ACA-MEDIA: A new dataset of media coverage of academics*

Ioanna Lachana† Fabrizio Core‡ Moqi Groen-Xu§

September 25, 2024

Media coverage is instrumental for the dissemination of academic research and its impact on decision making, e.g., (Peng et al., 2023), Eagleman (2013). As such, academic literature has examined the accuracy of coverage of medical news (Sumner et al., 2014) and documented various gaps between research and its media coverage, e.g., Peters (2013). Recently, the quality media coverage of academic research has been subject to discussions in the context of COVID (Mach et al., 2021). However, medical and indeed research in the hard sciences only constitutes a small fraction of media activity. Instead, most news are on politics, economy, business, and finance (Cagé et al., 2020). These news fall more into the domain of social sciences, which are much more based on debates rather than facts. This makes it even more important to document the nature of the media coverage on business, finance and economics researchers, and any of its biases.

Most current studies on the dissemination of academic research uses Altmetric, e.g., Chapman et al. (2022), Brigham (2014), Peng et al. (2024), Vasarhelyi et al. (2021). Altmetric is available for a wide cross-section of academic research, and links media coverage to publications. Our dataset focuses on the coverage of individuals which could include portraits[1], summary of research [2] and/or commentary on current affairs like rising inflation [3].

To do so, our data pipeline identifies the name and the gender of business, finance and economics researchers who are mentioned in media articles and, if applicable, the research articles cited in the same media article. We first identify newspapers and magazines articles from a pool of 2.5 million media items, provided by ProQuest for the 2010-2023 period, that mention PhD granting U.S. universities. Our sample includes major outlets for economic and financial news such as the Wall Street Journal, the New York Times, and the Economist, along with a large number of magazines. Then, within the university-filtered dataset we search for the names of researchers based at these universities as listed by the Academic Analytics Research Center (AARC). We then identify coverage instances that refer to specific research, and link them to any covered research studies (co)authored by the mentioned researchers. The resulting dataset contains 20,518 media articles with references

*The authors acknowledge generous support by AFFECT (Academic Female Finance Committee) and the Journal of Financial Economics that made this research project possible. We also thank Dr David Schroeder for his valuable feedback and guidance. We are grateful for the comments from the 8th MWZ-CEPR Text-as-Data Workshop and at Birkbeck University. The data will become available publicly upon publication of this paper.

†PhD student, Birkbeck College University of London, ilacha01@mail.bbk.ac.uk.

‡Erasmus University Rotterdam, core@ese.eur.nl.

§Queen Mary University of London, moqi.xu@qmul.ac.uk.

[1]"Nobel Prize in Economics Awarded to American Richard Thaler", Wall Street Journal, October 10, 2017

[2]"An Algorithm, an ETF and an Academic Study Walk Into a Bar", Wall Street Journal, May 11, 2017

[3]"How Has Inflation Hit Your Finances?", Wall Street Journal, May 22, 2024

to 4,164 researchers and 11,165 research papers in business, finance, and economics from 211 universities.

To assess the performance of the pipeline we draw 1,000 random articles from our dataset and ask human readers to identify them as research related or not and then compare the readers' output against the pipeline output. Within the 1,000 random articles, the readers identified 74 AARC researchers while our pipeline identified correctly only 69 of them, that results in 93% accuracy.

To our knowledge, our dataset ACA-MEDIA is currently the largest of its kind, both in terms of time-space coverage and scope of applications. The most comparable dataset is LYDIA Bautin et al. (2010), which provides media coverage of person-names between 2004 to 2009 but does not link to research specifically. Our dataset is open access and open source tools to help the community help on / extend our work. To make the database inter-operational, we linked each researcher and media information to standard identifiers such as the openAlex person ID and the DOI for the matched research papers.

Our project aims to support the growing and recent literature on the dissemination of academic research in business, economics and finance. Analysis of newspaper coverage include Ziegler (2021) who investigates the coverage of 10,902 NBER Working Papers between 2010-2019 period. Brogaard et al. (2018), explores how the number of publications changes for economists before and after they are awarded tenure, using an initial dataset of 2,763 faculty who were employed at any of the top 50 economics or finance departments in the United States from 1996 through 2014. Sarsons et al. (2021) study how co-authorship matters differently for tenure for male and female economists in academia, using a dataset they manually built based on the CVs of economists who came up for tenure between 1985 and 2014 at one of the top 35 US PhD-granting universities. Sherman and Tookes (2022) explores the gender gap in the academic finance profession by manually creating a dataset of finance faculty at top-100 U.S. business schools during 2009 to 2017. Ringgenberg et al. (2023) use citations by think tanks to finance papers to investigate dissemination. Compared to these papers, our data provides the deepest coverage of researchers in business, finance and economics with 211 institutions and 4,164 researchers.

More general literature on the dissemination of research uses Altmetrics data. This literature usually focuses on the hard sciences or does not distinguish social from natural sciences. For example, Peng et al. (2023) links tweets to scholars of all domains and document a gender-gap in self-promotion. Vasarhelyi et al. (2021) document a gender gap in Altmetrics coverage counts and Peng et al. (2024) an ethnicity gap.

Within economics, our data contributes more generally to the literature on media coverage. Gentkow and Shapiro (2010) propose a new index of ideological slant in news coverage and compute it for 429 U.S. daily newspapers, pulling data from ProQuest and NewsLibrary. Widmer et al. (2023) similar to Gentkow and Shapiro (2010) pull data from NewsLibrary and look into the causal effect of media slant between cable networks and local newspapers. Djourelova et al. (2024) uses NewsBank, a database containing the text and metadata of over 100 million articles published in over 800 newspapers in their sample, beginning in 1999 to investigate the effects of the launch of Craiglist on the content and the organisation of newspapers. Beattie et al. (2021) shows that newspapers report less about car safety recalls if the car firm buys advertising space, using Newslibrary.com database, collecting approximately 13,600 related articles from 115 daily newspapers for the period 2000-2014. Dell et al. (n.d.) create a new dataset by extracting text from nearly 20 million scans coming from the Library of Congress. The dataset can be used for topic modelling or training language models among other uses.

In business and finance, the most frequently used media datasets are Factiva and

LexisNexis. Chahrour et al. (2021) shows that media can be an independent source of business cycle fluctuations, using Factiva (2019) to source media content from six major US outlets (Wall Street Journal, the New York Times, USA Today, the Boston Globe, the Charleston Gazette, and the Atlanta Journal Constitution) from 1988 to 2018. Nimark and Pitschner (2019) look into the news coverage and their economic implications. They focus on two 3-month periods (August-October 2001 and August-October 2008) and use a dataset of approximately 15,000 articles sourcing Factiva. Other finance papers focus on the Wall Street Journal (WSJ), one of the most important newspapers in business, finance and economics, which allows access to its data via API (Bybee et al., 2021). Dixon-Fowler et al. (2013) use data from LexisNexis, for the 1991-2006 period looking to investigate how media reports on female CEOs might be different compared to the male CEOs' and the effect this could have on the stock price of other companies with female CEOs. The data coming from LexisNexis cover newspapers and magazines like Wall Street Journal and New York Times. Our data contains the WSJ as well as major newspapers covered by Factiva, NewsBank, and Newslibrary.com. Our data source, ProQuest, benefits from the possibility to perform text analysis on their server, whereas the other sources restrict large-scale access to texts.

Our dataset can be used for looking into the media coverage of economics or finance related academic research and topics like female representation in academia or any other topic that will require information about the gender of the researcher, the newspaper or magazine that cited this research and the actual research study cited on the media article.

References

Bautin, Mikhail, Charles Ward, Akshay Patil and Steven Skiena (2010), 'Access: News and blog analysis for the social sciences', *WWW '10: Proceedings of the 19th international conference on World wide web* pp. 1229–1232.

Beattie, Graham, Ruben Durante, Brian Knight and Ananya Sen (2021), 'Advertising spending and media bias: Evidence from newscoverage of car safety recalls', *Management Science* **67**(2), 661–1328.

Brigham, Tara J. (2014), 'An introduction to altmetrics', *Medical Reference Services Quarterly* **33**(4), 438–447.

Brogaard, Joathan, Joseph Engelberg and Edward Van Wesep (2018), 'Do economists swing for the fences after tenure?', *Journal of Economic Perspectives* **32**(1), 179–94.

Bybee, Leland, Bryan Kelly, Asaf Manela and Dacheng Xiu (2021), 'Business news and business cycles', *NBER Working Paper* .

Cagé, Julia, Nicolas Hervé and Marie-Luce Viaud (2020), 'The production of information in an online world', *Review of Economic Studies* **87**(5), 2126–2164.

Chahrour, Ryan, Kristoffer Nimark and Stefan Pitschner (2021), 'Sectoral media focus and aggregate fluctuations', *American Economic Review* **111**(12), 3872–3922.

Chapman, Colin A., Claire A. Hemingway, Diptof Sarkar, Jan F. Gogarten and Nils Chr. Stenseth (2022), 'Altmetric scores in conservation science have gender and regional biases', *Conservation and Society* **20**(3), 195–200.

Dell, Melissa, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin and Leander Heldring (n.d.), American stories: A large-scale structured text dataset of historical u.s. newspapers.

Dixon-Fowler, Heather R., Alan E. Ellstrand and Jonathan L. Johnson (2013), 'Strength in numbers or guilt by association? intragroup effects of female chief executive announcements', *Strategic Management Journal* **34**(12), 1488–1501.

Djourelova, Milena, Ruben Durante and Gregory J. Martin (2024), 'The impact of online competition on newspapers: evidence from the introduction of craigslist', *Review of Economic Studies* **forthcoming**.

Eagleman, David M. (2013), 'Why public dissemination of science matters: A manifesto', *The Journal of Neuroscience* **33**(30), 12147–12149.

Gentkow, Matthew and Jesse Shapiro (2010), 'What drives media slant? evidence from u.s. daily newspapers.', *Econometrica* **78**(1), 35–71.

Mach, Katharine J., Raúl Salas Reyes, Brian Pentz, Jennifer Taylor, Clarissa A. Costa, Sandip G. Cruz, Kerronia E. Thomas, James C. Arnott, Rosalind Donald, Kripa Jagannathan, Christine J. Kirchhoff, Laura C. Rosella and Nicole Klenk (2021), 'News media coverage of covid-19 public health and policy information', *Humanities and Social Sciences Communications* **8**.

Nimark, Kristoffer P. and Stefan Pitschner (2019), 'News media and delegated information choice', *Journal of Economic Theory* **181**, 160–196.

Peng, Hao, Misha Teplitskiy, Daniel M. Romero and Emöke-Ágnes Horvát (2023), 'The gender gap in scholarly self-promotion on social media', *arXiv* .
URL: *https://arxiv.org/abs/2206.05330*

Peng, Hao, Misha Teplitskiy and David Jurgens (2024), 'Author mentions in science news reveal widespread disparities across name-inferred ethnicities', *Quantitative Science Studies* **5**, 351–365.

Peters, Hans Peter (2013), 'Gap between science and media revisited: Scientists as public communicators', *PNAS* **110**, 14102–14109.

Ringgenberg, Matthew C., Chong Shu and Ingrid M. Werner (2023), 'The politics of academic research', *SSRN* .

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben and Arthur Schram (2021), 'Gender differences in recognition for group work', *Journal of Political Economy* **129**(1), 101–147.

Sherman, Mila Getmansky and Heather E. Tookes (2022), 'Female representation in the academic finance profession', *The Journal of Finance* **77**(1), 317–365.

Sumner, Petroc, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A. Venetis, Aimëe Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy and Christopher D. Chambers (2014), 'The association between exaggeration in health related science news and academic press releases: retrospective observational study', *BMJ* .

Vasarhelyi, Orsolya, Igor Zakhlebin, Stasa Milojevic and Emöke-´Ágnes Hórvat (2021), 'Gender inequities in the online dissemination of scholars' work', *PNAS* **118**(39).

Widmer, Philine, Sergio Galletta and Elliott Ash (2023), 'Media slant is contagious', *SSRN* .

Ziegler, Lennart (2021), What is the media impact of research in economics?
URL: *https://papersecon.univie.ac.at/RePEc/vie/viennp/vie2103.pdf*