Received: 23 August 2022 | Revised: 8 January 2024 | Accepted: 11 January 2024

DOI: 10.1002/jae.3057

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Nowcasting Euro area GDP with news sentiment: A tale of two crises

Julian Ashwin[1] | Eleni Kalamara[2] | Lorena Saiz[3]

[1]Maastricht University, Maastricht, The Netherlands

[2]Nomura, Tokyo, Japan

[3]European Central Bank, Frankfurt, Germany

Correspondence
Julian Ashwin, Maastricht University, Maastricht, The Netherlands.
Email: julian.ashwin@maastrichtuniversity.nl

Summary

This paper shows that newspaper articles contain signals that can materially improve real-time nowcasts of real GDP growth for the Euro area. Using articles from 15 popular European newspapers, which are machine translated into English, we create sentiment metrics that update daily and assess their value for nowcasting, comparing with competitive and rigorous benchmarks. We find that newspaper text is especially helpful early in the quarter before other indicators are available. We also find that general-purpose sentiment measures perform better than more economics-focused ones in response to unanticipated events and nonlinear supervised models can help capture extreme movements in growth but require sufficient training data to be effective.

KEYWORDS

business cycles, COVID-19, forecasting, machine learning, text analysis

1 | INTRODUCTION

Monitoring the economy in real time is crucial for making informed economic and policy decisions but macroeconomic fundamentals like gross domestic product (GDP) are measured at a quarterly frequency and released with a substantial delay. Market participants and policymakers typically rely on soft data such as business and consumer surveys to get a real-time assessment of economic conditions. For example, monetary policy communications frequently point to survey evidence when describing the current macroeconomic situation. However, even these survey-based metrics are often only available with a lag and at a monthly frequency. For this reason in the last years the interest in using text-based sentiment for economic forecasting has increased. Text-sentiment can be computed at a relatively low cost in relation to expensive surveys, at higher frequency (i.e., daily) and on a timely manner.[1] This paper builds daily sentiment metrics based on economic news media for the Euro area and shows that they can improve real-time nowcasting performance of GDP compared with the ECB's projections or nowcasts based on survey-based indicators and on other economic indicators.

This study contributes to this growing body of work which shows that non-traditional datasets and methods can improve economic forecasts at short forecast horizons.[2] This includes the use of observational text data where substantial progress has been made in recent years. The uncertainty indices of Baker et al. (2016) and the topic-based sentiment indicators of Thorsrud (2020) are prominent examples of how text can be useful for economic analysis. In a nowcasting and short-term

[1]Unfortunately for the Euro area, which is the focus of this work, the availability of alternative high-frequency indicators is rather limited (i.e., mostly financial).

[2]For a review of alternative datasets, we refer to Algaba et al. (2020) while a detailed application of a wide set machine learning (ML) methods to forecast US output is presented in Coulombe et al. (2020)

forecasting context, studies such as Larsen and Thorsrud (2019), Kalamara et al. (2022), Shapiro et al. (2022), Aguilar et al. (2021) and Barbaglia et al. (2023) show that text can improve forecasts of key macroeconomic variables including GDP, inflation and unemployment.[3] Similarly, Ardia et al. (2019) and Rambaccussing and Kwiatkowski (2020), using US and UK newspaper texts, respectively, combine expert judgement and linear ML methods to forecast economic growth. Within the Euro area, Aguilar et al. (2021) show that their sentiment indicator derived from Spanish newspapers is comparable with the sentiment index produced by the European Commission and more helpful in nowcasting GDP. Similarly, Aprigliano et al. (2022) build sentiment and uncertainty indices for Italy and provide evidence of sizeable gains in forecast accuracy, both in normal and turbulent times, when forecasting several macroeconomic aggregates. For France, Bortoli et al. (2018) find that monthly media sentiment improve GDP forecasts in relation to a linear autoregressive model and also when including a business climate indicator. Barbaglia et al. (2024) build news-based sentiment indicators for Germany, France, Italy, Spain and the United Kingdom using an aspect-based dictionary in English covering six topics (economy, unemployment, inflation, manufacturing, financial and monetary policy). They find that different topics are significant GDP predictors for different countries at different forecast horizons.

Our focus is on the use of text data derived from European newspaper articles, in several non-English languages, tagged as 'economic news', to nowcast quarterly real GDP growth in the Euro area. We transform the text into daily aggregate time series of news sentiment for the Euro area. We make use of well-known lexicon-based methods like the economics-oriented dictionaries of Correa et al. (2017), Loughran and McDonald (2011) and Barbaglia et al. (2022) but also more of general purpose such as VADER (Hutto & Gilbert, 2014) and AFINN (Nielsen, 2011). This allows us to create high-frequency text-based indicators which are able to capture the current economic conditions in a timely manner. Consequently, all data we use throughout in constructing our nowcasts and the benchmarks are data that were available in real time (although we assess performance relative to latest revisions of GDP growth at the time of writing). While the focus of this paper is on text, there are several other sources of high-frequency data that have promising applications in nowcasting such as credit card transactions (Galbraith & Tkacz, 2018) or internet searches (Nagao et al., 2019; Woloszko, 2020). These other sources are also promising avenues for future research.

We test the predictive ability of these daily time series with a range of commonly used nowcasting models. This include straightforward linear regressions, mixed data sampling (MIDAS) regressions and mixed-frequency dynamic factor models (DFMs).[4] We also consider less traditional non-parametric ML approaches given their flexibility to accommodate nonlinear relationships without needing to impose a structure on the nonlinearity.[5] In our exercise, we exploit the timeliness of text, which is one of the major advantages of this type of data, and train the models on a daily basis as new text information becomes available. As such, we do not induce sparsity by variable selection, for example, through lasso regression, or dimensionality reduction, for example, through partial least squares.

We find that our sentiment metrics provide substantial improvements in real-time nowcasting performance compared with demanding benchmarks such as the ECB's official GDP projections, benchmarks based on the composite Purchasing Managers' Index (PMI), which has been shown to be among the best performing 'soft' indicators of GDP (Gajewski, 2014), and benchmark models including a range of real-time macroeconomic and financial indicators. These gains are typically concentrated in the first half of the quarter, when other indicators are not yet available, and are particularly pronounced in the two major crisis periods in our sample: the Great Recession (2008–2009) and the Great Lockdown (2020).

In considering the link between text and economic activity, we explore which methods provide the best estimates of GDP in normal times but also in times of distress like the COVID-19 pandemic period. We find that standard linear methods work well when there are no big shifts on the economic outlook, but nonlinearities matter when extreme economic shocks occur and the nonlinear ML models can capture them better and filter out the noise.

In addition to demonstrating the value of news sentiment in nowcasting Euro area GDP at a daily frequency, we make several key contributions relative to the existing literature. First, we tackle the issue of translation. Our dataset consists of five million articles from 15 newspapers based in the four largest Euro area economies: Germany, France, Italy and Spain. The large majority of these articles are written in their country's native language which poses us two challenges. On the one hand, most of the natural language processing (NLP) literature has been developed specifically for English. On the other hand, we need to apply a consistent approach to all news articles in order to produce aggregated metrics

[3]A similar literature in finance finds that text data has predictive value (Casarin & Squazzoni, 2013; Casnici et al., 2015).

[4]A large macroeconometric literature has proposed a number of methodologies to nowcast GDP such as factor-based models (Bańbura et al., 2011), Bayesian vector autoregressions (Cimadomo et al., 2022), bridge equations (Baffigi et al., 2004) and mixed data sampling techniques or MIDAS models (Ghysels et al., 2004; Foroni & Marcellino, 2014).

[5]Many recent studies provide promising avenues for improving macroeconomic predictions with ML methods; see Kim and Swanson (2018) for a review.

for the Euro area. We therefore use the Google Translate API to translate the raw text into English and create sentiment metrics from the translated text.[6] We compare this translation methodology to alternatives in Section 2.3. Another paper that uses translated articles from a range of different European countries is Barbaglia et al. (2024), who assess the quality of their machine translations into English with a topic model but do not compare forecasting performance with different translation methodologies. We find that the important informational content for nowcasting GDP is still captured on the translated text which allows for the same sentiment analysis approach and comparisons across all different languages. Our results thus support previous evidence that sentiment analysis is robust to machine translation (De Vries et al., 2018; Shalunts et al., 2016), showing that this is also true in an economic context.

Second, we propose an original approach to construct truly real-time sentiment indicators on each day of every quarter by accumulating the informational content of the newspaper articles as soon as new information becomes available. This means that at each quarter, the indices are reset to reflect only signals occurring within the relative quarter, and all available news from the current quarter is used at any given time. We find strong correlations of these daily text metrics with GDP growth at a country level and at a Euro area aggregate level. Previous work mentioned above has generally focused on indicators at a monthly or quarterly frequency and so do not reap the true benefits of the timeliness of text.[7] We believe that this is particularly important given that our results show that the advantage of news-based indicators is higher earlier in the quarter and diminishes as other indicators become available. A high-frequency metric is therefore necessary to assess when and how text data might be useful.

Third, we document a key dimension in which the choice of the text analysis methodology matters for real-time nowcasting. For a given crisis, metrics that are tailored to the nature of the underlying shock will perform best. However, these more specific metrics will work less well when a crisis has a different cause, suggesting that general-purpose sentiment metrics offer greater robustness. In particular, we show that the financial stability-based dictionary of Correa et al. (2017) performs best during the Great Recession but fails during the COVID-19 crisis. On the other hand, general-purpose sentiment measures such as 'VADER' (Hutto & Gilbert, 2014) is more consistent across time and robust to 'black swan' crises.

Finally, we show how the nowcasting gains depend on the model with which these metrics are incorporated. Prior to the Great Lockdown period, we show that text information included in a linear model yields substantial improvements in performance, especially at the beginning of the quarter when survey-based data and updated projections are not available in real time. This is in line with the evidence found in Kalamara et al. (2022) where simple time-series text indicators improve forecasts when using linear autoregressive models for forecasting UK GDP growth. As regards alternative methodologies and specifications, ridge regressions deliver the highest forecast error reductions including the text-based information during normal times, but nonlinear ML models are proved necessary during periods of large shifts, provided that there are enough data available.

2 | DATA AND TRANSLATION

2.1 | Text data

We use articles that are tagged as economic, corporate or financial news from major print newspapers for the 'Big Four' Euro area economies from the Factiva database. Restricting the data to these tagged articles reduces noise in the sentiment series by excluding irrelevant articles. We have a total of five million articles covering the period from January 1998 to January 2021, from 15 separate sources. Our newspapers have a wide circulation and reflect a range of political leanings, to ensure that our metrics are not biased by newspaper's support for current governments. Table 1 shows the number of articles from each country and the newspapers from which they are taken, with more detail provided in Appendix S1.

2.2 | Daily sentiment metrics

A key advantage of news articles in nowcasting is that they are released at a daily frequency and are available in real time. To fully capitalise on these advantages, we create a daily sentiment series that corresponds to the quarterly frequency

[6]We use the python package 'googletrans'.

[7]One study that does construct a daily indicator is Algaba et al. (2023) who use daily text signals to produce daily nowcasts of Belgian GDP growth. However, they focus only on one country, one sentiment measure and one linear DFM and compare their results to less demanding benchmarks.

TABLE 1 Total number of articles per country.

	France	Germany	Italy	Spain	All
Sources	Les Échos	Die Welt	Corriere della Sera	Expansión	
	Le Figaro	Süddeutsche Zeitung	La Repubblica	El Mundo	
	Le Monde	Der Tagesspiegel	Il Sole 24 Ore	El País	
		German Collection	La Stampa	La Vanguardia	
Total articles	1,255,472	833,914	1,497,909	1,407,534	4,994,829

TABLE 2 English language sentiment dictionaries used.

Initials	Source	Units	Application
AFINN	Nielsen (2011)	Words (+5 to −5)	General purpose
CGLM	Correa et al. (2017)	Words (+1 or −1)	Financial stability
HIV	Tetlock (2007)	Words (+1 or −1)	General purpose
HL	Hu and Liu (2004)	Words (+1 or −1)	Opinion in reviews
LM	Loughran and McDonald (2013)	Words (+1 or −1)	Economics/finance
NKTGOS	Nyman et al. (2018)	Words (+1 or −1)	Finance
VADER	Hutto and Gilbert (2014)	Sentences (−1 to +1)	General purpose
ECONLEX	Barbaglia et al. (2022)	Words (−1 to +1)	Economics

of GDP data. In other words, we produce direct forecasts of quarterly GDP growth using accumulated daily information throughout the quarter. In our baseline case, we translate the articles from their native languages using Google Translate, described in detail in Section 2.3. We use a suite of English language sentiment measures, described in Table 2, to compute a daily sentiment metrics for each of the four countries.

The sentiment metrics we use vary along three dimensions. First, as shown in the Application column of Table 2, some are designed for economic or financial applications (e.g., LM, ECONLEX), while others are more general (e.g., AFINN, VADER). Second, seven of the methods work at the word level, classifying words in isolation, but VADER works at the sentence level to account for factors like negation and punctuation. Third, most of the methods classify selected words as either positive (+1), negative (−1), but two take a granular approach. More specifically, the AFINN dictionary classifies words on an integer scale from most positive (+5) to most negative (−5), and the VADER method classifies sentences on a continuous scale from −1 to 1 (see Appendix S1 for more information on VADER).

For each method, we obtain a sentiment score for each word/sentence. As our target variable has a quarterly frequency, we develop a daily sentiment metric that recognises this quarterly frequency of the target, although in Section 3.3, we show that our main results are robust to a range of alternative aggregations for the text metrics. Each day in our sample is thus associated with two indices: q indicates which quarter that day is from, and d indicates the day within that quarter. The index (q, d) thus denotes the dth day in quarter q. Let $N_{q,d}$ be the total number of words/sentences across all articles for a given country on that day of the quarter. We can then define $sent_{q,d,n}$ as the sentiment score for the nth word/sentence on that day.

For each day, we then use all the articles from that quarter up to (and including) that day's to calculate the sentiment metric, weighting each word/sentence equally. The sentiment score for day d in quarter q is calculated as

$$sent_{q,d} = \frac{\sum_{t \leq d} \sum_{n=1}^{N_{q,t}} (sent_{q,t,n})}{\sum_{t \leq d} N_{q,t}}, \tag{1}$$

where $sent_{q,t,n}$ and $N_{q,t}$ are defined as above. Of course, how $sent_{q,t,n}$ is calculated and whether n refers to words or sentences depends on the text method in question, but the daily metric can be constructed for all methods and countries.[8]

As the quarter progresses, our metric thus averages over more days and therefore over more articles.[9] At the beginning of the quarter, the measure is thus noisier as it relies on a smaller sample. As the quarter progresses more, articles become available until the end of the quarter and the metric becomes more stable. Furthermore, as the number of

[8]In this paper, we do not compare the performance of individual newspapers, and therefore, we sum the scores of words/sentences from different newspapers but the same country.

[9]To the best of our knowledge, constructing a daily metric in this way is novel. Previous work has either focused on monthly frequencies or computed the daily sentiment as the average value over known days of the month (Aguilar et al., 2021) or over a 30-day moving average (Barbaglia et al., 2023).

TABLE 3 Correlation of sentiment metrics with (quarterly) GDP growth.

Translation	Metric		France	Germany	Italy	Spain	Euro area
Article	CGLM	Economic	0.610	0.527	0.539	0.71	0.697
	LM	Economic	0.588	0.461	0.408	0.619	0.636
	NKTGOS	Economic	0.377	0.373	0.359	0.618	0.538
	ECONLEX	Economic	0.559	0.332	0.379	0.655	0.616
	AFINN	General	0.578	0.371	0.407	0.731	0.637
	HIV	General	0.575	0.317	0.251	0.524	0.576
	HL	General	0.509	0.403	0.326	0.664	0.597
	VADER	General	0.578	0.369	0.289	0.701	0.611
Dictionary	CGLM	Economic	0.402	0.376	0.483	0.624	0.591
	LM	Economic	0.323	0.214	0.441	0.571	0.527
	NKTGOS	Economic	0.151	0.145	0.391	0.478	0.365
	AFINN	General	0.329	0.268	0.469	0.628	0.535
	HIV	General	0.285	0.24	0.182	0.38	0.445
	HL	General	0.241	0.304	0.25	0.553	0.475
Own lang				0.355			

articles increases, the precision of the metric also increases, as shown by the standard error bands. As our main focus is on real-time nowcasting for quarterly GDP, this metric is a natural candidate: combining sentiment information available from the quarter so far.

2.3 | Translation methodology

The news articles we work with are written in the native languages of the Euro area's Big 4 economies (i.e., German, French, Italian and Spanish). As most methods in the NLP literature focus on English, we test three different approaches for extracting sentiment and find that translating the articles into English provides the most robust and reliable results.

1. *Translating the articles*. Using the Google Translate API, we translate all of the news articles in our sample into English.
2. *Translating sentiment dictionaries*. Perhaps the simplest approach in practice and less computationally expensive is to translate (again using the Google Translate API) the various sentiment dictionaries from English to each of the four languages and then use these translated dictionaries on the original text.
3. *Language-specific dictionaries*. Where possible, we use language-specific dictionaries at a country level. An economics/business-specific dictionary for the German language is publicly available (Bannier et al., 2019), BPW henceforth.

As the VADER metric is a hybrid approach that takes the context of terms into account, it is not possible to apply this to non-English articles, so we only consider this measure applied to translated articles.

The first two methodologies produce highly correlated results in most cases, particularly for the economics-focused dictionaries, as shown in Table A.2 in Appendix S1. These correlations vary a little across languages, probably due to a combination of factors including the performance of the translation software and inherent features of the respective languages. Table 3 shows the correlation of six sentiment metrics on the translated articles with GDP growth for each of the four economies and the Euro area as a whole.[10]

From Table 3, we see that the most promising measures are two based on the economics-focused lexicons (CGLM and LM) as well as the two general-purpose dictionaries that allow for varying strengths of positivity and negativity (AFINN and VADER). The average correlation across the translated articles metrics is 0.518 for the translated articles and 0.389 for the translated dictionaries. This provides further evidence that translating the articles into English is the best approach for extracting the economically relevant information. Furthermore, most of the metrics based on English language dictionaries provide a higher correlation with GDP that those based on language-specific German dictionary from Bannier et al. (2019). We therefore focus on the results using the translated articles in the remainder of the paper. Furthermore, in the interest of brevity, we present results with two of our sentiment metrics, CGLM and VADER, as these generally perform best and are widely used examples of economics-focused and general-purpose lexicons.

[10]We exclude the year 2020 from these correlations, as the GDP growth rates here are so extreme (in both directions) that a handful of observations here will determine the correlations.

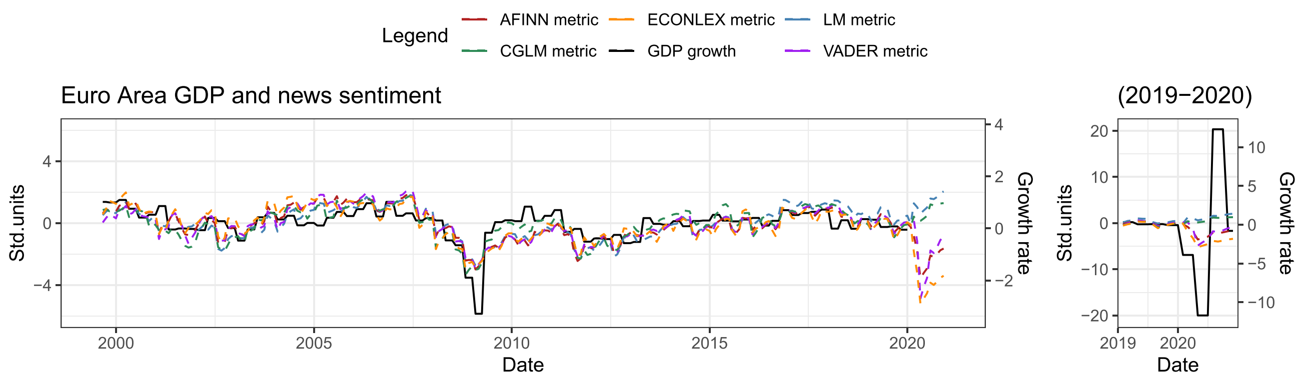

FIGURE 1 Euro area GDP growth and news sentiment.

2.4 | Constructing Euro area metrics

In order to forecast real GDP at the Euro area level, we compute news indicators for each country separately and then use Eurostat's GDP weights to compute Euro area aggregates.[11] We consistently found that models including these aggregated Euro area metrics outperformed models that incorporated the individual country series separately. To ensure that we only use data that were available in real time, we use weights from 2 years previously to construct the Euro area series.

Figure 1 shows three of the sentiment series alongside Euro area GDP, illustrating that the sentiment metrics co-move with economic activity.[12] Growth rates for 2020 are shown in a separate panel to account for the extremity of GDP data. The sentiment metrics are standardised so that the pre-2020 sample has zero mean and unit variance. These standardised units are given on the left-hand axis, with the quarter-on-quarter growth rates shown on the right-hand axis.

A first look at the sentiment metrics allows to draw two initial conclusions. First, the year 2020 shows a clear difference between the metrics based on economics-focused dictionaries (CGLM and LM) and the general-purpose dictionaries (AFINN and VADER), while the two types co-move throughout the rest of the sample and the Great Recession in particular. This supports our point that while the economics-focused dictionaries developed over the past decade perform well in response to the shocks they were designed to capture, they may not perform well in response to future shocks. An interesting case here is the ECONLEX dictionary (Barbaglia et al., 2022) which, in spite of being explicitly designed for economic applications, appears to be closer to the general-purpose dictionaries than the other economics-focused dictionaries. Second, while the series clearly co-move, this relationship appears to be nonlinear. In particular, during crisis periods where we see a large fall in GDP, the sentiment metrics do not fall proportionately to GDP. This is to be expected given that the sentiment metrics are naturally bounded by the methods used to generate them.

2.5 | Macroeconomic indicators

Our aim is to test whether sentiment from economic news articles can be valuable for nowcasting. We are therefore very careful to only use data that were available in real time, allowing us to assess the value of our sentiment metrics throughout the data release cycle. We judge performance by comparing with the latest vintage of GDP at the time of writing, as this gives the most reliable estimate of the underlying state of the economy that policymakers need to observe.[13]

The target series in the nowcasting exercise is Euro area real GDP growth, available at a quarterly frequency from Eurostat. In our benchmarking, we use the monthly PMI composite output index published by S&P Global which is one of the most watched survey-based economic indicators. The PMI aggregates activity in manufacturing and services for individual countries and the Euro area as a whole. We also consider historical official ECB macroeconomic projections in real time.

The publication of different indicators and their flash estimates varies across countries and over time. An example data release cycle is shown in Figure 2. For most of our sample period, the Euro area PMI composite index is released as a

[11] As we only have newspapers for four of the 19 Euro area countries, we first re-scale the weights so that they sum to one. As these four economies comprise around 75% of the Euro area's economic activity, this gives us a reliable picture of the Euro area as a whole.

[12] Series for individual countries are shown in Appendix S2.

[13] We show that results do not materially differ when evaluating according to the first vintage of GDP growth (i.e., flash estimates) in Appendix S3.4.

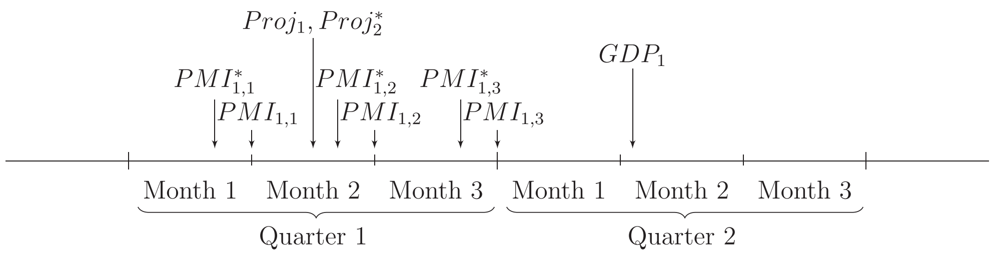

FIGURE 2 Example data release cycle. *Note*: GDP_q shows the release date of GDP growth data for quarter q. $PMI_{q,m}$ shows the release date of PMI composite index for month m of quarter q, and $PMI^*_{q,m}$ shows the corresponding flash estimate. $Proj_q$ denotes the ECB projection for quarter q that is released in quarter q, while $Proj^*_q$ denotes the projection for quarter q released in $q-1$.

flash estimate around the 24th of the month to which it refers (this is the case for Euro area, Germany and France) and the final estimate is released around the beginning of the following month. The ECB's official projections are finalised in the middle of each quarter. We focus on the GDP projections for both the current and next quarters. GDP data for a given quarter are first released during the following quarter and experiences frequent revisions. Since 2016, the first (flash) GDP estimate for the Euro area has been released 1 month after the reference period is over (i.e., 30 days), but before 2016, the publication was 1 month and a half (i.e., 45 days) after the end of the reference period.

For the PMI indicator, we also construct a real-time quasi-daily series that corresponds to the quarterly frequency of GDP. The PMI indicator is available at a monthly frequency, typically published at the start of the next month. However, for the Euro area, Germany and France a flash estimate is published a week before the end of the reference month. Throughout a quarter, there are therefore potentially six release dates.

Our quasi-daily PMI measure is constructed analogously to the daily sentiment measure. At the beginning of the quarter, before any PMI data for that quarter is available, we use the previous month's value. At the end of the quarter, when PMI measures for all three months have been released, we take the mean of these. In between, we take the mean of the latest estimates of all available PMI indicators for that quarter. For example, a few days before the end of the second month the final estimate for Month 1 and the flash estimate for Month 2 are available, so we take the mean of these. In Section 3.3, we also test alternative specifications for the PMI metric in which each month is included separately.

The focus of the analysis is on investigating if textual indicators convey information that complements survey-based sentiment indicators (e.g., PMI). To show that the results are not driven by omitted variables bias, we also consider other macroeconomic indicators which are typically used for nowcasting real GDP growth in the Euro area. In particular, we also include the following monthly Euro area indicators: industrial production, construction production, new passenger car registrations, retail trade, external trade, industrial orders, unemployment rate, surveys of the European Commission and the Purchasing Managers' surveys for services, manufacturing and construction, consumer confidence, loans to the private sector (deflated by HICP), monetary aggregate M1 (deflated by HICP) and two financial indicators: Eurostoxx and corporate spreads (i.e., difference between BBB non-financial corporate bond yields and AAA government bond yields). These macroeconomic and financial indicators are included in a DFM, as discussed in Section 3.4. At all times, we use real-time vintages of all variables.

3 | NOWCASTING SET-UP AND RESULTS

This section shows that text data can improve nowcasts of real GDP growth. Section 3.1 describes the general framework to create text-based and benchmark nowcasts. Section 3.2 illustrates this framework with an example of a simple linear model. Section 3.3 then shows that our news sentiment metrics add value compared with the PMI composite indicator, across a range of models and aggregation approaches. Section 3.4 then shows that the sentiment metrics add value over macroeconomic and financial indicators. Given the extreme fluctuations in GDP growth throughout 2020, we look at this period separately in Section 3.5.

Across models and settings, we can draw a number of conclusions. First, our sentiment metrics are particularly useful in the first half of the quarter, before traditional indicators become available. Second, the text data are useful in the two major crisis periods in our data, but of more limited value in normal times. Third, specialised economics or finance metrics perform well during the Great Recession but poorly in the Great Lockdown. More general-purpose metrics perform well in both periods. Fourth, nonlinearities are an important feature during turbulent times, and incorporating them improves performance during the Great Lockdown period. However, as nonlinear methods require a longer training sample to be reliable, they do not perform as well in the Great Recession.

3.1 | The econometric framework

While we test many different forecasting models, the framework that we use to assess the usefulness of the text at a daily frequency is the same across specifications. In every case we produce a daily nowcast using our text metrics and a range of benchmarks. Models are trained only using data that were available in real time and are then assessed on their ability to predict quarter-on-quarter real GDP growth (vintage as of 24 March 2021). This assessment is carried out separately for each day of the quarter, allowing us to show when in the data release cycle the text is most useful.

Our nowcasting models use a vector of indicators available on day d of quarter q ($x_{q,d}$) to produce daily nowcasts of GDP growth in quarter q ($\hat{y}_{q,d}$).

$$\hat{y}_{q,d} = g\left(x_{q,d}, \theta, \eta\right), \tag{2}$$

where θ is a vector of estimated parameters and η is a vector of hyperparameters that may be chosen by cross-validation. The function $g(\cdot)$ varies with the model at hand.

We use several real-time benchmarks to assess the usefulness of the text metrics. Most straightforwardly, we compare nowcasts for a model including text metrics to the same functional form but including only the PMI indicator and not the text metric. For the PMI model the hyperparameters η are cross-validated in the same way as the text-based model. PMI is often seen as the gold standard for soft indicators and, as such, is used as a competitive and relevant benchmark. Additionally, we compare nowcasts to the latest available ECB projection for GDP growth, as described in Section 2.5. These projections represent the synthesis of all available data and include expert judgement so are a very challenging benchmark. Finally, in Section 3.4, we compare models with text to those including macroeconomic and financial indicators.

Beginning 1 April 2006, as this is the first date for which we have real-time vintages of GDP and PMI, both the text and benchmark models are re-trained each day. For a given day, we ensure that data vintages that were available in real time are used, training on an expanding window of data for which at least one vintage of GDP has been published. Where data are standardised, we also take care to only use data available in real time for this standardisation. The latest vintages of GDP that were available on the day of the nowcast are used to compute the growth rates that the model is trained on.

We evaluate model's nowcasting performance throughout the quarter based on the error between each daily nowcast and the target variable. We compute a mean squared error (MSE) for each day of the quarter across the out-of-sample period (or a subset of the out-of-sample period). So if there are Q quarters in the out-of-sample period, we would compute the MSE for the dth day of each quarter as

$$MSE_d = \frac{1}{Q}\sum_{q=1}^{Q}\left(y_q - \hat{y}_{q,d}\right)^2, \tag{3}$$

where y_q is the target variable for quarter q (calculated using the latest vintages available on 24 March 2021) and $\hat{y}_{q,d}$ is the nowcast for y_q produced on the dth day of the quarter. We can thus assess the nowcasting performance of a given model on a daily basis throughout the quarter. To test whether the difference in performance between models is statistically significant, we use the Diebold and Mariano (1995) test with Harvey's correction for short samples (Harvey et al., 1997).

3.2 | An illustrative linear case

To illustrate how we evaluate the usefulness of sentiment metrics, we use a simple linear model. We exclude the year 2020 from this example as its extreme GDP growth rates dominate any comparison, so 2020 is examined in detail in Section 3.5. The text model is thus

$$g_{text}(x_d, \theta, \eta) = \theta_0 + \theta_1 PMI_{q,d} + \theta_2 sent_{q,d}, \tag{4}$$

where $sent_{q,d}$ is the sentiment metric and $PMI_{q,d}$ is the quasi-daily PMI metric. The PMI model used as a benchmark is therefore

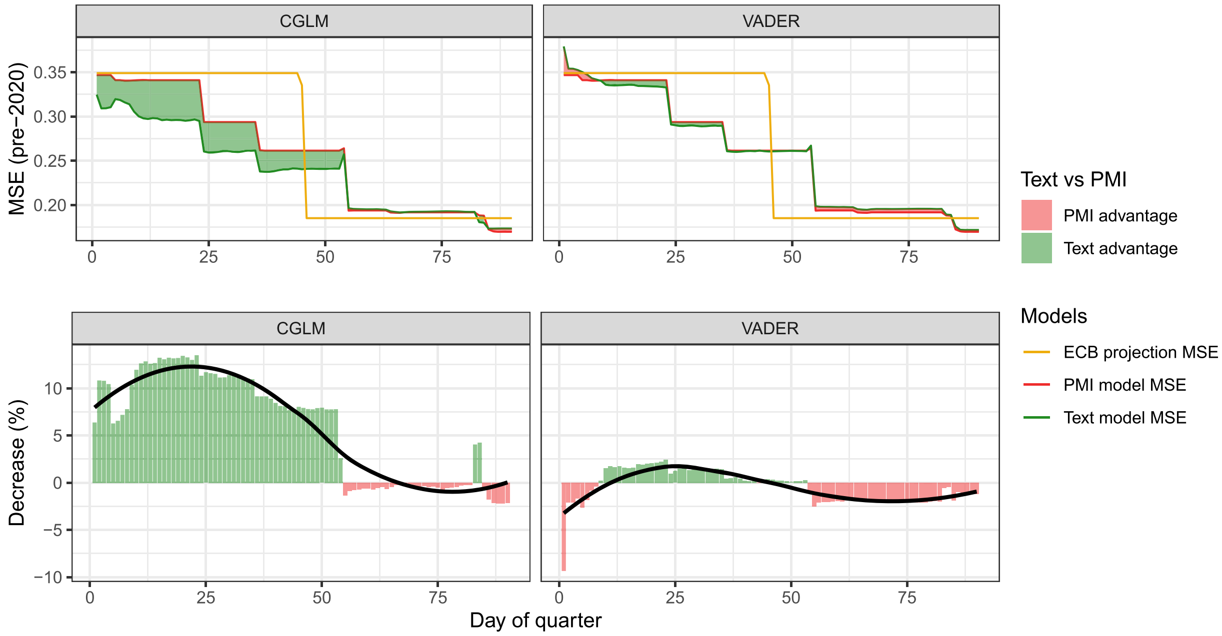

FIGURE 3 Illustrative example: pre-2020. *Note*: Upper panels show the MSE for each model for each day of the quarter, calculated as in Equation (3). This MSE is shown in green for the text model, red for the PMI benchmark and yellow for the latest available ECB projection. The difference between the MSE of the text model and PMI model is coloured green if the text model performs better at that stage and red if the PMI model performs better. The lower panels show the percentage improvement (i.e., decrease) in MSE of the text model compared with the PMI benchmark, for each day of the quarter. The black line on the lower panels shows a local polynomial regression of this decrease on the day of the quarter, estimated with the *stats* package in R. Results for CGLM are shown on the left with VADER on the right, and the sample period is April 2006 to December 2019.

$$g_{pmi}(x_d, \theta, \eta) = \theta_0 + \theta_1 PMI_{q,d}. \quad (5)$$

The model is estimated by ordinary least squares (OLS) so there are no further hyperparameters. In order to ensure that only comparable observations are used, we restrict the training set to observations at the same stage of the data release cycle (shown in Figure 2). For example, if the PMI for the first 2 months are available, we train the model only on observations from days on which PMI for the first 2 (of 3) months are available. This is intuitively similar to estimating a separate model for each day of the quarter but avoids the issue of data being released on slightly different days across quarters (e.g., because of weekends or leap years).

Figure 3 compares the forecasting performance of the text model, ECB GDP projections and the PMI benchmark for the period between April 2006 to December 2019. The left panel shows the performance with the CGLM metric and the right panel with the VADER metric. In each case, the upper panel shows the MSE for the ECB projection (in yellow), the PMI model (in red) and the text model (in green) for each day of the quarter across the sample period.[14] As expected, the nowcast error for all models generally decreases throughout the quarter as more data become available. Both text models and the PMI model outperform the first ECB projection, but once the second projection becomes available halfway through the quarter, they are less competitive. The text models perform better than the PMI model in the first half of the quarter, although this difference is more substantial for the CGLM text metric than for VADER.

As mentioned above, we find that text is particularly useful in crisis periods, while it has a more limited value when growth is fairly constant. Figure 4 illustrates this by showing the performance of the nowcasting models for the Great Recession, considering the period between April 2006 to December 2009. For both metrics, the improvements in this period are greater than for the pre-2020 period as a whole. The lower panels show the percentage improvement of the text model compared with the PMI benchmark. We get substantial improvements when using the text in the first half of the quarter, in particular until the flash estimate of PMI for the second month of the quarter is released.

[14]The PMI model is updated only periodically, as new PMI data become available, while the text model is updated on a daily basis as new articles are published, as well as when new PMI data become available.

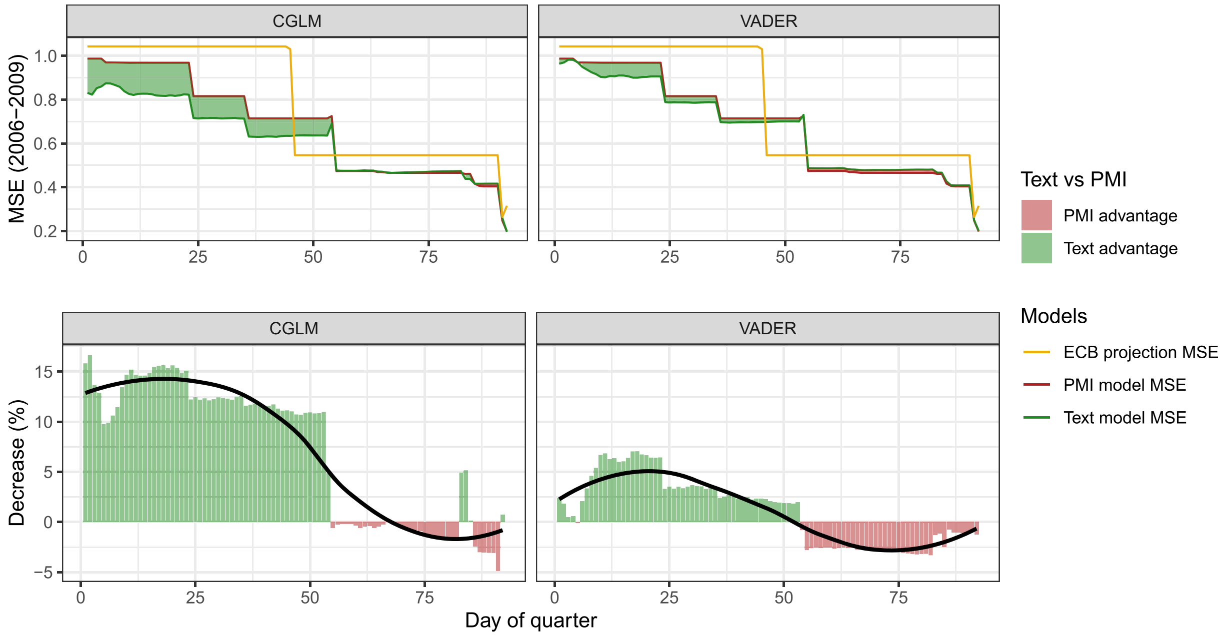

FIGURE 4 Illustrative example: Great Recession. *Note*: See Figure 3, but the sample period is restricted to April 2006 to December 2009.

3.3 | Text versus PMI indicator

In the previous illustrative example, we used the proposed cumulative aggregation approach to convert the daily text metric and monthly PMI indicator into quarterly indicators and simple linear regression models. However, our key results are robust to a range of alternative aggregation approaches for both the text data and the PMI indicator. They are also robust to a range of alternative modelling approaches, including MIDAS (Ghysels et al., 2004) and a range of ML models. We find that, with the exception of ridge regression, alternative models and aggregations do not add a great deal of value over the simple linear regression in our illustrative example. We conclude that the simple and intuitive cumulative average does miss any important dynamics of news sentiment necessary for nowcasting GDP, so focus on this metric going forward.

3.3.1 | Alternative aggregation approaches

In our linear regression models, the high-frequency text data are aggregated for the quarter implicitly applying equal weighting to each value in the average. Alternatively, the MIDAS regression (Ghysels et al., 2004) is popular approach for aggregating high-frequency indicators which allows for non-equal weights but keeping the number of regression coefficients tractable by fitting polynomial functions to the lags of the higher frequency indicators. We follow this approach and fit Legendre polynomials as in Babii et al. (2022), given better numerical properties, and include 90 lags of the daily news sentiment metrics.[15]

Figure 5 shows the MSE throughout the quarter of OLS and MIDAS models with alternative ways of aggregating the higher frequency variables (i.e., text and the PMI metrics). For the OLS regressions, in addition to the daily cumulative average discussed before, we test a 30-day rolling average, monthly averages for all available months of the current quarter, including lags and differences. Furthermore, in addition to the quasi-daily PMI metric described previously, we include the latest vintage of PMI for each month of the quarter as soon as they become available. Across all of these specifications, we find remarkably similar results for both the entire pre-2020 period (Figure 5) and the Great Recession (Figure C.7 in Appendix S3.1). In all cases, CGLM improves nowcast performance in the first half of the quarter relative to just the PMI indicators. For VADER, gains are smaller and more sensitive to the specification, but in most cases, we see small improvements during the Great Recession period.[16]

[15]Results are robust to alternative specifications using different number of lags, Almon polynomials and daily rather than monthly data for the PMI metric, as shown in Appendix S3.2.

[16]For the ECONLEX dictionary, we generally find that results are between those of CGLM and VADER, with results shown in Appendix S3.3.

WILEY — Journal of APPLIED ECONOMETRICS

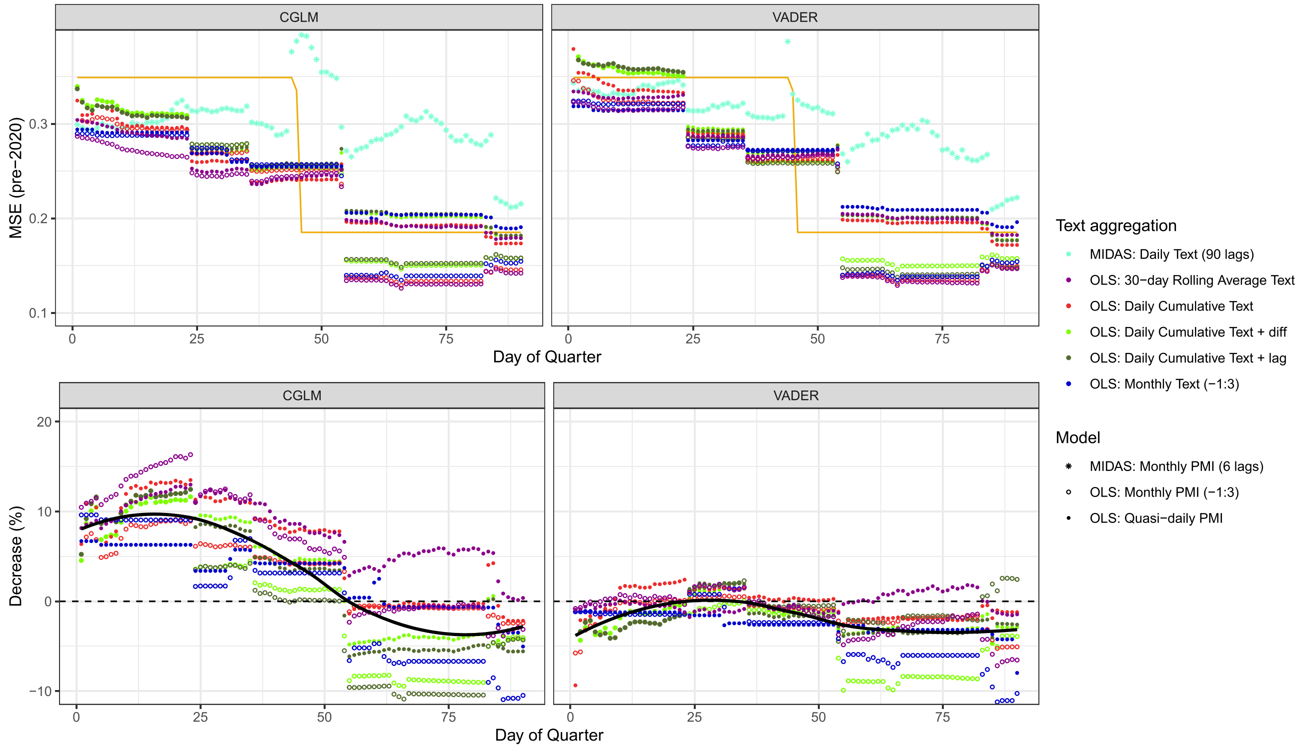

FIGURE 5 Alternative aggregation and modelling approaches for text and PMI (pre-2020). *Note*: Panels here are analogous to those in Figure 3. The upper panels show the MSE for each day of the quarter across the sample period (April 2006 to December 2019). The performance of the ECB projection (in yellow) is shown for reference. The colour of each point indicates how the news sentiment metrics are aggregated, while the shape of each point indicates the model and how the PMI metric is included. Further details on all metrics and models are given in Appendix S3.1. The lower panels show the decrease in MSE (in percentage terms) for the model including the text compared with the model with just the PMI metric. The MIDAS model is omitted from the lower panels due to the volatility of performance across the quarter but, as is shown in Appendix S3.2. The black line on the lower panels shows a local polynomial regression of this decrease on the day of the quarter. Results in the left-hand panels are for the CGLM metric while results in the right-hand panel are for VADER.

Nowcasting performance using the MIDAS approach throughout the quarter is shown as light blue asterisks in the upper two panels of Figure 5. The MIDAS models' performance is similar to our range of aggregated metrics in an OLS regression during the first 24 days of the sample, until the first release of PMI becomes available, but perform considerably worse as the quarter goes on. More specifically, CGLM does add value over the PMI in the first 24 days of the quarter where the MIDAS models are competitive with OLS, as does VADER in the Great Recession period. Most of the aggregated metrics in an OLS regression perform better than our MIDAS models with and without text throughout the whole quarter. Based on this, we conclude that our aggregation approaches likely do not miss any important dynamics of news sentiment that are relevant to nowcasting GDP.[17] To this end, we will focus on our daily cumulative metric for the remainder of the paper, given its simplicity and intuitive linking of daily and quarterly frequencies.

3.3.2 | Alternative modelling approaches

In addition to the OLS and MIDAS frameworks, we test several alternative models from the ML literature. A brief explanation of each model is given in Appendix S3.5, but they include a range of commonly used ML methods from a relatively straightforward ridge regression to random forests and neural networks that allow for arbitrary nonlinearity. We thus address two common critiques of previous studies using text data for forecasting: (i) that the modelling framework and benchmarks chosen were relatively simple and (ii) that it is likely that the relationship between economic activity and 'soft' indicators is nonlinear (Kalamara et al., 2022; Woloszko, 2020).

[17]Indeed it appears that that news sentiment is too volatile at a daily frequency for a MIDAS approach to be useful. Babii et al. (2022) find text data are useful in prediction when aggregated to a monthly frequency before estimating the MIDAS model.

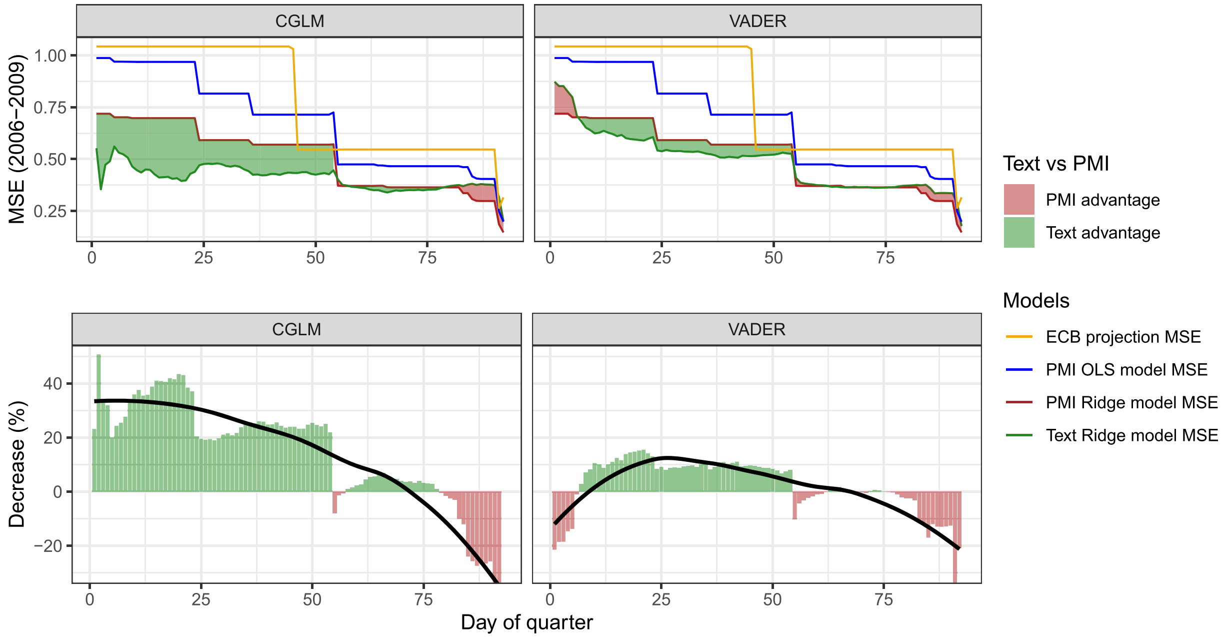

FIGURE 6 Ridge regression in Great Recession. *Note*: See footnote in Figure 3. Results here are for Ridge regressions and the sample period is April 2006 to December 2009.

For our pre-2020 sample, we find that the only ML model that performs better than the ECB projections is a ridge regression (see Appendix S3.5). This finding is in line with Giannone et al. (2021), who found that given pervasive model uncertainty in economics, denser predictive models such as ridge regression tend to outperform more sparse models. On the other hand, the relationship between our news sentiment indicator and GDP growth is certainly nonlinear, but it appears that our sample period is too short for models to learn this before 2020. Although the ML models do not perform very well in general, we do see that text models outperform their PMI only counterpart in all cases except for the neural network, so our finding that the text metrics add value over the PMI indicator is further supported. The above findings are further supported by a Diebold and Mariano test for statistical significance. We present the results for VADER and CGLM models in Appendix S3.7 for the whole pre-2020 period on a monthly basis.

3.3.3 | Great Recession and its aftermath

Investigating deeper, the advantage of the ridge regression including high-frequency text-based sentiment for nowcasting GDP seems to come from the Great recession period. Figure 6 shows the evolution of the MSE throughout quarters for the CGLM and the VADER text metrics using a ridge regression during the Great Recession period (1 April 2006 to 31 December 2009). The text-based ridge regression outperforms the ECB projections and the OLS benchmark. During the Great Recession period, the CGLM model consistently improves the daily nowcasts from the first days within the average quarter and up to the end of the second month. The advantage of the timeliness of the text is also apparent when using the VADER model but the magnitude of the MSE drop is smaller compared with the PMI model. In both cases, text-based models are quite helpful in the first 2 months before the releases of other indicators. Moreover, introducing regularisation through the ridge regression improves performance relative to OLS, and the added value of the CLGM metric is as high as 40% at the beginning of the quarter (and around 15% for VADER).

Furthermore, looking at the actual daily nowcasts in Figure 7, we observe that text is particularly helpful in predicting the start of the negative from 2008Q2 onwards, but less so in predicting the subsequent recovery, providing evidence that sentiment can provide early warning signals of future economic disruptions (Huang et al., 2019; Nyman et al., 2018). In fact, the negative jump on the first day of the 2008Q1 actually predicts that a recession will begin too early, as growth was actually still positive. We can see here that the superior performance of the CGLM metric compared with VADER is mainly due to it predicting a faster recovery. In the quarters with negative growth, VADER actually does slightly better than CGLM.

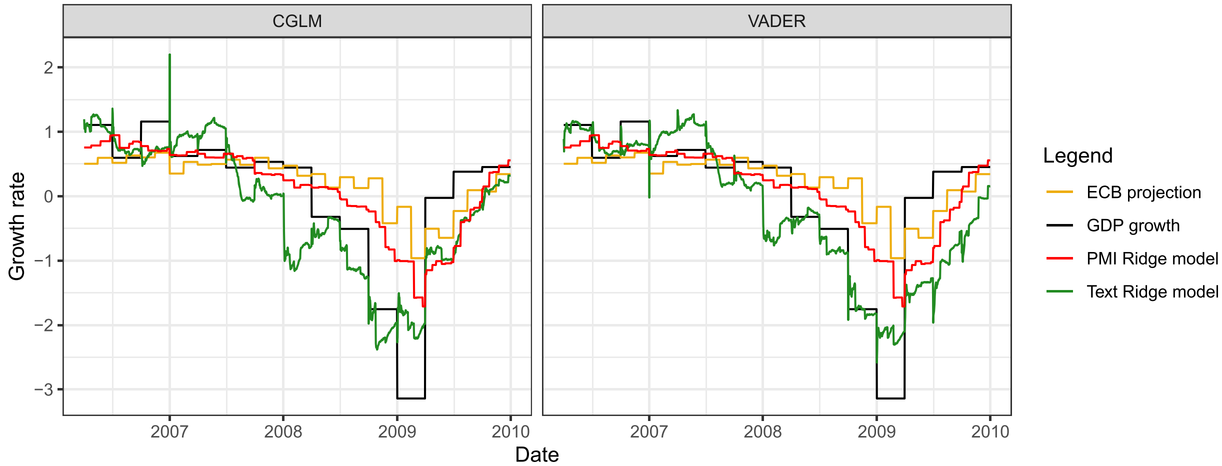

FIGURE 7 Nowcasts in the Great Recession. *Note*: This figure compares the real-time nowcasts of text-based PMI ridge regression models from April 2006 to December 2009.

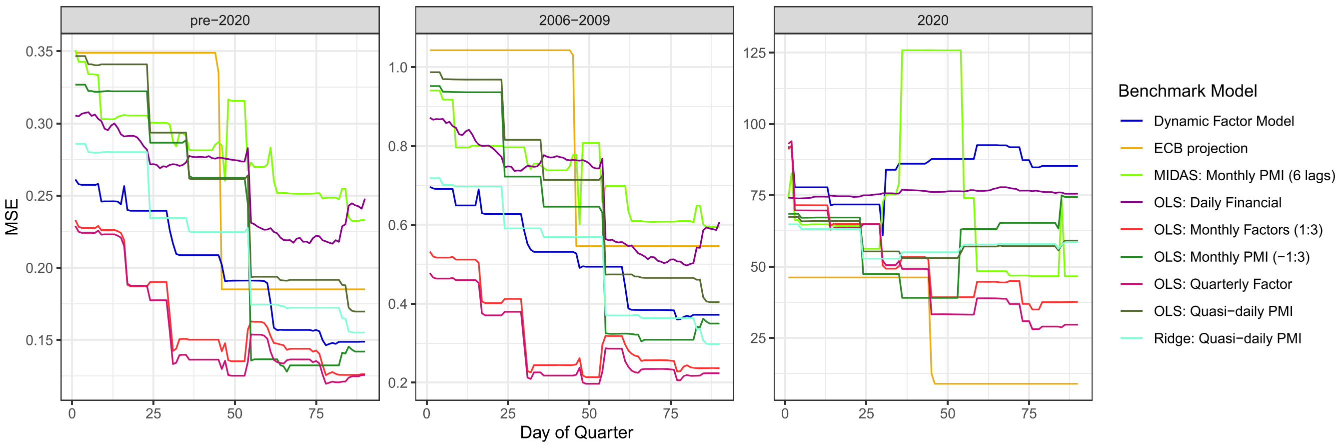

FIGURE 8 Comparing benchmark model performance. *Note*: Average MSE for each day of quarter shown for a range of benchmark models that do not include text variables. The left panel evaluates performance across the whole pre-2020 period, the central panel focuses on 2006–2009 and the right panel shows performance in 2020.

3.4 | Text versus macroeconomic and financial indicators

We have compared text nowcasts with benchmarks based on the ECB projection and the PMI indicator, but there are of course many sources of real-time information that may be useful in nowcasting and there is no guarantee that the ECB projections have historically made the best use of this information. We therefore include a range of macroeconomic and financial indicators in our benchmark models and demonstrate that text can add value even when all of this extra information is included. We use real-time vintages of all variables throughout.

Figure 8 shows the performance of a range of benchmark models that do not contain any information from the text data. We look at three periods separately, with the whole pre-2020 period on the left, the Great Recession period in the centre and the Great Lockdown on the right. In yellow, we have the performance of the ECB projections (as was also shown in previous figures). It is worth noting that in 2020, the ECB projections outperformed all of the benchmarks, as we will discuss in more detail in Section 3.5. A range of benchmark models are shown in the other colours. These are described in more detail in Appendix S3.6, but most importantly, we can note that the best performing benchmark across the pre-2020 and Great Recessions periods is an OLS regression that includes monthly factors. These monthly factors summarise the macroeconomic, financial and survey indicators (including PMI) described in Section 2.5 and are estimated with a mixed-frequency DFM.

The DFM model is based on that in Doz et al. (2011, 2012), and as in all of our analysis, we make sure to only use information that was available in real time. The factor model is specified as one common factor which follows an AR(2)

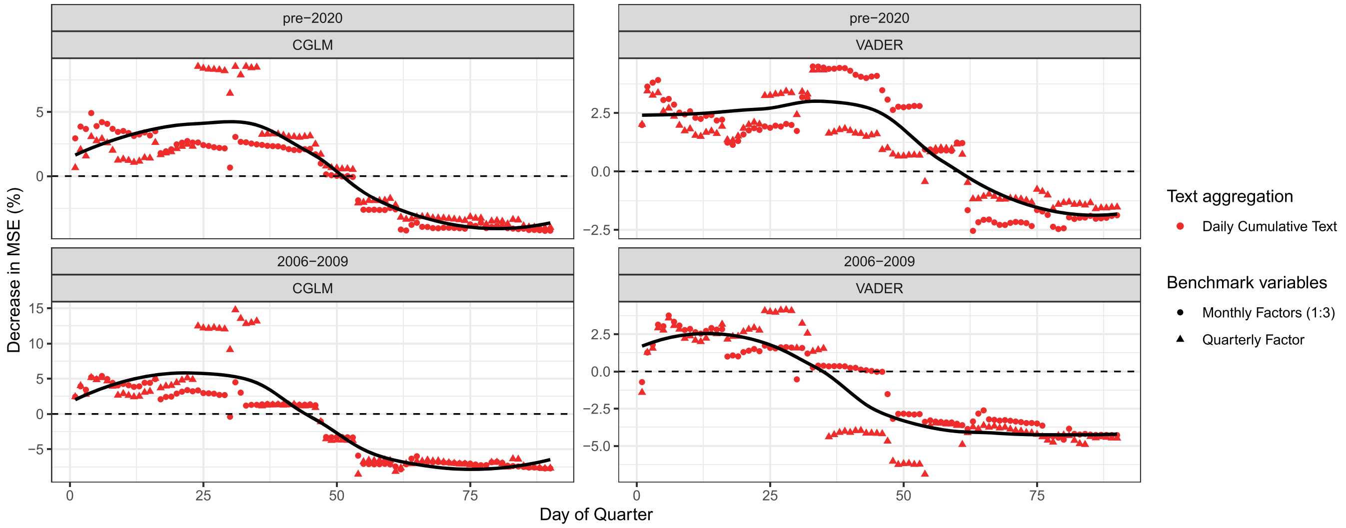

FIGURE 9 Comparing text (daily cumulative metric) and common factor. *Note*: Decrease in MSE (in percentage) of an OLS regression including both textual indicators and a common factor in relation to an OLS regression that does not include text variables. The decrease in MSE is shown for each day of quarter.

model, and the idiosyncratic components are serially correlated of order 1. The model is estimated with the expectation maximisation (EM) algorithm and the Kalman filter and smoother to deal with missing observations and handle the 'ragged edge' problem as in Bańbura and Modugno (2014). To avoid data leakage, the entire model including all of the factor loadings is re-estimated whenever a new vintage of data becomes available.

The errors that the DFM makes are correlated with the daily sentiment metrics throughout the sample period, at 0.281 for CGLM and 0.220 for VADER (both p-values $\ll 0.001$). This positive correlation is also consistent across the quarter, as shown in Figure C.13 (Appendix S3.6).

The performance of the DFM itself is shown in Figure 8 in blue, but we find that we get better performance in an OLS regression on this factor, as shown in red. This is likely explained by the fact that factor loadings can be rather unstable. However, as shown by Stock and Watson (2009), if the instability is sufficiently independent across series then the estimated factors could be well estimated even if individual relations between the observable series and the factors are unstable. This OLS regression on the factor is our best performing benchmark across the quarter in the pre-2020 sample.

The forecast based on the common factor includes a rich set of information which goes far beyond sentiment, and so alongside the ECB short-term projections, comprise a very demanding benchmark. To assess whether text data can add value compared with this benchmark, we compare an OLS regression with the factor to an OLS regression with the factor and our text metrics. Figure 9 illustrates this comparison for the daily cumulative text metric.[18] Although the gains are smaller than when compared with just PMI, there is still a decrease in MSE for most aggregation methods during the first half of the quarter for both CGLM and VADER. Across all of the benchmarks shown in Figure 8, both CGLM and VADER add value in most cases during the first half of the quarter.

3.5 | Nonlinearities and the Great Lockdown

We have seen that text metrics can be helpful in crisis periods when economic conditions can be subject to sudden and rapid changes. In this regard, the coronavirus outbreak serves as an interesting illustration to demonstrate the applicability of our text metrics and different methodologies. The COVID-19 pandemic was an unprecedented and unanticipated economic shock. Accordingly, we see in Figure 8 that our range of benchmark models perform poorly, even those based on a wide range of macroeconomic, financial and survey indicators. By far the best performing forecasts are the ECB projections, which are based on a combination of wide range of models and expert judgement. For 2020, we therefore focus on whether text models can outperform the ECB projection.

[18] Results are broadly similar for most alternative aggregations shown in Figure C.14.

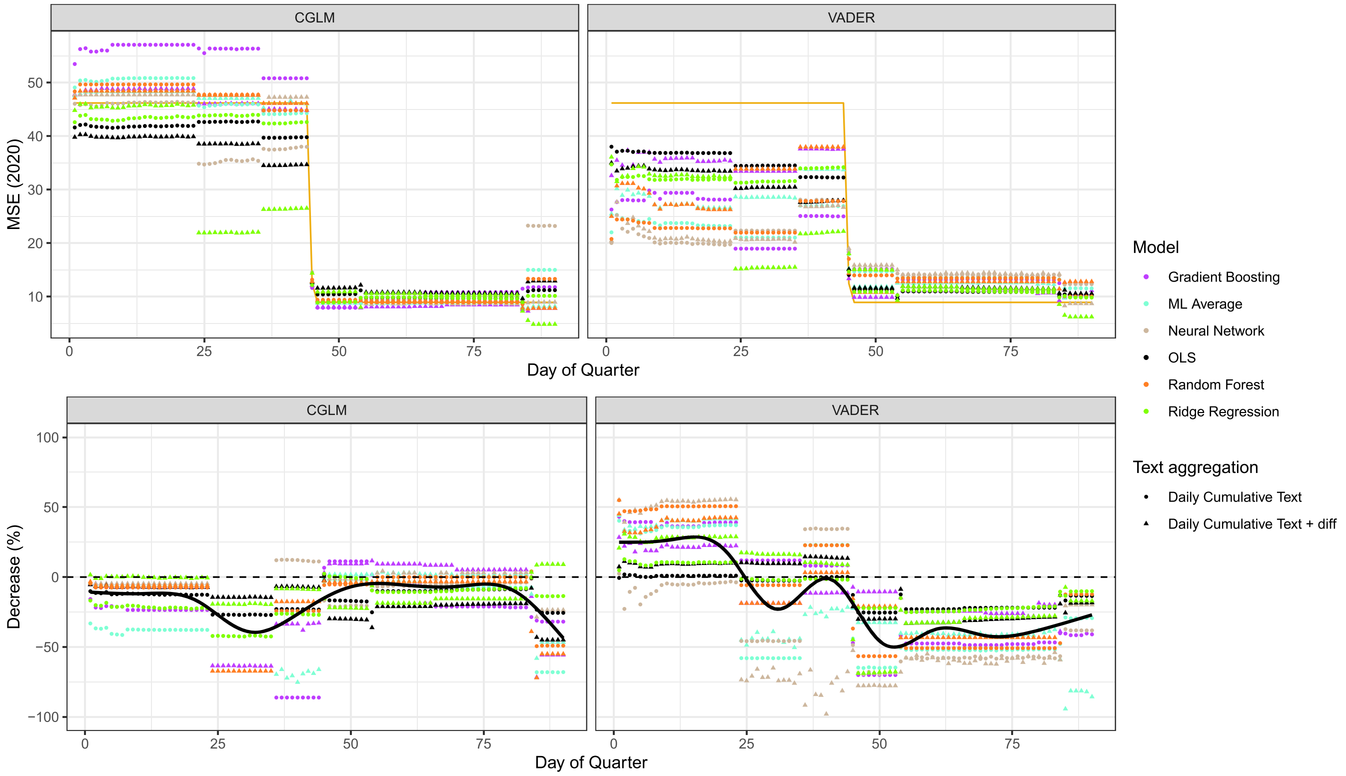

FIGURE 10 Comparing text and PMI in 2020, with ECB projection error target. *Note*: This figure has the same set-up as Figure 5. Across all panels, the colour points indicates the model used while the shape indicates how the text is aggregated. The decrease in MSE shown in the lower panels is for the text model relative to whichever performs best out of ECB projections and the same model with only the PMI metric.

When our nowcasting models target GDP growth directly, both our benchmarks and text models generally perform worse than the ECB projections, although the text metrics still add value over PMI (see Figure C.15 in Appendix S3.8). This poor performance relative to the ECB projections is largely due to the Q3 of 2020 where, after an extremely sharp contraction in Q2, GDP growth rebounded strongly reaching 12.5%. The ECB projections, due to their reliance on expert judgement and a suite of models, did capture at least part of this rebound. However, when we target the ECB projection forecast error rather than GDP growth directly, we find that text models outperform both their PMI counterparts and the ECB projections. Crucially, throughout all exercises we only use projections and data that would have been available at that date in real time.

Figure 10 shows the 2020 performance of various nowcasting models that target the ECB's projection error, using only information that was available in real time. Although our nonlinear ML models did not improve performance in the pre-2020 period, we find that they do add value in 2020. This is for two reasons. First, the fluctuations in GDP throughout 2020 were very extreme and so the relationship between indicators and GDP likely nonlinear. Second, as we use only real-time data in training our models, by 2020, our nonlinear models have a much longer sample period on which to be trained including a significant downturn.

Across a range of models, the general sentiment dictionary VADER shows notable improvements during the first month of the average quarter and the errors are consistently lower in relation to both the ECB projections and the corresponding PMI models. This contrasts with behaviour of the same metrics and specification during the Great Recession. The pandemic shock had a non-economic nature and triggered a global economic crisis and very strong policy support. As captured by the CGLM metric and in contrast to previous crises, the behaviour of the financial markets and the financing conditions remained favourable through 2020. For forecasting purposes, looking only at financially relevant terms is therefore not sufficient to capture the dynamics of this disruptive event.

To better understand the differing performance of the general-purpose and finance-focused dictionaries in 2020, we show the most important terms in 2020 for each of CGLM, ECONLEX and VADER in Figure 11. There are clearly different terms picked up across the different dictionaries, and although interpreting this is inherently subjective, a few key points stand out. Terms like ‘crisis’, ‘pandemic’, ‘emergency’ and ‘risk’ which drive the negative parts of ECONLEX and VADER

Negative	Negative	Negative	Positive	Positive	Positive
CGLM	ECONLEX	VADER	CGLM	ECONLEX	VADER
difficult	crisis	crisis	able	recovery	great
problem	pandemic	no	recovery	support	like
lost	risk	emergency	good	great	agreement
problems	virus	debt	better	good	hand
negative	against	risk	positive	agreement	support
losses	epidemic	difficult	benefit	aid	good
closure	difficult	negative	opportunity	positive	increase
fear	loss	war	success	able	well
forced	decline	problem	effective	strong	best
bad	war	bad	grow	addition	growth
challenge	unemployment	fear	improve	safety	united
concerns	collapse	unemployment	recover	opportunity	free

FIGURE 11 Most important terms within each metric in 2020. *Note*: This figure shows the 12 most important positive and negative terms for each of the CGLM, ECONLEX and VADER metrics in only 2020. The size of the text for each term indicated their frequency, weighted by the intensity assigned to that term by the relevant dictionary. For CGLM, words are either positive or negative, while for ECONLEX and VADER, terms are graded on a continuous scale, which is thus reflected in the weighting. The terms in the largest font size are thus those which have the greatest positive or negative impact on that metric in 2020 alone. For VADER, these weighting do not exactly reflect the importance of each term, as there are further rules applied around negation and intensity (see Appendix S1.4) but they are still a good guide to the terms that drive the overall score.

do not appear in CGLM. Similarly, words that drive the positive side of ECONLEX and VADER like ‘great’, ‘agreement’ and ‘support’ do also not appear in CGLM. While there may be some words that are misclassified in the finance-focused dictionaries (e.g., ‘positive’ and ‘recovery’ could be referring to coronavirus infections in many 2020 articles), we suspect that the comparatively poor performance is more likely explained by the fact that CGLM simply includes many fewer terms in its lexicon. For example, the version of VADER we use includes 7520 tokens and ECONLEX has 6670, while CGLM has only 391. The differing effect does not appear to be driven by terms having opposite signs across the different dictionaries: There are only three terms that are marked as negative in CGLM but positive in VADER (‘challenge’, ‘challenges’ and ‘challenging’) and only one term marked as positive in CGLM and negative in VADER (‘preventing’). An additional point of difference is that the general-purpose dictionaries weight terms along a continuous scale rather than a binary classification, so this additional nuance may explain the better performance as unambiguously terms (e.g., ‘crisis’) get a higher weighting than more ambiguous terms (e.g., ‘challenge’).

The daily nowcast series themselves for the PMI model and the average across ML models are shown in Appendix S3.8, where there are several points of note. All text metrics capture the drop of GDP growth in the beginning of 2020Q2 earlier than PMI models do. Additionally, the depth of the drop is larger when we include the text-based sentiment metric in levels, without imposing any further specification. This emphasises the timeliness advantage of text-based indicators compared with other indicators. However, as we progress throughout the year, it is the alternative specifications with projections and first differences that are able to capture the great rebound of GDP growth in Q3. This holds for models using both data types (i.e., text and non-text) suggesting that there is not a single standard mechanism which can describe how rapid changes occur but rather a combination of different specifications and data selection.

4 | CONCLUSIONS

This paper shows that machine-translated newspaper text can provide information about current economic outlook for the Euro area that is relevant to policymakers. Our results show that daily text signals can substantially improve GDP nowcasts, especially during crisis periods and over the first half of the quarter. This improvement is relative to competitive and rigorous benchmarks such as the ECB's official projections, models using the PMI composite index and models including a range of real-time macroeconomic and financial indicators. The results are robust to a range of alternative aggregation approaches and model types.

We also show how two commonly used sentiment metrics were able to add value to a variety of nowcasting models. As well as being useful proxies for sentiment, we show that the choice of the sentiment measure matters. While the financial stability-based dictionary of Correa et al. (2017) performs strongly during the Great Recession, it is not successful during the Great Lockdown. Nonetheless, during the COVID-19 pandemic, its evolution is consistent with the behaviour of the financial markets and the financing conditions which remained favourable in the context of very strong policy response. On the other hand, a general-purpose sentiment indicator such as VADER (Hutto & Gilbert, 2014) is more resilient especially in unexpected economic episodes such as the COVID-19 pandemic.

We also test the forecasting performance of a suite of nonlinear ML methods in a real-time setting. Nonlinear ML models respond more flexibly to the coronavirus outbreak and combining them with text information and expert judgement from the ECB projections provides the best combination on tracking the sudden drop on 2020Q2 and the asymmetric rebound on 2020Q3 in the Euro area. However, during normal times, their value in the nowcasts is not larger than using a linear model.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Beatrice Pierluigi, Grigor Stoevsky, Niccol Battistini, Roberto de Santis, Chiara Osbat, Michele Lenza, Eric Ghysels and participants at the ECB workshop on Applications of advanced analytics to monitoring economic conditions and the 16th International Conference on Computational and Financial Econometrics for valuable comments and suggestions. We are grateful to George Kapetanios and Stephen Hansen for helpful discussions at the beginning of this project. The views expressed in this paper are solely those of the authors and do not necessarily represent the views of the European Central Bank.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at https://doi.org/10.15456/jae.2024079.1245878012.

DATA AVAILABILITY STATEMENT

Replication data are available here: https://journaldata.zbw.eu/dataset/nowcasting-euro-area-gdp-with-news-sentiment-a-tale-of-two-crises-replication-data, with code available upon request.

REFERENCES

Aguilar, P., Ghirelli, C., Pacce, M., & Urtasun, A. (2021). Can news help measure economic sentiment? An application in COVID-19 times. *Economics Letters*, *199*, 109730.

Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys*, *34*(3), 512–547.

Algaba, A., Borms, S., Boudt, K., & Verbeken, B. (2023). Daily news sentiment and monthly surveys: A mixed-frequency dynamic factor model for nowcasting consumer confidence. *International Journal of Forecasting*, *39*(1), 266–278.

Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., & Monteforte, L. (2022). The power of text-based indicators in forecasting Italian economic activity. *International Journal of Forecasting*, *39*(2), 791–808.

Ardia, D., Bluteau, K., & Boudt, K. (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., Ghysels, E., & Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, *40*(3), 1094–1106.

Baffigi, A., Golinelli, R., & Parigi, G. (2004). Bridge models to forecast the euro area GDP. *International Journal of Forecasting*, *20*(3), 447–460.

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

Bańbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting with daily data. In *2012 Meeting Papers 555*, Society for Economic Dynamics.

Bańbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, *29*(1), 133–160.

Bannier, C., Pauls, T., & Walter, A. (2019). Content analysis of business communication: Introducing a German dictionary. *Journal of Business Economics*, *89*(1), 79–123.

Barbaglia, L., Consoli, S., & Manzan, S. (2024). Forecasting GDP in Europe with textual data. *Journal of Applied Econometrics*, *338–355*.

Barbaglia, L., Consoli, S., & Manzan, S. (2023). Forecasting with economic news. *Journal of Business & Economic Statistics*, *41*(3), 708–719.

Barbaglia, L., Consoli, S., Manzan, S., Pezzole, L. T., & Tosetti, E. (2022). Sentiment analysis of economic text: A lexicon-based approach. Available at SSRN 4106936.

Bortoli, C., Combes, S., & Renault, T. (2018). Nowcasting GDP growth by reading newspapers. *Economie et Statistique*, *505*(1), 17–33.

Casarin, R., & Squazzoni, F. (2013). Being on the field when the game is still under way. The financial press and stock markets in times of crisis. *PLoS ONE*, *8*(7), e67721.

Casnici, N., Dondio, P., Casarin, R., & Squazzoni, F. (2015). Decrypting financial markets through e-joint attention efforts: on-line adaptive networks of investors in periods of market uncertainty. *PLoS ONE*, *10*(8), e0133712.

Cimadomo, J., Giannone, D., Lenza, M., Monti, F., & Sokol, A. (2022). Nowcasting with large Bayesian vector autoregressions. *Journal of Econometrics*, *231*(2), 500–519.

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2017). Constructing a dictionary for financial stability. *Board of Governors of the Federal Reserve System*, *6*(7), 9.

Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? arXiv preprint arXiv:2008.12477.

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Doz, C., Giannone, D., & Reichlin, L. (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, *94*(4), 1014–1024.

Foroni, C., & Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. *International Journal of Forecasting*, *30*(3), 554–568.

Gajewski, P. (2014). Nowcasting quarterly GDP dynamics in the euro area: The role of sentiment indicators. *Comparative Economic Research*, *17*(2), 5–23.

Galbraith, J. W., & Tkacz, G. (2018). Nowcasting with payments system data. *International Journal of Forecasting*, *34*(2), 366–376.

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

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437.

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

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 168–177.

Huang, C., Simpson, S., Ulybina, D., & Roitman, A. (2019). News-based sentiment indicators. International Monetary Fund Working Paper 19/273.

Hutto, C., & Gilbert, C. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225.

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

Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339–354.

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

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

Loughran, T., & McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, *109*(2), 307–326.

Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the US unemployment rate using Google Trends. *Finance Research Letters*, *30*, 103–109.

Nielsen, F. A. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., & Smith, R. (2018). News and narratives in financial systems: Exploiting big data for systemic risk assessment. Bank of England Staff Working Papers 704.

Rambaccussing, D., & Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting*, *36*(4), 1501–1516.

Shalunts, G., Backfried, G., & Commeignes, N. (2016). The impact of machine translation on sentiment analysis. *Data Analytics*, *63*, 51–56.

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

Stock, J., & Watson, M. W. (2009). *The methodology and practice of econometrics: A Festschrift in honour of David F. Hendry*, pp. 1–57: Oxford University Press.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The 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 and Economic Statistics*, *38*(2), 1–17.

Woloszko, N. (2020). A weekly tracker of activity based on machine learning and Google Trends. (*Working Papers 1634*): OECD Economics Department: OECD Publishing.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Ashwin, J., Kalamara, E., & Saiz, L. (2024). Nowcasting Euro area GDP with news sentiment: A tale of two crises. *Journal of Applied Econometrics*. 1–19. https://doi.org/10.1002/jae.3057