# Israeli Media Monitor — Clustering Improvement Checklist (Prioritized)

*Generated: 2025-08-28 16:43*

## Top‑5 Quick Wins

* Normalize titles/summaries for CLUSTERING: strip CTA prefixes (e.g., 'צפו:', 'דעה:'), remove site suffixes ('| N12', '- mako'), delete nikud, unify quotes/dashes, lowercase, and collapse whitespace.
* Use title-weighted text\_for\_cluster (or title-only to start): e.g., title + title (+ summary).
* Use Hebrew word-level TF‑IDF with a proper token pattern and your Hebrew stopwords file.
* Tune the distance threshold for precision (start 0.86–0.90); adjust based on singleton rate and cross‑outlet mix.
* Cluster within tight time windows (per day / 48h) to avoid mixing same-topic stories from different days.

## Highest Impact (do these first)

* Normalize titles/summaries for CLUSTERING: strip CTA prefixes (e.g., 'צפו:', 'דעה:'), remove site suffixes ('| N12', '- mako'), delete nikud, unify quotes/dashes, lowercase, and collapse whitespace.
* Use title-weighted text\_for\_cluster (or title-only to start): e.g., title + title (+ summary).
* Use Hebrew word-level TF‑IDF with a proper token pattern and your Hebrew stopwords file.
* Tune the distance threshold for precision (start 0.86–0.90); adjust based on singleton rate and cross‑outlet mix.
* Cluster within tight time windows (per day / 48h) to avoid mixing same-topic stories from different days.

## Medium Impact (strong next steps)

* Try character TF‑IDF (3–5) for short Hebrew headlines; A/B vs word TF‑IDF; then consider combining both (hstack, scale char ~0.5).
* Deduplicate within each source before clustering (exact normalized titles or near-duplicates within short time).
* Precompute cosine distance and use AgglomerativeClustering(linkage='average', metric='precomputed'); avoid 'ward'.
* Refine stopwords/normalization using your corpus: remove generic glue tokens that appear in many titles.
* Adjust TF‑IDF df params: min\_df=1–2; tighten max\_df to 0.7–0.85 if generic clusters appear.
* Protect HH:MM as <TIME> and decide how to treat numbers (<NUM> vs keep digits) based on observed over‑merges/splits.
* Weight title more than summary to reduce style noise from summaries (they vary a lot by outlet).

## Lower Impact (polish and scale)

* Increase sample size for tuning: 150–300 articles (40–100 stories) within 48–72h; for analysis 500–1,500 over ~1 week.
* Evaluate with stable metrics: % clusters with ≥2 sources; % articles in ≥2‑source clusters; singleton rate; avg cluster size.
* Diagnose top terms per big cluster; if they’re generic (e.g., 'צפו', 'חדשות'), extend stopwords/CTAs list.
* Export both CSV and JSONL; keep timestamped runs so you can A/B compare thresholds and vectorizers.
* Cap per-source items per cluster (optional) to prevent same‑source floods from dominating a cluster view.
* Consider embeddings (Sentence‑Transformers) + cosine + HDBSCAN for a later upgrade once TF‑IDF baseline is solid.