Kap Code

RÉSEAUX SOCIAUX ET CORONAVIRUS : CARACTÉRISATION DES CONTENUS ÉCHANGÉS EN FRANCE SUR TWITTER PENDANT LA CRISE SANITAIRE DU COVID-19

Pierre Foulquié[1], Léa Châteauneuf[1], Paméla Voillot[1], Simon Renner[1], Adel Mebarki[1], Nathalie Texier[1] et Stéphane Schück[1]

[1]Kap Code, 28 rue d'Enghien 75010 Paris, France

INTRODUCTION

Suite aux premiers cas de coronavirus enregistrés sur son territoire en février 2020, **la France a pris différentes mesures sanitaires pour répondre à la crise mondiale**. Ces mesures, dont le confinement mis en place le 17 mars, ont impacté le quotidien des français de façon inédite. Les réseaux sociaux, notamment Twitter, source reconnue de données de vie réelle, ont permis aux français de communiquer et d'échanger des informations pendant cette crise. **Cette étude infodémiologique se propose d'étudier la nature et l'évolution des contenus échangés sur twitter à partir de plusieurs méthodes de fouille de texte.**

MATÉRIEL ET MÉTHODES

Un corpus de tweets français liés au Covid-19 a été constitué grâce à l'API Twitter. Les tweets extraits contenaient des mots-clés ou des hashtags évocateurs du coronavirus (*#coronavirusfrance*, *#covid19fr* etc.) et du confinement (*#restezchezvous*). Trois périodes d'intérêt ont été définies : du 10 au 31 mars, le mois d'avril et le mois de mai.

Le contenu des tweets est caractérisé par trois méthodes et les résultats de chaque période sont comparés.

Un modèle de sujet adapté aux textes courts (*biterm topic model* [1]) a été utilisé pour déterminer les thématiques des discussions. Ce modèle classe les messages selon les thématiques qu'ils abordent. Celles-ci se présentent sous la forme de listes de mots apparaissant ensemble dans les messages. Une classification manuelle des adresses web partagées a été opérée, basée sur la nature du site internet (presse, site gouvernemental etc.). Les proportions de ces regroupements dans le total des contenus web partagées est étudiée, ainsi que leur évolution au cours des trois périodes d'intérêt, dans le but de distinguer les types de contenus échangés.

Une identification des terminologies du dictionnaire MedDRA (*Medical Dictionnary for Regulatory Activities*), enrichie de vocabulaire patient [2], a permis d'extraire le vocabulaire médical et les éventuels symptômes du coronavirus et du confinement. Les terminologies détectées sont ensuite regroupées par catégories médicales, grâce à la structure arborescente du dictionnaire MedDRA. Par exemple, « expectorer » et « tousser » sont regrouper dans la catégorie « Toux ».

RÉSULTATS

Le corpus constitué via l'**API Twitter** contient **2,5 millions de tweets** : 933 481 en mars, 833 108 en avril et 774 778 en mai. La figure 1 montre l'évolution temporelle hebdomadaire nombre de messages. Le pic à la gauche du graphique correspond à la première semaine de confinement, qui a généré beaucoup de contenu. Le tableau 1 présente les cinq thèmes les plus abordés, chaque mois. Au mois de mars, ils correspondent à des inquiétudes et difficultés liées au confinement (37% des tweets), ainsi que des commentaires de la gestion de la crise par l'État (27%). Ces thèmes persistent en avril. Ils sont accompagnés par des thématiques plus positives décrivant des solidarités (4,68%) et des idées d'occupations (8,80%). Au mois de mai, les conditions du déconfinement sont largement discutées, avec par exemple la réouverture des écoles (18,09%) et le port du masque (22,55%).

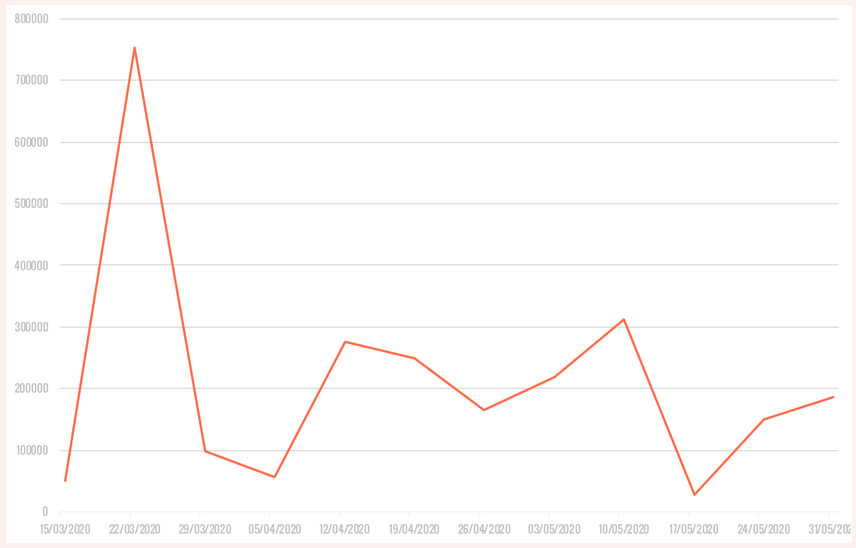

Figure 1 – Évolution hebdomadaire du volume de tweets

	Mars		Avril		Mai	
Thèmes	Thème	P	Thème	P	Thème	P
	Difficultés liées au confinement	37,48%	Difficultés liées au confinement	29,43%	Port du masque quotidien	22,55%
	La crise sanitaire en France	27,00%	Non-respect des mesures de santé publique	15,64%	Écoles	18,09%
	Initiatives suite au confinement	13,52%	La gestion de la crise par L'Etat	10,54%	L'utilité du masque	10,44%
	Mesures de santé pub. FR	9,16%	Positiver pdt le confinement	8,80%	Lutte contre l'épidémie	9,26%
	Le personnel soignant	6,37%	Relations sociales pdt le confinement	6,37%	Retour en entreprise	5,37%
URL	Type	P	Type	P	Type	P
	Hébergeurs de vidéos	20,90%	Hébergeurs de vidéos	19,25%	Presse quotidienne nat.	23,92%
	Réseaux sociaux	13,76%	Presse quotidienne nat.	15,78%	Télévision FR	14,49%
	Presse quotidienne rég.	12,59%	Presse quotidienne rég.	10,03%	Presse quotidienne rég.	11,94%
	Presse quotidienne nat.	11,31%	Télévision FR	6,92%	Hébergeurs de vidéos	8,42%
	Site Service Public	8,50%	Réseaux sociaux	5,83%	Station de radio	6,79%

Tableau 1
Thèmes de discussion abordés (haut) et type de sites partagés (bas)

Tableau 2
Catégories de terminologies médicales exprimées

Terminologies médicales	Mars		Avril		Mai	
	Concept	P	Concept	P	Concept	P
	Mort	9,33%	Mort	11,06%	Mort	10,26%
	Ennui	2,92%	Fatigue	2,25%	Anxiété	1,73%
	Fatigue	1,90%	Ennui	2,11%	Fatigue	1,69%
	Humeur dépressive	1,58%	Humeur dépressive	1,77%	Dépendance	1,37%
	Grippe	1,39%	Usage abusif de substances	1,29%	Etouffement	1,29%

Les sites de presse et d'informations constituaient un tiers des adresses web échangées en mars et en avril, et plus de la moitié en mai. Youtube, avec des vidéos d'ordre informatif, humoristique ou créative, est le site le plus partagé en mars (20% des liens). Cette proportion diminue à chaque période. Une tendance similaire est constatée pour les réseaux sociaux (autres que Twitter). Les sites de services publics sont présents au mois de mars, notamment du fait des décisions de confinement, du partage des gestes barrières et d'attestation de sorties. Cette proportion diminue dès avril. À chaque période, des terminologies médicales sont identifiées dans environ 10% des tweets. Les catégories les plus représentées sont regroupées dans le tableau 2 Les terminologies liées à la mort proviennent des bilans journaliers sur l'évolution de l'épidémie et relayées par les internautes. Si peu de symptômes du coronavirus sont observés, des termes évocateurs de troubles psychologiques sont identifiés : **Anxiété, Ennui, et Dépression**. Des comportements à risque sont également identifiables par les catégories Dépendance et Usage abusif de substances. Ces résultats suggèrent un effet du climat de crise et du confinement sur les individus.

CONCLUSION

L'étude du contenu échangé sur Twitter de mars à mai 2020 permet de qualifier l'usage dont en ont fait les français pendant la crise sanitaire. Les utilisateurs partagent leurs inquiétudes, leurs opinions, et partagent des supports d'information. Ils partagent également des troubles, notamment psychologiques et imputables au confinement. Chaque période analysée contient une ou plusieurs décisions de santé publique majeures (confinement, prolongation du confinement, déconfinement) ayant influencé le contenu des échanges.

Ces résultats souffrent de limites inhérentes à toutes études réalisées à partir des réseaux sociaux.

La représentativité d'abord, puisque seuls 34% des internautes français utilisent activement Twitter [3].

Dans un contexte d'épidémie, les réseaux sociaux sont néanmoins des indicateurs du niveau d'information de la population. Les analyser permet donc d'identifier et de répondre à ses attentes et inquiétudes et d'orienter les politiques publiques sanitaires.

RÉFÉRENCES

[1] Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In Proceedings of the 22nd international conference on World Wide Web (pp. 1445-1456).

[2] Abdellaoui, R., Schück, S., Texier, N., & Burgun, A. (2017). Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help?. JMIR public health and surveillance, 3(2), e36

[3] Hootsuite & We Are Social. Digital 2020: Global Digital Overview. DataReportal – Global Digital Insights https://datareportal.com/reports/digital-2020-global-digital-overview