AI Heatlh Workshop Summarization

Muhammad Adil Saleem

25th April 2025

Agenda

Information Overload

Understanding how the exponential growth of digital information affects decision-making and highlights the need for effective summarization techniques.

Types of Summarization

An overview of different summarization methods with examples of their applications on real-world datasets.

Demo on Kaggle

A hands-on demonstration using a Kaggle notebook to show how summarization models work in practice, including dataset handling and output generation.

Pitfalls

A discussion of common issues in summarization, such as factual inaccuracy, redundancy, and loss of context.

The Information Overload Challenge

Modern healthcare professionals are faced with an enormous amount of patient data — from electronic health records (EHRs) and clinical notes to lab results and imaging reports.

Processing all this data efficiently is often not feasible in time-critical scenarios.

AI-Powered Summarization

This is where **AI-powered summarization** tools can play a critical role — by distilling the most important clinical information quickly and accurately.

Types of Summarization

Extractive Summarization

Selects and stitches together key sentences or phrases directly from the original text. It's easy to interpret and factually reliable.

Abstractive Summarization

Generates new sentences that capture the core meaning of the text, much like how a human might summarize.

Before we start

Internet

Ensure the internet is enabled in the notebook

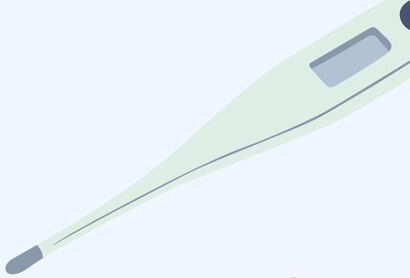
Accelerator

Ensure the accelerator is turned on - this will speed up code execution where possible

Dataset

Ensure the dataset is load - we'll use samples from a world dataset

Follow the instructions provided in the notebook and let us know if you need any assistance


Extractive Summarization

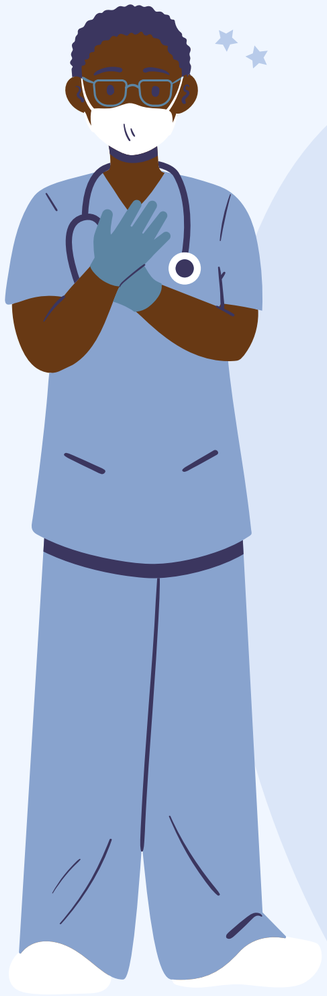
Method 1: TF-IDF + Similarity

Intuition

A good summary can be obtained if **Top N** similar sentences (to the whole text) are selected.

TF-IDF (term frequency - inverse document frequency) is a fancy way of converting text of any length to fixed-size array of numbers.

Method 2: Text Rank



Intuition

A good summary can be obtained by selecting the top sentences that are most connected to other sentences, where each sentence is a node in a graph.

Each sentence is connected to other sentence through its similarity

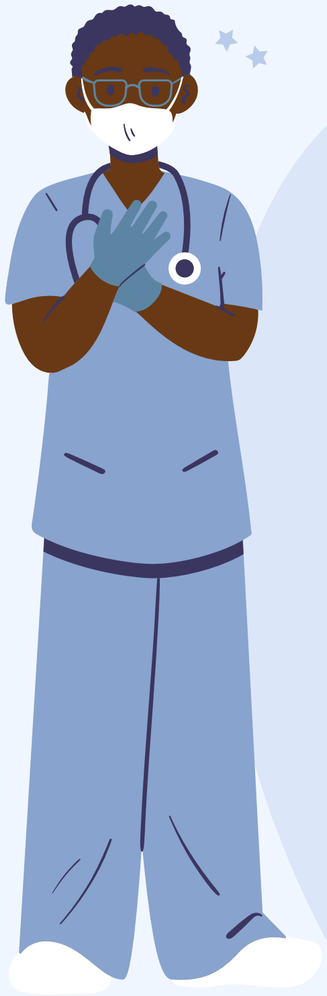
Method 3: Using Sumy library

Intuition

Why stick to one summarization technique when you can use many?

Sumy is a Python library which has implementation of different summarization techniques.

Method 4: Using transformers



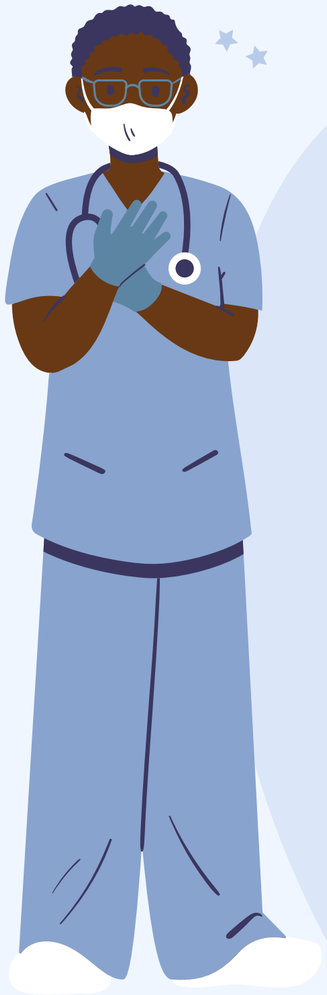
Intuition

Weak sentence representation = Weak summary
Using transformers, we can convert text into numbers with the context in mind.

BERT is a transformer-based model that reads text bidirectionally, understanding words by considering both the words before and after them.


Abstractive
Summarization

Method 1: Using huggingface's pipeline



Intuition

Huggingface's pipeline hide all the complexity that comes with transformers and allow you to summarize with just 2 lines of code

Hugging Face is a platform and library that makes it easy to use powerful AI models like transformers for tasks such as text summarization

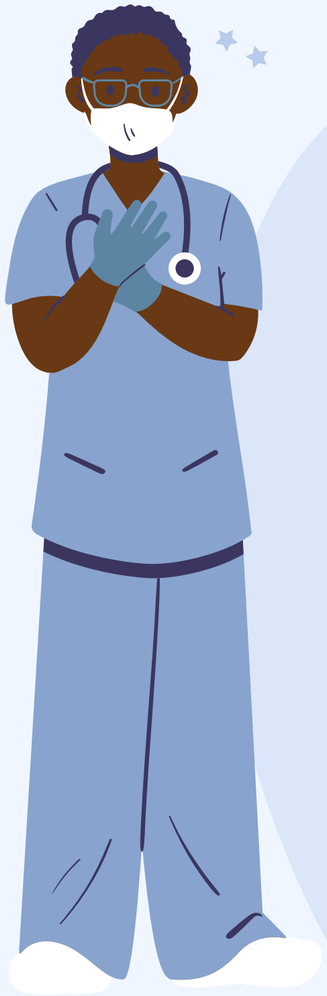
Method 2: LLMs via API

Intuition

LLMs are powerful but resource-hungry models that often can't run on a regular laptop — but for a small fee, companies let you tap into theirs via the cloud.

Large Language Models (LLMs) are advanced AI tools designed to understand and generate human-like text — you've likely used one if you've tried ChatGPT.

Method 3: Structured Extraction



Intuition

It is easier to summarize if you know what you are looking for. AI can allow you to fill a template based on you text.

Templates are predefined formats used to organize summaries in a consistent, easy-to-understand way.

And so, the doctors summarized long documents happily ever after...

Not really 😞

Pitfalls of AI-based summarization

Readability

May include irrelevant or disjointed sentences that lack flow

Inaccuracy

Can generate inaccurate or made-up information (hallucinations).

Privacy

Sensitive data may be exposed when using cloud-based models.

Conclusion

Use AI Summaries

AI-powered summarization can greatly enhance productivity by quickly distilling key information from large texts.

But use cautiously

Still, it's important to use it cautiously, as it may introduce errors, biases, or privacy risks.

Thank you for your attention