Les techniques bayésiennes appliquées à des données d'accidentologie

Claire Tissot

Institut pour la Maîtrise des Risques
Sûreté de Fonctionnement - Management - Cindyniques

Les techniques bayésiennes appliquées à des données d'accidentologie

Applying Bayesian techniques to occupational accident data

Claire Tissot
[1]INRS, 65, Bd Richard Lenoir 75011
Paris,
France
claire.tissot@inrs.fr
01 40 44 30 80

***Résumé*—Cet article présente l'application d'une classification naïve bayésienne à des données d'accidents du travail selon une variable textuelle « activité de la victime ». Les résultats des modélisations selon les variables « sexe », « âge » et « gravité des lésions » sont décrits et discutés.**

***Abstract*—This article presents the application of a Naive Bayes classifier of occupational accident data with a textual variable « victim's activity ». The models resulting from the learning on « gender of the victim », « age of the victim » and « severity of injury » are described and discussed.**

***Mots-clés*—textmining, retour d'expérience, traitement automatique du langage naturel, classifieur Naive Bayes, base de données EPICEA, analyse exploratoire, machine learning**

I. INTRODUCTION

La digitalisation du monde du travail s'accompagne d'un stockage exponentiel de données numériques. Dans le domaine du retour d'expérience industriel, de l'évaluation et de la maîtrise des risques, ces bases de données représentent une mémoire des signalements d'aléas de production, de dysfonctionnements, d'accidents, etc. Elles sont de mieux en mieux exploitées à mesure que les méthodologies d'analyse se vulgarisent sous l'appellation de *deep learning*, *machine learning* ou *datamining* incluant le *textmining*. En parallèle, l'évolution des algorithmes, des technologies informatiques et le développement de communautés d'utilisateurs rendent possible l'application et la diffusion de nouvelles approches descriptives ou prédictives à des domaines variés tels que l'accidentologie, ces approches pouvant être probabilistes et impliquer les statistiques bayésiennes [1,2].

II. CONTEXTE

La base de données EPICEA [1] [3] gérée par l'INRS[2] contient des analyses d'accidents du travail survenus à des salariés du régime général. Ces analyses sont réalisées par les contrôleurs de sécurité des quinze CARSAT[3], la CRAMIF[4] et les quatre CGSS [5]. Tous les accidents mortels, hors accidents du trajet, doivent donner lieu à une enquête. Parmi les malaises et les accidents non mortels, ceux jugés pertinents pour la prévention alimentent également la base de données. L'organisation et la structure d'EPICEA sont détaillées en [4].

Avec environ 200 à 300 nouveaux cas saisis chaque année, EPICEA n'est pas représentative de tous les accidents du travail survenant en France. Les résultats obtenus ne sont donc pas généralisables à l'ensemble des accidents du régime général et ne peut pas servir à définir des indices de sinistralité. Cependant, le contenu détaillé de chaque accident et la présence de données textuelles permettent la réalisation d'études exploratoires bénéficiant de la complémentarité de données structurées et textuelles, notamment lors de l'utilisation de méthodologies de TAL[6] et de *textmining*. Différentes techniques ont déjà été appliquées à ces données, par exemple l'analyse factorielle et la classification hiérarchique [5], LSA[7] et le *topic modeling* [6]. La présente communication met en œuvre une nouvelle approche basée sur une modélisation probabiliste et les statistiques bayésiennes.

Cette analyse est un extrait d'une étude concernant l'accidentologie selon le sexe regroupant un certain nombre de méthodologies visant à identifier des différences entre les situations de travail des hommes et des femmes ayant été

1 Études de prévention par l'informatisation des comptes rendus d'enquêtes accident
2 Institut national de recherche et de sécurité pour les accidents du travail et les maladies professionnelles
3 Caisse d'assurance retraite et de santé au travail
4 Caisse régionale d'assurance maladie Île-de-France
5 Caisse générale de sécurité sociale (départements d'outre-mer)
6 Traitement automatique du language naturel
7 *Latent semantic analysis*

accidentés, à travers notamment l'activité exercée au moment de l'accident. Cette étude a été l'occasion de découvrir et d'appliquer une technique de classification bayésienne. Elle a ensuite été élargie à deux autres variables structurées : l'âge de la victime et la gravité des lésions. L'un des objectifs de l'étude était d'identifier l'existence d'activités spécifiques aux différentes modalités[8] des trois variables structurées. Si certaines activités relèvent plus d'un sexe que d'un autre, d'une classe d'âges que d'une autre ou sont exercées dans des contextes pouvant s'avérer plus dangereux, la connaissance de ces différences doit être prise en compte dans la prévention des risques et être intégrée dans toute démarche d'évaluation des risques professionnels. Un deuxième objectif consistait à explorer et tester une technique de traitement automatique du langage naturel par une modélisation statistique.

III. MÉTHODOLOGIE ET TRAVAUX DÉVELOPPÉS

A. Méthodologie

Les analyses ont été effectuées avec le classifieur naïf bayésien, méthode simple et rapide qui s'applique très bien au traitement de données textuelles [7]. Un classifieur est un algorithme qui classe les données d'un fichier dans des groupes correspondant aux modalités d'une variable cible en fonction de la similarité des données. Trois variables cibles ont été utilisées : le sexe de la victime, l'âge de la victime et la gravité des lésions. Dans ce modèle, les variables explicatives sont les mots du corpus correspondant à l'activité de la victime au moment de l'accident. Si la variable cible est le sexe, l'objectif de la classification est de séparer les accidents des hommes et des femmes selon les mots décrivant leur activité.

Le classifieur naïf bayésien est un modèle probabiliste d'apprentissage supervisé basé sur le théorème de Bayes[9]. Les modèles probabilistes visent à décrire les processus qui génèrent des données. Avec la statistique classique (fréquentiste ou fréquentielle), les modèles s'appuient sur l'estimation de paramètres inconnus supposés fixes, estimés selon la loi des grands nombres. L'inférence fréquentiste énonce des règles de décision après l'acceptation ou le rejet de tests d'hypothèses. Elle est jugée fiable si elle mène rarement à des conclusions incorrectes. Les statistiques bayésiennes utilisent des paramètres aléatoires décrits par des distributions de probabilités *a priori* d'un phénomène, qui sous-tendent une notion de degré de croyance. L'inférence bayésienne repose sur la formule de Bayes (1) : si on connaît les effets d'un phénomène P(B\A), on peut remonter aux causes par une inversion des probabilités, en calculant la probabilité postérieure P(A\B). Le modèle pourra être affiné par la prise en compte de nouvelles probabilités observées ou de données plus complètes, constituant une consolidation de l'apprentissage et une « révision des croyances ». La contribution [8] présente une intéressante comparaison entre les deux approches statistiques fréquentiste et bayésienne.

$$P(A\setminus B)=\frac{P(B\setminus A)P(A)}{P(B)} \quad (1)$$

Les réseaux bayésiens (ou diagrammes d'influence) font référence à une représentation graphique des influences des variables entre elles. Pour une valeur de variable fixée, ils montrent comment se propage la mise à jour des probabilités jusqu'à l'événement final, permettant l'identification des variables les plus influentes [1,2]. Les réseaux bayésiens ne sont pas utilisés ici.

Le classifieur naïf est un apprentissage automatique supervisé dans le sens où l'apprentissage se fait à partir de données présentes dans le fichier et qui servent de guide à l'élaboration de règles de classement. Le terme « naïf » se réfère à l'hypothèse naïve que les variables explicatives (ici les mots) sont indépendantes. Cette hypothèse n'est en général pas respectée, notamment lorsque les données sont textuelles. Cela supposerait en effet que les mots du corpus sont indépendants les uns des autres et que leur position dans la phrase n'importe pas alors qu'en pratique la succession des mots est dictée par les règles grammaticales. Pourtant, différentes études ont montré que le non-respect de l'hypothèse d'indépendance n'avait pas ou peu d'incidence sur les résultats obtenus et cette méthode est largement utilisée [9].

Différentes méthodes d'apprentissage automatique existent, fréquentistes (régressions logistiques, arbres de décision, réseaux de neurones, etc.) comme bayésiennes (classification naïve bayésienne). En pratique, l'apprentissage automatique est effectué sur une partie du fichier de départ, en général 70 %, et les tests du modèle sur les 30 % restants[10]. La qualité du modèle est évaluée par l'application des règles sur le fichier test et la comparaison du résultat prédit avec les données réelles du fichier test. Avec le classifieur naïf, l'apprentissage automatique correspond à l'observation des probabilités dans le fichier d'apprentissage et au calcul des probabilités postérieures. Ces probabilités postérieures peuvent ensuite servir à :

- catégoriser une variable cible (le sexe de la victime) par les variables explicatives (mots décrivant l'activité) en mettant en évidence les associations les plus probables ;
- évaluer la qualité du modèle en comparant la prévision de la variable cible avec la valeur déjà présente dans le fichier test ;
- prédire la variable cible sur un nouveau jeu de données.

Appliquée aux données EPICEA avec, par exemple, la variable « sexe », l'inférence bayésienne calcule les probabilités postérieures du sexe de la victime connaissant la distribution des mots de l'activité (P(sexe\mot)) à partir de : la probabilité apprise (observée) des mots connaissant le sexe (P(mot\sexe)), la probabilité *a priori* du sexe (P(sexe) ou

8 Catégorie d'une variable structurée, par exemple la modalité "Femme" de la variable SEXE, la classe d'âges 20-29 ans de la variable AGE

9 Révérend Thomas Bayes, pasteur et mathématicien anglais (1702-1761)

10 D'autres méthodes existent comme le *bootstrap resampling*

proportion hommes/femmes[11]) et la probabilité des mots du corpus (P(mot)).
(1) peut ainsi être traduite :

$$P(sexe \setminus mot) = \frac{P(mot \setminus sexe)\,P(sexe)}{P(mot)}$$

Le fichier d'apprentissage est visualisé dans le tableau 1 et les prédictions illustrées dans le tableau 2. Un « 1 » correspond à la présence d'un mot dans le texte, un « 0 » à son absence.

TABLEAU 1 FICHIER D'APPRENTISSAGE

	Conduire	**Presse**	**Camion**	...	**Sexe connu**
Texte 1	1	0	1	...	Homme
Texte 2	1	1	0	...	Femme
...	...	...	...	...	…

TABLEAU 2 FICHIER DE TEST ET PRÉDICTIONS

	Vélo	**Conduire**	**Solvant**	...	**Sexe prédit**	**Sexe connu**
Texte 1	1	0	1	...	?	Femme
Texte 2	1	1	0	...	?	Homme
...	...	...	...	...	...	…

En fonction de la probabilité postérieure et d'un seuil de décision (0,5 par défaut), l'accident est classé dans l'une des catégories « homme » ou « femme ». L'évaluation de la modélisation se basera sur la comparaison des sexes connus et prédits à l'aide de la matrice de confusion.

B. Les travaux développés

Avant toute analyse, une préparation des données est nécessaire : élimination de certaines données manquantes en supprimant certains individus, regroupement de modalités rares comme les classes d'âges 60-65 ans avec les plus de 65 ans. De 23 800 lignes au départ, le fichier est réduit à 23 691 lignes dont 22 092 hommes et 1 599 femmes. L'activité de la victime au moment de l'accident est stockée sous deux formes dans la base : deux variables structurées (un verbe à l'infinitif et un complément d'objet direct – COD - prédéfinis) et deux variables texte (un verbe à l'infinitif et un complément d'objet direct ou de lieu, en texte libre). L'intérêt de prendre en compte les variables texte est de disposer d'un éventail le plus large possible de verbes et de COD non prédéfinis à l'avance. Ces deux variables texte ont été concaténées pour fournir une variable texte « activité ». Par exemple les deux modalités « Traverser » et « un passage piéton » sont transformées en activité « Traverser un passage piéton ». Cette nouvelle variable texte a subi les traitements habituels de *textmining*, à savoir la réduction de dimension par élimination des mots vides selon la liste standard des *stopwords* (verbes « être » et « avoir » conjugués, prépositions, déterminants, pronoms personnels ou possessifs, etc.) à laquelle ont été ajoutés les mots « non » « précisé » pour les activités non précisées. Aucune lemmatisation ou réduction à la racine n'a été effectuée, ce qui justifie la présence de mots au pluriel dans les résultats. Le corpus final contient 1 463 mots.

Une première étape a consisté à construire un tableau croisant les 23 691 documents (accidents) en ligne et les 1 463 mots en colonne. L'exploration du tableau permet de décrire les trois variables « sexe », « âge de la victime » et « gravité des lésions », notamment sous forme graphique avec les nuages de mots, les graphes des mots caractéristiques ou les réseaux de mots. L'environnement de programmation R [10] et le package quanteda [11] ont été utilisés. Cette étape consiste à se faire une première idée des associations des mots entre eux et des mots aux différentes modalités hommes/femmes, accidents mortels/graves/pas graves, classes d'âges.

En deuxième étape, le fichier de départ a été scindé aléatoirement en un fichier d'apprentissage de 16 583 accidents x 1 463 mots (70 % du fichier initial) et d'un fichier test de 7 108 accidents x 1 463 mots (30 % restant). La classification naïve bayésienne a été appliquée au fichier d'apprentissage. Les analyses ont été réalisées avec les packages quanteda [11] et quanteda.textmodels [12]. Le principe des traitements effectués sera détaillé avec la variable « sexe de la victime ». Les résultats aborderont les trois variables sexe, âge de la victime et gravité des lésions.

Différentes questions se posent au début des traitements : doit-on transformer le tableau individus x mots en lui appliquant une pondération ? Quelle probabilité *a priori* utiliser : par exemple pour le sexe, la fréquence observée hommes/femmes de l'échantillon d'apprentissage qui est ici respectivement de 93,4 %/6,6 % ou l'équiprobabilité 50 %/50 % ? Quelle distribution utiliser, multinomiale ou Bernoulli ? Ces choix ne sont pas sans influence sur les résultats. La pondération est un calcul sur chaque occurrence des mots du tableau visant à équilibrer l'influence éventuelle des mots selon leur fréquence. Les traitements ont été réalisés d'abord sans pondération, puis avec une pondération tf/idf et une pondération log/entropy. Les références [13,14] exposent les principes de la pondération et les résultats d'expérimentations selon différentes pondérations. Le package lsa [15] a été utilisé pour le calcul des pondérations. Pour la variable « sexe », la modélisation a été réalisée d'une part avec l'équiprobabilité hommes/femmes et d'autre part avec des probabilités correspondantes aux fréquences observées[12]. La modélisation basée sur une distribution « multinomiale » prend en compte l'occurrence des mots alors qu'une une distribution « Bernoulli » est basée sur la

11 Un choix ici est à faire entre la probabilité observée dans les données ou la probabilité d'être un homme ou une femme indépendamment des données

12 Par défaut dans quanteda, les probabilités *a priori* sont equi-probables car il est considéré que l'information portée par les proportions observées peut ne pas avoir de sens ou ne pas pouvoir être interprétée

présence ou l'absence d'un mot et est donc moins riche en information. Les deux distributions ont été testées.

Une fois la modélisation effectuée, il est possible de dresser une liste de mots les plus probables (probabilités les plus élevées) pour chacune des modalités d'une variable et ainsi définir une cartographie des activités les plus associées à chacune des modalités. Par exemple, pour la variable « sexe », avec une modélisation multinomiale, sans pondération et en utilisant les équiprobabilités hommes/femmes, la probabilité *a priori* que la victime soit un homme ou une femme est 0,5. Après la phase d'apprentissage et l'inférence bayésienne, la probabilité que la victime soit une femme sachant que son activité est décrite par le mot « javel » est de 0,95, celle que ce soit un homme n'est que de 0,05. Ou encore, sachant que l'activité de la victime concerne une « toiture », la victime a une probabilité de 0,03 d'être une femme et de 0,97 d'être un homme.

Les trois listes obtenues avec les trois variables cibles ont été comparées avec les modalités caractéristiques calculées par le logiciel SPAD [16] et avec l'identification de mots saillants proposée par quanteda. Il s'agit de deux calculs fréquentiels, une comparaison de fréquences dans SPAD et une mesure de similarité dans quanteda.

L'évaluation d'un modèle de classification automatique consiste à comparer les catégories obtenues avec celles présentes dans le fichier test. Les résultats sont regroupés dans une matrice de confusion qui croise les catégories observées et les catégories prédites. A partir de cette matrice sont calculés différents indicateurs de classement : le rappel qui correspond à une mesure de l'exhaustivité (aussi appelé sensibilité, en anglais recall ou sensitivity), la précision qui est une mesure de la qualité du modèle (specificity), le F-score qui est une moyenne harmonique du rappel et de la précision, l'exactitude (accuracy) qui correspond aux proportions de bien classés, et le taux d'erreur égal à 1-l'exactitude. Un modèle parfait aurait une précision de 1 (aucun faux positifs), un rappel de 1 (aucun faux négatifs) et un F-score de 1. Pour un modèle dans lequel certaines fréquences sont très déséquilibrées comme c'est le cas pour le sexe de la victime, il est conseillé de tenir compte du F-score qui évalue la classe positive (ici les femmes) plutôt que l'exactitude dont une valeur élevée peut simplement refléter une meilleure prédiction de la classe la plus fréquente (les hommes). Les matrices de confusion sont présentées dans les résultats pour les modèles basés sur les données non pondérées.
Une évaluation visuelle à l'aide des courbes ROC [13] et « précision/rappel » est couramment pratiquée. Les packages ROCR [17] et caret [18] ont été utilisés pour ces évaluations. La courbe ROC est un examen du compromis entre la détection de vrais positifs et l'évitement de faux positifs. Elle ne peut être appliquée que sur des variables binaires (à deux modalités). Elle a donc seulement été utilisée avec le sexe de la victime. La zone située sous la courbe peut être quantifiée par la valeur AUC [14]. Plus cette valeur est proche de 1, meilleur est le modèle. Dans le cas de probabilités très déséquilibrées, la courbe ROC peut masquer un fort taux de mauvais classement. Dans ce cas, il vaut mieux utiliser la courbe précision/rappel qui ne tient pas compte des vrais négatifs (les hommes classés en « homme »). Une courbe ROC peut donc montrer des résultats satisfaisants pouvant être démentis par la courbe précision/rappel.

IV. PRINCIPAUX RÉSULTATS

A. Description des trois variables cibles et de la variable textuelle

Le fichier analysé contient 23 691 lignes. La répartition des trois variables cibles selon leurs modalités est décrite dans les tableaux 3, 4 et 5. La variable « sexe » présente le plus fort déséquilibre de ses modalités. Seule variable binaire sur les trois, elle permettra une évaluation visuelle du modèle probabiliste.

TABLEAU 3 RÉPARTITION DU SEXE DE LA VICTIME

Femme	**Homme**
6,7 %	93,3 %

TABLEAU 4 RÉPARTITION DE L'ÂGE DES VICTIMES

14-19 ans	**20-29 ans**	**30-39 ans**	**40-49 ans**	**50-59 ans**	**60-88 ans**
2,9 %	24,0 %	24,9 %	25,9 %	20,4 %	1,9 %

TABLEAU 5 RÉPARTITION DE LA GRAVITÉ DES LÉSIONS

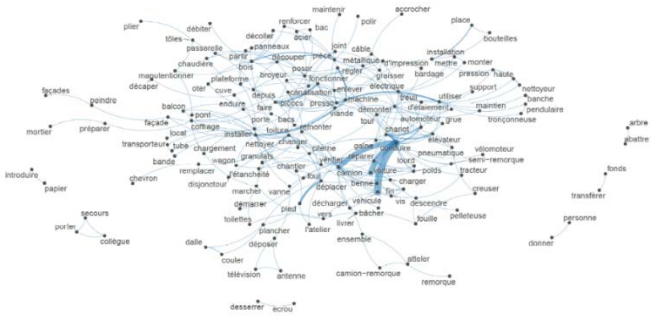
Décès	**Accident grave**	**Accident pas grave**
57,3 %	32,5 %	10,2 %

Pour la variable textuelle « activité », les mots de l'activité des hommes et des femmes sont représentés séparément dans les figures 1 et 2 par un réseau accidents x mots [15]. Pour plus de visibilité, seulement 1 000 accidents et les 150 mots les plus fréquents sont représentés. Les arcs en bleu relient un verbe et son complément mais peuvent également relier un autre verbe et/ou un autre complément en fonction de leurs cooccurrences. L'épaisseur d'un arc reflète la fréquence des liens.

13 *Receiver Operating Characteristic*
14 *Area Under Curve*
15 Réseaux réalisés avec quanteda

FIGURE 1 RÉSEAU DES MOTS DE L'ACTIVITÉ DES FEMMES (1 000 ACCIDENTS X 150 MOTS)

FIGURE 2 RÉSEAU DES MOTS DE L'ACTIVITÉ DES HOMMES (1 000 ACCIDENTS X 150 MOTS)



B. Catégorisation des variables par les mots

1) Le sexe de la victime

Les probabilités postérieures calculées par le modèle ont permis de dresser une liste de mots les plus associés aux hommes et aux femmes et d'en déduire leur activité. Par exemple, les quinze mots les plus associés aux hommes sont : « toiture », « béton », « échafaudage », « grue », « benne », « peindre », « camion », « décharger », « coffrage », « dépanner », « charpente », « tour », « remorque », « façade », « démonter ». Ces mots font clairement référence à des activités de chantier, de transport et de manutention. Les quinze mots les plus associés aux femmes sont : « aromatiques », « pâtes », « solvants », « javel », « sertir », « décollage », « cambrer », « personnel », « riveter », « repasser », « jambon », « administratifs », « phase », « sertisseuse », « reprendre ». Ces activités suggèrent diverses activités relatives au nettoyage, au secteur alimentaire, aux bureaux, aux usines.

Cette cartographie a été comparée à celles obtenues avec une caractérisation dans SPAD et aux mots saillants identifiés par quanteda. Les premiers mots des listes, les plus pertinents pour chaque méthode, sont généralement bien identifiés par les trois méthodes, d'autant mieux pour des modalités fréquentes (hommes). Des différences apparaissent dans la suite des listes. Certains mots identifiés dans SPAD ou quanteda ne sont pas identifiés par la classification bayésienne. Par exemple, les mots « presse », « utiliser », « voiture », « machine », « alimenter », « fonctionner », « pièces » sont associés aux femmes dans les méthodes SPAD et quanteda. Pour les hommes, ce sont les mots « chariot », « manutentionner », « régler », « charger », « place », « monter », « vérifier ». Ces mots non retenus par la modélisation peuvent être, de fait, assez communs aux deux sexes et avoir une tonalité plutôt générale.

2) L'âge de la victime

Les six classes d'âges sont décrites par des mots différents mais les associations ne sont pas aussi franches que pour le sexe. Les 14-19 ans sont associés à des activités en lien avec l'alimentation : « pâtisserie », « laminoir », « pâtes », « jambon », avec l'utilisation de deux-roues (« mobylette », « vélomoteur », « motocyclette »). Les chariots automoteurs, les plieuses ou les scies leur sont associés avec SPAD ou quanteda mais pas avec la modélisation. La modélisation bayésienne associe les 20-29 ans aux activités de manutention (« paquet », « ciment », « élévateur »), aux lignes de production (« meule », « mouler », « l'outil »), également aux deux-roues (« moto », « scooter »). SPAD et quanteda identifient également les presses, les chariots, les pièces ou le nettoyage. Les 30-39 ans sont associés à des activités de manutention (« soulever », « déménager », « planches », « plaques », « livraison »), également à des interventions (« vidanger », « joint », « bidon », « découpe », « traiter », « vanne »). Les mots « accompagner », « palette », « ridelles » ne sont pas retenus par la modélisation bayésienne mais le sont par SPAD ou quanteda. Les 40-49 ans sont associés à des activités de manutention ou de livraison (« remorque », « réceptionner », « chaîne », « accrocher », « brouette »), aux garages, à des activités de maintenance (« dépanner », « régler », « intervenir »), aux treuils et au ferraillage (SPAD et quanteda), aux hôtels (SPAD). Les 50-59 ans sont associés aux livraisons, aux travaux de bureau, aux déplacements (« accès », « rejoindre », « escalier », « arriver », « aller »). Les pauses et les reprises sont suggérées par les mots « café », « repas », « travail », « assis », « déjeuner ». La visite de client est également repérée. SPAD et quanteda identifient aussi clairement les déplacements à pied. Pour les plus de 60 ans, ce sont les déplacements (« piétons », « l'arrivée », « rentrer », « regagner ») qui sont signalés, ainsi que les lieux (« vestiaire », « réfectoire »). Les mots « traverser », « distribuer », « descendre », « alimentaires » sont de plus retenus par SPAD et quanteda.

3) La gravité des lésions

Les accidents mortels sont clairement liés aux accidents de la route, les accidents graves aux machines et les accidents « pas graves » à l'alimentation, à des activités administratives ou à l'utilisation de machine. Il y a une bonne homogénéité entre les trois méthodes pour les accidents mortels, catégorie la plus fréquente. Pour les accidents graves, la modélisation retient des types précis de machines comme les scies radiales, les diviseuses, les tenonneuses, et des verbes précis comme « plier », « extraire », « déligner ». Les mots saillants ou caractéristiques selon SPAD et quanteda sont plus généraux : « presse », « machine », « utiliser », « nettoyer », « pièce », « usiner », « scie », « régler », « plieuse ». Les mots associés aux accidents « pas graves » évoquent la chimie (« aromatiques », « solvants »), le nettoyage (« javel »), les bureaux (« administratifs »), l'alimentaire (« trancheur », « jambon », « pâtisserie », « sucre »), l'usinage (« poinçonneuse », « soudage », « conditionneuse », « rotative »). Les activités caractéristiques de SPAD ou saillantes pour quanteda

concernent l'utilisation de machine, les pièces, le réglage, le moulage, les scies, le fraisage.

C. Validation des modèles

1) Le sexe de la victime

L'observation des indicateurs[16] calculés à partir de la matrice de confusion (tableaux 6 et 7) montre un F-score moyen de 50,5 %, ce qui n'est pas un résultat très élevé. Le F-score des hommes (78,4%) est bien meilleur que celui des femmes (22,5%) ce qui traduit une meilleure capacité du modèle à identifier les hommes que les femmes à partir des mots à sa disposition. Ce meilleur résultat résulte notamment du fait qu'il y a beaucoup plus d'hommes que de femmes dans l'échantillon et que le modèle a plus de mots pour affiner son apprentissage.

TABLEAU 6 MATRICE DE CONFUSION POUR LA VARIABLE « SEXE »

Prédit / Observé	Femme		Homme		Total
	Nombre	*% total*	*Nombre*	*% total*	
Femme	347	**4,9**	154	2,2	501
Homme	2247	31,6	4360	**61,3**	6607
Total	2594		4514		7108

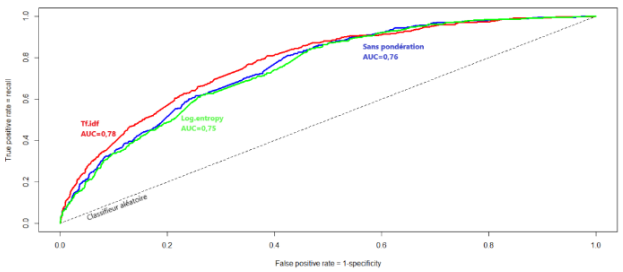
TABLEAU 7 INDICATEURS D'ÉVALUATION POUR LA VARIABLE « SEXE »

	PRÉCISION	RAPPEL	F-SCORE
Femme	13,4 %	69,3 %	22,5 %
Homme	96,6 %	66 %	78,4 %

L'exactitude du modèle (somme des % de la diagonale du tableau 6) est 66,2 %, le taux d'erreur de 33,8 % (1-exactitude). Une modélisation effectuée avec les fréquences hommes/femmes observées au lieu d'une équiprobabilité favorise la classification des hommes. Avec une exactitude plus forte (92,5 % au lieu de 66,2%), le F-score des femmes est néanmoins plus faible (12,4 % au lieu de 22,5 %). L'équiprobabilité a donc été privilégiée pour obtenir le plus d'information possible sur les femmes.

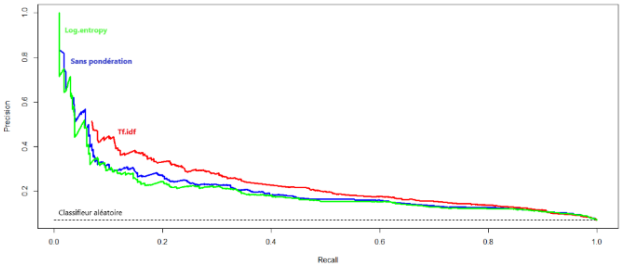
Le sexe de la victime étant une variable binaire, la courbe ROC a été calculée pour les trois choix correspondants aux différentes pondérations. (cf. figure 3).

FIGURE 3 PRÉDICTION DE LA VARIABLE « SEXE » – COURBE ROC (FREQUENCES DES SEXES 0,5/0,5)



Plus une courbe ROC se dessine du coin gauche/bas vers le coin gauche/haut puis vers le coin droit/haut, meilleur est le modèle. La courbe diagonale en pointillé correspond à un modèle qui classe au hasard les accidents en « homme » ou « femme ». La figure 3 ne montre pas de grandes différences entre les trois modèles. La pondération tf/idf produit le meilleur modèle, ce qui est également indiqué par une valeur plus élevée de l'AUC (0,78). La courbe ROC n'étant pas un bon indicateur quand les modalités sont très déséquilibrées, ce qui est le cas pour le sexe (les hommes représentent 93,6 % des accidents), les courbes précision/rappel ont été calculées et sont représentées figure 4.

FIGURE 4 PRÉDICTION DE LA VARIABLE « SEXE » – COURBE PRÉCISION/RAPPEL (FREQUENCES DES SEXES 0,5/0,5)



Le modèle le meilleur aurait une courbe se dessinant du coin gauche/haut vers le coin droit/haut puis vers le coin droit/bas, le moins bon étant proche de la situation aléatoire indiqué par la ligne pointillée. Ici encore, c'est la pondération tf/idf qui produit le meilleur modèle.

2) L'âge de la victime

Les tableaux 8 et 9 présentent les résultats de l'évaluation du modèle de l'âge en fonction des mots de l'activité de la victime. Le F-score moyen est de 18 %, valeur très faible. L'analyse des F-scores par classes d'âges montre les résultats les plus faibles pour les classes d'âges extrêmes, catégories ayant par ailleurs les plus faibles fréquences. Le F-score des classes d'âges intermédiaires oscillent entre 20 et 30 %.

TABLEAU 8 MATRICE DE CONFUSION POUR LA VARIABLE « AGE »

Prédit / Obs.	14-19 ans		20-29 ans		30-39 ans		40-49 ans		50-59 ans		60-88 ans		Total
	Nb	*% total*	*Nb*	*% total*	*Nb*	*% total*	*Nb*	*% total*	*Nb*	*% total*	*Nb*	*% total*	
14-20 a	56	**0,8**	54	0,8	31	0,4	23	0,3	16	0,2	20	0,3	200
20-29 a	298	4,2	480	**6,8**	339	4,8	213	3,0	203	2,9	157	2,2	1690
30-39 a	320	4,5	407	5,7	350	**4,9**	258	3,6	290	4,1	156	2,2	1781
40-49 a	320	4,5	368	5,2	362	5,1	285	**4,0**	328	4,6	185	2,6	1848
50-59 a	218	3,1	255	3,6	298	4,2	202	2,8	305	**4,3**	169	2,4	1447
60-88 a	31	0,4	23	0,3	24	0,3	16	0,2	26	0,4	22	**0,3**	142
Total a	1243		1587		1404		997		1168		709		7108

Exactitude moyenne = 21,1 % (somme des % de la diagonale)

16 Les indicateurs peuvent être exprimés en probabilité ou en pourcentage

TABLEAU 9 INDICATEURS D'ÉVALUATION POUR LA VARIABLE « AGE »

	PRÉCISION	**RAPPEL**	**F-SCORE**
14-20 ans	4,5 %	28 %	7,8 %
20-29 ans	30,2 %	28,4 %	29,3 %
30-39 ans	24,9 %	19,7 %	22,0 %
40-49 ans	28,6 %	15,4 %	20,0 %
50-59 ans	26,1 %	21,1 %	23,3 %
60-88 ans	3,1 %	15,5 %	5,2 %

L'âge des victimes n'étant pas une variable binaire, la courbe ROC ne peut pas être calculée.

3) La gravité des lésions

Les tableaux 10 et 11 récapitulent les différents indicateurs obtenus par la modélisation de la variable « gravité des lésions ». Le F-score moyen est de 50,4 %. Le modèle prédit le mieux les accidents mortels (F-score de 73,4%), classe ayant l'effectif le plus élevé. La courbe ROC ne peut pas non plus être calculée pour évaluer le modèle.

TABLEAU 10 MATRICE DE CONFUSION POUR LA VARIABLE « GRAVITÉ DES LÉSIONS »

Prédit / **Observé**	**Décès**		**Accident grave**		**Accident pas grave**		Total
	Nombre	*% total*	*Nombre*	*% total*	*Nombre*	*% total*	
Décès	2767	**38,9**	803	11,3	521	7,3	4091
Accident grave	548	7,7	1108	**15,6**	645	9,1	2301
Accident pas grave	127	1,8	277	3,9	312	**4,4**	716
Total	3442		2188		1478		7108

Exactitude = 59 %

TABLEAU 11 INDICATEURS D'ÉVALUATION POUR LA VARIABLE « GRAVITÉ DES LÉSIONS »

	PRÉCISION	**RAPPEL**	**F-SCORE**
Décès	80,4 %	67,6 %	73,4 %
Accident grave	50,6 %	48,2 %	49,4 %
Accident pas grave	21,1 %	43,6 %	28,4 %

V. DISCUSSION

Le classifieur naïf bayésien s'est révélé plus performant sur des variables ayant peu de modalités comme le sexe de la victime ou la gravité des lésions et pour les modalités ayant de fortes fréquences. Ce sont en effet pour les hommes et les accidents mortels que les indicateurs d'évaluation sont les meilleurs (F-scores respectivement de 78,4 et 73,4%). D'un point de vue qualitatif, les listes des mots les plus probables montrent que des modalités de plus faible effectif sont toutefois associées à certains mots, donc à certaines activités. Par exemple les femmes ont des activités liées à l'alimentation ou au nettoyage, les moins de 20 ans à l'utilisation de deux-roues (vélos, mobylette) alors que les 20-29 ans sont associés aux scooters ou aux motos. Il faut noter qu'il s'agit bien d'un critère différent d'une forte fréquence. En effet, la conduite d'une voiture très fréquente chez les femmes (cf. figure 1) n'est pas identifiée comme une activité très probable chez les femmes. Etant également très fréquente chez les hommes, cette activité ne départage pas les sexes lors de la modélisation bayésienne. Par contre, elle est associée aux femmes avec SPAD et quanteda. La classification bayésienne traitant les données de manière différente est donc à même d'apporter un autre type d'information.

La réglementation protège les jeunes salariés de moins de 18 ans en leur interdisant certains travaux dangereux comme les travaux de démolition, l'exposition à des risques d'effondrement dans les fouilles, les travaux en hauteur, les opérations sous tension électrique. La modélisation n'identifie, pour les moins de 20 ans, aucune de ces notions alors que les échafaudages sont associés aux 20-39 ans, les démolitions aux 30-39 ans et aux plus de 60 ans, le terme « électrique » aux 40-49 ans, les grues aux 30-49 ans, le décoffrage et les fouilles aux 50-59 ans. Malgré un score bas pour cette classe de jeunes travailleurs (7,8 %), on observe néanmoins que le classifieur « ne se trompe pas », ce qui amène à s'interroger sur la nature des liens existant entre l'interprétabilité d'un modèle et ses performances. Les interprétations des cartographies obtenues pourraient être affinées par la modélisation de variables croisées, par exemple le sexe et l'âge de la victime ou bien l'âge et la gravité des lésions, moyennant le risque d'un affaiblissement de la qualité des modèles dû à l'augmentation du nombre de modalités et la diminution corrélée de certaines fréquences.

Les données EPICEA analysées ici sont des données collectives stabilisées, de retour d'expérience sur des accidents survenus dans différentes entreprises. La description de l'activité de la victime est limitée à une courte phrase : les plus courtes comprennent deux ou trois mots (« passer l'aspirateur », « prendre un colis »), les plus longues entre 11 et 12 mots (« se reposer dans la couchette de la cabine du véhicule en circulation », « effectuer le tour du bus suite à une anomalie de frein »). Améliorer le modèle peut consister à augmenter le nombre d'individus du fichier d'apprentissage, par exemple en réactualisant l'analyse avec les nouveaux cas d'accidents enregistrés et, à l'occasion, en augmentant les catégories à fréquences faibles. L'amélioration peut également porter sur l'augmentation des mots du corpus d'apprentissage. Utiliser le récit de l'accident, variable texte également disponible dans la base de données, ne semble pas un choix pertinent car ce récit comprend plusieurs types d'information non homogènes concernant l'emploi, la nature de l'entreprise, les postes précédemment occupés, le processus de l'accident, la description des lésions, les causes identifiées, etc. Ces informations hétérogènes risquent de brouiller l'apprentissage automatique.

Au niveau d'une entreprise, des résultats équivalents de modélisation peuvent être resitués dans un contexte de poste de travail et complétés, par exemple, par des questionnaires aptes à fournir au modèle des données supplémentaires d'apprentissage.

Indépendamment des données de terrain, la qualité d'un modèle dépend d'un grand nombre de critères techniques : le choix des probabilités *a priori*, le type de pondération

appliquée au tableau individus x mots, le type de modélisation choisie (multinomiale ou Bernoulli). L'évaluation elle-même dépendra du choix des indicateurs (F-score, exactitude, rappel, etc.) ou des courbes (ROC ou précision/rappel). Les résultats obtenus avec différentes probabilités *a priori* et différents types de modélisation ne sont pas présentés ici mais leur interprétation ne s'est pas révélée fondamentalement différente.

Les résultats du classifieur bayésien ont été comparés à ceux de deux méthodes proposées par SPAD, avec des comparaisons de pourcentages, et quanteda, avec l'identification de mots saillants selon une mesure de chi deux. Les associations obtenues avec le classifieur semblent plus sélectives, identifiant par exemple des noms précis de machines (tenonneuse, diviseuse ou scie radiale) plutôt que des termes génériques comme « machine » ou « pièce » identifiés par SPAD et quanteda. Les mots retenus par ces deux dernières méthodes sont assez proches, bien que présentant elles-mêmes des différences non abordées dans cet article. Des comparaisons pourraient être faites avec d'autres méthodes d'apprentissage automatique comme les régressions logistiques, les arbres de décision et les forêts aléatoires comme présentées dans [19].

VI. CONCLUSION

Cette présentation décrit une application de *textmining* utilisant une classification naïve bayésienne. Le modèle obtenu par apprentissage automatique a servi à décrire trois variables structurées : deux variables d'identité l'âge et le sexe d'une victime d'accident, et la gravité des lésions, par une variable textuelle décrivant l'activité de la victime. L'évaluation des modèles indique une qualité assez moyenne des scores, la catégorisation des différentes catégories restant toutefois pertinente d'un point de vue qualitatif. Cette étude exploratoire montre l'intérêt d'analyser conjointement des données structurées et des données textuelles par la mise en évidence d'associations spécifiques. Les interprétations issues de telles modélisations permettent des éclairages nouveaux d'un jeu de données ou d'un sujet étudié et peuvent représenter une base d'orientation pour de futures études. Elles constituent toujours un enrichissement des connaissances.

L'objectif de ce travail consistait à appliquer une méthode de modélisation probabiliste dans le cadre du *textmining* tout en décrivant en détail les différentes étapes. De façon générale, les techniques d'apprentissage automatique et de modélisation probabiliste sont utilisées à des fins prédictives en appliquant les règles de classement apprises à de nouvelles données observées. Par exemple en médecine, pour prédire l'apparition d'une maladie en fonction de symptômes observés, en fiabilité pour prédire une panne en fonction de nouveaux événements enregistrés, en cybersécurité pour le tri de messages électroniques afin de bloquer d'éventuels spams. Ces méthodes s'appliquent aussi bien à des données textuelles que non textuelles. La composante prédictive de l'algorithme n'a pas été appliquée dans la présente étude. En effet, elle n'avait pas pour objectif de prédire le sexe ou l'âge d'une victime ou la gravité des lésions en fonction de nouvelles activités observées mais d'explorer un ensemble de données stabilisées à l'aide d'un modèle probabiliste. Ces travaux ont représenté l'opportunité d'explorer et approfondir des techniques bayésiennes de plus en plus mentionnées dans le monde scientifique sans être réellement explicitées d'un point de vue concret. L'approche bayésienne se différencie de l'approche fréquentiste dominante au XIXe siècle pour traiter les données industrielles, économiques ou administratives en pleine expansion à ce moment-là. L'approche bayésienne, pourtant antérieure à l'approche fréquentiste, a fait son retour au XXe siècle, facilitée par les ressources prodigieuses offertes par l'informatique, l'intelligence artificielle et leurs évolutions rapides. Dès les années 1980 elle a été utilisée dans le domaine de la maîtrise des risques et de la sûreté de fonctionnement pour exploiter ce qu'on appelle maintenant les *small data*.

Dans le domaine de l'analyse exploratoire de données industrielles ou d'accidentologie, le classifieur naïf bayésien est l'une des méthodes disponibles au même titre que les méthodes de classification, régression, arbres de décision, etc. Il présente néanmoins plusieurs avantages : sa simplicité d'utilisation, une quantification des résultats sous forme de probabilités, la rapidité des calculs même sur de très grands jeux de données et à l'inverse son application possible sur de très petits fichiers, la robustesse des modèles même si certaines hypothèses ne sont pas respectées. Toutefois, ces méthodes probabilistes nécessitent un investissement sur plusieurs plans : investissement méthodologique pour comprendre les fondements mathématiques des différents modèles et leurs spécificités d'application concrète, et pour choisir le modèle le plus approprié aux données à exploiter ; investissement prospectif pour trouver un logiciel implémentant le modèle, par exemple l'environnement de programmation R, libre et en pleine expansion depuis une dizaine d'années tant au niveau des packages proposés que du nombre d'utilisateurs. Ces investissements peuvent présenter un frein à la pratique de nouvelles méthodes. L'expérimentation de ces méthodes, la diffusion de résultats et de leur interprétation constituent un processus collaboratif de capitalisation d'expériences d'analyses de données, *small* ou *big data*, qui contribue à l'existence d'un fonds de connaissances en évolution permanente.

RÉFÉRENCES

[1] Lopez-Garcia J.R,, Mariscal Saldaña M.A., Garcia-Herrero S., Gutiérrez J.M. « Bayesian network analysis of the influence of labour market variables on accident rates of workers in Spain ». Risk, Reliability and Safety: Innovating Theory and Practice, 2016, https://www.researchgate.net/publication/313818622_Bayesian_network_analysis_of_the_influence_of_labour_market_variables_on_accident_rates_of_workers_in_Spain

[2] Chan A. P. C., Wong F. K. W., Hon C. K. H., Choi T. N. Y. « A Bayesian Network Model for Reducing Accident Rates of Electrical and Mechanical ». (E&M) Work. Int. J. Environ. Res. Public Health 2018, 15, 2496, https://www.mdpi.com/1660-4601/15/11/2496

[3] Base EPICEA et études réalisées consultables en ligne dans sa version limitée, http://www.inrs.fr/publications/bdd/epicea.html

[4] Tissot C., « EPICEA une base de données sur les accidents du travail au service de la prévention», Référence en santé au travail n°152, 2017, http://www.rst-sante-travail.fr/rst/dms/dmt/ArticleDMT/PratiquesMetiers/TI-RST-TM-43/tm43.pdf

[5] Tissot C. « Les accidents du tertiaire, des risques très divers pour une même activité de services », Hygiène et sécurité du travail, n°228, 2012, http://www.inrs.fr/accueil/dms/inrs/CataloguePapier/ND/TI-ND-2361/nd2361.pdf

[6] Tissot C., « Exploitation textuelle de données de retour d'expérience sous l'angle de la prévention des risque professionnels », Actes du congrès LambdaMu 20, 2016, DOI https://doi.org/10.4267/2042/61745,

[7] Jurafsky, D. & Martin, J.H. (2019). « From Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition ». Draft of October 16, 2019 (Chapter 4, Naive Bayes) https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

[8] Sprenger J. « Bayesianism vs. Frequentism in Statistical Inference », Tilburg Center for Logic and Philosophy of Science, TilburgUniversity, 2013, http://www.laeuferpaar.de/Papers/Bayes-vs-Freq-final-nodbl.pdf (traduction française : http://www.laeuferpaar.de/Papers/Sprenger_Bayes+Freq.pdf)

[9] Zhang H.,« The Optimality of Naive Bayes », Faculty of Computer Science, University of New BrunswickFredericton, New Brunswick, Canada, American Association for Artificial Intelligence (www.aaai.org), 2004, http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

[10] The Comprehensive R Archive Network https://cran.r-project.org

[11] Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). "quanteda: An R package for the quantitative analysis of textual data." Journal of Open Source Software, 3(30), 774. doi: 10.21105/joss.00774, https://quanteda.io

[12] Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Stefan Müller, Patrick O. Perry, Benjamin Lauderdale and William Lowe (2020). quanteda.textmodels: Scaling Models and Classifiers for Textual Data. R package version 0.9.1. https://CRAN.R-project.org/package=quanteda.textmodels

[13] Nakov P., Popova A., Mateev P., « Weight functions impact on LSA performance », Faculty of Mathematics and Informatics, Sofia University, Bulgaria https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwiy0qbDzO_oAhV_AGMBHeSBC5gQFjAAegQIBRAB&url=http%3A%2F%2Fciteseerx.ist.psu.edu%2Fviewdoc%2Fdownload%3Fdoi%3D10.1.1.69.9244%26rep%3Drep1%26type%3Dpdf&usg=AOvVaw04fpUWEj7noURBC3HWrGRx

[14] Ataa-Allah F., El Qadi A., Boulaknadel S., Aboutajdine D., « Calcul de similarité textuelle par le modèle d'Analyse de la Sémantique Latente », 2004, Conference: Institut des Nouvelles Technologies de l'Information et de la Communication At: Tanger, Morocco

https://www.researchgate.net/publication/260719435_Calcul_de_similarite_textuelle_par_le_modele_d%27Analyse_de_la_Semantique_Latente

[15] Fridolin Wild (2015). lsa: Latent Semantic Analysis. R package version 0.73.1. https://CRAN.R-project.org/package=lsa

[16] Coheris Analytics SPAD© - Système pour l'analyse des données – Versions 9, 1982-2017

[17] Sing T, Sander O, Beerenwinkel N and Lengauer T (2005). "ROCR: visualizing classifier performance in R." Bioinformatics, 21(20), pp. 7881. http://rocr.bioinf.mpi-sb.mpg.de.

[18] Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

[19] Al Mamlook R., Kwayu K.M., Alkasisbeh M.R., Frefer A.A., « Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity », Conference paper 2019, DOI: 10.1109/JEEIT.2019.8717393, https://www.researchgate.net/publication/333229225_Comparison_of_Machine_Learning_Algorithms_for_Predicting_Traffic_Accident_Severity