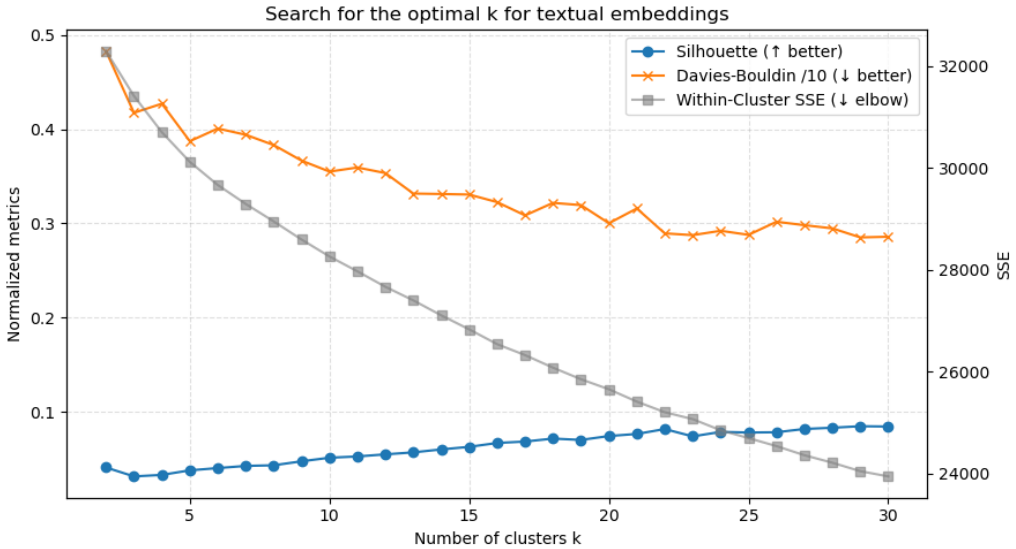
Giovanni Filomeno,
Katharina Hehenwarter,
Elīna Emīlija Ungure,
Kathrin Hofmarcher

Data Exploration

Machine Learning and Pattern Classification

JOHANNES KEPLER UNIVERSITY LINZ

Are text-clusters meaningful? (Question a)

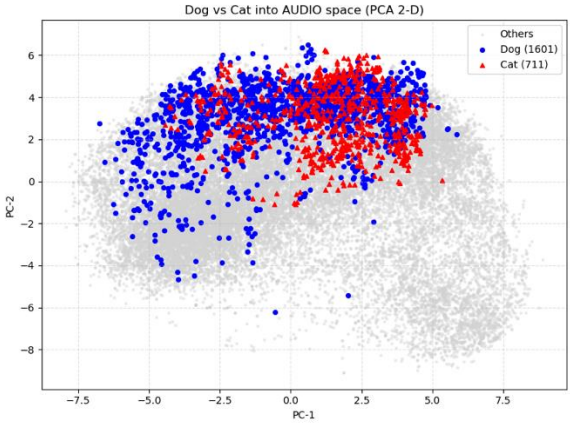


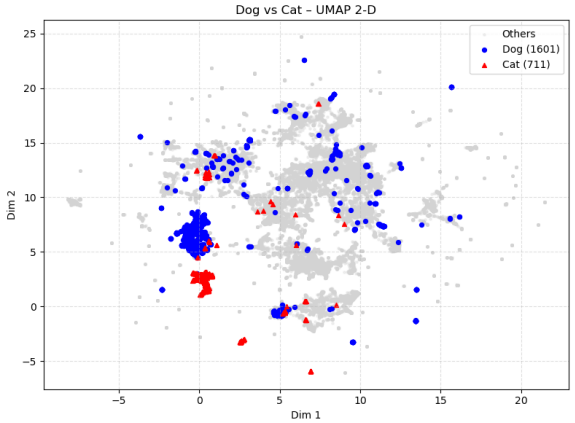
- 35 826 annotation embeddings
- Internal validity scan k ∈ [2…30]
  - Silhouette ↑, DB ↓, SSE elbow ↓
- Trade-off chosen ➔ k = 15
  - coherent keywords per cluster
  - still interpretable (< 4000 notes / cluster)

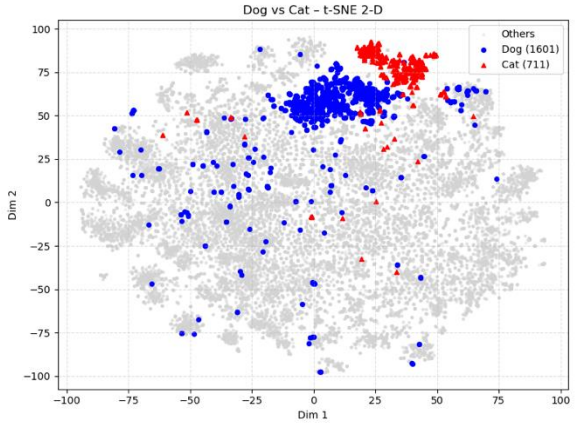
* DB – Davies-Bouldin index

** SSE – Sum of Squared Errors

Dog & Cat labelling function (Question b)







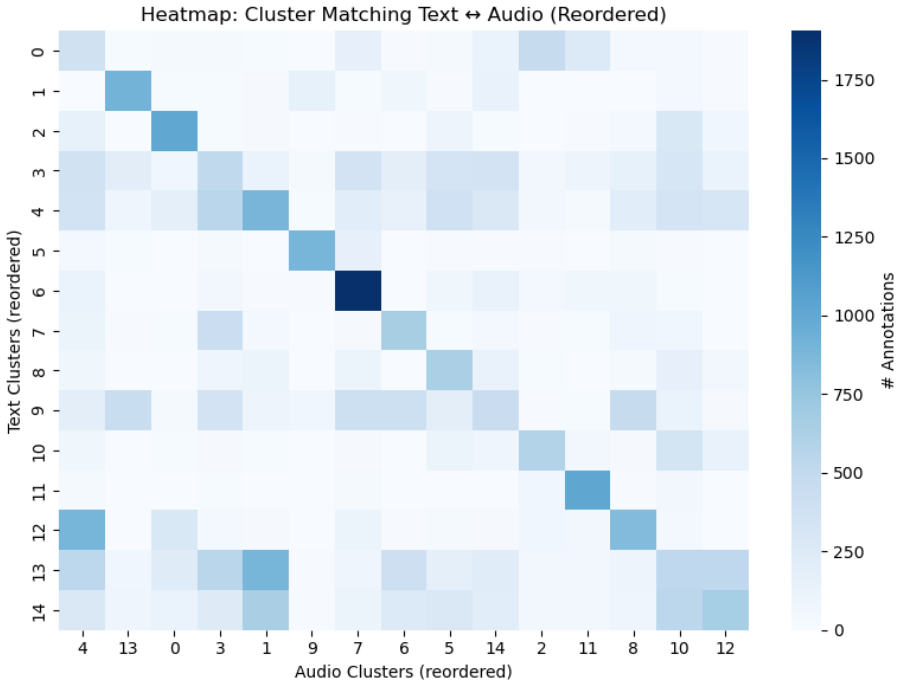
Regex Formula:
\b(?:dog|dogs)\b
\b(?:cat|cats)\b
⇒ 1 601 “dog” • 711 “cat”

- Silhouette global 0.16
  - Dog 0.15
  - Cat 0.19
- t-SNE: best local separation
- UMAP shows cat sub-clusters

JOHANNES KEPLER UNIVERSITY LINZ

Text ↔ Audio cluster alignment (Question c)

- Text & audio re-clustered with the same k = 15
  - Metrics ARI 0.13
  - NMI 0.26
- Good matches:
  - T6 ↔ A7 (74 %)
  - T11 ↔ A11 (81 %)
- Others scatter → semantic gap between narration & acoustics



* ARI – Adjusted Rand Index

** NMI – Normalized Mutual Information