Text Summarization

Mentor: Mr. Narendra Kumar

Group: 4

Introduction

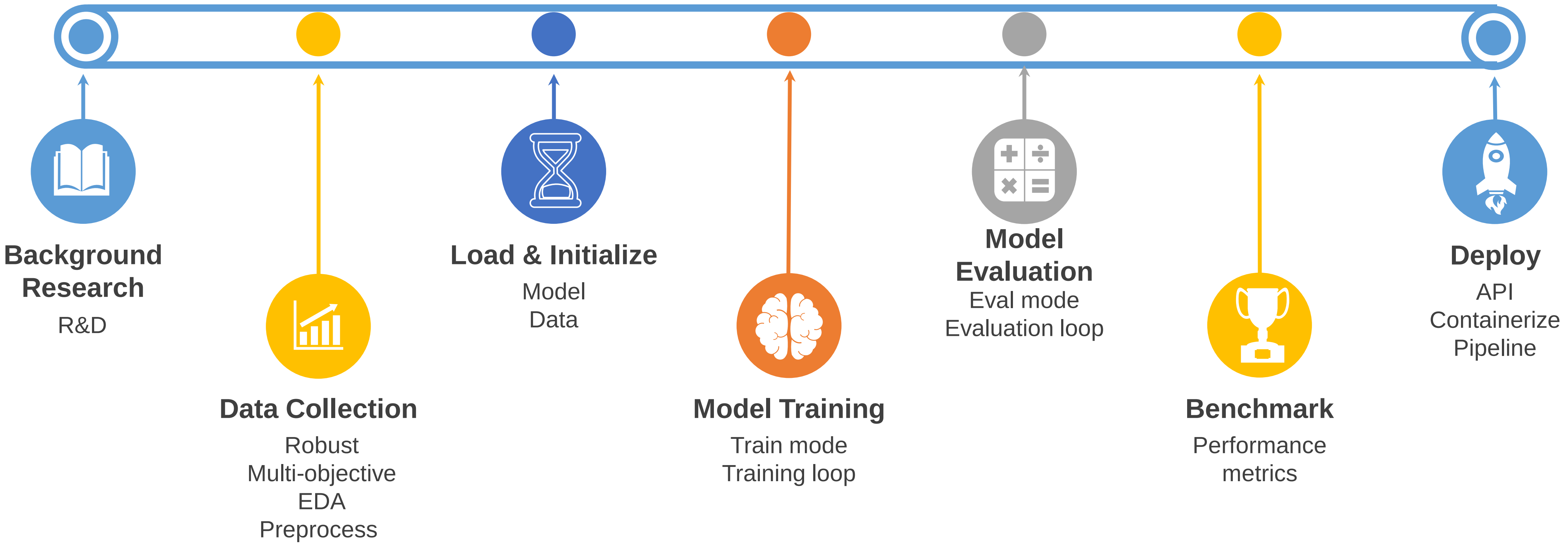
Problem Statement & Planning

Introduction

Problem Statement

- Developing an automated text summarization system that can accurately and efficiently condense large bodies of text into concise summaries is essential for enhancing business operations.
- This project aims to deploy NLP techniques to create a robust text summarization tool capable of handling various types of documents across different domains.
- The system should deliver high-quality summaries that retain the core information and contextual meaning of the original text.

INTENDED PLAN



Background Research

Literature Review & Findings

Background Research

Literature Review

S. No	Use-Case	Paper Title	Year	Method	Dataset	Results	Limitations
1	General text summarization	Text Summarization Using Deep Learning Techniques: A Review	2023	Deep Learning (Seq2Seq, Attention, Transformers)	CNN/Daily Mail, XSum	Improved performance in capturing semantic relationships, better coherence	Computationally expensive, requires large datasets

[1] Saiyyad, M.M.; Patil, N.N. *"Text Summarization Using Deep Learning Techniques: A Review"*. Eng. Proc. 2023, 59, 194.

Background Research
Literature Review

S. No	Use-Case	Paper Title	Year	Method	Dataset	Results	Limitations
2.	Implementation of the Transformer architecture	Attention is all you need	2023	Transformer	WMT 2014 English-German, WMT 2014 English-French	Introduced the Transformer architecture, significantly improving the performance of text summarization tasks.	Requires large datasets and computational resources for training.

[2] A. Vaswani, L. Jones, N. Shazeer, N. Parmar, A. N. Gomez, J. Uszkoreit, Ł. Kaiser, and I. Polosukhin, *"Attention Is All You Need,"* arXiv:1706.03762v7 [cs.CL], Aug. 2, 2023.

Background Research

Literature Review

S. No	Use-Case	Paper Title	Year	Method	Dataset	Results	Limitations
3.	Multi-document summarization	Surveying the Landscape of Text Summarization with Deep Learning	2023	Deep learning methods. Various techniques like RBMs and fuzzy logic employed for summarization.	CNN/Daily Mail	Incorporating transfer learning enhances summary quality and reduces data demand.	Complex models, high computational resources

[3] G. Wang and W. Wu, *"Surveying the Landscape of Text Summarization with Deep Learning: A Comprehensive Review,"* arXiv:2310.09411v1 [cs.CL], Oct. 13, 2023.

Background Research

Literature Review

S. No	U4se-Case	Paper Title	Year	Method	Dataset	Results	Limitations
4.	Abstractive summarizatio n	Pegasus: Pre-training with gap-sentences for abstractive summarization	2020	Transformer (Pegasus)	XSum, CNN/Daily Mail, and Reddit TIFU	Significant improveme nts in abstractive summarizat ion quality	Resource-intensive

[4] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, *"PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,"* arXiv:1912.08777v3 [cs.CL], Jul. 10, 2020.

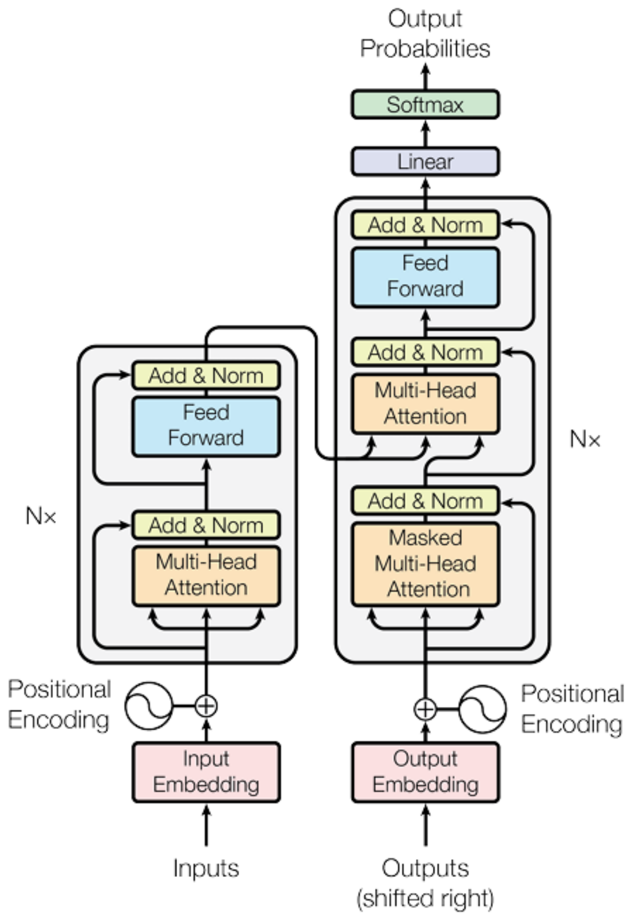
Background Research

Literature Review

S. No	Use-Case	Paper Title	Year	Method	Dataset	Results	Limitations
5.	Extractive summarization	Text Summarization with Pretrained Encoders	2019	Intersentence Transformer layers for summarization	CNN/Daily Mail, NYT, Xsum, DailyMail	BERT-based models outperformed other approaches in abstractive summarization.	High computational resources required

[5] Y. Liu and M. Lapata, *"Text Summarization with Pretrained Encoders,"* arXiv:1908.08345v2 [cs.CL], Sep. 5, 2019.

Research
Architecture



[2] Fig. :Transformer architecture:

- **Implementation methods:**
  - *From Scratch*
    - Build Model
      - NN
    - Initialize normalized W&B
    - Train model with extensive data
    - Hence,
      - Computationally Intensive
      - Sub-Optimal usage of resources
      - Out-of-scope
  - *Using Pre-trained model*
    - Load Model & its parameters
    - Re-Train with specific dataset
    - Evaluate
    - Hence,
      - Innovation can be done at intended tasks
      - Optimal utilization of resources

Proposal

Architecture, Findings & Team Details

Proposal
Workflow

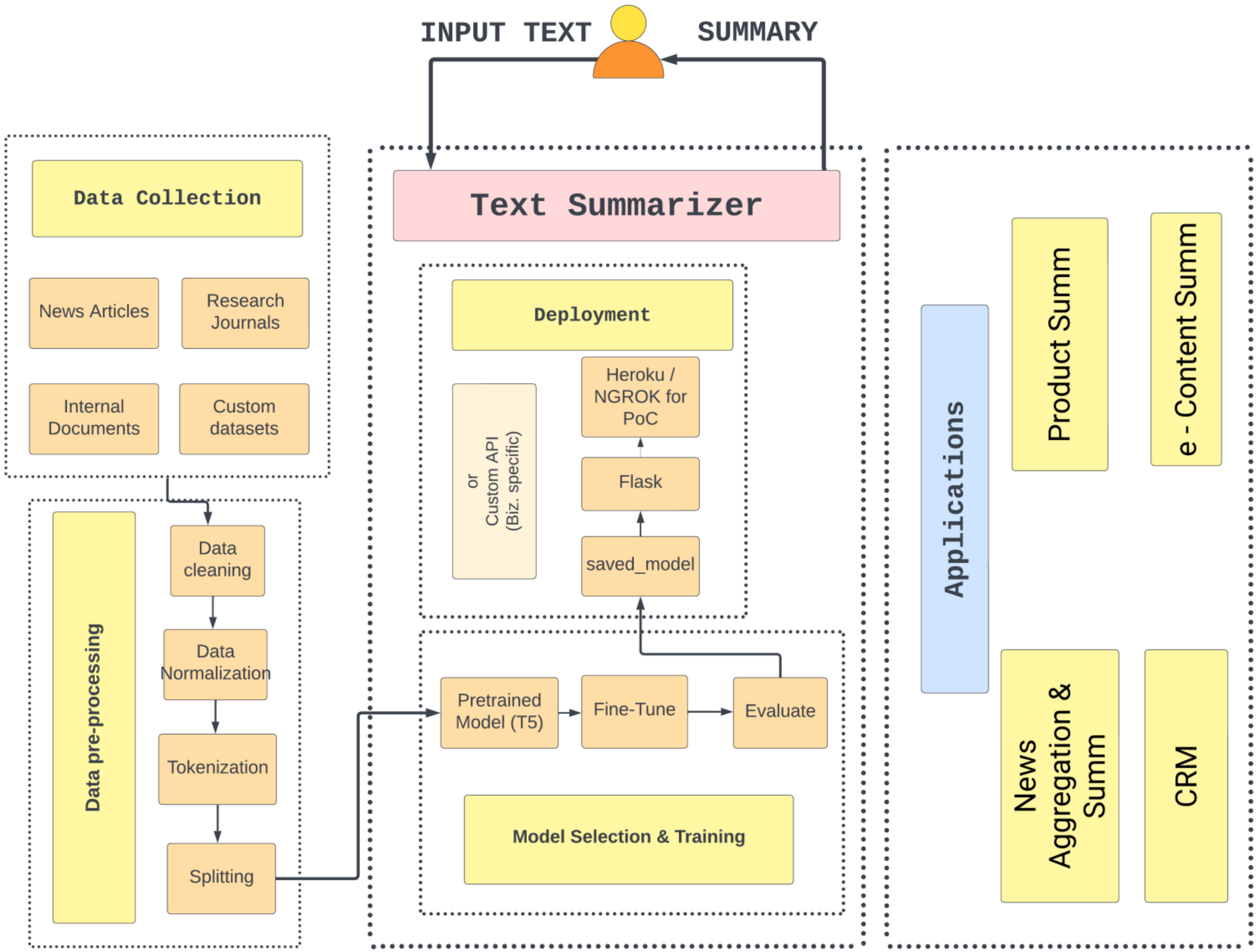


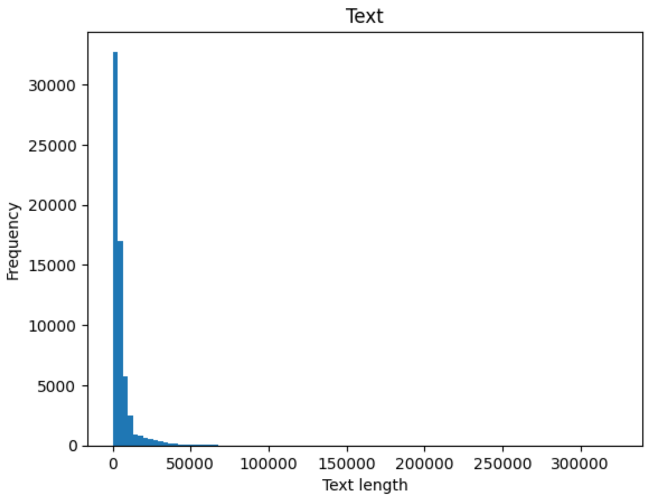
Fig. : Proposed Workflow

Proposal

Dataset

- Merged selective dataset from
  - CNN, Daily Mail : News,
  - BillSum: Legal,
  - ArXiv : Scientific
  - Dialoguesum
- Completed - data preprocessing
  - Removed
    - NULL records, punctation, stop-words
  - Lowercasing, lemmatization.

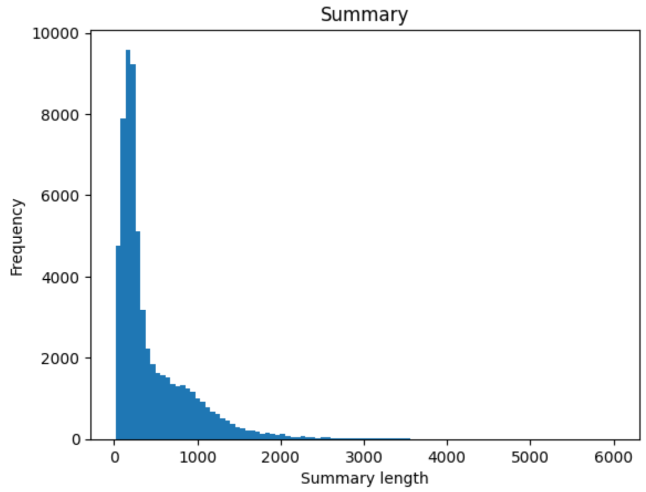
```
count     62707.000000
mean       5211.270975
std        7794.860686
min          83.000000
25%        1275.000000
50%        3176.000000
75%        5684.500000
max      323742.000000
Name: text, dtype: float64
```



	text	summary
0	section 1 liability business entity providing ...	shield business entity civil liability relatin...
1	section 1 short title act may cited human righ...	human right information act requires certain f...
2	section 1 short title act may cited jackie rob...	jackie robinson commemorative coin act directs...
3	section 1 nonrecognition gain rollover small b...	amends internal revenue code provide temporari...
4	section 1 short title act may cited native ame...	native american energy act sec 3 amends energy...
...	...	...
62702	person1 excuse mr green manchester arent perso...	tan ling pick mr green easily recognized white...
62703	person1 mister ewing said show conference cent...	person1 person2 plan take underground together...
62704	person1 help today person2 would like rent car...	person2 rent small car 5 day help person1
62705	person1 look bit unhappy today whats person2 w...	person2s mom lost job person2 hope mom wont fe...
62706	person1 mom im flying visit uncle lee family n...	person1 asks person2s idea packing bag visitin...

62707 rows × 2 columns

```
count    62707.000000
mean       448.081937
std        459.087443
min         16.000000
25%        154.000000
50%        255.000000
75%        618.000000
max       6014.000000
Name: summary, dtype: float64
```



* In characters.

https://drive.google.com/drive/folders/1yH89iZmARdc-R7QY6pwfE8tbOJI_n9K8?usp=sharing

Infosys_Text-Summarization/src/data_preprocessing.ipynb at main · MohanKrishnaGR/Infosys_Text-Summarization (github.com)

Proposal
Model Training

- Load pre-trained transformer
  - Facebook/bart
  - (or) Google/T5

- OOP implementation of Dataset
  - Feature, Target
  - Tokenize
  - Padding, Truncate
  - Convert to Tensor
  - Pass to: DataLoader – with batch size

- Training Loop
  - Adam optimizer
  - Forward pass & compute loss
  - Backward pass
  - Update params – compute gradient
  - Update LR
  - Zero the gradients
  - Update total loss

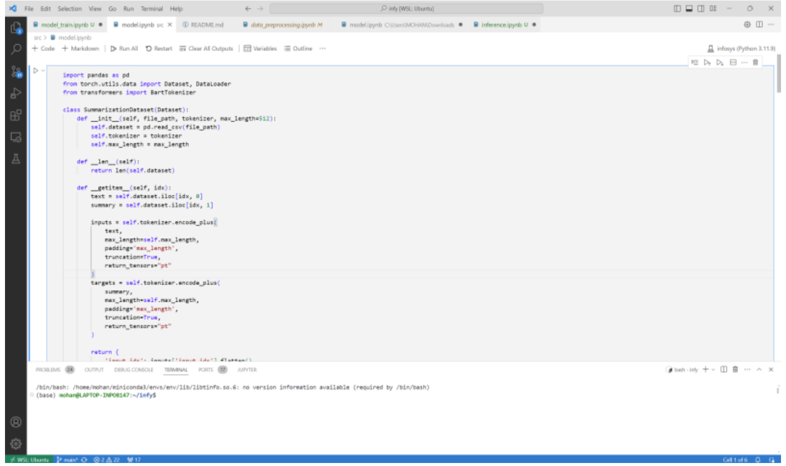


Fig. : Screenshot

Fig. : Fine-Tunning Overview

[Infosys_Text-Summarization/src/model.ipynb at main · MohanKrishnaGR/Infosys_Text-Summarization (github.com)]

Proposal
Model Validation

- Performance metrics – ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
  - Overlap between generated summary and reference summary.
  - Best suited : evaluating 'Text Summarization' tasks.
  - Other options : BLEU.

- Aimed to: implement custom evaluation function.
  - Calc. : ROUGE based on model's inference.

- Implementation:
  - Use: same Data loading methods - OOP.
  - Load the saved model & tokenizer.
  - Use 'ROUGE' metric from the Hugging Face's 'datasets' library.
    - To evaluate.
    - Calculates for all in dataloader

[Infosys_Text-Summarization/src/evaluation.ipynb at main · MohanKrishnaGR/Infosys_Text-Summarization (github.com)]