Médias traditionnels, médias sociaux : caractériser la réinformation

Cédric Maigrot, Ewa Kijak, Vincent Claveau

▶ **To cite this version:**

Cédric Maigrot, Ewa Kijak, Vincent Claveau. Médias traditionnels, médias sociaux : caractériser la réinformation. TALN 2016 - 23ème Conférence sur le Traitement Automatique des Langues Naturelles, Jul 2016, Paris, France. hal-01349871

HAL Id: hal-01349871

https://inria.hal.science/hal-01349871

Submitted on 29 Jul 2016

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.

Médias traditionnels, médias sociaux : caractériser la réinformation

Cédric Maigrot[1] Ewa Kijak[1] Vincent Claveau[2]
(1) IRISA - Université de Rennes (2) IRISA - CNRS, Rennes
{prénom}.{nom}@irisa.fr

RÉSUMÉ

Les médias traditionnels sont de plus en plus présents sur les réseaux sociaux, mais ces sources d'informations sont confrontées à d'autres sources dites de réinformation. Ces dernières ont parfois tendance à déformer les informations relayées pour correspondre aux idéologies qu'elles souhaitent défendre, les rendant partiellement ou totalement fausses. Le but de cet article est, d'une part, de présenter un corpus que nous avons constitué à partir de groupes Facebook de ces deux types de médias. Nous présentons d'autre part quelques expériences de détection automatique des messages issus des médias de réinformation, en étudiant notamment l'influence d'attributs de surface et d'attributs portant plus spécifiquement sur le contenu de ces messages.

ABSTRACT

Traditional medias, social medias : characterizing reinformation

Traditional media are increasingly present on social networks, but these usual sources of information are confronted with other sources called *of reinformation*. These last ones sometimes tend to distort the relayed information to match their ideologies, rendering it partially or totally false. The aim of this study is twofold : first, we present a corpus containing Facebook messages issued from both types of media sources. Secondly, we propose some experiments in order to automatically detect reinformation messages ; in particular we investigate the influence of shallow features versus features more specifically describing the message content.

MOTS-CLÉS : Réseaux sociaux, médias, hoax, réinformation, classification de textes.

KEYWORDS: Social network, medias, hoax, reinformation, text classification.

1 Introduction

À l'heure où les réseaux sociaux développent une capacité à remplacer les journaux [1], ces derniers tentent de garder leur place en augmentant leur diffusion et présence par l'intermédiaire des réseaux sociaux. Ils se retrouvent ainsi opposés, entre autres, à des sources d'informations dites de *réinformation*. Les médias de réinformation véhiculent des informations différentes des médias de masse traditionnels. Comme médias alternatifs [2], ils se démarquent par une ligne éditoriale prononcée et proposent ainsi des articles allant dans cette dernière, parfois à contre-courant des médias de masse (e.g *TF1*,*Le Monde*, ...).

1. www.lemonde.fr/pixels/article/2016/03/19/les-reseaux-sociaux-sont-plus-puissants-que-les-medias_4886122_4408996.html
2. https://fr.wikipedia.org/wiki/MÃl'dia_alternatif

Ces dernières se présentent comme des alternatives aux médias traditionnels, mais les sites d'analyses de rumeurs et de faux (eg. `hoaxbuster.com`, `hoax-slayer.com`) montrent leur tendance à diffuser des informations en les interprétant à leur manière, voire en les modifiant de telle sorte qu'elles défendent les opinions du média (e.g politiques, religieuses).

Dans cette étude, nous nous intéressons uniquement à déterminer dans quelle mesure les messages écrits sur les réseaux sociaux par les médias traditionnels diffèrent de ceux des groupes de réinformation. En effet, il existe des différences entre les deux structures, telles que le statut professionnel des médias traditionnels par exemple. Il s'agit de savoir si de telles différences existent également dans la manière de publier l'information. Ce travail de caractérisation de la source s'inscrit comme une étape intermédiaire dans un projet plus vaste de détection automatique des fausses informations sur les réseaux sociaux, en temps réel. Il est important de noter que l'objectif est donc de caractériser des messages des réseaux sociaux afin de déterminer automatiquement leur provenance (média traditionnel ou de réinformation) et non pas, à ce stade, leur véracité ou objectivité. Ce faisant, nous étudions l'influence de différents descripteurs en distinguant ceux de surface (e.g longueur du message, etc.) de ceux portant sur le contenu du message.

L'article est organisé comme suit. La section 2 présente les travaux réalisés en lien avec notre étude. La section 3 présente le corpus que nous avons constitué et les descripteurs utilisés. La section 4 présente le protocole expérimental et enfin la section 5, les résultats obtenus.

2 État de l'art

Ces dernières années, l'analyse des informations circulant dans les réseaux sociaux a donné naissance à plusieurs projets européens. Le projet *PHEME* (Derczynski & Bontcheva, 2014) a pour but de créer un programme de *fact checking* automatique. Plusieurs de leurs travaux étudient les liens entre messages sur les réseaux sociaux : (Declerck & Lendvai, 2015) pour une analyse des hashtags et (Derczynski *et al.*, 2015) pour une analyse des entités nommées présentes dans les tweets. Dans notre cas, nous souhaitons pouvoir définir la classe d'appartenance d'un message par son contenu et non par les liens qu'il possède avec d'autres messages classés.

Un second projet européen, nommé *Reveal Project* (Middleton, 2015b), porte sur l'analyse de différents médias tels que l'image (Zampoglou *et al.*, 2015), la vidéo (Middleton, 2015a), ainsi que le texte. Les études réalisées sur la sémantique du texte se basent sur l'ajout d'informations issues de bases de connaissances libres (Gottron *et al.*, 2014; Kordopatis-Zilos *et al.*, 2015). Pour notre part, nous souhaitons nous baser uniquement sur le contenu proposé par le média.

InVid[3] est un projet européen qui a pour but de créer un système de détection automatique de fausses vidéos, en travaillant sur des images issues des vidéos analysées. Nous nous concentrons sur les textes et les images, mais la comparaison de notre approche avec leurs résultats, basés sur des ressources différentes d'un même message de réseau social, pourrait apporter des caractéristiques complémentaires à celles étudiées dans cet article.

D'autres travaux étudient l'arrivée des médias traditionnels sur les réseaux sociaux, notamment sur le changement d'écriture que cela peut induire (Alejandro, 2010; Kwak *et al.*, 2010). Par exemple, (Sharma, 2015) s'intéresse à la spécificité du cyber-journalisme espagnol. L'auteur cherche à caractériser les messages écrits sur Facebook par des journalistes, par opposition à ceux écrits sur les sites

3. http ://www.invid-project.eu/

	Médias traditionnels	Médias de réinformation	*Total*
Francophones	11 sources, 157 885 messages	34 sources, 278 351 messages	352 319
Anglophones	14 sources, 194 434 messages	11 sources, 105 094 messages	383 445

TABLE 1 – Répartition des messages par langue et par type de médias

web officiels des journaux ou les éditions papier.

3 Corpus et descripteurs

Afin de réaliser cette tâche, nous travaillons avec des messages réels collectés sur *Facebook* qui se présentent comme dédiés à l'information. Nous décrivons ci-dessous ce corpus et les descriptions des messages adoptées lors de nos expériences.

3.1 Corpus

La base de données utilisée est constituée de messages *Facebook* sous forme textuelle (e.g contenu du message), multimédia (e.g image d'illustration du message) et sociale (e.g nombre de mentions *j'aime*). Ces messages proviennent de groupes appartenant à des médias traditionnels et à des médias de réinformation, en français ou en anglais. Les groupes *Facebook* ont été sélectionnés et classés selon plusieurs critères : l'appartenance à une société de presse identifiable ou l'association à une édition papier ou TV (indices de médias traditionnels), l'affichage explicite d'une volonté de réinformation (toutes variantes autour du thème de la révélation de la vérité cachée par les médias de masse ; indices de sources de réinformation), et enfin, l'existence d'articles classés comme faux dans des sites d'analyse (par exemple `hoaxbuster.com`, `hoax-slayer.com` ; indices de sources de réinformation). Pour ce faire, nous avons défini une tâche d'annotation manuelle en trois classes (*traditionnel*, *réinformation* ou *autre* dans le cas de sources qui ne sont pas jugées comme une source d'information) qui a été effectuée par trois annotateurs. Les accords inter-annotateurs sont élevés (κ de Fleiss (Davies & Fleiss, 1982) = 0.874 ; α de Krippendorff (Krippendorf, 1980) = 0.875) ; les divergences ont ensuite été discutées pour décider par consensus de la classe à attribuer. Les messages sont labelisés en fonction de leur provenance, c'est-à-dire selon que le groupe est jugé comme média traditionnel ou de réinformation. Ces messages sont souvent très courts, renvoyant fréquemment sur un autre contenu (site Web, image, vidéo).

Au total, la base contient à ce jour 735 764 messages, collectés depuis novembre 2015, se répartissant en deux corpus $\mathcal{C}_a$ et $\mathcal{C}_f$ pour respectivement l'anglais et le français, décrits dans le tableau 1. Cette base, sous la forme de listes d'URL, est mise à disposition de la communauté sur `https://www-linkmedia.irisa.fr/hoax-detection/` accompagnée des explications menant au classement de chacun des groupes Facebook en média traditionnel ou de réinformation, et de quelques chiffres caractérisant leurs contenus.

	$\mathcal{C}s_f$		$\mathcal{C}s_a$	
	réinformation	*traditionnel*	*réinformation*	*traditionnel*
Mots par message	$1173,64$	$4118,89$	$621,98$	$3067,14$
Hashtags par message	$10,27$	$26,89$	$1,48$	$10,48$
Majuscules par message (*)	$1,88\%$	$2,90\%$	$3,80\%$	$2,95\%$
Occurrence du symbole *?* (*)	$0,08\%$	$0,03\%$	$0,13\%$	$0,05\%$
Occurrence du symbole *!* (*)	$0,10\%$	$0,09\%$	$0,14\%$	$0,07\%$

TABLE 2 – Quelques caractéristiques de surface sur les corpus $\mathcal{C}s_f$ et $\mathcal{C}s_a$ (en moyenne par message)

3.2 Description des messages

3.2.1 Descripteurs de surface

Ces descripteurs, basés sur des informations de surface, permettent de caractériser la structure des messages. Pour nos travaux, nous nous inspirons de ceux proposés dans l'état de l'art pour la détection de fausses informations sur les réseaux sociaux (Boididou *et al.*, 2015). Treize descripteurs sont ainsi calculés pour chaque message, caractérisant : 1) la longueur du texte ; 2-5) la présence de signes de ponctuation particuliers : les signes " !" et ' ?' peuvent être des signes discriminants qui seront caractérisés par deux descripteurs chacun (présence ou non et occurrence du symbole dans le message) ; 6-8) l'orientation grammaticale du texte : trois descripteurs correspondant respectivement au nombre de pronoms de la première, deuxième et troisième personne dans les messages. Cette analyse permet de distinguer les textes qui évoquent un fait propre à l'auteur (utilisation des pronoms de la première personne), les textes interpellant le lecteur (pronoms de la deuxième personne) et les textes impersonnels (troisième personne) ; 9-13) la part de sentiments dans le message est représentée par le nombre d'occurrences (a) des émoticônes heureux et tristes, (b) des majuscules (les mots en majuscules accentuent la notion de sensationnel), (c) des mots positifs et mots négatifs (par calcul de la polarité des mots). Quelques-unes de ces caractéristiques sont données dans la table 2, mettant en évidence les différences de répartition des descripteurs en fonction des corpus. Les attributs notés avec un astérisque sont normalisés par la longueur des messages.

3.2.2 Descripteurs du contenu

Les descripteurs basés sur le contenu ont vocation à caractériser les médias par la présence de mots ou de séquences de mots spécifiques dans leurs messages. Pour ce faire, les messages sont lemmatisés avec TreeTaggger (Schmid, 1994) et les urls, hashtags et sources sont remplacés respectivement par les balises *[URL]*, *[HASHTAG]* et *[SOURCE]*, ensuite traités comme des mots. En plus de cela, si une URL est détectée dans le message initial (issu du réseau social), le contenu de la page pointée par cette URL est ajouté au message afin de compléter l'information apportée par le message. Ce choix est justifié par le fait que beaucoup de médias n'utilisent les réseaux sociaux que pour amener les utilisateurs à aller voir l'article complet sur leur site Web. Le message posté sur les réseaux sociaux peut alors être vu comme un texte d'accroche qui décrit brièvement le sujet de l'article entier. Nous conservons l'information que le message provenant du réseau social possédait une URL, c'est pourquoi nous gardons aussi un tag *[URL]*.

corpus $\mathcal{C}s_a$	*surface*		*contenu*		*surface + contenu*	
	F_1	$Taux\ BC$	F_1	$Taux\ BC$	F_1	$Taux\ BC$
Naive Bayes	$52,80\%$	$59,37\%$	$80,36\%$	$80,58\%$	$80,42\%$	$80,65\%$
J48	**65,04%**	**66,06%**	**85,14%**	**85,18%**	**84,94%**	**84,99%**
JRip	$64,25\%$	$65,34\%$	$68,80\%$	$70,31\%$	$75,78\%$	$76,15\%$
IB_1	$54,17\%$	$54,44\%$	$61,28\%$	$62,90\%$	$77,09\%$	$77,16\%$
Random Forest	$62,92\%$	$63,31\%$	$75,90\%$	$76,51\%$	$79,53\%$	$79,92\%$
SMO	$51,61\%$	$52,65\%$	$76,91\%$	$77,35\%$	$76,93\%$	$77,38\%$

TABLE 3 – F-mesure (F_1) et taux de bonne classification ($Taux\ BC$) des messages du corpus anglais

Pour chaque message, les valeurs TF-IDF (Sparck Jones, 1972) des n-grammes de taille 1 à 3 sont calculées. Les 1000 n-grammes les plus discriminants, selon leur gain d'information (Mitchell, 1997), sont alors retenus pour constituer le vocabulaire de description, et chaque message est donc représenté par un vecteur de dimension 1000.

4 Évaluation

Cette section décrit le protocole expérimental utilisé pour la tâche de classification visée, à savoir prédire l'origine (média traditionnel ou de réinformation) d'un message Facebook, les résultats obtenus, et une analyse de l'emploi des descripteurs dans les modèles de classifiction obtenus.

4.1 Contexte expérimental

Les tests sont réalisés sur des sous-ensembles $\mathcal{C}s_f$ et $\mathcal{C}s_a$ de la base de données (3.1), chacun constitué de $40\ 000$ messages issus pour moitié de médias traditionnels et pour moitié de médias de réinformation. Les descripteurs de surface et les descripteurs de contenu sont d'abord utilisés indépendamment, puis combinés.

La librairie Weka (Hall *et al.*, 2009) est utilisée pour la classification. Plusieurs classifieurs de différentes familles sont testés : classifieur bayésien naïf (*Naive Bayes*), règles propositionnelles (*JRip*), Arbre de décision (*J48*), forêts aléatoires (*Random Forest*) avec $N = 100$, SVM (*SMO*) avec noyau RBF et k-plus proches voisins (IB_k) avec $k = 1$. Les descripteurs ont été normalisés pour ces 2 dernières méthodes. Deux mesures complémentaires sont utilisées pour évaluer les résultats : la F-mesure (F_1) et le taux de bonnes classifications (*Taux BC*). Afin de tester les classifieurs, une méthode de validation croisée à 10 plis est utilisée.

4.2 Résultats

Les résultats obtenus sur les corpus anglais ($\mathcal{C}s_a$) et français ($\mathcal{C}s_f$) sont présentés respectivement dans les tableaux 3 et 4. La première constatation est que l'analyse du contenu retourne toujours des meilleurs résultats que l'analyse de surface seule. La combinaison des deux descripteurs améliore marginalement les résultats obtenus en utilisant les descripteurs sur le contenu seuls. On notera également que les résultats sur le corpus français $\mathcal{C}s_f$ sont légèrement meilleurs que sur le corpus

corpus $\mathcal{C}s_f$	*surface*		*contenu*		*surface + contenu*	
	F_1	$Taux\ BC$	F_1	$Taux\ BC$	F_1	$Taux\ BC$
Naive Bayes	$38,58\%$	$51,58\%$	$72,38\%$	$73,88\%$	$73,12\%$	$74,50\%$
J48	$61,62\%$	$61,76\%$	$82,67\%$	$82,81\%$	$83,41\%$	$83,54\%$
JRip	$59,28\%$	$59,41\%$	$81,86\%$	$82,02\%$	$81,08\%$	$81,33\%$
IB_1	$61,57\%$	$61,73\%$	$83,75\%$	$83,89\%$	$83,74\%$	$83,77\%$
Random Forest	**62,92%**	**63,31%**	$75,90\%$	$76,51\%$	$79,53\%$	$79,92\%$
SMO	$13,82\%$	$14,61\%$	**88,60%**	**88,62%**	**88,84%**	**88,86%**

TABLE 4 – F-mesure (F_1) et taux de bonnes classifications ($Taux\ BC$) des messages du corpus français

anglais. La supériorité des descripteurs basés contenu par rapport aux descripteurs de surface y est également plus marquée.

4.3 Analyse des descripteurs

L'analyse des modèles de classification obtenus (lorsqu'elle est possible) permet de comprendre les cas d'erreurs et de caractériser la pertinence des différents descripteurs.

Descripteurs de surface : La sélection des descripteurs les plus discriminants, effectuée par calcul de l'information mutuelle entre l'attribut et la classe, met en avant par ordre d'importance décroissante : 1) la longueur du texte ; 2) la présence de symboles de ponctuation *!* et *?* ; 3) l'orientation des pronoms personnels (i.e première, deuxième ou troisième personne). Ces résultats sont corroborés par l'étude des descripteurs effectivement utilisés dans les arbres de décision et des règles prépositionnelles.

Descripteurs de contenu : l'étude des classifieurs interprétables (comme JRip, Random Forest, J48) générés à partir de ces descripteurs montre que les modèles de décision cherchent à caractériser principalement les médias traditionnels, le message étant classé en média de réinformation par défaut, c'est-à-dire lorsque qu'aucun de ses descripteurs ne l'a amené à être classé comme traditionnel. Cela semble indiquer qu'il est plus facile de déterminer des caractéristiques communes à tous les messages traditionnels qu'aux messages des médias de réinformation.

De ce fait, les erreurs sont majoritairement commises par manque de règles décrivant les messages de médias de réinformation. Cependant, cela permet d'obtenir une précision élevée sur la classe *traditionnel* (i.e un message classé comme tel à une forte probabilité d'être bien classé) : par exemple, pour le corpus $\mathcal{C}s_f$, la précision de la classe *traditionnel* est de $94,75\%$ avec le classifieur *Naive Bayes* contre une précision de $66,29\%$ pour la classe *réinformation*.

Certains descripteurs discriminants de la classe *traditionnel* relèvent indirectement de l'aspect professionnel du site diffusant l'information ; il s'agit par exemple de la présence des termes *'RSS'* ou *'votre abonnement'* pour le français, et *accessiblity* ou *privacy* pour l'anglais. D'autres descripteurs notent le niveau de langue de certains sites de réinformation, faisant plus largement usage d'abréviation comme *WTF*, *DIY*, *pic*... Les médias traditionnels sont quant à eux caractérisés par une présence accrue de marques de citations, ou de mots comme *opinion*.

Combinaison des descripteurs : comme pour l'ensemble de descripteurs de contenu ci-dessus, les modèles de décision cherchent à détecter les messages de médias traditionnels. Cette fois aussi la

précision de la classe *traditionnel* vaut $94,75\%$ avec *Naive Bayes* corpus $\mathcal{C}s_f$ tout comme la précision de la classe *réinformation* qui obtient des résultats équivalents $66,8\%$ avec *Naive Bayes*.

5 Conclusion

Afin de différencier les messages postés sur les réseaux sociaux par les médias traditionnels et de réinformation, nous avons étudié l'influence d'attributs de surface, de contenu et l'association de ces deux ensembles d'attributs. Nous avons obtenus des résultats encourageants, notamment sur les deux approches utilisant le contenu du message.

Dans des travaux futurs, plusieurs améliorations sont envisagées. Il semble entre autre important de proposer de nouveaux descripteurs de surface et de contenu pour améliorer la tâche de classification. Notamment, un travail plus approfondi sur un descripteur portant sur le respect des normes orthographiques serait bénéfique. En effet, l'étude des descripteurs les plus utilisés dans les classifieurs fait ressortir que de nombreuses fautes sont présentes et sont vues comme des descripteurs parfois très discriminants : ainsi *repondre* à la place de *répondre*, est jugé comme très marquant des médias de réinformation.

Enfin, comme nous l'évoquions dans l'introduction, ce travail est une étape intermédiaire dans un but plus vaste de faire de la détection de fausses informations dans les réseaux sociaux, en temps réel. La caractérisation de la source doit être utilisée comme un indice, parmi d'autres, pour atteindre ce but. D'autres indices, portant notamment sur les aspects sociaux (étude de la propagation des informations dans le graphe social) et sur les médias autres que le texte (images, vidéos) seront étudiés.

Remerciements

Ces travaux sont soutenus par la Direction Générale de l'Armement (DGA) et l'Université Rennes 1.

Références

ALEJANDRO J. (2010). Journalism in the age of social media. *Reuters Institute Fellowship Paper, University of Oxford*, p. 2009–2010.

BOIDIDOU C., ANDREADOU K., PAPADOPOULOS S., DANG-NGUYEN D.-T., BOATO G., RIEGLER M. & KOMPATSIARIS Y. (2015). Verifying multimedia use at mediaeval 2015. In *Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*.

DAVIES M. & FLEISS J. (1982). Measuring agreement for multinomial data. *Biometrics*, p. 1047–1051.

DECLERCK T. & LENDVAI P. (2015). Processing and normalizing hashtags. *RECENT ADVANCES IN*, p. 104.

DERCZYNSKI L. & BONTCHEVA K. (2014). Pheme : Veracity in digital social networks. In *UMAP Workshops*.

DERCZYNSKI L., MAYNARD D., RIZZO G., VAN ERP M., GORRELL G., TRONCY R., PETRAK J. & BONTCHEVA K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, **51**(2), 32–49.

GOTTRON T., SCHMITZ J. & MIDDLETON S. (2014). Focused exploration of geospatial context on linked open data. In *Proceedings of the 3rd International Conference on Intelligent Exploration of Semantic Data-Volume 1279*, p. 1–12 : CEUR-WS. org.

HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter*, **11**(1), 10–18.

KORDOPATIS-ZILOS G., PAPADOPOULOS S. & KOMPATSIARIS Y. (2015). Geotagging social media content with a refined language modelling approach. In *Intelligence and Security Informatics*, p. 21–40. Springer.

KRIPPENDORF K. (1980). *Content Analysis : An Introduction to its Methodology*. Sage Publications.

KWAK H., LEE C., PARK H. & MOON S. (2010). What is twitter, a social network or a news media ? In *Proceedings of the 19th international conference on World wide web*, p. 591–600 : ACM.

MIDDLETON S. (2015a). Extracting attributed verification and debunking reports from social media : Mediaeval-2015 trust and credibility analysis of image and video.

MIDDLETON S. E. (2015b). Reveal project-trust and credibility analysis.

MITCHELL T. M. (1997). *Machine Learning*. New York, NY, USA : McGraw-Hill, Inc., 1 edition.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, p. 44–49, Manchester, UK.

SHARMA N. (2015). *Facebook journalism : An exploratory study into the news values and role of journalists on Facebook*. PhD thesis, INDIANA UNIVERSITY.

SPARCK JONES K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **28**(1), 11–21.

ZAMPOGLOU M., PAPADOPOULOS S. & KOMPATSIARIS Y. (2015). Detecting image splicing in the wild (web). In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*, p. 1–6 : IEEE.