A CASE STUDY

Using Machine Learning and Newspaper Text to Understand Journalists and Media Bias

An economics Ph.D. researcher uses TDM Studio to develop a machine learning model that analyzes media bias.

about.proquest.com

To talk to the sales department, contact us at **1-800-779-0137** or **sales@proquest.com**.

ProQuest®
Part of **Clarivate**

Levi Boxell earned a PhD in Economics from a top-ranking research university. His academic research on political polarization and media has been published in multiple peer-reviewed academic journals and cited in major national and international news outlets.

Introduction

Dr. Levi Boxell conducted his research with Jacob Conway using text and data mining methodologies while gaining his doctorate degree in Economics from a university consistently in the top 5 in world university rankings. We sat down with Boxell to discuss his research, process, and findings and below is a synopsis of his work. This case-study provides a summary of the research project that was joint work with Jacob Conway. Read the full paper with funding acknowledgments and additional references **here**.

As referenced in Boxell's paper, in the USA, 28 percent of journalists identify as a Democrat and 7 percent identify as a Republican (Weaver et al. 2019). This disproportionate ideological representation of journalists has been cited as a potential driver for an alleged overarching liberal bias in the media industry (Groseclose 2011).

"Anecdotal discussions about media bias and the role of journalists are common, particularly given the ideological distribution of journalists in the US. However, we wanted to examine these claims empirically. To do that, we fine-tuned a RoBERTa model (a transformer-based model of human language) using articles tweeted by politicians as our measure of slant and then applied the model to a large corpus of newspaper articles from TDM Studio, thus enabling us to track changes in article slant for a given journalist over time" says Boxell.

The Problem

Boxell seeks to better understand two competing arguments regarding media bias claims. If journalists have ideological preferences over the content they produce and can use their positions to influence the production of news, then the ideological distribution of journalists would have important effects on the average slant of news that gets produced and consumed in the United States today. On the other hand, the average journalist may not have ideological preferences over the content they produce or may be unable to influence it—due to editorial constraints driven by media owners or consumer demand.

This project examines these competing hypotheses about the role of journalists empirically by measuring how the average slant of content produced by journalists changes as they switch outlets.

"TDM Studio allowed us to upload our fine-tuned ML model to the TDM servers and gave us access to the computing resources required to efficiently apply our model to millions of full-text news articles."

– Dr. Levi Boxell

The Process

The intuition for the project is as follows. Consider a journalist who switches from writing for *The Wall Street Journal* (WSJ) to writing for *The New York Times* (NYT), which is traditionally viewed as a more left-leaning outlet than the WSJ. If the slant of the articles produced by the journalist becomes significantly more liberal after moving to the NYT, then that indicates outlets themselves have a large role in determining the slant that gets produced. On the other hand, if the slant of the articles produced by the journalist look the same at both outlets, then that is suggestive of journalists themselves playing an important role in the type of content that gets produced.

While this example is simplistic, Boxell and Conway developed a more comprehensive model of the journalist labor market and derive more precise implications for how changes in slant from journalists moving across news outlets can inform use about the role of journalist ideology in news production.

To test the hypotheses empirically using this methodology, Boxell and Conway:

1. Developed a measure of media slant that is sensitive enough to measure differences in slant across articles written by the same journalist, but potentially at different outlets.
2. Obtained a large text-based dataset with rich enough information on the articles and the article bylines to apply the measure of slant and implement the statistical tests.

For the measure of slant, they fine-tuned a machine learning model (RoBERTa) using articles tweeted by politicians. Given the full text of an article tweeted by a politician, the fine-tuned model can predict the party of the politician with over 80 percent accuracy.

For the dataset, they turned to ProQuest's TDM Studio and US Newsstream collection. They used full-text data between 2013 and 2018 from more than 300 newspapers, providing millions of news articles to which they applied their Machine Learning (ML) model of slant. "TDM Studio allowed us to upload our fine-tuned ML model to the TDM servers and gave us access to the computing resources required to efficiently apply our model to millions of full-text news articles," says Boxell.

Figure 1 reports basic trends in slant across major US newspapers over this period. Here, values closer to one indicate more conservative slant and values closer to zero indicate more liberal slant. Overall, they see a general trend towards content that their ML model predicts as being more likely to be shared by a Democratic politician. This trend accelerates after the 2016 election, at least for some outlets.

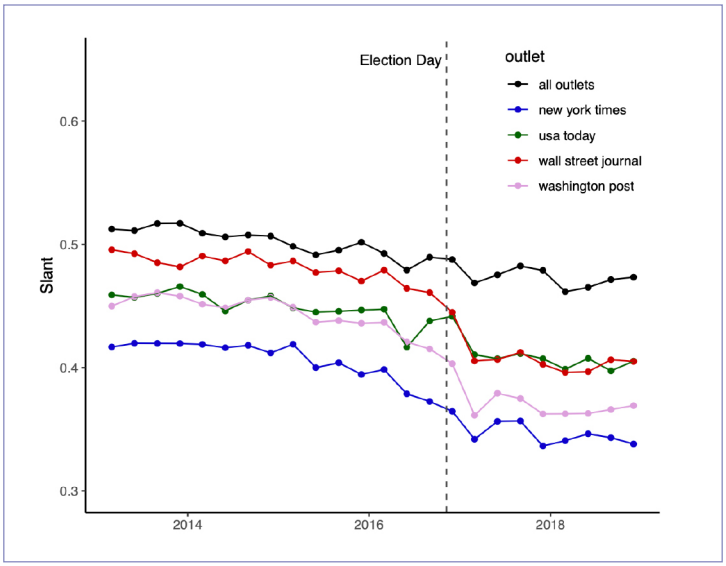

Figure 1: *Trends in Media Slant*

In their project, Boxell and Conway also merged author bylines to party registration data. This enabled them to measure differences in slant between journalists registered as Republicans or Democrats. **Figure 2** plots the distribution of journalist "fixed effects" (the average slant of journalists conditional on outlet and other factors) and reveals that journalist registered as Democrats tend to write more liberal content than journalists registered as Republicans, even when controlling for the news outlet.

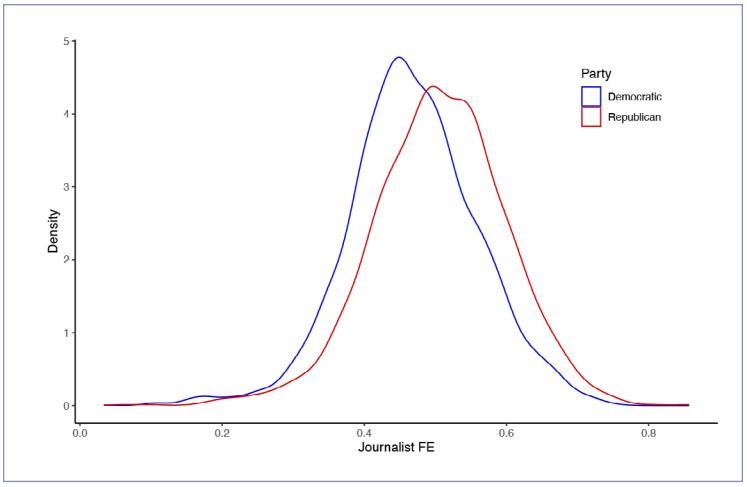

Figure 2: *Slant by Party Affiliation of Journalists*

Returning to the motivating example of a journalist that switches outlets, **Figure 3** reports a key test of the competing hypotheses. They use an event-study framework to estimate the extent to which journalists change slant when they switch outlets. "From the model we derived, the estimates allow us to reject the hypothesis that journalist preferences do not matter—because the post-move coefficient estimates are statistically different from one," says Boxell. In other words, consistent with the different distributions of average slant displayed in Figure 2, their estimates suggest that journalists (and journalist ideology) do play a role in the type of content that gets produced.

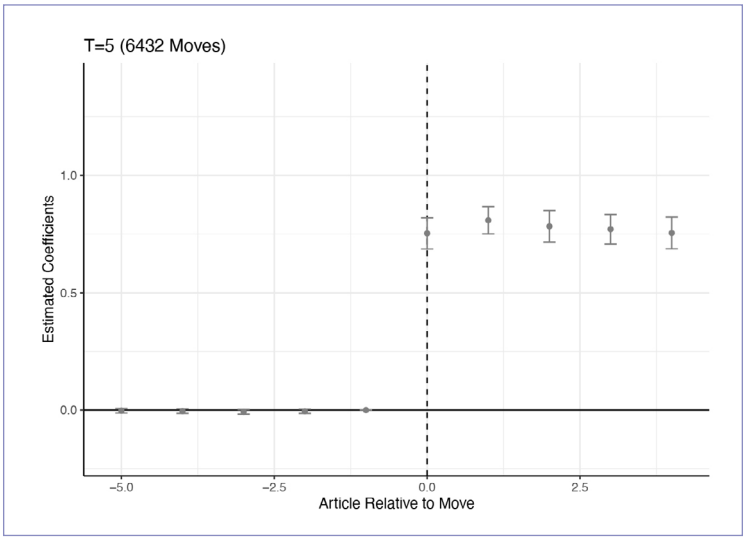

Figure 3: *Relative Change in Slant Upon Move*

Conclusion

Overall, this project provides new evidence on the role of journalists in driving the type of content that gets produced, and TDM Studio was a critical tool in the completion of the study. "By providing comprehensive, full-text data from hundreds of newspapers across multiple years, TDM Studio enabled us to track journalists as they changed outlets and measure media slant with a high degree of accuracy," says Boxell. "Moreover, TDM Studio provided an easy interface and responsive tech support for conducting this analysis via AWS Sagemaker."

About TDM Studio

ProQuest's workflow solution for text and data mining is designed for research, teaching and learning. TDM Studio provides access to sought-after content including current and historical newspapers, primary sources, scholarly journals, and dissertations and theses. It empowers researchers, students and faculty to analyze documents by uncovering connections and patterns that lead to career-defining discoveries.

References

Groseclose, Tim. 2011. *Left Turn: How Liberal Media Bias Distorts the American Mind.* St. Martin's Griffin: New York, NY.

Weaver, David H., Lars Willnat, G. Cleveland Wilhoit. 2019. *The American Journalist in the Digital Age: Another Look at U.S. News People.* Journalism & Mass Communication Quarterly. 96(1): 101–130

Learn more at proquest.com/go/tdm-studio
or contact your ProQuest representative today.