Title: Changes in total corpus word count over time appendix for "Quantifying the 'misinformation beat': 38 years of coverage in major U.S. daily newspapers"
Authors: Bryce Greene (1), Brian P. Harper (1), Christena E. Nippert-Eng (1)
Date: June 23rd, 2026
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix: Changes in total corpus word count over time

This appendix addresses the occurrences of misinformation terms relative to the total output of newspapers as it appears in the ProQuest archive for the months of our archived dataset. The ProQuest dataset may or may not be the same as the actual production and publication rates of these papers, depending on what material each paper submits to the ProQuest archive. This is typically limited by the contract between the publisher and the archive.

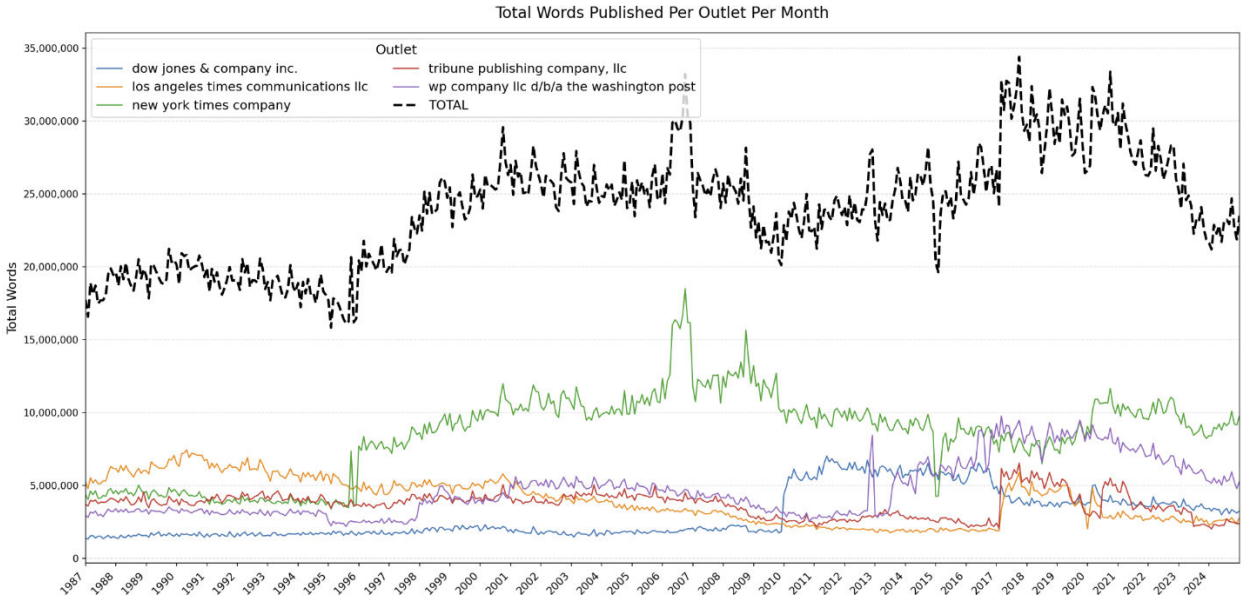

Figure A1. The total number of words published per month in total and also separated by newspaper outlet from 1987 to 2024.

Figure 1 shows the words published per month, both in total and separated by outlet over our timeline. There are plenty of granular changes, with word counts per month fluctuating between 1987 and 2024. The most dramatic changes appear to occur when an outlet changes the amount of data sent to ProQuest. We suspect these large discontinuities are indicative of either contract or technical changes in what is submitted to the archive rather than changes in journalistic output. For instance, there is a spike in total words published on January 1, 2017, near our time period of interest, stemming from a considerable increase in the number of words received from the Chicago Tribune and the L.A. Times. Such fluctuations are decoupled from the observed increases in misinformation terms, such as in November 2016 or July 2020.

As the charts below demonstrate, these alterations in the total word count do not affect the conclusions we reached in our investigation. The changes in the proportion of total words track very closely with the raw count of the key words, indicating that raw counts of key words and their changes over time are largely accurate indicators of the relative prominence of these terms in the press.

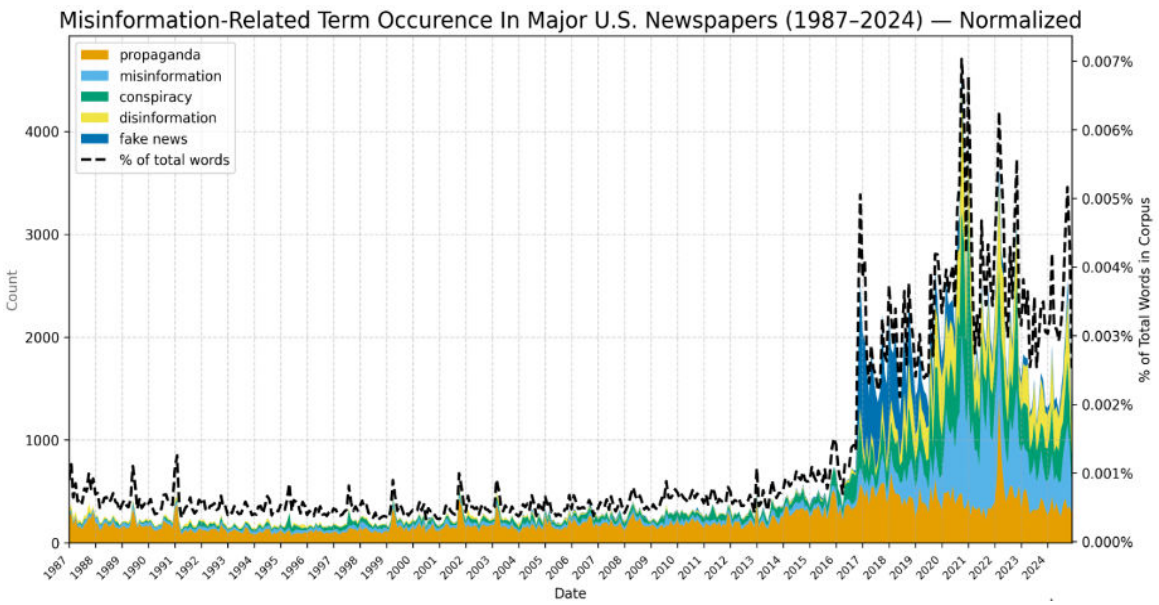

Figure A2. Our misinformation term counts alongside the percentage of the total words per month that were misinformation terms from 1987 to 2024.

The dotted line plotted in Figure 2 represents the word count of misinformation terms over the total number of words for that month, as can be seen in the secondary y-axis. This ratio has been placed atop our existing word counts at a scale to show their generally consistent parallel movement.

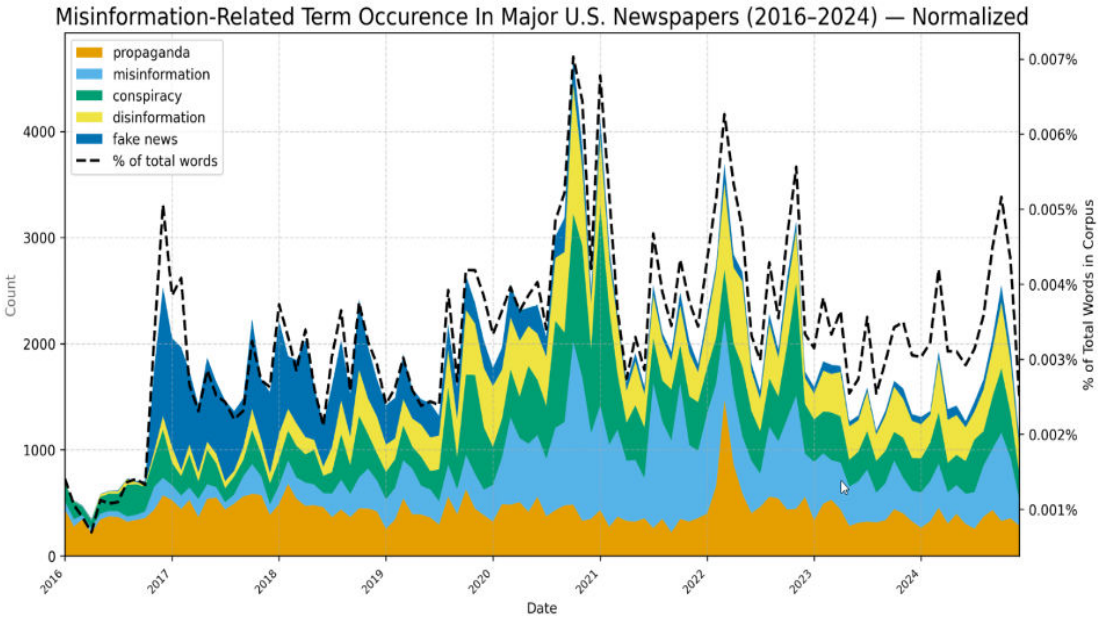

Figure A3. Our misinformation term counts alongside the percentage of the total words per month that were misinformation terms from 2016 to 2024.

Figure 3 focuses on our time period of acute interest (2016–2024). Overall, we see that the parallel pattern of total word count and our terms of interest continues in this period. Our analysis supports the conclusion that the patterns we see in the relative occurrence of misinformation terms over time are a reflection of their changing prominence in newspaper reporting, not in the changing amount of reporting that occurred during this same time.