L'analyse des opinions politiques sur Twitter
Défis et opportunités d'une approche multi-échelle

Marta Severo

EA Dicen – Université Paris Nanterre-IUF

msevero@parisnanterre.fr

Robin Lamarche-Perrin

ISC-PIF – LIP6 – CNRS

Robin.Lamarche-Perrin@lip6.fr

Remerciements

Cette article a été publié dans la *Revue française de sociologie*, 2018/3 (Vol. 59), p. 507-532. DOI 10.3917/rfs.593.0507

Ce travail a été partiellement financé par le programme H2020 FETPROACT 2016–2017 de la Commission Européenne sous le numéro de bourse 732942 (ODYCCEUS).

Résumé

Des blogs et forums aux pages Facebook et comptes Twitter, le récent déluge des données numériques du Web a fortement affecté la recherche en sciences sociales. Cette nouvelle catégorie d'information, utile à l'extraction des opinions politiques, se présente comme une alternative aux techniques traditionnelles telles que les sondages. Premièrement, en réalisant un état de l'art des études de l'opinion s'appuyant sur les données Twitter, cet article vise à mettre en relation les méthodes d'analyse utilisées dans ces études et les définitions de l'opinion politique qui y sont suggérées. Deuxièmement, cet article étudie la faisabilité de réaliser des analyses multi-échelles en sciences sociales concernant l'étude de l'opinion politique en exposant les mérites de plusieurs méthodes, allant des méthodes orientées contenus aux méthodes orientées interactions, de l'analyse statistique à l'analyse sémantique, des approches supervisées aux approches non-supervisées. Le résultat de notre démarche est d'ainsi identifier les tendances futures de la recherche en sciences sociales concernant l'étude de l'opinion politique.

1. Introduction

L'étude de l'opinion politique est un champ traditionnel de la recherche en sciences sociales. Des recherches sur les élections, les partis et les représentants politiques ont rempli les pages de nombre de revues en sciences sociales durant le siècle dernier. Elles reposent souvent sur des analyses empiriques, basées sur des méthodes traditionnelles comme, par exemple, sur des sondages d'opinion, des enquêtes, des groupes de discussions ou encore des entretiens. Récemment, le développement d'Internet a entraîné des changements importants dans ce domaine en offrant de nouvelles arènes d'expression pour les politiciens autant que pour les citoyens. Depuis le cas exemplaire de Barack Obama jusqu'à l'utilisation plus récente du big data dans le duel Trump-Clinton, Internet est devenu un outil central pour convaincre les électeurs, pour organiser les soutiens et pour diffuser les messages politiques. Si, il y a dix ans, les sites web, les blogs et les forums étaient les principaux espaces du débat politique, aujourd'hui Facebook, Twitter et YouTube sont désormais des lieux stratégiques pour exprimer et disséminer des opinions politiques, mais aussi pour débattre avec les candidats, entre amis ou avec des inconnus. L'ensemble de ces interactions en ligne entre politiciens et citoyens, mais aussi entre journalistes, représentants de partis et d'autres types d'influenceurs, génère une quantité importante de données numériques[1], appelées communément big data[2], qui a récemment intégré le champ de la recherche en sciences sociales en tant que source pour étudier les opinions politiques individuelles et collectives. D'après de nombreux chercheurs, les techniques d'analyse des opinions sur les médias sociaux peuvent être choisies comme alternatives aux méthodes traditionnelles.

Face à cette situation, cet article vise à étudier l'impact de l'utilisation des données des médias sociaux pour analyser les opinions politiques en sciences sociales, en s'intéressant en particulier au cas des études basées sur des données de Twitter. Comparé à d'autres médias sociaux, Twitter a été largement exploité à la fois par la facilité apparente de collecte de données[3] et par l'étendue large des sujets couverts par les échanges. Ces études varient en fonction de la taille des corpus, du small data au big data, du type de données, des méthodes et des approches – allant des approches basées sur le contenu à celles basées sur les interactions – et des méthodes. Concernant les méthodes, on présume généralement que la recherche dans

1 Si les données issues du Web ne constituent sûrement pas une nouveauté pour la recherche en sciences sociales, leur utilisation n'est pas si simple et demande une approche critique (Hogan, 2014). Plus particulièrement, il s'agit de ne pas négliger les problèmes techniques et commerciaux liés à l'accès aux données, les problèmes légaux liés à la protection de la vie privée et aux droits d'auteurs et, enfin, les problèmes de véracité et de représentativité de ces données (Severo *et al.*, 2016).

2 Dans cet article, le terme « big data » renvoie principalement aux « fichiers de données qui sont importants tant par leur taille que par leur complexité, et qui requièrent de nouvelles techniques algorithmiques afin d'en extraire des information utiles » (Holmes, 2017, p. 7, traduction). C'est pourquoi, nous faisons ensuite référence au défi de l'analyse multi-échelles du big data, qui renvoie à la capacité en informatique d'appliquer des algorithmes classiques à des problèmes plus grands. De ce fait, nous ne nous intéressons pas ici aux limites liées à la gestion des bases de données qui relèvent plus des contraintes de stockage que des contraintes d'analyse.

3 Twitter fournit des APIs pour la collecte des tweets. Ces APIs ont de nombreuses limites sur la taille, la durée et le type de corpus qui peut être collecté. Cependant, comme ils sont disponibles gratuitement, ils ont attiré l'attention de nombre de chercheurs.

ce domaine nécessite de dépasser l'opposition entre techniques quantitatives et qualitatives[4] et de trouver de nouvelles solutions pour une « redistribution des méthodes » (Marres, 2012). Alors que la les analyses en s'appuyant sur du big data ont rencontré de nombreuses difficultés dans la validation des résultats, celles basées sur du small data[5] semblent difficilement transposable à de grandes échelles. En conséquence, les chercheurs testent de nouvelles solutions méthodologiques mixtes afin de développer des approches « quali-quantitatives » (Venturini et Latour, 2010).

Étant donné la variété de ces études, notre premier objectif est de présenter un état de l'art de ce champ de recherche. L'intérêt de faire l'état de l'art d'une littérature aussi récente et vaste ne réside pas uniquement dans la possibilité de comparer les cas d'étude et les solutions méthodologiques, mais surtout dans l'exploration des différentes définitions de l'opinion politique utilisées par ces études. En définitive, il sera intéressant de voir si l'utilisation de nouvelles données provenant des médias sociaux – et spécifiquement l'exploitation de larges corpus considérés par les auteurs des articles étudiés du big data – permet la construction d'une approche qui soit réellement nouvelle pour l'analyse des opinions politiques, ou bien si l'approche de ces études ne diffère pas des recherches obtenues grâce aux techniques traditionnelles. Notre objectif final n'est donc pas de procéder à une analyse comparative de solutions empiriques, mais de révéler les théories (implicites ou explicites) sur lesquelles reposent ces analyses.

Dans un second temps, à partir de cet état de l'art, nous visons à vérifier si une nouvelle approche pour la recherche sur l'opinion politique en sciences sociales pourrait être basée sur des techniques multi-échelles. En effet, dans le corpus analysé, des approches multi-échelles ont émergé et elles constituent la nouveauté principale comparée aux techniques statistiques traditionnelles. Différemment des statisticiens, les *data scientists*[6] promettent aux chercheurs en sciences sociales la capacité de reproduire des analyses similaires à différentes échelles, du plus petit au plus grand et inversement, sans perdre en qualité. Partant de là, notre but est

4 Cette opposition a été définie de multiples façons. Une définition très simple pourrait être : « la recherche quantitative est une recherche empirique dans laquelle les données sont sous la forme de chiffres et de nombres... la recherche qualitative est une recherche empirique dans laquelle les données ne sont pas sous la forme de chiffres et de nombres » (Punch, 1998, p. 4, traduction). Ici, avec ces deux termes, nous essayons d'identifier deux approches différentes en sciences sociales. Les méthodes quantitatives sont utilisées pour quantifier l'objet de la recherche en générant des données chiffrées ou des données pouvant être transformées en statistiques utilisables. Les méthodes qualitatives, souvent basées sur des techniques non ou semi-structurées (groupes de discussion, entretiens, observations, *etc.*), sont principalement exploratoires et visent à étudier les raisons, les opinions et les motivations.

5 Par small data, nous entendons une donnée qui est suffisamment « petite » pour la compréhension humaine. Par son volume et son format, cette donnée est accessible, informative et exploitable. Il a été prouvé que ce type de donnée est très utile pour la recherche empirique. Cependant, de nombreux questionnements méthodologiques ont été soulevés lorsqu'on a essayé d'appliquer des méthodes et de vérifier des résultats obtenus sur du small data à des corpus de big data.

6 La science des données est utilisée comme un « concept pour unifier la statistique, l'analyse de données, l'apprentissage machine et les méthodes auxquelles ils renvoient » afin de « comprendre et [d'] analyser des phénomènes courants » à travers des données (Hayashi, 1998, traduction). Cependant, de nombreux chercheurs ont observé qu'il n'y a pas de différence entre les *data scientists* et les statisticiens. Nous utilisons ici le terme afin d'identifier une approche de la donnée plus interdisciplinaire incluant la statistique, mais aussi les mathématiques, les sciences de l'information et l'informatique, et utilisant des algorithmes automatisés.

d'explorer la faisabilité d'une sociologie numérique multi-échelle portant sur l'opinion politique afin d'appliquer des approches qualitatives à des corpus plus larges ou bien – et c'est là un défi – d'exploiter des approches du big data tout en répondant aux exigences épistémologiques de l'analyse qualitative. Au niveau de l'individu, la mesure et la quantification automatiques de l'opinion politique à partir des données des médias sociaux sont confrontées au contenu limité des échanges de microblogging, à sa nature souvent ambiguë, ironique ou paradoxale et à la nature multimédia d'un message qui peut contenir des images, des URL et des émoticônes. Au niveau du système, l'analyse des traces de l'activité de plusieurs millions d'individus requiert des innovations algorithmiques afin de rendre les méthodes traditionnelles efficaces à de plus grandes échelles. Cet article propose donc d'étudier les mérites de solutions méthodologiques visant à une analyse multi-échelle des opinions politiques sur Twitter, depuis des méthodes basées sur le contenu à celles basées sur l'interaction, depuis l'analyse statistique à l'analyse sémantique, depuis les approches supervisées aux approches non-supervisées[7]. Concernant la dernière opposition, nous discuterons en détail les statuts possibles d'une connaissance qualitative à l'intérieur d'une analyse quantitative, qu'elle soit responsable de la formalisation *a priori* du problème (approches supervisées) ou d'une interprétation *a posteriori* des résultats (approches non-supervisées), constituant ainsi un exemple concret de l'opposition épistémologique classique entre déduction et induction. Plus généralement, les défis actuels de l'apprentissage machine, c'est-à-dire l'utilisation de l'intelligence artificielle pour obtenir de l'ordinateur qu'il « apprenne » à partir des données, qu'il en déduise de nouveaux modèles et qu'il agisse sans être programmé explicitement, seront abordés comme des ré-actualisations pratiques de questionnements classiques dans le domaine des données numériques. Si ces questions ne sont pas nouvelles, le big data les rend actuelles.

Au final, nous souhaiterions identifier des tendances futures dans la recherche en sciences sociales portant sur l'opinion politique et soutenir l'adoption d'une méthodologie sociologique solide qui préserverait la sociologie empirique d'une « crise en devenir » (Savage et Burrows, 2007).

2. L'opinion politique sur Twitter : un état de l'art

2.1. Opinion politique et sondages d'opinion

Avant d'aborder la recherche sur Twitter, il semble utile d'examiner la vaste littérature sur les sondages d'opinion dans la mesure où cette approche pourrait être pertinente dans de nombreuses études impliquant Twitter. Si l'utilisation des sondages remonte à la fin du XIXe siècle (Cayrol, 2011), leur popularité a réellement décollé en 1936 lorsque George Gallup fut capable de prédire la victoire de Roosevelt. En France, Jean Stoetzel a établi les premiers sondages d'opinion en 1938 lorsqu'il fonda l'Institut français d'opinion publique. Depuis ce moment, la mesure de l'opinion publique est étroitement liée à l'organisation de tels sondages, basés sur le principe que l'opinion d'un échantillon de votants, sélectionnés avec attention,

7 Les approches supervisées sont basées sur une formalisation *a priori* d'un savoir expert, grâce par exemple à l'annotation manuelle des données, alors que les approches non-supervisées sont centrées sur une interprétation *a posteriori* de résultats statistiques généralement fournis par des données non-décrites.

peut être représentative de l'opinion d'une foule plus importante. Comme le note Loïc Blondiaux (1998), depuis plusieurs années, le succès des sondages d'opinion a été responsable d'une forme de négligence envers la définition de l'opinion d'un point de vue théorique.

Ces cinquante dernières années, les chercheurs ont développé des positions controversées sur la validité des sondages. Parmi les plus célèbres, citons l'essai de Pierre Bourdieu (1973) dans lequel l'auteur affirme que les opinions n'existent pas, afin de souligner le fait que les opinions réelles diffèrent de celles mesurées par les sondages. Bourdieu prétend que l'opinion telle qu'elle est mesurée dans les sondages est un artéfact créé par les chercheurs qui, par conséquent, introduisent leurs propres questions et, potentiellement, leurs propres réponses. À l'inverse, Page et Shapiro (1992) considèrent que les opinions sont réelles, mesurables et rationnelles. De leur point de vue, il est possible de se fier uniquement à des données d'enquête pour étudier l'opinion publique. Roland Cayrol (2011) souligne que l'objet des sondages n'est pas de connaître l'opinion des individus, mais d'agréger une opinion publique. Loïc Blondiaux (1998) propose une position intermédiaire. D'après lui, même si les sondages sont utiles pour étudier l'opinion publique, cela ne signifie pourtant pas que l'opinion et les sondages sont une seule et même chose. De la même manière, Ginsberg (1986) souligne que les réponses aux enquêtes sont le résultat d'une interaction entre une opinion et un outil de requête : par conséquent, les sondages modifient l'opinion.

Cet examen condensé de différentes positions à propos des sondages d'opinion est essentiel pour mettre en exergue deux points décisifs qui seront utiles à notre analyse. Tout d'abord, il y a une confusion entre le concept d'opinion publique agrégée (où « opinion » est toujours au singulier) et celui d'opinions individuelles (au pluriel). Appliquer l'une ou l'autre approche a d'importantes conséquences sur les techniques empiriques qui peuvent être utilisées pour analyser l'opinion politique. Par conséquent, si nous l'abordons du point de vue de la statistique en sciences sociales (Reynié, 1989) et si nous considérons que l'opinion de l'un est égale à celle d'un autre et que, donc, l'opinion publique est un phénomène agrégé, cela signifie que les sondages d'opinion doivent être considérés comme une technique de mesure efficace. En revanche, dans le cas où nous considérons que les opinions sont immanquablement liées aux individus, nous aurons alors besoin de méthodes capables de mesurer la contribution de chaque personne et l'interaction des uns avec les autres. Toutefois, s'il est souvent admis que les sondages d'opinion permettent uniquement d'examiner la dimension agrégée de l'opinion, la littérature sur le sujet ne propose pas directement d'alternatives lorsqu'il s'agit d'examiner les opinions individuelles. Cela nous amène au second point : l'importance des choix du chercheur et de la méthode adoptée. Il apparaît assez clairement que l'étude de la définition de l'opinion politique doit commencer précisément par l'analyse des choix empiriques, c'est-à-dire qu'il faut considérer les données sélectionnées et la manière dont elles sont traitées pour comprendre comment le chercheur définit l'opinion publique. Cela constitue justement l'objet des paragraphes qui suivent.

2.2. Une classification de l'opinion publique

Cette recherche vise à identifier les conceptions théoriques de l'opinion politique présentes dans la littérature utilisant des données provenant de Twitter. Pourtant, ce n'est pas une tâche

aisée, en particulier parce que ces recherches se concentrent principalement sur l'enquête empirique et ne définissent que rarement de façon explicite la manière dont elles définissent leurs concepts. Pour cette raison, comme étape préliminaire, il est utile de revenir à la classification proposée par Robert Entman et Susan Herbst (2001) qui identifient quatre formes de cadrage de l'opinion publique sur la base des méthodes utilisées dans les différentes études.

La première forme est l'« opinion de masse ». Selon cette forme, sans aucun doute la plus populaire, l'opinion publique peut être définie comme « l'agrégation ou la somme des préférences individuelles recoupées à travers des sondages d'opinion, des référendums ou des élections ». Dans cette forme, les opinions ne « reflètent pas des citoyens informés ou attentifs ». Elles sont plutôt les artéfacts des outils utilisés pour les collecter. La seconde forme d'opinion publique est l'« opinion publique latente » qui « sous-tend des opinions plus passagères et superficielles que nous pouvons trouver lorsque sont conduits des sondages de masse ». Cette forme dépend des considérations individuelles et des prédispositions politiques, c'est-à-dire « de caractéristiques stables au niveau de l'individu qui régulent l'acceptation ou non des communications politiques que les personnes reçoivent » (John R. Zaller, 1992, p. 22). La troisième forme, appelée « opinion publique activée », renvoie à l'opinion « de citoyens engagés, informés et organisés – ceux-là mêmes qui sont mobilisables pendant les périodes de campagne, mais aussi entre les élections. » Dans cette forme, nous pouvons à l'évidence inclure toutes les études qui insistent sur le rôle des leaders d'opinion et des « influenceurs » (Katz et Lazarsfeld, 1955; Lazarsfeld *et al.*, 1948; Merton, 1968), mais aussi des médias (Watts et Dodds, 2007) dans la formation des opinions politiques. La quatrième forme d'opinion publique, appelée « majorités perçues »[8], renvoie à des recherches où le terme d'opinion désigne « les perceptions de la plupart des observateurs, en incluant les journalistes, les personnages politiques et les membres du public en soi, de l'endroit où se situe la majorité sur un sujet ». Contrairement aux formes précédentes, ici, les opinions ne correspondent pas aux préférences de personnes considérées de façon agrégée ou individuelle, mais plutôt aux représentations de l'opinion produites par les médias. Elles ne sont pas un véritable sentiment, mais ce que les médias en rapportent. Par conséquent, les études sur ce sujet n'essaient pas d'évaluer l'opinion des personnes grâce à des sondages ou à d'autres techniques, mais sont centrées sur l'analyse de l'opinion médiée. Cette forme d'opinion publique inclut généralement des études centrées sur la configuration des intentions.

2.3. L'analyse des opinions politiques sur Twitter

Parmi les centaines d'articles[9] essayant d'analyser les opinions politiques sur Twitter, la plupart d'entre eux centrent leur attention sur l'utilisation de Twitter pour la prédiction de résultats de vote et considèrent les tweets comme une alternative aux sondages traditionnels.

8 Si nous appliquons la distinction entre 'opinion' et 'opinions' aux quatre formes, nous pouvons aisément voir que l'« opinion de masse » et les « majorités perçues » correspondent à une vue collective de l'opinion, alors que l'« opinion latente » et l'« opinion activée » sont basées sur l'idée d'opinions individuelles multiples.

9 L'état de l'art est basé sur l'analyse des 70 articles les plus cités selon Scopus et Google Scholar, correspondant à la requête « *political opinion* ET *twitter* » (Annexe 2).

Dans ce contexte, il semble intéressant d'établir si, dans ce corpus d'études, celles ayant des questionnements de recherche similaires partagent aussi des conceptions théoriques similaires de l'opinion politique.

En premier lieu, nous observons que les articles peuvent être divisés en deux groupes. Le premier groupe est constitué d'études qui adoptent une vision positive en validant l'utilisation de Twitter comme prédicteur de vote. C'est le cas du célèbre article de Tumasjan *et al.* (2010, cité 1487 fois d'après Google Scholar, mais absent dans Scopus) qui démontre qu'avec l'aide de l'analyse des sentiments[10] dans les tweets, il aurait été possible de prédire les résultats des élections fédérales allemandes de 2009 (étude basée sur environ 100 000 tweets). Selon les auteurs, « le nombre seul de tweets mentionnant un parti politique peut être considéré comme un reflet plausible de la répartition du vote et son pouvoir prédictif est proche de celui des sondages d'élection traditionnels ». On note que les auteurs définissent les opinions politiques comme des sentiments politiques ; pourtant, même s'ils utilisent l'analyse des sentiments, elle se réduit à un comptage très basique des mentions et les conclusions sont basées sur l'équivalence numérique entre les votes et les tweets. De la même manière, Livne *et al.* (2011) élaborent une technique pour prédire les résultats d'élection avec une précision de 88% (étude basée sur 460 000 tweets). Leur méthode de prédiction est fondée sur l'évaluation d'indicateurs multiples dérivés non seulement de la fouille de textes, mais aussi, et avant tout, de données structurelles basées sur l'analyse de réseaux. La position défendue par O'Connor *et al.* (2010) semble plus prudente puisqu'ils indiquent que Twitter peut être utilisé pour suivre l'opinion en temps réel, mais qu'un tel outil ne peut pas avoir un pouvoir de prédiction comparable à celui des sondages traditionnels. Leur article est le premier dans ce domaine à utiliser des techniques d'analyse des sentiments, même si elles demeurent basiques puisqu'elles reposent simplement sur des lexiques pour déterminer la polarité des tweets (étude basée sur 1 milliard de tweets).

Le second groupe est constitué d'articles qui rejettent ou relativisent l'utilisation de Twitter comme prédicteur de vote. Certains chercheurs ont concentré leurs efforts sur l'invalidation des études précédentes. Généralement, la critique de ces auteurs met l'accent sur la différence démographique entre les utilisateurs de Twitter et les électeurs probables et, par conséquent, sur la non-représentativité des échantillons induite par les données issues de Twitter. Metaxas *et al.* (2011) proposent la formule « prédire le présent » en soulignant le fait que les chercheurs qui prédisent les résultats d'élection le font après les élections, lorsqu'ils en connaissent déjà le résultat (étude basée sur environ 235 000 tweets). Daniel Gayo-Avello (2013) présente un état de l'art sur la question, mettant en exergue les faiblesses des études précédentes (notamment relatives à la reproductibilité des résultats). Jungherr *et al.* (2011) invalident l'analyse de Tumasjan *et al.* (2010) en montrant que la sélection des partis et la période choisie ont une grande influence sur les résultats obtenus. Skoric *et al.* (2012) montrent qu'il y a une certaine corrélation entre les discussions et les votes sur Twitter, mais qu'elle est insuffisante pour faire des prédictions exactes. Une critique quelque peu différente peut être trouvée chez Boyadjian (2014) qui ne se concentre pas sur les limites de l'analyse

10 Le terme d'« analyse des sentiments » est utilisé pour identifier un groupe de techniques variées ayant pour but d'extraire et d'analyser des états affectifs dans les textes.

empirique, mais qui examine plutôt la question plus générale de la représentativité sur Twitter, comparée aux sondages traditionnels.

D'autres chercheurs ont mis l'accent sur les limites des méthodes employées en offrant des solutions plus avancées. Certains articles suggèrent l'utilisation de l'apprentissage machine pour l'analyse des sentiments. Lors de leur étude de l'élection générale irlandaise de 2011 (32 578 tweets), Bermingham et Smeaton (2011) ont démontré la validité de l'utilisation de Twitter comme prédicteur grâce à une approche combinant analyse des sentiments (par apprentissage supervisé) et mesures de volume. Concernant l'apprentissage supervisé, les auteurs indiquent : « Nous avons donné comme instruction aux annotateurs de ne pas considérer comme sentiment la mention d'un fait positif ou négatif, mais que le sentiment peut être une émotion, une opinion, une évaluation ou une spéculation envers le sujet ciblé » (p. 5).

Lorsqu'on prend en compte l'ensemble de ces articles, on peut noter que personne n'examine l'équivalence entre les tweets et les opinions. Le fait que ces messages courts nous autorisent à connaître l'opinion de personnes est considéré comme légitime, d'une façon ou d'une autre. Même des chercheurs en sciences politiques, qui peuvent se montrer très critiques vis-à-vis du pouvoir prédictif de la plateforme, mettent l'accent sur les choix faits dans l'analyse empirique sans questionner la relation entre les données et l'objet étudié. Pourtant, en se concentrant sur les données et les méthodes sélectionnées, il est possible de distinguer quatre modèles conceptuels de l'opinion politique qui ne sont pas exclusifs et qui peuvent être combinés dans la même étude (Tableau 1).

Modèle conceptuel	**Forme d'Eldam et Herbst**	**Données**	**Méthodes**
Préférence	Opinion de masse	Le tweet comme unité	- Statistiques - Analyse basique des sentiments (lexique)
Sentiment	Opinion latente	Le tweet comme contenu	- Analyse avancée des sentiments (apprentissage non-supervisé et supervisé)
Interaction	Opinion activée	Le tweet comme interaction	- Analyse de réseaux
Agenda	Majorités perçues	Le tweet comme medium (définissant l'agenda)	- Analyse de discours - Fouille de textes

Tableau 1. Les quatre modèles conceptuels de l'opinion politique.

2.3.1. L'opinion comme préférence

Dans le premier groupe d'études (*préférence*), l'opinion politique est considérée comme une préférence agrégée relative à un objet spécifique sélectionné par le chercheur. Les chercheurs qui adoptent ce modèle construisent des corpus de tweets contenant des mots-clés ou des hashtags spécifiques, comme, par exemple, le nom d'un candidat ou d'un parti politique. Ils s'intéressent au tweet comme unité, dans son ensemble (sans considérer les co-occurrences en son sein ou les contextes d'usage) et ils observent principalement la variation des volumes de tweets selon différents paramètres (temps, espace, utilisateurs, sujet, *etc.*). Certains d'entre eux utilisent des techniques d'analyse des sentiments très basiques (lexiques construits manuellement) qui produisent un simple comptage de mots. Dans ce groupe, on trouve principalement des études cherchant à prédire des résultats d'élection (Tumasjan *et al.* 2010 ; Livne *et al.* 2011 ; Skoric *et al.* 2012). En utilisant des techniques quantitatives de nature différente, leur but est de vérifier s'il y a une corrélation entre le nombre de tweets mentionnant un·e candidat·e et le nombre de votes qu'il·elle reçoit. Dans ce type d'étude, l'opinion politique est clairement envisagée comme « opinion de masse ». Elle est traitée comme étant quantifiable, mesurable et comptable. Les tweets sont utilisés pour l'étudier comme un phénomène agrégé. Par conséquent, des similarités avec les sondages traditionnels peuvent être facilement identifiées. Le chercheur applique une question à un échantillon de données préexistant, en présupposant que cet échantillon contient la réponse.

2.3.2. L'opinion comme sentiment

Prenant en compte les limites d'une approche simplement basée sur le comptage de préférences, des études plus récentes se sont concentrées sur les attitudes individuelles. Les chercheurs qui adoptent cette définition peuvent également construire des corpus de tweets contenant des mots ou des hashtags spécifiques, mais ils s'intéressent au contenu du tweet plus qu'au tweet comme simple unité. Ils peuvent étudier les co-occurrences de mots ou des structures textuelles plus avancées en essayant d'interpréter le sentiment exprimé dans le texte. L'objectif final est d'obtenir un point de vue complexe qui prenne en compte les positions individuelles liées à l'objet de l'étude en accord avec la forme de l'« opinion latente ».

Dans ce contexte, un problème épineux est celui de la définition du concept de sentiment. Ces dernières années, les techniques d'« analyse des sentiments » sont devenues très populaires. D'après la définition de Wilson *et al.* (2005), le sentiment est une question de polarité contextuelle : « L'analyse des sentiments correspond à l'identification d'opinions, d'émotions et d'évaluations positives et négatives ». Les auteurs proposent un lexique des sentiments enrichi grâce à l'apprentissage supervisé. Au sein de notre corpus, O'Connor *et al.* (2010), cités 635 fois d'après Scopus et 1487 fois d'après Google Scholar, construisent leur analyse des sentiments sur le lexique d'OpinionFinder. Leur approche a été reproduite des dizaines de fois au cours des années suivantes. Par exemple, Conover *et al.* (2011, le troisième article le plus cité dans Scopus) ont développé une méthode d'analyse du contenu basée sur l'annotation manuelle (données décrites) pour l'investigation de 355 millions de tweets. Plus récemment, des chercheurs (Bermingham et Smeaton, 2011) ont proposé des approches

supervisées pour la construction de classeurs de sentiment. Enfin, l'approche de l'opinion comme sentiment inclut aussi des études qui qualifient le sentiment d'une manière plus qualitative, basées sur de petits corpus de données.

2.3.3. L'opinion comme interaction

Certains chercheurs ont élargi le champ de leur étude en mettant l'accent non pas sur le tweet, mais sur son contexte, afin d'identifier le réseau d'interactions lié à la formation et à la circulation des opinions. D'après ce modèle conceptuel, les opinions sont des sentiments individuels engendrés non seulement par les prédispositions d'une personne, mais aussi et avant tout influencés par la position de l'individu dans la société. La plupart des études dans ce champ se concentrent sur le rôle des influenceurs, en concordance avec la forme d'« opinion activée ». Ici, nous pouvons mentionner des études quantitatives cherchant à identifier des leaders d'opinion et basées sur des indicateurs de réseau. Grâce à l'analyse d'un corpus lié aux hashtags politiques #FreeIran, #FreeVenezuela et #Jan25, Bastos *et al.* (2013) ont étudié la structure du contrôle sur Twitter en analysant les retweets, les mentions et les réseaux de followers pour chaque hashtag. Ils ont rejeté l'idée selon laquelle il existerait des *hubs* assumant une fonction de contrôle et ils ont souligné plutôt l'importance de minorités engagées. Le point de vue développé par Park (2013) est sensiblement différent : il a mené une enquête mettant en exergue la différence entre le leadership d'opinion traditionnel basé sur la théorie de la communication à double étage (Katz et Lazarsfeld, 1955) et le leadership d'opinion sur Twitter.

Certains chercheurs combinent l'analyse de contenu avec l'analyse de réseaux. Xu *et al.* (2012), en se concentrant sur les réseaux d'activistes, ont exploré à la fois le leadership d'opinion grâce à des mesures statistiques de réseau et l'engagement politique grâce à l'analyse du profil et du contenu des tweets. Leurs résultats s'opposent à ceux obtenus par Bastos *et al*. (2013), montrant la connexion entre la centralité et le leadership. Afin de prédire l'affiliation politique des utilisateurs de Twitter, Conover *et al*. (2011) ont combiné des méthodes basées sur le contenu avec des analyses structurelles de l'information politique dans les réseaux de diffusion (réseaux de retweets et de mentions), et ont ainsi validé l'analyse de réseaux comme étant une solution plus efficace pour l'identification des alignements politiques des utilisateurs. De la même manière, Stieglitz et Dang-Xuan (2012) ont combiné l'analyse de sentiment (en utilisant le logiciel *Linguistic Inquiry and Word Count LIWC*[11]) avec l'analyse de réseau sur un corpus de 64 000 tweets. Leur intention était d'étudier si le sentiment articulé dans les tweets politiques avait un effet sur leur retweetabilité. Enfin, des études plus qualitatives ont analysé les tweets de catégories spécifiques d'utilisateurs considérés comme influents comme, par exemple, les journalistes (Molyneux, 2015).

2.3.4. L'opinion comme agenda

D'autres études se sont concentrées sur le rôle de médium de Twitter, configurant l'agenda de l'opinion publique. Dans ce cas, la relation entre le tweet et son auteur n'est pas pertinente.

11 Qui comprend des dictionnaires intégrés.

Les tweets n'équivalent pas aux opinions agrégées ou individuelles des personnes. Pourtant, elles transmettent à ces personnes des représentations sociales engendrées directement par la plateforme.

Par exemple, l'étude de Papacharissi et de Fatima Oliveira (2012, la deuxième plus citée dans Scopus) a retracé le rythme du récit sur Twitter des actualités liées au hashtag #Egypt. Les auteurs visaient à identifier l'évolution de la valeur médiatique des actualités, responsable de la sélection de ces actualités sur Twitter. Ici, les méthodes reposent sur une analyse textuelle automatisée combinée à une analyse de discours. L'attention n'est pas réellement portée sur la formation de l'opinion, mais plus sur Twitter en tant que médium, assurant le partage d'informations médiatiques.

Cette dernière approche est profondément différente des autres, car la recherche dans ce domaine n'utilise que rarement les tweets comme données. Elle cherche plutôt à examiner Twitter en tant qu'acteur social. De plus, elle met en exergue la question sous-jacente, mais essentielle, de la validité des tweets comme données ascendantes, représentant les opinions des personnes. En effet, selon ce modèle, les tweets seraient produits par la plateforme plus que par les personnes elles-mêmes (Marres, 2017).

3. Les défis méthodologiques de l'opinion politique

Une fois définis ces quatre modèles conceptuels, il est intéressant de les confronter sur le plan méthodologique. Plus spécifiquement, cet article vise à identifier les études proposant de passer du small au big data ou, mieux encore, de combiner différentes échelles d'analyse pour combler le fossé entre recherche qualitative et quantitative. Cependant, pour aborder cette question, il est important d'accorder en amont un peu d'attention au contexte scientifique. S'ils ne sont pas toujours d'accord sur la meilleure position à adopter, de nombreux chercheurs convergent en revanche sur le fait que les données numériques, et en particulier l'émergence du big data, constituent une rupture pour de nombreux domaines des sciences sociales. Elles devraient donc être considérées comme un véritable changement de paradigme, de manière toutefois prudente. Concernant le développement des plateformes numériques telles que Twitter, Venturini et Latour (2010) affirment qu'« [elles] offrent bien plus qu'un autre domaine où appliquer des méthodes existantes : elles offrent la possibilité de restructurer l'étude de l'existence sociale ». L'impact d'une telle révolution scientifique n'est pas de nature à affecter les sciences sociales uniquement *de l'intérieur* en invalidant des méthodes traditionnelles et en en promulguant de nouvelles, mais aussi *de l'extérieur* en modifiant les institutions, les pratiques et même certains objectifs épistémiques dans ce domaine. Parmi les effets externes du big data sur les pratiques des sciences sociales, les plus fréquents sont : les défis éthiques concernant l'accès et la protection de la vie privée, les défis politiques concernant la dépendance de la recherche scientifique à la production et aux conditions d'accès à ces données (Driscoll et Walker, 2014), ainsi que des défis institutionnels qui demandent de repenser complètement les communautés de recherche et d'éducation dans ce champ spécifique (Lazer *et al*., 2009).

Cependant, il ne s'agit pas ici d'aborder ces nombreuses questions liées aux pratiques des sciences sociales, mais plutôt de se concentrer sur l'impact qu'un tel changement de paradigme peut avoir sur les méthodes, c'est-à-dire sur les changements qui ont lieu *de l'intérieur*. Dans ce but, notre analyse s'articule autour de trois oppositions méthodologiques majeures liées à l'utilisation de l'informatique en sciences sociales : les approches uniquement micro et uniquement macro vs. les approches multi-échelles ; les approches exploratoires vs. les approches prédictives ; les approches supervisées vs. les approches non-supervisées.

3.1. Approches macro ou micro vs. approches multi-échelles

D'après Venturini et Latour (2010), l'un des défis épistémologiques majeurs que pose le big data aux sciences sociales implique de reconsidérer les visions dualistes de la donnée. L'utilisation classique d'outils statistiques – comme, par exemple, les méthodes agrégées – pour déchiffrer les structures sociales macroscopiques, introduit « une distinction fictive entre des micro-interactions et des macro-structures ». De manière analogique, la physique statistique a, très tôt, développé des modèles inter-niveaux pour distinguer les macro-mesures classiques des micro-mesures et a ainsi comblé l'écart causal entre individus et agrégats. Les sciences sociales développent actuellement des modèles ascendants similaires pour l'étude des phénomènes sociaux. Même si cela peut être attribué à la grande complexité des objets sociaux, il n'en demeure pas moins que cette posture dualiste conduit les approches traditionnelles à une simple « juxtaposition de l'analyse statistique avec l'observation ethnographique ». Poursuivant cet énoncé quelque peu perturbant, Venturini et Latour plaident en conséquence pour une approche véritablement « quali-quantitative », en se basant sur les traces numériques pour combiner et unifier la précision des enquêtes ethnographiques avec le champ très étendu de l'analyse statistique. Cette position épistémologique avait déjà été recommandée par les sciences de la complexité dans les années 1970, comme avec, par exemple, *Le Macroscope* imaginé par de Rosnay (1975). Ce nouvel instrument visait à observer les sociétés dans toute leur complexité grâce à la simulation informatique, anticipant de ce fait l'émergence des sciences sociales computationnelles qui utilisent des modèles sociologiques pour étudier la structure causale des phénomènes macroscopiques (Lazer *et al*., 2009; Cioffi-Revilla, 2016; Casini et Manzo, 2016).

Dans notre corpus, de nombreuses études s'intéressent uniquement à la question des agrégats. C'est le cas de la plupart des articles traitant de la prédiction des élections dans la mesure où leur objectif de recherche est motivé par la forme collective de l'opinion politique. Cet objectif est une justification suffisante pour l'utilisation de méthodes engendraint des mesures erronées des opinions individuelles. Par exemple, O'Connor *et al*. (2010) ont d'abord affirmé qu'ils « s'intéressaient uniquement au sentiment agrégé ». Ainsi, « un fort taux d'erreur [au niveau individuel] implique simplement [que] le détecteur de sentiment est un instrument de mesure bruité. Grâce à un grand nombre de mesures, ces erreurs se neutralisent au sein d'une opinion publique agrégée, relative à la quantité que nous choisissons d'estimer. » Dans cet état d'esprit, le seul critère pour la validation des méthodes employées est la capacité de prédire des valeurs agrégées, sans ainsi garantir leur intérêt pour des recherches orientées sur l'opinion individuelle.

Pourtant, d'après d'autres chercheurs, l'impact de certains individus très influents est crucial pour expliquer les dynamiques globales. Stieglitz et Dang-Xuan (2012) insistent sur le fait que « la discussion politique sur Twitter est conduite par un petit nombre d'utilisateurs très actifs, représentant environ un pour cent de l'ensemble des utilisateurs ». Ils ont ainsi procédé à une analyse détaillée de ce « un pour cent » et ont proposé des mesures cohérentes pour traiter simultanément les individus et les agrégats. De la même manière, Bermingham et Smeaton (2011), après avoir conduit une analyse des sentiments à grande échelle sur 32 578 tweets, ont essayé « d'identifier des termes qui ouvrent une voie pour l'exploration qualitative de l'ensemble des données ». Partant d'une observation macroscopique de l'opinion, ils s'intéressent ensuite à l'explication microscopique en examinant les facteurs individuels qui impactent les résultats agrégés de manière significative. Les travaux de J. Boyadjian (2014) relève aussi de l'approche multi-échelles – en termes de construction du corpus – dans la mesure où ils combinent une collecte de données détaillée, grâce à des enquêtes traditionnelles, et une collecte agrégée à grande échelle, grâce aux méthodes informatiques.

3.2. Approches exploratoires vs. approches prédictives

Le développement rapide du big data et des outils algorithmiques issus de l'apprentissage machine ont conduit à une autre crise épistémologique. Comme l'ont énoncé des chercheurs « néo-positivistes », ce changement de paradigme pourrait très bien conduire à une pratique des sciences sociales où les sociologues ne seraient plus nécessaires. Cette affirmation, tout aussi provocante que sincère, a été faite il y a dix ans par Anderson (2008) lorsqu'il a écrit que « la corrélation supplante [désormais] la causalité et [que] la science peut progresser sans modèles cohérents, sans théories unifiées et sans aucune explication mécanique. ». Alors que l'opposition épistémologique entre corrélation et causalité n'est ni nouvelle, ni close, la quantité colossale d'informations rendue disponible par les données numériques issues du Web semble avoir favorisé durant la dernière décennie l'application de méthodes purement statistiques. Ainsi, dans nombre d'articles de notre corpus, l'objectif principal est la prédiction des phénomènes sociaux, bien plus que leur explication. Pourtant, Metaxas *et al*. (2011) indiquent que « par le passé, des recherches ont traité les médias sociaux comme des boîtes noires : ils peuvent donner la bonne réponse, mais vous pouvez ne pas savoir pourquoi. Nous croyons que la recherche peut apporter une contribution intellectuelle, si les méthodes sont accompagnées d'au moins un modèle basique et raisonnable qui explique pourquoi elles peuvent faire des prédictions correctes. »

Pour contrer l'idée que le big data, accompagné par les méthodes d'apprentissage machine, peut conduire à une découverte automatique de corrélations significatives, des chercheurs argumentent, en s'appuyant sur des principes de théorie de l'information, que ces approches exclusivement basées sur l'analyse statistique décontextualisée sont vouées à l'échec (Calude et Longo, 2016). D'autres soulignent le fait que, même si la découverte de corrélations par ordinateurs est réalisable, cela ne remplit pourtant pas les objectifs des sciences sociales. Comme l'indique Pigliucci (2009), les sciences « n'ont pas pour objet la découverte de schémas – même si cela fait certainement partie du processus – mais elles visent à trouver des explications pour ces schémas. » La différence entre une loi statistique, qui peut être efficace

pour la prédiction, et un modèle, qui est capable de fournir une explication mécanique supplémentaire, est bien illustrée par Masad (2014) dans une expérience plutôt simple reposant sur le modèle de ségrégation spatiale de Schelling.

Cette distinction entre algorithmes, qui fonctionnent comme des boîtes noires, et approches exploratoires est une autre problématique cruciale pour notre domaine. On peut citer l'exemple de Stieglitz et Dang-Xuan (2012) qui ont essayé d'établir une corrélation entre l'expression de sentiments dans les tweets à caractère politique et leur retweetabilité. Pour ce faire, ils ont conçu un « modèle prédictif des retweets » utilisant différentes catégories de sentiment comme caractéristiques informatives de la retweetabilité. Tout en mentionnant le contexte théorique des sciences de la communication concernant « le rôle des sentiments dans la communication au sein de newsgroups, de forums de discussion ou dans d'autres contextes », les auteurs ne font jamais explicitement référence à la possibilité de comportements psychologiques sous-tendant la corrélation révélée par leurs conclusions. Il est donc possible que l'expression de sentiments ne soit pas, de fait, la cause de la retweetabilité, mais qu'elle lui soit uniquement corrélées à travers des variables cachées (comme, par exemple, la présence d'une URL dans le tweet). De la même manière, Ceron *et al*. (2014), dans leur étude sur la dynamique des élections, ont donné des interprétations « plutôt spéculatives » (*sic*) aux corrélations qu'ils ont trouvées, concernant la popularité des leaders politiques, entre des données d'enquêtes et des données issues de Twitter. Cependant, une analyse qualitative plus approfondie de ces élections serait nécessaire pour dépasser la simple corrélation statistique et pour mieux comprendre les relations causales qui peuvent exister entre la popularité en ligne et la popularité hors ligne, c'est-à-dire la manière dont l'une pourrait expliquer l'autre.

Inversement, d'après Bermingham et Smeaton (2011), « dans les mesures d'opinion et les analyses de médias sociaux, il est restrictif de simplement mesurer sans fournir les moyens d'expliquer les mesures ». Dans ce sens, afin de transférer les corrélations découvertes statistiquement à des modèles mécaniques intègrant une véritable compréhension sociologique et qualitative, Colleoni *et al*. (2014) ont proposé une interprétation éclairée des corrélations découvertes entre des éléments structurels du réseau de retweets et les penchants politiques trouvés au sein des comptes Twitter. Exploitant le modèle classique de l'homophilie politique dans les discussions en ligne et hors ligne, ils ont construit leur argumentation sur des typologies développées en sociologie politique (par exemple, les « penseurs politiques », les « activistes politiques », le « public général ») et dans les sciences de la communication (par exemple, les modalités distinctes de la communication politique dans les réseaux numériques, envisageant Twitter comme un « medium social » ou comme un « medium d'information »). Cette association permet aux modèles qualitatifs de fournir, en plus de corrélations, une explication sociologique adéquate. Chez Jürgens *et al*. (2011), de telles associations sont même réalisées au niveau formel. Des concepts-clés hérités de la recherche en communication de masse (comme les « *news gatekeepers* ») sont exprimés dans le cadre de la théorie des graphes, en se basant sur le *Key Player Problem* de Borgatti (2006). Ici, la formalisation mathématique de modèles issus de la sociologie qualitative est ensuite utilisée pour tester des hypothèses et pour ainsi détecter « les utilisateurs qui ont la plus grande possibilité de bloquer ou de perturber les flux d'information ».

3.3. Approches supervisées vs. approches non-supervisées

L'opposition entre approches prédictives et approches explicatives se retrouve dans la manière dont le savoir sociologique est intégré au sein de l'analyse quantitative : soit comme une manière d'interpréter *a posteriori* les résultats d'une corrélation basée sur des typologies qualitatives, soit comme un modèle *a priori* qui requiert d'être d'abord formalisé pour tester les hypothèses sociologiques. De ce fait, cette distinction semble être un exemple concret de l'opposition épistémologique classique entre les approches inductives, où l'interprétation qualitative vient après l'analyse statistique, et les approches hypothético-déductives. En particulier, la distinction entre apprentissage supervisé et non-supervisé au sein des méthodes d'apprentissage constitue un cas pratique de cette opposition. L'apprentissage supervisé implique, en effet, la déduction de catégories à partir de données annotées *a priori*, alors que l'apprentissage non-supervisé part de données non-annotées pour induire une classification qui sera interprétée *a posteriori*. Dans la littérature consacrée à l'utilisation de l'analyse des sentiments sur Twitter pour la prédiction de l'orientation politique des individus, on peut distinguer trois catégories : les lexiques d'opinions, l'apprentissage non-supervisé et l'apprentissage supervisé.

Tout d'abord, les méthodes traditionnelles pour l'analyse des sentiments sont basées sur des lexiques d'opinions comme, par exemple, ceux fournis par OpinionFinder (O'Connor *et al.*, 2010), par un corpus d'opinions (He *et al.*, 2012) ou par un logiciel linguistique utilisant un « dictionnaire interne validé psychométriquement » (Stieglitz et Dang-Xuan, 2012). Ces méthodes sont souvent appelées « méthodes non-supervisées » (Bermingham et Smeaton, 2011) dans le sens où, une fois le lexique défini, la classification des tweets est automatique : aucune interaction humaine supplémentaire n'est nécessaire. Toutefois, nous pensons que ce nom porte à confusion dans la mesure où, au contraire, une supervision importante est de fait requise en amont : un savoir d'expert est nécessaire au tout début de la tâche de classification. C'est pour cette raison que O'Connor *et al.* (2010) affirment que les lexiques d'opinion sont plus « transparents » que d'autres approches dans le sens où le résultat de la classification peut être expliqué simplement par les modèles linguistiques qui ont été utilisés pour construire le lexique. Cependant, nous notons aussi que ces « méthodes non-supervisées » ne sont pas des « méthodes d'apprentissage » puisque les catégories de sentiment ne sont pas apprises à partir des données, mais fournies *a priori*. Il en résulte que le processus est plutôt rigide et qu'il s'adapte peu à un contexte spécifique ou à un corpus particulier.

À l'extrême opposé, l'apprentissage non-supervisé vise à organiser les tweets selon leur contenu sans requérir aucun savoir linguistique. Ces approches construisent, par conséquent, des groupes cohérents de tweets à l'aide de similarités statistiques reposant sur des éléments génériques comme, par exemple, les mots qu'ils contiennent. Les catégories qui en résultent peuvent alors être décrites *a posteriori* en observant le contenu des tweets, puis être utilisées pour classer de nouveaux tweets. Comme aucun savoir préalable n'est requis, ces approches sont abordables en termes d'implémentation et peuvent produire des prédictions efficaces à l'intérieur des catégories ainsi découvertes. Cependant, il peut être difficile de donner du sens à ces catégories, respectant un cadre psychologique ou sociologique spécifique, puisqu'elles ont été construites de manière purement inductive. Au sein de notre corpus, quelques études

seulement reposent sur ce second type d'approche, mais nous pouvons citer au moins l'utilisation de *topic models* (Mei *et al*., 2007; He *et al*., 2012) et, plus généralement, de *latent space models* (Barberá *et al*., 2015).

Une troisième catégorie d'approches, largement utilisées, se situe entre les deux catégories citées précédemment. Il s'agit de l'apprentissage supervisé. Dans ce cas, un ensemble de catégories de sentiment est d'abord défini : soit par une typologie unidimensionnelle relativement simple (« positif », « négatif », « neutre »), soit par une typologie plus complexe liée aux émotions humaines classiques (par exemple, « joie », « excitation », « tristesse », « peur »). Ensuite, un ensemble de caractéristiques extraites du contenu des tweets sont définies comme potentiellement informatives. Il peut simplement s'agir de compter les séquences de mots (Bermingham et Smeaton, 2011), leurs fréquences (Colleoni *et al*., 2014), ou d'analyser des décompositions grammaticales plus complexes en *part-of-speech elements* (Pak et Paroubek, 2010). Le but de l'apprentissage machine est alors de construire un « classifieur », c'est-à-dire une fonction utilisant les caractéristiques de contenu des tweets pour identifier la catégorie de sentiment à laquelle ils appartiennent. Pour ce faire, une liste d'exemples obtenue par l'annotation manuelle d'un corpus significatif (l'ensemble d'entraînement) est utilisée pour initialiser le classifieur. C'est lors de cette phase principalement que le savoir humain est convoqué. Ensuite, une méthode statistique est appliquée afin d'identifier des corrélations entre les caractéristiques sélectionnées et les annotations fournies. De nombre de techniques différentes sont exploitées dans la littérature, comme, par exemple, les *Support Vector Machines* (SVM) chez Conover *et al*. (2011), l'apprentissage itératif (Bermingham et Smeaton, 2011), la classification passive-agressive (Colleoni *et al*., 2014), ou les classifieurs naïfs de Bayes (Pak et Paroubek, 2010). Enfin, les corrélations identifiées sont testées et validées en les comparant à d'autres corpus pour mesurer leur sensibilité (nombre de vrais positifs) et leur précision (nombre de faux positifs).

Par rapport à l'apprentissage non-supervisé, l'intérêt des lexiques d'opinions et de l'apprentissage supervisé réside dans le fait que les catégories de sentiments sont définies préalablement par les experts, en se basant, par exemple, sur des modèles psychologiques de la communication humaine. Cependant, dans le cas des lexiques d'opinions, les liens entre les éléments et les catégories sont exprimés *a priori* par l'expertise linguistique alors que, dans le cas de l'apprentissage supervisé, ces liens sont découverts de façon inductive grâce aux corrélations apprises. D'après Ceron *et al*. (2014), « les codeurs humains sont, bien entendu, plus efficaces et attentifs que les dictionnaires ontologiques », suggérant que la meilleure solution est donc un classifieur supervisé, plus flexible puisque basé sur des exemples pratiques réels. Pourtant, il est important de noter que, lorsque les caractéristiques sélectionnées sont très génériques (par exemple, n'importe quel mot peut être une caractéristique), il peut être difficile de démêler les corrélations indirectes (un mot est corrélé à un sentiment via une troisième variable cachée qui explique les deux) des relations causales (un mot est vraiment l'expression du sentiment). Même s'il peut prédire de manière efficace, le classifieur qui en résulte ne permet pas d'expliquer le résultat de la prédiction de façon claire pour l'analyse linguistique. Par conséquent, les chercheurs doivent choisir entre des approches explicatives, mais rigides, peu adaptées donc à la variété et à l'ambiguïté de

l'utilisation du langage sur Twitter, et des approches flexibles, mais uniquement prédictives, qui ne peuvent pas être mises en correspondance avec un modèle explicatif et causal.

4. Tendances futures relatives à l'émergence du big data

Après avoir examiné l'état de l'art et identifié les principales approches théoriques et méthodologiques développées par la recherche numérique en sciences sociales concernant l'étude de l'opinion politique, il est désormais temps de passer à notre second objectif, c'est-à-dire de proposer quelques orientations futures de ce champ. En termes pratiques, nous nous intéressons aux méthodes informatiques qui permettent une description multi-échelles des données, c'est-à-dire aux méthodes capables de décrire simultanément des dynamiques de long terme, des structures communautaires macroscopiques, d'une part, et des détails cruciaux, des micro-événements qui échappent au contrôle des méthodes agrégatives, d'autre part. Si nous essayons de croiser les approches conceptuelles de l'opinion politique avec les oppositions méthodologiques mentionnées plus haut, un certain nombre d'idées intéressantes émergent. Dans chaque cellule du Tableau 2, nous décrivons la situation actuelle et les perspectives futures de la recherche qui englobent les différentes définitions de l'opinion en lien avec les trois défis méthodologiques discutés dans la partie précédente. Ce tableau doit être considéré comme une catégorisation abstraite des travaux dans la mesure où la plupart des articles analysés combinent plusieurs modèles conceptuels. Cependant, cette abstraction vise à mettre en évidence les liens qui existent entre des choix conceptuels et des solutions méthodologiques.

Concernant le premier modèle, pour lequel l'opinion est considérée comme une préférence collective, les macro-structures suffisent à l'analyse. Les études dans ce domaine proposent des techniques statistiques adaptées aux agrégats et résultent presque exclusivement en des analyses prédictives. Cependant, le mérite d'un tel groupe de recherches est de mettre en exergue la question de la représentativité des traces d'activité obtenues grâce aux données numériques et, en particulier, celle de la comparabilité entre les analyses issues de Twitter et les sondages d'opinion traditionnels. Même s'il a été démontré qu'il n'y a aucune bijection entre les comptes Twitter et les individus (Boyd et Crawford, 2012) et que des biais socio-économiques importants dans l'utilisation actuelle des médias numériques peuvent invalider leur représentativité, de nombreux chercheurs émettent encore « des doutes sur le fait que de tels biais peuvent ou non affecter les aptitudes *prédictives* de l'analyse des médias sociaux en comparaison avec les enquêtes traditionnelles hors ligne » (Ceron *et al*., 2014).

	Multi-échelle	**Expliquatif**	**Supervisé**
Opinion comme préférence	Aucun intérêt pour le multi-échelle : la recherche est uniquement focalisée sur l'analyse agrégative.	Aucun intérêt pour les méthodes explicatives : la recherche est uniquement focalisée sur la prédiction (pas besoin de modèles).	Aucun intérêt pour l'apprentissage supervisé : les lexiques basiques sont satisfaisants.
Opinion comme sentiment	Vers le multi-échelle : On observe des agrégats, mais aussi certains individus cruciaux qui ont une forte activité (ou une grande influence) sur la plateforme.	Vers l'explication : Étudier les interactions causales entre les opinions en ligne et hors ligne (et non seulement leurs corrélations).	Vers le supervisé : Développer des catégories de sentiment plus flexibles qui peuvent être partiellement induites par les données, mais soutenues par des modèles psychologiques.
Opinion comme interaction	Vers le multi-échelle : Utiliser le réseau comme un outil formel pour mesurer les propriétés structurelles multi-échelles des interactions et leurs relations inter-niveaux (comment le micro influence le macro et vice versa).	Vers l'explication : Développer des modèles mécaniques de la diffusion de l'opinion pour expliquer comment la structure du réseau est responsable des opinions observées (individuelles et collectives).	Vers le supervisé : Utiliser les mesures structurelles des réseaux, justifiées par des modèles de communication, mais suffisamment générales pour découvrir de nouveaux schémas d'interaction.
Opinion comme agenda	Vers le multi-échelle : Étudier la contribution d'utilisateurs spécifiques des médias sociaux à la configuration de l'agenda des médias de masse.	Intérêt limité pour les méthodes explicatives : La recherche est centrée sur l'analyse de l'agenda existant et sur la prédiction de l'organisation future de cet agenda.	Aucun intérêt pour l'apprentissage supervisé : Les techniques traditionnelles de fouille de textes suffisent.

Tableau 2. Relations entre les approches conceptuelles de l'opinion politique et les oppositions méthodologiques liées aux défis du big data.

Les articles uniquement basés sur le modèle conceptuel de la préférence sont toutefois très rares. Ce modèle est souvent combiné avec le modèle du sentiment. Or, dans les deux cas, lorsque l'opinion est considérée soit comme un sentiment, soit comme une préférence, les méthodes visant à prédire le résultat d'élections ou de sondages politiques pourraient accroître leur efficacité en se concentrant aussi sur les utilisateurs avec l'activité la plus forte sur la plateforme (ou avec la plus grande influence), ainsi que le suggèrent Stieglitz et Dang-Xuan (2012). De tels individus pourraient, en effet, constituer des prédicteurs assez puissants des tendances globales à un niveau microscopique. La combinaison d'une telle analyse qualitative

avec la prédiction quantitative nécessiterait cependant l'utilisation d'approches basées sur les réseaux afin d'identifier automatiquement ces individus clés. De la même manière, passer de la simple prédiction des résultats d'élection à leur explication requiert une meilleure compréhension des interactions causales entre la formation de l'opinion en ligne et celle de l'opinion hors ligne (et non pas uniquement de leurs corrélations statistiques). Comme indiqué par Ceron *et al*. (2014), on doit considérer « la question de la direction de la causalité », c'est-à-dire : « l'opinion sur les médias sociaux devient-elle plus similaire à l'opinion publique générale ou, au contraire, les médias sociaux conduisent-ils (ou anticipent-ils) l'opinion publique générale ? ». Pour ce faire, le développement de catégories de sentiment pertinentes est crucial, même si cela peut se révéler difficile dans le contexte de la communication ambiguë qui a lieu sur Twitter. Développer des catégories de sentiment plus flexibles, qui puissent être partiellement induites par les données et sous-tendues par des modèles linguistiques et psychologiques, pourrait aider à réduire le fossé entre des lexiques d'opinion trop rigides et les approches faiblement supervisées de l'apprentissage machine.

Mélanger les échelles d'analyse est particulièrement pertinent pour les chercheurs qui examinent la structure en réseaux de l'opinion politique. De fait, les approches basées sur les interactions impliquent souvent la combinaison de mesures microscopiques et macroscopiques. En général, ces approches proposent de nombreux outils formels pour mesurer les propriétés structurelles des interactions à différents niveaux, depuis leur micro-structure, comme, par exemple, les *hubs* et les ponts (*bridges*), à leurs macro-structures, telles que les communautés ou autres schémas de connectivité particuliers. Ce qu'il manque actuellement est une compréhension claire, théorique et empirique, des interconnections entre ces mesures microscopiques et macroscopiques. Par exemple, comment la présence de *hubs* et de *bridges*, correspondant aux influenceurs potentiels du point de vue des sciences de la communication, est-elle corrélée avec la connectivité globale ou la polarisation des opinions dans le réseau ? Par conséquent, les modèles de diffusion d'opinion à base d'agents, en particulier ceux développés par les sciences sociales computationnelles (voir partie 4.1), constituent une voie de recherche prometteuse pour expliquer comment de telles structures en réseaux (à la fois micro et macro) peuvent être responsables des opinions observées (à la fois individuelles et collectives). Pour ce faire, les mesures structurelles choisies doivent être véritablement motivées par les modèles de communication, tels que décrits dans Jürgens *et al*. (2011). Elles doivent aussi demeurer suffisamment génériques pour découvrir de nouveaux modes d'interaction, réalisant ainsi un compromis entre les approches purement hypothético-déductives, qui formalisent et testent les modèles de communication, et les approches inductives qui sont capables d'adapter ces structures d'analyse à des utilisations variées des médias numériques.

Les études qui se concentrent sur les agendas engendrés par les tweets ne tentent que rarement un mélange des échelles d'analyse. Soit elles se concentrent sur l'échelle macroscopique en identifiant des facteurs généraux capables d'influencer l'agenda (comme, par exemple, dans le cas des études sur la valeur médiatique des actualités), soit, à l'inverse, sur l'échelle microscopique en identifiant des individus spécifiques, comme des journalistes ou des politicien·ne·s, afin d'étudier la manière dont les représentations qu'ils·elles produisent peuvent influencer l'agenda des citoyens ou des médias. À l'heure actuelle, ces approches

n'étudient pas les deux niveaux simultanément, en fournissant par exemple un modèle clair de la manière dont l'agenda microscopique suivi par des journalistes et des politicien·ne·s impacte les agendas collectifs et macroscopiques et, inversement, la manière dont les agendas macroscopiques peuvent ou non produire un *feedback* descendant sur l'agenda microscopique. Pourtant, il est intéressant de noter que la multiplicité des échelles est techniquement possible. En effet, les études dans ce domaine pourraient bénéficier d'une meilleure compréhension des interactions informationnelles entre les médias de masse (exprimant un agenda politique descendant) et les médias numériques (construisant un agenda politique ascendant), démontrant ainsi la manière dont les agendas individuels et collectifs s'entremêlent, via l'interaction d'espaces de discussion différemment structurés. Pour ce faire, il est nécessaire de développer des méthodes capables de prendre systématiquement en compte les opinions politiques dans leurs différents contextes de production et de diffusion, et de les interpréter comme différents niveaux de production de l'opinion.

Conclusion

Cet article s'est emparé d'une tâche difficile concernant la vérification de la possibilité d'une analyse multi-échelle de l'opinion politique en s'appuyant sur les données numériques du big data, et plus précisément sur les données issues de Twitter. Une première étape a consisté en l'adoption d'un point de vue théorique nous permettant d'identifier quatre approches conceptuelles principales lorsqu'on s'intéresse aux opinions politiques : l'opinion comme une préférence, comme un sentiment, comme une interaction et comme un agenda. Nous avons ensuite exploré l'éventail des méthodes proposées par le passé pour combler le fossé entre approches qualitatives et approches quantitatives.

En premier lieu, qu'il s'agisse de la collecte des données (des enquêtes individuelles par auto-déclaration aux collectes de traces numériques à grande échelle) ou de leur traitement (de l'analyse contextualisée d'individus ou d'événements spécifiques à l'interprétation de tendances macroscopiques ou de dynamiques à long terme), nous avons montré que le concept d'« analyse multi-échelle » peut être largement bénéfique à la manière dont nous pensons ce défi méthodologique, en encourageant l'analyse simultanée de détails cruciaux et d'agrégats significatifs. Nous avons vu en particulier que les approches conceptuelles du « sentiment » ou de l'« interaction » requièrent de telles techniques multi-échelles.

En second lieu, nous avons essayé d'apaiser l'une des principales critiques concernant les méthodes informatiques relatives au big data, c'est-à-dire le manque de pouvoir explicatif de l'apprentissage machine non-supervisé et des méthodes purement statistiques en général. En effet, il est essentiel que les résultats du traitement de données ne se limitent pas à la découverte automatique de corrélations entre les variables explorées, mais qu'il fournisse également un modèle causal pour la correcte interprétation de ces corrélations au sein des sciences sociales. Nous avons vu que cela était d'autant plus possible lorsque des équipes interdisciplinaires, constituées de chercheurs en sciences sociales et en informatique[12], étaient

12 En effet, les big data et les *data sciences* ne changent pas les méthodes des sciences sociales, des statistiques et de l'informatique. Le changement principal concerne la prise de conscience de la nécessité et de l'intérêt de travailler dans des cadres interdisciplinaires sans négliger l'une de ces disciplines.

capables de faire le lien entre leurs résultats empiriques et des modèles préexistants en sciences de la communication et/ou en sciences politiques.

En troisième lieu, ces questionnements méthodologiques sont à mettre en relation avec le statut du savoir expert au sein des approches informatiques. À cet égard, nous avons concentré notre discussion sur le domaine spécifique de l'analyse des sentiments où de nombreuses propositions ont été formulées pour intégrer (ou pour écarter) le savoir linguistique lors de l'analyse algorithmique des tweets, menant à différents avantages et inconvénients en ce qui concerne l'interprétation qualitative des résultats.

Il va de soi que les bonnes pratiques identifiées ici demandent du temps et des ressources. Cela implique un projet à long terme et de longs délais dans la production de publications scientifiques. De plus, comme indiqué précédemment, ces études sont confrontées à d'importants questionnements éthiques. Tout ceci explique pourquoi, même si l'on trouve des centaines d'articles examinant les opinions politiques à partir des tweets, peu d'entre eux s'appuient sur des méthodes multi-échelles qui pourraient à la fois satisfaire les normes des sciences informatiques et celles des sciences sociales. Cependant, notre espoir repose sur le fait que notre étude aidera à progresser dans ce domaine spécifique en mettant en exergue l'importance de la connexion entre les choix conceptuels et méthodologiques, ainsi que leurs conséquences.

Bibliographie

ALLPORT F. H., 1937, « Toward a science of public opinion », *Public Opinion Quarterly*, 1, 1, p. 7-23.

ANDERSON C., 2008, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired*, 16:7.

BARBERÁ P., JOST J. T., NAGLER J., TUCKER J. A., and BONNEAU R., 2015, « Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? » *Psychological Science*, 26:10, p. 1531-1542.

BASTOS, M. T., RAIMUNDO, R. L. G., and TRAVITZKI, R., 2013, « Gatekeeping Twitter: message diffusion in political hashtags », *Media, Culture & Society*, 35, 2, p. 260-270.

BERMINGHAM A. and SMEATON A. F., 2011, « On Using Twitter to Monitor Political Sentiment and Predict Election Results », *Workshop Sentiment Analysis where AI meets Psychology (SAAIP) at the International Joint Conference for Natural Language Processing (IJCNLP)*.

BLONDIAUX, L., 1998, *La fabrique de l'opinion. Une histoire sociale des sondages*, Paris, Le Seuil.

BORGATTI, S. P., 2006, « Identifying sets of key players in a network », *Computational, Mathematical and Organizational Theory*, 12, 1, p. 21-34.

BOURDIEU P., 1973, « L'opinion publique n'existe pas », *Les Temps Modernes*, janvier, 318.

BOYADJIAN, J., 2014, *Analyser les opinions politiques sur Internet: Enjeux théoriques et défis méthodologiques,* Doctoral dissertation, Montpellier 1, France.

BOYD D. and CRAWFORD K., 2012, « Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon», *Information, Communication & Society*, 15, 5, p. 662-679.

CALUDE C. and LONGO G., 2016, « The Deluge of Spurious Correlations in Big Data », *Foundations of Science*, p. 1-18.

CASINI L., and MANZO G., 2016, « Agent-based models and causality: a methodological appraisal", working paper, Linköping University. Available at http://www.diva-portal.org/smash/get/diva2:1058813/FULLTEXT01.pdf.

CAYROL R., 2011, *Opinion, sondages et démocratie*, Paris, Sciences Po Presses.

CERON A., CURINI L., IACUS S. M., and PORRO G., 2013, « Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France », *New Media & Society*, 16, 2, p. 340-358.

CIOFFI-REVILLA C., 2016, « Bigger Computational Social Science: Data, Theories, Models, and Simulations—Not Just Big Data », *8th International ACM Web Science Conference (WebSci'16)*.

COLLEONI E., ROZZA A., and ARVIDSSON A., 2014, « Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter using Big Data », *Journal of Communication*, 64, p. 317-332.

CONOVER M. D., GONÇALVES B., RATKIEWICZ J., FLAMMINI A., and MENCZER F., 2011, « Predicting the Political Alignment of Twitter Users », *Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT 2011) and the IEEE Third International Conference on Social Computing (SocialCom 2011)*, p. 192-199.

DE ROSNAY J., 1975, *Le Macroscope, vers une vision globale*, Paris, Le Seuil.

DRISCOLL K. and WALKER S., 2014, « Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data », *International Journal of Communication,* 8, p. 1745-1764.

ENTMAN, R. M. and HERBST, S., 2001, « Reframing public opinion as we have known it », in W. BENNETT, L. and ENTMAN R. M., *Mediated politics: Communication in the future of democracy*, Cambridge, Cambridge University Press, p. 203-225.

GAYO-AVELLO, D., 2013, « A meta-analysis of state-of-the-art electoral prediction from Twitter data », *Social Science Computer Review*, 31, 6, p. 649-679.

GINSBERG, B. 1986, *The Captive Public: How Mass Opinion Promotes State Power*, New York, Basic Books.

HABERMAS, J., 1962, *L'espace public*, Paris, Edition Payot.

HAYASHI C., 1998, « What is Data Science? Fundamental Concepts and a Heuristic Example», in C. Hayashi et al., *Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer Japan. p. 40–51.

HE Y., SAIF H., WEI Z., and WONG K.-F., 2012, « Quantising Opinions for Political Tweets Analysis », *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, p. 3901-3906.

HOGAN, B., 2014, "From invisible algorithms to interactive affordances: Data after the ideology of Machine Learning" in E. Bertino, S. A. Matei (eds.), *Roles, Trust, and Reputation in Social Media Knowledge 1 Markets, Computational Social Sciences*, DOI 10.1007/978-3-319-05467-4_7.

HOLMES, D.E., 2017, *Big Data. A very Short Introduction*, Oxford, Oxford University Press.

JUNGHERR A., JÜRGENS P., and SCHOEN H., 2011, « Why the Pirate Party Won the German Election of 2009 or the Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With 140 Characters Reveal About Political Sentiment" », *Social Science Computer Review,* p 1-6.

JÜRGENS P., JUNGHERR A., and SCHOEN H., 2011, « Small Worlds with a Difference: New Gatekeepers and the Filtering of Political Information on Twitter », *Proceedings of the Third International Web Science Conference (WebSci 2011)*.

KATZ, E. and LAZARSFELD, P. E., 1955, *Personal influence: The part played by people in the flow of mass communication*, New York, Free Press.

KEY, V. O., 1966, *The Responsible Electorate: Rationality in Presidential Voting, 1936-1960,* with the Assistance of Milton C. Cummings, Cambridge, Harvard University Press.

LAMARCHE-PERRIN R., DEMAZEAU Y., and VINCENT J.-M., 2014, « Building Optimal Macroscopic Representations of Complex Multi-agent Systems », *Transactions on Computational Collective Intelligence*, vol. XV, LNCS 8670, p. 1-27.

LAZARSFELD, P. F., BERELSON, B., and GAUDET, H., 1948, *The peoples choice: how the voter makes up his mind in a presidential campaign*, New York, Duell, Sloan and Pearce.

LAZER D., PENTLAND A., ADAMIC L., ARAL S., BARABÁSI A.-L., BREWER D., CHRISTAKIS N., CONTRACTOR N., FOWLER J., GUTMANN M., JEBARA T., KING G., MACY M., ROY D., and VAN ALSTYNE M., 2009, « Computational Social Science », *Science*, 323, 5915, p. 721-723.

LIPPMANN, W., 1925, *The Phantom Public*, New York, Harcourt Brace.

LIVNE A., SIMMONS M. P., ADAR E., and ADAMIC L. A., 2011, « The Party is Over Here: Structure and Content in the 2010 Election », *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, p. 201-208.

MASAD D., 2014, « Computational social science, data, and complexity », *Bad Networking,* available at http://davidmasad.com/blog/css-data-complexity/

MARRES, N., 2017, *Digital Sociology. The Reinvention of Social Research*, Wiley.

MARRES, N., 2012, « The redistribution of methods: on intervention in digital social research », *The sociological review*, *60*, S1, p. 139-165.

MEI Q., LING X., WONDRA M., SU H., and ZHAI C., 2007, « Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs », *Proceedings of the 16th international conference on World Wide Web (WWW'07)*, p. 171-180.

MERCIER, A., 2012, *Médias et opinion publique*, Les Essentiels d'Hermès, Paris, CNRS éditions.

MERTON, R. K., 1968, « Patterns of Influence: Local and Cosmopolitan Influentials », in R. K. MERTON (eds.), *Social Theory and Social Structure*, New York, Free Press, p. 441-474.

METAXAS, P. T., MUSTAFARAJ, E. and GAYO-AVELLO, D., 2011, « How (not) to predict elections », *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*.

MOLYNEUX, L., 2015, « What journalists retweet: Opinion, humor, and brand development on Twitter », *Journalism*, 16, 7, p. 920-935.

NOELLE-NEUMANN, E., 1984, T*he Spiral of Silence. Public Opinion – Our Social Skin*, Chicago and London, University of Chicago Press.

O'CONNOR B., BALASUBRAMANYAN R., ROUTLEDGE B. R., and SMITH N. A., 2010, « From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series », *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, p. 122-129.

PAGE, B. I., and SHAPIRO R.Y., 1992, *The Rational Public: fifty years of Trends of Americans' Policy Preferences*, Chicago, University of Chicago Press.

PAK A. and PAROUBEK P., 2010, « Twitter as a Corpus for Sentiment Analysis and Opinion Mining », *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, p. 17-23.

PAPACHARISSI, Z., and DE FATIMA OLIVEIRA, M., 2012, « Affective news and networked publics: The rhythms of news storytelling on# Egypt», *Journal of Communication*, 62, 2, p. 266-282.

PARK C. S., 2013, « Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement », *Computers in Human Behavior*, 29, 4, p. 1641-1648.

PIGLIUCCI M., 2009, « The end of theory in science? », *EMBO reports*, 10:6.

PUNCH K., 1998, *Introduction to Social Research: Quantitative and Qualitative Approaches*, London, Sage.

REYNIÉ D. 1989, « Le nombre dans la politique moderne », *Hermès*, 4, Paris, CNRS edition, p.159-164.

SAVAGE M. and BURROWS R., 2007, « The coming crisis of empirical sociology », *Sociology,* 41, 5, p. 885–899.

SEVERO M., FEREDJ A., and ROMELE A., 2016, « Soft Data and Public Policy: Can Social Media Offer Alternatives to Official Statistics in Urban Policymaking? », *Policy & Internet*, 8, 3, p. 354-372.

SKORIC M., POOR N., ACHANANUPARP P., LIM E. P., and JIANG J. 2012, « Tweets and Votes: A Study of the 2011 Singapore General Election », *Proceedings of 45th Hawaii International Conference on Systems Science (HICSS-45 2012), IEEE Computer Society*, Los Alamitos, CA, USA, p. 2583-2591.

STIEGLITZ S. and DANG-XUAN L., 2012, « Political Communication and Influence through Microblogging – An Empirical Analysis of Sentiment in Twitter Messages and Retweet Behavior », *Proceedings of the Forty-fifth Hawaii International Conference on System Sciences (HICSS 2012)*, p. 3500-3509.

TARDE G., 1989 (1901), *L'Opinion et la foule*, Paris, Les Presses Universitaires de France.

TUMASJAN A., SPRENGER T. O., SANDNER P. G., and WELPE I. M., 2010, « Predicting elections with Twitter: What 140 characters reveal about political sentiment », *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, p.178-185.

VENTURINI, T., and LATOUR, B., 2010, « The social fabric: Digital traces and quali-quantitative methods », *Proceedings of Futur en Seine 2009*, p. 87-101.

WATTS, D. J., and DODDS, P. S., 2007, « Influentials, networks, and public opinion formation », *Journal of consumer research*, *34,* 4, p. 441-458.

WILSON, T., WIEBE, J. and HOFFMANN, P., 2005, « Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis », *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing*, p. 347–354.

XU, W. W., SANG, Y., BLASIOLA, S., and PARK, H. W, 2014, « Predicting opinion leaders in Twitter activism networks: The case of the Wisconsin recall election», *American Behavioral Scientist*, 58, 10, p.1278-1293.

ZALLER, J. R., 1992, *The Nature and Origins of Mass Opinion*, Cambridge, Cambridge University Press.