How to predict economic growth thanks to press articles using NLP

Keywords: nowcasting, machine learning, NLP

1. INTRODUCTION

Most of the economic indicators are constructed on a monthly or a quarterly basis. They are usually available only at the end of the month or of the quarter. Nevertheless, knowing the evolution of activity even earlier can be crucial. This was especially essential during the Covid-19 health crisis, which caused sudden and large-scale economic movements. Press articles constitute an important source of knowledge to analyse the economic activity. They daily convey information on events which may have important impacts on economic activity, such as governmental measures or exceptional events such as strikes or natural disasters, etc.

The present work addresses the construction of NLP-based leading indicators of GDP, which uses daily press article tests. This work extends S. Combes & al. (2018) who calculated a media indicator from a count of positive or negative words on articles in the daily journal *Le Monde*. It goes further by extending this approach to another journal *Les Echos*, which is specialized in economics and by testing other techniques such as machine learning, NLP and sentiment analysis

2. METHODS

2.1 The database we used

The press articles we used come from two major French daily newspapers: *Le Monde* and *Les Echos*). *Le Monde* is a widespread French newspaper whereas *Les Echos* is specialised on economic facts. The text of the articles were cleaned in a quite usual way: removal of stop words, of punctuation marks, stemmatisation (we also tried lemmatisation) of words.

Our first goal was to replicate the media indicator of S. Combes & al.(2018). Indeed the authors proposed a media indicator based on the counting of positive and negative words in the articles such as defined in a dictionary build up for this purpose. However, a term may be positive in an economic context but may have a completely different meaning elsewhere. Hence we needed to select articles dealing with economic topic.

Some articles in the two daily newspapers are already labelled in categories such as Economics, Sports, Art, etc. but not all. A machine learning model (a Bayesian model and a logistic regression) was trained on a sample of labelled articles. The model was then used to categorize all the database. Only the articles predicted to be in the Economics category (or close[1]) were used in the rest of the analysis. We also tried to build indicators based on articles dealing with France.

1 The Economics category of *Les Echos* is not so well defined. Several categories were grouped to select the articles dealing with economic topics.

2.2 Indicators

We constructed an indicator based on a difference between the number of positive and negative key words in articles. This indicator is computed the same way as in S. Combes & al. (2018). The only differences stays in the source we used (an economic newspaper instead of a generalist one) and in the economic period. This enables us to analyse the behaviour evolution of the indicator during the current economic crisis.

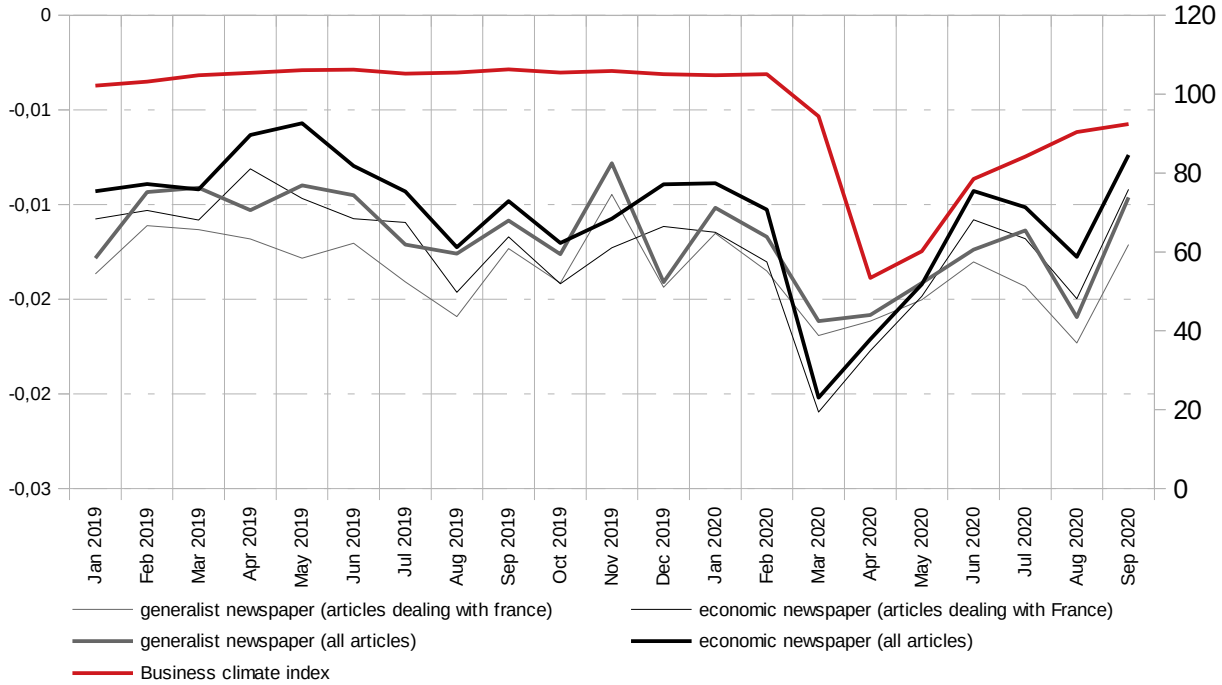
We also tried to use machine learning techniques. We tried to select the key words to use thanks to a random forest and to build an indicator by learning the link between these words and other leading economic indicators such as the business climate index.

3. RESULTS

The indicator based on *Les Echos*, the daily economic newspaper shows better results than the one calculated from the more generalist *Le Monde*. Its evolution seems more consistent with the business climate index and with the GDP. The decline during the health crisis of Covid-19 is more pronounced, which seems more consistent with GDP.

Evolution of indicators during the Covid-19 health crisis

For each newspaper (specialised in economy or not) we computed two set of indicators: one of them is based on a count of words in all articles whereas the other one is based on a count of terms in articles dealing with France only. All indicators seem to follow the evolution of the business climate index[2]. The restriction to articles dealing with France doesn't seem to have any effect. However indicators based on the economic newspaper are more accurate than the ones computed thanks to the more generalist and more broadly sold one.



2 We chose this index because it is a monthly index that is highly correlated with GDP. Unlike all the other series, the scale of this index is shown on the right.

Dictionary-based Indicators that count the difference between positive and negative terms in articles seem quite accurate. During the Covid-19 health crisis they declined sharply to recover a few month later.

These results were confirmed by basic time series models. We tried to use the indicators to predict GDP. The indicator based on the difference between positive and negative terms in articles of Les Echos, which is specialised in economic topics is the most performant. It improves the quality of the forecast compared to the indicator based on the articles of the more generalist newspaper, *Le Monde.*

Machine learning techniques are tricky to implement on our data. The number of articles is huge. We have several millions of them at our disposal but the number of points to predict is quite small. We only have monthly time series starting in the nineties. However the terms selected by this kind of techniques seem relevant to predict economic conditions. They could be used to build a model or to complete our dictionary.

4. CONCLUSIONS

The indicator based on the computation of terms in economic articles seems to lead to better results. It is useful to predict variations of GDP. Furthermore it is easy to understand. Its computation is quite straightforward. It is robust. The addition of new data does not change the past of the series. The dictionary it uses is static but it accurately predicted the evolution of GDP during the Covid-19 health crisis.

Machine learning technique could be able to select the suited terms to predict the GDP and to adapt to any abrupt change such as the Covid-19 health crisis. This is what our tests suggest. It might be possible to use the words selected by a machine learning technique into the dictionary used to compute positive or negative words in articles.

REFERENCES

[1] C. Bortoli, S. Combes and T. Renault, Comment prévoir l'emploi en lisant le journal, Note de conjoncture, French National Statistics Institute

[2] C. Bortoli, S. Combes and T. Renault, 2018 Nowcasting GDP Growth by reading Newspapers, Economie et statistique, 505-506,pp. 17-33

[3] F. Laurent, S. Simoni, When are Google data useful to nowcast GDP ? An approach via pre-selection and shrinkage, working paper series no. 717, Central bank of France,

[4] Shapiro, Adam Hale, Moritz Sudhof, and Daniel Wilson. 2017. "Measuring News Sentiment," Federal Reserve Bank of San Francisco Working Paper 2017-01.

[5]Ksenia Yakovleva, 2017. "**Text Mining-based Economic Activity Estimates**," Bank of Russia Working Paper Series wps25, Bank of Russia.

[6]A. Turrel, N. Anesti ans Silvia Miranda-Agrippino, 2019, *What's in the news ? Text-Based confidence indices and growth forecasts, bank of England*

[7]A. Hale Shapiro, M. Sudhof et D. Wilson, 2018, *Measuring News Sentiment,* Working Paper, Federal Bank of San Francisco

[8]L. Anders Thorsrud, ***2016***, *Words are the New Numbers: A Newsy Coincident Index of the Business Cycle,*

[9] M. E. Doms et N. J. Morin , 2004, *Consumer Sentiment, the Economy, and the News Media, Journal of Business & Economic Statistics*, *Paper No. 2004-09, FRB of San Francisco Working*

[10] D. Antenucci, M. Cafarella, M. Levenstein, C. Ré, M. D. Shapiro, ***2014***, *Using Social Media to Measure Labor Market Flows,Working Paper No. 20010, National Bureau of Economic Research*

[11]Ho, Tin Kam (1995), Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995.

[12]Breiman L (2001), "Random Forests". Machine Learning. 45(1): 5–32

[13]Robert Tibshirani, « Regression shrinkage and selection via the lasso », Journal of the Royal Statistical Society. Series B, vol. 58, no 1, 1996, p. 267-288