Étude des influences réciproques entre médias sociaux et médias traditionnels

Béatrice Mazoyer, Nicolas N. Turenne, Marie-Luce Viaud

Étude des influences réciproques entre médias sociaux et médias traditionnels

Béatrice Mazoyer*, Nicolas Turenne**, Marie-Luce Viaud*

*INA, 18 Avenue des frères Lumière, 94366 Bry-sur-Marne, France
bmazoyer@ina.fr, mlviaud@ina.fr
**Université Paris-Est, LISIS, INRA, 77454 Marne-La-Vallée, France
nturenne@u-pem.fr

Résumé. Cet article s'intègre dans un travail de recherche consacré à l'étude des médias traditionnels et des réseaux sociaux et de leurs influences mutuelles. Notre étude vise à mettre en place un outil de collecte en continu des tweets liés aux événements relayés par les médias traditionnels. L'objectif est d'obtenir pour chaque événement médiatique un corpus à la fois exhaustif et précis (minimisation du bruit) de tweets qui fassent référence à cet événement. Notre méthodologie se fonde sur un processus itératif : les tweets collectés dans un premier temps sont analysés pour affiner les collectes suivantes. Notre outil obtient de bons résultats en ce qui concerne la pertinence des tweets collectés vis-à-vis des événements, mais l'exhaustivité semble encore à perfectionner, ce pourquoi nous proposons des pistes d'amélioration.

1 Introduction

Les réseaux sociaux jouent de façon croissante le rôle d'intermédiaires entre les médias traditionnels (presse, radio, télévision) et leur audience. Les médias français s'adaptent à cette évolution : ainsi, de nombreux médias, parmi lesquels Le Parisien, Libération, FranceTVInfo ou L'Équipe, ont adopté en 2016 le format Facebook Instant Article [1] (qui permet une lecture facilitée sur mobile, sans quitter l'application Facebook) quitte à perdre une partie de leur contrôle sur leur format de distribution.

À cela s'associe une évolution des contenus publiés : les médias traditionnels s'inspirent des modes de communication des médias sociaux et relaient les informations les plus partagées sur ceux-ci. Cependant, il est difficile de quantifier et d'analyser de façon précise les influences réciproques entre ces deux sphères, notamment car on ne dispose pas d'un jeu de données associant événements médiatiques et réactions sur les réseaux sociaux sur une longue durée. C'est l'enjeu du travail que nous présentons ici : la construction d'un outil associant automatiquement à des événements d'actualité les tweets qui y font référence. Cet article présente des travaux en cours : nous travaillons actuellement sur la formalisation de l'évaluation des résultats, ce qui représente une tâche complexe car s'il est possible d'annnoter manuellement les tweets récoltés par l'outil, il est plus difficile d'estimer le volume de tweets non détectés.

1. https ://developers.facebook.com/docs/instant-articles

Le choix de Twitter comme média social étudié doit être justifié, étant donné la prédominance de Facebook en termes de nombre d'utilisateurs [2]. Pour notre étude, Twitter a l'intérêt d'être un réseau social où les interventions sont publiques, à l'inverse de Facebook, où la plupart des utilisateurs retreignent la visibilité de leurs publications à un cercle privé. C'est probablement la raison pour laquelle beaucoup de journalistes utilisent de façon privilégiée Twitter comme outil de travail [3]. Dans le cadre d'un travail sur les influences des médias, l'étude de Twitter nous a donc parue fondamentale.

2 Travaux antérieurs

D'autres travaux de recherche se sont consacrés à l'étude des liens entre tweets et événements médiatiques. Beaucoup travaillent cependant sur des corpus de tweets fixés, constitués au préalable, qu'ils cherchent à répartir en « sujets » ou « événements » décrits par les médias, soit de façon manuelle (Rieder et Smyrnaios, 2012), soit en utilisant des modèles thématiques de type LDA (Zhao et al., 2011; Hu et al., 2012; Hua et al., 2016). À l'inverse, notre démarche consiste à collecter en continu de nouveaux tweets associés à des événements d'actualité. Il s'agit d'une approche dynamique du traitement de l'information.

Des méthodes de type « First Story Detection » sont plus adaptées à des documents collectés en continu : elles consistent à attribuer chaque nouveau document d'un flux à la chaîne de documents avec laquelle sa similarité est la plus grande, en créant une nouvelle chaîne si la similarité est inférieure à un seuil s. Certains auteurs ont tenté ce type d'approches avec des tweets (Petrovic et al., 2010) mais leur modèle inclut tous les tweets collectés aléatoirement par l'API sample de Twitter, y compris ceux traitant de sujets non liés à l'actualité. Étant données les restriction de l'API (accès limité à 1% du flux mondial à un moment t), cette méthode ne peut pas garantir que l'on obtienne tous les tweets liés à un événement donné.

Enfin, un outil mis en place par le Centre Commun de Recherche de la Commission Européenne (Tanev et al., 2012) s'inscrit dans la même démarche que la nôtre puisqu'il vise à collecter en continu des tweets liés à des articles de presse. Les auteurs utilisent pour ce faire un corpus d'un million d'articles pour calculer le poids (tf-idf) à attribuer aux termes de chaque nouvel article, et utilisent ensuite les termes ayant les meilleurs scores pour formuler des requêtes à l'API search de Twitter. L'inconvénient d'une telle méthode est d'utiliser uniquement le vocabulaire de la presse pour interroger Twitter. Or les écarts sont souvent importants entre le vocabulaire des médias traditionnel et celui employé sur le réseau social, du fait de la nature différente de l'information transmise par les tweets (commentaires, opinions personnelles, rumeurs) (Hoang-Vu et al., 2014).

L'apport de notre approche consiste donc à reformuler les premières requêtes envoyées à l'API de Twitter en fonction des tweets obtenus grâce à celles-ci, afin de prendre en compte les spécificités de la communication sur le réseau social (abbréviations, hashtags, langage familier, fautes d'orthographe).

2. Selon un communiqué de presse de Médiamétrie, Facebook compte 35 millions de visiteurs uniques par mois (en comptabilisant uniquement les visites depuis un mobile), contre 16 millions pour Twitter en juillet 2016.

3. Selon une étude réalisée à la demande de la Commission Européenne auprès de 135 journalistes européens, les journalistes établissent une distinction entre Twitter, largement utilisé pour des raisons professionnelles, et Facebook, dont il est davantage fait un usage privé

3 Démarche adoptée

Le prototype mis en place [4], contrairement à l'outil de Tanev et al. (2012), n'intègre pas de connaissances extérieures et se fonde uniquement sur l'analyse des tweets collectés, dans un esprit robuste. Il prend en entrée les dépêches émises par l'AFP, qui constituent donc la représentation des « événements médiatiques » que l'on cherche à étudier. Pour chaque dépêche, on extrait les entités nommées (noms de lieux, de personnes et d'organisations) présentes dans le titre. Ce sont ces entités qui constituent les requêtes initiales envoyées à l'API search de Twitter. Ainsi, pour la dépêche AFP intitulée « Ziad Takieddine affirme avoir remis trois valises d'argent libyen à Nicolas Sarkozy et Claude Guéant », les termes automatiquement extraits sont « Takieddine », « libyen », « Sarkozy » et « Guéant ». La collecte s'effectue ensuite de la manière suivante : on collecte les tweets en français contenant les termes extraits (t1 AND t2 ... AND tn), dans une fenêtre de 24h autour de la date de la dépêche. Cette étape est répétée chaque jour avec les nouvelles dépêches. Pour les dépêches dont le titre ne contient aucune entité nommée, la requête envoyée à Twitter contient tous les mots du titre. Cette méthode permet d'obtenir les tweets contenant un lien url vers la dépêche AFP ou vers un article de presse en ligne ayant le même titre que la dépêche AFP (ce qui est relativement fréquent).

Dans un second temps, on procède à la reformulation des requêtes. On calcule un score tf-idf pour les mots des tweets collectés : on considère que l'ensemble des tweets associés à chaque événement constitue un "document" et que la "collection" est formée par tous les documents créés dans les 30 jours précédents. Ainsi, on peut attribuer un poids à chaque terme de chaque événement. On sélectionne alors les termes ayant un score tf-idf supérieur à certains seuils s1, s2 et s3 fixés manuellement actuellement (car l'évaluation des résultats n'est pas encore formalisée) pour les 1, 2 et 3-grammes. Ce sont ces termes qui sont utilisés pour formuler les prochaines requêtes à l'API. Dans l'exemple proposé ci-dessus, les termes utilisés pour les prochaines requêtes seraient le bigramme « argent lybien » et le hashtag « #Takieddine ».

Cette approche, si elle s'appuie sur le texte des tweets pour améliorer les requêtes, reste cependant très liée au titre des dépêches initiales, du fait de la brièveté des tweets. Le modèle word2vec permet d'identifier de nouveaux termes, différents de ceux présents dans le titre des dépêches mais employés dans un contexte similaire. Word2vec est un modèle prédictif proposé en 2013 permettant d'apprendre un « word embedding » à partir d'une large collection de textes (Mikolov et al., 2013). Il s'agit donc de représenter les mots d'une collection de textes par des vecteurs de réels, les termes employés dans un contexte similaire ayant des vecteurs proches. Cela permet par la suite d'obtenir, pour un terme donné, le ou les termes fréquemment employés dans le même contexte. Nos essais, réalisés en entraînant le modèle sur 2 millions de tweets en français collectés aléatoirement via l'API streaming, montrent qu'il permet effectivement de formuler des synonymes, des abbréviations (« fh » pour « François Hollande ») ou des orthographes alternatives (« Trierweiller » pour « Trierweiler »). Cependant, le modèle renvoie aussi d'autres types de résultats : des termes employés dans un contexte proche mais qui ne sont pas des synonymes. Ainsi, « Juppé », « Macron » ou « Valls » sont également des termes renvoyés par le modèle comme proches du terme « Hollande ». L'utilisation de word2vec nécessite donc un post-process de filtrage des termes pertinents pour les requêtes, fondé sur l'étude de la répartition temporelle des tweets.

4. L'outil se compose d'une partie extraction d'entités nommées, codé avec Knime, et d'une partie extension de requêtes, en Python. Cette partie est disponible sur Github : http://www.github.com/bmaz/quick_twython. L'outil nécessite l'installation d'Elasticsearch. Le modèle word2vec entraîné pour les tests est disponible dans le même dépôt.

4 Conclusion

L'outil mis en place permet la formulation de requêtes à l'API de Twitter en utilisant un vocabulaire spécifique aux usages du réseau social, grâce à l'identification des termes récurrents dans les corpus de tweets collectés. Par ailleurs, la recherche de synonymes via le modèle word2vec, si elle n'a pour l'instant été testée qu'à petite échelle, permet d'envisager la formulation de requêtes indépendantes des dépêches AFP initiales, à condition de mettre en place une étape de filtrage des termes utilisés. À cette fin, nous envisageons un filtre se fondant sur la répartition temporelle des tweets obtenus : une requête pertinente peut être identifiée si elle permet la collecte de tweets très concentrés autour de la date et l'heure de l'événement étudié. L'ajout de ce filtre à notre prototype devrait permettre la formulation de requêtes à la fois plus nombreuses et plus fiables pour collecter des corpus sans bruit.

Références

Hoang-Vu, T.-A., A. Bessa, L. Barbosa, et J. Freire (2014). Bridging vocabularies to link tweets and news. *WebDB*.

Hu, Y., A. John, F. Wangand, D. Duncan-Seligmann, et S. Kambhampati (2012). Et-lda: Joint topic modeling for aligning, analyzing and sensemaking of public events and their twitter feeds. *AAAI*, 59–65.

Hua, T., Y. Ning, F. Chen, C.-T. Lu, et N. Ramakrishnan (2016). Topical analysis of interactions between news and social media. *AAAI*, 2964–2971.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. *ICLR Workshop*.

Petrovic, S., M. Osborne, et V. Lavrenko (2010). Streaming first story detection with application to twitter. *HLT-NAACL*, 181–189.

Rieder, B. et N. Smyrnaios (2012). Pluralisme et infomédiation sociale de l'actualité : le cas de twitter. *Réseaux 6*, 105–139.

Tanev, H., M. Ehrmann, J. Piskorski, et V. Zavarella (2012). Enhancing event descriptions through twitter mining. *ICWSM*.

Zhao, W. X., J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, et X. Li (2011). Comparing twitter and traditional media using topic models. *ECIR*, 338–349.

Summary

This article is part of a research project focused on studying traditional media and social networks and their mutual influences. Our study aims at creating a tool collecting a continuous stream of tweets related to media events. Our goal is to get both exhaustive and precise (noise minimization) collections of tweets referring to each media event. Our methodology is based on an iterative process : tweets collected at the first step are analyzed in order to improve next collections. Out tool delivers good results concerning the relevance of collected tweets to the media events. However progress could be made to collect larger sets of tweets, that is why we suggest ways in which the tool could be improved.