Reproducibility Appendix
Project Report for NLP Course, Winter 2025

Daniel Tytkowski
Warsaw University of Technology
01161612@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION – We use pretrained sentence-level embeddings to represent news articles, followed by unsupervised clustering (k-means, HDBSCAN) for topic discovery. Narrative features (sentiment, subjectivity, bias, framing) are extracted using pretrained transformer-based classifiers and analyzed distributionally.

- LINK TO CODE –https://github.com/tytkowskid/NLP_2025W (project folder: Topic-Discovery-News-Narratives-Tytkowski)

- INFRASTRUCTURE – Experiments were run on a single NVIDIA RTX 3070 GPU (8 GB VRAM) with CUDA support.

  OS: Windows 11

  Python version: 3.10

- RUNTIME PARAMETERS – Semantic embedding extraction: 15 seconds per 1,000 articles

  Narrative feature extraction: 2–6 minutes per feature for 1,000 articles

  Clustering and evaluation: $<$ 1 minute

- PARAMETERS

Task	Model	# Parameters
Semantic Embedding	Sentence-BERT (all-MiniLM-L6-v2)	∼33M
Sentiment Analysis	twitter-roberta-base-sentiment	∼125M
Subjectivity Detection	mDeBERTa-v3-base (subjectivity classifier)	∼184M
Bias Detection	UnBIAS Classifier (RoBERTa-based)	∼125M
Framing Classification	DeBERTa-v3-small NLI (zero-shot)	∼44M
Dimensionality Reduction	UMAP	N/A

Table 1: Approximate number of parameters for models used in the study. All models are applied in inference mode without additional fine-tuning.

- VALIDATION PERFORMANCE – Not applicable. This work does not involve supervised training or validation splits.

- METRICS – Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), cluster purity, silhouette score, and Jensen–Shannon divergence are used. All metrics are defined mathematically in the Evaluation section of the report.

Multiple Experiments:

- NO TRAINING EVAL RUNS – Only the k-means algorithm was evaluated for 10 values of k.

- HYPER BOUND – $2 \leq k \leq 12$

- HYPER BEST CONFIG – $k = 6$

- HYPER SEARCH – 10

- HYPER METHOD – The silhouette score is used as a diagnostic criterion to assess cluster separability across different values of k when selecting the number of clusters for k-means.

- EXPECTED PERF – Clustering results are stable across runs with fixed random seeds; minor variations may occur due to non-deterministic GPU operations.

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS – 12,000 news articles sampled from the 2019 All the News dataset after filtering.
- DATA SPLIT – No train/validation/test split is used, as the study is fully unsupervised.
- DATA PROCESSING –
  - The corpus is restricted to articles published in 2019 to ensure temporal consistency.
  - Articles shorter than 1,500 characters are removed to retain sufficient discourse content.
  - Publisher-specific section labels are normalized into six unified categories.
  - A balanced subset of 2,000 articles per category (Business, Politics, World, Sports, Tech, Movies) is sampled for analysis.
- DATA DOWNLOAD – `https://huggingface.co/datasets/Tytanito/news-narratives-embeddings`
- NEW DATA DESCRIPTION – New data - in this case narrative features: subjectivity, bias, framing, sentiment were collected using pretrained models mentioned before. Semantic embeddings were also produced by pretrained transformer model. Every new feature is already present in the preprocessed dataset available under the link.
- DATA LANGUAGES – English