Indian Institute of Information Technology, Nagpur

Department of Computer Science and Engineering
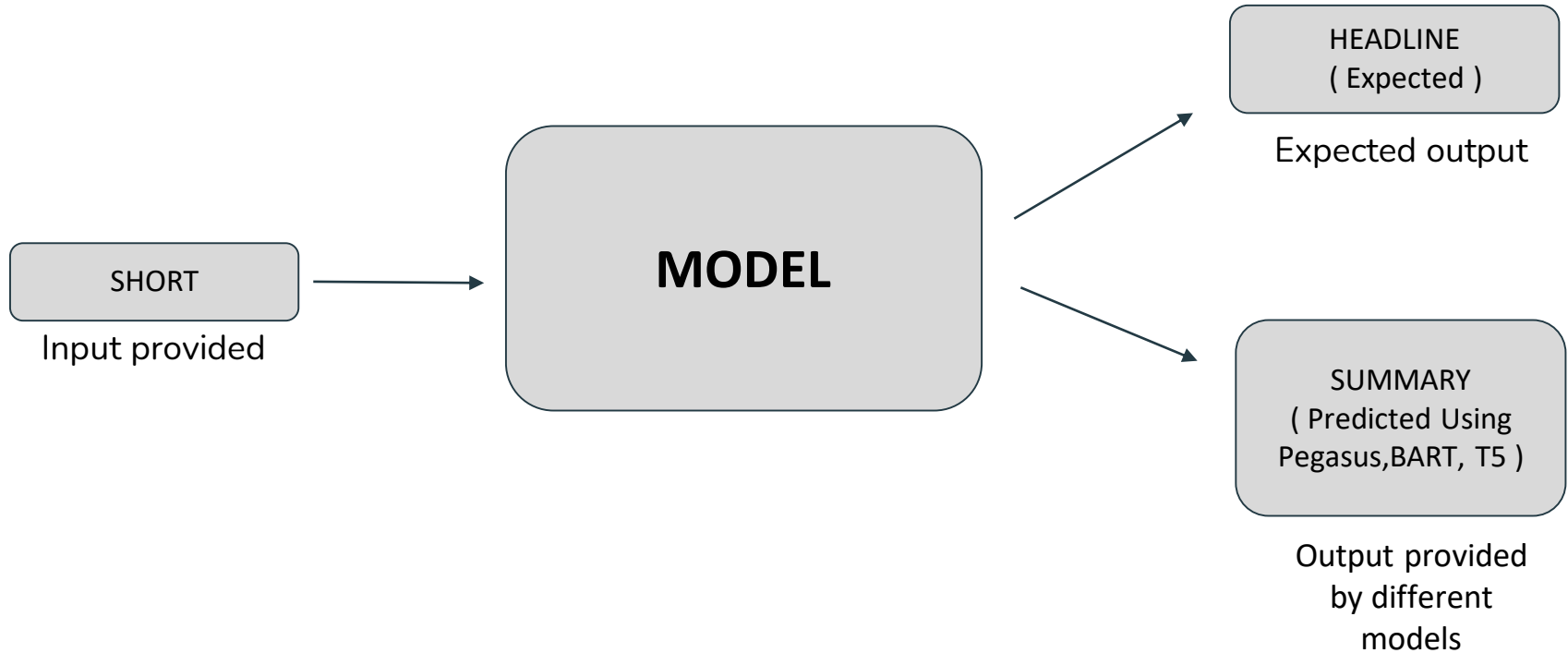
Text Extraction and Summarization using Transformer Model

Dataset Used

[Inshorts News Data | Kaggle](#)

WORKING OF THE MODEL

Pegasus,B5,Bart hugging face models

	Headline	Short	Pegasus	T5	BART
0	Mob attacks Radiohead fans in Turkey	About 20 people reportedly beat up customers a...	[{'summary_text': 'Turkish police are investig...	{'summary_text': 'the mob was reportedly angry...	[{'summary_text': 'The mob was reportedly angr...
1	Man City, Arsenal play out a 2-2 draw	Arsenal and Manchester City played out a 2-2 d...	[{'summary_text': 'Arsenal manager Arsene Weng...	{'summary_text': 'arsenal and manchester city ...	[{'summary_text': 'Manchester City and Arsenal...
2	Tanmay Bhat speaks on his video on Lata, Sachin	Stand-up comedian Tanmay Bhat, who is associat...	[{'summary_text': 'Comedian Tanmay Bhat has be...	{'summary_text': "drought hit farmers wonderi...	[{'summary_text': 'Comedian Tanmay Bhat shared...
3	Pro-Russian rebels cut trade ties with Ukraine...	The pro-Russian rebel group in Ukraine has dec...	[{'summary_text': 'Pro-Russian separatists in ...	{'summary_text': 'pro-Russian rebels cut econo...	[{'summary_text': 'Pro-Russian rebel group in ...
4	World may lose two-thirds of its wildlife by 2...	The world may lose two-thirds (68%) of its wil...	[{'summary_text': 'The world's wildlife popula...	{'summary_text': 'the world may lose two-third...	[{'summary_text': 'The world may lose two-thir...

TEXT - ' Arsenal and Manchester City played out a 2-2 draw at the Etihad Stadium in the Premier League on Sunday. Sergio Aguero and Kevin de Bruyne scored for the home side while Olivier Giroud and Alexis Sanchez scored for the Gunners. Notably, Man City will lose out on a UEFA Champions League spot if Manchester United win their next two games.'

Observed Output - 'Man City, Arsenal play out a 2-2 draw'.

Summary_text using Pegasus: 'Arsenal manager Arsene Wenger and Manchester City counterpart Pep Guardiola had a heated exchange at the end of their game on Sunday.'

Summary_text using T5   : 'arsenal and manchester city play out a 2-2 draw at the Etihad on sunday .man city will lose out on'

Summary_text using Bart   : 'Manchester City and Arsenal played out a 2-2 draw at the Etihad Stadium on Sunday. Sergio Aguero and Kevin de Bruyne'
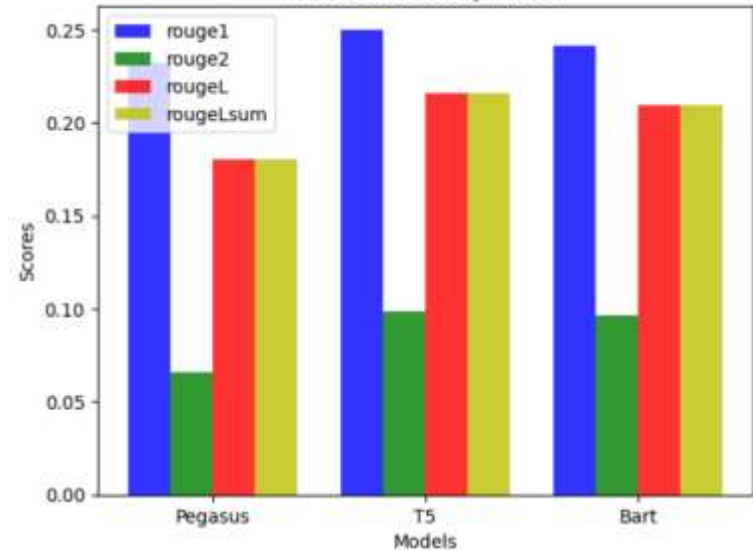
Rouge Values

	rouge1	rouge2	rougeL	rougeLsum
Pegasus	0.232123	0.065643	0.180665	0.180665
T5	0.250286	0.098951	0.216269	0.216269
Bart	0.241525	0.096538	0.209207	0.209207

ROUGE Scores by Model
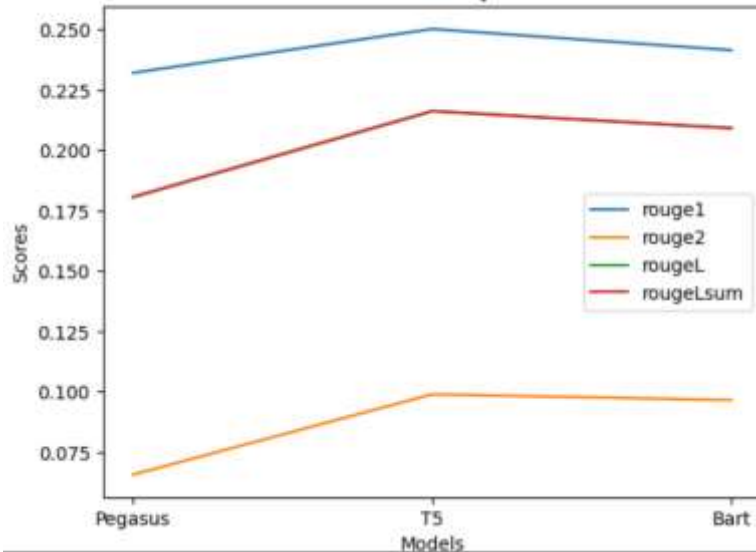
ROUGE Scores by Model

TRANSFORMER MODEL

The original Transformer diagram | A representation of a 4-layer Transformer
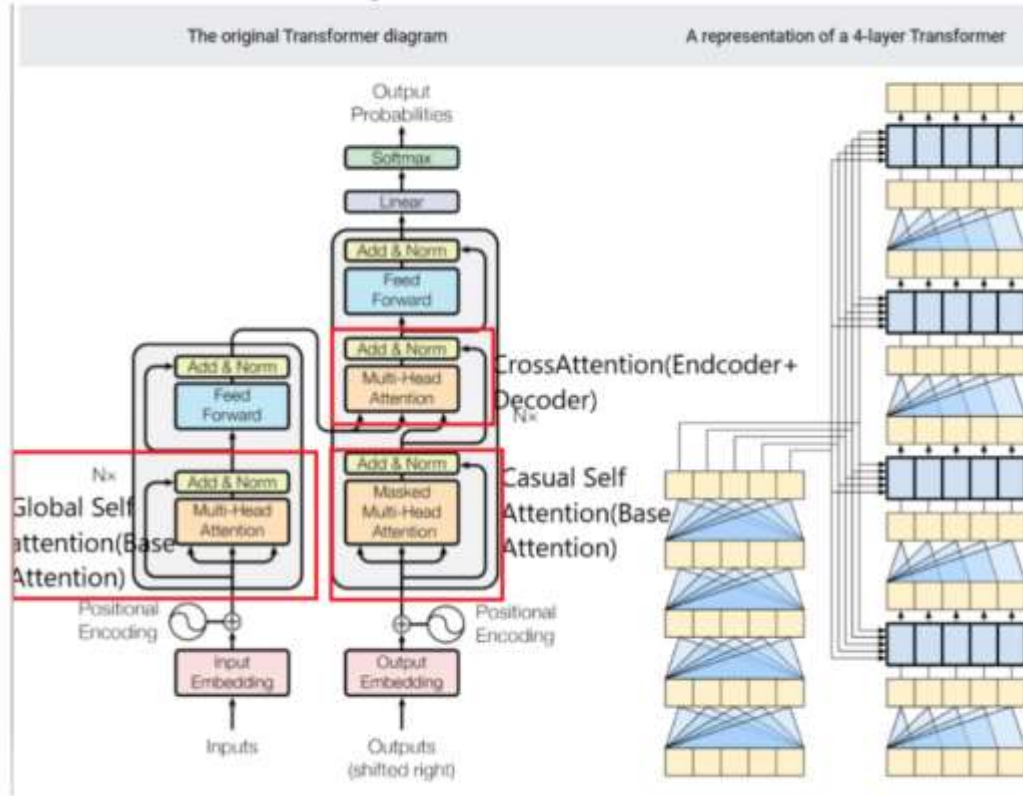
Embedding
Positional encoding
Encoder's self attention
Decoder's self attention
Cross Attention
Feed Forward
Normalization

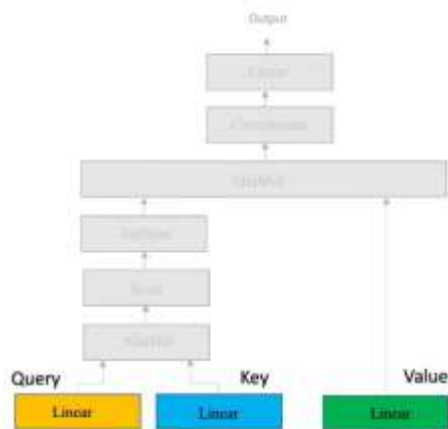**Build & train the Transformer:**
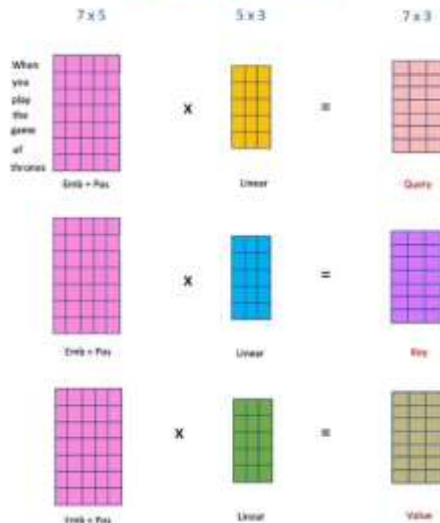Optimization
Loss metric

POSITIONAL ENCODING

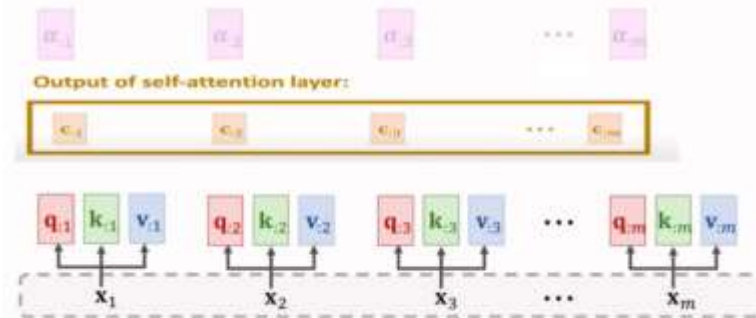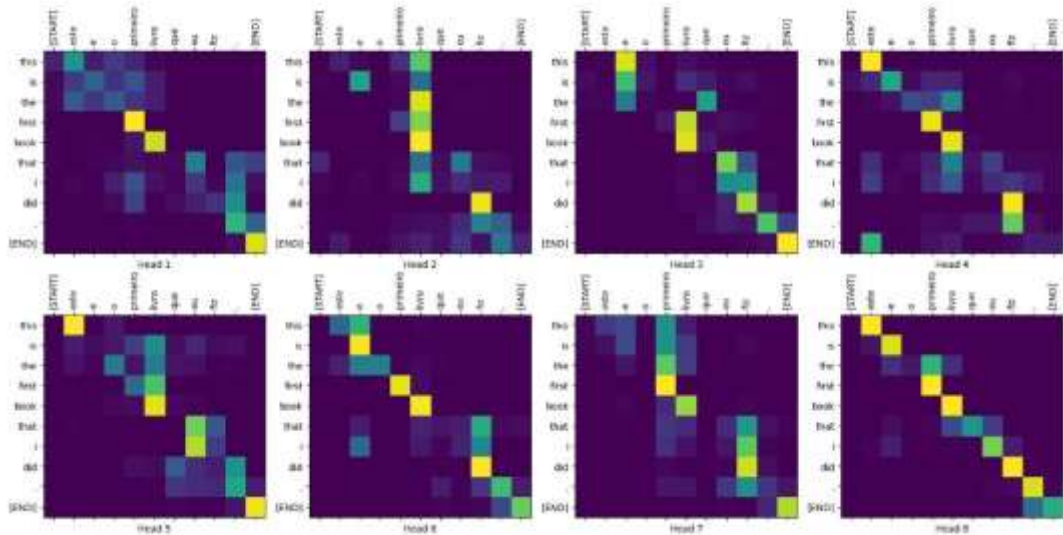| Sequence	Index of token, k	Positional Encoding Matrix with d=4, n=100			
		i=0	i=0	i=1	i=1
I	0	$P_{00}=\sin(0)$ = 0	$P_{01}=\cos(0)$ = 1	$P_{02}=\sin(0)$ = 0	$P_{03}=\cos(0)$ = 1
am	1	$P_{10}=\sin(1/1)$ = 0.84	$P_{11}=\cos(1/1)$ = 0.54	$P_{12}=\sin(1/10)$ = 0.10	$P_{13}=\cos(1/10)$ = 1.0
a	2	$P_{20}=\sin(2/1)$ = 0.91	$P_{21}=\cos(2/1)$ = -0.42	$P_{22}=\sin(2/10)$ = 0.20	$P_{23}=\cos(2/10)$ = 0.98
Robot	3	$P_{30}=\sin(3/1)$ = 0.14	$P_{31}=\cos(3/1)$ = -0.99	$P_{32}=\sin(3/10)$ = 0.30	$P_{33}=\cos(3/10)$ = 0.96

Multi-Head Attention

Weights: $\quad \alpha_{\cdot j} = \text{Softmax}(\mathbf{K}^T \mathbf{q}_{\cdot j}) \in \mathbb{R}^m.$

Context vector: $\quad \mathbf{c}_{\cdot j} = \alpha_{1j}\mathbf{v}_{\cdot 1} + \cdots + \alpha_{mj}\mathbf{v}_{\cdot m} = \mathbf{V}\alpha_{\cdot j}.$

ENCODER-DECODER
BASE SELF-ATTENTION

ATTENTION-WEIGHTS

MULTIHEAD

- Compute weights: $\boldsymbol{\alpha}_{:1} = \text{Softmax}(\mathbf{K}^T \mathbf{q}_{:1}) \in \mathbb{R}^m$.
- Compute context vector: $\mathbf{c}_{:1} = \alpha_{11}\mathbf{v}_{:1} + \cdots + \alpha_{m1}\mathbf{v}_{:m} = \mathbf{V}\boldsymbol{\alpha}_{:1}$.
- Here, $\mathbf{c}_{:j} = \mathbf{V} \cdot \text{Softmax}(\mathbf{K}^T \mathbf{q}_{:j})$.
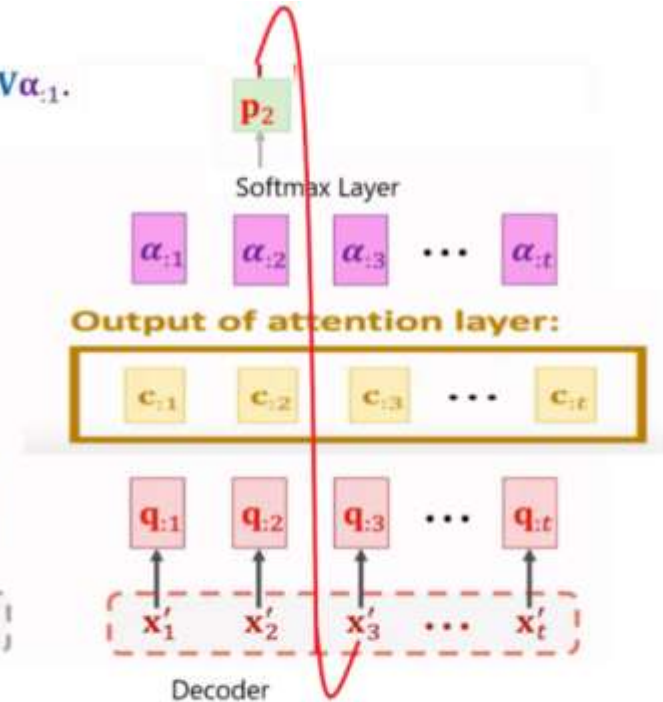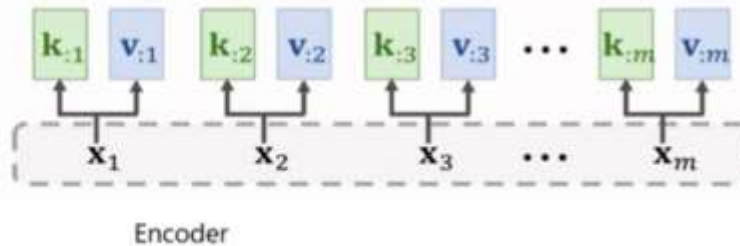
CROSS ATTENTION

TRAINING

```
Epoch 16 Loss 2.7961
Time taken for 1 epoch: 331.7034285068512 secs

Epoch 17 Batch 0 Loss 2.6423
Epoch 17 Batch 429 Loss 2.6650
Epoch 17 Batch 858 Loss 2.6944
Epoch 17 Loss 2.6946
Time taken for 1 epoch: 331.723192691803 secs

Epoch 18 Batch 0 Loss 2.5498
Epoch 18 Batch 429 Loss 2.5781
Epoch 18 Batch 858 Loss 2.6042
Epoch 18 Loss 2.6039
Time taken for 1 epoch: 331.5571415424347 secs

Epoch 19 Batch 0 Loss 2.5481
Epoch 19 Batch 429 Loss 2.4865
Epoch 19 Batch 858 Loss 2.5193
Epoch 19 Loss 2.5193
Time taken for 1 epoch: 332.1124782562256 secs

Epoch 20 Batch 0 Loss 2.2834
Epoch 20 Batch 429 Loss 2.4058
Epoch 20 Batch 858 Loss 2.4402
Saving checkpoint for epoch 20 at checkpoints/ckpt-4
Epoch 20 Loss 2.4405
Time taken for 1 epoch: 333.1804006099701 secs
```

OUTPUT

```
In [ ]: summarize("Chief Justice JS Khehar has said the Supreme Court will go paperless \
        in six to seven months in a bid to save funds and make the judiciary eco-friendly.\
         He further said the apex court will collect all the records electronically from the lower courts \
         and the high courts so that there is no need to file hard copies."
        )
```

```
Out[ ]: 'supreme court to go paperless in 6 months cji'
```

```
In [ ]: summarize("The Central Bureau of Investigation (CBI) has registered a case against\
        Oxfam India and its office-bearers for allegedly violating the provisions of India's\
         foreign funding rules and carried out searches at the offices of the non-governmental organisation, officials said on Wednesday.")
```

```
Out[ ]: 'cbi registers case against cbi over corruption'
```

All in one Website

Summariser

Paste the link of an article you'd like to summarise to get started

URL ⌄

20

https://www.tensorflow.org/text/tutorials/transformer

Summarise

Summariser

Instead of doing a hash lookup the attention layer combines the query and key vectors to determine how well they match the attention score. Given the Portuguese sequence the model would try to generate the English sequence.It's possible to train a model that way. A Transformer is a sequence-to-sequence encoder-decoder model similar to the model in the NMT with attention tutorial. You'd need to write out the inference loop and pass the model's output back to the input. The layer returns the average across all the values weighted by the attention scores.Each location the query-sequence provides a query vector. In an attention layer the query key and value are each vectors. In this tutorial you will:To get the most out of this tutorial it helps if you know about the basics of text generation and attention mechanisms. To implement these attention layers start with a simple base class that just contains the component layers. Neither 'triceratops' nor 'encyclopédia' are in the input dataset and the model attempts to transliterate them even without a shared vocabulary. A decoder then generates the output sentence word by word while consulting the representation generated by the encoder. This is a simple and efficient way to train a text generation model. The only difference is that the RNN layers are replaced with self attention layers. This tutorial demonstrates how to create and train a sequence-to-sequence Transformer model to translate Portuguese into English. Since the model doesn't contain any recurrent or convolutional layers. This layer connects the encoder and decoder. A single-layer Transformer takes a little more code to write but is almost identical to that encoder-decoder RNN model. At each location in the context sequence provides a key and value vector. For example:You have tested the model and the inference is working. The function below converts a dataset of text examples into data of batches for training. To learn more about tokenization visit this guide.This tutorial uses the tokenizers built in the subword tokenizer tutorial.

Summariser

Paste the link of an article you'd like to summarise to get started

Text Box

10

A data scientist is someone who practices Data Science to analyze data so that they can provide actionable insights by using various processes, methods, systems, and algorithms. A Data Scientist is a highly qualified professional with highly developed mathematical, statistical, scientific, analytical and technical skills who can prepare, clean, and validate structured and unstructured data for businesses to use to make better decisions. As a whole, they must fulfill several responsibilities, as follows:

Summarise

Summariser

A data scientist is someone who practices Data Science to analyze data so that they can provide actionable insights by using various processes methods systems and algorithms. As a whole they must fulfill several responsibilities as follows: Identify data sources and gather data. Create data strategies in collaboration with team members and leaders. Analyze structured and unstructured data effectively. Cleaning and verifying the data to ensure it is accurate and uniform. Use algorithms and modules to uncover trends and patterns. Formulate strategies to solve business challenges.