Sous la direction scientifique de
Nathalie de Marcellis-Warin – Benoit Dostie
Sous la coordination de
Genevieve Dufour

Le Québec économique 9

Perspectives et défis de la transformation numérique

Chapitre 11

SCIENCE DES DONNÉES, RÉSEAUX SOCIAUX ET POLITIQUES PUBLIQUES

BENOIT AUBERT, NATHALIE DE MARCELLIS-WARIN, THIERRY WARIN

CIRANO

2020CH-11
CHAPITRE DE LIVRE

CH

Chapitre 11

SCIENCE DES DONNÉES, RÉSEAUX SOCIAUX ET POLITIQUES PUBLIQUES

Benoit Aubert
Professeur titulaire à HEC Montréal, chercheur et fellow au CIRANO

Thierry Warin
Professeur titulaire à HEC Montréal, chercheur et fellow au CIRANO

Nathalie de Marcellis-Warin
Professeure titulaire à Polytechnique Montréal et présidente-directrice générale du CIRANO

Résumé

Dans le contexte de la transformation numérique, la fréquence d'utilisation des moteurs de recherche et des réseaux sociaux ne cesse d'augmenter, ce qui génère un grand nombre de données. La science des données permet de traiter, de visualiser et d'analyser de tels ensembles de données non structurées, incluant les données issues de textes. Les gouvernements ont accès à ces nouvelles sources de données, à ces nouveaux modèles d'analyse et à une puissance de calcul sans précédent qu'ils peuvent utiliser pour mieux comprendre certains phénomènes sociétaux actuels et ainsi renforcer l'efficacité des politiques publiques. Dans ce chapitre, nous présentons ces nouvelles sources de données et les nouvelles méthodologies d'analyse ainsi que les enjeux à prendre en compte. Plusieurs exemples d'analyses de données non structurées issus de travaux de recherche effectués au CIRANO sont présentés à titre d'illustration, notamment des exemples dans le contexte de la COVID-19.

Introduction

Selon Marco Iansiti et Karim Lakhani, deux chercheurs de la Harvard Business School, les données sont un atout stratégique à la fois pour la prise de décisions au sein des entreprises et pour l'efficacité organisationnelle, mais aussi pour la survie des entreprises dans le contexte de la révolution numérique (Lansiti et Lakhani, 2020). La science des données et la disponibilité de différents types de données – structurées et non structurées – modifient ce que nous pouvons mesurer et comment nous pouvons le mesurer. Les sources d'information se transforment rapidement et les données non structurées peuvent améliorer notre compréhension de certains phénomènes, ce qui favorise une meilleure prise de décision par exemple en matière de politiques publiques.

Les données structurées sont des données pouvant être clairement identifiées et codifiées. Les données d'un tableur sont typiquement des données structurées. On peut comprendre leur signification en croisant les titres de la ligne et de la colonne dans laquelle se trouvent les données. Ces données répondent à une codification qui permet de les classer et d'en tirer une information. Les systèmes d'analyse algorithmique ont depuis toujours été développés afin de traiter ce genre de données. L'ère des données massives permet surtout un traitement de grande ampleur et en temps réel de ces données (Alaoui, 2018).

Les données non structurées ne répondent pas à une codification permettant d'extraire mécaniquement une information. Il n'y a pas d'autres moyens que de lire les gazouillis (*tweets*) pour en extraire le sens. C'est ensuite en analysant le contenu des messages que l'on structure l'information. Des exemples de types de données non structurées sont des fichiers textes, des images, des fichiers audio ou vidéo et toute autre information issue d'un signal analogique. Les données non structurées représentent actuellement la grande majorité de l'information, mais encore la partie la moins exploitée. En mettant en place des outils de collecte et d'analyse de données massives efficaces, cette information devient exploitable (Warin et De Marcellis-Warin, 2014).

Par exemple, dans le contexte de la pandémie de COVID-19, il était important d'analyser en très peu de temps un grand volume de références scientifiques en épidémiologie. Pour ce faire, la première étape était de collecter les références et de structurer une base de données afin de la rendre analysable. Warin (2020) a développé un logiciel permettant d'avoir accès à plus de 85 000 publications médicales sur les coronavirus et d'intégrer ces données dans le langage d'analyse statistique *R*. Une analyse bibliométrique, une analyse de réseau et un calcul des indices de similitude peuvent être réalisés au moyen de ce logiciel. Il offre entre autres la possibilité d'associer des recherches sur un protocole médical à un réseau de chercheurs à la pointe sur ce protocole. Cela peut ainsi aider à gagner du temps dans cette course contre le virus.

Les réseaux sociaux de contenu favorisent la publication d'un contenu original, destiné à être partagé avec la communauté. Pour exister dans un environnement fortement concurrentiel, ces réseaux sociaux adoptent souvent un positionnement original. Twitter, par exemple, a misé sur la publication de messages courts. Ces messages sont un exemple de données qui peuvent être exploitées pour améliorer la prise de décision lorsqu'elles sont jumelées à d'autres sources d'information. Selon le NETendance 2018, 65 % des adultes québécois se sont connectés au moins une fois par jour aux réseaux sociaux[1], et selon le Baromètre CIRANO 2018, les réseaux sociaux sont la troisième source d'information des adultes québécois[2].

Nous allons présenter ces nouvelles sources de données et les nouvelles méthodologies d'analyse ainsi que des enjeux à prendre en compte. Plusieurs exemples d'analyses de données non structurées issus de travaux de recherche effectués au CIRANO seront présentés à titre d'illustration, notamment des exemples dans le contexte de la COVID-19.

Nouvelles sources de données pour les politiques publiques

Les nouvelles sources de données susceptibles d'être utilisées sont très variées, par exemple les recherches sur Internet, mais aussi de plus en plus des données issues de textes à l'image des conversations sur les réseaux sociaux, des textes de discours ou des rapports annuels. Pour les chercheurs en sciences sociales, l'information codée dans un texte est un complément riche aux types de données plus structurées traditionnellement utilisées dans la recherche, et ces dernières années ont vu une explosion de la recherche utilisant le texte comme données.

Ces données sous forme de textes peuvent être utilisées pour prévoir une variété d'événements ou pour mieux comprendre certains phénomènes (Gentzkow, Kelly et Taddy, 2019). Dans le domaine de la finance, les textes issus des actualités financières, des médias sociaux et des rapports déposés par les entreprises peuvent être utilisés pour prévoir les mouvements des prix des actifs et étudier l'impact causal des nouvelles informations (Warin et De Marcellis-Warin, 2014). Nyman et Ormerod (2020) donnent plusieurs autres exemples d'utilisation des données textes : en macroéconomie, le texte peut être utilisé pour prévoir les variations de l'inflation et du chômage et estimer les effets de l'incertitude politique. En analyse des médias, le texte des nouvelles et des médias sociaux est utilisé pour étudier les moteurs et les effets de l'orientation politique. En organisation industrielle et en marketing, les textes des publicités et des revues de produits sont utilisés pour étudier les facteurs de décision des consommateurs. En économie politique, le texte des discours des hommes politiques est utilisé pour étudier la dynamique des programmes et des débats politiques.

La principale différence entre le texte et les types de données souvent utilisées en économie est que le texte inclut plusieurs dimensions. Par exemple, un message sur Twitter comporte 280 caractères (depuis 2017), mais on peut aussi lui associer des métadonnées (nom de l'utilisateur, localisation GPS, etc.) ainsi que les mots-clics (*hashtags*) utilisés.

Données de recherches sur Internet

En raison de l'émergence des moteurs de recherche, les termes de recherche peuvent fournir un aperçu des intérêts actuels dans de nombreux domaines, tels que l'économie, la politique, la santé, etc. La recherche sur Google du mot « recession » a atteint un pic un mois avant le début de la récession américaine de septembre 2008 (Tkacz, 2013).

Choi et Varian (2012) ont montré comment utiliser les données des moteurs de recherche pour prévoir les valeurs à court terme des indicateurs économiques. Il s'agit par exemple des ventes d'automobiles, des demandes d'indemnités de chômage, de la planification des destinations de voyage et de la confiance des consommateurs. Wu et Brynjolfsson (2013) ont d'ailleurs utilisé les données des recherches sur Google pour prévoir les prix des logements, car les recherches sur des termes tels que « immobilier » ou « prix des logements » sont corrélées avec les intentions des acheteurs ou des vendeurs. Les données de recherche Google ont également été utilisées avec succès pour la production de prévisions sur le marché du travail, et leur utilité en tant qu'indicateurs économiques avancés a été étudiée par certaines banques centrales (Tkacz, 2013). Plusieurs recherches sont en cours au CIRANO afin d'analyser l'activité économique au Québec en utilisant les données de recherche sur les mots clés « emploi », « taux de chômage », « investissement », « consommation », etc.

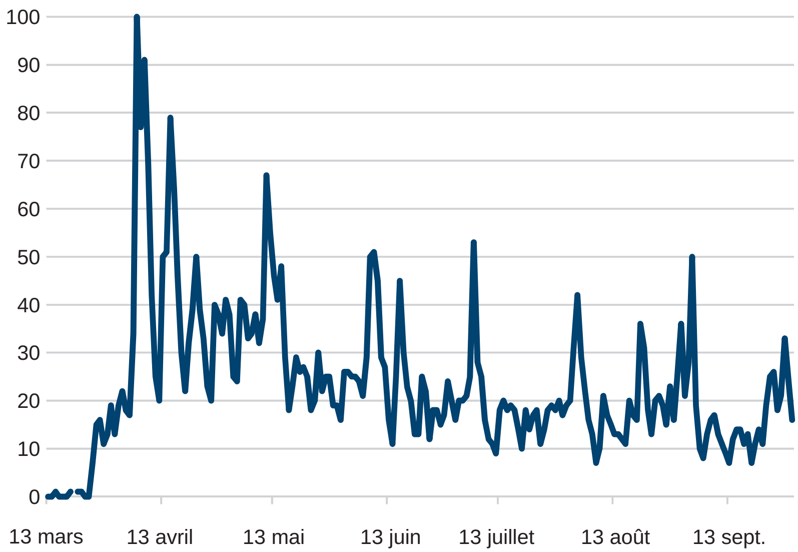
Un autre exemple d'intérêt pour la santé publique concerne les recherches sur les symptômes de grippe pour estimer le nombre de personnes atteintes. On présume qu'une personne faisant des recherches utilisant les termes « fièvre », « mal de tête » ou « grippe » est probablement atteinte par ces symptômes, ou fait la recherche pour un proche qui souffrirait de ces symptômes. Les données des recherches au Canada ont montré que cet algorithme permet d'avoir une idée de la propagation du virus bien avant les données officielles rapportées par les autorités médicales et peut donc améliorer les interventions publiques. L'algorithme est toutefois vulnérable. Une médiatisation de la saison de grippe tend à exagérer le nombre de cas (Tkacz, 2013). De plus, comme l'algorithme a été construit en utilisant les données de grippe saisonnière, il est moins précis pour prédire des souches inhabituelles, comme lors de l'épidémie de H1N1 (Eisenstein, 2018). Pour le système de santé, ces indicateurs peuvent néanmoins permettre d'anticiper des crises potentielles et de mieux se préparer à y faire face.

La plateforme statistique *Google Trends* permet de mesurer la proportion de recherches portant sur un mot clé donné dans une région et pour une période spécifique, par rapport à la région où le taux d'utilisation de ce mot clé est le plus élevé (valeur de 100). Ainsi, une valeur de 50 signifie que le mot clé a été utilisé moitié moins souvent dans la région concernée et une valeur de 0 signifie que les données pour ce mot clé sont insuffisantes. Dans le contexte de la crise de la COVID-19, de nombreux termes ont été recherchés sur les moteurs de recherche[3]. « Coronavirus » a fait partie des mots les plus recherchés dans le monde avec un pic entre les 13 et 15 mars 2020. Les recherches associées à ce mot clé soulevaient certaines questions, par exemple : *quels sont les symptômes du coronavirus à l'origine de la pandémie ayant débuté en 2019 ? Quand la COVID-19 va-t-elle se terminer ? Qu'est-ce que la COVID-19 ? Combien de personnes sont mortes du coronavirus ? Le coronavirus s'affaiblit-il ?* Aux États-Unis, les questions étaient surtout reliées aux symptômes et au vaccin : *la nausée est-elle un symptôme de la COVID-19 ? Le vomissement est-il un symptôme* *de la COVID-19 ? L'éternuement est-il un symptôme de coronavirus ? Le mal de gorge est-il un signe de la COVID-19 ? Existe-t-il un vaccin contre les coronavirus ?*

Les questions posées permettent de déterminer, dans ce contexte précis, quelles sont les préoccupations de la population. De nombreuses autres questions ont été posées dans les recherches concernant le port du masque et aussi les impacts économiques du confinement, notamment le chômage. Le nombre de recherches avec le mot « chômage » a considérablement augmenté au Canada après le 13 mars 2020 avec un pic observé durant la semaine du 22 mars 2020, soit après l'annonce du confinement[4]. À partir du 2 mai 2020 et jusqu'à la fin août 2020, le nombre de recherches relatives a été équivalent à celui de la période précédant le 13 mars 2020.

Pour le Canada, le mot clé « PCU » (pour « prestation canadienne d'urgence ») ainsi que le mot clé « CERB » (correspondant au terme en anglais, « Canada Emergency Response Benefit ») ont été aussi beaucoup recherchés. Le graphique 11-1 reprend l'évolution des recherches « PCU » sur le moteur de recherche Google avec sur l'axe des abscisses la date de la recherche sur le moteur de recherche et sur l'axe des ordonnées le taux de recherche du mot clé depuis le Canada.

Recherches effectuées depuis le Canada entre le 13 mars et le 30 septembre 2020 sur le moteur de recherche Google incluant le mot clé « PCU »[5]



Graphique g/2020-c11-1

Source : Auteurs.

Le graphique 11-1 compare le nombre relatif de recherches par jour au Canada. On observe notamment des pics qui suivent les annonces du gouvernement fédéral. Le premier pic est survenu durant la semaine du 5 au 11 avril 2020[6], et c'est celui avec le plus haut taux d'utilisation (valeur de 100). Ensuite, la recherche reste élevée avec d'autres pics. Les sujets les plus associés à cette recherche étaient « étudiant », « Service Canada », « ARC ». La province où le mot clé « PCU » a été le plus recherché est le Québec. Pour le mot clé « CERB », on observe presque exactement la même courbe, avec les mêmes pics en avril, mais des pics moins élevés par la suite[7]. La province où on a le plus consulté ce mot clé est Terre-Neuve-et-Labrador, suivie de la Colombie-Britannique et du Manitoba ; le Québec est la dernière pour la recherche du terme en anglais (valeur de 20).

Données issues des conversations sur les réseaux sociaux

Presque tous les journaux, chaînes de télévision, stations de radio et magazines publient leur contenu pertinent sur les réseaux sociaux et notamment Twitter, ce qui génère de nombreuses données et aussi des réactions. Il existe également sur Twitter une autre source d'information générées par des personnes (« la foule ») et portant sur des sujets qui n'ont peut-être pas (encore) été couverts par les médias traditionnels. Ces deux sources d'information constituent la première couche des gazouillis. Au-dessus de cette couche, nous trouvons des gazouillis qui peuvent être regroupés autour de certains sujets qui ont pris naissance dans la première couche. C'est ce qu'on appelle généralement le « *buzz* », ou battage médiatique. En fait, ces groupes sont des conversations. Les conversations peuvent être définies comme la dynamique du partage d'informations – éditorialisées ou non – sur un sujet spécifique. Nous avons représenté cette dynamique sur la figure 11-1.

Dynamique des conversations sur les réseaux sociaux

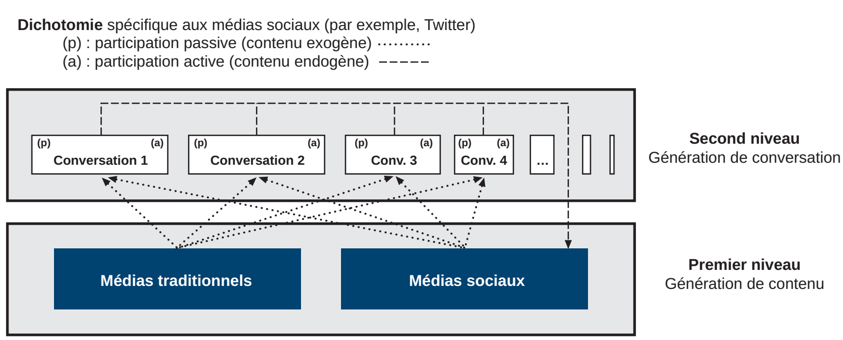


Figure f/2020-c**11-1**

Source : Auteurs.

En réalité, Twitter est une plateforme qui correspond précisément à cette définition. En effet, on trouve sur Twitter une première couche d'informations, matérialisée par des gazouillis, et les utilisateurs peuvent sélectionner un sujet, le lire et y contribuer de deux manières : (1) en commentant et/ou (2) en rediffusant les gazouillis (« *retweet* »). Au risque de trop simplifier, la première correspond à une participation active à la conversation, tandis que la seconde correspond davantage à une participation passive à la conversation. La participation passive est une contribution importante de Twitter. Qu'ils soient actifs ou passifs, ces gazouillis ajoutent de manière endogène du contenu à la première couche d'information.

Twitter a intégré sur sa plateforme des mots-clics et quelques fonctionnalités qui permettent aux chercheurs de décortiquer un échantillon de gazouillis en différentes dimensions. Par exemple, nous pouvons isoler les gazouillis provenant de médias traditionnels ou d'utilisateurs réguliers, ceux appartenant à la première couche, et ceux appartenant à la deuxième couche. Il y a également un autre aspect à prendre en compte : la dimension endogène par rapport à la dimension exogène. La couche de génération de contenu (la première couche) est d'abord composée d'informations exogènes provenant des médias traditionnels et sociaux. Ensuite, les utilisateurs génèrent les conversations de manière active (éditorialisée) et passive (rediffusion des gazouillis, par exemple). Ainsi, le contenu éditorialisé alimentera la première couche de notre taxonomie, car il correspond à une nouvelle génération de contenu – bien que connexe. La figure 11-1 illustre la dimension endogène de Twitter. Cette déconstruction est un outil puissant pour étudier la dynamique d'une conversation, en particulier lorsque cela ne peut se faire autrement.

Les messages publiés sur les réseaux sociaux fournissent une source d'information très intéressante pour connaître l'état des lieux dans plusieurs domaines. Par exemple, l'utilisation des commentaires rendus publics par des pêcheurs récréatifs permet de mieux comprendre l'état des écosystèmes. Les états des stocks de poissons peuvent être estimés en utilisant le nombre et la taille des prises qui ont été diffusés sur les réseaux sociaux (parfois avec des photos, des commentaires ou des mots-clics). Ces données sont disponibles, et peuvent être plus fiables que des données de sondage. Elles sont, de plus, gratuites (Monkman, Kaiser et Hyder, 2018).

L’analyse de textes sur les réseaux sociaux offre trois avantages (Nyman et Ormerod, 2020). Premièrement, il y a un avantage théorique. La théorie économique est construite sur le principe de la préférence révélée. Les enquêtes qui suscitent des opinions et des réponses à des questions hypothétiques ne sont pas aussi solidement fondées que les actions observées des agents. Les agents révèlent leurs préférences par leurs décisions. De la même manière, dans les conversations, les agents révèlent leurs émotions et leurs attitudes. Deuxièmement, elle peut être réalisée en temps réel plutôt qu’avec les décalages qu’impliquent les méthodes d’enquête classiques. Troisièmement, elle est beaucoup moins coûteuse à construire que les mesures basées sur les techniques d’enquête conventionnelles.

Il y a aussi des pièges à éviter qui, souvent, sont de la même nature que ceux des méthodes traditionnelles, mais qui peuvent aussi être d’une tout autre nature. Par exemple, il y a les biais d’échantillonnage. Il y a également les phénomènes d’amplification créés par les robots. Il est, en effet, possible pour une organisation de programmer un robot qui va émettre des gazouillis ou relancer des conversations sur des sujets qui intéressent cette organisation. Dans l’analyse des conversations sur les réseaux sociaux, il faut donc faire particulièrement attention à la collecte des données et au traitement des biais dans ces données. Il faut être particulièrement attentif aux phénomènes d’amplification créés artificiellement par des algorithmes.

Les premières analyses des conversations sur les réseaux sociaux ont été faites lors de campagnes électorales. Il est intéressant de comprendre si une thématique – jugée importante et prioritaire pour un pays – est bien comprise par la population. Lors des élections générales de 2015 au Canada, de nombreux messages ont été échangés sur Twitter. En utilisant une approche ordonnée d’un ensemble de données massives collectées sur Twitter (3,5 millions de gazouillis), Sanger et Warin (2017) ont développé deux méthodologies pour caractériser la façon dont les candidats et les différents partis ont été perçus sur les réseaux sociaux pendant la campagne électorale et notamment lors des débats télévisés. Tout d’abord, une analyse des sentiments a été réalisée pour chaque leader politique, puis, en utilisant l’ensemble des données, différents sujets électoraux ont été associés à chaque leader. Les conversations ont été analysées lors de deux débats en utilisant les mots-clics suivants : #polcan #Polcan2015, #Elxn42 et #elxn2015.

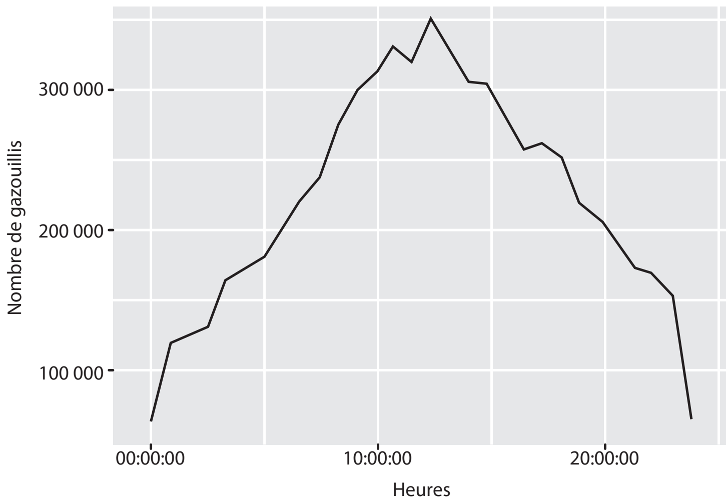
Pour chacun des débats, plus de 100 000 messages ont été collectés, avec des pics à 1 000 gazouillis par minute pour certaines parties du débat. Des thématiques ont été abordées par les utilisateurs de Twitter et nous pouvons déterminer si certains sujets électoraux provoquent plus de réactions que d’autres, et surtout si ces réactions sont restreintes dans le temps ou perdurent après la période accordée à ces sujets. Lors du débat des chefs en anglais, deux pics de messages correspondent à des mots-clics sur l’environnement (De Marcellis-Warin et Warin, 2017). En analysant un peu plus le contexte de ce débat et en essayant de trouver la raison des deux pics, nous avons remarqué qu’une campagne avait été préparée par l’organisme 350.org Canada[8] avant le débat pour inciter les utilisateurs de Twitter à communiquer un message spécifique pendant le débat. Cet organisme a fourni un exemple de gazouillis : « La science du climat n’est pas un débat. Les actions en faveur du climat signifient une économie propre et des sables bitumineux gelés. #ClimatELXN #globedebate[9] ». Plus de 860 000 personnes ont eu connaissance de cette campagne, toutes faisant partie des réseaux des 1 138 personnes ayant répondu à la demande de 350.org Canada.

Lors des élections fédérales de 2019, un groupe de chercheurs canadiens dirigé par le professeur Jean-François Savard de l’ENAP a analysé spécifiquement l’attention portée aux enjeux autochtones dans Twitter par les candidats au cours de la campagne[10]. Les mots « *indigenous* » et « autochtones » ont toujours conservé une place importante dans les discussions. Les gazouillis en français concernant les Premières Nations ont mis l’accent sur les enjeux climatiques, la réconciliation et les enjeux liés aux enfants à prendre en compte par le futur gouvernement.

D’une façon plus générale, l’émergence des plateformes de médias sociaux a permis aux citoyens ordinaires d’exprimer leurs inclinations idéologiques en adoptant le lexique des élites politiques. Les chercheurs disposent ainsi d’une nouvelle source de données très riche pour l’étude de l’idéologie politique (Temporão, Kerckhove, van der Linden, Dufresne et Hendrickx, 2018). Ainsi, que ce soit le contenu des conversations ou les moments des pics de messages, Twitter permet de bien comprendre les perceptions des citoyens et les dynamiques sociales. Lors du printemps 2012 et du conflit autour de la hausse des droits de scolarité au Québec, divers observateurs ont suggéré que les plateformes du Web 2.0, notamment Facebook et Twitter, auraient joué un rôle de premier plan dans cette crise (Latø -Toth, Pastinelli et Gallant, 2017). Toutefois, en analysant les traces d'activité archivées dans l'historique du compte Facebook de plusieurs jeunes, Latø -Toth et ses collaborateurs (2017) ont montré qu'en dépit de certains traits récurrents, les usages et les représentations de la plateforme sont loin d'être homogènes au sein de ce groupe d'âge et qu'ils reflètent les divers clivages sociaux.

Entre le 1er décembre 2019 et le 1er juin 2020, De Marcellis-Warin et Warin (2020) ont recueilli près de 6,5 millions de gazouillis liés aux coronavirus en utilisant un ensemble de termes de recherche prédéfinis (« COVID-19 », « coronavirus » ou « 2019-nCoV »)[11]. Six langues ont été sélectionnées pour la collecte des gazouillis : anglais, espagnol, chinois, allemand, français et italien. Les rediffusions de gaø uillis ont été repérées et supprimées de l'analyse, ainsi que la ponctuation, les mentions des utilisateurs de Twitter (@nom d'utilisateur), les chiffres, les liens HTML, les liens vers des photos et les petits mots tels que « *an* » et « *the* » (aussi appelés mots-stops)[12]. Les mots clés utilisés pour télécharger les gaø uillis ont également été supprimés afin de mieux mettre en évidence les mots les plus récurrents. En outre, différentes formes d'un même mot (par exemple « voyages », « *traveling* » et « *travel's* ») ont été converties en un seul mot principal (par exemple « *travel* ») à l'aide du logiciel en langage R intitulé « udpipe » (*package R udpipe*). Une analyse exploratoire des données des gazouillis postés en fonction de la date, de l'heure, du pays, de la langue et d'autres paramètres a été effectuée. On observe notamment que les pics de publication de messages étaient autour de 11 h et de 12 h, ce qui correspond, pour certains pays, aux heures des conférences de presse des gouvernements (graphique 11-2).

Nombre total de gazouillis par heure concernant le coronavirus entre le 1er décembre 2019 et le 1er juin 2020



Graphique g/2020-c**11-2**

Source : De Marcellis-Warin et Warin, 2020.

L'analyse des mots clés a aussi révélé une vue d'ensemble de l'appréhension de la pandémie dans le monde. Le nuage de mots (qui ne conserve pas les accents) a été représenté à la figure 11-2 à l'aide de la plateforme Nüance-R (Warin, 2019).

Mots clés les plus utilisés dans les gazouillis en français qui portaient sur le coronavirus entre le 1er décembre 2019 et le 1er juin 2020

Figure f/2020-c**11-2**

Source : De Marcellis-Warin et Warin, 2020.

Données issues de textes

Les chercheurs de différentes disciplines, notamment les politologues, les économistes et les chercheurs en marketing, utilisent de plus en plus les textes comme des données. Gentzkow et ses collaborateurs (2019) donnent un aperçu des applications en sciences sociales.

L'utilisation d'algorithmes qui permet l'analyse automatisée de textes réduit considérablement les coûts d'analyse des grands ensembles de textes. Cette évolution rapide est en partie due à l'évolution du stockage et de la diffusion des documents dans les bases de données textuelles électroniques. Ces méthodes textuelles automatisées deviennent un outil standard pour les politologues et les économistes (Grimmer et Stewart, 2013). Ces techniques permettent d'extraire et de recréer de l'information à partir d'un corpus de textes (classification, analyse, tendance, etc.).

Par exemple, ces méthodes ont permis d'analyser l'évolution de la communication de la Banque centrale européenne (BCE) à travers le temps, en considérant ses trois présidents successifs (W. Duisenberg, J. C. Trichet et M. Draghi) et la période précédant et suivant la crise financière de 2008. Warin et Sanger (2020) ont analysé les discours des banquiers centraux en utilisant des techniques de pointe d'apprentissage automatique et des analyses de sentiments afin de mesurer la polarité des messages afin de déterminer s'ils étaient plutôt positifs ou plutôt négatifs. Ces analyses ont permis de saisir l'évolution de la compréhension de la situation économique réelle par la BCE et de mesurer le niveau de stress à la BCE dans le temps. Un autre exemple d'analyse textuelle, pour des discours de l'Assemblée générale des Nations Unies, a permis d'identifier des enjeux de politique internationale ainsi que les positions adoptées par différents pays (Watanabe et Zhou, 2020).

L'analyse des données textuelles permet également de mieux comprendre comment les lois et les règlements sont appliqués au fil du temps. Par exemple, les données de transcription d'assemblées délibératives en Inde ont mis au jour les inégalités au sein des assemblées de villages ; notamment, les femmes, qui sont moins entendues et qui ont moins d'occasions de déterminer les sujets de discussions que les hommes, sont pénalisées (Parthasarathy, Rao et Palaniswami, 2019). Les jugements des cours chinoises ont aussi été numérisés. Grâce à l'analyse de ces documents, on peut voir comment les citoyens utilisent le système légal, et comment certaines poursuites permettent de contester certaines décisions de l'État (Liebman, Roberts, Stern et Wang, 2017). Les forums internationaux rassemblent des données améliorant notre compréhension des enjeux mondiaux.

L'utilisation d'algorithmes a aussi permis d'analyser le contenu de tous les résumés des brevets disponibles des 40 autorités mondiales de délivrance des brevets et des publications scientifiques en intelligence artificielle (IA) (Warin, Le Duc et Sanger, 2017). Cela a permis de voir l'évolution du nombre de brevets par année et par zone géographique, ainsi que d'extraire les catégories décrivant le mieux chaque sous-domaine de l'IA en les classant par secteur industriel. La base de données utilisée a rassemblé un total de 55 109 brevets et 29 225 articles scientifiques. Dans les deux cas, l'analyse d'une telle quantité d'informations ne serait pas possible sans une puissance de calcul dédiée. L'objectif était d'avoir une perspective et une compréhension plus approfondies des différents développements de l'IA dans le temps et l'espace. L'analyse géographique a fourni un portrait des principales régions et des principaux pays contribuant à l'innovation en matière d'IA et d'apprentissage automatique : la Chine, les États-Unis et le Japon.

Nouvelles méthodologies d'analyse

Les données non structurées ont à l'origine plutôt une nature qualitative. Pourtant, elles sont codées en format numérique pour pouvoir être utilisées sur les appareils électroniques (téléphones intelligents, tablettes, ordinateurs, etc.). Le fait que des informations qualitatives soient traduites en format numérique permet de les analyser de façon quantitative. Il faut pour cela structurer ces données. Un exemple plus complexe est celui de l'analyse sémantique : il s'agit par exemple de permettre la conversation dans une langue entre deux personnes et d'être capable d'analyser si cette conversation est de nature positive, négative, joyeuse, agressive, etc. Des modèles d'analyse des données quantitatives existent déjà depuis les premiers travaux sur les probabilités et l'analyse stochastique et n'ont eu de cesse d'être améliorés à travers le temps. Des outils d'analyse de très grosses bases de données avec une puissante capacité de calculs ont été développés.

De plus, pour aller explorer et chercher des données, les méthodes de fouille de données sont un point de départ évident, mais il faut aussi tenir compte des spécificités des données (Gama, Sebastião et Rodrigues, 2009). Les méthodes d'apprentissage automatique sont un autre point de départ évident. Très populaires dans les années 1970, elles ont été aussi améliorées par les modèles d'économétrie (Choi et Varian, 2012). La disponibilité d'une puissance de calcul importante avec le développement de nouveaux ordinateurs avait permis ce développement.

Gentzkow et ses collaborateurs (2019) donnent un aperçu des méthodes d'analyse des textes et des applications actuelles en sciences sociales et, plus spécifiquement, en économie. Considérant la nature des données et leur disponibilité massive, les méthodes issues de l'apprentissage automatique trouvent des domaines d'application intéressants ; par exemple, les arbres de régressions ou les analyses factorielles sont parfois plus efficaces que les modèles d'économétrie de première génération. Les modèles d'économétrie de deuxième génération avec des estimations non linéaires s'ajoutent à la panoplie des outils à la disposition des analystes de données massives. Un nombre croissant d'études ont appliqué les modèles d'apprentissage automatique à la prévision macroéconomique. Nakamura (2005) a montré l'utilité des réseaux de neurones pour la prévision de l'inflation. Les réseaux neuronaux surpassent les modèles autorégressifs univariés en moyenne pour des horizons courts d'un et de deux trimestres.

Néanmoins, pour rendre analysables de telles données, il faut développer des méthodes de structuration. En effet, il ne s'agit pas seulement d'utiliser une série de techniques économétriques ou d'algorithmes provenant des sciences de l'information, mais il faut aussi s'assurer que les meilleures méthodes des différents champs disciplinaires intéressés sont connues et utilisées pour analyser les données massives avec toutes leurs spécificités.

Analyse de la géolocalisation des messages et des textes

La géolocalisation des messages et des textes peut offrir un outil précieux d'analyse. En complément des recherches sur les moteurs de recherche que nous avons présentées précédemment, une étude a montré que l'utilisation de messages publiés sur Twitter a déjà permis d'améliorer le suivi en temps réel de l'évolution géographique de l'épidémie de grippe, notamment en établissant la propagation de la grippe et les foyers d'éclosion de neuf États américains (Reynard et Shirgaokar, 2019). Utiliser ces données localisées permet aux autorités publiques de réagir de façon plus ciblée.

Dans l'éventualité d'un désastre naturel ou d'une crise grave, il est souvent possible d'établir la position géographique de l'auteur d'un message émis sur les médias sociaux, que ce soit par une donnée GPS, par l'identification d'un point de repère ou par l'analyse du texte du message. Les messages relayés sur les réseaux sociaux permettent à la fois d'informer la population de l'état de la situation et de recueillir de l'information sur l'état du terrain. Les autorités publiques sont alors en mesure de recevoir, de la part des citoyens, des images ou des descriptions décrivant toute situation difficile (inondation, arbre tombé, panne électrique, etc.) (Reynard et Shirgaokar, 2019).

Analyse des réseaux

Des analyses de réseaux basées sur des données issues de réseaux sociaux peuvent permettre de repérer des communautés d'usagers au sein de Twitter et de les articuler avec les principaux thèmes mobilisés dans le cadre d'un débat. Par exemple, Smyrnaios et Ratinaud (2014) montrent comment articuler une analyse des réseaux avec une analyse des discours sur Twitter dans le contexte de la ratification du Traité du pacte budgétaire européen à l'automne 2012 en se basant sur la couverture française. Leurs résultats montrent que les conversations mêlent des messages informatifs avec une gamme d'expressions diverses (ironie, critique, humour, indignation, etc.). Les échanges sont dominés par ceux qui adoptent une position de forte opposition au traité, couplée, souvent, à des discours de dénonciation qui vont influencer les positions des internautes.

L'évaluation de l'influence sur les réseaux sociaux est au cœur d'une nouvelle tendance dans les sciences humaines et sociales. Twitter est devenu un espace privilégié pour le secteur de la finance. Les opérateurs de marché (*traders*) dits « 2.0 » s'informent sur Twitter. De Marcellis-Warin, Sanger et Warin (2017) ont évalué l'influence de certains utilisateurs concernant des conversations financières sur Twitter. Ces conversations sont très spécifiques, car les utilisateurs utilisent des mots-clics formés avec les noms des compagnies. Pour cette étude, les chercheurs ont constitué deux ensembles de données comprenant respectivement 489 000 et 280 000 gazouillis financiers. Si le fait d'obtenir des adeptes supplémentaires peut donner un aperçu d'une popularité croissante, cela ne se traduit pas nécessairement par une position plus influente. Les résultats de l'étude suggèrent que le nombre de suiveurs n'est qu'un élément de ce qui est considéré comme influent. En effet, le nombre de messages peut être biaisé par ce que l'on peut qualifier d'utilisateurs « bruyants ». Le nombre de rediffusions de gazouillis est un indicateur plus précis de l'influence. Enfin, par exemple, les mesures de centralité dans le réseau de rediffusions des gazouillis fournissent des informations privilégiées à toute personne qui suit ces utilisateurs. Il est aussi possible de comparer les impacts des actions mentionnées par chaque type d'utilisateur. À partir de ces résultats, de telles méthodes d'évaluation de l'influence sur Twitter permettent d'acquérir des signaux privilégiés dans un contexte financier.

En utilisant les mêmes méthodologies d'analyse de réseaux, Warin et Sanger (2018) se sont concentrés sur la connectivité et la proximité entre les institutions financières. Ils ont examiné les réseaux potentiels entre les institutions financières par le biais de leurs conseils d'administration. Pour cela, ils ont constitué un large échantillon de représentants de conseils d'administration (43 399 personnes) pour 2 209 institutions dans 52 pays en utilisant la base de données du Bureau van Dijk. L'article identifie des regroupements de conseils d'administration montrant – dans une certaine mesure – le niveau de concentration au sein du système financier dans des régions du monde en particulier. La principale contribution de cet article est de mettre en évidence des propriétés précises du système financier international qui pourraient être critiques, notamment en ce qui a trait au risque systémique.

Analyse de sentiments

L'analyse de sentiments vise à déterminer les sentiments des gens en analysant leurs messages et différentes actions qu'ils posent sur les réseaux sociaux. Elle consiste à classer les messages selon les différents sentiments opposés qu'ils transmettent, qui peuvent être positifs ou négatifs. Cette forme d'analyse pourrait être divisée en deux catégories principales (Warin et Sanger, 2018) : l'analyse lexicale, qui vise à calculer la polarité dans un document à partir de l'orientation sémantique des mots ou des phrases qu'il contient, et l'apprentissage automatique, qui vise à construire des modèles à partir des données d'apprentissage étiquetées (instances de textes ou de phrases) afin de déterminer l'orientation d'un document.

Bollen, Pepe et Mao (2009) ont utilisé des analyses de sentiments sur les messages publiés sur Twitter. Pendant toute la durée de leur étude, ils ont extrait des messages sous six dimensions de sentiments : tension, dépression, colère, vigueur, fatigue et confusion. Ils ont ensuite comparé leurs résultats aux fluctuations des marchés financiers, au prix du pétrole et à des perturbations ponctuelles créées par des situations exceptionnelles comme l'élection présidentielle américaine du 4 novembre 2008. Ils ont constaté que les événements dans les domaines sociaux, politiques, culturels et économiques ont un effet très significatif sur les six dimensions d'humeur de la population. Les auteurs suggèrent que l'information présente dans les réseaux sociaux peut être utilisée comme indicateur pour les opinions publiques.

Les firmes utilisent aussi l'analyse de sentiments pour évaluer l'image des marques qu'elles distribuent (Culotta et Cutler, 2016). En effet, elles considèrent que ces approches sont moins coûteuses que les méthodes traditionnelles (par exemple, les enquêtes), dont les résultats, de surcroît, peuvent rapidement devenir obsolètes. Ce type d'analyse peut être utilisée pour mesurer l'image associée à certaines politiques publiques, prévenant (ou à tout le moins prévoyant) ainsi des protestations vives qui pourraient venir de citoyens en fort désaccord avec une politique spécifique. Cela peut être très utile aux organismes gouvernementaux pour comprendre les besoins et les problèmes des sociétés et formuler des politiques publiques efficaces pour y répondre (Charalabidis, Maragoudakis et Loukis, 2015).

Modèles prédictifs

Les données des réseaux sociaux peuvent améliorer les modèles prédictifs. Ces derniers peuvent faire en sorte, par exemple, de comprendre les sentiments des investisseurs. Étant donné que les variations à court terme de la valeur des actions dépendent à la fois d'éléments fondamentaux et rationnels que d'éléments moins tangibles, dont la peur ou la confiance dans les marchés, cette analyse permet d'intégrer les sentiments des investisseurs et d'améliorer les modèles prédictifs (Attigeri, Manohara Pai, Pai et Nayak, 2015 ; Wang et Wang, 2016).

Les données publiées sur Twitter peuvent également être utilisées pour prédire les mouvements de foule. En effet, comme on l'a vu précédemment, les publications de gazouillis sont souvent géolocalisées. Leur analyse peut aider à anticiper, par exemple, les pics d'utilisation des transports (Ni, He et Gao, 2016). Les réseaux sociaux permettent également la prévision de mouvements à plus long terme. Grâce à de telles données, des études ont pu prévoir à l'avance les fluctuations du nombre de touristes sur une certaine période (Xin, Bing, Law et XianKai, 2017).

Stevanovic (2020) montre que l'accessibilité des données massives et l'avancement des techniques d'apprentissage automatique ont considérablement changé la façon d'approcher le problème de prévision macroéconomique. De nombreux chercheurs travaillent à améliorer les techniques utilisées (Coulombe, Leroux, Stevanovic et Surprenant, 2020).

Les enjeux à prendre en compte

Ce chapitre n'avait pas comme objectif de faire une revue exhaustive de la littérature sur le sujet. Toutefois, les exemples présentés montrent bien le potentiel des données non structurées et des nouvelles méthodes d'analyse, notamment l'apprentissage automatique. Cela va permettre entre autres de mieux éclairer les décisions publiques et, surtout, d'obtenir de l'information sur une base continue et de façon moins coûteuse. Toutefois, certains enjeux sont parfois soulevés quant à la nature des données, à la véracité des données ainsi qu'aux biais possibles dans les échantillons de données. Nous pourrions également ajouter une analyse des stratégies et des modèles d'affaires des plateformes de réseaux sociaux. Dans l'espace limité de ce chapitre, nous allons aborder les enjeux reliés aux biais dans les données. Il est important de les définir dès le début de l'analyse pour pouvoir trouver des solutions.

La nature et la qualité des données

Que ce soit pour les données structurées ou pour les données non structurées, la qualité des données reste toujours primordiale. Selon Statistique Canada, on mesure la qualité des données par le niveau (ou degré) de confiance à leur égard ainsi que par la capacité d'utilisation (la pertinence, l'exactitude, l'actualité, l'accessibilité, la possibilité d'interprétation et la cohérence)[13]. Pour les données relayées sur les réseaux sociaux, la qualité doit être évaluée en fonction de la provenance. On peut parfois avoir l'impression qu'une opinion issue de Twitter est fortement partagée, alors qu'elle provient d'un nombre limité de personnes, qui sont toutefois positionnées de manière stratégique dans le réseau, influençant ainsi fortement la conversation (Jungherr et Jürgens, 2014).

De plus, l'argument souvent apporté est que les utilisateurs de Twitter ne forment probablement pas un échantillon représentatif de la population, mais qu'ils sont plutôt une population en soi. L'avantage de cette situation est que l'on peut extraire tous les gazouillis sur un sujet, ce qui veut dire qu'avec suffisamment de puissance de calcul, il est possible d'utiliser toute la population sans faire d'échantillonnage. La question est ensuite de savoir si l'on peut néanmoins généraliser à l'ensemble de la population, incluant les conversations hors Twitter, en fonction de ce que l'on trouve dans Twitter. La littérature est partagée sur le sujet. Une des dynamiques sur les réseaux sociaux en général est la polarisation. Il n'y a pas un seul village mondial sur Internet, mais des villages de gens qui se sont regroupés par affinités culturelles, thématiques, démographiques, etc. Il y a donc un biais d'échantillon si l'on cherche à en tirer une quelconque généralisation. En revanche, il est intéressant d'analyser une conversation en particulier et de chercher les variables de contrôle de nature culturelle, thématique, démographique, etc. C'est à ce stade qu'il est possible de faire ressortir quelques lois générales sur la population au sens large du terme (De Marcellis-Warin et Warin, 2017).

La véracité des données : les « fausses » informations

Les réseaux sociaux servent à transmettre toutes sortes d'informations, vraies ou fausses. De fausses données ont été relayées dans toutes sortes de domaines comme les vaccins, la nutrition ou les informations boursières (Lazer *et al.,* 2018). Les fausses informations se répandent facilement sur le Web, pouvant devenir virales et influencer l'opinion publique et ses décisions (Bondielli et Marcelloni, 2019). Ces fausses nouvelles sont majoritairement relayées par des robots de recherche (*bots*) qui tentent de répandre largement ces faussetés (Al-Rawi, Groshek et Zhang, 2019). Dans certains cas, elles font leur chemin jusqu'aux médias traditionnels, qui vont alors jusqu'à les relayer. De plus, on remarque un risque d'écho. Une grande partie des personnes suivant les publications des journalistes sont d'autres journalistes (Bane, 2019). Il est donc possible que le même message soit répété à partir d'une seule source, simplement parce que les journalistes s'alimentent aux mêmes sources.

En parallèle à cette production de fausses nouvelles, l'étude de Grinberg, Joseph, Friedland, Swire-Thompson et Lazer (2019) suggère que la proportion de fausses nouvelles à laquelle un citoyen est exposé reste faible (même si elle est réelle). En utilisant les données de la campagne américaine de 2016, ces auteurs ont estimé que 6 % des utilisateurs relayant de l'information politique ont partagé du contenu faux. Cela aurait représenté un peu plus de 1 % du contenu consulté (Grinberg *et al.,* 2019). Il ne faut toutefois pas sous-estimer le problème. En réalité, les zones d'ombre ou de demi-vérité sont peut-être plus importantes que les fausses nouvelles.

Le défi de la véracité des données n'est pas seulement de savoir reconnaître les fausses nouvelles. Il s'agit aussi d'authentifier les vraies nouvelles. Plusieurs médias se sont vus accusés de publier de « fausses nouvelles » quand celles-ci ne plaisaient pas aux politiciens en place, et ce, même quand les faits publiés étaient avérés. Les médias sociaux servent alors à discréditer des sources « fiables » d'information qui doivent alors défendre leur légitimité (Lischka, 2019). Les recherches suggèrent par exemple qu'entre 250 000 et 300 000 membres du parti chinois alimentent les réseaux sociaux de messages choisis par le gouvernement en utilisant différents pseudonymes afin de créer l'illusion que ces messages sont légitimes (King, Pan et Roberts, 2017).

Par ailleurs, cet enjeu de l'authenticité est difficile à évaluer, puisqu'il n'existe pas de critère établi pour la détection de vraies ou de fausses nouvelles. L'intérêt pour les techniques de détection efficaces a donc augmenté très rapidement ces dernières années (Bondielli et Marcelloni, 2019).

L'endogénéité des données non structurées et les biais historiques

Certaines sources de données sont vulnérables à la manipulation. Par exemple, on estime que la moitié des messages transmis sur Twitter sont générés par des robots (Al-Rawi *et al.,* 2019). Ainsi dans certains cas, la manipulation des données pourrait venir « polluer » les algorithmes d'apprentissage automatique.

Les approches fondées sur les données mettent souvent à contribution l'intelligence artificielle. Ces algorithmes utilisent les données historiques afin de prédire le futur. Ils travaillent généralement en « boîte noire », une méthode ne permettant pas d'avoir une vue explicite des règles qui sont extraites de l'analyse des faits passés. Ce mode de fonctionnement présente des risques élevés de reproduction de biais historiques. Par exemple, un système utilisé par les cours de justice américaines pour prédire les risques de récidives prévoyait (erronément) un risque deux fois supérieur de récidive pour les prisonniers de race noire comparé aux prisonniers de race blanche (Buranyi, 2017). Les biais peuvent venir autant d'une représentation biaisée du problème à l'étude que de biais incrustés dans les données historiques (Roselli, Matthews et Talagala, 2019). Dans ce contexte, baser des politiques publiques sur de simples corrélations est très risqué. Il existe des pistes de solution pour limiter ces biais. On peut penser notamment à l'utilisation d'une tierce partie pour les évaluer, au développement d'un indice d'équité, ou à l'exploration des données à la recherche de ces modèles biaisés (Veale et Binns, 2017). Malheureusement, aucune solution n'est simple ou facile à automatiser. Bien que des techniques informatiques émergent pour répondre à certaines de ces préoccupations grâce à l'exploration de données en connaissance de cause ou encore à l'apprentissage machine de l'équité, de la responsabilité et de la transparence, la mise en œuvre pratique des solutions se heurte à des défis réels. Il faut donc procéder prudemment si ces systèmes sont utilisés de manière non supervisée.

Conclusion

Les réseaux sociaux peuvent servir de courroies de transmission pour les groupes qui veulent s'opposer à certaines politiques publiques, ou au contraire inciter un gouvernement à agir dans certaines circonstances (Larson, Nagler, Ronen et Tucker, 2019). Les réseaux sociaux deviennent également des outils pour recevoir des commentaires sur des éléments de politiques. Quand les sujets intéressent les citoyens, ceux-ci vont souvent commenter les rapports ou les actions gouvernementales publiquement, en utilisant les réseaux sociaux. Par exemple, sur l'environnement et les changements climatiques, la publication d'un rapport peut générer des conversations sur le contenu du rapport de même que sur les réponses du gouvernement par rapport aux enjeux soulevés dans le rapport (Pearce, Holmberg, Hellsten et Nerlich, 2014).

L'utilisation des médias sociaux n'augmente pas automatiquement la confiance que les citoyens ont envers leur gouvernement. Quand elle accroît la perception de transparence du gouvernement, à partir de ce moment seulement, il y a accroissement de la confiance (Song et Lee, 2016).

Une étude des communications de 75 gouvernements de proximité a montré que la communication se fait maintenant dans les deux sens (Bonsón, Torres, Royo et Flores, 2012). Les citoyens utilisent les médias pour informer les gouvernements, que ce soit à propos de leurs opinions, de leurs observations, de leurs activités quotidiennes ou même de leurs souhaits. Ils peuvent transmettre des commentaires, des photos ou des vidéos, et ils deviennent des collaborateurs actifs plutôt que simplement les utilisateurs d'un service. On remarque aussi que les médias traditionnels se fient aux comptes Twitter associés aux organisations publiques, et qu'ils les traitent maintenant comme des sources officielles (Moon et Hadley, 2014). Les politiciens font également partie du paysage des réseaux sociaux. Ils commentent les choix de politiques publiques et alimentent le débat. La recherche montre que les politiciens qui prônent des positions extrêmes sur le plan idéologique tendent à avoir plus d'abonnés que ceux qui sont plus modérés (Hong et Kim, 2016). Ceci rend plus difficiles l'atteinte de compromis et la discussion constructive.

Références

Alaoui, I. E. (2018). *Transformer les big social data en prévisions. Méthodes et technologies : application à l'analyse de sentiments* [thèse de doctorat en sciences de l'ingénieur, Université d'Angers; Université IbnTofail. Faculté des sciences de Kénitra]. https://tel.archives-ouvertes.fr/tel-02060594/document.

Al-Rawi, A., Groshek, J. et Zhang, L. (2019). What the fake? Assessing the extent of networked political spamming and bots in the propagation of #fakenews on Twitter. *Online Information Review, 43*(1), 53-71. https://doi.org/10.1108/OIR-02-2018-0065.

Attigeri, G. V., Manohara Pai, M. M., Pai, R. M. et Nayak, A. (2015). Stock market prediction: A big data approach. *TENCON 2015-2015 IEEE Region 10 Conference*. https://doi.org/10.1109/TENCON.2015.7373006 .

Bane, K. C. (2019). Tweeting the agenda. *Journalism Practice, 13*(2), 191-205. https://doi.org/10.1080/17512786.2017.1413587.

Bollen, J., Pepe, A. et Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011).* file:///C:/Users/vedes/AppData/Local/Temp/0911.1583.pd.

Bondielli, A. et Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences, 497,* 38-55. https://doi.org/10.1016/j.ins.2019.05.035.

Bonsón, E., Torres, L., Royo, S. et Flores, F. (2012). Local e-government 2.0: Social media and corporate transparency in municipalities. *Government Information Quarterly, 2*(29), 123-132. https://doi.org/10.1016/j.giq.2011.10.001.

Buranyi, S. (2017, 8 août). Rise of the racist robots – how AI is learning all our worst impulses. *The Guardian.* https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

Charalabidis, Y., Maragoudakis, M. et Loukis, E. (2015). Opinion mining and sentiment analysis in policy formulation initiatives: The EU-community approach. Dans E. Tambouris, P. Panagiotopoulos, Ø. Sæbø, K. Tarabanis, M. A. Wimmer, M. Milano et T. Pardo (dir.), *Electronic Participation* (p. 147-160). Springer International Publishing. https://doi.org/10.1007/978-3-319-22500-5_12.

Choi, H. et Varian, H. (2012). Predicting the present with Google trends. *Economic Record, 88*(s1), 2-9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Coulombe, P. G., Leroux, M., Stevanovic, D. et Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? *CIRANO Scientific Series, 2019s*(22). https://cirano.qc.ca/fr/sommaires/2019s-22.

Culotta, A. et Cutler, J. (2016). Mining brand perceptions from Twitter social networks. *Marketing Science, 35*(3), 343-362. https://doi.org/10.1287/mksc.2015.0968.

De Marcellis-Warin, N., Sanger, W. et Warin, T. (2017). A network analysis of financial conversations on Twitter. *International Journal of Web Based Communities, 13*(3). https://doi.org/10.1504/IJWBC.2017.086588.

De Marcellis-Warin, N. et Warin, T. (2017). Les sciences des données et la perception des risques naturels et climatiques : Analyse des conversations sur Twitter. Dans Bernard Motulsky, Jean Bernard Guindon et Flore Tanguay-Hébert (dir.), *Communication des risques météorologiques et climatiques.* Presses de l'Université du Québec.

De Marcellis-Warin, N. et Warin, T. (2020). *Twitter Analysis on Covid-19 Pandemic.* https://warin.ca/posts/article-tweets-covid/.

Eisenstein, M. (2018). Cloudy with a chance of flu. *Nature, 555*(7695). doi : 10.1038/d41586-018-02473-5.

Gama, J., Sebastião, R. et Rodrigues, P. P. (2009). Issues in evaluation of stream learning algorithms. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 329-338. https://doi.org/10.1145/1557019.1557060.

Gentzkow, M., Kelly, B. et Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535-574. https://doi.org/10.1257/jel.20181020.

Grimmer, J. et Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297. https://doi.org/10.1093/pan/mps028.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. et Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science (New York, N.Y.), 363*(6425), 374-378. https://doi.org/10.1126/science.aau2706.

Hong, S. et Kim, S. H. (2016). Political polarization on Twitter: Implications for the use of social media in digital governments. *Government Information Quarterly, 33*(4), 777-782. https://doi.org/10.1016/j.giq.2016.04.007.

Iansiti, M. et Lakhani, K. R. (2020, 1 janvier). Competing in the Age of AI. *Harvard Business Review, January-February 2020.* https://hbr.org/2020/01/competing-in-the-age-of-ai.

Jungherr, A. et Jürgens, P. (2014). Stuttgart's Black Thursday on Twitter: Mapping political protests with social media data. Dans M. Cantijoch, R. Gibson et S. Ward (dir.), *Analyzing Social Media Data and Web Networks* (p. 154-196). Palgrave Macmillan UK. https://doi.org/10.1057/9781137276773_7.

King, G., Pan, J. et Roberts, M. E. (2017). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review, 111*(3), 484-501.

Larson, J. M., Nagler, J., Ronen, J. et Tucker, J. A. (2019). Social networks and protest participation: Evidence from 130 million Twitter users. *American Journal of Political Science, 63*(3), 690-705. https://doi.org/10.1111/ajps.12436.

Latzo-Toth, G., Pastinelli, M. et Gallant, N. (2017). Usages des médias sociaux et pratiques informationnelles des jeunes Québécois : le cas de Facebook pendant la grève étudiante de 2012. *Recherches sociographiques, 58*(1), 43-64. https://doi.org/10.7202/1039930ar.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J. et Zittrain, J. L. (2018). The science of fake news. *Science (New York, N.Y.), 359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998.

Liebman, B. L., Roberts, M. E., Stern, R. E. et Wang, A. Z. (2017). *Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law*. https://doi.org/10.2139/SSRN.2985861.

Lischka, J. A. (2019). A badge of honor? *Journalism Studies, 20*(2), 287-304. https://doi.org/10.1080/1461670X.2017.1375385.

Monkman, G. G., Kaiser, M. J. et Hyder, K. (2018). Text and data mining of social media to map wildlife recreation activity. *Biological Conservation, 228,* 89-99. https://doi.org/10.1016/j.biocon.2018.10.010.

Moon, S. J. et Hadley, P. (2014). Routinizing a new technology in the newsroom: Twitter as a news source in mainstream media. *Journal of Broadcasting & Electronic Media, 58*(2), 289-305. https://doi.org/10.1080/08838151.2014.906435.

Nakamura, E. (2005). Inflation forecasting using a neural network. *Economics Letters, 86*(3), 373-378. https://doi.org/10.1016/j.econlet.2004.09.003.

Ni, M., He, Q. et Gao, J. (2016). Forecasting the subway passenger flow under event occurrences with social media. *IEEE Transactions on Intelligent Transportation Systems,* 1-10. https://doi.org/10.1109/TITS.2016.2611644.

Nyman, R. et Ormerod, P. (2020). Text as Data: Real-time Measurement of Economic Welfare. *arXiv:2001.03401 [econ, q-fin]*. http://arxiv.org/abs/2001.03401.

Parthasarathy, R., Rao, V. et Palaniswamy, N. (2019). Deliberative democracy in an unequal world: A *text-as-data* study of South India's village assemblies. *American Political Science Review, 113*(3), 623-640. https://doi.org/10.1017/S0003055419000182.

Pearce, W., Holmberg, K., Hellsten, I. et Nerlich, B. (2014). Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC working group 1 report. *PLOS ONE, 9*(4), e94785. https://doi.org/10.1371/journal.pone.0094785.

Reynard, D. et Shirgaokar, M. (2019). Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster? *Transportation Research Part D: Transport and Environment, 77,* 449-463. https://doi.org/10.1016/j.trd.2019.03.002.

Roselli, D., Matthews, J. et Talagala, N. (2019). Managing bias in AI. *Companion Proceedings of The 2019 World Wide Web Conference,* 539-544. https://doi.org/10.1145/3308560.3317590.

Sanger, W. et Warin, T. (2017). The 2015 canadian election on Twitter: A tidy algorithmic analysis. *2017 International Conference on Computational Science and Computational Intelligence (CSCI),* 910-915. https://doi.org/10.1109/CSCI.2017.158.

Smyrnaios, N. et Ratinaud, P. (2014). Comment articuler analyse des réseaux et des discours sur Twitter. L'exemple du débat autour du pacte budgétaire européen. *TIC & Société, 7*(2). https://doi.org/10.4000/ticetsociete.1578.

Song, C. et Lee, J. (2016). Citizens' use of social media in government, perceived transparency, and trust in government. *Public Performance & Management Review, 39*(2), 430-453. https://doi.org/10.1080/15309576.2015.1108798.

Stevanovic, D. (2020). Prévision macroéconomique dans l'ère des données massives et de l'apprentissage automatique. Dans N. De Marcellis-Warin et B. Dostie (dir.), *Le Québec économique 9 : Opportunités et défis de la transformation numérique.* Montréal, Québec : CIRANO.

Temporão, M., Kerckhove, C. V., van der Linden, C., Dufresne, Y. et Hendrickx, J. M. (2018). Ideological scaling of social media users: A dynamic lexicon approach. *Political Analysis, 26*(4), 457-473. https://doi.org/10.1017/pan.2018.30.

Tkacz, G. (2013). *Predicting recessions in real-time: Mining Google trends and electronic payments data for clues.* C.D. Howe Institute.

Veale, M. et Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society.* https://doi.org/10.1177/2053951717743530.

Wang, Y. et Wang, Y. (2016). Using social media mining technology to assist in price prediction of stock market. *2016 IEEE International Conference on Big Data Analysis (ICBDA),* 1-4. https://doi.org/10.1109/ICBDA.2016.7509794.

Warin, T. (2019). Nüance-R: A Technological Platform for Data Science in Higher Education. Figshare. https://figshare.com/articles/software/N_ance-R/8870174 .

Warin, T. (2020). Global research on coronaviruses: An R package. *Journal of Medical Internet Research, 22*(8), e19615. https://doi.org/10.2196/19615 .

Warin, T. et De Marcellis-Warin, N. (2014). *Un état des lieux sur les données massives* (Rapport Bourgogne nº 2014-RB 01). CIRANO.

Warin, T., Le Duc, R. et Sanger, W. (2017). Mapping innovations in artificial intelligence through patents: A social data science perspective. *2017 International Conference on Computational Science and Computational Intelligence (CSCI),* 252-257. https://doi.org/10.1109/CSCI.2017.40 .

Warin, T. et Sanger, W. (2018). Connectivity and closeness among international financial institutions: A network theory perspective. *International Journal of Comparative Management, 1*(3). https://doi.org/10.1504/IJCM.2018.094479 .

Warin, T. et Sanger, W. (2020). The speeches of the European Central Bank's presidents: An NLP study. *Global Economy Journal,* 1-31. https://doi.org/10.1142/S2194565920500098 .

Watanabe, K. et Zhou, Y. (2020). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review,* 0894439320907027. https://doi.org/10.1177/0894439320907027 .

Wu, L. et Brynjolfsson, E. (2013). The future of prediction: How Google searches foreshadow housing prices and sales (SSRN Scholarly Paper ID 2022293). *Social Science Research Network.* https://doi.org/10.2139/ssrn.2022293 .

Xin, L., Bing, P., Law, R. et XianKai, H. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59,* 57-66.

Notes

1. Les résultats complets du sondage NETendance sont disponibles en ligne : https://cefrio.qc.ca/fr/enquetes-et-donnees/netendances2018-medias-sociaux/ .
2. Le site Internet du Baromètre CIRANO regroupe l'ensemble des données des sondages effectués auprès de la population du Québec sur la perception des risques : https://barometre.cirano.qc.ca/.
3. La plateforme statistique Google Trends a mis en place une page spéciale pour décrire les recherches faites sur le moteur de recherche Google concernant la COVID-19 : https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en (page consultée le 8 août 2020).
4. La plateforme statistique Google Trends permet de voir le nombre de recherches associées au mot clé « chômage », mais aussi les recherches par région, par sujet associé, ou par requête associée : https://trends.google.ca/trends/explore?geo=CA&q=chomage (page consultée le 8 août 2020).
5. Le graphique 11-1 a été généré en utilisant le mot clé « PCU » : https://trends.google.com/trends/explore?date=2020-03-13%202020-09-30&geo=CA&q=PCU (page consultée le 30 septembre 2020).
6. La Loi sur la prestation canadienne d'urgence a été adoptée le 25 mars 2020. La prestation canadienne d'urgence pour étudiants (PCUE) a été instaurée le 28 avril 2020. La première période s'est étendue du 15 mars au 11 avril 2020 et la deuxième du 12 avril au 9 mai 2020. Plusieurs autres périodes se sont échelonnées par la suite jusqu'à l'automne 2020.
7. La page suivante permet de voir les recherches avec le mot clé « CERB » : https://trends.google.com/trends/explore?date=2020-03-13%202020-09-30&geo=CA&q=CERB (page consultée le 30 septembre 2020).
8. Cet organisme se définit, sur son site Web, comme construisant « un mouvement mondial pour le climat ».
9. La phrase originale était en anglais : « *Climate science isn't a debate. Climate actions mean a clean economy & freeźng tar sands. #ClimatELXN #globedebate* ».
10. Les auteurs ont mis en place un site Internet avec toutes les données : http://pcmg.enap.ca/elections2019/francais/
11. Le nombre total exact est de 6 664 956 gazouillis. Pour plus d'informations, voir : https://www.warin.ca/posts/article-tweets-covid/ (page consultée le 31 juillet 2020).
12. Les données textuelles contiennent des éléments qui ne donnent pas d'informations et augmentent la complexité de l'analyse. Il est important de les éliminer avant de les analyser.
13. https://www150.statcan.gc.ca/n1/pub/12-539-x/2019001/ensuring-assurer-fra.htm.