# Uncovering Research Trends in Computer Science: Large-Scale Topic Modeling of arXiv Abstracts (2019–2025)

Group 12

Thevamanoharan Vibeesshanan, Shamku Shane

Faculty of Computer Science, Wilfrid Laurier University

CP421: Data Mining

Professor Yang Liu

April 6th 2025

**Table of Contents**

## 1. Introduction

In today's world where information is constantly becoming known, navigating through academic literature can be extremely complex and challenging. Topic modeling is a solution, especially for datasets as broad as arXiv repository, which acts as a host for millions of academic papers across a wide range of fields.

The primary goal of this project was to apply topic modeling techniques to the arXiv dataset to identify key topics within these papers such as themes and structures, track topic evolution over time, and to understand how federated learning and blockchain have evolved in academic research. We focused on the computer science and statistics categories, using abstracts from over 350,000 papers as the input. To do this, we used BERTopic, a modeling framework that combines transformer-based language models with class based TF-IDF.

This allowed us to capture the context behind each abstract while still making the topics easy to label and interpret. Besides this, we also looked at how topics change over time, the connections to real-world trends, and papers that did not fit appropriately into one topic, which gave us interesting edge cases to explore.

## 2. Methodology

### 2.1 Dataset Overview

Due to the scale of the arXiv dataset, we initially narrowed down our focus to papers in four relevant categories: Computer Science, Physics, Mathematics and Statistic, which ended up resulting in about 938,000 records. Although we had access to T4 GPU on Google Colab,

running BERTopic on the full dataset proved to be unattainable as the model training was projected to take over seven hours, and repeated RAM overloads led to session crashes. Upon further inspection we found that Computer Science papers alone accounted for over half a million entries, which still exceeded our computational capacity. To create a balance between coverage and attainability, we further filtered by publication date, retaining only papers between 2019 to 2025. This allowed us to have a final dataset of 351,393 abstracts which was both topically relevant and computationally manageable within our hardware limits.

### 2.2 Data Preprocessing

As part of the preprocessing pipeline, we cleaned the abstract column to prepare the text for the topic modeling. This included first converting all the text to lowercase, then removing punctuation, numbers and special characters, and then normalizing extra whitespace, and finally trimming any leading or trailing spaces. We also excluded abstract that contained less than 20 words as they were typically too short to convey any important information.

### 2.3 Chunking Strategy

Given the size of the filtered data (roughly 351,000 rows), we then implemented a chunking strategy to avoid memory overloads in Google Colab. The data was processed in batches of 100,000 at a time, this allowed us to clean and encode the abstracts without exceeding our RAM limits. Each processed chunk was saved temporarily and later concatenated into a single DataFrame which was then exported as a CSV file to be used in the modeling phase.

| | title | clean_abstract |
|---|---|---|
| 65444 | A Deep Neural Network for Finger Counting and ... | paper present neurorobotics models deep artifi... |
| 130915 | Adaptive Inference through Early-Exit Networks... | dnns becoming less less overparametrised due r... |
| 251013 | Kernel Neural Optimal Transport | study neural optimal transport algorithm uses ... |
| 160816 | Code Representation Learning with Pr\"ufer Seq... | effective efficient encoding source code compu... |
| 141482 | Compositional Abstraction Error and a Category... | interventional causal models describe several ... |

### 2.4 Embedding and Modeling

After cleaning the abstracts, we turned each one into a numerical format using the model all-MiniLM-L6-v2 from SentenceTransformers. This model was chosen as it is lightweight, runs quickly, and is strong in understanding the meaning of sentences. These factors made it the best choice for topic modeling. A T4 GPU on Google Colab was used and we increased the batch size to 256, using the available GPU memory. Once all the abstract embeddings were collected, we saved them in a .npy file. This allowed us to reuse them later without repeating the embedding process. Next, we trained a BERTopic model using the English language setting. BERTopic works by combining the transformer-based embeddings with techniques such as reducing the size of data to make it easier to work with, and grouping similar abstracts together using clustering. The model returned a predicted topic label and a probability score that shows how well each abstract fits within a given topic.

The topic modeling results were then stored in a new DataFrame by combining the predicted topic labels and their confidence scores with the original cleaned abstracts. This updated dataset became the base for all the next steps in the analysis, including tracking topic frequency, how the topics are related to each other, studying trends over time, and subtopic exploration. Once the BERTopic model was trained and each abstract had a topic assigned, we

saved the resulting DataFrame, which contained the cleaned abstracts, topic labels, and confidence scores, to a CSV file called topic_results.csv. This helped save the results so further analysis could be done and helped to create visualizations without needing to run the model again. Additionally, we saved the trained BERTopic model itself using its built-in .save() method, storing it as bertopic_model.bin. This allowed us to reload the model at any time to review topic keywords, generate new visualizations, or apply it to new data without having to go through the training process again, which can be time-consuming. To continue the analysis in later sessions, we reloaded the three key files from disk: the embeddings (embeddings.npy), the trained model, and the saved topic results. This allowed us to quickly resume our workflow involving generating plots, interpreting topic trends, and exploring subtopics without any extra computing time.

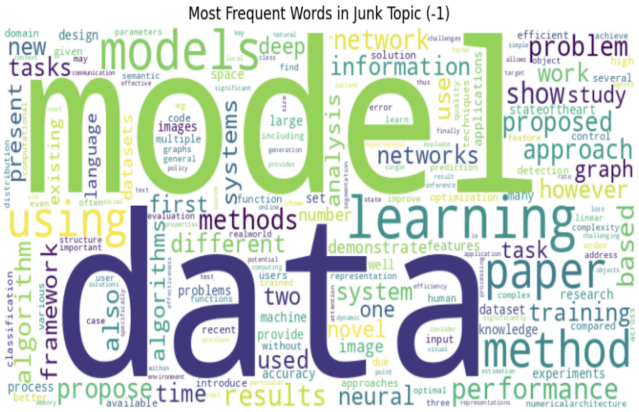
## 3. Results and Evaluation

### 3.1 Early Findings

To understand topic distribution we calculated the frequency of each topic and then visualized the top 15 using a bar chart, identifying the most prevalent research themes. However, one thing that stood out was: Topic -1: This “junk” topic accounted for over 160,000 documents, which was far more than any other cluster. Upon further inspection using a word cloud, we found that these abstracts were dominated by general academic terms such as “data”, “model”, “learning”, “methods”, and “paper”. This revealed a key insight: Topic -1 was not meaningless, but rather overly broad. The abstracts grouped here used highly generic and widely applicable language, which likely confused the topic modeling algorithm. Terms like “results,” “proposed,”

"algorithm," "task," and "system" appear across many subfields and lack the specificity needed to confidently assign them to a distinct semantic cluster. In essence, Topic -1 captures the linguistic overlap shared by papers from different domains, a limitation of even sophisticated models like BERTopic when faced with ambiguous or interdisciplinary text.

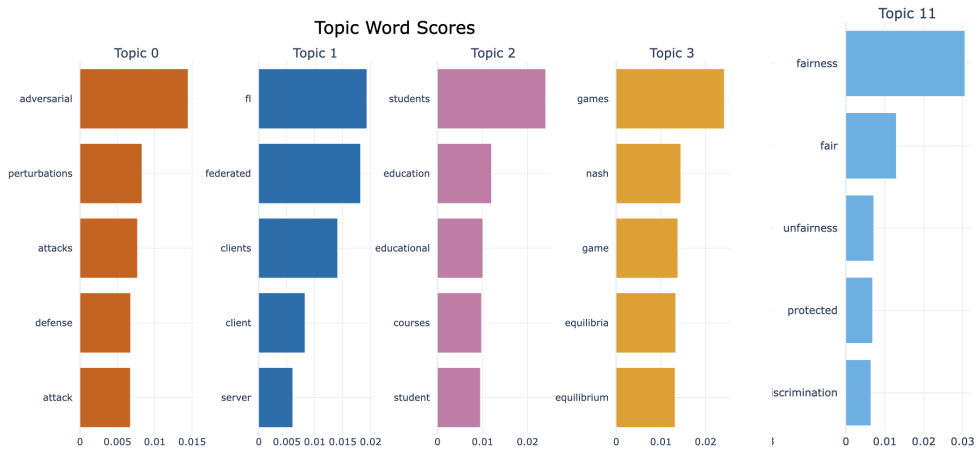
While we removed these documents from our main topic analysis to improve clarity and separation between distinct themes, we chose not to discard them entirely. In Section 3.8, we return to this cluster and apply further modeling in an attempt to uncover hidden subtopics within this group. A critical step for understanding the limits of transformer-based topic modeling and improving downstream interpretability.




### 3.2 Topic Representation

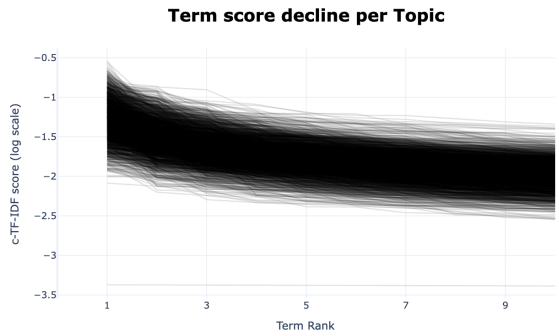
Having refined our dataset by removing the incoherent "junk" topic, we were left with a more semantically consistent set of clusters for deeper inspection. To further interpret the structure of each topic, we visualized the top-ranked terms based on their c-TF-IDF scores. Several topics, such as Topic 0 (adversarial attacks) and Topic 1 (federated learning), displayed highly concentrated vocabularies, with 4 - 5 dominant keywords accounting for the bulk of relevance.

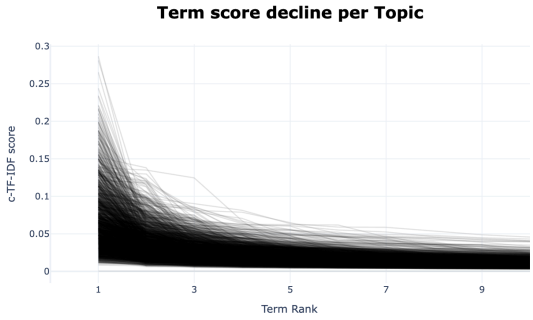
This pattern reflects semantically narrow domains with consistent terminology, allowing the model to form sharp, high-confidence clusters. In contrast, broader clusters like Topic 2 (education) or Topic 11 (fairness) featured more gradual declines in term relevance, indicating interdisciplinary themes that span across multiple application areas. Despite occasional conceptual overlap, the lack of term repetition across clusters highlights the model's ability to maintain linguistic separation and topic coherence even for thematically adjacent topics like causal inference and machine translation.




### 3.3 Term Score

To evaluate how concentrated each topic is around its most important terms, we visualized the decline in c-TF-IDF scores across the top 10 words per topic. As shown in the plots below, most topics follow a steep drop-off pattern from the first to third ranked terms, followed by a flatter slope. This "elbow" effect indicates that only a few keywords define most topics, reinforcing the tight clustering observed earlier. Such consistency across topics suggests strong semantic focus, where 3 - 5 core terms are typically sufficient to capture a topic's essence. This pattern further supports the interpretability and coherence of the topic model overall.



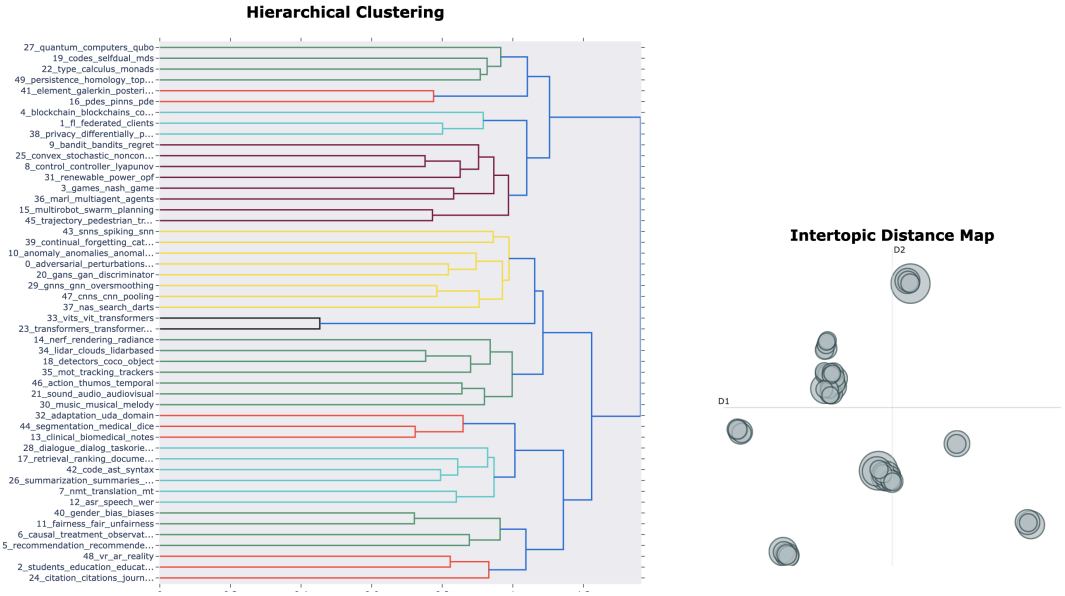

### 3.4 Topic Relationships and Clustering

To evaluate how semantically distinct or similar the generated topics are, we examined two complementary visualizations: the Intertopic Distance Map and a Hierarchical Clustering Dendrogram.

#### 3.4.1 Intertopic Distance Map &  Hierarchical Clustering

The Intertopic Distance Map shows how closely topics are positioned in a two-dimensional semantic space. Larger circles represent more frequent topics, while proximity between circles indicates thematic similarity. Most topics cluster tightly toward the center, suggesting that the text is largely cohesive, with moderate variance between thematic areas. However, some peripheral bubbles such as those related to sound, ethics, or niche domains hint at outlier or interdisciplinary research. To better interpret these relationships, we performed hierarchical clustering using the top 50 topics. This dendrogram uncovers clearer subtopic groupings. For instance, a dense cluster of computer vision models includes Topics 23 (Transformers), 33 (ViTs), and 47 (CNNs), reflecting shared vocabulary and application space. Adversarial learning techniques like GANs (Topic 20), Bandits (Topic 9), and Multi-agent systems (Topic 36) also cluster together, revealing a methodological alignment. More domain-specific branches such as

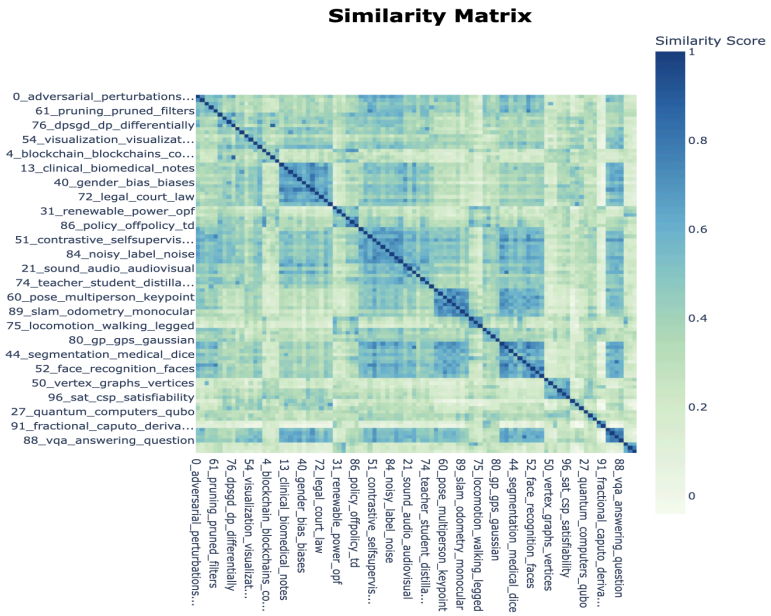
Topics 5 (Recommender systems), 2 (Education), and 24 (Citation networks) merge at higher semantic distances, signaling their distinct research agendas. Interestingly, Topics 11 (Fairness), 40 (Gender Bias), and 26 (Summarization) form a loose chain, suggesting a subtle interdisciplinary thread possibly around ethical or equitable information design in NLP. Finally, topics such as 27 (Quantum Computing), 22 (Type Theory), and 49 (Topological Data Analysis) appear at the fringes of both plots, affirming their role as semantic outliers with highly specialized vocabularies.




### 3.4.2 Topic Similarity Matrix

To complement the spatial and tree-based views of topic similarity, we also examined the topic similarity matrix. This matrix provides a pairwise overview of how semantically close or distant each topic is to the others, with darker cells indicating stronger connections. Medical AI clusters, Topics 13 (Clinical Notes), 44 (Medical Segmentation), and 84 (Noisy Labels) formed dense submatrices, reflecting their shared focus on healthcare applications. Topics tied to reinforcement

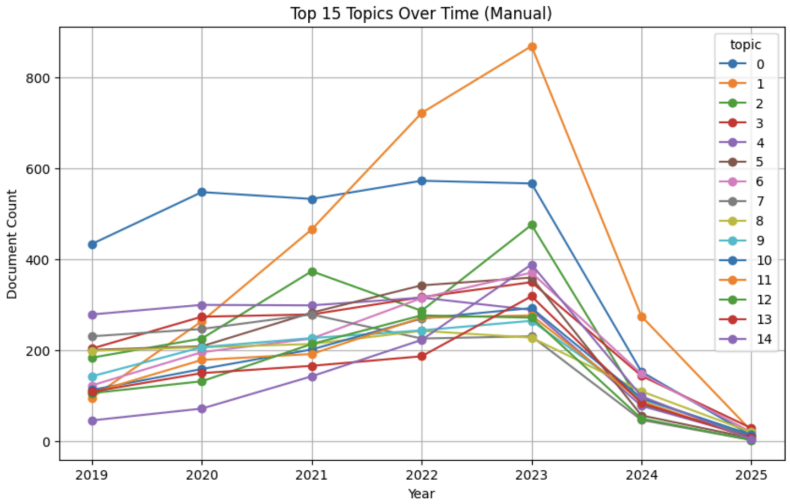
learning and visualization such as 0 (Adversarial Learning), 86 (Policy Learning), and 54 (Visualization), also showed tight inter-topic alignment. In contrast, outlier topics like 27 (Quantum Computing) and 91 (Fractional Calculus) appeared largely isolated, consistent with their niche vocabularies and earlier outlier status in other visualizations. Interestingly, Topics 40 (Gender Bias) and 72 (Legal/Court Law) revealed broad semantic overlap with a range of other topics, hinting at growing interdisciplinarity in AI, where fairness, ethics, and legal concerns are increasingly intertwined with mainstream ML research. These visualizations reinforce the semantic coherence of our clusters while also highlighting the nuanced diversity within the research text.




### 3.5 Topic Trend Over Time and Real-World Insights

The “Top 15 Topics Over Time” visualization shows meaningful shifts in academic focus from 2019 through 2025. Topic 0 (Adversarial Attacks and Defense) stood out early on, rising sharply until 2023 before tapering off. Its spike reflects growing interest in the robustness of machine learning models, especially for security-sensitive applications like autonomous driving and fraud detection. Even more prominent was Topic 1 (Federated Learning and Client Systems), which

surged consistently over the five-year span. Its growth aligns with increased concern around data privacy, driven by stricter regulations (e.g., GDPR) and public skepticism around centralized data collection. The trend also reflects the rise of on-device learning strategies in industries like healthcare and mobile AI.




Several other topics such as Topic 4 (Blockchain Consensus), Topic 5 (Recommendation Systems), and Topic 6 (Causal Inference) exhibited slower but steady growth. These topics represent well-established domains with enduring relevance across fintech, e-commerce, and scientific research. Most topics, however, experienced a noticeable decline in 2024–2025. This sharp drop is likely an artifact of incomplete or delayed indexing for more recent papers, rather than a genuine reduction in research activity.

### 3.6 Real-World Context and Cross-Topic Dynamics

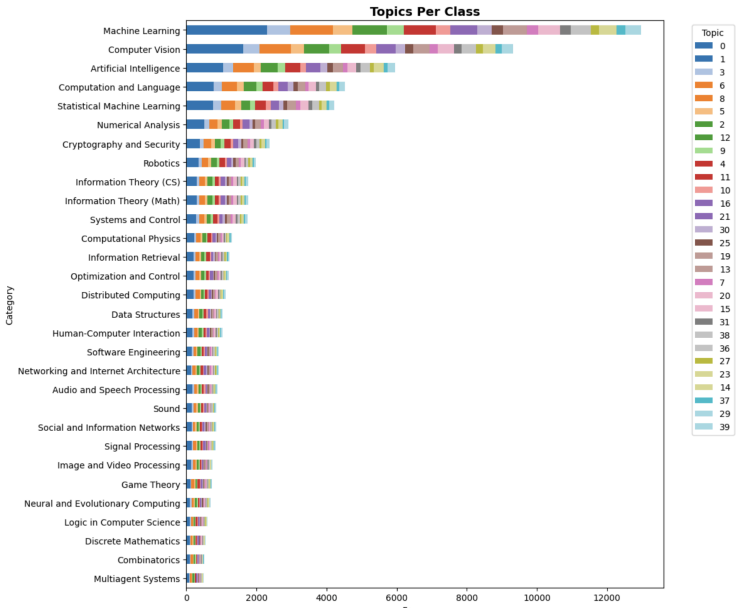
Among all the trends, Topic 1's rapid rise stood out not just statistically but contextually. When viewed alongside Topic 4 (Blockchain), we see both temporal and relative alignment. Between 2019 and 2023, blockchain technologies especially cryptocurrencies like Bitcoin and Ethereum

received massive global attention. Bitcoin hit an all-time high in April 2021, and NFTs, DeFi protocols, and smart contracts exploded into the mainstream. With this surge came real-world consequences: data breaches, ransomware attacks, and rising concerns about centralized systems. According to Chainalysis, over $14 billion in cryptocurrency was linked to illicit activity in 2021. These developments created academic demand for decentralized, privacy preserving solutions explaining the parallel growth in Topic 1. This relationship is further confirmed by Hierarchical Clustering, Topic 1 (Federated Learning) clusters closely with Topic 4 (Blockchain) and Topic 38 (Privacy), revealing shared semantic space around decentralization and trustless systems.


Together, these indicators illustrate a broader trend: academic research dynamically responds to global events, especially in applied AI domains. The intersection of federated learning, blockchain, and cybersecurity represents a fertile ground for interdisciplinary innovation, where shifts in the real world shape the trajectory of scholarly focus.

### 3.7 Topics Per Class

To understand how topic distributions vary by domain, we mapped the assigned BERTopic clusters to the original arXiv subject categories. After standardizing the category labels and exploding multi labeled entries, we grouped the data by category and topic to generate frequency counts. We then filtered the results to include only the top 30 categories and top 30 topics to ensure interpretability. The resulting stacked bar chart presents a high-level overview of how different research areas engage with key themes.

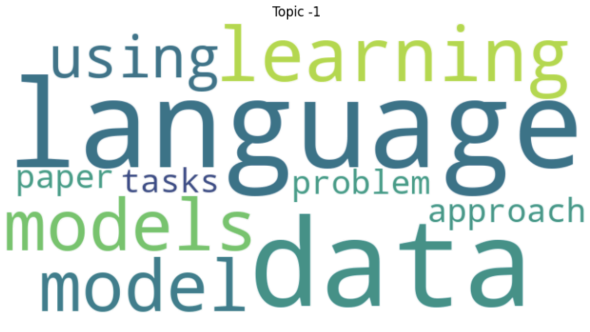


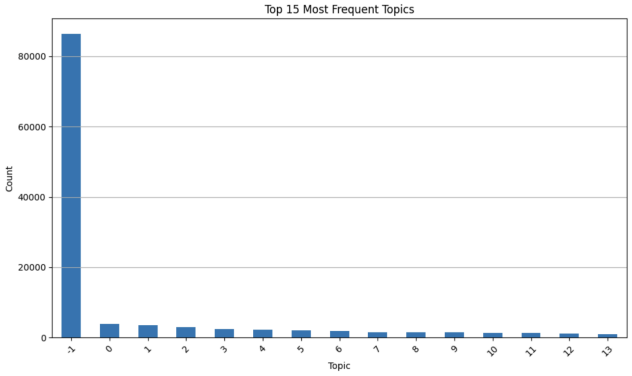

Topic 0, which captures broad machine learning terminology (e.g., *tasks*, *data*, *proposed*), is overwhelmingly present across nearly all categories. This suggests it represents introductory or framing language common in abstracts across disciplines. The Machine Learning category dominates in both volume and diversity. It spans highly technical areas like Topic 1 (adversarial attacks), Topic 6 (causal inference), and Topic 23 (transformers), underscoring the maturity and specialization within ML as a research field. Computer Vision and Artificial Intelligence exhibit similarly broad topical coverage. Vision-heavy categories lean toward Topic 18 (object detection) and Topic 33 (vision transformers), reflecting the field's shift from convolutional architectures to transformer-based pipelines. In contrast, Robotics, Speech Processing, and Numerical Analysis display tighter topical focus.

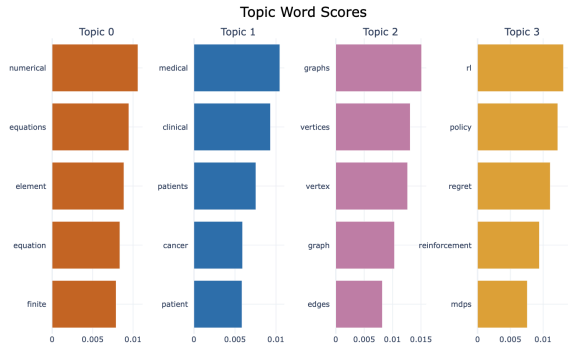
### 3.8 Junk Topic Analysis

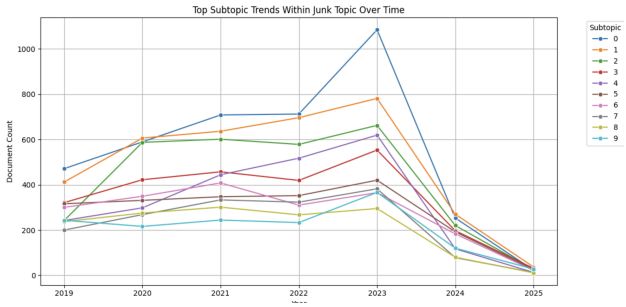
During early topic modeling, over 160,000 abstracts were assigned to Topic -1, a catch-all group BERTopic could not confidently classify. Initial word clouds showed high-frequency, generic

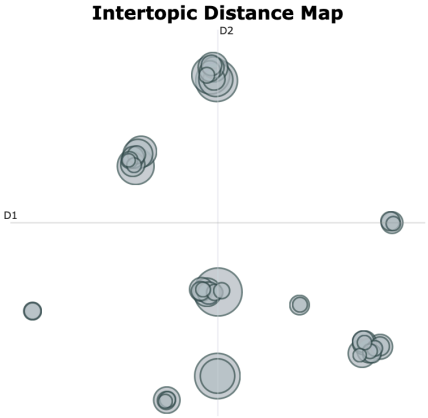
terms like "language," "learning," and "data," suggesting semantic ambiguity and lack of domain-specific signals. To explore potential hidden structure, we performed additional cleaning by removing stopwords and re-encoded the abstracts using the same SentenceTransformer. Re-running BERTopic uncovered several subtopics within the junk cluster, each with distinct vocabulary and thousands of associated documents, confirming that valuable content had previously gone undetected. However, a large number of documents (~80,000) still fell into a new -1 group, reaffirming that some texts remain too broad, interdisciplinary, or vague for reliable clustering. Term score decline plots for these subtopics showed flatter gradients compared to primary topics, implying a lack of tight lexical cohesion. Similarly, the intertopic distance map and hierarchical clustering revealed that even the recovered topics hovered at the semantic periphery of the text. Some subtopics (e.g., biomedical, graph theory, reinforcement learning, speech/audio) revealed clear themes but likely suffered from hybrid language or under-specified abstracts. This highlights a broader limitation of transformer-based models: without concise, distinctive terminology, even meaningful research can fall through the cracks.

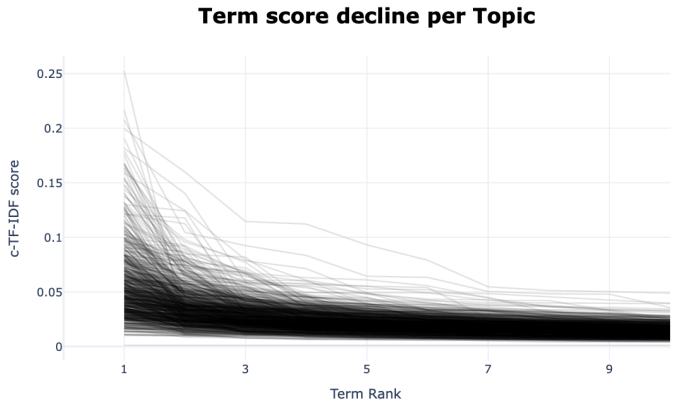





## 4. Conclusion

This project explored large-scale topic modeling on over 350,000 arXiv abstracts using BERTopic, focusing on recent research trends within Computer Science from 2019 to 2025. We implemented a robust pipeline involving data cleaning, transformer-based embedding, and unsupervised clustering, supported by comprehensive visual and semantic analyses. A key contribution was our investigation into the limitations of topic modeling at scale specifically, the handling of generic or ambiguous documents classified into a catch-all “junk” topic. Through stopword-free re-clustering and semantic exploration, we recovered meaningful subtopics previously obscured by lexical noise and embedding sparsity, demonstrating the value of

iterative refinement in NLP workflows. Our findings also offered real-world insights into research momentum across domains like federated learning, fairness, and blockchain, with clear connections to industry trends and societal needs. By analyzing topic dynamics over time and across academic categories, we provided a multidimensional view of the evolving research landscape. Future research directions could include, Integrating hierarchical topic models or hybrid approaches (semantic + citation graphs) to improve classification of interdisciplinary texts. Leveraging full-text data beyond abstracts for deeper context. Applying this pipeline to other domains (e.g., medicine, economics) to uncover field-specific innovation patterns.

Ultimately, our work highlights the potential of transformer-driven topic modeling to surface both dominant and fringe research narratives, provided careful attention is given to data preprocessing, model limitations, and post-modeling interpretability.

**5. Q&A Responses**

**Q:** How would you recommend these companies mitigate AI Bias moving forwards?

**A:** To reduce AI bias moving forward, companies should start by making sure their training data is diverse and representative of real-world populations. They also need to continuously test their models for fairness especially across different groups and retrain them when needed. Another big step is transparency: companies should be clear about how their models make decisions and involve ethicists or diverse teams when designing and reviewing AI systems. Lastly, they should stay updated on new regulations and best practices, so their models are both fair and responsible.

**Q:** What challenges did you face when defining and measuring bias?

**A:** One of the biggest challenges we faced was actually defining what bias looks like in this context. Bias can show up in a lot of subtle ways, and depending on what group you're looking at or even which metric you use it might show completely different results. So there's no one-size-fits-all definition. Another challenge was figuring out how to measure it in a meaningful way. With topic modeling, we weren't working with labeled data, so we had to look at things like the distribution of topics across categories and whether certain research areas were underrepresented. It's tough because some of the bias isn't super obvious it's more about what's missing than what's present.

**6. Task Division**

We split the work up accordingly: Shane completed the entire code, while Vibeesshanan did the entire presentation. Together Shane and Vibeesshanan did the report.

Shane - 100% Code + 50% Report

Vibeesshanan - 100% Presentation + 50% Report