Analyse de Données Aberrantes pour le Texte : Taxonomie et Étude Expérimentale

Jeremie Pantin*, Marie-Jeanne Lesot*, Christophe Marsala*

*Sorbonne Université, CNRS, LIP6, Paris F-75005, France
Prenom.Nom@lip6.fr

Résumé. La tâche de détection des données aberrantes consiste à étudier un ensemble de données afin d'estimer les individus singuliers. Cette tâche est souvent effectuée pour supprimer le bruit ou les données anormales lors de l'étape de nettoyage. Si la littérature est riche pour de nombreux types de données, les contributions pour le texte sont plus rares. Dans cette étude nous nous intéressons aux données aberrantes textuelles. L'apparition de ces aberrations peut survenir de manière indépendante ou contextuelle pour un sujet donné. Nous introduisons ainsi une taxonomie pour les identifier, et sur la base de celle-ci, nous proposons GenTO qui est une approche de génération d'observations aberrantes. Son utilisation permet d'identifier les forces et faiblesses d'une approche face à différentes aberrations. Une étude comparative des approches de l'état de l'art est finalement effectuée en utilisant GenTO. Les résultats montrent que les travaux récents ne sont pas significativement plus performants.

1 Introduction

La tâche de Détection des Données Aberrantes (DDA) concerne de nombreux domaines et applications. Cette tâche est formellement présentée dans de nombreux travaux (Hawkins, 1980; Hodge et Austin, 2004; Zhang, 2013; Aggarwal, 2017; Ruff et al., 2021). Auparavant, elle était réalisée dans le domaine des statistiques (Beckman et Cook, 1983) avec des données statiques et des méthodes paramétriques. Les progrès récents en fouille de données ont fait évolué la tâche, ainsi que son formalisme, pour la rendre applicable à un plus grand nombre de scénarios. Pour Aggarwal (2017), *une donnée aberrante est une donnée qui est significativement différente des autres données.* Les méthodes effectuant la DDA visent souvent à calculer un score de "normalité" d'une instance. Bien que ces approches produisent des résultats significatifs avec de nombreux types de données, leur application au texte manque de clarté.

Dans les travaux mentionnés, il existe un nombre suffisant de contributions qui offrent un aperçu de la tâche de DDA. Toutefois, elles étudient rarement le scénario dans lequel les données sont purement textuelles. L'intérêt pour la Détection des Données Aberrantes Textuelles (DDAT) a été récemment suscité par les réseaux neuronaux (Gorokhov et al., 2017; Ruff et al., 2019; Lai et al., 2020) et les méthodes de décomposition de matrices (Allan et al., 2008; Kannan et al., 2017). Bien que ces contributions

Sujet : Sport/Tennis. Documents normaux :
x_1 : Naomi Osaka came to Flushing Meadows to entertain and did not disappoint on Monday, overcoming a slow start to beat Czech Marie Bouzkova 6-4 6-1 and get her U.S. Open title defence under way in front of a roaring capacity crowd.
x_2 : World number one Novak Djokovic had to work hard for a three-set victory over Hungarian Marton Fucsovics at the Paris Masters on Tuesday in his first match since losing the U.S. Open final in September.

Sujet : Politique. Document aberrant :
x_3 : The Olympic Games have started and yet, all countries are [...] Japanese tennis star Naomi Osaka on Friday lit the Olympic cauldron to mark the formal start of Tokyo 2020, in an opening ceremony shorn of glitz and overshadowed by a pandemic but celebrated as a moment of global hope.

FIG. 1 – *Exemple de trois articles de Reuters et Eurosport qui porte sur le tennis et la politique. Les documents normaux traitent des résultats des joueurs de tennis tandis que le sujet aberrant porte sur la cérémonie d'ouverture des Jeux olympiques.*

se concentrent sur l'exécution de la tâche de DDA, ici appelée détection des anomalies, elles n'étudient pas la nature des aberrations textuelles. Une observation récurrente concernant ces contributions est qu'elles ne se comparent pas aux travaux récents. De plus, elles n'étudient pas les aberrations construites avec leur protocole expérimental.

La Figure 1 présente un exemple de DDAT avec trois articles de presse. Alors que le sujet principal est Sport/Tennis, nous observons qu'un article Politique apparaît à tort dans le corpus. Un tel scénario est courant, et la tâche de détection de spam présente des similitudes avec ce problème. La recherche bibliographique de DDAT dans la littérature est déroutante car plusieurs tâches telles que la détection d'anomalies, la détection d'aberrations et la détection de nouveautés sont souvent utilisées de façon interchangeable. Au vu de l'absence d'ensembles de données dédiés à la DDAT, la proposition d'un algorithme de génération de données aberrantes apparait nécessaire.

1.1 Contributions

Approcher la tâche de DDAT est difficile pour plusieurs raisons : le faible nombre d'études approfondies avec des approches de l'état de l'art, la définition de nouveaux problèmes pour une tâche similaire (détection de spam, détection de plagiat, ...), le manque de corpus dédiés et les protocoles expérimentaux divergents dans la littérature.

Étude approfondie des travaux correspondants. Notre première contribution est l'étude de la DDA pour les données textuelles avec une analyse approfondie des méthodes connexes. Alors que les approches de DDAT sont rares dans la littérature, nous étudions les méthodes populaires de détection de nouveautés, de détection d'anomalies, de détection de plagiats, de détection de spams et de classification à classe unique (CCU).

Taxonomie et génération d'aberrations. L'absence de jeux de données dédiés a conduit les approches de la littérature à préparer spécifiquement les corpus de classification de textes. Nous proposons une taxonomie qui identifie les différents types d'aberrations pour le texte. Nous formalisons celle-ci à partir des travaux de référence de la tâche de DDA. L'avantage de cette taxonomie est sa connexion avec des types d'aberrations bien connus de la littérature. À partir de celle-ci, nous proposons une approche qui génère des aberrations ponctuelles/indépendantes et des aberrations contextuelles/conditionnelles : GenTO. Notre algorithme peut être facilement appliqué à n'importe quel corpus possédant une hiérarchie de sujets.

Comparatif des configurations expérimentales. Dans la littérature, le type des aberrations construites n'est pas clairement identifié lors de l'étape d'évaluation, conduisant à une lecture incomplète des résultats. La proposition de GenTO est fortement motivée pour cette raison. En complément, nous définissons le taux de contamination d'un corpus (poids du nombre d'aberrations dans un corpus). Dans divers travaux, ce paramètre peut varier afin de donner une idée de la robustesse d'une approche par rapport aux données du monde réel. Nous proposons une étude comparative, basée sur GenTO, qui met en évidence ces caractéristiques.

1.2 Structure de l'article

La Section 2 passe en revue les travaux connexes et souligne les principales caractéristiques de la tâche. La Section 3 présente notre taxonomie et son application au texte, ainsi que l'algorithme de génération d'aberrations : GenTO. Une étude comparative est présentée dans la Section 4 sur tous les types d'aberrations identifiés. Enfin, nous concluons avec la Section 5 et abordons les perspectives de ce travail.

2 Préliminaires

Dans cette section, nous présentons les connaissances principales associées à l'analyse des données aberrantes, ainsi que les travaux connexes de la tâche de DDAT. Nous soulignons également les différences entre les termes anomalie et nouveauté qui, dans la pratique, sont souvent employés de manière interchangeable avec aberration.

2.1 Qu'est-ce qu'une aberration textuelle ?

Bien que la définition d'une observation aberrante soit claire dans le cadre des statistiques, les progrès récents avec les données plus complexes (image, texte, ...) impliquent de fournir de nouvelles définitions. Des termes tels que *anomalie* et *nouveauté* renvoient à des observations similaires à celles des données aberrantes. Concrètement, une anomalie est une observation d'intérêt, une nouveauté est un type d'observation qui nécessite la mise à jour des modèles et une aberration est une instance fréquemment considérée comme une donnée qui devrait être supprimée (Ruff et al., 2021).

Le texte possède des propriétés telles que la rareté et la grande dimensionnalité, la rendant difficile à traiter. En ce qui concerne la définition de Aggarwal (2017),

nous proposons qu'une aberration textuelle est *un document qui est significativement différent des autres documents.* Cette définition soulève un problème lié au texte qui est le niveau d'application : syntaxique ou sémantique. Nous nous concentrons sur cette dernière, et proposons la définition suivante : *"une aberration textuelle est un document associé à un sujet qui est significativement différent des sujets des autres documents."*

2.2 Détection de données aberrantes dans le texte

Dans cette partie, nous présentons quatre types d'approches qui effectuent la DDA. La plupart de ces contributions sont réalisées avec des données textuelles, mais pas nécessairement pour la tâche de DDAT. Premièrement, nous étudions les approches basées sur la distance qui calculent un score à partir de l'éloignement entre les observations. Le deuxième type d'approche est basé sur la densité, où les observations proches se concentrent autour d'une même partie de l'espace. Les approches SVM basées sur des noyaux séparent quant à eux les données via un hyperplan. Enfin, nous terminons par les approches par réseaux de neurones artificiels.

Approches basées sur la distance. Plusieurs travaux sont fondés sur l'utilisation d'approches par distance, tel que KNN : (Ramaswamy et al., 2000; Koppel et Seidman, 2013; Kannan et al., 2017). Koppel et Seidman (2013) propose l'utilisation de KNN avec une métrique de similarité de second ordre afin d'effectuer la détection du plagiat dans des textes. Wang et al. (2018), quant à eux, proposent trois modèles qui calculent la distance d'un document à un autre, d'un document à un cluster, et d'un document au reste du corpus. Les approches basées sur la distance sont très sensibles à la métrique de distance et peinent avec les matrices éparses (Aggarwal, 2017; Kannan et al., 2017).

Approches basées sur la densité. Nous pouvons observé que les approches basées sur la densité peuvent traiter le problème de DDA. L'hypothèse naturelle qui suit notre définition d'aberration textuelle est : *une aberration est situé en dehors de la zone de densité formée par les données normales.* Srivastava et Zane-Ulman (2005) proposent une approche basée sur le modèle de mélange gaussien (GMM) et l'Analyse en Composantes Principales (ACP) dans le but de réduire la matrice sac de mots (BOW). Des travaux tels que Tran Manh Thang et Juntae Kim (2011) utilisent DBSCAN pour détecter des anomalies via une optimisation des paramètres et de mécanismes dédiés. Local Outlier Factor (LOF) (Breunig et al., 2000) est un autre algorithme populaire qui estime le taux d'aberration d'une observation en fonction de sa proximité avec une densité locale. Walkowiak et al. (2020) l'utilise pour distinguer les documents incorrectement associés à des clusters. Récemment, Fouché et al. (2020) proposent également une approche pour l'extraction de données aberrantes dans des répertoires de documents. Leur méthode, kj-NN, utilise la similarité sémantique et estime au niveau des phrases la pertinence d'un document. Il est intéressant de noter qu'ils proposent également une taxonomie dédiée possédant deux types de d'aberrations : Type-O (documents hors distribution) et Type-M (documents mal classés).

Approches à noyaux. OCSVM (One-Class Support Vector Machine) est l'une des approches les plus populaires pour la détection d'aberrations et d'anomalies. Manevitz et Yousef (2001) utilisent OCSVM pour des données textuelles où ils étudient son efficacité sur quatre représentations. Ils effectuent leurs expérimentations avec des noyaux linéaires, polynomiaux, radiaux et sigmoïdes. Également, Shravan Kumar et Ravi (2017) ont récemment proposé d'utiliser OCSVM couplé à une représentation sémantique du texte avec LSA (Latent Semantic Analysis). Comparativement, ils ajoutent la réduction de dimension au travail de Manevitz et Yousef (2001).

Approches basées sur des réseaux de neurones artificiels. Les progrès récents des réseaux neuronaux artificiels ont eu un impact positif sur la tâche de DDAT. Deux types d'approches peuvent être observés : la reconstruction et la classification à classe unique. Mei et al. (2018) proposent une approche de détection de nouveautés qui utilise des auto-encodeurs. Bien que les résultats montrent que la méthode manque de stabilité au fil des expériences, elle parvient à éviter correctement les faux positifs. L'approche proposée par Lai et al. (2020) utilise une couche de récupération de sous-espace robuste (RSR). Cette couche cherche à extraire des sous-espaces significatifs où les aberrations seraient difficiles à localiser. Cette technique facilite l'étape de reconstruction du décodeur. Gorokhov et al. (2017) proposent une méthode de détection d'anomalies avec un réseau neuronal convolutif (CNN). On observe toutefois que l'efficacité des modèles de langage récents n'est pas pleinement exploitée dans la tâche. Ruff et al. (2019) proposent ainsi de les utiliser avec un Context Vector Data Description (CVDD).

3 Taxonomie des aberrations textuelles

Dans cette section, nous présentons deux des contributions de nos travaux : une taxonomie des aberrations textuelles et un algorithme de génération d'aberrations : GenTO. Une introduction des notations est effectuée au préalable.

3.1 Notations

Il existe plusieurs types de documents, et il est courant d'observer différentes sources de textes dans la littérature. Étant donné un corpus, *une aberration textuelle est un document qui est significativement différent des autres documents*. En suivant cette définition, nous pouvons l'appliquer aux articles de presse, par exemple, où une *aberration textuelle est un document associé à un sujet qui est significativement différent des sujets des autres documents*. Soit une séquence de M mots $x_1, ..., x_M$ avec un vocabulaire fixe $\mathcal{V}$ de taille $|\mathcal{V}| = V$. Chaque mot est représenté par $x_i \in [0, N]$ pour $i \in \{1, ..., M\}$ lorsque le modèle de représentation est un sac de mots (BOW). Ainsi, $\mathrm{x} \in \mathrm{X}$ où X est un corpus de N documents.

La DDA peut être associée à la classification à classe unique (CCU) où la plupart du temps, la sortie est un score tel que $s : \mathrm{X} \mapsto \mathcal{R}$. Les sorties de la classification binaire et de la CCU sont similaires, ainsi -1 désigne une aberration et $+1$ une donnée normale. Un classifieur de DDA vise à trouver le nombre optimal de données divergentes de X tout en minimisant autant que possible les faux positifs.

3.2 Taxonomie

Plusieurs définitions d'une aberration ont été proposées dans la littérature : aberration indépendante/ponctuelle, aberration conditionnelle/contextuelle et aberration collective/groupée. La nature du texte implique que lorsque plusieurs données aberrantes apparaissent, elles sont à la fois contextuelles et collectives.

Considérant que les documents contiennent des informations d'un ou plusieurs sujets, nous supposons qu'il existe une hiérarchie des sujets du corpus. Ce type de hiérarchie est nécessaire afin d'étudier les aberrations contextuelles sans ensemble de données dédié. Dans la tâche de classification, les approches hiérarchiques ne sont pas rares (Toutanova et al., 2001) et les structures à deux ou trois niveaux sont courantes.

Les sujets (catégories) sont organisés sous forme d'une structure arborescente où ils peuvent être associés, tout au plus, un parent. Nous ajoutons également la contrainte qu'une catégorie est unique, et qu'elle ne peut pas apparaître deux fois comme enfant ou parent. Dans cette configuration, les documents ne peuvent apparaître qu'au niveau des feuilles, et ne peuvent donc pas être parent d'un sujet. Soit H une hiérarchie qui a $l \in L$ niveaux et $y \in Y$ sujets. Chaque élément $h \in H$ admet un ensemble d'enfants qui mène au niveau $l + 1$ qui est soit un sujet, une feuille ou $\{\emptyset\}$.

Les aberrations indépendantes sont des observations dont le sujet ne partage aucune relation avec un autre sujet. Précisément, le sujet aberrant et le sujet définie comme normal ont un parent différent dans la structure hiérarchique des catégories. On note ζ la catégorie normale et son sous-ensemble correspondant $X_\zeta \subseteq X$. Nous définissons O le sous-ensemble de toutes les observations aberrantes, tel que $O \subset X$. On a :

$$O = X \backslash X_\zeta \tag{1}$$

En ce qui concerne O, nous pouvons distinguer deux contraintes différentes. Premièrement, une observation x_i est considérée aberrante si son sujet parent est différent du sujet parent normal, tel que :

$$O_{indépendante} = \{\text{parent}(y) \neq \text{parent}(\zeta) | \forall (o, y) \in O \times Y\} \tag{2}$$

La deuxième contrainte correspond aux documents qui ne se trouvent pas dans X_ζ mais qui partagent le même sujet parent que ζ. Ces observations sont identifiées comme un autre type d'aberration : aberrations contextuelles. Ainsi :

$$O_{contextuelle} = \{\text{parent}(y) = \text{parent}(\zeta) | \forall (o, y) \in O \times Y, O \backslash O_{point}\} \tag{3}$$

3.3 GenTO : Generation d'aberrations textuelles

À partir de l'Équation 2 et de l'Équation 3, nous proposons GenTO pour générer des aberrations textuelles. Dans la section précédente, nous avons défini deux types de d'aberrations : indépendantes et contextuelles. On note également que GenTO peut être appliqué pour chacun d'entre eux. L'Algorithme 1 décrit GenTO et suit l'Équation 2 pour la génération d'aberrations indépendantes et l'Équation 3 pour la génération d'aberrations contextuelles.

Algorithm 1 GenTO : Generation of Textual Outliers

Require: Sujet normal ζ, corpus X, taille l, taux de contamination ν, booléen *est_contextuelle*
Ensure: $0 < l \leq N$
  $c \leftarrow l\nu$
  $i \leftarrow 0$
  Initialisation de la matrice vide Z ▷ Matrice d'échantillonnage
  $O \leftarrow \{x_j \times y_j \in X \times Y | \forall j \in [0, N], y_j \neq \zeta\}$ ▷ Matrice de données aberrantes
  $X_\zeta \leftarrow \{X \backslash O\}$ ▷ Matrice de données normales
  while $|Z| < c$ **do**
    if *est_contextuelle* **then**
      Ajouter(x_i, y_i) dans Z sachant que Parent(y_i) = Parent(ζ)
    else
      Append(x_i, y_i) dans Z sachant que Parent$(y_i) \neq$ Parent(ζ)
    end if
    $i \leftarrow i + 1$
  end while
  Remplir Z avec X_ζ jusqu'à ce que $|Z| = l$
  return Mélanger(Z)

4 Expérimentations

Dans cette section, nous présentons les expériences menées sur la DDAT avec des aberrations générées via GenTO. Nous décrivons également les jeux de données et la manière dont ils sont utilisés, de l'étape de pré-traitement à l'étape de préparation. Les mesures d'évaluation utilisant Area Under the Receiver Operating Characteristics Curve (AUROC) et Area Under the Precision-Recall Curve (AUPRC) sont détaillées dans la deuxième partie. Les approches de référence sont introduites dans le troisième point, en plus de leur configuration. Enfin, nous présentons les résultats de notre expérimentation.

Données. Comme indiqué précédemment, les ensembles de données dédiés à la DDAT sont inexistants et la préparation de corpus compatibles est une étape importante. Même si il existe des groupes de jeux de données dédiés à la DDA (ODDS et UCI), ils fournissent principalement des données multidimensionnelles, des séries temporelles et des images. Dans ce contexte, les corpus textuels sont peu représentés. Toutefois, des tâches telles que la détection de spam et le détection de plagiat disposent d'un riche ensemble de corpus. Des travaux récents (Lai et al., 2020; Ruff et al., 2019; Kannan et al., 2017; Mahapatra et al., 2012) utilisent des ensembles de données de classification tels que Reuters-21578 [1] et 20 Newsgroups [2] dans le but d'évaluer leurs approches.

Pour 20 Newsgroups, nous transformons les documents en une matrice BOW après avoir appliqué la suppression des mots vides et la mise en minuscule du texte brut.

1. `http://www.daviddlewis.com/resources/testcollections/reuters21578/`
2. `http://qwone.com/~jason/20Newsgroups/`

Les termes qui apparaissent moins de trois fois ne sont pas pris en compte et le modèle BOW est entraîné sur le sous-ensemble d'entraînement. Nous préparons ensuite chaque classe avec GenTO pour générer des aberrations ponctuelles et contextuelles avec $\nu \in \{0.05, 0.10\}$. La motivation derrière ces valeurs est d'évaluer les approches contre plusieurs niveaux de contamination. Nous séparons les sous-thèmes entre sept thèmes principaux : informatique, vente, moteurs, politique, religion, science et sports.

Reuters-21578 est un corpus contenant des documents associés à plusieurs sujets. Nous supprimons tous ces documents afin de ne garder que ceux qui possèdent un unique sujet. Ensuite, le corpus est préparé de la même manière que 20 Newsgroups avec un modèle BOW entraîné sur le sous-ensemble d'entraînement. À partir de Toutanova et al. (2001), nous réorganisons les sujets pour obtenir une hiérarchie. Ainsi, quatre thèmes parents sont créés : matières premières, finance, métaux et énergie. Nous appliquons GenTO aux huit sujets qui ont le plus grand nombre de documents dans Reuters-21578.

Évaluation. La DDA est une tâche qui implique des données fortement déséquilibrées où les données normales (vrais négatifs) sont prédominantes. Par conséquent, la précision moyenne est souvent utilisée. Les différentes représentations sont comparées à l'aide de AUROC et de AUPRC. Ces mesures sont dérivées de la matrice de confusion et sont toutes deux souvent utilisées pour la tâche de détection des aberrations. La courbe ROC affiche le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) pour plusieurs seuils. L'augmentation ou la diminution de ce seuil influence les vrais positifs par rapport aux faux positifs. Elle aide à choisir le meilleur seuil pour un classifieur.

L'AUROC évalue les performances des méthodes de référence en fonction du score de l'aberration. Toutes les expériences ont été menées sur dix essais où l'AUROC et l'AUPRC sont calculées. L'étape d'évaluation est ensuite effectuée sur le sous-ensemble d'entraînement. Effectuer cette étape sur le sous-ensemble de test est similaire à évaluer la robustesse d'un modèle à la nouveauté. Dans les travaux futurs, l'étude de cette évaluation sera effectuée en raison de l'importance de l'information liée : est-ce qu'une approche est robuste contre les nouveaux types/générations de documents.

Méthodes de référence. Notre étape de préparation des données est similaire à Kannan et al. (2017); Lai et al. (2020); Ruff et al. (2019); Fouché et al. (2020), nous permettant de facilement conduire les expérimentations avec leur implémentation. Nous suivons de près leurs recommandations afin d'être le plus équitable possible. Pour TONMF (Kannan et al., 2017), nous avons fixé $\alpha = 15$ et $k = 30$ pour lesquels les résultats obtenus étaient les meilleurs. RSRAE (Lai et al., 2020) est configuré avec les paramètres décrits dans leurs travaux : la dimension latente est fixée à 10, le taux d'apprentissage à 0.00025 et le nombre d'époques à 200. Avec CVDD (Ruff et al., 2019), la configuration proposée par les auteurs est également rigoureusement respectée.

Pour la mise en œuvre de LOF, Isolation Forest (IForest), OCSVM (Manevitz et Yousef, 2001), KNN (Ramaswamy et al., 2000) et PCA (Shyu et al., 2003) nous utilisons l'outil PyOD (Zhao et al., 2019). La métrique de distance pour LOF est *cosine* et le nombre de voisins est fixé à 20, il en va de même pour KNN. Nous avons obtenu

		$\nu = 0.05$				$\nu = 0.10$			
		Indépendante		Contextuelle		Indépendante		Contextuelle	
Modèle		AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
LOF	20 Newsgroups	0.139	0.762	0.091	0.652	**0.217**	0.745	**0.153**	**0.637**
IForest		0.068	0.518	0.062	0.516	0.125	0.527	0.116	0.513
OCSVM		0.104	0.609	0.073	0.591	0.130	0.641	0.129	0.617
KNN		0.064	0.489	0.063	0.505	0.118	0.505	0.114	0.500
PCC		0.121	0.663	0.086	0.607	0.203	0.665	0.148	0.594
LSA_{l2}		0.105	0.614	0.079	0.592	0.172	0.700	0.148	0.687
NMF_{l2}		0.119	0.651	0.091	0.619	0.184	0.716	**0.150**	0.621
TONMF		0.072	0.548	0.073	0.514	0.147	0.619	0.127	0.547
CVDD		**0.176**	**0.783**	**0.107**	**0.702**	0.208	**0.822**	**0.150**	0.629
RSRAE		0.110	0.729	0.078	0.622	0.196	0.725	0.145	0.621
LOF	Reuters-21578	0.348	0.870	**0.302**	0.795	0.380	0.812	0.316	**0.751**
IForest		0.136	0.648	0.097	0.564	0.197	0.642	0.136	0.534
OCSVM		0.235	0.750	0.130	0.641	0.300	0.816	0.223	0.747
KNN		0.117	0.618	0.094	0.549	0.177	0.621	0.134	0.526
PCC		0.222	0.756	0.143	0.674	0.297	0.746	0.188	0.635
LSA_{l2}		0.229	0.756	0.155	0.719	0.268	0.709	0.245	0.610
NMF_{l2}		0.308	0.792	0.168	0.765	0.294	0.787	0.260	0.688
TONMF		0.109	0.632	0.102	0.587	0.140	0.647	0.148	0.693
CVDD		0.377	0.889	0.291	0.812	0.416	0.799	0.312	0.740
RSRAE		**0.390**	**0.898**	0.295	**0.825**	**0.449**	**0.866**	**0.346**	**0.752**

TAB. 1 – *Résultats des modèles de l'état de l'art pour les aberrations indépendantes et contextuelles, avec deux taux de contamination. Les mesures d'évaluation utilisées sont AUPRC et AUROC.*

de meilleurs résultats avec le nombre de voisins $n \in [20, 30]$, mais 20 semble être la valeur possédant le plus de stabilité. Isolation Forest (IForest) (Liu et al., 2008) est une approche basée sur le parcours d'arbres. Nous prenons la configuration par défaut proposée par les auteurs afin de calculer l'AUROC et l'AUPRC. Basé sur l'ACP, PCC est un classifieur de composantes principales robuste qui a été évalué à l'origine sur le jeu de données KDD'99. La méthode calcule un hyperplan de faible dimension construit par k vecteurs propres. Nous conservons toutes les composantes dans notre configuration. Également, nous utilisons le noyau polynomial pour OCSVM et définissons : degré = 2.

Nous proposons d'utiliser deux méthodes d'approximation de rang faible : analyse sémantique latente (LSA) et factorisation matricielle non négative (NMF). Nous prenons le rang le plus performant pour r en variant $r \in [20, 100]$. Afin d'aligner l'utilisation de NMF sur celle de TONMF, nous définissons tolerance $= 1e-6$. Également, β-loss est calculé avec la norme de Frobenius. Pour ces deux méthodes, nous calculons la norme l_2 sur la matrice réduite.

Résultats. Les résultats du tableau 1 montrent les performances des méthodes de référence avec AUROC et AUPRC, pour les aberrations indépendantes et contextuelles. Pour 20 Newsgroups, l'approche KNN est l'une des méthodes les moins performantes, en particulier avec $\nu = 0.05$. LOF avec la distance *cosine* et CVDD sont les méthodes les plus performantes tandis que RSRAE réussit à obtenir de bons résultats. Pour Reuters-21578, RSRAE est la méthode se démarquant le plus. LOF et CVDD ont également

de bons résultats tandis que TONMF, KNN et IForest rencontrent des difficultés, en particulier avec les aberrations contextuelles.

D'après les résultats, nous observons que les aberrations indépendantes enregistrent de meilleures performances que les aberrations contextuelles. De même, les approches sont plus performantes avec un taux de contamination plus élevé. Pour les réseaux de neurones artificiels, nous pouvons voir qu'ils sont plus robustes en fonction du taux de contamination. LOF semble également être beaucoup plus robuste que les autres approches basées sur la densité. Les approches à noyau et à distance peinent à obtenir des résultats similaires. Également, les méthodes d'approximation de rang faible obtiennent des résultats compétitifs dans la plupart des cas.

5 Conclusion

Nous avons appliqué la détection des aberrations aux données textuelles et nous avons également proposé une taxonomie dédiée. De plus, celle-ci peut être étendue à plusieurs corpus et applications. Une telle taxonomie peut aider à identifier de nouvelles caractéristiques dans un corpus. Sur la base de cette taxonomie, GenTO génère des aberrations indépendantes ou contextuelles à partir d'un corpus. Les expérimentations que nous avons menées montrent que les aberrations contextuelles sont plus difficiles à traiter que les aberrations indépendantes. Contrairement à la littérature, notre protocole expérimentale est transparent quant à la préparation des aberrations et des corpus.

Concernant les futurs travaux, nous souhaitons étudier le processus génératif pour les aberrations textuelles en absence d'étiquettes. De même, la robustesse par rapport au taux de contamination et aux types d'aberrations peut conduire à une métrique unique pour évaluer certaines approches. D'autres expériences sont prévues afin de proposer une étude comparative plus approfondie des approches pouvant réaliser la DDAT. En effet, nous souhaitons ajouter à l'étude des mécanismes propres au texte comme Kassab et Alexandre (2009).

Références

Aggarwal, C. C. (2017). *Outlier analysis* (Second edition ed.). Cham: Springer.

Allan, E. G., M. R. Horvath, C. V. Kopek, B. T. Lamb, T. S. Whaples, et M. W. Berry (2008). Anomaly Detection Using Nonnegative Matrix Factorization. In M. W. Berry et M. Castellanos (Eds.), *Survey of Text Mining II*, pp. 203–217. Springer.

Beckman, R. J. et R. D. Cook (1983). Outlier.....s. *Technometrics 25*(2), 119–149.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.

Fouché, E., Y. Meng, F. Guo, H. Zhuang, K. Böhm, et J. Han (2020). Mining text outliers in document directories. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 152–161. IEEE.

Gorokhov, O., M. Petrovskiy, et I. Mashechkin (2017). Convolutional Neural Networks for Unsupervised Anomaly Detection in Text Data. In H. Yin, Y. Gao, S. Chen, Y. Wen, G. Cai, T. Gu, J. Du, A. J. Tallón-Ballesteros, et M. Zhang (Eds.), *Intelligent Data Engineering and Automated Learning IDEAL 2017*, Volume 10585, pp. 500–507. Springer.

Hawkins, D. M. (1980). *Identification of Outliers.* Dordrecht: Springer.

Hodge, V. et J. Austin (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review 22*(2), 85–126.

Kannan, R., H. Woo, C. C. Aggarwal, et H. Park (2017). Outlier detection for text data. *SDM International Conference on Data Mining 17*, 489–497.

Kassab, R. et F. Alexandre (2009). Incremental data-driven learning of a novelty detection model for one-class classification with application to high-dimensional noisy data. *74*(2), 191–234.

Koppel, M. et S. Seidman (2013). Automatically Identifying Pseudepigraphic Texts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, pp. 1449–1454. Association for Computational Linguistics.

Lai, C.-H., D. Zou, et G. Lerman (2020). Robust subspace recovery layer for unsupervised anomaly detection. *ICLR International Conference on Learning Representations 8*, 1–28.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2008). Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422.

Mahapatra, A., N. Srivastava, et J. Srivastava (2012). Contextual anomaly detection in text data. *Algorithms 5*(4), 469–489.

Manevitz, L. M. et M. Yousef (2001). One-class SVMs for document classification. *Journal of Machine Learning Research 2*, 139–154.

Mei, M., X. Guo, B. C. Williams, S. Doboli, J. B. Kenworthy, P. B. Paulus, et A. A. Minai (2018). Using Semantic Clustering And Autoencoders For Detecting Novelty In Corpora Of Short Texts. In *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, pp. 1–8. IEEE.

Ramaswamy, S., R. Rastogi, et K. Shim (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00*, Dallas, Texas, United States, pp. 427–438. ACM Press.

Ruff, L., J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, et K.-R. Müller (2021). A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE 109*(5), 756–795.

Ruff, L., Y. Zemlyanskiy, R. Vandermeulen, T. Schnake, et M. Kloft (2019). Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4061–4071. Association for Computational Linguistics.

Shravan Kumar, B. et V. Ravi (2017). One-class text document classification with

OCSVM and LSI. In S. S. Dash, K. Vijayakumar, B. K. Panigrahi, et S. Das (Eds.), *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, Volume 517, pp. 597–606. Springer.

Shyu, M.-L., S.-C. Chen, K. Sarinnapakorn, et L. Chang (2003). A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables Fl Dept Of Electrical And Computer Engineering.

Srivastava, A. et B. Zane-Ulman (2005). Discovering recurring anomalies in text reports regarding complex space systems. In *2005 IEEE Aerospace Conference*, pp. 3853–3862. IEEE.

Toutanova, K., F. Chen, K. Popat, et T. Hofmann (2001). Text classification in a hierarchical mixture model for small training sets. In *Proceedings of the tenth international conference on Information and knowledge management - CIKM'01*, pp. 105. ACM Press.

Tran Manh Thang et Juntae Kim (2011). The anomaly detection by using DBSCAN clustering with multiple parameters. In *2011 International Conference on Information Science and Applications*, pp. 1–5. IEEE.

Walkowiak, T., S. Datko, et H. Maciejewski (2020). Utilizing local outlier factor for open-set classification in high-dimensional data - case study applied for text documents. In *Intelligent Systems and Applications*, Volume 1037, pp. 408–418. Springer.

Wang, F., R. J. Ross, et J. D. Kelleher (2018). Exploring Online Novelty Detection Using First Story Detection Models. In *Intelligent Data Engineering and Automated Learning IDEAL 2018*, Volume 11314, pp. 107–116. Springer.

Zhang, J. (2013). Advancements of Outlier Detection: A Survey. *ICST Transactions on Scalable Information Systems 13*(1), e2.

Zhao, Y., Z. Nasrullah, et Z. Li (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research 20*(96), 1–7.

Summary

Outlier detection is the task that studies a set of data and estimates its outlying observations. This task is often performed to remove noise or abnormal data during the cleaning stage. While the literature is rich for many types of data, contributions for text are poorly represented. In this study we are interested in textual outliers. The appearance of these outliers can occur independently or contextually for a given topic. We introduce a taxonomy to identify them, and based on it, we propose GenTO which is an approach to generate outliers. Its use allows us to identify the strengths and weaknesses of an approach against different type of outliers. A comparative study of the state of the art approaches is finally performed using GenTO. Results show that, although recent work tends to perform better, it does not perform significantly better than previous one.