Abstractive Text Summarization using Deep Learning

Team 7
SDS 384 Scientific Machine Learning
Aditya, Jaimik, Srishti, Yug

22nd April 2024

Introduction

- Our project focuses on text summarization using deep learning techniques applied to scientific papers.
- The primary goal was to develop models capable of automatically generating concise and informative summaries of scientific articles by leveraging deep learning algorithms.
- The project aims to extract key information from lengthy texts efficiently, facilitating faster comprehension and aiding researchers in identifying relevant literature.

Sample DataSet

PubMed (Source: https://huggingface.co/datasets/scientific_papers)

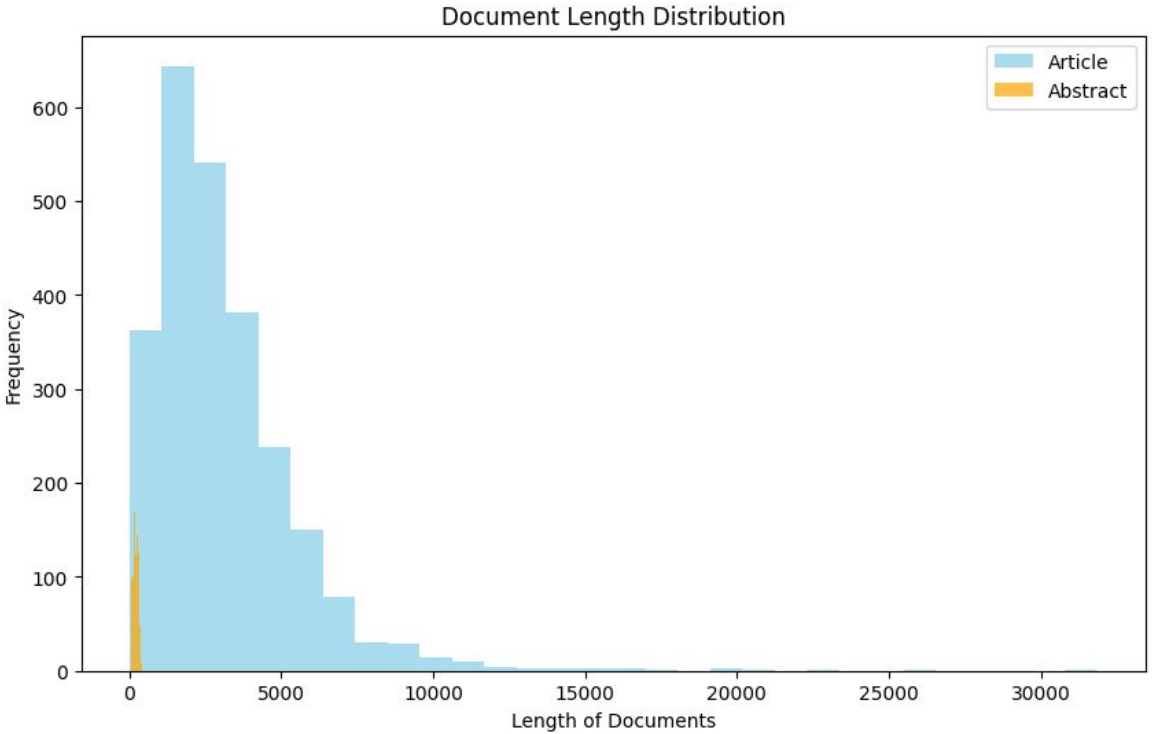
133k rows = 120k(Train) + 6.6k(Valid.) + 6.3k(Test)

1. article: the body of the document
2. abstract: the abstract of the document
3. section_names: titles of sections

	Unnamed: 0	article	abstract	section_names
0	73434	to review the presentation and histological di...	objective : to review the presentation and hi...	Objective:\nMaterials and Methods:\nResults:\n...
1	7459	ethylcellulose , a nonbiodegradable and biocom...	\n objective . \n the purpose of the recent s...	1. Introduction\n2. Materials and Methods\n3. ...
2	136	acute generalized exanthematous pustulosis (a...	acute generalized exanthematous pustulosis (...	Introduction\nCase Report\nDiscussion
3	76845	\n physical restraint is a coercive interventi...	\n background : considering the negative cons...	Introduction\nMethods\nResults\nDiscussion\nCo...
4	80361	disease registries are considered reliable sou...	the main aim of this study is to determine th...	1. Introduction\n2. Materials and Methods\n3. ...

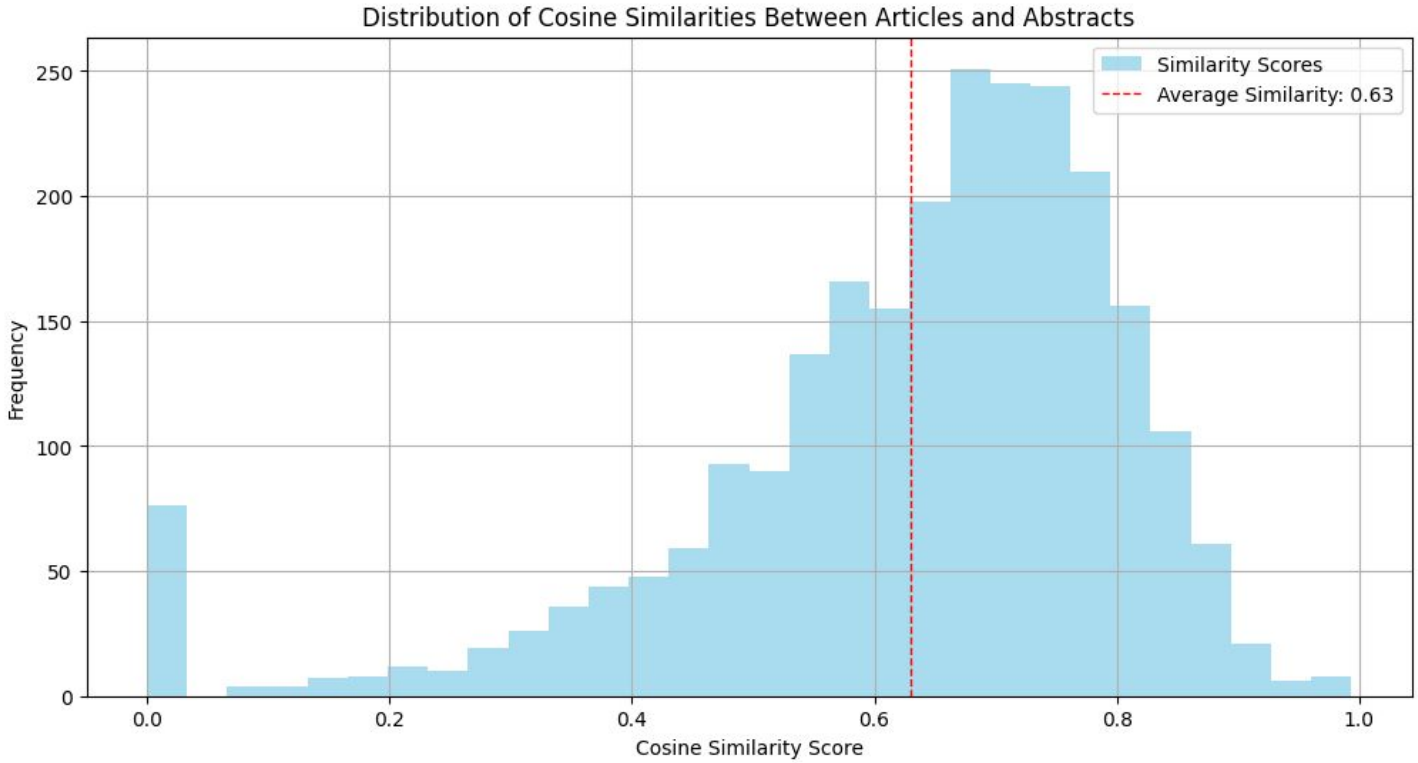
Exploratory Data Analysis

1. Length of the articles vs abstracts in the dataset



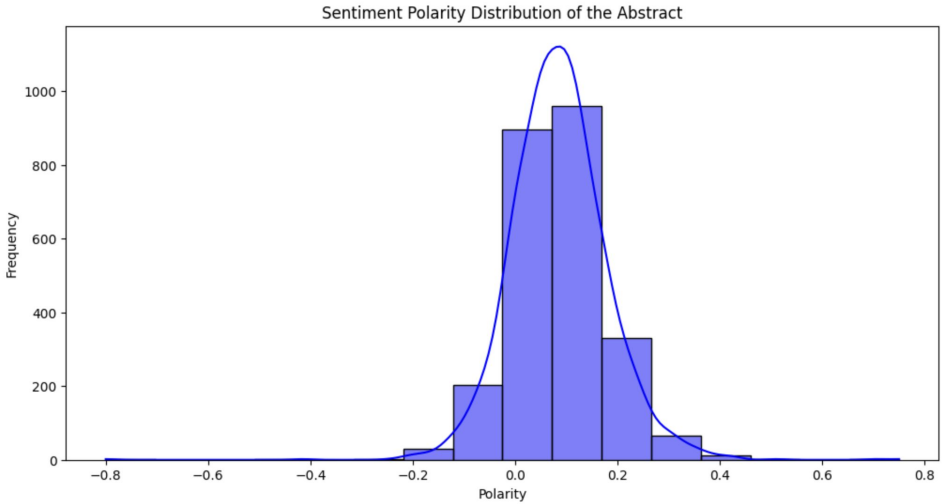
Exploratory Data Analysis

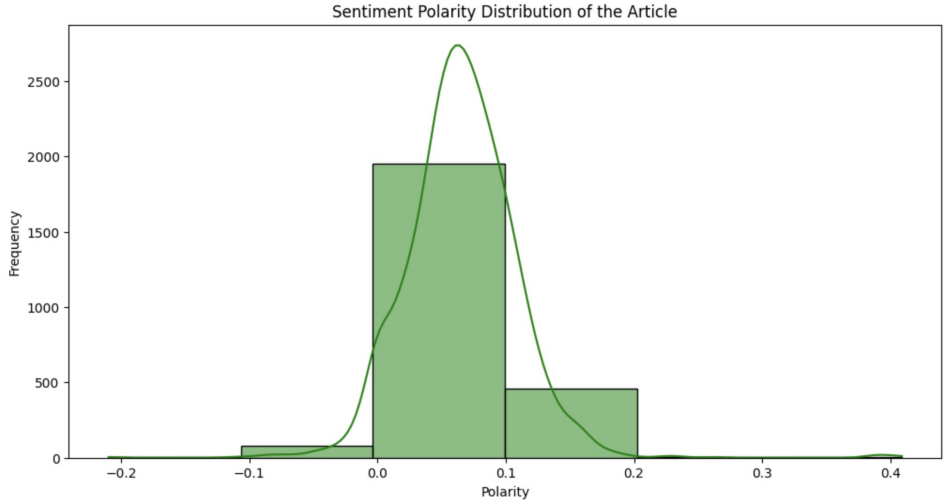
2. Similarity Analysis Between Abstracts and Articles



Exploratory Data Analysis

3. The sentiment analysis between the article and the abstract

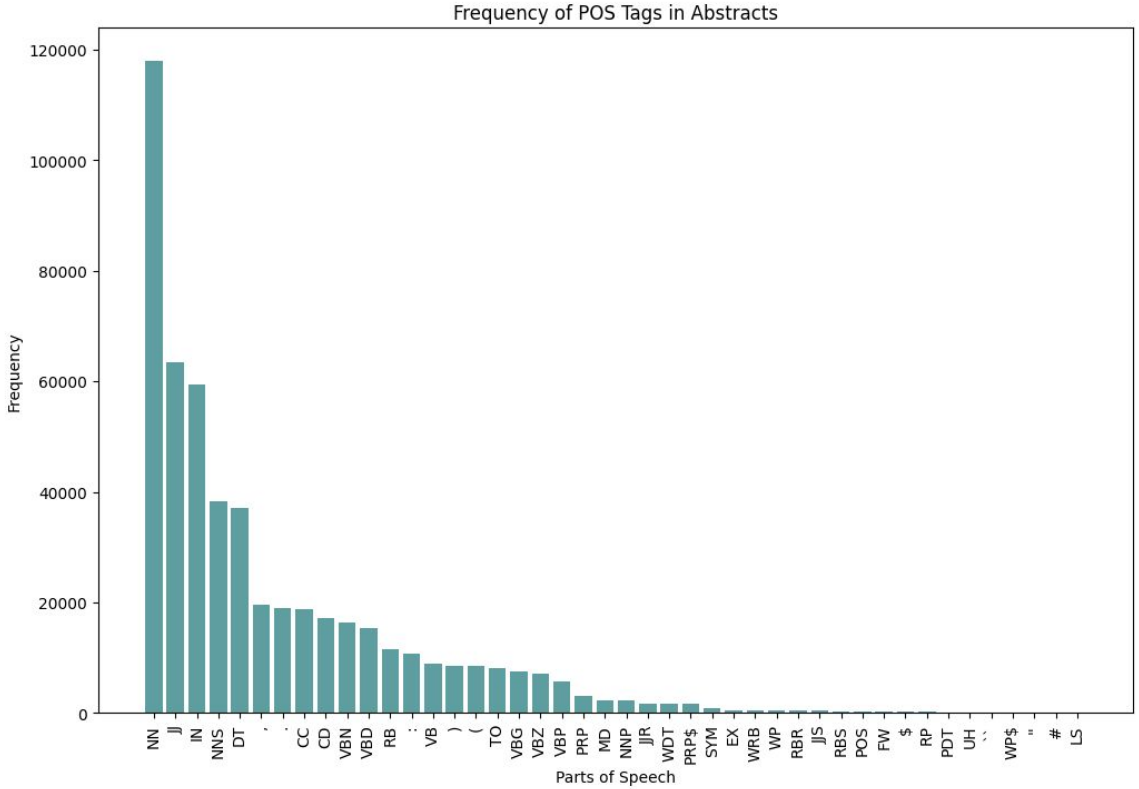




Exploratory Data Analysis

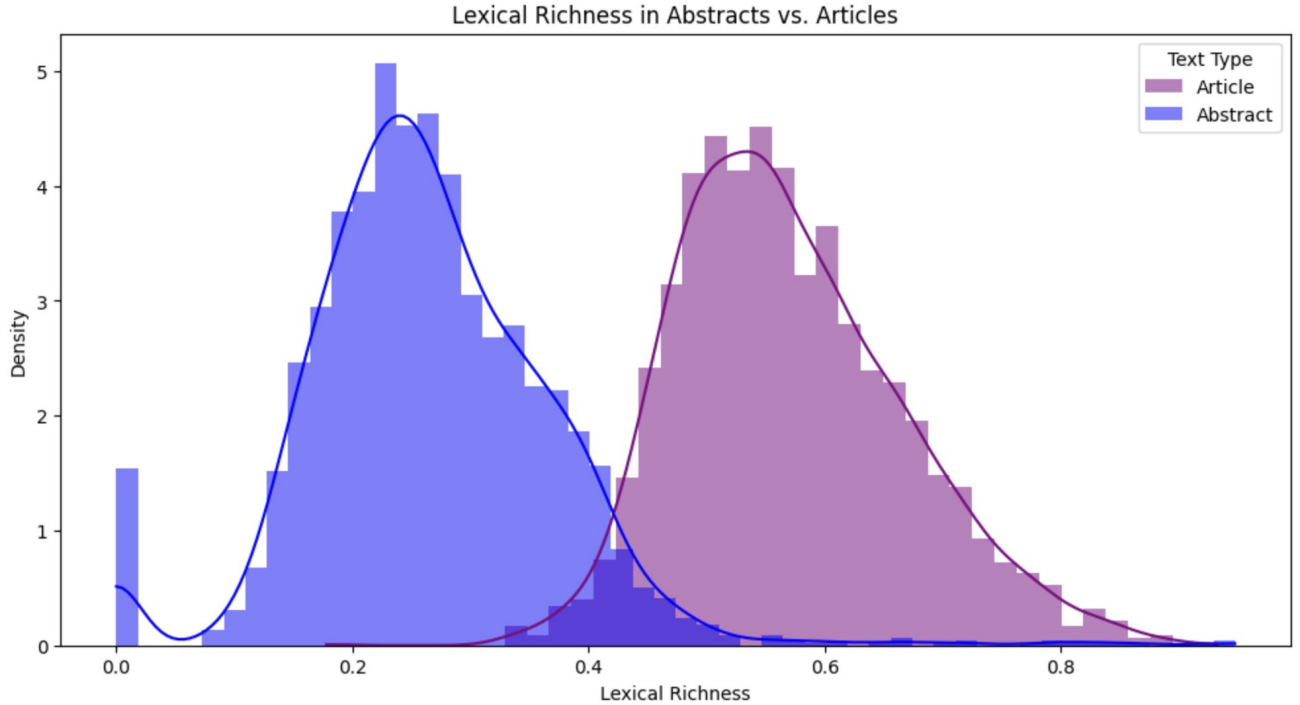
4. Most common words in the Abstract.

Exploratory Analysis



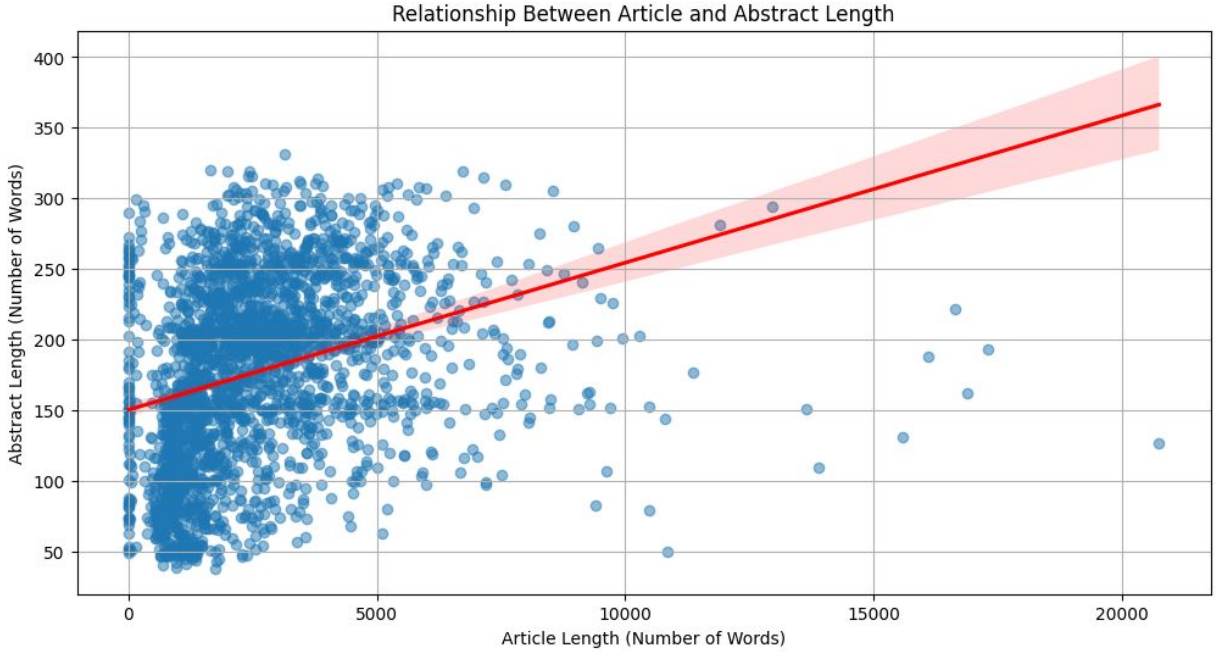
Hypothesis

1: The **lexical diversity** in the abstracts is significantly **lower** than in the full articles due to the concise and focused nature of abstracts.



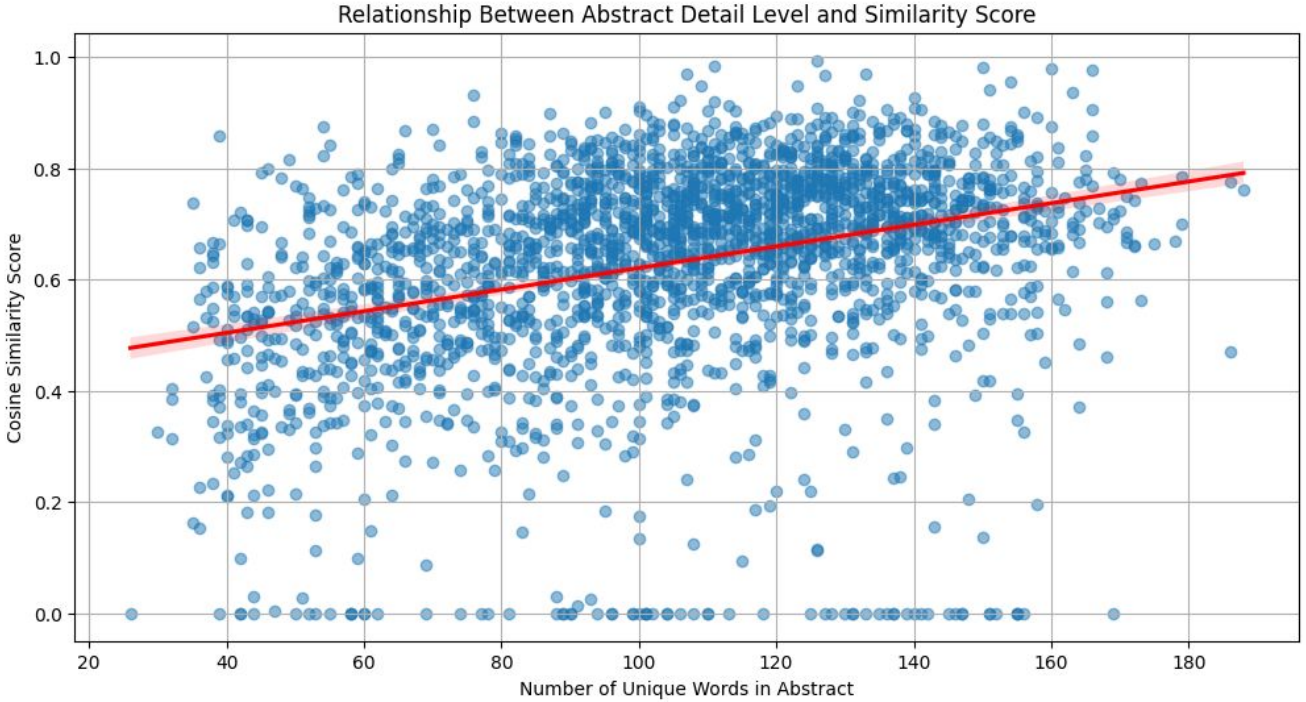
Hypothesis

2: There is a **positive correlation** between the **length of the article** and the **length of its abstract**.



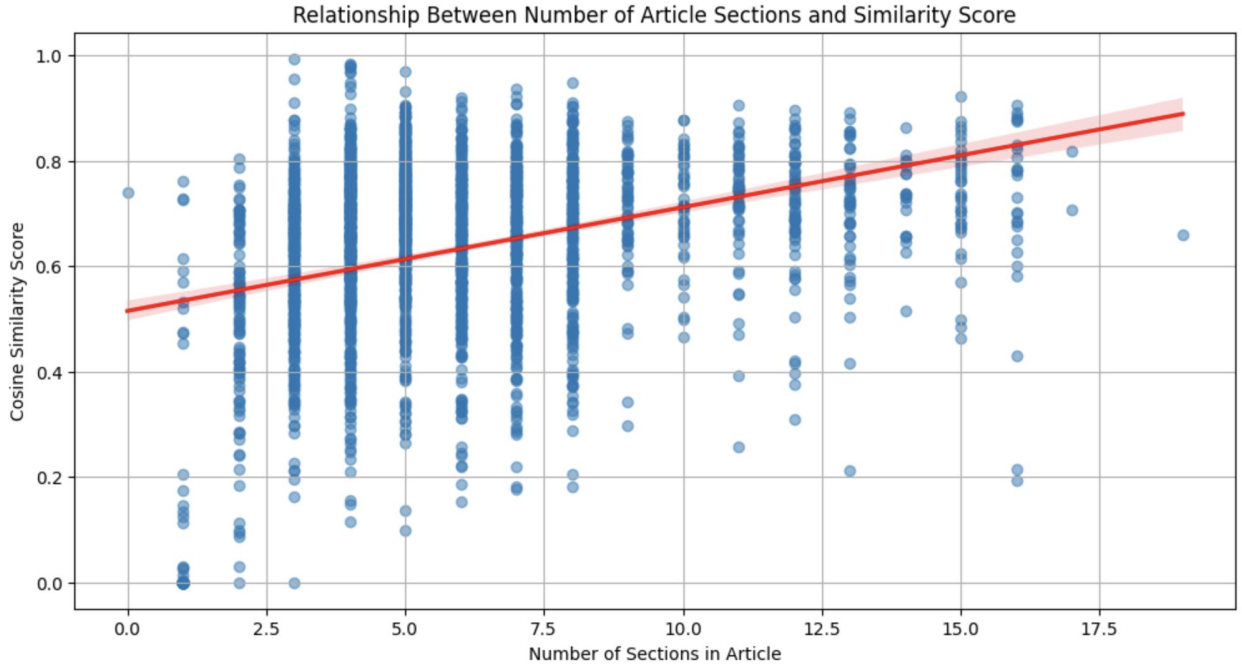
Hypothesis

3: The **detail level of abstracts** (measured by the variety of unique words) have **greater the similarity score** with the article (detailed abstracts are more closely aligned with the full content of the article).

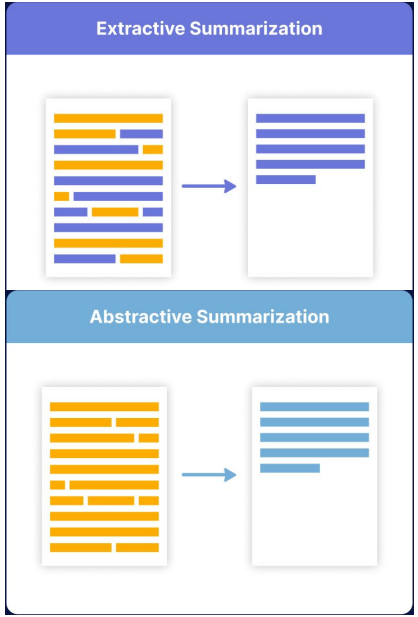


Hypothesis

4: Articles with **more sections** have a **lower similarity** score between their abstracts and articles.



Models for Text Summarization



Source:
https://www.abstractivehealth.com/extractive-vs-abstractive-summarization-in-healthcare

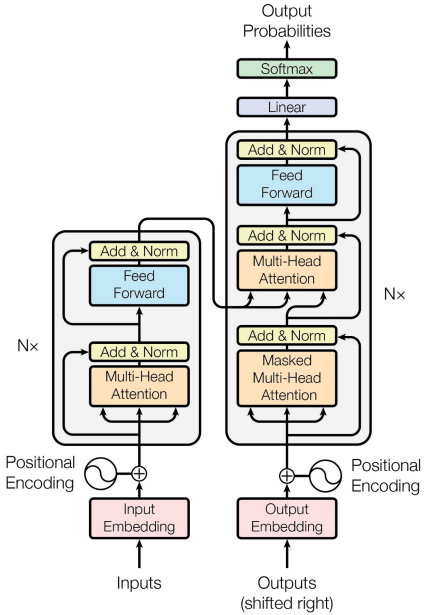
Extractive

- Select and combine essential sentences & phrases
- Summary is subset of original
- Simple algorithms: Frequency analysis, TF-IDF, LexRank

Abstractive

- Create new phrases & sentences
- May contain information not present in original data
- Complex algorithms: LSTM, Transformers, etc

Transformers



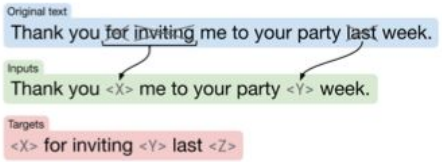
- Excels in Seq-2-Seq tasks
- Positional encoding along with self attention is used to enhance contextual information
- How they work:
  - Input text is tokenized and embedded into vectors.
  - Encoder processes the input through multiple layers using self-attention
  - Decoder then generates the output text, one word at a time using both self-attention and encoder-decoder attention

Source:
https://doi.org/10.48550/arXiv.1706.03762

Model Comparison

T5 Small Model

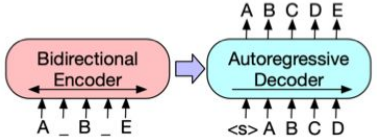
- Employs an encoder-decoder architecture designed to handle a "text-to-text" format
- Generally has around 60 million parameters



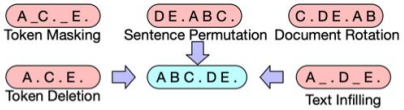
- Efficiently handles a wide range of tasks from translation to summarization by framing them all as text generation problems.

BART Base Model

- Also uses an encoder-decoder architecture



- Pre-trained primarily through a denoising task,



- Typically features around 110 million parameters

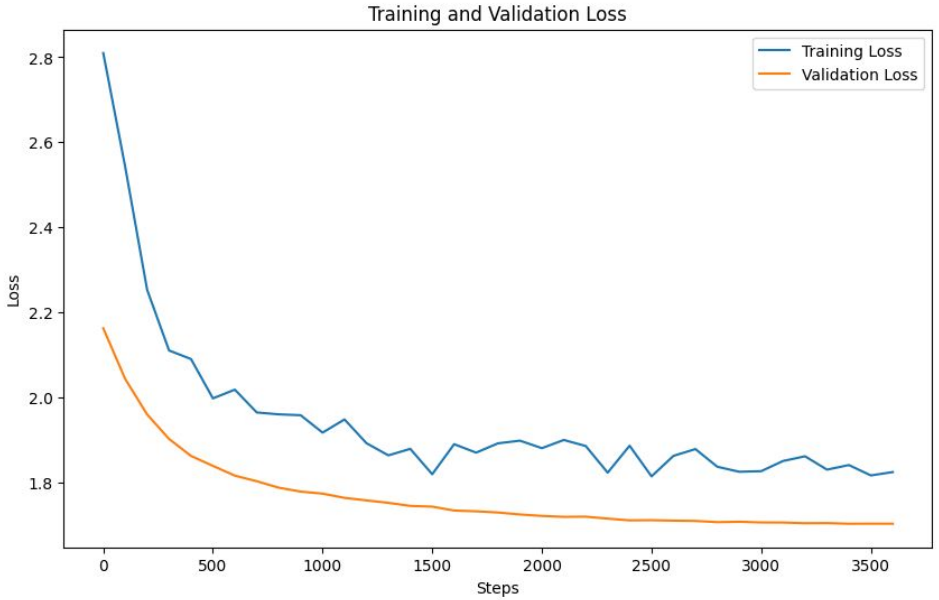
Image Source: https://www.cs.utexas.edu/~gdurrett/courses/fa2021/lectures/lec24-4pp.pdf

T5 Small Model

Data: 10% Train + 15% Val. + Test

Model Hyperparameters

```
num_train_epochs=3,
per_device_train_batch_size=8,
per_device_eval_batch_size=4,
warmup_steps=500,
weight_decay=0.01,
logging_steps=100,
save_steps=100,
eval_steps = 100,
save_total_limit = 3,
load_best_model_at_end=True,
evaluation_strategy = IntervalStrategy.STEPS,
metric_for_best_model="eval_loss",
greater_is_better=False
```

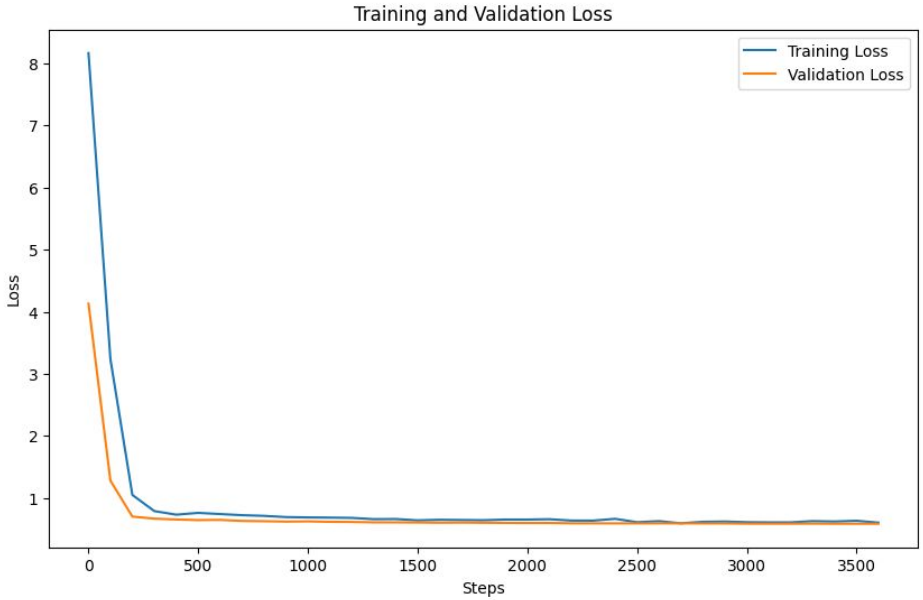


BART Base Model

Data: 10% Train + 15% Val. + Test

Model Hyperparameters

```
num_train_epochs=3,
per_device_train_batch_size=8,
per_device_eval_batch_size=4,
warmup_steps=500,
weight_decay=0.01,
logging_steps=100,
save_steps=100,
eval_steps = 100,
save_total_limit = 3,
load_best_model_at_end=True,
evaluation_strategy = IntervalStrategy.STEPS,
metric_for_best_model="eval_loss",
greater_is_better=False
```



Sample

Original Article

great progress has been made in neonatal care over the last few decades , reflected by improving survival rates and clinical outcomes of preterm infants . despite these advances , 45 years after its first description , bronchopulmonary dysplasia (bpd) remains a major complication of premature birth……pathogenesis of newborn lung diseases on grounds of pulmonary immaturity

(2670 words)

Given Summary

background . bronchopulmonary dysplasia (bpd) is the most common chronic lung disease……conclusion . rs1966265 in fgf receptor 4 is a possible genetic key variant in alveolar diseases of preterm newborns

(230 words)

T5 Summary

bronchopulmonary dysplasia (bpd) remains a major complication of premature birth. despite these advances, 45 years after its first description, the old bpd is mainly affected modestly premature newborns suffering from respiratory distress and therefore mechanically ventilated with high levels of supplemental oxygen. despite this advances, 45 years after its first description, prenatal maternal use of glucocorticoids, improved nutrition, and ventilator strategies the clinical course and pathology of bronchopulmonary dysplasia isoforms are important determinants for alveolarization.

(77 words)

BART Summary

background and objectives : bronchopulmonary dysplasia (bpd) remains a major complication of premature birth, causing ongoing morbidity and mortality. the purpose of this study was to identify genetic risk factors in a caucasian population of premature newborn with bpd and rds.materials and methods : we genotyped 27 polymorphisms within fourteen candidate genes for bpd : mmp-1, -2, -9, -12, and -16, fibroblast growth factor (fgf), -3, -4, -7, -18, and matrix metalloproteinase (mmp -9), which are important determinants for alveolarization in developing lungs and to protect from bpd.results:a total of 27 genotypes of bpd were identified, with a mean age ranging from 2 to 28 weeks of gestation ($p < 0.001$). of the genotypes, there was a significant association between mmp2 and -9 ($p = 0.002$) and bpd. in addition, there was no association between bpd or rds.

(145 words)

Evaluation Metrics

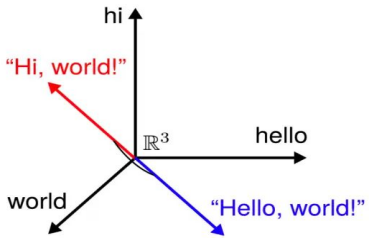
BLEU Score
Measures the similarity between machine-generated text and human-generated reference translations, based on n-gram overlap.
Doesn't capture fluency or grammatical correctness, sensitive to the choice of reference translations.

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

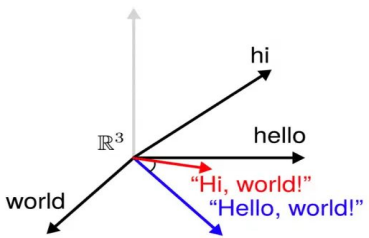
ROUGE Score
Measures the overlap between machine-generated text and human-generated summaries, based on n-gram recall and precision.
Sensitive to the choice of reference summaries, may not capture semantic similarity.

$$\frac{\text{number of n-grams found in model and reference}}{\text{number of n-grams in reference}}$$

Evaluation Metrics



Cosine Similarity



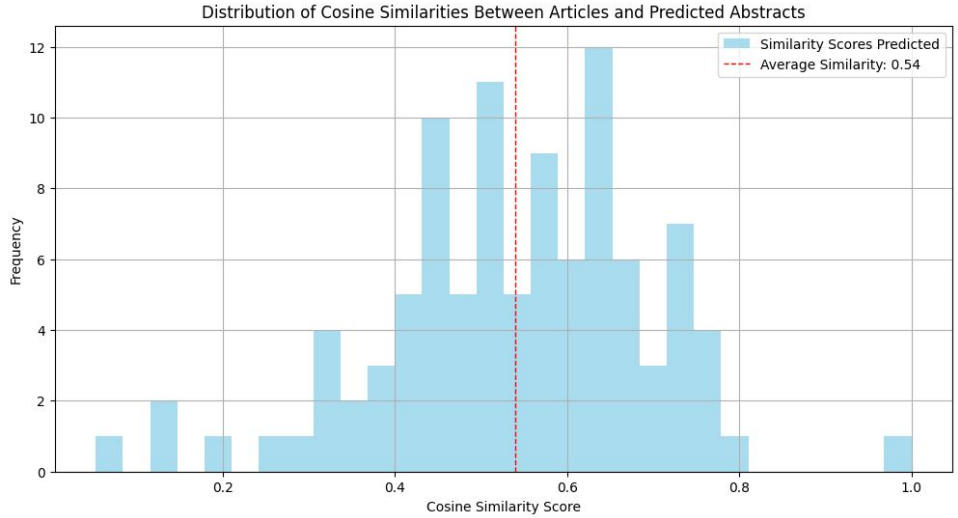
Soft Cosine Measure

Source: https://github.com/RaRe-Technologies/gensim/blob/develop/docs/notebooks/soft_cosine_tutorial.ipynb

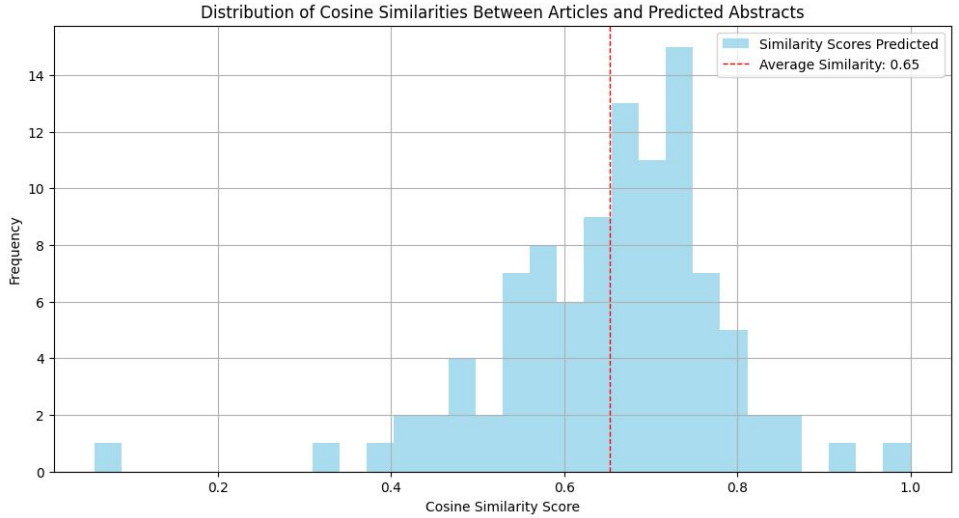
Cosine Similarity

- Cosine similarity is a mathematical metric used to measure the similarity between two vectors in a multi-dimensional space, particularly in high-dimensional spaces, by calculating the cosine of the angle between them.

Evaluation Metrics



T5

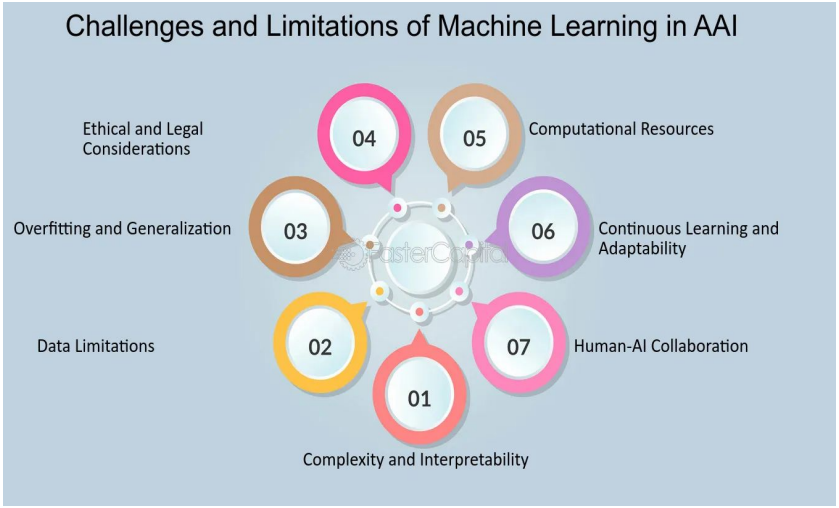


BART

Evaluation Metrics

Score/Model	Precision	Recall	F1	Score
Rouge-1 (T5)	0.41372	0.24552	0.29115	
Rouge-1 (BART)	0.46875	0.30029	0.34293	
BLEU (T5)				0.09512
BLEU (BART)				0.09538
BERT (T5)	0.83440	0.80646	0.81987	
BERT (BART)	0.85203	0.83545	0.84339	

Resource Limitation

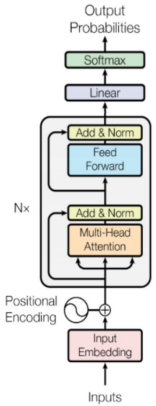


- Most of the limitations were computational in nature
- Platform with GPU access helped but had dataset limitations
- Larger models such as LongT5 & Pegasus resulted in “**RuntimeError: CUDA Out of memory**”
- Tried making a model from scratch but that did not work out

Conclusion

- Model Performance: The fine-tuned models showed significant improvements in generating coherent and contextually accurate summaries compared to their pre-trained versions.
- Data Adaptability: Fine-tuning allowed the models to adapt to specific characteristics of our dataset, including domain-specific language and stylistic nuances, which are often challenging for generic pre-trained models to capture.
- Computational Efficiency: Adjustments in the training process, including hyperparameter tuning and training epoch adjustments, optimized computational resources, reducing both training time and costs without compromising output quality.
- Practical Applications: The project has paved the way for practical applications in real-world scenarios where efficient and accurate summarization of large volumes of text is required, such as in legal document review, medical records summarization, and content curation for media.

Limitations & Future Work



Input:

Output: ⟶ Summary

- Our project was limited to around (10000*300 = 3 million tokens) and contained vocabulary mostly from medical articles.
- Although much work has been done in the fields of Text Summarization, better training of models can lead to increments in efficiency.
- We would like to make a Transformer model from scratch using JAX instead of PyTorch.

Thank You

Questions ?