PAR AMIRHOSSEIN MALEKZADEH *

L'empreinte digitale et ses pollutions

Savoir ce qui se dit de la marque sur le digital impose de bien isoler les messages en parlant réellement des spams homonymes et fantômes.

Comprendre ce qui se passe sur le digital commence souvent par une analyse de la quantité et de la qualité des messages mentionnant la marque, et par une comparaison face à ses concurrents. Élaborer à ce sujet un « brief » simplifié implique de savoir combien de messages il y a eu sur Twitter, Facebook, etc., qui mentionnent la marque. Cela semble simple... cependant, il n'est pas rare d'avoir des écarts importants entre différentes mesures de l'empreinte digitale. Isoler les messages concernant une marque est en effet un exercice complexe. Il arrive que le rapport entre le plus petit et le plus grand des chiffres pour une marque soit de 15. Ainsi, l'éventail de réponses à la question « *combien y a-t-il eu de messages sur moi ?* » peut aller de 10 000 à 150 000 sur une période de temps fixée, et ce en prenant le même ensemble de sources de données. Les écarts s'expliquent par deux séries de facteurs. En premier lieu, il y a la définition de la marque sur ces mêmes espaces : comment identifier tous les messages qui parlent d'une marque ? Puis, au sein de l'ensemble des éléments collectés, il faut nettoyer le bruit pour voir les contours de cette empreinte digitale.

Collecte des messages

La première étape est d'identifier et de rassembler dans une même base d'information tous les messages qui concernent une marque. Pour cela, la méthode standard

* *Président de Focusmatic.*

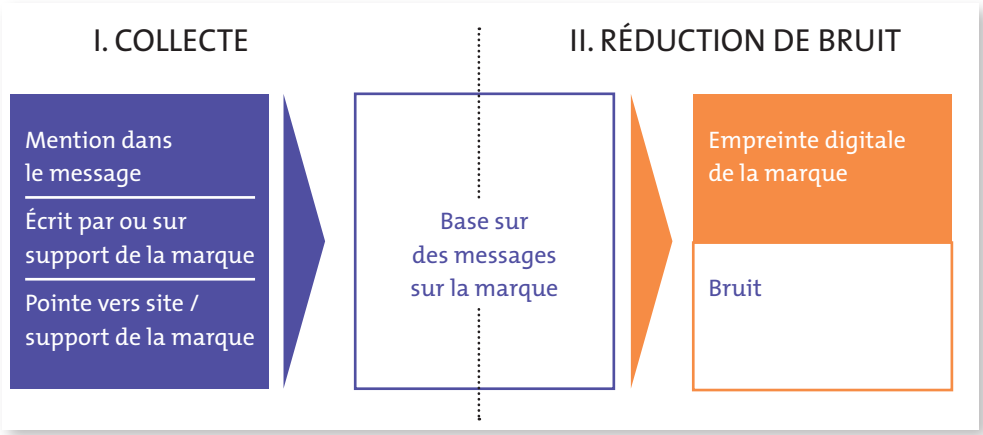

Figure 1 – Définition de l'empreinte digitale.

est de rechercher les mots-clés qui la définissent. Nous pourrons ainsi avoir tous les messages qui contiennent les « mots » de la marque dans leur texte. Si nous nous intéressons aux opérateurs télécoms, nous entrons « orange », « sfr » et « bouygues telecom » ainsi que des variantes d'orthographe (bouygues avec ou sans « s ») ou de concaténation de mots comme « bouyguestelecom ». Mais cela n'est pas suffisant. En effet, il convient de rajouter des messages selon qui les écrits ou encore où ils ont été publiés, sans se soucier du contenu. Le compte Twitter d'une marque ne se mentionne pas toujours dans ses messages. Ou encore, sur Facebook, les forums ou les blogs, les commentaires d'un post ne le reprennent pas forcément. Quand on répond à un post sur tel opérateur en donnant son avis, il est rare de reprendre l'ensemble du message initial. Or, sur une page de marque, les réactions sont typiquement riches et devraient être prises en compte et ajoutées à la base des messages de l'étude. Enfin, il convient d'ajouter les messages qui pointent vers la marque sans que celle-ci soit lisible dans les messages. Sur Twitter, il y a souvent des messages avec un lien raccourci comme « Belle approche http://bit.ly/fsprint ». L'URL http://bit.ly/fsprint est une version raccourcie d'une autre URL. Et celle-ci renvoie vers le site de l'une des marques qui vous intéresse.

L'éventail de réponses à la question « *combien y a-t-il eu de messages sur moi ?* » peut aller de 10 000 à 150 000 sur une même période et en prenant le même ensemble de sources de données.

La réduction de bruit

Une fois l'ensemble des messages collectés, il est nécessaire de bien analyser l'ensemble de la base pour la nettoyer du bruit. En effet, c'est à ce niveau que l'on va trouver des éléments souvent surprenants, mais qu'il est nécessaire de mettre de côté afin d'avoir une véritable image de la marque sur le digital. Nous catégorisons le bruit en quatre types : les homonymes, le spam, l'auditoire fantôme et les sujets inappropriés ou hors d'intérêt. Ces derniers dépendent de chacun. Vous souhaitez probablement voir votre empreinte en isolant des phénomènes comme les revendeurs illégaux de produits, les messages de dénigrement ou bien des événements associés à votre marque (comme un partenariat sportif) mais que vous souhaitez peut-être isoler. Les trois autres « bruits » nécessitent un travail important et des analyses similaires pour toutes les marques.

Les homonymes

Les homonymes nécessitent une compréhension fine du contexte pour les traiter. Sur notre exemple d'opérateurs télécom, l'un d'eux est particulièrement sujet à ce problème : « Orange » s'applique à la fois à l'entreprise, mais aussi à la couleur ou au fruit. Différentes approches permettent de traiter ce sujet, mais cela ne peut être uniquement sur la base du texte. En effet, on peut utiliser des associations de mots ou de la sémantique avancée, mais se référer uniquement au message est insuffisant. Il faut notamment tenir compte de qui écrit et où.

Les deux messages suivants illustrent ce propos. Le premier présente un message d'un auteur dont le nom contient « Orange », mais qui peut par le texte être identifié comme étant un homonyme : il y est fait mention de couleurs, de coucher du soleil, etc. Le second est moins évident, car il ne mentionne

que « Orange cloud », soit la marque, associée à l'un de ses produits/sujets d'expertise, puisque le cloud est une offre de service des opérateurs télécoms à leur clientèle entreprise.

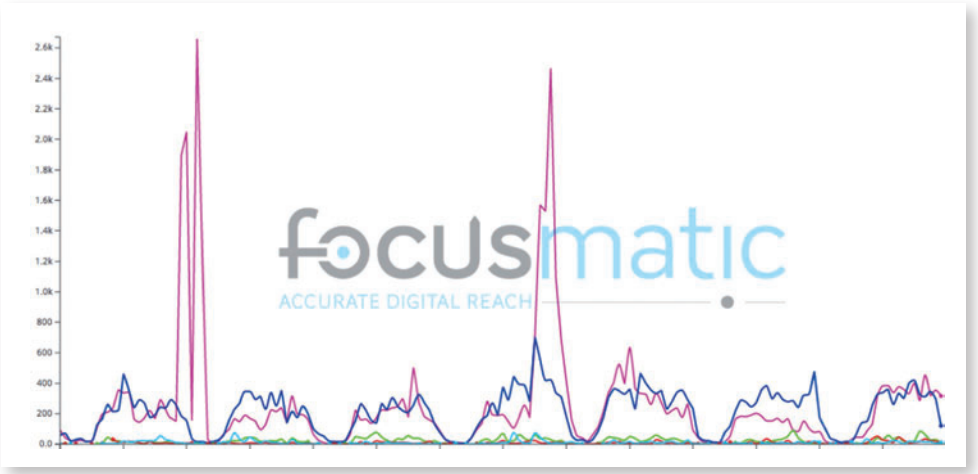

Figure 2 – Activité Twitter heure par heure sur une semaine.

Le spam

La problématique du spam est assez similaire à celle des mails. Des personnes vont chercher à profiter de l'audience d'une marque pour pousser leurs contenus, qui sont bien sûr sans rapport aucun avec celle-ci. Ainsi, il n'est pas rare de voir des pages Facebook avec de longs posts mentionner toute une série de marques. Sur Facebook, il y a souvent une connotation commerciale, tel ce Marabout dont la page mentionnait tous les opérateurs de cloud dans un post d'une longueur extrême.

Cette approche de spam se retrouve sur l'ensemble des réseaux sociaux, ainsi que sur le Web, notamment sur les blogs. Sur YouTube par exemple, les spamers mettent des tags de marques sur leurs vidéos. Ces vidéos ne sont probablement pas les premières que le moteur de recherche renvoie, mais un outil qui va systématiquement indexer toutes les vidéos relatives à une marque les remontera, et comptabilisera donc de l'inutile.

Au niveau des blogs, il y a aussi une surenchère pour être indexé par les moteurs de recherche. On y retrouve une déferlante de posts spams.

Une des méthodes consiste par exemple à industrialiser l'assemblage de textes qui s'appuient sur des contenus tiers. Des articles de l'encyclopédie Wikipedia sur un sujet quelconque sont modifiés en remplaçant ledit sujet (donc toutes les occurrences du mot du sujet) par des noms de marques. Pour un humain, cela n'a aucun sens, mais certains moteurs de recherche pensent qu'il s'agit de contenu original sur la marque et valorisent mieux le blog.

Les fantômes

Il y a enfin une dernière catégorie un peu plus subtile à isoler : l'auditoire fantôme. On trouve ici deux phénomènes aux effets équivalents, il y a tout d'abord des comptes qui sont des « bots » : des robots qui repostent des messages et donc ne peuvent être considérés comme faisant partie de votre audience (mais eux peuvent avoir une audience) ; et il y a des faux comptes ne correspondant ni à un robot ni à une personne réelle. Prenons deux exemples pour illustrer le cas précédent. L'un de nos clients semblait heureux de voir qu'il y avait un buzz autour du lancement de son application mobile. Malheureusement, tous les messages émis l'étaient par des

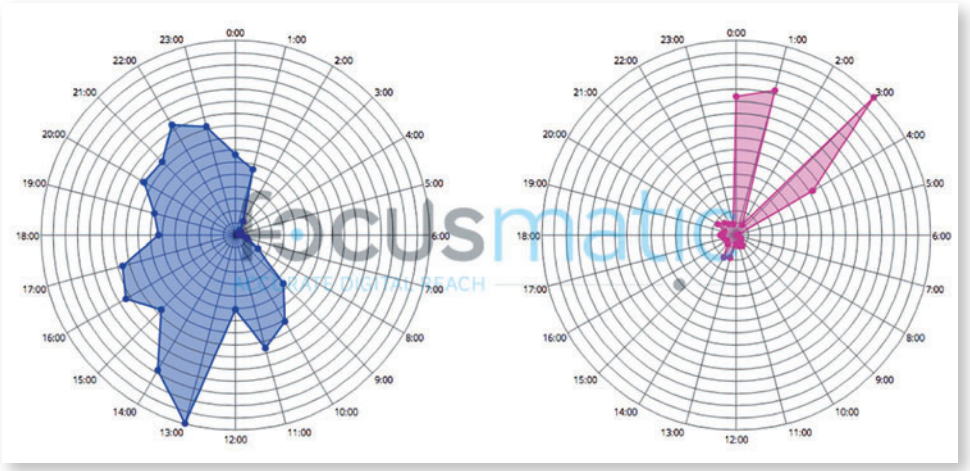

Figure 3 – Répartition des tweets selon les heures de la journée pour deux candidats.

> Les homonymes, les spams, mais aussi tous les faux nécessitent une capacité d'analyse fine des messages, afin de clarifier la vision de ce qu'est réellement la marque sur le digital.

comptes à peine ouverts sur Twitter, n'ayant qu'un ou deux « followers » (donc une ou deux personnes qui lisent leurs tweets), et chacun reprenait le message une fois par jour. Il est assez rare que tout un auditoire n'ait que quelques followers, ou bien qu'ils ne se suivent qu'entre eux. Il y avait effectivement production de tweets, mais en réalité ils n'étaient vus de personne et donc ne méritaient pas d'être comptabilisés. Il faut parfois une analyse encore plus fine pour identifier lesdits problèmes. Au cours de récentes élections, plusieurs hommes politiques se sont vus crédités de nombreux followers, qui reprenaient leurs messages. Une enquête plus approfondie permettait de comprendre que ces messages étaient eux aussi fantômes. La figure 2 montre l'activité heure par heure sur quelques candidats en lutte serrée pour un siège d'élu.

Les pics observés et correspondant à des buzz ne sont pas rares en période électorale, notamment au moment des meetings. Cependant, à y regarder de plus près, dans la lutte entre les deux principaux candidats, les pics de la semaine se sont déroulés en pleine nuit. La figure 3 montre la répartition des messages concernant ces deux candidats selon les heures de la journée, et pour l'un, il est flagrant de voir une écrasante activité entre minuit et quatre heures du matin. Environ 7700 personnes ont émis des tweets concernant ce candidat à ce moment-là. En y regardant de plus près, environ 80 % des comptes ayant tweeté sur ce candidat dans ce créneau horaire avaient moins de cinq followers. Quelques jours plus tard, Tweeter les a tous suspendus. Les homonymes, les spams, mais aussi tous les faux nécessitent une capacité d'analyse fine des messages, afin de clarifier la vision de ce qu'est réellement la marque sur le digital. Pour clarifier la vue, il est important de bien isoler les messages parlant réellement de la marque. Mais il est tout aussi important de retirer les messages et audiences qui ne sont que des simulacres. ■