Appendix B: The Selection of the Optimal Number of Clusters

Jianjun Yu[1]

[1] *Political science department, University of Iowa, Iowa City, IA, USA. Email: Jianjyu@uiowa.edu*

Selecting the optimal number of clusters is an important step in text clustering. In this research, two different approaches are used to determine the number of clusters. The first approach considers the dataset's actual number of manually labeled clusters. For the UCI data, there are four clusters, while for the Twitter data, there are six clusters. The second approach involves selecting the optimal number of clusters based on model performance.

To select the optimal number of clusters for STM, I use coherence and perplexity scores. For TLM-based models, the Elbow curve method and silhouette analysis are employed. Due to the computational complexity of silhouette analysis on large datasets, a random sample of 5000 data points is selected for calculating the Silhouette score for the UCI data.

Figure 1 presents the perplexity score (a) and coherence score (b) of STM on the UCI data. The performance of the STM model shows a gradual improvement until it reaches a plateau at around 15 to 25 topics. Based on these scores, a total of 25 topics is selected as the optimal number for this dataset. The perplexity score (c) and coherence score (d) of STM on the TTC data are shown in the last two pictures of Figure 4. Similarly, the performance of the STM model plateaus at around 5 to 10 topics. Therefore, ten topics are chosen as the optimal number for this dataset.

Figure 2 displays the results of the elbow curve method and silhouette analysis for BERT-based text clustering. The first row presents the results for the UCI data, while the second row shows the results for the TTC data. The first column illustrates the change in the sum of squared distances (SSD) across the number of topics. A smaller value indicates more coherent clusters. The second column depicts the first difference of the change, representing the improvement in coherence with the addition of one more cluster. A smaller value indicates a larger improvement in clustering coherence. The third column shows the silhouette scores across the number of topics. A higher value suggests a better number of topics for clustering.

For the UCI data, the improvement in SSD becomes relatively constant from 15 to 25 topics. Additionally, there is a significant jump in the silhouette score from 20 to 25 topics. As a result, 25 topics are selected as the optimal number for this dataset.

For the TTC data, the improvement in SSD becomes relatively constant from 10 to 15 topics. Therefore, 15 topics are chosen as the optimal number for this dataset. Since the silhouette score does not provide a clear pattern, it is not utilized to determine the optimal number of topics for this dataset.

Figure 3 illustrates the outcome of the elbow curve method and silhouette analysis for GPT-based text clustering on the two datasets. The first row presents the results for the UCI data, while the second row displays the findings for the TTC data. The three columns depict the Sum of Squared Distances (SSD) across the number of topics, the first difference in the change of SSD, and the silhouette score across the number of topics.

For the UCI data, as can be observed, the SSD improvement plateaus from 10 to 15 topics. Consequently, I select 15 as the optimal number of topics for GPT-based clustering on the UCI dataset.

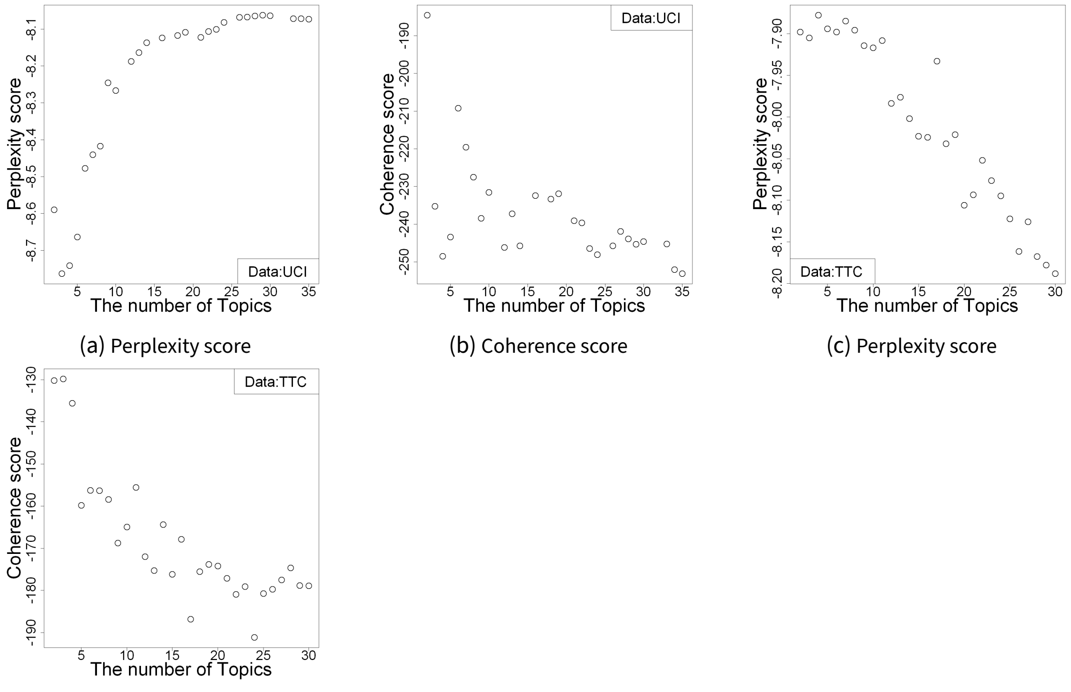
For the TTC data, the SSD improvement plateaus from 10 to 15 topics. Interestingly, the silhouette score peaks at around 15 topics. As such, I select 15 as the optimal number of topics for the TTC dataset. The selection of the optimal number of topics may vary based on individual researchers'

Political Analysis (2023)

DOI: 10.1017/pan.xxxx.xx

Corresponding author
Jianjun Yu

Edited by
John Doe



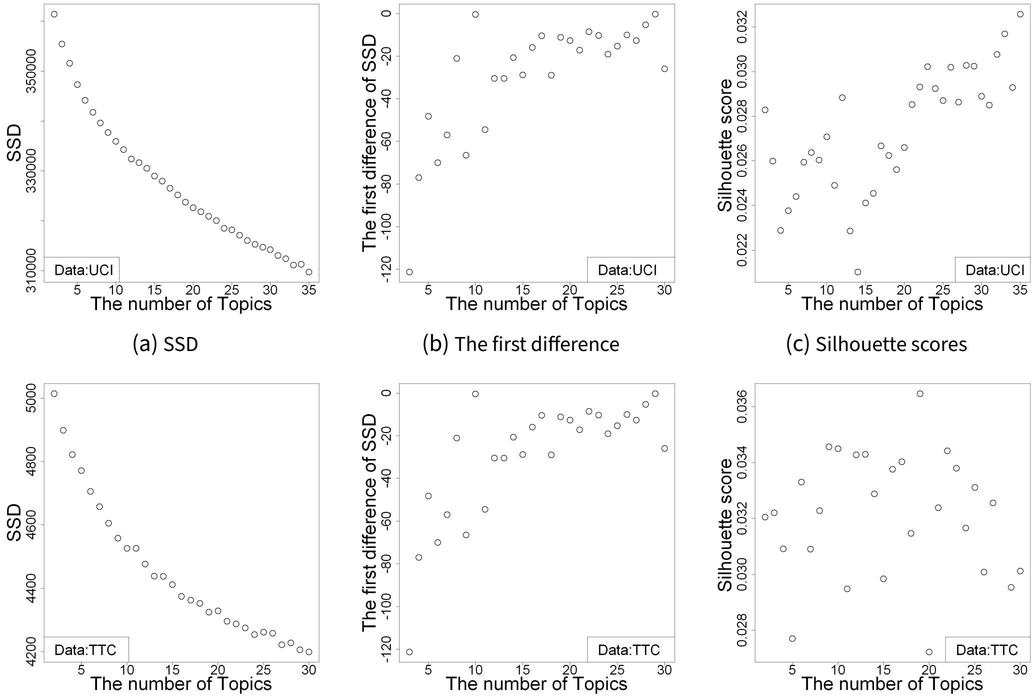
(a) Perplexity score

(b) Coherence score

(c) Perplexity score

(d) Coherence score

Figure 1. The optimal number of topics for STM on two datasets



(a) SSD

(b) The first difference

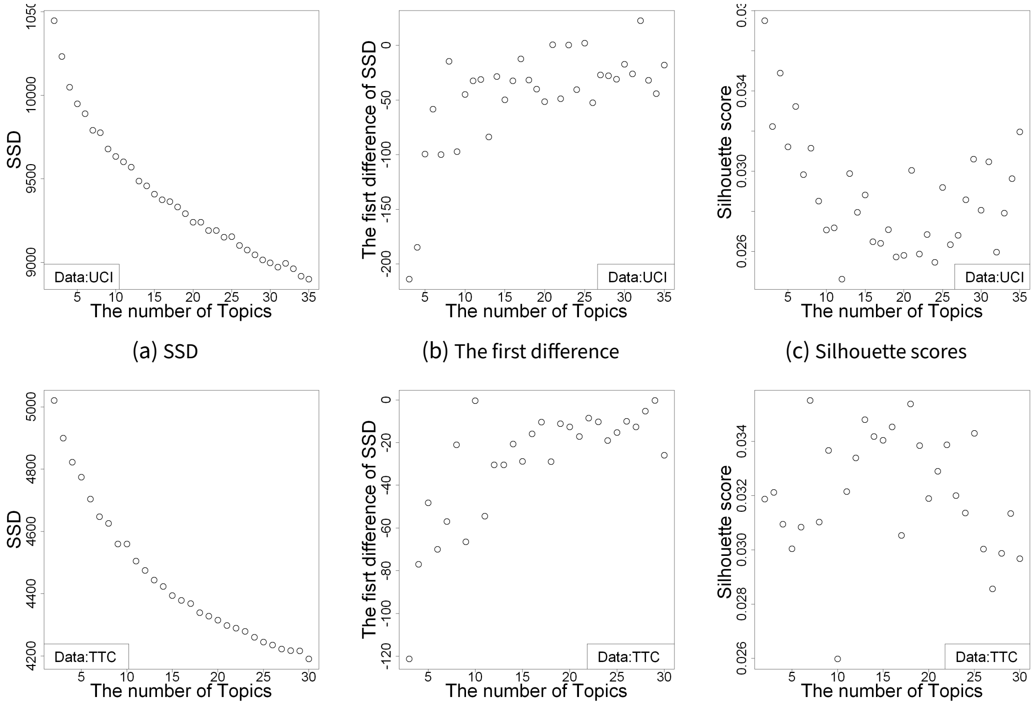
(c) Silhouette scores

(d) SSD

(e) The first difference

(f) Silhouette scores

Figure 2. The optimal number of clusterss for BERT-based text clustering on two datasets



(a) SSD

(b) The first difference

(c) Silhouette scores

(d) SSD

(e) The first difference

(f) Silhouette scores

Figure 3. Appendix B: The optimal number of clusters for GPT-based text clustering on two datasets

preferences, as the metrics only provide an approximate range for optimal topic numbers. Nonetheless, the principal objective of this paper is to illustrate that TLM-based text clustering outperforms the commonly used PTMs. Consequently, slight variations in the optimal number of topics should not substantially affect this primary aim.