BRAUDE
College of Engineering, Karmiel

25-1-R-19

Analysis of Medieval Arabic Creations

Yosra Fhamne & Omar Saleh
Supervisors: Prof. Zeev Volkovich & Dr. Renata Avros

Software and Information Systems Engineering Department

Problem Definition

Authorship attribution in medieval Arabic texts is challenged by complex morphology, inconsistent orthography, and a lack of labeled data. Traditional methods are subjective and not scalable. This study addresses the need for an automated, language-aware approach to detect falsely attributed works.

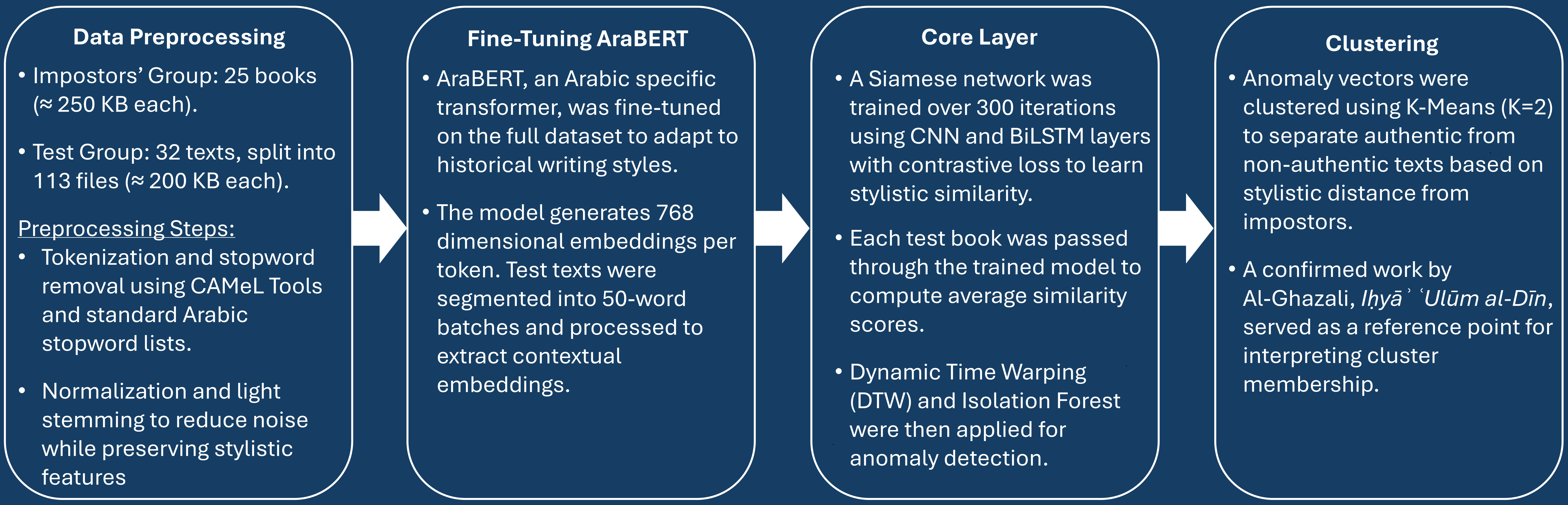
Design Goals

The system must operate without relying on labeled data, support the unique characteristics of classical Arabic, and generalize across diverse writing styles. It should scale to large historical corpora, produce interpretable similarity signals for analysis, and remain reproducible across different experimental setups.

Our Approach

We propose a scalable and reproducible framework for authorship verification of historical Arabic texts, focusing on works attributed to Imam Al-Ghazali. The approach combines AraBERT embeddings with a Siamese CNN-BiLSTM model to detect stylistic anomalies from internal linguistic features, enabling unsupervised attribution in low-resource, morphologically rich languages.

Algorithm Workflow



Data Preprocessing

- Impostors' Group: 25 books (≈ 250 KB each).
- Test Group: 32 texts, split into 113 files (≈ 200 KB each).

Preprocessing Steps:

- Tokenization and stopword removal using CAMeL Tools and standard Arabic stopword lists.
- Normalization and light stemming to reduce noise while preserving stylistic features

Fine-Tuning AraBERT

- AraBERT, an Arabic specific transformer, was fine-tuned on the full dataset to adapt to historical writing styles.
- The model generates 768 dimensional embeddings per token. Test texts were segmented into 50-word batches and processed to extract contextual embeddings.

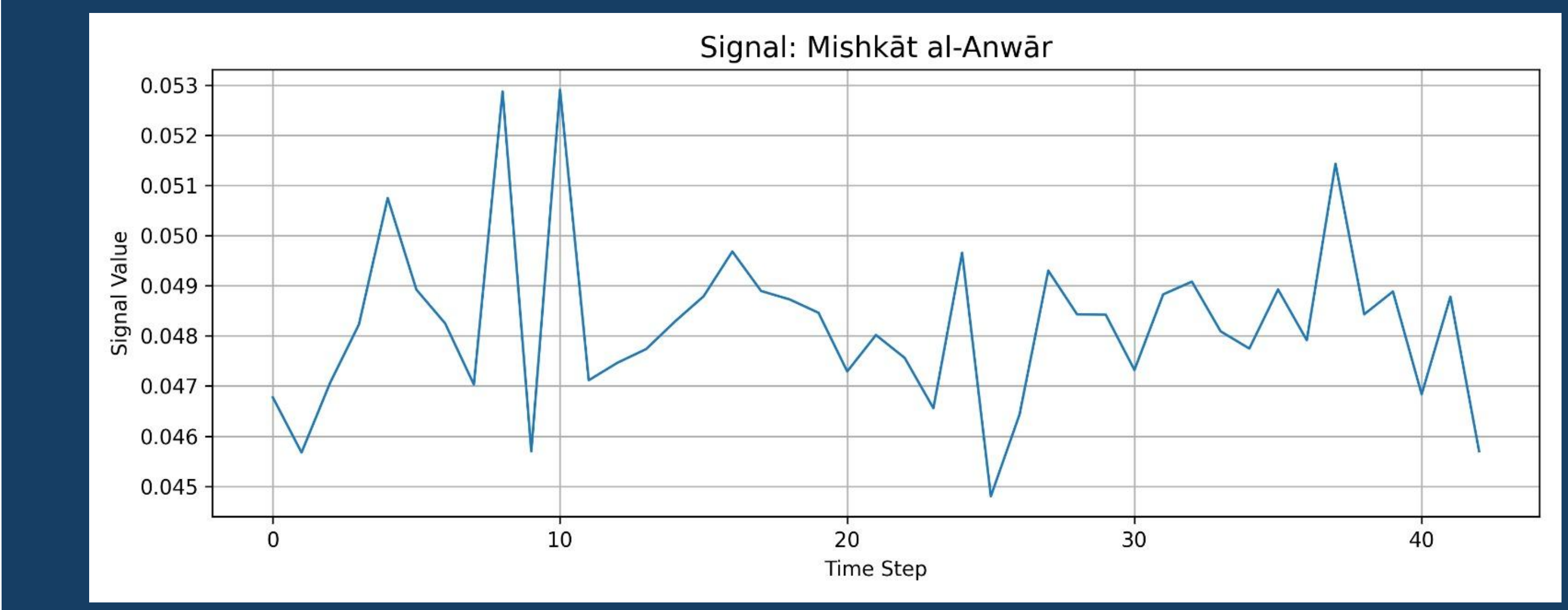
Core Layer

- A Siamese network was trained over 300 iterations using CNN and BiLSTM layers with contrastive loss to learn stylistic similarity.
- Each test book was passed through the trained model to compute average similarity scores.
- Dynamic Time Warping (DTW) and Isolation Forest were then applied for anomaly detection.

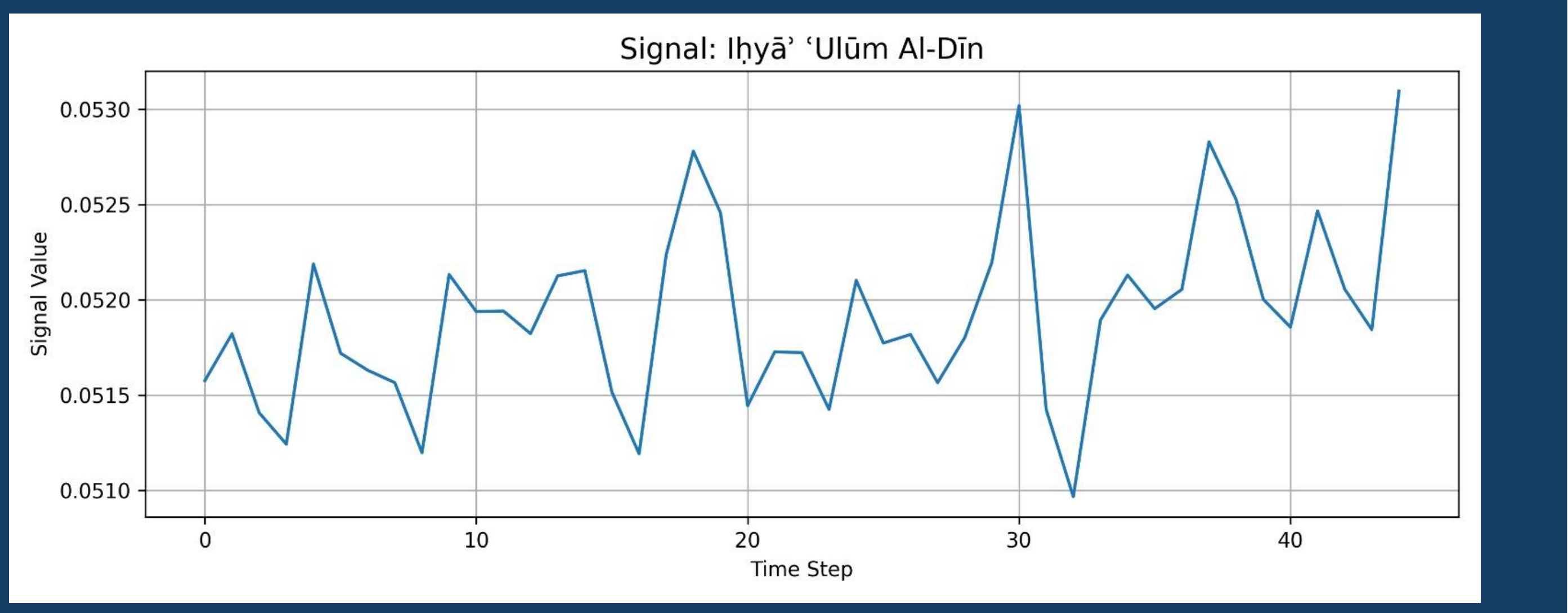
Clustering

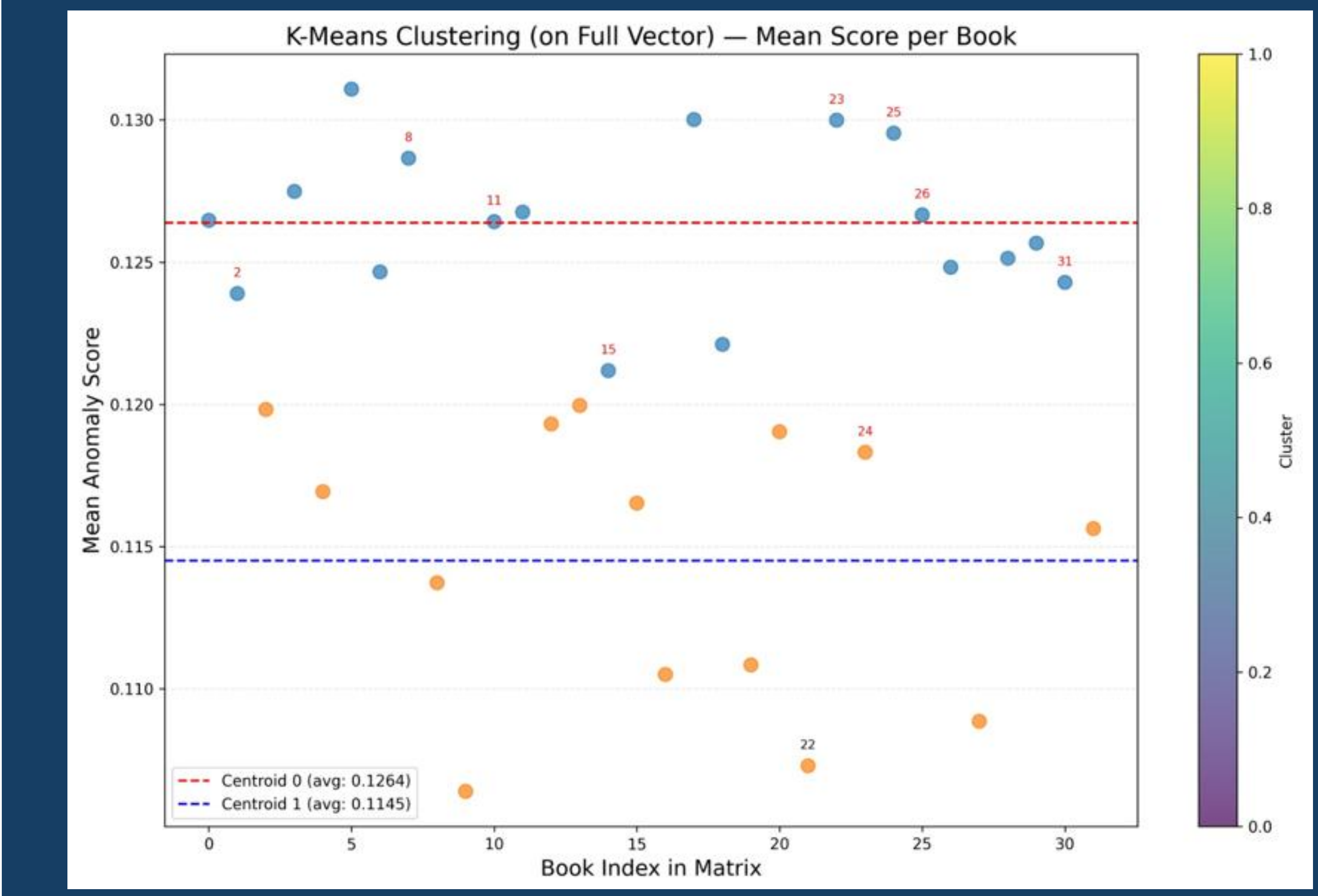
- Anomaly vectors were clustered using K-Means (K=2) to separate authentic from non-authentic texts based on stylistic distance from impostors.
- A confirmed work by Al-Ghazali, *Iḥyāʾ ʿUlūm al-Dīn*, served as a reference point for interpreting cluster membership.

Results

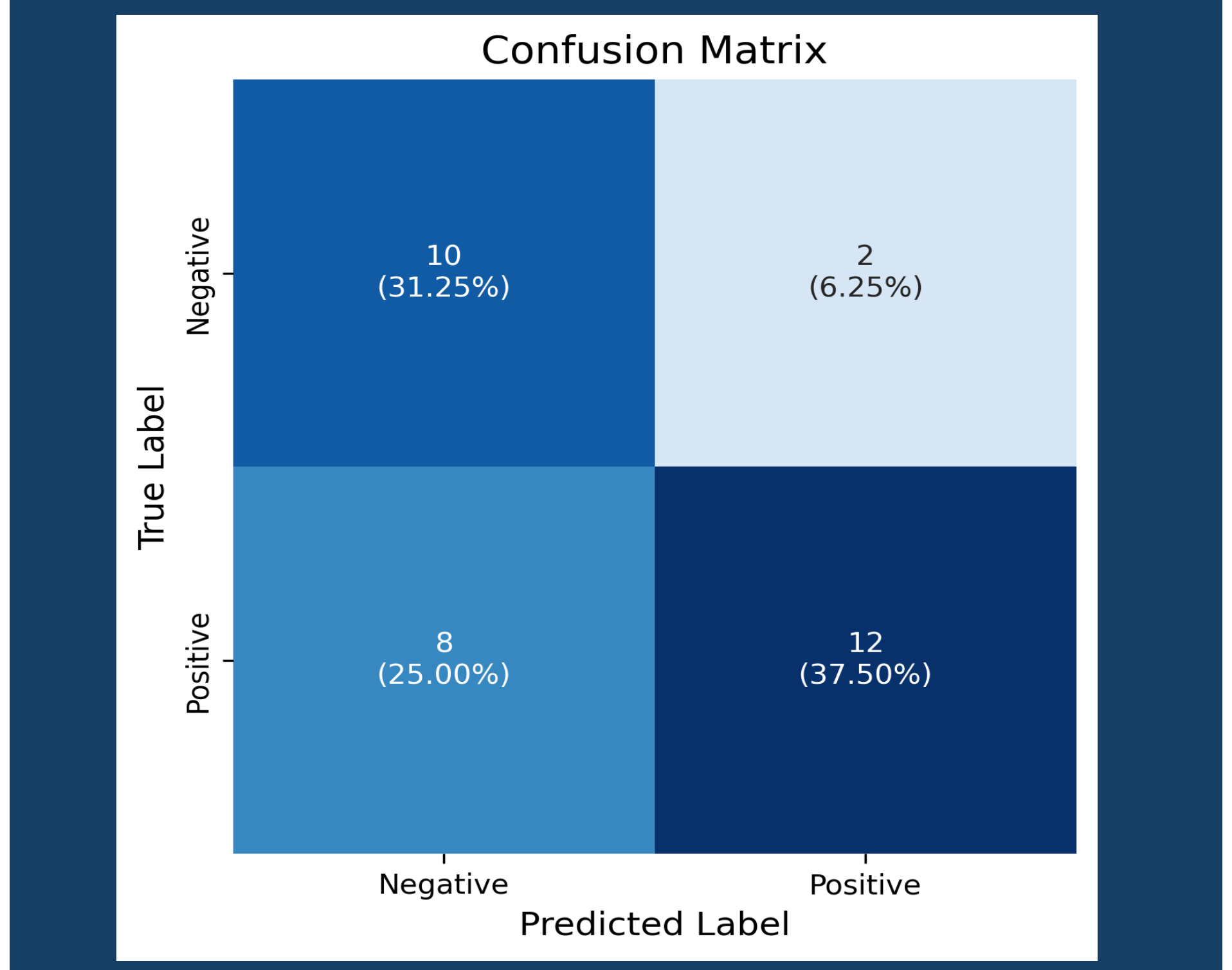


Signals vary in their smoothness and fluctuation patterns. Stable, coherent signals typically reflect authentic writing, while irregular or spiky signals may suggest stylistic inconsistency or disputed authorship.

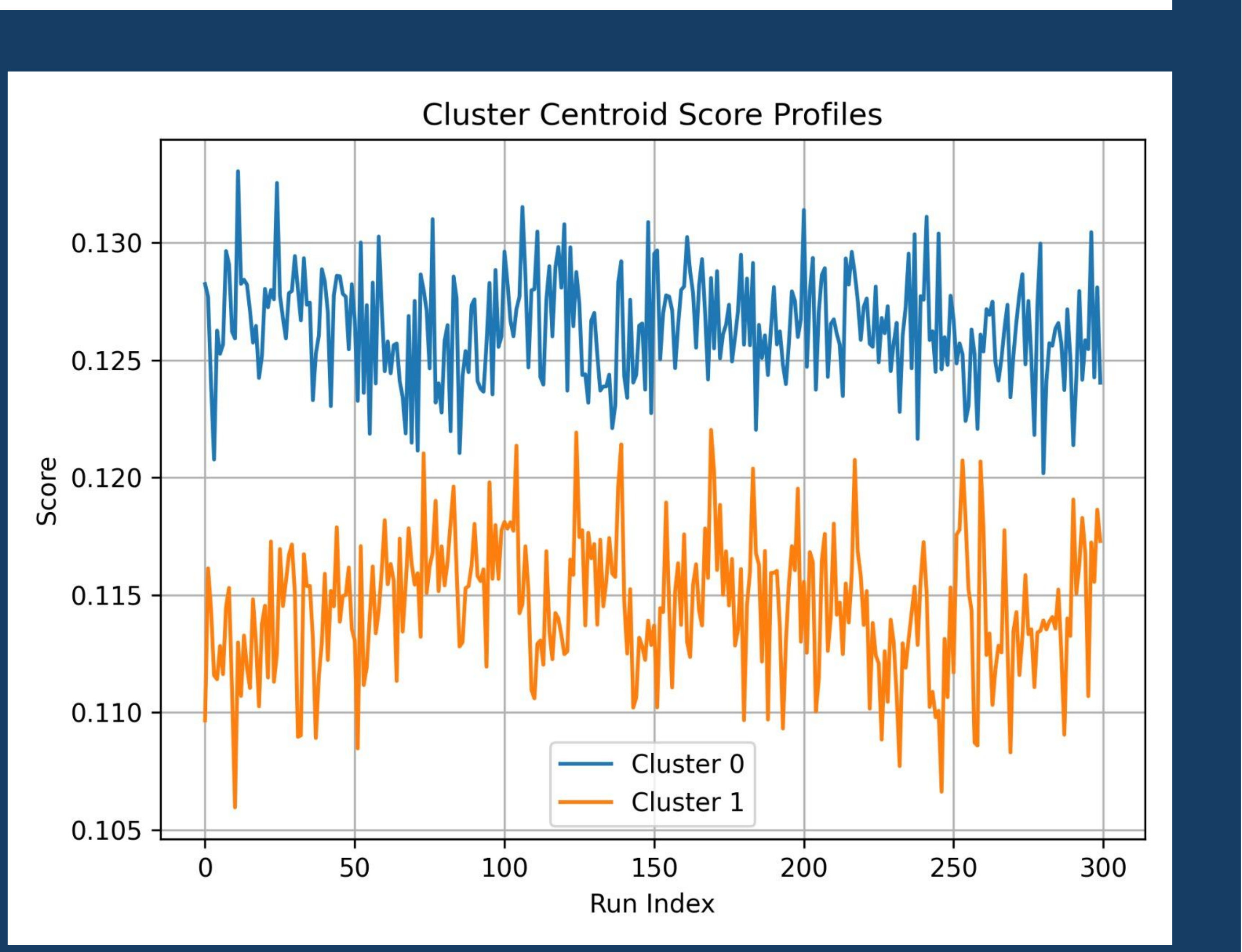




Book numbers in red indicate texts identified in prior research as non-authentic. Book 22 represents *Iḥyāʾ ʿUlūm al-Dīn*.



Each value shows the number of books and its percentage out of the full dataset (N=32).
TP (12) and TN (10) together account for 69% correctly classified texts.



Lower anomaly scores indicate stronger stylistic coherence, while higher scores suggest greater deviation from the learned reference style.

Conclusion

The results reveal a clear stylistic separation between authentic and disputed texts, with high alignment to scholarly classifications. This outcome highlights the reliability of our signal-based approach and its potential to support authorship studies where ground truth is uncertain.