

Opinion mining : des vertus de l'hybridation humain/machine

L'*opinion mining* consiste en une analyse automatisée des verbatims en ligne (messages, blogs, forums...) afin d'en extraire des informations subjectives (jugement, évaluation, émotion). Une approche qui nécessite toutefois une forte expertise humaine...

On ne peut en vouloir aux fournisseurs de solutions d'*opinion mining* d'avoir mis en avant les vertus de l'automatisation de leurs services puisque l'informatique et la linguistique avaient elles-mêmes adopté un vocable en français « TAL » (traitement automatique du langage, ce qui fait une différence avec « NLP » *Natural Language Processing*) qui laissait entendre qu'il existait des fondements scientifiques et techniques à cette automatisation. Cependant, la demande a augmenté de façon impressionnante depuis que les marques ont considéré qu'il devenait important – pour leurs investisseurs avant tout – de suivre leur réputation sur le web et sur les réseaux sociaux, tant les volumes de verbatims produits dans ces plates-formes étaient massifs et contagieux.

L'AUTOMATISATION ET SES LIMITES

On ne peut comprendre l'explosion de l'offre de l'*opinion mining* (280 services en 2012) si l'on ne renvoie pas à cette anxiété du marketeur en charge du suivi de la réputation d'une marque, de la répercussion des événements qu'il génère, des conversations des communautés qu'il soutient, des risques

mais il se trouve que ces méthodes gagnent même les approches scientifiques malgré leurs limites remarquables.

Cependant, il faut faire la distinction entre ce que les promoteurs de ces méthodes annoncent lorsqu'ils vantent leurs systèmes automatisés tout au long de la chaîne de traitement et leurs pratiques réelles qui s'appuient sur une hybridation beaucoup plus marquée entre approches automatisées, *machine learning* et expertise humaine. Il serait donc plus juste de parler d'« *opinion mining* assisté par ordinateur », comme cela se faisait pour la traduction, mais tout se passe comme si ces formules révélaient une faiblesse des ordinateurs en question et ne suffisaient à vendre l'anxiolytique qu'est l'*opinion mining*.

Selon les objectifs poursuivis, cet équilibre ou cette composition peuvent être différents :

- **centrer le suivi sur des agrégats en masse** autour de quelques mots-clés et suivre leur évolution sur un grand nombre de sources reste nécessairement très grossier et peut donner lieu à une automatisation plus avancée ;
- **suivre à la trace un énoncé qui fait événement** peut aussi être automatisé largement si l'on accepte la réduction de l'étude des dérivations faites à partir des énoncés d'origine ;

- **monitorer en permanence un milieu ou des communautés** qui échangent sur des thèmes qui leur sont en partie propres requiert nécessairement des expertises humaines plus importantes parce que les bases de dictionnaires standards n'existent pas toujours et que les énoncés peuvent être fortement contextualisés.

Un simple contrôle de qualité de ces prestations devrait entraîner à examiner la qualité des personnels, des process et des produits. Or, sur toute cette chaîne, de nombreux maillons sont conçus comme des boîtes noires, car ce sont les résultats qui comptent. Ceux-ci sont souvent présentés à un tel niveau d'agrégation à l'aide de tableaux de bords si séduisants que la discussion est rarement possible sur les conditions de leur production. La combinaison avec un choix d'exemples judicieux rend la démonstration toute puissante : « *les clients focalisent leurs critiques sur la vitesse du processeur et n'aiment*

L'automatisation est intéressante si elle s'appuie sur du *machine learning* dans lequel un expert validera les retours de verbatims sélectionnés.

de crise qui ferait dévisser le cours de l'action en quelques minutes. C'est pourquoi toutes les solutions ont trouvé un moment leur place sur le marché en dépit des faiblesses criantes de leurs approches¹. C'est pour contrer cette angoisse que les outils proposés doivent avant tout permettre la réaction et non la réflexivité, qui supposerait modèles sophistiqués ou discussion entre causalités. Le *big data* est convoqué avec ses puissances de calcul et ses approches de *matching* (correspondance) générant des corrélations entre séries de données pour traiter tous ces verbatims, qui remplissent les conditions de volume, de variété mais surtout de vélocité. Tout autre usage de ces techniques peut se faire dans des conditions de scientificité différentes,

[1] Documentées en détail dans : Dominique Boullier, Audrey Lohard, *Opinion mining et sentiment analysis*, Open Editions Press, 2012, <http://books.openedition.org/oep/198>

pas l'image renvoyée par l'acteur X dans les pubs du produit » est un énoncé qui vaut programme d'action immédiate et qui ne nécessite pas de remonter dans le traitement qui a été effectué. Or, cette mise en forme est elle-même déjà humaine et fortement orientée par des stratégies de recommandations considérées par le prestataire comme essentielles ou plus lisibles. Les données ne sont pas données, elles sont obtenues et dans ce dernier maillon elles sont mises en scène.

UNE EXPERTISE INDISPENSABLE EN AMONT...

Si l'on prend la chaîne de traitement dans l'autre sens, le *sourcing*² constitue un moment décisif pour lancer les robots qui vont récupérer les verbatims. Or, il ne peut quasiment jamais être constitué automatiquement. Il faut au minimum choisir quelques mots-clés, ce qui ne peut être fait qu'avec un minimum d'expertise humaine sur le domaine. Il faut aussi orienter l'exploration et la collecte vers quelques sources de statuts distincts, car les techniques elles-mêmes seront très différentes.

On ne récupère pas les tweets comme les verbatims de blogs, on ne peut pas traiter les commentaires d'un journal à forte audience comme ceux d'un blog confidentiel, il faut donc décider d'une stratégie d'exploration et de collecte et pondérer ces sources. Cette méthode ne peut pas être automatisée à partir d'un indicateur *a priori* car il faut tenir compte des spécificités de chaque domaine ou de chaque problème exploré.

Le *sourcing* est une activité qui ne peut jamais se conduire sans l'assistance marquée d'un expert du domaine. Le nettoyage reste une phase essentielle – les robots collectent beaucoup de bruit – souvent automatisée dès lors qu'un expert a été capable de repérer les biais de la requête effectuée. Puis, les premiers traitements de base peuvent être le plus souvent automatisés s'ils restent fondés sur des chaînes de caractères, des cooccurrences, voire des entités nommées que l'on sait bien identifier désormais. Tout dépend de la qualité du dictionnaire constitué car chaque domaine demande le sien et les grandes sociétés de service d'*opinion mining* ont dû constituer leurs propres ressources internes au fur et à mesure pour devenir de plus en plus pertinentes. L'automatisation peut devenir intéressante si elle s'appuie sur du *machine learning* dans lequel un expert jouera un rôle en validant les retours de verbatims sélectionnés et leur qualification et/ou catégorisation. La machine pourra alors reprendre ces mêmes critères et les étendre à tout le corpus pour développer ses propres inférences.

Certains prestataires offrent des outils d'analyse sémantique plus sophistiqués qui permettent de traiter des ambiguïtés, des synonymies, des formes complexes comme des doubles négations, etc. Pour

cela, il a fallu que des experts se mobilisent sur des corpus divers pour constituer l'équivalent d'ontologies d'un domaine particulier, et cela domaine par domaine. Ces techniques sont notamment utiles pour effectuer l'analyse de tonalité (ou *sentiment analysis*) qui donne lieu encore très souvent à des approximations (positif/négatif/neutre), mais qui présente l'avantage d'alerter plus aisément sur les points problématiques dans les verbatims pour pouvoir réagir rapidement. Une phase de tests de robustesse est souvent oubliée alors que l'apparente validité d'un résultat de traitement ne peut être établie sans comparaison des méthodes sur des corpus tests de façon à identifier la production de silence ou de bruit. Or, toutes ces approches nécessitent une validation par des experts. La relation entre l'expert du domaine et l'expert linguistique est essentielle et peut constituer un retour de pertinence qui met les demandeurs des suivis dans la boucle : c'est en réalité à ce moment que les parties prenantes comprennent mieux les qualités et les limites des systèmes automatiques et peuvent voir reconnues leurs propres compétences.

... ET EN AVAL

Enfin nous l'avons vu, la présentation finale des résultats est elle-même une activité stratégique de communication qui fait intervenir des spécialistes humains : la mise à disposition de volumes de données dites brutes (ce qu'elles ne sont jamais) ou d'indicateurs agrégés à partir des masses de verbatims constitue en fait une démarche d'intimidation des demandeurs, qui ne pourront quasiment jamais s'orienter dans cette prolifération de données ou pire qui en feront des interprétations abusives en se focalisant sur des cas non significatifs. Tout travail d'ouverture des boîtes noires est utile de ce point de vue, quand bien même il serait moins séduisant et plus exigeant, et même peu efficace, pour calmer l'angoisse du marketeur évoquée plus haut.

Toutes ces méthodes sont en cours de construction de conventions qui demanderaient une intervention plus argumentée des sciences humaines et sociales comme nous le proposons avec nos sciences sociales de troisième génération³.

DOMINIQUE BOULLIER

Professeur des universités à Sciences Po Paris
Directeur exécutif de Forccast (projet de formation innovante Idefi)
dominique.boullier@sciences-po.fr



↑ Illustration de Charis Tsevis pour la couverture de *Communication of the ACM*, avril 2013, revue publiée par l'Association for Computing Machinery, avec un dossier sur l'analyse de tonalité (*sentiment analysis*).

[2] Ensemble des opérations, préalables à la collecte de données, qui visent à identifier des sources (sites web, blogs, forums, etc.) contenant ou susceptibles de contenir de l'information.
[3] Voir le site <http://shs3g.hypotheses.org>