EXPLOITER LE BIG DATA POUR MIEUX COMPRENDRE LES COMPORTEMENTS DES INTERNAUTES : UNE APPLICATION AU PARTAGE DE L'INFORMATION SUR TWITTER

Sophie BALECH
Institut d'Administration des Entreprises Amiens, Université Picardie Jules Verne – Laboratoire CRIISEA

sophie.balech@gmail.com

Résumé :

Cet article s'intéresse aux comportements de partage de l'information des utilisateurs au sein d'une plateforme de micro-blogging, Twitter. En nous appuyant sur le modèle ELM, nous proposons un modèle explicatif de la performance d'un message en prenant en compte des éléments relevant de la route périphérique et nous le testons empiriquement, sur un corpus de près de 800 000 tweets originaux émis par environ 235 000 utilisateurs sur une période de 7 mois concernant l'épidémie de Covid-19 en France. Nous montrons ainsi l'importance de la crédibilité de la source du message et de sa stratégie sur la plateforme, mais aussi celle de la forme du message, sa composition et son degré d'élaboration. Ces résultats sont obtenus grâce à une méthodologie mixte de traitement des big data, utilisant des outils de text mining pour créer les indicateurs nécessaires pour tester notre modèle.

Mots-clés : bouche-à-oreille électronique ; plateforme de micro-blogging ; C2C ; Twitter ; NLP ; ELM ; partage d'information

UNDERSTANDING THE BEHAVIOUR OF INTERNET USERS THANKS TO BIG DATA: AN APPLICATION TO INFORMATION SHARING ON TWITTER

Abstract :

This paper focuses on the information sharing behaviour of users within a micro-blogging platform, Twitter. Based on the ELM model, we propose an explanatory framework of post's performance by taking into account the peripheral cues and we test it empirically, on a corpus of nearly 800,000 original tweets sent by about 235,000 users over a period of 7 months concerning the Covid-19 epidemic in France. We thus show the importance of the source's credibility and its strategy on the platform, but also of the form of the post, its composition and its degree of elaboration. These results are obtained through a mixed methodology of big data processing. Indicators used in model testing are created with text mining tools.

Keywords : eWOM ; micro-blogging platform ;C2C ; Twitter ; NLP ; ELM ; information sharing

Exploiter le big data pour mieux comprendre les comportements des internautes : une application au partage de l'information sur Twitter

INTRODUCTION :

Twitter est le 5ème réseau social en France avec 12 millions d'utilisateurs. Cette plateforme est prisée pour le partage d'information en temps réel sur des sujets d'actualité (Asselin, 2021). L'efficacité d'un message sur cette plateforme se mesure principalement par le nombre de partages que reçoit ledit message, les retweets. En effet, un message retweeté par un utilisateur a été reconnu comme digne d'intérêt et a suscité suffisamment d'engagement pour que ce dernier souhaite le partager à sa communauté (ses followers, les membres de son audience) (Boyd, Golder & Lotan, 2010). Plus un message est viral, plus il montre ses capacités à susciter de l'engagement et plus son audience s'accroît, ce qui fait partie des objectifs des campagnes de marketing viral sur les réseaux sociaux (Sohn, Gardner & Weaver, 2013).

Cet article a un double objectif : il vise d'une part à présenter les outils disponibles aux chercheurs en marketing pour l'exploitation des big data en s'appuyant sur les données contenues dans les réseaux sociaux. D'autre part, d'un point de théorique, il cherche à mettre en évidence les éléments qui, au-delà du contenu même du message, en favorisent la diffusion, en s'appuyant sur les éléments périphériques du modèle ELM (Elaboration Likelihood Model) (Petty & Cacioppo, 1986; Petty, Cacioppo, & Schumann, 1983).

La source et la forme comme déterminants de la diffusion de l'information

La source du message

Les nombreuses études qui se sont intéressées à la diffusion d'information sur Twitter ont mis en évidence le rôle clé de certaines caractéristiques de la source dans l'efficacité de son message, notamment le rôle du nombre de personnes qui suivent le compte ou l'expérience de l'utilisateur de la plateforme (Berman, Melumad, Humphrey, & Meyer, 2019 ; Nesi, Pantaleo, Poali, & Zaza, 2018 ; Van De Velde, Meijer, & Homburg, 2015). Si ces variables semblent fondamentales pour expliquer le nombre de retweets, nous pouvons aussi nous intéresser à la stratégie des utilisateurs postant le message vis-à-vis de la plateforme et de sa communauté. En effet, certains utilisateurs font preuve d'une grande proactivité pour étendre leur communauté, et ainsi leur sphère d'influence, marquant à la fois des formes de réciprocité importante ou des niveaux d'activité forts. À l'inverse, certains utilisateurs ont un comportement beaucoup plus passif, que l'on peut qualifier comme de recherche de contenus. Nous formulons donc l'hypothèse suivante :

- H1 : les stratégies proactives (vs réactives) de l'émetteur vis-à-vis de la plateforme favorisent la diffusion du message.

La crédibilité de la source est un signal important pour la persuasion s'appuyant sur la route périphérique. Dans une revue de littérature sur l'application du modèle ELM aux avis en ligne, Soulard (2015) montre que les recherches se sont intéressées à deux points particuliers : la réputation de l'auteur de l'avis en ligne et la divulgation des informations personnelles. Ces deux

dimensions se retrouvent sur Twitter, où un utilisateur peut connaître l'identité du compte émetteur d'un message via son nom, mais aussi avoir accès aux informations que ce dernier choisit de divulguer à travers sa photo et sa description de profil, et peut savoir si ce dernier est un utilisateur que Twitter certifie (à travers la procédure de vérification de compte). Nous nous attendons donc à voir un effet de la crédibilité de la source dans l'efficacité que rencontre un message :

- H2 : plus l'émetteur est jugé crédible (vs non-crédible), plus son message est diffusé.

La forme du message

Les études portant sur la viralité des tweets se sont intéressées, en plus des variables d'audience et d'expérience de la source, aux caractéristiques de la forme du message. Au-delà des sujets de conversation (*topics*) traités dans les messages, certaines études se sont intéressées aux éléments de composition (inserts d'image, de lien, de mot-clé renvoyant à un sujet précis, d'interpellation d'autres utilisateurs ou encore d'émojis) : par exemple, Jenders, Kasneci, & Naumann (2013) et Van De Velde, Meijer, & Homburg (2015) mettent en évidence le rôle de la présence d'url, de hashtag ou de mentions dans la probabilité d'être retweeté, tandis que Quesenberry & Coolsen (2019) ont ajouté la présence d'émojis dans leur modèle. Aussi, nous formulons l'hypothèse suivante :

- H3 : plus le message contient d'éléments de formes supplémentaires (vs uniquement du texte), plus il est diffusé.

À ces variables de composition s'ajoutent aussi des variables relevant plutôt de la forme stylistique du message, ce que nous appelons le degré d'élaboration. Berger & Milkman (2012) ont mis en évidence le rôle des sentiments exprimés dans les contenus pour la diffusion d'informations en ligne, élément que l'on retrouve sur Twitter. La longueur des messages a également été prise en compte dans les modèles explicatifs de la viralité des tweets, comme c'est le cas dans l'étude de (Lahuerta-Otero, Cordero-Gutiérrez, & De la Prieta-Pintado, 2018). Pour compléter ces variables traditionnellement mobilisées pour expliquer la diffusion des tweets, nous ajoutons une mesure de lisibilité qui permet de rendre compte de la facilité de compréhension d'un texte (DuBay, 2004), ainsi qu'une mesure de diversité lexicale qui rend compte de la richesse du vocabulaire utilisé. Formellement, nous chercherons à tester l'hypothèse suivante :

- H4 : plus le degré d'élaboration du message est important (vs faible), plus il est diffusé.

Ces différents éléments sont formalisés dans le modèle proposé dans l'annexe **Erreur ! Source du renvoi introuvable.**.

DONNÉES ET MÉTHODES :

Jeu de données :

Le jeu de données utilisées est extrait du corpus de tweets constitué par Banda et al. (2020). Les mots-clés utilisés pour collecter les tweets en temps réel sont différentes variations autour des hashtags Coronavirus et Covid. Dû aux conditions de service de Twitter, le corpus contient uniquement les identifiants de plus de 194 millions de tweets collectés entre janvier et décembre 2020. Nous avons réhydraté cet ensemble pour obtenir les contenus des tweets et de nombreuses

variables disponibles via l'API de Twitter. Ces variables concernent les caractéristiques des tweets, celles liées à leurs auteurs et à leurs performances sur la plateforme. Puis, nous avons extrait un sous-corpus en langue française. Le jeu de données utilisé comprend un total de 5,3 millions de tweets émis par 1,1 millions de contributeurs, entre le 31 janvier et le 1er septembre 2020. La répartition dans le temps et entre les différentes formes de post (tweet original, retweet, *reply* et *quote*) est présenté dans l'annexe 2. On constate l'importance des retweets dans ce sujet de conversation qui représentent environ 70 % des tweets émis sur la période. Seuls les contenus originaux ont été retenus pour l'analyse, ce qui représente un total de 797 374 tweets émis par 234 887 comptes. Les données ont été traitées en utilisant l'environnement R (R Core Team, 2014) et les bibliothèques du tidyverse (Taylor, 2017).

Opérationnalisation des variables :

La performance des tweets est mesurée par le logarithme du nombre de retweets obtenus pour chaque contenu original. L'expérience des utilisateurs sur la plateforme a été appréhendée par le nombre total d'actions réalisé par un utilisateur (*statuses*) et par le nombre de tweets qu'un utilisateur a mis dans ses favoris (*favourites*). L'audience d'un utilisateur est mesurée par le nombre de personnes suivant son compte (*followers*) et par le nombre de personnes suivis par l'utilisateur (*friends*). Une transformation en log10 a été appliquée à ces différentes mesures. Pour rendre compte des stratégies proactives ou réactives des émetteurs vis-à-vis de la plateforme, nous ajoutons des effets d'interaction :

- Stratégie proactive de création de contenus : interaction entre le nombre de *followers* et le nombre de *statuses* ;
- Stratégie proactive de réciprocité : interaction entre le nombre de *followers* et le nombre de *friends*, ainsi qu'entre le nombre de *followers* et de *favourites* ;
- Stratégie réactive de consommation de contenu : interaction entre le nombre de *friends* et le nombre de *favourites*.

Nous avons créé deux variables pour représenter le niveau de qualité d'un émetteur. Pour approcher la crédibilité d'un utilisateur, nous nous sommes appuyés sur son comportement de dévoilement sur la plateforme, en élaborant un indicateur de transparence du profil. Nous avons appliqué un algorithme de détection des entités nommées (Rinker, 2017) sur le nom de l'utilisateur et sa description, pour identifier la présence de noms de personnes, d'organisations ou de lieux. À ces informations, nous avons ajouté la présence d'une photo de profil, l'absence d'émojis dans la description de profil et si le compte a été vérifié par Twitter. L'indicateur ainsi créé varie de -1 à +9. De plus, les comptes identifiés comme ayant une activité très importante (un tweet par jour en moyenne) ont fait l'objet d'une détection de *bot* grâce à la bibliothèque Tweetbotornot (Kearney, 2018), afin de distinguer les comptes automatisés des comptes opérés par un individu. Un compte opéré par un *bot* est considéré comme non crédible.

Les variables de composition des tweets correspondent aux éléments de forme insérés dans le contenu du tweet. Nous avons créé 4 variables binaires codées 1 si le tweet contient une mention à

un autre compte, un hashtag, un lien url ou un média (photo, vidéo, gif). De plus, nous avons compté le nombre d'émojis présents dans les tweets.

Le degré d'élaboration des tweets est mesuré par 4 indicateurs. Pour la mesure de lisibilité, qui indique la facilité de compréhension d'un texte (plus la valeur est élevée, plus le niveau de formation nécessaire pour comprendre le texte est élevé), nous avons choisi l'indicateur ARI (*Automated Readability Index*) (DuBay, 2004)□. Nous avons choisi l'indicateur CTTR (*Caroll's Corrected Text-Type* Ratio) pour la mesure de diversité lexicale (Torruella & Capsada, 2013)□. Nous avons réalisé une analyse de sentiment grâce à l'annotateur du LIWC (Piolat, Booth, Chung, Davids, & Pennebaker, 2011 ; Tausczik & Pennebaker, 2010)□ pour créer un indicateur de sentiment positif et un indicateur de sentiment négatif. Nous avons ajouté les effets d'interaction entre ces indicateurs de sentiments, pour rendre compte des formes d'hyper-expressivité.

Le processus de collecte et de traitements des données est présenté dans l'annexe 3.

Méthodes :

Pour tester la pertinence de notre modèle général pour expliquer le nombre de retweets, nous utilisons des modèles linéaires emboîtés :

- le modèle de base comprend les variables d'expérience et d'audience des comptes émetteurs ;
- le modèle 2 ajoute les effets d'interaction entre ces variables pour rendre compte des stratégies des acteurs sur la plateforme ;
- le modèle 3 introduit les variables liées à la qualité de l'émetteur ;
- le modèle 4 intègre les variables de composition des tweets ;
- le modèle 5 prend en compte le degré d'élaboration des tweets.

RÉSULTATS :

Qualité d'ajustement des modèles :

Les principaux indicateurs de performance des modèles testés sont présentés dans l'annexe 4. Les résultats des tests ANOVA confirment l'existence de différences entre les modèles (modèle1:modèle2, F= 16 364, *p-value* < 0,001 ; modèle2:modèle3, F= 1973,9, *p-value* < 0,001 ; modèle3:modèle4, F= 4099,2, *p-value* < 0,001 ; modèle4:modèle5, F= 480,96, *p-value* < 0,001).

Nous constatons un accroissement de la qualité d'ajustement à mesure que les modèles sont plus élaborés. Les interactions entre les variables d'expérience et d'audience améliorent la qualité de prédiction ($R^2_{modèle\ 1}$ = 0,171 ; $R^2_{modèle\ 2}$ = 0,234 ; gain relatif de R^2 de 37 %). Les variables de qualité du profil, cherchant à approcher la crédibilité accordée à la source montrent un effet positif sur les critères d'ajustement du modèle. L'ajout des variables de composition améliore sensiblement la qualité d'ajustement ($R^2_{modèle\ 3}$ = 0,238 ; $R^2_{modèle\ 4}$ = 0,257 ; gain relatif de R^2 de 8 %). Enfin, le degré d'élaboration des tweets augmente encore un peu la qualité de la prédiction.

Etude des paramètres

L'annexe 5 reproduit les coefficients standardisés des 5 modèles. Les variables représentant les stratégies proactives des émetteurs sur la plateforme ont un impact positif sur le nombre de retweets ($\beta_{followers:friends}$ = 0,02 – *p-value* < 0,001 ; $\beta_{followers:favourites}$ = 0,03 – *p-value* < 0,001 ; $\beta_{followers:status}$ = 0,01 – *p-value* < 0,001), tandis que celle représentant une stratégie réactive a un impact négatif ($\beta_{favourites:friends}$ = -0,03 – *p-value* < 0,001). L'hypothèse H1 est donc vérifiée.

L'impact de l'indicateur de transparence du profil a un effet positif sur le nombre de retweet ($\beta_{profil\_transparency}$ = 0,01 – *p-value* < 0,001), tandis que la variable *bot* a un effet négatif (β_{bot}= -0,04 – *p-value* < 0,001). L'hypothèse H2 est ainsi vérifiée.

Le nombre de retweets est impacté positivement par la présence d'un média (β_{medias} = 0,08 – *p-value* < 0,001), par le nombre d'émojis (β_{emoji} = 0,04 – *p-value* < 0,001) et par la présence de hashtags ($\beta_{hashtag}$ = 0,02 – *p-value* < 0,001). A contrario, la présence de mention et d'url ont un impact négatif sur le nombre de retweets ($\beta_{mention}$ = -0,07 – *p-value* < 0,001 ; β_{url} = -0,05 – *p-value* < 0,001). L'hypothèse H3 n'est donc que partiellement vérifiée.

Concernant les variables représentant le degré d'élaboration des tweets, nous constatons un impact positif de l'indicateur de diversité lexicale (β_{CTTR} = 0,02 – *p-value* < 0,001), et un impact légèrement négatif de l'indicateur de lisibilité (β_{ARI} = -0,001 – *p-value* < 0,001). Par contre, les coefficients des variables de sentiment ne sont pas significatifs. L'hypothèse H4 n'est donc que très partiellement confirmée.

CONCLUSION :

Une première contribution empirique de ce travail est de s'intéresser aux comportements de partage de l'information réellement observés sur une plateforme et non à des intentions de partage mesurées en conditions expérimentales en exploitant des données massives. Nous montrons l'importance des signaux extérieurs aux caractéristiques intrinsèques des messages pour favoriser leur diffusion.

Nous avons en particulier montré le rôle important des stratégies des acteurs sur la plateforme (interaction entre les variables d'expérience et d'audience) dans la diffusion de l'information. Bien maîtriser les règles et les normes de la plateforme est donc un pré-requis pour que les messages soient transmis de pairs à pairs, mais il ne faut pas oublier de se constituer une large audience.

L'effort de dévoilement de l'émetteur sur le réseau social est aussi une stratégie payante pour voir ses messages diffusés largement, même si c'est dans une moindre mesure que l'effet d'une stratégie proactive. La crédibilité de la source, approchée par l'effort de dévoilement et par le fait d'opérer manuellement le compte, est donc un élément important dans la diffusion de l'information sur une plateforme de micro-blogging, ce qui confirme les résultats mis en évidence sur l'utilité des avis en ligne.

Au niveau de la forme, notre étude empirique montre le rôle important de la présence de médias, d'émojis et de hashtags dans la diffusion de l'information, tandis que le fait de renvoyer à une source extérieure à la plateforme (par la présence d'url) ou de s'adresser personnellement à d'autres utilisateurs (présence de mentions) ne sont pas favorables à un partage des contenus. La diffusion de

l'information au sein de la plateforme est donc favorisée par la présence d'éléments internes à la plateforme et par des communications s'adressant à tous, ce qui constituent des résultats attendus.

Par contre, les mesures employées pour représenter le degré d'élaboration d'un tweet ne sont pas concluantes (hormis pour la diversité lexicale), ce qui constitue une surprise, notamment concernant le rôle des sentiments exprimés. Ce résultat va à l'encontre des travaux de Berger & Milkman (2012)□, mais peuvent s'expliquer par la nature des messages autorisés par la plateforme, de longueur maximale de 280 caractères, ce qui ne laisse pas beaucoup de place pour transmettre des émotions nuancées ou des textes complexes. Il serait pertinent d'intégrer d'autres éléments, comme la ponctuation ou le recours aux majuscules pour approcher le sentiment exprimé. Par contre, nous pouvons conclure que les messages comprenant de nombreux mots différents les uns des autres sont favorisés dans les comportements de diffusion de l'information, ce qui laisse penser que les messages au contenu complexe sont plus diffusés.

Ainsi, au niveau des éléments de la route périphérique qui favorisent le partage d'information, notre étude suggère que la source du message est l'élément le plus important, que ce soit au niveau de son audience ou de sa crédibilité, alors que les éléments de forme qui impactent le plus la diffusion de l'information sont ceux immédiatement visibles et interprétables comme les images ou les émojis. Ces éléments laissent supposer que le modèle ELM n'est peut-être pas le plus pertinent pour expliquer le partage d'information sur Twitter, et que la théorie du signal est une piste concurrente à explorer. Une recherche future s'intéressera au rôle du contenu de message et de la route centrale du modèle ELM dans la diffusion de l'information sur Twitter afin d'apporter des éclairages sur ce point.

D'un point de vue méthodologique, le principal apport de ce travail repose sur l'approche mixte que nous avons employé pour traiter les données. Nous avons utilisé différents outils de *text mining* pour extraire des indicateurs de sentiments exprimés, repérer les émojis, détecter les comptes automatisés et créer un indicateur de transparence des profils grâce notamment aux algorithmes de détection d'entités nommées.

Les apports managériaux de cette recherche s'adressent principalement aux *community managers* qui utilisent Twitter, à qui nous pouvons conseiller de bien maîtriser les sujets de conversation sur lesquels ils interviennent pour diffuser plus largement leur message. Nous pouvons aussi leur conseiller de mettre en œuvre une stratégie proactive de réciprocité sur la plateforme, favorisant les messages fréquents et la création de relations bi-directionnelles avec leur audience, tout en affichant clairement leur identité. Pour la forme des messages émis, la présence d'un média et l'utilisation des émojis et des hashtags (à bon escient) sont des éléments pertinents pour être remarqué et diffusé. Enfin, nous pouvons conseiller à Twitter d'intégrer un aperçu des liens présents dans les messages pour favoriser leur visibilité et le partage, ce qui s'apparenterait à la présence d'un média, et ce qui est fait sur de nombreuses autres plateformes de réseaux sociaux.

Cette recherche participe au développement des travaux sur les comportements de mimétisme dans un environnement digital, en particulier celui du partage de l'information et aux travaux visant à exploiter les big data. Les procédures d'analyse employées peuvent encore être améliorées, notamment au niveau de l'analyse des sentiments pour lequel l'approche par dictionnaire peut être dépassée par la mise en œuvre d'algorithme de machine learning après annotations manuelles des sentiments des tweets.

RÉFÉRENCES :

Araujo, T., Neijens, P., & Vliegenthart, R. (2017). Getting the word out on Twitter: The role of influentials, information brokers and strong ties in building word-of-mouth for brands. *International Journal of Advertising*, *36*(3), 496-513. https://doi.org/10.1080/02650487.2016.1173765

Asselin, C. (2021). Twitter : les chiffres essentiels France et Monde pour 2021. Consulté à l'adresse Digimind website: https://blog.digimind.com/fr/tendances/twitter-chiffres-essentiels-france-monde-2020#sources

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., & Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research -- an international collaboration. *arXiv:2004.03688 [cs]*.

Barnier, V. De. (2006). Le modèle ELM : bilan et perspectives. *Recherche et Applications en Marketing*, *21*, 61-83.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192-205. https://doi.org/10.1108/sd.2012.05628haa.014

Berman, R., Melumad, S., Humphrey, C., & Meyer, R. (2019). A Tale of Two Twitterspheres: Political Microblogging During and After the 2016 Primary and Presidential Debates. *Journal of Marketing Research*, *56*(6), 895-917. https://doi.org/10.1177/0022243719861923

Boyd D., Golder S. & Lotan G. (2010), Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter, *2010 43rd Hawaii International Conference on System Sciences*, pp. 1-10, doi: 10.1109/HICSS.2010.412.

DuBay, W. (2004). *The Principles of Readability*.

Jenders, M., Kasneci, G., & Naumann, F. (2013). Analyzing and Predicting Viral Tweets. *Proceedings of the 22nd International Conference on World Wide Web*, 657-664. https://doi.org/10.1145/2487788.2488017

Kearney, M. W. (2018). *R Package: Tweetbotornot*.

Lahuerta-Otero, E., Cordero-Gutiérrez, R., & De la Prieta-Pintado, F. (2018). Retweet or like? That is the question. *Online Information Review*, *42*(5), 562-578. https://doi.org/10.1108/OIR-04-2017-0135

Langley, D. J., Hoeve, M. C., Ortt, J. R., Pals, N., & Vecht, B. Van Der. (2014). Patterns of Herding and their Occurrence in an Online Setting. *Journal of Interactive Marketing*, *28*(1), 16-25. https://doi.org/10.1016/j.intmar.2013.06.005

Li, Q., & Liu, Y. (2017). Exploring the diversity of retweeting behavior patterns in Chinese microblogging platform. *Information Processing and Management*, *53*(4), 945-962. https://doi.org/10.1016/j.ipm.2016.11.001

Li, X., Wu, C., & Mai, F. (2019). The effect of online reviews on product sales : A joint sentiment-topic analysis. *Information & Management*, *56*, 172-184. https://doi.org/10.1016/j.im.2018.04.007

Nesi, P., Pantaleo, G., Poali, I., & Zaza, I. (2018). Assessing the reTweet proneness of tweets : predictive models for retweeting. *Multimedia Tools and Applications*, *77*, 26371-26396.

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLoS ONE*, *10*(12), 1-22. https://doi.org/10.1371/journal.pone.0144296

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology*, *19*(C), 123-205. https://doi.org/10.1016/S0065-2601(08)60214-2

Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and Peripheral Routes to Advertising Effectiveness: The Moderating Role of Involvement. *Journal of Consumer Research*, *10*(2), 135. https://doi.org/10.1086/208954

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation. *Psychologie Française*, *56*(3), 145-159. https://doi.org/10.1016/j.psfr.2011.07.002

Pressgrove, G., & Mckeever, B. W. (2018). What is Contagious ? Exploring why content goes viral on Twitter : A case study of the ALS Ice Bucket Challenge. *International Journal of Nonprofit and Voluntary Sector Marketing*, *23*, 1-8. https://doi.org/10.1002/nvsm.1586

Quesenberry, K., & Coolsen, M. (2019). Twitter Posts That Are Engaging: a Content Analysis of Twitter Brand Post Text That Increases Retweets, Replies and Favorites in Twitter Brand Posts To Influence Organic Viral Reach. *American Academy of Advertising Conference Proceedings*, (1), 120.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rinker, T. (2017). *R Package: Entity*.

Rui, H., Liu, Y., & Whinston, A. B. (2010). Chatter Matters: How Twitter Can Open The Black Box of Online Word-of-Mouth. *Thirty First International Conference on Information Systems*. Saint-Louis.

Sohn, K., Gardner, J. T., & Weaver, J. L. (2013). Viral marketing–more than a buzzword. *Journal of Applied Business and Economics*, *14*(1), 21-42.

Soulard, O. (2015). La crédibilité des avis en ligne : une revue de littérature et un modèle intégrateur. *Management & Avenir*, *82*(8), 129. https://doi.org/10.3917/mav.082.0129

Sunder, S., Kim, K. H., & Yorkston, E. A. (2019). What Drives Herding Behavior in Online Ratings? The Role of Rater Experience, Product Portfolio, and Diverging Opinions. *Journal of Marketing*, *83*(6), 93-112. https://doi.org/10.1177/0022242919875688

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24-54. https://doi.org/10.1177/0261927X09351676

Taylor, A. (2017). A Tidy Data Model for Natural Language Processing using cleanNLP. *The R Journal*, *9*(2), 1-20.

Torruella, J., & Capsada, R. (2013). Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. *Procedia - Social and Behavioral Sciences*, *95*, 447-454. https://doi.org/10.1016/j.sbspro.2013.10.668

Van De Velde, B., Meijer, A., & Homburg, V. (2015). Police message diffusion on Twitter: Analysing the reach of social media communications. *Behaviour and Information Technology*, *34*(1), 4-16. https://doi.org/10.1080/0144929X.2014.942754

Yoo, E., Gu, B., & Rabinovich, E. (2019). Diffusion on Social Media Platforms: A Point Process Model for Interaction among Similar Content. *Journal of Management Information Systems*, *36*(4), 1105-1141. https://doi.org/10.1080/07421222.2019.1661096

ANNEXES :

Annexe 1 : Modèle conceptuel

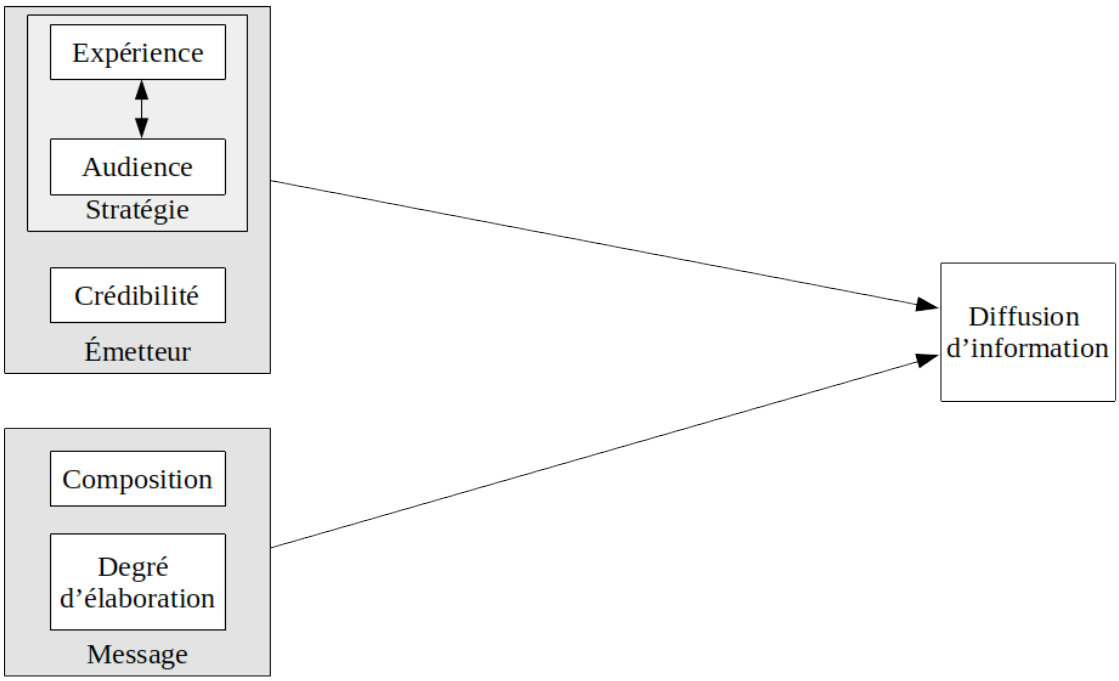

Figure 1 : Modèle testé : variables de la route périphérique agissant sur la diffusion d'information

Annexe 2 : Répartition des tweets dans le temps

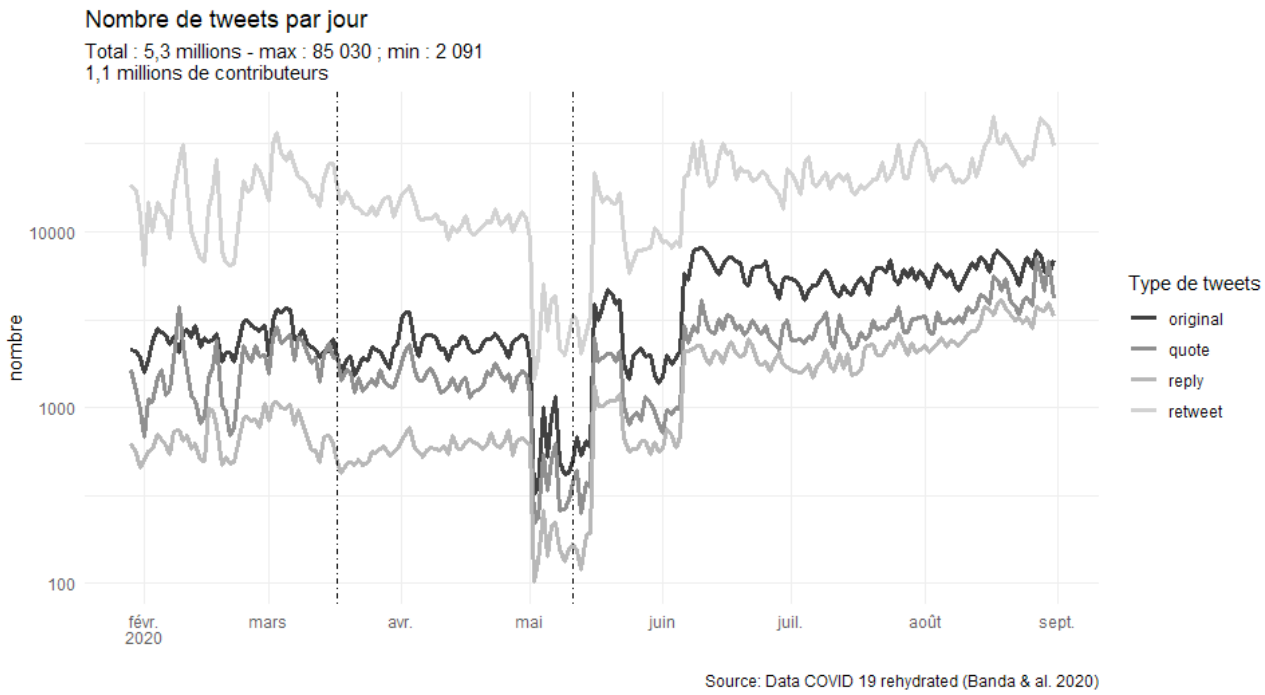

Figure 2 : Distribution des tweets dans le temps. Les lignes verticales représentent la période du confinement en France.

Annexe 3 : Processus de collecte et de traitement des données

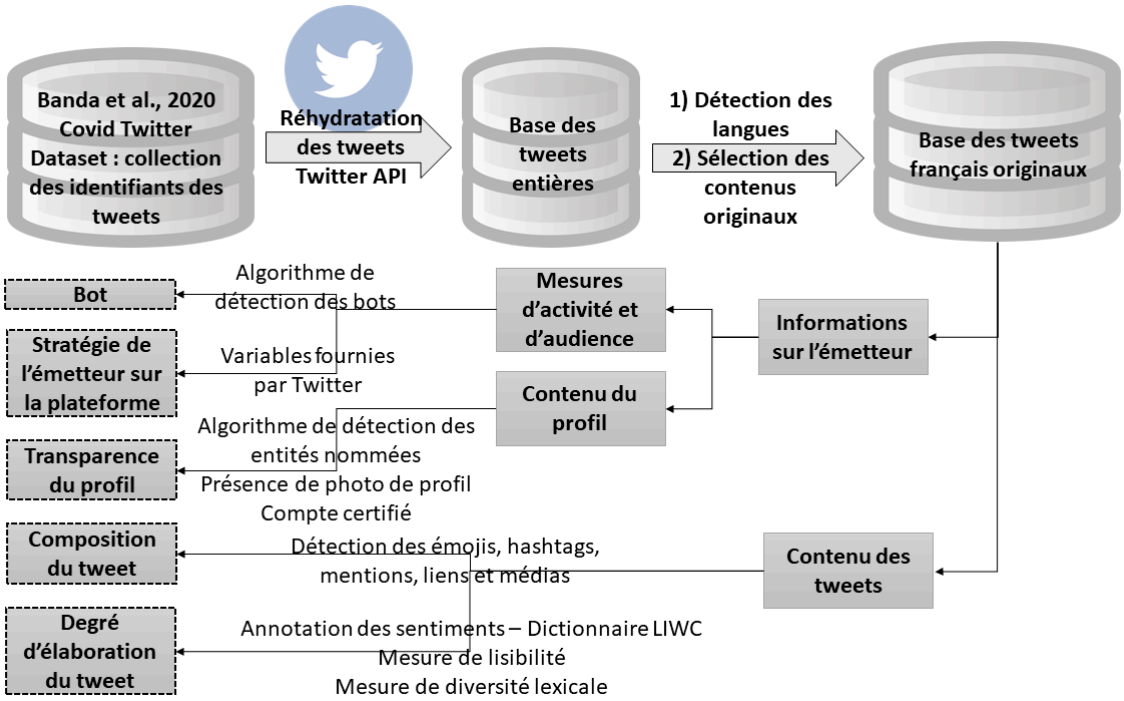

Figure 3 : Processus de collecte et d'opérationnalisation des données

Annexe 4 : Critères d'ajustement des modèles testés

Figure 4 : Indicateurs de qualité d'ajustement des modèles testés

Annexe 5 : Résultats des modèles emboîtés

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
(Intercept)	-0.05***	0.24***	0.23***	0.25***	0.21***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
followers	0.13***	-0.05***	-0.05***	-0.04***	-0.04***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
friends	-0.03***	-0.04***	-0.04***	-0.04***	-0.04***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
favourites	0.02***	0.01***	0.01***	0.01***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
status	-0.03***	-0.07***	-0.06***	-0.05***	-0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
followers:favourites		0.04***	0.04***	0.03***	0.03***
		(0.00)	(0.00)	(0.00)	(0.00)
followers:friends		0.02***	0.02***	0.02***	0.02***
		(0.00)	(0.00)	(0.00)	(0.00)
favourites:friends		-0.03***	-0.03***	-0.03***	-0.03***
		(0.00)	(0.00)	(0.00)	(0.00)
followers:status		0.01***	0.01***	0.01***	0.01***
		(0.00)	(0.00)	(0.00)	(0.00)
profil_transparency			0.01***	0.01***	0.01***
			(0.00)	(0.00)	(0.00)
bot			-0.06***	-0.04***	-0.04***
			(0.00)	(0.00)	(0.00)
emoji				0.05***	0.04***
				(0.00)	(0.00)
mention				-0.06***	-0.07***
				(0.00)	(0.00)
hashtag				0.02***	0.02***
				(0.00)	(0.00)
url				-0.07***	-0.05***
				(0.00)	(0.00)
medias				0.07***	0.08***
				(0.00)	(0.00)
émopos					-0.00
					(0.00)
émonég					0.00
					(0.00)
ARI					-0.00***
					(0.00)
CTTR					0.02***
					(0.00)
émopos:émonég					0.00
					(0.00)
R^2	0.17	0.23	0.24	0.26	0.26
Adj. R^2	0.17	0.23	0.24	0.26	0.26
Num. obs.	797374	797374	797374	797374	797374

***p < 0.001; **p < 0.01; *p < 0.05

Statistical models

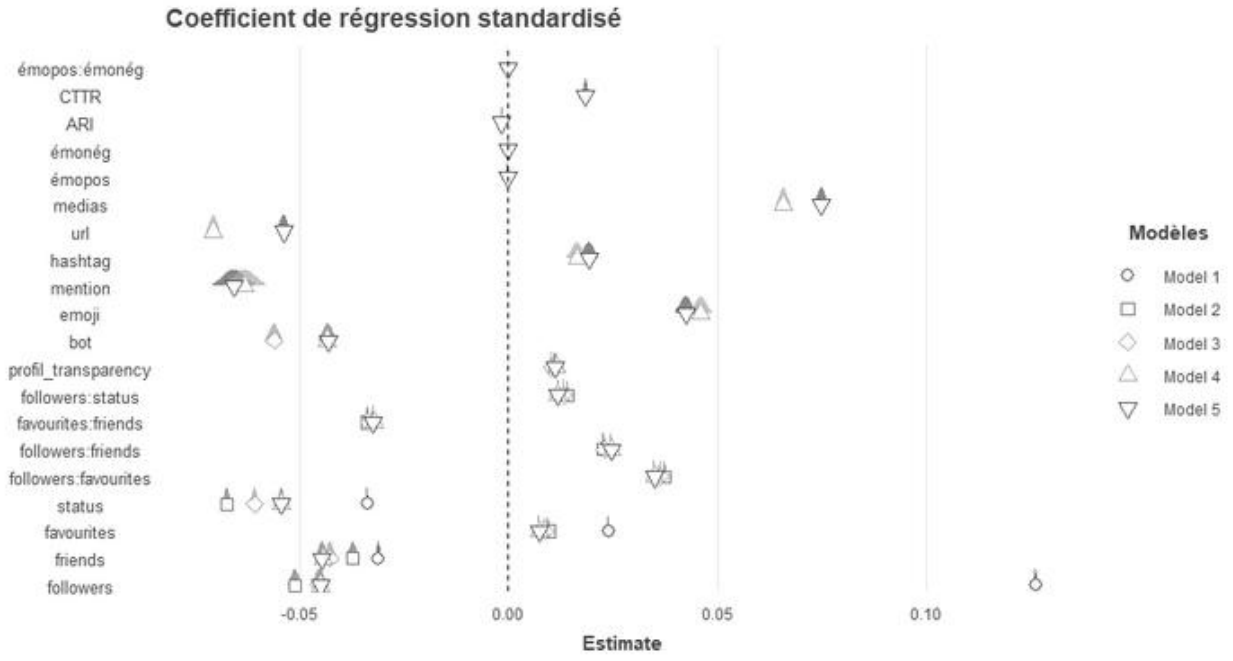

Figure 5 : Coefficients de régression standardisés des modèles emboîtés