When the News Matters: Newspaper Text and Real-Time Forecasts of U.S. Growth*

Aleksandr Shneider

October 2026

Click here for the latest version

Abstract

Does newspaper text help to forecast the U.S. economy in real time, and when? I build monthly text series, topic shares, policy-uncertainty counts and sentiment, from 11.76 million articles of seven U.S. newspapers over 1990–2020, and race them against 101 FRED-MD series in unrestricted MIDAS regressions estimated with four machine-learning methods on expanding real-time windows, scored as RMSE relative to an autoregression over 2003Q1–2020Q1. First, on all articles text alone is behind the hard data for GDP at the nowcast and for investment, and adding it to the hard data with the method held fixed gives 14 significant improvements and 7 deteriorations in 144 comparisons, about 7 of each being expected by chance; the improvements are for consumption. Second, which articles are used matters: an automatic filter for the economic articles, checked against hand labels, improves the text forecasts in 27 of 144 comparisons and worsens them in 4, in normal times as well as in the recession, under every selection rule tried. Third, the gain over the benchmark is a matter of timing: four quarters of 2008–2009 hold between a third and a half of the benchmark's squared error; in those quarters text beats the benchmark by 10–13% for GDP and 20–35% for consumption, and outside them neither text nor hard data beat it. The same-method gain from adding text survives scoring against first releases; the gain over the benchmark at the nowcast does not.

Keywords: Text as data, nowcasting, mixed-frequency regression, news media, machine learning
JEL: C53, C55, E37

*Department of Economics, University of Minnesota. Email: shnei003@umn.edu. I thank my advisers and seminar participants for comments. All errors are my own.

Introduction

Newspapers report on the economy every day, long before the statistical agencies publish their quarterly numbers, and what they report is read by the households and firms whose decisions those numbers summarise. Does this flow of text carry a signal that a forecaster of the macroeconomy could use in real time, and when over the business cycle is the signal there? This paper asks both questions on a corpus of seven U.S. print newspapers, the Boston Globe, the Chicago Tribune, the Los Angeles Times, the New York Times, USA Today, the Wall Street Journal and the Washington Post, drawn from a cleaned ProQuest U.S. Newsstream archive of 11.76 million articles over 1990–2020. It builds monthly text series from the articles, races them against the monthly hard-data series of FRED-MD in real-time forecasts of quarterly GDP, consumption and investment growth, and evaluates the forecasts by period as well as on the full sample. The closest antecedent is Ellingsen et al. (2022), who find on the Dow Jones Newswires that news topics forecast U.S. growth about as well as the hard data and add to them, most of all for consumption; the forecasting design here follows theirs so that the numbers can be set beside theirs.

The newspapers hold 10.55 million of the 11.76 million articles. Most of what they print is not about the economy: by a clustering of the articles into 250 groups, economy and business account for 22% of the articles, and the share moves with the news cycle and the sports calendar. I therefore build an automatic filter that selects the economic articles. Each article's title and lead are embedded with a published sentence encoder, the embeddings are grouped into 250 clusters, clusters are labelled economic by a rule based on their characteristic words and their closeness to a set of anchor sentences, and an article is kept when its cluster is economic or when it is itself close to the economic anchors. The filter keeps 24.7% of the corpus. Against 300 articles I labelled by hand, 71% of what it keeps is economic and it keeps 76% of what is economic. A keyword rule of the kind used to build news-based indices flags 7.9% of the articles; the filter keeps 87% of those and adds 2.1 million more.

From the economic articles I build monthly series for each newspaper: the shares of 80 topics from a latent Dirichlet allocation model re-fitted on the economic articles, policy-uncertainty counts, and dictionary and FinBERT sentiment. Each measure is summarised across the seven papers as a consensus and as a dispersion, which gives 206 monthly series. The hard data are the 101 FRED-MD series present in every real-time vintage from 1999 to 2024, screened for outliers.

The forecasting design follows Ellingsen et al. (2022) so that the results can be set beside theirs: quarterly growth of real GDP, real consumption and real non-residential investment is regressed on six monthly lags of every predictor, the unrestricted mixed-data-sampling form, with no lag of the outcome. The regression is estimated with four methods, Elastic Net, principal components followed by least squares, a random forest and gradient-boosted trees, and with the three methods

of their paper, on an expanding window of first-release data, separately for the nowcast and for one, two and three quarters ahead and for each month of the quarter. Forecasts are made for 2003Q1–2020Q1, 69 quarters, after training from 1997Q1 when five of the seven papers enter the corpus, and scored as RMSE relative to a direct autoregression on the real-time history of the outcome. Three predictor sets are raced: hard data, news, and both.

Three features of the evaluation matter for reading the results. First, two comparisons are reported rather than one: news against hard data, and hard data plus news against hard data with the method held fixed. The second comparison involves no choice of a best model after the fact. Second, every ratio is reported by period as well as on the full sample, with the periods defined on the quarter being forecast and the best model never re-chosen inside a period. Third, the test of equal accuracy allows for the serial correlation of multi-step forecast errors, with a Newey–West variance and the small-sample correction of Harvey et al. (1997); with it, the chance level for a table of 36 cells is about 2 significant cells in each direction and for 144 comparisons about 7.

The first result is on all articles of the seven papers. Text alone is behind the hard data for GDP at the nowcast and for investment at every horizon, and within a few percent of them for consumption and for GDP beyond the nowcast; across the 36 cells, text is significantly worse than hard data in 8 and significantly better in 3. Added to the hard data with the method held fixed, text gives 14 significant improvements and 7 significant deteriorations in 144 comparisons, against about 7 of each expected by chance. The improvements are concentrated in consumption, with 12 improvements and no deterioration in 48 comparisons.

The second result is that which articles are used matters. The same method on the economic articles does significantly better than on all articles in 27 of 144 comparisons and worse in 4, and the improvement is there in normal times as well, with 30 improvements and 7 deteriorations when the recession quarters are left out. Investment gains most, with its RMSE 10–14% lower at the nowcast and the next two horizons. The filter brings the news forecasts from behind the benchmark, significantly worse in 12 of 36 cells on all articles, to level with it, significantly worse in 1. The result does not depend on the particular selection rule: relabelling the doubtful clusters, halving the kept share or raising it by a quarter leaves the news-only ratios within 0.03 of the main specification for GDP and consumption and the increment counts within three of each other. On the economic articles, the main specification, text added to the hard data with the method held fixed lowers the RMSE in two thirds of the 144 comparisons, with a median ratio of 0.98, and significantly in 15 of them against 2 significant deteriorations. Ten of the 15 improvements are for consumption, which gains about 3% inside and outside the recession; for GDP text adds nothing outside the recession; for investment it adds nothing or makes the forecast worse.

The third result is about timing. On the full sample the best news-only model has an RMSE of 0.94, 0.96 and 0.95 of the autoregression's for GDP at the nowcast and one and two quarters

ahead, 0.87, 0.95 and 0.88 for consumption, and 1.04, 1.03 and 1.08 for investment; Ellingsen et al. (2022) report 0.94, 0.95 and 0.96, 0.89, 0.89 and 0.92, and 0.91, 0.94 and 1.02 on their corpus. But four quarters, 2008Q3–2009Q2, hold 41–51% of the benchmark's total squared error for GDP, 30–38% for consumption and 51–60% for investment, and the full-sample ratios are a statement about those quarters. Inside 2008–2009 the best news-only model beats the autoregression by 10–13% for GDP, 20–35% for consumption and 6–23% for investment, and beyond the nowcast it beats it by more than the hard data do. Outside those four quarters it is 3–6% worse than the autoregression for GDP, level for consumption and 27–31% worse for investment, and the hard data do no better. In normal times neither text nor hard data beat a simple autoregression of quarterly growth. On a two-paper sample that reaches back to 1996, the 2001 recession shows no gain either. These ratios score the forecasts against the vintage of August 2026; against the first published numbers the nowcast gain of text over the benchmark disappears (GDP 0.94 to 0.99, consumption 0.87 to 0.94, no cell significant), while the gain from adding text to the hard data with the method held fixed is unchanged (see Which actuals under robustness).

The contribution is threefold. First, the paper builds a cross-paper corpus of seven newspapers with several text measures side by side, topic shares, policy-uncertainty counts and two kinds of sentiment, and an automatic economic-article filter checked against hand labels, and shows that the filter improves the text forecasts. Second, it evaluates by period, with a test that allows for multi-step errors, and shows that the gains over the benchmark are concentrated in 2008–2009. Third, by adopting the regression form and the benchmark of Ellingsen et al. (2022), it gives an independent real-time check of their full-sample ratios on a different archive, ProQuest U.S. Newsstream rather than the Dow Jones Newswires, which reproduces them for GDP and consumption and not for investment. The sentiment comparison and the policy-uncertainty counts are investigations, not the contribution.

The paper sits in the literature on text as data in economics (Gentzkow et al., 2019), on news topics as time series for macroeconomic measurement and forecasting (Larsen and Thorsrud, 2019; Thorsrud, 2020; Bybee et al., 2024), on nowcasting with mixed-frequency data (Giannone et al., 2008; Foroni et al., 2015), and on the real-time evaluation of forecasts (Croushore and Stark, 2001). The section on related literature reviews it. The section on data describes the corpus, the filter, the text series, the hard data and the targets. The section on the forecasting design sets out the regression, the methods, the benchmarks and the tests. The results section reports the results in the order in which the exercise was built, from all articles to the main specification and from the full sample to the two recessions. The robustness section varies the selection rule, the methods, the news series, the outlier screen, the actuals and the combination, and the conclusion concludes. Three appendices hold the full tables, the details of the filter and the implementation details.

Related Literature

This paper draws on four literatures: newspaper text as a forecasting input, with the nowcasting and real-time tradition it sits in; topic models and machine-learning forecasts, which are its tools; the measurement of news sentiment and policy uncertainty, which are among the series it builds; and the political economy of the news media, which bears on the differences between newspapers. Each is reviewed briefly, in that order.

News, text and macroeconomic forecasting

Gentzkow et al. (2019) survey the use of text as data in economics and organise it around the methods that reduce high-dimensional text to a few series: dictionary counts, supervised learning and unsupervised models such as topic models; Varian (2014) makes the broader case for alternative data and machine-learning tools in econometrics. The line of work closest to this paper turns news into time series for macroeconomic measurement and forecasting. Larsen and Thorsrud (2019) convert the topics of a Norwegian business newspaper into daily series and document their predictive content, and Thorsrud (2020) builds a daily news-based coincident index of the business cycle from them. Ellingsen et al. (2022) carry the approach to the United States: they regress quarterly growth of GDP, consumption and investment on topic series from the Dow Jones Newswires archive and on the hard-data series of FRED-MD, estimate the regressions with machine-learning methods in real time, and find that the news topics carry information beyond the hard data, most of all for consumption. Their regression form and benchmark are adopted here so that the results can be set beside theirs. Bybee et al. (2024) estimate a topic model on about 800,000 Wall Street Journal articles and show that news attention tracks economic activity and adds to standard predictors; Kalamara et al. (2022) show that newspaper text forecasts U.K. real activity; Ardia et al. (2019) forecast U.S. growth from thousands of news-sentiment series with sparse methods; and Ulbricht et al. (2017) ask whether media data help to predict German industrial production. Newspaper text, then, has been found to carry a signal for real activity in several countries. What is less settled, and what this paper asks, is when in the sample the signal is there.

The forecasting design belongs to the large-dataset and nowcasting tradition. Stock and Watson (2002) forecast with principal components extracted from many predictors, Bai and Ng (2002) give criteria for the number of factors, and McCracken and Ng (2016) provide the FRED-MD panel that serves as the hard data here. Nowcasting, the use of the within-quarter flow of monthly information to predict a quarterly aggregate, follows Giannone et al. (2008) and the survey of Bańbura et al. (2011), and is the reason the forecasts are made in each month of the quarter. Monthly predictors and quarterly outcomes are bridged by the mixed-data sampling regression of Ghysels et al. (2004), in the unrestricted form of Foroni et al. (2015); Clements and Galvão (2008) show

that MIDAS forecasts U.S. output well, and Kuzin et al. (2011) compare it with mixed-frequency VARs. Because macroeconomic data are revised, Croushore and Stark (2001) make the case for real-time vintages, which the hard data and the targets of this paper follow.

Why news should carry forward-looking information at all is the subject of a smaller literature. Carroll (2003) shows that households update their expectations from news coverage, which is one reading of the consumption result below; Sims (2003) gives the rational-inattention reason why agents rely on summarised, news-mediated information; Nimark and Pitschner (2019) model the media as editors who choose which states of the world agents attend to; and Larsen et al. (2021) find that news predicts the inflation expectations of households more than those of professional forecasters.

Topic models and machine-learning forecasts

The latent Dirichlet allocation model of Blei et al. (2003) treats each document as a mixture of topics and each topic as a distribution over words; Blei (2012) gives an exposition, and Griffiths and Steyvers (2004) develop the collapsed Gibbs sampler and the choice of the number of topics, which bears on the 80-topic specification here. Chang et al. (2009) show that the topics correspond closely to the categories people would assign, while held-out likelihood and interpretability can diverge. The model discards word order, so it separates the senses of a word only through document-level co-occurrence. Hansen et al. (2018) use it to study deliberation in the Federal Open Market Committee; the news-topic series of Larsen and Thorsrud (2019) and Thorsrud (2020) descend from it, as does the pipeline of this paper.

The regressions are estimated with off-the-shelf methods: among them the LASSO of Tibshirani (1996) and the random forest of Breiman (2001). Medeiros et al. (2021) find random forests dominant in data-rich inflation forecasting, and Babii et al. (2022) propose a sparse-group LASSO suited to text-augmented nowcasting. Forecasts are compared with the test of Diebold and Mariano (1995) with the small-sample correction of Harvey et al. (1997). Nested comparisons against an autoregressive benchmark are the natural place for the adjustment of Clark and West (2007); this paper reports the Diebold–Mariano test throughout and treats the comparison of hard data plus news with hard data alone as a comparison of two estimated models rather than as a nested test. Forecast combinations follow Timmermann (2006).

Measuring the news: sentiment and policy uncertainty

Two of the text measures have literatures of their own. News sentiment begins with Tetlock (2007), who shows that media-sentiment word counts predict stock-market dynamics, and with the finance dictionary of Loughran and McDonald (2011), the dictionary baseline of this paper.

García (2013) documents that the predictive value of news sentiment is concentrated in recessions, a pattern the timing result below echoes, and Shapiro et al. (2022) build a daily news-sentiment index from combined lexicons. Neural classifiers promise context-sensitive sentiment: the transformer of Devlin et al. (2019) underlies FinBERT (Araci, 2019), the classifier used here, and Huang et al. (2023) develop a large-language-model variant that outperforms the dictionary at the sentence level. Dictionary and FinBERT sentiment are both among the text series of the main specification (see Text measures in the data section); whether the sentence-level advantage of neural sentiment survives aggregation to the publisher–month panel is not tested on its own. A model pretrained on later text and applied to historical articles in a real-time experiment raises a look-ahead concern (Sarkar and Vafa, 2024), which the data section states for FinBERT and for the filter's encoder.

News-based counts of policy uncertainty enter as predictors. Baker et al. (2016) construct the Economic Policy Uncertainty index from scaled counts of articles that meet economy, policy and uncertainty term criteria, validated by audits and against the VIX; their term sets, scaling and multi-paper aggregation define the policy-uncertainty series computed here from the economic articles of each newspaper and month. The channel from uncertainty to real activity is that of Bloom (2009). No validation of a count from this corpus against the published index is attempted, and no claim is made about the forecasting value of the uncertainty series on their own.

Media slant and publisher heterogeneity

Newspapers differ, and the political economy of the news media measures how. Gentzkow and Shapiro (2010) build a language-based index of slant and find that reader demand, not owner identity, accounts for most of its variation; Groseclose and Milyo (2005) measure slant from think-tank citations instead; Mullainathan and Shleifer (2005) give the theory of how competition and reader preferences generate slant; Martin and Yurukoglu (2017) show that outlet slant has measurable persuasion effects; and Gentzkow et al. (2019) generalise the measurement of group differences in high-dimensional choices such as partisan speech. Publisher heterogeneity enters this paper through the dispersion of every text measure across the seven newspapers, which is a predictor in the main specification; whether the forecasting value of an outlet's news moves with its political lean is left for later work.

Data

Three kinds of data enter the forecasts: the newspaper corpus and the monthly text series built from it, the monthly hard-data series of FRED-MD in real-time vintages, and the quarterly outcomes. This section describes each, and the filter that selects the economic articles from the corpus.

The newspaper corpus

The text comes from ProQuest U.S. Newsstream. After cleaning, the corpus holds 11,760,354 articles under ten source labels, dated 1990 to 2020. Seven of the ten labels are print newspapers: the Boston Globe, the Chicago Tribune, the Los Angeles Times, the New York Times, USA Today, the Wall Street Journal and the Washington Post. Together they hold 10,546,732 articles, and they are the cross-section used in the forecasts. The other three labels are the online editions of the Los Angeles Times, the Wall Street Journal and the Washington Post, which appear from 2010; they are left out so that the cross-section is the same kind of outlet throughout. The distinction between a print paper and its online edition is the vendor's. Each article carries ProQuest's publication title, the masthead under which ProQuest files it, and the source label is built from that title; the online editions are separate publication titles in ProQuest, so no article is assigned to print or online by its content or its web address. For each article the corpus holds the publisher, the title, the date, the text and the source label; it holds no ProQuest subject or section metadata, so nothing in what follows uses an editorial classification of the articles.

Until 1995 the corpus holds about 100 thousand articles a year, almost all from the New York Times and the Wall Street Journal (Figure 1 gives the volume by year and Table 1 the counts by newspaper for selected years). Five of the seven papers have essentially no articles before 1997: the Boston Globe, the Los Angeles Times, the Chicago Tribune, the Washington Post and USA Today each hold a few thousand articles or fewer in 1996 and tens of thousands in 1997. The break is the boundary of the vendor's full-text coverage, not a gap in digitisation: in ProQuest U.S. Newsstream the full text of the Los Angeles Times and the Chicago Tribune begins in December 1996, that of the Washington Post in April 1996 and that of the Boston Globe in January 1997, while the New York Times runs from 1980 and the Wall Street Journal from 1984.[1] The earlier years of those papers exist in ProQuest Historical Newspapers, a separate product not included in this corpus. From 1997 the seven papers hold 397 thousand articles, reach a peak of 508 thousand in 2006, and hold 319 thousand in 2020. With the three online editions the total stays between 400 and 500 thousand a year after 2010. The break in 1997 is the reason the main specification trains from 1997Q1 (see Methods and training in the design section): before that date the cross-paper averages and dispersions described below are computed over two papers, afterwards over seven. The two-paper window of the results subsection Two recessions, two papers uses the New York Times and the Wall Street Journal alone and therefore reaches back to 1990.

[1] ProQuest title list for U.S. Newsstream as reproduced by Columbia University Libraries, https://guides.library.columbia.edu/news/usnewspapers, accessed 14 October 2026.

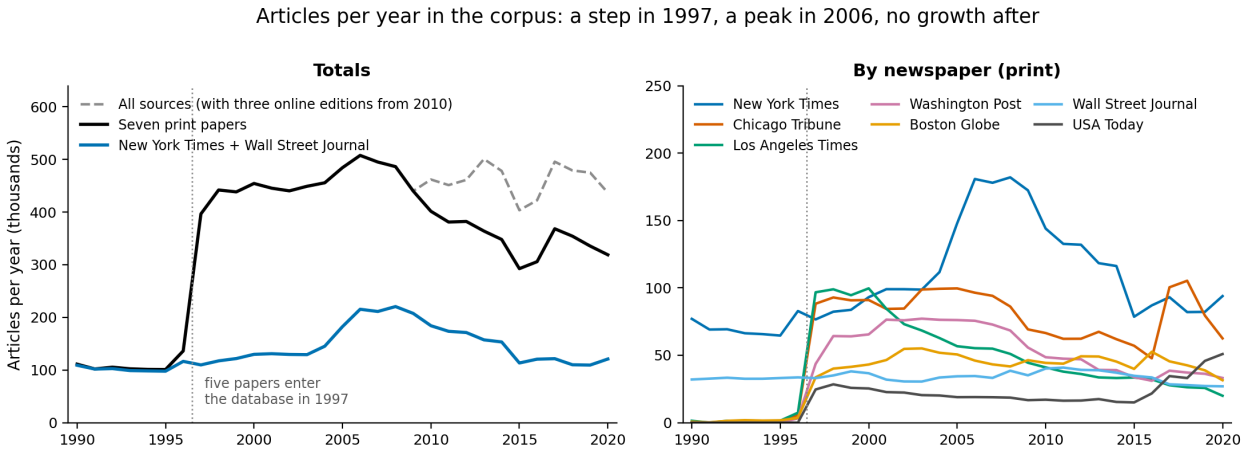


Figure 1: Articles per year in the seven print newspapers, 1990–2020, and in the ten source labels including the three online editions. The dotted line marks 1997, when five of the seven papers enter the database.

Table 1: Articles per newspaper in selected years, seven print newspapers.

Newspaper	1992	1994	1996	1997	1998	2005	2015	2020
New York Times	69,308	65,698	82,832	76,605	82,329	147,960	78,596	93,927
Wall Street Journal	33,344	32,573	33,595	33,058	35,024	34,394	34,764	26,956
Boston Globe	1,342	1,542	3,513	33,649	40,106	50,606	39,878	31,460
Los Angeles Times	1,339	1,361	7,352	96,738	98,952	56,712	33,479	19,942
Chicago Tribune	19	5	5,782	88,249	92,891	99,626	57,088	62,520
Washington Post	0	0	3,152	43,612	64,262	76,136	33,762	33,159
USA Today	0	0	4	24,617	28,426	18,921	15,001	50,921

Note: Counts of articles in the cleaned corpus by source label and year. Five of the seven papers have essentially no articles before 1997, so the cross-section is the New York Times and the Wall Street Journal until 1996 and seven papers from 1997.

What newspapers print

Most of what a newspaper prints is not about the economy. To see how much is, I grouped the articles into 250 clusters by the method described in the next subsection: the clusters are fitted on a random sample of 300,000 articles, every article of the corpus is then assigned to its nearest cluster, and each cluster is given a theme by hand from its most characteristic words. Figure 2 and Table 2 show the result for all 11.76 million articles. Economy and business, the 57 clusters the filter below labels economic, account for 21.9% of the articles. Sports account for 16.0%, arts, entertainment and media for 14.5%, and food, fashion, travel, family and religion for 9.3%; world affairs, local news, national politics, crime, health and science, and obituaries make up most of the rest. All themes other than economy and business together hold 78.1% of the articles. The shares on the 300,000-article fitting sample differ from these by at most 0.2 percentage points.

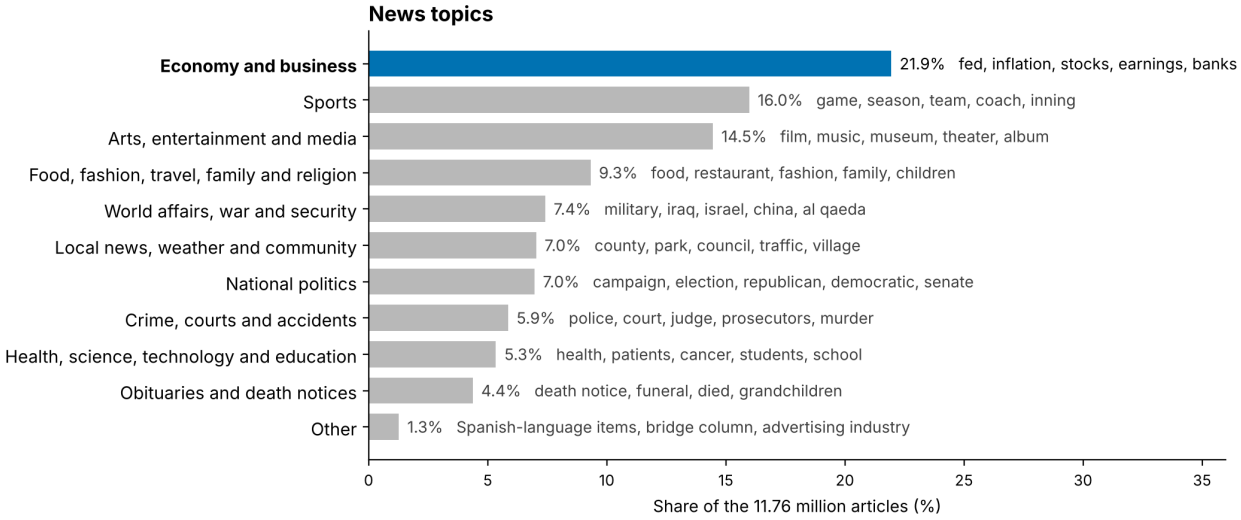


Figure 2: News topics: share of the 11.76 million articles by theme, 1990–2020.

The economic share is not only small, it moves over time (Figure B.2 in Appendix B). By the embedding filter it is 42–45% in 1990–1995, when only two papers are in the corpus, 22–25% in 1997–2009 and 16–19% in 2016–2020. By a keyword rule it falls from 8% in 2010 to 4.7% in 2019. Papers also write more about the economy in recessions: by the keyword rule the share rises from 6% to 10% in 2008–2009 and from 6% to 7% in 2001, and by the embedding filter from 23% to 25% in 2008–2009. A topic model fitted on all articles therefore produces topic shares that move with the other three quarters of the paper, with the sports season and the film calendar as much as with the economy. For a macroeconomic forecast that movement is noise, and it is the reason the main specification works with the economic articles only.

Table 2: Themes of the 250 article clusters and their shares of all articles.

What the articles are about	Share	Among their most characteristic words
Economy and business	21.9%	fed, inflation, stocks, earnings, banks, oil prices, mortgage, jobs
Sports	16.0%	game, season, team, coach, inning, quarterback
Arts, entertainment and media	14.5%	film, music, museum, theater, album, book
Food, fashion, travel, family and religion	9.3%	food, restaurant, fashion, family, children, church
World affairs, war and security	7.4%	military, iraq, israel, china, al qaeda, minister
Local news, weather and community	7.0%	county, park, council, traffic, village, hurricane
National politics	7.0%	campaign, election, republican, democratic, senate, white house
Crime, courts and accidents	5.9%	police, court, judge, prosecutors, murder, crash
Health, science, technology and education	5.3%	health, patients, cancer, students, school, scientists
Obituaries and death notices	4.4%	death notice, funeral, died, grandchildren
Other	1.3%	Spanish-language items, bridge column, advertising industry
All other themes together	78.1%	

Note: 250 clusters of sentence embeddings fitted on 300,000 articles sampled from the whole corpus, 1990–2020; shares are of all 11,760,354 articles by nearest cluster. Economy and business are the 57 clusters the filter labels economic; the other 193 clusters are grouped into themes by hand from their most characteristic words. The full article rule, which also admits articles by their own anchor margin, keeps 24.7% of all articles.

The economic-article filter

The filter selects the economic articles automatically, without subject metadata and without reading the articles. It works in four steps.

First, the sentence encoder `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019; Wang et al., 2020), a published model used as it is, embeds the title and the lead of each article. Second, a uniform random sample of 300,000 articles from the whole corpus is grouped into 250 clusters, and every article in the corpus is then assigned to its nearest cluster (the settings are in Appendix C). Third, a cluster is labelled economic by a rule, not by hand: either its most characteristic words score at least 0.25 on a list of economic vocabulary, or the cluster's centre is closer to ten economic anchor sentences than to ten non-economic ones by a margin of at least 0.10. The twenty anchor sentences were written by hand. Of the 250 clusters, 57 qualify, all of them by the anchor margin, 21 also by their words and none by their words alone. Fourth, an article is labelled economic if its nearest cluster is economic or if its own anchor margin is at least 0.10. The second route admits articles that sit in a non-economic cluster but read as economic themselves. Figure B.1 in Appendix B shows the 250 clusters on the two scores.

The filter keeps 2,908,967 articles, 24.7% of the 11,760,354 in the corpus. In the seven print papers it keeps 2,444,945 of 10,546,732 articles, and in the seven papers from 1997 to 2020, the period of the main specification, 2,124,570 of 9,787,639. The 80-topic model of the next subsection is re-fitted on the kept articles and every monthly text series is rebuilt from them.

A keyword rule serves as the comparison. It flags an article when the title and the lead contain terms from two of twenty economic term groups, or when the title alone contains a term from one

group; it is the kind of rule used to build news-based indices such as that of Baker et al. (2016). The keyword rule flags 923,363 articles, 7.9% of the corpus. Table 3 crosses the two rules. Of the articles the keywords flag, 87% are also kept by the embedding filter. The embedding filter adds 2,105,994 articles that the keywords miss, 72% of what it keeps; only the keywords flag 120,390 articles, 1% of the corpus. Table 4 shows four randomly drawn headlines from each of the three groups. The keywords alone misfire on words out of context, a team "hiring" a coach or a museum "lending" a sculpture; the embedding filter picks up company, deal and budget news that uses none of the listed words, and has false alarms of its own. The full samples of fifteen headlines per group are in Appendix B.

Table 3: The keyword rule against the embedding filter: articles flagged as economic by each.

Flagged as economic by	Articles	Share
Both rules	802,973	6.8%
Embedding filter only	2,105,994	17.9%
Keywords only	120,390	1.0%
Neither	8,730,997	74.2%
Keyword rule, total	923,363	7.9%
Embedding filter, total	2,908,967	24.7%

Note: All 11,760,354 articles of the ten source labels, 1990–2020. Keyword rule: terms from two of twenty term groups in the title and lead, or from one group in the title. Embedding filter: nearest cluster economic, or own anchor margin of at least 0.10.

To measure the filter's precision and recall, I drew 300 articles of the seven papers from 1997–2020 from 18 strata, defined by whether the article's nearest cluster is one of the 21 word-qualified clusters, one of the other 36 economic clusters, or a non-economic cluster, crossed with six bands of the article's anchor margin, and weighted the strata by their size. I labelled 116 of the 300 as economic, 182 as not economic and 2 as borderline, which are counted as not economic. Weighted by stratum size, 20.5% of all articles are economic by this standard. Against these labels the filter has a precision of 71%: that share of the articles it keeps are economic. Its recall is 76%: that share of the economic articles are kept. The 90% intervals from a bootstrap within strata are 66–77% and 66–86%. By their characteristic words, 8 of the 57 economic clusters are doubtful; they are about war, the military, immigration, schools, foreign policy or Washington politics. The robustness subsection The selection rule reports what happens to the forecasts when these clusters are relabelled and when the rule is tightened or loosened, and Appendix B gives the precision and recall of ten rules.

Table 4: Headlines flagged by the keyword rule only, by the embedding filter only, and by both.

Keywords only	**Embedding filter only**	**Both rules**
Raiders Close to Hiring Payton (Los Angeles Times, 2004)	How debt-ceiling talks came apart (Chicago Tribune, 2011)	Strong I.B.M. Report Helps Propel Dow to Another High (New York Times, 1999)
Italy lending Bernini art to the Getty (Los Angeles Times, 2008)	Investors say no to bid from Genzyme (Boston Globe, 2007)	Bank to Buy Credit Card Business for $6.45 Billion (New York Times, 2005)
Starbucks dips toe into beer and wine sales (USA Today, 2009)	Web Ads Hit Rock Bottom: Some Are Free (Wall Street Journal, 2002)	Bush Plan Would Hold Increase In 2004 Spending to Tight 4% (Wall Street Journal, 2003)
Trump May Choose Critic of Fannie and Freddie to Lead Agency Overseeing Them (New York Times, 2018)	A spin around the old vinyl (Boston Globe, 2014)	Dell, EMC Complete Merger (Wall Street Journal, 2016)

Note: Four of fifteen headlines drawn at random from each group among the articles of the seven print papers, 1997–2020. Headlines only; the newspaper and year are in brackets.

Text measures

Every text measure is computed for each newspaper and month from the economic articles of that paper in that month, and then summarised across the seven papers. The measures are of four kinds.

The first is topic shares. A latent Dirichlet allocation model (Blei et al., 2003) with 80 topics is fitted on a random sample of the 2.9 million economic articles (the implementation, the sample and the vocabulary are in Appendix C). Every economic article is then given its topic shares by the fitted model, and the paper-month topic share is the mean over the paper's articles in that month. The second is policy-uncertainty counts in the manner of Baker et al. (2016), computed from fixed term lists. The third is dictionary sentiment with the word lists of Loughran and McDonald (2011). The fourth is sentiment from FinBERT (Araci, 2019), a published classifier used as it is.

Each paper-month measure is then summarised in two series: the consensus, the equal-weighted mean across the papers, and the dispersion, the standard deviation across them. Months in which fewer than two papers have articles get no dispersion. The topic measures give 80 consensus and 80 dispersion series, 160 in all; with 18 policy-uncertainty and 28 sentiment series, the full set holds 206 series. Before the summary, a paper-month with too few articles relative to that paper's median monthly count is dropped, so that a month with a handful of articles does not enter the average or the dispersion as if it were a full month; a dropped paper-month is a missing value for that paper, and the cross-paper summary uses the remaining papers.

None of this is real time in the strict sense. The sentence encoder and FinBERT are published models pretrained on text that includes years after the dates being forecast. The clusters and the topic model are fitted once on articles from all of 1990–2020, so the definition of a cluster or a topic uses later text even though each month's shares use only that month's articles. The trimming threshold uses each paper's median over the whole period, and the word lists were compiled from text of the sample period. Each month's series uses only that month's articles, but the instruments that turn the articles into series were built with the whole corpus in view.

Hard data and targets

The hard data are the monthly series of FRED-MD (McCracken and Ng, 2016) in real-time vintages: 300 monthly vintage files from August 1999 to July 2024. At each forecast origin the vintage dated to the standing month is used; before August 1999 the earliest vintage is used, truncated at the standing month minus one, which is pseudo-real-time. The predictor set is the 101 series present in every one of the 300 vintages, so that the columns of the design are the same at every origin. Each series is transformed as FRED-MD prescribes, and each is known through the standing month minus one, the one-month publication lag. Following McCracken and Ng (2016), a transformed observation more than ten interquartile ranges from the series median, within each vintage, is set to missing; in the latest vintage this screen affects 151 observations in 53 series. The robustness subsection The outlier screen and the walk from all articles shows that the screen changes no conclusion.

The outcomes are the quarterly growth rates of real GDP, real personal consumption expenditures and real private non-residential fixed investment, each as the log difference of the levels times 100, with both levels taken from the same vintage; the series are not per capita (the series identifiers are in Appendix C). The models are trained on first-release growth, the value published in the quarter after the one it refers to. Vintages of real GDP start in December 1991, so the training outcomes for 1990Q1–1991Q3 in the two-paper window are taken from the earliest vintage. Forecasts are scored against the growth rates in the vintage of 26 August 2026, the most recent one, in every table unless stated otherwise; the same forecasts are also scored against the vintage that contains 2020Q1, the choice closest to that of Ellingsen et al. (2022), and against the first releases (see Which actuals under robustness). The evaluation ends in 2020Q1, so the collapse of 2020Q2 is outside the sample. The Survey of Professional Forecasters, the third benchmark of the design section, is the mean forecast of the survey of the Federal Reserve Bank of Philadelphia, converted from levels to growth rates for the matching quarter.

Forecasting Design

The forecasting exercise follows Ellingsen et al. (2022): quarterly growth is regressed on six monthly lags of every predictor, the regression is estimated with machine-learning methods on an expanding window of real-time data, and the forecasts are scored out of sample as root-mean-squared errors (RMSE) relative to an autoregressive benchmark. This section sets out the regression, the methods and the training scheme, the benchmarks, and the way the tables are read. The data that enter are described in the Data section above.

The regression

For each outcome, horizon $h \in \{0,1,2,3\}$, month $m \in \{1,2,3\}$ of the origin quarter q, and predictor set $S \in \{H, N, H{+}N\}$, the forecasting equation is

$$y_{q+h} \;=\; \alpha \;+\; \sum_{j \in S} \sum_{k=0}^{5} \beta_{j,k}\, x^{(m)}_{j,\,t_q-k} \;+\; \varepsilon_{q+h}, \tag{1}$$

where y_{q+h} is the growth rate of the quarter being forecast, $x_{j,t}$ is monthly series j, and t_q is the last month of quarter q. Six monthly lags of every series enter as separate regressors, so the equation covers the origin quarter and the one before it. This is the unrestricted mixed-data-sampling (U-MIDAS) form of Foroni et al. (2015), with $p = 6$ as in Ellingsen et al. (2022); no lag search is made. The equation contains no lag of the dependent variable. H is the set of hard-data series, N the set of news series, and $H{+}N$ their union; the sets are listed in the data subsections Text measures and Hard data and targets.

The superscript (m) marks the information set. The forecast is made in month m of quarter q, so the later months of that quarter are not yet observed. The lag layout is anchored at the quarter's third month and is the same in all three months; what changes from month 1 to month 3 is which slots are filled with data. News series for the standing month are available in that month; hard-data series arrive with a one-month publication lag and are known through the standing month minus one. A slot not yet observed at the standing month, or missing, is filled with the predictor's real-time mean (Appendix C).

The hard-data set holds 101 FRED-MD series, so $101 \times 6 = 606$ regressors. The topic-only news set holds 80 topic shares, each as a cross-paper average and a cross-paper dispersion series, so $160 \times 6 = 960$ regressors, and 1,566 together with the hard data. The main specification's news set adds the policy-uncertainty and sentiment series, 206 series in all, so 1,236 regressors alone and 1,842 together with the hard data.

Methods and training

Equation (1) has far more regressors than the training samples have quarters, so it is estimated with four methods that either shrink the coefficients or summarise the predictors. Each method gives its own forecast.

- *Elastic Net* (Zou and Hastie, 2005): a linear regression with a penalty that shrinks the coefficients and sets most of them to zero. The mix of the two penalties, the ℓ_1 ratio, and the penalty strength are chosen by cross-validation with time-ordered folds inside each training sample. In the main specification the cross-validation picks an ℓ_1 ratio of one, so that the Elastic Net forecast equals the LASSO forecast, in 62% of the fits.

- *PCA–OLS*: the predictors are summarised by their first five principal components and the outcome is regressed on them by ordinary least squares, in the tradition of Stock and Watson (2002).

- *Random forest* (Breiman, 2001): the average of many regression trees, each grown on a random part of the data.

- *XGBoost* (Chen and Guestrin, 2016): regression trees added one after another, each fitted to the residuals left by the previous ones.

The specification of Ellingsen et al. (2022) is estimated alongside as the replication specification: the LASSO of Tibshirani (1996), PCA–OLS and a random forest. Only the Elastic Net is tuned; the other three methods use fixed settings, listed with the software in Appendix C.

One model is estimated for every combination of outcome, horizon, month and predictor set, and each gives a direct forecast of y_{q+h}: no forecast is iterated from a shorter one. The training window expands: a forecast made at origin o uses every earlier origin $r \leq o-h-1$ whose outcome is known, so the latest outcome in the training sample is that of quarter $o-1$, which is the quarter whose first release normally appears in the first month of quarter o. The training cutoff is the same in the three months of a quarter.[2] A model is estimated only when at least 20 training quarters are available. Models are trained on first-release outcomes and scored against later vintages, as described under Hard data and targets in the data section.

In the main specification training starts in 1997Q1, when five of the seven newspapers enter the corpus, and forecasts are evaluated from 2003Q1 to 2020Q1: 69 quarters. The first nowcast

[2]The cutoff does not look at release dates. Two first releases inside the evaluation sample came out after the first month of the following quarter, both after government shutdowns: the advance estimate for 2013Q3 on 7 November 2013 and the initial estimate for 2018Q4 on 28 February 2019. The month-1 forecasts made in 2013Q4 and in 2019Q1 therefore train on one outcome that was published a few weeks later; this affects 288 of the 29,592 forecasts behind the main tables.

is trained on 24 quarters and the last on 92. Each further quarter of horizon removes two training quarters, because the forecast of a given quarter is made a quarter earlier and because every training outcome is paired with predictors dated h quarters before it. With the minimum of 20 quarters, three-quarter-ahead forecasts start in 2003Q3, so that horizon has 67 evaluation quarters and the others 69. The forecasts are made in each of the three months of the quarter. Across four methods, three outcomes, three months, four horizons and 69 or 67 quarters, each predictor set yields 9,864 forecasts; the three sets reported in the tables yield 29,592. A second window uses only the New York Times and the Wall Street Journal, the two papers with articles in every year: training from 1990Q1 and evaluation from 1996Q1 to 2020Q1, 97 quarters, the window of Ellingsen et al. (2022).

Benchmarks

Three forecasts that use neither news nor the FRED-MD series serve as benchmarks. The main one is an autoregression (AR). At each origin the real-time first-release history of the outcome through quarter $o-1$, the same information the models train on, is used to fit a direct $(h+1)$-step autoregression, and the lag order is chosen by the Bayesian information criterion over one to four lags at each origin. The AR benchmark captures the persistence of growth and nothing else. The second benchmark is constant growth: the expanding mean of real-time growth through quarter $o-1$, computed once at least eight observations are available. The third is the mean forecast of the Survey of Professional Forecasters (Croushore, 1993), whose level forecasts are converted to growth rates for the matching horizon.[3]

Evaluation

The forecast error of predictor set S at horizon h and month m over the T evaluation quarters is

$$\mathrm{RMSE}^{S}_{h,m} = \sqrt{\frac{1}{T}\sum_{q}\left(y_{q+h}-\hat{y}^{S}_{q+h\,|\,q,m}\right)^{2}}. \tag{2}$$

Every number in the tables is a ratio of two such RMSEs, computed on the same quarters with the same real-time information and the same estimation window. A ratio below one means that the forecast in the numerator is more accurate. Ratios are used because GDP, consumption and investment growth differ several-fold in volatility, which makes a ratio comparable across outcomes and horizons. A ratio also isolates the predictor set when everything else is held equal, equal accuracy can be tested, and Ellingsen et al. (2022) report the same ratios, so the tables can be read

[3]The real-time target history has no value for 1959Q1 for GDP and for 1950Q1 for consumption. The autoregression takes its lags by position across the missing quarter. An autoregression that leaves out the pairs whose lag falls on the missing quarter changes a forecast by at most 0.020 growth points for GDP and 0.298 for consumption.

against theirs. The ratio has one drawback: squared errors give volatile quarters a large weight, so a full-sample ratio can be decided by a few quarters. Every ratio is therefore also reported by period.

Three comparisons are made. The first puts news against hard data: $\text{RMSE}^{N}/\text{RMSE}^{H}$, which asks whether text can stand in for the hard data. The second puts hard data plus news against hard data alone: $\text{RMSE}^{H+N}/\text{RMSE}^{H}$, which asks whether text adds anything to the hard data. The third puts each forecast against a benchmark, with the benchmark's RMSE in the denominator.

The "best" model in a table is the method with the lowest RMSE over the evaluation sample within the named predictor set, for each outcome, horizon and month. It is chosen after the fact, the convention of Table 1 of Ellingsen et al. (2022). Because the choice is made ex post, it flatters the side on which it is made: a table of the best news-only model against the best hard-data model gives both sides their most favourable method, and a table of the best model against a benchmark gives the model that advantage and the benchmark none. The same-model comparisons involve no such choice: the ratio $\text{RMSE}(H+N, \text{method})/\text{RMSE}(H, \text{method})$ compares the same method with and without text. Both kinds of table are reported.

Sub-samples are defined on the quarter being forecast, not on the origin: the full sample; the sample without the four quarters 2008Q3–2009Q2; the quarters through 2007Q4; 2008Q1–2009Q4; and 2010Q1–2019Q4. Inside a sub-sample the best model is the full-sample pick and is never re-chosen. Where a table averages over the three months, or over methods or horizons, the average carries no test.

Equal accuracy is tested with the test of Diebold and Mariano (1995). Because a forecast $h+1$ steps ahead has serially correlated errors, the long-run variance of the loss differential is estimated with the Newey–West estimator (Newey and West, 1987) with truncation lag h, the statistic is scaled by the small-sample factor of Harvey et al. (1997), and it is referred to a Student-t distribution with $n-1$ degrees of freedom (the details are in Appendix C). Stars mark significance at the 10, 5 and 1% levels. No test is reported on fewer than 20 quarters, so the eight quarters of 2008–2009 carry no stars. The tests are two-sided at the 10% level, so about 5% of the cells in a table are expected to be significant in each direction by chance alone: about 2 of the 36 cells of a best-model table (three outcomes, four horizons, three months) and about 7 of the 144 cells of a same-model table (four methods times 36). Every count of significant cells below is to be read against these numbers. The tests adjust neither for the ex post choice of the best model nor for the number of cells.

In the tables, H0 denotes the nowcast of the current quarter and H1 to H3 the forecasts one to three quarters ahead; M1 to M3 denote the month of the quarter in which the forecast is made. Tables in the main text average the ratio over the three months; the appendix shows the months separately. Where a table gives three or four numbers separated by slashes in one cell, they are the horizons H0, H1, H2 (and H3) in that order.

Results

The results are presented in the order in which the exercise was built. The first subsection starts from all articles of the seven papers and asks the two questions of the design section: can text stand in for the hard data, and does it add to them. The second shows what the economic-article filter changes. The third reports the main specification, on economic articles, against the hard data, against the autoregressive benchmark and against the numbers of Ellingsen et al. (2022). The fourth asks when the gains over the benchmark are made, and the last compares the two recessions in the sample. Throughout, a ratio below one favours the forecast in the numerator, every main-text table averages over the three months of the quarter, and the chance level for counts of significant cells is about 2 of 36 and 7 of 144 in each direction (see Evaluation in the design section).

Baseline: all articles

The baseline specification uses every article of the seven papers: 80 topics fitted on all articles, the policy-uncertainty and sentiment series, the 101 FRED-MD series with the outlier screen, and the four methods, on the window 2003Q1–2020Q1. Table 5 reports the two comparisons with the best model on each side. The left block divides the RMSE of the best text-only method by the RMSE of the best hard-data method; the right block does the same for the best hard-plus-text method.

Table 5: All articles: text against hard data, and hard data plus text against hard data, best model on each side.

	Text only vs. hard only				Hard + text vs. hard only			
Outcome	H0	H1	H2	H3	H0	H1	H2	H3
GDP	1.07	1.00	0.98	0.99	1.02	1.00	0.98	0.98
Consumption	0.99	1.02	0.97	0.93	0.99	1.01	0.98	0.94
Investment	1.29	1.17	1.13	1.08	1.02	1.06	1.05	1.02

Note: Seven papers, all articles, 80 topics, the methods and text measures of the main specification; evaluation 2003Q1–2020Q1. Each cell is the RMSE of the best text-only method (left block) or of the best hard-plus-text method (right block) divided by the RMSE of the best hard-data method, where the best method is chosen ex post per outcome, horizon and month and the ratio is averaged over the three months. Below one: more accurate than hard data alone. The cells behind the averages are 3 outcomes $\times$ 4 horizons $\times$ 3 months $=$ 36.

Text alone is behind the hard data for the GDP nowcast, with a ratio of 1.07, and for investment at every horizon, with ratios from 1.29 at the nowcast to 1.08 three quarters ahead. For consumption and for GDP beyond the nowcast the two sides are within a few percent of each other, with the

text side ahead only for consumption three quarters ahead, at 0.93. Across the 36 cells, text is significantly worse than hard data in 8 and significantly better in 3. Adding text to the hard data changes little: the right block stays within 0.02 of one for GDP and consumption except at H3, where it reaches 0.94 for consumption, and is above one for investment; the best hard-plus-text model is significantly better than the best hard-data model in 3 of 36 cells and worse in 2, both counts at the chance level.

With the method held fixed, the average ratio of the hard-plus-text forecast to the hard-data forecast is near one at every horizon, from 1.02 at the nowcast to 0.97 three quarters ahead (Table 6). Of the 144 comparisons, 14 are significant improvements and 7 are significant deteriorations, against about 7 of each expected by chance. The improvements are concentrated in consumption, with 12 improvements and no deterioration in 48 comparisons; GDP and investment together have 2 improvements and 7 deteriorations. By method, the Elastic Net is the one that loses from text, with no improvement and 4 deteriorations in 36. The full set of 144 comparisons, the news-only forecasts against the three benchmarks and the comparison by period for this specification are in Appendix A (Tables A.16 to A.19).

Table 6: All articles: adding text to the hard data with the method held fixed.

Model	H0	H1	H2	H3	Sig. better	Sig. worse
Elastic Net	1.03	1.02	1.04	0.96	0 of 36	4 of 36
PCA–OLS	1.05	1.02	0.99	0.96	4 of 36	1 of 36
Random forest	1.00	1.02	0.99	0.99	5 of 36	2 of 36
XGBoost	0.99	1.00	0.93	0.95	5 of 36	0 of 36
All models	1.02	1.01	0.99	0.97	14 of 144	7 of 144
GDP	1.03	1.00	0.98	0.96	1 of 48	4 of 48
Consumption	0.98	0.99	0.99	0.96	12 of 48	0 of 48
Investment	1.05	1.05	1.00	0.97	1 of 48	3 of 48

Note: Each row compares the same method with and without text: the RMSE of the hard-plus-text forecast divided by the RMSE of the hard-data forecast, averaged over the three months and, in the lower blocks, over methods or outcomes. Below one: text helps. "Sig." counts the comparisons in the row that are significant at the 10% level: 36 per method (3 outcomes $\times$ 4 horizons $\times$ 3 months), 48 per outcome, 144 in all. About 7 of 144 are expected in each direction by chance.

What the filter changes

The main specification differs from the baseline in one respect: it is built on the economic articles selected by the filter of the data section, with the 80 topics re-fitted on them. Table 7

compares the two directly, method by method: the RMSE of a method on economic articles divided by the RMSE of the same method on all articles, with the same hard data and the same window. Only the articles change.

Table 7: The filter: the main specification against the baseline, same method, same hard data, same window.

	News only						Hard data + news					
Outcome	H0	H1	H2	H3	Sig. < 1	Sig. > 1	H0	H1	H2	H3	Sig. < 1	Sig. > 1
GDP	0.93	0.97	0.99	0.99	8 of 48	2 of 48	0.98	0.97	0.97	0.99	9 of 48	0 of 48
Consumption	1.00	1.00	0.96	1.03	4 of 48	2 of 48	1.00	0.99	0.95	1.02	5 of 48	4 of 48
Investment	0.87	0.86	0.90	1.01	15 of 48	0 of 48	1.00	0.93	0.98	1.04	10 of 48	1 of 48
All outcomes	0.94	0.95	0.95	1.01	27 of 144	4 of 144	0.99	0.96	0.97	1.02	24 of 144	5 of 144

Note: Each cell is the RMSE of a method on economic articles divided by the RMSE of the same method on all articles, averaged over the four methods and the three months. Below one: the filter helps. "Sig." counts the comparisons significant at the 10% level in a Diebold–Mariano test, of 48 per outcome and 144 in all; about 7 of 144 are expected in each direction by chance. Left block: news-only forecasts. Right block: hard data plus news.

With news alone, the filter lowers the RMSE in 27 of 144 comparisons significantly and raises it in 4, against about 7 of each by chance. The gain is largest for investment, whose RMSE falls by 10–14% at H0–H2, with 15 significant improvements in 48 comparisons and no deterioration; the GDP nowcast improves by 6–7%; consumption is unaffected, with ratios of 1.00 at H0 and H1. With news added to the hard data the pattern is the same and smaller: 24 significant improvements and 5 deteriorations. The effect is not a recession effect: without the four quarters 2008Q3–2009Q2 the news-only count is 30 improvements and 7 deteriorations. With the three methods and the topics-only set of Ellingsen et al. (2022) the filter gives 11 improvements and 1 deterioration in 81 comparisons (Table A.20).

Against the AR benchmark the filter moves the news forecasts from behind the benchmark to level with it. On all articles the best news-only model is significantly worse than the AR in 12 of 36 cells and significantly better in 4; on economic articles it is significantly worse in 1 and better in 2. The levels are in Table A.21: for GDP the full-sample ratio to the AR falls from 1.01, 1.00 and 0.99 at H0–H2 on all articles to 0.94, 0.96 and 0.95 on economic articles, and for investment from 1.23, 1.17 and 1.17 to 1.04, 1.03 and 1.08. In 2008–2009 the GDP ratio falls from 1.00, 0.97 and 0.94 to 0.88, 0.90 and 0.87. Outside the four worst quarters the filter changes little for GDP: 1.05, 1.06 and 1.06 on all articles against 1.03, 1.04 and 1.06 on economic articles. Against the hard data, text on all articles is significantly worse in 8 of 36 cells and better in 3, as reported above; on economic articles no cell differs significantly. The filter therefore improves the text forecasts, and does so in normal times as well as in the recession, but it does not by itself make text beat the benchmark outside the recession.

Main specification

The main specification uses the seven print papers, the economic articles only, 80 topics refitted on them, all text measures, the 101 FRED-MD series with the outlier screen and the four methods of the design section, with training from 1997Q1 and evaluation over the 69 quarters 2003Q1–2020Q1. Table 8 repeats the two best-model comparisons of Table 5 for it.

Table 8: Economic articles: text against hard data, and hard data plus text against hard data, best model on each side.

	Text only vs. hard only				Hard + text vs. hard only			
Outcome	H0	H1	H2	H3	H0	H1	H2	H3
GDP	0.99	0.96	0.95	0.97	0.99	0.97	0.96	0.97
Consumption	0.98	1.00	0.95	1.00	0.96	0.97	0.96	1.00
Investment	1.08	1.03	1.04	1.01	1.03	1.02	1.02	0.99

Note: Main specification. The best model on each side is chosen ex post per outcome, horizon and month; the ratio is averaged over the three months. Left block: RMSE of the best text-only method over the RMSE of the best hard-data method. Right block: the best hard-plus-text method over the best hard-data method. No cell of 36 differs significantly from one in either comparison.

No cell of 36 differs significantly from one in either comparison. Text alone is ahead of the hard data for GDP beyond the nowcast, at 0.96, 0.95 and 0.97 one to three quarters ahead, near one for consumption, at 0.95–1.00, and behind for investment, at 1.08 at the nowcast and 1.01–1.04 beyond. The best hard-plus-text model is within 0.04 of the best hard-data model everywhere. The same comparison method by method, which involves no ex post choice, is in Table A.5.

With the method held fixed, adding text to the hard data lowers the RMSE in two thirds of the 144 comparisons, with a median ratio of 0.98 (Table 9); 15 of the comparisons are significant improvements and 2 are significant deteriorations, against about 7 of each by chance. Ten of the 15 improvements are for consumption, with no deterioration in its 48 comparisons and an average ratio of 0.94 at H2; eleven of the 15 are at H2, where the average ratio across outcomes is 0.96. GDP has 4 improvements and 2 deteriorations and investment 1 improvement and none. By method, PCA–OLS gains most, with 6 improvements and no deterioration and ratios of 0.92–0.93 at H1–H3. XGBoost has 5 improvements and 1 deterioration, the random forest 3 and 0, and the Elastic Net 1 and 1. By period (Table A.7), the average ratio over methods, horizons and months is 0.95 in 2008–2009, 1.00 in 2003–2007 and 1.02 in 2010–2019; for consumption it is 0.96–0.98 in every period, for investment 0.94 in 2008–2009 and 1.11 in 2010–2019. The gain from text, where there is one, is a consumption gain and a recession gain. All 144 comparisons are in Table A.6, and Figure A.2 shows them over time.

On the full sample the best news-only model reproduces the ratios to the AR benchmark of

Table 9: Economic articles: adding text to the hard data with the method held fixed.

Model	H0	H1	H2	H3	Sig. better	Sig. worse
Elastic Net	1.02	0.97	1.02	1.05	1 of 36	1 of 36
PCA–OLS	1.00	0.93	0.92	0.93	6 of 36	0 of 36
Random forest	1.00	1.00	0.98	0.99	3 of 36	0 of 36
XGBoost	1.02	0.99	0.90	0.96	5 of 36	1 of 36
All models	1.01	0.97	0.96	0.98	15 of 144	2 of 144
GDP	1.01	0.97	0.95	0.95	4 of 48	2 of 48
Consumption	0.98	0.97	0.94	0.98	10 of 48	0 of 48
Investment	1.04	0.98	0.98	1.00	1 of 48	0 of 48

Note: Main specification. Each row compares the same method with and without text: the RMSE of the hard-plus-text forecast over the RMSE of the hard-data forecast, averaged over the three months and, in the lower blocks, over methods or outcomes. Below one: text helps. "Sig." counts the comparisons significant at the 10% level in a Diebold–Mariano test with Newey–West variance (lag h) and the small-sample correction: 36 per method, 48 per outcome, 144 in all; about 7 of 144 are expected in each direction by chance.

Ellingsen et al. (2022) for GDP and consumption and not for investment (Table 10). For GDP the main specification gives 0.94, 0.96 and 0.95 at H0–H2, against 0.94, 0.95 and 0.96 in their Table 1; for consumption 0.87, 0.95 and 0.88 against 0.89, 0.89 and 0.92. For investment this paper finds the best news-only model behind the benchmark, at 1.04, 1.03 and 1.08, where they find 0.91, 0.94 and 1.02. With their three methods and the topics only (last column) GDP is level with the benchmark, consumption is similar and investment is further behind. The windows differ: theirs runs from 1996Q1, this one from 2003Q1.

Table 10: Best news-only model against the AR benchmark on the full sample, next to the numbers of Ellingsen, Larsen and Thorsrud (2022).

Outcome	Ellingsen et al. (2022) Table 1	This paper, main specification	This paper, with the methods of Ellingsen et al.
GDP	0.94 / 0.95 / 0.96	0.94 / 0.96 / 0.95	1.01 / 0.99 / 0.98
Consumption	0.89 / 0.89 / 0.92	0.87 / 0.95 / 0.88	0.92 / 0.96 / 0.89
Investment	0.91 / 0.94 / 1.02	1.04 / 1.03 / 1.08	1.13 / 1.11 / 1.16

Note: RMSE of the best news-only model over the RMSE of the AR benchmark, mean over the three months, at H0 / H1 / H2. The best model is chosen ex post: of three methods in Ellingsen et al. (2022), of four in the main specification and of three in the last column. First column: Ellingsen et al. (2022), Table 1, their corpus, evaluation 1996Q1–2020Q1. Second and third columns: this paper, seven newspapers, economic articles, evaluation 2003Q1–2020Q1. Stars for the cells behind the averages are in Tables A.2 and A.10.

Behind the averages, the best news-only model is below the AR benchmark for GDP in all 12 cells, with ratios of 0.91–0.98, and for consumption in 11 of 12, with ratios of 0.84–1.01 (Table A.2). Few of the cells are significant. With the test of the design section, which allows for the serial correlation of multi-step errors, the best news-only model is significantly better than the AR in 2 of 36 cells and significantly worse in 1; the best hard-plus-news model is significantly better in 4 and worse in 2; the best hard-data model is significantly better or worse in none (Table A.1). All six of the significant improvements are for consumption at the nowcast or one quarter ahead. Against the constant-growth benchmark the counts are 4 better and 2 worse for news only. The Survey of Professional Forecasters is ahead of the best news-only model in all 36 cells, with ratios of 1.01–1.37, and significantly so in 30. The full tables against the three benchmarks for the three predictor sets are Tables A.2 to A.4.

Timing: four quarters hold half of the error

The full-sample ratios of Table 10 are averages of squared errors over 69 quarters, and squared errors weight volatile quarters heavily. Table 11 shows how heavily. The four quarters 2008Q3–2009Q2 hold 41–51% of the AR benchmark's total squared error for GDP over 2003Q1–2020Q1, depending on the horizon, 30–38% for consumption and 51–60% for investment. They are 4 of the 69 evaluation quarters. Every full-sample ratio is therefore largely a statement about those four quarters.

Table 11: Share of the AR benchmark's squared forecast error that falls in 2008Q3–2009Q2.

Outcome	H0	H1	H2	H3
GDP	41%	47%	50%	51%
Consumption	32%	30%	38%	37%
Investment	51%	57%	59%	60%

Note: For each outcome and horizon, the AR benchmark's squared errors for the target quarters 2008Q3–2009Q2, summed over origins and months, divided by the same sum over the whole evaluation sample 2003Q1–2020Q1 (69 quarters). The four quarters are 4 of the 69.

Three samples defined on the quarter being forecast split the comparison with the benchmark: all 69 quarters, the 65 quarters without 2008Q3–2009Q2, and the eight quarters 2008Q1–2009Q4 (Table 12). The best model is the full-sample pick and is not re-chosen inside a sub-sample. The ratios are averaged over the nowcast and the next two quarters; the horizon-by-horizon version is

Table A.8, and the version with five samples Table A.9.

Table 12: Against the AR benchmark, by period: full sample, without the four worst quarters, and 2008–2009.

Outcome	Best model	All 69 quarters	Without the four worst quarters	2008–2009 only
GDP	News only	0.95	1.04	0.88
	Hard data only	0.99	1.02	0.97
	Hard data + news	0.96	1.05	0.91
Consumption	News only	0.90	0.98	0.74
	Hard data only	0.92	1.00	0.84
	Hard data + news	0.89	0.98	0.75
Investment	News only	1.05	1.29	0.85
	Hard data only	1.00	1.06	0.98
	Hard data + news	1.02	1.18	0.95

Note: Main specification. RMSE of the best model within the named predictor set over the RMSE of the AR benchmark, averaged over the three months and over H0, H1 and H2. Below one: more accurate than the benchmark. The best model is chosen ex post on the full sample and is not re-chosen inside a sub-sample. The four worst quarters are 2008Q3–2009Q2; 2008–2009 is 2008Q1–2009Q4. Periods are defined on the quarter being forecast. Eight quarters are too few for the test.

Without the four worst quarters no predictor set beats the AR benchmark for GDP: news alone is at 1.04, hard data alone at 1.02, and both together at 1.05. Horizon by horizon (Table A.8) the news-only ratio is 1.03–1.06 at H0–H2, the hard-data ratio 1.01–1.03 and the combined ratio 1.01–1.09. For consumption outside the recession the three sets stay within a few percent of the benchmark: 0.96–1.02 for news, 0.99–1.02 for hard data and 0.94–1.00 for both. For investment the best news-only model is far above the benchmark outside the recession, at 1.23–1.31 at H0–H3, and the hard data are at 1.00–1.09.

In 2008–2009 everything beats the benchmark, and text by more than hard data. News alone is 12% below the AR for GDP, at 0.88, and 26% below for consumption, at 0.74; hard data alone are 3% and 16% below, at 0.97 and 0.84. By horizon the best news-only model is 10–13% below the AR for GDP, at 0.88, 0.90 and 0.87 at H0–H2, and 20–35% below for consumption, at 0.65, 0.80 and 0.76; the hard-data model is 10% below for GDP at the nowcast and level with the benchmark at H1 and H2, at 0.99 and 1.01. For investment news alone is at 0.85 and hard data at 0.98. These are eight quarters and carry no test.

The running gain over the benchmark, quarter by quarter, shows the same thing (Figure 3). Each line is a running total of the AR benchmark's squared error minus the model's, for the best model in each predictor set, averaged over the three months; the line rises in a quarter in which the model was more accurate than the benchmark and falls in one in which the benchmark was. Until 2007 the lines stay at or below zero: there is no gain. In 2008–2009 they jump; after 2010 they are roughly flat. At the nowcast the three lines jump together: hard data, text and both all

see the recession. One quarter ahead the lines with news jump and the hard-data line less: there the gain comes from text. The step at the right edge of the figure is 2020Q1, the first quarter of the pandemic. The version with the second quarter ahead, and the same figure for investment, are Figures A.1 and A.3.

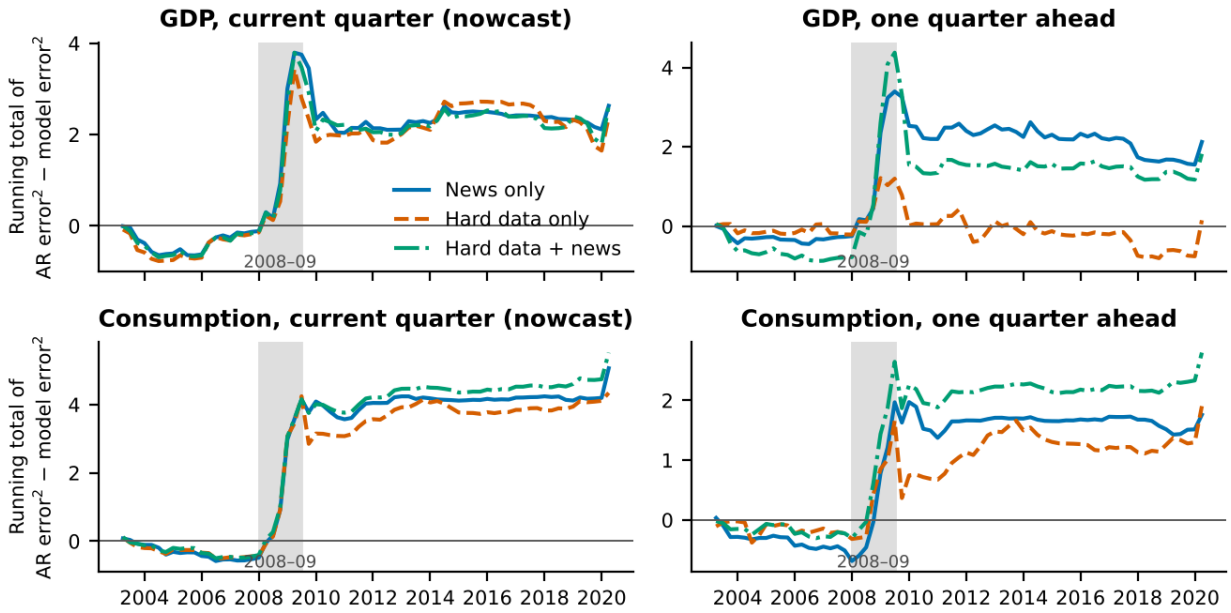


Figure 3: When the gains over the AR benchmark are made: cumulative squared-error gain of the best model in each predictor set, nowcast and one quarter ahead, main specification, 2003Q1–2020Q1.

The timing result says what the full-sample ratios measure. Text beats the autoregressive benchmark on the full sample because it beats it in the Great Recession, by a wide margin, and the full-sample ratio is dominated by those quarters. In normal times neither text nor the hard data beat a simple autoregression of quarterly growth.

Two recessions, two papers

The evaluation sample of the main specification contains one deep recession. To see a second one, I use the two papers that have articles in every year, the New York Times and the Wall Street Journal, train from 1990 and evaluate from 1996Q1 to 2020Q1, 97 quarters, with the main specification otherwise unchanged. The cross-section is two papers, so the dispersion series are the dispersion between two papers. This sample contains the 2001 recession, 2001Q1–2001Q4, as well as 2008–2009.

On this sample the cumulative gain over the benchmark, drawn as in Figure 3, falls or stays flat around 2001 and jumps in 2008–2009 for text and hard data alike (Figure 4). Table 13 gives the

numbers: the RMSE of the best model over the RMSE of the AR benchmark inside each recession, averaged over the three months and over H0–H2.

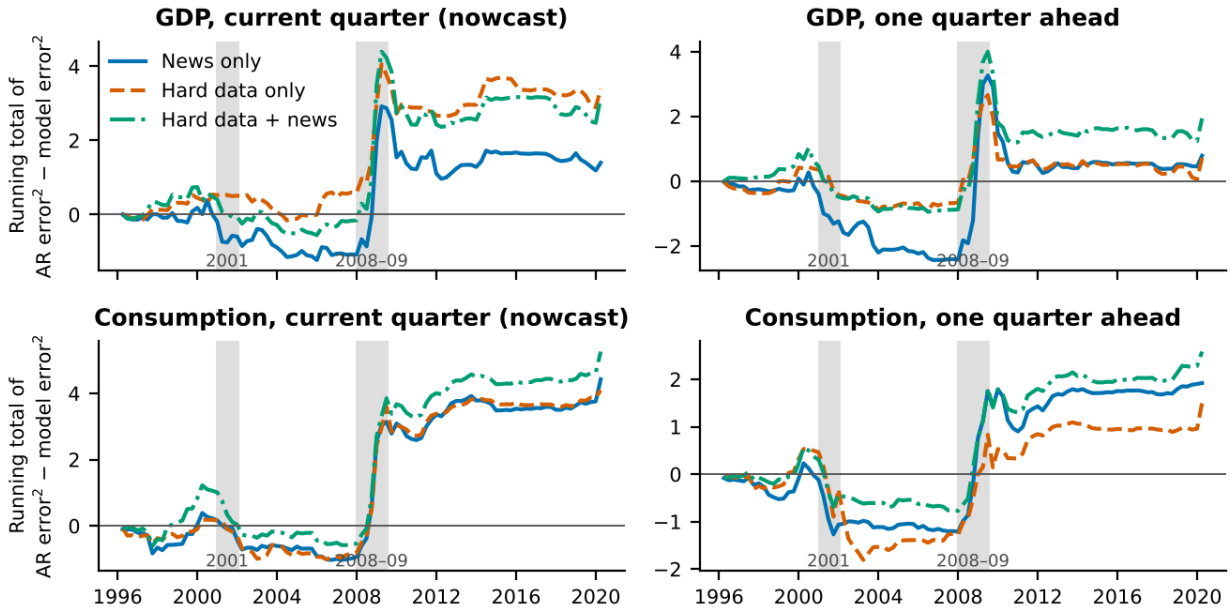


Figure 4: Two recessions: cumulative squared-error gain over the AR benchmark of the best model in each predictor set, New York Times and Wall Street Journal, evaluation 1996Q1–2020Q1, nowcast and one quarter ahead.

In the 2001 recession nothing beat the benchmark: news alone is at 1.13 for GDP and 1.20 for consumption, hard data alone at 1.09 and 1.16, and both together at 1.13 and 1.23. By horizon the news-only ratio is 1.10–1.15 for GDP and 1.16–1.27 for consumption, and the hard-data ratio 0.94–1.19 and 1.11–1.23 (Table A.33). In 2008–2009 everything did: news alone is at 0.86 for GDP and 0.74 for consumption, hard data alone at 0.89 and 0.77, both together at 0.87 and 0.76. With two papers text and hard data are level in the Great Recession: by horizon 0.85–0.87 against 0.84–0.95 for GDP.

On the full 97 quarters the best news-only model is at 0.98–0.99 of the benchmark for GDP and 0.91–0.96 for consumption at H0–H2, and one cell of 36 is significant (Table A.31). Before 2008 it is above one, at 1.05–1.12 for GDP and 1.05–1.17 for consumption at H0–H3; in 2008–2009 it is at 0.85–0.87 and 0.70–0.77 at H0–H2 (Table A.32). In 2010–2019 the news-only model is above the benchmark for GDP, at 1.03–1.08, and a few percent below it for consumption, at 0.91–0.98 at H0–H2, where the hard-data model is at 0.89–0.94. The gain of text over the benchmark is one episode: the 2001 recession, a shallower one, shows no such gain.

Table 13: Two recessions: best model against the AR benchmark inside 2001 and inside 2008–2009, two papers.

	2001 recession		2008–2009	
Best model	GDP	Consumption	GDP	Consumption
News only	1.13	1.20	0.86	0.74
Hard data only	1.09	1.16	0.89	0.77
Hard data + news	1.13	1.23	0.87	0.76

Note: New York Times and Wall Street Journal only; training from 1990, evaluation 1996Q1–2020Q1 (97 quarters), main specification otherwise. RMSE of the best model within each predictor set over the RMSE of the AR benchmark inside the recession, averaged over the three months and over H0, H1 and H2. Below one: more accurate than the benchmark. The best model is chosen ex post on the full sample and not re-chosen. 2001 recession: 2001Q1–2001Q4; 2008–2009: 2008Q1–2009Q4. Four and eight quarters are too few for the test.

Robustness

This section varies the main specification one ingredient at a time: the rule that selects the economic articles, the forecasting methods, the construction of the news series, the outlier screen on the hard data, the actuals the forecasts are scored against, and the way the forecasts are combined. The tables behind the summaries are in Appendix A unless shown here.

The selection rule

The filter of the data section is one rule among many. To see whether the forecasts depend on it, I put three other rules through the same forecasting exercise. The three were picked from the ten rules of Table B.4 in Appendix B on what they keep, their precision and recall against the 300 hand labels, before any of their forecasts was computed. The cleaned rule relabels the 8 doubtful clusters, so that 49 clusters are economic, and keeps 22.8% of the articles with a precision of 75% and a recall of 76%. The strict rule raises the anchor margin to 0.20 at both the cluster and the article level, so that 32 clusters are economic; it keeps 13.3% of the articles, about half of the main rule, with a precision of 87% and a recall of 49%. The loose rule lowers the margin to 0.05, so that 64 clusters are economic; it keeps 31.0%, a quarter more than the main rule, with a precision of 62% and a recall of 81%. The cleaned and strict rules reuse the main specification's topic model, since they keep a subset of its articles; the loose rule re-fits it.

Against the AR benchmark the best news-only model stays within 0.03 of the main specification for GDP and consumption, and within 0.04 for investment, at every horizon under every rule (Table 14). Adding news to the hard data with the method held fixed gives 14–18 significant im-

provements of 144 under every rule, against 15 under the main rule, and 2–3 deteriorations. The loose rule hurts investment: 4 of its 12 investment cells are significantly worse than the AR, against 1 under the main rule. Method by method against the main specification, no rule is uniformly better: the cleaned rule is significantly better in 10 of 144 comparisons and worse in 18, the strict rule 10 and 8, the loose rule 6 and 5. Relabelling the doubtful clusters, halving the kept share or raising it by a quarter leaves the forecasts essentially where they are.

Table 14: Does the result depend on the selection rule? Four rules through the same forecasting exercise.

	Rule				Labels (%)		News only vs AR				Hard + news	vs main
	Clusters	Margin	Kept (%)	Topics	Prec.	Rec.	GDP	Consumption	Investment	sig. + / −	sig. + / −	sig. + / −
Main	57	0.10	24.7	re-fitted	71	76	0.94/0.96/0.95	0.87/0.95/0.88	1.04/1.03/1.08	2 / 1	15 / 2	–
Cleaned	49	0.10	22.8	held fixed	75	76	0.96/0.97/0.95	0.88/0.95/0.88	1.07/1.03/1.04	3 / 0	18 / 2	10 / 18
Strict	32	0.20	13.3	held fixed	87	49	0.95/0.97/0.97	0.88/0.98/0.90	1.05/1.02/1.05	3 / 0	15 / 3	10 / 8
Loose	64	0.05	31.0	re-fitted	62	81	0.95/0.97/0.96	0.88/0.95/0.90	1.05/1.06/1.09	2 / 4	14 / 3	6 / 5

Note: Kept: share of the 11.76 million articles. Precision: share of the kept articles that are economic; recall: share of the economic articles that are kept; both against 300 hand-labelled articles of the seven papers, 1997–2020. Cleaned and strict reuse the main specification's topic model; loose re-fits it. News only vs AR: best news-only model over the AR benchmark, H0 / H1 / H2, mean over the three months, final-vintage actuals; sig. + / −: cells significantly better / worse than the AR at the 10% level, H0–H3, of 36. Hard + news: the same method with and without news, significantly better / worse, of 144. Vs main: the RMSE of the news-only forecast under the rule over its RMSE under the main rule, same method, significantly better / worse, of 144.

The methods of Ellingsen, Larsen and Thorsrud

The main specification uses four methods and all text measures. With the three methods of Ellingsen et al. (2022), the LASSO, PCA–OLS and a random forest, and the 160 topic series alone, on the same articles, hard data and window, the best news-only model against the AR benchmark is at 0.89–0.96 for consumption, 0.95–1.01 for GDP and 1.10–1.19 for investment across the 27 cells (Table A.10). It is significantly better than the AR in 2 cells and worse in 7, the worse cells all investment; the best hard-data model is significantly better in 5 and worse in 1 (Table A.11). With the method held fixed, adding the topics to the hard data gives 9 significant improvements and 8 deteriorations in 81 comparisons, a median ratio of 0.99 and a ratio below one in 59% of the comparisons (Table A.13); the improvements are consumption (7 of 27), the deteriorations investment (6 of 27). The three-method specification is the last row but one of Table 18 below. The step from it to the main specification, four methods and all text measures, is what brings GDP below the benchmark on the full sample, from 1.01, 0.99 and 0.98 to 0.94, 0.96 and 0.95, and raises the same-method improvements from 9 of 81 to 15 of 144.

Variants of the news series

Three variants of the topic series are raced with the three methods in the main specification, one change each against the 160-series set with six free monthly lags (Table 15). Detrending replaces each topic series by its log deviation from its own mean over the previous 36 months (at least 12; months before January 1997 are not used). Quarterly averaging replaces the six monthly lags by two quarterly averages per series, the origin quarter's observed months and the previous quarter, 320 regressors in all. Consensus only drops the 80 cross-paper dispersion series and keeps the 80 consensus series, 480 regressors.

Table 15: Variants of the news series: detrended topics, quarterly averages, consensus series only.

	News only vs. AR, H0 / H1 / H2			Hard + news vs. hard, same model		
News measures	GDP	Consumption	Investment	Median	Below one	Sig. < 1 / > 1
Baseline: 160 series, six free lags	1.01 / 0.99 / 0.98	0.92 / 0.96 / 0.89	1.13 / 1.11 / 1.16	0.99	59%	9 / 8
Detrended topic series	1.02 / 1.00 / 0.97	0.99 / 1.04 / 0.99	1.03 / 1.03 / 1.03	0.99	53%	12 / 13
Quarterly averages	1.02 / 0.99 / 0.98	0.92 / 0.95 / 0.89	1.12 / 1.09 / 1.14	0.99	65%	9 / 5
Consensus only (80 series)	1.01 / 0.98 / 0.96	0.91 / 0.96 / 0.89	1.13 / 1.12 / 1.15	0.99	64%	11 / 4

Note: The three methods of Ellingsen et al. (2022), main specification otherwise. Left block: best news-only model over the AR benchmark, H0 / H1 / H2, mean over the three months. Right block: the same method with and without news, 81 comparisons; the median ratio, the share of ratios below one, and the comparisons significantly below / above one at the 10% level.

Quarterly averages and consensus-only move no ratio in Table 15 by more than 0.02. Detrending does more. It removes most of the investment losses of 2010–2019, from 1.34 to 1.06 against the AR in Table 16, and all of the 2008–2009 gain for consumption, from 0.88 to 1.01; with the method held fixed it gives 12 improvements and 13 deteriorations in 81. Drift in the topic shares therefore accounts for the investment losses in the later years, and the recession signal for consumption lives in the levels of the topic shares, not in their deviations from a rolling mean. No variant improves on the baseline overall.

Table 16: Variants of the news series by evaluation sample: best news-only model against the AR benchmark.

	GDP			Consumption			Investment		
News measures	Full	2008–09	2010–19	Full	2008–09	2010–19	Full	2008–09	2010–19
Baseline: 160 series, six free lags	0.99	0.98	1.03	0.92	0.88	0.97	1.13	1.06	1.34
Detrended topic series	1.00	1.01	1.00	1.01	1.01	0.98	1.03	1.06	1.06
Quarterly averages	1.00	0.99	1.03	0.92	0.89	0.97	1.12	1.03	1.31
Consensus only (80 series)	0.99	0.98	1.00	0.92	0.87	0.96	1.13	1.05	1.33

Note: The three methods of Ellingsen et al. (2022), main specification otherwise. RMSE of the best news-only model over the RMSE of the AR benchmark, mean over H0–H2 and the three months, on the full sample, in 2008Q1–2009Q4 and in 2010Q1–2019Q4. Samples are defined on the quarter being forecast; the best model is the full-sample pick.

The outlier screen and the walk from all articles

The main specification screens the FRED-MD series for outliers (see Hard data and targets in the data section). Table 17 compares the main specification with and without the screen. The best hard-data model against the AR moves by at most 0.04 in any cell of the table, and the same-method comparison of hard data plus news against hard data gives 15 significant improvements and 2 deteriorations with the screen and 12 and 3 without, with the same median ratio of 0.98 and the same share of 66% below one. Without the screen the hard-data LASSO and Elastic Net nowcasts of consumption for 2008Q2 come out at about $+16\%$ against an actual of $+0.3\%$; those cells are excluded from the unscreened summary, hence 141 comparisons rather than 144. The screen changes no conclusion.

Table 17: With and without the outlier screen on the hard data.

	Hard only vs. AR		Hard + news vs. hard, same model			
Hard data	GDP	Consumption	Comparisons	Median	Below one	Sig. <1 / >1
Not screened	0.94 / 1.00 / 1.01 / 1.01	0.89 / 0.91 / 0.95 / 0.98	141	0.98	66%	12 / 3
Outliers screened	0.95 / 1.00 / 1.01 / 1.00	0.89 / 0.95 / 0.93 / 0.99	144	0.98	66%	15 / 2

Note: Main specification with and without the McCracken–Ng outlier screen. Left block: best hard-data model over the AR benchmark, H0 / H1 / H2 / H3, mean over the three months. Right block: the same method with and without news, counted with the test of the design section at the 10% level. Without the screen the outlier-driven hard-data cells for 2008Q2 are excluded, hence 141 comparisons.

Six specifications lead from the first one estimated on this corpus to the main specification, one ingredient at a time (Table 18). The first row uses all articles of the seven papers, the topics fitted on them, a 17-series subset of FRED-MD and the three methods of Ellingsen et al. (2022). The second row keeps only the economic articles, the third re-fits the topics on them, the fourth returns to all articles but uses the full 101-series FRED-MD without the screen, the fifth is the three-method specification with the screen, and the sixth is the main specification. Every row uses the seven papers, training from 1997Q1 and evaluation over 2003Q1–2020Q1.

Two steps matter. Selecting the economic articles and re-fitting the topics on them moves investment from 1.21, 1.13 and 1.20 against the AR to 1.13, 1.11 and 1.16 and GDP from 1.03, 1.01 and 1.00 to 1.01, 0.99 and 0.98, and turns the same-method count from 1 improvement and 13 deteriorations into 4 and 7. Moving from the three methods and the topics to the four methods and all text measures brings GDP to 0.94, 0.96 and 0.95 and the same-method count to 11 improvements and 2 deteriorations at H0–H2. Widening the hard data from 17 to 101 series leaves the news-only ratios unchanged, as it must, and moves the best-against-best comparison by at most 0.03.

Table 18: From all articles to the main specification, one ingredient at a time.

	News only vs. AR, H0 / H1 / H2			Hard + news vs. hard, same model			News vs. hard,
Specification	Consumption	GDP	Investment	Median	Below one	Sig. $< 1 / > 1$	best vs. best
7 papers, all articles, 17 hard series	0.90 / 0.96 / 0.89	1.03 / 1.01 / 1.00	1.21 / 1.13 / 1.20	0.99	54%	1 / 13	1.12 / 1.04 / 1.00
+ economic articles	0.94 / 0.98 / 0.90	1.04 / 1.00 / 0.99	1.17 / 1.09 / 1.08	0.99	57%	5 / 5	1.12 / 1.03 / 0.97
+ topics re-fitted on them	0.91 / 0.96 / 0.89	1.01 / 0.99 / 0.98	1.13 / 1.11 / 1.16	0.99	53%	4 / 7	1.09 / 1.02 / 0.98
All articles, full FRED-MD	0.90 / 0.96 / 0.89	1.03 / 1.01 / 1.00	1.21 / 1.13 / 1.20	1.00	50%	6 / 10	1.10 / 1.05 / 1.03
Main spec, paper's three models	0.91 / 0.96 / 0.89	1.01 / 0.99 / 0.98	1.13 / 1.11 / 1.16	0.99	59%	9 / 8	1.08 / 1.03 / 1.01
Main spec, four models, all text	0.87 / 0.95 / 0.88	0.94 / 0.96 / 0.96	1.04 / 1.03 / 1.08	0.98	64%	11 / 2	1.02 / 1.00 / 0.98

Note: Rows 1–5: the three methods of Ellingsen et al. (2022), topics only; row 6: four methods, all text measures. All rows: seven papers, training from 1997Q1, evaluation 2003Q1–2020Q1. Left block: best news-only model over the AR benchmark, H0 / H1 / H2, mean over the three months. Middle block: the same method with and without news; the median ratio, the share below one, and the comparisons significantly below / above one at the 10% level (81 in rows 1–5, 144 in row 6). Right: best news-only over best hard-data model, H0 / H1 / H2, mean over outcomes and months. The full tables behind each row are Tables A.22 to A.29.

Which actuals

Every table so far scores the forecasts against the growth rates in the vintage of 26 August 2026. The same forecasts are also scored against the vintage that contains 2020Q1, the choice closest to that of Ellingsen et al. (2022), who evaluate against the data available at the end of their sample, and against the first releases, the numbers a forecaster would have been judged on at the time. The three are re-scores of the same forecasts, so the models, the training samples and the benchmarks' real-time histories are identical across them; only the realised values in the denominator of the ratio and in the loss differential change.

The vintage containing 2020Q1 changes little in the two summaries of the main specification (Table 19): no full-sample ratio of the best news-only model moves by more than 0.02, the model is significantly better than the AR in 4 of 36 cells and worse in 1 instead of 2 and 1, and the same-method comparison gives 13 improvements and 3 deteriorations in 144 instead of 15 and 2, with a median ratio of 0.97 against 0.98. First releases move the news-only ratios towards one. GDP at the nowcast goes from 0.94 to 0.99, consumption from 0.87, 0.95 and 0.88 to 0.94, 0.98 and 0.93, and no news-only cell differs significantly from the AR in either direction; GDP one and two quarters ahead stays at 0.95, and investment stays at or above one, at 1.03, 1.00 and 1.05. The same-method comparison does not move: against first releases text added to the hard data gives 17 significant improvements and 3 deteriorations in 144, with a median ratio of 0.97. The full-sample gain of text over the autoregression is therefore smaller when the forecasts are judged against the first published numbers, most of all at the nowcast, while the gain of text added to the hard data with the method held fixed is the same under every set of actuals.

Forecast combination

Ellingsen et al. (2022) also combine their forecasts and read the weight on the news-based

Table 19: Three sets of actuals: the main specification scored against the vintage of 26 August 2026, the vintage containing 2020Q1, and the first releases.

	Best news-only model vs AR, H0 / H1 / H2			News only	Hard + news vs hard	
Actuals	GDP	Consumption	Investment	sig. + / − of 36	sig. + / − of 144	median
Vintage of 26 August 2026	0.94 / 0.96 / 0.95	0.87 / 0.95 / 0.88	1.04 / 1.03 / 1.08	2 / 1	15 / 2	0.98
Vintage containing 2020Q1	0.93 / 0.95 / 0.95	0.87 / 0.96 / 0.89	1.05 / 1.04 / 1.07	4 / 1	13 / 3	0.97
First releases	0.99 / 0.95 / 0.95	0.94 / 0.98 / 0.93	1.03 / 1.00 / 1.05	0 / 0	17 / 3	0.97

Note: Main specification, evaluation 2003Q1–2020Q1; the forecasts are identical across the rows, only the realised values change. Left block: RMSE of the best news-only model over the RMSE of the AR benchmark, H0 / H1 / H2, mean over the three months. News only: cells significantly better / worse than the AR at the 10% level, H0–H3 and three months, of 36. Hard + news vs hard: the same method with and without news, significantly better / worse at the 10% level, of 144, and the median ratio. The first row is the convention of every other table.

members as a measure of their information content. I repeat this on the main specification. The members are the news-only and the hard-data forecasts of each of the four methods, eight in all. At each origin the weights are chosen by least squares on the past forecast errors, constrained to be non-negative and to sum to one, on an expanding window; 24 quarters of forecasts are used for the first weights, so the combination is evaluated from 2009Q1 to 2020Q1, after the Great Recession. The mean weight on the news members for GDP, averaged over the three months, is 0.40 at the nowcast, 0.86 one quarter ahead and 1.00 and 0.99 two and three quarters ahead; for consumption 0.42, 0.56, 0.94 and 0.22; for investment 0.34, 0.73, 0.65 and 0.87. The weight is not the same as a gain. The optimal combination of all eight members against the optimal combination of the four hard-data members is at 1.00, 0.95, 0.92 and 0.96 for GDP at H0–H3, 0.97, 1.00, 0.95 and 0.97 for consumption, and 1.10, 1.09, 1.06 and 1.13 for investment. The all-member combination is below the hard-only one in 18 of 36 cells, in none of them significantly, and significantly above it in 2, both investment three quarters ahead. After 2009 admitting the news members to the combination does not lower the error significantly anywhere.

Conclusion

This paper asked whether newspaper text helps to forecast U.S. quarterly growth in real time, on a corpus of seven print newspapers, and when over the sample any help arrives. The design follows Ellingsen et al. (2022), so that the results can be set beside theirs: six monthly lags of every text and hard-data series in a regression estimated with machine-learning methods on expanding real-time windows, scored against an autoregression. The paper's own parts are a broad cross-paper corpus, several text measures side by side, an automatic filter for the economic articles checked against hand labels, and an evaluation by period with a test that allows for multi-step forecast errors.

First, on all articles of the seven papers, text alone is behind the hard data for the GDP nowcast and for investment, and adding it to the hard data with the method held fixed improves the

consumption forecast and little else.

Second, which articles are used matters. The same method on economic articles does better than on all articles, significantly in 27 of 144 comparisons and worse in 4, in normal times as well as in the recession, and the improvement survives relabelling, tightening and loosening of the selection rule. The filter brings the text forecasts from behind the benchmark to level with it. It does not, by itself, make them beat it outside the recession. On the economic articles, text added to the hard data with the method held fixed gives a small gain for consumption, inside and outside the recession, and no gain for GDP outside it. Investment is the exception throughout: text alone is far behind the benchmark in normal times and, added to the hard data, leaves the forecast where it is or makes it worse.

Third, text was a turning-point signal in the Great Recession. In 2008–2009 the best news-only model beat the autoregressive benchmark by a wide margin for GDP and consumption, and one quarter ahead and beyond it beat the benchmark by more than the hard data did. That episode holds between a third and a half of the benchmark's squared error over the whole sample for GDP and consumption, and more than half for investment, so the full-sample ratios, which reproduce those of Ellingsen et al. (2022) for GDP and consumption, are a statement about it. Outside it, in the years before 2008 and the decade after 2009, neither text nor the hard data beat a simple autoregression of growth, and the 2001 recession on a two-paper sample shows no gain either.

For the use of newspaper text in real-time forecasting the lesson is a conditional one. The information is in the text when the economy turns sharply, and in those quarters it is visible before it is visible in the monthly hard data. In the long stretches between such quarters the text series are mostly noise for this purpose, and no choice of articles, method or series construction tried here changes that. The recursive combination of the robustness section, which can put any weight on text, puts most of it on text beyond the nowcast after 2009 and gains nothing in accuracy by doing so; the weight follows the recession fit, not a gain in normal times.

Two things would sharpen the answer. A sample with a second deep recession would tell whether 2008–2009 is the pattern or the exception, and a time-stamped language model behind the filter would remove the one component of the design whose knowledge of later text could matter.

References

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Ardia, D., K. Bluteau, and K. Boudt (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting 35*(4), 1370–1386.

Babii, A., E. Ghysels, and J. Striaukas (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics 40*(3), 1094–1106.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics 131*(4), 1593–1636.

Bańbura, M., D. Giannone, and L. Reichlin (2011). Nowcasting. In *The Oxford Handbook of Economic Forecasting*. New York: Oxford University Press.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica 77*(3), 623–685.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (2024). Business news and business cycles. *Journal of Finance 79*(5), 3105–3147.

Carroll, C. D. (2003). Macroeconomic expectations of households and professional forecasters. *Quarterly Journal of Economics 118*(1), 269–298.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pp. 288–296. Cambridge, MA: MIT Press.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*(1), 291–311.

Clements, M. P. and A. B. Galvão (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. *Journal of Business & Economic Statistics 26*(4), 546–554.

Croushore, D. (1993). Introducing: The survey of professional forecasters. *Federal Reserve Bank*

of Philadelphia Business Review (November/December), 3–15.

Croushore, D. and T. Stark (2001). A real-time data set for macroeconomists. *Journal of Econometrics 105*(1), 111–130.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263.

Ellingsen, J., V. H. Larsen, and L. A. Thorsrud (2022). News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics 37*(1), 63–81.

Foroni, C., M. Marcellino, and C. Schumacher (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A 178*(1), 57–82.

García, D. (2013). Sentiment during recessions. *Journal of Finance 68*(3), 1267–1300.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? evidence from U.S. daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2004). The MIDAS touch: Mixed data sampling regression models. Technical report, CIRANO Working Papers.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*(Suppl. 1), 5228–5235.

Groseclose, T. and J. Milyo (2005). A measure of media bias. *Quarterly Journal of Economics 120*(4), 1191–1237.

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics 133*(2), 801–870.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 13*(2), 281–291.

Huang, A. H., H. Wang, and Y. Yang (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research 40*(2), 806–841.

Kalamara, E., A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics 37*(5), 896–919.

Kuzin, V., M. Marcellino, and C. Schumacher (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting 27*(2), 529–542.

Larsen, V. H. and L. A. Thorsrud (2019). The value of news for economic developments. *Journal of Econometrics 210*(1), 203–218.

Larsen, V. H., L. A. Thorsrud, and J. Zhulanova (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics 117*, 507–520.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *Journal of Finance 66*(1), 35–65.

Martin, G. J. and A. Yurukoglu (2017). Bias in cable news: Persuasion and polarization. *American Economic Review 107*(9), 2565–2599.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Medeiros, M. C., G. F. R. Vasconcelos, A. Veiga, and E. Zilberman (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics 39*(1), 98–119.

Mullainathan, S. and A. Shleifer (2005). The market for news. *American Economic Review 95*(4), 1031–1053.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

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

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Reimers, N. and I. Gurevych (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992.

Sarkar, S. K. and K. Vafa (2024). Lookahead bias in pretrained language models. Working paper.

Shapiro, A. H., M. Sudhof, and D. J. Wilson (2022). Measuring news sentiment. *Journal of Econometrics 228*(2), 221–243.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics 50*(3), 665–690.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance 62*(3), 1139–1168.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics 38*(2), 393–409.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58*(1), 267–288.

Timmermann, A. (2006). Forecast combinations. In *Handbook of Economic Forecasting*, Volume 1, pp. 135–196. Amsterdam: Elsevier.

Ulbricht, D., K. A. Kholodilin, and T. Thomas (2017). Do media data help to predict German industrial production? *Journal of Forecasting 36*(5), 483–496.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Wang, W., F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33*.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67*(2), 301–320.

Appendix A: Additional Tables

This appendix holds the full tables behind the summaries of the Results and Robustness sections: every horizon and every month, every method, and the three benchmarks. The conventions are those of the subsection Evaluation of the design section. H0 is the nowcast and H1–H3 the forecasts one to three quarters ahead; M1–M3 is the month of the quarter in which the forecast is made; AR is the direct autoregression, RW the constant-growth benchmark and SPF the Survey of Professional Forecasters. Stars mark a Diebold–Mariano test with Newey–West variance (lag h) and the small-sample correction, at the 10 / 5 / 1% levels (one, two and three stars). The best model in a predictor set is chosen ex post per outcome, horizon and month. Sub-samples are defined on the quarter being forecast, and the best model is the full-sample pick.

Main specification: four methods, all text measures

Table A.1 counts the significant cells of the main specification against the three benchmarks; Tables A.2 to A.4 give the cells. Tables A.5 and A.6 are the two same-method comparisons for every method and month, and Table A.7 the second one by period. Tables A.8 and A.9 are the comparison with the AR benchmark by period, horizon by horizon. Figures A.1 to A.3 show the cumulative squared-error gains over time.

Table A.1: Main specification: cells significantly below / above one, by predictor set and benchmark.

	Cells significantly below / above one, against		
Best model	AR	Constant growth	SPF
News only	2 / 1	4 / 2	0 / 30
Hard data only	0 / 0	4 / 1	0 / 31
Hard data + news	4 / 2	6 / 2	0 / 33

Note: Best model per predictor set, chosen ex post; 36 cells (3 outcomes × H0–H3 × M1–M3), counted at the 10% level. About 2 cells are expected in each direction by chance.

Table A.2: Main specification: best news-only model against the three benchmarks.

Outcome	Benchmark	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	0.97	0.95	0.91	0.98	0.97	0.93	0.96	0.95	0.96	0.97	0.98	0.95
	RW	0.88	0.86	0.83*	0.94	0.94	0.90	0.95	0.94	0.95	0.98	0.98	0.96
	SPF	1.23**	1.20**	1.16*	1.13*	1.12*	1.07	1.06*	1.05	1.06*	1.03*	1.04	1.01
Consumption	AR	0.89	0.87*	0.84**	0.97	0.96	0.93	0.90	0.88	0.87	1.01	1.00	0.97
	RW	0.86*	0.84**	0.81**	0.92	0.91	0.88	0.93	0.91	0.90	0.99	0.98	0.95
	SPF	1.21***	1.19***	1.14**	1.16***	1.15***	1.11***	1.12**	1.10**	1.08**	1.15	1.14	1.10*
Investment	AR	1.05	1.03	1.03	1.07	1.04	0.99	1.09	1.08	1.08	1.08	1.07	1.13*
	RW	0.91	0.89	0.89	1.02	0.99	0.94	1.07	1.05	1.05	1.07	1.07*	1.12*
	SPF	1.37***	1.35***	1.35***	1.31**	1.27**	1.21**	1.23***	1.22***	1.22*	1.17***	1.17***	1.23**

Note: RMSE of the best news-only model over the RMSE of the benchmark. Stars: Diebold–Mariano test, Newey–West variance (lag h), small-sample correction; 10 / 5 / 1%.

Table A.3: Main specification: best hard-data model against the three benchmarks.

Outcome	Benchmark	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	0.95	0.97	0.92	1.01	0.99	1.00	1.02	1.01	1.00	1.01	1.01	0.99
	RW	0.87*	0.89	0.84	0.97	0.96	0.97	1.01	1.01	0.99	1.01	1.01	0.99
	SPF	1.21	1.24**	1.17**	1.16*	1.14	1.15*	1.12	1.12	1.10*	1.07**	1.07**	1.05*
Consumption	AR	0.89	0.91	0.86	0.96	0.95	0.94	0.96	0.92	0.90	0.99	0.99	1.00
	RW	0.86	0.87	0.83*	0.91	0.89	0.89*	0.99	0.95	0.93	0.97	0.98	0.99
	SPF	1.22**	1.24***	1.17**	1.15**	1.13*	1.13*	1.19*	1.14*	1.13*	1.13**	1.13**	1.14**
Investment	AR	0.99	0.96	0.92	1.02	1.00	0.98	1.08	1.05	1.00	1.09	1.09	1.07
	RW	0.85	0.83*	0.80	0.97	0.96	0.93	1.06*	1.02	0.98	1.08	1.08	1.06
	SPF	1.29**	1.26*	1.21	1.24**	1.22*	1.20*	1.22*	1.18*	1.13**	1.19*	1.18**	1.16*

Note: RMSE of the best hard-data model (101 FRED-MD series, outliers screened) over the RMSE of the benchmark. Stars as in Table A.2.

Table A.4: Main specification: best hard-data-plus-news model against the three benchmarks.

Outcome	Benchmark	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	0.96	0.95	0.93	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.99	0.96
	RW	0.87	0.86	0.85	0.94	0.94	0.93	0.97	0.96	0.95	0.97	0.99	0.96
	SPF	1.21^{**}	1.20^{**}	1.18^{**}	1.12^{*}	1.12^{*}	1.10	1.08^{*}	1.07^{*}	1.06^{*}	1.03^{*}	1.05^{*}	1.02
Consumption	AR	0.87^{*}	0.86^{**}	0.83^{**}	0.94	0.93	0.91^{*}	0.90	0.89	0.88	1.00	0.99	0.99
	RW	0.84^{**}	0.83^{**}	0.80^{**}	0.88	0.88^{*}	0.86^{*}	0.93	0.92	0.91	0.99	0.97	0.98
	SPF	1.19^{***}	1.17^{***}	1.13^{***}	1.12^{***}	1.11^{***}	1.09^{**}	1.12^{**}	1.11^{**}	1.10^{**}	1.14^{*}	1.13^{*}	1.13^{*}
Investment	AR	1.02	1.00	0.95	1.02	1.02	0.99	1.06	1.08^{*}	1.05	1.07	1.06	1.10^{*}
	RW	0.88	0.86	0.82^{*}	0.98	0.98	0.95	1.04	1.06	1.03	1.06	1.05^{*}	1.09^{**}
	SPF	1.33^{**}	1.31^{**}	1.25^{**}	1.25^{**}	1.25^{**}	1.22^{**}	1.20	1.23^{**}	1.19^{***}	1.16^{***}	1.15^{***}	1.20^{***}

Note: RMSE of the best hard-data-plus-news model over the RMSE of the benchmark. Stars as in Table A.2.

Table A.5: Main specification: news only against hard data only, same method.

Outcome	Model	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	Elastic Net	1.05	0.99	1.14*	1.13*	1.07	1.00	1.10*	1.09*	1.08	0.86	0.88	1.05
	PCA–OLS	0.99	0.97	0.99	0.92	0.94	0.90	0.92	0.93	0.92	0.89**	0.90	0.92
	Random forest	0.99	0.99	1.00	0.97	0.96	0.98	0.96	0.94	0.96	0.96	0.97	0.96
	XGBoost	1.05	0.99	1.02	0.96	0.99	0.99	0.83*	0.85	0.95	0.90	0.90	0.89
	Best vs. best	1.02	0.97	0.99	0.97	0.99	0.93	0.94	0.93	0.96	0.96	0.97	0.96
Consumption	Elastic Net	1.09	0.97	0.92	1.01	1.02	0.94	0.86	0.88	1.07	0.88	0.92	0.88
	PCA–OLS	1.00	0.96	0.97	1.00	1.02	0.98	0.95	0.99	0.96	0.92	0.91	0.90
	Random forest	1.01	1.00	0.96	1.02	0.99	0.98	0.95	0.92*	0.94*	1.02	1.00	0.99
	XGBoost	1.04	1.04	0.98	1.02	1.00	0.97	0.88**	0.86**	0.92**	1.02	0.95	0.95
	Best vs. best	1.00	0.96	0.97	1.01	1.02	0.98	0.94	0.96	0.96	1.02	1.00	0.97
Investment	Elastic Net	1.22	1.12	1.30*	0.90	0.94	1.02	1.14	1.07	1.14	1.46	1.06	1.14
	PCA–OLS	1.05	1.07	1.08	0.91	0.89	0.86	0.95	0.90	0.89	0.97	0.92	0.96
	Random forest	1.14**	1.16**	1.18**	1.11	1.09	1.11	1.03	1.08*	1.13*	0.99	0.99	1.05
	XGBoost	1.03	1.10	1.30**	0.93	1.21	1.24*	0.76	0.74	1.03	0.93	0.92	1.10
	Best vs. best	1.06	1.08	1.11	1.05	1.04	1.01	1.01	1.03	1.08	0.99	0.99	1.05

Note: RMSE of the news-only forecast over the RMSE of the hard-data forecast for the same method; "best vs. best" is the ex post best model on each side. Stars as in Table A.2.

Table A.6: Main specification: hard data plus news against hard data only, same method.

Outcome	Model	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	Elastic Net	1.08**	0.98	1.03	1.04	1.04	0.96	1.01	1.02	1.04	0.90	1.03	1.03
	PCA–OLS	0.97	0.97	1.01	0.91	0.94	0.93	0.90*	0.92	0.95	0.90***	0.93	0.96
	Random forest	0.99	0.99	0.99	0.97	0.97	0.98	0.96	0.95	0.96	0.96	0.98	0.97
	XGBoost	1.06*	1.00	1.03	0.96	1.00	1.00	0.90*	0.88*	0.94	0.90	0.96	0.91
Consumption	Elastic Net	0.99	0.96	0.89	0.91	1.02	0.90	0.87**	0.90	0.96	1.05	1.00	0.98
	PCA–OLS	0.98	0.94	0.97	0.97	0.98	0.97	0.93**	0.97	0.97	0.92**	0.92*	0.93
	Random forest	1.01	1.00	0.97	1.00	0.99	0.98	0.95*	0.94**	0.96*	1.01	0.99	0.99
	XGBoost	1.04	1.02	0.99	1.01	1.00	0.96	0.93*	0.93*	0.94*	1.04	0.98	0.97
Investment	Elastic Net	1.12	0.98	1.16	0.98	0.89	1.02	1.29	1.01	1.10	1.39	0.92	1.10
	PCA–OLS	1.02	1.08	1.07	0.87	0.88	0.90	0.86	0.87	0.91	0.91**	0.91	0.98
	Random forest	1.03	1.04	1.02	1.06	1.07	1.02	1.01	1.04	1.05	0.98	0.97	1.03
	XGBoost	0.94	1.06	0.99	0.95	1.05	1.02	0.82	0.82	0.95	0.96	0.90	1.00

Note: RMSE of the hard-data-plus-news forecast over the RMSE of the hard-data forecast for the same method; below one: text helps. Stars as in Table A.2.

Table A.7: Main specification: does text add to the hard data? By period.

Evaluation sample	Quarters	All	GDP	Consumption	Investment
Full sample	69	0.98	0.97	0.97	1.00
Without 2008Q3–2009Q2	65	1.00	1.00	0.97	1.04
Through 2007	20	1.00	0.98	0.98	1.03
2008–2009	8	0.95	0.95	0.96	0.94
2010–2019	40	1.02	0.99	0.97	1.11

Note: The same method with and without text; ratio below one: text helps. Mean over the four methods, the four horizons and the three months. Samples are defined on the quarter being forecast.

Table A.8: Main specification: against the AR benchmark by period, every horizon.

Outcome	Best model	Full sample (69 q.)	Without 2008Q3–2009Q2 (65 q.)	2008–2009 (8 q.)
GDP	News only	0.94 / 0.96 / 0.95 / 0.97	1.03 / 1.04 / 1.06 / 1.01	0.88 / 0.90 / 0.87 / 0.93
	Hard data only	0.95 / 1.00 / 1.01 / 1.00	1.01 / 1.03 / 1.03 / 1.03	0.90 / 0.99 / 1.01 / 0.98
	Hard data + news	0.94 / 0.97 / 0.97 / 0.97	1.02 / 1.09 / 1.03 / 1.01	0.89 / 0.92 / 0.92 / 0.94
Consumption	News only	0.87 / 0.95 / 0.88 / 0.99	0.96 / 1.02 / 0.97 / 1.00	0.65 / 0.80 / 0.76 / 0.99
	Hard data only	0.89 / 0.95 / 0.93 / 0.99	0.99 / 1.00 / 1.00 / 1.02	0.74 / 0.92 / 0.87 / 0.95
	Hard data + news	0.85 / 0.93 / 0.89 / 0.99	0.94 / 1.00 / 1.00 / 0.99	0.67 / 0.81 / 0.77 / 1.00
Investment	News only	1.04 / 1.03 / 1.08 / 1.09	1.30 / 1.31 / 1.27 / 1.23	0.77 / 0.83 / 0.94 / 0.99
	Hard data only	0.96 / 1.00 / 1.04 / 1.08	1.00 / 1.09 / 1.08 / 1.08	0.96 / 0.97 / 1.03 / 1.09
	Hard data + news	0.99 / 1.01 / 1.06 / 1.08	1.08 / 1.22 / 1.22 / 1.16	0.92 / 0.94 / 0.99 / 1.02

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2 / H3. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast. Eight quarters are too few for the test.

Table A.9: Main specification: against the AR benchmark by evaluation sample, five samples.

Outcome	Best model	Full sample (69 q.)	Without 2008Q3–2009Q2 (65 q.)	2003–2007 (20 q.)	2008–2009 (8 q.)	2010–2019 (40 q.)
GDP	News only	0.94 / 0.96 / 0.95 / 0.97	1.03 / 1.04 / 1.06 / 1.01	1.02 / 1.04 / 1.11 / 0.98	0.88 / 0.90 / 0.87 / 0.93	1.02 / 1.08 / 1.06 / 1.05
	Hard data only	0.95 / 1.00 / 1.01 / 1.00	1.01 / 1.03 / 1.03 / 1.03	1.02 / 1.04 / 1.06 / 1.07	0.90 / 0.99 / 1.01 / 0.98	1.01 / 1.06 / 1.02 / 1.07
	Hard data + news	0.94 / 0.97 / 0.97 / 0.97	1.02 / 1.09 / 1.03 / 1.01	1.02 / 1.13 / 1.05 / 1.00	0.89 / 0.92 / 0.92 / 0.94	1.02 / 1.03 / 1.04 / 1.04
Consumption	News only	0.87 / 0.95 / 0.88 / 0.99	0.96 / 1.02 / 0.97 / 1.00	1.10 / 1.13 / 1.09 / 1.09	0.65 / 0.80 / 0.76 / 0.99	0.99 / 1.06 / 0.94 / 0.92
	Hard data only	0.89 / 0.95 / 0.93 / 0.99	0.99 / 1.00 / 1.00 / 1.02	1.09 / 1.06 / 1.05 / 1.12	0.74 / 0.92 / 0.87 / 0.95	0.87 / 0.93 / 0.98 / 1.01
	Hard data + news	0.85 / 0.93 / 0.89 / 0.99	0.94 / 1.00 / 1.00 / 0.99	1.08 / 1.06 / 1.02 / 1.14	0.67 / 0.81 / 0.77 / 1.00	0.90 / 0.99 / 0.98 / 0.88
Investment	News only	1.04 / 1.03 / 1.08 / 1.09	1.30 / 1.31 / 1.27 / 1.23	1.20 / 1.11 / 1.18 / 0.94	0.77 / 0.83 / 0.94 / 0.99	1.32 / 1.37 / 1.32 / 1.38
	Hard data only	0.96 / 1.00 / 1.04 / 1.08	1.00 / 1.09 / 1.08 / 1.08	1.09 / 0.94 / 0.95 / 0.87	0.96 / 0.97 / 1.03 / 1.09	0.88 / 1.10 / 1.11 / 1.16
	Hard data + news	0.99 / 1.01 / 1.06 / 1.08	1.08 / 1.22 / 1.22 / 1.16	1.11 / 1.16 / 1.00 / 0.91	0.92 / 0.94 / 0.99 / 1.02	1.05 / 1.12 / 1.28 / 1.29

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2 / H3. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast.

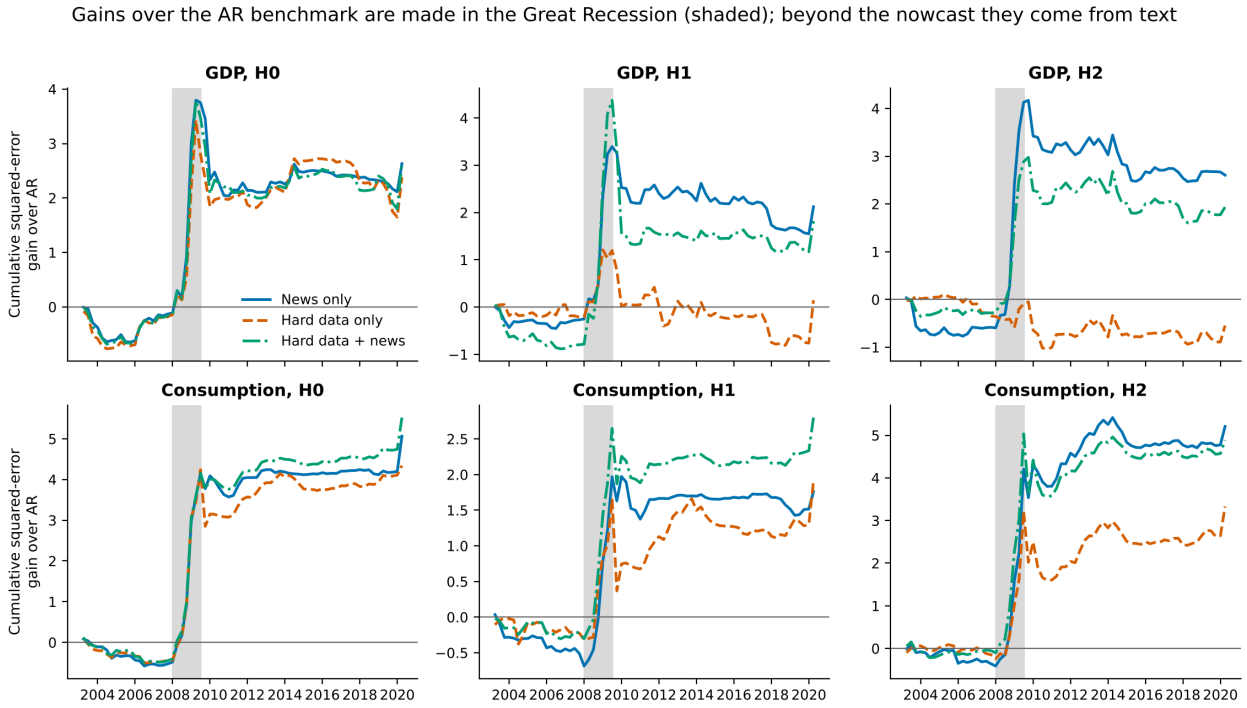


Figure A.1: Main specification: cumulative squared-error gain over the AR benchmark, nowcast to two quarters ahead. Cumulative sum of the AR benchmark's squared error minus the model's, best model per predictor set, mean over the three months; shaded: 2007Q4–2009Q2.

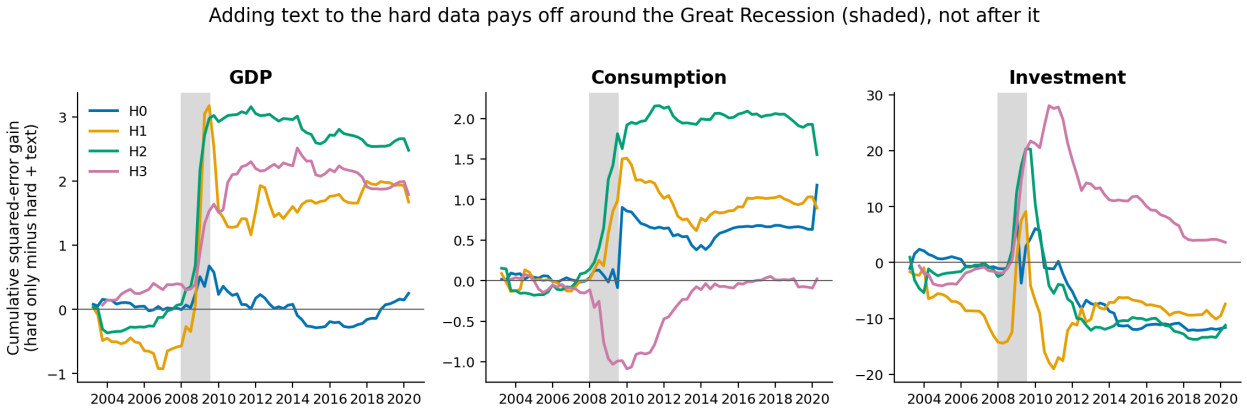


Figure A.2: Main specification: hard data plus text against hard data only over time. Cumulative sum of the hard-data model's squared error minus the hard-data-plus-text model's, best model on each side, mean over the three months.

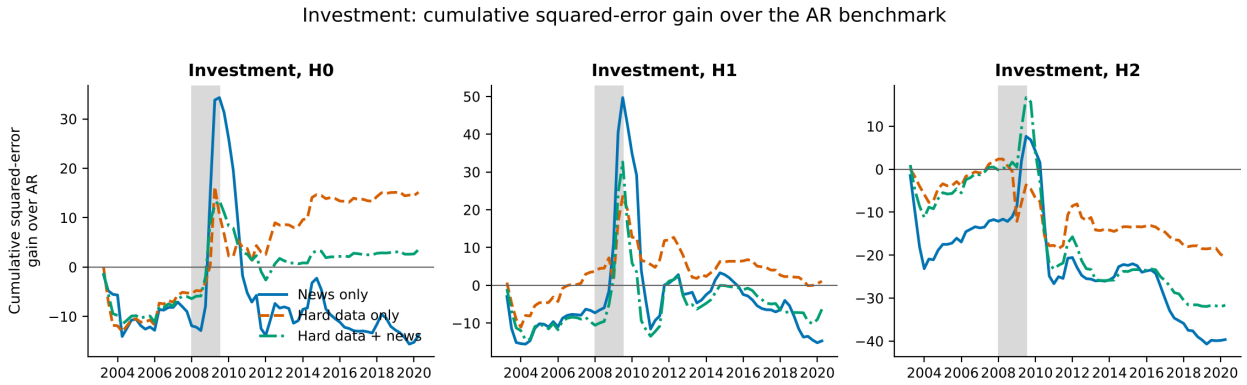


Figure A.3: Main specification: investment against the AR benchmark over time. Cumulative sum of the AR benchmark's squared error minus the model's, best model per predictor set, mean over the three months; shaded: 2007Q4–2009Q2. Outside the recession the best news-only model is at 1.23–1.31 of the benchmark.

The three methods of Ellingsen, Larsen and Thorsrud, topics only

The tables of this subsection use the LASSO, PCA–OLS with one factor and a random forest of 500 trees, with the 160 topic series as the news set, on the main specification's articles, hard data and window (the robustness subsection on the methods of Ellingsen, Larsen and Thorsrud). Tables A.10 and A.11 are the comparison with the benchmarks, Table A.12 the comparison with the AR by period, Tables A.13 and A.14 the same-method increment by method and by period, Table A.15 text against hard data with the best model on each side, and Figure A.4 the cumulative gain over the AR.

Table A.10: Three methods, topics only: best news-only model against the three benchmarks.

Outcome	Benchmark	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	1.01	1.01	1.01	0.99	1.00	0.98	1.00	0.98	0.95
	RW	0.92	0.92	0.92	0.96	0.96	0.95	0.99	0.97	0.95
	SPF	1.28^{**}	1.28^{*}	1.29^{*}	1.15^{*}	1.15^{*}	1.13^{*}	1.11^{*}	1.09	1.06
Consumption	AR	0.91	0.92	0.91^{*}	0.96	0.96	0.95	0.89	0.89	0.89^{*}
	RW	0.88^{*}	0.89^{**}	0.88^{**}	0.90^{*}	0.90^{*}	0.90^{*}	0.92	0.92	0.92
	SPF	1.24^{***}	1.26^{**}	1.24^{**}	1.15^{**}	1.15^{*}	1.14^{*}	1.11^{*}	1.10^{*}	1.11
Investment	AR	1.14^{***}	1.15^{***}	1.11^{*}	1.10	1.13^{*}	1.10^{**}	1.15	1.19^{*}	1.15^{*}
	RW	0.98	0.99	0.95	1.05	1.08	1.05^{*}	1.13	1.17^{*}	1.12
	SPF	1.49^{***}	1.50^{***}	1.45^{**}	1.34^{*}	1.39^{**}	1.34^{*}	1.31^{*}	1.35^{***}	1.30^{**}

Note: LASSO, PCA–OLS (one factor), random forest; topics only; main specification otherwise. Across the 27 cells the ratio to the AR is 0.89–0.96 for consumption, 0.95–1.01 for GDP and 1.10–1.19 for investment. Stars as in Table A.2.

Table A.11: Three methods, topics only: cells significantly below / above one, by predictor set and benchmark.

	Cells significantly below / above one, against		
Best model	AR	Constant growth	SPF
News only	2 / 7	6 / 2	0 / 24
Hard data only	5 / 1	8 / 1	0 / 21
Hard data + news	5 / 2	7 / 2	0 / 25

Note: Best model per predictor set, chosen ex post; 27 cells (3 outcomes $\times$ H0–H2 $\times$ M1–M3), counted at the 10% level.

Table A.12: Three methods, topics only: against the AR benchmark by period.

Outcome	Best model	Full sample (69 q.)	Without 2008Q3–2009Q2 (65 q.)	2008–2009 (8 q.)
GDP	News only	1.01 / 0.99 / 0.98	1.03 / 1.06 / 1.05	1.05 / 0.97 / 0.93
	Hard data only	0.99 / 1.00 / 1.01	0.99 / 1.01 / 1.04	1.02 / 1.01 / 1.01
	Hard data + news	1.00 / 0.99 / 0.99	1.00 / 1.03 / 1.05	1.03 / 0.98 / 0.96
Consumption	News only	0.92 / 0.96 / 0.89	0.95 / 0.95 / 0.92	0.85 / 0.96 / 0.82
	Hard data only	0.91 / 0.96 / 0.95	0.95 / 0.96 / 0.96	0.87 / 0.97 / 0.91
	Hard data + news	0.91 / 0.96 / 0.90	0.96 / 0.96 / 0.98	0.85 / 0.95 / 0.80
Investment	News only	1.13 / 1.11 / 1.16	1.19 / 1.17 / 1.34	1.09 / 1.07 / 1.03
	Hard data only	0.94 / 1.00 / 1.05	0.98 / 1.07 / 1.10	0.94 / 1.00 / 1.03
	Hard data + news	1.00 / 1.06 / 1.12	1.06 / 1.17 / 1.24	0.98 / 0.99 / 1.06

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast.

Table A.13: Three methods, topics only: does text add to the hard data? By method and outcome.

Model	H0	H1	H2	Sig. better	Sig. worse
LASSO	1.07	0.96	0.98	3 of 27	3 of 27
PCA–OLS	0.98	0.96	0.98	5 of 27	0 of 27
Random forest	1.02	1.01	1.00	1 of 27	5 of 27
All models	1.03	0.98	0.98	9 of 81	8 of 81
GDP	1.00	0.99	0.99	0 of 27	0 of 27
Consumption	1.02	0.96	0.94	7 of 27	2 of 27
Investment	1.05	0.98	1.02	2 of 27	6 of 27

Note: The same method with and without text; ratio below one: text helps. Mean over the three months and, in the lower blocks, over methods or outcomes. "Sig." counts the comparisons significant at the 10% level: 27 per method and per outcome, 81 in all. Stars as in Table A.2.

Table A.14: Three methods, topics only: does text add to the hard data? By period.

Evaluation sample	Quarters	All	GDP	Consumption	Investment
Full sample	69	1.00	1.00	0.97	1.02
Without 2008Q3–2009Q2	65	1.02	1.03	0.99	1.03
Through 2007	20	1.00	0.99	0.98	1.03
2008–2009	8	0.97	0.97	0.95	0.99
2010–2019	40	1.05	1.04	1.03	1.07

Note: The same method with and without text; mean over methods, horizons and months, by evaluation sample defined on the quarter being forecast.

Table A.15: Three methods, topics only: text against hard data, best model on each side.

	Text only vs. hard only			Hard + text vs. hard only		
Outcome	H0	H1	H2	H0	H1	H2
GDP	1.02	0.99	0.97	1.00	0.99	0.98
Consumption	1.01	0.99	0.94	1.00	0.99	0.95
Investment	1.20	1.11	1.11	1.06	1.05	1.07

Note: Best model on each side chosen ex post, per outcome, horizon and month; mean over the three months.

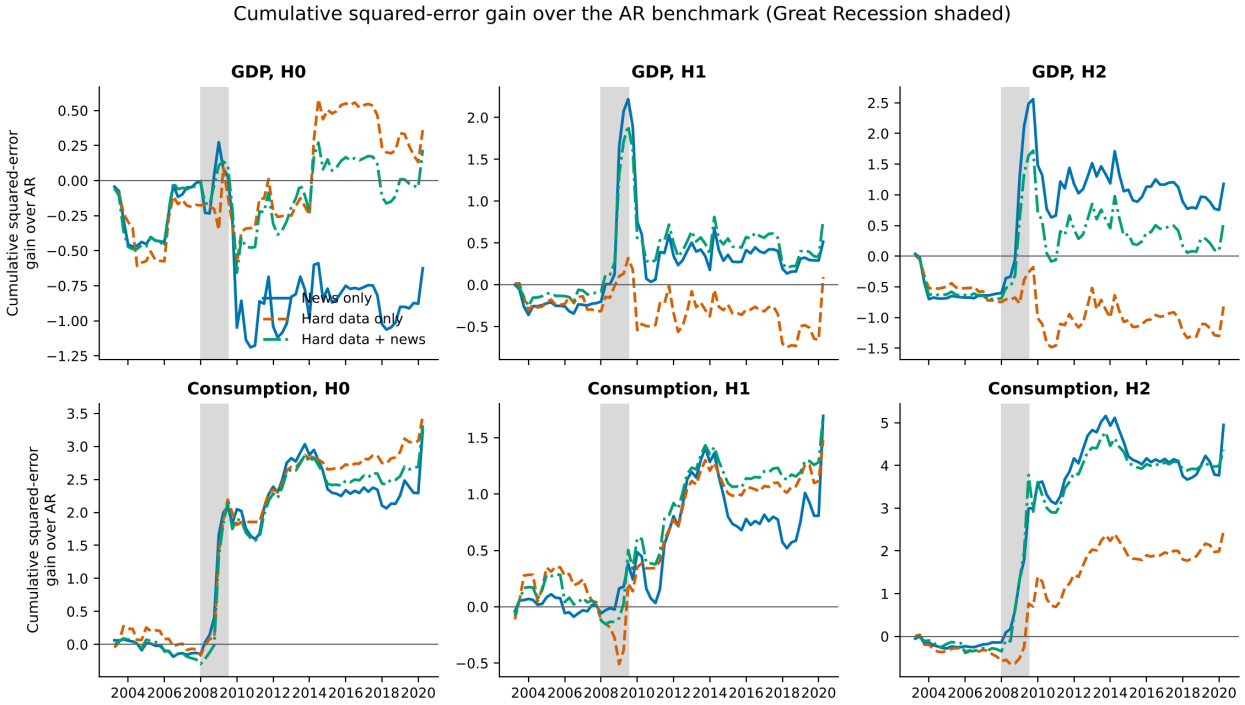


Figure A.4: Three methods, topics only: cumulative squared-error gain over the AR benchmark. Cumulative sum of the AR benchmark's squared error minus the model's, best model per predictor set, mean over the three months; shaded: 2007Q4–2009Q2.

Baseline: all articles

The tables of this subsection are the baseline of the results section: all articles of the seven papers, the four methods and all text measures, the 101 screened FRED-MD series, evaluation 2003Q1–2020Q1. Table A.16 is the comparison with the benchmarks, Tables A.17 and A.18 the two same-method comparisons, Table A.19 the comparison with the AR by period, and Figure A.5 the cumulative gain over the AR.

Table A.16: All articles: best news-only model against the three benchmarks.

Outcome	Benchmark	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	1.02	1.02	1.00	1.00	0.99	1.01	1.00	0.98	0.99	0.98	0.99	0.99
	RW	0.93	0.93	0.91	0.96	0.96	0.98	0.99	0.97	0.98	0.99	0.99	0.99
	SPF	1.30^{**}	1.29^{**}	1.27^{**}	1.15^{*}	1.15^{*}	1.16^{**}	1.11^{**}	1.08^{**}	1.10^{**}	1.05	1.05^{*}	1.05
Consumption	AR	0.88	0.91	0.85^{*}	0.97	0.98	0.97	0.90	0.90	0.89^{*}	0.93^{*}	0.92^{*}	0.93
	RW	0.85^{*}	0.87	0.82^{**}	0.92	0.92	0.91^{*}	0.93	0.94	0.92	0.91^{**}	0.91^{*}	0.92^{*}
	SPF	1.20^{**}	1.23^{***}	1.16^{**}	1.17^{**}	1.17^{**}	1.16^{**}	1.12	1.13^{*}	1.11^{*}	1.06	1.05	1.06
Investment	AR	1.25^{***}	1.24^{***}	1.20^{***}	1.19^{**}	1.16^{*}	1.15^{*}	1.17^{*}	1.16^{*}	1.19^{**}	1.19^{**}	1.19^{**}	1.14^{**}
	RW	1.08	1.07	1.04	1.13	1.11	1.10	1.14^{*}	1.14^{*}	1.17^{*}	1.18^{***}	1.18^{***}	1.13^{**}
	SPF	1.64^{***}	1.63^{***}	1.58^{**}	1.45^{***}	1.42^{***}	1.41^{***}	1.32^{***}	1.31^{***}	1.35^{***}	1.30^{***}	1.29^{***}	1.24^{**}

Note: RMSE of the best news-only model over the RMSE of the benchmark. Stars as in Table A.2.

Table A.17: All articles: news only against hard data only, same method.

Outcome	Model	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	Elastic Net	1.15***	1.11***	1.17***	1.05	1.04	0.96	0.95	0.93	1.00	0.82	0.86	0.97
	PCA–OLS	1.13**	1.11*	1.13*	1.00	1.03	1.00	0.95	0.96	0.98	0.86*	0.87	0.91
	Random forest	1.01	1.01	1.00	0.99	0.98	1.01	0.98	0.96	0.99	0.97	0.98	1.01
	XGBoost	1.09*	1.06	1.07	1.00	1.06	1.10	0.95	0.99	1.04	0.94	1.00	0.98
	Best vs. best	1.07	1.05	1.09	0.99	1.01	1.01	0.98	0.96	0.99	0.97	0.98	1.00
Consumption	Elastic Net	1.06	1.05	0.93	0.88	0.97	0.93	0.92	1.01	1.12	0.94	0.90	0.86
	PCA–OLS	1.09	1.04	1.04	1.05	1.06	1.05	0.95	0.99	0.99	0.87**	0.85	0.88
	Random forest	0.98	0.97	0.93*	1.01	0.98	0.97	0.94*	0.94	0.95	0.94**	0.93**	0.93*
	XGBoost	0.97	0.98	0.92	1.01	1.04	0.99	0.88	0.95	0.99	1.01	0.95	0.96
	Best vs. best	0.99	1.00	0.99	1.01	1.03	1.02	0.94	0.98	0.99	0.94**	0.93**	0.93*
Investment	Elastic Net	1.27**	1.18	1.30*	0.99	1.09	1.24*	1.13	1.07	1.26*	1.06	0.85	0.92
	PCA–OLS	1.37**	1.42**	1.37**	1.18	1.14	1.11	1.11	1.03	1.07	0.99	0.90	1.04
	Random forest	1.27***	1.29***	1.31***	1.17**	1.16***	1.18*	1.07	1.11*	1.19*	1.09	1.10*	1.11
	XGBoost	1.25	1.35**	1.43***	1.02	1.42	1.42*	0.91	1.00	1.17	1.07	1.09	1.22*
	Best vs. best	1.27***	1.29***	1.30*	1.17**	1.16***	1.18*	1.08	1.11*	1.19*	1.09	1.09	1.06

Note: RMSE of the news-only forecast over the RMSE of the hard-data forecast for the same method; "best vs. best" is the ex post best model on each side. Stars as in Table A.2.

Table A.18: All articles: hard data plus news against hard data only, same method.

Outcome	Model	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	Elastic Net	1.12**	1.04*	1.13*	1.01	1.04	0.99	0.97	1.01	1.05*	0.89	0.97	1.01
	PCA–OLS	1.00	1.02	1.04	0.97	1.01	0.97	0.93	1.00	1.00	0.90**	0.95	0.96
	Random forest	1.00	1.00	0.99	0.99	0.99	1.00	0.98	0.97	0.99	0.97	0.98	1.00
	XGBoost	1.05	0.98	1.00	1.02	1.03	1.00	0.96	0.94	0.98	0.96	1.00	0.96
Consumption	Elastic Net	0.97	0.99	0.93	0.93	0.92	0.95	1.03	1.07	1.17	1.08	1.06	1.03
	PCA–OLS	1.02	0.98	1.00	1.00	1.06	1.06	0.93**	1.00	1.01	0.91**	0.92*	0.93
	Random forest	0.99	0.99	0.96	1.01	0.99	0.98	0.95*	0.95	0.96*	0.95***	0.95***	0.94**
	XGBoost	0.98	0.98	0.99	0.99	0.98	0.97	0.92*	0.95	0.98	0.95*	0.92***	0.94**
Investment	Elastic Net	1.06	0.96	1.07	1.07	1.15	1.09	1.03	1.00	1.02	0.99	0.82	0.84
	PCA–OLS	1.10	1.19*	1.15	1.04	1.06	0.99	0.97	1.07	1.02	0.94	1.04	1.09
	Random forest	1.04	1.04	1.03	1.07*	1.08**	1.03	1.04	1.04	1.07	1.03	1.04	1.06
	XGBoost	0.96	1.01	0.94	0.95	1.03	1.00	0.88*	0.80	1.00	0.95	0.90	0.96

Note: RMSE of the hard-data-plus-news forecast over the RMSE of the hard-data forecast for the same method; below one: text helps. Stars as in Table A.2.

Table A.19: All articles: against the AR benchmark by period, every horizon.

Outcome	Best model	Full sample (69 q.)	Without 2008Q3–2009Q2 (65 q.)	2008–2009 (8 q.)
GDP	News only	1.01 / 1.00 / 0.99 / 0.99	1.05 / 1.06 / 1.06 / 1.03	1.00 / 0.97 / 0.94 / 0.95
	Hard data only	0.95 / 1.00 / 1.01 / 1.00	1.01 / 1.03 / 1.03 / 1.03	0.90 / 0.99 / 1.01 / 0.98
	Hard data + news	0.97 / 1.00 / 0.99 / 0.98	1.07 / 1.05 / 1.05 / 1.03	0.92 / 0.97 / 0.95 / 0.94
Consumption	News only	0.88 / 0.97 / 0.90 / 0.93	0.96 / 0.99 / 0.95 / 0.93	0.71 / 0.93 / 0.84 / 0.93
	Hard data only	0.89 / 0.95 / 0.93 / 0.99	0.99 / 1.00 / 1.00 / 1.02	0.74 / 0.92 / 0.87 / 0.95
	Hard data + news	0.88 / 0.96 / 0.91 / 0.94	0.97 / 0.98 / 0.96 / 0.95	0.72 / 0.94 / 0.84 / 0.92
Investment	News only	1.23 / 1.17 / 1.17 / 1.17	1.41 / 1.44 / 1.41 / 1.35	1.07 / 0.96 / 1.00 / 1.05
	Hard data only	0.96 / 1.00 / 1.04 / 1.08	1.00 / 1.09 / 1.08 / 1.08	0.96 / 0.97 / 1.03 / 1.09
	Hard data + news	0.98 / 1.06 / 1.09 / 1.10	1.09 / 1.23 / 1.21 / 1.17	0.92 / 0.95 / 1.01 / 1.06

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2 / H3. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast.

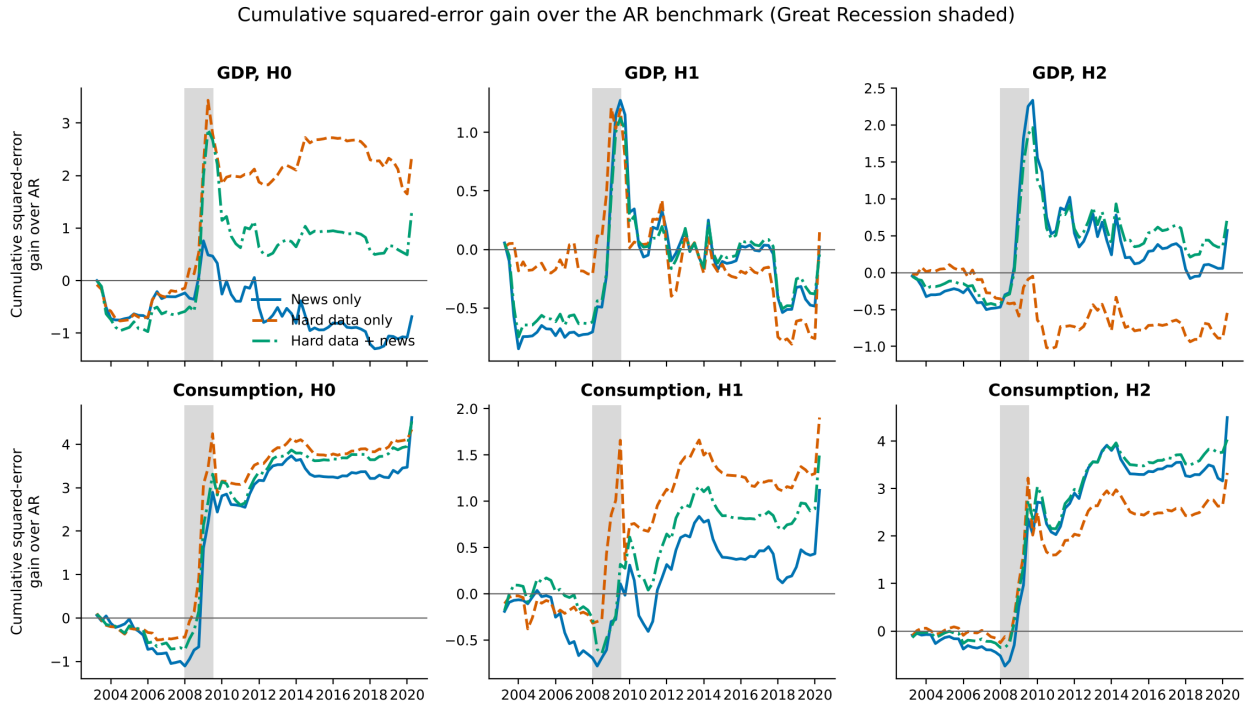


Figure A.5: All articles: cumulative squared-error gain over the AR benchmark. Cumulative sum of the AR benchmark's squared error minus the model's, best model per predictor set, mean over the three months; shaded: 2007Q4–2009Q2.

The filter in every specification

Table A.20: The filter effect in all four specifications: all articles against economic articles.

	News only vs. AR, H0 / H1 / H2			Hard + news vs. hard, same model			News vs. hard,
Specification	Consumption	GDP	Investment	Median	Below one	Sig. < 1 / > 1	best vs. best
Three models, all articles	0.90 / 0.96 / 0.89	1.03 / 1.01 / 1.00	1.21 / 1.13 / 1.20	1.00	52%	5 / 8	1.10 / 1.05 / 1.03
Three models, economic articles	0.91 / 0.96 / 0.89	1.01 / 0.99 / 0.98	1.13 / 1.11 / 1.16	0.99	59%	9 / 8	1.08 / 1.03 / 1.01
Four models, all articles	0.88 / 0.97 / 0.90	1.01 / 1.00 / 0.99	1.23 / 1.17 / 1.17	1.00	49%	5 / 7	1.12 / 1.06 / 1.03
Four models, economic articles	0.87 / 0.95 / 0.88	0.94 / 0.96 / 0.96	1.04 / 1.03 / 1.08	0.98	64%	11 / 2	1.02 / 1.00 / 0.98

Note: Same hard data (101 FRED-MD series, outliers screened), same window. Three methods: LASSO, PCA–OLS (one factor), random forest, topics only. Four methods: Elastic Net, PCA–OLS (five factors), random forest, XGBoost, all text measures. Left block: best news-only model over the AR benchmark, H0 / H1 / H2, mean over the three months. Middle block: the same method with and without news, H0–H2; median ratio, share below one, comparisons significantly below / above one at the 10% level. Right: best news-only over best hard-data model, H0 / H1 / H2.

Table A.21: What the filter changes against the AR benchmark: all articles against economic articles, full sample and 2008–2009.

	Full sample		2008–2009	
Outcome	All articles	Economic articles	All articles	Economic articles
GDP	1.01 / 1.00 / 0.99	0.94 / 0.96 / 0.95	1.00 / 0.97 / 0.94	0.88 / 0.90 / 0.87
Consumption	0.88 / 0.97 / 0.90	0.87 / 0.95 / 0.88	0.71 / 0.93 / 0.84	0.65 / 0.80 / 0.76
Investment	1.23 / 1.17 / 1.17	1.04 / 1.03 / 1.08	1.07 / 0.96 / 1.00	0.77 / 0.83 / 0.94

Note: Best news-only model over the AR benchmark, mean over the three months, at H0 / H1 / H2; main specification, with the best model chosen ex post on the full sample. All articles: the baseline of the results section; economic articles: the main specification.

The walk from all articles to the main specification

The tables of this subsection are the specifications of the first four rows of Table 18: the three methods of Ellingsen et al. (2022) and the topic series, seven papers, training from 1997Q1, evaluation 2003Q1–2020Q1. Steps 1–3 use a 17-series subset of FRED-MD; step 4 returns to all articles with the full 101-series FRED-MD without the outlier screen. For each step the first table is the best news-only model against the benchmarks and the second the same-method increment (Tables A.22 to A.29); Table A.30 is step 4 by period.

Table A.22: Walk, step 1: all articles, 17 hard-data series. Best news-only model against the benchmarks.

Outcome	Benchmark	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	1.03	1.02	1.03	1.01	1.01	1.01	1.00	1.00	1.00
	RW	0.94	0.93	0.94	0.98	0.98	0.98	0.99	0.99	1.00
	SPF	1.31^{**}	1.30^{**}	1.31^{**}	1.17^{*}	1.17^{*}	1.17^{*}	1.10^{*}	1.10	1.11^{*}
Consumption	AR	0.89^{*}	0.90^{*}	0.92^{*}	0.96	0.96	0.95	0.89	0.89	0.89
	RW	0.86^{**}	0.87^{**}	0.88^{**}	0.91^{*}	0.90^{*}	0.90^{*}	0.92	0.92	0.92
	SPF	1.22^{***}	1.23^{**}	1.25^{**}	1.16^{*}	1.15^{*}	1.14^{*}	1.11^{*}	1.10^{*}	1.11^{*}
Investment	AR	1.20^{***}	1.21^{***}	1.22^{***}	1.09	1.15^{*}	1.15^{**}	1.25^{***}	1.21^{**}	1.15^{*}
	RW	1.03	1.05	1.05	1.04	1.10^{*}	1.10	1.23^{***}	1.19^{**}	1.13^{*}
	SPF	1.57^{**}	1.59^{***}	1.60^{***}	1.33^{*}	1.40^{**}	1.41^{***}	1.42^{***}	1.37^{***}	1.31^{***}

Note: RMSE of the best news-only model over the RMSE of the benchmark. Stars as in Table A.2.

Table A.23: Walk, step 1: all articles, 17 hard-data series. Hard data plus news against hard data, same method.

Outcome	Model	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	LASSO	1.12**	1.14**	1.08	0.89	1.00	1.08**	0.96	0.96	0.97
	PCA–OLS	0.97	0.98	1.10	0.96	0.96	1.01	0.99	0.98	0.98
	Random forest	1.01	0.99	1.02	0.99	1.01	0.99	1.00	0.99	0.98
Consumption	LASSO	0.92	0.87	0.88	0.97	0.98	0.93	0.96	0.96	0.89
	PCA–OLS	0.98	0.99	0.98	0.96*	0.95	0.96	0.95	0.93	0.90
	Random forest	1.01	1.02	1.00	1.01	1.02	1.01	0.97	0.97	0.96
Investment	LASSO	1.14**	1.06	1.13*	0.80	1.02	1.14	1.35**	0.98	1.22**
	PCA–OLS	0.95	0.95	1.26*	0.98	0.94	1.09	1.07	1.03	1.06
	Random forest	1.16***	1.19***	1.19***	1.08	1.11**	1.15**	1.03	1.00	1.05

Note: RMSE of the hard-data-plus-news forecast over the RMSE of the hard-data forecast for the same method; below one: text helps. Stars as in Table A.2.

Table A.24: Walk, step 2: economic articles. Best news-only model against the benchmarks.

Outcome	Benchmark	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	1.04	1.04	1.03	1.00	1.00	0.99	1.00	0.99	0.98
	RW	0.94	0.94	0.94	0.97	0.96	0.96	0.99	0.99	0.97
	SPF	1.32*	1.31*	1.31*	1.15*	1.15*	1.14*	1.11*	1.10*	1.08
Consumption	AR	0.94	0.94	0.95	0.97	0.97	0.99	0.89	0.89	0.91*
	RW	0.90*	0.90*	0.91**	0.91*	0.92*	0.93*	0.92	0.92	0.94
	SPF	1.27**	1.28**	1.29***	1.16*	1.17*	1.18*	1.11*	1.11	1.13
Investment	AR	1.17*	1.17	1.18***	1.10*	1.09*	1.07*	1.11	1.08	1.05
	RW	1.01	1.01	1.02	1.05	1.04	1.03	1.09*	1.05	1.03
	SPF	1.53**	1.53*	1.54**	1.34**	1.34**	1.31**	1.26**	1.22**	1.19***

Note: Economic articles, topics fitted on all articles, 17 hard-data series. Stars as in Table A.2.

Table A.25: Walk, step 2: economic articles. Hard data plus news against hard data, same method.

Outcome	Model	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	LASSO	1.05	1.06	1.05	0.89	0.99	1.02	0.94	0.97	1.00
	PCA–OLS	0.98	0.99	1.12	0.97	0.97	1.02	0.99	0.98	0.98
	Random forest	1.01	1.00	1.02	0.95*	0.98	0.98	0.96	0.95**	0.95
Consumption	LASSO	1.06	0.99	0.95	0.99	0.94	0.98	0.93	0.92	0.89
	PCA–OLS	0.99	1.01	1.01	0.97	0.97	0.99	0.96	0.94	0.92
	Random forest	1.05**	1.05*	1.04	1.01	1.04	1.03	0.98	0.96*	0.95**
Investment	LASSO	1.08	1.03	0.97	0.70	0.96	1.10	1.08	0.79	1.12
	PCA–OLS	0.95*	0.94	1.24	0.98	0.93	1.07	1.06	1.01	1.02
	Random forest	1.14**	1.17***	1.14**	1.01	1.03	1.08	0.94	0.93	0.97

Note: Below one: text helps. Stars as in Table A.2.

Table A.26: Walk, step 3: topics re-fitted on the economic articles. Best news-only model against the benchmarks.

Outcome	Benchmark	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	1.01	1.01	1.01	0.99	1.00	0.98	1.00	0.98	0.95
	RW	0.92	0.92	0.92	0.96	0.96	0.95	0.99	0.97	0.95
	SPF	1.28**	1.28*	1.29*	1.15*	1.15*	1.13*	1.11*	1.09	1.06
Consumption	AR	0.91	0.92	0.91*	0.96	0.96	0.95	0.89	0.89	0.89*
	RW	0.88*	0.89**	0.88**	0.90*	0.90*	0.90*	0.92	0.92	0.92
	SPF	1.24***	1.26**	1.24**	1.15**	1.15*	1.14*	1.11*	1.10*	1.11
Investment	AR	1.14***	1.15***	1.11*	1.10	1.13*	1.10**	1.15	1.19*	1.15*
	RW	0.98	0.99	0.95	1.05	1.08	1.05*	1.13	1.17*	1.12
	SPF	1.49***	1.50***	1.45**	1.34*	1.39**	1.34*	1.31*	1.35***	1.30**

Note: Economic articles, topics re-fitted on them, 17 hard-data series. Stars as in Table A.2.

Table A.27: Walk, step 3: topics re-fitted on the economic articles. Hard data plus news against hard data, same method.

Outcome	Model	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	LASSO	1.03	1.04	1.07	0.96	1.02	1.02	1.05	1.07	0.97
	PCA–OLS	0.98	0.98	1.10	0.97	0.97	1.02	0.99	0.98	0.98
	Random forest	0.99	0.99	1.00	0.94	0.97	0.97	0.97	0.95	0.93*
Consumption	LASSO	1.04	0.98	1.01	1.05	1.00	0.92	0.90	0.86**	0.82*
	PCA–OLS	0.99	0.99	0.97	0.96	0.95	0.96	0.96	0.93	0.91
	Random forest	1.02	1.03	1.03	1.01	1.03	1.03	0.98	0.94	0.94
Investment	LASSO	1.25*	1.35	1.20**	0.66	0.91	0.99	1.01	0.82	1.12
	PCA–OLS	0.95*	0.94	1.24*	0.98	0.94	1.08	1.06	1.01	1.05
	Random forest	1.12**	1.16***	1.15***	1.06	1.08	1.11*	1.03	1.01	1.05

Note: Below one: text helps. Stars as in Table A.2.

Table A.28: Walk, step 4: all articles, full FRED-MD without the outlier screen. Best news-only model against the benchmarks.

Outcome	Benchmark	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	1.03	1.02	1.03	1.01	1.01	1.01	1.00	1.00	1.00
	RW	0.94	0.93	0.94	0.98	0.98	0.98	0.99	0.99	1.00
	SPF	1.31**	1.30**	1.31**	1.17*	1.17*	1.17*	1.10*	1.10	1.11*
Consumption	AR	0.89*	0.90*	0.92*	0.96	0.96	0.95	0.89	0.89	0.89
	RW	0.86**	0.87**	0.88**	0.91*	0.90*	0.90*	0.92	0.92	0.92
	SPF	1.22***	1.23**	1.25**	1.16*	1.15*	1.14*	1.11*	1.10*	1.11*
Investment	AR	1.20***	1.21***	1.22***	1.09	1.15*	1.15**	1.25***	1.21**	1.15*
	RW	1.03	1.05	1.05	1.04	1.10*	1.10	1.23***	1.19**	1.13*
	SPF	1.57**	1.59***	1.60***	1.33*	1.40**	1.41***	1.42***	1.37***	1.31***

Note: All articles, 101 FRED-MD series, no outlier screen. Stars as in Table A.2.

Table A.29: Walk, step 4: all articles, full FRED-MD without the outlier screen. Hard data plus news against hard data, same method.

Outcome	Model	H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	LASSO	1.09*	1.03	0.97	1.02	1.01	1.01	1.04	1.02	1.04
	PCA–OLS	0.98	0.98	1.02	0.98	0.97	0.95	0.99	0.98	0.96
	Random forest	1.01	1.00	1.02	1.02	1.03	1.02	1.03	1.00	1.01
Consumption	LASSO	0.80†	0.25†	0.80†	0.91*	0.97	0.98	0.96	1.04	0.99
	PCA–OLS	0.99	0.98	0.96	0.97*	0.96*	0.92*	0.96	0.95	0.90
	Random forest	1.00	1.00	0.98	1.02	1.00	0.98	0.97	0.96	0.98
Investment	LASSO	1.04	0.94	0.99	0.97	1.23*	1.02	1.05	1.18	1.02
	PCA–OLS	0.97*	0.95*	1.02	0.98	0.96	0.90	1.02	1.01	0.97
	Random forest	1.07***	1.07***	1.08***	1.07**	1.07**	1.05**	1.05	1.08**	1.06*

Note: Below one: text helps. Stars as in Table A.2. †: the hard-data forecast for 2008Q2 is outlier-driven; the cell is excluded from the summaries.

Table A.30: Walk, step 4: all articles, full FRED-MD without the outlier screen. Against the AR benchmark by period.

Outcome	Best model	Full sample (69 q.)	Without 2008Q3–2009Q2 (65 q.)	2008–2009 (8 q.)
GDP	News only	1.03 / 1.01 / 1.00	1.08 / 1.04 / 1.04	1.04 / 0.99 / 0.97
	Hard data only	0.99 / 1.00 / 1.01	0.99 / 1.02 / 1.03	1.02 / 1.01 / 1.00
	Hard data + news	1.00 / 1.02 / 1.01	1.02 / 1.06 / 1.07	1.03 / 1.00 / 0.97
Consumption	News only	0.90 / 0.96 / 0.89	0.95 / 0.94 / 0.93	0.83 / 0.99 / 0.80
	Hard data only	0.92 / 0.96 / 0.94	0.95 / 0.96 / 0.99	0.87 / 0.97 / 0.88
	Hard data + news	0.91 / 0.96 / 0.91	0.95 / 0.97 / 0.95	0.85 / 0.94 / 0.84
Investment	News only	1.21 / 1.13 / 1.21	1.35 / 1.22 / 1.44	1.12 / 1.06 / 1.04
	Hard data only	0.94 / 1.00 / 1.04	0.97 / 1.06 / 1.08	0.94 / 1.00 / 1.04
	Hard data + news	0.99 / 1.06 / 1.11	1.04 / 1.18 / 1.23	0.97 / 1.01 / 1.04

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast.

Two papers, 1996–2020

The tables of this subsection use the New York Times and the Wall Street Journal only, training from 1990Q1 and evaluation over 1996Q1–2020Q1 (97 quarters), with the main specification otherwise (see Two recessions, two papers in the results section). Tables A.31 to A.33 and Figure A.6 use the four methods and all text measures; Tables A.34 and A.35 use the three methods of Ellingsen et al. (2022) and the topic series.

Table A.31: Two papers, four methods: best news-only model against the three benchmarks.

Outcome	Benchmark	H0			H1			H2			H3		
		M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
GDP	AR	1.00	0.98	0.95	1.00	0.99	0.98	0.99	0.98	0.97	1.00	1.01	1.02
	RW	0.92	0.90	0.87	0.97	0.96	0.95	0.98	0.97	0.96	1.01	1.01	1.02
	SPF	1.18**	1.16**	1.11*	1.07*	1.06	1.05	1.03	1.02	1.01	1.01	1.01	1.02
Consumption	AR	0.94	0.91	0.89*	0.97	0.95	0.97	0.94	0.93	0.92	0.97	1.02	1.04
	RW	0.91*	0.88*	0.86**	0.92	0.90	0.91	0.98	0.97	0.96	0.96	1.01	1.02
	SPF	1.14**	1.09*	1.07	1.08*	1.05	1.07	1.08*	1.06	1.06	1.02	1.07	1.08**
Investment	AR	1.04	1.03	1.01	0.99	0.97	0.97	1.03	1.03	0.97	1.07	1.05	1.07
	RW	0.88*	0.87*	0.85	0.95	0.93	0.92	1.03	1.03	0.97	1.08	1.06	1.09
	SPF	1.36***	1.35***	1.32***	1.21***	1.18**	1.18**	1.15**	1.15**	1.08	1.13	1.11	1.14*

Note: New York Times and Wall Street Journal; training from 1990Q1, evaluation 1996Q1–2020Q1 (97 quarters). Stars as in Table A.2.

Table A.32: Two papers, four methods: against the AR benchmark by period.

Outcome	Best model	Full sample (97 q.)	1996–2007 (48 q.)	2008–2009 (8 q.)	2010–2019 (40 q.)
GDP	News only	0.98 / 0.99 / 0.98 / 1.01	1.05 / 1.10 / 1.10 / 1.12	0.87 / 0.86 / 0.85 / 0.91	1.03 / 1.08 / 1.05 / 1.08
	Hard data only	0.95 / 0.99 / 0.97 / 1.00	0.97 / 1.03 / 1.13 / 1.09	0.89 / 0.95 / 0.84 / 0.88	0.99 / 1.05 / 1.04 / 1.14
	Hard data + news	0.95 / 0.97 / 0.98 / 1.00	1.01 / 1.04 / 1.11 / 1.10	0.86 / 0.91 / 0.86 / 0.92	1.02 / 1.05 / 1.05 / 1.05
Consumption	News only	0.91 / 0.96 / 0.93 / 1.01	1.05 / 1.07 / 1.06 / 1.17	0.70 / 0.76 / 0.77 / 0.95	0.91 / 0.98 / 0.92 / 1.05
	Hard data only	0.92 / 0.97 / 0.89 / 1.00	1.04 / 1.07 / 0.99 / 1.14	0.72 / 0.87 / 0.74 / 0.89	0.89 / 0.94 / 0.90 / 0.98
	Hard data + news	0.90 / 0.95 / 0.92 / 1.02	1.03 / 1.04 / 1.04 / 1.17	0.69 / 0.80 / 0.77 / 0.91	0.87 / 0.93 / 0.92 / 1.00
Investment	News only	1.03 / 0.98 / 1.01 / 1.06	0.96 / 0.89 / 0.93 / 0.96	0.88 / 0.91 / 0.97 / 0.96	1.33 / 1.29 / 1.32 / 1.49
	Hard data only	0.88 / 0.93 / 0.98 / 0.98	0.88 / 0.86 / 0.92 / 0.96	0.87 / 0.95 / 1.02 / 0.98	0.87 / 1.06 / 1.06 / 1.04
	Hard data + news	0.91 / 0.94 / 1.00 / 1.04	0.94 / 0.87 / 0.93 / 0.96	0.84 / 0.91 / 0.95 / 0.92	0.96 / 1.17 / 1.29 / 1.46

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2 / H3. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast. 2001 recession: 2001Q1–2001Q4.

Table A.33: Two papers, four methods: the two recessions, every horizon.

	2001 recession (4 q.)		2008–2009 (8 q.)	
Best model	GDP	Consumption	GDP	Consumption
News only	1.10 / 1.15 / 1.13	1.17 / 1.27 / 1.16	0.87 / 0.86 / 0.85	0.70 / 0.76 / 0.77
Hard data only	0.94 / 1.14 / 1.19	1.14 / 1.23 / 1.11	0.89 / 0.95 / 0.84	0.72 / 0.87 / 0.74
Hard data + news	1.12 / 1.17 / 1.11	1.28 / 1.21 / 1.20	0.86 / 0.91 / 0.86	0.69 / 0.80 / 0.77

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast. 2001 recession: 2001Q1–2001Q4. Four and eight quarters are too few for the test.

Table A.34: Two papers, three methods: against the AR benchmark by period.

Outcome	Best model	Full sample (97 q.)	1996–2007 (48 q.)	2008–2009 (8 q.)	2010–2019 (40 q.)
GDP	News only	1.00 / 1.00 / 0.99	1.04 / 1.08 / 1.09	0.99 / 0.94 / 0.91	0.99 / 1.03 / 1.03
	Hard data only	0.96 / 1.00 / 0.99	1.01 / 1.06 / 1.12	0.90 / 0.96 / 0.88	0.97 / 1.02 / 1.05
	Hard data + news	0.97 / 1.00 / 1.00	1.02 / 1.05 / 1.10	0.95 / 0.94 / 0.91	0.93 / 1.06 / 1.04
Consumption	News only	0.93 / 0.97 / 0.93	1.06 / 1.06 / 1.04	0.76 / 0.88 / 0.80	0.91 / 0.94 / 0.91
	Hard data only	0.92 / 0.99 / 0.91	1.02 / 1.06 / 0.99	0.76 / 0.96 / 0.82	0.84 / 0.93 / 0.87
	Hard data + news	0.92 / 0.97 / 0.92	1.07 / 1.04 / 1.02	0.71 / 0.90 / 0.81	0.86 / 0.94 / 0.90
Investment	News only	1.07 / 1.03 / 1.04	0.98 / 0.93 / 0.95	1.10 / 1.01 / 1.04	1.20 / 1.32 / 1.28
	Hard data only	0.89 / 0.93 / 1.00	0.89 / 0.83 / 0.88	0.90 / 0.99 / 1.05	0.87 / 1.00 / 1.17
	Hard data + news	0.92 / 0.97 / 1.03	1.00 / 0.87 / 0.90	0.81 / 1.00 / 1.05	0.92 / 1.11 / 1.31

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast. 2001 recession: 2001Q1–2001Q4.

Table A.35: Two papers, three methods: the two recessions, every horizon.

	2001 recession (4 q.)		2008–2009 (8 q.)	
Best model	GDP	Consumption	GDP	Consumption
News only	1.23 / 1.09 / 1.13	1.17 / 1.10 / 1.21	0.99 / 0.94 / 0.91	0.76 / 0.88 / 0.80
Hard data only	0.86 / 1.06 / 1.08	1.05 / 0.96 / 1.10	0.90 / 0.96 / 0.88	0.76 / 0.96 / 0.82
Hard data + news	1.11 / 1.06 / 1.10	1.08 / 1.03 / 1.16	0.95 / 0.94 / 0.91	0.71 / 0.90 / 0.81

Note: Relative RMSE against the AR benchmark, mean over the three months, at H0 / H1 / H2. The best model is chosen ex post on the full sample and not re-chosen. Samples are defined on the quarter being forecast. 2001 recession: 2001Q1–2001Q4.

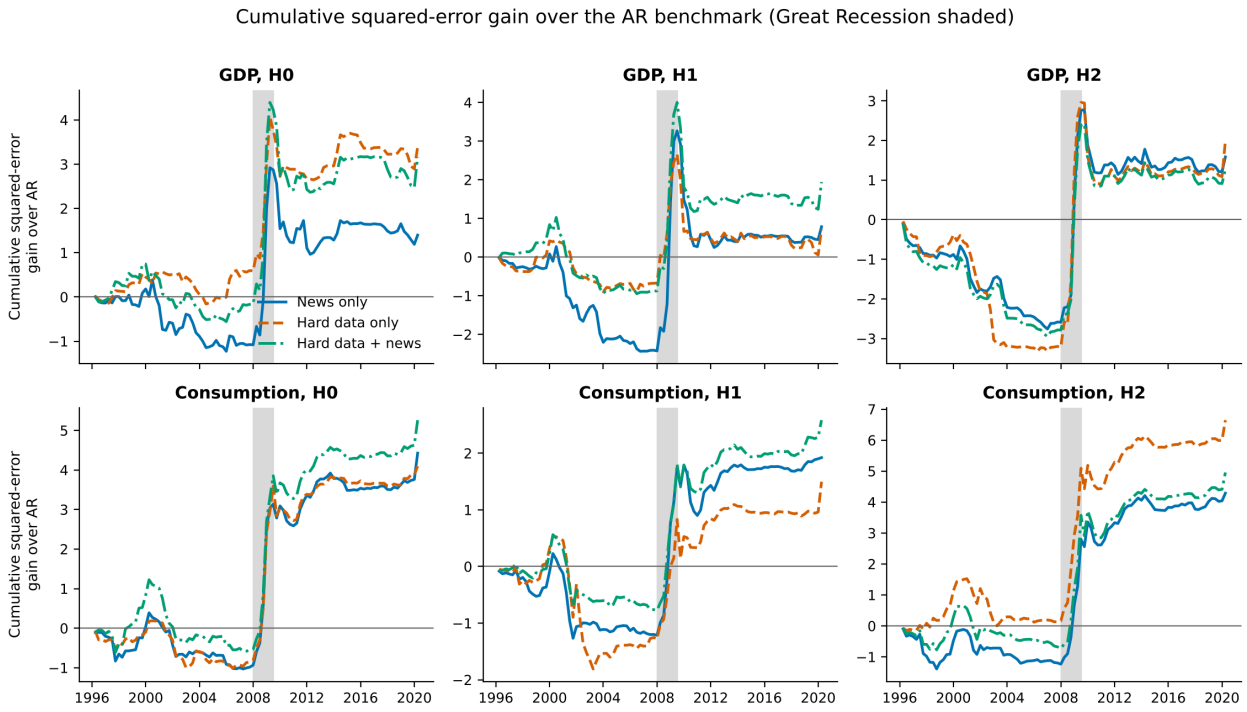


Figure A.6: Two papers, four methods: cumulative squared-error gain over the AR benchmark, 1996–2020, nowcast to two quarters ahead. Cumulative sum of the AR benchmark's squared error minus the model's, best model per predictor set, mean over the three months; shaded: 2007Q4–2009Q2.

Appendix B: The Economic-Article Filter

This appendix gives the details of the filter of the data section: the cluster map, the economic share over time, the keyword rule and the sampled headlines, and the ten selection rules against the hand labels.

The cluster map

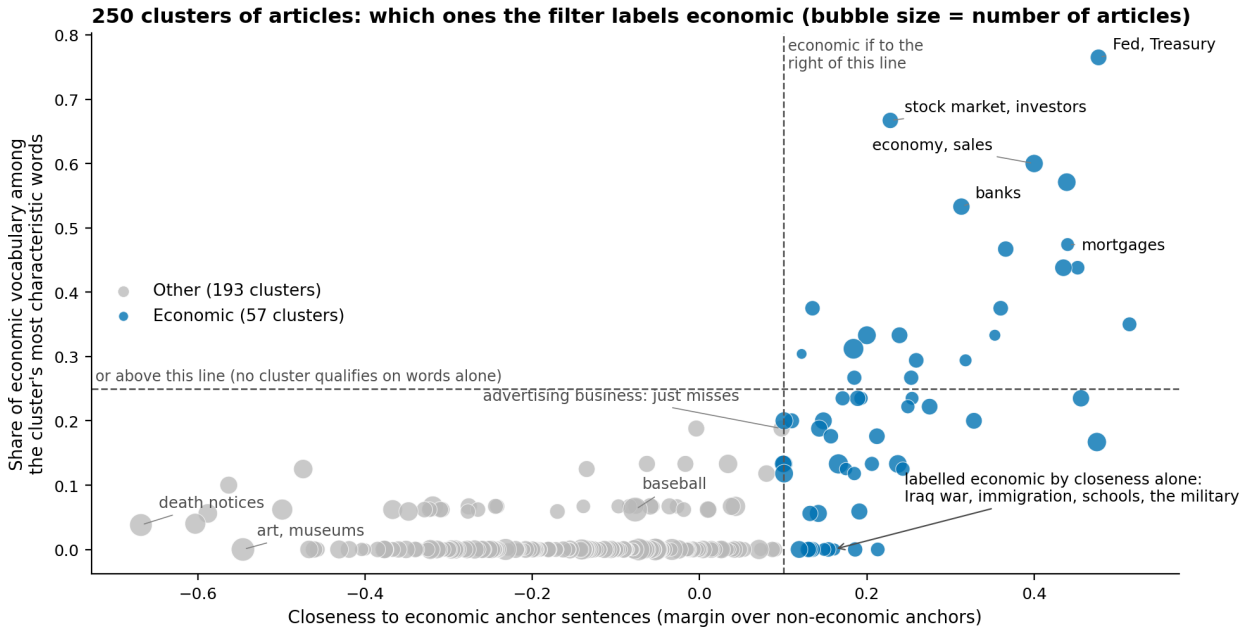


Figure B.1: Which clusters the filter labels economic. Each bubble is one of the 250 clusters, placed by its anchor margin (horizontal) and its word score (vertical); a cluster is economic to the right of the vertical line or above the horizontal line.

Figure B.1 places each of the 250 clusters by its two scores. The horizontal axis is the anchor margin, the difference between the cluster centre's similarity to ten economic anchor sentences and to ten non-economic ones; the vertical axis is the word score, the share of the cluster's most characteristic words that are on a list of economic vocabulary. A cluster is labelled economic when it is to the right of the vertical line, a margin of at least 0.10, or above the horizontal line, a word score of at least 0.25. All 57 economic clusters qualify by closeness to the anchors, and 21 of them also by their words. The 36 that qualify by closeness alone include clusters on the Iraq war, immigration, schools and the military, which is where the 8 doubtful clusters of the data section sit.

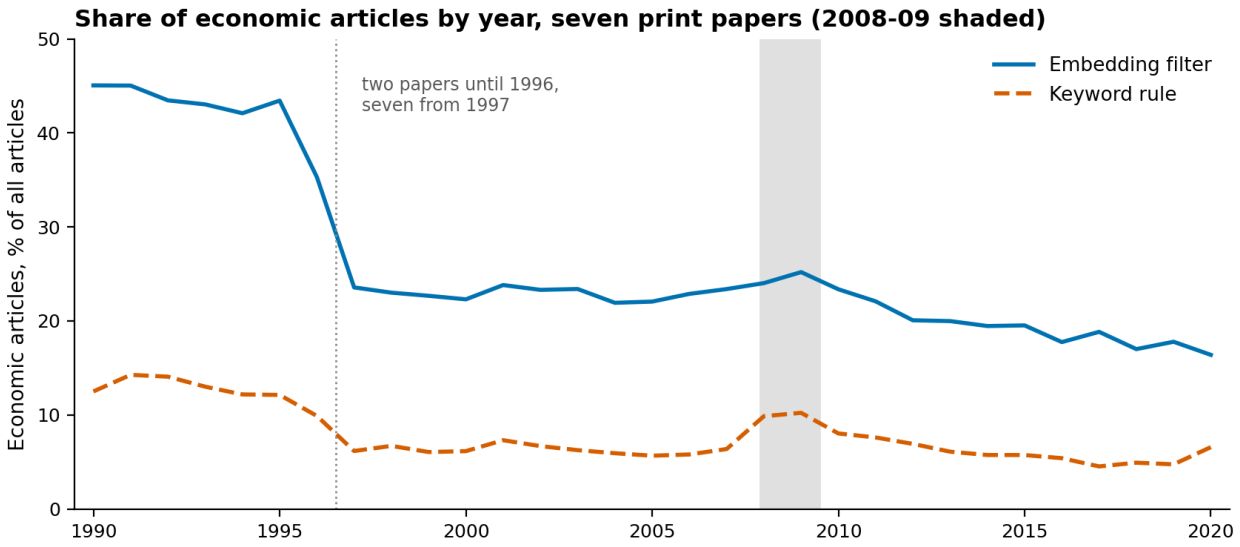


Figure B.2: The share of articles flagged as economic by the keyword rule and by the embedding filter, by year, 1990–2020, seven print newspapers.

The economic share over time

Figure B.2 shows the share of articles each rule flags, by year. Over the whole corpus the keyword rule flags 7.9% and the embedding filter keeps 24.7%. Neither share is stable. By the embedding filter the share is 42–45% in 1990–1995, when only the New York Times and the Wall Street Journal are in the corpus, 22–25% in 1997–2009 and 16–19% in 2016–2020; by the keyword rule it falls from 8% in 2010 to 4.7% in 2019. Both rules rise in recessions: the keyword share from 6% to 10% in 2008–2009 and from 6% to 7% in 2001, the embedding share from 23% to 25% in 2008–2009. The two rules therefore agree that the newspapers write more about the economy in a recession, and disagree by a factor of three on how much of what they write is about the economy at all.

The keyword rule and the sampled headlines

The keyword rule is the first stage of an earlier version of the filter and is kept as the comparison. It reads the title and the first 1,500 characters of an article and flags the article when they contain terms from two of twenty economic term groups, or when the title alone contains a term from one group. It flags 6–10% of the articles in a given year. Table 3 in the data section crosses it with the embedding filter.

To see what each rule catches, I drew fifteen headlines at random, with a fixed seed, from each of the three groups of articles of the seven print papers over 1997–2020: the 93,665 articles that only the keyword rule flags (Table B.1), the 1,576,849 that only the embedding filter keeps

(Table B.2), and the 547,721 that both rules flag (Table B.3). Only the headlines are shown. The keyword-only group holds sports, arts and lifestyle items that use an economic word in another sense; the embedding-only group holds company, deal, budget and market news without the listed words, and some items that are not economic; the group flagged by both is business and policy news.

Table B.1: Sampled headlines: articles flagged by the keyword rule only.

Newspaper	Year	Headline (cut at 54 characters)
Los Angeles Times	2006	Red Cross Chapter Chief Resigns Amid Complaints; Cont…
Chicago Tribune	2017	American Dental Association Foundation Announces Hiri…
Los Angeles Times	2020	Artists are essential too; Stop assuming that content…
Los Angeles Times	2004	AROUND THE NFL; Raiders Close to Hiring Payton
Wall Street Journal	2000	The Homestretch
New York Times	2018	Trump May Choose Critic of Fannie and Freddie to Lead…
Boston Globe	2018	New England colleges have one big worry: 2025
USA Today	2009	Starbucks dips toe into beer and wine sales
Los Angeles Times	2008	Italy lending Bernini art to the Getty; Major exhibit…
New York Times	2014	In Detroit, Judge Takes Offbeat Tack in Hiring Help
Wall Street Journal	2012	Next: Planning & Living The New Retirement (A Special…
Los Angeles Times	1998	Don't Rush to Raise Teacher Pay
Los Angeles Times	2008	HARD TIMES: CULTURE AT A CROSSROADS; Panic and anger…
Boston Globe	2009	Gerald Scully; applied economics to professional base…
Chicago Tribune	2014	King's abdication could be end of royal line in Spain

Note: Fifteen headlines drawn at random from the 93,665 articles of the seven print papers, 1997–2020, that only the keyword rule flags. Headlines are cut at 54 characters.

Ten selection rules against the hand labels

The 300 hand-labelled articles were drawn from the seven papers over 1997–2020 in 18 strata: whether the article's nearest cluster is one of the 21 word-qualified clusters, one of the other 36 economic clusters, or a non-economic cluster, crossed with six bands of the article's own anchor margin. The strata are weighted by their size in the corpus. I labelled 116 articles as economic, 182 as not economic and 2 as borderline; the borderline articles count as not economic. Weighted by stratum size, 20.5% of all articles are economic by this standard. Table B.4 applies ten rules to the same 300 articles. Each rule is a different combination of which clusters count as economic and what margin an article needs on its own; the kept share is the share of the 11.76 million articles each rule keeps; precision is the share of the kept articles that are economic and recall the share of the economic articles that are kept, with 90% intervals from a bootstrap within strata.

Table B.2: Sampled headlines: articles kept by the embedding filter only.

Newspaper	Year	Headline (cut at 54 characters)
Chicago Tribune	2011	How debt-ceiling talks came apart
Washington Post	2006	WORKING
Boston Globe	2007	Investors say no to bid from Genzyme; After rare defe…
Wall Street Journal	2014	Overheard
Boston Globe	2003	HYDROGEN HOT AIR
Wall Street Journal	2002	Web Ads Hit Rock Bottom: Some Are Free
Wall Street Journal	1997	Party Line: The Battle for MCI Takes Another Twist: N…
Boston Globe	2014	A spin around the old vinyl
New York Times	1999	Newport News Joins in a New Venture
Boston Globe	2003	OUR DREAM OF A GREAT SOCIETY
New York Times	2010	An Unlikely Path to a Great Workplace
Los Angeles Times	2008	SMALL BUSINESS; Shoemaker in step with clients; J&A S…
Chicago Tribune	2002	‘Are you serious? That’s him? ; Former Enron executiv…
Los Angeles Times	2011	Amy Baer out as head of CBS Films; The executive left…
Chicago Tribune	2001	A few simple steps key to short storage

Note: Fifteen headlines drawn at random from the 1,576,849 articles of the seven print papers, 1997–2020, that only the embedding filter keeps. Headlines are cut at 54 characters.

Table B.3: Sampled headlines: articles flagged by both rules.

Newspaper	Year	Headline (cut at 54 characters)
New York Times	1999	THE MARKETS: STOCKS; Strong I.B.M. Report Helps Prope…
New York Times	2005	Bank to Buy Credit Card Business for $6.45 Billion
Los Angeles Times	2010	MONDAY BUSINESS; SMALL BUSINESS / Q & A; Getting more…
Wall Street Journal	2003	Bush Plan Would Hold Increase In 2004 Spending to Tig…
New York Times	2020	Kodak C.E.O. Got Stock Options Day Before News of Loa…
USA Today	2002	Enron workers get $29M ’beacon of hope’ ; 4,200 forme…
Los Angeles Times	1999	China Budgets Red Ink to Meet 7% Growth Plan
Wall Street Journal	2003	Beiersdorf Sharpens Its Brand Focus
Washington Post	1999	Save Our Steel
Los Angeles Times	2000	WALL STREET, CALIFORNIA; MONEY MAKE-OVER / Southern C…
Wall Street Journal	2016	Dell, EMC Complete Merger
Wall Street Journal	2018	Universal Income, Incentives and Well-Being
Wall Street Journal	2006	Currency Games
Wall Street Journal	2019	World News: China Counts the Costs Of Its Big Bet on…
Chicago Tribune	1997	PRICIEST OF ALL POSSIBLE WORLDS INFLATION MUST BE A T…

Note: Fifteen headlines drawn at random from the 547,721 articles of the seven print papers, 1997–2020, that both rules flag. Headlines are cut at 54 characters.

Table B.4: Ten selection rules against 300 hand-labelled articles.

Rule	Kept (%)	Precision	Recall	F1
economic cluster (57) or margin ≥ 0.10 (main specification)	24.7	71 (66–77)	76 (66–86)	73
economic cluster (49: 8 doubtful clusters relabelled) or margin ≥ 0.10	22.8	75 (70–81)	76 (67–86)	76
as main, but never from the 8 doubtful clusters	22.0	75 (70–81)	72 (63–82)	74
economic cluster (57) only	21.9	77 (71–83)	72 (63–81)	74
economic cluster (64) or margin ≥ 0.05	31.0	62 (57–67)	81 (72–91)	71
economic cluster (32) or margin ≥ 0.20	13.3	87 (82–93)	49 (42–56)	62
economic cluster (57) and margin ≥ 0.10	13.2	84 (77–91)	45 (39–52)	59
margin ≥ 0.10 only	16.0	74 (68–80)	49 (43–57)	59
core clusters only (21: economic by their words)	7.9	91 (87–95)	29 (25–34)	44
as main, and flagged by the keyword rule	6.8	91 (83–97)	26 (20–32)	40

Note: Kept: share of the 11.76 million articles. Precision: share of the kept articles that are economic; recall: share of the economic articles that are kept; in brackets, 90% intervals from a bootstrap within strata. Borderline articles count as not economic. Doubtful clusters: 8 of the 57 whose characteristic words are about war, the military, immigration, schools, foreign policy or Washington politics. The main specification is the first row.

Precision rises and recall falls as the rule tightens, and no rule has both above 80%. The main rule, 71% and 76%, sits in the middle. Relabelling the 8 doubtful clusters raises precision to 75% at the same recall; requiring both the cluster and the article margin, or the cluster and the keyword rule, raises precision to 84–91% and drops recall to 26–45%. The three rules that were run through the forecasting exercise in the robustness section, cleaned, strict and loose, were picked from this table before their forecasts were computed.

Details of the filter

The filter has four stages. In the first, an earlier keyword stage reads the title and the first 1,500 characters of each article against 20 term groups, flagging an article when two groups appear or one appears in the title; it flags 6–10% of the articles and is kept only as the comparison above. In the second, the title and the first 1,200 characters of each article are embedded with `all-MiniLM-L6-v2`, 250 k-means clusters are fitted on a sample of 300,000 articles after a reduction to 50 principal components, 57 clusters are labelled economic by the two scores, and an article is labelled economic when its nearest cluster is economic or its own anchor margin is at least 0.10; this keeps 24.7% of the articles, 2.9 million. In the third, the paper-month panels of every text measure are rebuilt from the flagged articles alone. In the fourth, the latent Dirichlet allocation model with 80 topics is re-fitted on the 2.9 million economic articles, and every monthly topic series is rebuilt from it. Against the 300 hand labels the filter has a precision of 71% and a recall of 76%; the three other rules give the same forecasts (see The selection rule under robustness).

Appendix C: Implementation Details

This appendix collects the settings behind the Data and Forecasting Design sections: the text pipeline, the hard data and the targets, the estimators, and the test. None of them is a result; each is what a replication would need.

Text pipeline. The sentence encoder `all-MiniLM-L6-v2` embeds the title and the first 1,200 characters of each article, with at most 256 tokens and mean pooling. Before clustering, the embeddings of the 300,000-article sample are reduced to 50 dimensions by principal components, and the 250 clusters are fitted by mini-batch k-means. The keyword rule of the data section reads the title and the first 1,500 characters of each article. The topic model is the online variational Bayes implementation of latent Dirichlet allocation in scikit-learn, run in five passes on a random sample of 400,000 economic articles, the sample cap, from the 2.9 million; the vocabulary consists of Snowball stems that appear in at least ten documents and at most half of them, at most 80,000 terms, counted over all economic articles. Before the cross-paper summary, a paper-month whose article count is below 15% of that paper's median monthly count is dropped; in the main specification this removes 96 of the 2,360 paper-months in which the seven papers have economic articles, 53 of them for the Los Angeles Times, 25 for the Boston Globe, 15 for the Chicago Tribune and 3 for USA Today.

Hard data and targets. Each FRED-MD series is transformed with the transformation code given in its vintage file. A transformed observation that the outlier screen sets to missing is filled like any other missing value, by the rule below. The targets are the ALFRED series GDPC1, PCECC96 and PNFIC1. For the final-vintage actuals of investment the annualised percent change A008RL1Q225SBEA is used and converted to a quarterly log growth rate as $g = 100\log(1 + r/100)/4$.

Estimation. In the regression, a slot whose month lies after the standing month, or whose value is missing, is filled with the predictor's real-time mean, the mean of its observations up to the standing month (zero if there are none); because of this fill rule no missing value reaches the estimator. The Elastic Net, PCA–OLS and random forest are the implementations of scikit-learn (Pedregosa et al., 2011). The random seed is fixed at 42 throughout. For the Elastic Net the predictors are standardised, the ℓ_1 ratio is chosen from $\{0.5, 1\}$ and the penalty strength from a grid of 20 values, by cross-validation with three time-ordered folds inside each training sample. PCA–OLS uses the standardised predictors. The random forest averages 300 trees of depth at most six, with at least five observations per leaf. XGBoost uses 300 trees of depth four, learning rate 0.05, row and column subsampling of 0.8, and ℓ_1 and ℓ_2 penalties of 0.1 and 1 on the leaf weights. In the replication specification of Ellingsen et al. (2022), the LASSO is the Elastic Net with the ℓ_1 ratio fixed at one, otherwise the same cross-validation; PCA–OLS uses one factor; and the random forest has 500 trees with a third of the predictors tried at each split.

Evaluation. The Diebold–Mariano test is applied to the difference in squared errors. The Newey–West estimator of the long-run variance uses Bartlett weights with truncation lag h, and the small-sample factor of Harvey et al. (1997) is the one for an $(h+1)$-step forecast.