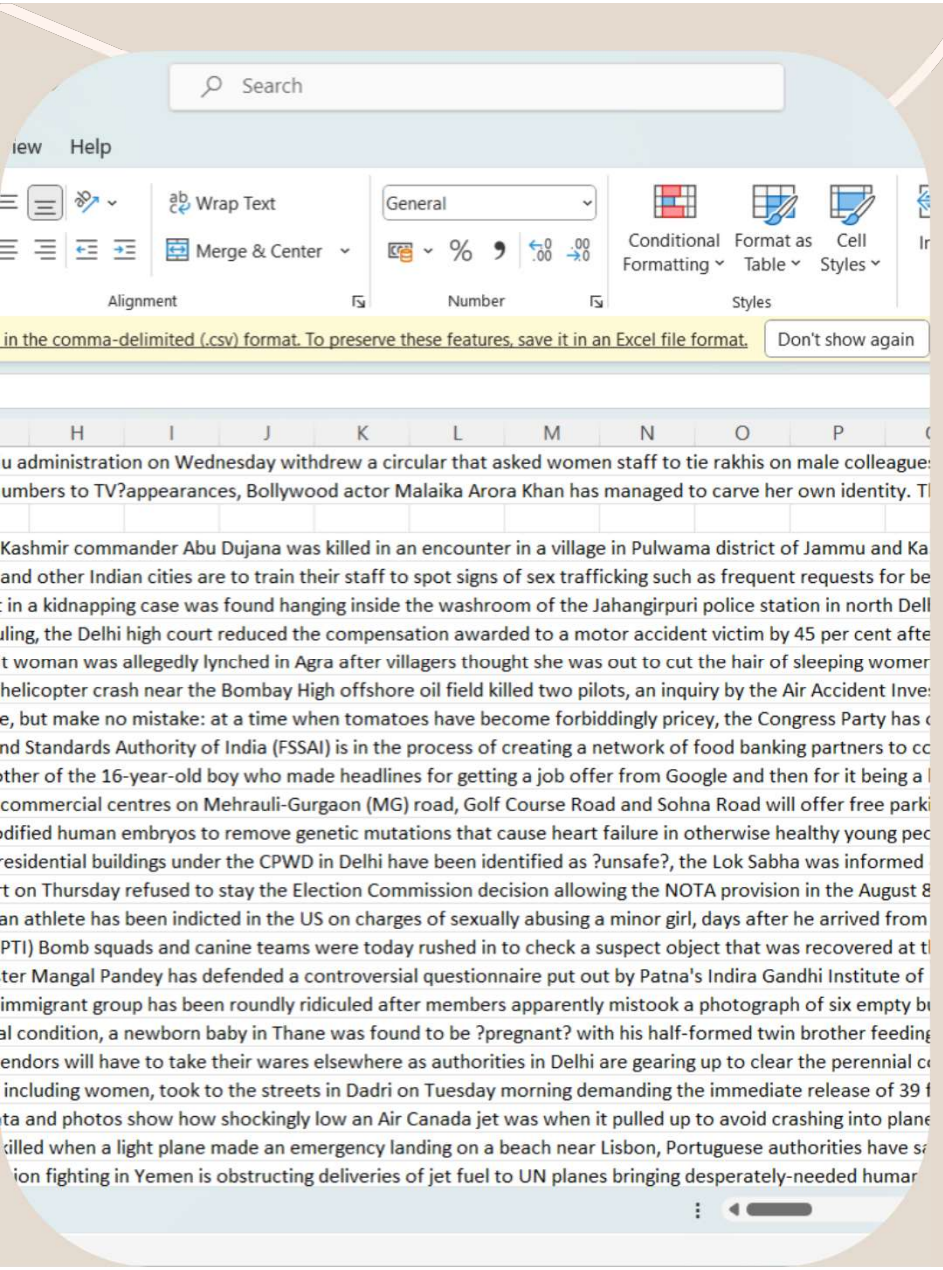
2.BART INFERENCE

```
Map:   0%|          | 0/3384 [00:00<?, ? examples/s]
Map:   0%|          | 0/847 [00:00<?, ? examples/s]

# Padding
data_collator = DataCollatorForSeq2Seq(tokenizer=tokenizer, model=model_name)

# model
BARTmodel = AutoModelForSeq2SeqLM.from_pretrained(model_name).to(device)

# set up hyper-parameters
training_args = Seq2SeqTrainingArguments(
    output_dir="bart-news",
    evaluation_strategy="epoch",
    learning_rate=1e-6,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    weight_decay=0.01,
    save_total_limit=2,
    num_train_epochs=3,
    predict_with_generate=True,
    fp16=True,
    report_to="none"
)

# setup trainer
trainer = Seq2SeqTrainer(
    model = BARTmodel,
    args = training_args,
    train_dataset = tokenized_data["train"],
    eval_dataset = tokenized_data["val"],
    tokenizer = tokenizer,
    data_collator = data_collator,
    compute_metrics = compute_metrics
)

trainer.train()
```

[1269/1269 35:10, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Rouge-1	Rouge-2	Rouge-l
1	No log	1.381600	{'r': 0.493402097017617, 'p': 0.4594609602926955, 'f': 0.4715771714382409}	{'r': 0.2649264284253257, 'p': 0.2433900077220443 3, 'f': 0.2509761138222017}	{'r': 0.4487789578814032 6, 'p': 0.4180886314302871, 'f': 0.4290059855052 78}
2	1.508000	1.344802	{'r': 0.5042939780973792, 'p': 0.4575656949749346 6, 'f': 0.4758068268439658}	{'r': 0.2764344043923761, 'p': 0.2446058869662 99, 'f': 0.2569999817531 5217}	{'r': 0.4598105394584861 6, 'p': 0.4173289838475831, 'f': 0.4338942773258532}
3	1.368500	1.337948	{'r': 0.5111880166210206, 'p': 0.4602218674394551, 'f': 0.4807087798356639 6}	{'r': 0.2813799244356644, 'p': 0.2474484319842643 3, 'f': 0.2609590540760679 5}	{'r': 0.4658045451659364 5, 'p': 0.4196197934453683, 'f': 0.4381741038720 02}

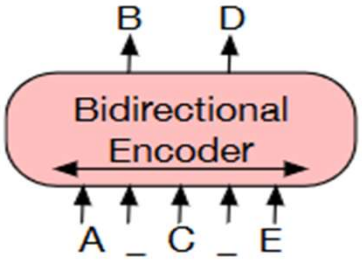
```
TrainOutput(global_step=1269, training_loss=1.4170112820287875, metrics={'train_runtime': 2110.6515, 'train_samples_per_second': 4.81, 'train_steps_per_second': 0.601, 'total_flos':
1.0970069731442688e+16, 'train_loss': 1.4170112820287875, 'epoch': 3.0})

# save the model
model_path = "bart-news"
trainer.save_model(model_path)
tokenizer.save_pretrained(model_path)

('bart-news/tokenizer_config.json',
 'bart-news/special_tokens_map.json',
 'bart-news/vocab.json',
 'bart-news/merges.txt',
 'bart-news/added_tokens.json')

model = BartForConditionalGeneration.from_pretrained("bart-news")
tokenizer = AutoTokenizer.from_pretrained("bart-news")

#for i in range(len(test_df['news'])):
for i in range(5):
    print("original_news: ",test_df['news'].iloc[i])
    summarizer = pipeline("summarization", model = model,tokenizer = tokenizer, max_length = 100)
    summary = summarizer(test_df['news'].iloc[i])
    print(summary[0])
    print()

original_news:  summarize: taapsee pannu opts out of event organised by fairness cream. taapsee pannu, who was catapulted to the big league after her film pink became a runaway hit, has

{'summary_text': 'actress taapsee pannu has joined the likes of kangana ranaut and ranbir kapoor by taking a stand against fairness creams. the actor was supposed to be a part of an even

original_news:  summarize: sachin attends rajya sabha after questions on attendance. former cricketer sachin tendulkar was spotted in the rajya sabha today a couple of days after absence

{'summary_text': 'samajwadi party mp naresh agarwal had on tuesday questioned the absence of celebrity members of the house and advocated them being disqualified from rajya sabha members

original_news:  summarize: kohli will catch up with dhoni as captain, says ravi shastri. virat kohli?s ascendancy since becoming the india test captain in 2014 has been startling. his ag

{'summary_text': 'cricket coach ravi shastri has said that virat kohli will soon catch up with mahendra singh dhoni\'s achievements as captain of india\'s test cricket team. dhoni remain

original_news:  summarize: punjab cm announces ?5 lakh reward for harmanpreet kaur. punjab chief minister capt amarinder singh on sunday announced a cash reward of rs five lakh for crick

{'summary_text': "punjab cm captain amarinder singh on sunday announced a cash reward of ?5 lakh for cricketer harmanpreet kaur for her performance that steered india to the finals of ic

original_news:  summarize: upset with govt, deaflympics team refuses to leave airport. the indian contingent of hearing impaired athletes, returning to the country after its best ever pe

{'summary_text': 'the indian contingent of hearing impaired athletes, returning to the country after clinching five medals at the deaflympics, refused to leave the indira gandhi internat
```



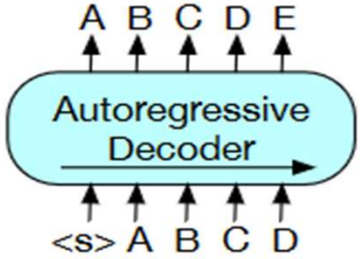
DATASET USED

- NEWS SUMMARIZATION DATASET ON KAGGLE:
- INCLUDES NEWS ARTICLES PAIRED WITH SUMMARIES.
- USEFUL FOR EXPERIMENTING WITH SUMMARIZATION TECHNIQUES

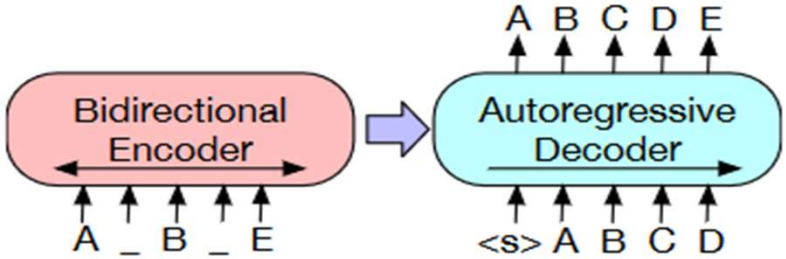
BERT & BART



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

CONCLUSION

- Abstractive summarization using the BART model offers a promising approach to distilling the essence of news articles. By leveraging advanced natural language generation capabilities, BART generates concise and coherent summaries that capture the key points of the original content. With its ability to paraphrase and rephrase information, BART can produce summaries that are not only informative but also fluent and human-like. This empowers readers to quickly grasp the main ideas of news articles, enhancing comprehension and accessibility.
- On the other hand, extractive summarization utilizing the BERT model provides a robust and efficient method for summarizing news articles. By identifying and selecting important sentences or passages from the original text, BERT extracts salient information to create summaries that closely reflect the content of the source material. Leveraging contextual embeddings and attention mechanisms, BERT ensures that the extracted summaries maintain the coherence and relevance of the original article. This approach offers a reliable and effective means of condensing news articles into concise summaries, facilitating quick and accurate information retrieval.
- In summary, both abstractive summarization with the BART model and extractive summarization with the BERT model offer valuable tools for distilling news articles into digestible formats. While abstractive summarization focuses on generating novel summaries by rewriting the content, extractive summarization emphasizes retaining the original wording while condensing the information. Depending on the specific requirements and preferences, either approach can be employed to produce informative and succinct summaries tailored to the needs of readers and applications.

Results:

Abstractive summarization using the BART model showcases impressive capabilities in generating concise and coherent summaries of news articles. By leveraging advanced natural language generation techniques, BART effectively captures the essence of the original content by paraphrasing and rephrasing key points, producing summaries that are both informative and fluently articulated. On the other hand, extractive summarization employing the BERT model demonstrates robust performance in extracting salient information from news articles to create concise summaries. By identifying and selecting important sentences or passages, BERT ensures that the extracted summaries maintain the coherence and relevance of the original content while retaining the original wording. Leveraging contextual embeddings and attention mechanisms, BERT offers a reliable and efficient means of quickly distilling news articles into succinct summaries, facilitating accurate information retrieval.

thank you

ASHISH-521116
KARISHMA TAJ.M-521159