Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Atefeh FARZINDAR, Diana INKPEN. Natural Language Processing for Social Media, Second Edition. *Morgan & Claypool publishers*. 2017. 175 pages. ISBN 978-1-68173-614-3.

Lu par **Pascal VAILLANT**

Université Paris 13 / LIMICS – UMR INSERM 1142

Atefeh Farzindar et Diana Inkpen, spécialistes de traitement automatique des langues et de sciences des données, ont coécrit cet ouvrage en langue anglaise sur le traitement des contenus textuels des médias sociaux. Le livre consacre une grande partie de ses pages aux nouvelles méthodes que l'on doit concevoir pour adapter les techniques existantes de traitement automatique des langues aux matériaux spécifiques de ce domaine en pleine expansion. Il aborde ensuite les différents domaines d'application de l'analyse textuelle des médias sociaux avec leurs problématiques spécifiques. Enfin, il traite des différentes questions que posent la collecte et l'annotation de ce type de données.

Un volume ahurissant de données est créé chaque jour sur les médias sociaux : en juin 2019, à chaque minute qui s'écoule, trois millions de messages sont postés sur Facebook à travers le monde, et un demi-million sur Twitter. Cette sphère de données est en phase d'expansion rapide.

Une partie importante de l'information de ces messages est constituée de texte en langue naturelle (pas nécessairement la plus importante en termes de quantité d'octets, mais certainement la plus importante en termes de clés de compréhension du contenu des messages). Ces flux de données gigantesques sont porteurs d'informations sur la manière dont les foules réagissent en temps réel aux stimuli de leur environnement et aux nouvelles qui leur parviennent. Chaque utilisateur individuel, en envoyant son message, même insignifiant, contribue plus ou moins consciemment à un vaste hypertexte qui fournit des informations économiques, écologiques, sociologiques et politiques.

De nombreux acteurs ont donc un vif intérêt à voir se développer la possibilité d'analyser ces flux pour détecter les informations qui les concernent. Les champs d'application sont nombreux, de l'analyse de tendances marketing à la détection de menaces terroristes, en passant par la recherche de signaux d'alerte pour la pharmacovigilance.

Les technologies de la fouille de textes (ingénierie linguistique, apprentissage automatique) sont donc de plus en plus souvent sollicitées pour analyser les corpus de médias sociaux.

Des problèmes spécifiques pour l'analyse linguistique

Les méthodes développées pendant cinq décennies par l'informatique linguistique pour le traitement automatique des langues ont été soumises à de nouvelles contraintes lorsqu'il s'est agi de les appliquer aux médias sociaux. De nombreuses tâches (reconnaissance d'entités, résolution de chaînes d'anaphores...) impliquent des phases de prétraitement qui fonctionnent de manière optimale avec des corpus de textes constitués de phrases complètes, écrits dans une langue homogène formelle, et avec une orthographe normalisée. Ces conditions ne sont en général pas présentes dans les contenus engendrés par les utilisateurs des médias sociaux. L'exactitude d'une tâche comme l'étiquetage en parties du discours (pour prendre l'exemple d'une tâche éprouvée) tombe de 97 % à 80 % lorsque l'on passe d'un corpus d'articles de journaux à un corpus de *tweets*, microgenre textuel (*microblogging*) limité à cent quarante signes (à l'époque où l'étude a été réalisée), et où pullulent les ellipses et les abréviations.

Pourtant les utilisateurs (habitués) comprennent les *tweets*. Ce qui leur permet de les comprendre est que l'information qui n'est pas présente dans le texte lui-même est présente dans son entour hypertextuel et dans ses composantes multimodales. Un *tweet*, par exemple, est souvent inséré dans une conversation, et le contexte de la conversation n'est pas repris, car il est supposé connu du lecteur. Il est émis par un interlocuteur qui occupe une certaine position dans un graphe, par rapport au lecteur autant que par rapport à d'autres interlocuteurs – que l'on cite, que l'on mentionne, auxquels on répond. Il contient des « mots-dièses » (*hashtags*) qui sont autant d'ancres hypertextuelles dont le but est de se positionner dans un espace de discussion dynamique. Il contient des « émoticones », pictogrammes numériques qui permettent en un seul caractère *Unicode* d'exprimer des sentiments ou des prises de position. Il est accompagné d'images qui véhiculent une partie de l'information, qui certes ne prend sens qu'avec le texte, mais dispense le texte de la développer.

Dans leur chapitre 2, Farzindar et Inkpen ont passé en revue différentes tâches élémentaires qui constituent les « briques de base » de l'ingénierie linguistique (segmentation, étiquetage, *chunking*, analyse syntaxique, détection d'entités nommées, identification de langue) et ont caractérisé, pour chacune d'elles, les adaptations qu'elles doivent subir pour être adaptées aux corpus de textes de médias sociaux.

Ces adaptations peuvent consister en prétraitement des textes eux-mêmes (normalisation de la ponctuation), en annotation de corpus d'entraînement (pour réentraîner les étiqueteurs ou les analyseurs à la « syntaxe Twitter »), en redéfinition des catégories de sortie (adaptation de l'ensemble des catégories de parties du discours pour prendre en compte des éléments non linguistiques tels que mots-dièses, mentions nommées, émoticones, images, URL), ou en réentraînement des paramètres des algorithmes d'apprentissage (par exemple pour l'identification de langue).

Après ce tour d'horizon des phases d'adaptation des techniques existantes pour les textes des réseaux sociaux, le chapitre 3 aborde la question de leur interprétation. Il s'agit, bien sûr, dans ce contexte d'interprétation automatique, du processus consistant à inférer à partir des textes, des représentations formelles qui pourront être utilisées pour l'analyse automatique agrégée de grandes quantités de ces textes. Dans le chapitre 3, les auteurs exposent donc les informations que l'on peut en extraire. Elles montrent, par exemple, que des algorithmes d'apprentissage neuronal profond peuvent localiser, avec un certain degré d'exactitude, les utilisateurs à partir de ce qu'ils écrivent, même lorsque ceux-ci n'autorisent pas leur appareil à partager leurs coordonnées géographiques exactes. Elles expliquent ensuite dans quelle mesure certaines des tâches les plus courantes en analyse d'information peuvent être menées sur des textes de médias sociaux.

L'annotation sémantique consiste à attribuer aux entités détectées dans les textes des étiquettes correspondant aux éléments connus d'une source de connaissances « contrôlée » (thesaurus ou ontologie) : on parle d'*entity linking*. Pour atteindre une certaine efficacité sur des textes de type *microblog*, il faut faire usage de tous les indices possibles (contexte conversationnel, mots-dièses, mentions nommées...) ; encore que le score maximal atteint dans l'état de l'art, celui du système YODIE développé par l'équipe GATE, plafonne-t-il à 0,45 (en termes de F-mesure).

D'autres tâches, comme la détection d'opinion, d'émotion, ou de sarcasme, nécessitent également des adaptations du même type, c'est-à-dire, en résumé, l'utilisation de techniques éprouvées de TAL et d'apprentissage, mais en élargissant le spectre des variables d'entrée pour prendre en compte les liens extérieurs et les relations connues dans le graphe des utilisateurs (par exemple, le sens d'un « bravo ! » est à interpréter différemment s'il est adressé en réponse au message d'un utilisateur contre lequel il existe un historique d'opposition polémique). L'état de l'art sur les tâches de détection d'événements ou de sujets, de résumé automatique, et de traduction automatique, est également exploré.

Une large gamme d'applications

Dans leur chapitre 4, les auteurs passent en revue les différentes applications possibles de l'analyse de corpus de médias sociaux. Celles auxquelles est consacrée une section sont celles qui font l'objet d'un corpus de recherche abondant et dynamique.

Dans le domaine de la santé, l'analyse des réseaux sociaux peut être utile pour la pharmacovigilance (en permettant de détecter les effets secondaires dont se plaignent spontanément les utilisateurs) ; une nouvelle partie a également été ajoutée dans cette deuxième édition de l'ouvrage pour présenter l'application de détection de signaux d'alerte de la dépression ou d'envies de suicide.

Dans le domaine de l'analyse financière, ce que disent les utilisateurs sur Twitter peut être un indicateur du moral des consommateurs, ou de celui des investisseurs. Les opinions exprimées sur des entreprises sont également corrélées à leur valeur d'investissement, qu'elles en soient le reflet ou – en partie – la cause : une étude a

montré que des jugements (positifs ou négatifs) exprimés sur Twitter se reflétaient dans un délai d'un à dix jours sur la valorisation boursière des entreprises.

Les textes des réseaux sociaux peuvent également être utilisés pour prédire les intentions de vote, faire de la mercatique en ligne, ou plus généralement modéliser certaines caractéristiques de la personnalité de l'utilisateur pour toutes sortes de buts. Enfin, la surveillance des sujets émergents de façon massive et plus ou moins soudaine dans une zone précisément localisée peut donner des indices forts sur la survenue d'événements catastrophiques : raz-de-marée, tremblement de terre, pollution, ou attaque terroriste. Ces informations peuvent être corrélées à d'autres sources de données (sismographes, capteurs de qualité de l'air...), mais n'en restent pas moins utiles : la pénétration des *smartphones* dans l'Inde du Nord est bien plus élevée que celle des capteurs de qualité de l'air. Quant à la surveillance du terrorisme, l'analyse des productions des utilisateurs sur les réseaux sociaux ne permet pas seulement de savoir quand une attaque survient, elle peut également servir à détecter des signes précurseurs de la radicalisation d'un utilisateur (focalisation sur certains types de sujets, changement dans la typologie des utilisateurs fréquentés, adoption d'un vocabulaire marqué).

Enfin, le chapitre 5 examine les problèmes spécifiques que posent le recueil de corpus de textes dans les médias sociaux (problèmes techniques de débit et de volume, mais aussi problèmes de vie privée et de confidentialité des données), puis l'annotation, et enfin l'évaluation, avec un tour d'horizon des différents *benchmarks* d'évaluation qui ont été conçus spécialement pour les systèmes d'analyse des réseaux sociaux.

Un état de l'art utile en 2018

L'ouvrage de Farzindar et Inkpen offre, en somme, un état de l'art particulièrement utile et bien informé de l'avancée des travaux sur le domaine de l'analyse des médias sociaux (tout particulièrement centré sur Twitter). De nombreuses références, soigneusement mises à jour pour cette deuxième édition, donnent un aperçu panoramique des travaux actuels.

Le lecteur qui y chercherait une réflexion plus approfondie, en termes d'analyse linguistique des nouveaux genres textuels créés par ces pratiques d'échanges, risquerait de rester sur sa faim. Il y aurait beaucoup de choses à dire sur les modes d'interprétation de ces genres émergents, comportant moins de contenu textuel *in praesentia* et beaucoup plus de références hypertextuelles et intermodales. Le livre *Natural Language Processing for Social Media* se cantonne aux technologies du langage telles qu'elles étaient fin 2017. Il a sur ce plan, en tant qu'état de l'art à une époque bien précise, une utilité incontestable.