Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post

Julien Velcin, Jean-Claude Soulages, Solange Kurpiel, Luis Otávio Dias, Myrian del Vecchio, Frédéric Aubrun

▶ **To cite this version:**

Julien Velcin, Jean-Claude Soulages, Solange Kurpiel, Luis Otávio Dias, Myrian del Vecchio, et al.. Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post. Extraction et Gestion des Connaissances (EGC), Jan 2017, Grenoble, France. hal-01571265

HAL Id: hal-01571265
https://hal.science/hal-01571265

Submitted on 1 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post

Julien Velcin*, Jean-Claude Soulages**, Solange Kurpiel**,****,
Luis Otávio Dias****, Myrian Del Vecchio**,****, Frédéric Aubrun***

* Université de Lyon (Université Lyon 2, ERIC EA 3083)
Julien.Velcin@univ-lyon2.fr
http://mediamining.univ-lyon2.fr/velcin
** Université de Lyon (Université Lyon 2, Centre M. Weber)
Jean-Claude.Soulages@univ-lyon2.fr
solange.kurpiel@gmail.com
*** Université de Lyon (Université Lyon 2)
aubrunf@gmail.com
**** Université Fédérale du Paraná (Brésil)
myriandel@gmail.com
fototavio@yahoo.com.br

Résumé. Cet article présente un processus d'analyse mis en place dans le cadre d'une collaboration entre des chercheurs en informatique, en sociologie et en sciences de l'information et de la communication, à l'occasion du projet Journalisme A l'heure Du Numérique. Le processus consiste pour le moment en un recodage manuel de thématiques extraites de manière totalement non supervisée à partir des données textuelles publiées sur le site du Huffington Post. Cette démarche rend possible une analyse comparée d'un corpus d'articles publiés durant l'été 2016 dans trois éditions différentes du journal (française, américaine, brésilienne). Les premiers résultats présentés permettent de valider la démarche tout en interrogeant sur les améliorations possibles, en particulier une automatisation plus importante des étapes qui composent le processus.

1 Introduction

L'arrivée des nouvelles formes de communication virtuelle, qu'il s'agisse des blogs ou des médias sociaux, a révolutionné la manière dont l'information est publiée et diffusée. De nombreux métiers s'en trouvent bouleversés, comme celui de journaliste (Charon et Papet, 2014). La diversité de ces nouveaux lieux d'expression et leur caractère hautement concurrentiel a multiplié les points de vue sur l'actualité, ce qui conduit à nous interroger sur les choix qui y sont faits, que ceux-ci soient explicites (politique éditoriale) ou implicites (culture locale). Cette interrogation se pose avec d'autant plus d'acuité lorsque l'on observe l'information diffusée par un même média global, comme c'est le cas du Huffington Post abordé dans ce tra-

vail, développant des stratégies éditoriales locales (version américaine, française, brésilienne, arabe, etc.). Au-delà de la tension entre les stratégies éditoriales globales et locales, déjà mise en lumière par le biais de la publicité (Aubrun, 2015), d'autres critères peuvent intervenir à différents degrés dans le choix final qui sera réalisé, comme par exemple le coût élevé de la création d'un reportage original par rapport à la reproduction *ad nauseam* d'articles publiés ailleurs. Cet article présente un travail pluridisciplinaire pour capturer ces différences à l'aide d'outils de fouille de données et en faire une première analyse critique.

Afin d'être en mesure d'expliciter les biais dans l'information véhiculée par les médias, il est nécessaire de parvenir à comparer l'information à partir d'un référentiel commun malgré une différence parfois très importante dans les sujets traités et la langue employée. Dans ce travail, nous proposons d'articuler une méthode automatique qui extrait des thématiques de manière non supervisée directement à partir des données, spécifiques à chacun des corpus étudiés, avec un recodage à l'aide de catégories transversales, pour le moment réalisé de manière manuelle par des sociologues. L'intérêt est double : a) conserver une certaine flexibilité en permettant de ne pas être trop collé aux catégories proposées par les sites des médias, et donc permettre à des catégories nouvelles d'apparaître, b) garantir un ensemble de catégories de qualité validées par les experts et permettant de rendre les corpus comparables. Il s'agit donc d'une manière de construire un résumé associé à chaque média, et plus précisément dans notre cas à chaque version du même média, et ce de manière semi-automatique.

Ce travail représente la première étape du projet collaboratif Journalisme A l'heure Du Numérique [1] (JADN) qui mobilise des sociologues, spécialistes de l'analyse des médias, et des informaticiens, spécialistes de fouille de données. Une première application visée est de fournir aux journalistes ou aux chercheurs un outil permettant de visualiser la couverture de l'actualité. Un deuxième enjeu est d'ordre culturel ou géo-culturel puisqu'il s'agit de mettre au jour les points de convergence ou de divergence dans les différentes éditions d'un même média (ici, le Huffington Post). Les tout premiers résultats obtenus sur un corpus de 18 555 articles de presse, publiés sur le site du Huffington Post en trois langues (anglais, français, portugais) et sur une durée d'environ trois mois durant l'été 2016, sont présentés dans cet article. Ils permettent de donner un aperçu du "paysage informationnel" (voir Appadurai (2011) pour la notion de *mediascape*) des trois corpus et d'en réaliser une première analyse comparée. Outre une brève interprétation qualitative de ces résultats, ils mettent en lumière différentes problématiques que nous souhaitons développer dans nos travaux futurs. Nous prévoyons en particulier d'automatiser une grande partie de ce processus, encore très manuel, telles que la construction des catégories et la comparaison des résumés.

La suite de cet article est divisée en quatre parties. La première section décrit l'approche systémique que nous avons suivie, dans laquelle nous combinons des outils d'apprentissage automatique avec une analyse qualitative. La deuxième section présente les données récoltées, donnant une brève motivation concernant l'étude d'un média en particulier, avant de décrire le protocole expérimental et les premiers résultats obtenus. La troisième section donne quelques éléments bibliographiques, à la fois en analyse comparée des médias et en fouille de données à base de modèles thématiques. Enfin, la dernière section est consacrée à la conclusion et aux perspectives de ce travail.

1. http://jadn.univ-lyon2.fr

2 Approche proposée

La démarche décrite ci-dessous est le fruit d'un consensus qu'il a fallu trouver entre les deux équipes, comme c'est souvent le cas dans les projets pluridisciplinaires. Elle repose principalement sur l'idée de pouvoir remettre en question chacune des étapes, par exemple le nombre de thématiques choisies ou les catégories construites pour subsumer ces thématiques. Elle est composée de cinq étapes décrites ci-dessous :

— Constitution de plusieurs corpus composés d'articles publiés durant la même période. A ce stade, aucune hypothèse n'est formulée et l'intégralité des articles est récupérée pour les analyses ultérieures.

— Extraction automatique des thématiques en adoptant une approche de bas vers le haut, c'est-à-dire totalement non supervisée. A ce stade, plusieurs modèles sont proposés dans la littérature (voir section 4). Dans notre cas, nous avons choisi un modèle reconnu et pour lequel il convient de fixer le nombre de thématiques attendu (modèle LDA de Blei et al. (2003)). L'avantage est que l'on peut s'attendre à obtenir des thématiques de granularité comparable, même si ce ne sera pas toujours le cas comme nous l'avons constaté lors des expérimentations.

— Détection des thématiques jugées erronées car trop difficiles à interpréter. Cette détection se fait uniquement sur la base des 10 mots-clefs les plus importants. Il peut s'agir, par exemple, de thématiques regroupant des mots outils spécifiques au corpus ou des erreurs dans l'acquisition des données (par exemple, nous avions une thématique regroupant des balises javascript non détectées par l'algorithme d'acquisition).

— Attribution d'une catégorie à chaque thématique conservée, plusieurs thématiques pouvant partager la même catégorie (par exemple "politique étrangère" pour une thématique sur le coup d'état en Turquie ou sur le Brexit). Les catégories sont fixées par le sociologue à la lumière de la lecture des thématiques, pour le moment fournies sous la forme d'une liste des 10 mots-clefs les plus importants (donc ceux qui maximisent la probabilité $p(w/z)$, w étant un mot du dictionnaire et z la thématique en question). Il a été question de fournir les articles les plus centraux, donc ceux qui maximisent $p(d/z)$, mais il a été décidé de ne pas utiliser ce surcroît d'information pour le moment et d'étudier si les mots en tête de liste pouvaient s'avérer suffisants dans la compréhension du contenu abordé par la thématique. Pour fixer les catégories, la méthode suivante a été appliquée :

1. Sur la base des 10 mots-clefs d'une thématique, le spécialiste choisit une catégorie qui lui semble assez générale pour couvrir plusieurs thématiques (le flou sur ce que recouvre le terme "générale" est conservé à dessein). A ce stade, celui-ci n'est pas restreint et peut parfaitement créer de nouvelles catégories.
2. Le spécialiste peut revenir sur l'attribution de catégories à des thématiques précédentes, voire supprimer certaines catégories si celles-ci s'avèrent inappropriées (la catégorie "élections américaines" a finalement été exclue du corpus US).
3. Le spécialiste modifie les catégories des différents corpus afin de converger vers une unique liste de catégories qui permet d'encoder chaque thématique.
4. Il est possible à présent de compter, pour chaque corpus, le nombre de thématiques et d'estimer le volume d'articles pour chacune des catégories. Pour une catégorie donnée c, on connaît le nombre de thématiques associées et on peut estimer son

importance en calculant $f(c) = \sum_{z \in Z_c} \sum_{d \in D} p(z/d)$, où Z_c est l'ensemble des thématiques pour la catégorie visée c et d est un document de notre corpus D. On peut interpréter cette valeur comme le nombre d'articles publiés dans cette catégorie, bien qu'il s'agisse d'une estimation étant donné qu'un document est associé à plusieurs thématiques dans ce genre de modèle. Dans les résultats présentés dans cet article, nous avons arrondi cette estimation à l'entier le plus proche afin d'être cohérent avec cette interprétation que nous en donnons.

— Puisque les différents corpus sont codés à l'aide du même ensemble de catégories, il est à présent possible de les comparer sur la base des différentes valeurs de $f(c)$, comme nous le verrons dans le diagramme de la figure 3.

3 Expérimentations

3.1 Données

Dans ce travail, nous avons choisi de nous concentrer sur des articles de presse publiés sur le site du Huffington Post [2]. Ce média a l'avantage de répondre à un ensemble de préoccupations méthodologiques et thématiques en lien avec l'évolution du journalisme à l'ère du numérique. The Huffington Post est un *pure player* gratuit fondé aux États-Unis par Ariana Huffington, Kenneth Lerer et Jonah Peretti en 2005, racheté par AOL pour 315 millions de dollars américains en 2011 et décliné en 12 éditions à travers le monde, dont les éditions française (23 janvier 2012), et brésilienne (29 janvier 2014). Présenté à la presse et au public comme un média indépendant qui vise à révolutionner le journalisme grâce à une offre numérique « alternative » du point de vue technologique comme politique, le Huffington Post a fait de sa structure algorithmique, de son modèle économique comme de son panel de rédacteurs triés sur le volet, les leviers de son succès américain puis international. A mi-chemin entre un agrégateur d'informations et un producteur de nouvelles, il compte dans ses rangs des journalistes salariés rompus au travail de curation et des contributeurs extérieurs bénévoles, jouissant d'une grande notoriété. L'originalité de son modèle éditorial repose sur un spectre très large d'articles d'actualité mais surtout sur des articles d'opinion sous forme de blogs signés par des experts ou des personnalités connues comme Barak Obama, David Cameron, Michael Moore, Rachida Dati, Madona, etc. Son positionnement à la fois global et local constitue une formidable opportunité pour étudier le poids de la globalisation dans la production de l'information à travers la distribution et la reprise de nouvelles et de blogs par chacune des éditions nationales.

Afin d'étudier ces données, nous avons mis en place un aspirateur de flux RSS sur quatre versions du Huffington Post (américaine, française, brésilienne, arabe) à partir de juin 2016, à la fois sur les articles de presse que sur les articles de blog. Chaque article est associé à un titre, au corps de l'article, à une date et à un auteur. Les résultats présentés dans cet article correspondent à presque trois mois de collecte (plus précisément du 20 juin au 8 septembre 2016) et se concentrent uniquement sur les trois premières versions et sur les articles de presse. Le tableau de la figure 1 donne des statistiques simples sur les corpus étudiés.

2. http://huffingtonpost.com

Version	langue	#articles	longueur	#mots
US	anglais	12 067	454.4	5 482 661
FR	français	4 133	369.6	1 527 416
BR	portugais	2 355	429.5	1 011 373

FIG. 1 – *Brève description des trois corpus constitués pour notre analyse. #articles donne le nombre d'articles publiés, leur longueur moyenne et #mots le nombre total de mots dans le corpus brut (c'est-à-dire non pré-traité, voir section 3.2).*

3.2 Protocole expérimental

Pour le moment, l'approche est divisée en deux étapes : une étape automatique, dans laquelle on extrait les thématiques pour chacun des corpus, et une étape manuelle qui consiste à étudier les thématiques pour leur associer une catégorie transversale à tous les corpus avant de les analyser. Nous discutons dans la section 5 le projet d'intégrer davantage ces deux étapes.

Concernant la partie automatique sur l'extraction des thématiques, nous avons utilisé le modèle LDA (Blei et al., 2003) dans son implémentation parallèle disponible via la librairie MALLET [3]. Les hyper-paramètres du modèle α et β, correspondant respectivement à l'à-priori sur les thématiques et sur les documents, sont symétriques et ont été estimés automatiquement à partir des données (Mccallum et al., 2009). Après plusieurs tentatives manuelles, nous avons choisi de fixer le nombre k de thématiques à 100. D'autres valeurs peuvent évidemment s'avérer pertinentes, mais nous pensons que ce n'est pas tellement important dans la mesure où nous effectuons une analyse comparée de plusieurs corpus. Le nombre d'itérations pour l'estimation avec un échantillonnage de Gibb's a été fixé à 2000, comme préconisé par les créateurs de la librairie. L'algorithme a été exécuté sur une version légèrement nettoyée des données : lettres mises en minuscules, suppression de la ponctuation au début et à la fin des mots (on conserve ainsi "qu'il" ou "sang-froid"), suppression des mots n'apparaissant que dans un seul document et suppression des mots-outils habituels.

La partie concernant l'annotation manuelle a été réalisée par trois chercheurs en sociologie et en sciences de l'information et de la communication. Le processus a été réalisé de manière individuelle dans un premier temps (une personne sur le corpus en anglais et en français, les deux autres sur le corpus brésilien) puis de manière collective afin de converger vers un consensus, à la fois sur les catégories choisies et sur l'annotation. A savoir qu'il était possible qu'une catégorie ne soit présente que pour un ou deux corpus. Finalement, le processus collectif a convergé vers la définition de 15 catégories pour coder l'ensemble des corpus.

3.3 Résultats obtenus

Le tableau de la figure 2 présente un extrait de 5 thématiques, sur les 100 thématiques extraites à partir de chacun des trois corpus.

Nous constatons tout d'abord qu'une majeure partie des thématiques a été conservée car jugée suffisamment pertinente par les sociologues. Plus précisément : 15 ont été écartées en français, 16 en anglais et 16 en portugais. En effet, la plupart des thématiques peuvent être

3. http://mallet.cs.umass.edu/

topic	#doc	cat.	mots les plus probables
			en français (sur 4133 articles) :
z_{18}	28	1	manifestation, paris, police, travail, loi, contre, syndicats, place, bastille, 2016
z_{19}	36	1	loi, travail, gouvernement, l'état, texte, l'assemblée, d'urgence, mois, projet, conseil
z_{25}	39	2	jeux, rio, olympiques, olympique, août, jo, athlètes, 2016, brésil, cérémonie
z_{47}	18	3	morandini, jean-marc, inrocks, catherine, l'animateur, lui, qu'il, europe, comédiens, plainte
z_{73}	47	4	nice, 14, l'attentat, anglais, promenade, camion, attentat, police, soir, christian
			en anglais (sur 12067 articles) :
z_{14}	92	5	refugees, children, refugee, people, countries, world, syrian, rights, million, year
z_{21}	74	2	gymnastics, biles, olympic, team, simone, olympics, gymnast, gold, rio, hernandez
z_{3}	46	6	pokemon, game, pokémon, playing, players, catch, «pokemon, go», pizza, play
z_{50}	56	7	muslim, religious, muslims, faith, church, god, christian, religion, hate, american
z_{27}	140	8	clinton, voters, trump, poll, polls, americans, election, support, vote, relationships
			en portugais (sur 2355 articles) :
z_{44}	52	8	dilma, presidente, impeachment, senado, senadores, processo, senador, rousseff, julgamento, defesa
z_{58}	7	9	sexo, menstruação, durante, rao, mccane, comédia, realmente, corpo, riso, menstruada
z_{71}	11	7	negros, brancos, negras, pessoas, racial, negra, racismo, país, movimento, black
z_{37}	57	2	brasil, vôlei, jogo, medalha, vitória, ouro, seleção, set, brasileiras, torcida
z_{99}	20	7	lgbt, gay, preconceito, violência, sexual, direitos, família, orgulho, estupro, aborto

FIG. 2 – *Extrait des thématiques trouvées dans les trois langues. Les catégories attribuées ici (cat.) correspondent à : 1- Economie / Social, 2- Sport / JO, 3- Show business / people, 4- Sécurité / attentats, 5- Politique étrangère, 6- Technologie / science, 7- Société, 8- Politique nationale, 9- Santé. Les 15 catégories sont visibles dans le graphique de la figure 3.*

interprétées aisément sur la base des premiers mots clefs, comme illustré dans le tableau (nous avons omis les scores pour des raisons de lisibilité). Ce résultat n'est pas surprenant en lui-même car les données en entrée sont de bonne qualité, si on les compare par exemple à des tweets. Il s'agit d'articles de presse souvent bien écrits et d'une longueur suffisante pour en déduire des régularités statistiques. Un examen plus attentif des thématiques rejetées montre cependant qu'une partie d'entre elles (7 dans le cas français) pourrait en réalité être conservées en retournant consulter les données pour faciliter leur interprétation. Dans notre cas, la thématique relative à l'explosion ayant eu lieu à l'aéroport Atatürk en Turquie a ainsi été écartée peut-être un peu rapidement. Nous constatons également une certaine diversité dans les thématiques qui traitent parfois de sujets très génériques (la famille, le cancer ou les images au sens de photographie) ou d'événements très précis (l'attentat de Nice ou la rumeur au sujet du dernier Iphone). Certains sujets traités de manière prépondérante (par exemple les Jeux Olympiques) se retrouvent découpés en plusieurs thématiques (dans notre exemple, une médaille d'or pour la natation, les jeux paralympiques, etc.), ce qui pose des questions sur le caractère homogène du niveau de granularité de l'information analysée.

Ensuite, nous pouvons calculer la distribution de chacun des corpus sur les catégories mises en place. La figure 3 donne une représentation graphique de cette distribution qui a été normalisée afin de gommer la différence dans le volume des articles publiés (trois fois plus d'articles pour la version US que pour la version française, six fois plus que la version brésilienne). Il

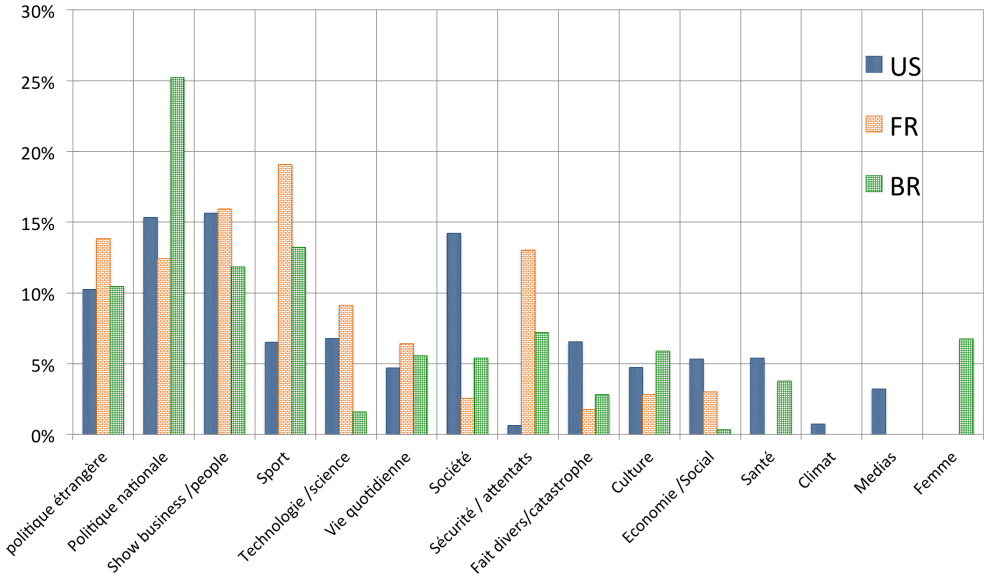

FIG. 3 – *Distribution normalisée des trois éditions sur les 15 catégories, chaque catégorie pouvant être associée à plusieurs thématiques.*

faut rappeler que chaque catégorie est composée de plusieurs thématiques. Par exemple, la politique étrangère est composée de 13 thématiques dans les éditions US et française, alors que le sport est composé de respectivement 12 et 5 thématiques pour ces mêmes éditions.

On peut tirer de ce graphique plusieurs observations préliminaires qui pourront être étudiées plus en détails par la suite. Par exemple, on observe sans grande surprise que la formule mise en place au Huffington Post conduit à mettre en tête la catégorie 'Show business / people' en tête, ou presque si on ne prend pas en compte la catégorie Sport qui est conjecturelle avec la couverture des Jeux Olympiques. L'autre caractéristique de la ligne éditoriale est le spectre très large de la couverture de l'actualité. Sur la politique étrangère, un examen détaillé permet de constater qu'elle est orientée vers les mêmes thématiques : la guerre en Syrie et Irak, les attentats pour la France.

Les écarts structurels entre les deux éditions française et US, qui ne sont que des hypothèses à ce stade de notre travail, sont les suivants. Tout d'abord, il n'existe pas de rubrique Santé dans l'édition française, ainsi que de catégories dédiées au climat ou aux médias, contrairement à l'édition US. Un premier examen des articles associés aux thématiques correspondantes fait ressortir une attention spéciale portée, aux Etats-Unis, aux maladies et, de manière générale, à toutes les manières de "mieux vivre". L'écart constaté sur la rubrique 'Fait divers / catastrophe' s'explique en partie par une place importante des faits divers violents (fusillade, etc.) dans l'actualité américaine. Ces analyses méritent, bien sûr, d'être approfondies et de nombreuses améliorations possibles à court terme sont discutées en section 5.

Comme pour les autres éditions, l'agenda du Huffington Post brésilien respecte l'agenda

médiatique événementiel traditionnel. Les thématiques les plus traitées sont la politique nationale et le sport. La période est marquée par la couverture du processus de destitution de la présidente Dilma Rousseff et des Jeux Olympiques. Au Brésil, la catégorie Show business / people apparaît aussi en tête de liste. La programmation télévisuelle et ses célébrités s'avèrent la principale source de *softnews* du Huffington Post. Notons aussi la place donnée aux problématiques autour du genre (catégorie "femme", mais également "société"), marquées par un discours de dénonciation des violences et infractions des droits.

Cette indexation supervisée par les chercheurs en SHS met en évidence l'arbitraire de toute démarche de catégorisation. Ainsi, les rubriques 'Vie quotidienne' et 'Société' qui cherchent à discriminer délibérément "monde vécu" et "espace public", en suivant Habermas (1987), représentent *de facto* des catégories par défaut, générées par le regroupement d'événements ou de situations échappant à des classifications socio-politiques stables (politique, économie, étranger, etc.). Ces deux catégories génériques sont à développer à moyen terme pour les distribuer en nouvelles thématiques en recourant à des allers-retours entre le niveau macro et les articles eux-mêmes et en isolant éventuellement certains marqueurs sémantiques.

4 Travaux connexes

4.1 Analyse de corpus médiatiques

Les recherches portant sur l'analyse du discours des médias confrontent les chercheurs en sciences humaines à un certain nombre de problèmes méthodologiques. Ils s'accordent sur le fait que la question de la définition d'un évènement ou d'un *thème* évènement pose d'emblée le problème de leur labilité et de leur volatilité ; complexité à laquelle il faut bien ajouter tous les phénomènes de sérialisation qui constituent une des principales caractéristiques de la communication de masse.

Tout corpus qui se voudrait un tant soit peu représentatif prend rapidement l'apparence d'une nébuleuse d'occurrences qu'il s'avère tout simplement difficile de circonscrire. Si l'on écarte une attitude volontariste qui viserait à construire ce dernier en fonction d'hypothèses externes suffisamment fortes, comme c'est le cas pour certaines controverses (Chateauraynaud, 2011), il reste à scruter ce qui demeure observable et tenter alors d'extraire de ce flux des données objectivables. Et, pour ce faire, s'il n'en existe pas, d'envisager l'élaboration de nouveaux instruments d'observation et d'assumer l'option d'une interdisciplinarité "focalisée" avec d'autres sciences (Charaudeau, 2010).

C'est cette instrumentalisation de l'observation et l'étayage de l'analyse à partir de données objectivées à travers de vastes corpus représentatifs qui ont constitué un des principes fondateurs de certaines des recherches précédentes dans le cadre de l'Inathèque de France portant sur les émissions de paroles ou le discours de l'information (Soulages, 2015; Soulages et Lochard, 2016). Ces travaux ont donné lieu à la création de base de données relationnelles dédiées à différents corpus de productions médiatiques et à l'élaboration d'un outil informatique destiné aux chercheurs en SHS (Médiacorpus).

Cette démarche, au départ non supervisée qui repose sur le traitement automatisé de flux de news, met au jour des corrélations ou des facteurs récurrents invisibles à l'observation non appareillée de l'analyste de discours. De plus, des aller retour fréquents entre les résultats produits par l'indexation automatique des données et des séquences d'analyse contextualisées

et compréhensives à l'initiative du chercheur en SHS, visent à la fois à améliorer l'efficacité de l'outil de fouille des données mais aussi à affiner l'analyse qualitative des discours médiatiques étudiés.

4.2 Agrégation par modèles thématiques

L'analyse comparée de plusieurs corpus médiatiques a donné lieu à de nombreux travaux en SHS. Lorsqu'on cherche à faire une analyse un tant soit peu quantitative, celle-ci nécessite l'annotation manuelle de nombreuses ressources (Powers et Benson, 2014). L'utilisation de thématiques peut être vue comme un *proxy* permettant de faciliter cette comparaison. En effet, il est clairement moins coûteux d'annoter un nombre limité de thématiques en lieu et place de l'intégralité du corpus, voire même un sous-ensemble de ce corpus qui nous ramènerait à l'apprentissage supervisé dont nous avons déjà souligné les limitations. Ces techniques automatiques ont déjà été testées dans la pratique du journalisme (Rusch et al., 2013; Günther et Quandt, 2016) mais, à notre connaissance, jamais dans une perspective inter-corpus, qui plus est rédigés en plusieurs langues.

L'extraction des thématiques est associée à une littérature importante en informatique depuis le projet *Topic Detection and Tracking* initié par Allan et al. (1998). Elle peut se baser sur des méthodes de clustering géométrique similaires aux k-moyennes (Velcin et Ganascia, 2007), mais plus souvent sur des factorisations de matrices, comme dans le cas de LSA (Deerwester et al., 1990) ou NMF (Paatero et Tapper, 1994), et sur des approches probabilistes, avec des modèles comme pLSA (Hofmann, 1999) ou LDA (Blei et al., 2003). Bien que voisines des techniques de clustering plus traditionnelles qui attribuent une unique catégorie à un texte, il ne faut pas confondre les deux tâches (Velcin et al., 2016). Lorsque les textes sont suffisamment longs, comme c'est le cas ici avec les articles de presse, il nous semble important de pouvoir associer plusieurs thématiques à un même texte. Cela ouvre également des perspectives très intéressantes pour étudier les liens existants entre les thématiques, et ainsi estimer une structure entre celles-ci et améliorer la compréhension du corpus.

Dans notre travail, nous avons choisi le modèle LDA pour sa popularité et sur la base de ses bons résultats obtenus dans de nombreuses applications (Hall et al., 2008), bien que d'autres modèles pourraient être choisis sans remettre en cause l'approche que nous proposons. De plus, des variantes permettent de prendre en compte des dimensions additionnelles, telles que les auteurs (Rosen-Zvi et al., 2004) ou le temps (Wang et McCallum, 2006). Cela ouvre donc de nombreuses perspectives pour prendre en compte ces dimensions complémentaires dans notre analyse. Malgré cela, nous souhaitons mettre au point une démarche qui ne dépend pas de la nature du modèle employé, que ce dernier soit basé sur des similarités, de l'algèbre linéaire ou des modèles probabilistes.

5 Conclusion et perspectives

Dans cet article, nous présentons une première analyse comparée des trois éditions différentes d'un média international en ligne. Les travaux étant encore dans leur première phase, nous avons mis l'accent sur la méthodologie déployée et donné uniquement quelques résultats préliminaires qui demandent à être approfondis. On constate cependant clairement plusieurs éléments saillants, telles que la couverture de nombreuses thématiques quelque soit l'édition,

les catégories communes (la politique, le show business), mais aussi les spécificités locales (les attentats ou le sport en France, la santé ou les thèmes de société aux USA, les questions de genre au Brésil). Ce travail nous a permis de mettre au jour plusieurs points faibles qui sont autant de perspectives à court terme que nous envisageons d'explorer.

Tout d'abord, l'élaboration des catégories reste un processus manuel qui ne peut être totalement traité comme une tâche de classification supervisée. En effet, il nous paraît évident qu'il est risqué de se reposer sur les catégories à priori proposées sur le site du (ou des) média(s), en particulier dans une optique diachronique. Pour un même média, ces catégories changent régulièrement en fonction de l'actualité, mises à part peut-être certaines d'entre elle, et celles-ci sont rarement les mêmes d'un média à l'autre. D'un autre côté, il est nécessaire d'avoir un socle commun pour réaliser une comparaison et l'opération ne peut pas être totalement réalisée sur la base des thématiques. Les modèles proposés par Paul et Girju (2009) ou Chen et al. (2015) vont dans ce sens mais ils ne sont pas aisément extensibles et, pour le moment, dédiés à une seule langue. Une approche semi-supervisée, dans laquelle il serait possible de combiner des catégories à priori avec des catégories émergentes semblerait plus appropriée, dans un esprit intégré discuté par Chuang et al. (2014). La prise en compte d'à priori provenant des experts sous la forme de collections de mots-clefs, comme dans les travaux de Newman et al. (2011) ou Hu et al. (2014), paraît une bonne piste pour cela. Nous envisageons également de recourir à des techniques d'alignement entre thématiques, rendues plus difficiles ici du fait de la diversité des langues employées. Bien que les résultats sur trois éditions aient été présentés ici, nous visons toujours l'intégration de l'édition arabe du Huffington Post. Contrairement aux travaux réalisés par Mimno et al. (2009), nous n'avons pas ici les mêmes articles accessibles dans plusieurs langues et une plus grande hétérogénéité en fonction des sources.

Ensuite, une autre perspective qui nous paraît importante consiste à aider l'expert à comprendre le sens des thématiques, ce qui lui permet ensuite d'attribuer plus facilement des catégories et de réaliser ainsi son analyse. C'est pourquoi nous travaillons actuellement à l'intégration de techniques de nommage des thématiques (*topic labeling*, voir Mei et al. (2007)). Cet étiquetage automatique des catégories thématiques, qu'il opère au niveau des thématiques initiales ou des grandes catégories, peut reposer par exemple sur des n-grams. Ainsi, la thématique z_{18} (France) du tableau 2 peut assez facilement être étiquetée par les termes "loi travail" et/ou "manifestation contre la loi". De la même façon, la thématique z_3 (US) peut être étiquetée par "augmented reality game" et/ou "pokemon go game". Au-delà d'un titre compréhensible (Lopez et al., 2014), différentes techniques de recherche d'information et de résumé automatique peuvent être mobilisées pour donner à l'expert une bien meilleure compréhension de la cohérence issue des régularités statistiques.

Enfin, nous prévoyons d'automatiser l'identification des thématiques non pertinentes en ayant recours aux indices de qualité développés dans la littérature (Röder et al., 2015). Nous avons également comme objectif de comparer l'information issue des articles de presse de celle publiée sur les billets de blogs hébergés par le Huffington Post, mais également d'intégrer l'édition arabe et d'étendre l'analyse pour prendre en compte la dimension temporelle (comment les thématiques ou les catégories évoluent-elles dans le temps ?).

Références

Allan, J., J. G. Carbonell, G. Doddington, J. Yamron, et Y. Yang (1998). *Topic detection and tracking pilot study final report*.

Appadurai, A. (2011). Disjuncture and difference in the global cultural economy 1990. *Cultural theory : An anthology 2011*, 282–295.

Aubrun, F. (2015). *Crise(s), publicité et marque : L'émergence de nouveaux modèles*. Ph. D. thesis, Université Lumière Lyon 2.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*, 993–1022.

Charaudeau, P. (2010). Pour une interdisciplinarité "focalisée" dans les sciences humaines et sociales. *Questions de communication* (17), 195–222.

Charon, J.-M. et J. Papet (2014). *Le journalisme en questions : réponses internationales*. Paris : L'Harmattan.

Chateauraynaud, F. (2011). *Argumenter dans un champ de forces. Essai de balistique sociologique*. Paris : éditions Petra.

Chen, C., W. Buntine, N. Ding, L. Xie, et L. Du (2015). Differential topic models. *IEEE transactions on pattern analysis and machine intelligence 37*(2), 230–242.

Chuang, J., J. D. Wilkerson, R. Weiss, et al. (2014). Computer-assisted content analysis : Topic models for exploring multiple subjective interpretations. In *Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning*.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, et R. A. Harshman (1990). Indexing by latent semantic analysis. *JASIS 41*(6), 391–407.

Günther, E. et T. Quandt (2016). Word counts and topic models : Automated text analysis methods for digital journalism research. *Digital Journalism 4*(1), 75–88.

Habermas, J. (1987). *Théorie de l'agir communicationnel*. Paris : Fayard.

Hall, D., D. Jurafsky, et C. D. Manning (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 363–371. Association for Computational Linguistics.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on Uncertainty in Artificial Intelligence*, pp. 289–296. Morgan Kaufmann.

Hu, Y., J. Boyd-Graber, B. Satinoff, et A. Smith (2014). Interactive topic modeling. *Machine learning 95*(3), 423–469.

Lopez, C., V. Prince, et M. Roche (2014). How can catchy titles be generated without loss of informativeness ? *Expert Systems With Applications (ESWA) 41*(4), 1051–1062.

Mccallum, A., D. M. Mimno, et H. M. Wallach (2009). Rethinking lda : why priors matter. In *Advances in Neural Information Processing Systems*, pp. 1973–1981.

Mei, Q., X. Shen, et C. Zhai (2007). Automatic labeling of multinomial topic models. In *ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pp. 490–499.

Mimno, D., H. M. Wallach, J. Naradowsky, D. A. Smith, et A. McCallum (2009). Polylingual topic models. In *Proceedings of the ACL Conference on Empirical Methods in Natural*

Language Processing : Volume 2, pp. 880–889.

Newman, D., E. V. Bonilla, et W. Buntine (2011). Improving topic coherence with regularized topic models. In *Advances in neural information processing systems*, pp. 496–504.

Paatero, P. et U. Tapper (1994). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*(2).

Paul, M. et R. Girju (2009). Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the ACL conference on Empirical Methods in Natural Language Processing : Volume 3*, pp. 1408–1417.

Powers, M. et R. Benson (2014). Is the internet homogenizing or diversifying the news ? External pluralism in the US, Danish, and French press. *The International Journal of Press/Politics 19*(2), 246–265.

Röder, M., A. Both, et A. Hinneburg (2015). Exploring the space of topic coherence measures. In *Proceedings of the ACM international conference on Web Search and Data Mining*.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, et P. Smyth (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494. AUAI Press.

Rusch, T., P. Hofmarcher, R. Hatzinger, K. Hornik, et al. (2013). Model trees with topic model preprocessing : An approach for data journalism illustrated with the wikileaks afghanistan war logs. *The Annals of Applied Statistics 7*(2), 613–639.

Soulages, J.-C. (2015). Le traitement télévisuel du sida en france : apprivoisement du fléau, instrumentation des médias. In I. Pânzaru et D. Popescu-Jourdy (Eds.), *Nouvelles approches de la rationalité. Défis contemporains des sciences humaines et sociales*, pp. 229–257. PUL.

Soulages, J.-C. et G. Lochard (2016). Comment la télévision traite la laïcité. In *La laïcité dans l'arène médiatique. Cartographie d'une controverse sociale*, pp. 95–115. INA éditions.

Velcin, J. et J.-G. Ganascia (2007). Topic extraction with AGAPE. In *Advanced Data Mining and Applications*, pp. 377–388. Springer.

Velcin, J., M. Roche, et P. Poncelet (2016). Shallow text clustering does not mean weak topics : how topic identification can leverage bigram feature. In *Proceedings of DMNLP, Workshop at ECML/PKDD*, Riva del Garda, Italy, pp. 25–32.

Wang, X. et A. McCallum (2006). Topics over time : a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433. ACM.

Summary

This article presents a joint work undertaken by a group of researchers in computer science, sociology and communication and information sciences, in the context of the digital journalism initiative. In the design of a new analysis process, we build a manual encoding upon a first automatic, unsupervised topic extraction from textual data provided on the Huffington Post website. This approach makes possible a comparative study of news published during the summer 2016 in three editions of the journal (French, English, Brazilian). The first presented results validate the whole process but gives room for improving many steps of the process.