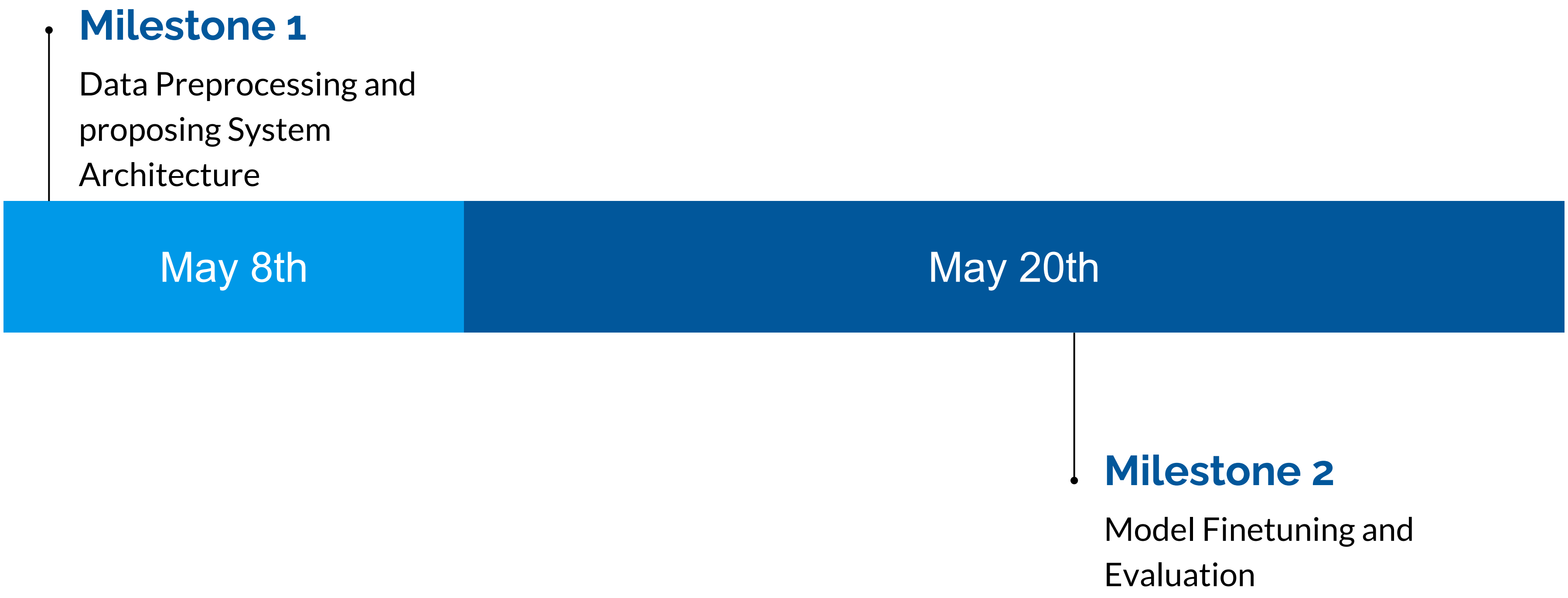
NLP Project: Text Summarization

By
Ahmed Serry, Farida Helmy and Seif Maged
Supervised by:
Dr. Mervat Abuelkheir
Mayar Ossama

Motivation

With the breakthrough of Chat-GPT, the Text summarization task caught our attention of how accurate such language models can capture the essence of information from raw large corpus.

Milestones



Milestone 1

Data Preprocessing and proposing System Architecture

May 8th

May 20th

Milestone 2

Model Finetuning and Evaluation

1. Our Dataset

We imported the CNN Daily News Dataset from Huggingface, which is the essential dataset for training our language model.

➔ **Training Dataset**

Will be preprocessed and cleaned for our model training.

➔ **Validation Dataset**

Used for tuning the hyperparameters of our model.

➔ **Testing Dataset**

Used to evaluate our finetuned model.

A limitation that we faced

There are 198,000,000 words in the original dataset: sampling for simplicity

After Sampling

49,000,000 words are present after sampling.

Data Exploration

Data Preprocessing

- Removing stop words and punctuation

Data Analysis

- Articles and highlights columns, and better understanding of both features

Quotes for illustration purposes only

Our chosen Architecture

BART

Abstractive models use advanced NLP (i.e. word embeddings) to understand the semantics of the text and generate a meaningful summary.

Tip

Experimenting with fine tuning the model on our dataset to preview the Text summarization Task.

Thank you!

Any questions or suggestions?