Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Intégration de l'apprentissage automatique au codage du Recensement canadien de 2021 à l'aide de fastText

par Andrew Stelmack

Date de diffusion : le 05 novembre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Intégration de l'apprentissage automatique au codage du Recensement canadien de 2021 à l'aide de fastText

Andrew Stelmack[1]

Résumé

Dans le cadre du traitement du recensement canadien de 2021, les réponses en toutes lettres aux 31 questions du recensement doivent être codées. Jusqu'en 2016, il s'agissait d'un processus en trois étapes, dont une deuxième étape de « codage interactif (humain) ». Cette étape de codage humain est à la fois longue et coûteuse, s'étalant sur de nombreux mois et nécessitant le recrutement et la formation d'un grand nombre d'employés temporaires. Dans cette optique, pour 2021, cette étape sera soit complétée ou entièrement remplacée par des modèles d'apprentissage automatique à l'aide de l'algorithme « fastText ». Dans cette présentation, nous discuterons de la mise en place de cet algorithme ainsi que des défis et des décisions prises en cours de route.

Mots clés : traitement du langage naturel, apprentissage automatique, fastText, codage

1. Introduction

1.1 Motivation

Le Recensement canadien de 2021 pose 31 questions qui peuvent donner lieu à une réponse écrite qui ne correspond pas à l'une des options présentées au répondant dans les « cases à cocher ». Par exemple, la question « Quelle(s) langue(s) la personne parle-t-elle régulièrement à la maison? » a des centaines de réponses possibles, mais le répondant ne verra que trois options : *Anglais*, *français* et *autre(s) langue(s) – précisez*. Il incombe à l'équipe de traitement du codage du recensement d'attribuer une valeur codée aux réponses en toutes lettres à l'option *Autre(s) langue(s) – précisez*. À titre d'exemple fictif, les réponses écrites « allemand » peuvent devoir être attribuées au code 12345 et « italien » au code 67890.

À première vue, cette tâche ne semble pas extrêmement compliquée. Cependant, les répondants n'ont pas d'obligation sur la façon de répondre à une telle question. Leurs réponses comprendront probablement différentes orthographes ou même des réponses qui n'ont aucun rapport avec la question posée. De plus, le nombre de réponses nécessitant un codage varie d'environ 2 000 à 23 000 000, selon la variable codée. En raison de ces facteurs, le processus de codage est une entreprise d'une très grande envergure qui prend environ 10 mois. En raison de cette complexité, Statistique Canada a pris la décision de compléter son processus de codage au moyen d'applications d'apprentissage automatique pour le cycle du recensement de 2021. Le présent article décrit l'ajout de l'apprentissage automatique au processus actuel et indique les nombreuses différences entre le projet et le flux de production classique de l'apprentissage automatique ainsi que les décisions prises pour surmonter ces différences.

1.2 Processus de codage traditionnel

Traditionnellement, le processus se déroulait en trois étapes : le codage automatique, le codage interactif et la correction de code. Le codage automatique consiste à apparier la réponse écrite à un fichier de référence de réponses attendues créées par des experts en la matière, à savoir l'orthographe correcte de réponses ou des orthographes incorrectes très courantes de ces réponses. Si une réponse correspond au fichier de référence, elle reçoit le code inscrit dans le fichier de référence et passe à l'étape de correction de code. Pour la plupart des variables, la grande majorité des réponses écrites sont codées à cette étape. De plus, on considère que le niveau d'exactitude de cette étape est très

[1] Andrew Stelmack, 100, promenade Tunney's Pasture, Ontario, Canada, K1A 0T6, andrew.stelmack@statcan.gc.ca

élevé et, de fait, pour la plupart des variables, si le fichier de référence est bien entretenu, elle devrait avoir une exactitude presque parfaite. Ces deux facteurs donnent un niveau de qualité global élevé au processus de codage.

Tout ce qui ne correspond pas au fichier de référence est envoyé au codage interactif. Le codage interactif est une étape de codage manuel où un être humain code chaque réponse l'une après l'autre. Bien que des codeurs expérimentés travaillent pour Statistique Canada dans différents programmes, en raison du volume même des données du recensement et du temps entre les cycles, Statistique Canada doit embaucher des centaines d'employés temporaires qui passent environ 10 mois à terminer le codage de cette étape. De même qu'avec le codage automatique, tous les enregistrements effectués dans le codage interactif passent ensuite à l'étape de correction de code.

La correction de code est une étape continue tout au long de l'étape de production du codage (qui se déroule en fait parallèlement au codage interactif) pendant laquelle des experts en la matière peuvent interroger la base de données pour corriger les réponses qu'ils jugent nécessaire de corriger. Ils peuvent le faire en interrogeant un enregistrement à la fois ou, le plus souvent, en envoyant une requête en bloc quand l'expert en la matière peut, par exemple, trouver toutes les réponses contenant un certain mot et les corriger en bloc en un seul code.

1.3 Passer à l'apprentissage automatique

Comme cela a été indiqué dans la section précédente, l'étape de codage interactif exige le recrutement de nombreux employés temporaires pour plusieurs mois. Ce processus est à la fois très coûteux sur le plan financier et chronophage. Ajoutons que comme ces employés sont temporaires et reçoivent une formation limitée, ils sont plus susceptibles de commettre des erreurs que les codeurs expérimentés de Statistique Canada. Selon la variable, on a estimé que le codage interactif a un taux d'erreur d'environ 5 % pour une variable facile à coder, comme le lieu de naissance, qui nécessite une réponse à un seul mot (ou très peu de mots), à environ 30 % d'erreur pour une variable comme l'industrie, pour laquelle le répondant donne une longue description de sa profession. Dans le but d'automatiser le plus possible le processus, on a déterminé que l'apprentissage automatique pourrait offrir une solution de rechange moins coûteuse et plus rapide que le codage interactif.

Le principal coût associé à l'apprentissage automatique provient de l'élaboration des modèles, qui a été réalisée par des employés déjà en poste. S'il est vrai que la mobilisation des services de ces employés à cette fin représente un coût, les modèles d'apprentissage automatique sont un investissement à long terme comparativement au caractère temporaire de l'embauche de codeurs interactifs à chaque cycle. Pour ce qui est de l'efficacité en matière de temps, l'apprentissage automatique se produit presque instantanément pendant la production : dès qu'une réponse est introduite dans le système de traitement, l'apprentissage automatique est en mesure d'attribuer un code qui peut ensuite être analysé au moment de la correction de code. Dans le codage traditionnel, les experts en la matière auraient à attendre que les enregistrements soient traités en codage interactif avant de pouvoir commencer à les examiner. Le temps supplémentaire qui peut ainsi être consacré à la correction de code devrait améliorer dans l'ensemble la qualité des données. En tenant compte de ces facteurs, à savoir le temps, le coût et la qualité, il a été décidé que pour 2021, l'apprentissage automatique remplacerait la totalité ou une partie de l'étape de codage interactif pour la plupart des variables (l'étape de codage automatique serait conservée).

1.4 fastText

Après une première période d'étude des possibilités, il a été déterminé que l'algorithme « fastText » serait l'algorithme sélectionné pour le Recensement de 2021. FastText est un algorithme de traitement du langage naturel développé par Facebook dans les 10 dernières années. En termes simples, la classification utilisant fastText utilise un réseau neuronal pour transformer une chaîne d'entrée en « plongement de mots », c'est-à-dire une représentation vectorielle numérique de la chaîne qui peut ensuite être transformée en probabilités de classe au moyen d'une régression softmax. L'algorithme est exposé de façon plus complète dans Joulin, 2016. FastText fonctionne à l'aide de la ligne de commande et son utilisation est officiellement prise en charge en Python. Aux fins de son utilisation à Statistique Canada, une interface utilisateur graphique (IUG) pour fastText a été créée à l'intérieur du système interne de codage généralisé de Statistique Canada : G-Code. Cette IUG est essentiellement un enveloppeur qui fournit à l'utilisateur un outil permettant de sélectionner les hyperparamètres dans un processus automatisé beaucoup plus simple qu'une création à partir de rien au moyen de la ligne de commande. Malheureusement, les limitations de l'interface ne lui permettent pas de s'adapter à une situation qui différerait d'un flux de production d'apprentissage automatique

typique. Comme nous le verrons dans une section ultérieure, en raison des nombreuses particularités du projet, l'interface n'était pas appropriée. C'est pourquoi il était préférable d'utiliser fastText à partir d'un appel de ligne de commande dans l'environnement R.

2. Élaboration du modèle

2.1 Préparation des données

Comme nous l'avons dit plus haut, ce projet comportait de nombreux défis qui nécessitaient une attention particulière, notamment le fait que deux variables n'étaient pas nécessairement traitées de la même façon. Cela dit, dans une situation idéale, le processus de préparation des données se déroulerait comme suit. Idéalement, les modèles seraient construits à partir des données du cycle de recensement précédent, c'est-à-dire les données du Recensement de 2016. Ces données seraient divisées en ensembles d'entraînement et de test selon un ratio d'environ 75 % d'entraînement et 25 % de test. La validation croisée k-fois serait alors effectuée sur la portion d'entraînement des données pour sélectionner l'ensemble optimal d'hyperparamètres (fastText en ayant beaucoup).

Un premier problème nous a éloignés de cet idéal : les données de 2016 n'étaient pas toutes codées au moyen du même processus. Certaines données ont été codées automatiquement et d'autres par des codeurs interactifs. Comme cela a été mentionné précédemment, l'apprentissage automatique visait à remplacer uniquement l'étape du codage interactif et non l'étape du codage automatique. Par conséquent, le fait d'avoir des enregistrements codés automatiquement dans les blocs (sous-ensembles) de test (ou le bloc de validation de la validation croisée) donnerait une évaluation incorrecte du modèle, puisque dans un environnement de production, le modèle ne coderait jamais ces enregistrements. Étant donné que les enregistrements codés automatiquement sont habituellement plus faciles à coder que leurs équivalents codés interactivement; les inclure dans la portion de test ou de validation donnerait une vision biaisée optimiste des performances du modèle. Cela dit, les enregistrements codés automatiquement sont tout de même précieux comme données d'entraînement, car il est utile pour le modèle d'apprendre les orthographes parfaites et les fautes d'orthographe courantes. La répartition appropriée serait donc la suivante : tous les enregistrements codés automatiquement font partie de l'ensemble d'entraînement, tandis que les enregistrements codés interactivement sont divisés en entraînement et test à un ratio de 75/25. Cependant, cela rend impossible la sélection des hyperparamètres dans l'enveloppeur G-Code. G-Code reçoit simplement un ensemble de données d'entraînement et le divise aléatoirement en k nouveaux ensembles de données avec une portion aux fins d'entraînement et une portion aux fins de validation. Cette nouvelle subdivision ne peut pas tenir compte de ce qui est codé automatiquement par rapport à ce qui est codé de façon interactive et, par conséquent, certains enregistrements codés automatiquement se retrouveraient inévitablement dans l'ensemble de validation, ce qui pourrait mener à la sélection d'un ensemble d'hyperparamètres sous-optimaux.

Comme cela a été mentionné précédemment, cette difficulté a été atténuée par l'utilisation d'un script R écrit à l'interne qui permettait de diviser les données de façon appropriée. Concernant les variables pour lesquelles nous disposons seulement d'une petite quantité de données, une procédure de validation croisée emboîtée a été utilisée avec une boucle interne à 5 blocs et une boucle externe à 5 blocs. Pour ces petits ensembles de données, cela a permis d'évaluer le modèle selon le plus grand ensemble possible. Pour les ensembles de données plus volumineux, nous avons utilisé une division semblable à celle de la situation idéale, et pour les variables ayant un très grand nombre d'enregistrements (de l'ordre de quelques millions), nous avons utilisé une seule division des données entre entraînement et test de 75/25, suivie d'une deuxième division de l'ensemble d'entraînement de 75/25 pour former l'ensemble de validation plutôt que de recourir à une validation croisée, car la puissance de calcul pose problème avec cette quantité de données.

2.2 Modification de l'ensemble de codes

Une autre complication provenait du fait que les données de 2016 qui devaient être utilisées comme données d'entraînement ne convenaient pas telles quelles à cet usage. Pour bon nombre des variables qui devaient être codées, l'ensemble de codes (la liste des codes qu'il est possible d'attribuer) a été considérablement modifié par rapport au cycle précédent. Différentes raisons l'expliquent, d'un changement de portée de la variable à un changement des

normes de l'industrie. Cela pose évidemment problème, car le modèle apprend seulement à partir des données d'entraînement, et il est illogique d'utiliser des données d'entraînement avec un ensemble de codes de 2016 tout en attendant des résultats qui utilisent un ensemble de codes de 2021. Par conséquent, il faut « corriger » les données de 2016 pour qu'elles contiennent les valeurs du nouvel ensemble de codes de 2021. Le cas le plus simple pourrait être un changement d'étiquette, par exemple le code d'« enseignant » est passé de 1000 à 1001. Il suffit d'utiliser un fichier de concordance qui remplace les codes de 2016 par leur version de 2021. Un cas plus complexe, mais tout aussi facile à corriger, serait un cas où une réponse écrite devient plus générale, par exemple si « enseignant au primaire » – 1001 et « enseignant au secondaire » – 1002 en 2016 deviennent tous deux « enseignant » – 1000 en 2021. Encore une fois, cela peut être corrigé au moyen d'un fichier de concordance. Le cas le plus difficile à corriger est le cas où une étiquette devient plus granulaire, par exemple si « enseignant » voit son code changer de 1000 à « enseignant au primaire » – 1001 ou « enseignant au secondaire » – 1002. Ce « fractionnement » de code ne peut pas être corrigé au moyen d'un simple fichier de concordance et, essentiellement, ne peut être corrigé que par l'intervention manuelle d'experts en la matière.

La première étape pour recoder ces enregistrements selon l'ensemble de codes de 2021 a consisté à soumettre les réponses écrites de 2016 au processus de codage automatique de 2021. Le code de 2021 approprié était attribué à tous les enregistrements qui avaient pu être codés automatiquement. Malheureusement, il restait ceux qui avaient échoué au codage automatique avec les codes de 2016. Les experts en la matière avaient alors deux solutions : soit recoder ces enregistrements selon le code de 2021 au moyen d'un fichier de concordance quand c'était possible ou d'une intervention manuelle, soit créer un modèle en utilisant seulement ces enregistrements codés automatiquement. La première solution n'était pas réalisable dans tous les cas, car dans le cas des codes où la concordance était d'un à plusieurs, il aurait fallu entièrement réeffectuer le processus de codage. Cependant, la deuxième solution produit nécessairement un modèle moins bon que la première, car non seulement il n'a pas les données de la première solution, mais il n'a pas non plus les données exactes que nous essayons de modéliser. Autrement dit, le modèle n'apprendrait qu'à partir de l'orthographe parfaite ou presque parfaite des mots, mais il tenterait de coder le contraire par la suite. Cette méthode présente également l'inconvénient de potentiellement sélectionner un ensemble d'hyperparamètres sous-optimal. Étant donné que le modèle est entraîné uniquement sur les enregistrements codés automatiquement, la recherche d'hyperparamètres est également validée uniquement au moyen des enregistrements codés automatiquement. L'ensemble d'hyperparamètres qui prédit le mieux les mots parfaitement orthographiés n'est pas nécessairement l'ensemble qui prédit le mieux les mots incorrectement orthographiés.

Mis à part ces problèmes, pour la plupart des variables, les experts en la matière ont choisi la méthode la plus simple consistant à entraîner le modèle exclusivement à partir des enregistrements codés automatiquement. Après évaluation de tous les modèles sur un échantillon de réponses écrites codées à la main pour 2016 qui avaient échoué au codage automatique, leur qualité a été jugée adéquate.

2.3 Nouvelles variables pour 2021

Un autre scénario dans lequel les données de 2016 ne convenaient pas à l'entraînement était l'absence de données concernant une variable, quand celle-ci avait été nouvellement ajoutée pour le cycle de 2021. Ce cas s'est présenté pour quelques-unes des variables nécessitant des modèles en 2021. La première solution à ce problème consisterait à entraîner le modèle à l'aide du fichier de référence (le fichier des réponses courantes et attendues utilisé aux fins du codage automatique). Il est certain que le codage au moyen du fichier de référence serait une option viable. Toutefois, comme ces questions n'étaient pas posées auparavant, il aurait peut-être été difficile pour les experts en la matière d'envisager la liste de toutes les réponses attendues possibles. Ainsi, un modèle entraîné uniquement sur le fichier de référence n'aurait peut-être pas vu toute la population de réponses possibles.

Une autre solution, que nous avons utilisée en complément du fichier de référence, consistait à utiliser les données d'une autre enquête comme données d'entraînement. Par exemple, la question sur le genre avait été nouvellement ajoutée pour le Recensement de 2021, mais elle avait été posée dans d'autres enquêtes fournies par Statistique Canada. Comme dans le cas d'un changement d'ensemble de codes, nous avons dû traiter les réponses écrites de ces enquêtes au moyen de notre procédure de codage automatique de 2021 pour obtenir un code valide dans notre ensemble de codes. Tous les enregistrements ayant échoué au codage automatique ont alors dû être codés manuellement par des experts en la matière ou, comme dans le cas d'un changement d'ensemble de codes, ont dû utiliser un modèle fondé uniquement sur les enregistrements codés automatiquement. Chaque fois que cette situation s'est produite, les experts

en la matière ont choisi de coder manuellement les enregistrements restants. Cela était plus réalisable que dans le cas d'un changement type d'ensemble de codes, puisque la quantité d'enregistrements obtenus d'autres enquêtes (~ 1 000 à 30 000 enregistrements) était d'un ordre de grandeur inférieur à la quantité tirée des données du recensement de 2016.

3. Évaluation

3.1 Taux d'appariement et exactitude

Dans un flux de production d'apprentissage automatique typique, les enregistrements dans les ensembles de test et de validation ont la valeur attribuée par le modèle comparativement à la vraie valeur afin d'arriver à une mesure d'évaluation. Dans le cadre du processus de codage interactif, on constitue un échantillon d'assurance de la qualité afin d'estimer la qualité des codes attribués par les codeurs interactifs. Pour la plupart des variables, l'exactitude estimée se situe entre 80 et 90 %, mais pour quelques variables compliquées comportant des réponses écrites plus longues (industrie et profession), l'exactitude estimée chute à environ 65 à 75 %. En raison du niveau d'inexactitude des codes dans les données de 2016, une comparaison entre le code attribué par le modèle et le code attribué de façon interactive de 2016 (ci-après appelée « taux d'appariement ») conduit à une mauvaise évaluation du modèle, car il n'y a aucun moyen de déterminer si la non-correspondance du modèle avec le code attribué est due à une erreur du modèle ou à une erreur du codeur interactif.

Prenons un exemple où l'exactitude et le taux d'appariement sont relativement élevés, soit une exactitude du codeur interactif estimée à 90 % et un taux d'appariement de 85 %. L'exactitude réelle de ce modèle pourrait se situer entre 75 % et 95 %, selon que le modèle correspond au codeur interactif quand le code du codeur interactif était exact ou quand il était inexact. Cette fourchette large (qui augmente à mesure que l'exactitude estimée du codeur interactif et le taux d'appariement diminuent) rend cette méthode inefficace aux fins d'évaluation des modèles ou, à tout le moins, exige que des mesures secondaires soient utilisées simultanément. C'est pourquoi il fallait étudier d'autres moyens d'évaluer nos modèles.

3.2 Enregistrements de correction de code

L'un de ces moyens consistait à prendre en compte les enregistrements qui avaient été réalisés auparavant à l'aide de corrections de code à des fins d'évaluation. Il s'agit d'enregistrements qui auraient été réalisés par codage interactif (et qui existeraient donc dans les ensembles de test et de validation), mais qui auraient été examinés par un expert en la matière pendant la correction de code. On peut supposer que ces réponses comportent moins d'erreurs que celles provenant uniquement d'un codage interactif, puisqu'elles ont été effectuées par un expert en la matière.

Toutefois, comme nous l'avons dit plus haut, la correction des codes des enregistrements est souvent réalisée en bloc et ils ne sont donc pas exempts d'erreur. En raison de contraintes de temps, un expert en la matière peut interroger plusieurs milliers d'enregistrements qui contiennent tous une réponse écrite semblable, contenant potentiellement tous un mot en commun, et les changer en bloc pour le même code, en sachant qu'il se peut que ce code ne doive pas être attribué à certains des enregistrements. Comme nous l'avons vu dans la section précédente, une petite erreur dans les étiquettes suffit à entraîner une grande variance dans les estimations de l'exactitude du modèle, ce qui peut poser un problème avec cette méthode. De plus, les enregistrements qui ont été évalués dans le cadre d'une correction de code ne sont pas représentatifs de la population. Au-delà de l'examen en bloc, certains enregistrements sont également marqués comme étant « à transmettre à un expert en la matière » parce que le codeur interactif ne sait pas quel code leur attribuer. Cela signifie que la distribution des enregistrements de correction de code présentera un plus grand nombre de réponses écrites plus difficiles que dans l'ensemble de la population, ce qui donnera une vision pessimiste du rendement de l'apprentissage automatique.

Enfin, il est difficile de comparer les performances du modèle par rapport au codage interactif à l'aide de ces enregistrements, car, par nature, la correction du code de l'enregistrement est susceptible de changer le code par rapport au code attribué de façon interactive. Ainsi, une exactitude du modèle de 80 % parmi les enregistrements de

correction de code comparativement à une exactitude du codage interactif de 5 % parmi ces enregistrements ne reflète pas nécessairement des résultats 16 fois meilleurs de l'apprentissage automatique par rapport au codage interactif.

3.3 Estimation du taux d'erreur des codeurs interactifs

Une autre solution envisagée consistait à utiliser des enregistrements qui faisaient partie du processus d'estimation du taux d'erreur des codeurs interactifs évoqué précédemment. Pendant le codage, un échantillon d'enregistrements est effectué par deux codeurs interactifs. Si le deuxième codeur est d'accord avec le premier (sans connaître le code qui a été attribué par le premier codeur), le code est jugé correct. Sinon, il est envoyé à un arbitre qui a un niveau de connaissance plus élevé qu'un codeur interactif type. Si cet arbitre est d'accord avec le premier codeur, le code est jugé correct et incorrect sinon. Cet échantillon est à la base des taux d'erreur des codeurs interactifs mentionnés tout au long du présent article. Comme l'algorithme avait déjà été déterminé à ce stade, le processus d'évaluation vise principalement à savoir dans quelle mesure l'apprentissage automatique peut ou non offrir une amélioration par rapport à un codeur interactif. Dans cette optique, l'utilisation des enregistrements ayant servi à déterminer le taux d'erreur des codeurs interactifs semblait un choix logique.

À cette fin, le code d'apprentissage automatique était considéré comme le code du « premier codeur ». Si le deuxième codeur était d'accord avec l'apprentissage automatique, l'apprentissage automatique était considéré comme correct. S'ils étaient en désaccord, mais que l'arbitre était d'accord avec l'apprentissage automatique, là encore, l'apprentissage automatique était considéré comme correct. Il est difficile d'utiliser cette méthode dans les cas où l'apprentissage automatique et le premier codeur sont en désaccord, mais que le deuxième codeur est d'accord avec le premier codeur. Dans ce cas, il n'y avait pas de valeur d'arbitre à comparer à l'apprentissage automatique qui aurait servi si l'apprentissage automatique avait vraiment été le premier codeur. Cette question donne une vision pessimiste et biaisée des performances du modèle.

Un autre inconvénient est l'hypothèse selon laquelle en cas d'accord de deux codeurs interactifs, le code est correct. Les codeurs interactifs ont reçu la même formation, codent souvent près les uns des autres et se posent des questions pour avoir l'avis des autres. Ils sont par conséquent susceptibles de commettre les mêmes erreurs. De fait, une petite étude s'est intéressée à cette question : elle examinait un échantillon d'enregistrements pour lequel les deux codeurs interactifs étaient d'accord et les donnait à recoder à une personne plus compétente. Cet « expert en la matière » était d'accord avec les deux codeurs interactifs seulement dans 70 % des cas environ. Étant donné que l'évaluation de l'apprentissage automatique à l'aide de ces enregistrements repose également sur cette hypothèse, l'évaluation pourrait ne pas être exacte. Cette méthode (ainsi que la méthode de correction de code décrite précédemment) est également inutilisable dans le cas des changements d'ensemble de codes pour lesquels la concordance n'a pas été réalisée, car le code des modèles sera attribué au moyen de l'ensemble de codes de 2021 et ne peut être comparé au deuxième codeur dont le code a été attribué avec l'ensemble de codes de 2016.

3.4 Étalon de référence

La dernière méthode d'évaluation envisagée était celle utilisant un « étalon de référence ». Il s'agit d'un échantillon d'enregistrements qui ont d'abord été codés de façon interactive, puis recodés un par un par un expert en la matière. Idéalement, cet échantillon ne comporte pas d'erreur et permettrait ainsi d'atténuer les problèmes décrits plus haut et de comparer le code attribué par le modèle au code attribué de façon interactive. Bien que cette méthode semble idéale en raison de sa simplicité et de sa qualité, elle présente aussi des problèmes. Premièrement, elle peut être très chronophage. Selon la variable, un échantillon de quelques milliers d'enregistrements seulement prendrait de quelques jours à quelques semaines. Il n'est pas facile de trouver un expert en la matière ayant à la fois la capacité de recoder correctement les enregistrements et la possibilité de passer des semaines à le faire.

Deuxièmement, même si les dossiers sont recodés par des experts en la matière, il est possible que la réponse écrite soit quelque peu ambiguë si bien qu'il n'est pas garanti que deux experts différents lui attribuent le même code. Ou plus simplement, l'expert en la matière peut se tromper, ce qui nous ramène au problème initial de comparaison avec un enregistrement erroné codé interactivement. Par conséquent, cet « étalon de référence » peut aussi comporter des erreurs. Cela dit, toutes les solutions citées présentent des inconvénients et, peut-être pour sa simplicité de compréhension, l'étalon de référence a été notre méthode d'évaluation préférée. Il permettait en outre que les enregistrements utilisés aux fins de l'étalon de référence soient ajoutés aux données d'entraînement après la fin de

l'évaluation du modèle dans les modèles qui ont été entraînés seulement au moyen d'enregistrements codés automatiquement.

4. Évaluation préproduction

Bien que des étalons de référence aient été créés pour toutes les variables, leur capacité de fournir de l'information utile était limitée selon qu'on disposait d'une concordance de code interactif de 2016 à 2021, comme nous l'avons vu plus haut. Il est possible de calculer une mesure de l'exactitude au moyen de l'étalon de référence pour toutes les variables. Toutefois, en l'absence de concordance entre les codes de 2016 et de 2021, la seule mesure à laquelle cette exactitude peut être comparée est l'exactitude estimée des codeurs interactifs de 2016.

Comme cela a été mentionné précédemment, cette estimation de l'exactitude repose sur l'hypothèse selon laquelle si deux codeurs interactifs arrivent au même code, le code est exact, ce qui rend cette estimation non fiable. De plus, l'estimation de l'exactitude du codage interactif de 2016 est fondée sur un processus qui utilise l'ensemble de codes de 2016. Pour les mêmes raisons qu'une concordance entre 2016 et 2021 n'était pas possible, une comparaison de l'exactitude du codage interactif de 2016 et de l'exactitude de l'étalon de référence de 2021 pourrait ne pas être valide. En effet, le changement d'ensemble de codes peut être dû à un changement de portée de la question, ce qui signifie que les deux ne sont plus comparables. Le plus souvent, quand on évaluait la pertinence du modèle d'apprentissage automatique avant de passer à la production, on présentait aux experts en la matière les résultats de l'étalon de référence et on leur demandait de déterminer la pertinence des apprentissages automatiques indépendamment des estimations antérieures de l'exactitude des codeurs interactifs. Pour toutes les variables sauf une, les experts en la matière étaient suffisamment satisfaits du niveau d'exactitude des modèles d'apprentissage automatique pour lancer la production.

Concernant les variables pour lesquelles l'ensemble de codes a peu changé – c'est-à-dire que les experts en la matière ont pu fournir une concordance entre les codes de 2016 et de 2021 – l'étalon de référence a permis d'ajouter une comparaison directe entre l'apprentissage automatique et le codage interactif. Pour l'une de ces variables, « pays » (un modèle composé de quelques questions dont la réponse était un pays) parmi les 2 910 réponses écrites uniques fournies à l'étalon de référence, l'apprentissage automatique donnait un résultat correct 84,33 % du temps, tandis que le codage interactif donnait un résultat correct à un taux légèrement plus élevé de 84,58 %. Bien que le codage interactif ait donné un résultat légèrement supérieur à celui de l'apprentissage automatique pour les réponses écrites concernant cette variable, les économies de temps et de coûts réalisées par le modèle d'apprentissage automatique permettaient aux experts en la matière de faire plus que combler la différence dans la correction de code.

5. Conclusion et perspectives

L'apprentissage automatique par l'algorithme fastText a été mis en œuvre avec succès à l'étape du traitement du codage du Recensement canadien de 2021 pour toutes les variables pour lesquelles il a été examiné, sauf une. À l'heure actuelle, à l'étape de la production, nous sommes en train de terminer un test d'assurance de la qualité qui permettra d'obtenir un « étalon de référence » des réponses codées par l'apprentissage automatique recodées par des experts en la matière. Après le cycle de 2021, les recherches viseront à améliorer le processus pour le prochain cycle du recensement en 2026. Elles s'intéresseront notamment à la possibilité de faire coder par l'apprentissage automatique des enregistrements déjà codés automatiquement, d'autres mesures d'évaluation, ainsi que des techniques d'apprentissage actif pour les cas où la concordance entre les ensembles de codes est impossible. Selon nous, l'apprentissage automatique jouera un rôle plus important dans le processus de codage dans les prochains cycles du recensement.

Bibliographie

Joulin, A., E. Grave, P. Bojanowski et T. Mikolov (2016), « Bag of Tricks for Efficient Text Classification », *arXiv Preprint arXiv:1607.01759v3*.