Covid19 Data Mining (CDM)

NLP Search Engine as a Service

Context

- ✓Our **CDM platform has been tested on a small datasets (<100 PDFs) for Covid19**; built on existing data mining models (BERT, RoBERTa, XLNet, etc.) with Python scripts running on MS Azure Cloud. We want to test the scale and execution speed of these models on an **HPC platform**

- ✓As we intend to use multiple of data mining frameworks on identified Covid19 datasets to **increase the reliability and trust from these AI results**, we intend to leverage the compute infrastructure of C-DAC

- ✓Our objective is to **deploy and test to run CDM as a service and leverage multiple data mining frameworks** on 40k+ Covid19 datasets to aid in medical community and act as a *Covid19's NLP Search Engine as a Service*

CDM - AI/ML + Cloud + REST API

AI/ML Models

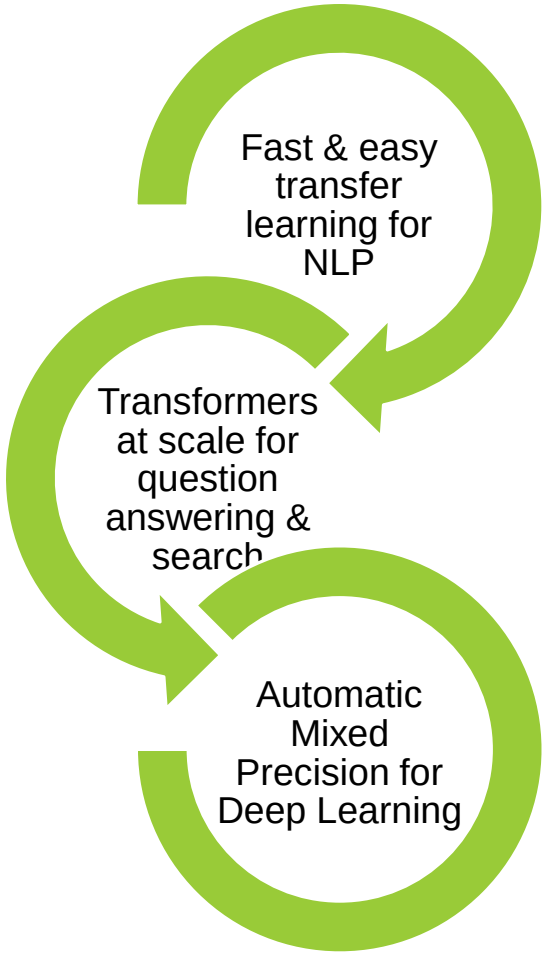
Utilize all latest transformer-based models (BERT, ALBERT, RoBERTa)

Data Mining Framework

Leverage multiple Data Mining Frameworks – FARM, Transformer, HayStack

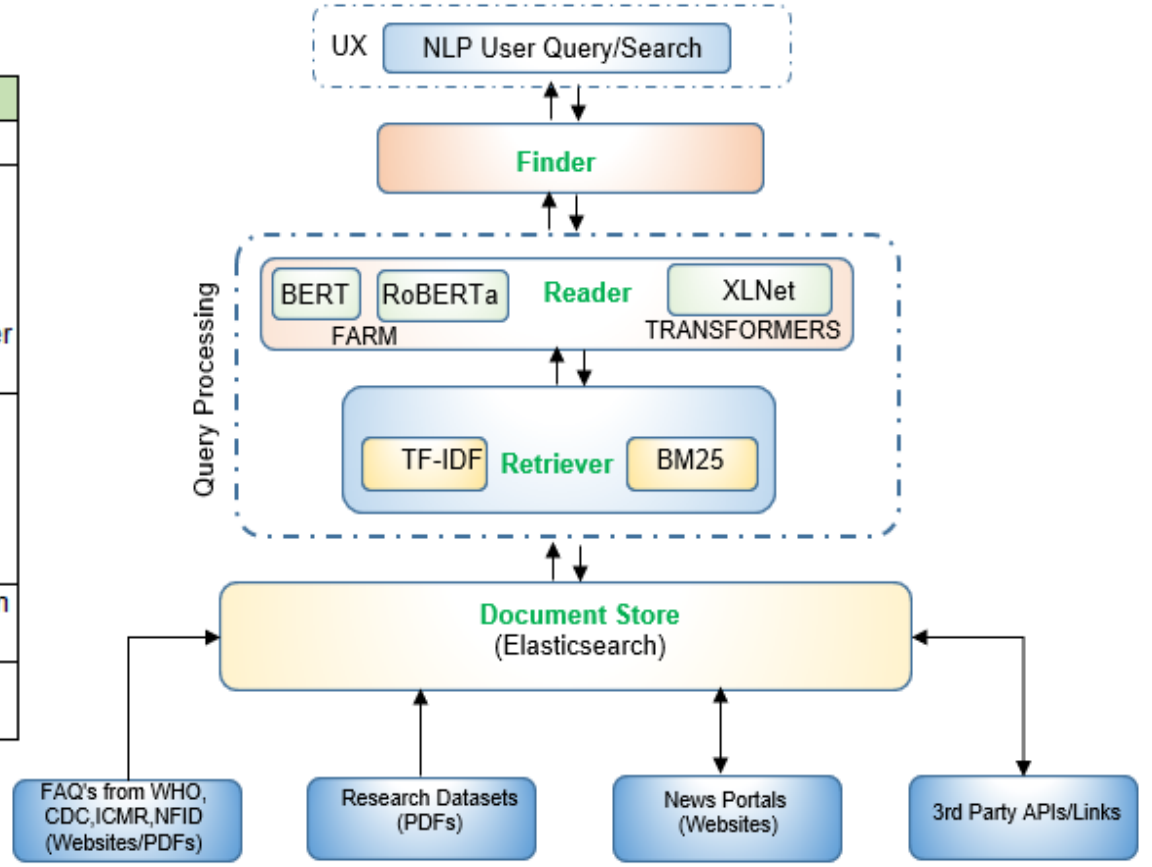
Production Ready

➢Production-ready deployments via Elasticsearch backend & REST API



CDM - Current Stack

CDM Component	CDM Component Description
Document Store	Elasticsearch Database storing the documents for our search
Retriever	Quick algorithms that identifies candidate passages from a large collection of documents Algorithms include TF-IDF or BM25, custom Elasticsearch queries, and embedding-based approaches The Retriever helps to narrow down the scope for Reader to smaller units of text where a given question could be answered
Reader	Neural model that reads through texts in detail to find an answer Use diverse models like BERT, RoBERTa or XLNet trained via FARM or Transformers on SQuAD like tasks The Reader takes multiple passages of text as input and returns top-n answers with corresponding confidence scores
Finder	Glues together a Reader and a Retriever as a pipeline to provide an intuitive Q&As interface
REST API	Exposes a simple API for running QA search, collecting feedback and monitoring requests



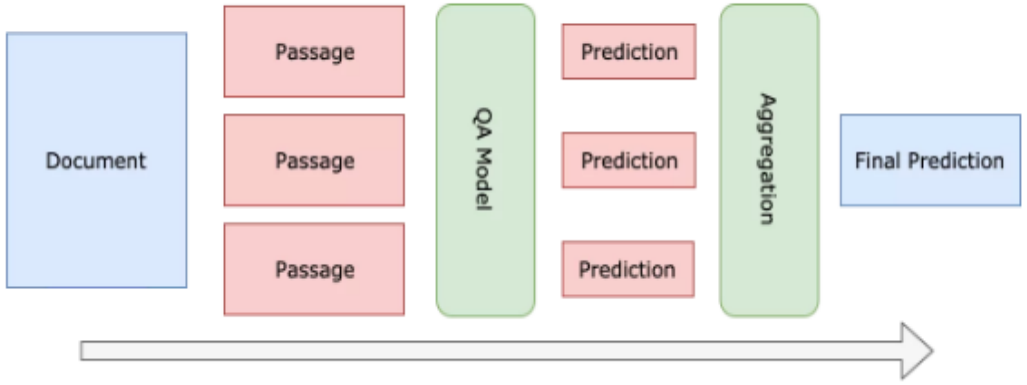
CDM - Core Features

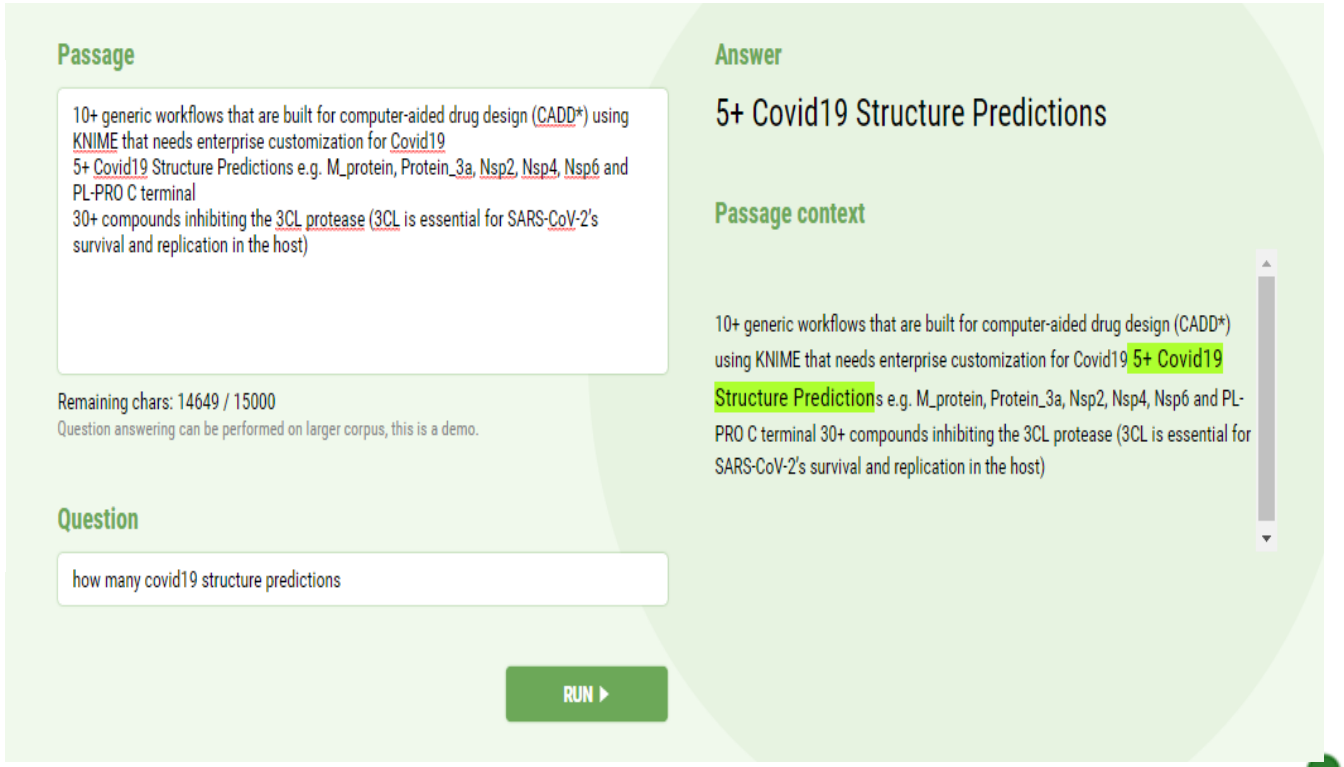
- **Powerful ML models**: Utilize all latest transformer-based models (BERT, ALBERT, RoBERTa)
- **Modular & future-proof**: Easily switch to newer models once they get published
- **Developer friendly**: Easy to debug, extend and modify
- **Scalable**: Production-ready deployments via Elasticsearch backend & REST API
- **Customizable**: Fine-tune models to your own domain & improve them continuously via user feedback

S #	Tasks	BERT	RoBERTa	XLNet	ALBERT	DistilBERT	XLMRoBERTa
1	Text classification	x	x	x	x	x	x
2	NER	x	x	x	x	x	x
3	Question Answering	x	x	x	x	x	x
4	Language Model Fine-tuning	x					
5	Text Regression	x	x	x	x	x	x
6	Multilabel Text classif.	x	x	x	x	x	x
7	Extracting embeddings	x	x	x	x	x	x
8	LM from scratch (beta)	x					
9	Text Pair Classification	x	x	x	x	x	x
10	Passage Ranking	x	x	x	x	x	x

CDM - Working Model

Document➔ Passage➔ QnAs➔ Prediction





CDM's Benefits

1

Plug-in to any Covid19 Datasets

Stack that can be easily integrated with based datasets for end users Q&As

2

Integrates with 5+ Data Mining Models

Deploy multiple Data Mining frameworks/ models (BERT, RoBERTa or XLNet)

3

Run Data Mining as a Service

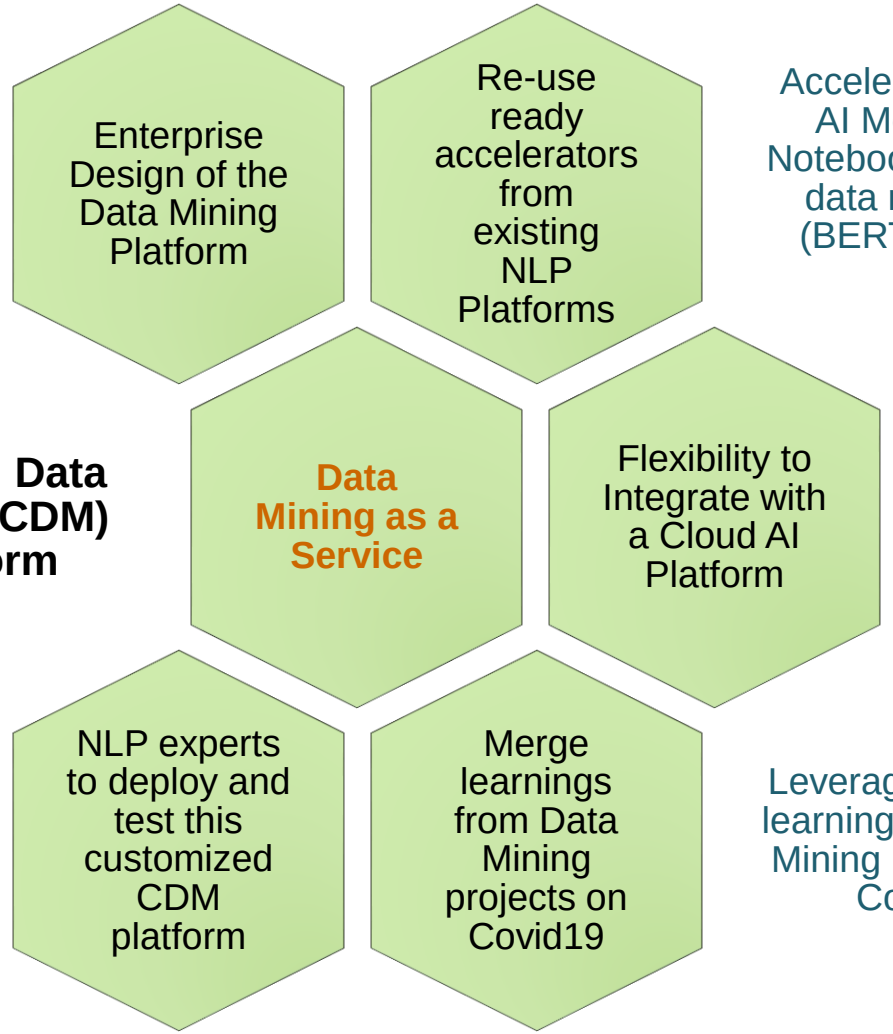
Data Mining as a Service on identified sources (PDFs, news sites, blogs, FAQs, APIs, etc.)

- ✓ Helps to deliver the **Proof of Concept (POC) in 2-3 weeks**
- ✓ Easily scales up for enterprise adoption by using the **industry best practices in Data Mining**
- ✓ Improves reliability by leveraging existing **Data Mining Models**

CDM's Enterprise Scale

* Computer Aided Drug Design/Discovery

Open and Enterprise Architecture based on multiple data mining models



Accelerators in form of AI Models, Python Notebooks from existing data mining models (BERT, RoBERTa or XLNet)

CDM can be easily extended and integrated with MS Azure/AWS/GCP

NLP experts to engage who have worked with diversified data mining models like BERT, RoBERTa or XLNet trained via FARM or Transformers

Leverage available learnings from Data Mining Projects on Covid19

Thank You

Greenojo provides Automation, Analytics and AI solutions to enterprise customers

For RFPs, Solutions and Sales/Partner enquiries, connect us at - *sales@greenojo.com*

Sales Offices
Houston, TX, USA |Burlington, ON, Canada |Dubai, UAE | Lagos, Nigeria

Delivery Offices - India
Bhubaneswar, Odisha |Hyderabad, Telangana | Trivandrum, Kerala