Mapping the Research Frontier: Multigranular Topic Modeling and Interactive Visualizations for Interdisciplinary Science

Stephen Liu, Trisha Gite, Dr. Aditi Mallavarapu, Jason Ronallo

NC STATE UNIVERSITY

NC STATE University Libraries

NC STATE UNIVERSITY Computer Science

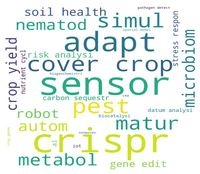
Introduction

Motivation:

- Many social and global problems require **interdisciplinary perspectives** that combines ideas and methods from different fields to **solve complex problems together.**
- Mapping **thematic evolution** and **collaboration patterns** can uncover gaps and opportunities across disciplines.

Problem:

- Analyzing and documenting interdisciplinary research requires accounting for **diverse disciplines**, supporting **multiple granularities**, and needs.
- Different needs demand **flexibilities** in the kinds of **metrics** and **topic granularities.**



Unigram word cloud of the machine detected topics

Rich and contextual manually created topic model by N.C. PSI staff

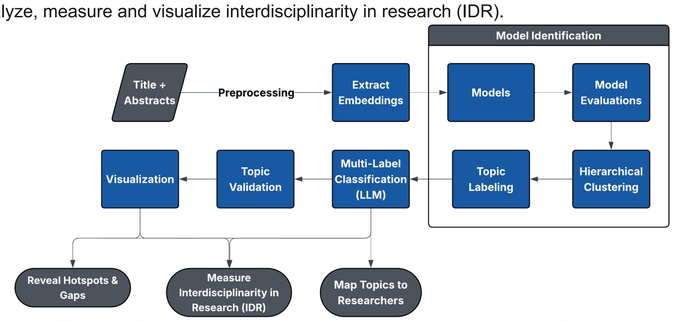
- Several platforms offer interfaces that highlight thematic patterns and provide metrics but operate as blackbox and do not cover many granularities.

Significance:

- Programs like **Office of University Interdisciplinary Programs** (OUIP) at NC State University, and initiatives like NSF RETTL motivate intentional interdisciplinarity in research
- However, researchers, administrators and students affiliated with these programs **struggle to find interdisciplinary partners** and **gauge interdisciplinarity** of their efforts.

Contributions

A flexible and multi-granular **framework** and complementary **computation pipeline** to map, analyze, measure and visualize interdisciplinarity in research (IDR).



Proposed schematic of the computational pipeline to fine-tune machine learning and AI models

The pipeline:

- Identifies contextual parameters for the embeddings and fine-tune machine learning and AI models.
- Compares outputs from multiple models to find best-fit-flexibility in granularity.
- Names machine derived and validated topics with LLM guided by humans.

Potential Implications:

- Map topics to researchers and identify collaboration patterns
- Measure and characterize IDR in terms of expertise diversity and intensity
- Reveal knowledge topic hotspots, gaps and evolutions overtime.

Case Study

- N.C Plant Sciences Institute (PSI) established in 2021 envisions to tackle grand challenges in agriculture by fostering partnerships for developing workforce and cutting edge technology in unique cross-disciplinary domains.
- We analyze publication and funding proposal texts from researchers affiliated with **N.C. PSI** as our first case-study.

Data Description:

- > 3000 publications were retrieved from OpenAlex
- Data was preprocessed to retrieved abstracts for open-source papers
- Usable data included ~2500 titles and abstracts
- The titles and abstracts were used to **extract embeddings**

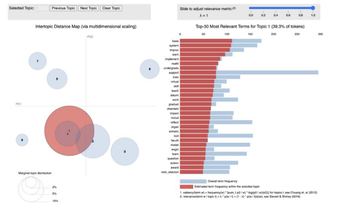
Authors	Name of People Involved with the Particular Publication
Title	Name of Publication
Abstract	Summary Description of the Publication
DOI	Unique and Persistent Alphanumeric String Assigned Each Paper
Year	Year the Paper was Published

PSI Dataset's Data Dictionary

Model Identification

Classical Machine Learning:

- **Latent Dirichlet Allocation (LDA):** Assumes each document is a mix of topics and each topic is a mix of words to identify model .
- **Latent Semantic Analysis (LSA):** Identifies words based on their semantic meaning



Circles represent clusters indicative of topics and bar chart represents frequency counts for lemetized unigram keywords in the highlighted cluster as modelled by LDA

Analysis:

- LDA: Data was preprocessed with, lemmatization, stemming, and n-grams
- LSA: Cleaned data, tested unigram, bigram, trigram patterns

Results:

- LDA: Produced moderately interpretable clusters and captured some low-frequency but meaningful terms, often misaligned with human-coded themes.
- LSA: Generated too many topics for effective validation.

Results highlight **coherence scores and variance metrics alone are insufficient for finding optimum number of clusters** in complex interdisciplinary settings calling for more sophisticated and context-aware topic modeling approaches.

We compare general-purpose and scientific-domain embedding models to assess generalizability across disciplines and establish a framework applicable beyond scientific publications.

General Topic Embedding Models:

- **Nvidia - Llama-Embed-Nemotron (8B):** Fine-tuned from Llama-3.1, this 8B-parameter bi-encoder model uses contrastive learning to produce high-quality embeddings.
- **Qwen - Qwen3-Embedding (0.6B):** Efficient mid-size embedding model balancing computational efficiency with strong performance across multilabel classification and clustering tasks.
- **Google - Embedding-Gemma (300m):** General-purpose embedding model capturing contextual meaning and word relationships across large corpora.

Rank	Model	Num Parameters	Multilabel Classification	Clustering
1	nvidia/llama-embed-nemotron-8b	7B	29.86	54.35
5	Qwen/Qwen3-Embedding-0.6B	595M	24.59	52.33
9	Google/EmbeddingGemma-300m	307M	24.82	51.17

Models under consideration based on the Massive Text Embedding Benchmark (MTEB) Hugging Face leaderboard

Scientific Topic Embedding Models:

We selected the following three embedding models to compare performance:

- **AllenAI- SciBERT:** Transformer model trained on scientific text for capturing domain-specific language and terminology. [1] This model is not part of the benchmark models.
- **Malteos - SciNCL:** Contrastive learning model fine-tuned on scientific literature to enhance semantic similarity between related research topics. [3]
- **AllenAI- Specter2:** Embedding model designed for scientific papers, leveraging citation and abstract context to model document-level relationships. [4]

Model	SciRepEval In-Train	SciRepEval Out-of-Train	SciRepEval Avg
SciNCL	55.6	73.4	67.5
Specter2 Adapters	62.3	74.1	71.1

Scientific Representation Evaluation (SciRepEval) Benchmark [4]

BERTopic:

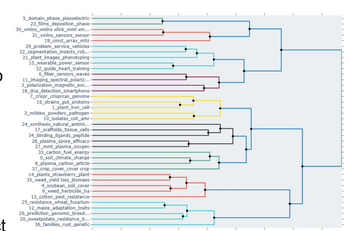
Combines embeddings with clustering to generate interpretable topics [2].

Model Evaluations:

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDB-SCAN): Groups documents based on density, identifying meaningful topic structures and filtering out noise. We use this to find thresholds to create manageable number of relatable topic clusters.

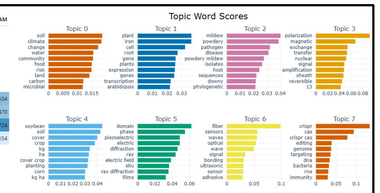
Token-level Topic Distribution: The keyword level probability distribution is used to find threshold for how many topics could be accommodate for multi-class labeling.

Topic Word Scores: Comparison of the most distinct terms within and across topics to validate topic interpretability.



Topic hierarchy visualized through hierarchical clustering showing the organization of machine-identified topics at multiple granularities.

Token-level topic contribution visualization. Each cell represents a token's relative weight toward a specific topic.



Bar charts show the top words and their word scores for several topics. Showing the most representative terms and word importance across topics.

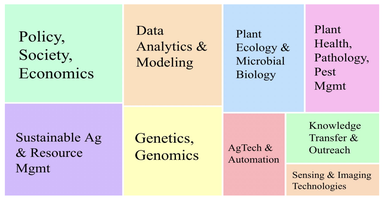
Topic Validation:

- **Human-AI Validation:** A small subset labeled by human domain experts to directly compare human and model-generated topic label for accuracy and interpretability.
- **Two Coder Validation:** Multiple independent humans review model-generated topic label for accuracy and interpretability

Visualization

To enhance the value of the information extracted from the analyses, currently we are working with PSI program manager to create human-centric interactive tools that present the data in a usable manner.

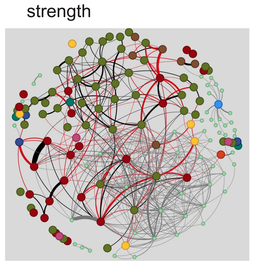
We detail some prototypes below:



Tree map of PSI research by publication volume based on manually created topic model

Interactive race chart highlights evolving priorities:

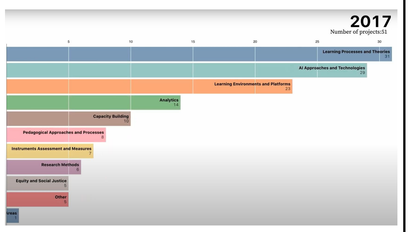
- Identify expertise that changed priority over time
- Identify expertise that has constantly been a strength



PSI Collaboration Graph

Interactive tree map to identify hotspots and coldspots:

- Coldspots (smaller boxes) are potential areas of recruitment and external collaborations
- Hotspots (larger boxes) are stronger areas of contribution
- Being designed as a potential tool for people to find intentional interdisciplinary collaborators.



Prototype from another dataset for interactive bar chart showing changing priorities through the decade [6].

Interactive network map showing collaboration between people affiliated with PSI, nodes represent people and edges represent strength of collaborations over the years

- Identifies teams and people that have reinforced their interdisciplinary expertise over time and those that are constantly bringing in new members and forming new collaborations, demonstrating what can be called versatile expertise.

Future Work

- Apply lemmatization to normalize keywords and reduce morphological variations for more robust topic representations.
- Integrate large language models to augment and contextualize topic keywords
- Explore using funding solicitations and local research abstracts to perform a more targeted gap analysis, identifying where existing expertise aligns or doesn't with funding opportunities.

Acknowledgements

Rob Dunn, Lauren Maynard, N.C. PSI for guiding this effort. Trisha Gite is supported by PI-SUES (CSC) REU.

References

[1] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. ACL.
[2] Grootendorst, M.R. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv, abs/2203.05794.
[3] Ostendorff, M., Risch, J., Rehm, G., Gipp, B., & Sataslaatti, K. (2022). Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings (SciNCL). Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[4] Singh, S. et al. (2023). SciRepEval: A multi-format benchmark for scientific document representations.
[5] Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833
[6] Interactive race chart. Retrieved from: https://observablehq.com/@aditimallavarapu/scrubber-interaction-nsf-categories#chart_time