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

Besoin de vitesse : Utilisation de fastText (apprentissage automatique) afin de coder l'Enquête sur la population active

par Justin Evans et Javier Oyarzun

Date de diffusion : le 5 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

Besoin de vitesse : Utilisation de fastText (apprentissage automatique) afin de coder l'Enquête sur la population active

Justin Evans et Javier Oyarzun[1]

Résumé

L'Enquête sur la population active (EPA) de Statistique Canada joue un rôle fondamental dans le mandat de Statistique Canada. L'information sur le marché du travail fournie par l'EPA est l'une des mesures les plus actuelles et les plus importantes du rendement global de l'économie canadienne. Le codage de l'industrie du répondant selon le Système de classification des industries de l'Amérique du Nord (SCIAN), de la profession selon le Système de classification nationale des professions (CNP) et de la principale catégorie de travailleurs (PCDT) fait partie intégrante du traitement mensuel des données de l'EPA. Chaque mois, jusqu'à 20 000 enregistrements sont codés manuellement. En 2020, Statistique Canada a travaillé au développement de modèles d'apprentissage automatique utilisant fastText afin de coder les réponses au questionnaire de l'EPA selon les trois classifications mentionnées précédemment. Le présent article donnera un aperçu de la méthodologie développée et des résultats obtenus à partir d'une application potentielle de l'utilisation de fastText dans le processus de codage de l'EPA.

Mots clés : apprentissage automatique; Enquête sur la population active; classification de texte; fastText.

1. Introduction

L'Enquête sur la population active (EPA) est une enquête essentielle à la mission de Statistique Canada, car elle joue un rôle fondamental dans l'estimation des conditions du marché du travail du Canada. L'EPA fournit des estimations de l'emploi selon l'industrie, la profession, les secteurs public et privé et les heures travaillées, entre autres. Les résultats de l'enquête servent également à prendre des décisions concernant la création d'emplois, l'éducation et la formation, les régimes de pension et le soutien du revenu. L'une des principales composantes de l'EPA est le codage mensuel des enregistrements du Système de classification des industries de l'Amérique du Nord (SCIAN), de la Classification nationale des professions (CNP) et de la principale catégorie de travailleurs (PCDT). Chaque enregistrement contient des descriptions numériques et textuelles, recueillies auprès des répondants à l'enquête, qui servent à attribuer un code dans chaque classification. À Statistique Canada, les codeurs de la Division des opérations et de l'intégration codent manuellement entre 15 000 et 20 000 enregistrements en 7,5 jours. En raison du temps limité et du grand volume de codage requis, une solution d'apprentissage automatique a été recherchée.

Les organismes nationaux de statistique (ONS) utilisent l'apprentissage automatique pour diverses tâches, y compris l'imputation, l'estimation et la classification de texte (Équipe de l'apprentissage automatique de la Commission économique pour l'Europe des Nations Unies, 2018). Dans la classification de texte, les descriptions fournies par les répondants à l'enquête et associées à une variable d'intérêt (classifications du SCIAN, de la CNP et de la PCDT) peuvent être utilisées dans les tâches d'apprentissage automatique supervisé. Dans l'ensemble des ONS, les algorithmes utilisés pour les tâches d'apprentissage automatique supervisé – forêt aléatoire, bayésien naïf, fastText, machine à vecteur de support (SVM pour l'anglais *Support Vector Machine*), réseau neuronal convolutif, perceptrons multicouches, etc. – et les étapes de prétraitement peuvent varier selon le type et la disponibilité des données (Sthamer, 2020).

À Statistique Canada (StatCan), la majorité des activités de codage automatisé sont effectuées au moyen d'une application de codage développée à l'interne, G-Code. L'intégration récente à G-Code de fastText , une bibliothèque

[1]Justin Evans, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (justin.evans@statcan.gc.ca); Javier Oyarzun, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 (javier.oyarzun@statcan.gc.ca)

de représentation et de classification de texte créée par le Laboratoire de recherche sur l'IA de Facebook (Joulin et coll., 2016), et la connexion à la plateforme de codage de StatCan Environnement de codage et de correction (ECC) ont facilité l'utilisation de modèles d'apprentissage automatique dans la production de statistiques officielles. Le présent article examinera en détail les classifications de l'EPA qu'il faut coder, le cadre d'apprentissage automatique pour ce faire, et l'évaluation des résultats.

2. Classifications de la CNP, du SCIAN et de la PCDT

2.1 Classification nationale des professions (CNP)

La CNP[2] a été élaborée et est tenue à jour par Emploi et Développement social Canada (EDSC) et Statistique Canada. Elle se fonde sur le cadre organisationnel accepté à l'échelle nationale des professions sur le marché du travail canadien. Les critères utilisés pour regrouper les professions de la CNP sont généralement les tâches, les fonctions et les responsabilités de la profession et ils peuvent comprendre le degré de responsabilité et la complexité du travail, ainsi que les produits et services fournis. Ainsi, la CNP a été établie selon une structure hiérarchique à quatre niveaux, allant de grandes catégories à des groupes de base, ce qui donne 500 classes distinctes.

Figure 2.1-1. Structure hiérarchique des classes de la CNP

0 – Gestion
- 00 – Cadres supérieurs/cadres supérieures
    - 001 – Membres des corps législatifs et cadres supérieurs/cadres supérieures
        - 0011 – Membres des corps législatifs
        - 0012 – Cadres supérieurs/cadres supérieures – administration publique
        - 0013 – Cadres supérieurs/cadres supérieures – services financiers, communications et autres services aux entreprises

2.2 SCIAN

Le SCIAN[3] a été conçu en collaboration par les organismes statistiques du Canada, du Mexique et des États-Unis. Le SCIAN est articulé autour des principes de l'offre ou de la production, afin de s'assurer que les données sur les industries qui sont classées en fonction du SCIAN se prêtent à l'analyse de questions liées à la production, comme le rendement industriel. Les critères utilisés pour regrouper les industries dans le SCIAN sont les structures des facteurs de production, les qualifications de la main-d'œuvre et les processus de production. Tout comme la CNP, le SCIAN a été établi selon une structure hiérarchique dans laquelle le niveau le plus élevé divise l'économie en 20 secteurs et le niveau le plus bas divise les industries par activité économique, ce qui donne 324 classes distinctes.

2.3 Principale catégorie de travailleurs (PCDT)

La PCDT sert à classer les travailleurs comme employés dans le secteur public ou privé. La PCDT a donc deux classes (1 – Public, 2 – Privé).

[2]Classification nationale des professions (CNP) 2016 version 1.3;
https://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVD&TVD=1267777
[3] Système de classification des industries de l'Amérique du Nord (SCIAN) Canada 2017 version 3.0 :
https://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVD&TVD=1181553

3. Cadre d'apprentissage automatique

3.1 Algorithme

Il a été démontré que fastText permet d'obtenir plus rapidement des résultats comparables à ceux des classificateurs classiques de l'apprentissage profond au moyen de représentations en n-grammes de sous-mots et d'un softmax hiérarchique. FastText permet de représenter chaque mot comme un sac de n-grammes de caractères. Par exemple, « texte » avec n=3, est converti en « <te, tex, exte, xte> ». Les vecteurs de mots peuvent donc comprendre des mots incorrectement orthographiés ou inventés et des mots concaténés (Joulin et coll., 2016). Dans les réponses d'enquête, qui contiennent souvent des mots incorrectement orthographiés et une terminologie propre à une industrie, nous avons constaté que fastText est en mesure de produire des prédictions de grande qualité. Nous avons aussi constaté que, compte tenu du grand nombre d'enregistrements et de grandes classes de l'EPA (SCIAN-4, CNP-4), le fait que fastText utilise un softmax hiérarchique, une approximation de la fonction de perte softmax, a contribué à réduire le temps d'entraînement du modèle.

3.2 Entraînement

Pour évaluer l'ajustement du modèle, nous avons divisé les données déjà codées disponibles en un ensemble de données d'entraînement et de validation. Pour tenir compte des variations mensuelles des professions déclarées, une année de l'EPA a été retenue comme ensemble de données de validation (juin 2019 à mai 2020, enregistrements = 200 000). Comme l'EPA examine ses données historiques chaque fois qu'une nouvelle année de classification est introduite, par exemple, au passage du SCIAN 2012 au SCIAN 2017v1, nous avons pu inclure les données d'entraînement de janvier 2013 à mai 2019 (enregistrements = 1 500 000). La sélection des variables auxiliaires de texte d'entrée dans l'ensemble de données d'entraînement a été évaluée à l'aide du chi carré et du gain d'information. Un prétraitement minimal du texte a été effectué lors de la concaténation de certaines variables de texte, à savoir les majuscules, la suppression d'accent et le masquage. Le réglage des hyperparamètres de recherche aléatoire par k-fois une validation croisée, a été effectué sur toutes les variantes du modèle avant d'être exécuté sur l'ensemble de données de validation.

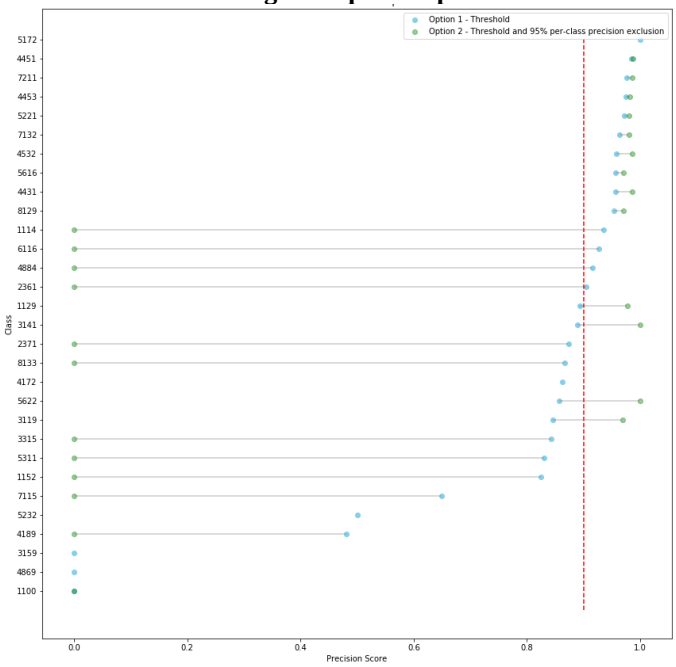
3.3 Seuil et sélection de classe

Dans la prédiction par modèle, fastText renvoie la classe ayant le score d'activation le plus élevé après la couche softmax, ce qui donne une cote de confiance située entre 0 et 1. Bien que cette cote soit faiblement associée à la probabilité que le modèle prédise une classe positive, des méthodes de calage sophistiquées peuvent servir à régler les cotes de confiance du classificateur (Zadrozny et Elkan, 2002). Ici, nous montrons que, bien que l'exclusion de toutes les prédictions d'apprentissage automatique avec une cote de confiance inférieure à un seuil donné puisse entraîner un taux de codage et un score de précision global élevés (tableau 3.3-1, option 1), de nombreuses classes situées au-dessus du seuil tombent sous le niveau acceptable leur permettant d'être utilisées dans des statistiques officielles. (figure 3.3-1, option 1). Par conséquent, afin d'éviter les erreurs systématiques et d'optimiser le taux de codage, nous avons sélectionné un seuil après avoir exécuté une boucle sur toutes les cotes de confiance disponibles tout en maintenant une précision de 95 % par classe dans notre ensemble de données de validation. Les prédictions d'apprentissage automatique dans les classes dont la précision était inférieure à 95 % ont été exclues du processus de codage proposé (option 2). De futurs travaux porteront sur l'effet du calage des cotes de confiance de fastText sur le nombre de classes acceptées dans notre processus.

Tableau 3.3-1.
Analyse des options aux fins de seuillage dans une prédiction par modèle

Options	Seuil de confiance	Taux de codage	Précision globale	Classes > précision de 95 %
1 – Seuil	0,88	62,08 %	95,10 %	100
2 – Seuil et exclusion de précision de 95 % par classe	0,96	41,41 %	98,29 %	143

Figure 3.3-1.
Score de précision d'un sous-ensemble de classes du SCIAN dans le seuillage d'une prédiction par modèle. Chaque point représente une classe. Le trait rouge indique une précision de classe de 95 %.



4. Résultats

À partir du modèle décrit à la section 3, les modèles du SCIAN, de la CNP et de la PCDT ont été créés et mis à l'essai sur une année de l'EPA (ensemble de données de validation) (tableau 4-1). Dans les sous-sections suivantes, nous examinerons l'incidence de l'utilisation de ces modèles dans le processus de codage et sur les estimations publiées de l'EPA.

Tableau 4-1.
Taux de codage automatique global et score de précision au niveau de la classification et de l'enregistrement, dans l'ensemble de données de validation.

Classification	Niveau de la classification		Niveau de l'enregistrement	
	Taux de codage automatique	Score de précision	Taux de codage automatique	Score de précision
SCIAN	40,9 %	98,3 %	18,5 %	99,0 %
CNP	30,7 %	97,5 %	18,5 %	98,2 %
PCDT	100 %	97,6 %	18,5 %	98,6 %

4.1 Comparaison avec le codage manuel

L'évaluation de la qualité d'un processus de codage peut être longue et coûteuse. Ici, nous avons pu utiliser la migration de l'EPA de sa plateforme de codage précédente (CODCO) à sa nouvelle plateforme de codage (CCE) pour comparer les enregistrements codés manuellement avec ceux codés par apprentissage automatique. Pour ce faire, nous avons comparé les classes de sortie finales d'un enregistrement codé dans une plateforme de codage au même enregistrement dans une plateforme de codage différente ou aux classes prédites au moyen de l'apprentissage automatique, où le taux d'erreur est le taux de désaccord sur la classe de sortie finale. Nous montrons que sur le sous-ensemble d'enregistrements qui auraient été entièrement codés par apprentissage automatique (~ 18,5 % des enregistrements), le taux d'erreur entre deux mois codés manuellement (CODCO vs CCE) est le même que si l'apprentissage automatique (avant le contrôle de la qualité) avait été utilisé dans l'une ou l'autre des plateformes de codage (AA vs CODCO, AA vs CCE). Nous montrons également que les enregistrements d'apprentissage automatique sélectionnés aléatoirement dans le processus de contrôle de la qualité (CQ) de l'ECC présentent le même faible taux d'erreur (CQ de l'apprentissage automatique dans l'ECC) (tableau 4.1-1). Ces résultats montrent que le seuil et les critères de sélection des classes (section 3.3) produisent des prédictions pour le SCIAN, la CNP et la PCDT de la même qualité que si seul le processus manuel était employé.

Tableau 4.1-1
Comparaison des taux d'erreur du codage manuel (sous-ensemble de l'AA) dans différentes plateformes de codage (CODCO, CCE)

	CODCO vs CCE	AA vs CODCO	AA vs CCE	CQ de l'AA dans CCE
Année-Mois	2020-09	2020-09	2020-09	2021-06
SCIAN	2,2 %	2,2 %	1,7 %	0,9 %
CNP	2,3 %	2,3 %	2,0 %	1,2 %
PCDT	1,9 %	1,9 %	1,7 %	1,3 %

4.2 Analyse des ruptures de séries chronologiques

Afin d'évaluer si l'apprentissage automatique a une incidence sur les tendances de l'EPA, qui sont des résultats clés pour les utilisateurs, nous avons cherché à déterminer s'il y a des changements observables dans la répartition des classes au fil du temps. Chaque classe du SCIAN, de la CNP et de la PCDT a été exprimée sous forme d'une série de points de données ordonnés dans le temps (série chronologique). Pour une série chronologique, tout changement abrupt à un moment donné est appelé rupture structurelle ou rupture de série chronologique. Parce que les ruptures structurelles indiquent un changement important dans les données des séries chronologiques de l'EPA, elles doivent être communiquées aux utilisateurs des données.

Pour chaque mois de la période de l'année, un nombre d'unités par classe a été calculé pour le codage manuel (sans AA) et le codage intégré manuel et AA (avec AA) afin d'évaluer les éventuelles ruptures des séries chronologiques. En raison du faible taux d'erreur des modèles d'apprentissage automatique, la répartition des classes ne semble pas radicalement différente quand les prédictions de l'apprentissage automatique sont intégrées au processus manuel (tableau 4.2-1). Ici, la plus grande différence dans les nombres de classes est dans les « restaurants à service complet et établissements de restauration à service restreint » (SCIAN – 7225), six enregistrements en plus ayant été attribués à cette classe.

Figure 4.2-1. Modification de la répartition des classes lors de l'intégration des prédictions de l'AA. Période de référence : mai 2020.

Garnisseurs/garnisseuses de tablettes, commis et préposés/préposées aux commandes dans les magasins

Serveurs/serveuses au comptoir, aides de cuisine et personnel de soutien assimilé

Concierges et surintendants/surintendantes d'immeubles

CNP-4	Nbre codage manuel	Nbre codage intégré AA	Différence absolue
6622	258	263	5
6711	388	392	4
6733	218	214	4
0631	127	130	3
4112	73	76	3
6731	283	286	2
7452	177	174	2
6341	88	90	2
6511	63	61	2
…	…	…	…

SCIAN-4	Nbre codage manuel	Nbre codage intégré AA	Différence absolue
7225	874	880	6
7224	15	13	2
4471	61	63	2
5616	91	93	2
5617	399	401	2
7139	266	268	2
8113	72	70	2
9120	274	272	2
1112	33	34	1
…	…	…	…

Restaurants à service complet et établissements de restauration à service restreint

L'EPA utilise la méthode d'estimation de la variance bootstrap pour produire un rapport mensuel sur le total de l'emploi et le chômage (Statistique Canada, 2017). C'est pourquoi nous avons cherché à déterminer s'il y avait une rupture structurelle dans ces estimations publiées si l'apprentissage automatique avait été utilisé. Deux critères différents (indiqués ci-dessous) ont servi à déterminer les différences potentielles entre les estimations manuelles et les estimations intégrées codage manuel et AA . Une méthode semblable a été adoptée aux fins de l'évaluation des ruptures de séries chronologiques au niveau du SCIAN à six chiffres quand l'Enquête sur les dépenses en immobilisations (EDI) est passée au Programme intégré de la statistique des entreprises (PISE) (Oyarzun, 2016).

- Critère 1 : Signaler l'estimation si l'intervalle de confiance de 90 % pour l'estimation manuelle ne chevauche pas l'intervalle de confiance de 90 % pour l'estimation intégrée codage manuel et apprentissage automatique.
- Critère 2 : Signaler l'estimation si l'estimation ponctuelle du codage manuel ne tombe pas dans l'intervalle de confiance (IC) de 90 % pour l'estimation intégrée codage manuel et apprentissage automatique.

Ici, nous avons comparé les estimations des classes à chaque niveau de chiffres (p. ex. SCIAN-4, SCIAN-3, SCIAN-2, SCIAN-1) sur une période de 12 mois (d'octobre 2019 à septembre 2020). Nous constatons que l'intégration de l'AA ne donne pas lieu à des estimations de la variance qui se situent en dehors des limites de confiance de l'estimation du codage manuel (critère 1) ni à une estimation ponctuelle du processus manuel hors de sa limite de confiance (critère 2). Bien que nous n'ayons pas observé de ruptures dans les 13 665 estimations comparées (SCIAN : 5 332; CNP : 8 309; PCDT : 24), nous pouvons nous attendre à une certaine volatilité dans les classes où le nombre d'enregistrements est faible, mais cela ne devrait pas poser problème aux utilisateurs de données.

4.3 Analyse du contrôle de la qualité

Le contrôle de la qualité est une technique consistant à s'assurer que la qualité est supérieure à un niveau établi par la mesure de la qualité des caractéristiques d'intérêt, par sa comparaison à une cible de qualité, et par des mesures correctives si la norme n'est pas atteinte (Statistique Canada, 2003). Dans CCE, le codage manuel et l'apprentissage automatique font l'objet d'un contrôle de la qualité au moyen d'un échantillon aléatoire simple (dénommé « CQ régulier » dans le tableau 4.3-1). Le taux de CQ de chaque codeur est déterminé en fonction du taux d'erreur du mois précédent et a une incidence directe sur le taux d'erreur sortant. Selon le nombre moyen d'enregistrements reçus pour

le codage et les taux de CQ attribués, les codeurs sont en mesure d'effectuer un nombre maximal d'enregistrements chaque mois (environ 24 200 enregistrements). Ici, nous présentons deux scénarios dans lesquels l'utilisation de l'apprentissage automatique peut faire passer les codeurs manuels du codage de première ligne à la vérification. De plus, en réduisant le CQ pour l'apprentissage automatique par le passage du scénario 2 au scénario 3 (indiqué par « Total des codes manuels attribués (sans CQ régulier) »), nous pouvons augmenter le CQ pour les enregistrements plus difficiles et plus sujets à l'erreur qui restent codés par des codeurs manuels (indiqué par « Charge de travail totale restante (CQ régulier) », sans AA), ce qui réduit le taux d'erreur de sortie. Si ces gains d'efficacité de l'AA sont réinvestis dans un CQ de l'AA à court terme de 50 % et un CQ de l'AA à long terme de 10 %, le taux d'erreur sortant baisserait à 10,4 % et 9,5 % respectivement (tableau 4.3-1).

Tableau 4.3-1
Scénarios de réinvestissement des gains d'efficacité de l'AA dans le CQ du codage manuel et incidence sur le taux d'erreur estimé de sortie.

Nbre	Mesure	Scénario 1 (pas d'AA)	Scénario 2 (AA à 50 % de CQ)	Scénario 3 (AA à 10% de CQ)
1	« Budget » d'enregistrements du centre de codage	24 200	24 200	24 200
2	Nombre estimé de codes par AA	0	3 000	3 000
3	Total des codes manuels attribués (sans CQ régulier)	22 000	20 545	19 155
4	Charge de travail totale restante (CQ régulier), sans AA	2 200	3 655	5 046
5	Enregistrements ne faisant pas partie du CQ, sans AA	12 929	8 889	7 896
6	Enregistrements ne faisant pas partie du CQ, avec AA	0	1 500	2 850
7	Estimation du nombre d'erreurs non liées à l'AA	2 586	1 778	1 579
8	Estimation du nombre d'erreurs de sortie, avec AA	0	45	85,5
9	**Taux d'erreur estimé**	**14,78 %**	**10,42 %**	**9,51 %**

5. Conclusion

Nous avons fait la démonstration d'un cadre de modèle conçu pour coder des prédictions de haute qualité du SCIAN, de la CNP et de la PCDT pour l'EPA. Grâce à l'établissement de seuils de confiance réglables et à des règles d'exclusion de classes, nous pouvons garantir des codes produits par apprentissage automatique ayant la qualité requise pour les statistiques officielles. De plus, notre examen du processus de CQ et de codage donne à penser que l'apprentissage automatique modifiera la nature du travail de codage manuel et améliorera le taux d'erreurs de sortie de l'EPA. Enfin, notons que le codage par apprentissage automatique a été mis en œuvre avec succès dans le processus de production de l'EPA en octobre 2021.

Remerciements

Les auteurs remercient les personnes suivantes pour leur contribution : Laura Wile, Khushnood Khan, Anthony Yeung, Danielle Lebrasseur, Julie Portelance, Isaac Ross, Kartik Vashisth, Sreenija Koya, Sylvie White, Meghan Fulford, David Menyah et Charles Mitchell.

Bibliographie

Joulin, E. Grave, P. Bojanowski, T. Mikolov. (2016), « Bag of Tricks for Efficient Text Classification », *arXiv* : 1607,01759.

Oyarzun, J. (2016). « *Une aiguille dans une botte de foin – La façon de détecter un bris dans une série : l'expérience de l'Enquête sur les dépenses en immobilisations dans le Programme intégré de la statistique des entreprises* », article présenté au congrès annuel de la SSC – Recueil du groupe des méthodes d'enquête.

Statistique Canada (2017), « Méthodologie de l'Enquête sur la population active du Canada », nº 71-526-X2017001 au catalogue.

Statistique Canada (2003), « Méthodes et pratiques d'enquête ». 309-319. Nº 12-587-XPE au catalogue.

Sthamer, C. (2020) « UNECE–HLG-MOS Machine Learning Project Classification and Coding Theme Report », rapport.

Équipe de l'apprentissage automatique de la Commission économique pour l'Europe des Nations Unies (2018) « L'utilisation de l'apprentissage automatique dans les statistiques officielles », rapport.

Wood, S., R. Muthyala, Y. Jin, Y. Qin, N. Rukadikar, A. Rai et H. Gao (2017), « Automated Industry Classification with Deep Learning », article présenté à la conférence internationale de l'IEEE sur les mégadonnées.

Zadrozny, B. et C. Elkan (2002), « Transforming Classifier Scores into Accurate Multiclass Probability Estimates », KDD, p. 694-699.