BLENS: Biomedical Literature Extraction & Scoring System

Tyler Simone, Department of Data Science, University of New Hampshire

University of New Hampshire

Introduction

- No publicly available tool exists for automated, cross-database literature quality scoring for systematic reviews
- BLENS retrieves from PubMed, ClinicalTrials.gov, medRxiv and bioRxiv
- Applies unsupervised + semi-supervised ML (Machine learning), text converted to numerical vectors via Sentence-BERT, recalibrated using expert-validated gold standard studies as anchors
- Validated on Third Molar Extraction (TME) and relevant keywords as the target domain

Objectives

- Build a unified retrieval pipeline across PubMed, ClinicalTrials.gov, medRxiv, and bioRxiv with normalized output
- Score relevance using Sentence-BERT (dense vector representations of text) and cosine similarity (geometric distance between vectors)
- Recalibrate scores using gold-standard studies from a published TME meta-analysis as semi-supervised anchors
- Validate by confirming all gold-standard studies as correctly classified as High tier

Acknowledgements

Special thanks to Marek Petrik and Mark Wingertzahn for their assistance guiding and advising on this project

Methods

- Data Collection: Custom Python Application (B-LENS Query App) fetching records via PubMed Entrez API, ClinicalTrials.gov API v2, and bioRxiv/medRxiv REST APIs, results normalized into a unified schema (.xlsx)
- Text from titles and abstracts encoded into numerical vectors using Sentence-BERT (all-MiniLM-L6-v2); each record scored by cosine similarity to TME topic query.
- Clustering through HDBSCAN (density-based clustering, no preset k) groups semantically similar records; Isolation Forest (statistical outlier detection) flags low-quality or off-topic records using text + numerical metadata
- Semi-supervised recalibration of results using gold-standard studies, records scored by distance to gold cluster in embedding space
- PDF rubric: pattern-matching against weighted criteria from published meta-analysis: RCT design, TME pain model, etc.
- Composite Score: Orig. Relevance (35%) + Gold anchor (40%) + Rubric (25%) normalized 0-1; tiered High (>0.66) / Medium (0.33-0.66) / Low (<0.33)

Results

- 1,795 total records retrieved: PubMed (499), ClinicalTrials (560), medRxiv (442), bioRxiv (294), 2000-2026
- All 5 gold standard studies correctly classified as High tier after recalibration, the primary validation metric. Tightened classification (792 → 645 High)
- Clustering: 3 groups – "molar/extraction/pain", all gold studies, avg composite 0.7; "genome/genetic" n=121, Rxiv only); Noise is 630, 96% preprints
- Anomalies: 144 flagged – bioRxiv 25.2%, medRxiv 15.6%, PubMed 0.2%, ClinicalTrials 0%
- RCTs in recommended set: 327/1058 (30.9%)

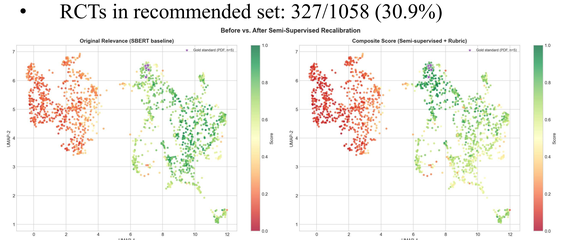

Figure 1: Comparison between original relevancy score baseline and the composite score after gold study semi-supervision is applied. Scores closer to gold studies are closer to 1.0, more relevant

More Results

Figure 2: Displays density of records, records closer to purple stars are weighted higher, axis is for spacing

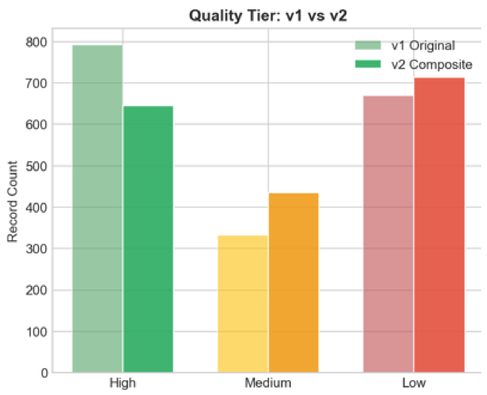

Figure 3: Displays the bins from unsupervised modeling compared against semi-supervised; Semi reduces the number of false-positives (Less High records)

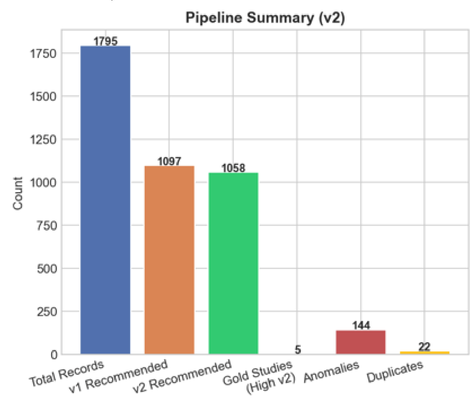

Figure 4: Total count of records is y-axis, x-axis is different groups of studies; recommended is High+Med bins, gold studies is the meta-doc studies, anomalies are records deemed irrelevant, duplicates are studies that overlap from each database

Conclusions

- Composite score is the primary screening metric; records should be reviewed in descending composite score order
- Recall = 100% on gold standard studies – BLENS correctly identifies all expert-validated studies as High quality, confirming the pipeline works as intended
- Sentence-BERT embeddings significantly outperformed TF-IDF, gold study baseline improved from 0.18-0.3 to 0.59—0.72 without any labeled data
- Preprint API limitations are a fundamental retrieval constraint, medRxiv/bioRxiv lack field-level query support, impossible to pre-filter by study design or condition, so irrelevant content is retrieved.
- BLENS reduced 1,795 raw records to: 645 High, 436 Medium, 714 Low records
- BLENS demonstrates successful proof-of-concept performance (100% recalls all gold studies) but requires stricter calibration before deployment, as 645 High-tier records (36% of total) suggest scoring weights are too loose; stronger cutoffs are needed to improve precision

Next Steps

- Implement analysis directly into BLENS application, rather than a separate python script
- Test framework on different models by swapping rubric criteria/anchor set, compare weight changes
- Implement/test stricter rubric thresholds and composite score cutoffs to limit results, reducing manual labor required
- Integrate a large language model (LLM) to automatically generate summaries of top-ranked articles after analysis
- Manual precision audit: choose a sample of recommended records among RCTs and top-ranked records to establish a precision baseline for future weight calibration/model improvement

References

NCBI Entrez Programming Utilities (E-utilities). National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/books/NBK25501/

ClinicalTrials.gov API v2. National Library of Medicine. https://clinicaltrials.gov/data-api/api

Rxivist / bioRxiv & medRxiv REST API. https://api.biorxiv.org/