Authorship Attribution

Presented By:

Hazim Bukhari
Ahmed Aljmiai
Abdultawwab Safarji

Introduction

King Fahad library is looking for a way to help visitors find arabic authors based on the author's writing style and recommend books with similar writing style.

Methodology

EDA

Books per author, text length, etc… .

Preprocessing

Remove tashkeel, sentence length, etc… .

Building Models

transfer learning , built from scratch.

Model serving

Serve the model on the web.

Data Descriptions

Authors

80 unique Authors

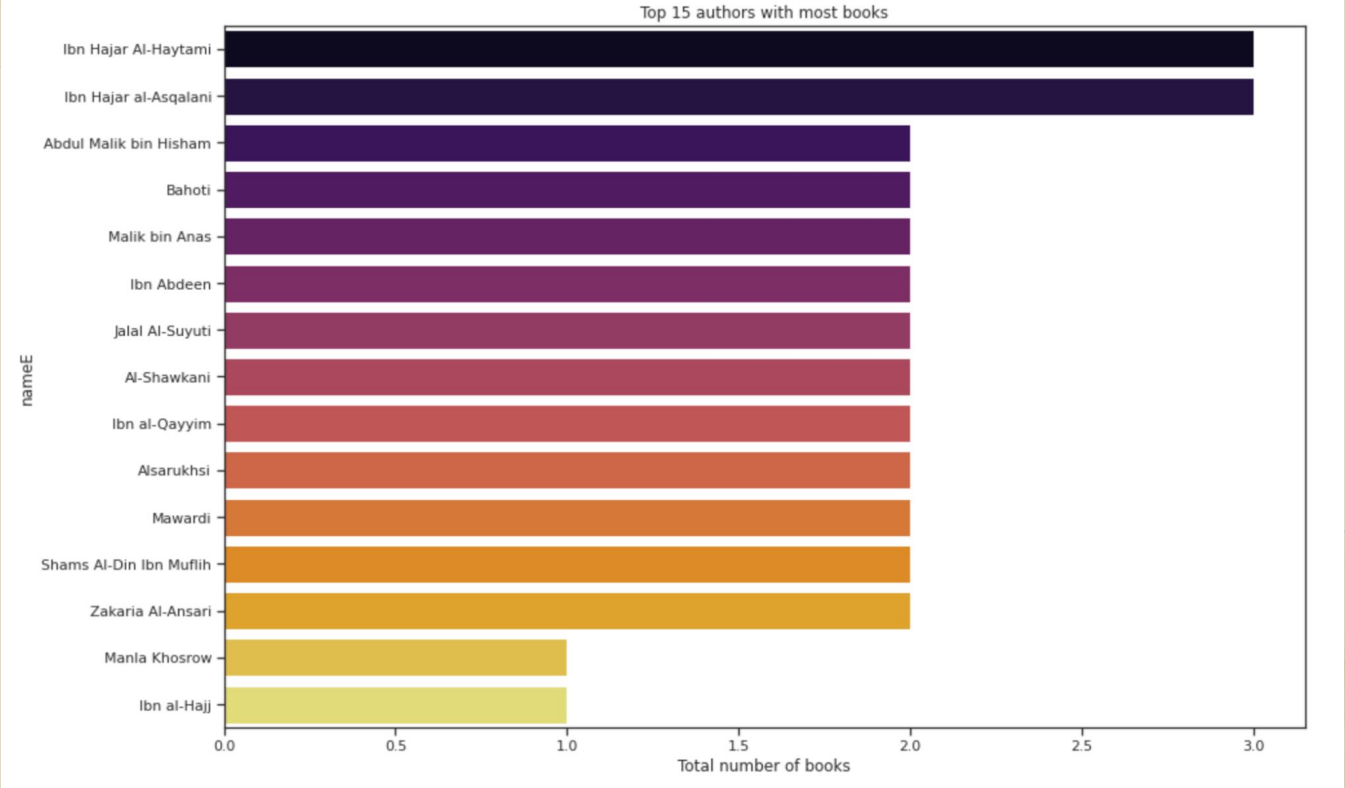
Books

97 Book

Data Source

Provided by SDAIA

EDA:


Top 15 authors with most books
Ibn Hajar Al-Haytami
Ibn Hajar al-Asqalani
Abdul Malik bin Hisham
Bahoti
Malik bin Anas
Ibn Abdeen
Jalal Al-Suyuti
Al-Shawkani
Ibn al-Qayyim
Alsarukhsi
Mawardi
Shams Al-Din Ibn Muflih
Zakaria Al-Ansari
Manla Khosrow
Ibn al-Hajj
nameE
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Total number of books

Challenges

Unbalanced Classes

Some authors have more books than others.

Training Time

One epoch takes up to an hour to complete.

Labeling

Finding the author of each book.

Corpus Size

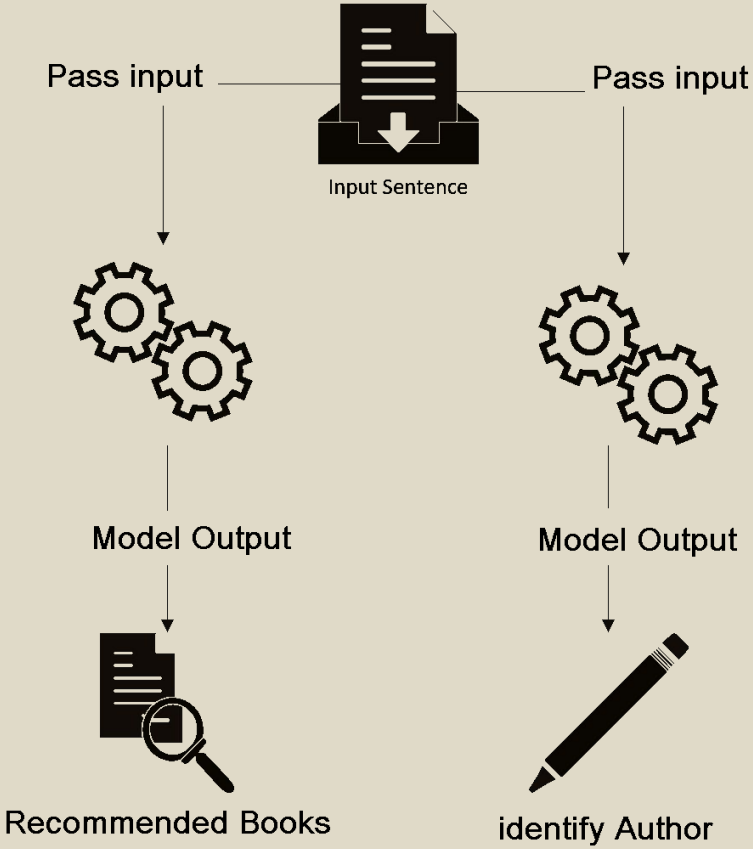
-exceeding 1gb.
-Min doc size ~30mb.

Distributed deep learning using TPUs:

- Reducing computing time of ~42gb per 32 step per epoch.
- Handling Big Data.
- Using bert transformers
- Google platforms: Google cloud platform and colab TPU back end.

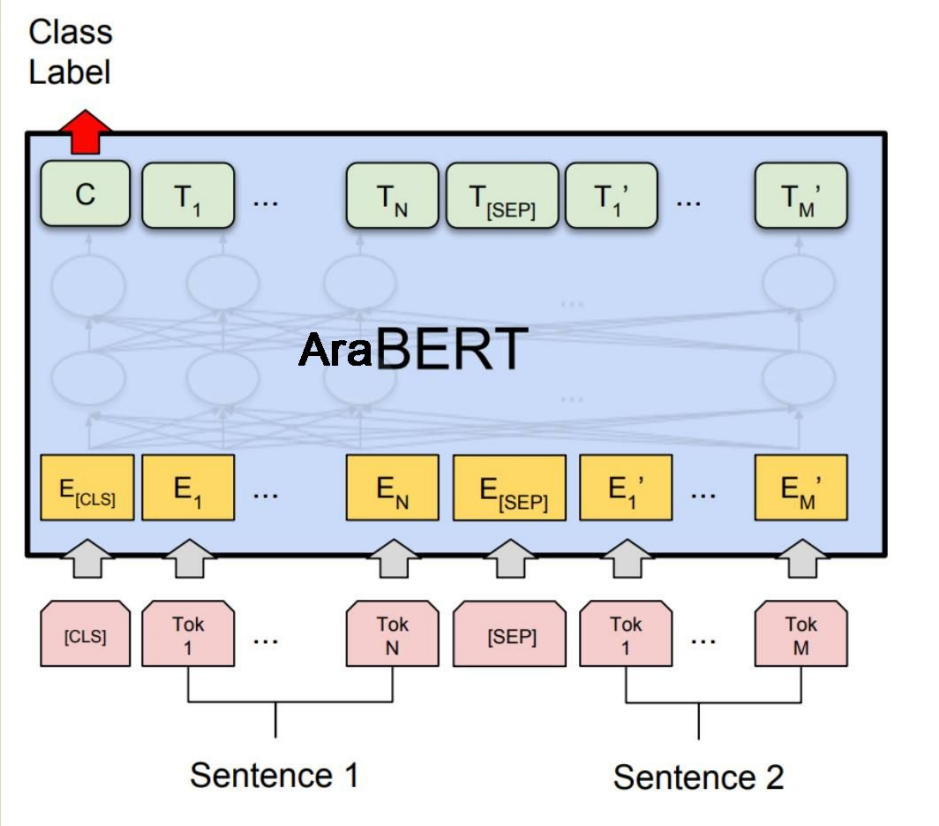
Processing unit	CPU	GPU	TPU
Time	Train time>10 hrs	4 hrs> Train time >3hrs	8min >Train time> 13 min

System Architecture:

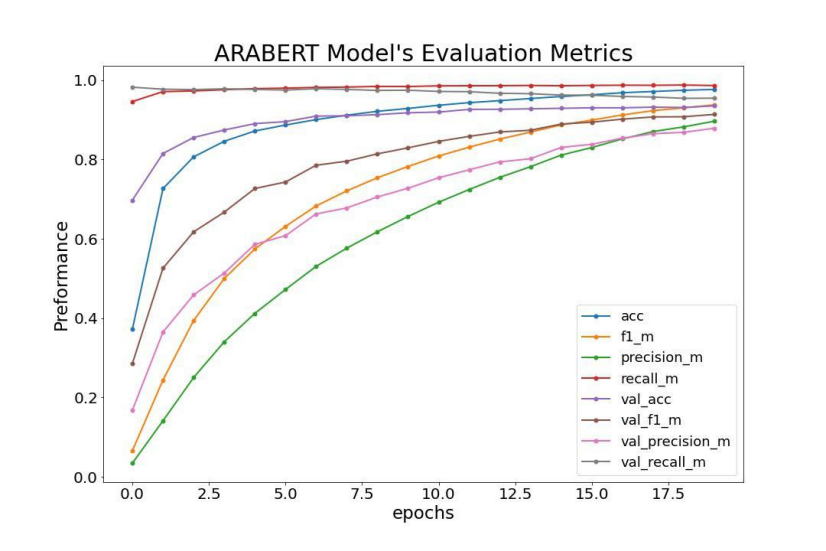


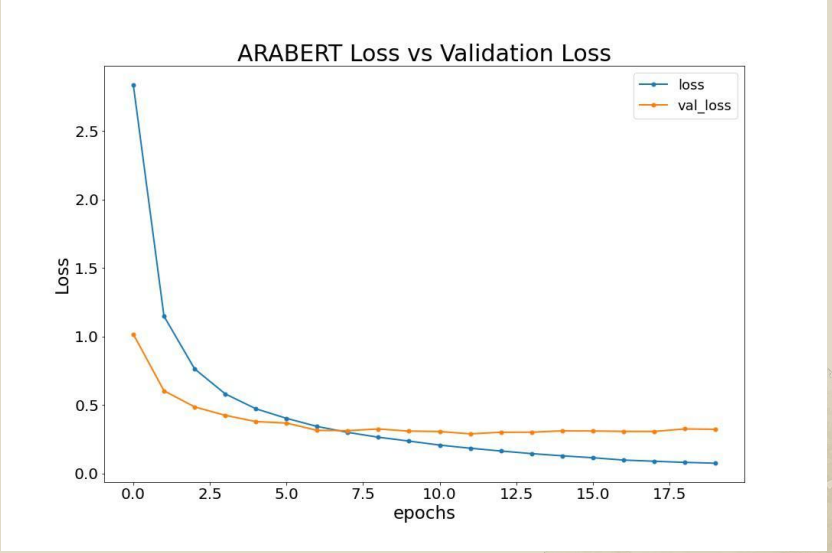
AraBERT Model:

- Sequence Length 512.
- Added layers: 1 Flatten, 2 Dense, 1 Output.



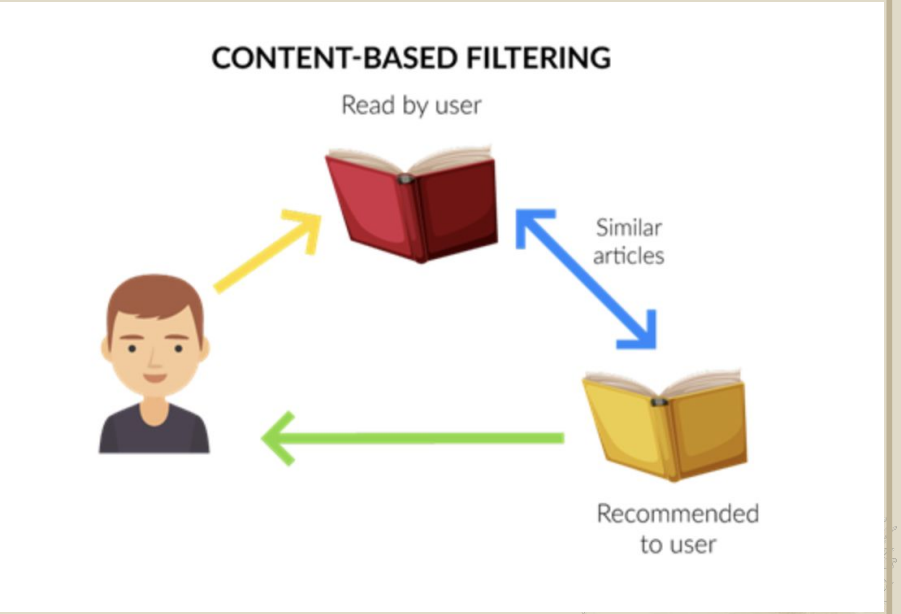
Classification Performance:





Recommender:

- Recommend similar books based on a list of books read.
- Act as an external service for the AraBERT model.
- Recommend based on model output.



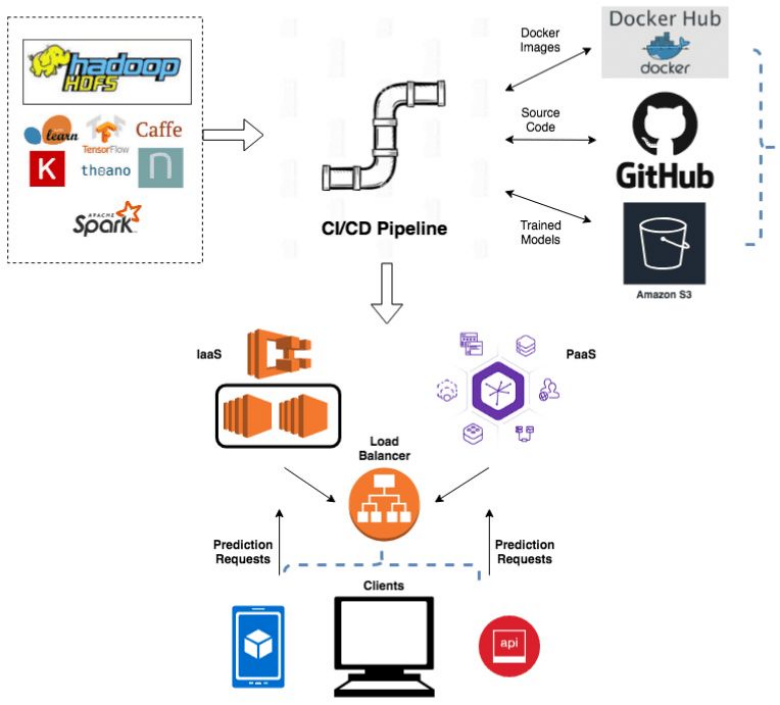
Challenges Deploying Deep Learning Models to Production:

Data & Experimentation:

- Large data size (load & store)
- Large trained transformer size

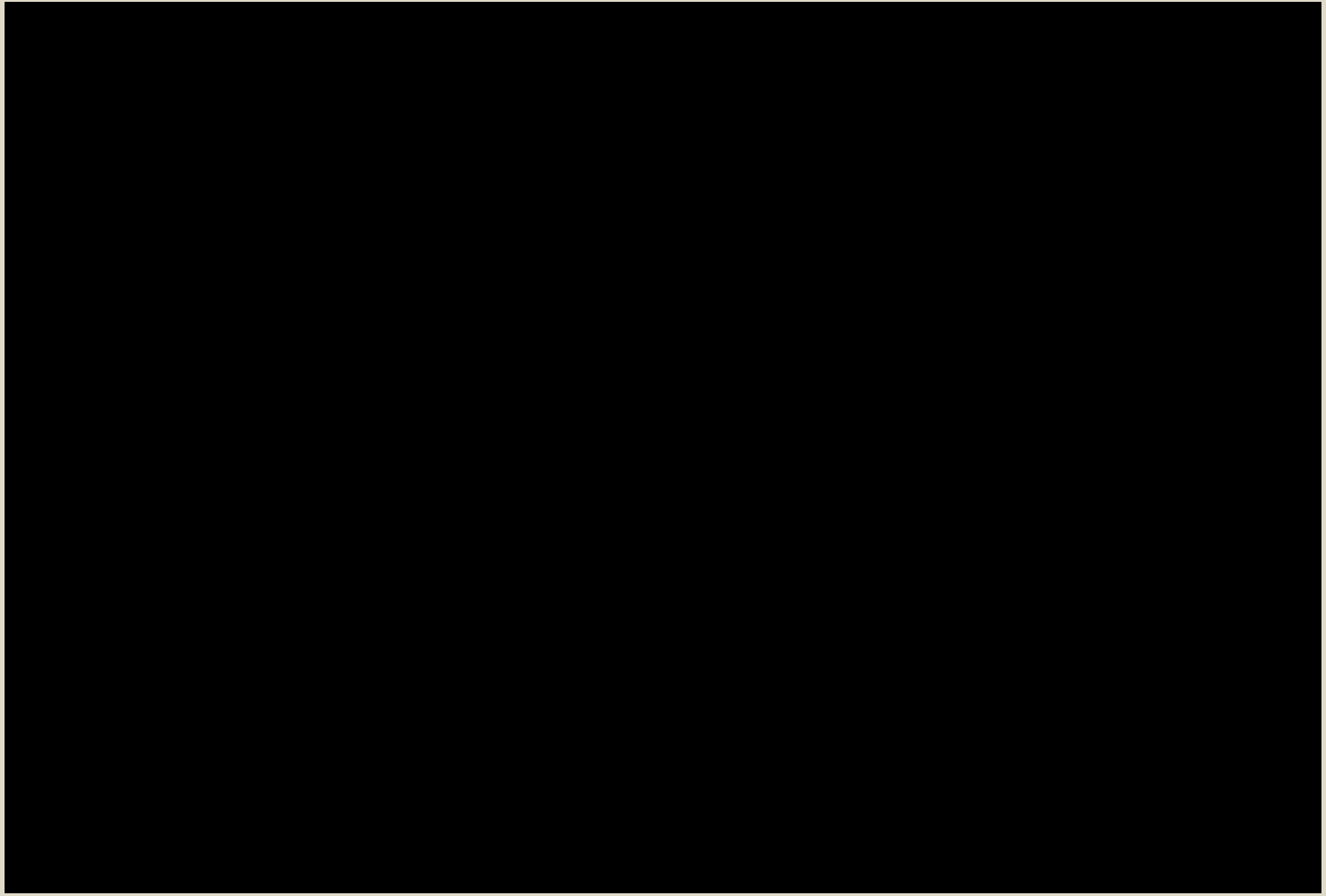
Production Deployment Workflow:

- Prediction Time
- Troubleshooting different services
- Infrastructure Requirements
- Code Quality



Demo:

SCAN ME



Future Work

01. Summaries Books

Show a summary of the recommend books.

02. New Domains

Add authors from different domains.

03. Enhance Recommender

Use Autoencoder to elevate the recommender system performance

04. Text Generation

Generate text similar in style to the input sentence.

Thanks For Listening!
QA

Appendix

Tools

TensorFlow

Streamlit

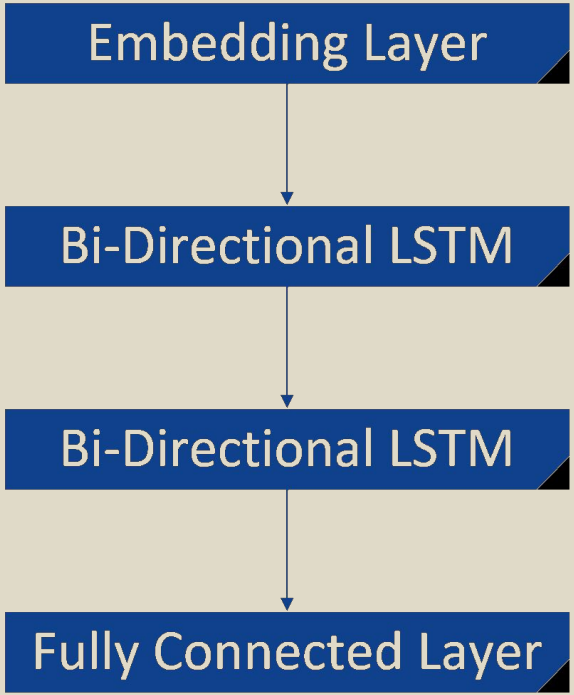
PyArabic

Google Cloud Platform

CO

GitHub

BI-LSTM Model:



Classification Performance: