Comparison of Embedding Models for Clustering Findlay Newsroom Articles

Rohith Chakinarapu

Master of Science in Applied Security & Analytics

ABSTRACT

This study analyzed over 1,000 archived articles from the University of Findlay Newsroom to evaluate the effectiveness of various text embedding and clustering techniques in uncovering communication themes and audience engagement patterns. Word2Vec, GloVe, and BERT embeddings combined with K-Means and Hierarchical Clustering were compared to assess cluster quality using Silhouette Score and Davies–Bouldin Index. The combination of Word2Vec and K-Means produced the most coherent and interpretable clusters, offering insights for enhancing institutional communication strategies.

INTRODUCTION

Data were scraped from the University of Findlay Newsroom – News Releases using *BeautifulSoup* and *Requests* Python libraries. The corpus collected included article titles, dates, view counts and content. The articles were cleaned and processed using natural language processing techniques including removing stopwords, standardizing word forms (lemmatization), and handling names consistently (using a custom dictionary). The processed texts were then represented numerically using three text embedding methods—Word2Vec, GloVe, and BERT—to capture different levels of word meaning and context. These representations were grouped using two clustering techniques: K-Means, which partitions data into defined groups, and Hierarchical Clustering, which builds a tree-like structure of related topics. Cluster performance was evaluated using standard quality measures: Silhouette Score (how well articles fit their assigned group) and Davies–Bouldin Index (how distinct the groups are). Analyzing article view counts across these clusters revealed which topics drew the highest audience attention.

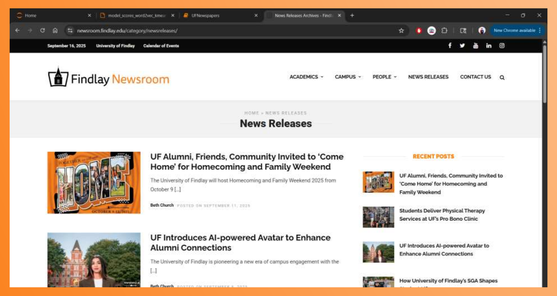

Figure 1: Image showing the Findlay Newsroom from which the data were extracted for analysis.

METHOD

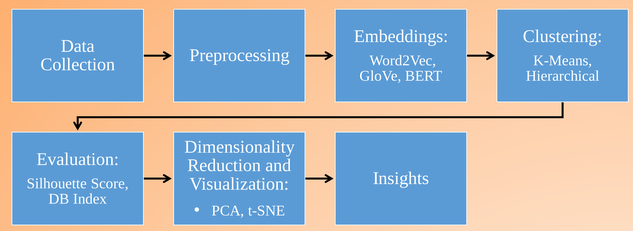

Figure 2: Workflow of Analysis Steps

After data collection and preprocessing, the three embedding models tested for representation were:

- Word2Vec – captured local context
- GloVe – learned global co-occurrence patterns
- BERT – conducted deep contextual transformer embedding

Each embedding was clustered using:

- K-Means with cluster sizes k = 2–15
- Hierarchical Clustering with n = 2,7,13,15

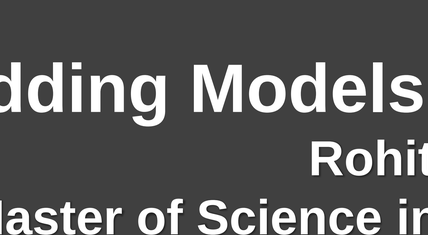

Figure 3: Scatter Chart of Clustering Quality

Silhouette Score (higher is better) and Davies–Bouldin Index (lower is better) were used to select the best k values for Word2Vec + K-Means. Both metrics favored using K-Means with 9 clusters, as seen in Figure 3.

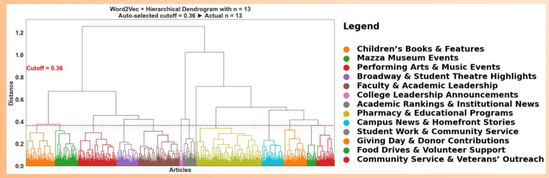

Figure 4: Hierarchical Clustering (Word2Vec representation, cluster size = 13)

The dendrogram in Figure 4 groups UF Newsroom articles into 13 topic clusters using Word2Vec embeddings and hierarchical clustering. The red line shows the distance-based cutoff. Colors represent distinct clusters, matched with the legend for interpretation.

RESULTS

S.No	Embedding	Clustering	Best k / n	Silhouette Score (↑)	Davies–Bouldin Index (↓)
1	Word2Vec	K-Means	9	**0.1543**	**1.7199**
2	GloVe	K-Means	2	0.1417	2.2936
3	BERT	K-Means	3	0.0804	3.1578
4	Word2Vec	Hierarchical	13	0.1276	1.7258
5	BERT	Hierarchical	12	0.0510	3.2127
6	GloVe	Hierarchical	12	0.0600	2.4294

Figure 5: Clustering Results for Each Embedding and Clustering Combination

Word2Vec + K-Means (k = 9) produced the clearest, most interpretable clusters with strong quantitative scores (Silhouette = 0.42, DBI = 0.79).

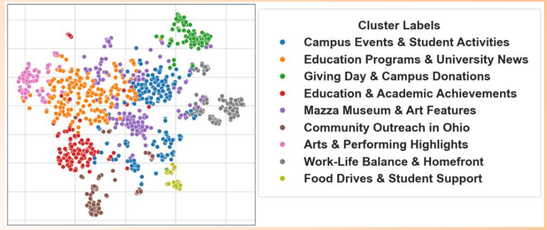

Figure 6: Overall Best Clusters of UF Newsroom Articles

Each dot represents a UF Newsroom article positioned in 2D space using dimensionality reduction. Articles are grouped into 9 color-coded clusters based on semantic similarity using Word2Vec embeddings and K-Means (k=9). Cluster labels were assigned using top keywords and reflect dominant themes such as campus events, education, and community outreach.

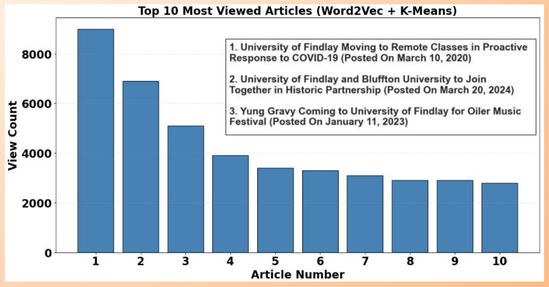

Figure 7: Top 10 Most Viewed Articles Across All Clusters

This chart shows the 10 UF Newsroom articles with the highest view counts, based on clustering with Word2Vec + K-Means. It highlights which topics drew the most audience attention. These high-performing articles can help inform content strategy by revealing what kinds of news stories engage readers most.

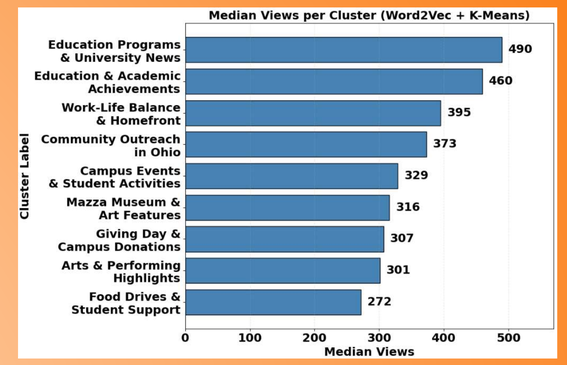

Figure 8: Median Views per Cluster (Word2Vec + K-Means)

Median article views revealed Education Programs & University news - content themes attracted the most readers across UF Newsroom clusters.

CONCLUSIONS

- Lightweight embeddings like Word2Vec outperform heavier models such as BERT on small institutional datasets.
- K-Means delivers compact, well-separated clusters, while Hierarchical Clustering aids interpretability.
- Quantitative metrics aligned with human interpretability, confirmed cluster validity.
- Integrating text analytics with engagement data can help the University tailor news toward the themes that attract the most readers and strengthens data-driven communication strategy.

REFERENCES

- Asudani, R., et al. (2023). Impact of Word Embedding Models on Text Analytics in Deep Learning Environment: A Review– https://pmc.ncbi.nlm.nih.gov/articles/PMC9944441/pdf/10462_2023_Article_10419.pdf
- Zhang, Y., et al. (2024). From Word Vectors to Multimodal Embeddings: Techniques, Applications, and Future Directions for Large Language Models – https://arxiv.org/pdf/2411.05036
- Wang, R. (2023). Revisiting GloVe, Word2Vec, and BERT: On the Homogeneity of Word Vectors – https://www.cs.toronto.edu/~rwang/files/embeddings.pdf
- University of Findlay Newsroom – https://newsroom.findlay.edu/

University of Findlay®