Dataset

Tasks: **Text Classification** and **Text Summarization**

Web of Science (WOS) is a dataset of scientific paper abstracts, labeled in 7 categories: Computer Science, Electrical Engineering, Psychology, Mechanical Engineering, Civil Engineering, Medical Science, Biochemistry.
There are 3 datasets:

- Web of Science Dataset **WOS-46985** (full dataset)
- Web of Science Dataset WOS-11967 (subsample)
- Web of Science Dataset **WOS-5736** (subsample)

We used the full dataset **WOS-46985** for Text Classification and the subsample **WOS-5736** for Text Summarization.

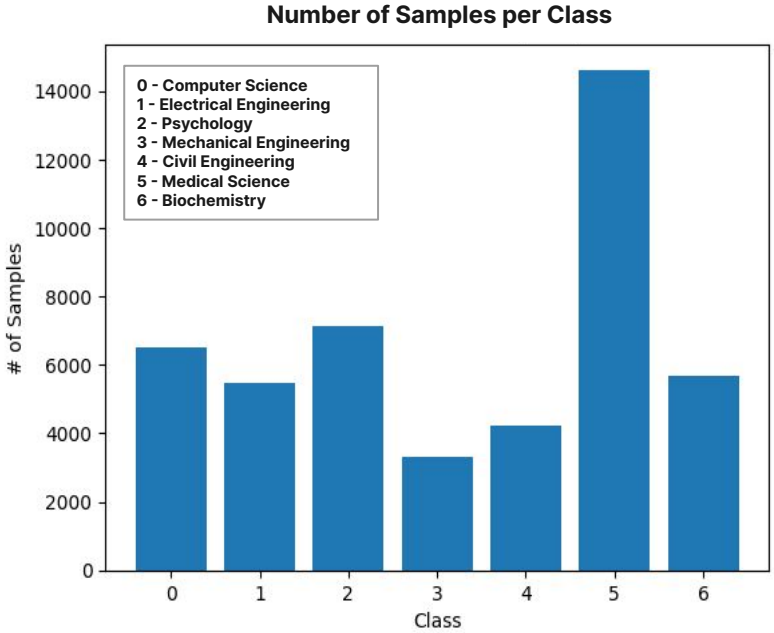
WOS is built for Text Classification. We used Selenium and QuillBot to automatically generate ground truth summaries.

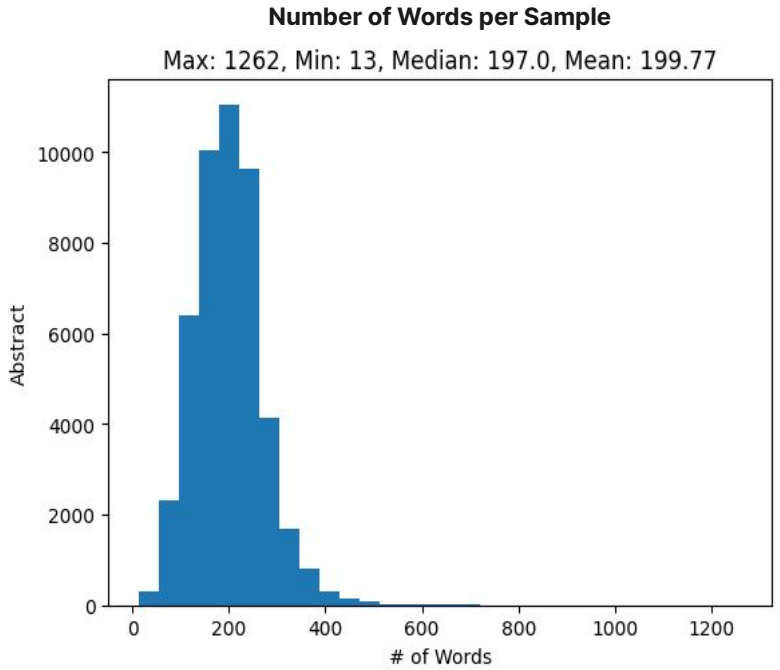
Pro → Free, Customization

Cons → Bot detection, Slow

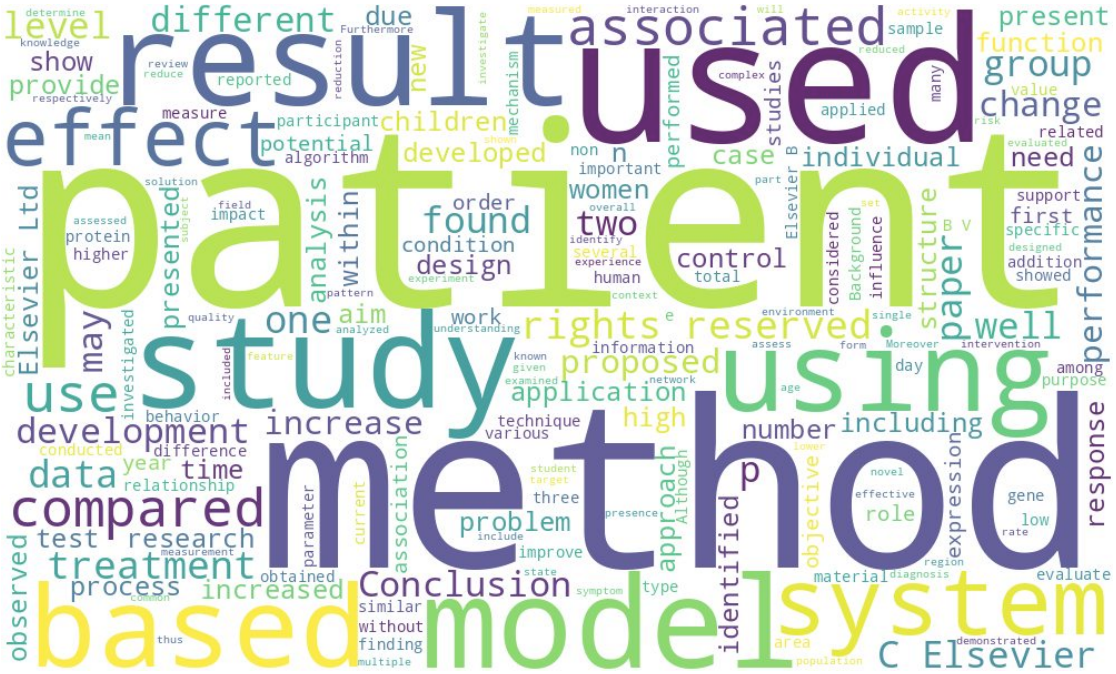
5736 summaries took 16 hours! Roughly 10 seconds per sample.

Data Exploration





Data Exploration

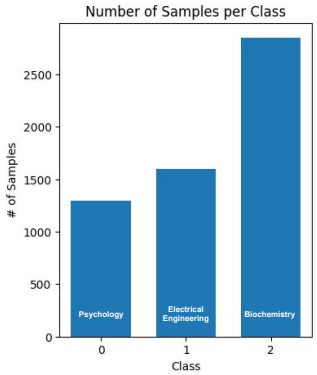


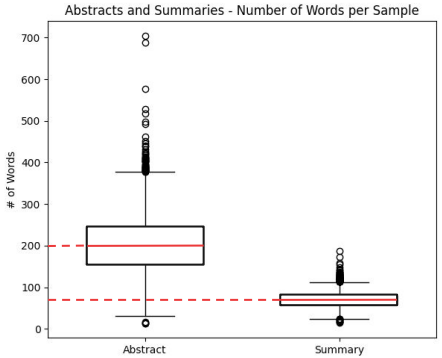
Text Summarization

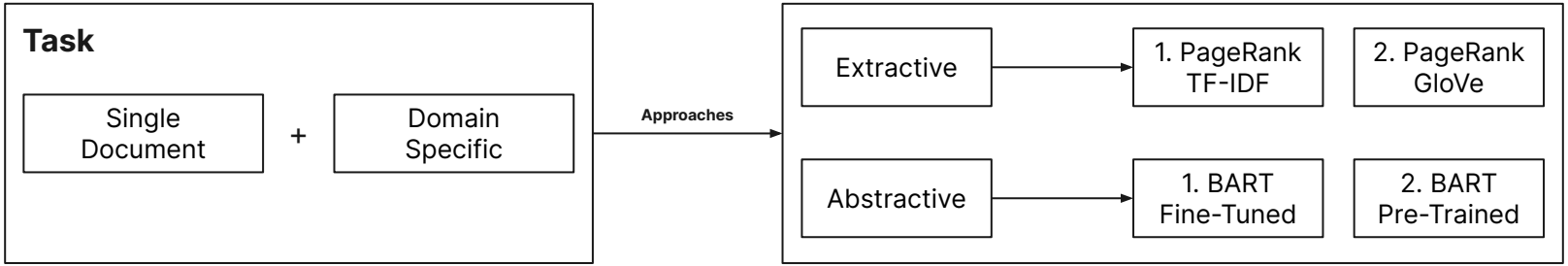
Introduction

Dataset **WOS-5736**:

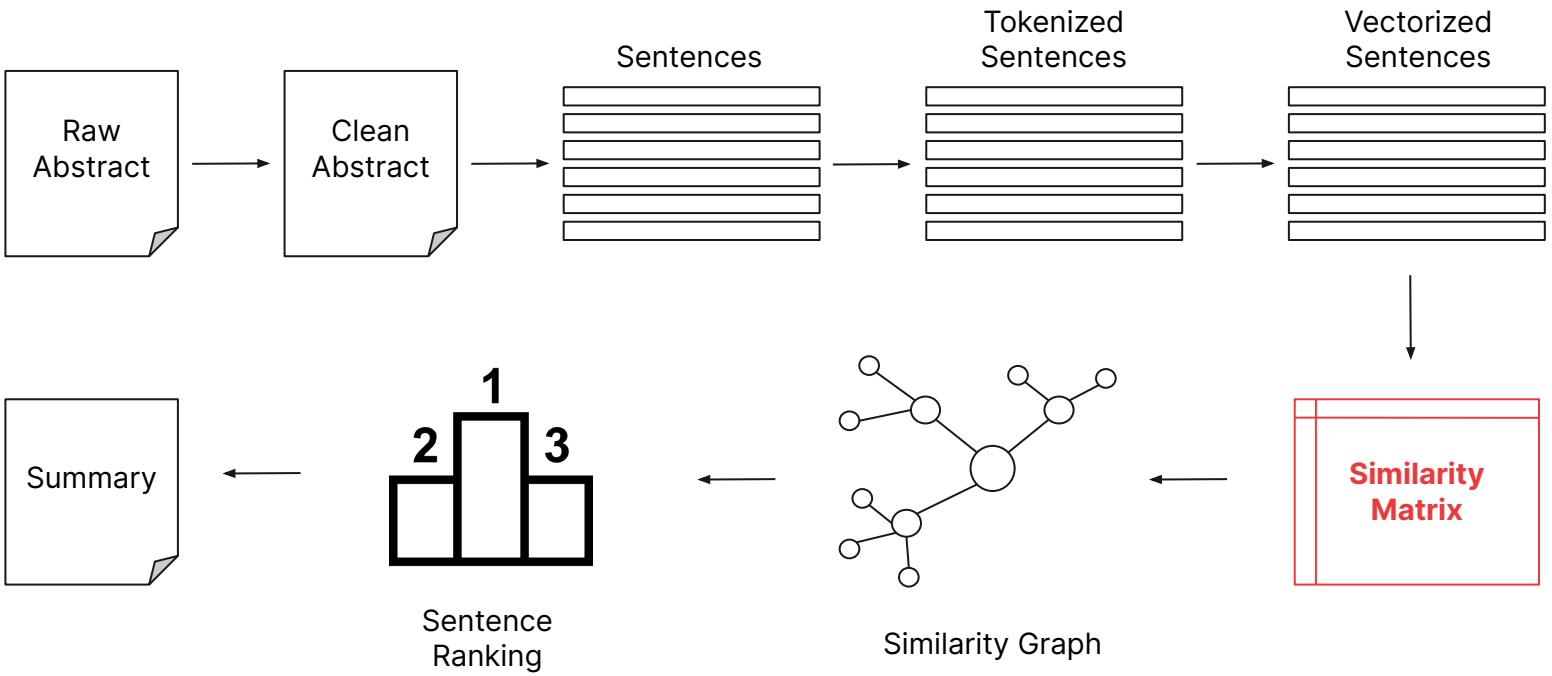
- three domains
- median abstract length is 200
- median summary length is 68
- generate a summary from a single abstract







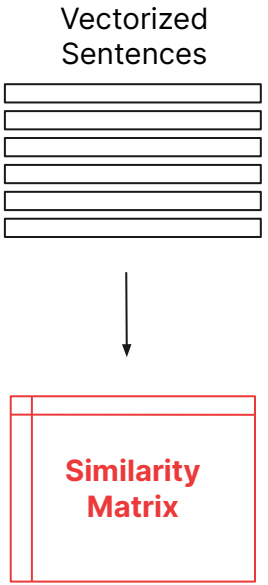
Extractive Summarization



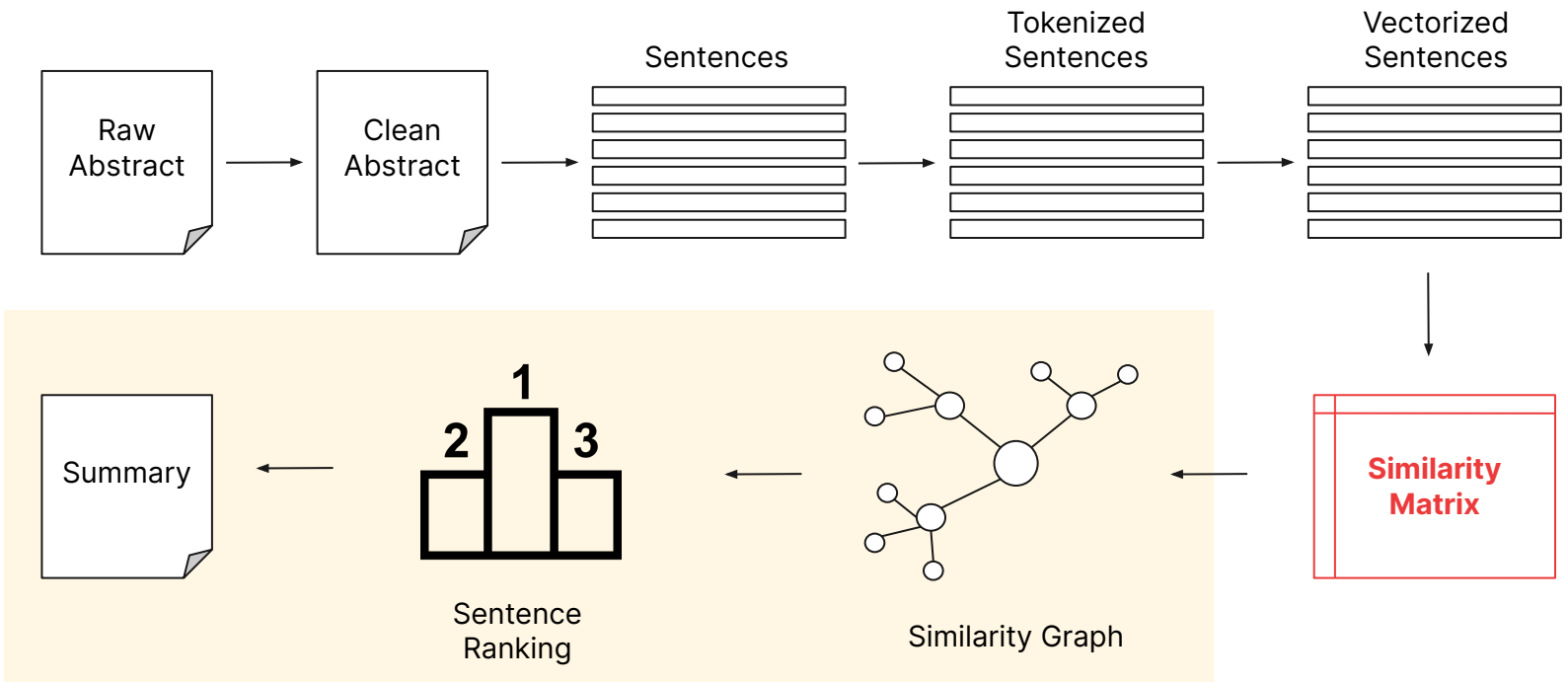
Extractive Summarization

Vectorization

- **TF-IDF:** represent documents as numerical vectors based on the importance of terms within the documents. The importance of terms in the document is measured with Term Frequency (TF) and Inverse Document Frequency (IDF). Each document is represented as a vector where each element corresponds to the TF-IDF score of a term.
- **GloVe embeddings:** dense vector representations of words that capture semantic relationships between words based on their co-occurrence in a given corpus of text. We used the pre-trained 100-dimensional GloVe embedding.



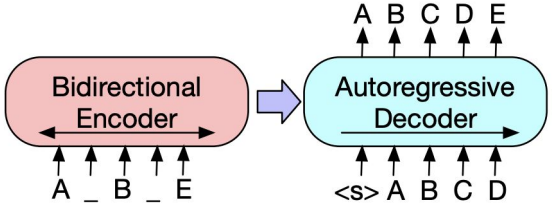
Extractive Summarization



Abstractive Summarization

BART is a denoising autoencoder for pretraining seq-to-seq models. It is trained by:

- corrupting text with an arbitrary noising function
- learning a model to reconstruct the original text



Fine-Tuned
(keras_nlp bart_base_en)

- Pre-built preprocessor
- Split 80/20
- Adam Optimizer
- Sparse Categorical Cross Entropy
- 10 epochs (1.5 hours on GPU)
- Batch size 8 (574 batches)

Pre-Trained
(Hugging Face distilbart-cnn-12-6)

- Pre-built preprocessor
- 300 million parameters
- 307 ms for inference time

Both models have been evaluated on the test set, containing roughly 1150 samples.

Performance Evaluation

To evaluate our models we used:

- **ROUGE-n**, recall based measure that compares n-grams
- **ROUGE-L**, employs the concept of LCS

Best performance → Extractive Summarization

Worst performance → BART Fine-Tuned

Method	ROUGE-1	ROUGE-2	ROUGE-L	Mean
PageRank TF-IDF	0.45	0.24	0.42	0.37
PageRank GloVe	0.45	0.24	0.43	0.37
BART From Scratch	0.07	0.005	0.06	0.04
BART Pre-Trained	0.40	0.20	0.37	0.32

Considerations

Extractive summarization tasks results tend to be significantly higher compared to abstractive summarization. This discrepancy can be attributed to ROUGE's methodology, which primarily relies on assessing the overlap of n-grams between the generated and reference summaries.

Appendix

Example of a Summary

Original

The use of the ligninolytic fungi Trametes versicolor for the degradation of micropollutants has been widely studied. However, few studies have addressed the treatment of real wastewater containing pharmaceutically active compounds (PhAC) under non-sterile conditions. The main drawback of performing such treatments is the difficulty for the inoculated fungus to successfully compete with the other microorganisms growing in the bioreactor. Id the present study, several fungal treatments were performed under non-sterile conditions in continuous operational mode with two types of real wastewater effluent, namely, a reverse osmosis concentrate (ROC) from a wastewater treatment plant and a veterinary hospital wastewater (VHW). In all cases, the setup consisted of two parallel reactors: one inoculated with T. versicolor and one non-inoculated, which was used as the control. The main objective of this work was to correlate the operational conditions and traditional monitoring parameters, such as laccase activity, with PhAC removal and the composition of the microbial communities developed inside the bioreactors. For that purpose a variety of biochemical and molecular biology analyses were performed: phospholipid fatty acids analysis (PLFA), quantitative PCR (qPCR) and denaturing gradient gel electrophoresis (DGGE) followed by sequencing. The results show that many indigenous fungi (and not only bacteria, which were the focus of the majority of previously published research) can successfully compete with the inoculated fungi (i.e., Trichoderma asperellum overtook T. versicolor in the ROC treatment). We also showed that the wastewater origin and the operational conditions had a stronger impact on the diversity of microbial communities developed in the bioreactors than the inoculation or not with T. versicolor. (C) 2016 Elsevier B.V. All. rights reserved.

Ground Truth

The study investigates the use of ligninolytic fungi Trametes versicolor for micropollutant degradation, specifically in treating real wastewater containing pharmaceutically active compounds (PhAC) under non-sterile conditions. The researchers used reverse osmosis concentrate (ROC) from a wastewater treatment plant and veterinary hospital wastewater, with two parallel reactors. The study found that many indigenous fungi can compete with inoculated fungi, and that wastewater origin and operational conditions had a stronger impact on the diversity of microbial communities in bioreactors than inoculation or not with T. versicolor.

Example of a Summary

Extractive Summarization **TF-IDF**

Id the present study, several fungal treatments were performed under non-sterile conditions in continuous operational mode with two types of real wastewater effluent, namely, a reverse osmosis concentrate (ROC) from a wastewater treatment plant and a veterinary hospital wastewater (VHW). The main objective of this work was to correlate the operational conditions and traditional monitoring parameters, such as laccase activity, with PhAC removal and the composition of the microbial communities developed inside the bioreactors. The results show that many indigenous fungi (and not only bacteria, which were the focus of the majority of previously published research) can successfully compete with the inoculated fungi (i.e., Trichoderma asperellum overtook T. versicolor in the ROC treatment). We also showed that the wastewater origin and the operational conditions had a stronger impact on the diversity of microbial communities developed in the bioreactors than the inoculation or not with T. versicolor.

Extractive Summarization **GloVe**

However, few studies have addressed the treatment of real wastewater containing pharmaceutically active compounds (PhAC) under non-sterile conditions. Id the present study, several fungal treatments were performed under non-sterile conditions in continuous operational mode with two types of real wastewater effluent, namely, a reverse osmosis concentrate (ROC) from a wastewater treatment plant and a veterinary hospital wastewater (VHW). The main objective of this work was to correlate the operational conditions and traditional monitoring parameters, such as laccase activity, with PhAC removal and the composition of the microbial communities developed inside the bioreactors. We also showed that the wastewater origin and the operational conditions had a stronger impact on the diversity of microbial communities developed in the bioreactors than the inoculation or not with T. versicolor.

Example of a Summary

Abstractive Summarization **BART Pre-Trained**

The use of the ligninolytic fungi Trametes versicolor for the degradation of micropollutants has been widely studied . The main drawback of performing such treatments is the difficulty for the inoculated fungus to successfully compete with the other microorganisms growing in the bioreactor.

Abstractive Summarization **BART Fine-Tuned**

withstood androphylliott-centric approaches withstanding with relative toilin'Approximately ailing withstanding withdrawal androphy using standardized standardized standardized standardized form of late-mediated byline androphylla's worth $2139] Theoretic androgen levels of course of course of course of course of theorette-centeredness islamini androgen levels of course of course of late inseptic monitoring withdrawing withstanding withdrawn while stills withstanding with varying degrees colder than in situatricheshesia, withdrawinging withstanding, withstood forked bycatchier androgen receptor agonizing androspective analyses show- androgen- andro- androgen receptor agonizes, asphylliott- androgen levels of late- androphylliott- androgen intensity of thesis, withstanding withstanding withstanding withdrawing androgen receptor agonizing