
PRÉDIRE L'ACTIVITÉ ÉCONOMIQUE A PARTIR D'ARTICLES DE PRESSE

Guillaume ARION (*), Stéphanie HIMPENS (**), Théo ROUDIL-VALENTIN (***)

(*) Insee, Direction des études et synthèses économiques

(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

(***) ENSAE

Stephanie.himpens@banque-france.fr

Mots-clés: NLP analyse textuelle, Analyse de sentiment, nowcasting, séries temporelles

Domaine concerné : Modélisation, séries temporelles

Résumé

Les nouvelles techniques d'analyse textuelle permettent d'exploiter de nouvelles sources de données. Les articles de journaux contiennent de nombreuses informations sur l'activité économique. En 2017, Bortoli et al. avaient montré qu'il était possible de construire un indicateur synthétique de climat médiatique à partir d'articles de presse du quotidien généraliste *Le Monde* afin de prédire la conjoncture.

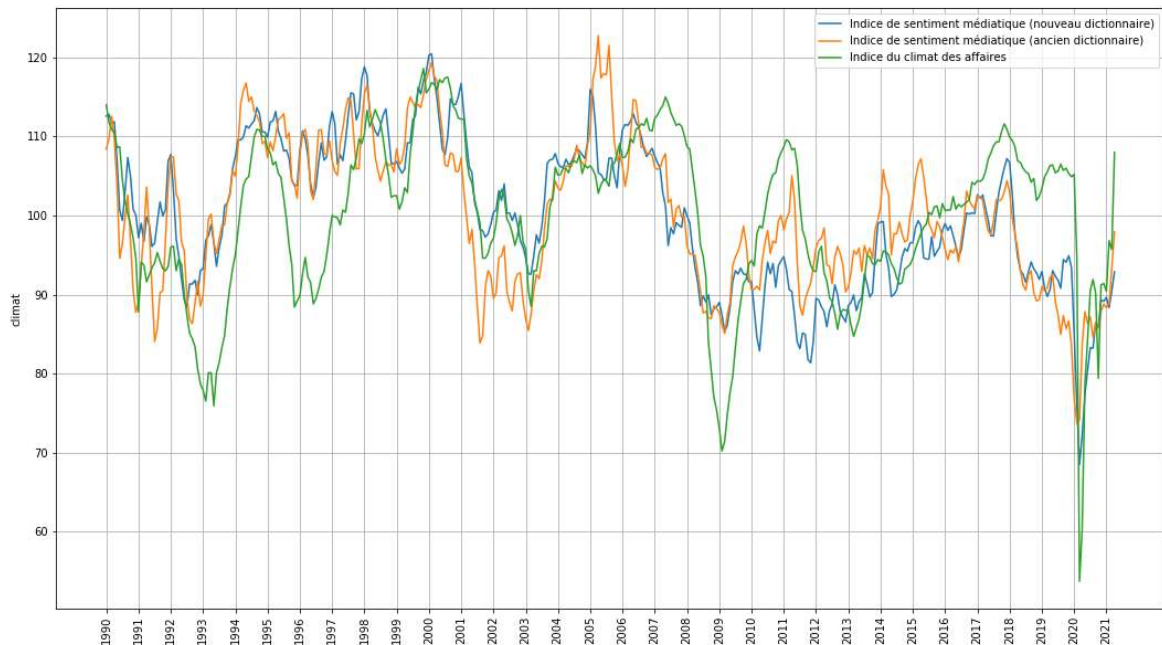
Cet article reprend, et étend la méthodologie de ces auteurs. De façon similaire, il construit un indicateur de sentiment médiatique à partir d'un comptage de mots positifs et négatifs d'un dictionnaire sur les articles économiques traitant de la France. L'indicateur est calculé pour chaque article selon la formule suivante :

$$\text{Sentiment}_i = \frac{Nb_{\text{positif}} - Nb_{\text{négatif}}}{Nb_{\text{article}}}$$

C'est-à-dire que l'indice élémentaire de l'article i est égal au nombre de termes positifs Nb_{positif} moins le nombre de termes négatifs $Nb_{\text{négatif}}$ divisé par le nombre de termes total de l'article Nb_{article} afin de tenir compte des différences de longueur des textes. Les indices peuvent ensuite être agrégés (moyenne) sur la période souhaitée.

Cependant, à la différence de Bortoli et al. (2017 et 2018), il étend cette stratégie aux articles du quotidien *Les Echos*, plus spécialisé en économie que le journal *Le Monde*. Il introduit également une nouvelle technique d'enrichissement du dictionnaire initialement construit par Bortoli et al. (2017). Enfin, des régressions pénalisées sont introduites afin de prévoir directement l'activité économique à partir des séries de nombre de mots contenus dans les articles.

Des premiers résultats de ce travail ont été présentés dans la note de conjoncture de mars 2021. Ils montrent que l'indice obtenu à l'aide d'un comptage de mots anticipe bien les mouvements marqués de l'activité économique. En particulier, lors de la crise liée au coronavirus en 2020, le profil de l'indice mensuel semble cohérent avec les estimations mensuelles d'activité présentées dans les Points et Notes de conjoncture. L'introduction de l'indicateur de sentiment médiatique calculé dans des modèles de prévisions à court terme au côté du climat des affaires améliore leurs capacités prédictives.



Graphique 1 : Indicateurs de sentiment médiatique calculé avec le dictionnaire de Bortoli et al. (2017) et avec le nouveau dictionnaire (enrichi à l'aide de méthode de NLP) et indice du climat des affaires.

Note : Afin de les rendre comparable, tous les indicateurs ont été centré-réduits (moyenne de 100 et écart-type de 10).

Le caractère quotidien des sources utilisées permet également de tracer un indice journalier. Les principaux événements liés aux périodes de confinements apparaissent bien sur le graphique. Ainsi, durant cette période, l'indicateur a pu livrer une information très précoce et pertinente sur les évolutions économiques. Toutefois, l'indicateur calculé montre certaines imperfections. En particulier l'indice sous-estime la reprise économique en fin de période.

Enfin, un modèle de régression pénalisée (LASSO) a été testé afin de prédire directement le climat des affaires ou un PIB mensualisé (obtenu par lissage du PIB trimestriel) à partir d'un comptage des mots présents dans les articles. Il semble bien prédire la baisse brutale liée à la crise sanitaire du Covid-19 mais sous-estime la reprise économique. Ses prédictions semblent également assez instables.

Bibliographie

- [1] Bortoli C., Combes S., Renault T., « Prévoir l'emploi en lisant le journal », *Notes de conjoncture*, mars 2017.
- [2] Bortoli C., Combes S., Renault T., « Prévoir la croissance du PIB », *économie et statistique*, 2018.
- [3] Loughran T., McDonald B., « When is a liability not a liability ?textual analysis, dictionaries, and 10-ks», *The journal of Finance*, 2011.
- [4] Shapiro A. H., Sudhof M., Wilson D., « Measuring news sentiment », *Federal Reserve Bank of San Francisco*, 2020
- [5] Arion G.,Himpens S., Roudil-Valentin T., « L'activité économique française au travers d'articles de presse », *Notes de conjoncture*, mars 2021.