Nowcasting GDP Growth by Reading Newspapers

Clément Bortoli*, Stéphanie Combes et Thomas Renault*****

Abstract – GDP statistics in France are published on a quarterly basis, 30 days after the end of the quarter. In this article, we consider media content as an additional data source to traditional economic tools to improve short-term forecast / nowcast of French GDP. We use a database of more than a million articles published in the newspaper *Le Monde* between 1990 and 2017 to create a new synthetic indicator capturing media sentiment about the state of the economy. We compare an autoregressive model augmented by the media sentiment indicator with a simple autoregressive model. We also consider an autoregressive model augmented with the Insee Business Climate indicator. Adding a media indicator improves French GDP forecasts compared to these two reference models. We also test an automated approach using penalised regression, where we use the frequencies at which words or expressions appear in the articles as regressors, rather than aggregated information. Although this approach is easier to implement than the former, its results are less accurate.

JEL Classification: E32, E37, C53
Keywords: economic analysis, nowcasting, GDP, media, Big Data, sentiment analysis, machine learning, natural language analysis

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

* *Insee, Economics department, at the time of writing this article (clement.bortoli@gmail.com)*
** *Insee, Statistical methods department, at the time of writing this article (stephanie.combes@gmail.com)*
*** *Paris 1 Panthéon Sorbonne University, CES & LabEx ReFi, IÉSEG School of Management (thomas.renault@univ-paris1.fr)*

Received on 20 September 2017, accepted after revisions on 18 May 2018
Translated from the original version: "Prévoir la croissance du PIB en lisant le journal"

To cite this article: Bortoli, C., Combes, S. & Renault, T. (2018). Nowcasting GDP Growth by Reading Newspapers. *Economie et Statistique / Economics and Statistics*, 505-506, 17–33. https://doi.org/10.24187/ecostat.2018.505d.1964

Because macroeconomic data are only known after a certain time lag, it is vital for policy makers to have tools enabling them to forge a real-time analysis of the economic situation. For example, GDP statistics in France are published on a quarterly basis, with a delay of 30 days after the end of the quarter. To get a forecast (or a nowcast) of GDP growth before its publication, forecasters traditionally use business climate surveys, conducted by different institutes, as the main information source. These are questionnaires made up of qualitative questions sent every month to a sample ranging from several hundred to several thousand business leaders. The answers are summarised as "opinion balances", i.e. by calculating the difference between positive and negative responses. Some synthetic indicators are also calculated from these opinion balances, such as the business climate that considers the economic situation overall or by sector. These various indexes are sometimes called "leading indicators", since they are available before official figures are published. It is also possible to forecast the GDP for the current quarter (which is obviously not known before the end of the quarter): this type of forecast is referred to as "nowcast". It can also be interesting to forecast GDP for the coming quarter, which is also possible *via* business climate surveys that contain prospective balances.

The current explosion of Internet content, as well as the technical progress associated with the "Big Data" offer the possibility of building alternative economic indicators in real time. From this point of view, media content is particularly interesting, because its properties are quite similar to those of surveys on business climate. Indeed, this information is available instantaneously and contains qualitative details about the economic situation several weeks before official data are published.

Thus, the aim of this paper is to use the content of a major media Internet site to improve real-time GDP growth forecasting (or nowcasting). It will be particularly interesting to compare the predictive power of this information with that from business surveys traditionally used, in order to determine if the two approaches substitute for each other, complement each other, or if one of them appears more accurate than the other.

For the purpose of this article, we chose the French newspaper *Le Monde* website. The content of this website is covering a time scale rare in France, in particular including many articles published in the paper edition before the advent of the Internet. In addition, it is the leading information website in France. We therefore built a database containing more than a million articles published in this newspaper from 1990 to today. We first sorted this database by combining statistical models and textual analysis, retaining only articles covering the French economic situation; this results in a sample of approximately 200,000 news items. We then use the information contained in this reduced database, following two different strategies.

The first strategy requires the use of a sentiment dictionary, in other words a list of terms with positive or negative connotations from an economic viewpoint. Such dictionaries are widespread in English, less so in French. We therefore built one containing 548 positive terms and 1,295 negative terms. These terms are then identified in each article in the database, which are allocated a "sentiment score" based on the number of positive and negative terms it contains. In this way it is possible to summarise the information contained in the database as a unique numerical indicator, called media sentiment. This can then be used in simple regression models (autoregressive or AR, augmented).

We then carry out a real-time[1] forecasting exercise over the period 2000-Q2 - 2017-Q3, which means that we make forecasts for a given timescale every quarter from the second quarter of 2000 to the third quarter of 2017, each time only using the data available up to that date. We compare the accuracy of each model by calculating the RMSFEs (Root Mean Square Forecast Error) from the series of forecast errors relative to the real value calculated in this way. Thus, we find that a model combining "Media Sentiment" and "Business Climate" provides, over certain timescales, significantly greater accuracy compared to an augmented AR model of business climate surveys alone.

The use of a "handmade" dictionary may appear to be somewhat arbitrary, costly, and imprecise since all the available information is summarised in a single indicator. A way to deal with these drawbacks is to consider automatic

[1] *Strictly speaking, one should talk about* pseudo real time*, as the sentiment dictionary is constructed* ex-post *by experts. However, for ease of language, we will speak of* real time *in the rest of this article.*

methods, which would avoid prior judgment on the terms to be selected or their connotation, while also keeping the information in a disaggregated format. The automatic methods used here also have the advantage of being relatively inexpensive to implement. It involves constructing the series for the frequency of appearance (or a weighting similar to the concept of frequency) for each term and combinations of two terms (or bigrams): to do so, the terms are first stemmed to bring singular and plural to the same form. These time series are then used for forecasting as part of penalised regressions (Elastic-Net). Penalisation ensures a selection of regressors and therefore the parsimony of the model, which helps to prevent a risk of over-adjustment, especially high given the large number of variables available. However, the calculation of RMSFEs suggests that an automatic method for selecting words does not significantly improve the forecast compared to the autoregressive model augmented with the business climate indicator.

The structure of this article is organised as follows: A brief literature review is presented in the first section. The second section describes the data used and the way they are handled. The econometric models used are then described in the third section. The fourth section presents the results obtained. The fifth section concludes.

Literature Review

It is possible to separate the literature dealing with GDP nowcasting into two major categories. On the one hand, a part of the literature is dealing with techniques aimed at choosing the best forecasting model from a predefined "traditional" set of variables. These papers are generally devoted mainly to comparing the predictive performance of different approaches: bridge models, state space model, mixed-data-sampling, blocking, etc. Among others, we can refer to Baffigi *et al.* (2004) and Foroni & Marcellino (2014). More recently, Bec & Mogliani (2015), in a paper devoted to comparing combinations of models and combinations of information, offer an instructive summary of the different techniques that can be used to make a macroeconomic forecast. On the other hand, a part of the literature, using a predefined model, deals with improving the forecast by considering the addition of new explanatory variables. Here we focus our attention on this second segment of the literature.

Four main types of variables are used in the literature: (1) "quantitative" variables (industrial production, retail sales, etc.), published monthly with a time delay of 30 to 45 days; (2) "qualitative" variables (surveys, polls, etc.), available at the end of every month; (3) "financial" variables (interest rate, stock market index, etc.) available in real time; and (4) "alternative" variables (Google Trends, media sentiment, etc.) often available in near real time.

There is a consensus regarding the contribution of adding "qualitative" variables, mainly when the "quantitative" information about the current quarter is not yet available. For example, by analysing the contribution of each variable depending on the timing of the GDP forecast for the current quarter (1st month, 2nd month or 3rd month), Angelini *et al.* (2011) showed that "qualitative" information carried greater weight for the first estimates, while the predictive power of "quantitative" information becomes predominant for estimates in the 3rd month. This change is explained quite simply by the fact that "quantitative" information about the current quarter starts to become available during the 3rd month (e.g. industrial production for January 2016 was published on 15 March 2016 and can therefore be used to nowcast GDP for the 1st quarter of 2016 conducted during the 3rd month of the same quarter): this "quantitative" information is used by national accountants in order to construct the GDP on a quarterly basis. The contribution of qualitative information was confirmed by Darné (2008), among others, for the specific case of France.

The conclusions are more mixed regarding the contribution of financial variables. According to Andreou *et al.* (2013), adding financial variables improves the accuracy of the model, while the opposite findings are presented by Banbura *et al.* (2013). This difference is explained by the fact that Andreou *et al.* (2013) do not use the high frequency of indicators by extrapolating the monthly data over the quarter (unlike Banbura *et al.*, 2013), making it difficult to compare the two studies.

Finally, more recently, different studies have focused on the contribution of "alternative" variables. For example, Choi & Varian (2012), McLaren & Shanbhogue (2011), Fondeur & Karamé (2013), and D'Amuri & Marcucci (2017) showed that the change in the search volume on Google Trends for certain keywords ("jobless claims", "unemployment benefits") improved the forecast for the change in unemployment rate. Regarding the contribution of *Google Trends* to predicting the French economic situation, more mixed findings were put forward by Bortoli & Combes (2015).

Content published in the media has also been used extensively in finance to forecast the change in financial markets (Tetlock, 2007; Garcia, 2013). One possible approach is to calculate a sentiment score for a press article, then to construct a time series for "sentiment" by aggregating the scores for articles published over a given period (e.g. every month). To do so, a dictionary containing a list of "positive" keywords and a list of "negative" keywords, either generic (Harvard IV dictionary) or specific to the study area (e.g. Loughran & McDonald's financial dictionary, 2011), is used: the sentiment of each article is then simply defined using the frequency of words from the dictionary in the body text weighted by their score (in the simplest case, +1 for a word with positive connotations and -1 for a word with negative connotations).

The approach founded on a dictionary or sentiment score is not always based on a binary positive/negative approach: Baker *et al.* (2016) looked at the change in the number of articles containing at least one keyword linked to a sentiment of uncertainty and dealing with economic policy, in order to create a new index (Economic Policy Uncertainty Index).[2] The authors of this new index show that an increase in media uncertainty helps to forecast changes in GDP.

Another possible approach using "media" data consists of analysing the change in the frequency of appearance of different subjects detected automatically using an unsupervised approach such as Latent Dirichlet Allocation. Applying this methodology to Norway, Larsen & Thorsud (2015) showed that the variation in frequency of appearance of certain subjects may be used to improve the forecast of economic fluctuations.

In this article, we focus on the forecast at the end of the 1st month, the 2nd month and the 3rd month of the current quarter and the previous quarter. We compare the accuracy of a simple AR model with an AR model augmented by the business climate and an AR model augmented by alternative "media" data (summarised or disaggregated).

Data

The Original Database

Among the different French media whose content can be used to construct a media sentiment indicator, *Le Monde* has interesting features. It is one of the leading French newspapers: the printed edition has the second highest national circulation behind *Le Figaro* (approximately 260,000 copies per day) and its website lemonde.fr is the most-visited information site in France, just ahead of the website of *Le Figaro*. In addition, the media content available online covers a remarkable time scale for France, including many articles published in the paper edition before the advent of the Internet. Thus, it was possible to create a database of 1,405,038 online articles published since 1990.

It might also be interesting to use articles from specialised economic newspapers such as *Les Echos* or *La Tribune*. In fact, *Les Echos* website also has interesting properties, as articles have been available since 1991. However, this newspaper has a lower media outreach than *Le Monde* (whether in terms of number of printed copies sold or visits to the Internet site): for this article, we chose to favour the "general public" source. It could therefore be interesting for a future study to estimate if "specialist" information has stronger predictive power than generalist media. On the other hand, the use of *La Tribune* seems more problematic: the risk of a break in the series over a long period is high in relation to the predictive power of online content, due to the radical change in editorial line that occurred in 2012.

The number of articles contained in the database varies strongly depending on the period, most of the time between 2,000 and

[2] *www.policyuncertainty.com. For France, the EPU index is based only on articles from the newspapers* Le Monde *and* Le Figaro *(which justifies the comparison we are making* infra *between this indicator and our media sentiment index). However, for the United States, the EPU indicator is based on three components, one of which relates to the media.*

6,000.[3] This limit is exceeded between 2000 and 2002, where the series reached its maximum (11,000 in March 2001), then more briefly in 2012. Since 2013, the number of articles per month has oscillated around 4,000.

Construction of a Restricted Database

First, it is necessary to sort this database to keep only the articles relevant to our study, i.e. those covering economic topics and dealing mainly with the situation in France. In fact, keeping more articles could interfere with summarising media information and its use in forecasting. It is also necessary to remove articles from the database dealing with information published by statistical institutes (Insee, Dares – the ministry of Labor statistical services, *Pôle emploi* – the French employment agency, etc.), because the information we are looking for in the media content has to be independent from these sources. Moreover, some articles are reserved for subscribers: in this case, only the title and first lines are freely available. We restrict our analysis to articles where there are at least 50 characters freely accessible.

We first discard all articles not dealing with economics. The more recent articles (since 2005) are already classified into categories (economics, international, politics, sport, etc.) by journalists at *Le Monde*. This classification is registered in the metadata for each article, and can therefore contribute to identifying articles dealing with economics among the older texts that have not been pre-classified by the journalists. A supervised learning algorithm is calibrated using a sample of 25,000 articles from the "economics" category and 25,000 articles from other categories: the algorithm calculates the probability of an article belonging or not belonging to the "economics" category, based on the frequency of appearance of words contained in it for the two sets of the training dataset. In this way, the presence of the word "employment" in an article will increase its probability of belonging to the "economics" category, as in the training dataset this word is most frequent in articles covering economics than in the others. Such an algorithm, which can be qualified as "naive Bayesian" (Kotsiantsis *et al.*, 2006), makes it possible to classify all the oldest texts in the database very rapidly. By analysing the accuracy of the classification on 10,000 articles (out-of-sample), we get a classification accuracy of 89.7%; this supports our use of this type of approach to categorise all the articles in our database.

In parallel, the articles that focus mostly on France are identified by another procedure. Two lists containing the names of geographical entities are used: one is made up of French toponyms (names of towns, *départements*, regions) and the other international toponyms (names of countries and capitals). We retain only the articles that include at least as many French entities as foreign entities.

The final sample contains 194,848 articles. The proportion of articles retained each month oscillates between 10% and 20%. This proportion seems to follow a falling trend over the recent period: it was 18% in 2009 and not more than 13% in 2016.

The Traditional Economic Indicators: Insee Business Surveys

One of the important ideas of the article is to compare the information contained in the media content with that synthetized in more traditional economic variables such as business surveys.

Business climate surveys are used to follow the recent and current economic situation, and to forecast short-term changes. They are conducted every month among company managers. They provide an overview of a given business sector, highlighting the fields that are not covered, or covered more belatedly, by traditional statistics. The information gathered in business climate surveys are referred to as qualitative, because the respondents are asked to assign qualities and not quantities to variables about which questions are asked.

For France, the three main producers of business climate surveys are Insee, *Banque de France* and Markit (PMI surveys). For this paper, we relied only on Insee business surveys, and more particularly on the synthetic Business Climate indicator. This is the common component, extracted using factor-based analysis techniques, of 26 business surveys in five different sectors (industry, services, construction, retail and wholesale trade). The Business Climate indicator is normalised: over the long period, its mean is 100 with a standard deviation of 10.

[3] The database is visibly atypical in 2006, where the number of articles per month was highly discontinuous compared to prior and subsequent periods (barely more than 1,000 articles per month).

The Variable to Be Forecast: GDP Growth

The variable we aim to forecast is the real quarterly volume growth of the French GDP (chain-linked), seasonally and working-day adjusted, published by Insee. There are three publications for each quarter (two before 2016): a first estimate 30 days after the end of the quarter, a second estimate 60 days after the end of the quarter and "detailed figures" 85 days after the end of the quarter.[4] The quarterly growth figures may still change further over three years, until the national accountants publish the final account for the year considered. After this date, GDP growth for a given quarter no longer changes beyond the normal fluctuations associated with corrections for seasonal variations.

Knowing whether it is best to measure the performance of a forecasting model over the series of first publications of GDP or over a given recent vintage ("final" series) is a question with no obvious answer. As stated by Bec & Mogliani (2015), it is possible to defend an economic forecast having the main purpose of giving political decision-makers the best possible estimate of business activity: from this viewpoint, it would be better to test our models using given historical data, preferably the most recent possible ("final" series). In fact, GDP growth values in this case closely match the best possible measurement of economic activity, once all the information is available. Thus, Mogliani & Ferrières (2016) show that, in the case of France, revisions of the GDP are generally not biased, but that the first growth estimates do not efficiently use all the available macroeconomic and financial information.

Nonetheless, from a pragmatic viewpoint, it is true that the performance of one forecasting method is *de facto* judged in the light of the first GDP figures published. In this article we have thus chosen to adopt a real-time approach, i.e. using historical first publication data. In particular, this is justified by the fact that we are using time lags in GDP as explanatory variables in our model: so, we are using the information that was available during the quarter to be forecast. Nonetheless, as a precaution, all the estimations in this paper have also been made using a recent vintage of GDP growth: the results are very similar to those presented here.

Models

We propose two different strategies for extracting media information from the database and for using it in forecasts. The first consists in constructing a "Media Sentiment" index that offers a numerical figure for the general tone of the articles, following a procedure similar to that applied in Bortoli *et al.* (2017). The second uses all the available information by calculating the change over time in the occurrence frequencies of the terms in the database. These time series are then used in forecasts as part of penalised regressions.

Constructing an Indicator of "Media Sentiment" and Using it in a Forecast Model

A first strategy for extracting the media information from the database consists in constructing a Media Sentiment indicator that gives a score for the general tone of the articles in the database. The main advantage of this method is that it provides a tool very similar to more traditional economic indicators, such as Business Climate indicators: thus, it will be possible to compare the predictive performance of our Media Sentiment indicator to the Business Climate constructed by Insee. In addition, this is an easily interpreted indicator: a simple reading discloses the cyclical position of the economy as established by the indicator.

Choice of Frequency for the Media Sentiment Indicator

The first strategic choice to be made for the media sentiment indicator is its frequency. Given the database created, it would be possible to create a quarterly, monthly, weekly or even daily index. We have chosen to ignore the last two possibilities:

- A daily indicator would risk appearing too volatile, even more since the number of articles published is likely to vary significantly from one day of the week to another (with a particular drop at weekends, especially Sunday);

- A weekly indicator would be problematic to use to forecast a quarterly variable such as GDP, given that these two frequencies do not "fit" one

[4] *Before 2016, the first results were published 45 days after the end of the quarter and there was no additional publication before the detailed figures.*

inside the other (a quarter does not contain a fixed whole number of weeks). In addition, such an indicator could present a risk of volatility that would still be too high.

It therefore remains to choose between quarterly and monthly frequencies. The first solution would have the advantage of minimising the noise contained in the indicator. However, it would be necessary to wait for the end of the quarter to calculate it. Conversely, a monthly indicator offers a forecasting model from the first month of the quarter, without waiting for its end. Thus, the monthly indicator appears to offer the best compromise between volatility and frequency/speed of availability (in theory, from the end of every month). Furthermore, this frequency is also chosen by major institutions to publish their main economic and business climate statements.

Construction of the Sentiment Dictionary

Calculating a media sentiment indicator requires being able to quantify the positive or negative tone of the articles selected; to do so, we use a "sentiment dictionary". This is a list of terms that may have positive or negative connotations. Many dictionaries already exist in English to analyse texts: the Harvard IV-4 Psychological Dictionary is the main one, but other dictionaries are used for specific research fields, such as the Loughran & McDonald dictionary (2011) in the field of finance. However, this type of pre-existing list is much rarer in French: it is therefore necessary to construct one for the needs of this study.

We began by stemming all the terms encountered in the corpus studied using the Snowball algorithm adapted for the French language (Porter, 2001). We then assigned a sentiment to all stems appearing more than 500 times in the corpus (i.e. 5,575 stems), based on three possible ratings: positive, neutral or negative.

However, building a dictionary comprising exclusively unique stems (or unigrams) could prove problematic. In reality, a stem such as "increase" does not have the same value depending on whether one is discussing an increase in growth or unemployment. To overcome this type of ambiguity, we supplemented the dictionary with a list of bigrams, i.e. pairs of stems. Similar to what we did for the unigrams, we identified the commonest 5,000 bigrams from the corpus, then we classified them according to the same three ratings. In total, the dictionary contains 840 terms, 281 positive and 559 negative.[5]

Allocating a Score to Each Article and Calculating the Media Sentiment Indicator

From the dictionary created, a "sentiment score" is attributed to each article *i*, depending on the number of positive and negative terms that it contains. Several scoring systems can be considered. The simplest coding consists in adopting a discreet score for each article (discrete coding). The attributed score is 1 if the article contains more positive than negative terms, -1 if it contains more negative than positive terms and 0 if the two categories are equal. Discrete coding has the merit of simplicity, but it does not distinguish articles where the overall connotation is very marked from those where it is more subtle. It may therefore be interesting to consider an alternative scoring system, where the score can be established on a continuous scale between 1 and -1 (continuous coding). To do this, we calculate for each article the difference between the number of positive words and number of negative words, then we normalise for the number of words in the article.

The value of the sentiment indicator for month *t* is then a simple arithmetic mean of the sentiment scores obtained for each article *i* published during the month. Labelling *n(t)* the number of articles published in month *t*, $S_{i,t}$ the sentiment associated with each article *i* published during month *t*, we therefore define a monthly sentiment variable $MediaSent_t$, such that:

$$MediaSent_t = \frac{1}{n(t)} \sum_{i=1}^{n(t)} S_{i,t}$$

It is thus possible to calculate two monthly media sentiment indicators: one based on continuous coding and the other based on discrete coding. These two indicators are obviously very similar over the period[6] (Figure I): this result is already reassuring in itself, as it shows that our method makes it possible to extract from the articles database an overall media sentiment that does not depend too much on the parameters chosen to do so. We also note that the indicator is always negative,

[5] The dictionary is available online: http://www.thomas-renault.com

[6] In Figures I, II, III and IV, the media sentiment indicators are smoothed for reasons of legibility (3rd order moving averages). However, the raw indicators are used in the forecasting models.

irrespective of the chosen coding, which denotes a generally pessimistic bias across the articles selected by the filter. Furthermore, we note that using continuous coding leads to obtain a less volatile indicator than discrete coding and better considers the nuances developed in the text of these articles. In the rest of this article, we select the continuous indicator as it provides better forecasting results.

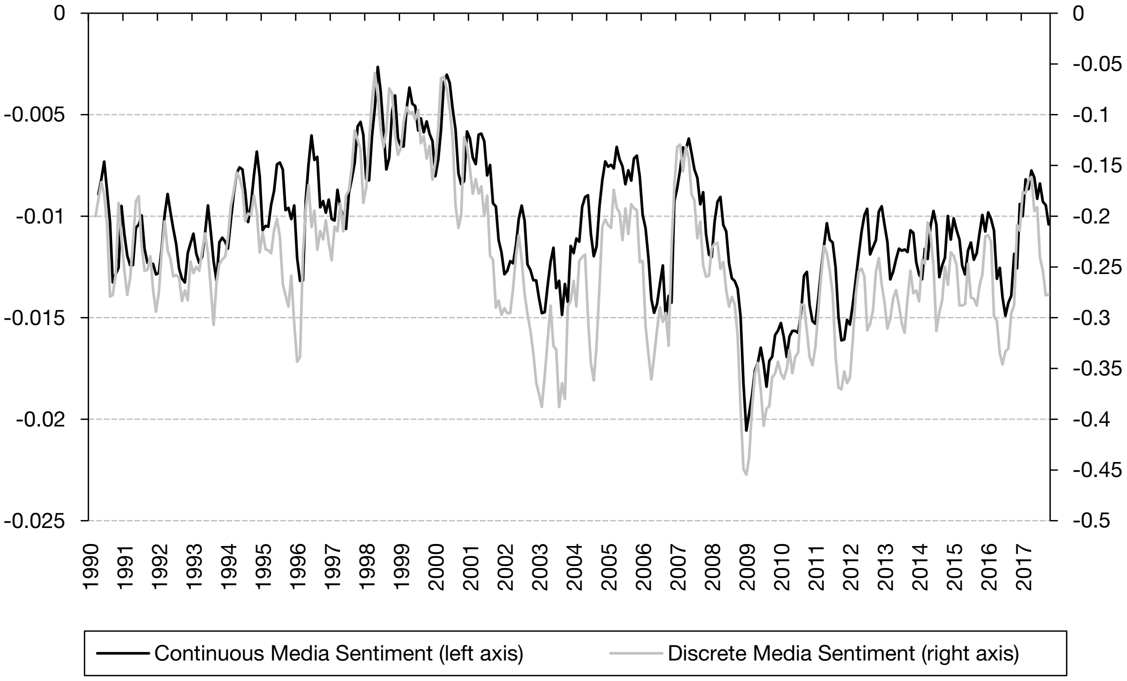
Over the whole period, the Media Sentiment indicator also appears to follow the main growth trends closely (Figure II), even if it does not fit very well the quarterly ups and downs, especially over the recent period. However, this does not disqualify it, as the sudden quarterly variations in GDP may be due to specific phenomena that an economic indicator does not always capture. Nonetheless we observe two significant divergences between our indicator and business activity. First, the indicator diverged abruptly in 2006, while business activity experienced no particularly noticeable deviation in that year (apart from a weak third quarter). Second, at the end of the crisis, the indicator only recovers gradually after having reached a low point in 2008-2009, although business activity rebounded vigorously over the same period. This created a divergence between the two series, which only disappears in 2011, when business activity slumped again following the Eurozone sovereign debt crisis.

In addition, our indicator is obviously quite similar to the Business Climate published by Insee (Figure III). However, we note that, while the two series follow identical major trends, the Insee Business Climate reveals short cycles lasting one or two years (particularly visible at the beginning of the period), absent from the Media Sentiment indicator. In the same way, the divergences already observed by comparing our indicator with business activity (in 2006 and post-crisis) are also visible here.

Finally, an overall similarity can be observed between our Media Sentiment indicator and (the opposite of) the "Economic Policy Uncertainty" (EPU) indicator described by Baker *et al.* (Figure IV).[7] Once again, two significant exceptions can be noticed. First, the media sentiment

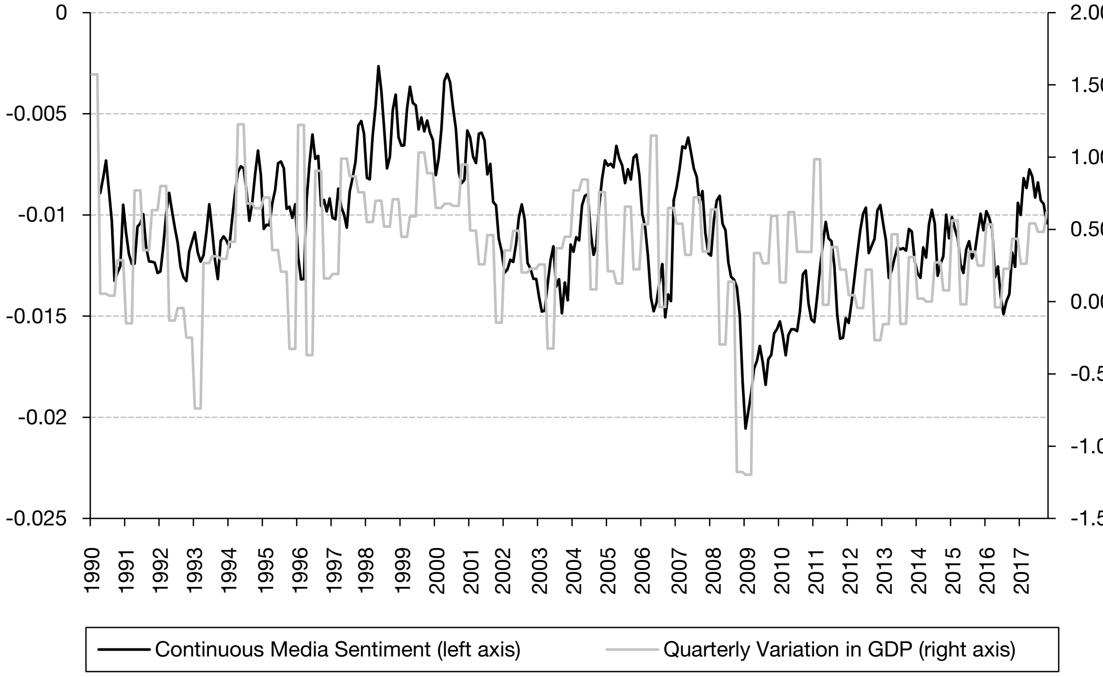
[7] *As the EPU indicator is an index of uncertainty, we have reversed the scale for the latter in order to compare it with our media sentiment, so as to make the graph easier to read (increasing uncertainty is actually consistent with decreasing sentiment)*

Figure I
Discrete and Continuous Media Sentiment Indicators – 3-Month Moving Average



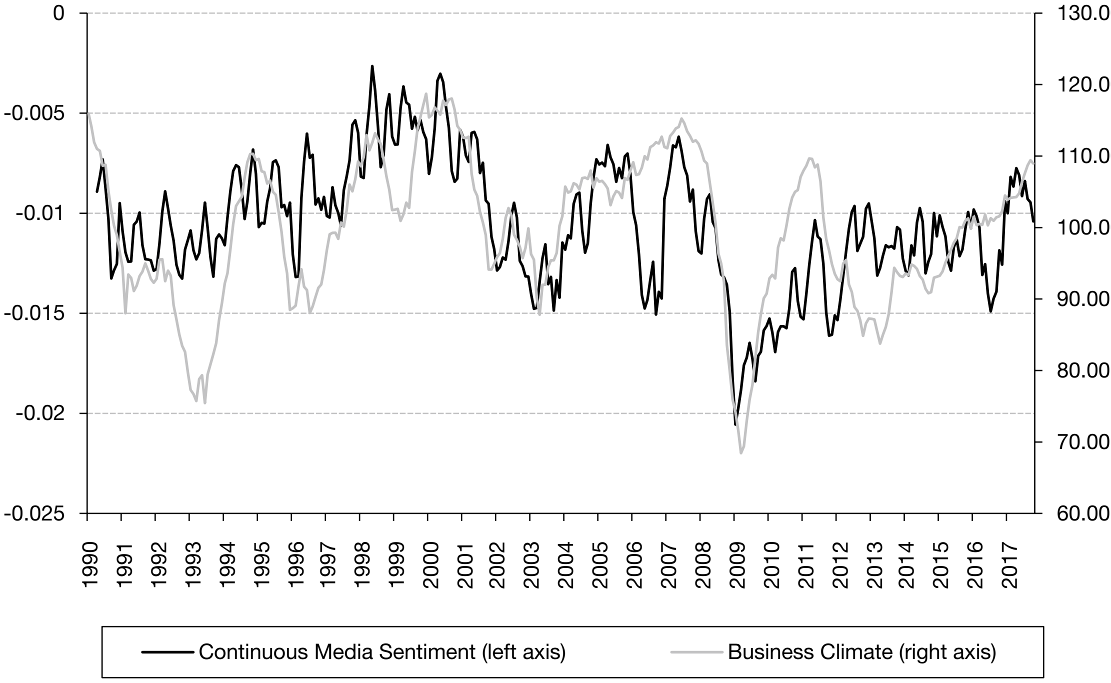
Note: This graph illustrates the change in the media sentiment indicator (3-month moving average), calculated on the basis of continuous coding and discrete coding.
Source: *Le Monde* authors' database.

Figure II
Continuous Media Sentiment Indicator and Quarterly Variation in French GDP



Notes: This graph illustrates the change in the media sentiment indicator (3-month moving average) and quarterly variation in French GDP.
Sources: *Le Monde* authors' database; Insee.

Figure III
Continuous Media Sentiment Indicator and Insee Business Climate



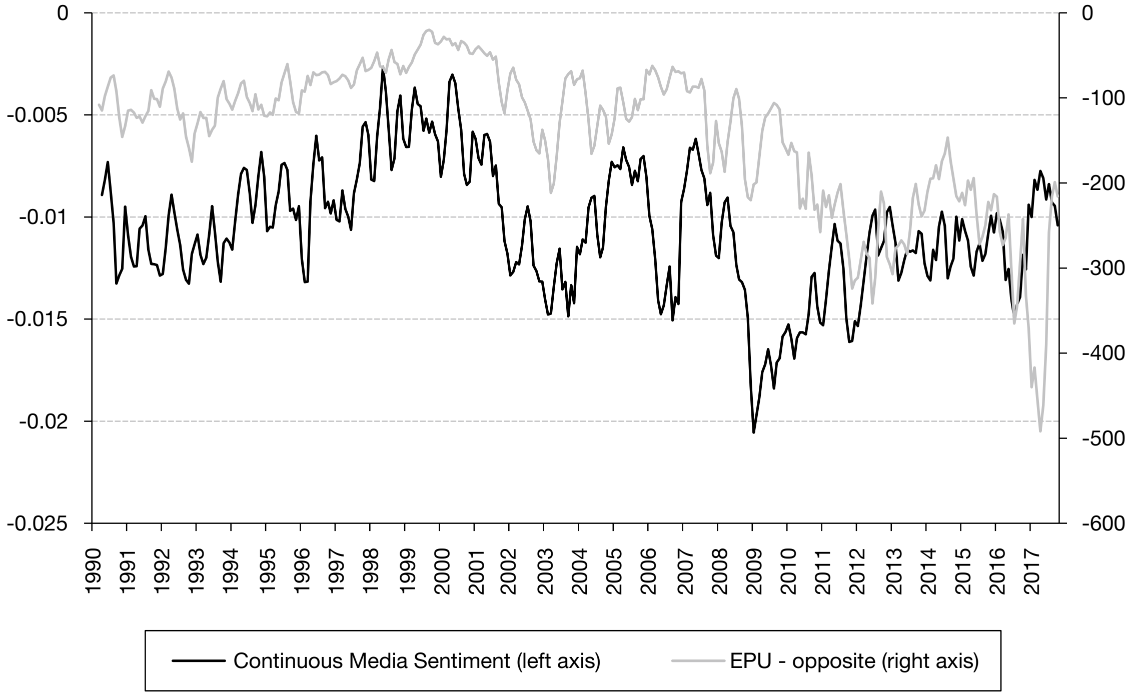
Notes: This graph illustrates the change in the media sentiment indicator (3-month moving average) and the Insee Business Climate indicator.
Sources: *Le Monde* authors' database; Insee.

indicator diverges more quickly and more significantly than the EPU of Baker *et al.* at the time of the 2009 financial crisis. Conversely, the latter shows a significant rise in uncertainty during 2016-2017, certainly due to the elections in France and rising influence of the *Front National* (perhaps with a Brexit effect), while our media sentiment indicator is fairly stable.

In both cases, our media sentiment indicator experiences changes more similar to economic activity than the EPU of Baker *et al.*: thus, we can expect *ex-ante* that the EPU is less effective than ours for forecasting.

Our graphical observations are confirmed by a simple analysis of correlations of the different series considered. The Insee Business Climate indicator is slightly more correlated to GDP growth than our media sentiment indicator, which may be a sign of better forecasting performance. Moreover, the Insee climate and sentiment indicator are fairly well correlated with each other. Finally, correlations of the EPU of Baker *et al.* with the other variables (and in particular with GDP growth) are weaker, which confirms our suggestion of lesser predictive capability (Table 1). Nonetheless, we can see that it is slightly better correlated to

Figure IV
Media Sentiment Indicator and (opposite) Economic Policy Uncertainty Indicator of Baker *et al.* for France



Note: This graph illustrates the change in the media sentiment indicator (3-month moving average) and the Economic Policy Uncertainty of Baker *et al.* (3-month moving average, opposite).
Sources: *Le Monde* authors' database; Baker *et al.* (2016).

Table 1
Correlations Between GDP Growth, Media Sentiment Indicator, Insee Business Climate and the EPU of Baker *et al.* (Opposite)

	Media Sentiment	Insee Business Climate	EPU (opposite)
GDP growth	0.469	0.547	0.268
Media Sentiment	-	0.575	0.389
Insee Business Climate	-	-	0.253

Note: The figure at the intercept of row *i* and column *j* corresponds to the correlation between the variable displayed in row i and that displayed in column j. For parsimony each correlation is shown only once.
Sources: *Le Monde* authors' database; Insee; Baker *et al.* (2016).

our media sentiment than to the other two variables, which seems to show a certain specificity of the media information. The statistics describing the different indicators are presented in an appendix.

Using Media Sentiment Indicators in Forecasting

The continuous monthly media sentiment indicator is used to forecast GDP growth for the current quarter. Several techniques can theoretically be considered to handle the difference in frequency between the variable to be forecast (quarterly) and the explanatory variables (monthly). A first possibility would be to use the MIDAS method (see, among others, the work of Ghysels *et al.*, 2005; 2007) that is designed to forecast a low frequency variable using high frequency explanatory variables. For this paper, we rather opted for an approach similar to "blocking", commonly used by forecasters (e.g. see Bec & Mogliani, 2015), which consists in using a different forecasting model (or "calibration") for each month of the quarter, each time using all the information available at the date considered. Thus, the "month 1", "month 2" and "month 3" calibrations use, respectively, all the information available at the end of the first, second and third months of the quarter. In practice, for example for the Business Climate (for which we consider the first difference) we will label $Climate_t$ the regressor that will correspond, in forecast "month 1", to the variation between the value of the business climate for the 1st month relative to the mean of the values taken for the three months of the previous quarter. In "month 2", we will consider the mean value for the two months of the current quarter relative to the value of the previous quarter. In "month 3", we then have all the information. The same logic is adopted for the variable $MediaSent_t$, except the fact that it is taken as level and not as first difference.[8] The first lag of GDP growth is also used as explanatory variable, when it is available (which is not the case, for example, in month 1).[9] However, we do not use the EPU indicator of BBD as explanatory variable: actually, our first graphic and correlation analyses were confirmed by the fact that this indicator does not improve the predictive performance of our models.

Since one of the aims of the article is to compare the respective performance of Insee Business Climate and the "Media Sentiment indicator", four models are considered for each month in the quarter: the first only uses the past variation in GDP (simple AR with the first lag of GDP growth when it is available, otherwise the second), the second includes the first lag of GDP growth and the Media Sentiment indicator, the third the first lag of GDP growth and the Business Climate, finally the fourth includes both the first lag of GDP growth, the Media Sentiment indicator and the Business Climate in France. The forecasting performance of these models are measured in real-time conditions. The models are estimated from the first quarter 1990 and up to a sliding date from the second quarter 2000 to the third quarter 2017, which supplies a list of forecasting errors from which we can calculate an RMSFE for each model.

To nowcast the current quarter, the models can be formalised as follows (to forecast the next quarter, only the index of the dependent variable changes).

$$\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \varepsilon_t$$

$$\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \beta_2 \cdot \Delta Climate_t + \varepsilon_t$$

$$\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \beta_2 \cdot MediaSent_t + \varepsilon_t$$

$$\begin{aligned}\Delta GDP_t = \alpha + \beta_1 \cdot \Delta GDP_{t-1} + \beta_2 \cdot \Delta Climate_t \\ + \beta_3 \cdot MediaSent_t + \varepsilon_t\end{aligned}$$

We present the estimates in full sample for equations 1 to 4 in Appendix 2. The media sentiment variable is significant at the 1% threshold in all models.

Using Penalised Regression for Forecasting

Constructing a media sentiment indicator provides a simple and readable tool comparable with more traditional economic indicators such as the business climate. However, it also has disadvantages. Firstly, it depends largely on the researcher's preconceptions: on the one hand, the terms in the sentiment dictionary are classified by experts and therefore based on presuppositions, on the other hand choices

[8] *This choice provides the best fit of the data in-sample and offers the best forecasting performance out-of-sample.*

[9] *Longer time lags of the growth of GDP were rarely significant in samples and did not substantially improve the performance of forecasting models. In general, adding them only modified the models at the margin: in the end, we therefore chose not to include them and to keep the models lean.*

have to be made about scoring the articles and aggregating the scores, for which there is no "natural" method. In addition, calculating a simple summary indicator does not allow full use of the richness of the database and therefore creates the risk of ignoring part of the information that could turn out to be useful in forecasting.

Thus, we offer a second forecasting method, leaving less space for the researcher's preconceptions and making better use of the diverse information contained in the database. The regressors used in this approach are the weightings of each term of the vocabulary (i.e. all the terms used at least once in the corpus of articles): however, we exclude the so-called "stopwords", i.e. words very often used (determinants, certain adverbs) and therefore in principle not discriminating. Similarly, we have also eliminated the commonest terms (present in more than 90% of documents) and the rarest terms (less than 5% of the time). Furthermore, as previously, the terms are stemmed and the combinations of two consecutive terms, or bigrams, are also considered to take better account of expressions such as "labour market".

We calculate the weightings associated with each term of the vocabulary using the tf-idf approach (term frequency-inverse document frequency) used extensively in the literature on information retrieval (e.g. see Breitinger *et al.*, 2015).[10] This weighting has proven more relevant than the frequency of terms when the documents handled (here, articles) are long. Using the frequency of the word in the document and the inverse of the frequency of documents containing this word, it is possible to make better use of a frequent word within an article if it is little used elsewhere. The weightings for each word from each article of the corpus can then be averaged by month or quarter, so that regressors are available at the same frequency as the dependent variable.

Once these variables have been obtained, we can apply the usual transformations to them: thus, we also retain their first lag, their growth rate and moving average over two quarters. In total, we obtain approximately 6,000 potential regressors. As this is a very large number, even greater than the number of points in the series to be forecast, it is necessary to select a sub-set of regressors. Actually, it is better for the forecast to focus on parsimonious models, i.e. that only use a limited number of variables. This is necessary to avoid overlearning phenomena: selecting too many explanatory variables generally degrades the predictive performance of the model outside the estimation sample. To do so, we use one of the most commonly-used techniques for automatic variable selection: penalised regression.

Penalised regression is a simple linear regression, to which we add a constraint (or penalty) regarding the amplitude of the coefficients associated with each regressor. This amplitude can be measured using different norms: we talk about Lasso regression when the amplitude is measured using norm L1 (sum of absolute values of coefficients) and Ridge regression when norm L2 (Euclidean) is used. As the Lasso penalty has the property of being quite abrupt and often leads to models that are too parsimonious, we use a combination of the Lasso penalty and the Ridge penalty: this is referred to as Elastic-Net regression.

Penalised regressions offer greater robustness than iterative techniques such as *stepwise*, and they have the advantage of being configurable, the hyper-parameters corresponding to the size of the penalty. By seeking parameters optimising forecasting performance, we can favour the selection of regressors with better predictive power. More precisely, hyper-parameters are optimised by "grid search": for different values of the parameters, we use a sliding window and produce a listing of forecasting deviations, from which we calculate an RMSFE. We then select hyper-parameters minimising the RMSFE.[11]

Results

In this section, we present the results using the Media Sentiment indicator computed from our dictionary with a continuous coding as well as those supplied by the automatic penalised regression method.

We present the RMSFEs of the different models depending on the month of the quarter at which

[10] *In information retrieval, tf-idf weighting is used to represent documents (e.g. web pages) in the form of numerical vectors that can then be compared with the numerical vector corresponding to a query; it is then possible to put documents in order based on their relevance to the query (e.g. a query from a user in the search engine).*

[11] *So as not to bias the results towards this approach, the sliding window used is not the same as that from which RMSFEs are produced from the different methods compared in this study. The RMSFEs are therefore produced over the period from the 1st quarter of 1998 to the last quarter of 1999.*

Table 2
RMSFE of Models for Forecasting GDP Growth Rate in Quarter Q for Different Forecast Time Scales

	Forecast month	Month 1 (Q-1)	Month 2 (Q-1)	Month 3 (Q-1)	Month 1 (Q)	Month 2 (Q)	Month 3 (Q)
	Month before publication	6	5	4	3	2	1
[1]	AR(1)	0.4057	0.3941	0.3941	0.3927	0.4039	0.4039
[2]	AR(1) + Sentiment	0.3968	0.3951	0.3931	0.3798	0.3727	0.373
[3]	AR(1) + Elastic-Net	0.3781	0.3955	0.3904	0.3793	0.3672	0.3820*
[4]	AR(1) + Climate	0.3434*	0.3475*	0.3459*	0.3406*	0.3689	0.3712
[5]	AR(1) + Elastic-Net + Climate	0.3642	0.3879	0.3835	0.3755	0.3552	0.3749
[6]	AR(1) + Sentiment + Climate	**0.3357**	**0.3446**	**0.3403**	**0.3281**	**0.3331***	**0.3326***

Note: This table presents the RMSFEs from models [1] to [6]. For each time scale (each column), the lowest RMSFE is shown in bold. For each month of the quarter and each model, the asterisk * indicate that, according to the Harvey *et al.* (1997) test, the Root Mean Square Forecast Error (RMSFE) of the model is significantly less than for the benchmark model (at the threshold of 10%). Models [2], [3] and [4] are compared to model [1]. Models [5] and [6] are compared to model [4]. For example at month 2 in Q, the RMSFE of model [6] (AR(1) + Sentiment + Climate) is significantly lower than that of model [4] (AR(1) + Climate).
Sources: *Le Monde* authors' database; Insee; authors' calculation.

the forecast is made (Table 2). We test the assumption that the model combining Media Sentiment and Business Climate provides a significantly better forecast than the other models using the Harvey *et al.* (1997) test.

Individually, model [2] (AR + sentiment) provides slightly better accuracy than model [1] (simple AR) for the current quarter (*nowcasting*), but this improvement is not significant. Model [4] (with climate) has superior properties. Nonetheless, when we combine climate and sentiment, the predictive performance of the model is superior (model [6]) to that for climate use alone (model [4]). This is particularly sensitive with effect from month 2 of the current quarter. For all time scales, the forecast from model [6] is more accurate than the other models. The Harvey *et al.* (1997) test shows us that this difference is significant for months 2 and 3 of the current quarter at a 10% threshold.

This result tends to show that individually, the Insee Business Climate remains a more reliable economic indicator than our Media Sentiment. Nonetheless, the Media Sentiment contains information in addition to that contained in the business climate, improving the forecast of French GDP.

Model [3] (penalised regression) also demonstrates superior performance compared to the autoregressive model [1] for some time scales. However, when we add business climate, a variable already having great predictive power, the disaggregated approach [5] does not give better performance than the simple autoregressive model augmented by the Insee Business Climate [4]. It should be stressed that despite its robustness when using large-scale data, this approach doubtless suffers here from the very small number of observations in comparison (one hundred for 60 times more variables). However, this disaggregated approach remains interesting, in the sense that it is easier to implement, automatically calibrated, and not involving compiling lists of terms, which is both laborious and debateable.

* *
*

We have therefore shown that media information was a promising tool for economic analysis. The systematic treatment of articles published by *Le Monde* since 1990 using textual analysis techniques enabled us to measure this potential for forecasting or nowcasting French GDP. More precisely, we considered two different strategies: the first consisted in constructing a synthetic indicator, the second in using more extensively all the information available in the database. These two approaches each have their advantages and drawbacks. The first offers the possibility to construct a readable media sentiment indicator with theoretical properties similar to other more traditional economic tools (business climate). However, such an indicator takes into account only a fraction of the information contained in the database and, in addition, its construction is based on a certain number of choices and questionable bias. Conversely, a variables selection technique (penalised regression) has the advantage of using all the information from the database in an exhaustive and "agnostic" way: it is easy to implement and does not rely on any preconception. However, it provides inferior results to the approach using a predefined sentiment dictionary.

Nonetheless, this generally favourable observation should be somewhat tempered. At all time scales, the Insee Business Climate indicator appears to be a more effective tool than media information. Similarly, adding media information does not always enable a significant gain in predictive power: it therefore currently appears to play a greater role as complement than substitute. Finally, it should be recalled that economic institutes have to continue to develop their activity producing indicators: media sentiment indicators would not be able to replace them since economists and public authorities need an independent and controlled source to measure the business climate. □

BIBLIOGRAPHY

Andreou, E., Ghysels, E. & Kourtellos, A. (2013). Should Macroeconomic Forecasters Use Daily Financial Data and How? *Journal of Business & Economic Statistics*, 31(2), 240–251.
https://doi.org/10.1080/07350015.2013.767199

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L. & Rünstler, G. (2011). Short-term forecasts of euro area GDP growth. *The Econometrics Journal*, 14(1), C25–C44.
https://doi.org/10.1111/j.1368-423X.2010.00328.x

Baffigi, A., Golinelli, R., & Parigi, G. (2004). Bridge models to forecast the euro area GDP. International Journal of forecasting, 20 (3), 447–460.
https://doi.org/10.1016/S0169-2070(03)00067-0

Baker, S. R., Bloom, N. & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593–1636.
https://doi.org/10.1093/qje/qjw024

Bańbura, M., Giannone, D., Modugno, M. & Reichlin, L. (2013). Now-Casting and the Real-Time Data Flow, *Handbook of Economic Forecasting*, vol. 2 (Part A), 195–237.
https://doi.org/10.1016/B978-0-444-53683-9.00004-9

Bec, F. & Mogliani, M. (2015). Nowcasting French GDP in real-time with surveys and "blocked" regressions: Combining forecasts or pooling information? *International Journal of forecasting*, 31 (4), 1021–1042.
https://doi.org/10.1016/j.ijforecast.2014.11.006

Bortoli, C. & Combes, S. (2015). Apports de Google trends pour prévoir la conjoncture française: des pistes limitées. Insee, *Note de conjoncture*, mars 2015.
https://www.insee.fr/fr/statistiques/1408926?sommaire=1408931

Bortoli, C., Combes, S. & Renault, T. (2017). Comment prévoir l'emploi en lisant le journal. Insee, *Note de conjoncture*, mars 2015.
https://www.insee.fr/fr/statistiques/2662520?sommaire=2662600

Breitinger, C., Gipp, B. & Langer, S. (2015). Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4), 305–338.
https://doi.org/10.1007/s00799-015-0156-0

Choi, H. & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, 88 (1), 2–9.
https://doi.org/10.1111/j.1475-4932.2012.00809.x

Darné, O. (2008). Using business survey in industrial and services sector to nowcast GDP growth: The French case. *Economics Bulletin,* 3(32), 1–8.
https://ideas.repec.org/a/ebl/ecbull/eb-08c50137.html

D'Amuri, F. & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816.
https://doi.org/10.1016/j.ijforecast.2017.03.004

Fondeur, Y. & Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30, 117–125.
https://doi.org/10.1016/j.econmod.2012.07.017

Foroni, C. & Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting Euro

area macroeconomic aggregates. International Journal of Forecasting 30(3), 554–568.
https://doi.org/10.1016/j.ijforecast.2013.01.010

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267–1300.
https://doi.org/10.1111/jofi.12027

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2005). There is a risk-return trade-off after all. *Journal of Financial Economics*, 76(3), 509–548.
https://doi.org/10.1016/j.jfineco.2004.03.008

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, 26(1), 53-90.
http://dx.doi.org/10.2139/ssrn.885683

Harvey, D., Leybourne, S. & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2), 281–291.
https://doi.org/10.1016/S0169-2070(96)00719-4

Kotsiantis, S. B., Pintelas, P. E. & Zaharakis, I. D. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3), 159–190.
https://doi.org/10.1007/s10462-007-9052-3

Larsen, V. H. & Thorsrud, L. A. (2015). The value of news. BI Norwegian Business School, *Working Papers* N° 6/2015.
ttps://ideas.repec.org/p/bny/wpaper/0034.html

Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66 (1), 35–65.
https://doi.org/10.1111/j.1540-6261.2010.01625.x

McLaren, N. & Shanbhogue, R. (2011). Using Internet search data as economic indicators. *Bank of England Quarterly Bulletin* N° 2011-Q2.
http://dx.doi.org/10.2139/ssrn.1865276

Mogliani, M., Darné, O. & Puyaud, B. (2017). The new MIBA model: Real-time nowcasting of French GDP using the Banque de France's monthly business survey. *Economic Modelling*, 64, 26–39.
https://doi.org/10.1016/j.econmod.2017.03.003

Mogliani, M. & Ferrière, T. (2016). Rationality of announcements, business cycle asymmetry, and predictability of revisions. The case of french GDP. *Banque de France, Working Papers Series* N° 600.
https://publications.banque-france.fr/en/economic-and-financial-publications-working-papers/rationality-announcements-business-cycle-asymmetry-and-predictability-revisions-case-french-gdp

Porter, M. F. (2001). Snowball: A language for stemming algorithms.
http://snowball.tartarus.org/texts/introduction.html

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.
https://doi.org/10.1111/j.1540-6261.2007.01232.x

APPENDIX 1

DESCRIPTIVE STATISTICS

Table A1

	Frequency	Average	Median	Min	Max	Standard deviation	Skewness	Kurtosis
GDP growth	Quarterly	0.3383	0.3456	-1.1967	1.2270	0.4218	2.0606	-0.7953
Media Sentiment	Monthly	-0.0105	-0.0104	-0.0228	-0.0011	0.0037	0.1955	-0.2251
Insee Business Climate	Monthly	99.47	100.35	68.43	118.71	10.13	-0.0877	-0.4747

Sources: *Le Monde* authors' database; Insee.

APPENDIX 2

COEFFICIENTS OF ECONOMETRIC MODELS

Table A2-1

	Month 1 (Q)	Month 1 (Q)	Month 1 (Q)	Month 1 (Q)
α	0.2537***	0.7207***	0.2514***	0.6228***
ΔGDP_{T-2}	0.2700***	0.1456	0.2942***	0.1935**
ΔGDP_{T-1}				
$\Delta Climate_T$			0.0605***	0.0560***
$\Delta MediaSent_T$		40.4608***		32.1605***
Adjusted R^2	0.070	0.145	0.258	0.303

Table A2-2

	Month 2 (Q)	Month 2 (Q)	Month 2 (Q)	Month 2 (Q)
α	0.2642***	0.8672***	0.2980***	0.8402***
ΔGDP_{T-2}				
ΔGDP_{T-1}	0.2430*	0.0908	0.1593	0.0283
$\Delta Climate_T$			0.0467***	0.0431***
$\Delta MediaSent_T$		51.95***		46.9353***
Adjusted R^2	0.055	0.169	0.196	0.288

Table A2-3

	Month 3 (Q)	Month 3 (Q)	Month 3 (Q)	Month 3 (Q)
α	0.2761***	1.0301***	0.3118***	0.9987***
ΔGDP_{T-2}				
ΔGDP_{T-1}	0.2139*	0.0036	0.1190	- 0.0645
$\Delta Climate_T$			0.0423***	0.0384***
$\Delta MediaSent_T$		64.4305***		58.9808***
Adjusted R^2	0.037	0.206	0.190	0.331

Note: The table shows the results from the equation $\Delta GDP_T = \alpha + \beta_1 * \Delta GDP_{T-1} + \beta_2 * \Delta Climat_T + \beta_3 * MediaSent_T + \varepsilon_t$ (ΔGDP_{T-2} at month 1, as the GDP for the next quarter has not been published yet) over the whole sample (1990-Q1 to 2017-Q4). ***, **, * indicate significance of the coefficients at 1%, 5% and 10%, respectively. The standard deviations are robust to heteroscedasticity.
Sources: *Le Monde* authors' database; Insee; authors' calculation.