Journées de méthodologie statistique de l'Insee 2022

# *PREDIRE L'ACTIVITE ECONOMIQUE 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** : 9 et 10

## 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.

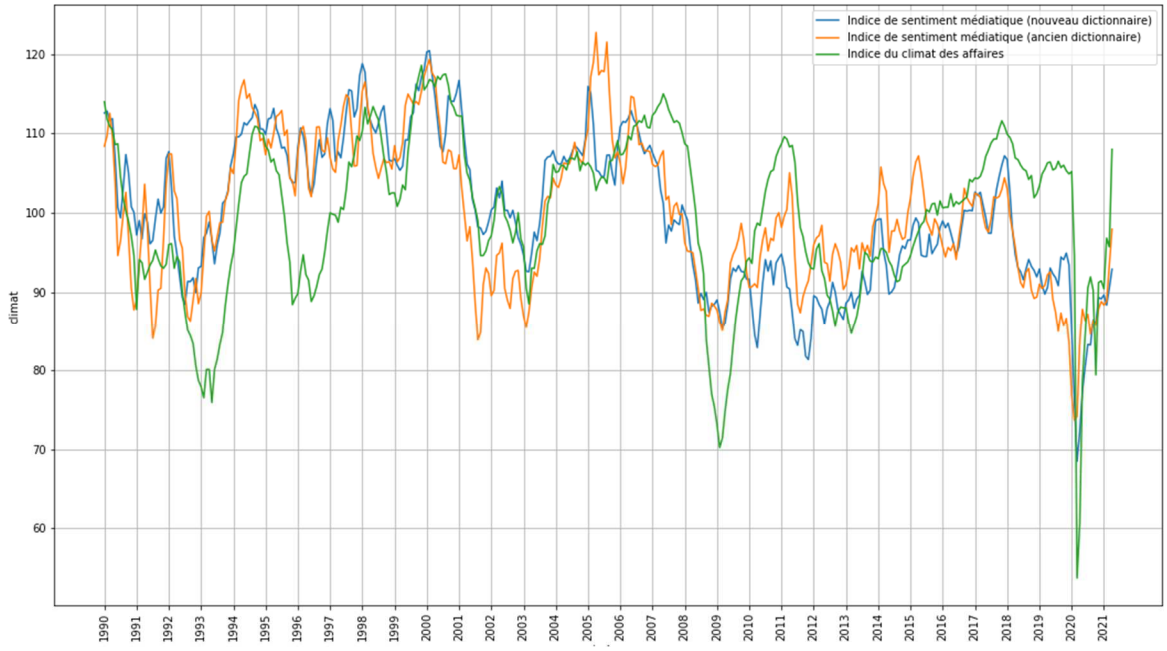
Cet article reprend, et étend la méthodologie de ces auteurs. De façon similaire, il construit un indicateur de sentiment médiatique à partir d'un comptage de mots positifs et négatifs d'un dictionnaire sur les articles économiques traitant de la France. L'indicateur est calculé pour chaque article selon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

$$Sentiment_i = \frac{Nb_{positif} - Nb_{négatif}}{Nb_{article}}$$

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 $Nb_{négatif}$ divisé par le nombre de termes total de l'articles $Nb_{article}$ 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 liées 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.

## Abstract

Natural language processing (NLP) techniques enables to process new sources of data. Newspaper articles contain a lot of information about economic activity. In 2017, Bortoli et al. had shown that it was possible to construct a synthetic indicator of business activity from newspaper articles of Le Monde, a generalist media source.

This article repeats and extends the methodology of these authors. In a similar way, it constructs an indicator of media sentiment from a count of positive and negative words of a dictionary in economic articles about France. It can be shown that the indicator is useful to predict the major events in the covid 19 crisis.

**Introduction**

Les exercices de prévisions réalisés trimestriellement à l'Insee, dans le cadre des *Notes de Conjoncture*, utilisent des indicateurs disponibles seulement à la fin du mois ou du trimestre. L'information produite par les médias présente l'avantage d'être réactive, fournie et d'aborder de nombreux domaines économiques. Les articles de presse sont disponibles deux à trois semaines avant une bonne partie des indicateurs conjoncturels usuels (enquêtes de conjoncture notamment) et traduisent le contexte économique courant. Grâce à l'émergence des techniques récentes d'analyse textuelle (*text mining*), de collecte automatisée des données en ligne (*web scraping*) et d'apprentissage supervisé (*machine learning*), des indicateurs de sentiment peuvent être élaborés afin de traduire cette richesse.

Cette approche d'analyse de sentiment médiatique s'est généralisée ces dernières années. Par exemple Shapiro et al. [2020] analysent le sentiment provenant d'un ensemble de médias américains et l'appliquent à la prévision d'indicateurs économiques. Loughran and McDonald [2011] notent que l'analyse lexicale de textes peut aider à comprendre comment les informations financières ont un impact sur les séries financières. Fraiberger [2016] utilise également l'approche de construction d'un indice de sentiment médiatique pour prévoir le PIB avec des résultats très prometteurs. D'autres, comme Tetlock [2007], Garcia [2013] s'attachent à prévoir les séries financières via les colonnes financières du Wall Street Journal et du New York Times, respectivement. Le premier trouve que son indice de sentiment permet de prévoir les retournements de tendance et leur retour à la valeur fondamentale. Le deuxième quant à lui met en exergue la prégnance de la relation en temps de récession. Ainsi, de nombreux travaux tentant à partir de journaux de prévoir des séries financières, et parfois même le PIB, existent. Deux publications récentes de l'Insee (Bortoli et al., 2017 et Bortoli et al., 2018) ont montré qu'il est possible d'utiliser l'information des articles en ligne du *Monde* pour améliorer la prévision du PIB français.

Ce travail reproduit l'indice de sentiment médiatique selon la méthodologie initiale de Bortoli et al. [2017 et 2018] et permet ainsi d'analyser son comportement lors de la crise sanitaire du covid-19. Alors que Bortoli et al. s'étaient exclusivement appuyés sur les articles du quotidien généraliste Le Monde, ce travail va plus loin et utilise les données d'un quotidien plus spécialisé en économie : Les Echos. Le dictionnaire de termes "positifs" et "négatifs" des auteurs a également été élargi via des techniques d'analyse textuelle et de *machine learning*.

Une première partie détaille la construction de la base d'articles. Une deuxième partie détaille la méthodologie de calcul de l'indice de sentiment médiatique. Une troisième partie montre les résultats obtenus. La dernière partie évoque une approche expérimentale cherchant à prédire l'activité économique directement à partir des séries de nombre de mots contenus dans les articles.

## 1. Les données

### 1.1. Obtention des données

Pour ce travail, nous avons récupéré la base du journal *Le Monde* scrapée par les auteurs de Bortoli et al. [2017 et 2018] que nous avons complétée avec des articles plus récents. Elle couvre maintenant la période de janvier 1990 à septembre 2020. Les données les plus récentes (à partir de janvier 2017) ont été obtenues en scrapant la page de recherche dans les archives du *Monde* (https://www.lemonde.fr/recherche).

La base des *Echos* est constituée de deux parties. Une première, jusqu'en août 2019, a été mise à disposition par le journal. La deuxième partie a été scrapée à partir du site du quotidien (https://www.lesechos.fr/) . La base finale s'étend de janvier 1994 à octobre 2020 et comporte plus d'un million d'articles.

### 1.2. Le nettoyage des articles

Les données textuelles des articles (titre, sous-titre et contenu) sont nettoyées : suppression des caractères spéciaux, des espaces (harmonisation de la casse), des retours à la ligne, des accents et de la ponctuation. Bortoli et al. [2017, 2018] avait réalisé une simple racinisation. Dans ce travail, une lemmatisation, en théorie plus précise, est réalisée à l'aide du package Spacy de python (utilisation du modèle de langue *fr_core_news_sm* entraîné sur Wikipédia). Au final, seuls les noms, les verbes et les adjectifs sont conservés.

### 1.3. Sélection des articles pertinents

De nombreux articles ne sont pas labellisés. Tout comme dans l'article de Bortoli et al. [2017,2018], nous entraînons un modèle de *machine learning* sur les articles catégorisés afin de déduire la catégorie des autres articles. Pour chacun des journaux, une base d'apprentissage est constituée pour moitié d'articles « économiques » c'est-à-dire classés dans une catégorie liée à l'économie par le journal comme (indicateur économique) ou (production industrielle) et pour moitié d'articles non économiques c'est-à-dire dans une catégorie non liée à l'économie par le journal comme arts, spectacle etc (cf. Tableau 1.3-1). Une régression logistique est ensuite entraînée sur cette base à prédire la catégorie des articles. Sur un échantillon test, la régression des Echos parvient à bien classer 96,8 % des articles. Le modèle du Monde classe quant à lui correctement 94,2 % des documents. Au final, 26 % des articles des Echos traitent d'économie. C'est le cas de seulement 14 % des articles du Monde (cf. Tableau 1.3-2).

**Tableau 1.3-1** : Thèmes de la base d'entrainement

| Journal | Thèmes économiques | Thèmes non-économiques |
|---|---|---|
| Les Echos | Indicateur économique, production industrielle, banques centrales, économie, emploi, balance commerciale | médias, services télécoms, assurances, arts, culture, santé, sport, management, éducation |
| Le Monde | économie | Sport, politique, société, culture et planète |

**Tableau 1.3-2 : Structure de la base finale d'articles**

| | Total | Le Monde | Les Echos |
|---|---|---|---|
| Nombre d'articles | 2 650 177 | 1 643 818 | 1 006 359 |

| | | | |
|---|---|---|---|
| Nombres d'articles « économiques » | 487 840 | 226 914 | 260 926 |
| Proportion dans le total | 100 % | 62 % | 38 % |
| Proportion dans le total « économique » | 18 % | 46 % | 54 % |

**Source :** *Le Monde* et *Les Echos*, **calculs** : Insee

**Lecture :** parmi l'ensemble des articles, 62 % proviennent du *Monde*. Parmi l'ensemble des articles, 18 % sont catégorisés comme « économiques ». Enfin, parmi l'ensemble des articles « économiques », 54 % proviennent des *Échos.*

## 2. Construction de l'indicateur de sentiment médiatique

### 2.1. Méthodologie de calcul

Le présent travail reprend et étend la méthodologie introduite par (Bortoli et al [2017,2018]). 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 :

$$Sentiment_i = \frac{Nb_{positif} - Nb_{négatif}}{Nb_{article}}$$

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 $Nb_{négatif}$divisé par le nombre de termes total de l'articles $Nb_{article}$ afin de tenir compte des différences de longueur des textes. Les indices peuvent ensuite être agrégés (moyenne) sur la période souhaitée. Il est ainsi possible de connaître le sentiment médiatique pour des fréquences trimestrielle, mensuelle mais aussi quotidienne.

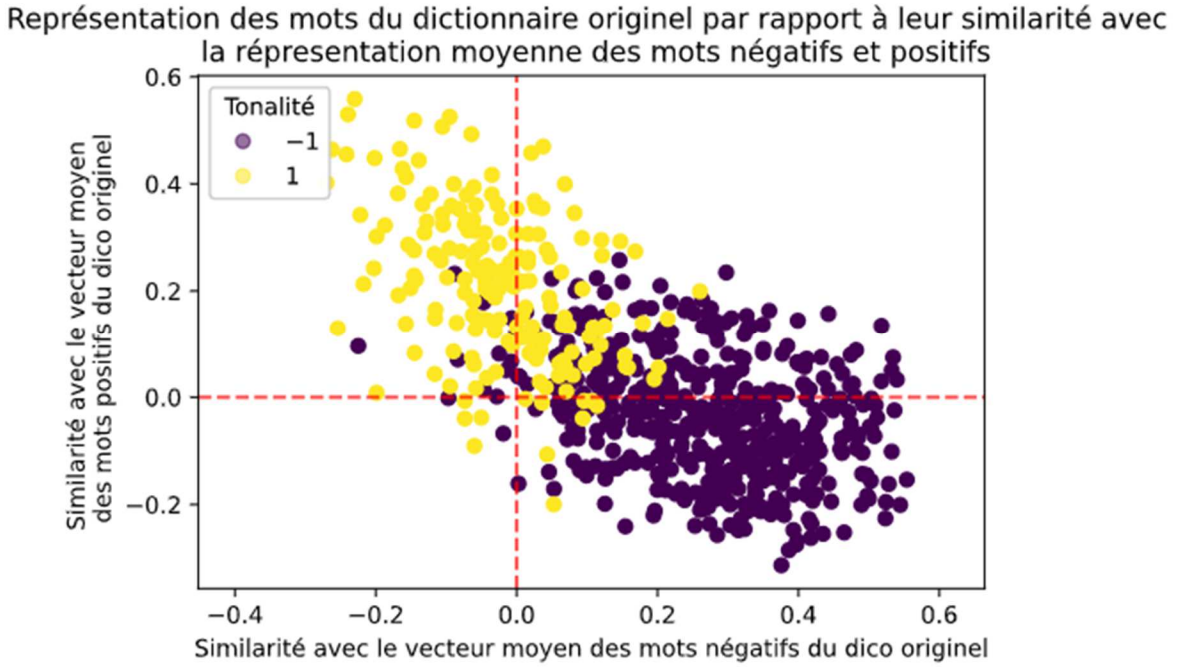
### 2.2. Dictionnaire utilisé

Le calcul d'un tel indicateur nécessite de choisir un dictionnaire capable de détecter les mots et de les associer à un sentiment (positif) ou (négatif). Loughran and McDonald [2011] conseillent d'utiliser un dictionnaire adapté au domaine défini, car les dictionnaires plus généraux captent mal les subtilités de chaque domaine. Nous reprenons le dictionnaire créé par Bortoli et al. [2018, 2018]. Ces derniers ont d'abord racinisé l'ensemble des termes du corpus. Ils ont ensuite sélectionné les racines qui apparaissent plus de 500 fois et leur ont assigné une tonalité : positif, neutre ou négatif. Afin d'améliorer leur dictionnaire, ils ont ensuite intégré des bigrammes (c'est-à-dire des suites de deux mots) afin de capter plus précisément les ambiguïtés au sein des articles et ainsi obtenir plus d'informations lors de l'analyse lexicale des articles. Ce dictionnaire a été retravaillé par nos soins afin

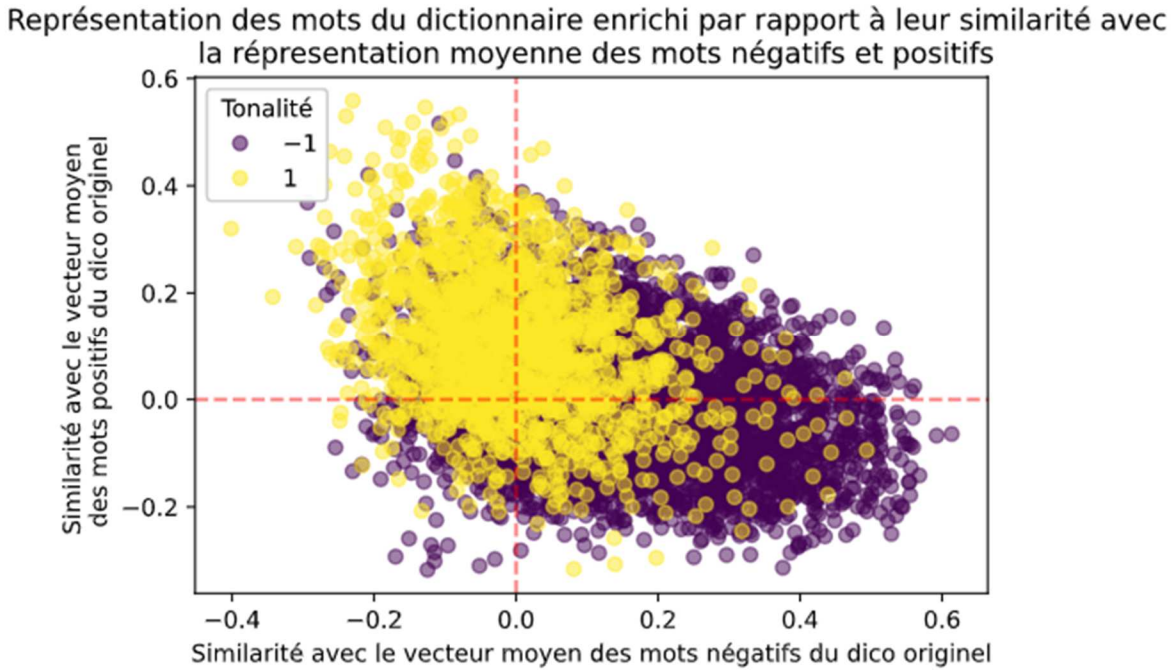
d'en obtenir une version lemmatisée. Cela a permis d'éliminer certains mots parasites, n'ayant que peu de rapport avec l'économie, tels que (cris).

Deux approches ont permis d'enrichir le dictionnaire de Bortoli et al. [2017 et 2018]. La première repose sur un modèle Word2Vec développé par google (Mikolov [2013]) entrainé sur notre base d'articles. Ce type de modèle repose sur un réseau de neurones à deux couches et cherche à apprendre des représentations vectorielles des mots. Dans cet espace vectoriel, des mots utilisés dans des contextes similaires apparaîtront comme proches. C'est cette propriété qui est utilisé ici. Tous les mots proches, c'est-à-dire ayant une similarité cosinus supérieur à 0,7[1], d'un mot du dictionnaire existant sont utilisés pour l'enrichissement avec la même tonalité. Un modèle Word2Vec est entrainé pour chaque année sur les données des deux journaux afin de tenir compte d'éventuelles modifications du vocabulaire utilisé au cours du temps. À des fins de vérifications, les mots sont projetés dans l'espace constitués par les vecteurs moyens de mots « positifs » et « négatifs » du dictionnaire originel (cf. graphique 2.2-1). Si une séparation apparaît pour les mots du dictionnaire initial : les mots négatifs (violets) sont majoritairement dans le cadran sud-est et les mots positifs (jaunes) sont dans le cadran nord-ouest, l'ajout des nouveaux mots floute le résultat. Une nouvelle étape de nettoyage serait peut-être la bienvenue ici. Par exemple, sélectionner les nouveaux mots que s'ils respectent la règle de la cohérence (cadran nord-ouest pour les mots positifs, et sud-est pour les négatifs). Un examen manuel du dictionnaire serait peut-être souhaitable.

**Graphique 2.2-1**: Représentation vectorielle (via la similarité cosinus) des mots des deux versions du dictionnaire par rapport au vecteur positif moyen et au vecteur négatif moyen.




[1] Ce seuil est arbitraire. Un comptage de mots permet de vérifier que ce seuil permet de récupérer presque l'ensemble des mots parmi les 50 plus proches d'un mot du dictionnaire.




Un deuxième enrichissement consiste à utiliser des forêts aléatoires cherchant à prédire le climat des affaires, bien corrélé à l'activité économique. Cette première étape permet d'obtenir une liste de mots pertinents. Ces mots sont ensuite utilisés dans une régression pénalisée (LASSO) afin de prédire le climat des affaires. Les coefficients associés à chacun des mots retenus servent pour catégoriser la tonalité de chaque mot.

À l'issu de ces deux enrichissements, le dictionnaire passe de 1 596 lemmes, dont 549 lemmes positifs et 1 047 lemmes négatifs, à 7 433 lemmes, dont 2 183 mots positifs et 5 250 mots négatifs. 5 418 des nouveaux mots viennent de l'enrichissement Word2Vec contre seulement 904 pour les forêts aléatoires.

## 3. Résultats

### 3.1. Une première analyse graphique

Sur l'ensemble de la période d'étude, l'indicateur de sentiment médiatique suit les évolutions de l'activité économique, y compris pendant la crise sanitaire du covid 19 (cf. graphique 3.1-1). L'enrichissement du dictionnaire permet d'améliorer la corrélation de l'indice calculé avec l'indice de climat des affaires (cf. Tableau 3.1-1). L'utilisation simultanée des deux journaux (le quotidien généraliste *Le Monde* et le quotidien spécialisée *Les Echos*) donne les meilleurs résultats. Cette version permet effectivement de réduire la volatilité de l'indice obtenu.

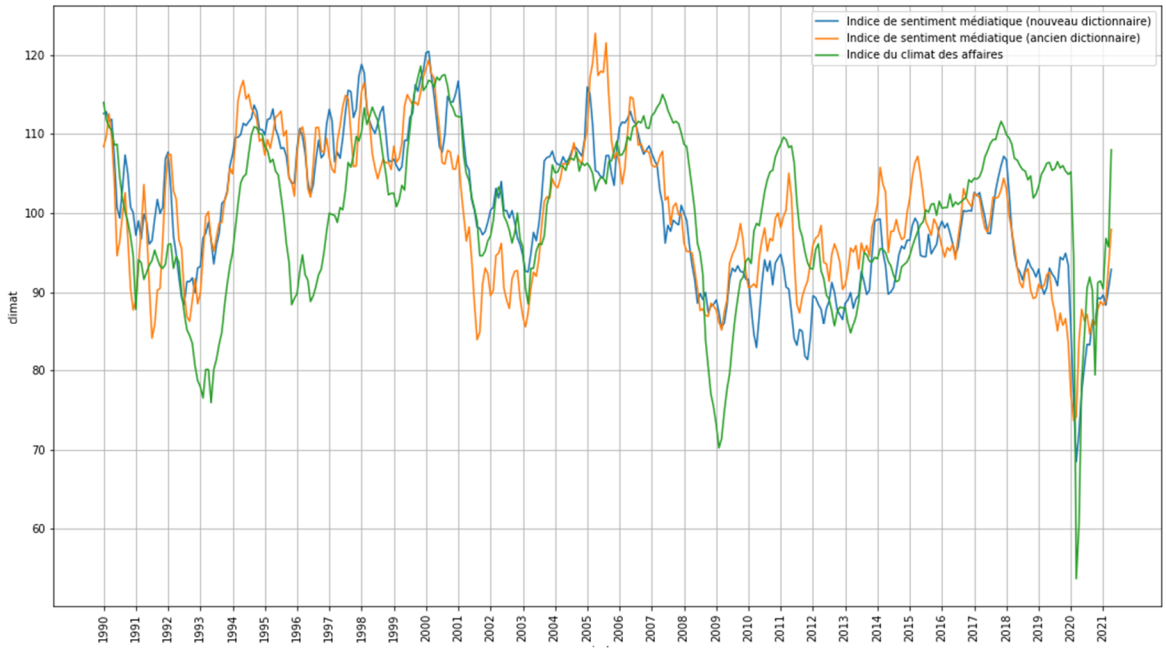
**Tableau 3.1-1**: corrélation des indicateurs (lissés avec une moyenne mobile d'ordre 3) avec l'indice de climat des affaires

| Base | *Les Echos* | *Les Echos* et *Le Monde* |
|---|---|---|

| Dictionnaire | Ancien | Nouveau | Ancien | Nouveau |
|---|---|---|---|---|
| Corrélation avec le climat des affaires | 0,488 | 0,559 | 0,624 | 0,685 |

**Note :** Les résultats restent identiques en excluant la fin de la période (après 2019) montrant des évolutions très importantes, sans commune mesure avec le reste de la série.

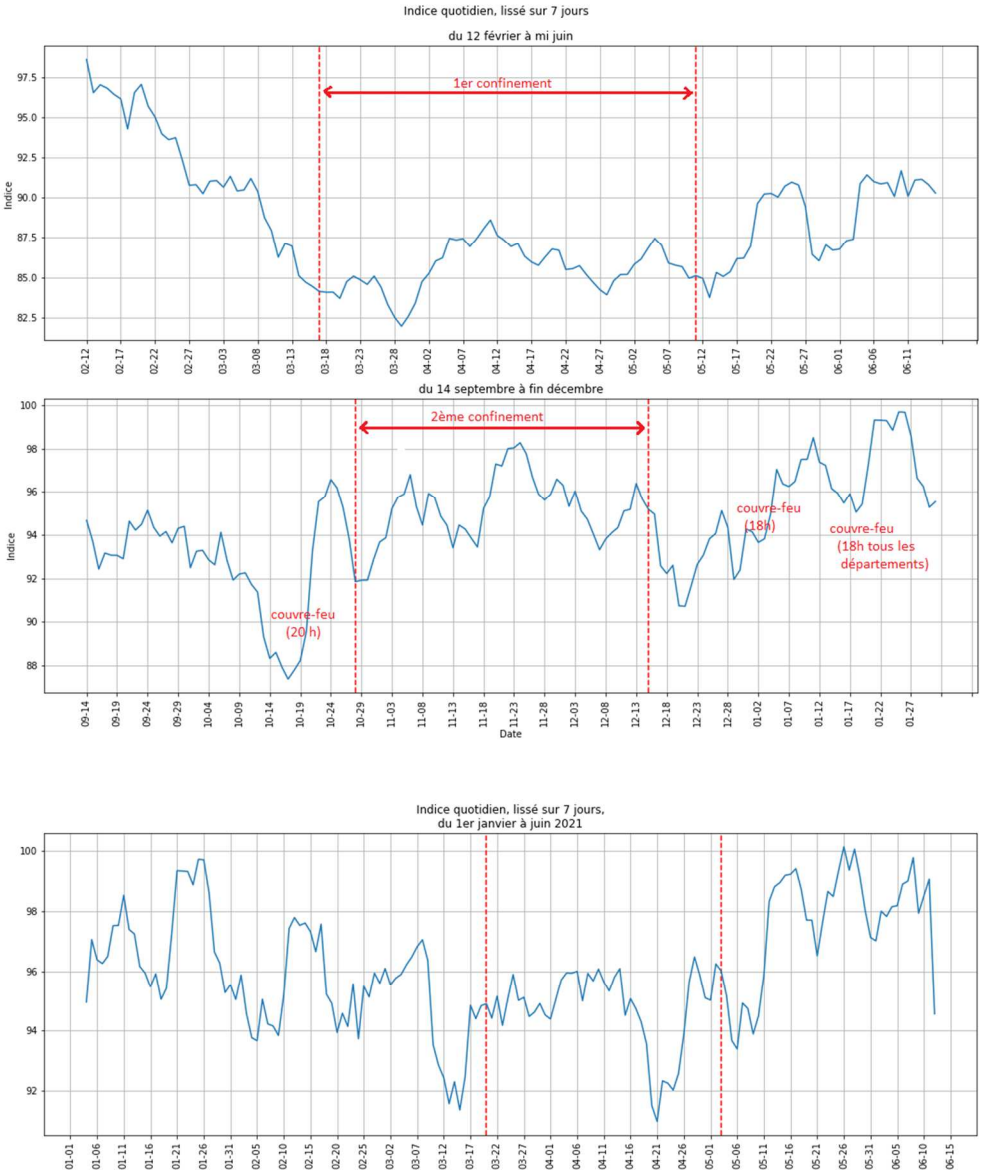
**Graphique 3.1-1:** Indicateur de sentiment médiatique (moyenne 100 et écart-type 10) et indice du climat des affaires.




*Remarque : les données sentiment ci-dessus ont été lissées avec une moyenne mobile d'ordre 3. Elles ont ensuite été centrées-réduites afin d'avoir un écart-type de 10 et une moyenne de 100.*

L'indice calculé peut être calculé à une fréquence quotidienne. Il permet ainsi une analyse très fine de périodes particulières, intéressantes d'un point de vue économique. Ainsi, un indice quotidien est calculé sur la période de la crise sanitaire du covid 19 de février 2020 à juin 2021 (graphique 3.1-2). Cet indice est standardisé de façon indépendante de l'indice mensuel. Son niveau et l'ampleur de ses évolutions ne sont donc pas comparables mais il reste informatif. Par exemple, dès début mars 2020, l'indice quotidien de sentiment médiatique commence à diminuer de manière importante, s'éloignant de plus de 10 points de sa moyenne de long terme (100). Cet période correspond en effet au premier confinement. Il atteint ensuite un plancher, auquel il reste pendant la totalité du premier confinement, pour remonter ensuite à partir de la troisième semaine de mai. En juin, il retrouve des niveaux plus comparables à sa moyenne de long terme, bien que dégradés d'une dizaine de points. Ainsi, dans le contexte très particulier du début de la crise sanitaire où les indicateurs conjoncturels usuels étaient soit non encore disponibles soit peu opérants, l'indice de sentiment médiatique a livré une information pertinente avant et pendant le premier confinement.

**Graphique 3.1-2:** Indice de sentiment médiatique quotidien, zoom sur les deux périodes de confinement (lissé sur 7 jours), et sur le premier semestre 2021




### 3.2. Introduction de l'ISM dans un modèle de prévision

Il est possible d'introduire l'indicateur précédemment calculé dans une régression afin de prévoir les évolutions du PIB (en variations trimestrielles) en utilisant les retards du PIB, l'indicateur de climat des affaires et l'indice de sentiment médiatique. Afin de gérer la différence de fréquence entre les variables (trimestrielle pour le PIB et mensuelle pour le climat des affaires et l'indice de sentiment), l'approche retenue consiste à proposer un étalonnage différent selon le mois du trimestre, de manière à exploiter chaque mois l'intégralité de l'information disponible. Pour chaque mois du trimestre étudié, le but est d'utiliser le maximum d'information disponible en proposant des étalonnages différents. Ainsi, les étalonnages "mois 1", "mois 2" et "mois 3" utilisent respectivement l'intégralité de l'information disponible à la fin du premier, du deuxième et du troisième mois du trimestre. Le régresseur pour le climat des affaires correspond, lors du premier mois du trimestre, à la variation entre la valeur de l'indicateur du 1$^{er}$ mois du trimestre par rapport à la moyenne du trimestre précédent. Au "mois 2", il correspond à la variation entre la valeur moyenne des deux premiers mois par rapport à la valeur du trimestre précédent. Au "mois 3", l'ensemble de l'information est utilisé. Pour la variable Sentiment (ISM), la même logique est adoptée, à l'exception du fait qu'elle est prise

en niveau et non en différence, s'inspirant ainsi de Bortoli et al. (2018). L'introduction de retards de l'indice de sentiment a été testée en partant du principe que l'indice reflète la croissance contemporaine et celle des trimestres récents. Cependant cette méthode n'a pas fourni de bons résultats. Enfin, le retard du PIB est également utilisé en tant que variable explicative.

Quatre modèles sont estimés sur la période allant du T1 1993 au T4 2019 afin de comparer les performances prédictives du climat des affaires et de l'indicateur de sentiment médiatique. L'année 2020, pour laquelle les méthodes habituelles de prévision à l'aide des enquêtes de conjoncture n'ont pas été opportunes, n'a pas été prise en compte dans l'estimation. Le modèle 1 ne comprend qu'une variable explicative: le PIB retardé. Ce modèle est identique pour les trois mois du trimestre. Le modèle 2 combine le climat des affaires et le retard du PIB. L'indice de sentiment médiatique se substitue au climat des affaires dans le modèle 3. Enfin, le modèle 4 intègre simultanément ces trois variables explicatives.

Les différents étalonnages estimés sont les suivants :

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \epsilon_T \text{(Modèle 1)}$$

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \alpha_3 \Delta Climat_T + \epsilon_T \text{(Modèle 2)}$$

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \alpha_3 Sentiment_T + \epsilon_T \text{(Modèle 3)}$$

$$\Delta PIB_T = \alpha_1 + \alpha_2 \Delta PIB_{T-1} + \alpha_3 \Delta Climat_T + \alpha_4 Sentiment_T + \epsilon_T \text{(Modèle 4)}$$

**Tableau 4.2-1 : $R^2$ ajustés des étalonnages**

| | Mois 1 | Mois 2 | Mois 3 |
|---|---|---|---|
| Modèle 1 | 0,300 | 0,300 | 0,300 |
| Modèle 2 | 0,315 | 0,321 | 0,308 |
| Modèle 3 | 0,342 | 0,342 | 0,300 |
| Modèle 4 | 0,336 | 0,339 | 0,302 |

Pour le premier et le deuxième mois du trimestre, l'indicateur de sentiment médiatique semble être un meilleur prédicteur que le climat des affaires. Aussi, lorsque le l'indicateur de sentiment médiatique est combiné au climat des affaires, le $R^2$ ajusté des modèles est supérieur à ceux des modèles contenant uniquement le climat des affaires. Au troisième mois du trimestre l'apport d'une variable explicative (climat des affaires et/ou indicateur de sentiment) est marginal.

L'indicateur calculé à partir d'un simple comptage de mots possède certains avantages : l'ajout de quelques points en fin de série ne provoque pas de modification de tous les points antérieurs. La méthode est assez simple, ce qui lui assure une certaine robustesse. Elle reproduit bien les évolutions du climat des affaires.

L'indicateur est cependant très bruité. Bortoli et al. avaient choisi de le traiter avec une moyenne mobile afin de le rendre plus lisible. L'usage d'une moyenne mobile pose certaines questions : quels ordres utiliser ? L'utilisation d'une moyenne mobile symétrique fait entrer un peu du futur de

la série dans son passé, et donc dans un potentiel échantillon d'apprentissage. Nous disposons d'un très grand nombre d'articles. Cette dimension est finalement très peu utilisée. Après agrégation, elle disparaît. Afin de contourner ces obstacles, une approche de prédiction du PIB directement à partir des séries de comptage de mots des articles a été testée.

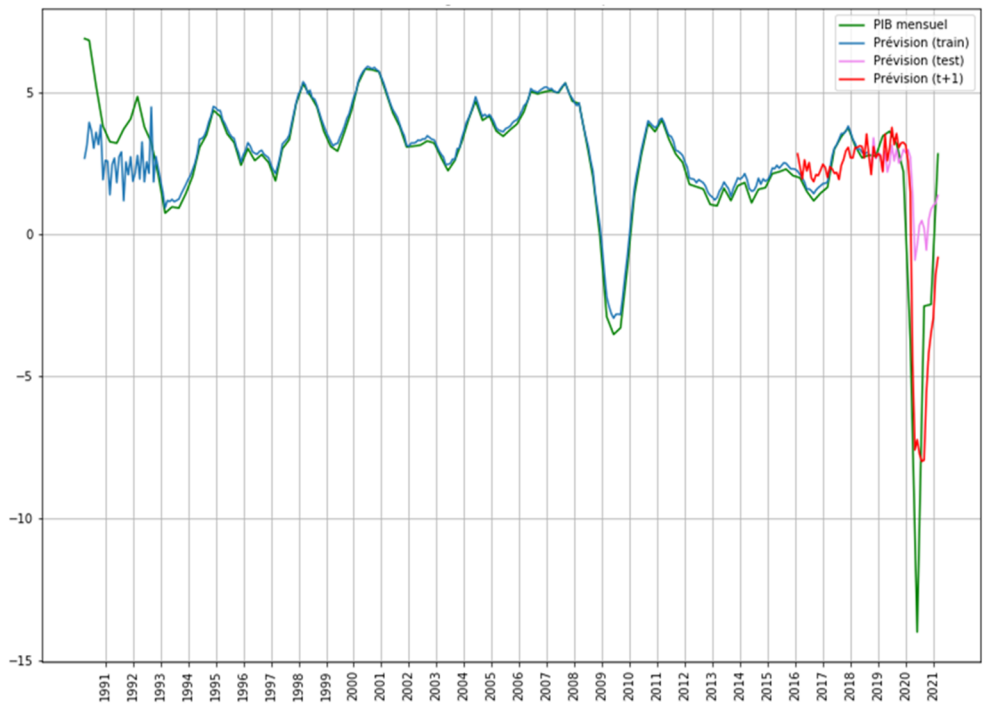
## 4. Une approche exploratoire : prévision du climat des affaires à l'aide des séries de mots

Une autre approche consiste à essayer de prédire un indice important pour la conjoncture (le PIB ou un indice qui lui est corrélé) mais disponible avec un certain retard directement à l'aide des séries de nombre de mots. Effectivement, la fréquence d'apparition de certains mots tels que « crise » ou « grève » est liée à l'activité économique. Un modèle de machine learning (forêt aléatoire ou régression pénalisée) peut être utilisée afin de prédire le niveau d'activité du mois ou du trimestre.

La variable la plus importante à prédire est sans conteste le PIB. Cet indicateur est trimestriel ce qui diminue fortement le nombre de points disponibles pour estimer le modèle. Il a donc été mensualisée : le PIB trimestriel a été lissé afin d'obtenir des données mensuelles. Cette méthode est rudimentaire. Elle présente notamment le désavantage de faire rentrer de l'information future dans les points présents via la moyenne mobile. Des tests ont également été réalisés avec l'indice du climat des affaires.

Parmi les méthodes possibles de sélection des variables explicatives, c'est une régression pénalisée de type elasticnet qui a été privilégiée, car sélectionnant automatiquement les variables pertinentes. Un premier test a été réalisé en statique. Une période d'apprentissage allant jusqu'en décembre 2015 a été sélectionné. Les résultats obtenus ont ensuite été évalué sur la période restante (échantillon test). La courbe rose ci-dessous (figure n°) montre bien une forte baisse en 2021.

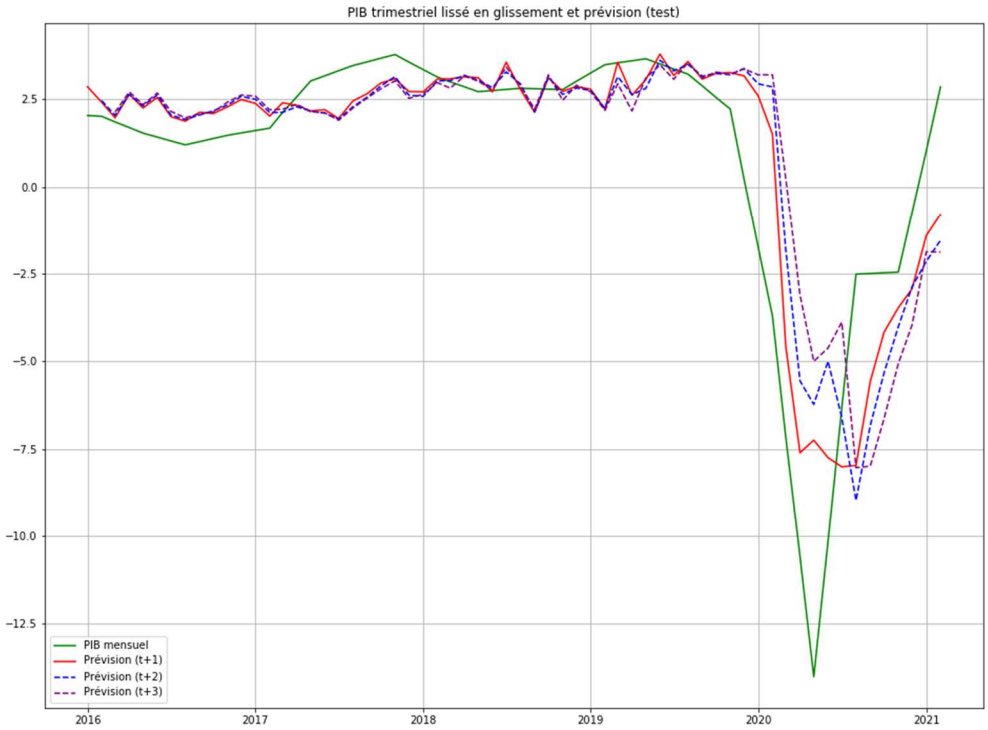
**Graphique 4.1-1:** Prévisions du PIB mensualisé à l'aide d'un modèle de machine learning




**Note :** Le modèle est entrainé sur la période allant de janvier 1991 à décembre 2015. Il est testé sur la période allant de janvier 2016 à janvier 2021. Pour la courbe rose, le modèle n'est pas réentrainé, c'est-à-dire que les coefficients de la régression pénalisée ne sont pas réestimés.

Une évaluation de la performance du modèle a également été réalisée sur des fenêtres glissantes. La période initiale s'étend de janvier 1991 à décembre 2015. La prévision hors échantillon commence en 2016 et se situe sur une fenêtre temporelle croissante : chaque mois supplémentaire de prévision conduit à une ré-estimation sur l'ensemble de la période de l'échantillon, intégrant dès lors les nouvelles informations mensuelles disponibles. La prévision qui en résulte est donc effectuée en pseudo-temps réel. La régression pénalisée, en utilisant les variations de l'occurrence relative des mots au cours des mois, permet de bien prévoir le PIB mensualisé sur la période d'apprentissage, c'est-à-dire entre 1990 et 2015 (graphique 4.1-2). Le modèle est estimé sur cette période et ajuste donc parfaitement les données, son $R^2$ s'élève à 0,96. Sur l'ensemble de l'échantillon de test, avec des données nouvelles, le $R^2$ s'élève à 0,58, et ce en utilisant uniquement les séries d'occurrences relatives des mots. Sur la partie hors de l'échantillon, soit depuis 2016, la prévision en t+1 (la courbe rouge) suit les mouvements de l'activité économique ainsi que leur ampleur. En particulier, lors de la chute très brutale d'avril, le modèle réussit une prévision très précise. En sélectionnant de manière automatique des termes instructifs, comme « crise », « quarantaine » et « épidémie », il réussit à prévoir cet effondrement alors même qu'aucune baisse aussi forte n'a été observée depuis le début de l'échantillon.

**Graphique 4.1-2:** Prédictions du modèle sur des fenêtres glissantes




**Note :** Le modèle (régression pénalisée de type ElasticNet) est initialement entrainé sur la période allant de janvier 1991 à décembre 2015. Les données de janvier 2016 à janvier 2021 sont progressivement intégrées dans l'échantillon d'apprentissage mois par mois. Les courbes rouge, bleue et violette correspondent aux résultats obtenus sur les données de test à différents horizons de prévisions.

D'autres algorithmes de sélection de mots sont possibles. En particulier, des méthodes mobilisant des forêts aléatoires et un réseau de neurones ont été expérimentées. Elles ne donnaient cependant pas de bons résultats. Pour le réseau de neurones, la fréquence insuffisante des données (mensuelles) empêche le réseau de généraliser correctement une fois en phase de prévision.

Le travail de cette partie possède certaines limites. Il s'attache à prédire le PIB en glissement annuel. La série est mensualisée de façon assez simpliste. Le modèle néglige l'aspect temporel des données.

## Conclusion

Les articles de journaux constituent une nouvelle source de données qu'il est possible de mobiliser afin de prédire les évolutions de la conjoncture économique. Les nouveaux articles peuvent être facilement récupérés afin d'actualiser le calcul de l'indice. Toutefois, certaines évolutions peuvent être difficiles à interpréter : baisse ou augmentation subite du nombre de journaux, modification des mots employés, etc. Ces modifications peuvent avoir un impact sur les indicateurs calculés.

Les tests menés dans cet article semblent confirmer les résultats de l'article de Bortoli et al. (2017). L'indicateur s'appuyant sur le décompte de mots positifs et négatifs contenus dans les articles améliore la prévision du PIB. L'enrichissement du dictionnaire proposé ici améliore de plus les résultats.

L'indicateur calculé à partir du journal des *Echos*, plus spécialisé en économie, semble plus performant que celui calculé à partir du journal du *Monde*, plus largement diffusé mais généraliste. En pratique, il semble toutefois préférable d'empiler les différentes sources. Ici, aucune réflexion n'a été mené sur les pondérations des articles des différents journaux.

Les techniques de machine learning introduites ici afin de prédire le PIB directement à partir des mots des articles semblent prometteuses. Elles doivent cependant encore être approfondies. Dans cette perspective, bien que nous disposons d'un nombre d'articles conséquent, le nombre de points à prédire est faible (les séries ne débutent que dans les années 90 ce qui fait au plus 400 points en considérant des séries mensuelles). Les nombres moyens de mots sont des séries temporelles. Ceci ouvre des perspectives mais apporte aussi son lot de difficultés. Les méthodes de machine learning usuelles doivent être repensées afin de s'adapter à des données en entrée dépendantes du temps. Si elles permettent parfois un relâchement des contraintes des modèles classiques, leur usage n'est pourtant pas sans risque. Une forêt aléatoire, par exemple, ne saurait prédire une valeur absente de son échantillon d'apprentissage. Il lui sera donc impossible de prédire des données avec une tendance. Il serait possible de lui faire prédire des évolutions (c'est-à-dire différencier les séries). Mais que se passe-t-il dans ce cas lors d'évolutions extraordinaires comme dans le cas de la crise sanitaire du Covid-19 ?

**Bibliographie**

[1] Arion G., Himpens S., Roudil-Valentin T., « L'activité économique française au travers d'articles de presse », *Notes de conjoncture*, mars 2021.
[2] Bortoli C., Combes S., Renault T., « Prévoir l'emploi en lisant le journal », *Notes de conjoncture*, mars 2017.
[3] Bortoli C., Combes S., Renault T., « Prévoir la croissance du PIB », *économie et statistique*, 2018.
[4] Fraiberger, S. P., « News sentiment and cross-country fluctuations. », *Association for Computational Linguistics*, 2016
[5] Garcia, D. « Sentiment during recessions. », *The journal of Finance*, 68(3) :12671300, 2013
[6] Loughran T., McDonald B., « When is a liability not a liability ?textual analysis, dictionaries, and 10-ks», *The journal of Finance*, 2011.
[7] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). « Efficient estimation of word representations in vector space ». *arXiv preprint arXiv:1301.3781*.
[8] Shapiro A. H., Sudhof M., Wilson D., « Measuring news sentiment », *Federal Reserve Bank of San Francisco*, 2020
[9] Tetlock, P. C. «Giving content to investor sentiment : The role of media in the stock market. » The journal of finance, 62(3) :11391168, 2007.