Thematic Classification of Articles Using Graph Representations

Robert B Barrett[1], Haeun Lee[1], Paul Nagy PhD[1]

[1]Biomedical Informatics and Data Science, Johns Hopkins University, Baltimore, MD, USA

Background

The Observational Health Data Sciences and Informatics (OHDSI) community has significantly advanced observational research by developing open-source tools and methods, including the OMOP Common Data Model (CDM), analytical packages within the Health Analytics Data-To-Evidence Suite (HADES) ecosystem, and comprehensive methodological frameworks [1]. However, systematically identifying scientific articles related to OHDSI and tracking their impact through traditional literature searches is challenging. Keyword-based retrieval often fails to capture relevant studies that do not explicitly mention OHDSI, while manual curation is inefficient and not scalable.

Automated classification of scientific literature has been extensively explored in prior work, using a variety of methodological approaches. For instance, Chowdhury and Schoen [2] employed supervised machine learning methods, such as Naive Bayes, Support Vector Machine (SVM), and Random Forests, leveraging TF-IDF representations of article abstracts to categorize publications into academic disciplines. In contrast, Small et al. [3] utilized unsupervised graph-based clustering techniques on citation networks to identify emerging research topics. Xu et al. [4] further compared multiple unsupervised algorithms—including latent Dirichlet allocation (LDA), k-means, and hierarchical clustering—to detect emerging research areas within large bibliometric datasets.

Topic-specific classification of articles often requires the full text, which may not be available through common retrieval methods (e.g., APIs), often being limited to the abstract alone. To address this, we leveraged a graph-based supervised learning framework that captures both content features and structural relationships that may provide context beyond semantic features alone. Our team aimed to identify scientific articles influenced by the OHDSI community, allowing for more flexible search criteria to capture related articles within a broader reach.

Methods

A list of 704 known OHDSI articles and 877 non-OHDSI-related articles, spanning 2010 to 2025, was used for this study. Articles were retrieved from PubMed databases and enriched for citations, Digital Object Identifier (DOI), and metadata using Entrez and Crossref APIs. Non-OHDSI-related articles were manually reviewed from a search for articles in PubMed using the search criteria: (observational health) AND (patient). All articles were represented in BibTeX format.

From these articles, a graph was constructed using the DOI, author name, and journal name as nodes. Directed edges were created for cited articles, authorship, and the journal they were published. PageRank and degree were calculated as features for articles within the graph. Scores for articles with OHDSI-related authors, citations, and keywords were calculated. Finally, TF-IDF was used to include features from the articles' abstracts. An XGBoost model was used to classify articles, due to its robust performance and ability to extract feature importance.

To evaluate the benefit of graph-extracted features, we further explored the information value of semantic, non-spatial features in abstracts alone by combining classical TF-IDF-based logistic regression with transformer-based SciBERT embeddings and evaluating their effectiveness through stratified cross-validation. Several preprocessing strategies were tested: (a) numeric masking (replacing digits with <YEAR> or <NUM> tokens to avoid trivial numeric cues), (b) removal of numeric tokens, (c) stemming (Porter algorithm), (d) lemmatization (WordNet-based), and (e) stop-word removal (standard NLTK stopword list).

For evaluation against both methods, we withheld 150 positive and 150 negative articles (a roughly 80:20 split), using the remaining documents for the training set. Standard performance metrics including precision, recall, accuracy, and F1-score were calculated from classification predictions.

Results

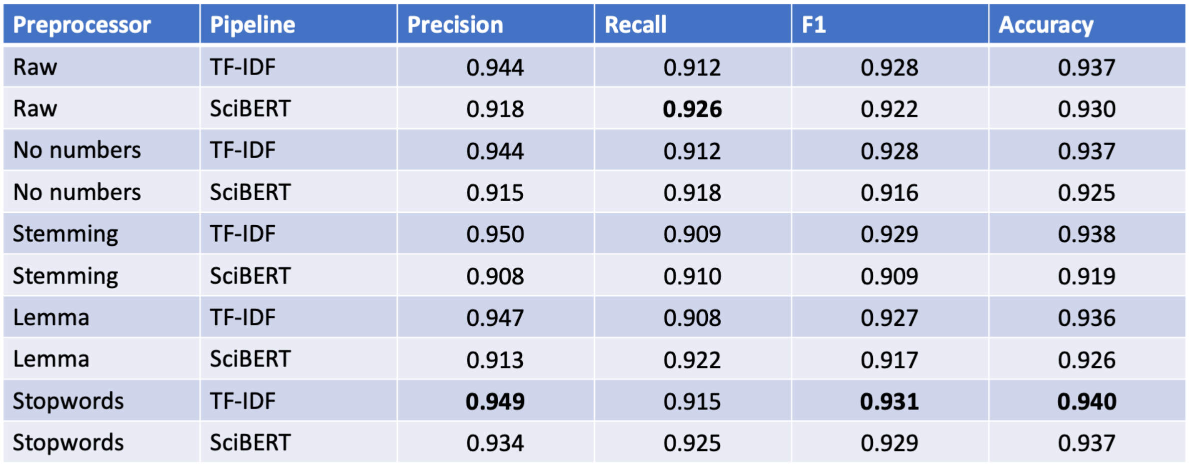
On the 300-article hold-out set the model achieved:

- **Accuracy:** 0.953
- **Precision (PPV):** 0.986
- **Recall (Sensitivity):** 0.920
- **F1-score:** 0.952
- **Specificity:** 0.987 (**TN = 148, TP = 138, FP = 2, FN = 12**)

An analysis of feature importance showed that OHDSI article classification was primarily driven by OHDSI-related-author count and citation-overlap features, as well as TF-IDF-derived signals identifying domain terms such as “OMOP” and “common data model”.

The analysis comparing preprocessing variants of article abstract data with TF-IDF and SciBERT, with logistic regression for prediction produced the following:

Table 1: Performance comparison of abstract only preprocessing strategies and pipelines



Preprocessor	Pipeline	Precision	Recall	F1	Accuracy
Raw	TF-IDF	0.944	0.912	0.928	0.937
Raw	SciBERT	0.918	**0.926**	0.922	0.930
No numbers	TF-IDF	0.944	0.912	0.928	0.937
No numbers	SciBERT	0.915	0.918	0.916	0.925
Stemming	TF-IDF	0.950	0.909	0.929	0.938
Stemming	SciBERT	0.908	0.910	0.909	0.919
Lemma	TF-IDF	0.947	0.908	0.927	0.936
Lemma	SciBERT	0.913	0.922	0.917	0.926
Stopwords	TF-IDF	**0.949**	0.915	**0.931**	**0.940**
Stopwords	SciBERT	0.934	0.925	0.929	0.937

Among the five text normalization strategies, stopword removal consistently improved the discriminative power of lexical models, giving the TFIDF + LogisticRegression pipeline the strongest overall scores. Averaged over five stratified folds, the stopword configuration yielded a precision of 0.949, recall of 0.915, F1score of 0.931, and accuracy of 0.940.

Conclusion

This study demonstrates that a supervised pipeline, including both TF-IDF and graph-derived features, with XGBoost can reliably identify OHDSI-related publications while retaining high specificity. The inclusion of graph-extracted features demonstrated significant improvement over simpler TF-IDF and SciBERT embedding-based predictions alone, with greater precision (0.986 vs. 0.949), greater accuracy (0.953 vs. 0.940) and greater overall F1 (0.952 vs 0.915). These results highlight the value of meta-relationships between articles in their authorship and citation network context.

Limitations for this study existed in the availability of data through PubMed/Entrez APIs, where not all articles have retrievable full-text data. For this reason, the dataset was limited to metadata (e.g., authors, title, keywords) and the abstract. For this reason, OHDSI-specific terminology may not be observed within the available context. Moreover, citation network capture is an imperfect process, and was not always available from the process used.

This work provides a practical and scalable approach for automated tracking of scientific contributions relevant to specific research communities, such as OHDSI. The demonstrated effectiveness of simple NLP pipelines, with observed benefit from graph-based features, suggests broad applicability to similar classification tasks in other specialized domains. This application is particularly promising for topics with well-defined and unique vocabularies. Additionally, our analysis highlights the importance of vocabulary coverage and preprocessing methods in domain-specific literature classification tasks.

Acknowledgments

This research was supported by the NLM, part of the NIH, under award number T15LM013979]

References

1. Voss, Erica A., et al. "Feasibility and utility of applications of the common data model to multiple, disparate observational health databases." *Journal of the American Medical Informatics Association* 22.3 (2015): 553-564.

2. S. Chowdhury and M. P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," *2020 Intermountain Engineering, Technology and Computing (IETC)*, Orem, UT, USA, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249211.

3. Small, Henry, Kevin W. Boyack, and Richard Klavans. "Identifying emerging topics in science and technology." *Research policy* 43.8 (2014): 1450-1467

4. Xu, Shuo, et al. "Emerging research topics detection with multiple machine learning models." *Journal of Informetrics* 13.4 (2019): 100983.

5. Akritidis, Leonidas, and Panayiotis Bozanis. "A supervised machine learning classification algorithm for research articles." *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. 2013.