Multilabel Text Classification of Research Articles

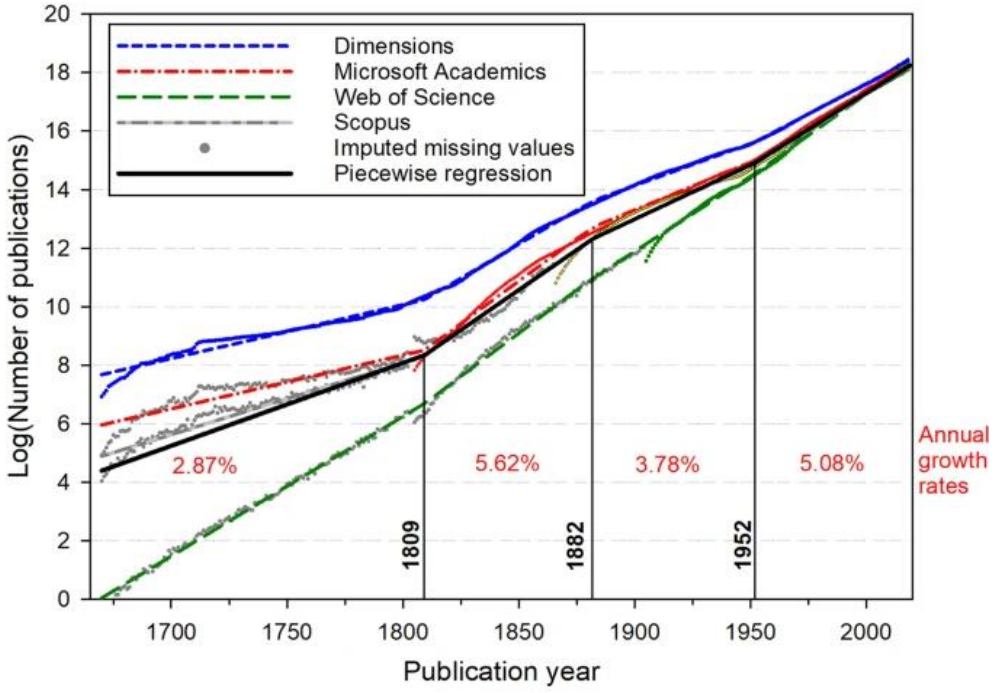
Rachel Chiang

August 2023

Scholarly Publishing Overview

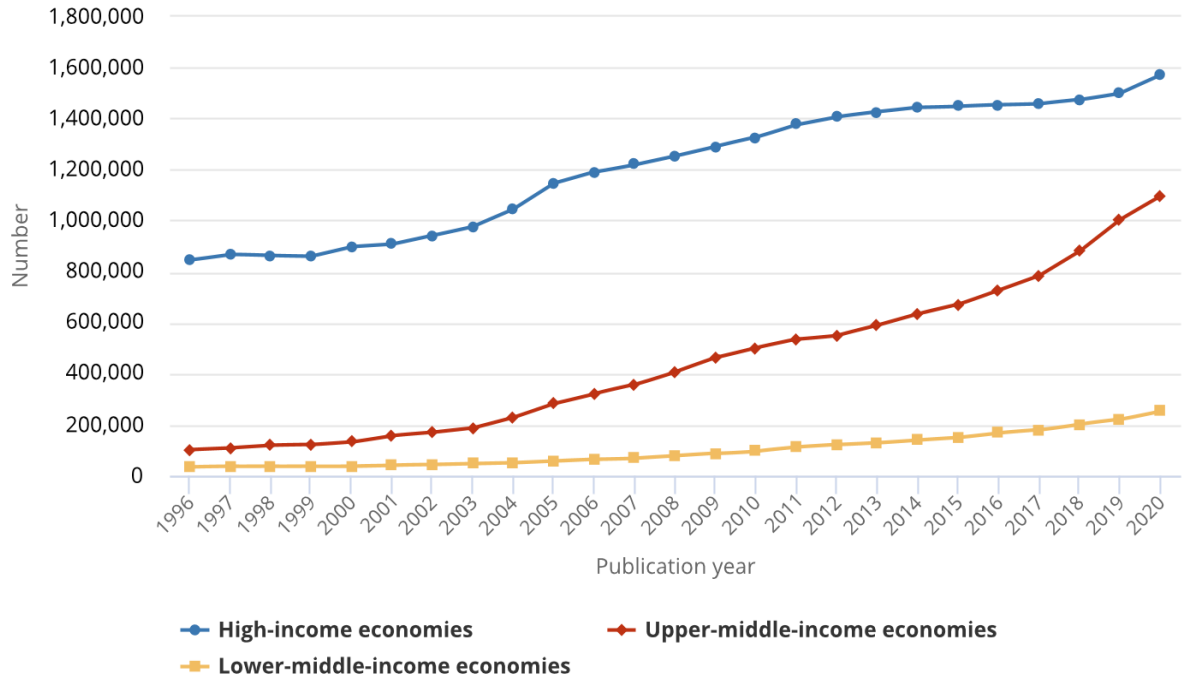
- Number of research articles increases
  - Growth rate of 3-4% [1]
  - Knowledge doubles every 15-17 years [1, 2]
- Voice of the researchers
  - Desire for accessibility and transparency [3]
  - Trends: Open Access, crowdsourcing, and open sources [4, 5]
- Internet and technology
  - Changes in publishing models [5]
  - Availability, affordability, and speed

Scientific article growth (log publications vs year)



Plot for segmented unrestricted growth of scientific articles from four bibliographic databases [1]

Science and engineering articles published (count vs year)



Science and engineering articles published per year by income group in 1996-2020 [6]

Project Objective

Problems

- Sheer volume and increased digitization of articles
- Researchers desire availability and transparent, sophisticated search engines [3]

Solution

- (Or at least one part of a solution)
- Automated categorization with at least 70-80% accuracy using titles and abstracts

Who may be interested?

Who can use the tool?

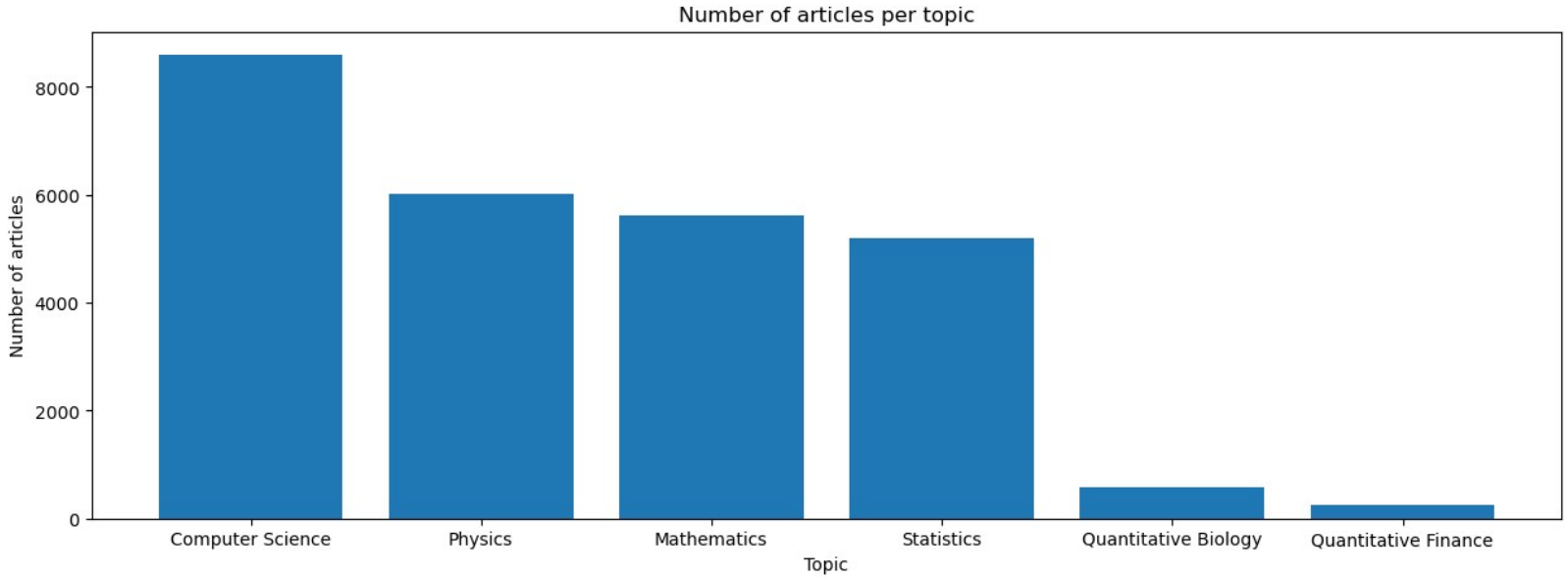
Digital journals, archives, libraries, search engines

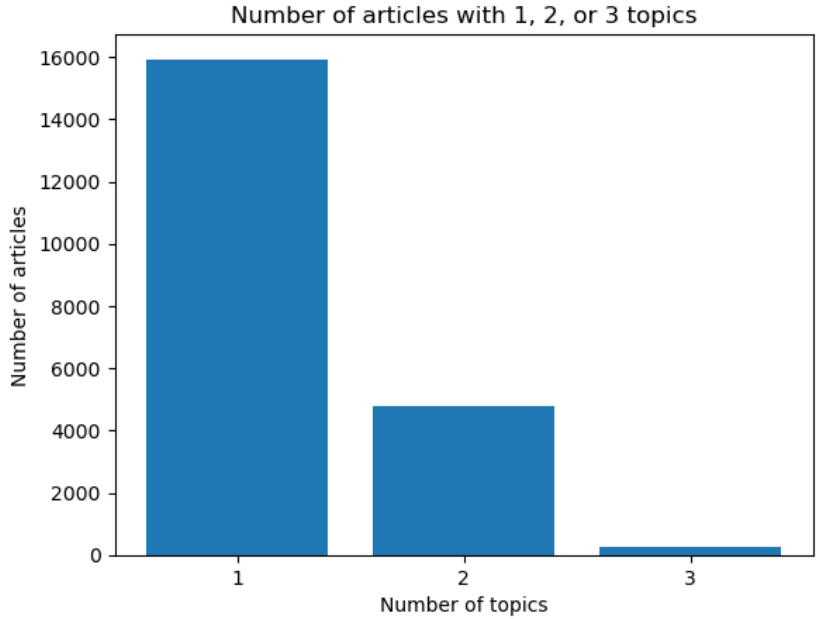
Who does this benefit?

Researchers, students, and librarians

The Data

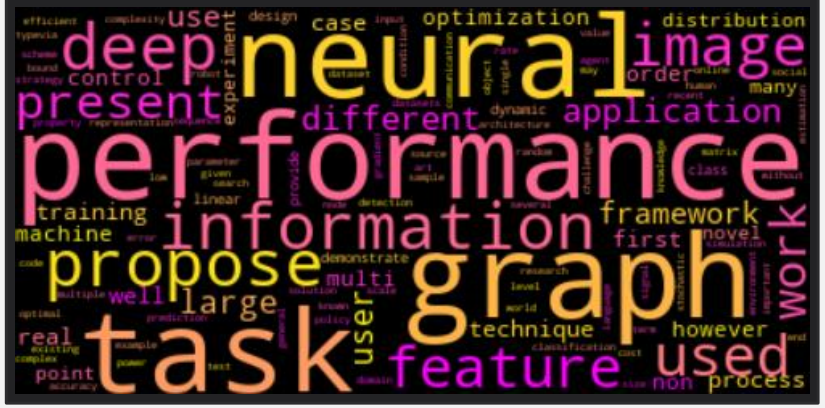
- 20,972 articles labeled with one to three of the categories:
  - Computer Science
  - Physics
  - Mathematics
  - Statistics
  - Quantitative Biology
  - Quantitative Finance



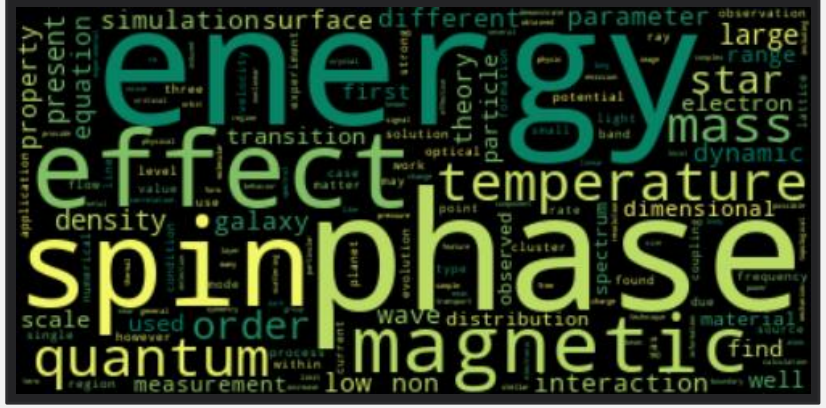


Filtered Word Clouds

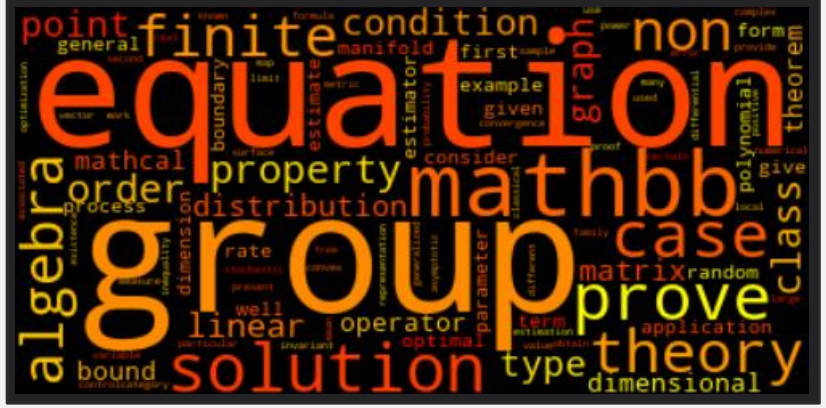
Computer Science



Physics

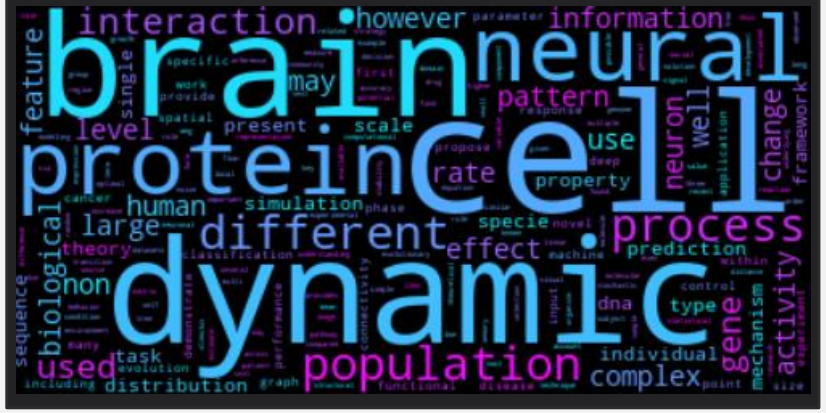


Mathematics

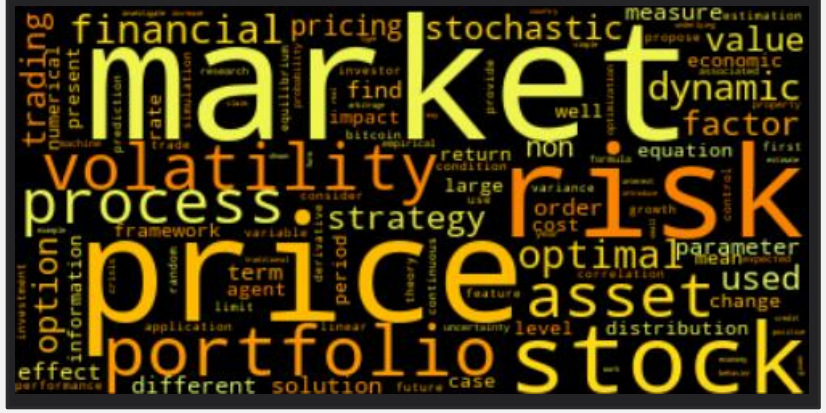


Statistics

Quantitative Biology



Quantitative Finance



Final Model: Unigram Bag-of-Words Model

Problem:

- Multilabel, four categories

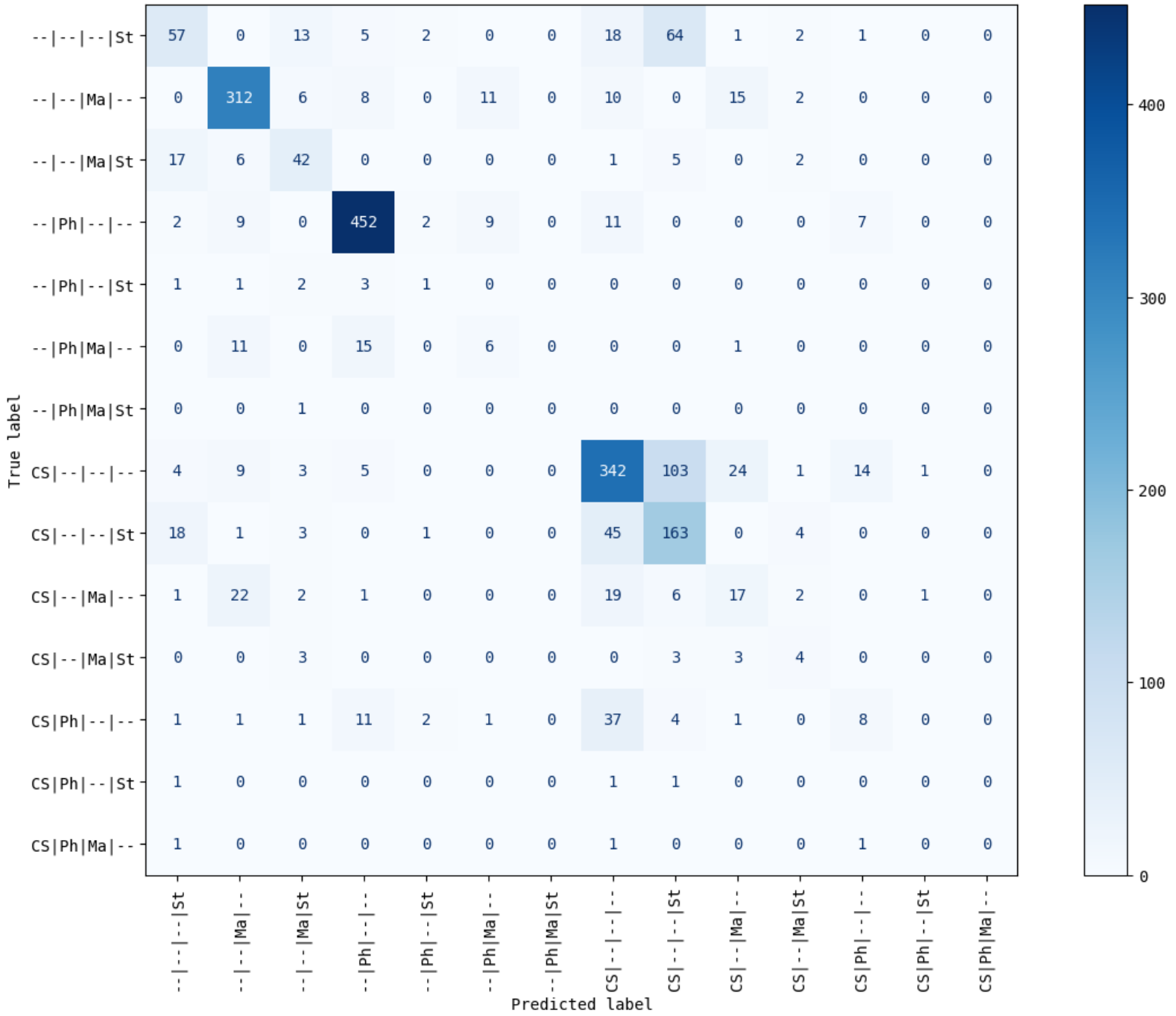
spaCY English pipeline with regex:

- Convert to lowercase
- Tokenize and lemmatize
- Remove stop words, numbers, punctuation, special characters, and empty tokens

spaCY Multilabel Text Categorizer:

- Unigrams only
- Scorer threshold of 0.45 (to 0.50)
- Always returns one category

	Precision	**Recall**	**F1-Score**
Computer Science	0.85	0.90	0.87
Physics	0.91	0.85	0.88
Mathematics	0.84	0.84	0.84
Statistics	0.74	0.83	0.78
Micro Avg	0.84	0.86	0.85
Macro Avg	0.84	0.85	0.84
Weighted Avg	0.84	0.86	0.85
Samples Avg	0.88	0.89	0.86



Confusion Matrix for 14 Label Combinations

Top 10 Tokens from 100 Samples by Mean SHAP Values

Computer Science	Physics	Mathematics	Statistics
• Reachability • Bit • Inability • Improvement • Communication • Robot • Attempt • Segmentation • Principled • Validate	• Mechanic • Calculation • Sky • Symmetry • Detector • Molecular • Galaxy • Hydrodynamic • Spacecraft • Removal	• Prove • Operator • Sharp • Article • Homotopy • Theorem • Perturb • Mathematical • Category • Conic	• Bayesian • Sequential • Statistical • Recommender • Clustering • Approximate • Trial • Parametric • Learning • Explanation
-2.47 and +[0.05, 0.06]	-3.52 and +[0.04, 0.07]	-9.97 and +[0.05, 0.06]	-10.24 and +[0.05, 0.1]

How the model can be used

Automatically tag new articles with the four categories

- Overall good precision/recall for the major categories, especially when articles have one label
- May tag too many articles with Computer Science and too few articles with Statistics
- General multilabel fuzziness

Future Work: Addressing the two problems

Problem: Category imbalance

- Resample datasets: downsample Computer Science, upsample Statistics
- Use different categories, thus becoming both an unsupervised and supervised problem

Problem: Scoring

- Adjust prediction interpretation to be more sophisticated, such as using present score distribution
- Convert to a recommendation system problem

Thank you.

References

[1] Bornmann, L., Haunschild, R., & Mutz, R., "Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases". *Humanit Soc Sci Commun* 8, 224 (2021). https://doi.org/10.1057/s41599-021-00903-w .

[2] Amnet, "Scholarly Publishing: Challenges, Opportunities, and Trends". *Amnet* (2022). https://amnet-systems.com/scholarly-publishing-challenges-opportunities-and-trends/ .

[3] Sandusky, R., Tenopir, C., & Casado, M., "Figure and table retrieval from scholarly journal articles: User needs for teaching and research". *Asis&t* 44: 1-33 (2008). https://doi.org/10.1002/meet.1450440390 .

[4] ScienceEditor, "The Future of Academic Publishing: Emerging Trends You Should Know". *ServiceScape* (2021). https://www.servicescape.com/blog/the-future-of-academic-publishing-emerging-trends-you-should-know .

[5] Singh, S., "2022 in review: Key developments shaping scholarly publishing". *Editage insights* (2023). https://www.editage.com/insights/2022-in-review-key-developments-shaping-scholarly-publishing .

[6] White, K., "Publication Output by Country, Region, or Economy and Scientific Field". *Science and Engineering Indicators,* National Science Board, National Science Foundation (2021). https://ncses.nsf.gov/pubs/nsb20214/publication-output-by-country-region-or-economy-and-scientific-field .