ABSTRACT

ZHANG, ANGELA. PTNScan: Personalized Topic and Novelty Scan, a Hybrid Clustering and LLM Framework. (Under the direction of Christopher G. Healey.)

Determining publication readiness is often a complex decision-making process that requires a lengthy, manual, and time-consuming literature search and review. We propose a multi-stage analysis to provide a topic-based readiness assessment, asking, "Which topics did the author potentially overlook?" and "Which new topics did the author propose that are novel?" Our approach combines unsupervised text clustering with large language model (LLM) reasoning to provide concise, qualitative data that supports the subjective decision-making process of evaluating publication novelty and completeness. To the best of our knowledge, our work is the first to combine both literature and peer reviews. A multi-stage pipeline: (1) hierarchically clusters published paper abstracts from a target research domain; (2) extracts topic keyword sets for each cluster; (3) compares the abstract for a paper under consideration for publication to match its abstract to the most semantically similar published paper cluster; and (4) provides a "personalized" assessment of topic coverage and novel topics in the user's abstract, allowing them to decide whether they have appropriately considered relevant topics in the research domain, and whether sufficient novelty is present to warrant submission. A novel topic clustering algorithm based on minimum spanning tree clustering and Jaccard similarity is presented to reduce cluster size and minimize outliers while maintaining the semantic content within each cluster. Customized LLM prompting is developed to compare topic keywords extracted from a cluster to a user-supplied abstract. Abstracts from the most similar cluster are used to personalize an assessment of the topic coverage of the user's abstract, identifying appropriate acknowledgement of current, state-of-the-art approaches through topic overlap and potentially novel contributions reported as unique topics in the user's abstract. Our approach ensures a comprehensive review of a target research domain and identifies the potentially novel contributions of the user's publication. Results from both experimental and anecdotal evaluations are positive, suggesting our personalized topic and novelty scan can significantly reduce the time and effort required to determine whether a paper is ready for submission, and to highlight its novel contributions.

PTNScan: Personalized Topic and Novelty Scan, a Hybrid Clustering and LLM Framework

by
Angela Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2026

APPROVED BY:

James Lester	Jung-Eun Kim
Rada Chirkova	Christopher G. Healey Chair of Advisory Committee

DEDICATION

To the wonderful community in my hometown of Portland, Indiana: your decades of support mean the world to my family, and the Jay County education system was the first home to cultivate my inspiration.

BIOGRAPHY

Angela Zhang earned her bachelor's degree in Industrial Engineering from Purdue University and pursued a career in financial technology. Her professional experience motivated her to pursue a Ph.D., and in 2019 she enrolled in the Computer Science Ph.D. program at North Carolina State University, focusing on artificial intelligence and text analytics.

ACKNOWLEDGEMENTS

I would like to express sincere gratitude and appreciation to my advisor Professor Christopher Healey for his feedback, guidance, and support. I'm thankful for all he taught me within and beyond this thesis.

I am also deeply thankful to my committee members, Professor James Lester, Professor Rada Chirkova, and Professor Jung Eun Kim for their valuable feedback, advice, and inspirational research.

Lastly, I'd like to acknowledge my dearest family and friends for their instrumental emotional and logistical support: My family (especially my parents), Mark, Fogo, James C., Jim and Debbie W., and Bartholomew.

AUTHORSHIP STATEMENT

The author, Angela Zhang, conducted all the research components within this thesis: literature review, data collection, experiments, result interpretation, and writing the original draft of the manuscript. The author's advisor, Dr. Christopher Healey, provided research and experiment guidance and editorial feedback for this thesis. Generative artificial intelligence tools were used in the contexts described explicitly in each relevant chapter. This thesis did not use generative AI to write any portion of the final submission.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Research in this thesis falls within the broad categories of machine learning, text analytics, and large language models. We present PTNScan, a Personalized Topic and Novelty Scan framework to identify both common and unique aspects of an individual document relative to a collection of reference documents, all from a common domain. The practical problem we address is assisting an author's attempt to determine: Does a potential submission to a conference or journal: (1) properly address important topics or areas of current interest in the domain? and (2) contain new or novel work that would support a recommendation for publication? PTNScan is not designed to return a Yes–No answer or a probability that estimates the likelihood of acceptance. Instead, our framework highlights topics in previously accepted papers that are not addressed in the author's work, as well as findings in the candidate article that are not commonly found in the existing literature. The author can then utilize their domain expertise and knowledge of the research area to assess whether critical material is missing from their submission and whether its unique aspects represent the novel contributions the author is proposing. Implementing a mixed initiative approach avoids attempts to automatically categorize a submission, allowing text analytics methods and a user to collaborate in ways that leverage their unique strengths: the ability of a computer algorithm to rapidly analyze massive amounts of text based on guidance from the user (e.g., an LLM prompt), and the user's ability to quickly address subjective questions (e.g., is this unique topic a novel contribution or

noise in the submission), manage ambiguity, identify patterns that are difficult to represent in a structured fashion, and apply domain knowledge in an efficient and effective manner. The result is an approach that combines the strengths and avoids the limitations of the participants, computer algorithm and human user, producing results that exceed what either participant would be capable of achieving on their own. Evaluating whether research is publication-ready can be highly subjective, as researchers must assess, often without structured guidelines, prerequisites to ensure the paper is suitable for a top-tier venue. This includes evaluating the novelty and relevance of the proposed research to the current domain of the paper, the soundness of the methodology, and the impact of its findings. Such an evaluation requires identifying relevant past work, but conventional methods (e.g., search engines) can introduce subject bias due to an over-reliance on citation counts or manually tagged keywords. This can over- or under-weight an existing paper's relevance regardless of recency or affiliation of the author(s). As an example, younger authors who produce excellent work but do not yet have a strong citation record may be less frequently cited than older, more established researchers. Sites like `connectedpapers.com` exist, but they sort by cited references, often excluding important, relevant work.

PTNScan represents an analysis pipeline that leverages automated topic modeling to streamline the current, mostly manual, literature search and review process. It is designed to efficiently address the challenge of identifying related research and comparing it to a potential submission. We combine unsupervised clustering with post-processing methods, using Jaccard similarity, to analyze large collections of papers and uncover common patterns and topics. These cluster topics are then used to match against topics present in a user-provided abstract.

Several approaches are used to perform the cluster–user abstract matching, including large language model (LLM) zero-shot topic prompting. An LLM is prompted to identify topics and corresponding topic keywords. Two techniques are used to address the issues of scale and token limits within an LLM. First, papers are represented by their abstracts, acting as human-generated abstractive summaries of a paper's content and its most important contributions. Second, we developed a novel algorithm, MST-Jaccard sampling, to select a small subset of representative abstracts from large clusters.

Once topic clusters are constructed and topic keywords are assigned to each cluster, a submission's abstract can be compared to the clusters for semantic similarity. This allows us to identify overlaps and outliers: topics common to both the published paper abstracts and the proposed submission's abstract, topics unique to the cluster abstracts but not included in the submission abstract, and topics unique to the submission abstract but not included in the cluster's abstracts. By automatically identifying relevant and important topics in previously published work, as well as novel topics in the abstract of a paper under consideration, our technique streamlines the self-

evaluation process and provides feedback to support a determination of whether the proposed work is publication-ready. It addresses our goal of efficiently enhancing a researcher's ability to gauge the novelty and relevance of their findings against an existing body of research.

Our framework is completed in the following research steps.

1. Convert published article abstracts from a target research domain into semantic text vector representations, then cluster the vectors to form common topics within the research area.
2. Minimize outliers and manage cluster size by constructing topics at both global and local granularities using hierarchical clustering approaches.
3. Intelligently compress large clusters using MST-Jaccard sampling, a technique that selects a subset of representative abstracts from large clusters.
4. Perform semantic modeling of a potential submission's abstract.
5. Compare the semantic representation of the submission abstract to the cluster hierarchy to identify the most similar topic cluster.
6. Apply an LLM-RAG analysis to generate a "personalized" summary of the overlaps and outliers between the most similar topic cluster and the submission abstract.
7. Report the LLM's findings to the user, allowing them to assess their submission's publication readiness.

Our work provides the following novel contributions.

- A method to identify semantic topic overlaps and outliers between a document corpus and a target document,
- Methods to choose cluster sizes that balance topic coherence, topic inclusion, and outlier count within a document corpus,
- A novel algorithm, MST-Jaccard sampling, to select a subset of representative abstracts and compress large clusters while maintaining topic representations,
- A topic-based RAG method to construct input for an LLM that focuses its output on specific keywords or topics during the summarization of similarity and uniqueness between a document abstract cluster and a user-provided abstract, and

- A combined literature scan and personalized feedback on a potential paper submission, presented in a way that allows an author to apply domain expertise and an understanding of the research area to evaluate a potential submission's strengths and limitations.

The remainder of this thesis includes the following material. First, we review background research on LLMs, dimensional reduction, Jaccard undersampling, hierarchical topic clustering, semantic similarity, and RAG-based summarization of similarity using LLMs. Next, we present methods to cluster a corpus of published paper abstracts, together with large cluster post-processing experiments that compress clusters as needed to achieve single-topic precision. We present and evaluate several topic-matching algorithms designed to calculate the similarity between keywords representing a cluster's topics and a user-provided abstract. Finally, we investigate multiple topic-enhanced LLM prompts mean to perform topic coverage assessment and to detect novelty in the user abstract. We conclude with an analysis of the strengths and limitations of PTNScan, together with recommendations for future work.

CHAPTER

2

BACKGROUND

Automating the interpretation of domain-specific terminology is often a challenge in many domains. Methods such as LLMs often overlook these terms, which lead to inaccurate semantic retrieval of related literature. In the scientific realm, this over-generalization of similar articles can cause bias when interpreting novelty and relevance for research publications. The following background is intended to introduce recent LLM techniques that act as an automated research aid (e.g., automated literature review peer review). Then, we explain concepts for each step of our end to end steps: semantic document cluster retrieval, topic analysis, and automated, RAG-guided LLM topic identification of novel and relevant peer review publications.

We provide background on the technical challenges resulting from processing unstructured text with domain-specific terms: data sparsity and class imbalanced datasets. Next, we introduce Jaccard similarity for a novel class balancing algorithm: Jaccard undersampling. Unsupervised learning (clustering) is used with dimension reduction and topic labeling techniques to semantically cluster relevant research papers. This chapter concludes by incorporating key topics into a RAG-based LLM prompt designed to return novel and relevant topic contributions. We formulate our definitions for follow-on assessment of topic novelty and relevance in scientific literature publications.

2.1 Automated Literature and Peer Reviews

Effective literature reviews enable efficiency in identifying relevant topics, and constructive peer reviews provide guidance with meaningful feedback for publication improvement. There is recent interest in using LLMs for both tasks [Guo23; Mu24]. Experiments on automating literature reviews test an LLM's effectiveness in organizing research articles into hierarchies, scanning related literature against specified criteria, and authoring literature surveys. Research on automating peer reviews leverage LLMs to comment on the clarity and depth of research using constructive, polite tones. To the best of our knowledge no existing work has combined both automation approaches in a single pipeline.

2.1.1 LLM Automation in Literature Reviews

LLM-automated literature reviews enable efficient scanning of hundreds of thousands of candidate documents through prompt engineering with relevant phrases and keywords. Some authors have created curated datasets to test specific features such as ability to categorize or generate hierarchical structures. Significant focus has been placed on prompt engineering with key phrases to find related literature for more abstract ideas, and screening against formal guidelines (e.g., PRISMA). Recent literature also evaluates an LLM's relevance for primary or sole authorship in literature reviews [Zhu25].

Hsu et al. curated a dataset, CHIME, focused on biomedicine for studying an LLM's potential for organizing scientific research papers into hierarchical organizations [Hsu24]. They define hierarchical organization as "tree structures where nodes refer to topical categories and every node is linked to the studies assigned in that category." Their findings, using human-in-the-loop evaluators, were that while LLMs are capable of generating the appropriate hierarchical structures, it is less accurate at assigning studies to the correct categories. Thus, they further developed a corrector model using human feedback to improve performance and released data and models to encourage further work in this area.

Noe-Steinmuller et al. seek to find a representation of theoretical views of pain-related suffering, cross-evaluated with topic modeling and the use of GPT LLMs [NS24]. "Topic modeling was applied to the text corpus [of verbatim quotes regarding pain-related suffering] to cross-validate the different dimensions of pain-related suffering..." Twenty-five words before and after the keyword "suffering" were captured, representing approximately two sentences before and after the term, to develop 8,910 matching strings. GPT was then used to define suffering in an a-priori approach based on these matching strings, forming definitions of the current string using prior string sequences. The end

result of their systematic topic modeling and GPT-generated definition of suffering validated a conceptual framework with eight discrete dimensions: social, physical, personal, spiritual, existential, cultural, cognitive, and affective.

Scherbackov et al. used GPT to evaluate literature reviews that were written in part or in whole by an LLM [Sch24]. They retrieved 3,788 articles and through GPT-based exclusion prompting removed articles that did not explicitly use LLMs for writing the review. This produced 172 articles that were retained for their study. They noted, of the articles retained, only twenty-six captured reviews that explicitly used LLMs for help in writing. Metrics were presented on how these reviews were conducted, specifically, whether abstract capture correctly matched the objective topic: "Of the 172 citations, 79 (45.9%) reported common metrics like Accuracy, Precision/Recall, and F1, while 36 (20.9%) used less common metrics like G-score and Jaccard similarity. The remaining 57 publications (33.1%) relied on qualitative assessments."

Li et al. explored LLMs' effectiveness (including ChatGPT v3.5, ChatGPT v4.0, ChatGPT v3.5 turbo, ChatGPT v4.0 turbo, Google PaLM 2, Google Gemini 1.0 pro, Meta Llama 2, Meta Llama 3, and Claude 3) in performing abstract screening [Li24]. Their prompt provides a specific topic with inclusion and exclusion rules based on the PRISMA format, together with an abstract. GPT is asked to classify whether the abstract belongs to the target category, either Yes or No. They found that ChatGPT v4.0 was more than 90% accurate. However, this is limited to situations where it is possible to provide specific topics and assess each abstract for each topic.

Wilkins developed GPTscreenR to automatically screen abstracts using a chain-of-thought technique and a Population, Concept, and Context (PCC) framework [Wil23]. "GPT-4 is instructed to summarize the inclusion criteria for the scoping review, compare the title and abstract against these summarized criteria, and make a final recommendation on whether to include or exclude the source." GPTscreenR achieved an accuracy of 83% compared to Li's 90%. Thus, it is unclear whether providing a chain of commands for literature reviews has any benefit over a zero-shot prompt.

2.1.2 LLM Automation in Peer Reviews

Several criteria are normally combined to formulate a constructive peer review. Feedback clarity provides the author with a better understanding for course correction. Polite communication encourages collaborative environments that respect the author's work. In addition, subject matter understanding and novelty evaluation are critical for proper content evaluation. Several previous works have evaluated these components.

Biswas et al. describe several methods where authors can use GPT to form polite and appropriate responses for paper reviews [Bis24]. LLMs can offer responses that are stated in respectful

tones to ensure that every review point is addressed, reword sensitive explanations so they are not confrontational, and design a plan for how to incorporate reviewer comments. Ethical limitations are also discussed such as ensuring academic integrity, determining intellectual authorship, and maintaining originality in responses.

Biswas et al. also evaluated ChatGPT's performance in acting as a peer reviewer [Bis23]. While GPT was noted for its efficiency and its structured arguments for improving the writing's organization, "ChatGPT might struggle with in-depth comprehension and contextual understanding of highly specialized or nuanced scientific topics." Furthermore, nuanced discussion regarding specific feedback clarity is better exchanged between the author and reviewer, as GPT may not have the same perspective towards subjective components compared to actual reviewers.

Lin et al. introduce a benchmark to evaluate an LLM's ability to assess novelty in scholarly papers [Lin25]. They use a temporal assumption that novel papers will be more recently published. With this assumption, they formed 15,000 pairs of papers across six fields sampled from the arXiv dataset with publication dates spanning from two to ten years apart. Their expectation was that an LLM would declare a recent paper as more novel. Similar to our work using unsupervised learning to cluster similar documents, they developed RAG-Novelty, which leverages the retrieval of similar papers to assess novelty. Their experiments were performed across various LLMs and showed significant disparities across the different models. The primary finding stated in their work is that the limited context provided by a paper's abstract may prevent accurate novelty evaluation.

2.2 Outlier Challenges

Clustering scientific data without specialized preprocessing can result in the majority of samples being classified as outliers. This can be attributed to data sparsity issues caused by unique, domain specific terminology. The more specialized the term, the less likely it will appear across multiple abstracts. Too many specialized terms in an abstract forces it to be classified as an outlier. This impacts the performance of an LLM, as there is less available public training data to capture each domain's specialty.

2.2.1 Data Sparsity in Scientific Literature

To meet the "novelty" criteria for research publication, a paper's description of the most important contribution, interesting subtopic, or unique problem statement may result in minimal overlap with other papers. This underscores the data sparsity problem commonly found in text modeling, where the majority of a term–document matrix's entries are zero due to minimal overlap in vocabulary.

This difficulty also extends to the creation of new training data for models that analyze scientific publications, as a productive research environment is continually innovating with new vocabulary and growing subdomains [Don24].

2.2.2 Web Crawl Limitations

One potential source of specialized vocabulary is web resources that focus on the target domain. Web crawling is a common method for locating and extracting text contained in these types of online repositories. For example, Common Crawl is a free, open source web crawl dataset that is one of the most commonly used web-specific resources [Com]. Even so, Common Crawl is limited in learning esoteric or recent vocabulary from emerging domains due to two factors: paywalls and focus. Most research publications are blocked by paywalls, restricting any free form web crawl from accessing its content. There are sites like arXiv that are free, but this limits content to the fields and authors who prioritize publishing on that site in addition to (or in place of) the target conferences [Arx]. Also, arXiv is not peer reviewed. A web crawl typically scans for information that contributes to general conversations and vocabulary: blogs, news articles, and social media conversations, for example. It does not focus on capturing information from a target research domain. Thus, the vocabulary captured by Common Crawl may be more suitable for general conversation interpretation rather than scientific knowledge.

2.2.3 Embedding Model Limitations

Embeddings are a common method of representing text as a high-dimensional vector, where vector simliarity corresponds to semantic similarity. Unfortunately, high dimensional data produces challenges when a model is trying to identify patterns, relationships, and context between words. Sparse data can also hinder dimension reduction techniques. This leads to a model producing less accurate embeddings, or embeddings biased towards limited contexts it can understand and overfitting of more general, high frequency context and vocabulary.

Context overfitting limitations where a model “over learns” and generates specific contexts that are biased by over-representation in the training data causes failure or poor generalization of real-world or testing data. This can be a challenge for scientific terms that have broader themes or applications. For example, “genetic algorithms” is used in many fields beyond the standard biotechnology context. Researchers and industries have learned how to utilize this in improving network efficiency, social science, and even financial simulations. However, if the training data is of a specific subject (e.g., natural selection), then it may fail to capture the use of this phrase in other subjects. Even with adequate training data, models may struggle to accurately embed domains that

have many rare or specialized terms.

2.2.4 Scientifically Trained Fine-Tuning Models and LLM Limitations

Bidirectional Encoder Representations from Transformers (BERT) is a deep bidirectional transformer with word tokenization [Dev19]. It learns text predictions through masked language modeling (MLM) and next sentence prediction. The bidirectional transformers will predict masked words based on surrounding context. BERT is trained with more general vocabulary from Wikipedia and BookCorpus. RoBERTa is an updated variant of BERT that is trained with more data (adding CC-News, OpenWebText, and Stories), uses byte pair encoding, increases batch sizing for stronger generalization, removes next sentence prediction, and adds dynamic masking where different subsets of tokens are masked during pretraining [Zhu21]. SciBERT [Bel19]and INDUS [Bha24] are LLMs based on BERT and RoBERTa that are fine-tuned with scientific literature.

SciBERT updates BERT with a scientific corpus of a size similar to the original training dataset (3.17B articles compared to BERT's 3.3B): 18% from computer science papers and 82% from biomedical sciences. Fine-tuning did increase accuracy for computer science retrievals by 7% in relation classification, but biomedical science retrievals had marginal improvements. INDUS follows the RoBERTa architecture with a customized byte pair encoding tokenizer trained from its curated scientific corpora, a retriever, and scientific evaluation datasets. INDUS marginally improved performance over SciBERT but not by more than 5% in any evaluation. These small improvements, despite great efforts on retriever and dataset customization, highlight a foundational problem with LLMs. Despite fine-tuning and custom curation of training datasets, LLMs will always rely, even for specialized vocabulary, on text frequency. Thus, specialized vocabulary that are not common compared to more general terminology may always be missed when fine-tuning LLMs, making it difficult for text personalization tasks such as summarization or retrieval in specialized domains. It is important to note that for smaller models, specialized vocabulary in fine-tuning does make a notable impact. In this type of environment, INDUS improved 38.6% in document classification over TinyBERT's performance.

2.3 Text-to-Vector Representations

In unsupervised learning of text representations, text is first processed into numerical, high dimensional vectors. These vectors can either be frequency counts for each word (e.g., bag-of-words), phrases (e.g., n-grams), or embedding spaces that capture semantic comparisons across the entire document or document collection. Pre-trained embeddings such as GloVE and Word2Vec are

commonly used to capture meaningful patterns. These approaches, however, are limited by the need for high term frequency. Thus, adequate processing of scientific text requires embedding transformations from larger models (e.g., BERT or GPT).

2.3.1 n-Grams, Bag of Words, and Word Embeddings

Bag-of-words divides sentences or phrases into individual words. n-grams represent a multiple-word phrase with n representing the specified count of words. Using n-grams is most beneficial if context is dependent on phrases contain n consecutive words, for example, "machine learning" is different than the individual words "machine" and "learning." This is especially impactful for non-sequential models. These considerations are most relevant for foundational tools such as term frequency-inverse document frequency (TF-IDF) or Latent Dirichlet Allocation (LDA). Trained LLM embeddings typically use a tokenizer like Byte Pair Encoding (BPE). BERT uses wordpiece tokenization, which breaks words into subwords before embedding. This enables a better understanding of specialized vocabulary by learning representations from subword patterns (i.e., by identifying "word roots.")

2.3.2 Unsupervised Pre-Trained Embeddings

Global Vectors for Word Representations (GloVe) is a log-bilinear regression model that combines global matrix factorization with local context window methods [Pen14]. It compacts a sparse matrix into a dense "word–word co-occurrence matrix" and trains only on the nonzero elements. "The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition."

Word2Vec extends the skip-gram model by subsampling frequent words and introducing negative sampling [Mik13]. A feature of this algorithm focuses on improving vector representations for frequent phrases instead of individual words. The skip-gram model identifies phrases by capturing frequent co-occurring words and distributed representations within a context window. Negative sampling focuses on smaller, more frequent subsets of large pool negative samples used to compare against positive samples. Increasing the sample frequency of negative samples adds model complexity during training, as negative samples are less distinctive from positive samples and often resemble text from the real world.

Both GloVe and Word2Vec prioritize word frequency in their solution to data sparsity. GloVe utilizes a weighted least squares objective to prioritize frequent co-occurrence over rare co-occurrence during training, assuming that rare co-occurrences could be more indicative of noise. Word2Vec

utilizes the negative subset samples to select more frequent words for gradient computation and loss approximation, based on the assumption that frequency can correlate with less distinction between more positive terms. This forces the model to learn more distinctive qualities within the data distribution.

Since both GLoVE and Word2Vec prioritize words that are more frequent in general conversations, blogs, and articles, they would not be optimal for handling less frequent, domain-specific vocabulary. General terminology such as “cancer” or “technology” may be better recognized over new algorithm names or subtopics. Since both models use word co-occurrences, however, they could be used as a first high level scan to separate higher level domains.

2.4 Dimension Reduction for Document Clustering

High-dimension data such as scientific abstracts require advanced techniques for automated analysis. One approach is to use unsupervised machine learning techniques such as dimension reduction with clustering to group data that are highly similar. Topic clustering methods that use word embeddings are forced to operate in a high dimension related to the length of the word embeddings. This leads to the *curse of dimensionality*, which can hinder data interpretation and clustering effectiveness.

2.4.1 Dimension Reduction of Embedding Models

Contributing factors for increasing dimensionality are word counts or n-grams. For algorithms that consider semantic similarity, related features can also contribute to the overall dataset dimensionality.

Several dimension reduction approaches apply synonymous word replacement techniques to reduce feature dimensions and model complexity while minimizing information loss and maintaining classification utility and accuracy. For example, Singh et al. developed a dimensional reduction technique to first apply GloVe and then cluster vectors that have high similarity measurements. The term with the highest TF-IDF score is used to represent each cluster [Sin22].

2.4.2 Traditional Dimension Reduction Algorithms

Traditional methods use statistical algorithms to reduce dimensionality prior to clustering: principal component analysis (PCA) [Hot33], Latent Dirichlet allocation (LDA) [Ble03], and multidimensional scaling (MDS) [You38] are common examples. PCA and MDS are common techniques for dimension reduction and visualization. The goal is to find principal components, or reduced orthogonal axes,

that retain a required amount of variance present in the original data space. Reduced axes enable improved clustering performance and maintain data patterns in the original data space.

While PCA seeks to estimate the original variances in the dataset with a minimum number of axes, MDS attempts to preserve important pairwise relationships between data points. MDS first calculates similarity distances between every pair of points. It then seeks an optimal embedding in a lower-dimensional space that most accurately preserves the similarity distances between the original samples. MDS does not reduce individual features, rather it focuses specifically on maintaining data relationships.

LDA estimates topic-word and document-topic distributions across a document collection with the objective of understanding latent topic prevalence. Several iterative or variational inference algorithms can be applied until the model converges. Input to LDA requires a predefined number of topics, which is often challenging as it requires data and subject expertise and can introduce bias. Moreover, LDA does not return concept labels explicitly and is computationally expensive. Many approaches apply PCA before LDA to reduce computation complexity, or use MDS as a post processing step to LDA for data visualization.

2.4.3 Uniform Manifold Approximation and Projection

Most recently, Uniform Manifold Approximation and Projection (UMAP) has found wide success in dimension reduction applications for a variety of text analytic applications [McI18]. It is the official dimension reduction algorithm implemented within BERTopic, a popular and effective topic labeling algorithm [Gro22]. UMAP is preferred over PCA for specialized vocabulary because it assumes non-linearity and uses graph based, manifold learning techniques to capture both local and global relationships. It has been used in a variety of scientific publications.

2.5 Clustering

Unsupervised learning is particularly useful for many domains that do not have labeled data. This represents an ongoing challenge for research as emerging topics, perspectives, discovery, and solutions are the expectation of quality publications. Clustering has often been used to group similar research papers, usually by citation or abstract metadata [Wen22; Che24; Yin23]. Accurate clusters can also enable search and retrieval by improving specialized vocabulary analysis, a capability that is particularly useful in domains that frequently contain vocabulary not sufficiently covered in LLM training data. Some clustering techniques are effective for identifying outliers, which could indicate new or emerging topics. We will introduce three methods: k-means based on centroid clustering,

DBSCAN based on point-to-point distances, and HDBSCAN, a hierarchical extension of DBSCAN.

2.5.1 k-Means Partitioning

k-means is a widely used clustering algorithm, dependent on requesting k clusters. k can be manually selected or algorithmically optimized. The elbow method, for example, is a heuristic used to determine a locally optimal number of clusters, and is often used to select k [Iko23; Yin23].

Due to the convergence goal of balancing data points between centroids, clusters in k-means are spherical. Every point will be assigned to its closest centroid. With no outlier allowance, accuracy is highly sensitive to noise, as it allows zero outliers and is dependent on a set of k cluster centers. k-means is often successful with structured data with clear distinctions between samples. k-means clustering for bibliographies, as an example, has success due to the concise text of the bibliography entries and strong similarity between common references. k-means uses a simple iterative algorithm to converge to a set of k clusters using the following steps.

1. Select k to specify a request for k clusters. Initialize the centroids by randomly selecting k data samples to assign as the initial centers of each cluster.

2. Calculate a distance (e.g., Euclidean distance) between a data point and every cluster center. Assign the data point to the cluster with the closest center.

3. Once all data points are assigned, find the data point closest to the center of each cluster, and assign it as the new cluster center.

4. Repeat steps 2 and 3 either for a specified set of iterations or until a measure of convergence is met.

2.5.2 Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) groups data points by their proximity to one another [Est96]. It is not reliant on a specified k. The input parameters to DBSCAN are a ε distance and a *minPts* constant. ε specifies the maximum distance two points can be from one another for the points to be considered locally neighboring. *minPts* defines a minimum number of data points required for a set of neighboring points to be considered a cluster. If a point does not satisfy these criteria, it will be considered an outlier. In this way, density-based algorithms allow non-spherical shapes and do not require balanced clusters. This can reflect better accuracy of datasets that have non-balanced groups. Clusters generated by DBSCAN are not limited by a

predetermined k or spherical shape, and are more resistant to noise due to the ability to include outliers.

2.5.3 Hierarchy Density-Based Spatial Clustering of Applications with Noise

Hierarchy Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) extends DBSCAN's capabilities by focusing on clusters of varying densities, rather than a constant density as used in DBSCAN [Cam13]. This allows HDBSCAN to capture nested clusters and proves effective in a variety of real world datasets and classes. It does this by completing the following steps.

1. Establish mutual reachability distances (MRDs) between data points, where

   - $d(x, y)$ is a distance metric between points x and y,
   - $\text{core}_k(x)$ is the *core distance* between point x and its k-nearest neighbor, and
   - $\text{MRD}(x, y) = \max(\text{core}_k(x), \text{core}_k(y), d(x, y))$

2. Use MRD to develop a minimum spanning tree (MST) by connecting data points with the minimum total edge weight as defined by MRD. This ensures nearby points are grouped together, increasing density.

3. Prune large distance edges in the MST to optimize cluster stability by removing excess mass. Stability is the summed difference between the most dense and the least dense regions. This intuitively defines a large difference as a cluster with various densities, and therefore not a strong candidate. A smaller difference means the density within the cluster is more unified.

4. Extract clusters based on a minimum cluster size *minPts*. In this way HDBSCAN starts at the core point of the MST and expands outward until *minPts* is reached.

2.6 Jaccard Similarity for Balancing Data

Challenges of unbalanced data can result in classification bias. Datasets that have mixed classes may have repetitive features across several classes. This causes separate classes to merge into a single large cluster. Jaccard similarity has been used as a method to address this issue, either by identifying nearly duplicate data or discovering data that have exact features. We discuss Jaccard similarity, related sampling approaches, and the effect of cross entropy loss on dataset balance.

2.6.1 Jaccard Similarity

The Jaccard similarity, also commonly referred to as the Jaccard index, measures the similarity of items between two sets. It is defined as the absolute value of the intersection of the sets divided by the absolute value of the union of the sets (Eq. 2.1). Verma et al. provide a more detailed explanation and abstraction of the original formula [Ver20]. The range of Jaccard should always be between $0 \ldots 1$, with higher Jaccard values indicating higher similarities. For example, in text data mining, Jaccard is useful for measuring the similarity of syntax, exact text (e.g., bag of words,) or exact phrases (e.g., n-grams). Jaccard similarity uses a binary determination of whether an item is or is not in the set, so it differs from other similarity measures such as TF-IDF that requires word counts.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.1}$$

For label classifications Jaccard's binary nature can be useful. As an example, Jaccard similarity can identify whether a word is common in some subjects and an outlier in others by observing its binary occurrence. If a set of documents produce a high Jaccard similarity score, further analysis can be performed to identify words that occur within most documents. This can be useful for feature engineering and pre-processing prior to other algorithms such as LDA for concept or subject refinement and identification. Words that appear across multiple documents can indicate those documents may contain similarities in their subjects. For example, the word *fabric* is more likely to appear in the subject *textile* and is more likely to be considered an outlier for a subject like *computer science education*.

2.6.2 Jaccard Similarity for Document Categorization

Jaccard similarity is used in many domains for document categorization. Different algorithms combine Jaccard similarity measurements with supervised data classification or unsupervised clustering. Notably, the similarity measures have shown advantages in keyword information retrieval and in domains with limited training data.

Across various domains from text, image, and pairwise sequence alignments Jaccard is commonly used for data de-duplication efforts [Ert22; Gov23; Abi20; Bah20]. In [Ert22], Jaccard similarity is used in conjunction with local sensitivity hashing and min-hash algorithms to reduce near-duplicates in large datasets. Baharav also proposed an approximate min-hash Jaccard algorithm, spectral Jaccard similarity, to improve pairwise sequence alignments for uneven k-mer distributions [Bah20]. [Gov23] developed an improved Jaccard similarity to evaluate a weighted bag-of-visual-word features within a pixel-based feature extraction process to detect near duplicate images. [Abi20]

showed Jaccard outperformed conventional text classifiers like multi-label and cosine similarity for duplicate news detection across various preprocessing methods.

Suphakit et al. demonstrated the use of Jaccard for information retrieval from documents. Specifically, they measure Jaccard similarity performance between keywords and index terms through Prolog programming [Niw]. A small set of keywords (100 words from thesauruses of Thailand agriculture) is measured against "crashed words," "over-typed words," and mispelled words." The results show very accurate and stable performance. More recent applications of Jaccard include [Cas22] demonstrating that Jaccard similarity can be effectively used to address TF-IDF's limitation of building features from co-occurrence between similar terms within court documents.

2.6.3 Unbalanced Data Bias

Unbalanced data is common for many text categorization applications. In the field of wireless sensor networks, [Pat20] describes the critical issue of unbalanced data and its implications. They state: "The term imbalance refers to uneven distribution of data into classes that severely affects the performance of traditional classifiers, that is, classifiers become biased toward the class having larger amounts of data." They also express the issue of unbalanced data to be common across many other fields such as the health care sector, modeling of cultures, and fraud detection. Solutions such as such as SMOTE (Synthetic Minority Oversampling Technique), Tomek links, and various k-nearest neighbor (KNN) algorithms have demonstrated success in balancing datasets where labeled data is available. However, in domains such as document clustering where labeled data is limited or mixed with unlabeled data, these solutions cannot be applied.

Solutions that can categorize unlabeled data often do not consider the balance of the dataset. [Kim19] describes a popular technique that combines LDA and TF-IDF to find relevant topics for non-categorized research papers by combining LDA's ability to identify latent topics with TF-IDF's term importance weights. Recently, transformers and deep neural networks have been implemented for document classification [Adh19; Pap19]. Although these models can incorporate additional modeling properties such as fine-tuning a cross entropy loss function to improve dataset balance, its performance can also be enhanced through data preparation. Attention-transformers often require significant labeled training data to be effective. Because of this, there are still opportunities for exploring solutions that maintain balanced datasets with limited or no labeled training data.

2.6.4 Sampling Approaches for Unbalanced Datasets

[Kul20] compares performance for the most popular undersampling, oversampling, and hybrid methods for unbalanced datasets. Undersampling removes data points from majority classes. Over-

sampling generates additional data points for minority classes. Hybrid methods combine both oversampling and undersampling. The coverage for undersampling algorithms is most specific to our research: random undersampling, Tomek links, and edited nearest neighbors (ENNs) [Kul20]. Although these approaches cannot be applied without labeled training data, learning from these methods can inform critical considerations for algorithm development and preprocessing of unlabeled training sets.

Random sampling selects arbitrary data points from majority classes until there is balance between all classes. [Wal14] demonstrates undersampling via random sample in an experiment to determine vulnerable files in Moodle. They classify vulnerable files as "positive" and clean files as "negative." Since the majority of files are not corrupt, all "positive" files are kept and randomly select "negative" files are removed until the dataset is balanced. Tomek linking selects pairs of samples from two different clusters that are closest to one another. The data point in the majority class is then removed. ENN compares each majority point with k neighbors. If the majority of the neighbors are from minority classes the majority point is removed. [Nur23] uses both Tomek links and ENN to balance datasets while detecting depression in tweets. After oversampling the dataset, Tomek links pair samples of opposite classes. Different from the original Tomek link algorithm that only removes points from the majority classes, their work removes all paired data points.

2.6.5 Cross Entropy Loss for Balancing Datasets

The cross entropy loss function is frequently used to balance datasets for classification models, mostly through supervised learning. Lower cross entropy loss generally indicates better model performance. Differing weights are assigned to each class during the computation of cross entropy loss. The cross entropy loss function can also be customized to assign more importance to correctly predicting instances of minority classes.

Typically, loss is computed for every batch and modeling adjustments are made to minimize losses for the next batch, where a batch is a subset of the training data processed during a single training step. Minority classes, for example, could be assigned higher weights, implying a higher impact to minority member misclassifications and encouraging the model to make additional training adjustments. Alternatively, errors from majority classes can be assigned lower weights, leading to fewer training modifications for subsequent batches.

Cross entropy loss cannot be applied where there is no labeled data, so it does not work for unsupervised document clustering. There are some exceptions when unsupervised learning creates pseudo-labels during training, transitioning into semi-supervised learning. Several approaches achieved success with variations of this approach. [Hat21] proposed Cformer, a semi-supervised

model to cluster text documents that only have a small population of training data using consistency or unsupervised loss that "constrains the model predictions to be invariant to input noise by forcing augmented samples to have the same labels as the original data samples." [Wei22] trains with both labeled and unlabeled data in a two step approach known as Deep Contrastive Semisupervised Clustering (DCSC). First, the model is trained with cross entropy and supervised contrastive loss based on labeled data alone. Next, a classifier is built to produce pseudo-labels. The algorithm then trains with both labeled and pseudo-labeled data. [Lim21] developed a classification approach via clustering and data self-labeling "based on a deep autoencoder combined with a self-labeled technique that takes into consideration cross-entropy." Using the stochastic gradient descent algorithm to calculate cross entropy loss, weights are adjusted according to clustering performance within the reduced dimensionality space.

2.7 Text-to-Embedding Vector Transformation

In order for text to be processed for machine learning models such as text generation, recommender systems, or information retrieval, it must be encoded in vector form for semantic interpretation. This section details the common challenges of data sparsity and general vocabulary overfitting, presenting dimension reduction techniques as solutions. A background on common conversion techniques (selecting n-grams versus bag-of-words) is presented and examples of the most common pre-trained embedding models are introduced.

2.7.1 Topic Modeling for Scientific Publications

Topic modeling is often used to interpret latent concepts among collections of documents. Traditional approaches involve probabilistic modeling such as LDA and linear matrix factorization such as non-negative matrix factorization (NMF) and latent semantic analysis (LSA). These techniques are used to find common themes among groups of non-sequential text. Topic visualization can be performed to improve explanation of latent topic relationships. Recent work combining traditional approaches (e.g., LDA or TF-IDF) with fine-tuned language model embeddings such as SciBERT show improved topic coherency.

- **Non-negative matrix factorization (NMF).**
  NMF decomposes a term–document matrix X into two non-negative matrices N and M representing a contribution probability of topics to document (N) or terms to topic weighting (M) [Lee99] The objective is to minimize the Forbenius norm of the difference between X and the product of the decomposed matrix $N \times M$ (representing the loss from decomposing

X). NMF imposes non-negative constraints, meaning features can only add and not subtract. NMF optimizes for data sparsity by penalizing for non-sparse and non-smooth values until convergence. Several methods have been used for initialization of NMF: randomization (assigning each entry in M with small random values), singular value decomposition, and hybrids of both approaches. Proper initialization is key for a successful NMF implementation. By assigning random values to all words, weak word to topic contributions will iteratively drop to zero. An example would be a word that comes from a contradicting topic having its value canceled in favor of retaining words with stronger relationships. Since NMF relies on word frequency and co-occurrence, similar to LDA, results of uniform word distributions (e.g., characteristics of scientific abstracts) might still be poor despite overcoming the data sparsity challenge.

- **Latent semantic analysis (LSA).**
  LSA is based on singular value decomposition (SVD) [Alt16]. Similar to dimension reduction algorithms that reduce the original data to the most strongly contributing features, LSA's objective is to identify the most significant orthogonal vectors that capture latent topic relationships. Semantic similarity is calculated between the remaining terms. LSA is often used in isolation to identify the strongest topic–word relationships, or with other constraint optimization algorithms (e.g., combinatorial optimization or n-grams) to capture the most significant phrases. As with LDA and NMF, LSA's reliance on frequent terminology and supporting terms greatly limits its performance with shorter text such as research abstracts.

- **Latent Dirichlet allocation (LDA).**
  [Che21] present a methodology to demonstrate how different topics are related to one another through quantitative modeling. Their work performs topic modeling from publications related to Big Data and Dye-Sensitized Solar Cells drawn from World of Science. The paper addresses existing limitations of the original LDA algorithm by incorporating meta data with discovered topics. They further extend cluster analysis by applying a post-topic modeling process to identify and quantify relationships between topics.

  [Kim19] compare LDA and NMF for classifying research papers. They evaluate performance across a collection of 1,740 papers from the New York University website. Results show that LDA had better average coherence scores and optimal topic numbers. Twenty-two topics were originally found using LDA but were reduced to ten to improve categorization. LDA was still found to be the better performer.

2.7.2 Modeling Emerging Topics

Instead of direct topic modeling on abstracts or longer text forms, [Sma14] uses time and citation-based models on the Scopus dataset to identify clusters of emerging topics using two methods: a co-citation model and a direct citation model. The co-citation model looks for new *threads* that represent connections between small clusters of papers from a common year that focus on the same science problems. The direct citation model uses the video object segmentation (VOS) algorithm to "maximize the ratio of links within clusters to links between clusters." Small et. al's work showed that the direct citation method produced clusters that are broader conceptually and cover a longer time duration [Sma14]. Several other topic models were introduced to forecast emerging trends [Xu21; Che06]

Different from static models like LDA that cannot differentiate time periods in a collection, [Ble06] developed probabilistic time series models to represent topics sequentially from large document collections. They use state space models to approximate posterior inference over latent topics. The data is divided by a time slice (e.g., annually) to reflect order in an evolving set of topics.

2.7.3 Topic Visualization

[Jeo24] developed a method using BiTerm topic models (BTM) with a latent space item response model (LSIRM) to visually interpret relationships between research topics. Topic–word distributions are reflected as item–response data, enabling visualization of the space between word frequency and latent topics to better interpret word–document relationship. Their work further proposes methods to interpret variations within the latent space that can result from "variations in distances between word pairs due to rotation, translation, and reflection."

2.8 RAG + LLM Topic Modeling

Several retrieval augmented generation (RAG) techniques have been explored to further improve Q&A responses for domain specific subjects. These techniques range from text cleaning and preprocessing [Lak24] , to unsupervised learning approaches to define temporal representations, to trend forecasting. Topic enhanced re-ranking techniques are also compared to more popular re-ranking models to evaluate trade-offs between performance and resource efficiency.

Tran proposes a topic modeling approach that improves retrieval and response generation performance [Tra24]. Each cluster is trained on individual encoders. Then the top k documents are retrieved from each cluster based on topic distribution similarities from an input query. RAG-topic and RAG-context-topic are validated to select relevant turns in a dialogue history. Overall,

the RAG-based models achieve improvements in both retrieval and generation, complementing models that focus on building dialogue history representations. The authors show that ChatGPT can benefit from the improved retrieval performance to produce improved generation results, generating its best responses when external knowledge is provided. Based on the studies that were run RAG-context-topic is the best performer.

Kumar et al. uses BERT-based topic modeling with generative AI and RAG for semantic insights and trend forecasting [Kum24]. Topic modeling is performed on both patent data and publication data over three topics: crypto-currencies, quantum computing, and reinforcement learning. Kumar et al. benchmark emerging trends' effectiveness with the Gardner hype cycle.

Finally, Ampazis proposes a re-ranking framework, Topic Enhanced RAG (TER), that utilizes latent topics to improve the precision and relevance of RAG results [Amp24]. He contrasts the performance and efficiency of TER with traditional cross-encoder re-ranking models that are theoretically more accurate but much slower. Cross-encoder models process a data pair (e.g., two documents) jointly with the encoder that results in less information loss. Unfortunately, this approach is very time consuming. Ampazis's work shows that TER improves baseline RAG but falls short of COLBert, a state of the art cross-encoder re-ranker, by a small margin.

CHAPTER

3

HIERARCHICAL CLUSTERING FOR AUTOMATED LITERATURE SCAN

Scientific work should acknowledge domain-relevant topics and provide sufficient novel contributions to be considered meaningful for publication. The rapid growth of research literature makes it increasingly difficult to determine publication readiness. The use of LLMs for automated literature search and peer review is an area of significant interest, but to the best of our knowledge, no existing techniques combine aspects of both functions. This separation restricts LLMs to either literature review summaries or automated peer reviews for writing improvements. To address this gap, our investigations provide a scalable, automated approach using unsupervised methods with LLM abstractive reasoning to draft evaluations of a paper's topic novelty and completeness compared to published papers in the same domain. Clustering and topic modeling with semantically similar research will simulate a high-level literature search. The resulting topic map is used with LLM abstractive reasoning to provide personalized feedback, for example, topics from the broader research that should be considered but are missing from a paper's abstract, or potentially novel topics discussed in the paper. To achieve this goal, we first improve clustering effectiveness on domain-specific text. The unique vocabulary in scientific manuscripts creates high data sparsity, which leads to large outlier populations. To minimize noise while keeping "true" outliers, we focus

on improving cluster quality to generate clusters with a single dominant topic. For any remaining large clusters, Jaccard undersampling is applied to further refine both topic coherence and inclusion, and to reduce a cluster's size to a small collection of representative samples.

3.1 High-Quality Clustering

High-quality clustering is essential for producing coherent, semantically relevant topic groups. Domain-specific corpora often vary widely in vocabulary and scope, making it difficult to maintain consistent cluster granularity. For our use case, we define a cluster as "high quality" if it has both high topic coherence and high topic inclusion. Topic coherence measures the similarity between keywords mined from a cluster's dominant topic. Topic inclusion prevents the omission of critical subtopics. We anticipated that smaller local clusters should allow for both because of each cluster's focus on one dominant topic. This is controlled using a minimum cluster size hypertuning parameter. Unfortunately, limiting cluster size can result in an excessive number of outliers. At the opposite end of the spectrum, global clusters can include documents that may have been falsely classified as outliers, but this is a trade-off versus generating excessively large clusters (e.g., > 100 documents). Our use of Jaccard undersampling is meant to refine large clusters by creating semantic barriers between documents with high Jaccard similarity. Hypertuning Jaccard similarity at this scale is an important challenge that we were required to explore. Prior experiments that consisted of well-defined boundaries with forty documents in each group were manually tuned, suggesting that a Jaccard similarity threshold of 0.15 was sufficient during preliminary experiments. Given this, we developed a pipeline that supports the following.

- We leverage automated topic modeling to streamline document comparisons for literature reviews. To achieve this, we investigated the following open problems: (1) grouping related documents through semantic text vector representations, and (2) topic clustering.
- We minimize outliers by constructing level-of-detail hierarchies at both local and global topic granularities with dimension reduction and HDBSCAN clustering.
- We maintain effective cluster sizes to generate relevant cluster topics. Large clusters may indicate that a cluster's dominant topic is too broad. Alternatively, restricting all clusters to a maximum size can produce a significant number of outliers.

Based on this, the novel contribution of this stage of our research includes the following.

- A method to define "true" outliers for research classification by optimizing a cluster's topic coherence and topic inclusion.

- Evaluation of ways to select optimal cluster sizes that balance topic coherence, topic inclusion, and outlier count over a collection of documents.
- Implementation of Jaccard undersampling to compact large clusters into a smaller set of representative documents.

3.1.1 Related Document Clustering

Our first goal was to group related documents through semantic text vector representations and topic clustering. If done correctly, this will allow us to enumerate topics within a document collection at different levels of detail. This forms a foundation to evaluate which topics exist, or do not exist, within a user-provided abstract for a paper under consideration for submission. We extract our data from the dataset [Kow17], specifically the abstracts from the Web of Science Computer Science category, which focuses on seventeen subcategories. The two most common methods of representing text are term frequency vectors and word embeddings. Large language models are normally built on general semantic embeddings using neighboring terms to predict a target term (CBOW) or a target term to predict its neighboring terms (skip grams). Regardless of which approach is chosen, the embeddings are used as input for clustering documents. For example, *paraphrase-MiniLM-L6-v2*, a sentence embedding transformer-based model, generates a semantic representation of a sentence as an embedding vector of over 300 dimensions. Since word embeddings are vectors, pairs of embeddings can be compared using methods like cosine similarity to estimate the semantic similarity of the text represented by the embeddings. The pairwise similarity scores form a similarity matrix act as input to a clustering algorithm. We used HDBSCAN, since it can return level-of-detail clusters. This allowed us to investigate how well clusters at different levels of detail capture a dominant topic, as well as the number of outliers generated. As noted above, our goal is a level of detail that identifies a single dominant topic in each cluster while minimizing the number of outliers. Input parameters to HDBSCAN, like epsilon distance (to define neighborhood size) and minimum cluster size, were varied to determine their effect on achieving topic cluster coherence and inclusion.

3.2 Minimizing Outliers with PCA and UMAP

High-dimensional text vectors require dimension reduction prior to clustering. Projecting to a lower dimension reduces noise and irrelevant features, enabling better boundary detection. The lower the dimension, the more likely clusters will represent a single dominant topic, as there are fewer variables to create overlap. We explored both linear and non-linear dimension reduction techniques

for their ability to distinguish strong features that contribute to a core topic and reduce weaker variables that contribute to overlapping clusters.

3.2.1 Principal Component Analysis

Principal component analysis (PCA) is a linear dimension reduction technique that identifies the principal components, or linear regression directions, that capture the most variance in a high-dimensional dataset. These directions are defined by the eigenvectors of the covariance matrix of the dataset. Higher variance indicates combinations of the original dataset's features that create "optimal" separation boundaries between data points. In our environment, this is analogous to creating an improved isolation of a dominant topic. We ignore lower variance components because they are likely to have similar qualities for most or all the data points, generating noise that contributes to multi-label clusters. Ideally, we want to reduce data to the lowest dimension that prioritizes one key topic for each cluster. The goal is to retain the most critical information that effectively identifies semantic cluster boundaries. We applied PCA in experiments that first generate BERT embeddings, project those embeddings with PCA, and then cluster the embeddings with HDBSCAN. A high-quality PCA result occurs when there are only a few eigenvectors with large eigenvalues, suggesting only a few principal component dimensions are needed to effectively separate the data. Our experiments explored PCA's ability to generate a small number of dimensions that capture the majority of the variance in our embeddings. The results were suboptimal, as the reduced dimensions did generate smaller clusters at the expense of significant outliers ($> 85\%$ were outliers).

3.2.2 Uniform Manifold Approximation and Projection

Non-linear structures have the advantage of constructing clusters through manifold learning techniques that use curved structures with concave and convex boundaries. Text characteristics that can contribute to produce non-linear structures include varying densities between differing authorship volume, popularity within growing subtopics, and time window presence within the same domain. These varying shapes and cluster characteristics may mislead linear transformation techniques like PCA to falsely classify data points that should be clustered as individual outliers. To explore possible solutions for better refining outlier classification, we investigated Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensional reduction technique that preserves neighboring structures without flat plane limitations. Theoretically, UMAP should better capture data inside curved structures to generate "true" clusters while at the same time reducing false outliers. UMAP calculates pairwise distances between each point in the high-dimensional space. Using a continuous function, it transforms these distances into fuzzy probability representations of the likelihood and

degree of similarity between points. Points closer to each other have higher fuzzy probabilities. This is used to build a lower-dimensional graph of the dataset. The graph is further optimized with stochastic gradient descent (SGD) to improve cross entropy loss compared to the higher dimension graph. This balances the construction of local clusters while maintaining sufficient global scaffolding. Connecting data through fuzzy probabilities captures hierarchical relationships. This can better distinguish true outliers from data points that should belong to a cluster through the manifold learning structures. Unfortunately, there is a trade-off when using hierarchical clustering and dominant topic identification. Unlike PCA where dominant topics can often be clearly differentiated with a few high-value eigenvectors, multiple hierarchical levels in UMAP can add more noise and produce multi-subtopic connections. This is normally evident in more abstract global clusters. Thus, choosing UMAP over PCA requires a trade-off between inappropriate outliers and cluster overlap reduction. UMAP uses a k-nearest-neighbor (KNN) approach based on its fuzzy probability scores, where k is defined by the user. Smaller k will produce a smaller neighborhood, while larger k will allow for more global topics. We explored the performance of UMAP and the impact of different values of k using the BERTopic library. BERTopic chooses a default k assumed to work across many use cases. This provided a good starting point from which to explore and investigate whether a common k is effective, whether custom k are needed for different datasets, and if so what methods exist to choose an appropriate k. We found that UMAP was successful at reducing outliers compared to PCA (65% outlier reduction versus 15% with PCA). However, this was at the expense of adding global clusters with more than 100 documents that may be connected to several topics. The section "Improving Global Cluster Precision" describes our plan to improve single topic precision for these potentially broadly focused clusters.

3.3 Jaccard Undersampling

Jaccard sampling uses binary lexicon interpretations that leverage semantic similarity vectors. For our research, Jaccard is critically related to the following two properties:

1. **Clear distinction between Jaccard similarity levels.**
   Scientific abstracts are written for conciseness, focusing on key functions: introduction to a problem space, novel methodology, and results. Due to similar terminology introducing or concluding these components, we expect to find common Jaccard similarity trends across all abstracts. With exact word binary matching with one hot vectors (a binary "match-–no match" decision), we expect that any matched terms are likely to either be part of the formal introduction phrases or to have an elevated likelihood of belonging to similar subtopics. In our

preliminary Jaccard undersampling experiments, we measured Jaccard similarity across all documents and plotted the maximum similarity for each document. There were two distinct patterns across maximum Jaccard similarity levels.

(a) Topics that have a linear trend line.

(b) Topics that have a linear trend line with noticeable peaks.

Further examination revealed an additional characteristic that differentiated these topic groups into two categories: topics that are narrow in scope and topics that span multiple domains. Accuracy was negatively impacted in topic groups with linear trends and improved in topic groups that contained peaks in the similarity trends. We hypothesize that topics that span multiple domains could have common keywords across those topics, explaining the peaks we observed (e.g., "wireless sensors" could be used in both the "computer science" and "robotics" topics). Our preliminary experiments did not choose a subset of groups when applying Jaccard undersampling. However, our results clearly identified topic characteristics where clustering consistently improved with Jaccard undersampling. For our current work, we believe that global clusters may represent topics that span multiple domains or have similar subgroups. To test this hypothesis, we performed an identical experiment to confirm that global topics have noticeable peaks in their maximum Jaccard similarity trends, and the results did show this pattern. Even though the Jaccard maximum similarity trend line was produced as expected, the Jaccard undersampling implementation was unsuccessful. There was no significant impact found in either outlier reduction or an increase in local clusters. We hypothesize this is because HDBSCAN develops clusters with empirical neighbor statistics, and as such, is less impacted by undersampling techniques that rely on lexical similarities (as compared to our preliminary experiments that use TF-IDF and DBSCAN).

2. **Risk aversion in document removal.**
By leveraging exact word matching, any elevated Jaccard similarity value is likely to indicate higher topic similarity content. Binary scores offer clear distinctions between hypertuning levels by observing any peaks in similarity. Using exact word matching would place more emphasis on subtopic matching that can address multi-label noise when clustering. Using this level of precision reduces the risk of unnecessarily removing documents that may not contribute to subtopic overlaps. Semantic similarity, especially when used for vector comparison across full abstracts, can obscure whether the documents contain similar subtopics. In mean pooling, for example, the continuous similarity score is averaged across terms. The overall document similarity may be high if significant noise is present in the abstracts, producing a

higher average similarity, but that may not be indicative of topic-related similarities. Semantic similarity also has less extreme variance, making it more difficult to choose boundaries for document undersampling. To test this theory, we performed an experiment in which documents were semantically embedded with *all-MiniLM-L6-v2* and compared by cosine distance. As expected, we found the general consensus of maximum cosine distances to be similar, making it impossible to accurately differentiate data points with very high similarity to undersample.

3.3.1 Tuning Jaccard Similarity Distance

Jaccard similarity is calculated between document vectors. After the Jaccard values are generated, we evaluate appropriate ranges of Jaccard similarity. It is expected that documents with a common topic will have higher Jaccard similarity. Consequently, for smaller datasets, Jaccard similarity values are expected to be lower as it is less likely to include documents with overlapping words between categories. We identified an appropriate Jaccard similarity range based on both semantic similarity and dataset size.

3.3.2 Jaccard Undersampling

For unlabeled datasets, determining whether the dataset is balanced can be nuanced and domain specific. In terms of document clustering, it is assumed that document features (words) should be equally important. The binary nature of Jaccard assumes that all words have equal weight. Thus, if there are pairs of documents with high Jaccard similarities, significant feature imbalance can lead to inaccurate clustering of different topics. Removing one of the paired documents can hypothetically generate more accurate cluster boundaries. To reduce dataset feature bias, for document pairs with higher Jaccard similarity (lower Jaccard distance), one of the paired documents was removed as part of the Jaccard undersampling process. As stated in above section “Clear Distinction Between Jaccard Similarity Levels,” we found that the clustering algorithm used is a major factor in whether Jaccard undersampling can be successful. To summarize, Jaccard undersampling works best with lexical-based clustering and was not as successful with semantic-based clustering using HDBSCAN’s mutual reachability distance.

3.4 Falsely Classified Outliers

Falsely classified outliers are documents that should belong to a cluster, as their content is semantically similar to the dominant topic of that cluster. When documents are falsely tagged as outliers, relevant information is excluded from the cluster they belong to. This decreases topic modeling

accuracy, especially for smaller clusters. On the other hand, correcting this by generating global clusters is unlikely to be successful since global clusters have poor single-topic precision. If a cluster contains more noise than single-topic precision due to assigning false outliers to the cluster and thereby increasing its size, it will have a negative impact on topic coherence. It is important to understand how outliers belonging to a local or a global cluster influence the decrease in topic coherence. We examine how assigning false outliers impacts topic coherence with the following approaches.

- **Do most outliers align more closely to local or global clusters?**
  Prior experiments that applied different dimension reduction techniques (PCA versus UMAP) decreased outliers but increased the number of global clusters. We hypothesize that the outliers found during dimension reduction were assigned to global clusters, decreasing their probability of containing a single, strongly connected topic.

- **What factors impact outlier assignments? Are these stochastic or deterministic?**
  We focus on the impact of HDBSCAN, which uses Mutual Reachability Distance (MRD) likelihood to assign membership probabilities. This is based on how likely sets of documents belong to the same cluster and which are considered outliers.

  During HDBSCAN, two documents can have a strong likelihood of belonging to the same cluster without having direct similarities. Since similarity likelihood is less strict than direct similarity comparison, this offers more clustering opportunities and decreases false outliers.

  However, outlier assignments following dimension reduction leads to stochastic outlier selection (e.g., the pruning of documents that border cluster edges is stochastic because it is not always clear how specific features are combined or modified during dimension reduction). This stochastic characteristic makes it difficult to determine the impact on topic coherence based solely on outlier documents or their observed characteristics. To eliminate this risk, we increased our scope of false outlier impact by assuming that any document could be considered an outlier.

- **What is the impact of false outliers on topic coherence within global clusters?**
  We assume that any document within a global cluster can be identified as a false outlier. To test this hypothesis, we selectively remove documents that have a higher likelihood of being "representative" or core documents to assess their impact on topic coherence. Jaccard similarity is used to represent the likelihood of being a core document, since it is less likely to be impacted by duplicate vocabulary in scientific documents unless the vocabulary itself is

domain-specific. Section 3.5.1.2 Jaccard Undersampling reflects the impact of increasing false outliers.

3.5 Improving Global Cluster Precision

Global clusters built using HDBSCAN can capture broadly related documents and their subtopics based on MRD. Clustering via probabilistic similarity produces fewer outliers by maximizing the opportunities through which documents can be connected. It is not necessary that there are semantic or lexical similarities to more than one document. In the worst case, every connection in a global cluster could have equal weight, corresponding to each document having a single similarity connection. HDBSCAN improves recall by decreasing false outliers, better ensuring that all documents with any connected subtopic are included in a common cluster. Relaxing the criteria for how documents are connected, rather than strict similarity distance clustering like DBSCAN, can also introduce numerous subtopics. As a result, the precision for single-topic clustering is reduced. We explored the possibility of using Jaccard similarity to: (1) undersample, and (2) find the most representative samples (documents) in a global cluster to improve precision by increasing topic diversity and topic coherence.

$$\text{Topic Diversity} = \frac{\text{number of unique keywords}}{\text{total number of keyword occurrences}} = \frac{|\{k_1, k_2, \ldots, k_N\}|}{\sum_{i=1}^{N} n(k_i)} \tag{3.1}$$

Measuring the topic diversity of an entire dataset uses Eqn. 3.1. When we decompose a single global cluster into more local clusters the total number of topic keywords increases based on the number of new local clusters multiplied by expected keywords per cluster (ten keywords for each cluster in our experiments). To calculate the diversity of the original global cluster, we combine the topic keywords from the newly generated local clusters and apply the topic diversity equation to the keyword set.

Measuring the topic coherency of a cluster is performed with Eqn. 3.3.

$$\text{Coherence}(T) = \frac{1}{\binom{|T|}{2}} \sum_{\substack{w_i, w_j \in T \\ i<j}} \text{sim}(w_i, w_j) \tag{3.2}$$

$$\text{Topic Coherence} = \frac{1}{Z} \sum_{z=1}^{Z} \text{Coherence}(T_z) \tag{3.3}$$

where Z is the number of topics and T_z is the set of keywords for topic z.

3.5.0.1 Defining Topics with Jaccard Similarity and RAG

We evaluated whether Jaccard similarity can be useful for constructing subclusters from global clusters to enhance RAG efficiency. Traditional RAG finds the most similar documents, usually through embeddings and cosine distance, to a user query. Since clustering maintains common themes across document groups, it is assumed that the resulting topic keywords represent a common concept. If a topic keyword is found in one document, then performing RAG on that document should be sufficient to generate an acceptable definition of the keyword. Documents with the highest Jaccard similarities would indicate a higher probability of those documents containing relevant topic keywords. This property is important in understanding that, while topic clustering can be useful for capturing the most dominant topic keywords, a cluster or subcluster containing the topic keywords at least once is sufficient to provide context and topic definition.

3.5.0.2 Isolated Global Cluster Subsampling

We explored the following refined clustering approaches to improve the single-topic precision of a global cluster. Popular methods include reducing global clusters to local clusters and selecting representative samples from large global clusters. Our approaches use Jaccard similarity, BERTopic, and RAG in the following experiments.

1. **Assessing the Jaccard similarity trend line.**
   Prior experiments have identified an exponential trend line for single topics spanning multiple domains. Since global clusters are known to connect multiple subgroups, if there are sufficient keyword overlaps, this should result in an exponential Jaccard similarity trend. Global clusters, though broadly defined by nature, can also exhibit characteristics of single topics that span multiple domains.

2. **Performing Jaccard undersampling.**
   Undersampling decreases the size of a global cluster while retaining documents containing the most connected, underlying single topic. Undersampling reduces noise generated by less representative topics and improves precision for the most relevant topic. Reducing noise while focusing on the most highly connected topics should improve topic coherence and single-topic precision.

3. **Rerunning BERTopic on global clusters.**
   To further isolate documents with the highest potential to be false outliers, we reapply BERTopic to increase the number of clusters and decrease cluster sizes. Increasing clusters should improve single topic precision, as clusters are refined into more local topics. We do expect a slight

decrease in topic diversity, as we are starting from the highest topic diversity score possible of 1 for global clusters.

4. **Minimum Spanning Tree with top-k edge removals.**
   For global clusters, we calculated the cosine distance between pairs of embeddings to use as edge weights for implementing a minimum spanning tree clustering algorithm. Removing k edges could increase single topic precision by forming new subclusters. In this experiment, pruning is selectively applied to the top-k edges to obtain specific target cluster counts.

5. **Minimum Spanning Tree with Jaccard distance and top-k centroids.**
   The MST-Jaccard approach is designed to select the most representative samples from large clusters. The nodes with the highest centrality score (i.e., the nodes with the minimum distance to local nodes) are assumed to contain the most representative topic keywords in their local neighborhoods. Using Jaccard distance as a weight, we select the nodes with the lowest total neighborhood distance as representative samples.

3.5.1 Approach and Experiment

3.5.1.1 Jaccard Similarity Trend

We construct a Jaccard similarity trend line to determine whether a non-linear trend line is formed. From our Jaccard undersampling experiments, we observed non-linear trends for a single topic that spans multiple domains. This could be explained by having a small number of documents with a high similarity in exact vocabulary, but this is rare, especially after preprocessing and stop word removal, for scientific documents.

Experiment Design

Though there is no official metric that defines a cluster as local or global, we noted a sharp increase in the number of documents in clusters containing more than 100 documents. This was used to identify a small number of clusters as global. Based on the original Jaccard undersampling experiments, each cluster is evaluated independently. Jaccard similarity is measured between all documents in the cluster. Then, for each document, the maximum Jaccard similarity score is obtained. We sort these scores in ascending order and plot them in a line graph to determine the similarity trend.

Results and Discussion

We noted that the top three global clusters have an exponential trend line (Fig. 3.1), suggesting that there is at least one common theme across these documents that may span multiple domains. Based on this finding, experiments were continued to explore Jaccard similarity for refining local

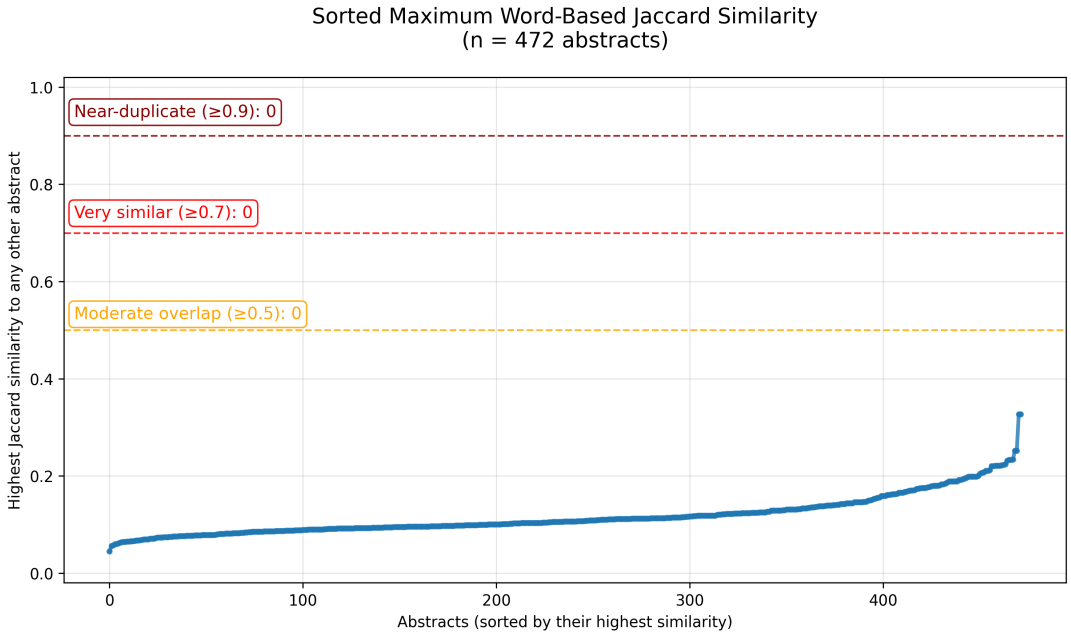

Figure 3.1 Sorted maximum word-based Jaccard similarity scores for all abstracts. Each point represents one abstract's highest similarity to any other abstract in the dataset.

subclusters.

3.5.1.2 Jaccard Undersampling

Jaccard undersampling was explored to select a smaller subset of the global cluster for more precise single topic representation. This was tested by removing nodes that we identified as false outliers. An assumption is made that, depending on the selected dimension reduction and clustering techniques, all data points are at risk of becoming a false outlier. To identify the most impactful removals, we used Jaccard similarity as an indicator of core nodes. The higher the similarity, the more core a node (document) is for topic representation. We perform incremental runs of Jaccard undersampling, decreasing Jaccard similarity thresholds on each run, until we remove half of the cluster's nodes.

BERTopic is run to identify the topic keywords for a global cluster. Next, Jaccard undersampling is performed incrementally with a lower-bound similarity threshold of at least 0.01. Lower Jaccard scores are more likely to contain random similarity. Finally, BERTopic is run again over the full dataset, but excluding undersampled documents. The resulting topic keywords are then identified.

Results

Removing documents until a Jaccard similarity of 0.01 was met resulted in approximately 50% of the documents being removed from a global cluster. It was expected that such a large document removal would result in significant changes to topic keywords, but the undersampled cluster's topic keywords had only minor changes. This suggests the keywords reflect global concepts common across the document groupings. If a large number of documents can be removed without significant impact on topic keywords, we wondered whether further selective sampling could be performed with a smaller subset while still retaining the majority of the keywords. This question motivated further experiments on refining document clusters.

3.5.1.3 BERTopic on Isolated Global Clusters

It is expected that rerunning BERTopic, and therefore HDBSCAN (since it is part of BERTopic), on a small document subset should generate new subsets. An important question is whether these subsets are useful. To explore this, we measured the change in topic diversity after multiple runs of BERTopic. Running BERTopic on a global cluster further separates it into subclusters. Topic quality could decrease, however, if there is less differentiation between the subclusters. If a global cluster does, in fact, represent a high-level single topic, then we expect topic diversity to decrease when we further subdivide the global cluster into subclusters as topic keywords would repeat themselves.

Experiment Design

We first isolate each global cluster by removing it from the dataset. Next, we run BERTopic with the same parameters that were applied to the entire dataset. If this generates multiple clusters we evaluate whether these subclusters generate meaningful subtopics or are simply clusters with little diversity and less differentiable subtopics. This is done by measuring topic diversity and comparing it to the average topic diversity produced by BERTopic on the full dataset. We performed this experiment on four clusters, as these clusters contained more than 100 documents and were therefore classified as global.

Results and Discussion

When evaluating a global cluster in isolation, we assigned it a diversity score of 1, representing the full set of ten topic keywords. The diversity scores after rerunning BERTopic individually on each global cluster are shown in Table 3.1, with an average diversity score of 0.689 over the four global clusters.

Identifying topic keywords is required to match a user's abstract to an appropriate cluster of previously published documents. Given an average topic diversity score below 0.7, we do not recommend separating global clusters in this manner, as we do not believe it will lead to improved

Table 3.1 Global cluster diversity statistics.

Cluster ID	Total Rows	Subclusters	Unique Words	Diversity
1	460	7	47	0.67
2	392	4	29	0.725
3	385	5	34	0.68
4	151	3	21	0.7

cluster–abstract matching. Instead, doing so may limit correct matching, as there is insufficient separation between clusters to expect a single correct match.

3.5.1.4 Minimum Spanning Tree with Top-k Edge Removal

In this approach we select a representative document from each subcluster within a global cluster. The MST clustering algorithm is effective in this regard, utilizing a replacement of the traditional edge distance threshold step with the top-k edge removal method. This ensures that all subclusters are represented in $k+1$ clusters and can be further refined to select representative samples.

Experiment Design

1. Embed each abstract with *paraphrase-MiniLM-L6-v2*.
2. Calculate the cosine distance for each embedding.
3. Run Kruskal's algorithm to construct a minimum spanning tree.
4. Remove the top nineteen edges based on edge weights.

Results and Discussion

MST clustering indicated that the majority of the new subclusters (> 90%) are outliers. This is expected, as many outliers were also observed in prior experiments with PCA dimension reduction. If those outliers were clustered using a different dimension reduction technique, they would be the most likely candidates for the most distant points within a subcluster.

3.5.1.5 Minimum Spanning Tree with Top-k Centroid Subsampling

Undersampling based on MST with Jaccard distance was explored to selectively optimize the data needed for topic coverage. The "highest centrality" refers to nodes with the lowest connected edge weights. Lower connected distances indicate a node with more centrality within its local

neighborhood. The goal is to select the minimum centrally located documents (i.e., similar to the largest number of neighboring documents) needed to define relevant topic keywords.

Through semantic clustering, it is assumed that representative topic keywords are both unique and common in meaning across a cluster. This implies that scientific terminologies have ubiquitous definitions throughout the cluster and are not relying on the semantics of supporting text. If this is true, capturing the keywords from one undersampled document should produce a general keyword definition all documents within the cluster.

Experiment Design

1. Perform preprocessing for each global cluster.
2. Calculate a cluster's Jaccard distance matrix.
3. Construct each cluster's MST.
4. Compute node centrality scores by summing the total edge weight of each node.
5. Extract the top twenty highest centrality abstracts, labeled as the MST-Jaccard top-twenty set.
6. Retain each of the topic keywords observed in the global cluster if it appears in the MST-Jaccard top-twenty set.
7. Steps 1–6 are performed for every global cluster, and for some mid-sized clusters to evaluate performance across varying cluster densities.

Results and Discussion

MST-Jaccard performed very well (Table 3.2) as most clusters captured 100% of the relevant topic keywords. As noted above, it is expected that the topic keywords have common terminology within a cluster. This property will be especially useful for combining the MST-Jaccard top-twenty set with RAG to guide an LLM to generate a summary with appropriate topic keywords.

3.6 Conclusion

This chapter focused on creating the most effective single topic clusters using hierarchical clustering and topic modeling. After clustering, we experiment with several post processing methods that target both the reduction of outliers and reducing noise and redundancy from large clusters ($>$ 100+ abstracts in each cluster). The best approach, MST-Jaccard, was found to be successful in reducing

Table 3.2 Distribution of Abstracts Across Selected Clusters

Cluster ID	Abstract Count	Captured
1	463	90%
2	422	100%
3	173	100%
4	161	100%
5	144	100%
40	31	100%
41	31	100%
42	31	100%
43	30	100%
44	30	100%

these large clusters to 20 core documents while still containing representation of all topic keywords ($>$ 98% average across all trials). We will use these core documents in Chapter 5 to derive definition context for the topic keywords while performing topic coverage assessments.

CHAPTER

4

TOPIC MATCHING

PTNScan requires that cluster topic keywords produced in Chapter 3 are accurately matched to a user-provided abstract. In this chapter we evaluate three text matching techniques to accomplish this goal: (1) using a keyword frequency algorithm (e.g. TFIDF), (2) applying semantic embedding models to user abstracts and topic keyword list, and (3) performing LLM topic prompting before embedding prompt outputs and topic keyword lists. The best-matched topic keyword list should capture the user abstract's conventionality, which is a requirement for the novelty definition we employ.

4.1 Conventionality and Novelty Definitions for Topic Matching

After evaluating z-scores of content overlap and uniqueness across 17.9 million research articles, Uzzi et al. discovered that high-impact novelty combines a high level of conventionality with a high subtopic distinctness. In their study, Uzzi also states single-discipline journals were more often found to have high impact. In translation to topics and subtopics, matching the user's abstract to the document cluster with the highest similarity should capture conventional aspects. Scientific research that is clustered could indicate similarity in conventional, long-standing research questions. For example, transformers, graph theory, and most recently, large language models all serve to

answer scientific questions that derive from a common origin: natural language processing (NLP). Traditional top-k RAG approaches select the highest-ranked documents, but they cannot provide insight into what foundational research components are most relevant. Topic keyword frequency is a popular method for capturing foundational discussions. The most cited topics show progress on research questions that are central to a communities' interests. The overlap of a user's abstract with the most influential keywords across a selected cluster can highlight relationships with the domain's most important and popular ideas. Thus, the highest similarity between a user abstract and its most similar cluster identifies which domain conventionality is most likely to be impactful. Local cluster matches have the best conventionality results, based on Uzzi's claim that higher-impact journals often have similarities within the same domain, driven by their goal of capturing single-topic precision. [Uzz13]

Evaluating differentiating novelty factors.
As discussed above, novelty requires high conventionality and high subtopic distinctness. For our work, we define high conventionality in scientific abstracts as a high similarity with other common or popular abstract topics. From this perspective, a user abstract can be considered conventional if it clusters with a group of similar abstracts. We define the degree of conventionality as how often the cluster's topic keywords are reflected in the user abstract (i.e., higher topic keyword overlap corresponds to higher conventionality). Our goal is to correctly assign the user abstract to a cluster of published documents with a single topic focus. From this perspective, the related keywords should represent relevant subtopics discussed in prior work. If BERTopic captures the most relevant keywords (by frequency weighted c-TFIDF), then some or all of those keywords, regardless of whether they appear explicitly in the user's abstract, can be used to evaluate the conventional components of the abstract. Given this, we establish the following goals.

1. **Construct a gold standard baseline for document matches.**
   Clustering is performed using full document similarity. Establishing a baseline using full document similarity enables us to estimate the best scores for matching a document to a target cluster. Follow-on experiments will only use the topic(s) for matching, and it is anticipated that using less context will decrease the similarity scores compared to the gold standard baseline.

2. **Perform topic keyword matching to a user-provided abstract.**
   It is important to note that a document's abstract maps to the appropriate topic keywords and not simply to a cluster's content, since topic keywords are used in the LLM prompts to construct personalized feedback. If the mapping is incorrect, it could indicate that clustering, the embedding model, or topic modeling itself requires improvement.

3. **Evaluate whether an LLM can provide better abstract topic modeling to improve topic matching.**
   LLMs can provide more universal domain terminology, which in theory could aid in matching topic keywords that are selected from a large document collection versus exact keyword or phrase matching against text in a user's abstract. LLM prompt outputs can also be used to extract the most important concepts in an abstract and to reduce noise.

Scoring

When we compare an abstract to a cluster of documents, we obtain two scores: an *absolute match score* and a *relative match score*. The absolute match score is a percentage of the number of abstracts we tested that correctly matched their target cluster. The score is binary since it is 1 if the assignment is correct, and 0 otherwise.

$$\frac{\text{\# of correctly assigned abstracts}}{\text{Total number of abstracts}}$$

Alternatively, the relative match score uses a normalized distance error equation. Specifically, it normalizes the size of the error based on the maximum possible error for a given abstract and target cluster. Intuitively, it assigns a higher partial accuracy when an abstract is matched to a cluster that is "closer" to the correct cluster. The relative math score for a document is computed as

$$\text{Relative Match Score} = 1 - \left(\frac{d_{\text{assigned}} - d_{\text{min}}}{d_{\text{max}} - d_{\text{min}}} \right), \tag{4.1}$$

where

- d_{assigned} is the distance from an abstract to its original cluster's embedded list of topic keywords,
- d_{min} is the minimum distance from the abstract to any cluster's embedded list of topic keywords, and
- d_{max} is the maximum distance from the abstract to any cluster's embedded list of topic keywords.

4.2 TF-IDF

We apply a cluster averaged version of TF-IDF on all preprocessed abstracts to construct a gold standard baseline. Traditional TF-IDF computes a text frequency vector for individual documents. Our version extends the original TF-IDF by calculating each cluster's average TF-IDF score. We expect this version to maintain better cluster distinction, as the IDF scores still reflect representation

of an individual abstract with respect to the document collection. In theory, this should lead to a more accurate representation of the individual abstracts within a cluster, as opposed to other popular cluster-related TF-IDF algorithms, for example, c-TFIDF used in BERTopic where all the abstracts within the cluster are combined as a single document.

In c-TFIDF, there is less distinction for generic terms. If a generic term occurs often in a cluster, it could be recognized as a specialized term since it appears more in that cluster than in other clusters. Cluster averaged TF-IDF will assign the generic term a lower weight unless it is used frequently in only a small number of abstracts. We expect that our baseline can identify the potential accuracy we can achieve with topic matching when using keyword frequency mapping.

4.2.1 Test Dataset

HDBSCAN clustering is performed by BERTopic across a dataset. For each cluster, we extract an abstract to use to simulate a test abstract that a user could submit. A true positive is assigned for that abstract if it is matched to its original the cluster during our experiments.

4.2.2 Procedure

We construct a TF-IDF matrix for the entire dataset. Next, we remove the TF-IDF vectors for the test documents described above. The TF-IDF values of the remaining documents are then averaged by cluster. We perform a cosine distance comparison of each test document's TF-IDF vector to each cluster's averaged TF-IDF vector. The test document is matched to the cluster with the most similar TF-IDF vector. We run the experiment five times, selecting a different test document for each run.

4.2.3 Results and Discussion

Since each test document was used by BERTopic's HDBSCAN to cluster documents in the original document collection, we expect these accuracy scores to represent a "gold standard" for the best matching outcomes. Both absolute and relative scores reflect this expectation (Table 4.1). Relative accuracy scores averaged 93% over our five experiment runs, and never fell below 89%. Even absolute accuracy, which only scores above zero when a test document is matched to the exact cluster it was extracted from, averaged to 71% and never fell below 59%.

4.3 Topic Keywords to User Abstract Matching

To better simulate a real-world environment, a test document is selected to ensure that cluster topic keywords are assigned independently of any test abstracts matched to the clusters. To do this,

Table 4.1 Cluster averaged TF-IDF document matching performance (relative and absolute scores)

Run	Relative Score	Absolute Score
1	89.3%	58.9%
2	94.2%	70.8%
3	93.8%	72.4%
4	93.9%	74.3%
5	94.5%	78.1%
Average	93.1**%**	70.9**%**

the full dataset is run through BERTopic twice. The first run creates clusters and extracts one test abstract per cluster, exactly as we did in the previous experiments. The test abstracts are removed from the document collection, and BERTopic is run a second time on the remaining documents to generate potentially new clusters for use during cluster–abstract matching. This ensures the test abstracts do not participate in cluster creation.

4.3.1 Test Dataset

The second run of BERTopic is applied to a document collection with the test documents removed. We are measuring accuracy based on how well a test document is matched to its original cluster, but these were clusters from the first run of BERTopic, not the second. To ensure cluster consistency between the first and second BERTopic runs, we measure the similarity of the document groupings (i.e., the document clusters BERTopic creates) between the two runs. We selected a threshold similarity of 95% to guarantee the ability to accurately determine whether a test document is mapped to its "correct" cluster.

When we measured cluster consistency, we found that cluster size has a significant impact on how documents are assigned to clusters. For original clusters that were global and small in size, document groupings changed significantly between BERTopic runs. Given this, it would be unreliable to measure accuracy for test documents extracted from these clusters. To address this, test documents were only extracted from the five "middle" clusters with median cluster sizes. Our consistency checks confirmed that these document groupings remained unchanged between the first and second BERTopic runs.

Table 4.2 Topic keyword to user abstract matching performance (relative and absolute Scores)

Run	Relative Score	Absolute Score
1	71.0%	20.0%
2	82.2%	0.0%
3	78.5%	20.0%
4	81.2%	0.0%
5	67.3%	0.0%
Average	76.0**%**	8.0**%**

4.3.2 Procedure

We used the embedding model *all-MiniLM-L6-v2* to construct embedding vectors for each of the five abstracts selected to act as test documents. The same model is used to construct a sentence embedding vector for the concatenation of the topic keywords assigned to each cluster from the second BERTopic run. This allowed us to use the same cosine distance strategy from the original experiment to determine which cluster's topic keyword embedding vector most closely matched the test abstract's embedding vector. Once the five test abstracts were assigned to clusters from the second run of BERTopic, average absolute and relative accuracies were calculated exactly as in the original experiment.

4.3.3 Results and Discussion

Our results are promising. The average relative accuracy across the three experiment runs was 76% (Table 4.2). Absolute accuracies were much worse, however, averaging 8% and never achieving a correct match in three of the five experiment runs. Although further testing is needed to confirm and better understand why the absolute scores are so poor, we hypothesize that this is due to the characteristics of mean-pool embedding typically used for sentence embedding models, including the one we used. To obtain exactly the right match between the topic keyword list and the user abstract's embedded vector, the keywords that represent topics in the user abstract should be clearly distinguishable. With mean pool embedding, however, uniqueness in the topic keywords could be averaged out with other text, leading to less differentiable topics and lower absolute match scores.

This also highlights the importance of relative accuracy scores. They show that, even when an abstract is not matched to its "correct" cluster, it is matching a cluster that is similar to the correct cluster. Further tuning of clustering and topic modeling parameters may improve performance. Future work could explore ways to extend our test document selection to global and small clusters.

Although they had different document groupings between BERTopic runs, determining a way to include more than the five median clusters while still calculating representative test abstract to cluster accuracies would significantly improve the justification for use of our proposed approach.

4.4 LLM Prompted Topics with Embedding

Mu demonstrated that if an LLM is given common topics, subtopics and similar topics can be extracted from a body of text. Similar to our work, [Mu24] also experimented with several prompting strategies to extract topics from articles and twitter posts using an LLM. We investigated a combination of these two prompting options: (1) entering a basic prompt to extract three topics, and (2) providing seed topics. Following, we performed a variation of LLM-prompted topic modeling by asking the LLM to extract a core topic and related subtopics.

4.4.1 Experiment Design

We constructed our test document collection and document collection clusters in the same manner as described in the previous section. For each of the five test document abstracts, we experimented using the following: "What is the main topic and its related subtopics from this abstract? For each topic and subtopic, extract a compact set of keywords only. Output nothing else." The goal of the prompt was described to Grok for refinement, so the final prompt (above) was produced by Grok. The prompt outputs are generated using the *gpt-4o model*. Embedding vectors for each test abstract's prompt output are embedded with *all-MiniLM-L6-v2*. The resulting vectors can then be compared to the topic keyword sets of the document collection's clusters using cosine distance, exactly as in the previous experiments.

4.4.2 Results and Discussion

Relative and absolute accuracy results were even higher than those found in prior experiments (Table 4.3). We achieved a relative average accuracy of nearly 100%, correctly matching all five test abstracts to their original clusters in four of the five experiment runs. This produces a very high average absolute accuracy of 96%, since exact matches were found in almost all cases, producing an absolute accuracy score of 1 and very few scores of 0. Our results should be viewed through the constraint of only selecting abstracts from median-sized clusters, however. These clusters each contain approximately thirty documents. In future work, we could select the best thirty documents using our MST-Jaccard algorithm for each cluster. This would allow us to compare clusters between the first and second runs of BERTopic (prior to and after extracting a test document from each

Table 4.3 Topic keyword to user abstract matching performance (relative and absolute scores)

Run	Relative Score	Absolute Score
1	100.0%	100.0%
2	100.0%	100.0%
3	99.0%	80.0%
4	100.0%	100.0%
5	100.0%	100.0%
Average	99.7**%**	96.0**%**

cluster) to determine whether using clusters that are similar in size produces similar document groupings. If so, it would allow us to extract test documents from many more clusters, which in turn would provide a much more comprehensive investigation of our abstract to topic keyword matching approach.

4.5 Conclusion

This chapter introduces topic modeling algorithms that apply a "novelty" definition that requires a topic to be both highly conventional and highly unique. In terms of topic modeling, we represent conventionality as the overlap between the frequent topic keywords extracted from the document collection's clusters and a user-provided abstract. Our results show that while all of the three approaches tested (cluster averaged TF-IDF, embedded abstract to topic keyword matching, and LLM prompting to suggest topics prior to prompt output to topic keyword matching) produced stable results, the prompt output achieved the best performance, with a 99.7% relative accuracy. The topic keyword list from this third approach will be used in follow-on experiments as a topic guide for LLM prompts that analyze topic coverage assessment and novelty detection in the next chapter.

CHAPTER

5

PERSONALIZATION

The last stage of our PTNScan pipeline is personalizing the topic coverage assessment of an abstract for a paper under consideration for submission. In prior chapters, we explored the best approaches to cluster documents in a document collection, capture a single topic for each cluster, and match an abstract to the most semantically similar cluster. In this chapter, we describe how the topic keywords-user abstract pair is used to personalize topic coverage and novelty assessment.

Topic mining personalization provides topic coverage analysis for a user abstract. LLMs apply topic keywords from the most relevant cluster to infer domain context for the user's abstract. Specialized definitions for topic keywords are used to represent the topic domain of each cluster. These definitions provide clarification for specific subdomain interests (e.g., "computer science education" versus "education" or "language" in "natural language" versus "foreign language").

From these definitions, an LLM infers novelty contexts in a structure relevant for scientific publications. A traditional novelty definition is described as a combination of both highly conventional and unique content [Uzz13]. The conventional aspect maps the user abstract to historical relevance. Examples of this can be a problem statement in a popular subdomain, traditional benchmarking datasets, or a methodology related to a common technology such as transformers. Unique aspects are factors that differentiate the abstract from the documents in the matching cluster (e.g., using a novel technique or exploring a niche angle for a popular problem space).

The novelty components of a scientific abstract are typically found in either the "Problem Statement" section by introducing a new subdomain or the "Methodology" section where a novel solution is proposed. Our work focuses on indirectly extracting the key features that distinguish the uniqueness of these sections through a paper's abstract. We represent these key features with topic keywords for each document cluster.

Different LLM prompts were tested to extract context between a topic keywords-user abstract pair. The goal is to maximize the accuracy and clarity of the topic keyword definitions, topic coverage, and novelty extraction. The strengths and limitations of the prompts are evaluated through three approaches.

1. **Provide definitions for the topic keywords that apply to the majority of the documents within the matching cluster.**
   In scientific abstracts, we expect topic keywords to be present as core components of an abstract (e.g., problem scopes, methodologies, or commonly used benchmark datasets). We assume that the same keywords can also appear across multiple core components (e.g., appearing in the problem scopes of some documents and in the methodologies of others). We observed terminology ambiguity for topic keywords when context is missing (e.g., "owl" as a nocturnal animal versus as an acronym for Web Ontology Language). Our goal is to identify the context that provides topic keyword definitions that are accurate for the majority of the documents in the matching cluster.

   We address this by prompting an LLM to define topic keywords using three different context perspectives: (1) No additional context, (2) additional context based on the additional topic keywords BERTopic assigned to the matching cluster, and (3) additional context based on the complete documents within the matching cluster, supplied in a RAG-based manner. Clusters with large document collections (i.e., global clusters) can contain too much noise to obtain clear definitions. We wondered if representative subsets could reduce noise while still providing sufficient context for constructing accurate topic keyword definitions. Prior experiments in Chapter 3 confirm that the MST-Jaccard algorithm can successfully select the most representative full document samples. We proceeded with the MST-Jaccard technique to select the samples for our RAG-based approach.

2. **Provide a topic coverage assessment for the user-provided abstract.**
   We attempt to leverage an LLM to generate a personalized topic coverage assessment. This assessment compares the matching cluster's topics to a user-provided abstract. Our best topic matching results use an LLM to prompt topics and subtopics for a user abstract before

semantic cluster matching. We select our abstract's topic keywords by leveraging this same technique.

Next, we prompt an LLM to draft a topic comprehension analysis. The analysis compares topic keywords by determining: (1) Which topics are covered in the user abstract? and (2) Which topics are not covered but are recommended for coverage? Several prompt types are investigated to evaluate different factors that can influence the accuracy of our assessments.

Results from our analysis identified topic temporal evolution as a specific characteristic that contributes to inaccurate LLM output. Document clusters are more temporally agnostic compared to LLM-prompted assessments. Regardless of recency, the cluster is primarily dependent on the main subtopics present in its documents. LLMs contain more temporal bias, as several studies describe an LLM's tendency to favor more recent content [Fan25].

Our experiments tested the how topic coverage assessments are impacted by factors such as temporal bias. For example, if temporal bias is present, it can lead to incorrect topic coverage assessments that fail to adequately consider more historical work. Here, the LLM is applying its most recent knowledge of a subdomain instead of correctly assessing the actual relevance of a paper to the subdomain. An example would be computer algorithms that were developed primarily for centralized architectures in the past, but have recently been extended to focus on distributed computing. We observed that older, seminal abstracts on popular algorithms that are still common today are ignored by an LLM, which instead focuses on more recent distributed computing articles due to temporal recency bias. This occurs even when distributed computing is not an important focus of the underlying algorithm.

3. **Optimize the performance of novelty detection by including topic keywords with LLM prompts.**
   Our final goal is to optimize the performance of novelty detection in user abstracts. Using an LLM to detect novelty in user-provided abstracts is non-trivial, since the assessment requires accurate analysis of several subjective factors [Wu25]. We attempt to improve an LLM's ability to identify novelty by guiding it from a topic-oriented perspective. This attempts to detect novelty by: (1) translating the traditional novelty definition to a topic-driven context, and (2) applying the topic novelty function to detect novel subtopics.

   For our research, novelty topic translation consists of the following steps.

   (a) There exist topics that have high similarity to at least one cluster. This similarity occurs because these topics are either a continuation or a new combination of prior research top-

ics. This continuation or combination of historical work draws a parallel to the traditional novelty definition by addressing the "highly conventional" criteria.

(b) Topics within a user abstract are considered unique if they are not explicitly or implicitly included in a cluster's topic keywords. This corresponds to the "uniqueness" criteria of traditional novelty definition.

We test whether our translation is useful for LLM novelty prompts by initially evaluating whether it improves novelty detection during zero-shot prompting without additional context. Our observations demonstrate a high potential for applying novelty in this topic-driven context. We observed a statistical improvement in accuracy when using topic-guided prompts. Given this, we manually reviewed which of the LLM-discovered novel subtopics were useful. We found that simply prompting the LLM for novel subtopics without additional context delivers output that can be somewhat random and often not helpful (e.g., declaring an abstract as novel with no explanation or obvious reason). On the other hand, applying our novelty topic translation framework does enable the LLM to select subtopics with higher novelty influence (e.g., selecting keywords focused on identifying novel factors within the user abstract's methodology).

5.1 Generating Standard Definitions for Topic Keywords

We begin our novelty assessment by generating a single definition for each topic keyword that is applicable to the majority of documents in a cluster. We conducted three different LLM prompting experiments to provide definitions using the following context: (1) other keywords within the topic, (2) without additional context, and (3) additional context provided through a RAG-driven LLM approach that uses a representative subset of the documents from the matching cluster. The best technique should allow the LLM to capture the appropriate breadth and depth in its topic definitions (i.e., topic definitions that are neither too broad nor too constrained for the specific subdomain).

Our prompts were generated by Grok after describing our intentions. We found this method to be more successful than manually generating our own prompts. *GPT-4o* was selected for LLM output generation using the Grok-generated prompts. The basic methodology of these approaches is to select certain polysemous keywords (e.g., "system", "activity", "control"), then compare the prompt output to manually constructed definitions. Future work could include more quantitative approaches that employ subject matter expertise to determine accuracy for a larger set of topic keywords and definition pairings.

We found that using RAG with representative subsamples, the third approach described above,

provided the best topic keyword definitions. The topic keyword sets remained consistent with the majority of the keywords included in the corresponding full document text. The topic keywords and definition pairings were broadly defined but still relevant to a parent cluster.

5.1.1 Experiment Design

Documents and topic keywords for clusters 40-45 were used from our previous experiments. These are the same clusters that were selected to evaluate topic matching performance. They were chosen because their document groupings were stable compared to larger and smaller clusters over subsequent runs of BERTopic that included or excluded documents acting as test abstracts. Based on these clusters, the following polysemous keywords were selected for manual evaluation: "control", "system", "genes", "expression", "owl", and "activity".

5.1.2 LLM Prompting Procedures

We explored the three different approaches for using an LLM to generate definitions for topic keywords, then compared their results to select the semantically most accurate method.

Approach One. Prompt an LLM to provide a definition of a keyword given a context of the other keywords from the topic cluster.
The Grok interface was used as our LLM. Since only a few examples were required with selected keywords acting as target and context, the cost to use Grok was not prohibitive. The following prompt was issued to Grok.

*Given this set of topic keywords, provide the definition of "**[target keyword]**" in relation to the surrounding words: **[additional topic cluster keywords]***

For example:

Given this set of topic keywords, provide the definition of "activity" in relation to the surrounding words: protein, proteins, peptides, peptide, strains, antimicrobial, amps, activity, bacterial, bioinformatics.

Approach Two. Prompt an LLM to provide a scientific definition for each keyword without additional context.
The following prompt was issued to Grok.

Provide a concise, accurate definition (1-2 sentences) for the term "[target keyword]" in the context of

scientific research. Focus on its most common meaning in research papers.

Approach Three. Prompt an LLM to provide a scientific definition for each keyword while providing a representative subset of documents using a RAG-based approach.
The MST-Jaccard algorithm (Chapter 3) was used to select the top twenty documents from each cluster. These documents were used as RAG support to provide the LLM with context for defining topic keywords. The following prompt was issued to Grok to provide context.

You are an expert in scientific research terminology. Here are highly representative abstracts/documents from a research topic cluster: ***[twenty documents selected by MST-Jaccard]***

Next, the definition prompt was issued.

*This cluster has "****[target keyword]****" as one of its top representative keywords. Provide a concise (1–2 sentences), accurate definition of the term "****[topic keyword]****" as it is used in this specific research context. Focus on how the term is applied in these papers. Do not give a generic dictionary definition.*

5.1.3 Results and Discussion

Approach One.
When only providing context in the form of topic keywords, the LLM would hallucinate interpretations of the relationships between the keywords' contexts. For example, the LLM connected "activity" to "antimicrobial activity" without instructions to do so. We expect these randomly generated connections will lead to inaccurate topic definitions. Therefore, this approach is not recommended.

Approach Two.
Prompting for topic keyword definitions without additional context resulted in definitions that are not related to the topic. For example, "owl" is an acronym for "Web Ontology Language" in computer science. Without context, however, the LLM treats the keyword as a nocturnal bird of prey. Given the LLM's inability to properly identify which context to assign to a keyword, we do not recommend this approach.

Approach Three.
Using an MST–Jaccard–RAG approach provided the most accurate topic keyword definitions. The definitions returned from our prompt were relevant to the cluster's topic scope, and they did not encourage hallucinations of unstated keyword relationships. This approach proved to be the most effective in generating topic keyword definitions that are applicable to the underlying abstracts.

5.2 Topic Coverage for User-Provided Abstracts

Once definitions for each cluster's topic keywords are generated, they can be compared to the user-provided abstract to select the document cluster that is most semantically similar. This allows us to identify topics in the matching cluster that are included or excluded in the abstract. We applied LLM prompting to address this goal. First, we translate the abstract to address the novelty requirement of "highly conventional" (as described above). This generates an LLM prompt that correctly identifies overlap between the matching cluster's topic keywords and the user's abstract. Initial investigations revealed that an LLM will attempt to extend its response beyond simple keyword matching. For example, the LLM will infer general knowledge of a subdomain it thinks the user's abstract is referencing. While this can be advantageous for topic coverage accuracy if the subdomain is correctly interpreted, it can also be unreliable, as even minor nuances in properties like naming conventions can lead to inaccurate topic matching.

To avoid this, we explicitly direct the LLM in our prompts to provide transparency about why a topic is considered relevant, rather than allowing fully automated decision-making. Importance ratings are also requested to provide the level of topic relevance inferred by the LLM.

5.2.1 Experiment Design

We manually selected a subset of abstracts that contain common publication nuances to explore an LLM's behavior toward these abstract characteristics, since these are the types of subtle properties that were previously seen to be difficult to manage correctly. Note that these are in addition to the core goal of identifying topic overlap and novelty. The selections are also driven by what we assume would be expected use cases of our tool.

1. A user who prefers abstract brevity but still wants to ensure topic completeness.
2. A user submitting an abstract with a niche scope that can include many acronyms.
3. A user preparing for publication resubmission after a significant amount of time has passed since the original submission.

From these use cases, we manually selected three abstracts that cover the following characteristics to explore the limitations of an LLM.

1. Very short abstracts (i.e., less than 3 sentences).
2. Acronyms in context with seemingly rare subdomain terminologies.
3. Older abstracts.

5.2.2 Procedure

We experimented with both the Grok and ChatGPT interfaces to compare prompt output responses to abstract nuances. We would first prompt using Grok and then ask ChatGPT to comment on the accuracy of Grok's output, and vice versa. We also provide both LLMs with the same prompt to identify conflicting topic coverage assessments. We found that, although not common, at times both LLMs provided differing assessments (e.g., ChatGPT saying Grok was wrong, or different responses from Grok and ChatGPT for a common prompt). Since situations where significant disagreements occurred were rare, and because the responses to our prompts were considered to be of a high quality even in these situations, we do not believe this approach needs to be excluded from our pipeline. Future work should conduct additional experiments to investigate why outputs are not always stable across different LLMs, however.

Our final evaluation of an LLM's ability to address our core goal of identifying topic overlap and novelty was performed using the following prompt.

*You are an academic writing assistant helping an author enhance the novelty of their abstract. Given the following list of topics in the research field: "**[list of topic keywords]**". They are topics that come from a cluster of related papers. Please do not assume these topics are inherently related to the actual abstract below.*

*And the user's abstract: **[user abstract]***

Do the following:
*1. **Included Topics**: Identify which of the given topics are addressed in the abstract.*
*2. **Missing Topics**: For the topics not included, provide: - **State the reason why including this would underscore conventionality aspects of this work** why inclusion could enhance the novelty or relevance. - **Suggested angle or approach** for incorporating it into the abstract. - **Importance of inclusion (1-5)**, where 5 = most important to highlight.*
***Output Format**: Provide results as clear tables or bullet points for easy reference. Consider how the missing topics that have at least an inclusion score of 5 or higher contribute to explaining the prior state of the field.*

5.2.3 Results and Discussion

The LLM provided its topic coverage output in a clear and easy-to-follow format. Interpreting accuracy is more challenging. The reasons the LLM states to include missing topics are not always reliable, as both LLM peer checking (e.g., ChatGPT peer checking Grok) and manual analysis identified inaccurate matchings.

- LLM confusion occurred when interpreting correct subdomains and attempting to manage details related to acronyms and naming conventions. An example is a naming convention that uses acronyms to differentiate unique algorithm characteristics from their broadly related algorithmic family. Inaccurate subtopic selection can lead to further errors in defining the data structure or the common methodology being applied.

- LLMs can hallucinate relationships between topic keywords. Our experiments show that including the matching cluster's topic keyword list in the prompt, regardless of our instruction not to assume keyword relationships, still biases the LLM towards results that imply the cluster's keywords are related to the user's abstract. We hypothesize that this is due to an LLM interpreting information via token sequencing. If true, simply adding the keywords might direct the LLM towards training text that is related to the keywords.

- LLMs are biased towards more recently published paper abstracts rather than material that is more relevant to the user abstract's publication period, for example, in traditional algorithms that have undergone continuous development. We investigated this issue by studying whether an LLM can accurately detect missing topics using temporally differentiating features (e.g., older abstracts would be missing topics that are more recent).

  An older abstract was used to test whether more recent topic keywords are detected as "missing." A correct prompt response would identify that an older abstract is missing many of the more recent topics and recommend them for inclusion. For example, an article about two-server optimization algorithms in 2015 would not include recent keywords such as "distributed computing," so they should be detected as "missing to be included." The result, however, showed the opposite and inaccurately recognized "distributed computing" as covered. Although not conclusive, this provided further evidence of an LLM biasing towards recent topics because it is projecting modern trends onto older text even when it is not applicable.

Our overall conclusion is that while an LLM is useful in providing feedback on missing topics to include, the nuances specific to a particular domain can lead to errors. We recommend that the results LLMs return should be supplemented, when possible, by a domain expert, but they should not be replaced by a human-generated topic assessment. Although there is an opportunity for improvement, the LLMs' responses are generally of a high quality, so these types of errors do not outweigh the overall accuracy and efficiency of LLM-reported topic overlap.

5.3 Topic Novelty for User-Provided Abstracts

We next evaluated whether an LLM can accurately identify novel topics in a user-provided abstract. To do this, we conducted experiments to determine whether including topic keywords improves an LLM's ability to detect novelty. Previously, we discovered that an LLM is biased towards including relationships with keywords that are embedded in a prompt. We recognized that this could, in fact, be advantageous if the topic keywords are used to guide the LLM towards novelty topics relative to common topics in peer publications. In theory, this should reduce randomness in novelty direction (e.g., selecting random words solely for uniqueness) and provide more useful feedback.

We tested our hypothesis using two approaches.

1. We examine the limits of zero-shot prompting for "novelty" without additional context.
2. We investigate whether incorporating topic keywords into zero-shot prompting improves performance for identifying novel topics in the user's abstract.

The output of both prompts are topic keywords, since novel additions can be identified as new subtopics that historically relate to common topics. The cosine distance is calculated between each prompt output and the topic keywords of a target cluster, similar to the topic matching experiments described in Chapter 4.

5.3.1 Experiment and Design

The topic keywords from the median clusters 40–44 were used in our experiments. This choice of clusters matches our prior experiments in Chapters 3 and 4. Five random documents were selected from each cluster, resulting in a dataset of twenty-five document abstracts.

5.3.2 Procedure

An explanation of both approaches is provided to Grok, which produced prompts that we refined manually. The result is input to *GPT-4o* together with each test document. GPT returns novel subtopics as its output. We calculate the maximum cosine distance (i.e., the minimum similarity) between the output subtopic list and the cluster's topic keywords, similar to our approach for topic matching. Once distances are obtained for all test documents using both prompts, a t-test was run to identify statistically significant differences between the two approaches. Results confirm that a significant difference exists when including topic keywords. To confirm the statistical findings,

we performed a follow-on manual evaluation to determine whether the discovered subtopics are useful and accurate by comparing them to the original abstract.

Approach One: Prompt without topic keywords.
Identify up to 5 subtopics that are the most NOVEL. Extract a compact set of keywords only. Output nothing else.

Approach Two. Prompt with topic keywords.
You are an expert in scientific topic analysis. Given the following abstract and the main topic keywords that already represent its broad category, identify up to 5 subtopics from the abstract that are not represented in the main topic keywords. Output only keywords.
Main topic keywords: ***[Matching cluster's topic keywords]***
Abstract: ***[User-provided abstract]***

5.3.3 Results and Discussion

A t-test score of $p = 0.098$ was obtained, indicating moderate statistical significance for $\alpha = 0.1$, a threshold considered acceptable for experimental analysis. Manual observations indicate that including topic keywords enables the LLM to focus on unique aspects that contribute to novelty (e.g., specific tool modifications or methodologies). Without topic keywords, the LLM more often selected information that was random and irrelevant to novelty. We believe this is because the topic keywords encourage the LLM to search for context relevant to the cluster, since that is what the keywords represent. This means that the LLM may not search for novelty only from the user abstract, but also from text patterns that include "novelty" and from the topic keyword list. This would result in the LLM applying those patterns to the user's abstract.

We also noted in prior prompting experiments that the worst results occur when no additional context is provided. Results of prompts with no topic context resulted in the LLM incorrectly hallucinating topic and keyword relationships or selecting phrases at random. Our manual examination of approach one confirmed similar behavior. It is recommended that topic keywords be included as additional context when prompting for novel subtopics.

5.4 Conclusions

In this chapter, we evaluated several methods for assessing topic coverage and identifying novelty subtopics.

1. Methods to define topic keywords associated with clusters that are semantically closest to a

user-provided abstract.

2. Prompting experiments that provide topic coverage clarity while avoiding negative effects on accuracy due to factors such as LLM temporal bias.

3. Evaluation of the improvements provided when topic keywords are included in an LLM prompt for novel subtopics contained in a user's abstract.

We found that the best approach to define topic keywords was an MST–Jaccard–RAG–based LLM prompt technique. MST–Jaccard successfully extracted core documents from each cluster, reducing the number of documents in large clusters and allowing us to use the full document text as background context for topic keyword definitions. The LLM generated definitions that focused on the cluster's content without overspecializing. These definitions could then be broadly applied to additional abstracts. Due to time constraints, we were unable to include experiments that also included the derived topic keyword definitions in additional follow-on steps (e.g., topic coverage analysis or topic novelty for user abstracts). These experiments were restricted to the topic keyword list without definitions. Future work can include these definitions in the follow-on steps by replacing the topic keyword list with the topic keywords and corresponding definitions.

While we found that the LLM provided clarity and structure during topic coverage assessment (i.e., the prompt output format made it clear to the user why a topic was deemed necessary or not through descriptive reasoning and ratings), we discovered LLM limitations in accurately providing topic coverage. We discussed the limitations for analyzing the relationship between keywords (e.g., temporal bias or limitations in understanding the utility of naming conventions or acronyms). It is recommended that the LLM's output be used as a recommendation to inform the user, rather than a complete replacement for user judgment.

Finally, we investigated the impact of including topic keywords in prompts to discover novel subtopics. We found that including topic keywords provided a moderate statistical difference ($p < 0.1$) which is significant for exploratory studies. Our manual assessments on novelty usefulness confirmed that including topic keywords reduced selecting random keywords and provided the LLM with guidance to select subtopics that have actual novelty. These results suggest that including topic keywords can improve an LLM's prompting in novelty subtopic detection. Future work is encouraged to extend testing to include more quantitative assessments.

This chapter completes the personalization process of topic coverage assessment and novelty subtopic detection. Our personalization approach defines topic keyword terminology, explores prompting techniques for providing topic coverage assessments, and compares prompting techniques for novelty subtopic detection.

CHAPTER

6

FUTURE WORK

The methods developed in this thesis demonstrate success in clustering research abstracts to obtain appropriate single-topic precision and enable accurate topic matching to user-provided abstracts. Our results show reliable performance in obtaining relevant topics to act as a literature review for these abstracts. Topics were then defined with MST-Jaccard-RAG and LLM to achieve accurate topic keyword definitions. The definitions were used to provide transparent topic inclusion/exclusion recommendations and novelty detection assessments via LLM prompting. These initial findings offer the potential to further explore improvements that can manage edge cases and improve accuracy by: (1) more fully accounting for outliers, (2) expanding a test dataset by improving clustering and document stability limits, (3) investigating improvements in prompting that includes topic keyword definitions, and (4) allowing an LLM to express ambiguity when reporting whether a topic is covered.

In Chapter 3, our experiments evaluated abstracts that were initially included in a target cluster. In the future it is recommended to explore new approaches that handle novelty detection for outliers. A reasonable approach, for example, would answer the question: “Can LLM prompting be used to extract topics from the outliers and then match to the best semantic topic keyword list?” If there are no reliable matches, we could then ask: “Should a new topic keyword list be created and included in the list of topic keywords?” This can lead to supporting topic keyword matches with outliers, where currently only successful clustered abstracts are explored.

In many of our experiments, we only used the five median-sized clusters due to the inability to guarantee stable document groupings in smaller or larger clusters. We encourage expanding the number of clusters that can provide test documents by better evaluating and managing how cluster size impacts to grouping stability. Future studies could also examine the topic matching performance for both larger and smaller clusters.

Finally, we designed prompts to generate reliable topic keyword definitions. However, these definitions were not used in the follow up prompting for topic coverage assessment and topic novelty detection. We recommend replacing the section in the prompt that provides topic keywords with a list of topic keyword-topic definition pairs. New evaluation metrics would need to be developed to reflect accuracy improvements resulting from including these definitions. Furthermore, our topic coverage analysis limits the LLM to suggesting topic inclusion or exclusion, but does not allow for reporting ambiguity. Future work could experiment with prompts that allow the LLM to express uncertainty about "unknown topic placement" with transparency in the reasoning behind its decisions. We believe allowing additional flexibility has potential to increase the reliability of LLM topic coverage accuracy, since it will not force uncertain decisions to be excluded from an assessment.

CHAPTER

7

CONCLUSIONS

Our work introduced PTNScan: Personalized Topic and Novelty Scan as an effective topic coverage and novelty detection framework. This involved combining hierarchical clustering and topic modeling with a novel MST-Jaccard sampling algorithm and topic-guided LLM prompts. The result is a topic-centered analysis which guides a user by identifying conventional topics that are common in previously published work but not included in the user's abstract, together with topics that are unique and potentially novel contributions.

Clustering and topic modeling of abstracts from previously published papers are used to identify relevant topic keywords. The keywords are semantically compared to a user's abstract, producing the recommended list of missing and novel topics. Our work proposes a novel algorithm, MST-Jaccard, to select representative abstract subsets to enable efficient identification of core topics while reducing noise in large clusters.

The abstracts selected through the MST-Jaccard algorithm are then used in a RAG-LLM prompt to define topic keywords. We concluded with experiments that evaluate an LLM's ability to perform topic coverage classification and topic novelty detection. The results of our framework show the potential to automatically provide qualitative topic coverage evaluation. Recommended future work focuses on improvements in managing outliers, expanding the test datasets used during our experiments, and improved prompting to increase an LLM's topic keyword-guided coverage

assessment and novelty detection.

BIBLIOGRAPHY

[Abi20] Abid, A. et al. “Semi-Automatic Classification and Duplicate Detection From Human Loss News Corpus”. *IEEE Access* **8** (2020), pp. 97737–97747.

[Adh19] Adhikari, A. et al. “DocBERT: BERT for Document Classification” (2019). arXiv:1904.08398 [cs].

[Alt16] Altszyler, E. et al. “Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database” (2016).

[Amp24] Ampazis, N. “Improving RAG Quality for Large Language Models with Topic-Enhanced Reranking”. en. *Artificial Intelligence Applications and Innovations*. Ed. by Maglogiannis, I. et al. Cham: Springer Nature Switzerland, 2024, pp. 74–87.

[Bah20] Baharav, T. Z. et al. “Spectral Jaccard Similarity: A New Approach to Estimating Pairwise Sequence Alignments”. en. *Patterns* **1**.6 (2020), p. 100081.

[Bel19] Beltagy, I. et al. “SciBERT: A Pretrained Language Model for Scientific Text”. en. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3613–3618.

[Bha24] Bhattacharjee, B. et al. “INDUS: Effective and Efficient Language Models for Scientific Applications”. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Ed. by Dernoncourt, F. et al. Miami, Florida, US: Association for Computational Linguistics, 2024, pp. 98–112.

[Bis23] Biswas, S. et al. “ChatGPT and the Future of Journal Reviews: A Feasibility Study”. *The Yale Journal of Biology and Medicine* **96**.3 (2023), pp. 415–420.

[Bis24] Biswas, S. S. “AI-Assisted Academia: Navigating the Nuances of Peer Review With ChatGPT 4”. en. *The Journal of Pediatric Pharmacology and Therapeutics* **29**.4 (2024), pp. 441–445.

[Ble06] Blei, D. M. & Lafferty, J. D. “Dynamic topic models”. *Proceedings of the 23rd international conference on Machine learning*. ICML ’06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 113–120.

[Ble03] Blei, D. M. et al. “Latent Dirichlet Allocation”. *Journal of Machine Learning Research* **3** (2003), pp. 993–1022.

[Cam13] Campello, R. J. G. B. et al. “Density-Based Clustering Based on Hierarchical Density Estimates”. en. *Advances in Knowledge Discovery and Data Mining*. Ed. by Pei, J. et al. Berlin, Heidelberg: Springer, 2013, pp. 160–172.

[Cas22] Castro, A. P. et al. “Weighting construction by bag-of-words with similarity-learning and supervised training for classification models in court text documents”. en. *Applied Soft Computing* **124** (2022), p. 108987.

[Che06] Chen, C. “CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature”. en. *Journal of the American Society for Information Science and Technology* **57**.3 (2006), pp. 359–377.

[Che21] Chen, H. et al. “Identify Topic Relations in Scientific Literature Using Topic Modeling”. *IEEE Transactions on Engineering Management* **68**.5 (2021), pp. 1232–1244.

[Che24] Cherradi, M. & El Haddadi, A. “Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques”. *Seminars in Medical Writing and Education* **3** (2024), p. 68.

[Dev19] Devlin, J. et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Burstein, J. et al. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[Don24] Dong, F. et al. “Once Read is Enough: Domain-specific Pretraining-free Language Models with Cluster-guided Sparse Experts for Long-tail Domain Knowledge”. en. *Advances in Neural Information Processing Systems* **37** (2024), pp. 88956–88980.

[Ert22] Ertl, O. “ProbMinHash – A Class of Locality-Sensitive Hash Algorithms for the (Probability) Jaccard Similarity”. *IEEE Transactions on Knowledge and Data Engineering* **34**.7 (2022), pp. 3491–3506.

[Est96] Ester, M. et al. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, 1996, pp. 226–231.

[Fan25] Fang, H. et al. “Do Large Language Models Favor Recent Content? A Study on Recency Bias in LLM-Based Reranking”. en. *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. Xi’an China: ACM, 2025, 85–94.

[Gov23] Governor, K. et al. “Near duplicate detection of images with area and proposed pixel-based feature extraction”. en. *Concurrency and Computation: Practice and Experience* **35**.2 (2023).

[Gro22] Grootendorst, M. “BERTopic: Neural topic modeling with a class-based TF-IDF procedure” (2022).

[Guo23] Guo, E. et al. “Automated paper screening for clinical reviews using large language models”. *Journal of Medical Internet Research* (2023).

[Hat21] Hatefi, A. et al. “Cformer: Semi-Supervised Text Clustering Based on Pseudo Labeling”. en. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Virtual Event Queensland Australia: ACM, 2021, pp. 3078–3082.

[Hot33] Hotelling, H. “Analysis of a Complex of Statistical Variables into Principal Components”. *Journal of Educational Psychology* **24**.6 (1933), pp. 417–441.

[Hsu24] Hsu, C.-C. et al. *CHIME: LLM-Assisted Hierarchical Organization of Scientific Studies for Literature Review Support*. arXiv:2407.16148. 2024.

[Com] *https://commoncrawl.org*. en.

[Arx] *https://info.arxiv.org/about/index.html*.

[Iko23] Ikotun, A. M. et al. “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data”. *Information Sciences* **622** (2023), pp. 178–210.

[Jeo24] Jeon, Y. et al. “Network-based topic structure visualization”. en. *Journal of Applied Statistics* (2024), pp. 1–15.

[Kim19] Kim, S.-W. & Gil, J.-M. “Research paper classification systems based on TF-IDF and LDA schemes”. en. *Human-centric Computing and Information Sciences* **9**.1 (2019), p. 30.

[Kow17] Kowsari, K. et al. “HDLTex: Hierarchical Deep Learning for Text Classification”. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2017, pp. 364–371.

[Kul20] Kulkarni, A. et al. “5 - Foundations of data imbalance and solutions for a data democracy”. en. *Data Democracy*. Ed. by Batarseh, F. A. & Yang, R. Academic Press, 2020, pp. 83–106.

[Kum24] Kumar, D. et al. “Artificial Intelligence Driven Trend Forecasting: Integrating BERT Topic Modelling and Generative Artificial Intelligence for Semantic Insights”. en (2024).

[Lak24] Lakatos, R. et al. “Designing Prompts and Creating Cleaned Scientific Text for Retrieval Augmented Generation for More Precise Responses from Generative Large Language Models”. *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*. 2024, pp. 1–6.

[Lee99] Lee, D. D. & Seung, H. S. “Learning the parts of objects by non-negative matrix factorization”. en. *Nature* **401**.6755 (1999), pp. 788–791.

[Li24] Li, M. et al. “Evaluating the effectiveness of large language models in abstract screening: a comparative analysis”. *Systematic Reviews* **13**.1 (2024), p. 219.

[Lim21] Lima, B. V. A. et al. “Deep semi-supervised classification based in deep clustering and cross-entropy”. en. *International Journal of Intelligent Systems* **36**.8 (2021), pp. 3961–4000.

[Lin25] Lin, E. et al. “Evaluating and Enhancing Large Language Models for Novelty Assessment in Scholarly Publications”. en. *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*. Albuquerque, New Mexico, USA: Association for Computational Linguistics, 2025, pp. 46–57.

[McI18] McInnes, L. et al. “UMAP: Uniform Manifold Approximation and Projection”. en. *Journal of Open Source Software* **3**.29 (2018), p. 861.

[Mik13] Mikolov, T. et al. “Distributed Representations of Words and Phrases and their Compositionality”. *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013.

[Mu24] Mu, Y. et al. “Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling” (2024).

[Niw] Niwattanakul, S. “Using of Jaccard coefficient for keywords similarity.” *Proceedings of the international multiconference of engineers and computer scientists* **Vol. 1. No. 6. 2013** ().

[NS24] Noe-Steinmüller, N. et al. “Defining suffering in pain: a systematic review on pain-related suffering using natural language processing”. en. *Pain* **165**.7 (2024), pp. 1434–1449.

[Nur23] Nurfadhila, B. & Girsang, A. S. “Identifying Indication of Depression of Twitter User in Indonesia Using Text Mining”. en. *International Journal of Intelligent Systems and Applications in Engineering* **11**.2 (2023), pp. 523–530.

[Pap19] Pappagari, R. et al. “Hierarchical Transformers for Long Document Classification”. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019, pp. 838–844.

[Pat20] Patel, H. et al. “A review on classification of imbalanced data for wireless sensor networks”. en. *International Journal of Distributed Sensor Networks* **16**.4 (2020), p. 155014772091640.

[Pen14] Pennington, J. et al. “Glove: Global Vectors for Word Representation”. en. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543.

[Sch24] Scherbakov, D. et al. *The emergence of Large Language Models (LLM) as a tool in literature reviews: an LLM automated systematic review*. arXiv:2409.04600. 2024.

[Sin22] Singh, K. N. et al. "A novel approach for dimension reduction using word embedding: An enhanced text classification approach". *International Journal of Information Management Data Insights* **2**.1 (2022), p. 100061.

[Sma14] Small, H. et al. "Identifying emerging topics in science and technology". *Research Policy* **43**.8 (2014), pp. 1450–1467.

[Tra24] Tran, N. & Litman, D. "Enhancing Knowledge Retrieval with Topic Modeling for Knowledge-Grounded Dialogue". *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Calzolari, N. et al. Torino, Italia: ELRA and ICCL, 2024, pp. 5986–5995.

[Uzz13] Uzzi, B. et al. "Atypical Combinations and Scientific Impact". en. *Science* **342**.6157 (2013), 468–472.

[Ver20] Verma, V. & Aggarwal, R. K. "A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective". en. *Social Network Analysis and Mining* **10**.1 (2020), p. 43.

[Wal14] Walden, J. et al. "Predicting Vulnerable Components: Software Metrics vs Text Mining". *2014 IEEE 25th International Symposium on Software Reliability Engineering*. ISSN: 2332-6549. 2014, pp. 23–33.

[Wei22] Wei, F. et al. "Semi-Supervised Clustering with Contrastive Learning for Discovering New Intents" (2022). arXiv:2201.07604 [cs].

[Wen22] Weng, M.-H. et al. "Identification and Visualization of Key Topics in Scientific Publications with Transformer-Based Language Models and Document Clustering Methods". en. *Applied Sciences* **12**.21 (2022), p. 11220.

[Wil23] Wilkins, D. "Automated title and abstract screening for scoping reviews using the GPT-4 Large Language Model" (2023). arXiv:2311.07918.

[Wu25] Wu, W. et al. "Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge". en. *Journal of the Association for Information Science and Technology* **76**.11 (2025), 1452–1469.

[Xu21] Xu, S. et al. "A topic models based framework for detecting and forecasting emerging technologies". *Technological Forecasting and Social Change* **162** (2021), p. 120366.

[Yin23] Yin, B. et al. “Sentence-BERT and k-means Based Clustering Technology for Scientific and Technical Literature”. *2023 15th International Conference on Computer Research and Development (ICCRD)*. 2023, pp. 15–20.

[You38] Young, G. & Householder, A. S. “Discussion of a set of points in terms of their mutual distances”. *Psychometrika* **3**.1 (1938), pp. 19–22.

[Zhu21] Zhuang, L. et al. “A Robustly Optimized BERT Pre-training Approach with Post-training”. eng. *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Ed. by Li, S. et al. Huhhot, China: Chinese Information Processing Society of China, 2021, pp. 1218–1227.

[Zhu25] Zhuang, Z. et al. “Large language models for automated scholarly paper review: A survey” (2025). arXiv:2501.10326.