Recueil du Symposium de 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Les défis de la production d'estimations nationales de la maltraitance des enfants à l'aide de données administratives provenant de différents secteurs de compétence

David Laferrière et Catherine Deshaies-Moreault[1]

Résumé

Il a été demandé à Statistique Canada de réaliser une étude de faisabilité sur la façon d'élaborer un système de surveillance de la maltraitance infantile. Le Système canadien de surveillance des signalements d'enfants victimes de maltraitance (SCSSEVM) intégrerait les données des organismes de protection de l'enfance de chaque province et territoire pour calculer des estimations annuelles de la maltraitance des enfants dans cinq catégories : la violence physique, la violence psychologique, l'abus sexuel, la négligence et l'exposition à la violence conjugale. Afin de réduire le fardeau des travailleurs des services de protection de l'enfance, la principale source de données serait un recensement des données administratives. Nous discutons des défis à surmonter en vue de mettre en œuvre le SCSSEVM, y compris du fait que chaque secteur de compétence possède ses propres lois qui définissent et catégorisent la maltraitance des enfants, ainsi que des systèmes différents de détermination et de suivi des cas de maltraitance, ce qui entraîne une variation considérable du contenu et de la structure de leurs données administratives et des textes narratifs connexes.

Toutefois, pour le SCSSEVM, des techniques d'apprentissage automatique provenant du traitement du langage naturel seront examinées afin de déterminer si les cas de mauvais traitements pourraient être automatiquement repérés et classés à partir des rapports narratifs et des ensembles de données existants. Nous discutons des défis pratiques et techniques liés à l'utilisation des approches traditionnelles ainsi que des techniques plus modernes pour créer des estimations nationales cohérentes à partir des données administratives de 13 secteurs de compétence infranationaux.

Mots-clés : Maltraitance des enfants; apprentissage automatique; données administratives.

1. Introduction

1.1 Contexte

Les gouvernements, les organismes non gouvernementaux et les médecins, entre autres, ont en commun un intérêt à avoir des données exactes et actuelles afin de comprendre la nature (type de maltraitance) et l'étendue (comptes et caractéristiques démographiques) de la maltraitance des enfants (ME) dans les régions géographiques qu'ils desservent (Potter, Hovdestad, & Tonmyr, 2013). La maltraitance des enfants fait référence à la violence physique ou psychologique, à l'abus sexuel, à la négligence ou à l'exposition à la violence conjugale d'une personne âgée de moins de 18 ans.

Malgré que l'on en reconnaisse la nécessité, les sources de données sur la maltraitance des enfants disponibles au Canada sont très limitées. De plus, chaque province et chaque territoire est responsable de l'adoption et de l'application de leurs propres lois sur la protection de l'enfance, ce qui exacerbe les problèmes de collecte à l'échelle du Canada de données sur la maltraitance des enfants et de l'utilisation de ces données pour produire des estimations.

[1]David Laferrière, Statistique Canada, 100 promenade Tunney's Pasture, Canada, K1A 0T6 (david.laferriere2@canada.ca); Catherine Deshaies-Moreault, Statistique Canada, 100 promenade Tunney's Pasture, Canada, K1A 0T6 (catherine.deshaies-moreault@canada.ca)

1.2 Statistiques sur la maltraitance des enfants actuellement disponibles[2]

Bon nombre de provinces et de territoires publient des rapports annuels sur la maltraitance des enfants, mais les données contenues dans ces rapports sont habituellement sous la forme de tableaux agrégés avec peu de granularité. Par exemple, il n'y a souvent aucune ventilation des cas de maltraitance des enfants par type (violence physique, négligence, etc.) ou par groupe d'âge.

L'Étude canadienne sur l'incidence des signalements de cas de violence et de négligence envers les enfants (ECI) est une enquête périodique[3] qui vise à fournir un profil national des enfants et des familles qui reçoivent des services d'aide à l'enfance. Elle porte sur un échantillon de travailleurs des services de protection de l'enfance dans chaque province et territoire qui remplissent un questionnaire pour chacun des cas dont ils sont responsables et qui ont été nouvellement ouverts au cours d'une période de référence donnée. Parmi ses principaux objectifs, l'ECI a été conçue de manière à « déterminer le taux des cas de violence physique, d'abus sexuel, de négligence, de violence psychologique et d'exposition à la violence conjugale corroborés et ayant fait l'objet d'une enquête ainsi que les multiples formes de maltraitance » (Agence de la santé publique du Canada, 2010). L'ECI est une source précieuse de données sur la maltraitance des enfants, mais sa nature périodique limite considérablement son actualité.

Il existe d'autres sources de données sur la maltraitance des enfants au Canada, mais, à l'heure actuelle, elles forment une mosaïque quelque peu incomplète (Potter, Hovdestad, & Tonmyr, 2013) qui ne permet pas facilement aux chercheurs d'estimer des statistiques actuelles à l'échelle du pays.

1.3 Étude de faisabilité

En 2017, l'Agence de la santé publique du Canada (ASPC) a approché Statistique Canada afin de mener une étude de faisabilité sur la mise au point d'un système de surveillance[4] des rapports de maltraitance des enfants. Ce système s'appellerait le Système canadien de surveillance des signalements d'enfants victimes de maltraitance (SCSSEVM). Son principal objectif consisterait à intégrer les données des organismes de protection de l'enfance de chaque province et territoire pour calculer les comptes annuels du nombre d'enfants faisant l'objet d'une enquête en lien avec la maltraitance et le nombre total d'enquêtes dans cinq catégories : violence physique, violence psychologique, abus sexuel, négligence et exposition à la violence conjugale (EVC). Ces chiffres incorporeraient les caractéristiques démographiques des familles mentionnées dans les données recueillies.

Une composante principale de l'étude de faisabilité comportait une série de consultations, lancées en 2018, entre Statistique Canada, l'ASPC et les ministères provinciaux et territoriaux responsables de la protection de l'enfance, dans le but de terminer les consultations avec chaque secteur de compétence d'ici le début de 2019. Un des principaux objectifs de ces discussions visait à déterminer de quelle manière chaque secteur de compétence procédait à la saisie et au stockage des données se rapportant à leurs rapports et enquêtes en lien avec l'aide à l'enfance. Un autre objectif consistait à déterminer dans quelle mesure chaque secteur de compétence était disposé à participer au SCSSEVM selon divers scénarios de collecte de données. Ces scénarios comprenaient une enquête auprès des travailleurs des services de protection de l'enfance et un recensement administratif.

À la suite de consultations et d'un examen des sources de données existantes, Statistique Canada a préparé une ébauche de l'étude de faisabilité dans laquelle il comparait la viabilité et l'efficacité des deux options pour la collecte de données sur la maltraitance des enfants et les a comparées : une enquête auprès des travailleurs des services de protection de l'enfance et un recensement administratif. En raison du fardeau sur les travailleurs des services de protection de l'enfance et des coûts associés à une enquête, l'étude de faisabilité a recommandé la tenue d'un recensement au moyen des données sur la maltraitance des enfants recueillies par les organismes de protection de l'enfance dans chaque province et territoire. Il y a cependant plusieurs défis liés à la collecte et à la combinaison des données administratives des 13 secteurs de compétence du Canada.

---

[2] Pour un rapport plus complet sur les sources de données sur la maltraitance des enfants disponibles au Canada, voir Potter, Hovdestad et Tonmyr, 2013.

[3] L'ECI a été menée en 1998, en 2003 et en 2008.

[4] Les termes « système de surveillance » sont utilisés ici dans un sens épidémiologique et signifient la collecte, l'analyse et l'interprétation de l'action pour les données sur la santé (Potter, Hovdestad, & Tonmyr, 2013).

2. Défis liés aux données administratives

Les données administratives représentent une excellente opportunité pour les organismes nationaux de statistique (ONS). En effet, l'utilisation de données administratives dans les statistiques officielles diminue le fardeau de réponse, peut offrir des données plus actuelles et peut améliorer la qualité en réduisant ou en éradiquant les erreurs d'échantillonnage. Par contre, elles comportent leurs propres défis. Par définition, ces données ne sont pas recueillies à des fins statistiques et la portée ou la définition de l'information recueillie pourrait ne pas être tout à fait en accord avec les objectifs de surveillance pour lesquels on examine l'utilisation de ces données administratives. La population cible des données administratives et celle de l'enquête pourraient ne pas être exactement la même non plus. De plus, certains prétraitements des données pourraient être effectués par l'organisme qui recueille des renseignements sans que l'ONS en soit informé. En plus de ces défis habituels avec les données administratives, le SCSSEVM a ses propres défis : en particulier, les données d'intérêt sont régies par des lois infranationales et se trouvent dans un nombre inconnu de systèmes de données.

2.1 Lois infranationales

Chaque province et chaque territoire est responsable de ses propres lois sur la protection de l'enfance, et les catégories de mauvais traitements définies dans une province donnée peuvent ne pas correspondre directement à celles du SCSSEVM. Nous devrons également collaborer avec des partenaires provinciaux et territoriaux afin de rester au courant de tous les changements apportés aux lois sur la protection de l'enfance de chaque province et territoire, étant donné que ces changements peuvent avoir une incidence sur les données saisies par un secteur de compétence donné.

2.2 Systèmes de données

Les 13 secteurs de compétence infranationaux (10 provinces et 3 territoires) du Canada possèdent différents outils d'évaluation de la maltraitance des enfants et systèmes de données, et même à l'intérieur d'un secteur de compétence donné, il peut y avoir des incohérences dans l'utilisation des outils et des systèmes. Ainsi, même les données combinées pour une province ou un territoire donné peuvent être hétérogènes. La qualité de l'information saisie (dans le contexte de la production de statistiques officielles) peut varier entre secteurs de compétence et à l'intérieur de ceux-ci. Par conséquent, nous nous attendons à ce que des efforts importants soient nécessaires pour harmoniser les données et dériver des chiffres nationaux.

Les systèmes informatiques ne sont pas statiques, les outils d'évaluation de la protection de l'enfance non plus, car ils évoluent au fil du temps en fonction des modifications à la législation ou des améliorations aux méthodes d'évaluation. Par conséquent, le projet du SCSSEVM nécessitera la surveillance des changements qui peuvent avoir une incidence sur la manière de saisir ou de soumettre les données, et ses systèmes et ses processus devront être mis à jour en conséquence.

Enfin, nous devons tenir compte de la manière dont les deux unités d'intérêt du SCSSEVM (l'enfant et l'enquête) sont saisies dans les bases de données. Par exemple, les secteurs de compétence peuvent recueillir des données au niveau de l'enfant ou de la famille. D'un système à l'autre, la saisie de cas où le même enfant fait l'objet de plusieurs enquêtes variera vraisemblablement. Nous devrons évaluer attentivement chaque système afin de nous assurer que nous ne faisons pas un surdénombrement des enfants ou des enquêtes.

3. Codage des textes narratifs

En règle générale, les travailleurs des services de protection de l'enfance saisissent les données de leur cas au moyen d'un formulaire électronique comportant une combinaison de menus déroulants, de cases à cocher et de champs de texte. Ces champs de texte comprennent habituellement de l'espace pour des textes narratifs longs donnant les détails d'un cas. Comme les renseignements nécessaires pour saisir les cinq catégories de maltraitance d'enfants peuvent ne

pas être tous disponibles dans les variables du menu déroulant, nous allons probablement devoir utiliser l'information contenue dans les textes narratifs qui accompagnent chaque cas pour nous assurer de ne pas sous-estimer les cas signalés de maltraitance. Par exemple, nous devrons presque certainement utiliser les textes narratifs pour remplir la catégorie EVC dans certains secteurs de compétence. La détermination des variables nécessaires à partir de textes narratifs non structurés exigera un codage, c'est-à-dire qu'à partir de chaque texte narratif, il faudra saisir des données pour chacune des cinq catégories de maltraitance. Des codeurs humains, des algorithmes d'autocodage ou une combinaison des deux peuvent être utilisés pour obtenir les données nécessaires des textes. Il est pertinent à ce moment-ci de discuter des aspects suivants de l'exercice de codage : l'obtention des données de référence, les codeurs humains et l'apprentissage automatique.

3.1 Données de référence

Que l'on ait recourt au codage automatique, humain ou à une combinaison des deux pour coder les textes narratifs, des données étiquetées de référence, ou données de référence, seront nécessaires pour former les codeurs et/ou les algorithmes. Idéalement, les données étiquetées seraient codées par les personnes qui connaissent le mieux les données et les systèmes – les travailleurs des services de protection de l'enfance. Toutefois, avant qu'ils ne codent les données de référence, ils devraient également recevoir une formation sur les cinq catégories de maltraitance (et les variables connexes) afin de s'assurer que les renseignements sont appariés uniformément dans l'ensemble du pays.

Des discussions sont en cours entre Statistique Canada et l'Agence de la santé publique du Canada afin de déterminer l'approche idéale pour obtenir des données de référence. Deux approches sont envisagées : chaque secteur de compétence offre des exemples pratiques dans les cinq catégories, ou un groupe centralisé d'experts sur les catégories nationales code les données provenant de chaque secteur de compétence. Un défi de la première approche réside dans le fait qu'il est peu probable que les travailleurs des services de protection de l'enfance aient du temps à consacrer pour coder des données. Il serait donc très difficile d'obtenir des données de référence de tous les secteurs de compétence codées par des travailleurs des services de protection de l'enfance actuellement employés. Il pourrait toutefois être possible d'embaucher d'anciens travailleurs des services de protection de l'enfance, surtout ceux qui ont récemment pris leur retraite, pour effectuer le codage. La deuxième approche comporte également des défis, dont la plus évidente consiste à trouver et à former des gens pour qu'ils deviennent des experts des 5 catégories de maltraitance des enfants et des lois régissant les 13 secteurs de compétence.

Enfin, il sera important de garder à l'esprit que, même si les personnes qui créent des données de référence sont très bien informées, des erreurs humaines sont possibles, ce qui fait qu'il est possible que les données de référence contiennent des erreurs.

3.2 Codeurs humains

Statistique Canada a un département qui se consacre au codage, y compris le codage effectué par des humains. Le processus habituel de codage humain est le suivant : tout d'abord, des spécialistes possédant une expertise en codage des données élaborent le matériel de formation. Ensuite, à l'aide du matériel de formation et d'exemples (provenant potentiellement des données de référence), les codeurs apprennent la stratégie de codage. Puis, ils codent des données pour lesquelles les résultats sont déjà connus, et leurs taux de réussite et d'uniformité pour le codage de ces données de référence sont évalués. Enfin, la formation et l'information fournies aux codeurs sont ajustées au besoin, et le processus est répété, jusqu'à ce que l'on atteigne un certain niveau de satisfaction à l'égard du codage.

L'élaboration du matériel de formation, la formation des codeurs, leur évaluation et leur emploi au codage des données exigent beaucoup de temps et sont très coûteux. Il y a aussi des sources d'erreurs qui sont difficiles à atténuer, et cela est particulièrement vrai lorsque les concepts devant être codés sont très complexes (comme c'est le cas pour le SCSSEVM). Il est particulièrement difficile de s'assurer que les codeurs aient une approche uniforme entre eux, c'est-à-dire que deux codeurs liront le même texte narratif et le coderont de manière identique.

En 2016, l'ASPC a mené une étude pour examiner dans quelle mesure les codeurs humains seraient efficaces pour coder des données administratives sur la maltraitance des enfants à partir de données administratives et de textes narratifs (Tonmyr, et al., 2018). Dans leur étude, 12 travailleurs des services de protection de l'enfance (TSP) d'une province ont reçu une formation portant sur la définition de maltraitance des enfants selon l'ASPC et ont fourni des

données fondées sur leurs enquêtes sur la maltraitance d'enfant concernant 187 enfants. Ces données ont été utilisées comme données de référence. Deux codeurs ont ensuite reçu une formation sur les cinq catégories de maltraitance, et, par la suite, ils ont codé les mêmes renseignements que les TSP avaient soumis au système de données administratives, sans avoir accès aux connaissances directes des travailleurs sur le cas. Les résultats des codeurs ont alors été évalués en les comparant aux données de référence. Les résultats étaient prometteurs : la classification de la violence physique, de l'abus sexuel et de la négligence par les deux codeurs correspondaient largement à celles des TSP. Toutefois, les codeurs avaient de la difficulté à déterminer de manière uniforme et exacte l'abus sexuel, et le codage de l'exposition à la violence conjugale n'a pu être évalué.

Deux des défis liés aux codeurs humains, à savoir le coût et l'incohérence ou inexactitude, peuvent être atténués quelque peu en utilisant des techniques de codage automatique. Il est peu probable que l'on puisse éviter complètement le recours aux codeurs humains, mais le codage automatique à l'aide de techniques d'apprentissage automatique pourrait au moins réduire le besoin en codage manuel.

3.3 Codage automatique

Le codage automatique fait référence au codage effectué au moyen d'algorithmes qui utilisent un processus automatisé, par opposition à un codage manuel qui exige la prise de décisions par des humains. Le codage automatique comporte plusieurs avantages : il est habituellement plus rapide, moins coûteux et plus uniforme au fil du temps. En outre, le codage automatique permet d'exploiter différents types de renseignements. En particulier, les textes narratifs et les variables catégorielles peuvent être utilisés par un algorithme de codage, ce qui serait une tâche exigeant beaucoup de temps pour un codeur humain.

La plupart des algorithmes de codage automatique (aussi appelé autocodeurs) actuellement utilisés par Statistique Canada sont fondés sur des règles, c'est-à-dire qu'ils utilisent une forme quelconque d'arbre de décision ou un ensemble de règles préétablies pour choisir les codes résultants. Toutefois, en raison de la complexité des concepts de maltraitance des enfants et du fait que tout texte qui accompagne un enregistrement que nous recevons sera d'une longueur et d'une complexité substantielle, il est presque certain qu'un autocodeur fondé sur des règles ne sera pas suffisant pour le SCSSEVM. À la place, nous visons à utiliser des algorithmes d'apprentissage automatique pour coder au moins certains des enregistrements.

3.3.1 Apprentissage automatique pour l'identification d'événements/la classification de texte

D'importants travaux ont déjà été effectués à l'aide de l'apprentissage automatique afin d'interpréter des textes narratifs dans diverses applications, comme le codage des causes de décès à partir de rapports d'autopsies verbaux (Danso, Atwell, & Johnson, 2014), l'analyse de sentiments de critiques de film (Zhang, Marshall, & Wallace, 2016) et le codage des données sur les accidents du travail provenant des demandes d'indemnisation des travailleurs (Measure, 2014). Les principales techniques utilisées dans ces applications ont un fondement commun : les méthodes sont basées sur la conversion du texte en un vecteur numérique où chaque rangée du vecteur représente un mot.
De nombreuses techniques de traitement du langage naturel peuvent contribuer à ce processus de vectorisation, comme le retrait des mots vides (mots courants qui n'ajoutent pas de signification importante) et la radicalisation (représentant des mots ayant une base commune comme « blessé » et « blessure » au moyen d'un seul mot radical « blesser »). Ces techniques réduisent la dimension des vecteurs représentant les textes narratifs tout en préservant, en théorie, la signification des textes. Après le retrait de mots vides et la radicalisation, le texte narratif est converti en un vecteur où chaque rangée représente le nombre de fois qu'un mot apparaît dans le texte. Les textes narratifs de référence sous forme de vecteur sont les données utilisées pour former les algorithmes d'apprentissage automatique, les mettre à l'essai et les valider. Une fois qu'un algorithme est choisi, il est utilisé pour classer de nouvelles données.

La représentation de cette manière de textes par des vecteurs comporte l'important avantage qu'elle permet d'incorporer directement des données catégorielles ou numériques qui accompagnent le texte dans le vecteur lui-même en ajoutant simplement les dimensions appropriées pour représenter ces données. Pour le SCSSEVM, on s'attend à ce que les textes narratifs de nombreux secteurs de compétence soient accompagnés de certaines données catégorielles et numériques, qui peuvent être incorporées dans n'importe quel algorithme que nous développons.

Une quantité importante de recherches ont été menées sur l'utilisation de l'apprentissage automatique afin d'identifier les événements provenant de textes narratifs inclus dans les demandes d'indemnisation des travailleurs. Plusieurs algorithmes d'apprentissage automatique ont été mis en œuvre avec succès dans ce domaine, y compris la machine à vecteurs de support (SVM), les classificateurs naïfs et flous bayésiens et la régression logistique régularisée (Marucci-Wellman, Corns, & Lehto, 2017). Ces algorithmes ont été comparés les uns avec les autres, et Marucci-Wellman et coll. ont observé que, bien que la régression logistique fonctionnait mieux que les autres algorithmes individuellement, le meilleur classificateur qu'ils ont pu construire utilisait en fait un mélange de classificateurs naïfs bayésiens et de SVM. Plus précisément, un cas donné serait classé automatiquement uniquement si le classificateur naïf bayésien et le SVM avaient classé le cas dans la même catégorie. Autrement, il serait renvoyé à des codeurs humains aux fins de codage manuel.

Parmi les articles que nous avons examinés, il semble y avoir un consensus selon lequel les algorithmes d'apprentissage automatique n'ont pas encore progressé au point où ils peuvent remplacer tout à fait les codeurs humains et que la manière la plus efficace d'utiliser ces algorithmes consiste à coder automatiquement certains cas et à retourner les cas plus « difficiles » aux codeurs manuels. La plupart des algorithmes de classification de textes fonctionnent en estimant la probabilité qu'un cas se situe dans une catégorie donnée, alors une manière évidente d'utiliser un tel algorithme consiste à choisir un seuil acceptable de probabilité prévue au-dessus duquel un cas est codé automatiquement. Par exemple, on pourrait former un classificateur naïf bayésien qui code automatiquement tous les cas où la probabilité prévue de se situer dans une catégorie donnée est supérieure à 80 % et renvoie aux codeurs tous les cas qui se situent en deçà de cette valeur. La valeur optimale de ce seuil dépend du contexte, et il faut être très prudent au moment de la choisir.

Parmi la recherche actuelle sur l'identification d'événement, les travaux effectués sur les textes narratifs portant sur les accidents du travail semblent s'appliquer le mieux à notre projet, et nous utiliserons cette recherche comme un tremplin pour nos méthodes. En fait, le code informatique et les données utilisées par de nombreux chercheurs dans ce domaine ont été rendus publics, ce qui a déjà grandement accéléré notre propre travail.

4. Travaux actuels et futurs

Il y a deux principales tâches qui comprennent les travaux futurs visant à coder les données sur la maltraitance des enfants pour qu'elles soient utilisées par l'ASPC et ses partenaires. À court terme, nous allons poursuivre notre travail sur l'élaboration d'un modèle en langage de programmation Python, que nous utiliserons pour mettre à l'essai divers algorithmes sur les données des textes narratifs provenant de différentes sources. À plus long terme, le développement continu sera nécessaire pour mettre à jour le codeur automatique au fur et à mesure que la législation et les systèmes de données changeront et que les données seront disponibles.

Nous avons utilisé une version préliminaire de notre modèle en langage de programmation Python pour mettre en œuvre avec succès une régression logistique régularisée, une machine à vecteurs de support et les algorithmes naïfs bayésiens afin de classer automatiquement les textes narratifs sur les accidents en milieu de travail publiquement disponibles en ligne. Au fur et à mesure que nous recevrons des données sur la maltraitance des enfants en provenance des provinces et des territoires, nous élaborerons un algorithme (ou une combinaison d'algorithmes) pour chacune des configurations uniques des données, plutôt qu'un algorithme unique pour toutes les provinces et tous les territoires. Si nous devions élaborer uniquement un seul algorithme pancanadien, il est probable qu'il classerait efficacement uniquement les textes narratifs provenant des plus grandes provinces, ce qui n'est évidemment pas idéal. Cependant, le calendrier pour la réception des données nécessaires n'a pas encore été établi.

Les données sur la maltraitance des enfants sont de nature extrêmement sensible, et le fait que les lois et les systèmes de données varient entre les provinces et les territoires complique particulièrement le processus d'acquisition des données par Statistique Canada. De plus, nous aurons besoin de données qui ont déjà été classées par les travailleurs des services de protection de l'enfance comme données de références pour la formation de nos algorithmes, et nous ne savons pas encore très bien comment ce travail sera fait. Entre temps, Statistique Canada étudie la possibilité d'acquérir des données d'enquêtes et des textes narratifs sur la maltraitance des enfants d'autres sources que nous pourrions utiliser pour faire les travaux préliminaires de formation des algorithmes et déterminer le vocabulaire et les phrases propres à ce sujet.

5. Conclusion

Il y a un besoin évident pour des données nationales complètes sur la maltraitance d'enfants. Afin de répondre à ce besoin, Statistique Canada a recommandé à l'ASPC d'entreprendre un recensement administratif annuel des organismes de protection de l'enfance. Comme chaque province et territoire du Canada est responsable de ses propres lois en matière de protection de l'enfance, la structure et le contenu des données provenant de chaque secteur de compétence devraient varier considérablement, ce qui représente un défi de taille pour le codage des données après leur réception. Nous estimons que nous pouvons atténuer des difficultés liées au codage humain des données sur la maltraitance des enfants en incorporant des techniques d'apprentissage automatique au processus de codage. Une fois que nous commencerons à recevoir des données, nous allons élaborer, mettre à l'essai et maintenir divers algorithmes d'apprentissage automatique afin de créer des codeurs automatiques efficaces pouvant servir pour réduire le besoin en codage manuel.

Remerciements

Nous tenons à remercier Lil Tonmyr (ASPC) et Wendy Hovdestad (ASPC) pour leurs idées et leurs précieux commentaires sur ce projet.

Bibliographie

Agence de la santé publique du Canada (2010), *Étude canadienne sur l'incidence des signalements de cas de violence et de négligence envers les enfants 2008 (ECI-2008) : Données principales*.

Danso, S., E. Atwell, et O. Johnson (2014), « A comparative study of machine learning methods for verbal autopsy text classification », préimpression, (extrait de arxiv.org/abs/1402.4380).

Marucci-Wellman, H. R., H. L. Corns, et M. R. Lehto (2017), « Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review », *Accident Analysis and Prevention*, p. 359-371.

Measure, A. (2014), « Automated coding of worker injury narratives », *JSM Proceedings, Government Statistics Section, American Statistical Association*, p. 2124-2133.

Potter, D., W. Hovdestad, et L. Tonmyr (2013), « Sources of child maltreatment information in Canada », *Minerva Pediatr*, 65, p. 37-49.

Tonmyr, L., A. Asokumar, W. E. Hovdestad, M. Shields, J. Laurin, et L. Burnside (2018), « Can coders abstract child maltreatment variables from child welfare administrative data and case narratives for public health surveillance in Canada? », document de travail présenté à la International Society for the Prevention of Child Abuse and Neglect, Prague, République Tchèque.

Zhang, Y., I. Marshall, et B. C. Wallace (2016), « Rationale-augmented convolutional neural networks for text classification », *Proceedings of the Conference on Emperical Methods in Natural Language Processing*, p. 795-804.