

NEWS SUMMARIZATION

S a r a t h M a n o j ,
A k s h u l M i t t a l ,
A m i t G u p t a ,
D e v a n s h S h r e s t h a

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 dataset

Build Better Models

Setup auto scrapping

Project, Containers,
Deployment Setup

Build final
app/website

TIME LINE

Proof of concept

1 week

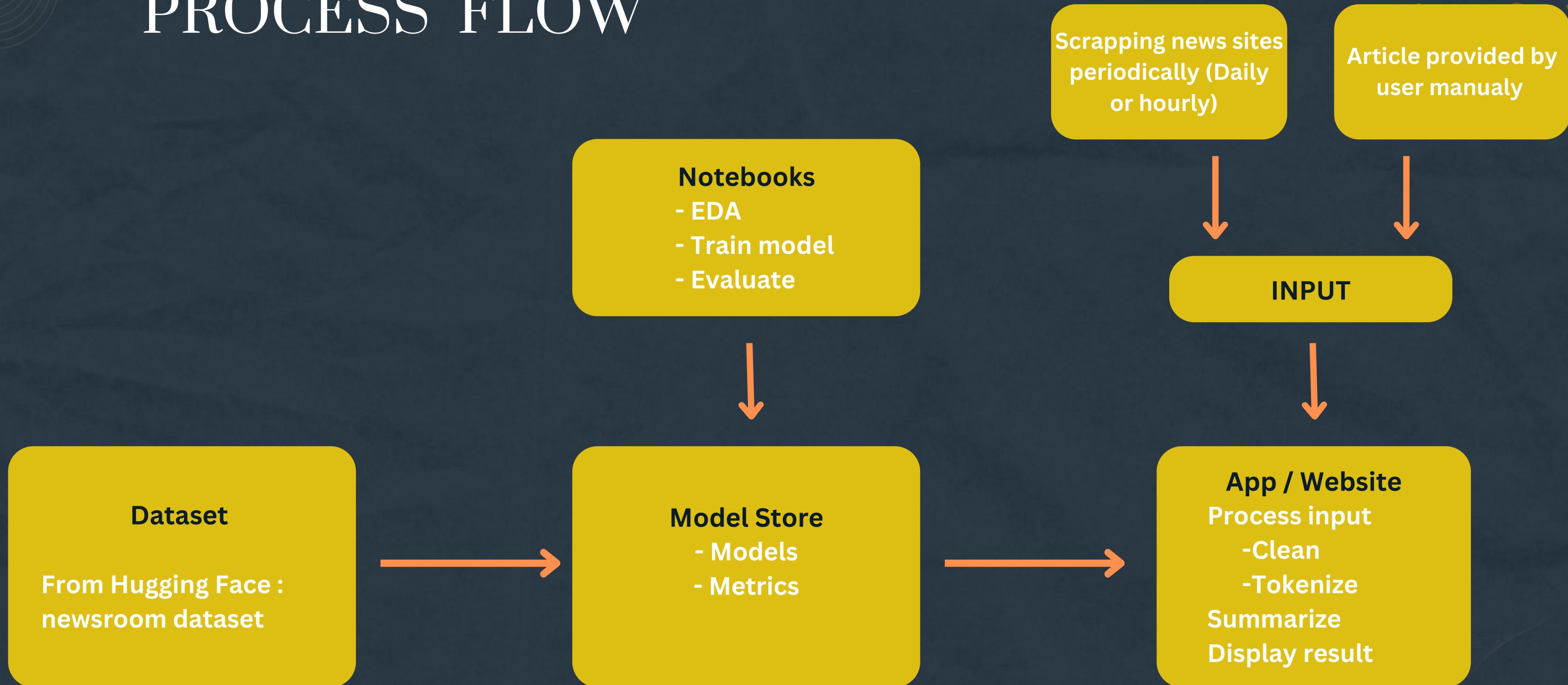
Protoype

2 weeks

**Minimum Viable
Product**

2 - 3 weeks

PROCESS FLOW



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

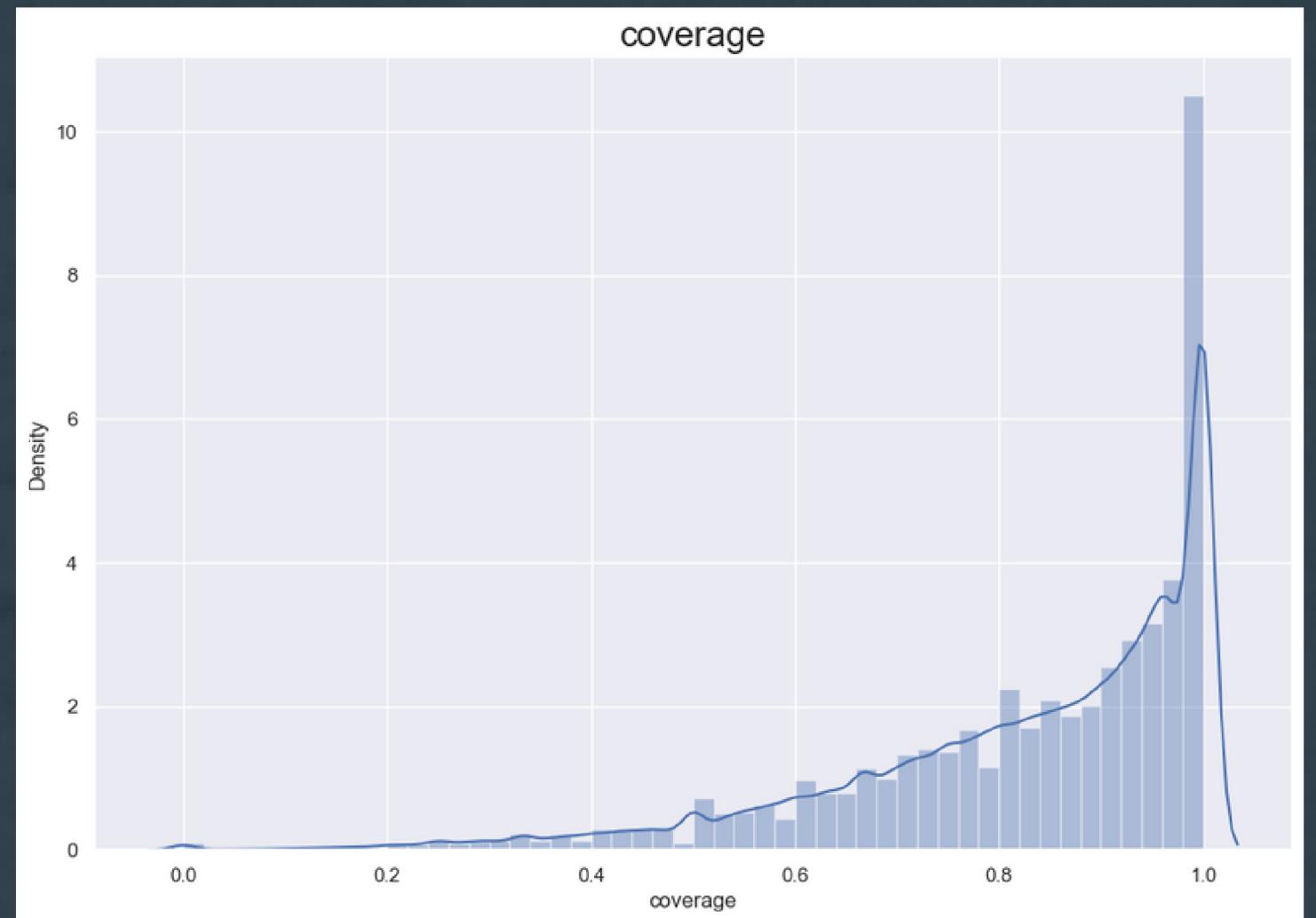
Dataset link: <https://lil.nlp.cornell.edu/newsroom/index.html>

<https://huggingface.co/datasets/newsroom>

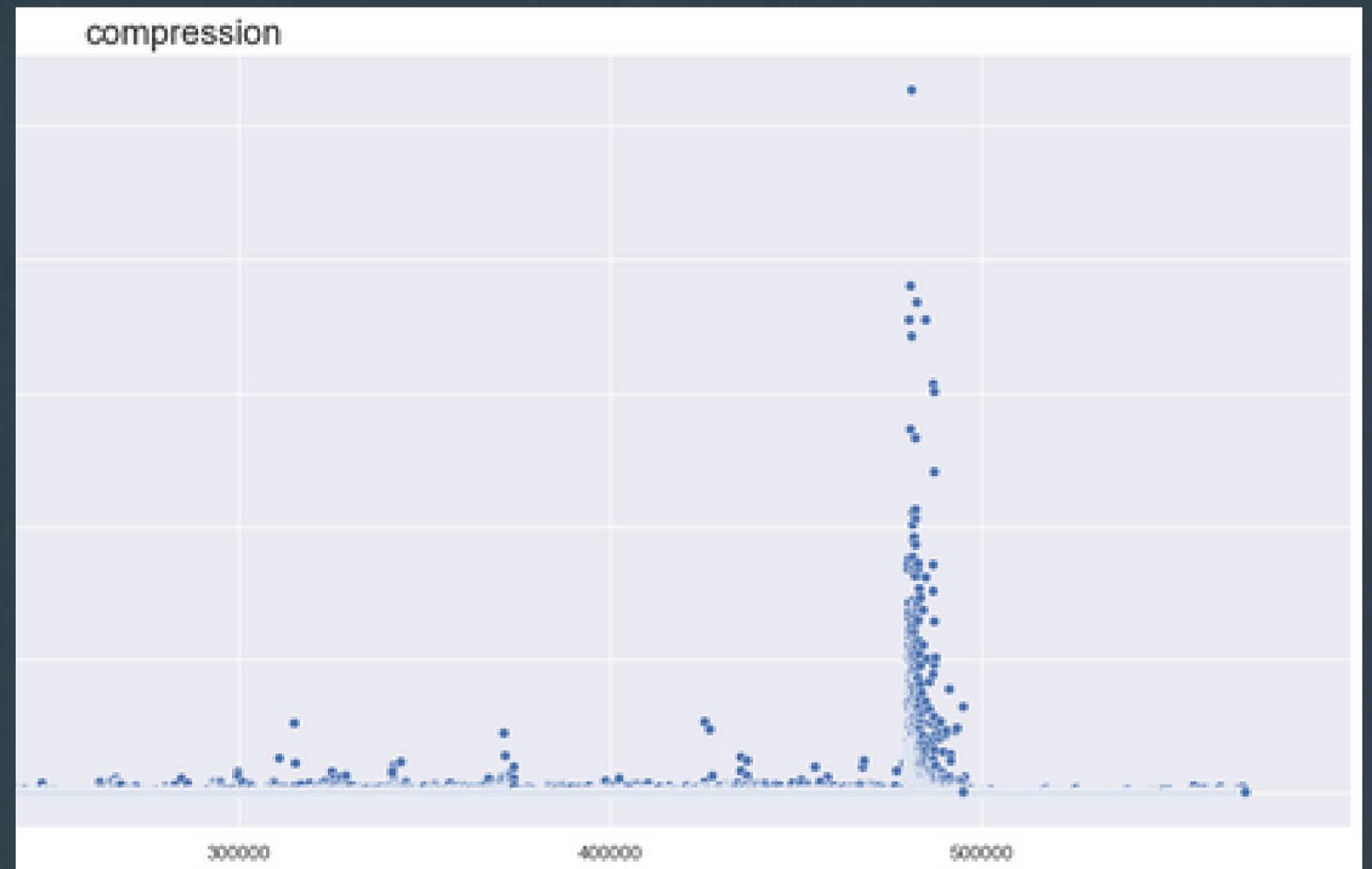
MODELS

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

EDA



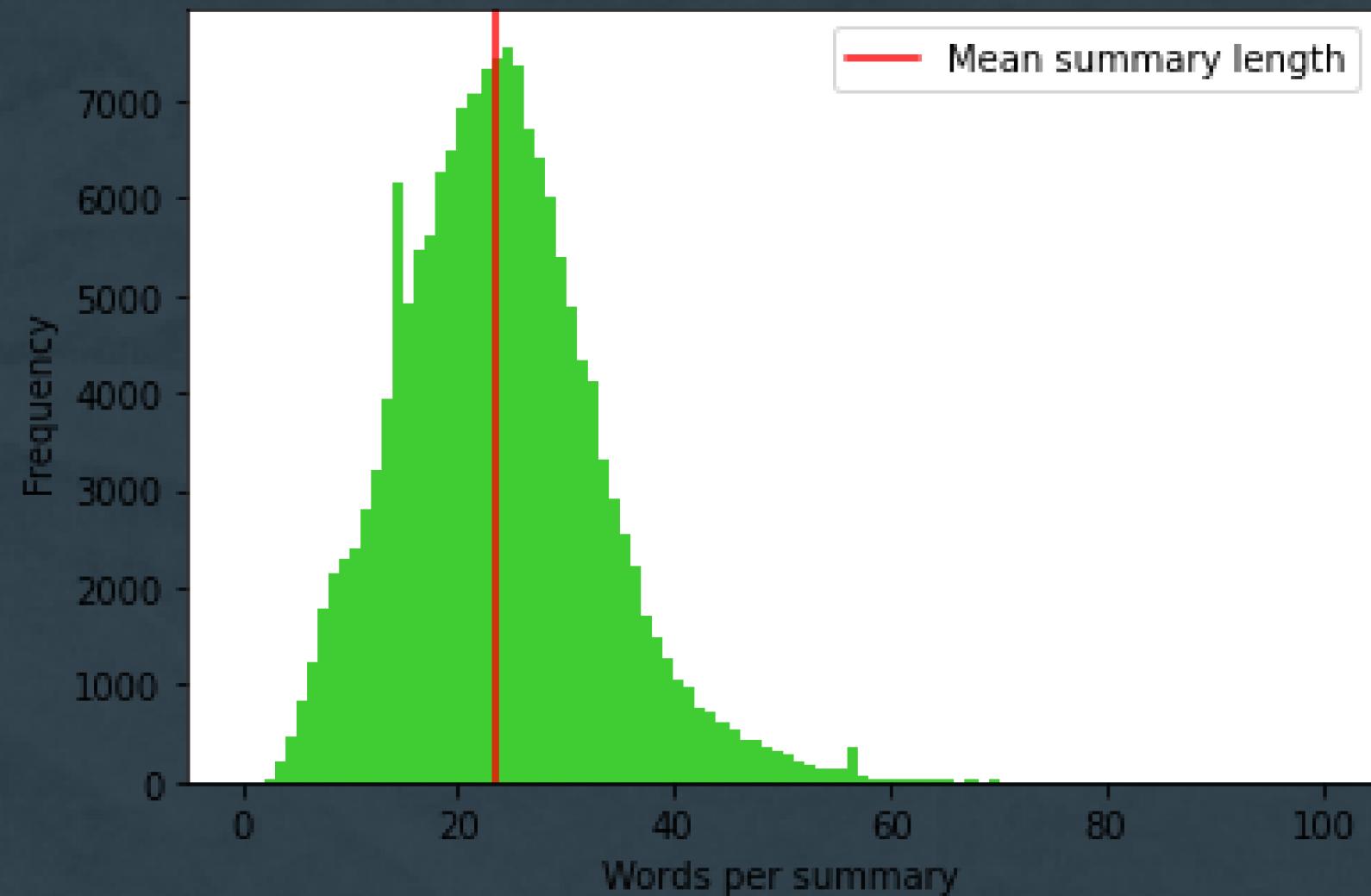
Coverage



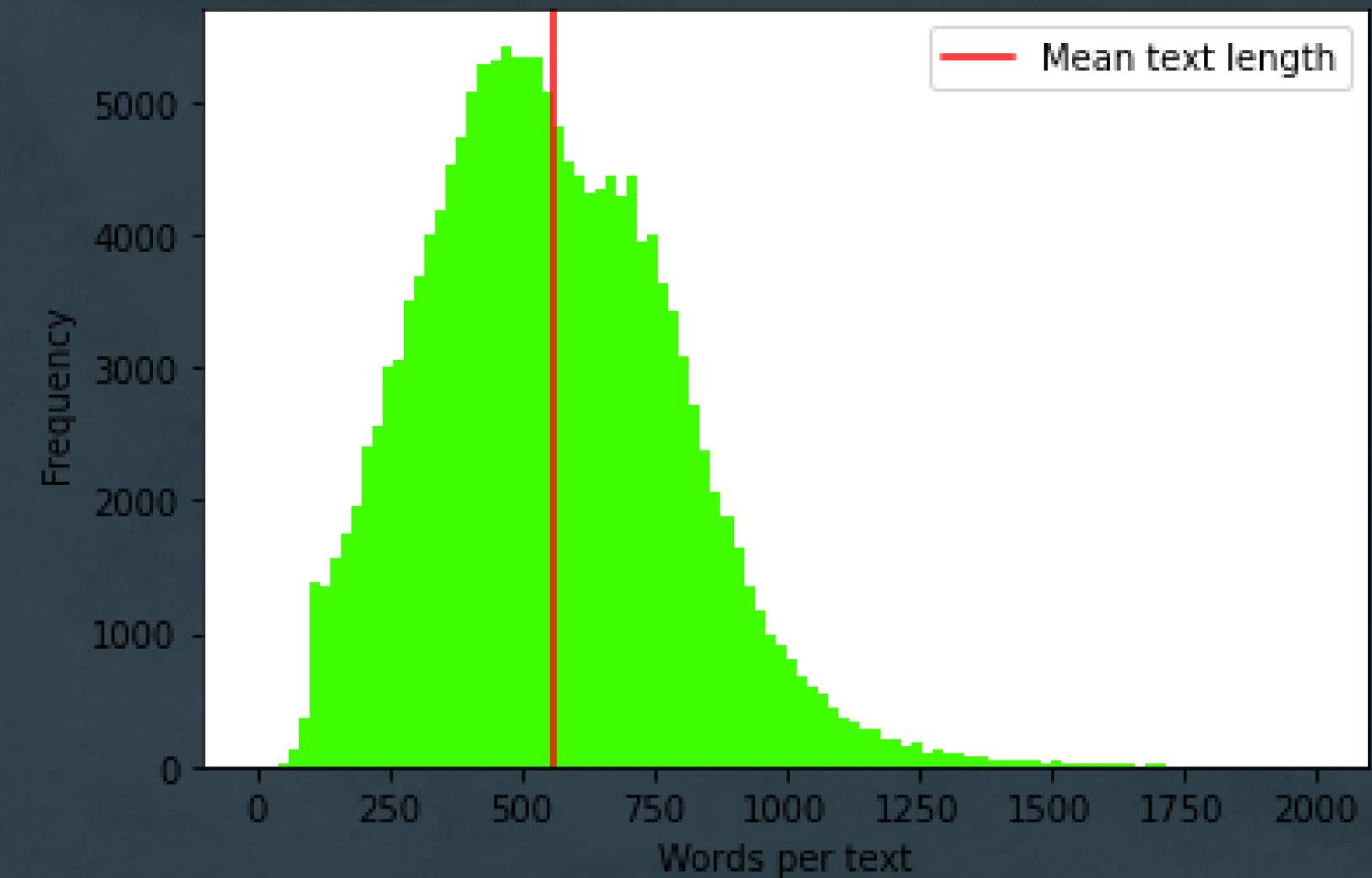
Compression

EDA

Distributions of sentence lengths



Summary lengths



Text lengths

MODELS ON SUBSET OF DATA

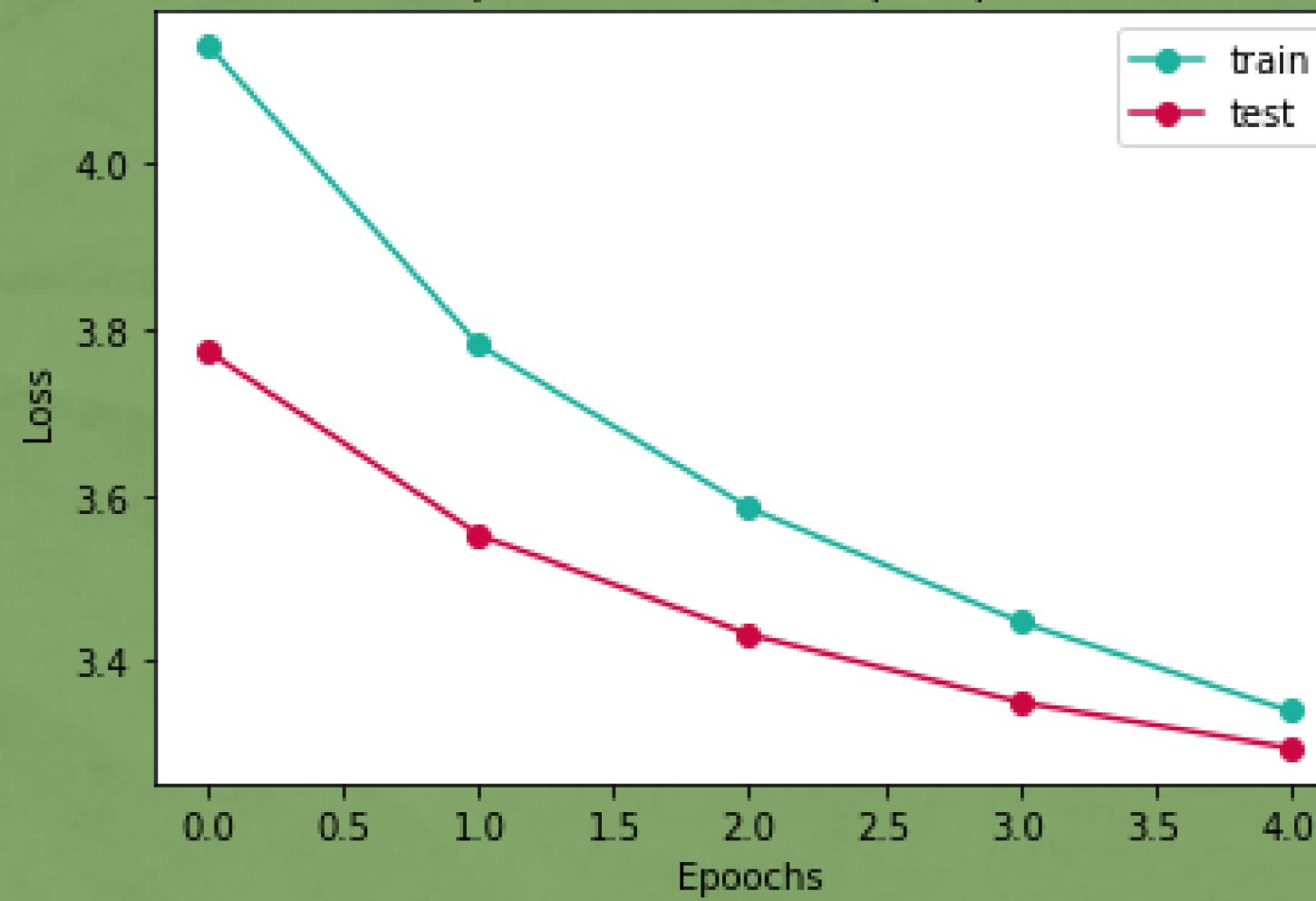


Models	Execution time	Train loss	Validation loss
Seq2seq with 2 Encoder LSTM	58 minutes	3.3393	3.2945
Seq2seq with 3 Encoder LSTM	1 hour 43 minutes	3.6283	3.5911

*Here loss is sparse categorical crossentropy

LOSS PLOTS

Loss plot for 2 LSTM seq2seq model



Loss plot for 3 LSTM seq2seq model

