Ashwin, Julian; Kalamara, Eleni; Saiz, Lorena

Working Paper

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

ECB Working Paper, No. 2616

Provided in Cooperation with:
European Central Bank (ECB)

Suggested Citation: Ashwin, Julian; Kalamara, Eleni; Saiz, Lorena (2021) : Nowcasting euro area GDP with news sentiment: A tale of two crises, ECB Working Paper, No. 2616, ISBN 978-92-899-4869-2, European Central Bank (ECB), Frankfurt a. M., https://doi.org/10.2866/240669

This Version is available at:
https://hdl.handle.net/10419/249889

EUROPEAN CENTRAL BANK

EUROSYSTEM

Working Paper Series

Julian Ashwin, Eleni Kalamara, Lorena Saiz

Nowcasting euro area GDP with news sentiment: a tale of two crises

No 2616 / November 2021

Abstract

This paper shows that newspaper articles contain timely economic signals that can materially improve nowcasts of real GDP growth for the euro area. Our text data is drawn from fifteen popular European newspapers, that collectively represent the four largest Euro area economies, and are machine translated into English. Daily sentiment metrics are created from these news articles and we assess their value for nowcasting. By comparing to competitive and rigorous benchmarks, we find that newspaper text is helpful in nowcasting GDP growth especially in the first half of the quarter when other lower-frequency soft indicators are not available. The choice of the sentiment measure matters when tracking economic shocks such as the Great Recession and the Great Lockdown. Non-linear machine learning models can help capture extreme movements in growth, but require sufficient training data in order to be effective so become more useful later in our sample.

Keywords: Text analysis, Forecasting, Machine learning, Business cycles, COVID-19

JEL Classification: C43, C45, C55, C82, E37

Non-technical summary

Monitoring the economy in real time is crucial for making informed economic and policy decisions. However, macroeconomic fundamentals like Gross Domestic Product (GDP) are typically measured at a quarterly frequency and released with a substantial delay. Market participants and policymakers have traditionally tracked soft data, particularly business and consumer confidence surveys, in order to get a real time assessment of economic conditions. This is widely evidenced by monetary policy communications, which frequently point to survey evidence when describing the current macroeconomic situation. On the academic side, recent advances in data availability and computing power have promoted the use of alternative sets of predictors and novel methods often originated from the machine learning literature to gauge economic activity or detect turning points. These new data and models can act as complements to the current econometric tools used by policymakers and provide substantial aid especially in times of distress.

In this paper we build text-based sentiment indicators for the euro area derived from newspaper articles in the largest four euro area countries in their native languages. These indicators are available at daily frequency and contain timely economic signals which are comparable to those from well-known sentiment indicators such as the Purchasing Managers' index (PMI). In a second step, we test their predictive ability across the quarter and find that they prove to be highly beneficial in the first month of the quarter when other soft indicators are not available. Their power is diminishing as we proceed within the quarter and new data are released. The sentiment indices are based on counts of words in news articles translated into English and rely on several well-known English language dictionaries.

When it comes to detecting turning points, it appears that the choice of the dictionary matters. A commonly used finance-oriented dictionary captures very well the Great Recession, unsurprisingly given the financial nature of this crisis, but fails to capture the COVID-19 pandemic crisis. By contrast, the general-purpose dictionary is more consistent and robust across time. Therefore, the nature of economic shocks plays a significant role in identifying the most appropriate text dictionary to be used.

We test the predictive ability of these daily time series with a number of forecasting models,

from straightforward linear regressions to more sophisticated non-parametric machine learning algorithms. We find that our sentiment metrics provide substantial improvements in nowcasting performance compared to both the ECB's official projections and benchmarks based on the Purchasing Managers composite index (PMI). These gains are typically concentrated in the first half of the quarter, when other indicators are not yet available. These gains are particularly pronounced in the two major crisis periods in our sample: the Great Recession (2008-2009) and the Great Lockdown (2020). More specifically, we find that standard linear methods work well when there are no big shifts on the economic outlook but non-linearities matter when extreme economic shocks occur and the non-linear machine learning models can capture them more fully.

1 Introduction

Monitoring the economy in real time is crucial for making informed economic and policy decisions. However, macroeconomic fundamentals like Gross Domestic Product (GDP) are typically 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 confidence surveys to get a real time assessment of economic conditions. This is widely evidenced by monetary policy communications, which frequently point to survey evidence when describing the current macroeconomic situation. On the academic side, recent advances in data availability and computing power have promoted the use of alternative sets of predictors and novel methods often originated from the machine learning literature to answer many economic questions Hansen et al. (2017); Baker et al. (2016); Azqueta-Gavaldon et al. (2020)[1]. This study contributes to this growing body of work which shows that non-traditional datasets and methods can improve macroeconomic forecasts at short forecast horizons in the Euro area.

In this paper, we focus on the use of textual data derived from European newspaper articles 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) and Loughran and McDonald (2011) but also more of general purpose like VADER (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.

We test the predictive ability of these daily time series with a number of forecasting models, from straightforward linear regressions to more sophisticated non-parametric machine learning approaches. We find that our sentiment metrics provide substantial improvements in nowcasting performance compared to both the ECB's official GDP projections and benchmarks based on the Purchasing Managers' index (PMI). 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).

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

The economic literature on text analysis has made substantial progress in recent years. The uncertainty indices of Baker et al. (2016) and the topic-based sentiment indicators of Thorsrud (2018) are prominent examples of how text can be useful for economic analysis. In a nowcasting and short-term forecasting context, studies such as Larsen and Thorsrud (2019); Kalamara et al. (2020); Shapiro et al. (2020); Aguilar et al. (2021) show that text can significantly improve forecasts of key macroeconomic variables including GDP, inflation, and unemployment. Similarly, Ardia et al. (2019) and Rambaccussing and Kwiatkowski (2020), using US and UK newspaper text respectively, combine expert judgement and linear machine learning methods to forecast economic growth. Within the Euro area, Aguilar et al. (2021) show that their sentiment indicator derived from Spanish newspapers is comparable to the sentiment index produced by the European Commission and more helpful in nowcasting GDP. Similarly, Aprigliano et al. (2021) 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.

All of the studies mentioned above base their analysis on a monthly frequency and so do not reap the true benefits of the timeliness of text. Our nowcasting approach is closest to the one proposed by Algaba et al. (2021) who explicitly exploit the daily text signals and produce daily nowcasts of Belgian GDP growth. Our application differs from theirs in several aspects, in addition to focusing on a different region. This study produces nowcasts using a linear dynamic factor model, while we apply a range of linear and non-linear supervised machine learning algorithms paying particular attention to the aggregation of the metrics in real-time. In addition, Algaba et al. (2021) develop a new sentiment dictionary for Belgium, while we aim to compare the performance of a set of distinct lexicon-based approaches using a new dataset for the four largest Euro area economies.

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 non-linearities matter when extreme economic shocks occur and the non-linear machine learning models can capture them better and filter out the noise.

While a large literature on macroeconometrics has proposed a number of methodologies to nowcast GDP such as factor-based models (Bańbura et al., 2011), Bayesian vector autoregressions (Cimadomo et al., 2021), bridge equations (Baffigi et al., 2004) and mixed data sampling techniques or MIDAS models (Ghysels et al., 2004; Foroni and Marcellino, 2014), many recent studies provide promising avenues for improving macroeconomic predictions with machine learning (ML) methods; see Kim and Swanson (2018) for a review. However, most attention has been given to dimensionality reduction approaches at lower frequencies rather than the use of high frequency data for real time forecasting, which is our focus here. In our exercise, we exploit the timeliness of text, which is one of the major advantages of this type of data, and train a set of ML models on a daily basis as new text information becomes available. The advantage of these ML methods in our context is therefore their flexibility to accommodate non-linear relationships, rather than an ability to trim a large pool of data. As such, we do not induce sparsity by variable selection, e.g. through Lasso Regression, or dimensionality reduction, e.g. through Partial Least Squares.

We make several key contributions relative to the existing literature. First, we tackle the issue of translation. Our dataset consists of 5 million articles from fifteen 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 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.[2] To the best of our knowledge, although text analysis has received a growing attention in economics, there are no studies examining the role of translation of non-English text and its impact on constructing economic sentiment metrics. We find that the translated sentiment indicators appear to be strongly correlated with the indices derived from the raw, untranslated text. The important informational content is still captured on the translated text which allows for the same sentiment analysis approach and comparisons across all different languages.

[2] We use the *python package "googletrans"*. We compare this translation methodology to alternatives in Section 2.3.

Second, we 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 but also at a Euro area aggregate level.

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, a general-purpose sentiment measures such as "VADER" (Gilbert, 2014) is more consistent across time and robust in the context of large "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. (2020) 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 machine learning models are proved necessary during periods of large shifts, provided that there are enough data available.

The rest of the paper is organised as follows: Section 2 describes the data used in the nowcasting exercise and, in particular, introduces the raw dataset, the strategy followed for the non-English articles and the methods used to convert text into daily time-series indicators. Section 3 describes the nowcasting setup and provides a brief model overview. Section 4 presents the main results

and Section 5 concludes. A more detailed description of all the models used in our empirical exercise as well as supplementary material is provided in the Appendix.

2 Data and Translation

2.1 Text Data

We use articles from major print newspapers for the "Big Four" Euro area economies from the Factiva database. In each case, we restrict the data to articles that are tagged as either economic, corporate or financial markets news. This reduces the noise in our sentiment measures by excluding articles focused on topics such as sport and lifestyle. We then have a total of 5 million articles covering a period from January 1998 to January 2021, from 15 separate sources. We choose newspapers that have wide circulation and reflect a broad spectrum of political leanings. Table 1 shows the total number of articles from each country and the newspapers from which they are taken. More detail on these sources and the number of articles for each source and country over time is found in Appendix A.

Table 1: Total number of articles per country

	France	Germany	Italy	Spain	All
Total articles	1,255,472	833,914	1,497,909	1,407,534	4,994,829
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	

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 still corresponds to the quarterly frequency of the GDP data that we are nowcasting.

In our baseline case, we translate the articles from their native languages using Google Translate. This is described in detail in Section 2.3 below, which also describes our alternative translation methodologies. We then use a suite of English language sentiment measures, described in Table

2, to compute a daily sentiment metrics for each of the four countries.

Table 2: English language sentiment dictionaries used

Initials	**Source**	**Description**
AFINN	(Nielsen, 2011)	Classifies words on a scale from +5 to -5, general purpose.
CGLM	(Correa et al., 2017)	Classifies words as positive (+1) or negative (-1), focus on financial stability.
HIV	(Tetlock, 2007)	Classifies words as positive (+1) or negative (-1), general purpose.
HL	(Hu and Liu, 2004)	Classifies words as positive (+1) or negative (-1), developed to capture opinion in reviews
LM	(Loughran and McDonald, 2013)	Classifies words as positive (+1) or negative (-1), focus on economics and finance.
NKTGOS	(Nyman et al., 2018)	Classifies words as excited (+1) or anxious (-1), developed for finance applications.
VADER	(Hutto and Gilbert, 2014)	Classifies sentences on a scale from -1 to 1, developed social media text.

Note that these methods differ in three main dimensions, all of which have an effect on nowcasting performance. First, some have been designed specifically with an economics application in mind, while others are much more general. Second, six of the seven methods work at the word level, so classify each word in isolation, but VADER works at the sentence level and is therefore able to account for factors like negation and punctuation that can affect meaning. Third, most of the methods classify words as either positive (+1), negative (-1) or neither, but two of them have a more 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.

Overall, for each method we obtain a sentiment score for each word/sentence which can then be aggregated. As our target variable has a quarterly frequency, we develop a daily sentiment

metric that recognises this quarterly frequency of the target. 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 $sent_{q,t,n}$ 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

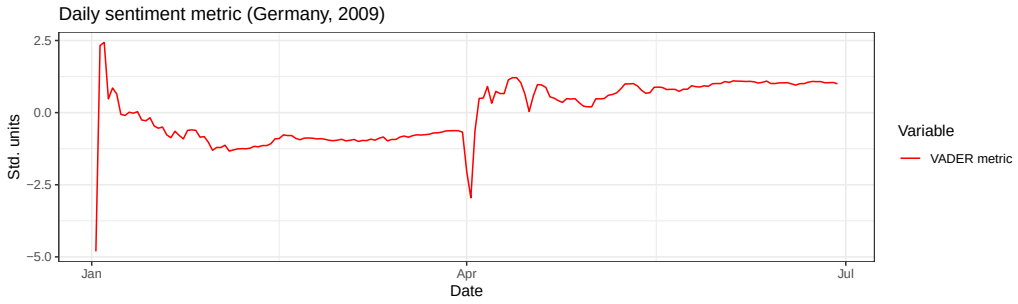
$$sent_{q,d} = \frac{\sum_{t \leq d} \sum_{n=1}^{N_{q,t}} (sent_{q,t,n})}{\sum_{t \leq d} N_{q,t}}$$

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.[3]

This means that as the quarter progresses, the metric averages over more days and therefore over more articles. Other works have 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. (2020)). Figure 1 shows the German daily metric constructed using the VADER dictionary for the first two quarters of 2009. This Figure illustrates that at the beginning of the quarter the measure is noisier as it relies on a smaller sample, and large movements are possible in the first few days. As the quarter progresses more articles become available until the end of the quarter and the metric becomes more stable. This is a plausible strategy to follow, as our main focus is on real-time nowcasting.

[3]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.

Figure 1: Daily sentiment metric



VADER. Unlike polarity-based methods which classify terms or phrases as either positive or negative, valence-based measures take the intensity of the expressed sentiment into account. Valence Aware Dictionary and sEntiment Reasoner (VADER) is based on a lexicon of over 7,500 lexical features, including commonly used abbreviations and emojis. These features are then rated by workers on Amazon Mechanical Turk on a scale from -4 (Extremely Negative) to +4 (Extremely Positive). In addition to this lexicon, VADER also applies five general heuristics that affect either intensity or polarity of a sentence. For example, degree modifiers such as "extremely" increase the intensity of sentiment and negation switches the polarity of a sentence. VADER is thus different from the other pure lexicon based sentiment measures we implement, as it takes the context of each word into account. As this is done in a rules-based way, it is not especially computationally demanding and so still realistic for our corpus of 5 million articles.

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 natural language processing 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. Appendix B shows an example translation for each of the four languages at hand.

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.

Table 3 shows that the first two methodologies produce highly correlated results in most cases, particularly for the economics-focused dictionaries. This varies very 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: Correlation of daily metrics across two translation methodologies

sent_metrics1	France	Germany	Italy	Spain	Euro area
CGLM	0.670	0.650	0.737	0.845	0.821
LM	0.622	0.576	0.824	0.871	0.813
AFINN	0.691	0.702	0.817	0.814	0.828
HIV	0.654	0.209	0.712	0.573	0.652
NKTGOS	0.482	0.611	0.595	0.715	0.664
HL	0.602	0.792	0.817	0.777	0.817

Table 4 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. 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. As the VADER metric is a hybrid approach that takes the context of terms into account, it is not straightforward to apply this to non-English articles, so we only consider this measure applied to translated articles.

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 0.519 for the translated articles and 0.391 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).

Table 4: Correlation of sentiment metrics with (quarterly) GDP growth

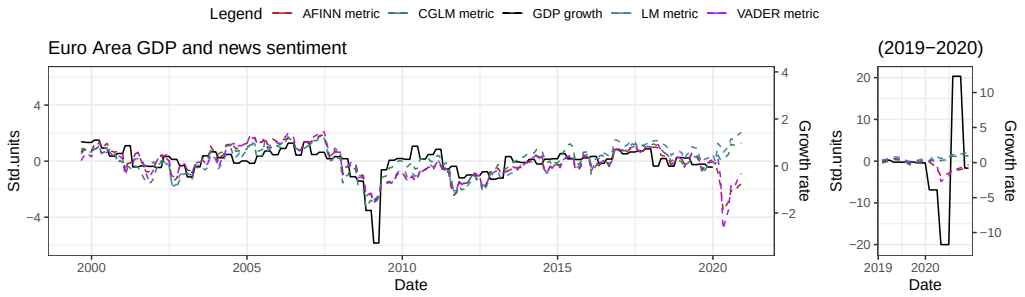
Translation	Metric		France	Germany	Italy	Spain	Euro area
Article	CGLM	Economic	0.615	0.527	0.542	0.71	0.698
	LM	Economic	0.593	0.461	0.412	0.619	0.636
	NKTGOS	Economic	0.38	0.373	0.364	0.618	0.538
	AFINN	General	0.583	0.371	0.41	0.731	0.637
	HIV	General	0.58	0.317	0.254	0.524	0.575
	HL	General	0.513	0.403	0.33	0.664	0.597
	VADER	General	0.583	0.369	0.292	0.701	0.611
Dict	CGLM	Economic	0.407	0.376	0.487	0.624	0.593
	LM	Economic	0.329	0.214	0.445	0.571	0.528
	NKTGOS	Economic	0.157	0.145	0.389	0.478	0.365
	AFINN	General	0.334	0.268	0.472	0.628	0.536
	HIV	General	0.288	0.24	0.187	0.38	0.446
	HL	General	0.246	0.304	0.253	0.553	0.478
Own lang				0.355			

We therefore focus on the results using the translated articles in the remainder of the paper. The results with the untranslated articles are somewhat similar and for the sake of space not included, but they are available upon request.

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.[4] Figure 2 shows three of the sentiment series (plotted at a monthly frequency for clarity) alongside Euro area GDP, illustrating that the sentiment metrics co-move with economic activity.[5] We plot the growth rates for 2020 on a separate panel with a different scale as they are so extreme that visualising the developments in the pre-2020 period is difficult. 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.

Figure 2: Euro area GDP growth and news sentiment



A first look at the sentiment metrics allows to draw two initial conclusions. Firstly, 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 comove 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, this does not guarantee that they will perform well in response to future shocks. In section 3.4.2 we will provide more details. Secondly, while the series clearly co-move, this relationship appears to be non-linear. In particular, during crisis periods where we see a large fall in GDP, the sentiment metrics also

[4] 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. To ensure that we only use data that was available in real time, we use the previous year's weights in each period to construct the Euro area series.

[5] These series for the individual countries are also shown in Appendix C.

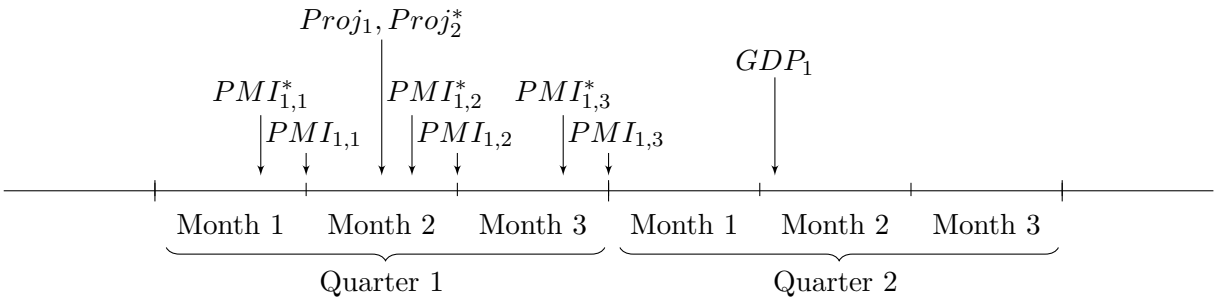
fall, but not proportionately to GDP. This is to be expected given that the sentiment metrics are naturally bounded by the methods used to generate them. In Section 3 we will therefore consider both linear and non-linear nowcasting models.

2.5 Macroeconomic indicators

The target series in the nowcasting exercise is the Euro area real GDP growth, which is available at a quarterly frequency from Eurostat. In our benchmarking, we use the monthly Purchasing Managers' Index (PMI) which is published by IHS Markit and is one of the most watched survey-based or economic sentiment indicators in the world. We focus on the PMI composite output index which aggregates activity in manufacturing and services, for the big four Euro area countries and the Euro area as a whole. We also consider historical official ECB macroeconomic projections. We make sure we only use data that was available in real time for our nowcasts, allowing us to assess the value of our sentiment metrics throughout the data release cycle. As we want our nowcasts to be as accurate as possible, we judge performance by comparing to the final revision of GDP (i.e. latest vintage), while ensuring that only data that were available in real time are used for training the predictive models.

The publication of different indicators and their flash estimates varies across countries and over time. An example data release cycle is shown in Figure 3. For most of our sample period, the Euro area PMI composite index is released as a 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 released around the middle of each quarter. We focus on the GDP projections for both the current and next quarters. GDP data for a given quarter is first released during the following quarter and experiences frequent revisions. Since 2016 the first (flash) GDP estimate for the Euro area is released one month after the reference period is over (i.e. 30 days), but before 2016 the publication was one month and a half (i.e. 45 days) after the end of the reference period.
For the PMI indicator we also construct a quasi-daily series that corresponds to the quarterly frequency of GDP. The PMI indicator is available at a monthly frequency, and is typically published at the beginning of the next month. However, in some cases (i.e. 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 three relevant values of PMI, and potentially

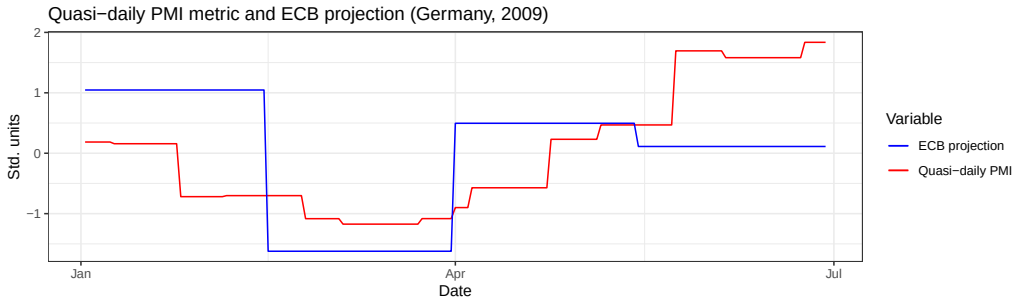
Figure 3: Example data release cycle



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 $PMI_{q,m}^*$ 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 $q-1$.

six release dates.

Figure 4: Daily PMI metric



Our quasi-daily PMI measure is constructed as follows. 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 month's 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 one and the flash estimate for month two are available, so we take the mean of these. The quasi-daily PMI indicator is thus constructed in the same vein as our daily sentiment indicator. Figure 4 shows this PMI metric for Germany during the first two quarters of 2015, illustrating that it is updated 6 times throughout the quarter.

3 Nowcasting Setup and Results

The aim of this section is to show that text data contains useful information that can improve nowcasts of real GDP growth especially at times when other key indicators are not available because of publication lags. But it also shows that the nowcasting model matters, being machine learning methods the ones that make a more efficient use of the information while accommodating possible non-linearities during times of distress.

This section is structured as follows. Section 3.1 describes the general framework to create daily nowcasts and the benchmarks used to assess the value of the sentiment metrics. Section 3.2 then illustrates this framework with an example of a simple linear model. Section 3.3 describes in detail the alternative nowcasting models and specifications used to assess the value of the sentiment metrics. Section 3.4 presents the results, focusing on the pre-2020 period and the year 2020, respectively.

Across models and settings, we can draw a number of conclusions. Firstly, our sentiment metrics are particularly useful in the first half of the quarter, before more traditional indicators become available. Secondly, the text data is useful in the two major crisis periods in our data, but of more limited value in normal times. Thirdly, while specialised economics or finance metrics perform well during the Great Recession, they perform poorly in the Great Lockdown. More general purpose metrics perform well in both periods. Finally, non-linearities are an important feature during turbulent times and incorporating them help significantly the predictability of GDP growth during the Great Lockdown period. However, these non-linear methods understandably require a longer training sample in order to work well, and so don't perform as well as simpler methods 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 also for each of our benchmarks. The 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.

The baseline nowcasting models that we use have the general form:

$$\hat{y}_{q,d} = g\left(x_{q,d}, \theta, \eta\right), \tag{1}$$

where $\hat{y}_{q,d}$ is the nowcast of quarter-on-quarter real GDP growth on the dth day of quarter q; $x_{q,d}$ is a vector of predictors for that day (for example, the daily sentiment and PMI metrics); θ is a vector of estimated parameters; and η is a vector of hyperparameters used to train the model. The function $g(\cdot)$ varies with the model at hand, and more detail on the models used is given in Section 3.3.

We assess the value of including the text metrics by comparing a model with text to two classes of benchmark. The first benchmark is a model of the same functional form $g(\cdot)$ as the text model, but includes only the quasi-daily PMI metric 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. The second benchmark is the latest available ECB projection for real 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. In what follows we will focus on results for two of our sentiment metrics, CGLM and VADER, as these are widely used examples of economics-focused and general-purpose lexicons. Furthermore, sentiment metrics based on these lexicons have a high correlation with real GDP growth.

Beginning on 1 April 2006, as this is the first date for which we have real time vintages of GDP and PMI available, both the text and PMI models are re-trained each day. We use an expanding window of data, starting in January 2002, for which at least one vintage of GDP has been published.[6] For a given day, we ensure that we only use data vintages that were available in real time. So only the latest available vintages of GDP are used to compute the growth rates that the model is trained on. In order to ensure that only comparable observations are used,

[6] Where data are standardised, we take care to only use data available in real time for this standardisation.

we restrict the training set to observations at the same stage of the data release cycle (shown in Figure 3). For example, if the PMI for the first two months are available, we train the model only on observations from days on which PMI for the first two (of three) 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).

To compare the text model to the benchmarks we compute the error between each daily nowcast and the target variable. We can them 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, for instance, we compute the MSE for the first day of each quarter as

$$MSE_1 = \frac{1}{Q}\sum_{q=1}^{Q}(y_q - \hat{y}_{q,1})^2 \tag{2}$$

where y_q is the target variable for quarter q (calculated using the latest vintages available on 24th March 2021) and $\hat{y}_{q,1}$ is the nowcast for y_q produced on the first day of the quarter. This allows us to compute the nowcasting performance for 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

The purpose of this section is to illustrate with a model as simple as possible how we will evaluate the usefulness of sentiment metrics. We also exclude the year 2020 from our example as if included this dominates any comparison due to the extreme GDP growth rates. Year 2020 is examined in detail in Section 3.4.2.

Our text model is

$$g_{text}(x_d, \theta, \eta) = \theta_0 + \theta_1 PMI_{q,d} + \theta_2 sent_{q,d}, \tag{3}$$

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

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

In this case, we estimate the model using Ordinary Least Squares (OLS) where no hyperparameters η are required.

Figure 5 compares the forecasting performance of the text model, ECB GDP projections and the PMI benchmark for the period between April 2006 to December 2019. Panel (a) shows the performance with the CGLM metric and panel (b) with the VADER metric. In each case the upper panel shows the mean squared error (MSE) for the ECB projection (in red), the PMI model (in green) and the text model (in blue) for each day of the quarter across the sample period.[7] 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 lower panels show the percentage improvement of the text model compared to the PMI benchmark. 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.[8] Figure 6 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.

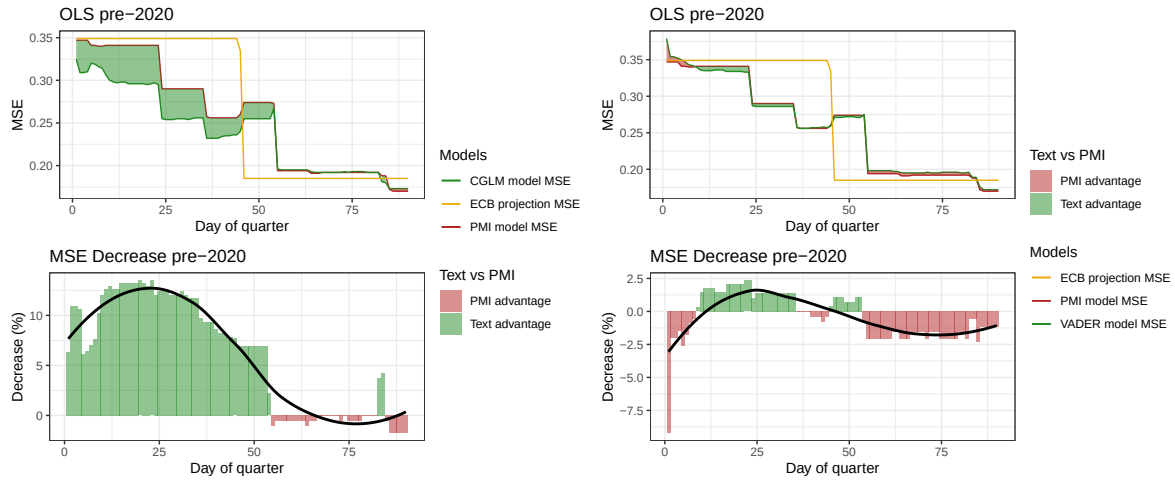
[7] Note that the PMI model is updated only periodically, as new PMI data become available, while the text model is updated on a daily basis as new articles are published, as well as when new PMI data becomes available.

[8] Intuitively, we attribute this to the fact that our metrics are inevitably noisy given shifts in coverage across newspapers and news that affect sentiment but are unrelated to output. Therefore, small movements in the metrics may be hard to interpret.

Figure 5: Linear example: pre-2020

(a) CGLM metric

(b) VADER metric

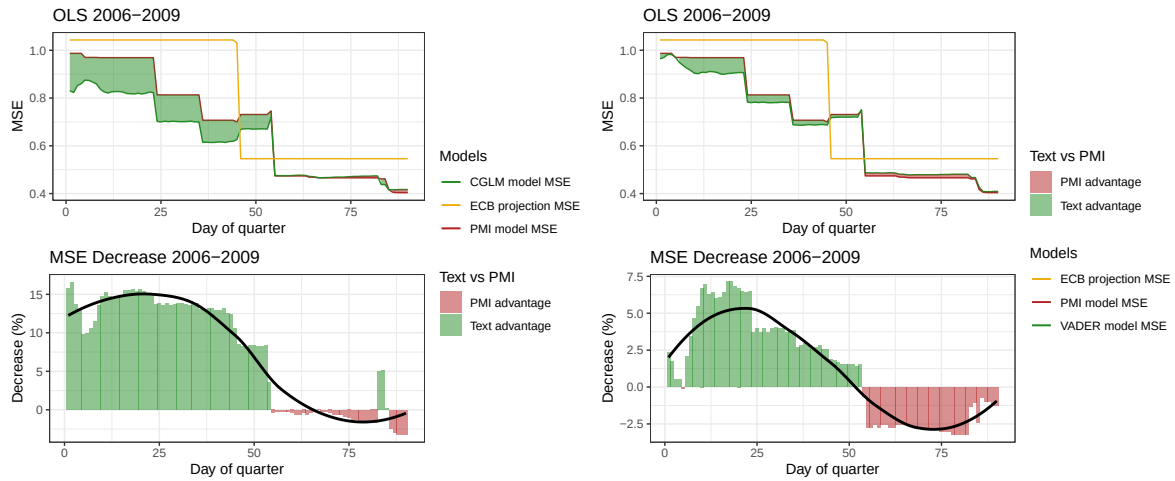


Notes: Upper panels show the average MSEs for each model for each day of the quarter, calculated as in Eq. 2. This MSE is shown in green for the text model, red for the PMI benchmark and yellow for the latest available ECB projection. The difference between the MSE of the text model and PMI model is 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 to the PMI benchmark, for each day of the quarter. The black line here 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.

Figure 6: Linear example: Great Recession

(a) CGLM metric

(b) VADER metric



Notes: see Figure 5 for full explanation. The results shown here are for April 2006 to December 2009

3.3 Overview of Alternative Models

A common critique of previous studies using textual data for forecasting is that the modelling framework and benchmarks chosen were relatively simple (e.g. simple linear regressions as in the example above). While conventional econometric techniques such as linear regression models often work well and are a good starting point, there is well-established evidence that the relationship of economic activity with "soft" indicators is non-linear (Woloszko, 2020; Kalamara et al., 2020). For this reason, we test a range of alternative models and specifications. These include both, models that can accommodate non-linearities as well as alternative specifications of the predictors $x_{q,d}$ (e.g. first differences of the sentiment indicators, ECB GDP projections).

3.3.1 Models

The models we consider in addition to an OLS linear regression are: Ridge Regression (Ridge), Random Forests (Forest), Neural Networks (NN) and Boosting Algorithm (Boosting). The latter three are well-known machine learning algorithms which allow for non-linearities in a data-driven way. Next, we present a brief overview of the models and their basic properties. A more technical description is provided in the Appendix E.

Ridge Regression. Ridge regression is a shrinkage method that produces linear combinations of the original regressors, where those coefficients that do not carry any predictive power for the target variable are assumed to approach zero according to a shrinkage parameter η, which differs across models. This methodology shrinks the coefficients of predictors that contribute little to the predictive ability of the model towards zero, albeit they never become exactly zero—it is therefore a dense model which draws on all available information although to different degree. In the case of no shrinkage, i.e. $\eta = 0$, ridge regression becomes equivalent to a OLS linear regression.

Random Forest and Boosting. Both Random Forests and Boosting are ensemble methods and their model architecture is centred on regression trees. Generally, regression trees are based on consecutively splitting the in-sample dataset until an assignment criterion with respect to the target variable into a "data bucket" (leaf) is reached. This is a powerful idea, since it can fit various functional relationships between the dependent variable and a set of explanatory

variables, say $f(x_{q,d})$, without imposing linearity or additivity, which are commonly assumed in standard linear regression models.

There are several ways to further improve the performance of regression trees. Bagging is perhaps the most commonly used; the method essentially generates multiple versions of a predictor and uses these to get an aggregated predictor.[9] In the context of regression trees, bagging averages across trees estimated with different bootstrapped samples to create a "forest". More specifically, a Random Forest grows a large collection of de-correlated trees (hence the name forest) and then average them. This is achieved by bootstrapping a random sample at each node of every tree. In order to induce "decorrelation" of trees, when growing trees, before each split it selects a subset of the input variables at random as candidates for splitting. This prevents the "strong" predictors imposing too much structure on the trunk of the tree. An alternative way to improve the performance of regression trees is via boosting.

Boosting focuses on the predictive power of individual predictors one at a time. In an economic context, boosting has been applied by Bai and Ng (2009) and Ng (2014)). For example, Ng (2014) uses boosting in order to screen a number of potentially relevant predictors and their lags and give warning signals of recessions. The method focuses on the predictive power of individual regressors instead of considering all covariates together [10]. In this context, regressors are chosen sequentially based on their individual ability to explain the dependent variable. Based on an iterative procedure, the misclassified observations are given increasing cost in each estimation repetition. Overall, the idea is to consider regressors one by one in a simple regression setting, and successively selecting the best fitting ones (Friedman, 2001).

Neural Networks. Neural networks (NN) can also incorporate nonlinearity and interaction of variables through a flexible functional form. The structure of a neural network can be described by three components: input layer, hidden layer and output layer. Each layer is collegiated by synapses which deliver signals from the neurons in the preceding layer to the succeeding one. In our setting, the input layer corresponds to predictor variables so that the number of neurons in the input layer is the same as the dimension of predictors. The hidden layer converts an output

[9] See Breiman (1996) for an overview.

[10] This approach has led to a variety of alternative specification methods sometimes referred to collectively as "greedy methods".

from the preceding layer (including the input layer) through an activation function. Finally, the output layer summarizes the output from the hidden layer.

We use multilayer perceptrons (MLP), a form of feed-forward network, as NN architecture. The activation function $g(x_{q,d}, \eta)$ acts as a gate for signals and introduce non-linearity into the model. Its functional form is subject to hyperparameter tuning. The variables $x_{q,d}$ in the input layer are multiplied by weight matrices η at each layer, then transformed by an activation function in the hidden layers and passed on through the network until the linear output layer is reached resulting in a prediction $\hat{y}_t$ [11].

Neural network becomes more complex and flexible when we increase the number of units in a hidden layer (wider neural network) or increase the number of hidden layers between input and output layers (deeper neural network).They are generally more accurate but also require more data to train them due to the larger number of parameters in the weight matrices. The number of hidden layers, i.e. the depth of the network, and the number of neurons in each layer as well as appropriate weight penalisation in our ANN are hyper-parameters, and are determined by cross-validation as discussed below.

3.3.2 Hyperparameter Tuning

Cross validation strategy. All of the models except OLS require some form of hyperparameter selection prior to estimation. In cases where the derivation of the traditional and widely used information criteria is not feasible, such as the Ridge Regression and machine learning tools, we use a cross-validation procedure. Here, the strength of regularisation parameters, the number of nodes and layers for the NN, or the maximum depth of the leaves for the Forest and Boosting is chosen among others. Our nowcasts are updated daily. Performing cross validation on a daily basis would be computationally intensive and thus we opt to update the hyperparameters of the methods at every quarter. We take care of not including any future information and perform cross validation only on the in-sample data at each quarter step.

In particular, the procedure we follow can be summarised as follows: For each time step, we

[11] We use the rectified linear unit activation *ReLu* function applied element-wise. While also other functions may be considered, *ReLu* remains a preferred choice due to its simple form of its gradient which facilitates the estimation procedure (Bianchi et al., 2020; Farrell et al., 2021).

split the in-sample data in k=5 folds as the train set and the k + 1-th fold as test set.[12] This is consistent with our expanding window evaluation of the out-of-sample test forecasts. As a performance metric, we consider the average mean squared error (MSE) over the test set.

3.4 A Tale a Two Crises

This section provides the main findings with regard to the two major economic crises that the Euro area has experienced in the last two decades: the Great Recession and the Great Lockdown. As shown before, the text models perform better in the first month of the quarter when other soft indicators are not available. Here we want to show that text models have a better forecast performance in crisis periods than in normal times. But there is no text indicator or model than performs best in all crises.

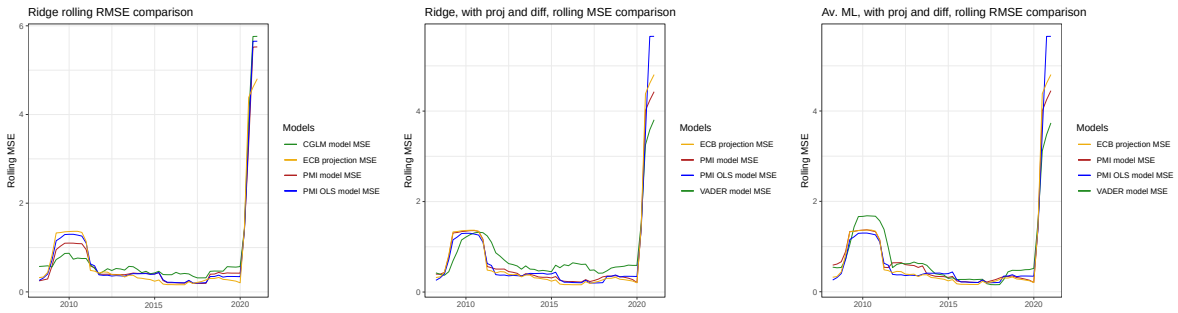
Figure 7 shows the evolution of the Mean Squared Error (MSEs) for our main model specifications over a rolling window of 8 quarters in the period from 2006 until 2020. The GDP forecasts to compute the forecast errors are those obtained at the end of the first month of the reference quarter, which is when the text models perform best. Plot (a) provides the rolling MSEs for the text-based model using the GCLM metric (green line) and the PMI-only model (red line), both of them estimated using Ridge Regression. Yellow and blue lines are the rolling MSEs of the ECB projections and of the OLS PMI-only model, respectively. During the Great Recession period, the text-based model produces consistently more accurate nowcasts compared to all the other models and ECB projections. Interestingly, this is not the case during the recent COVID-19 pandemic. This is expected to some extent given that the pandemic is caused by an inherently different shock from the global financial crisis and the GCLM metric is explicitly designed to capture the sentiment related to the financial environment. This finding is further supported by Plot (b) which shows the rolling MSEs using the VADER indicator, a general-purpose sentiment metric instead, as the main text-based model. Even though the model does not capture properly the crisis in 2008, it performs remarkably well during the current pandemic. The main takeaway is that the choice of the dictionary used is crucial especially during extraordinary periods and depends on the nature of the shock occurring in the economy.

[12] More specifically, we use Time Series Split which is a variation of k-fold targeted to time series. In this approach successive training sets are super-sets of those that come before them.

In Plot (c) we focus on the performance of the VADER metric combined with the non-linear machine learning methods. We take the average of the three models we use, namely the neural network, the random forest and the boosting algorithm as a measure of the non-linear component of the models. We find that including non-linearities improves substantially the nowcasts of GDP growth during both turbulent periods. This is particularly clear during the Great Lockdown given that the rolling MSEs are the lowest across all other model specifications. It is also important to note that the PMI-only ML based model outperforms its linear counterpart and ECB projections during this period emphasising the importance of incorporating non-linearities. The next two subsections describe our results for each of these two crises in more detail.

Figure 7: MSE using an 8-quarter rolling window

(a) CGLM Ridge (b) VADER Ridge (c) VADER AvML metric



Notes:Rolling MSEs using a 8-quarter window for the whole period considered. Green lines denote the text-based model (GCLM or VADER) while red lines are for the PMI-only model. Yellow lines denote the ECB projections errors and blue lines are the linear PMI-only benchmark.

3.4.1 Great Recession and its aftermath

In this section we assess the value added of incorporating the high–frequency text-based sentiment in nowcasting GDP growth across a range of models and specifications from April 2006 to December 2019. We find that both CGLM and VADER contain useful information beyond that in the PMI metric, and that these benefits are greatest at the beginning of the quarter and during the Great Recession. Introducing regularisation through the Ridge regression improves performance relative to a simple OLS, but further non-linearities do not. We suspect that this is because these more complex approaches require more data in order to reap the benefits of their flexibility.

In Section 3.2 above, we showed results for this time period in the special case where the

nowcasting model is the simple OLS regression in Eq 3. In what follows, we examine the performance of the models described in Section 3.3, comparing in each case the text model with a PMI model of the same form, and OLS regression with only PMI (as in Section 3.2) and the ECB projections. As above, we display results for both CGLM and VADER as we find these to be the best performing economics-focused and general sentiment metrics, respectively.

Figure 8 shows the daily average evolution of the MSEs across quarters for the CGLM and the VADER text metrics using a Ridge regression. The upper panels focus on the overall out-of-sample period up to 2020 (i.e. pre-2020 or 2006-2019) while the lower panels zoom in on the Great Recession period (1 April 2006 to 31 December 2009).

Generally, Ridge regression outperforms the linear survey-based model for both periods considered. This suggests that the contribution of the predictors is different across the quarter and penalisation is necessary to allocate appropriately the weights associated to the predictors. While the forecast accuracy gain with the text-based Ridge regression is more limited for the overall period considered, the improvement during the Great Recession is significant. This evidence is stronger for the CGLM metric than the VADER metric. Note that the CGLM metric is based on a financial dictionary and is meant to capture the sentiment related to financial stability[13], while VADER captures a general-purpose sentiment. This supports the intuitive idea that not all the sentiment lexicons provide the same informational content and their power is linked with the nature of the economic shock at hand.

Incorporating alternative specifications, i.e. including first differences of the sentiment indicators and ECB projections as predictors, the forecast accuracy gains are even higher for Ridge regression suggesting a clear advantage of this approach for the linear case. During the Great Recession period (lower panels), 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 present when using the VADER model but the magnitude of the MSE drop is smaller compared to the PMI model. In both cases, text-based models are quite helpful in the first two months before the official releases of other indicators.

[13]For more details see, Correa et al. (2017).

Furthermore, looking at the actual daily nowcasts in Figure 9, we observe that text is particularly helpful in predicting the start of the crisis in 2008 rather than the subsequent recovery. This provides supporting evidence that this alternative type of information is able to provide early warning signals of future economic disruptions (Nyman et al., 2018; Huang et al., 2019).

Even though we are not in a "big" data setting, we test the ability of the non-linear machine learning models (Forest, Boosting, NN) to nowcast GDP growth on a daily basis. We do this by training each of the nonlinear models separately following the same data specifications and then taking the average of the predictions as indicative for the non-linear performance. Figure 10 shows the results for the pre-2020 period and the Great Recession. ML-text models are more competitive to its ML-survey counterpart and, in line with the previous findings, the advantages are greater for the CGLM metric compared to the VADER metric. However, several observations arise.

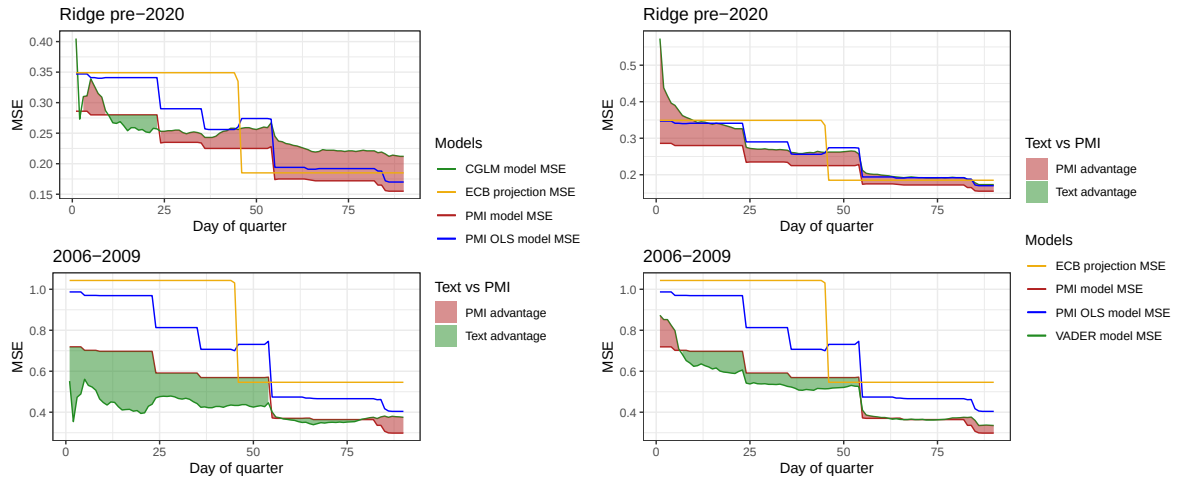
First, none of the ML-based specifications are able to improve the daily nowcasts when we compare results with the linear benchmark, especially at the beginning of the average quarter. This is mainly attributed to the low data availability as these methods are designed to perform well on rich data environments. Despite this fact, all ML-based models show a gradual MSE drop as we progress within the quarter and new information becomes available.

Second, the ML-based models seem to capture the updates of the information which is inherent to the construction of the soft data sets. For example, in the beginning of the average quarter, the text indicators are noisy and become more informative as we accumulate more articles. The daily GDP nowcasts follow a similar pattern, and therefore, the MSE improvements are larger at the end of the quarter rather at the beginning.

All the above findings are further supported by a Diebold and Mariano test for statistical significance. We present the results for VADER and CGLM models on Tables 7 and 8 respectively in the Appendix for the whole pre-2020 period on a monthly basis.

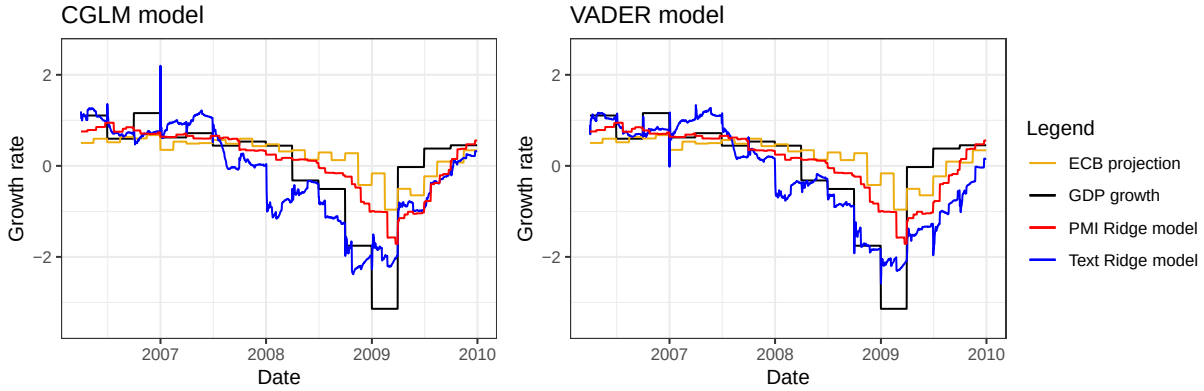
Figure 8: Ridge regression in levels: pre-2020 and Great Recession

(a) CGLM metric (b) VADER metric



Notes: The interpretation of this Figure is very similar to that of the upper panels in Figure 5. As before, the nowcast MSE across quarter for the text model is shown in green, the corresponding PMI model in red (in this case a Ridge Regression) and the ECB projections in yellow. In addition we show the MSE for an OLS regression including only the PMI metric in blue. The green and red lines therefore represent the competition between PMI and the text given a particular model, while the blue and yellow lines show benchmarks that use a different modelling framework.

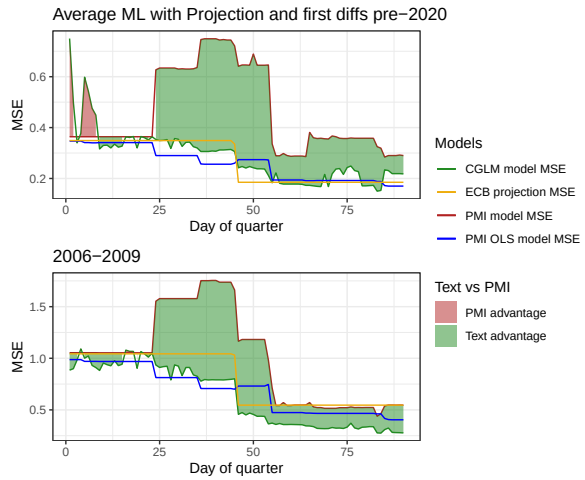
Figure 9: Nowcasts in the Great Recession



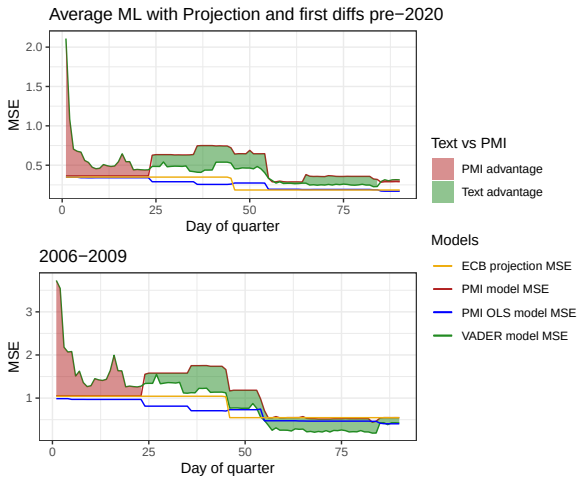
Notes: This Figure compares the real-time nowcasts of various text models and PMI models from April 2006 to December 2009.

Figure 10: Average ML with projections and first differences: pre-2020 and Great Recession

(a) CGLM metric



(b) VADER metric



Notes: see Figure 8 for full explanation. Results here are for the averaged ML models with the ECB projections and first differences included.

3.4.2 Results for year 2020: the Great Lockdown

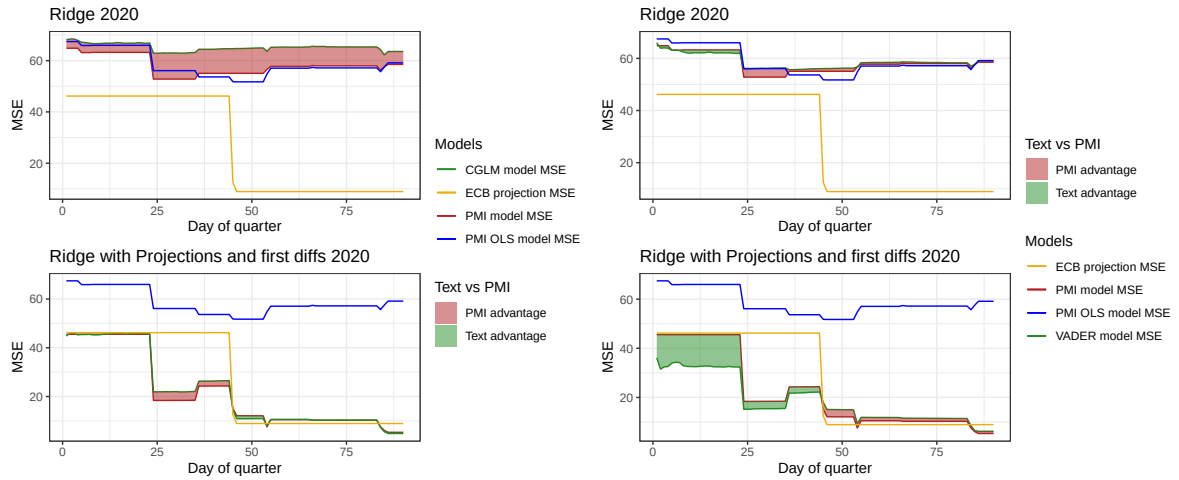
So far, we have seen that text can be helpful in crisis periods when economic indicators can be subject to sudden and rapid changes. In this regard, the coronavirus outbreak serves as an interesting illustration to demonstrate the applicability of our text metrics and different methodologies. As previously mentioned, we deem that the COVID-19 pandemic is an unprecedented economic shock which should be examined on its own. Thus this section is explicitly devoted to the year 2020 when the Covid-19 crisis took place.

We start by testing the linear cases using the text-based information. Figure 11 shows the evolution of the daily MSEs for the year 2020 including sentiment indicators in levels (upper panels) and including the projections and sentiment indicators in first differences (lower panels). Various interesting observations emerge. First, as it may be expected, the linear PMI-benchmark completely misses the contraction. Second, projections and first differences boost significantly the performance of the different models no matter the soft information we use. Finally, The GCLM metric (plot (a)) with Ridge appears less informative irrespective of the specification compared to its PMI benchmark counterpart. On the other hand, the general sentiment dictionary VADER shows some 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 PMI-OLS model. This is contradicting with the behaviour of the same metrics and specification during the financial crisis in 2008 suggesting that the selection of the lexicon method should be consistent with the nature of the economic shock. The pandemic is a shock of a non-economic nature that has triggered a global economic crisis and very strong policy support. As captured by the GCLM metric and in contrast to previous crises, the behaviour of the financial markets and the financing conditions have remained favourable. For forecasting purposes, looking only on financially relevant terms is therefore not sufficient to capture the dynamics of this disruptive event.

Figure 11: Ridge with and without projections: 2020

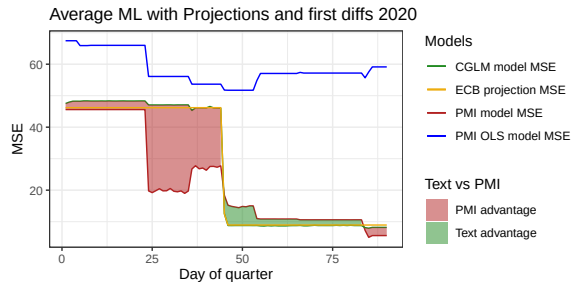
(a) CGLM metric (b) VADER metric



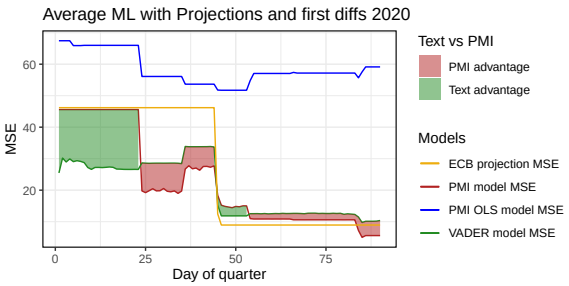
Notes: see Figure 8 for full explanation. Results here are for a Ridge regression without ECB projections and first differences (in the upper panels) and with (in the lower panels), for the year 2020 only.

Figure 12: Average ML with ECB projection and diffs: 2020

(a) CGLM metric



(b) VADER metric



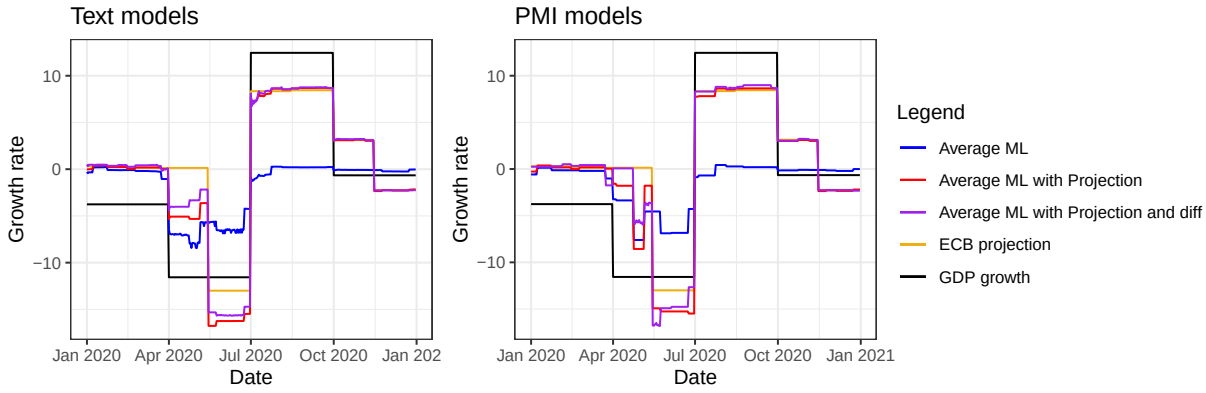
Notes: see Figure 8 for full explanation. Results here are for the averaged ML models regression with ECB projections and first differences, for the year 2020 only.

Figure 12 shows the average-ML predictions within the average quarter of 2020 [14]. The results with respect to the lexicon choices remain qualitatively the same - the GCLM sentiment does not produce significant daily error improvements while VADER appears stronger especially in the beginning of the average quarter. Interestingly, the use of the non-linear methods seem to have a higher positive effect on the PMI model which suggests that the non-linearities are key for the period considered no matter the soft indicators at hand.

Zooming in further on year 2020, we plot the daily nowcasts with text and PMI models and ML-Averages for alternative specifications in Figure 13. Several interesting features arise. First, all text metrics manage to 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 to 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.

[14] For the sake of space the results for the different machine learning methods are not shown. Curiously, the best-performing models during the pandemic have been the neural networks, which were the worst-performing models during the financial crisis.

Figure 13: Nowcasts in 2020 (VADER)



Notes: This Figure compares the real-time nowcasts of various text models and PMI models throughout 2020.

4 Conclusions and further work

This paper shows that 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 the competitive and rigorous benchmarks of the ECB's official projections and models using the PMI composite index.

Metrics derived from translated text appear highly correlated with the relative non-translated counterparts and are still able to capture in a timely manner some of the information that policymakers might usually get from other proxies. This is a strong evidence as it allows to make direct comparisons of the text metrics across different multilingual countries.

We also show how two commonly used dictionaries in the literature were able to add value to a variety of nowcasting models. Much attention has previously focused on the extraction of information about uncertainty from text. However, we find that text-based sentiment metrics are better correlated with GDP growth and seem to be more informative inputs for the nowcasts.

As well as being useful proxies for sentiment, we however show how the choice of the lexicon used matters when it comes to detect turning points in economic activity. While the financial stability-based dictionary of Correa et al. (2017) performs strongly during the Great Recession,

unsurprisingly given the financial nature of this crisis, it is 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 have remained favourable in the context of very strong policy response. On the other hand, a general purpose sentiment indicator such as VADER (Gilbert (2014)) is more resilient especially in unexpected economic episodes such as the COVID-19 pandemic. The nature of economic shocks therefore plays a significant role in identifying the most appropriate text dictionary to be used.

We also test the forecasting performance of a suite of non-linear machine learning methods in a real-time setting. Unlike other studies which usually seek to tackle the curse of dimensionality by applying those methods, we focus on their functional form to capture non-linearities using high frequency data. Non-linear machine learning models respond more flexibly to the coronavirus outbreak and combining them with text information provides the best combination on tracking the sudden drop on 2020Q2 but also 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.

There are several avenues for future work. First, the analysis can be extended to other macroeconomic fundamentals and components of GDP (e.g. investment, consumption). Second, the forecasting exercise conducted in this paper focused on a small set of predictors. Moving to a big data environment would allow to exploit the advantages of machine learning methods to extract and select relevant information from large volumes of data, and get further nowcasting improvements. Third, an interesting application would be to develop alternative country-specific non-English dictionaries and estimate machine learning models to investigate the best approach for nowcasting GDP in each country. Fourth, another interesting area of work is the decomposition of sentiment indicators into drivers using a topic-modelling technique such as a Dynamic Topic Model (Blei and Lafferty, 2006) and analyse their effects on the economy in a SVAR environment.

References

Aguilar, P., Ghirelli, C., Pacce, M., and 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., and 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., and Verbeken, B. (2021). Daily news sentiment and monthly surveys: A mixed–frequency dynamic factor model for nowcasting consumer confidence. *National Bank of Belgium Working paper*, 396.

Aprigliano, V., Emiliozzi, S., Marcucci, J., Luciani, A., and Libero, M. (2021). The power of text-based indicators in forecasting the italian economic activity. *Banca d'Italia working paper*, 1321.

Ardia, D., Bluteau, K., and 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.

Azqueta-Gavaldon, A., Hirschbuehl, D., Onorante, L., and Saiz, L. (2020). Economic policy uncertainty in the euro area: an unsupervised machine learning approach. *European Central Bank working paper*, 2359.

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

Bai, J. and Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 24(4):607–629.

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

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

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

Barbaglia, L., Consoli, S., and Manzan, S. (2020). Forecasting with economic news. *Available at: https://ssrn.com/abstract=3698121.*

Bianchi, D., Buchner, M., and Tamoni, A. (2020). Bond Risk Premia with Machine Learning. *The Review of Financial Studies*, 34(2):1046–1089.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

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

Bühlmann, P., Hothorn, T., et al. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical science*, 22(4):477–505.

Cimadomo, J., Giannone, D., Lenza, M., Monti, F., and Sokol, A. (2021). Nowcasting with large Bayesian vector autoregressions. *CEPR Discussion Paper No. DP15854*.

Correa, R., Garud, K., Londono, J. M., Mislang, N., et al. (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., and Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? *arXiv preprint arXiv:2008.12477*.

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

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

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

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Ghysels, E., Santa Clara, P., and Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models. *CIRANO Working papers*.

Gilbert, C. H. E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf*.

Hansen, S., McMahon, M., and Prat, A. (2017). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

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

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Huang, C., Simpson, S., Ulybina, D., and Roitman, A. (2019). News-based sentiment indicators. *International Monetary Fund working paper*, 19/273.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., and Kapadia, S. (2020). Making text count: economic forecasting using newspaper text. *Bank of England Staff Working Paper*, 865.

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

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

Ng, S. (2014). Boosting recessions. *Canadian Journal of Economics/Revue canadienne d'économique*, 47(1):1–34.

Nielsen, F. Å. (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., and 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. and Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting*, 36(4):1501–1516.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics, in 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. (2018). 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. Technical report, OECD Economics Department Working Papers 1634, Paris: OECD Publishing.

Appendix A News article sources

We choose sources with a wide circulation in their respective countries, covering a broad range of political leanings (see Table 5). The news coverage varies over time (Figure 14). It is relatively good for France, Italy and Spain over the sample period, while it is a bit poor for Germany, since there are news articles from only two sources (only one before 2004).

Table 5: News sources

Country	Source	Total articles	Daily Circulation	Political leaning
France	Les Échos[1]	691,287	120,546	economic liberal
	Le Figaro[1]	341,925	313,541	centre-right
	Le Monde[1]	222,260	302,624	centre-left
Germany	Süddeutsche Zeitung[2]	514,284	361,507	centre-left
	Die Welt[2]	180,694	165,686	centre-right
	Der Tagesspiegel[2]	71,095	113,716	liberal
	German Collection[3]	67,841	-	-
Italy	Corriere della Sera[4]	412,944	258,991	liberal
	La Repubblica[4]	263,339	176,010	progressive
	Il Sole 24 Ore[4]	605,480	145,685	liberal
	La Stampa[4]	216,146	115,870	social liberal
Spain	Expansión[5]	634,659	50,180	liberal conservative
	El Mundo[5]	174,651	248,463	liberal conservative
	El País[5]	354,613	359,809	centre-left
	La Vanguardia[5]	243,611	180,939	liberal

1: Circulation data from `https://en.wikipedia.org/wiki/List_of_newspapers_in_France`, as of 3-Feb-2021.

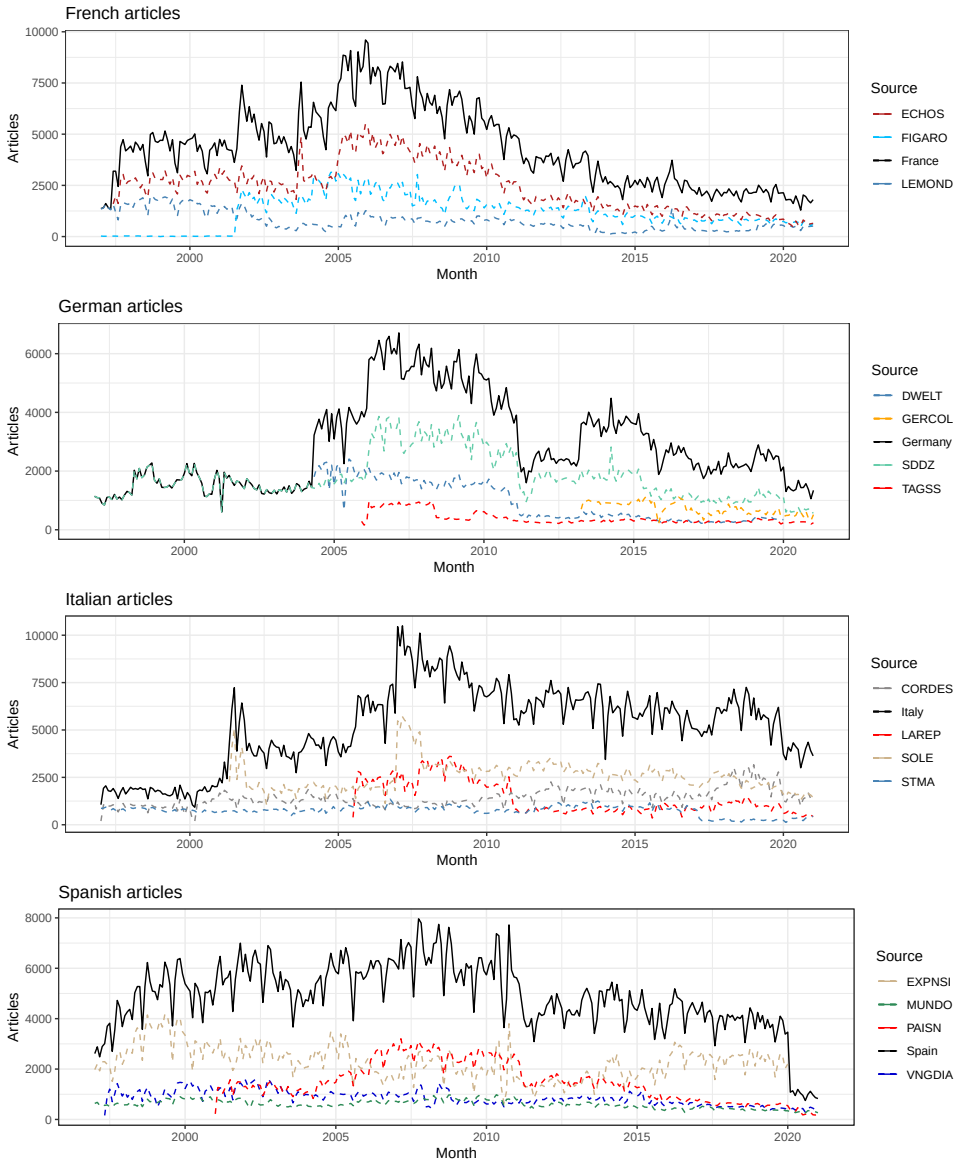
2: Circulation data from `https://en.wikipedia.org/wiki/List_of_newspapers_in_Germany`, as of 3-Feb-2021.

3: Collection of abstracted company, industry and financial news from the leading German general, business and financial newspapers including *Boersen-Zeitung*, *Handelsblatt*, *Süddeutsche Zeitung* and *Frankfurter Allgemeine Zeitung*.

4: Circulation data from `https://en.wikipedia.org/wiki/List_of_newspapers_in_Italy`,as of 3-Feb-2021.

5: Circulation data from `https://en.wikipedia.org/wiki/List_of_newspapers_in_Spain`, as of 3-Feb-2021.

Figure 14: Articles across time



Appendix B Translation

Table 6: Correlation of monthly metrics across two translation methodologies

sent_metrics1	France	Germany	Italy	Spain	Euro area
CGLM	0.75	0.704	0.862	0.89	0.872
LM	0.686	0.602	0.881	0.867	0.843
AFINN	0.746	0.793	0.878	0.839	0.87
HIV	0.729	0.176	0.764	0.662	0.719
NKTGOS	0.578	0.707	0.581	0.789	0.725
HL	0.644	0.888	0.902	0.81	0.867

Figure 15: French Translation Example

(a) Original French

Aquoi ressemblerait un groupe français qui résulterait de la fusion de Renault, d'Accor, de Capgemini, de Danone et d'Arcelor? Probablement à un monstre ingouvernable et rapidement promis au déclin. A l'heure où même une entreprise beaucoup moins diversifiée comme Veolia s'interroge sur son avenir, le profil de Tata laisse songeur.

(b) English Translation

What would a French group look like resulting from the merger of Renault, Accor, Capgemini, Danone and Arcelor? Probably to an ungovernable monster and quickly on the verge of decline. At a time when even a much less diversified company like Veolia is wondering about its future, Tata's profile leaves one wondering.

Figure 16: German Translation Example

(a) Original German

Bundeskanzlerin Angela Merkel lässt offen, ob von deutschen und europäischen Sanktionen gegen Russland wegen der Vergiftung des Oppositionspolitikers Alexej Nawalnyj auch russische Gaslieferungen oder das Pipeline-Projekt Nord Stream 2 betroffen sein könnten. Am Donnerstag äußerte sie sich in Berlin nach einem Treffen mit dem schwedischen Ministerpräsidenten Stefan Löfven. Forderungen, die umstrittene Pipeline Nord Stream 2 nicht fertigzustellen, wurden aus mehreren Parteien laut. Sie kamen von den Grünen, der FDP und der CDU

(b) English Translation

Chancellor Angela Merkel leaves open whether the German and European sanctions against Russia for poisoning opposition politician Alexej Navalnyj could also affect Russian gas deliveries or the Nord Stream 2 pipeline project. On Thursday, she made a statement in Berlin after a meeting with the Swedish Prime Minister Stefan Löfven. Demands not to complete the controversial Nord Stream 2 pipeline were voiced by several parties. They came from the Greens, the FDP and the CDU

Figure 17: Italian Translation Example

(a) Original Italian

Per migliorare efficienza e distribuzione servirebbero 80 euro per abitante: ora sono soltanto 34; Deficit di investimenti del 60%. Il surplus di polemiche politiche e il deficit di investimenti sono le due caratteristiche strutturali della gestione dell'acqua in Italia, e aprono le falle di una rete idrica che ormai arriva a perdere il 40% dell'acqua immessa nei tubi e di una rete di depurazione che ancora dimentica circa il 20% degli italiani.

(b) English Translation

To improve efficiency and distribution, 80 euros per inhabitant would be needed: now there are only 34; Investment gap of 60%. The surplus of political controversies and the investment deficit are the two structural characteristics of water management in Italy, and open the holes in a water network that is now losing 40% of the water fed into the pipes and in a network of purification that still forgets about 20% of Italians.

Figure 18: Spanish Translation Example

(a) Original Spanish

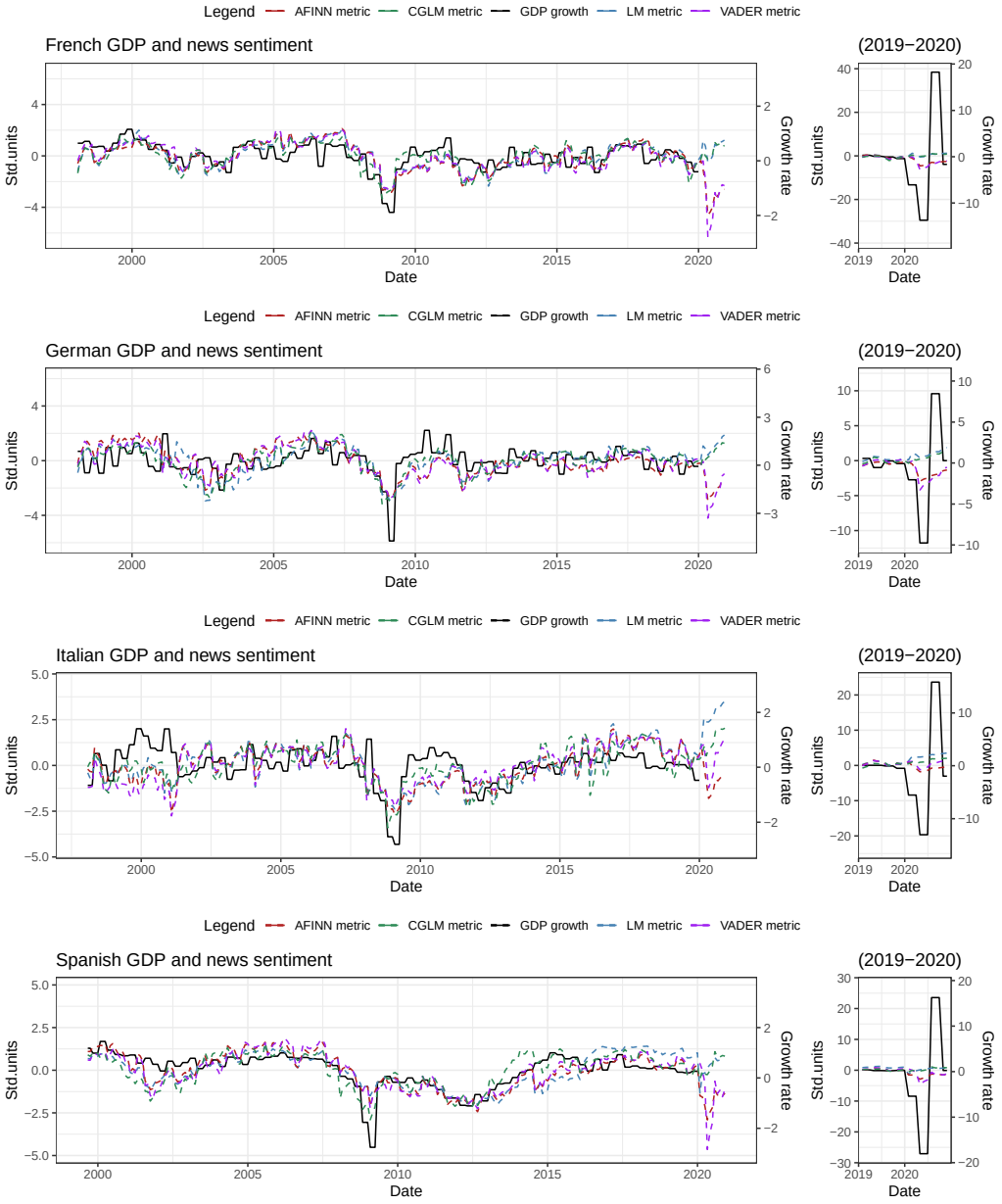
El grupo farmacéutico suizo Novartis registró un beneficio neto de 1.477 millones de dólares (1.128 millones de euros) en el primer trimestre, un 16% más. Sus ventas aumentaron un 11% hasta 7.341 millones de dólares. En el mismo periodo, la norteamericana Merck ganó 1.370 millones de dólares, con una caída del 15,4%. Sus ventas se redujeron un 5%, hasta 5.360 millones de dólares.La también estadounidense Schering-Plough obtuvo un beneficio neto de 127 millones de dólares hasta marzo, frente a las pérdidas de 73 millones de dólares del mismo periodo del año anterior.La cifra de negocio fue de 2.369 millones de dólares, un 21% más.

(b) English Translation

The Swiss pharmaceutical group Novartis posted a net profit of 1,477 million dollars (1,128 million euros) in the first quarter, up 16%. Its sales increased 11% to $7,341 million. In the same period, the North American Merck earned 1.37 billion dollars, with a fall of 15.4%. Its sales fell by 5%, reaching 5.36 billion dollars. The also American Schering-Plow obtained a net profit of 127 million of dollars until March, compared to the losses of 73 million dollars in the same period of the previous year. The turnover was 2.369 million dollars, 21% more.

Appendix C Sentiment metric plots

Figure 19: Country-level sentiment series



Appendix D Machine Learning Nowcasts

Table 7: Nowcasting results on a monthly basis for GDP growth using VADER - pre-2020 period

		Levels		First differences		Projections		Projections and First Diff.	
Model	Month	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$
AverageML	1	1.004	1.042	0.891**	0.943*	0.997	1.118	0.991	1.094
AverageML	2	1.033	1.063	0.906**	0.962	0.986	0.987	0.978	0.988
AverageML	3	1.008	1.126	0.883	0.953	0.956*	1.026	0.901	0.965
Boosting	1	1.038	1.068	0.889**	0.965**	1.011	1.123	1.013	1.102
Boosting	2	1.006	1.088	0.856***	0.977	0.983	1.035	0.98	1.016
Boosting	3	0.930**	1.14	0.786***	0.982	0.923	1.087	0.911**	1.044
Forest	1	0.999	1.038	0.881	0.943	0.989	1.111	0.999	1.081
Forest	2	1.005	1.052	0.915	0.996	0.977	0.995	0.974	0.983
Forest	3	0.909*	1.088	0.908	0.999	0.936	1.076	0.917	0.975
NN	1	1.046	1.126	0.985	1.023	0.997	1.159	0.965	1.128
NN	2	1.134	1.131	1.025	1.040	1.017	0.987	0.995	0.998
NN	3	1.384	1.361	1.026	1.034	1.067	1.009	0.898**	0.965
OLS	1	0.981*	1.059	0.918	1.100	1.005	1.005	1.012	1.053
OLS	2	0.997	0.969	0.893	1.132	1.007	1.007	1.007	0.952**
OLS	3	0.984	0.946**	0.81	1.317	1.006	1.006	1.003	0.924
Ridge	1	0.954	0.977	0.943*	0.960	1.036	0.980	0.979	1.008
Ridge	2	1.097	1.130	1.011	1.008	1.027	1.013	0.987	1.049
Ridge	3	1.222	1.294	1.004	1.045	1.053	1.006	0.976	1.127

Notes: Relative RMSEs using the VADER metric and the PMI series as predictors to predict GDP growth compared to only PMI-based model. Results are compared to the relative ML counterpart and the linear model. For the case of the OLS, the columns relative to the 'OLS_ben' correspond to the case where the benchmark is the PMI in levels. Significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. $***\backslash**\backslash*$ indicates significance at $10\%, 5\%, and 1\%$, respectively.

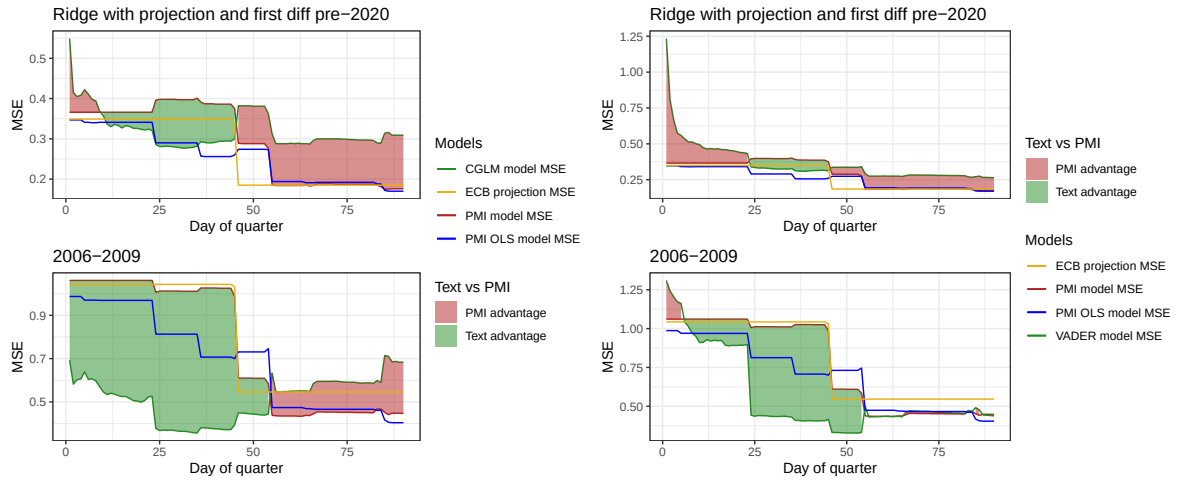
Table 8: Nowcasting results on a monthly basis for GDP growth using CGLM - pre-2020 period

		Levels		First differences		Projections		Projections and First Diff.	
Model	Month	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$	$\frac{RMSE_{text}}{RMSE_{ML\_ben}}$	$\frac{RMSE_{text}}{RMSE_{OLS\_ben}}$
AverageML	1	0.965**	1.082	0.864***	0.914	0.955	1.054	0.918	0.953
AverageML	2	0.987	0.988	0.815	0.866**	0.989*	0.999	0.966	0.994
AverageML	3	1.010	1.085	0.776	0.838**	0.950*	1.017	1.033	1.154
Boosting	1	0.982	1.091	0.857	0.931	0.971	1.056	0.935	0.962
Boosting	2	0.970	1.021	0.760	0.868**	0.975	1.011	0.974	1.054
Boosting	3	1.011	1.191	0.700	0.874**	0.946	1.083	0.998	1.223
Forest	1	0.973	1.092	0.885	0.947*	0.985	1.066	0.918	0.954
Forest	2	0.985	1.003	0.831	0.905**	1.009	1.019	0.932*	0.976
Forest	3	1.079	1.241	0.881	0.970	1.054	1.121	1.108	1.326
NN	1	0.937	1.089	0.883***	0.916**	0.907	1.060	0.903	0.972
NN	2	1.015	0.985	0.892**	0.905*	0.981	0.983	1.010	1.008
NN	3	1.000	0.946*	0.820**	0.826**	0.888**	0.954*	1.048	1.031
OLS	1	0.972	0.972	0.844**	1.011	0.975	1.015	0.921	0.995
OLS	2	1.016	1.016	0.836**	1.060	1.019	0.963	0.970	0.944
OLS	3	1.009	1.009	0.702**	1.143	1.014	0.933*	0.917*	0.881***
Ridge	1	0.990	0.936	0.831***	0.847*	0.903	0.930	0.900*	0.922
Ridge	2	1.074	1.059	0.913*	0.910	1.003	1.066	1.100	1.133
Ridge	3	1.164	1.111	0.880*	0.915*	0.966	1.115	1.323	1.401

Notes: Relative RMSEs using the stability metric and the PMI series as predictors to predict GDP growth compared to only pmi-based model. Results are compared to the relative ML counterpart and the linear model. Significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. $***\backslash**\backslash*$ indicates significance at $10\%, 5\%, and 1\%$, respectively.

Figure 20: Ridge regression with projections and first differences: pre-2020 and Great Recession

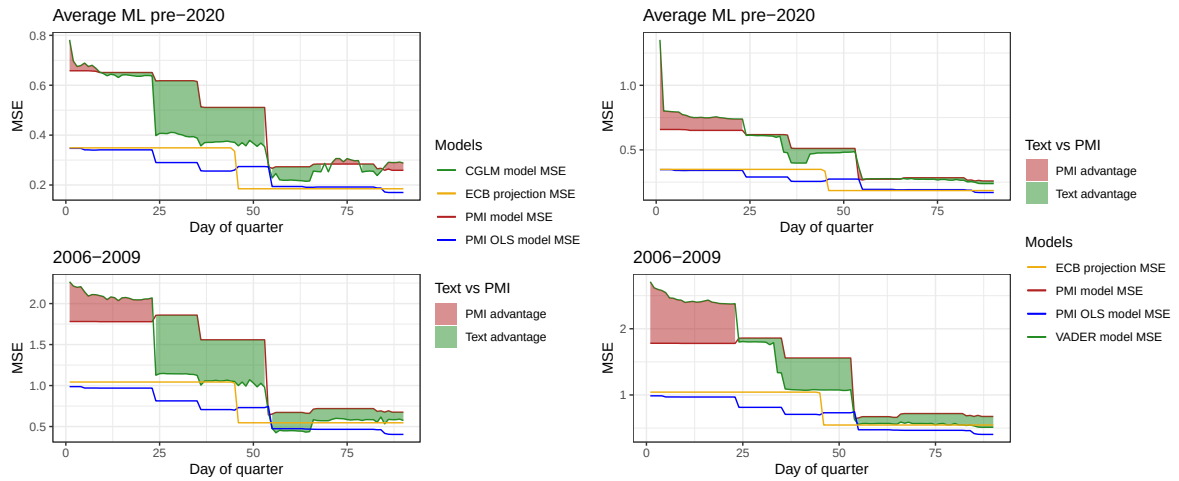
(a) CGLM metric (b) VADER metric



Notes: see Figure 8 for full explanation. Results here are for a Ridge Regression with the ECB projections and first differences included.

Figure 21: Average ML in levels: pre-2020 and Great Recession

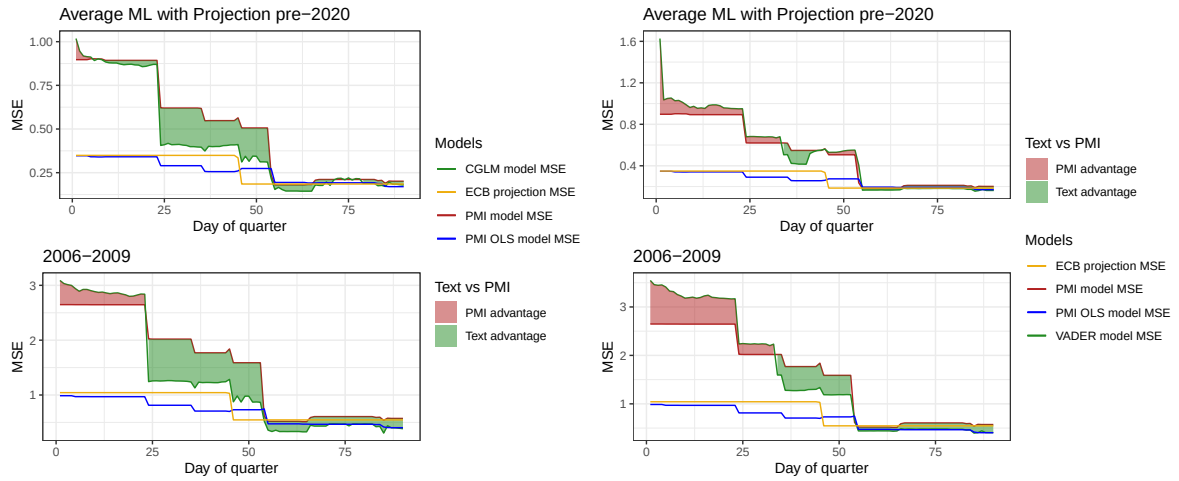
(a) CGLM metric (b) VADER metric



Notes: see Figure 5 for full explanation. The results shown here are for ....

Figure 22: Average ML with projections: pre-2020 and Great Recession

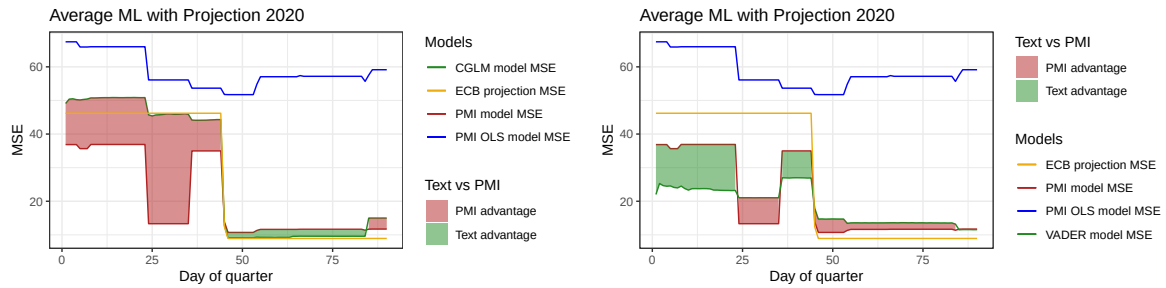
(a) CGLM metric (b) VADER metric



Notes: see Figure 5 for full explanation. The results shown here are for ....

Figure 23: Average ML with ECB projection: 2020

(a) CGLM metric (b) VADER metric



Notes: see Figure 5 for full explanation. The results shown here are for ....

Appendix E Alternative models: Technical details

E.1 Ridge Regression

Ridge Regression is a shrinkage method that penalises the residual sum of squares (RSS) with the sum of squared coefficients (L2-norm). This shrinks the coefficients of those predictors with a minor contribution in terms of predictive ability of the model towards zero, albeit they never become exactly zero. As such, the Ridge regression is a dense modelling technique—it uses the full range of predictors, although assuming that the contribution of many of them might be

small. Under our framework, the optimisation problem can be written as:

$$\beta^{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{d}^{T} (y_{q,d} - \alpha - \sum_{j}^{N} \beta x_{q,d,j})^2 + \eta \sum_{j}^{N} \beta_j^2 \right\} \tag{5}$$

for given values of α and $\eta \geq 0$. It is common practice to centre the values of predictors around the mean first, and not to include the constant term.[15]

The parameter η stands for the penalty imposed on coefficients and controls its overall magnitude. We have $\hat{\beta}^{Ridge} \to \beta^{\hat{O}LS}$ as $\eta \to 0$ which is the no penalty case, and $\hat{\beta}^{Ridge} \to 0$ as $\eta \to \infty$. Selecting a good value for the tuning parameter η is crucial and is done via cross-validation.

E.2 Non-Linear Machine Learning Models

Tree Models and Random Forests. Tree models are a non-parametric methods for both regression and classification problems. Their basic idea is to consecutively split the training dataset until an assignment criterion with respect to the target variable into a "data bucket" (leaf) is reached. The algorithm minimises the objective function within areas of the target space, i.e. these "buckets", conditioned on the input $x_{q,d}$. Splitting the vector of predictors $x_{q,d}$ to M subspaces i.e $P = \{P_1, \dots P_M\}$, the optimal estimates of β coefficients is just the average of the estimated Y in each region in the training sample. The regression function is

$$y_{t+h} = \sum_{m=1}^{M} \beta_m I(x_{q,d} \in P_m) + \varepsilon_t, \quad \text{with} \quad \beta_m = 1/|P_m| \sum_{y^{tr} \in P_m} y^{tr},\ m \in \{1, \dots, M\}. \tag{6}$$

A disadvantage of regression trees is that they are not identically distributed: they are built adaptively to reduce the bias. This may lead to severe over-fitting. "Random Forest" (Breiman, 2001), or similar ensemble approaches, are routinely used to overcome this problem. A random forest contains a set of uncorrelated trees which are estimated separately. The predictions of the individual trees are averaged for a single prediction reducing variance. A general drawback of random forests, as compared to single trees, is that they are hard to interpret due to the built-in randomness with causes the differences between individual trees.

[15] The reason for this is that the ridge regression coefficients estimates can substantially change when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the Ridge Regression objective function.

Tree models are usually sparse models as their hierarchical structure acts like a filter. That is, only variables which actually improve the fit are chosen during construction or growth. This makes them particularly suitable to our high-dimensional setting.

Gradient Boosting Regressor. As an alternative to including many regressors simultaneously in a penalised regression setup, a number of papers have developed methods that focus on the predictive power of individual regressors instead of considering all N covariates together. This approach has led to a variety of alternative specification methods sometimes referred to collectively as "greedy methods". In this context, regressors are chosen sequentially based on their individual ability to explain the dependent variable. Perhaps the most widely known of such methods, developed in the machine learning literature, is "boosting" whose statistical properties have received considerable attention, see Friedman (2001).

Boosting is an iterative procedure where misclassified observations are given increasing cost in each estimation repetition. The idea is to consider regressors one by one in a simple regression setting, and successively selecting the best fitting ones, giving rise to 'greedy' algorithms. More details on boosting algorithms for linear models, and their theoretical properties can be found in Bühlmann et al. (2007). The algorithm can be described described as follows.

1. (Initialisation). Let $x_{q,d} = (x_{1t}, ..., x_{Nt})'$, $\mathbf{X} = (x_1, ..., x_N)$ and $\mathbf{e} = (e_1, ..., e_T)$. Define the least squares base procedure:

$$\widehat{g}_{\mathbf{X},\mathbf{e}}(x_{q,d}) = \widehat{\delta}_{\widehat{s}} x_{\widehat{s}t}, \quad \widehat{\delta}_i = \frac{\mathbf{e}'\mathbf{x}_i}{\mathbf{x}_i'\mathbf{x}_i}, \quad \widehat{s} = \min_{1 \leq i \leq N} \left(\mathbf{e} - \widehat{\delta}_i \mathbf{x}_i\right)' \left(\mathbf{e} - \widehat{\delta}_i \mathbf{x}_i\right)$$

2. Given data $\mathbf{X}$ and $\mathbf{y} = (y_1, ..., y_{q,d})'$, apply the base procedure to obtain $\widehat{g}^{(1)}_{\mathbf{X},\mathbf{y}}(x_{q,d})$. Set $\widehat{F}^{(1)}(x_{q,d}) = \upsilon \widehat{g}^{(1)}_{\mathbf{X},\mathbf{y}}(x_{q,d})$, for some $\upsilon > 0$. Set $\widehat{s}^{(1)} = \widehat{s}$ and $m = 1$.

3. Compute residuals $\mathbf{e} = \mathbf{y} - \widehat{F}^{(m)}(\mathbf{X})$ where $\widehat{F}^{(m)}(\mathbf{X}) = (\widehat{F}^{(m)}(\mathbf{x}_1), ..., \widehat{F}^{(m)}(x_{q,d}))'$ and fit the base procedure to the current residuals to obtain the fit $\widehat{g}^{(m+1)}_{\mathbf{X},\mathbf{e}}(x_{q,d})$ and $\widehat{s}^{(m)}$. Update

$$\widehat{F}^{(m+1)}(x_{q,d}) = \widehat{F}^{(m)}(x_{q,d}) + \upsilon \widehat{g}^{(m+1)}_{\mathbf{X},\mathbf{e}}(x_{q,d}).$$

4. Increase the iteration index m by one and repeat step 3 until the stopping iteration M is achieved. The stopping iteration is given by

$$M = \min_{1 \leq m \leq m_{\max}} AIC_c(m),$$

for some predetermined large $m_{\max}$. $m_{\max} = 500$ and $\upsilon = \{0.1, 1\}$ values can be used as suggested in the literature.

In an economic context, boosting has been applied by Bai and Ng (2009) and Ng (2014). For example, the latter uses boosting in order to screen a large number of potentially relevant predictors and their lags and give warning signals of recessions.

Neural networks. Neutral Networks are similar to linear and non-linear least squares regressions and can be viewed as an alternative statistical approach to solving the least squares problem. A standard architecture of ANN are multilayer perceptrons (MLP), a form of feed-forward network. The variables $x_{q,d}$ in the input layer are multiplied by weight matrices, then transformed by an activation function in the first hidden layer and passed on to the next hidden or the output layer resulting a prediction $y_{q,d}$. The number of hidden layers L determines the depth of a network, with deeper networks being generally more accurate but also needing more data to train them. Formally, this can be described as

$$y_{q,d} = G(x_{q,d}, \beta) + \varepsilon = g_L(g_{L-1}(g_{L-2}(\dots g_1(x_{q,d}, \beta_0), \dots, \beta_{L-2}), \beta_{L-1}), \beta_L) + \varepsilon \qquad (7)$$

The activation functions $g(\cdot)$ act as gates for signals and introduce non-linearity into the model. Common choices are rectified linear unit functions (ReLU) or the hyperbolic tangent. The activation of the last layer L mostly reflects the type of problem and is a linear matrix multiplication for our regression problem. Note that ANN can handle multiple input with multiple output situations, i.e. several time steps can be modelled using the same model, i.e. $H = \{1, \dots, 12\}$. Determining the number of layers L and the number of neurons in each layer as well as appropriate weight penalisation in our ANN is addressed by cross-validation discussed in the next subsection.

Acknowledgements

We would like to thank Beatrice Pierluigi, Grigor Stoevsky, Niccolò 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 for valuable comments and suggestions. We are grateful to George Kapetanios and Stephen Hansen for helpful discussions at the beginning of this project.

Julian Ashwin
European Central Bank, Frankfurt am Main, Germany; University of Oxford, Oxford, United Kingdom;
email: julian.ashwin@economics.ox.ac.uk

Eleni Kalamara
European Central Bank, Frankfurt am Main, Germany; King's College London, London, United Kingdom;
email: eleni.kalamara@kcl.ac.uk

Lorena Saiz
European Central Bank, Frankfurt am Main, Germany; email: lorena.saiz@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

PDF ISBN 978-92-899-4869-2 ISSN 1725-2806 doi:10.2866/240669 QB-AR-21-107-EN-N