**Lexical vs Semantic Features — Methodology (TextLab)**

**What this note covers**

• Definitions with examples from evaluation/education texts

• How TextLab uses each feature type (Word/Document/Summary pages)

• Tuning knobs you control (and safe defaults)

• Validation patterns to avoid false positives

**1) Quick definitions**

• \*\*Lexical features\*\*: treat text as \*words\*. Count or weight exact tokens/phrases (after lowercasing & accent‑folding).

• Examples: TF‑IDF, BM25, keyword counts

• \*\*Semantic features\*\*: capture \*meaning\* beyond exact words using embeddings.

• Examples: SBERT sentence vectors, cosine similarity, BERTopic clusters

**2) Strengths & weaknesses**

\*\*Lexical (TF‑IDF/BM25)\*\*

• ✅ Precise on exact phrases (“teacher retention”) and rare terms

• ✅ Transparent counts for auditing

• ⚠️ Misses synonyms (“in‑service training” vs “teacher PD”)

• ⚠️ Sensitive to morphology/hyphenation (mitigated in TextLab by accent‑fold & hyphen→space)

\*\*Semantic (SBERT/BERTopic)\*\*

• ✅ Finds paraphrases & cross‑language similarities (with multilingual models)

• ✅ Clusters themes even if wording differs

• ⚠️ Can over‑match if threshold too low

• ⚠️ Requires models; slightly heavier compute

**3) Where each is used in TextLab**

• \*\*Word (Keyword Analyzer)\*\*

• \*Exact mode\*: lexical whole‑word match with accent/hyphen normalization

• \*AI/Semantic\*: SBERT → RapidFuzz fallback → simple inflections; builds a \*semantic vocab\* from your corpus (1–3‑grams)

• Outputs: per‑doc counts + validation \*snippets\* (sentence ±1)

• \*\*Document (Topics)\*\*

• BERTopic (embeddings + clustering) with safe fallbacks to NMF/LDA

• BM25 \*prefilter\* option to focus on keyword‑relevant items before modeling

• \*\*Summary\*\*

• Retrieval uses BM25+TF‑IDF, optional SBERT; summaries strip TOC/heading noise first

**4) Tuning recipes**

• \*\*Exact keyword counts (high precision)\*\*

• Keep \*Exact (whole‑word)\* on; add domain variants manually

• Use validation snippets to confirm

• \*\*Semantic discovery (higher recall)\*\*

• Switch to \*AI/Semantic\*; start threshold at \*\*0.55\*\* (0.45 = broader, 0.65 = stricter)

• Increase \*top expansions\* for more variants; raise \*min phrase count\* to cut noise

• \*\*Focus corpus around a theme\*\*

• Enable \*\*BM25 prefilter\*\*, set Keep top‑N to 400–800

• Then run BERTopic/NMF for cleaner topics

• \*\*Small corpora\*\*

• Prefer NMF with 6–12 topics; ensure min\_df < max\_df (TextLab clamps automatically)

• \*\*Multilingual\*\*

• Use a multilingual SBERT (e.g., paraphrase‑multilingual‑MiniLM) if many non‑EN docs

**5) Validation patterns (avoid false positives)**

• Remove TOC/indices, dotted leaders (“…..”), page numbers, Roman numerals

• Normalize diacritics (Côte d’Ivoire → cote d ivoire); hyphen→space (school‑based → school based)

• For counts, inspect \*validation\_snippets\* column before reporting

**6) Interpreting scores**

• \*\*BM25 scores\*\*: relative ranking, not probabilities; higher = better lexical match

• \*\*Cosine similarity (SBERT)\*\*: 0–1 after normalization; >0.55 often indicates a good semantic match

• \*\*Topic quality\*\*: mean cosine to centroid per topic (higher = tighter theme)

**7) When to choose what**

• Need auditable counts in a specific section → \*\*Lexical\*\*

• Need to catch paraphrases / varied wording → \*\*Semantic\*\*

• Want theme discovery across many docs → \*\*Semantic topics\*\* (BERTopic/NMF) possibly after \*\*BM25 prefilter\*\*

• Best practice: run \*\*both\*\* and compare; lexical for counts, semantic for discovery

**8) Troubleshooting quickies**

• “No hits but I’m sure they exist” → Lower semantic threshold to 0.45; add synonyms; analyze \*\*full\_text\*\* column

• “Too many loose matches” → Raise threshold to 0.65; increase \*min phrase count\*; turn off cross‑language model if N/A

• “Vectorizer error: max\_df 1.001” → Handled by TextLab clamps; if custom code, set `max\_df<=1.0` & `min\_df<max\_df`

**9) Glossary (1‑liners)**

• \*\*TF‑IDF\*\*: word importance scaling by rarity

• \*\*BM25\*\*: search ranking over bag‑of‑words

• \*\*SBERT\*\*: sentence embeddings for meaning

• \*\*BERTopic\*\*: topic discovery using embeddings + clustering

• \*\*Cosine similarity\*\*: angle between vectors, measure of semantic closeness

**10) Actionable checklist**

• Pick the right text column (Findings or full\_text)

• Decide: Exact vs AI/Semantic; set threshold

• (Optional) BM25 prefilter to focus the corpus

• Validate with snippets; export counts/topics for sharing