Capstone3 Final Report

Preliminary work on better understanding the scientific literature

Deanna M. Church

2025-04-18

Executive summary

This project was an exploratory approach to understand what is required to programmatically assess the scientific literature. There are over 1 million papers published each year (https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html). To address this, we used an annotated dataset describing over 136,000 papers in arXiv (https://www.kaggle.com/datasets/sumitm004/arxiv-scientific-research-papers-dataset/data). I used three different approaches to explore this dataset.

- Paper categorization by tuning a more general language model
- Trend prediction using time-series analysis
- Identifying high leverage individuals using network analysis

While the results obtained are not useful for production work, the results point to additional approaches that could be taken. The most successful analysis was the trend prediction analysis. Precise forecasting was not achieved, but identification of overall trends in the data was possible. Results from the paper categorization and identifying high leverage individual analysis were not successful, but deeper analysis identified issues that could be addressed to improve these analyses.

Recommendations for next steps

I recommend continuing this work. There are several approaches that could be implemented to improve these analysis.

Data improvements:

- Use of full text rather than just abstracts.

- Use of authors complete names rather than initials.
  - ORCID ids would be even better.
- Increasing the size of the dataset.
  - While 136,000 papers seems like a large set, it is a small percentage of the entire corpus, even if restricting to a subtopic.
- Test additional models
  - PaperQA: https://github.com/Future-House/paper-qa
  - BixBench (for evaluation): https://github.com/Future-House/BixBench

Technical improvements:

- Migrate this analysis to a cloud environment with access to GPUs.
  - The paper categorization model training took a week running on a MacBook M1 64Gb laptop.

Methods

Below are some more details on the various analyses performed.

Paper classification

Results

Despite attempts to more precisely tune the base model, performance is only minimally improved.

Table 1: Table 1. Overall model performance

	accuracy	average confidence	Macro F1	Weighted F1
Base SciBert	0.755	0.877	0.227	0.746
Tuned SciBert	0.755	0.877	0.227	0.746
Hypertuned SciBert	0.77	0.834	0.181	0.757

Table 2: Table 2. Model performance by bin size

	base model correct	tuned model correct	hypertuned correct
Very small (1-10)	0.0488	0.0488	0.0
Small (11-100)	0.182	0.182	0.102
Medium (101-1000)	0.414	0.414	0.412
Large (1001-1000)	0.512	0.512	0.541
Very large (10001 +)	0.836	0.836	0.854

Only the two largest bins saw any improvement with tuning, suggesting that more data is needed to improve performance. I suspect it may be more important to add the full text rather than just adding more examples with abstracts and titles only. It is quite likely that data imbalance also influenced the results of this analysis.

Details

The data had been analyzed and cleaned in a previous notebook. However, category codes containing less than 5 rows were removed as this was a requirement for the model to execute. The data was split into train, test sets and tokenized. The model was then tuned and hyperparameter tuning was performed.

Trend analysis

Results

The goal of the trend analysis was to determine if we can understand which categories of papers are growing (suggesting there is more interest in this top), or decreasing (so less interest in a topic). It is clear that the number of total papers is growing year over year.

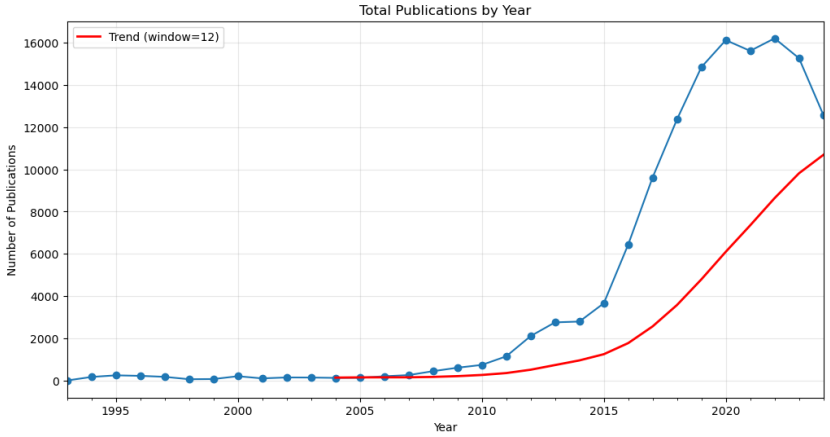

Figure 1: Plot of papers submitted to ArXiv by year

However, there is more variability between individual categories.

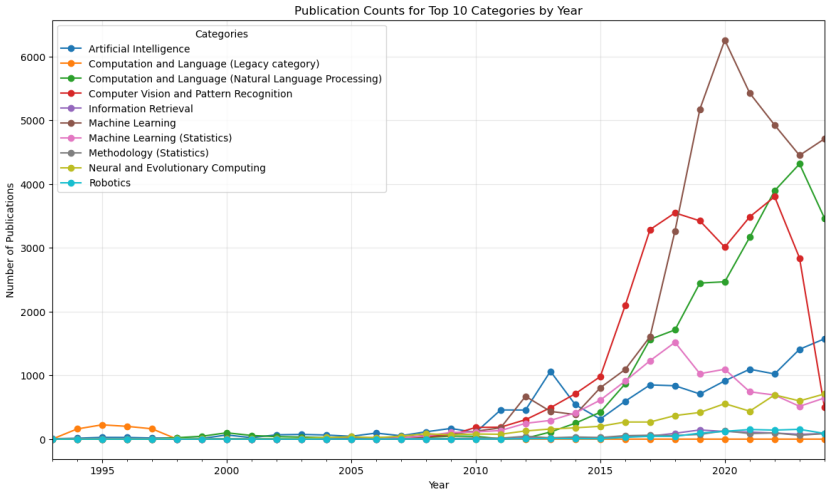

Figure 2: Plot of papers submitted to ArXiv by year and category

An attempt was made to forecast the number of paper in a given area, but this was unsuccessful. I suspect one factor contributing to this is that each category had different characteristics. While all were non-stationary to begin with, different approaches had to be used to make the category datasets stationary. For example, for 'Machine Learning' the best method was log_difference but for 'Artificial Intelligence' it was twice_difference.

However, looking at overall trends in the data (which categories are strongly increasing vs. relatively stable) could be gleaned from this data.

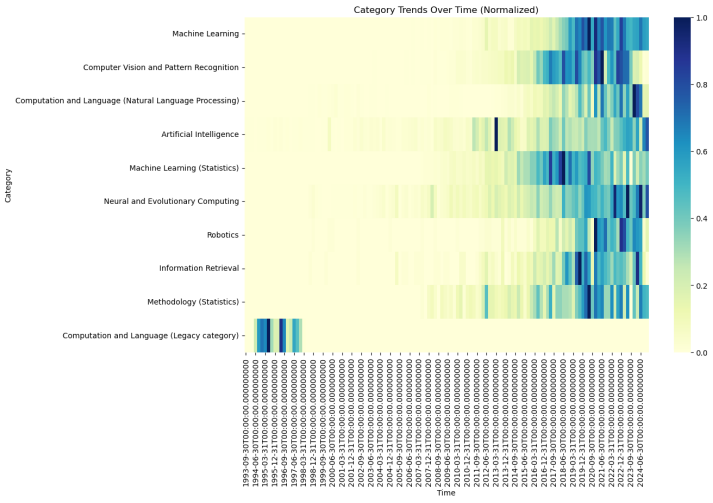

Figure 3: Heatmap of category trends

Details

Overall, little cleaning had to be done to this data. The biggest adjustment was removing all data from 2025 as this year contained only 1 month of data. This one month of data had a large impact on interpretation. When testing the entire data set for stationarity, with the 2025 data retained that dataset looked stationary, which was unexpected. However, removing this bit of data showed the data was not stationary.

Network analysis

Results

The last analysis was to identify highly influential authors in this list. To achieve this, the author list was used to build a network, and centrality measures were calculated three ways: Degree, Eigenvector and Pagerank. Betweeness was not used due to the computational load of that analysis. All three analysis gave similar results.

Table 3: Table 3. Top 10 authors by different centrality measures.

Degree	Eigenvector	Pagerank
Y. Wang	Y. Wang	Y. Wang
Y. Li	Y. Zhang	Y. Zhang
Y. Zhang	Y. Li	Y. Li

Degree	Eigenvector	Pagerank
Y. Liu	Y. Liu	Y. Liu
Z. Wang	Z. Wang	Z. Wang
Y. Chen	X. Wang	Y. Chen
X. Wang	J. Wang	X. Wang
J. Wang	J. Li	J. Wang
J. Li	Y. Chen	J. Li
J. Zhang	J. Zhang	X. Li

The networks themselves were too dense to make sense of, even when using only last authors to build the network. Improving name handling, ideally using ORCID ids but at a minimum using full names, and doing more focused network analysis, perhaps by category, may improve clarity.

Details

Author names were not regularized in the original data. Some came in as initials + last name and others were full name. To regularize the data, author names were cleaned and returned as 'Initials. LastName'. While this was prudent, as many of the authors were already in the list this way. However, it potentially caused different people to be lumped together.

Additionally, without additional information, like citation counts, it is unclear if these authors are truly influential, or generating lots of papers via papermills.