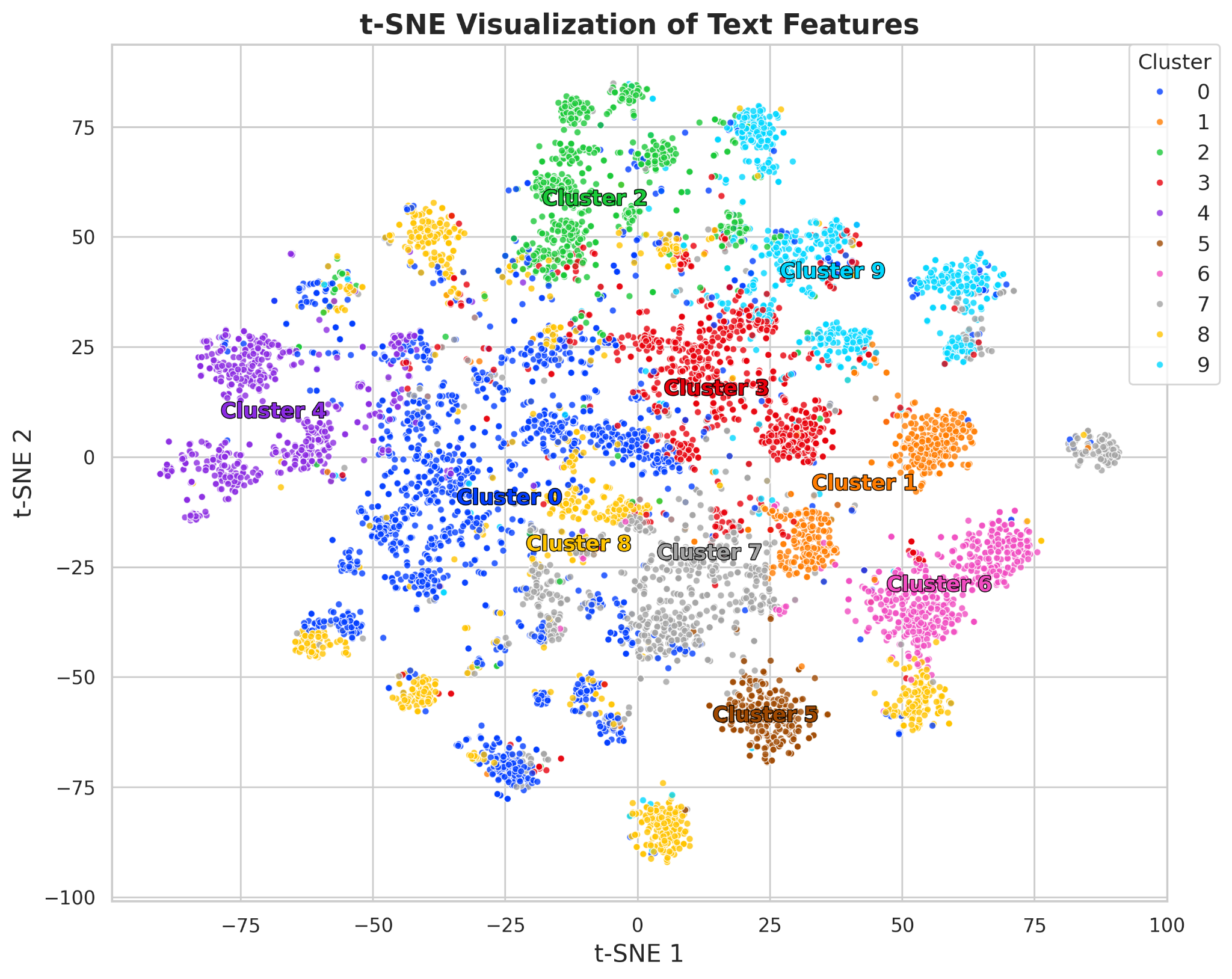
4. Text Features (6 points): Load and analyze the text features:

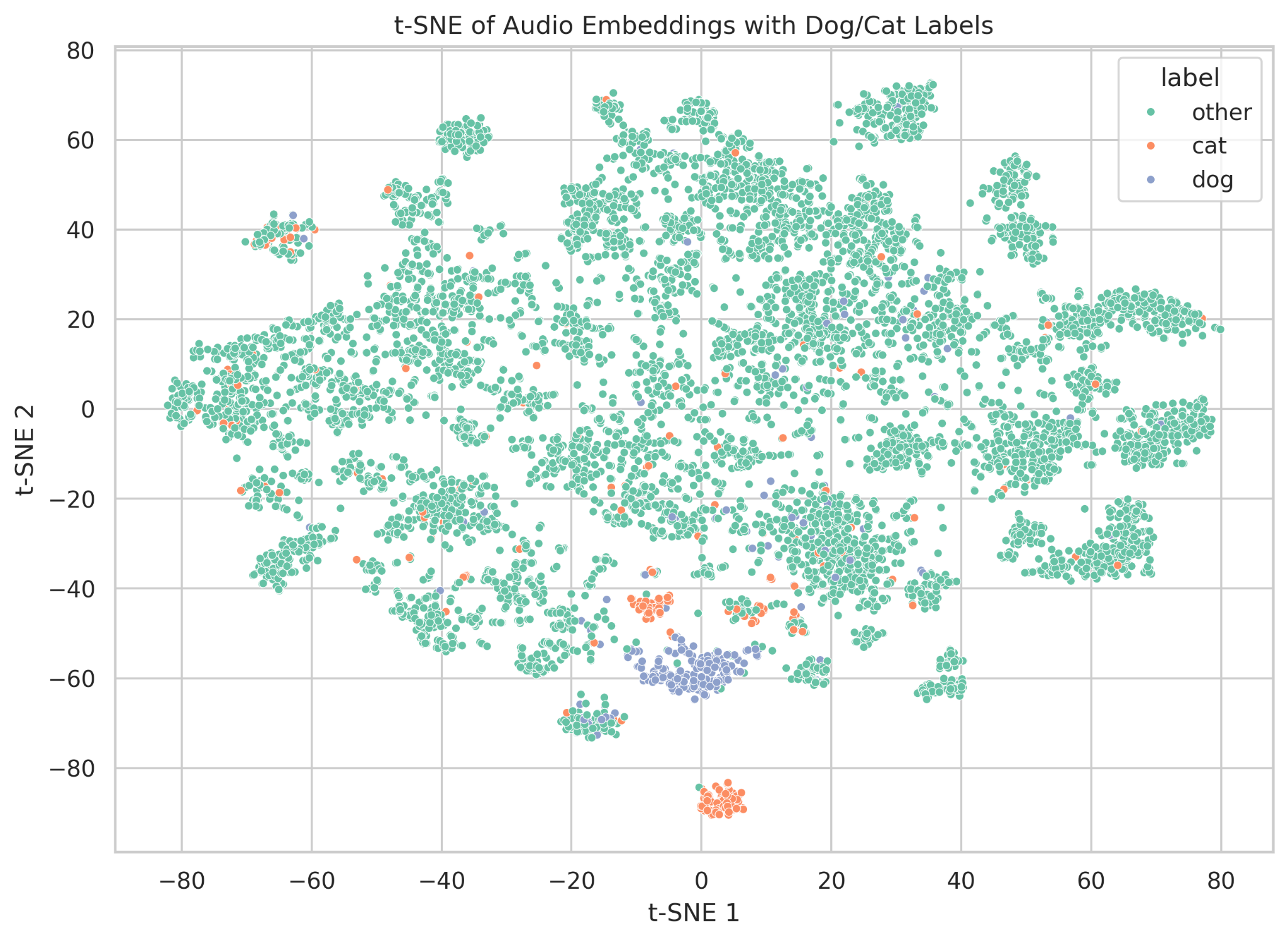
(a) Cluster the text features. Can you find meaningful clusters?



The above figure presents a two-dimensional t-SNE projection of high-dimensional text feature embeddings, created from concatenating the keyword and title embeddings, colored by cluster labels obtained via k-means clustering. Cluster centoroids are marked by the text “Cluster X”.

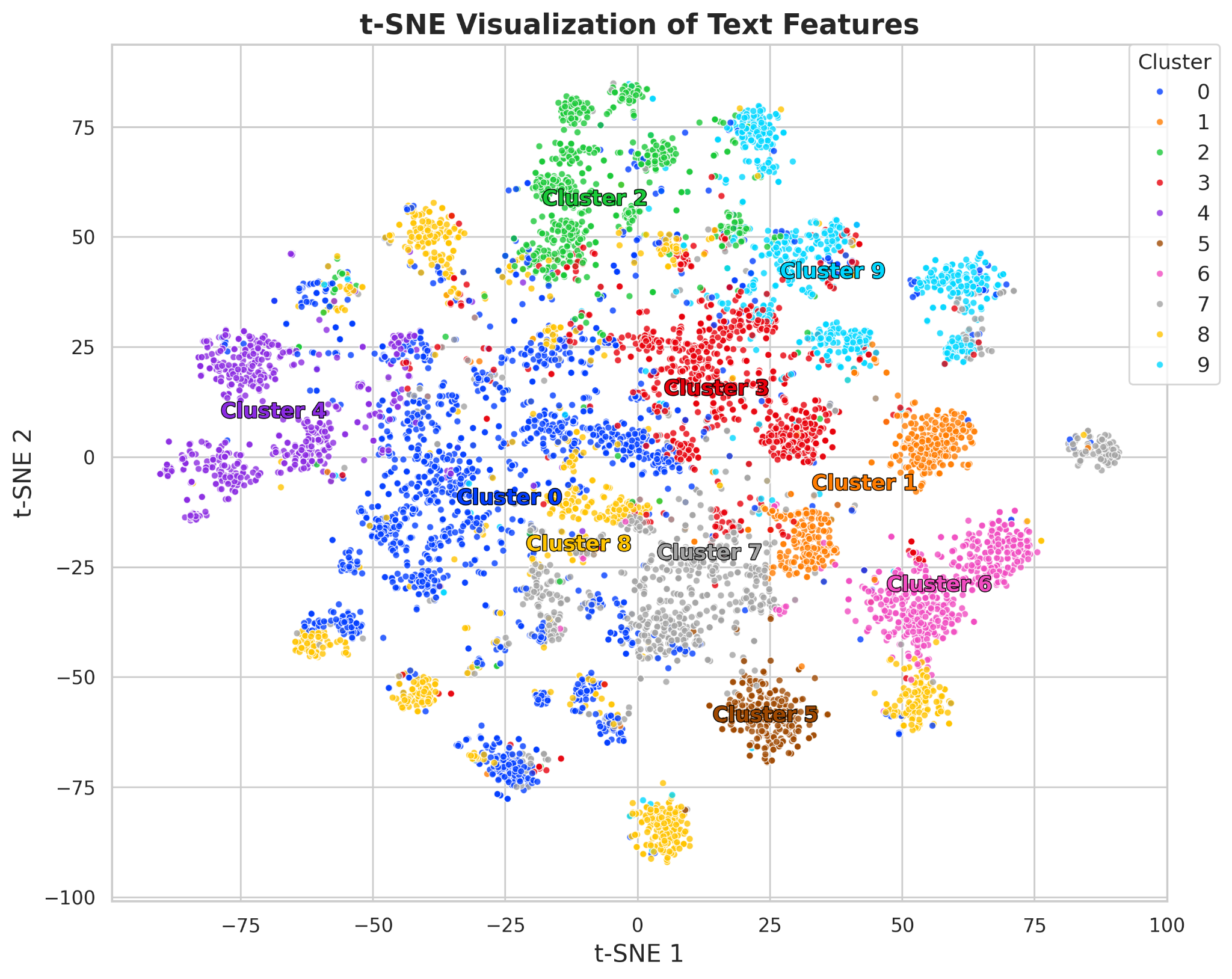
The spatial separation and density of the clusters suggest that the embedding space captures semantically meaningful distinctions. Several clusters exhibit high intra-cluster similarity and low inter-cluster overlap, supporting the validity of the clustering model and the discriminative capacity of the underlying features.

(b) Design a labeling function for classes dog and cat. Do the annotations labeled as dog or cat sounds form tight clusters in the text and audio feature space?



The above figures show t-SNE visualizations for classes dog and cat in the text and audio feature spaces

* Text Annotations Embedding Space (Left Plot):  
  The "dog" and "cat" labels form fairly well-separated clusters from each other and from the "other" class. The "dog" cluster (blue) and the "cat" cluster (orange) are compact and localized, suggesting that the text descriptions for these labels are semantically coherent and distinctive.
* Audio Embedding Space (Right Plot):The clustering is less pronounced. While "dog" and "cat" annotations (blue and orange) still appear more concentrated than the "other" class, they are more interspersed with noise and overlap. This suggests that the audio features for these classes are more heterogeneous or less separable (using the current feature representation).

(c) How well do the audio feature clusters align with text clusters?

As seen in Figures from 4(a) and 3(c), there is a notable alignment between certain clusters in the text and audio feature spaces. For instance, Cluster 4 in the text feature t-SNE plot (left) shares a similar position in the 2D embedding space with Cluster 2 in the audio feature t-SNE plot (right). This alignment suggests that the clustering structures across modalities are partially consistent — that is, certain groups of samples cluster together in similar ways whether represented by text or audio features. While t-SNE is primarily a visualization tool and may distort global relationships, the local alignment of these clusters indicates that similar underlying semantics are being captured in both modalities.