Réduire les biais dans la collecte de tweets

Béatrice Mazoyer*, Nicolas Hervé*,
Céline Hudelot **, Julia Cagé***

*Institut National de l'Audiovisuel
bmazoyer@ina.fr, nherve@ina.fr
**CentraleSupélec
celine.hudelot@centralesupelec.fr
***SciencesPo Paris
julia.cage@sciencespo.fr

Les réseaux sociaux sont une source de données potentielle pour de nombreuses disciplines en sciences humaines et sociales (sociologie, sciences politiques, sciences de l'information et de la communication, économie, etc.). Malgré la variété des domaines concernés et des approches choisies, tous ces travaux sont confrontés à des difficultés de captation, liées aux restrictions d'accès à leurs données qu'imposent les plateformes, et qui rendent délicate la constitution de jeux de données représentatifs du phénomène étudié.

Nous présentons dans ce résumé des solutions pour répondre à ce problème, développées dans le cadre d'une thèse sur la propagation de l'information entre Twitter et les médias traditionnels. Même si le sujet de cette étude est spécifique, nos propositions sont d'intérêt pour de nombreux autres domaines d'études qui se fondent sur les contenus issus de Twitter.

1 Contraintes imposées par Twitter

Twitter ne donne pas accès à l'intégralité de ses données. Les API proposées par la plateforme limitent fortement le volume de tweets disponibles et la durée pendant laquelle ils peuvent être collectés. Pour accéder aux API, il est nécessaire d'avoir une clef d'accès associée à un compte Twitter et à un numéro de téléphone, ce qui limite fortement le nombre de clefs que l'on peut obtenir. Beaucoup d'études sur des données Twitter sont donc fondées sur de petits jeux de données de tweets contenant certains mots-clefs ou émis par des comptes sélectionnés manuellement. Ces études utilisent vraisemblablement l'API "Filter" de Twitter, qui permet de capter en continu les tweets contenant certains termes ou provenant de certains comptes. Cette méthode permet d'avoir un accès complet à tous les tweets qui contiennent ces mots, avec une limitation à 1% du volume mondial de tweets. D'autres utilisent l'API "Sample", donnant accès à 1% aléatoire de tous les tweets émis dans le monde, ce qui leur fournit très peu de données concernant leur sujet.

Notre problématique étant l'analyse de la diffusion de l'information médiatique en France, nous ne pouvons pas nous limiter à un petit nombre de mots qui biaiseraient l'étude vers certaines thématiques. Nous ne pouvons pas non plus utiliser l'API "Sample", car n'avoir accès qu'à 1% de ce qui est émis ne ferait pas apparaître certaines "petites" thématiques, qui n'apparaîtraient même pas dans l'échantillon.

2 Approche choisie pour la constitution d'un jeu de données

Nous proposons un autre angle d'approche : au lieu de collecter les tweets contenant certains mots-clefs ou émis par certains comptes, nous utilisons des mots neutres (les mots français les plus couramment utilisés dans un échantillon collecté avec l'API Sample) pour constituer nos requêtes, avec "fr" comme paramètre de langue. Comme le français est une langue peu parlée sur Twitter (1,8% des tweets), la restriction du volume à 1% du total émis est finalement peu contraignante et nous permet d'avoir accès à une part très importante des tweets. Théoriquement, cette API peut fournir jusqu'à 55% ($\frac{1\%}{1.8\%}$) des tweets émis en français.

Pour maximiser le volume de tweets captés et limiter les biais de captation, nous avons fait varier différents paramètres dans nos requêtes à l'API : nombre de clef d'accès à l'API, nombres de mots-clefs utilisés, répartition des mots clefs sur les clefs d'accès. En effet, il est inutile d'utiliser chaque fois les mêmes mots-clefs "neutres" avec différentes clefs d'accès : dans ce cas, tous les accès à l'API renvoient les mêmes tweets. Nos tests ont montré qu'il était plus utile de grouper ensemble dans un même accès à l'API des mots fréquemment co-occurrents, de façon à réduire au maximum l'intersection entre les ensembles de tweets. Nous avons comparé les résultats obtenus en faisant varier le nombre de clefs d'accès de 1 à 3, et les nombre de mots-clefs de 50 à 400. Nos meilleurs résultats, aussi bien en termes de volume collecté que de similarité avec l'échantillon aléatoire, ont été obtenus avec 3 clefs d'accès et 200 mots.

3 Présentation du jeu de données obtenu

Au total, nous avons capté 4,5 millions de tweets par jour pendant un an pour obtenir une collection de plus 1,6 milliards de tweets, en associant notre méthode de collecte et l'API "Sample". Cette méthode a des limitations qui nous paraissent inévitables : d'une part, nous ne captons pas les tweets qui ne contiennent pas de mots. Cela est impossible avec l'API "Filter", et nous avons observé que même les tweets en français renvoyés par l'API "Sample" contiennent tous au moins 4 caractères. D'autre part, l'identification de la langue des tweets n'est pas parfaite : certains tweets contenant un seul terme ("excellent", "vote !") sont renvoyés à tort par l'API alors que leurs auteurs sont manifestement anglophones, ce qui mène à une sur-représentation des comptes américains (puisqu'ils sont bien plus nombreux sur Twitter). Cependant, la comparaison de nos données à celles collectés par d'autres chercheurs utilisant des mots propres à l'actualité française nous a permis d'éprouver la solidité de notre jeu de données.

Summary

Social networks are a potential source of data for many research fields. All of them are confronted with the data access limitations imposed by the social media platforms, which make it difficult to build representative datasets.

We present solutions to this problem, developed as part of a PhD thesis on the spread of information between Twitter and traditional media. Our proposals are of interest for many other fields of study that rely on Twitter content.