Classification automatique des faits techniques pour la conformité des lanceurs spatiaux

Michal Kurela, Mathilde Bacqué, Remi Laurent

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

Classification automatique des faits techniques pour la conformité des lanceurs spatiaux
Automatic classification of technical events for space launchers conformity

Michal Kurela
Centre Spatial Guyannais
Direction des Lanceurs, CNES
Kourou, France
michal.kurela@cnes.fr

Mathilde Bacqué
Université Toulouse Jean Jaurès
Toulouse, France
mathilde.bacque@etu.univ-tlse2.fr

Remi Laurent
Centre Spatial Guyannais
Direction des Lanceurs, CNES
Kourou, France
remi.laurent@cnes.fr

Résumé **- L'évaluation des conséquences des faits techniques (FT) permet au CNES d'assurer la conformité des lanceurs spatiaux vis-à-vis de la Loi sur les Operations Spatiales (LOS). Dans ce but la Classification Automatique de la gravité des risques des FT a été appliqué, testé et optimisé.**

Abstract **– CNES evaluate the consequences of the technical events to assure the conformity of space launchers with French Space Operations Act (FSOA). The Automatic Classification of risk severity associated to those events has been applied, tested and optimized.**

Mots clés — LOS, TAL, apprentissage supervisé, classification, gravité

I. INTRODUCTION

Dans le cadre de la Loi relative aux Operations Spatiales (LOS) [7] l'État subordonne sa garantie à une autorisation pour laquelle il s'assure en particulier que l'ensemble des risques liés à l'opération spatiale est maîtrisé par l'opérateur. Le traitement de flux des faits techniques (FT) est un des éléments constituant l'évaluation de la conformité technique des lanceurs spatiaux avec la LOS par le CNES pour le compte de l'État.

Cette démarche inclut l'évaluation des conséquences potentielles ou avérées des défaillances provoquées par ces FT. Les procédures strictes de production des lanceurs au Centre Spatial Guyanais (CSG) amènent les opérateurs à déclarer tout écart, même les plus minimes, via l'émission d'un Fait Technique. Ceci contribue à la grande qualité des lanceurs du CSG mais rend la tâche de contrôle des faits techniques à impact LOS beaucoup plus ardue. La majorité écrasante des FT est anodine, relevant de l'activité normale de l'opérateur du lancement, de ses industriels sous-traitants et ne concerne pas la LOS. Par contre leur diversité et la quantité importante (plusieurs centaines) dans le cycle de production des lanceurs et de vie d'une campagne de lancement conduit au besoin d'échantillonner et de diversifier la profondeur de l'analyse « manuelle » par une petite équipe d'ingénieurs du CNES. D'autre part l'objectif est de maximiser la détection des évènements potentiellement catastrophiques ou graves (décès et blessures irréversibles des personnes, destruction des biens) pour réduire leur probabilité d'occurrence. Le traitement automatique des langues (TAL) a été identifié comme un des outils d'optimisation de cette analyse.

II. DETECTION DES FAITS TECHNIQUES D'UN SYSTEME DE LANCEMENT AVEC LA GRAVITE LOS

Dans cette étude, l'amélioration de la détection des faits techniques de gravité « LOS » était notre objectif principal. Habituellement cette gravité est établie par l'étude des Faits Techniques soumis à la LOS par l'opérateur de lancement (Arianespace pour Ariane 5 par exemple). Les ingénieurs de la conformité LOS analysent les conséquences potentielles d'apparition de ce FT en établissant une chaine des évènements redoutés. Si ceux-ci correspondent à un des Evènements Redoutés (ER) finaux identifiés dans l' Etude des Dangers du système de lancement (EDD, exigée par article 7 du [7]) une action est prise auprès de l'opérateur de lancement afin de dédouaner ou éliminer un risque LOS potentiel. L'EDD spécifie les ER correspondant aux particularités d'un système de lancement dans ses différentes phases de fonctionnement.

La Fig. 1 présente une vision schématique de l'application de la LOS pendant différentes phases de vol d'un lanceur. Après décollage de son pas de tir, le lanceur sépare ses étages, la coiffe qui protège une ou plusieurs charge utile (CU) et enfin il sépare ses charges utiles. Pendant ces phases, pour les éléments prévus de retomber, l'opérateur de lancement doit garantir la maitrise des risques liés à la retombée du lanceur et de ces fragments telle que spécifiée dans l'article 23 du [7]. Conformément à l'article 27 l'équipe des ingénieurs de la « Sauvegarde vol » du CNES suit l'évolution du lanceur pendant toute sa phase de visibilité puis un algorithme embarqué prend éventuellement le relais via un diagnostic interne de bonne santé du lanceur (article 18 de [7]). L'objectif étant d'optimiser les capacités de neutralisation en cas de lanceur devenu dangereux.

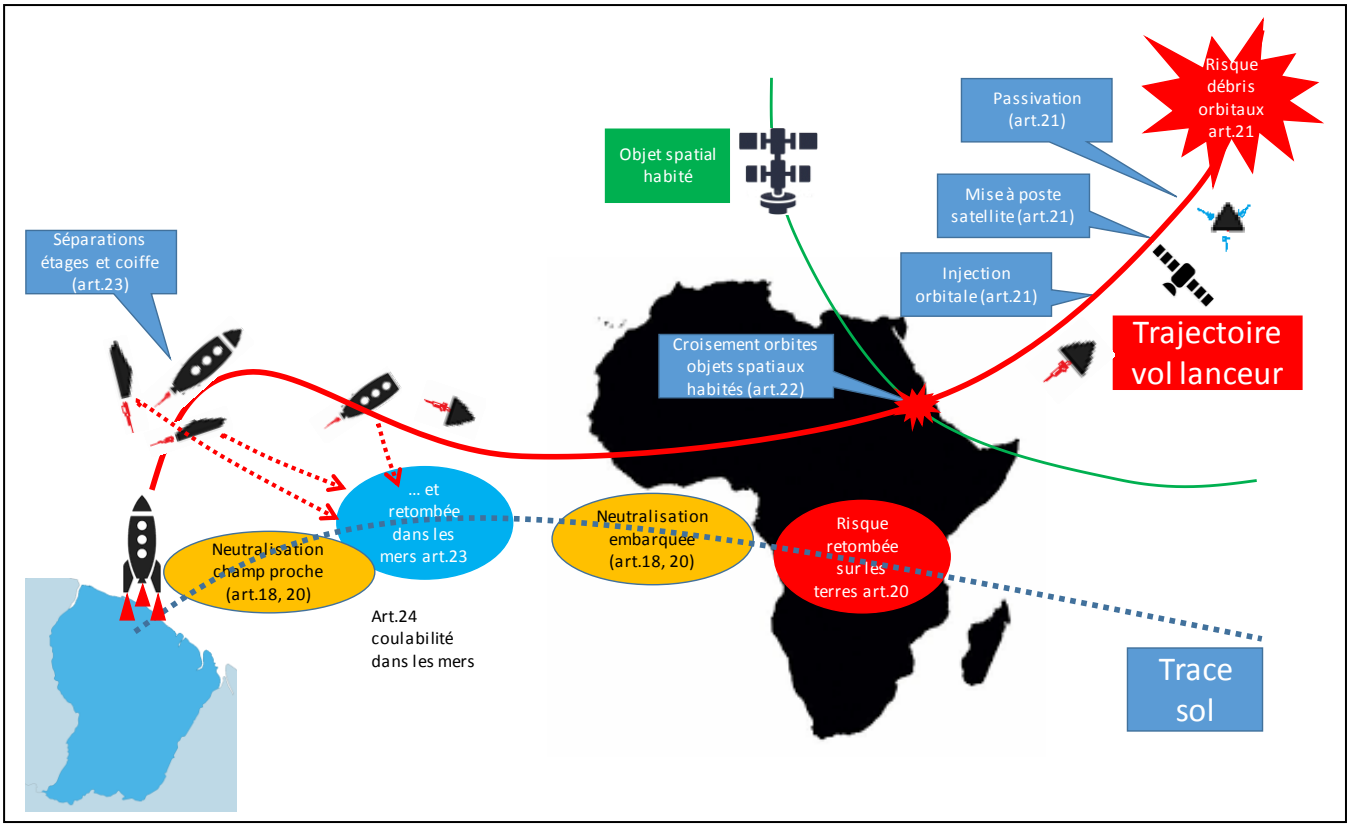

Fig. 1. logique d'application de la LOS art.7 du [7] dans une Etude des Dangers d'un système de lancement en Guyane

Pour les autres phases de vie du lanceur, d'autres exigences qualitatives sont à respecter (e.g. la coulabilité dans les mers des fragments des retombées selon art. 24, la passivation des sources d'énergie et la désorbitation selon art.21 concernant les risques des débris spatiaux) et quantitatives (e.g. art. 20 pour l'exigence quantitative vis-à-vis du risque au lancement et à la rentrée et l'art.21 pour les exigences relatives aux débris spatiaux).

Les Faits techniques pouvant conduire à des risques à impact LOS qui auraient été observés en vol et non détectés préalablement au lancement sont analysés lors de l'exploitation des paramètres post-lancement. L'objectif final étant de garantir la maîtrise des risques en prenant en compte le retour d'expérience pour minimiser la probabilité d'occurrence ou la gravité des Faits Techniques observés en vol.

Le nombre de faits techniques, même nominaux, dans la vie de production d'un lanceur est très important entre les modifications implémentées pour différentes raisons (optimisation des couts, de la sureté de fonctionnement ou de la performance), les impacts de l'exploitation des lancements précédents lors de leur analyse en niveau 0 (art. 19 de [7]), les particularités de l'analyse de mission (art. 17 de [7]) et enfin les anomalies ou non qualités détectés en production et en campagne de lancement. La grande majorité de ces FT est anodine du point de vue de la LOS. Ceci rend leurs détections difficiles. Il est donc nécessaire d'effectuer une sélection des Faits Techniques à analyser afin d'échantillonner le contrôle approfondi de l'impact de ceux-ci sur les risques LOS identifiés dans les études de danger. Typiquement tous les faits techniques adressant la chaine de neutralisation du lanceur sont analysés comme décrit dans l'article 27 de [7] via le contrôle du respect des termes de la réglementation de l'exploitation des installations du CSG [9]). D'autres FT sont échantillonnés pour l'analyse approfondie sur la base du retour d'expérience des causes et effets des FT passés (échecs de lancement concurrents, points critiques du développement des lanceurs, autres faits techniques saillants passées) et selon l'expérience technique des ingénieurs conformité LOS.

Les faits techniques sont disponibles en format informatique (fichiers PDF, Excel, Word) qui peut être traité en utilisant des algorithmes et méthodes du Traitement Automatique des Langues, notamment pour la classification automatique en gravité du risque.

III. ETAT DE L'ART

La méthode étudiée fait suite aux travaux [1], qui ont permis de démontrer l'utilité des méthodes du TAL (clustering par l'utilisation de la distribution latente du Dirichlet, recherche des similarités par l'application des méthodes TF-IDF et LSA) pour le retour d'expérience Sûreté de Fonctionnement (SdF) des lanceurs, notamment en appliquant les prétraitements pour simplifier la vectorisation du texte constituant les FT.

Les études telles que [2] et [4] ont inspiré l'évaluation d'une méthode plus particulièrement adaptée aux besoins de la LOS : le classement automatique des métadonnées. Il peut être défini par rapport à sa capacité de détection de la vraie classe d'une observation (dans notre cas des FT) :

TABLE I. CLASSEMENT AUTOMATIQUE PAR RAPPORT À L'ÉTAT « VRAI » D'UNE OBSERVATION

	classe prédite par le modèle	autre classe prédite par le modèle
la vraie classe	Vrai positif (VP)	Faux négatif (FN)
une autre classe	Faux positif (FP)	Vrai négatif (VN)

Sur cette base nous avons étudié des taux suivants de la classification automatique :

- Justesse : taux des observations correctement prévues sur le total de toutes les observations = (VP+VN) / (VP+VN+FP+FN). Sa valeur globale ne discrimine pas une grande asymétrie dans le nombre des observations par classe car une ou plusieurs classes statistiquement prépondérantes peuvent « écraser » les autres classes moins nombreuses (i.e. modèle détectant très bien une classe « commune » et très mal une classe « rare »).
- Précision : le taux de bon classement des observations dans la classe « vraie » par rapport à tous les classements dans cette classe par le modèle = VP / (VP+FP). Cette mesure traduit la « performance » du classement, car un grand nombre de FT « faux positifs » en terme de gravité LOS consommerait le temps d'analyse des experts en charge du contrôle de conformité afin de valider les logiques de dédouanement mises en place par l'opérateur de lancement.
- Rappel : taux des bons classements des observations dans la classe « vraie » par rapport à toutes les observations vraies = VP / (VP+FN) traduisant le taux de couverture du classement. Les « faux négatifs » traduisent donc les faits techniques non identifiés comme graves, alors qu'ils pourraient être graves. On peut retrouver ici notamment des « cas d'espèces », des cas de FT initialement jugés comme anodins dans le cadre de campagnes précédentes et qui sont devenus graves suite à un changement de contexte d'un FT.
- Score F1 : moyenne pesée des rappels et précisions associées à une classe ou l'intégralité du modèle = 2*(Rappel * Précision) / (Rappel + Précision)

Pour les besoins de la conformité LOS le taux de rappel est le plus significatif (augmenter la détection des vrais risques graves), bien que sa maximisation à 100% conduirait à devoir tout analyser de telle manière que tous les risques soient connus sans limites. Ceci n'est pas réalisable, et il s'agit donc de faire des compromis méthodologiques qui permettent de maximiser cette métrique avec un effort raisonnable, ce qui revient à optimiser le score F1 et la précision de la classe de gravité LOS avec un taux de rappel maximal. Enfin la justesse peut être utilisée pour une évaluation globale du modèle, ce que nous avons fait dans les analyses de sensibilité. Cette dernière valeur est à confronter avec la distribution des scores F1 et rappels des classes évaluées.

IV. BASES DE DONNEES DES FT TRAITES

Pour nos expérimentations avec la classification automatique nous avons utilisé les 3 BDD suivantes :

La BDD des 2k (2200) FT de référence a été établie au cours des 2 dernières années par 3 ingénieurs conformité avec un référentiel précis et l'importance fondamentale de la classification du risque LOS. Elle concerne uniquement des FT à impact sur le lanceur.

La BDD de 8k (~8000) FT du lanceur sur un périmètre technique partiel et diversifié par des périodes (nombreuses personnes dans un cadre organisationnel et industriel changeant dans le temps) classé en gravité avec l'intention de maitriser la disponibilité du système de lancement principalement.

La BDD de 42k (~42000) FT du segment sol a bénéficié du travail de nombreuses personnes pendant plusieurs années afin de réaliser une classification cohérente et sans trop insister sur l'importance de cette classification (subjectivité) vis-à-vis des risques LOS. La classification en gravité adressait avant tout les risques de l'absence de disponibilité des moyens sol pour les campagnes de lancement. Elle traite un ensemble plus large de faits techniques, souvent très anodins de l'environnement des ensembles de lancement (par exemple « fontaine d'eau non alimentée au bâtiment d'assemblage du lanceur »).

Toutes ces bases de données contiennent les mêmes champs de données textuelles, le classement en gravité « C3 » décrit ci-dessous et certaines métadonnées permettant de faire des tests supplémentaires de classement. Nous avons testé également d'autres sous-ensembles de FT avec biais (par exemple une BDD avec les FT mixte en anglais et en français) qui sont mentionnés ci-après.

V. GRAVITE DU FAIT TECHNIQUE EN TANT QUE METADONNEE DE CLASSIFICATION

Dans le processus de conformité, les FT sont classés en gravité du risque. Elle se divise en plusieurs niveaux définis par le créateur de la BDD des FT. La LOS ne définit pas directement des catégories de risques, uniquement les événements redoutés des systèmes de lancement et le type de dommage (art.1 et 7 de [7]), typiquement le « Dommage catastrophique » : perte de vie humaine, immédiate ou différée, ou blessures graves aux personnes (lésions corporelles, autres atteintes à la santé, invalidité ou maladie professionnelle, permanente ou temporaire) et le « Débris spatial » : tout objet spatial non fonctionnel d'origine humaine, y compris des fragments et des éléments de celui-ci, en orbite terrestre ou rentrant dans l'atmosphère terrestre. La LOS demande à l'opérateur de lancement de mettre en place des moyens de « protections des personnes, des biens, de la santé publique et de l'environnement » (art.7 [7]). A partir de ces prescriptions et de la pratique industrielle du secteur des lanceurs européens il est possible de créer plusieurs classifications plus ou moins détaillées :

TABLE II. GRILLES DE LA CLASSIFICATION EN GRAVITE

C1	C2	C3	C4
G0A catastrophique	G0A	LOS	LOS
G0B débris spatiaux	G0B		
G0B environnement			
G0B destruction moyens sol			
G1 perte de mission	G1		
C client	C	C	Non LOS
G2 mission dégradée			
G2 disponibilité lancement	G2	G2	
G3 autres	G3	G3	

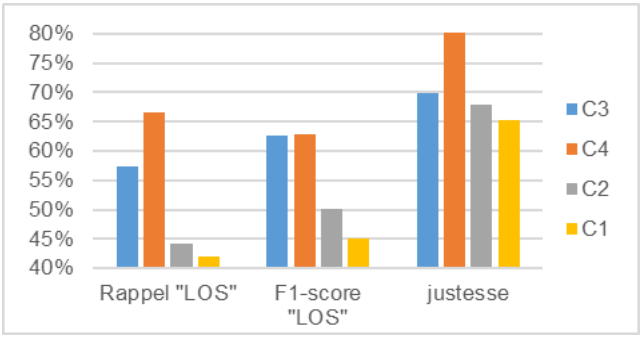

Fig. 2. comparaison des taux BDD 2k selon système de gravité

Les ingénieurs conformité LOS évaluent la gravité des FT en regard de la grille « C1 » en justifiant chaque gravité avec une chaine descriptive des causes et effets menant aux évènements redoutés. Il s'agit ici de la gravité de risque « a priori » correspondant à une perception de conséquences potentielles sur la base de l'état qualifié connu du système de lancement et de l'énoncé initial d'un FT avant son traitement. Ce dernier est traité pendant la campagne de lancement entre l'équipe de conformité LOS et l'opérateur de lancement afin de consolider la logique de dédouanement des évènements redoutés identifiés (actions de réduction de risque par mise en place des barrières matérielles, logicielles, organisationnelles, analyse détaillée permettant de démontrer l'acceptabilité du risque résiduel identifié initialement ou, le cas échéant, les justifications probabilistes permettant de démontrer la conformité avec les objectifs quantitatifs de la LOS, par ex. articles 20 et 21 [7]). A la fin de la phase de préparation de la campagne de lancement lors de la revue finale avant lancement l'ensemble des risques résiduels LOS doit être maîtrisé et rendu acceptable avant l'émission de l'avis de conformité CNES vers le ministère chargé de l'espace.

Comme mentionné dans §III l'impact de la distribution par classe d'une métadonnée sur les résultats de sa classification automatique est très important (homogénéité dans le classement : mêmes critères, mêmes personnes classifiant le corpus d'entrainement). Les classes avec très peu d'occurrences auront tendance à ne pas être bien détectées et l'entrainement du classificateur automatique avec les classifications « C1 » et « C2 » souvent ne convergeront pas. Typiquement il n'y a pas beaucoup de faits techniques à impact directement catastrophique. Par contre les risques de perte de mission du lanceur « G1 » sont très souvent porteurs du risque potentiel LOS. D'un autre coté la classification « C4 » est trop générale. Ceci nous a conduit à utiliser la classification « C3 » pour la gravité en « classification automatique » proposant le meilleur compromis entre le besoin de décrire la conséquence finale d'un FT et les scores de précision et rappel constatés.

Hormis les FT à impact « LOS », des FT avec d'autres niveaux de gravité peuvent intéresser les ingénieurs conformité par échantillonnage. Typiquement les risques du type « G2 disponibilité » peuvent potentiellement indiquer un risque LOS potentiel, si leur scénario d'occurrence est analysé dans un contexte diffèrent (par exemple la panne d'origine du FT révélé dans le segment sol transposé sur un système équivalent coté lanceur). Par contre les FT avec la gravité « G3 » ne font généralement pas l'objet d'une investigation plus approfondie relevant la plupart du temps de problème de planification, de coût ou d'autres sujets importants pour les industriels du système de lancement, mais sans impact sur le périmètre de la LOS.

VI. NETTOYAGE DES DONNEES D'ENTRAINEMENT

La préparation des données et le prétraitement joue un rôle prépondérant dans le rendu de la classification automatique. D'une part, de nombreux tokens (mots, expressions) identifiés dans les textes analysés font partie du « signal de bruit » et d'autre part l'information superflue consomme l'espace de mémoire de l'ordinateur et allonge son temps de traitement. Ceci a une importance, si les textes traités sont nombreux et relativement longs. Par exemple, la BDD de 2k FT était constituée d'un tableau de textes divisés en plusieurs données de type « plain text » correspondants à la description du fait technique, ses causes, effets et actions planifiées et exécutées du traitement. Un fait technique moyen contient en moyenne 725 signes correspondant à 130-180 mots.

Dans le cas de nos bases de données (BDD) d'entrainement il s'agissait de supprimer définitivement pour chaque FT les expressions trop présentes et non significatives. Par exemple le texte « fait technique clos » a été supprimé dans une colonne décrivant les dispositions de qualité car d'une part cette expression n'apportait aucune information pertinente pour l'évaluation de la gravité et d'autre part les FT de la base d'entrainement étaient tous résolus. Un autre cas semblable était une mention « à déterminer » / « AD » dans une colonne décrivant la cause de fait technique. Des signes des caractères sans utilité ont été également supprimés (typiquement signes comme « * », « / » ou d'autres placés dans les cellules vides par les personnes, qui remplissait la base de donnée à l'origine).

Des gains significatifs peuvent ainsi être faits sur la taille de la BDD à traiter et donc impacter positivement le nombre des tokens significatifs et la durée du traitement.

Dans un souci de cohérence il est également important de filtrer les textes écrits dans des langues étrangères, qui peuvent polluer la modélisation de classification. Des tests conduits sur une BDD de 1k FT à moitié en anglais et à moitié en français ont conduit à des résultats médiocres en terme de justesse et une forte instabilité des scores F1 et des taux de rappel.

VII. PRETRAITEMENT DE LA BASE D'ENTRAINEMENT ET DES REQUETES POTENTIELLES

Pour appliquer un modèle mathématique de classification automatique, les textes de la BDD doivent être mis en forme matricielle compatible avec ces modèles. Un jeu de prétraitements habituels du TAL est appliqué sur le corpus à chaque entrainement, tels que le remplacement des sigles par des formes développées, de la tokenisation, de la lemmatisation et de la vectorisation (cf. [1],[4]). Cette dernière a été soumise à une étude de sensibilité (validation croisée de la BDD 2k) en variant le nombre des n-grams, de niveau des filtres des expressions trop rares et trop communes pour choisir une configuration la plus intéressante possible pour maximiser le F1 score obtenu et limiter le nombre de tokens pour ne pas trop pénaliser la durée du traitement.

Une particularité de notre prétraitement est le remplacement des valeurs numériques par les noms significatifs des unités physiques associées en utilisant des

expressions régulières (regex) [10]. Les expressions régulières sont une notation symbolique permettant de modéliser le contenu d'une chaine de texte. Ce type de notation est utilisé notamment dans le langage de programmation Python. Par exemple le texte « 1,2mA » peut être remplacé par « courant » suivant la détection dans le texte avec l'expression régulière de type « \b\d+[.,]?\d*mA\b » pour détecter toutes les valeurs chiffrées associées à l'unité physique de microampère. Cette expression détectera « 0,1mA », « 1 mA », « 1mA » ou « 0,0005mA », mais elle ne détectera pas des « mots » comme « m0,1mA », « 0,1mAbcd », « 0/1mA ». La raison de cette détection est de simplifier le texte en terme des mots différents présents. Si la notion du courant électrique symbolisé par la présence de « mA » est très présente dans le texte ceci signifie qu'elle est importante pour ce texte. Plusieurs indications du genre « 0,2mA », « 5mA » et « 15 mA » sans prétraitement proposé seront vues comme 4 « mots » différents (donc très rares, car présents une seule fois) et avec le prétraitement proposé ils seront considérés comme le même mot présent 4 fois dans le texte.

Le même genre de traitement peut s'appliquer au préfixe comme par exemple en détectant la référence d'un lancement comme par exemple « VA250 » étant la 250eme campagne de lancement d'Ariane 5 pour le remplacer par le mot « lancement ». Sans ce prétraitement les « mots » résultants sont identifiés comme des mots uniques avec très peu de fréquence dans l'ensemble du corpus d'entrainement.

Ces modifications améliorent unitairement à l'échelle de +0,1% la justesse du modèle de classification automatique. L'amélioration peut être significative en identifiant de nombreuses règles de remplacement dans les textes permettant de réduire le nombre de tokens concernant le même concept ou objet. Néanmoins il faut faire attention à ne pas appliquer de règles de remplacement par des mots ambiguës (par exemple l'abréviation « PTE » signifie « protection thermique étage » et « production technical event » : le faire remplacer par une des versions de ce sigle conduira à la dégradation de la classification automatique). La définition des dictionnaires demande donc un nombre de tests importants avec la même BDD d'entrainement et le même modèle pour caractériser l'impact sur la classification automatique de chaque nouvelle règle introduite dans le prétraitement. Cette démarche représente un risque en cas de modification de la BDD d'entrainement (ajout de nouveaux textes) qui peut invalider l'apport positif de certaines règles identifiées précédemment. En ajoutant trop des règles il y a également un risque de sur-ajustement du modèle par rapport à de nouvelles données à classifier. D'autre part la construction et validation des très nombreuses règles sont consommatrices du temps.

VIII. CHOIX DU MODELE DE CLASSIFICATION AUTOMATIQUE

Nous avons testé de nombreux modèles accessibles via la bibliothèque Python « scikit learn » [5] pour caractériser leur justesse, écart type et le temps de traitement. Les paramètres du modèle retenu peuvent être optimisés par des méthodes mathématiques, plus formelles, tel que présenté dans [6],[8] par exemple. Pour nos besoins opérationnels nous avons déroulé une étude de sensibilité multicritère plus sommaire par l'itération des paramètres des modèles avec la validation croisée sous 10 plis sur la même base de données 2k et avec l'application du même prétraitement. Dépendant de leur paramétrage, certains modèles peuvent donner des résultats avec des temps de traitement très longs. L'exemple du modèle gaussien avec un kernel rbf a convergé au bout d'une demi-heure alors que la régression logistique liblinéaire en 1 seconde. D'un autre côté, des essais de sensibilité sur les réseaux de neurones ont montré des temps de traitement très variables sur une échelle de 3s à 400s selon le choix du nombre de neurones et de couches ou autres paramètres (avec une justesse bien inférieure à la régression logistique).

Ensuite nous avons retenu des modèles avec des durées de traitement plus faibles (inférieurs à 10s) pour vérifier la stabilité des scores F1 de l'ensemble des gravités non anodines (notamment de la gravité « LOS ») sur 2500 entrainements de 10% des FT tirés d'une manière aléatoire dans la BDD 2k.

TABLE III. EXEMPLES COMPARAISON DES MODELES D'ENTRAINEMENT SUR LA BDD 2K

Modèle	Justesse Moyenne (%)	Justesse Ecart type (%)	Temps calcul (s)	F1 gravité "LOS " (%)	Ecart type F1 "LOS " (%)	F1 hors gravité "G3" (%)
LR (liblinear, balanced class weight)	70	2	1,19	62,6	4,4	71
SVM sigmoid	70	3	130,15			
Ridge Classifier	70	4	5,91	63,7	4,3	72
SVM linear	69	3	1,32	63,1	4,2	71
SVM rbf	69	4	221,55			
Random forest	68	3	39,71			
K Neighbors Classifier	67	4	27,05			
Reseaux neurones lbfgs (100 neurones)	66	3	107,20			
BernoulliNB	66	5	1,25	60,3	4,6	69
Passive Agressive Classifier	66	5	14,05			
Perceptron	65	4	6,16	61,2	4,3	68
SVM poly	59	4	229,56			
GaussianNB	56	4	2,26	52,5	4,4	58
Decision tree	53	5	34,91			
Linear Discriminant Analysis	33	6	114,35			
Quadratic Discriminant Analysis	28	3	29,43			

La classification en gravité LOS a un score F1 inférieur aux autres gravités dans tous les modèles examinés alors que la gravite la plus nombreuse « G3 » avait tendance à avoir le meilleur score. Vu que la gravité anodine « G3 » n'est pas du tout la priorité de notre démarche nous avons calculé des scores F1 pondérés pour toutes les autres gravités hormis celle-ci.

Le Ridge Classifier obtient le meilleur résultat en terme des scores F1 pour la gravité LOS et l'ensemble des gravités non « G3 ». Néanmoins pour la suite des investigations pour

cet article, nous avons sélectionné la régression logistique avec équilibrage des classes. Les résultats obtenus pour ce modèle sont relativement positifs et le temps de traitement est le plus court obtenu sur un poste bureautique. L'optimisation plus fine du choix de modèle fait partie des perspectives pour améliorer l'efficacité de la classification automatique.

IX. Impact du nombre et de la qualite des donnees d'entrainement sur la qualite de la classification automatique

Une première impression peut conduire à considérer qu'en augmentant le nombre de données dans une BDD d'entrainement, le score de classification devrait s'améliorer. Dans le principe c'est vrai, si les critères de classement sont stables.

Des tests sur les classes de gravité dans les BDD 8k et 42k exploitées depuis plus que 20 ans, ont révélé l'influence du temps, des changements d'organisation et des personnes qui classifient : le sens de classification est modifié, de nouvelles valeurs d'une métadonnée ont été introduites dans la BDD, d'autres ont été abandonnées menant à une incohérence de classement (typiquement l'apparition d'une nouvelle classe qui décrit le même concept que les classes préexistantes) et donc la diminution de taux de rappel et de la précision. Typiquement la BDD 42k utilise beaucoup la gravité « G1 perte de mission » par rapport à la gravité anodine « G3 », alors qu'elle adresse le risque de panne au sol : ceci déséquilibre complètement le sens de classement en fournissant une information biaisée. Ces BDD n'ont jamais fait l'objet d'une relecture a posteriori pour uniformiser la définition des gravités de risque contrairement à la BDD 2k. Elles prenaient en temps réel des faits techniques telles que constatés au moment de leur occurrence.

Si la manière de classifier a changé dans le temps ou si le critère de classification est trop général ou encore trop subjectif l'augmentation du nombre de données ne peut pas améliorer la précision. Au contraire, celle-ci peut être dégradée avec les cas contradictoires présents dans le corpus. La durée de traitement et le besoin en mémoire augmenteront également.

Pour vérifier cela, on peut comparer la BDD 2k de référence avec les résultats obtenus en classification automatique de gravité des risques dans d'autres BDD plus grandes (validation croisée de la justesse avec 10 plis, classification des FT de la BDD 2k avec d'autres BDD en tant que bases d'entrainement pour les scores F1) :

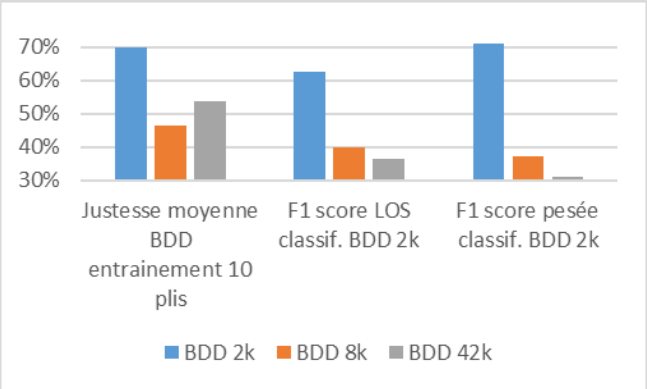

Fig. 3. Justesse des BDD FT testés, comparaison de classification de la BDD 2k par d'autres BDD

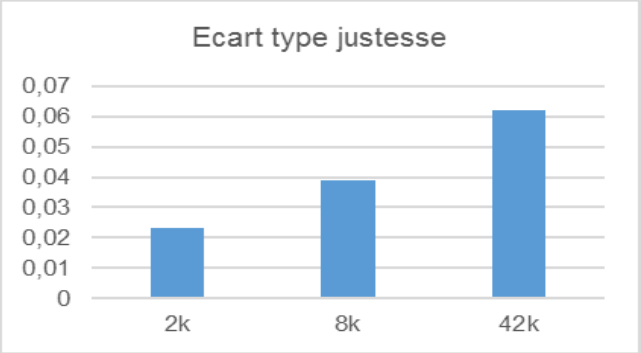

Fig. 4. Ecart type de la justesse des BDD FT testés, comparaison de classification de la BDD 2k par d'autres BDD

On peut voir une différence significative de la justesse obtenue dans les trois cas, malgré le même prétraitement appliqué et le même périmètre de description (métadonnées). Il y a également une forte différence sur la dispersion de la justesse. Par ailleurs un test de classification des FT du BDD 2k sur la base d'entrainement constituée des autres BDD (8k, 42k) avait donné des F1 scores très inférieurs aux tests de la BDD 2k elle-même (2500 entrainements sur 10% de corpus de test).

Vu que le classement en gravité peut être subjectif (sous-pondéré ou surpondéré vis-à-vis de l'évènement redouté), une vérification supplémentaire est possible en testant la classification automatique par un critère plus objectif, comme par exemple la « spécialité technique » concernée par le fait technique avec deux possibilités : mécanique/fluide ou électro/pyrotechnique. Dans les deux cas le vocabulaire est bien distinct et le risque d'erreur humaine en classification initiale est moindre que pour la gravité du risque. L'entrainement sur cette métadonnée pour toutes les BDD testés ont conduit à un résultat très cohérent de 85-88% du score F1 et de taux de rappel.

L'augmentation du nombre des faits techniques dans le corpus d'entrainement augmente en principe le vocabulaire particulier correspondant à un classement spécifique et diminue l'écart type associé à la justesse de la classification automatique en supposant que le classement manuel de la BDD d'entrainement suit les mêmes règles. Une étude de sensibilité permet de visualiser l'évolution de la justesse et des scores F1 et leur écart type selon nombre des FT dans la BDD d'entrainement (300 tirages pour chaque point en supprimant 100 FT en boucle à chaque fois).

Comme nos BDD sont organisées en ordre chronologique nous avons testé cette hypothèse sur une BDD 1,3k FT étant un sous ensemble de la BDD 2k de référence. Par rapport à la BDD2k nous avons supprimé des FT antérieurs à 2 ans pour avoir un corpus parfaitement équilibré contenant uniquement des campagnes de lancement entières : pas de biais provenant de FT antérieurs concernant des campagnes parfois très éloignées dans le temps, ajoutées à posteriori pour enrichir la BDD d'entrainement. De cette manière on peut mesurer à posteriori l'effet d'ajout des faits techniques d'une nouvelle campagne à la base d'entrainement.

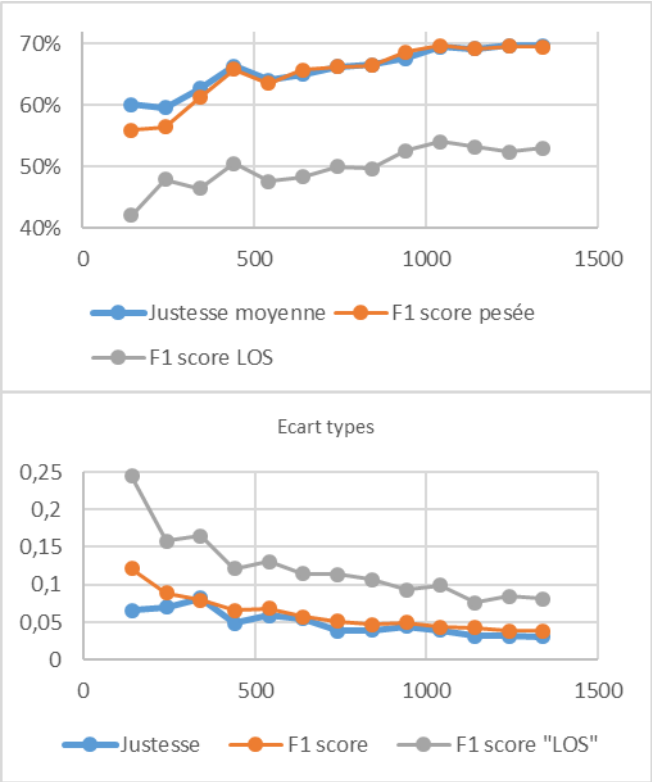

Fig. 5. Sensibilité des métriques de classification automatique de la BDD 1,3k au nombre des données d'entrainement

Après une zone instable avec peu d'éléments d'entrainement, la justesse converge vers une valeur moyenne associée à une dispersion inférieure à 5% pour à peu près +1000 FT. En terme de score de rappel, avec la croissance de la BDD, la classe « G3 » la plus nombreuse semble être pénalisée à un niveau très limité. En conséquence l'enrichissement de la BDD avec les campagnes à venir devrait continuer à améliorer le rappel et le F1 score de l'ensemble du modèle moyennant le maintien de la rigueur dans le classement en gravité des FT introduits dans la BDD d'entrainement.

Les métriques de la BDD 2k complète prenant en compte une sélection de FT passés et classifiés selon les mêmes critères, sont encore meilleures. Pour améliorer encore plus les résultats de notre BDD, nous avons testé l'introduction de nos données de retour d'expérience non issues de l'exploitation d'Ariane 5 : un sous ensemble de 160 points critiques de développement d'Ariane 5 (en détail décrits dans [1]) et 560 retours d'expérience concernant les échecs et les reports de lancements concurrents dans le monde issus de la veille technologique. Ces deux ensembles ont été classés en gravité et justifiés par les responsables de la Sûreté de Fonctionnement de la Direction des Lanceurs du CNES d'une manière cohérente avec la classification des ingénieurs conformité LOS. La particularité concerne leur classement en gravité, généralement « LOS » ou encore « G2 », qui modifie donc la distribution globale en gravité de la base d'entrainement en diminuant le poids des faits techniques « C » et « G3 ». En passant de 2200 à 3000 FT le classement automatique a été significativement amélioré en s'approchant de F1 score de 75% pour la gravité LOS, G2 et la moyenne pesée de toutes gravités sans trop dégrader les scores de la gravité « G3 » ni « C ».

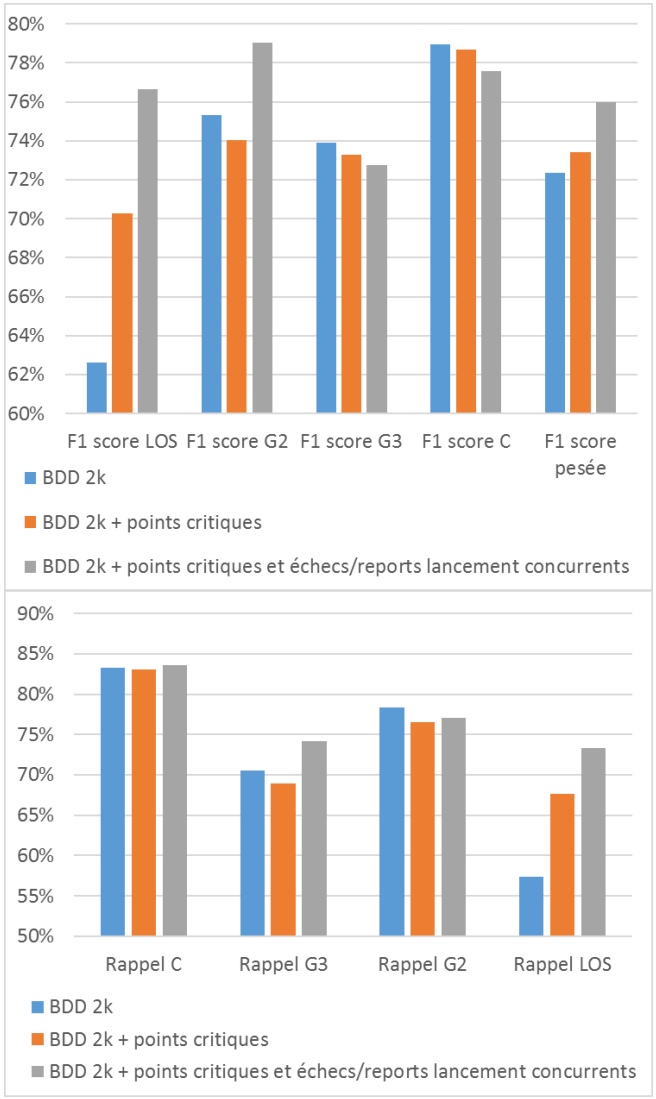

Fig. 6. effet d'ajout des données du REX dans la BDD 2k sur les scores F1 et rappel des différentes classes de la gravité

Cette démarche montre qu'il est intéressant d'enrichir une BDD d'entrainement avec des données venant d'autres sources que l'objet direct de la BDD. Néanmoins il est primordial de maitriser les critères de classement pour assurer la cohérence du classificateur automatique.

Un test similaire avec la base de 42k cas conduit à un résultat très dispersé provoqué par les changements, qui sont apparu pendant une très longue période de temps dans la classification. Au tout début, la justesse diminue fortement avec la croissance des FT dans la BDD. Puis l'écart type de la justesse augmente très fortement entre 10000eme et 20000eme FT. On remarque tout de même une convergence à très long terme sur une moindre dispersion associée à une justesse dégradée.

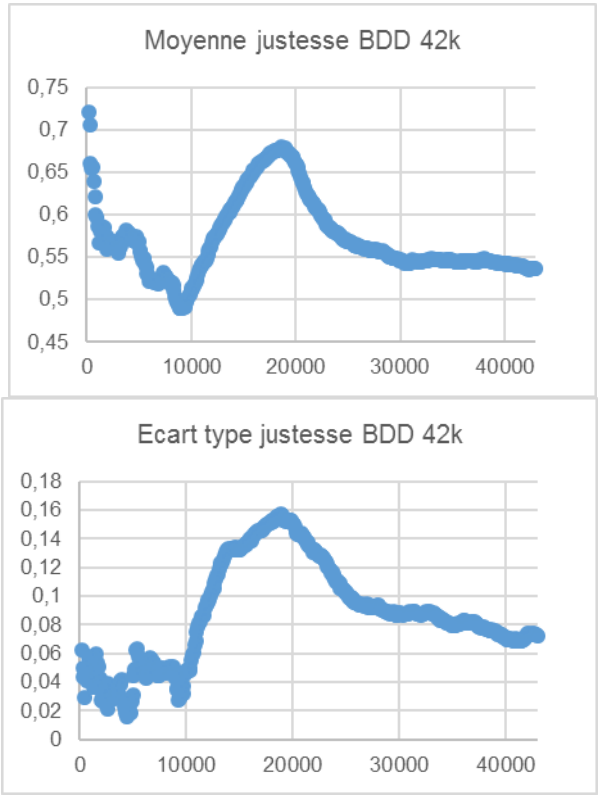

Fig. 7. sensibilité de la justesse de classification automatique de la BDD 42k selon nombre des données d'entrainement

Malgré une très grande quantité de données, la BDD 42k n'a pas beaucoup d'intérêt en soi pour la classification en gravité. De plus, pour obtenir uniquement une distribution de sa justesse dépendant de la longueur de la base, il aura fallu 3 jours de temps de calcul pour un ordinateur de bureautique.

La quantité de données a évidemment un impact sur la durée de traitement et la consommation en mémoire de l'ordinateur. L'utilisation de BDD d'entrainement au-delà de quelques milliers observations est rédhibitoire pour une exploitation récurrente en classement (besoin d'attendre plusieurs minutes pour le résultat aussi bien en déroulant l'entrainement qu'en chargeant un modèle prétraité, enregistré). Par ailleurs l'exemple ci-dessus montre qu'il est plus intéressant de travailler sur la qualité et la stabilité du classement manuel de la base d'entrainement que sur la quantité de données dans l'absolu.

En augmentant la longueur et la qualité de texte d'entrainement pour chaque Fait Technique classifié, on aura tendance à améliorer le score de précision et taux de rappel en apportant une information pertinente. Vu que le texte de la BDD 2k est divisé en plusieurs colonnes, nous avons entrainé le modèle en partant d'une seule colonne et en augmentant le nombre de colonnes « plain text » correspondant au titre, à la description du fait technique, à ses causes, à ses effets et à son traitement comme suit :

TABLE IV. RESULTATS D'ENTRAINEMENT BDD 2K EN MODIFIANT LE NOMBRE DES COLONNES PLEIN TEXTE (2500 TIRAGES)

TEST	Modification	Rappel LOS (%)	F1 score LOS (%)	F1 score hors G3 (%)	Nb tokens (mots)	Nb signes texte par FT
1	Titre	49,26	58,41	64,37	866	47
2	+Description	57,24	62,42	70,60	1769	220
3	+Effet	57,59	62,43	70,90	1798	225
4	+Traitement réalisée	57,43	62,58	70,99	2850	512
5	+Causes	57,41	62,63	71,27	3062	584
6	+Traitement proposée	59,48	63,96	71,79	3303	725
7	+Activité de campagne	60,42	65,15	73,14	3363	762
8	+Disposition qualité	60,55	64,67	73,10	3389	794
9	+spécialité - dispo qualité	60,47	65,02	73,12	3362	788
10	+classe cause - spécialité	60,33	64,73	73,14	3371	804

Certaines données « plain text » auront tendance à dégrader le résultat de classification. Dans le test 8 ci-dessus, il s'agissait d'ajouter à l'entrainement une colonne décrivant les dispositions de type qualité après sa résolution : le contenu n'apportait pas beaucoup d'information technique permettant de différencier les gravités. Le résultat en terme de score F1 était moins bon que dans le test 7 précèdent, ne prenant pas en compte cette colonne. Dans les tests 9 et 10 nous avons également testé l'ajout de certaines métadonnées textuelles, telles que la classification de la cause du FT ou encore la spécialité technique concernée. Cette démarche s'est avérée contreproductive car elle n'ajoutait pas de mots significatifs permettant de bien différencier les classes de gravité : les scores F1 et rappels ont été dégradés par rapport au meilleur test 7. A noter que dans le cadre d'article les études de sensibilité présentées ci-avant ont été réalisées en configuration du Test 5.

X. EFFET D'AJOUT D'UNE METADONNEE DESCRIPTIVE A LA GRAVITE DES RISQUES

Comme décrit dans les § précédents, le niveau de gravite en tant que métadonnée est assez général, agrégeant de nombreux faits techniques de natures très différentes. Pour plus de clarté pour l'utilisateur, nous avons ajouté le classement de tous nos faits techniques par des classes d'évènement redoutées intermédiaires, par exemple « une fuite interne d'un circuit fluide », qui lui-même peut mener à un risque LOS dans le sens d'EDD présenté en §II.

Généralement il est possible d'identifier plusieurs évènements redoutés associés à un fait technique. Pour cette raison, le résultat d'entrainement d'une métadonnée complexe permet de proposer à l'utilisateur une formulation alternative à la sienne pour décrire le risque associé à un fait technique. Parfois la classification automatique donne un résultat divergeant du jugement expert. Il est dû au fait que l'outil cherche à rapprocher les FT le plus corrélés par rapport à des mots (tokens) exprimant une classification. Le modèle statistique évalue la fréquence de ces mots vis-à-vis de la distribution statistique des mots dans les autres catégories. Dans la grande majorité des FT ces mots expriment des concepts techniques significatifs et donc des raisons pertinentes pour l'outil pour classer un FT dans une catégorie donnée. D'autre part le classement manuel

préalable de la BDD d'entrainement constitue un savoir latent de l'expert qui l'avait classifié. La qualité et cohérence de la classification manuelle apporte donc le sens à la classification automatique dérivé de celui-ci. Ceci peut conduire à la révision de la gravité et/ou évènement redouté initialement identifié par l'ingénieur conformité. C'était notamment le cas lors de la relecture de l'ensemble de la base d'entrainement BDD 2k.

Dans cette BDD nous avons identifié environ 40 « évènements redoutés », qui peuvent être associés avec une ou plusieurs gravités, créant ainsi une métadonnée combinée de la gravité et l'évènement redouté associé. Dans le cas de cet exercice on passe de 4 gravités à 53 couples « gravité + évènement redouté ». D'une part ce type de classement est plus descriptif pour l'ingénieur conformité LOS et d'autre part, il permet de mieux compartimenter les faits techniques de nature similaire au sein d'une telle métadonnée combinée. Pendant la mise à jour manuelle des données d'entrainement, l'ingénieur conformité LOS décide quels faits techniques sont similaires aux autres. Le défaut de cette démarche est le besoin de maitriser un très grand nombre de critères et règles de classement pour les 40 évènements redoutés identifiés par rapport aux 4 gravités.

TABLE V. EXEMPLE DE CLASSIFICATION DE CORPUS D'ENTRAINEMENT

Ref. FT	**Gravité**	**Evènement Redouté**	**Source classif. Auto**
FT1	LOS	Impact Sauvegarde Vol (MSI)	LOS Impact Sauvegarde Vol
FT2	LOS	Objets non identifiés	LOS Objets non identifiés
FT3	G3	Défauts superficiels matériel	G3 Défauts superficiels matériel
FT4	G2	Mesures indispo en chrono	G2 Mesures indispo en chrono
FT5	LOS	OVNI	LOS OVNI
FT6	G2	Mesures indispo en chrono	G2 Mesures indispo en chrono
…	…	…	… …
FTn	…	…	… …

L'effet direct sur la classification automatique est négatif : les taux de rappel et la précision sont dégradés d'environ 10% par rapport au classement des 4 gravités en forme « C3 », car les corpus sous-jacents de chaque catégorie combinée sont plus petits. Par contre, après entrainement avec cette métadonnée combinée, il est possible de tronquer l'évènement redouté pour revenir aux 4 gravités en forme « C3 ».

TABLE VI. EXEMPLE DE RESULTAT DE CLASSIFICATION DE M FAITS TECHNIQUES SOUMIS A LA CLASSIFICATION AUTOMATIQUE

Ref. FT	**Résultat classification Auto**		**Gravité tronqué**
FTa	C Exploitation TM degrade	→	C
FTb	LOS Perte intégrité structural	→	LO
FTc	G3 Défauts superficiels matériel	→	G3
FTd	G3 Configuration fluide remis en sécurité (MCA)	→	G3
FTe	LOS Pilotage degrade	→	LO
FTf	G3 Défauts superficiels matériel	→	G3
…	… …	→	…
FTm	… …	→	…

En procédant ainsi, il est possible de calculer le taux de rappel et la précision de détection de la gravité du risque pour les deux approches : par troncature d'une métadonnée complexe « gravité + évènement redoutée » et par simple classification de la gravité. Une amélioration peut être constatée sur le rappel de la plupart des gravités en utilisant une métadonnée combinée, notamment « LOS » tout en maintenant le niveau semblable de précision. Le rappel et la précision de la classe de gravité « G3 » concernant les évènements anodins, sont dégradés, ce qui indique que la classe la plus nombreuse est pénalisée par cet approche.

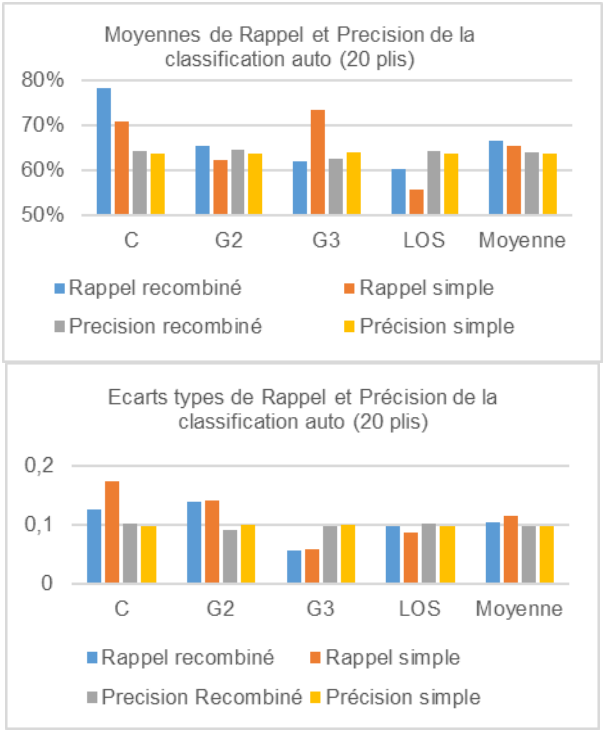

Fig. 8. métriques d'entrainement de la BDD 2k avec 20 plis avec métadonnée complexe et simple.

XI. SIGNAUX FAIBLES NON DETECTABLES AVEC LA CLASSIFICATION AUTOMATIQUE

La relecture des résultats de la classification automatique a démontré que les faits techniques inhabituels ne sont pas correctement classifiés. Par exemple, un problème anodin d'étalonnage des capteurs peut cacher un problème bien plus important du système concerné. Vu que les problèmes anodins avec des capteurs arrivent plus souvent, ces textes sont classés comme tels dans le corpus d'entrainement. Quand un problème majeur avec les mêmes symptômes est révélé, il sera la plupart de temps classé comme « anodin » à cause des mots semblables aux problèmes anodins dans l'énoncé du fait technique. La classification automatique aura tendance à « gommer » les phénomènes divergents dans le temps, tant qu'ils contiennent des tokens similaires aux évènements du passé.

Des « cas d'espèce » (faits techniques uniques dans leur genre) constituent un autre problème, puisque la base d'entrainement ne contient pas de cas semblables. La classification aura tendance à être aléatoire dans une telle situation.

Tout cela conduit à considérer la classification automatique uniquement comme une technique de productivité, permettant de confronter l'avis d'un

responsable de la conformité LOS à la vision statistique plus large de son texte, sans la capacité de révéler un « signal faible » d'une manière efficace.

XII. Pistes envisages pour la recherche des signaux faibles avec la classification automatique

Pour adresser la notion du « signal faible » pour la détection des faits techniques potentiellement graves, une alternative consiste à explorer la notion de « gravité intrinsèque » : il s'agit de définir la gravité a posteriori du traitement d'un fait technique, en ayant connaissance de son traitement. Les faits techniques a priori identifiés comme graves peuvent devenir anodins après une analyse démontrant l'absence d'impact réel. A l'inverse, le fait d'exécuter des actions importantes en réduction de risque (par exemple l'échange d'un équipement) renforce la gravité importante initialement identifiée. Dans notre cas, cette démarche a conduit à réduire fortement le nombre de faits techniques à impact LOS dans notre corpus d'entrainement. Une démarche supplémentaire pourrait donc consister à utiliser une telle classification plus ciblée conjointe à l'utilisation d'une métadonnée complexe présenté en §X.

Une autre piste concerne la détection de la causalité dans les phrases des textes techniques, soit par une annotation syntaxique, soit par les méthodes d'apprentissage profond basées sur un corpus d'entrainement de référence. Enfin un travail plus approfondi pourrait être mené sur l'utilisation et optimisation des modèles, notamment des réseaux de neurones.

XIII. Conclusion

Les résultats de notre étude sont très sensibles à l'équilibre de la distribution des différentes gravités dans la base d'entrainement des FT. Les évènements redoutés de nature anodine constituent la grande majorité de la base d'entrainement des faits techniques qui apparaissent dans l'exploitation des lanceurs. En variant différents paramètres de l'ensemble des étapes menant à la classification automatique, nous avons démontré des opportunités intéressantes pour l'amélioration de ses métriques.

Il faut tout d'abord travailler sur la qualité du classement manuel de la base d'entrainement et sur son prétraitement, qui semblent avoir une influence prépondérante sur les résultats. Cela a pu être constaté lors de la comparaison de notre base d'entrainement de référence avec d'autres bases de données classées moins rigoureusement. In fine, la classification automatique permet ainsi de déterminer la cohérence et la qualité des classificateurs passés, quand une BDD de référence, validée par les experts, existe. D'autre part, nous avons pu observer l'intérêt d'inclure les retours d'expérience dans la base d'entrainement pour améliorer les résultats de la classification automatique, mais sous condition d'une classification manuelle rigoureuse de ces données.

In fine pour la base d'entrainement, un taux de rappel de l'ordre de 75% pour l'ensemble des classes de gravité et notamment la gravité « LOS » a été obtenu en utilisant la régression logistique, le F1-score de 75%, alors qu'au début cet indicateur se trouvait à un niveau bien inférieur, proche de 40%. Actuellement en utilisant l'ensemble des optimisations étudiées ci-avant nous arrivons à remonter ce score à 78%.

L'ajout d'une métadonnée supplémentaire, décrivant l'« Évènement Redouté » à la base d'entrainement permet d'améliorer la classification automatique des gravités des FT. Le défaut principal de cette démarche est l'impossibilité de détecter de nouveaux évènements graves sans antériorité, car les classes correspondantes ne sont pas identifiées dans la base d'entrainement.

Par contre la confrontation des résultats d'analyse automatique avec l'avis des experts nous a conduits à réviser des avis antérieurs d'experts en améliorant ainsi le taux de rappel et la précision par principe de précaution. Pour un fait technique spécifique, l'expert prenait en compte une vision plus globale traduite par la base d'entrainement.

En perspective, nous allons adresser le problème de la détection des nouveaux FT graves sans antériorité directe, notamment par des méthodes issues de l'apprentissage profond, par la détection automatique de la causalité et l'exploration de la notion de la gravité intrinsèque d'un FT. Nous allons également continuer à chercher des optimisations du modèle appliqué et des prétraitements pour améliorer les métriques de notre classement.

References

[1] Galand, L., Kurela, M., Clavijo, H. (2018). Techniques de TAL pour la recherche des « signaux faibles » et catégorisation des risques dans le REX SDF des lanceurs spatiaux. Communication présentée au 21e Congrès Lambda, « Maîtrise des risques et transformation numérique : opportunités et menaces », Reims, France.

[2] Tulechki, N. (2015). Natural language processing of incident and accident reports: application to risk management in civil aviation (Thèse de doctorat). Université Toulouse Jean Jaurès.

[3] Dechy, Nicolas et Jouniaux, Pierre et Haduda, David et al. (2013). Détection et pertinence d'un signal faible dans le traitement d'un retour d'expérience,

[4] Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C. (2016). Natural language processing for aviation safety reports: from classification to interactive analysis. Computers in Industry, 78, 80-95

[5] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011

[6] Mathieu Stricker. Réseaux de neurones pour le traitement automatique du langage: conception et réalisatin de filtres d'informations. domain_other. ESPCI ParisTECH, 2000. English. pastel-00000488

[7] Arrêté du 31 mars 2011 relatif à la Réglementation Technique en application de la Loi N° 2008-518 du 3 juin 2008 relative aux opérations spatiales (parution au J.O. du 31/05/2011)

[8] IMdR (2013) Méthodes d'analyse textuelle pour l'interprétation des REX humains, organisationnels et techniques. Synthèse du projet P10-5, 2013

[9] REI (09/12/2010), Règlementation d'Exploitation des Installations du CSG

[10] A.M. Kuchling, Guide des expressions régulières, https://docs.python.org/fr/3/howto/regex.html