Analysis of Authorship in Arabic sources

Project by : Mayar Salih , Bolos Khoury
Supervisors : Dr. Renata Avros , Prof. Zeev Volkovich

25-2-R-7

Background & Motivation

Authorship verification aims to determine whether a text was written by a claimed author.
In classical Arabic literature, this task is particularly challenging due to:

- Lack of labeled training data
- Stylistic overlap between authors
- Historical editing and compilation processes
- Internal stylistic variation within long texts

Traditional supervised approaches are unsuitable in this setting, motivating the need for an unsupervised, style based verification framework.

Project Goal

Our goal is developing an unsupervised system for authorship verification in classical Arabic texts that identifies stylistic consistency and deviation without training on the target author.

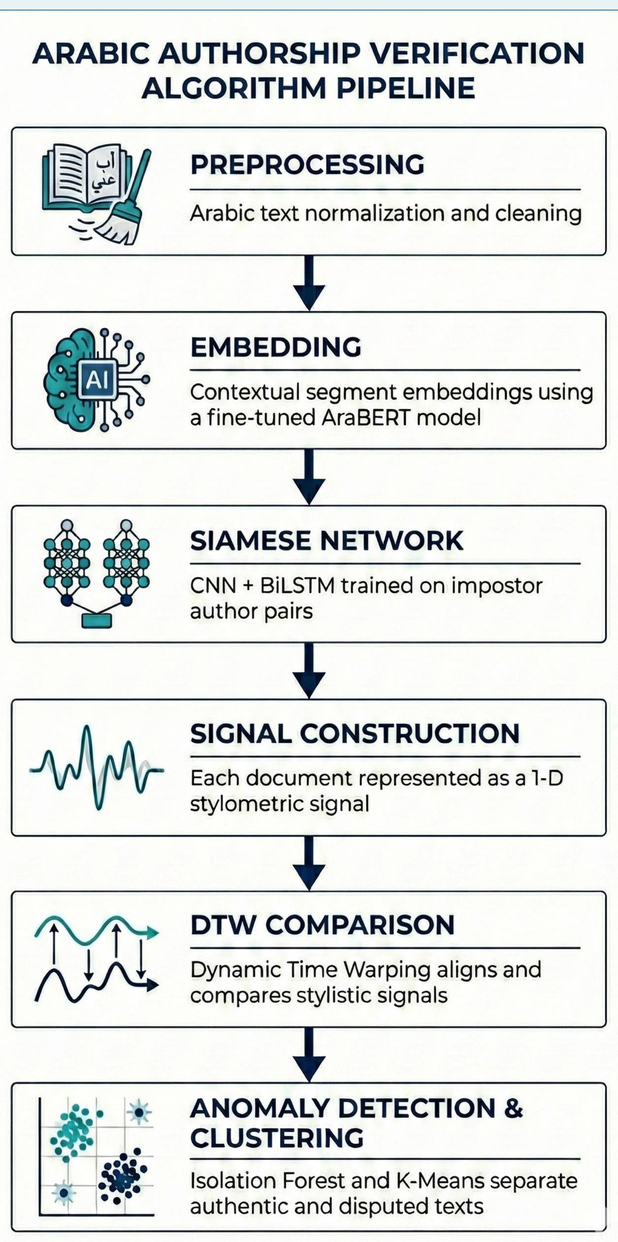
Test Author: Al-Jahiz

Al-Jahiz (776–868 CE) was a major figure of classical Arabic prose during the Abbasid era. His works are known for stylistic richness, rhetorical complexity, and thematic diversity.
We selected Al-Jahiz as the test author due to the size and diversity of his corpus and the existence of historically disputed attributions among works associated with him. His writing exhibits internal stylistic variation, making him an ideal case for evaluating unsupervised authorship verification using the Deep Impostors framework.

Our Solution

We implement the Deep Impostors methodology, which reframes authorship verification as a signal comparison problem.
Instead of training on Al-Jahiz works, the system learns stylistic boundaries using pairs of impostor authors and evaluates how test texts behave relative to these boundaries.

Pipeline



Data Collection

We collected our data for our test author Al-Jahiz and other 25 impostor authors from similar historical or literary contexts from these data sources:

- OpenITI Repository
- Al-Shamela Digital Library

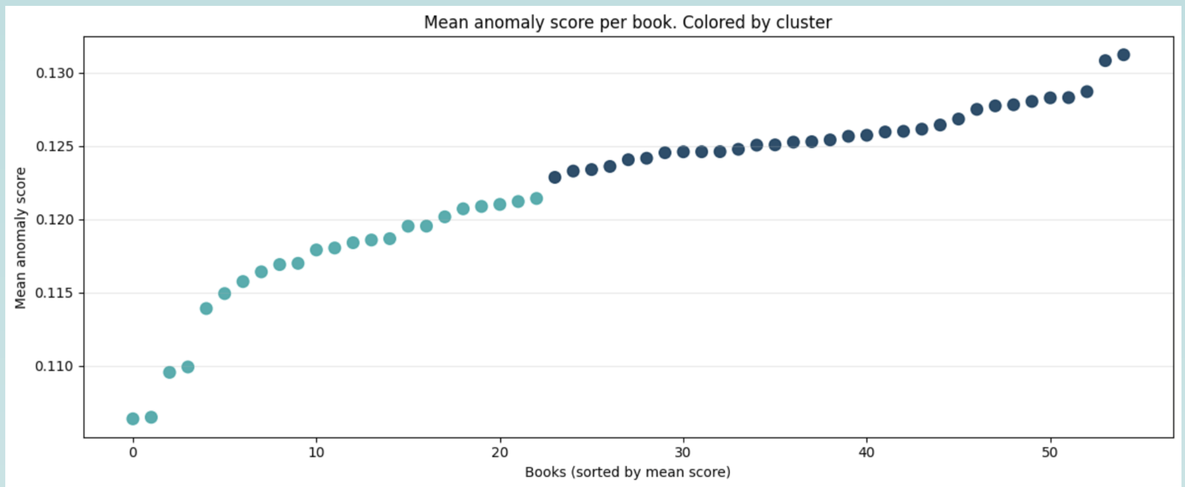
AraBERT Fine-Tuning

We fine-tuned AraBERT using masked language modeling with PyTorch and HuggingFace on a corpus of classical Arabic texts, including works attributed to Al-Jahiz and impostor authors, to adapt the model to Abbasid-era writing styles.

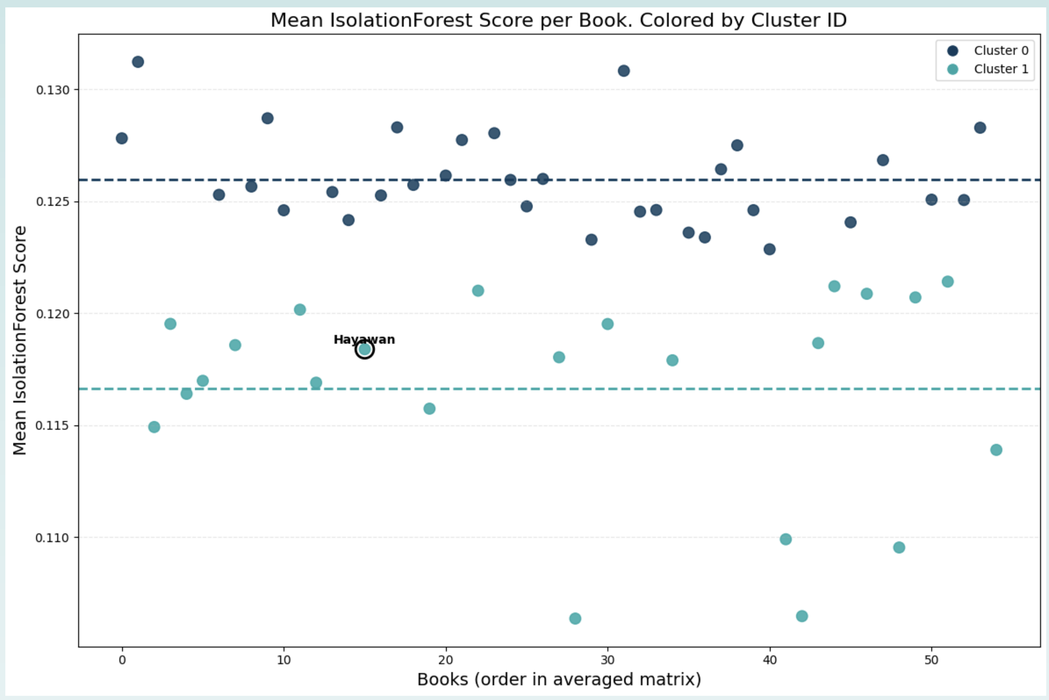
Technologies & Tools

- Python
- AraBERT
- PyTorch
- CAMeL Tools
- Dynamic Time Warping
- Isolation Forest
- K-Means Clustering
- Google Colab & Lambda

Results

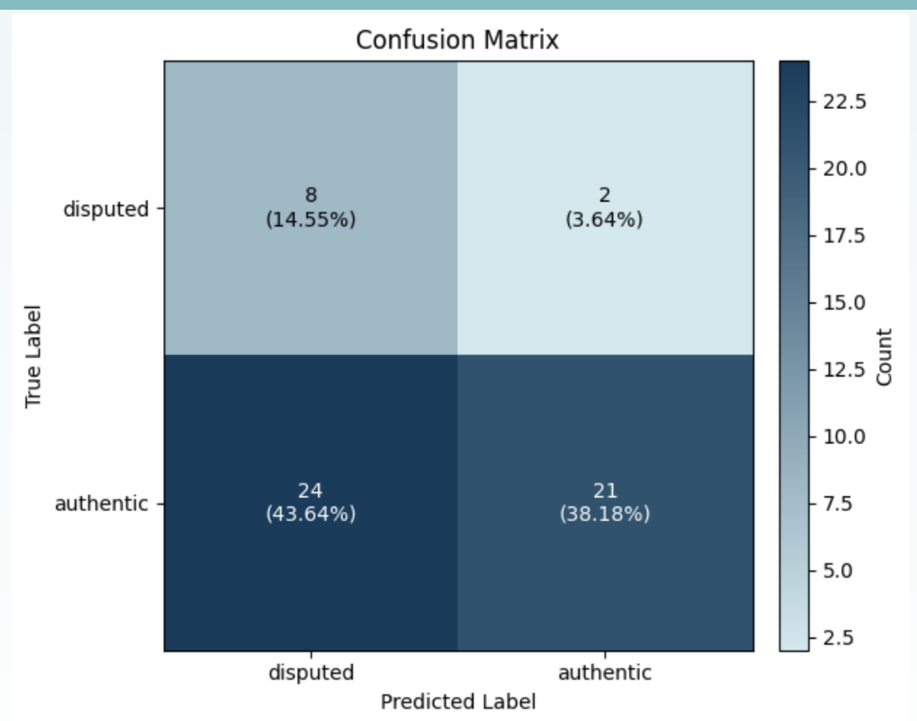


This figure presents the mean anomaly score of each test book, averaged across all impostor pairs and sorted in ascending order.
Higher scores indicate stronger stylistic deviation from impostor baselines, providing a global measure for distinguishing potentially disputed texts.



Each point represents a book, colored by its K-Means cluster assignment based on the full anomaly-score vector. The separation between clusters reflects two dominant stylistic groups, with higher score clusters corresponding to stylistically disputed texts and lower score clusters corresponding to authentic works.

Conclusion



Most works traditionally attributed to Al-Jahiz. the majority of which are labeled authentic compared to a smaller set of disputed texts . concentrate in a single stylistically consistent cluster. The dispersion of some works reflects the known stylistic breadth of Al-Jahiz's corpus rather than sharp authorial boundaries. This result highlights both the internal diversity of his writing and the gradual nature of stylistic variation in classical Arabic prose.