NEWS SUMMARIZATION

Sarath Manoj,
Akshul Mittal,
Amit Gupta,
Devansh Shrestha

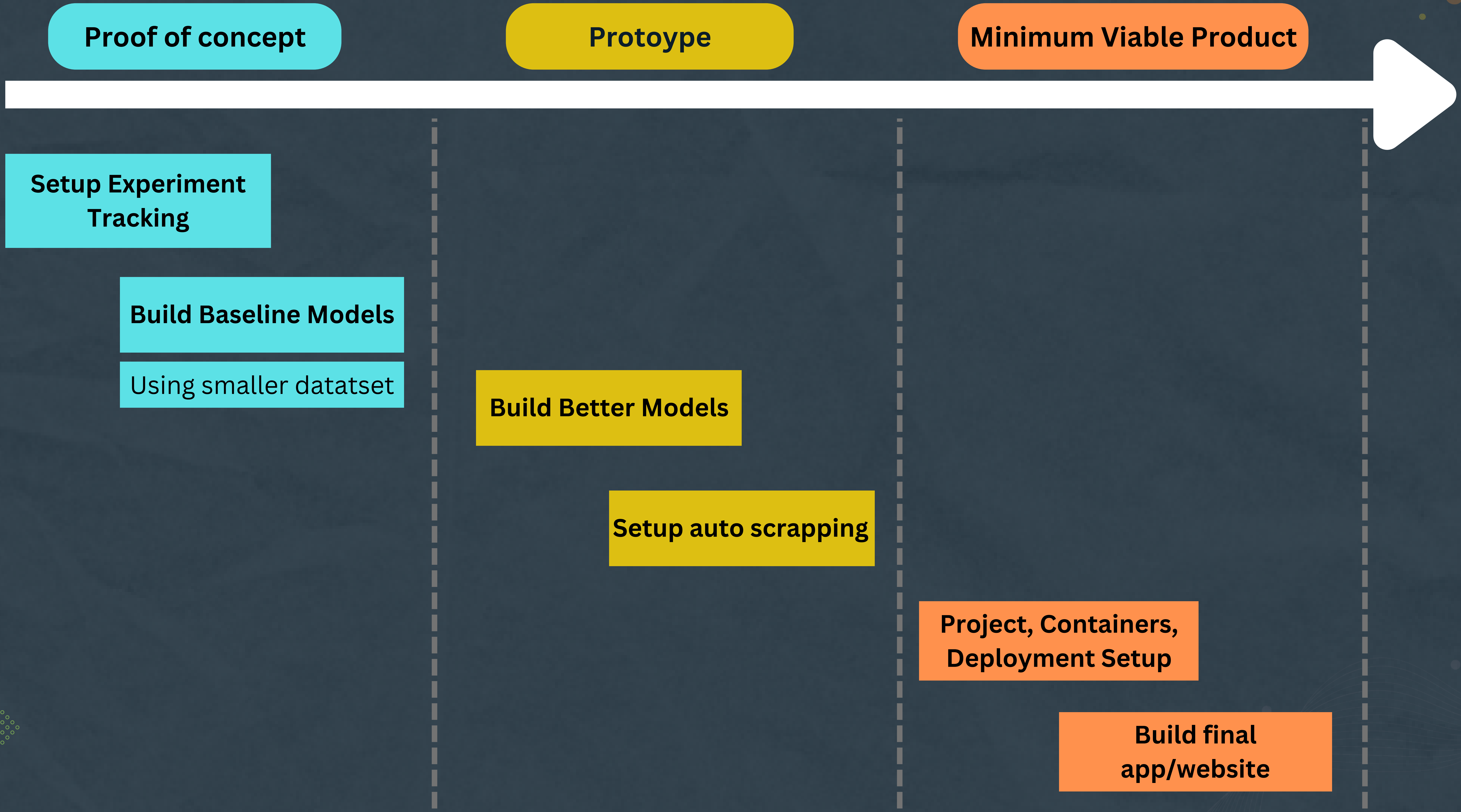
PROBLEM DEFINITION

Joseph wants to be updated on the current affairs to be an informed individual . But the large influx of information , shortened attention span and lack of time poses huge obstacles in this pursuit. Further, he is unable to consume all the wisdom that past literature works offer. Despite these challenges , he will continue to build his reading habits if they are more rewarding to match the more challenges he has to overcome.

Proposed Solution

We can lighten the burden of people like Joseph by providing meaningful summaries of news and other articles .
They can read from the selection of news summaries from scrapped news articles or input the article that they want to be summarized .
We do this by training an algorithm to identify the key insights and remove unnecessary description.

PROJECT WORKFLOW



Proof of concept

Protoype

Minimum Viable Product

Setup Experiment Tracking

Build Baseline Models

Using smaller datatset

Build Better Models

Setup auto scrapping

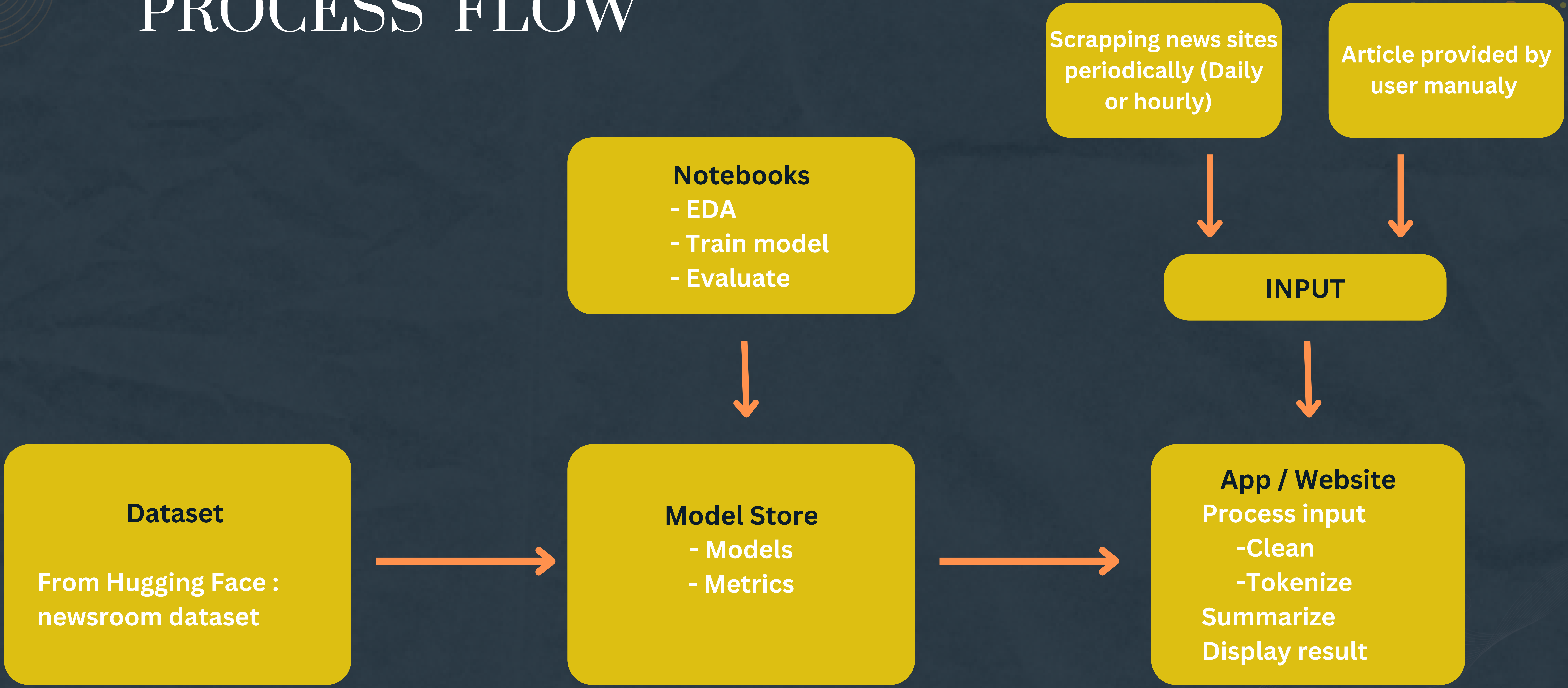
Project, Containers, Deployment Setup

Build final app/website

TIME LINE

Proof of concept	Protoyope	Minimum Viable Product
1 week	2 weeks	2 - 3 weeks

PROCESS FLOW



Scrapping news sites periodically (Daily or hourly)

Article provided by user manualy

Notebooks
- EDA
- Train model
- Evaluate

INPUT

Dataset

From Hugging Face : newsroom dataset

Model Store
- Models
- Metrics

App / Website
Process input
-Clean
-Tokenize
Summarize
Display result

DATASET

CORNELL NEWSROOM is a large dataset for training and evaluating summarization systems. It contains 1.3 million articles and summaries written by authors and editors in the newsrooms of 38 major publications. The summaries are obtained from search and social metadata between 1998 and 2017

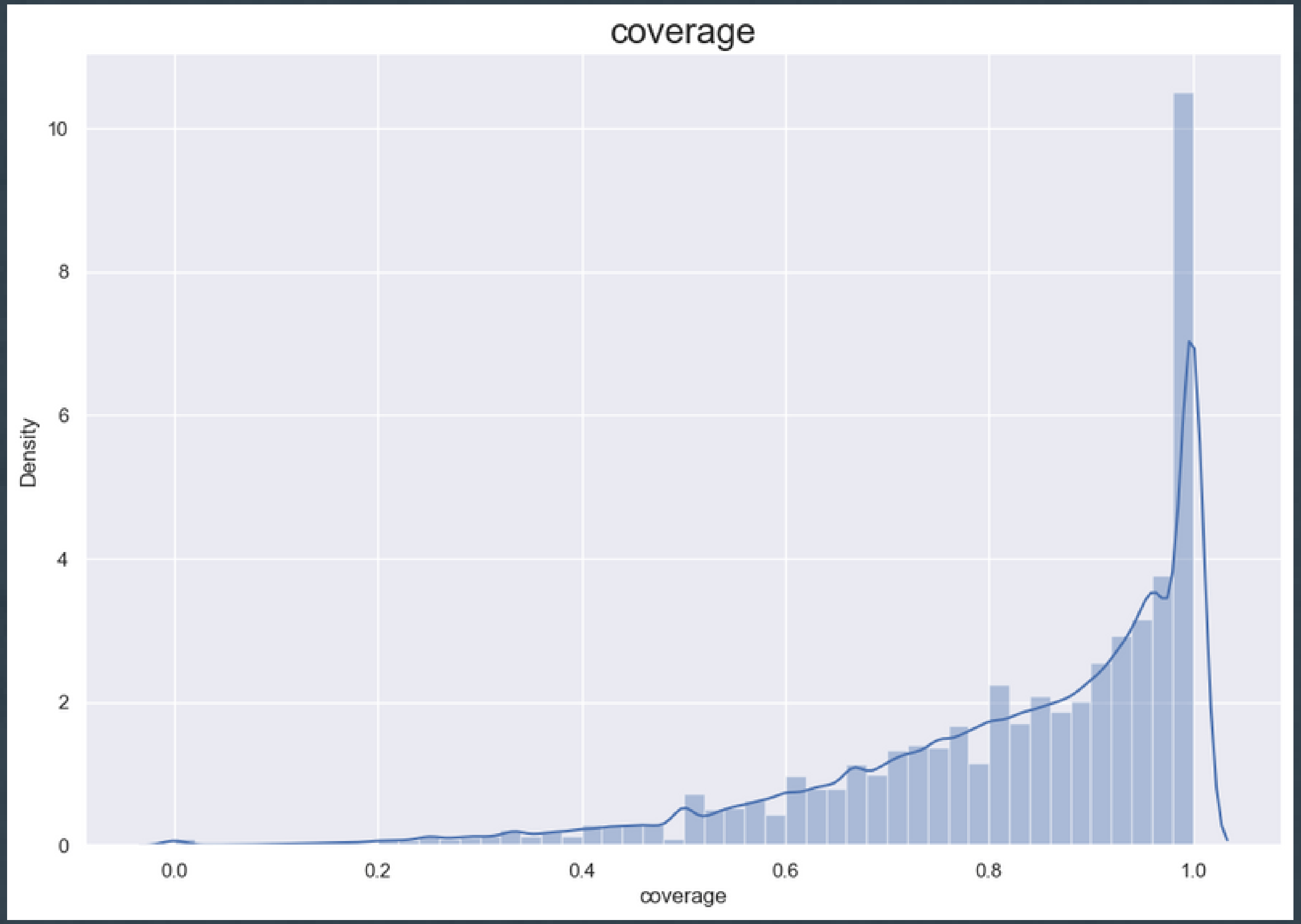
- Text
- Summary
- Compression
- Date
- Density
- URL
- Title

Dataset link: https://lil.nlp.cornell.edu/newsroom/index.html
https://huggingface.co/datasets/newsroom

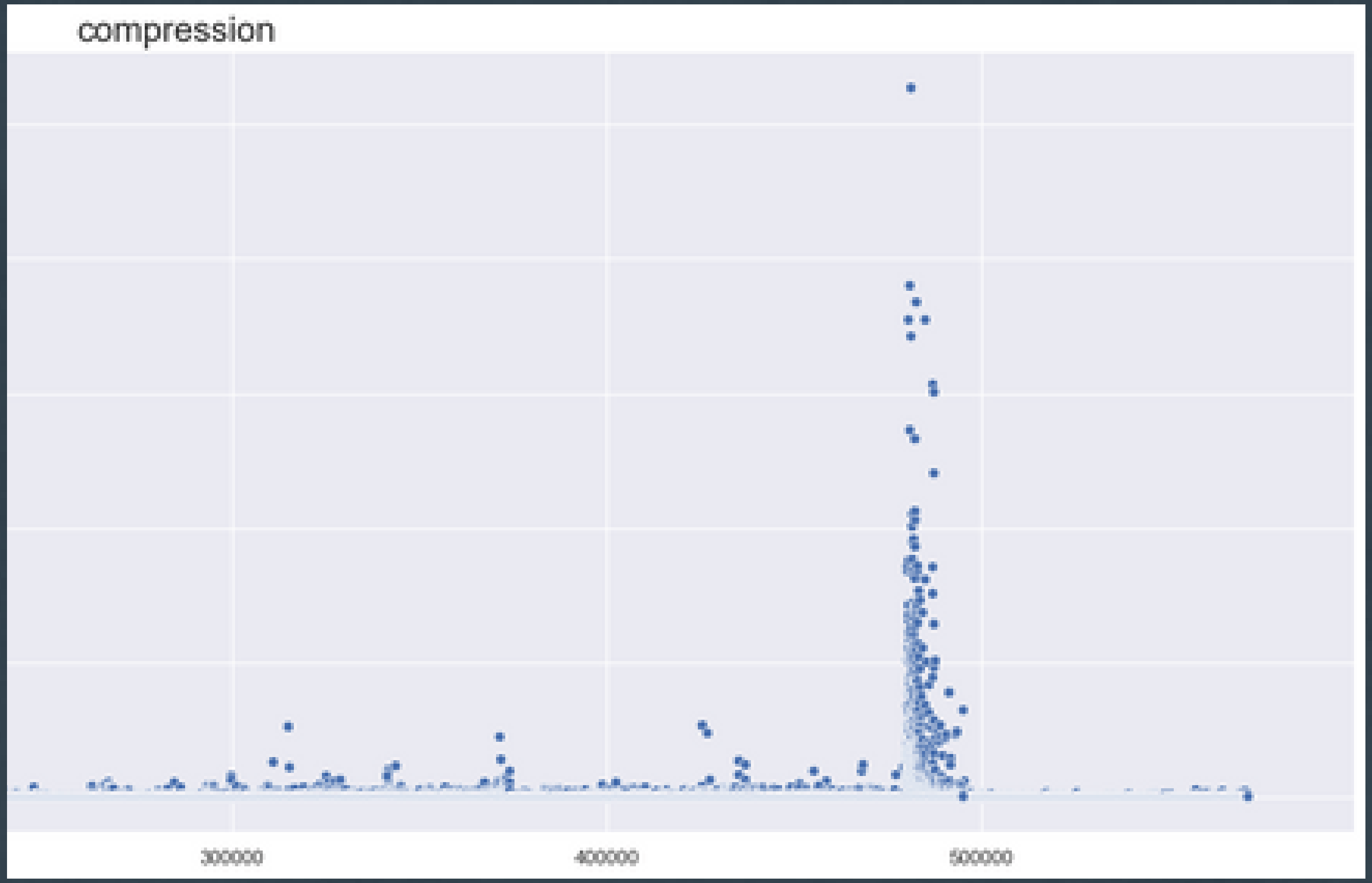
MODELS

- Seq2seq
- Seq2seq + attention
- BERT
- Word2vec Embeddings
- Transformers

EDA



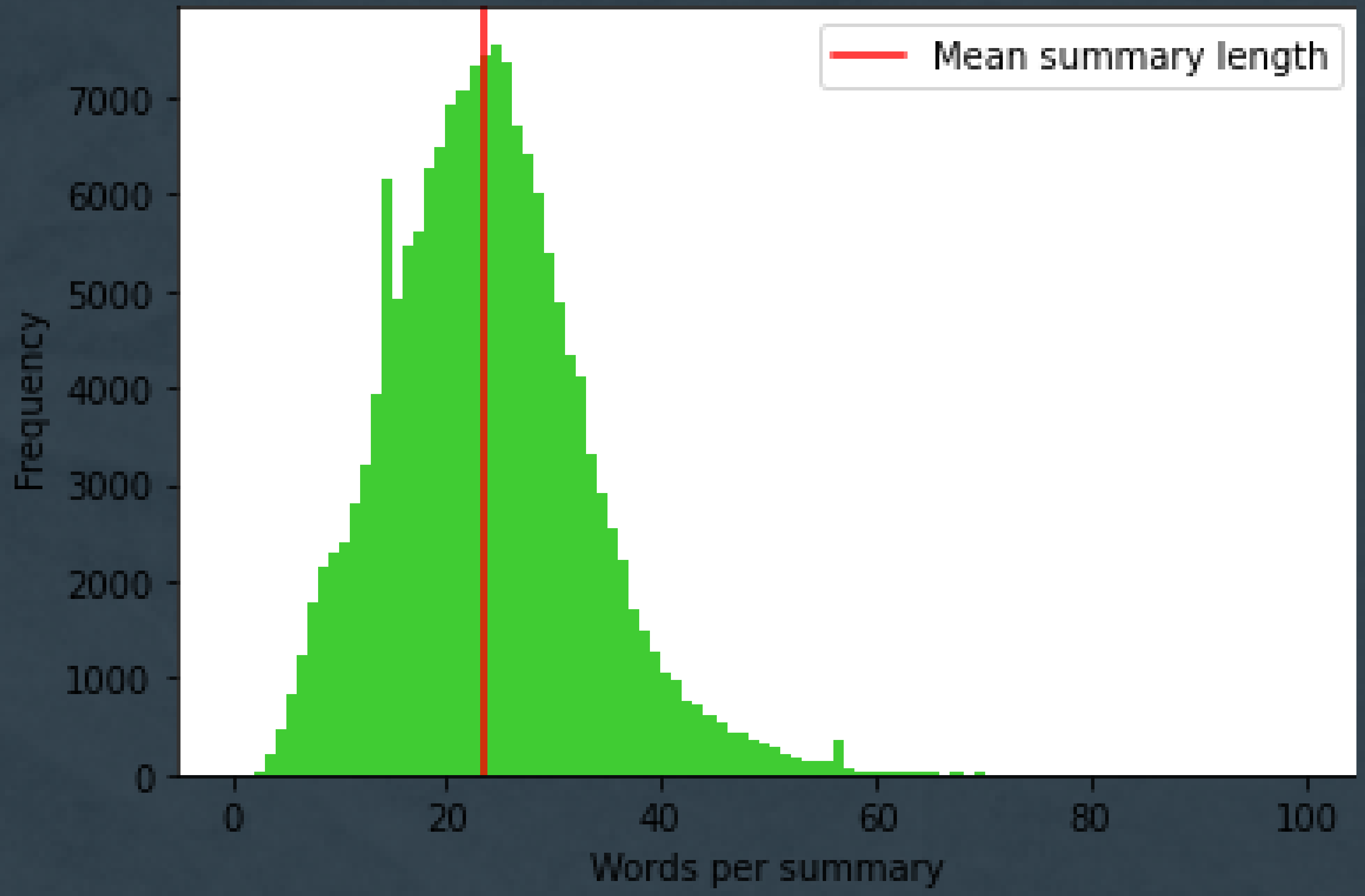
Coverage



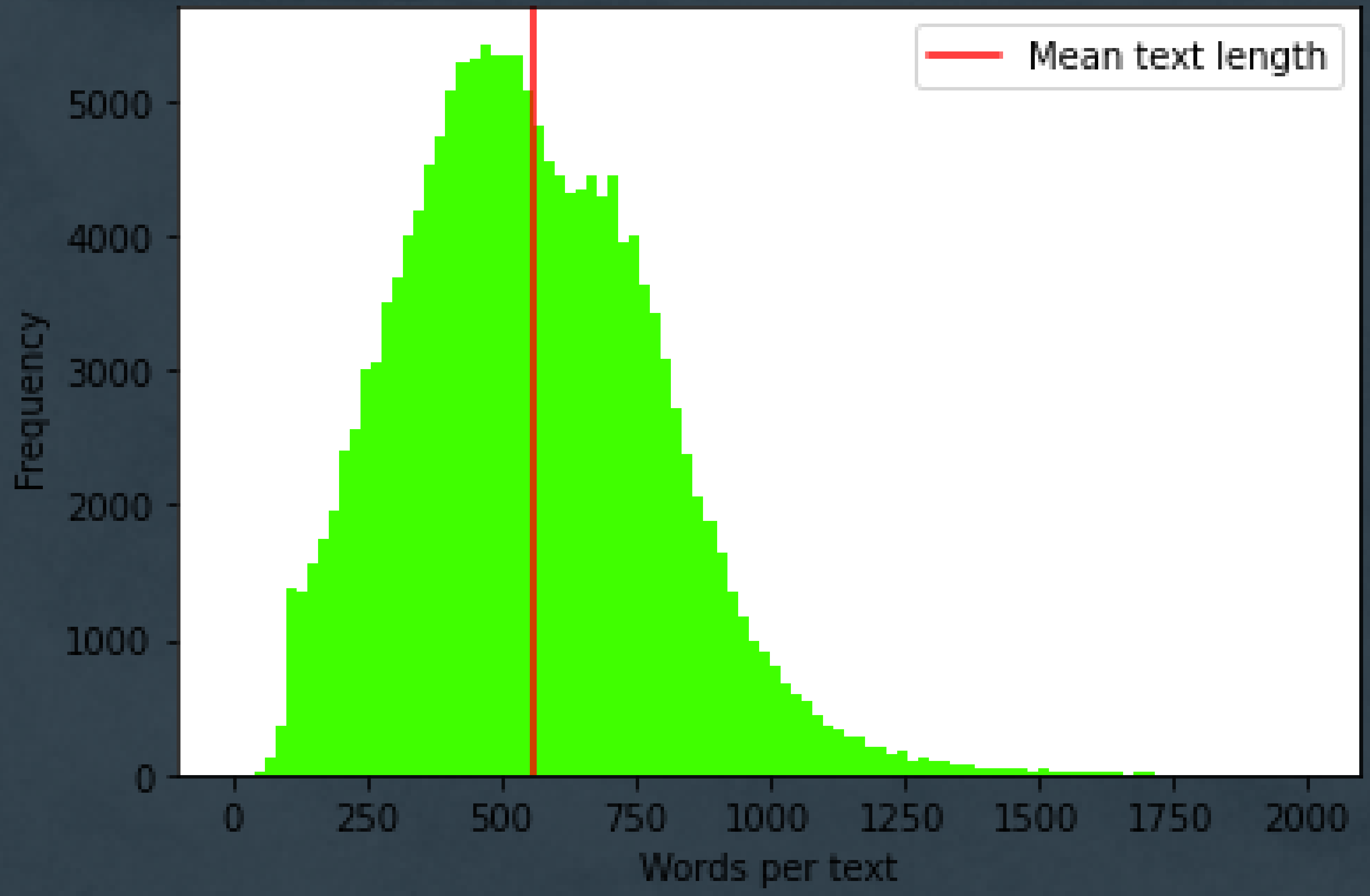
Compression

EDA

Distributions of sentence lengths



Summary lengths



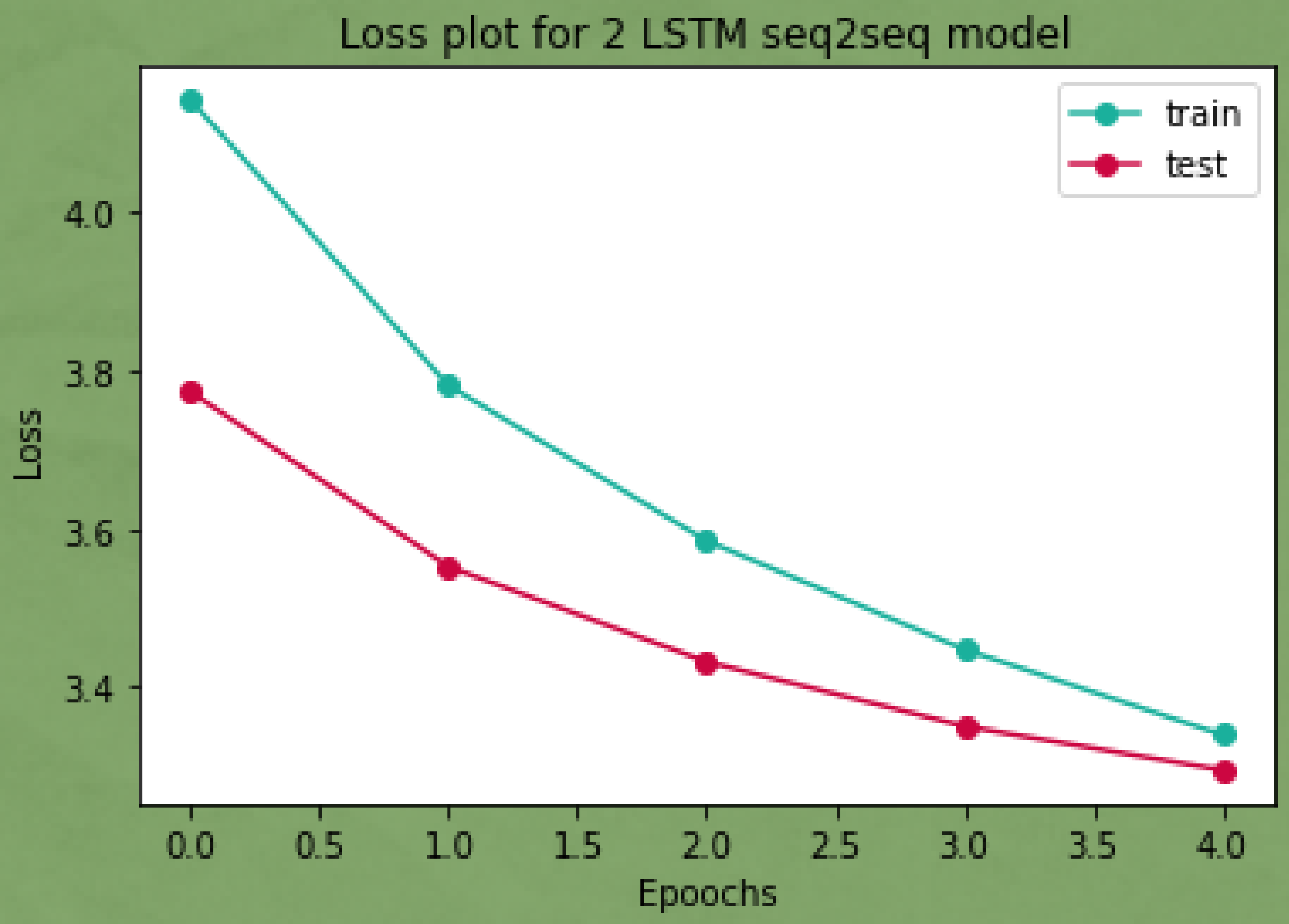
Text lengths

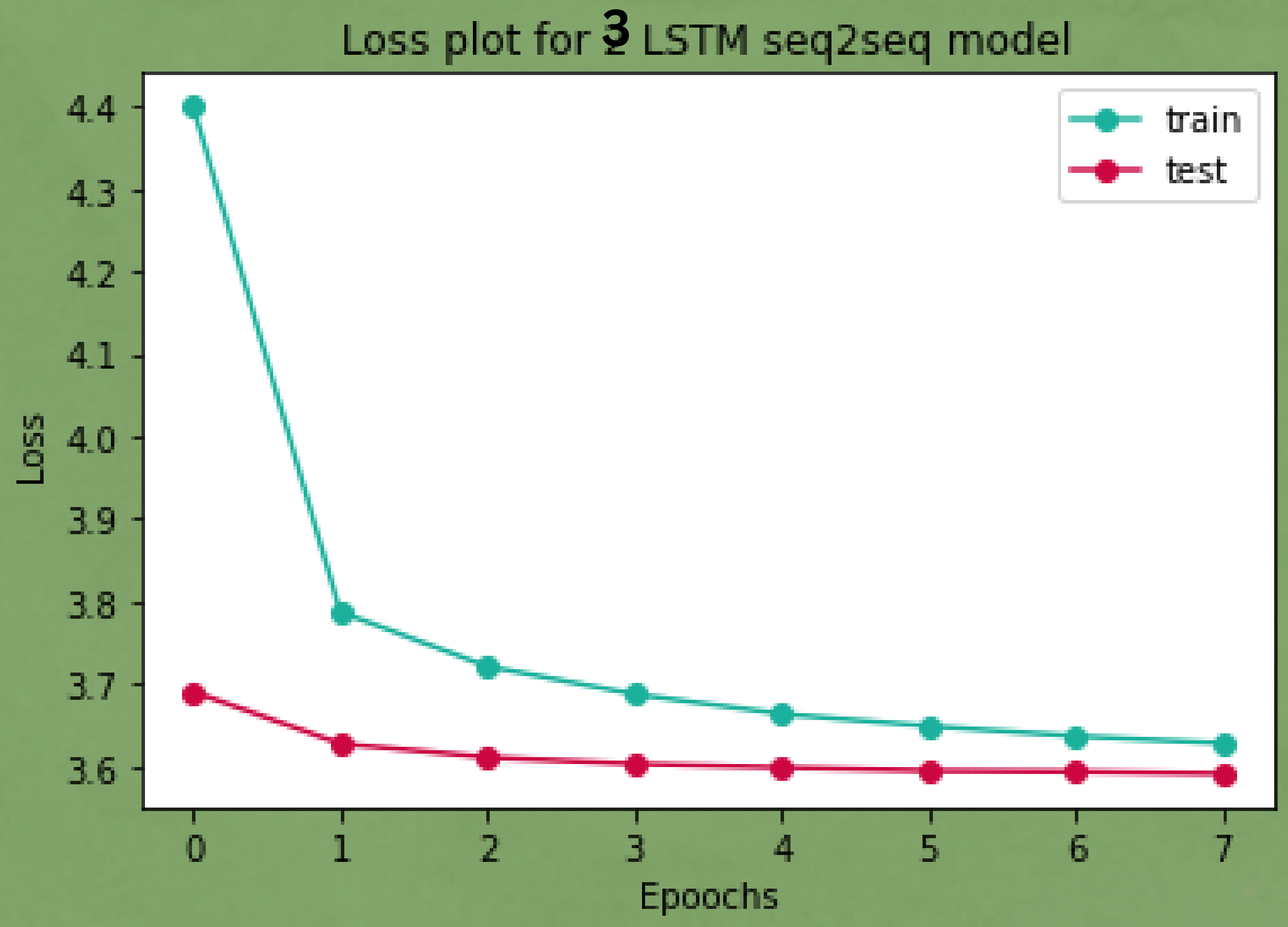
MODELS ON SUBSET OF DATA

Models	Execution time	Train loss	Validation loss
Seq2seq with 2 Encoder LSTM	58 minutes	3.3393	3.2945
Seq2seq with 3 Encoder LSTM	1 hour 43 minutes	3.6283	3.5911

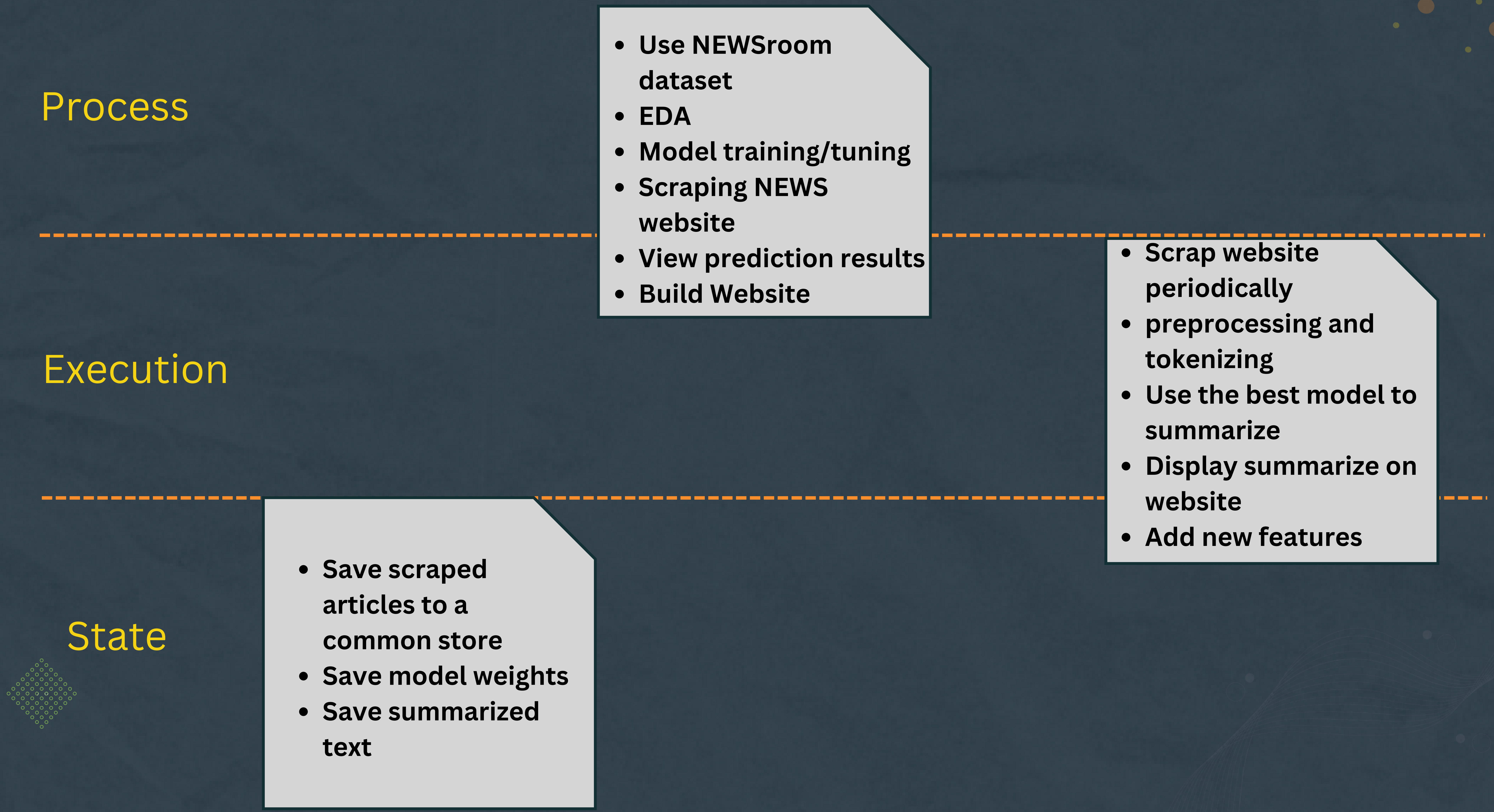
*Here loss is sparse categorical crossentropy

LOSS PLOTS


Loss plot for 2 LSTM seq2seq model
train
test
Loss
4.0
3.8
3.6
3.4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epoochs


Loss plot for 3 LSTM seq2seq model
train
test
Loss
4.4
4.3
4.2
4.1
4.0
3.9
3.8
3.7
3.6
0
1
2
3
4
5
6
7
Epoochs

News summarization



Process

- Use NEWSroom dataset
- EDA
- Model training/tuning
- Scraping NEWS website
- View prediction results
- Build Website

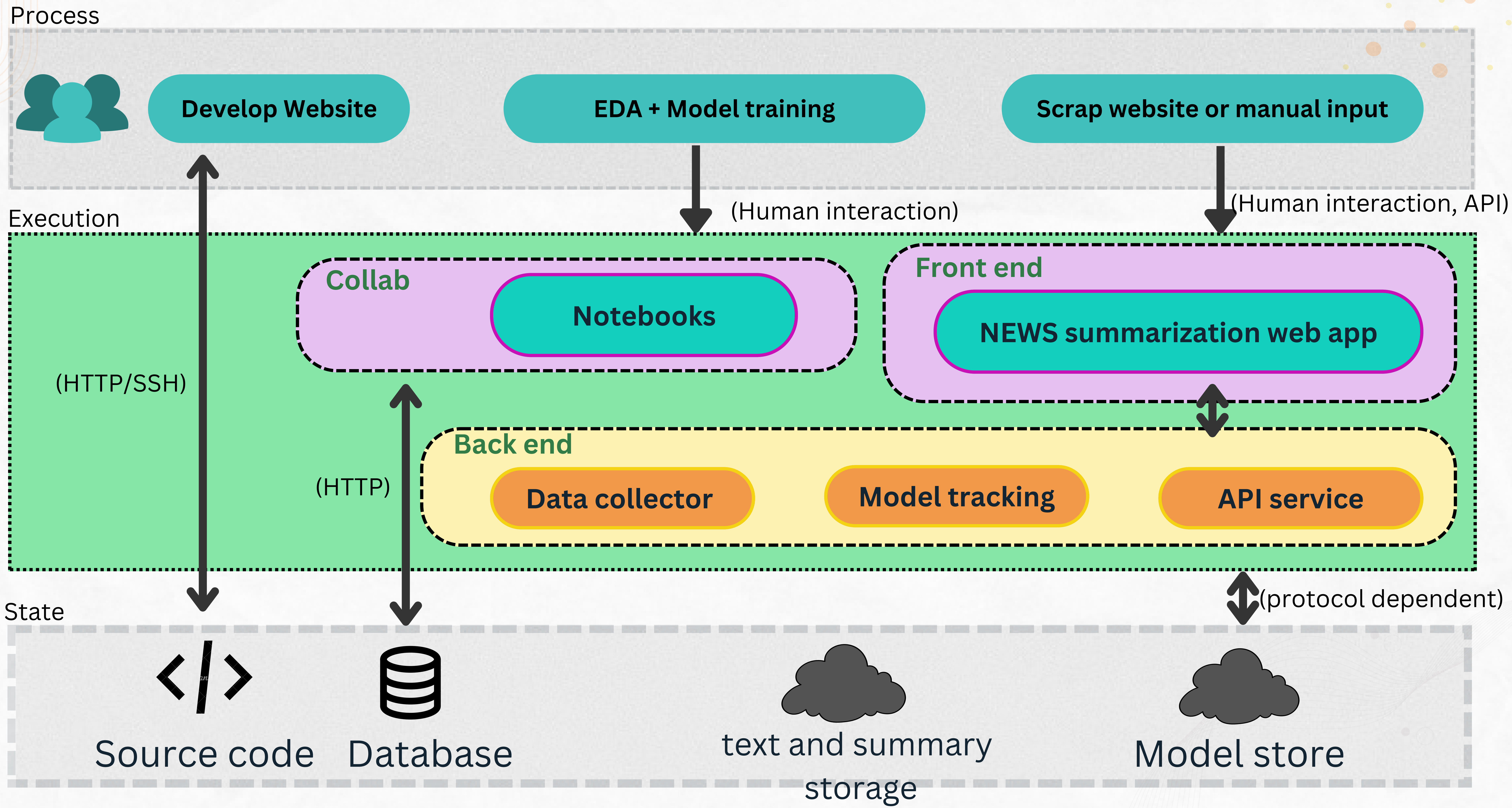
Execution

- Scrap website periodically
- preprocessing and tokenizing
- Use the best model to summarize
- Display summarize on website
- Add new features

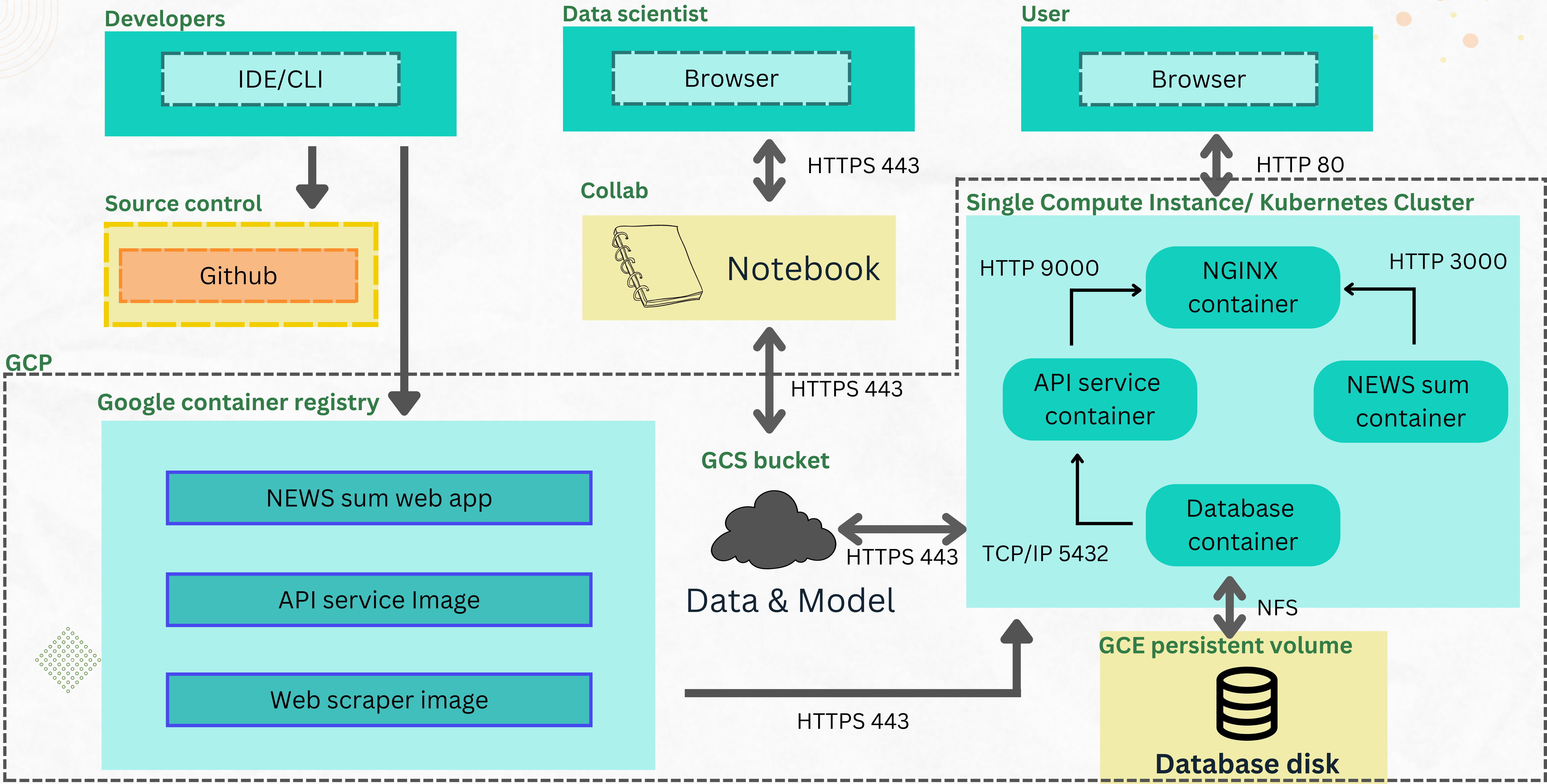
State

- Save scraped articles to a common store
- Save model weights
- Save summarized text

SOLUTION ARCHITECTURE



TECHNICAL ARCHITECTURE


Developers
IDE/CLI
Data scientist
Browser
User
Browser
HTTPS 443
HTTP 80
Source control
Github
Collab
Notebook
Single Compute Instance/ Kubernetes Cluster
HTTP 9000
NGINX container
HTTP 3000
GCP
API service container
NEWS sum container
HTTPS 443
Google container registry
GCS bucket
NEWS sum web app
API service Image
Web scraper image
HTTPS 443
TCP/IP 5432
Database container
Data & Model
NFS
GCE persistent volume
Database disk
HTTPS 443