Processus d'identification automatique d'articles de presse dédiés à la veille syndromique

Rosalie Nkounghawe[1], Paulin Melatagia[2], Isabelle Pieretti[3,4], Carlène Trévenec[5,6], Mathieu Roche[7,8]

[1] Université Sorbonne Paris Nord, France
[2] Université de Yaoundé I, Cameroun
[3] CIRAD, UMR PHIM, F-34398 Montpellier, France.
[4] PHIM, CIRAD, INRAE, Institut Agro, IRD, Université de Montpellier, Montpellier, France.
[5] INRAE, UMR ASTRE, F-34398 Montpellier, France.
[6] ASTRE, CIRAD, INRAE, Université de Montpellier, Montpellier, France.
[7] CIRAD, UMR TETIS, F-34398 Montpellier, France.
[8] TETIS, Université de Montpellier, AgroParisTech, CIRAD, INRAE, Montpellier, France

corine.nkounghawe@facsciences-uy1.cm, paulin.melatagia@facsciences-uy1.cm,
isabelle.pieretti@cirad.fr, carlene.trevennec@cirad.fr, mathieu.roche@cirad.fr

Résumé

Le projet SURSY (SURveillance SYndromique) vise à améliorer la détection précoce d'événements sanitaires en santé végétale à partir de sources textuelles issues du Web. L'un des principaux défis de ce domaine réside dans la rareté et l'hétérogénéité des données disponibles, rendant difficile l'apprentissage de modèles robustes. Dans ce travail, nous proposons une approche fondée sur plusieurs méthodes d'intelligence artificielle pour structurer, classifier et enrichir des données textuelles liées à la veille syndromique.

Notre contribution s'articule en trois phases. La première consiste à classifier les articles selon trois thématiques (santé animale, santé végétale, santé publique) à l'aide d'embeddings SBERT et d'un perceptron multicouches, après un équilibrage du corpus réalisé via une combinaison UMAP + K-means. La deuxième phase porte sur la détection automatique des articles de veille syndromique, en comparant une approche lexicale (TF-IDF + XGBoost) à une approche sémantique reposant sur le fine-tuning de SBERT. Enfin, une troisième phase exploratoire évalue l'impact de l'augmentation de données par rétro-traduction sur la classe minoritaire.

Mots-clés

veille syndromique, fouille de texte, traitement automatique du langage, classification de textes, évaluation des modèles

Abstract

The SURSY (SYndromic SURveillance) project aims to improve the early detection of plant health events using textual information extracted from Web sources. One of the main challenges in this domain lies in the scarcity and heterogeneity of available data, which makes it difficult to train robust models. In this work, we propose an approach based on several artificial intelligence methods to structure, classify, and enrich textual data related to syndromic surveillance.

Our contribution is organized into three phases. The first phase focuses on classifying articles into three thematic categories (animal health, plant health, public health) using SBERT embeddings and a multilayer perceptron, after balancing the corpus through a combination of UMAP and K-means clustering. The second phase addresses the automatic detection of syndromic surveillance articles by comparing a lexical approach (TF-IDF + XGBoost) with a semantic approach based on fine-tuned SBERT. Finally, a third exploratory phase evaluates the impact of data augmentation through back-translation on the minority class.

Keywords

syndromic surveillance, text mining, natural language processing, text classification, model evaluation

1 Introduction

La veille sanitaire repose de plus en plus sur l'exploitation automatique de sources d'information issues du Web afin de détecter précocement l'émergence ou la réémergence de maladies. Dans les domaines de la santé animale et végétale, plusieurs plateformes de surveillance ont été développées pour analyser des flux continus d'articles de presse et de contenus en ligne [20]. Parmi elles, ESA[1] (Epidémiosurveillance en Santé Animale) et ESV[2] (Epidémiosurveillance en Santé Végétale) constituent des dispositifs opérationnels permettant de suivre l'apparition d'événements sanitaires à partir de sources ouvertes [2]. Dans ce cadre, la plateforme PADI-web[3], dédiée à la surveillance événe-

1. https ://plateforme-esa.fr/fr
2. https ://plateforme-esv.fr/
3. https ://www.padi-web-one-health.org

mentielle en santé animale et végétale, collecte automatiquement des articles de presse multilingues et applique un filtrage thématique pour identifier des signaux sanitaires pertinents [21].

Cependant, ces dispositifs atteignent leurs limites lorsqu'il s'agit d'identifier des phénomènes émergents pour lesquels la cause est encore inconnue ou mal caractérisée. La veille syndromique vise à répondre à ce besoin en s'appuyant sur l'analyse de signaux faibles issus de descriptions de symptômes, indépendamment de l'identification explicite d'une maladie. Si cette approche est largement étudiée en santé publique et animale, notamment via des systèmes tels que HealthMap ou ProMED [3], son application à la santé végétale reste encore peu explorée, alors même que les enjeux liés aux maladies émergentes, aux nouvelles plantes hôtes et au changement climatique sont croissants.

Dans ce contexte, l'un des principaux défis de la veille syndromique en santé végétale réside dans la rareté et la diversité limitée des données textuelles disponibles, rendant difficile l'apprentissage de modèles de classification robustes. Le projet SURSY s'inscrit dans cette problématique en proposant d'étendre les systèmes de veille existants à la détection de signaux syndromiques en santé végétale, tout en consolidant les liens avec la veille en santé animale.

Dans cet article, nous proposons une approche de classification automatique de textes appliquée à des articles issus de sources Web et de la plateforme PADI-web. Le travail est structuré en trois phases principales : une première phase dédiée à la classification thématique, une seconde phase centrée sur l'identification des articles relevant de la veille syndromique, et une troisième phase exploratoire portant sur l'augmentation de données textuelles. Nous présentons les méthodes mises en œuvre, une synthèse de certaines expérimentations réalisées et les résultats obtenus, en mettant l'accent sur leur pertinence pour des applications opérationnelles de veille syndromique en santé végétale.

2 État de l'art

2.1 Veille syndromique et systèmes existants

La veille syndromique repose sur l'analyse de signaux faibles décrivant des symptômes avant la confirmation d'un diagnostic. En santé publique, plusieurs systèmes exploitent des sources ouvertes pour détecter précocement des événements émergents. Parmi eux, HealthMap [3], ProMED [4] et GPHIN [7] constituent des références majeures pour la surveillance mondiale des maladies infectieuses.

En santé animale et végétale, des plateformes telles que ESA, ESV permettent la surveillance d'événements sanitaires à partir de sources médiatiques [2]. Toutefois, ces systèmes restent principalement centrés sur des maladies connues et ne couvrent que partiellement les signaux faibles caractéristiques de la veille syndromique. À ce jour, peu de travaux portent sur la veille syndromique végétale, malgré des similarités fortes avec les descriptions de symptômes en santé animale.

2.2 Classification automatique de textes

Les approches classiques de classification textuelle reposent sur des représentations lexicales telles que TF-IDF [1], souvent combinées à des modèles supervisés comme les SVM [11] ou XGBoost [5]. Ces méthodes, bien que robustes et rapides, restent limitées par leur dépendance au vocabulaire et leur difficulté à capturer la sémantique profonde des textes.

Les modèles de langue pré-entraînés ont profondément transformé le traitement automatique du langage naturel (TALN). BERT [6], SBERT [18] et RoBERTa [14] ont démontré une capacité supérieure à représenter le sens des phrases, notamment dans des contextes où les formulations sont variées, implicites ou synonymiques. Leur adaptation par fine-tuning permet en général d'obtenir des résultats de bonne qualité même sur des corpus spécialisés.

2.3 Données rares et déséquilibrées

La classification en contexte de données déséquilibrées constitue un défi majeur, en particulier lorsque la classe d'intérêt est fortement minoritaire. Dans ce cas, les modèles supervisés ont tendance à privilégier la classe majoritaire, ce qui dégrade les performances sur la classe cible. L'utilisation de métriques adaptées est essentielle : l'AUC-PR est notamment recommandée pour évaluer les performances sur des classes minoritaires [19], car elle met l'accent sur la précision et le rappel de la classe positive.

Plusieurs stratégies permettent de limiter les effets du déséquilibre, notamment la pondération des classes, la sélection de données ou l'utilisation de représentations sémantiques plus robustes. Une synthèse des méthodes d'apprentissage sur données déséquilibrées est proposée dans [9], qui souligne l'importance de combiner des techniques d'équilibrage avec des modèles capables de capturer des structures complexes dans les données.

2.4 Augmentation de données textuelles

L'augmentation de données vise à enrichir artificiellement les classes minoritaires afin d'améliorer la robustesse des modèles. Les techniques courantes incluent la synonymie, la suppression ou l'insertion de mots, ainsi que la rétrotraduction [8], qui consiste à traduire un texte vers une langue pivot avant de le retraduire vers la langue d'origine. Ces méthodes permettent de générer des variantes textuelles tout en préservant le sens global.

Des approches plus récentes, telles que EDA (Easy Data Augmentation) [22], proposent des transformations simples mais efficaces pour diversifier les données textuelles. Par ailleurs, les modèles génératifs modernes ouvrent des perspectives pour produire des textes synthétiques plus variés et plus cohérents. Parmi eux, l'approche Retrieval-Augmented Generation (RAG) [13] combine un module de recherche d'information et un modèle génératif afin de produire des textes s'appuyant sur des documents réels. Cette méthode permet de générer des contenus plus diversifiés et plus fidèles au domaine étudié.

2.5 Contributions

Les contributions principales de ce travail sont les suivantes :

— la construction d'un corpus textuel structuré pour la veille syndromique en santé végétale, mis à disposition publiquement ;
— une approche de classification thématique (santé animale, santé végétale, santé publique) fondée sur SBERT et un équilibrage *UMAP + K-means*. UMAP est donc utilisé pour projeter les embeddings SBERT dans un espace de dimension réduite tout en préservant la structure locale des données. Cette projection facilite l'application de K-means, qui permet ensuite de sélectionner de manière contrôlée des échantillons représentatifs dans chaque cluster. Cette combinaison assure un équilibrage par thématique qui préserve la diversité sémantique du corpus tout en limitant les biais liés aux thématiques sur-représentées ;
— un modèle de détection d'articles de veille syndromique reposant sur le fine-tuning de SBERT, surpassant les approches lexicales. Plusieurs variantes de SBERT ont été expérimentées, mais une seule retenue que nous présentons dans la section 4 ;
— une exploration de l'augmentation de données par rétro-traduction et son impact sur la classe minoritaire.

3 Données

Les données utilisées dans ce travail proviennent principalement de **PADI-web**, une plateforme de veille automatisée collectant quotidiennement des articles de presse multilingues liés à la santé animale et végétale. Les corpus ont été constitués et annotés différemment selon les besoins des trois phases méthodologiques. Cette section décrit précisément leur origine, leur structure et leurs caractéristiques quantitatives.

3.1 Corpus pour la classification thématique (Phase 1)

Le premier corpus regroupe des articles issus de PADI-web, organisés initialement par *maladie* (ex. Avian Influenza, African Swine Fever, Xylella Fastidiosa) et par *type de maladies* (santé animale (SA), santé publique (SP) et santé végétale (SV)). Les fichiers fournis contenaient les colonnes *title*, *text*, *host_type*, *disease*, *year* et *relevance*.

Plusieurs opérations ont été réalisées pour pré-traiter le corpus : suppression de doublons (environ 4200 articles), correction des titres manquants, suppression des textes corrompus (détection via *langdetect* et ratio de lettres), suppression des noms de maladies, fusion des fichiers en un corpus unique.

La taille du corpus initial est répartie de la manière suivante :

— SA : 93 398 textes
— SP : 8 819 textes
— SV : 8 168 textes

La longueur des textes est résumée ci-dessous :

— Moyenne : 230 mots
— Médiane : 180 mots
— Min/Max : 40–1200 mots

Après équilibrage UMAP + K-means détaillé dans les sections suivantes, chaque classe (SA, SP, SV) contient 8168 textes.

3.2 Corpus pour la détection de la veille syndromique (Phase 2)

Ce corpus vise à distinguer les articles VS (veille syndromique) des articles NVS (non veille syndromique). Les annotations ont été réalisées manuellement par des experts du CIRAD et de l'INRAE selon un protocole interne basé sur la présence de descriptions de symptômes, l'absence de diagnostic explicite et la détection de signaux faibles.

Les 2250 articles composant le corpus sont répartis de la manière suivante :

— VS : 225 textes ($\approx$ 10%)
— NVS : 2025 textes ($\approx$ 90%)

La longueur des textes est résumée ci-dessous :

— Moyenne : 260 mots
— Médiane : 210 mots
— Min/Max : 50–1500 mots

3.3 Annotations thématique et pertinence

Les annotations de données pour ce projet ont été réalisées par des expertes du domaine qui sont co-autrices de cet article et spécialisées en veille épidémiologique au sein des plateformes ESV[4] et ESA[5]. Un protocole multi-annotateurs avec mesure d'accord inter-annotateurs est envisagée dans nos futurs travaux.

3.4 Corpus augmenté (Phase 3)

Pour enrichir la classe minoritaire VS, une augmentation par rétro-traduction (anglais $\rightarrow$ allemand $\rightarrow$ anglais) a été appliquée et détaillée dans la section suivante.

La taille du corpus après augmentation est :

— NVS : 2241 textes
— VS : 498 textes après suppression des doublons une fois l'augmentation réalisée.

La longueur moyenne des textes augmentés est de 270 mots, avec une variabilité légèrement accrue due aux reformulations.

4 Méthodologie

La méthodologie adoptée dans ce travail est structurée en trois phases principales, correspondant aux objectifs du projet SURSY. Chaque phase produit un jeu de données spécifique, mis à disposition publiquement [15]. Le code

4. La Plateforme ESV (Plateforme d'Épidémiosurveillance en Santé Végétale) est en charge d'améliorer la surveillance de la santé des végétaux en France en associant 7 acteurs publics et privés nationaux reconnus en leur haut niveau d'expertise.

5. La Plateforme nationale d'Épidémiosurveillance en Santé Animale apporte un appui méthodologique et opérationnel aux services compétents de l'État et aux autres gestionnaires de dispositifs de surveillance pour la conception, le déploiement, l'animation, la valorisation et l'évaluation des dispositifs de surveillance sanitaire et biologique du territoire.

associé à l'ensemble des expérimentations est disponible en accès libre [17]. L'ensemble du pipeline méthodologique s'appuie à la fois sur des approches lexicales classiques et sur des modèles de langage pré-entraînés, conformément aux recommandations de travaux récents en classification de textes [6, 18].

Protocole d'évaluation : Le protocole d'évaluation appliqué dans les trois phases est adapté à la nature de chaque modèle. Pour les classifieurs ne nécessitant pas de sélection de modèle en cours d'entraînement (Random Forest, Naive Bayes, régression logistique), nous utilisons une séparation *train*/*test* stratifiée, avec une graine aléatoire fixée pour la reproductibilité. Pour les modèles entraînés de façon itérative avec arrêt anticipé ou sélection du meilleur *checkpoint* (MLP, XGBoost avec arrêt anticipé, et fine-tuning de modèles pré-entraînés tels que SBERT ou RoBERTa), nous utilisons une séparation *train*/*validation*/*test* stratifiée, où l'ensemble de validation sert à la sélection du meilleur modèle pendant l'entraînement et l'ensemble de test est réservé à l'évaluation finale. Dans les deux cas, la stratification préserve la distribution des classes entre les ensembles, ce qui est particulièrement important pour la Phase 2 où la classe VS est fortement minoritaire. Pour garantir la comparabilité entre les approches évaluées au sein d'une même phase, les indices des ensembles sont sauvegardés et réutilisés à l'identique pour chaque modèle comparé. Les métriques rapportées (accuracy, F1-score, AUC multi-classes pour la Phase 1, AUC-PR pour la Phase 2) sont choisies en fonction de la nature de la tâche et du déséquilibre éventuel des classes.

4.1 Phase 1 : Classification des articles par thématique

Cette première phase vise à structurer les données textuelles selon trois thématiques : santé publique (SP), santé animale (SA) et santé végétale (SV). Cette étape permet de réduire les biais liés à des sources ou maladies sur-représentées et constitue un prérequis essentiel pour la veille syndromique.

Prétraitement et vectorisation : Les articles issus de PADI-web sont nettoyés (suppression des balises HTML, normalisation, suppression des doublons). Deux types de représentations textuelles sont explorés : TF-IDF, couramment utilisé dans les approches lexicales [11], et SBERT (`all-MiniLM-L6-v2`), modèle de plongement sémantique performant pour la classification [18]. Cette version de SBERT a été retenue comme vectorizer dans ce projet pour son bon compromis entre qualité des représentations sémantiques et coût de calcul : six couches Transformer, 22M de paramètres, et des performances compétitives sur les tâches de similarité sémantique [18].

Équilibrage des données : La distribution initiale des thématiques (SA, SP, SV) étant fortement déséquilibrée, un équilibrage est nécessaire pour éviter que le modèle n'apprenne des biais structurels. Nous avons comparé deux stratégies :

- **Sous-échantillonnage par maladie** : limitation du nombre d'articles par couple thématique $\times$ maladie, afin de garantir une représentation équivalente de chaque maladie au sein de chaque thématique.
- **Combinaison UMAP + K-means** : UMAP (Uniform Manifold Approximation and Projection) réduit la dimensionnalité des embeddings SBERT tout en préservant la structure locale des données [16] ; K-means est ensuite appliqué dans l'espace réduit pour identifier des clusters cohérents, et un nombre équilibré d'articles est sélectionné dans chaque cluster afin de préserver la diversité sémantique tout en homogénéisant la distribution entre SA, SV et SP.

La stratégie d'équilibrage **UMAP + K-means** est particulièrement adaptée aux corpus textuels hétérogènes, comme le montrent des travaux similaires en classification thématique [1].

Ce type de combinaison « réduction de dimension + clustering » pour l'équilibrage de corpus textuels déséquilibrés est proche de travaux récents. Comme dans les travaux de Last *et al.* [12], qui proposent un ré-échantillonnage guidé par K-means (K-Means SMOTE) afin de mieux exploiter la structure du corpus, nous utilisons K-means pour regrouper les articles en zones sémantiques cohérentes avant d'opérer une sélection équilibrée. De manière similaire, Jamil *et al.* [10] ont récemment évalué l'association entre UMAP et des méthodes de rééquilibrage pour la classification de textes médicaux, confirmant l'intérêt d'une projection manifold-learning en amont de la gestion du déséquilibre. Notre approche s'inscrit dans cette lignée, tout en se distinguant par le fait que nous ne générons pas d'échantillons synthétiques : nous sélectionnons uniquement des articles réels, ce qui garantit la plausibilité linguistique des textes retenus. Cette stratégie évite les écueils bien connus de SMOTE dans le cadre textuel, où la génération d'échantillons synthétiques par interpolation dans l'espace des embeddings ne correspond pas nécessairement à des textes linguistiquement valides.

Modélisation de notre solution : Les embeddings SBERT équilibrés servent d'entrée à un perceptron multicouches (MLP) composé de deux couches denses avec activation ReLU. L'entraînement utilise un arrêt anticipé basé sur la perte de validation. Les performances sont évaluées via l'accuracy, le F1-score et l'AUC multi-classes.

Comparaison des approches : Pour justifier les choix méthodologiques décrits ci-dessus (combinaison UMAP + K-means pour l'équilibrage et SBERT + MLP pour la classification), nous avons croisé les deux stratégies d'équilibrage avec trois combinaisons vectorisation/classifieur. Les résultats obtenus sur l'ensemble de test sont rassemblés dans le Tableau 1. Les performances sont systématiquement meilleures avec UMAP + K-means qu'avec le sous-échantillonnage par maladie, et SBERT + MLP s'est révélé le plus stable sur des phrases courtes ou des formulations s'écartant du vocabulaire d'entraînement, ce qui a motivé notre choix final.

Équilibrage	Approche	Acc.	F1	AUC
Par maladie	TF-IDF + RF	0.92	0.92	0.98
Par maladie	SBERT + RF	0.93	0.93	0.98
Par maladie	SBERT + MLP	0.93	0.93	0.99
UMAP + K-means	TF-IDF + RF	0.98	0.98	1.00
UMAP + K-means	SBERT + RF	0.97	0.97	1.00
UMAP + K-means	SBERT + MLP *(retenu)*	**0.98**	**0.98**	**1.00**

TABLE 1 – Comparaison des approches testées pour la classification thématique (Phase 1), sur l'ensemble de test. F1 et AUC sont calculés en moyenne macro (*one-vs-rest* pour l'AUC sur les trois thématiques SA, SP, SV). La combinaison UMAP + K-means avec SBERT + MLP a été retenue pour son comportement plus stable sur des formulations n'appartenant pas strictement au vocabulaire d'entraînement.

4.2 Phase 2 : Détection des articles de veille syndromique

La deuxième phase constitue le cœur du projet et vise à distinguer les articles de veille syndromique (VS) des articles non veille syndromiques (NVS). Cette tâche est formulée comme une classification binaire fortement déséquilibrée (environ 10% VS).

Prétraitement et vectorisation : Les textes sont normalisés puis vectorisés selon deux approches complémentaires :

- **TF-IDF + XGBoost**, approche rapide et assez robuste pour des tests préliminaires, mais limitée par sa dépendance au vocabulaire [5].
- **SBERT fine-tuné**, permettant de capturer des formulations implicites et des synonymes, particulièrement utiles dans la veille syndromique où les signaux sont rarement explicites.

Plusieurs combinaisons vectorisation/classifieur ont été évaluées à titre comparatif avant d'arrêter le choix du modèle SBERT `all-MiniLM-L6-v2` ; les résultats obtenus sur l'ensemble de test pour la classe minoritaire VS sont résumés dans le Tableau 2. Deux constats se dégagent : d'une part, les approches hybrides SBERT + classifieur superficiel exploitent mal la richesse des embeddings sur un corpus aussi déséquilibré ; d'autre part, TF-IDF + XGBoost obtient de bons scores mais reste très sensible aux synonymes, comme le confirment les tests qualitatifs sur phrases courtes (voir Section 5). Le fine-tuning complet de SBERT, où tous les poids du modèle sont mis à jour via la classe `Trainer` de Hugging Face, est l'approche retenue pour le déploiement. RoBERTa a également été fine-tuné à titre exploratoire, avec des performances proches (F1-macro $\approx$ 0.93), mais pour un coût d'entraînement plus élevé qui ne justifiait pas son adoption.

4.3 Phase 3 : Augmentation de données textuelles

La troisième phase explore l'augmentation de données pour enrichir la classe VS. L'objectif est d'améliorer la robustesse du modèle face à la variabilité lexicale.

Rétro-traduction : Nous utilisons la rétro-traduction (anglais $\rightarrow$ allemand $\rightarrow$ anglais), technique couramment employée pour générer des paraphrases [8]. Afin de choisir la langue pivot, nous avons comparé deux configurations : l'anglais $\rightarrow$ français $\rightarrow$ anglais et l'anglais $\rightarrow$ allemand $\rightarrow$ anglais. Les deux configurations améliorent les performances par rapport au modèle de la Phase 2 sans augmentation, mais l'allemand offre un F1-score légèrement supérieur sur la classe minoritaire VS (voir Section 5), et a donc été retenu comme langue pivot. Cette méthode permet d'introduire des variations lexicales tout en conservant le sens.

Réentraînement et évaluation : Les textes augmentés sont intégrés au corpus d'entraînement du modèle SBERT. L'impact est mesuré via l'évolution de l'AUC-PR et la stabilité des prédictions sur des phrases tests.

Approche	Préc. (VS)	Rapp. (VS)	F1 (VS)	AUC-PR (VS)
TF-IDF + Naive Bayes	0.25	0.06	0.10	0.80
TF-IDF + Rég. logistique	0.96	0.55	0.70	0.80
SBERT + Rég. logistique	0.52	0.83	0.64	0.72
SBERT + XGBoost	0.86	0.52	0.65	0.81
TF-IDF + XGBoost	1.00	0.87	0.93	0.98
SBERT *fine-tuné* (retenu)	**0.95**	**0.87**	**0.91**	**0.95**

TABLE 2 – Comparaison des approches testées pour la détection des articles de veille syndromique (Phase 2), sur le sous-ensemble de test (225 articles, dont 23 VS et 202 NVS). Les valeurs reportées correspondent à la précision, au rappel, au F1-score et à l'AUC-PR de la classe minoritaire VS. Le fine-tuning complet de SBERT est l'approche retenue pour sa robustesse aux reformulations, confirmée par les tests qualitatifs (Section 5).

5 Résultats

Cette section présente les principaux résultats obtenus lors des différentes phases du projet, en cohérence avec la méthodologie décrite précédemment. Les résultats sont présentés de manière synthétique, en mettant l'accent sur ceux qui ont été retenus durant les expérimentations.

5.1 Résultats de la classification thématique

Pour la classification thématique des articles (santé publique, santé animale et santé végétale), nous avons retenu l'approche ayant fourni les meilleures performances, fondée sur l'apprentissage par transfert en utilisant les données équilibrées avec UMAP et K-means. Les textes sont d'abord représentés à l'aide du modèle SBERT *all-MiniLM-L6-v2*, puis les embeddings obtenus sont utilisés comme entrée d'un modèle séquentiel léger de type perceptron multicouches.

Le modèle converge après 67 époques (Figure 1a), avec des performances élevées et stables sur les ensembles d'entraînement et de validation. Les scores d'accuracy et de F1-score dépassent 97% sur l'ensemble de validation. L'évaluation sur l'ensemble de test confirme ces résultats, avec une AUC multi-classes de 0.99 (Figure 1b) et une très faible confusion entre les thématiques (Figure 1c).

Enfin, un test de prédiction sur une phrase courte (*"A new disease affecting animals was discovered this year."*)

montre une probabilité de 0.98 pour la classe correcte (Figure 1d), confirmant la robustesse du modèle.
L'ensemble de ces résultats est synthétisé en Figure 1.

5.2 Résultats de la détection des articles de veille syndromique

L'objectif de cette deuxième phase est de classifier automatiquement les articles selon leur type : veille syndromique (VS) ou non veille syndromique (NVS). Cette tâche est rendue particulièrement complexe par le fort déséquilibre des données, reflétant la réalité opérationnelle, avec une proportion d'environ 10% d'articles VS contre 90% d'articles NVS.
Parmi les multiples approches abordées durant cette phase, nous avons retenu deux méthodes, bien qu'une seule soit véritablement fiable. Nous les avons retenues pour des raisons particulières :

- **TF-IDF + XGBOOST** (Figure 2) : vectorisation avec TF-IDF et classification avec XGBoost. Cette approche est rapide en termes de calcul et donne d'assez bonnes performances, mais elle reste limitée par l'effet statique du vectorizer, très sensible aux synonymes qui ne figurent pas dans son vocabulaire.
- **Fine-tuning de SBERT** (Figure 3) : après un temps d'entraînement plus conséquent, cette approche produit un excellent rapport de classification ainsi qu'un score AUC-PR de 0.95. De plus, le meilleur modèle sauvegardé parvient à bien classer les phrases contenant des synonymes ou des formulations variées liées à la veille syndromique, ce qui n'était pas le cas de TF-IDF + XGBoost.

Nous avons conservé les deux approches car, pour un test rapide, la première est suffisante, tandis que la seconde est plus adaptée à un déploiement opérationnel nécessitant davantage de robustesse.
Des tests complémentaires réalisés sur des phrases courtes confirment que le fine-tuning de SBERT offre une meilleure capacité de généralisation. Contrairement aux modèles basés sur TF-IDF, le modèle fine-tuné sur SBERT parvient à identifier des articles relevant de la veille syndromique même lorsque les expressions employées reposent sur des synonymes ou des formulations implicites. Cette capacité à capturer le sens global des textes constitue un atout majeur pour la veille syndromique, où les signaux recherchés sont rarement explicitement formulés.
Le détail des performances des deux approches est présenté dans les Figures 2 (TF-IDF + XGBoost) et 3 (SBERT fine-tuné).

5.3 Résultats de l'augmentation de données textuelles

La troisième phase du projet a pour objectif d'explorer l'impact de l'augmentation de données textuelles sur la détection des articles de veille syndromique. Cette étape vise à enrichir la classe minoritaire (VS), fortement sous-représentée, afin d'améliorer la robustesse et la stabilité des modèles de classification.

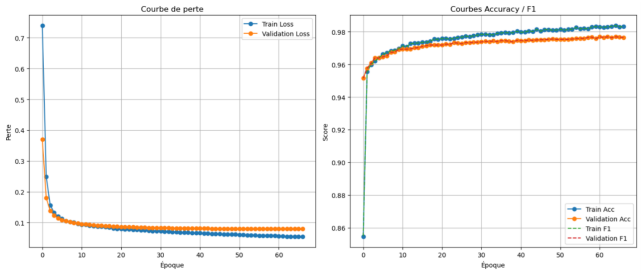
Configuration	AUC-PR	F1 (VS)
Sans augmentation (Phase 2)	0.95	0.91
Augmentation, pivot français	0.97	0.91
Augmentation, pivot allemand *(retenu)*	**0.98**	0.91

TABLE 3 – Comparaison de l'effet de la langue pivot pour la rétro-traduction (Phase 3), évalué via le fine-tuning de SBERT sur l'ensemble de test. Les deux configurations améliorent les performances par rapport à la Phase 2 sans augmentation, l'allemand offrant l'AUC-PR le plus élevé.

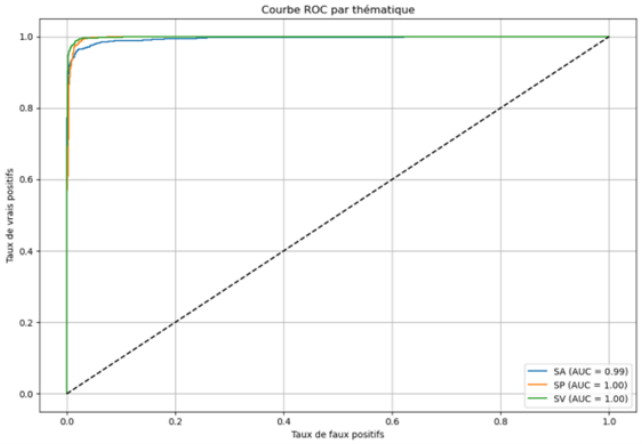
Après avoir obtenu le jeu de données par la méthode de rétro-traduction, le modèle SBERT a été réentraîné sur ces derniers via le fine-tuning. Nous avons comparé deux langues pivots : le français et l'allemand (Tableau 3). Les deux configurations améliorent les performances par rapport au modèle de la Phase 2 sans augmentation et augmentent globalement les probabilités de prédiction pour la classe VS, ce qui confirme l'apport de l'augmentation. L'allemand offre en particulier l'AUC-PR le plus élevé (0.98 contre 0.97 avec le français, voir Figure 4a), et a donc été retenu pour la suite. Le détail des métriques par classe sur l'ensemble de test est présenté en Figure 4b. Les tests réalisés sur des phrases courtes confirment une meilleure assurance du modèle dans ses décisions, suggérant un effet positif de l'augmentation sur la capacité de généralisation.
Ces premiers résultats indiquent que l'augmentation de données textuelles constitue une piste prometteuse pour renforcer la détection des articles de veille syndromique. Cette phase reste néanmoins exploratoire et ouvre des perspectives vers l'utilisation de techniques plus avancées, telles que l'exploitation de plusieurs langues pivots ou le recours à des modèles génératifs récents comme le RAG.
L'ensemble des résultats de cette phase est synthétisé en Figure 4.

6 Discussion

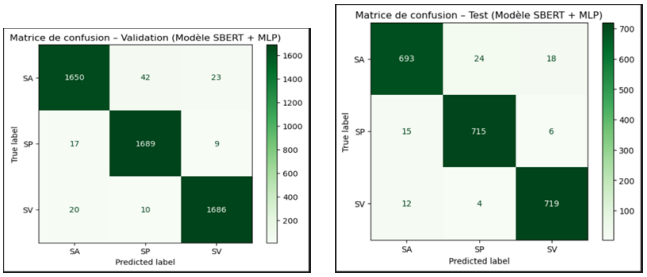
Les résultats obtenus au cours des différentes phases du projet mettent en évidence plusieurs enseignements importants pour le développement de systèmes de veille syndromique en santé végétale.
Tout d'abord, la phase de classification thématique confirme l'intérêt de structurer les données en amont afin de réduire les biais liés à des maladies, des sources ou des contextes dominants. L'utilisation de représentations sémantiques basées sur SBERT, combinée à une stratégie d'équilibrage des données (UMAP et K-means), permet d'obtenir des performances élevées et stables. Cette étape apparaît essentielle pour garantir une meilleure généralisation des modèles dans un contexte multi-domaines, et constitue un prérequis solide pour les phases ultérieures de veille syndromique.
La comparaison menée lors de la phase de détection des articles de veille syndromique met clairement en évidence les limites des approches purement lexicales, telles que TF-IDF combiné à XGBoost. Bien que rapides et efficaces pour des tests préliminaires, ces méthodes restent fortement dépendantes du vocabulaire observé lors de l'apprentissage



(a) Courbes d'apprentissage (loss / accuracy).



(b) Courbes ROC multi-classes (one-vs-rest).

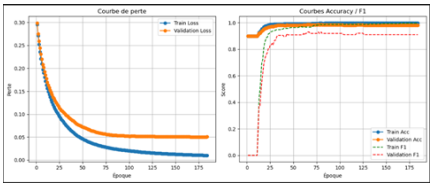


(c) Matrices de confusion : validation (gauche) et test (droite).

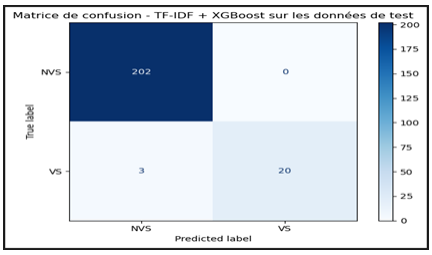
Texte	*A new disease affecting animals was discovered this year*
Thématique prédite	SA
Score de confiance	0.9870
Probabilités	
SA	0.9870
SP	0.0102
SV	0.0028

(d) Exemple de prédiction.

FIGURE 1 – Synthèse des résultats de la phase 1 (classification thématique).



(a) Courbe d'apprentissage (TF-IDF + XGBoost).

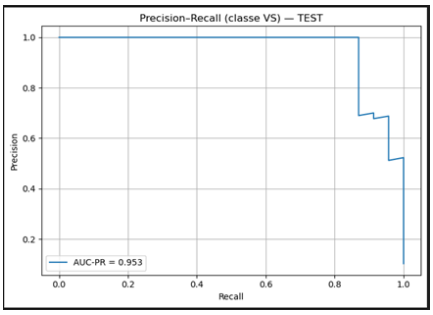


(b) Matrice de confusion (test).

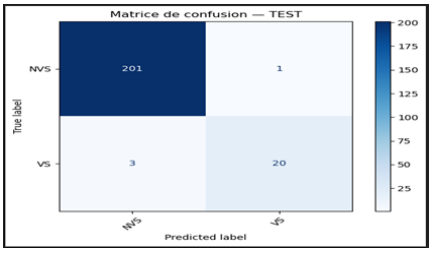
Classe	Prec.	Rapp.	F1	Supp.
NVS	0.99	1.00	0.99	202
VS	1.00	0.87	0.93	23
Acc.		0.99		
Macro	0.99	0.91	0.96	225
Weighted	0.99	0.99	0.99	225

(c) Synthèse des métriques (TF-IDF + XGBoost).

FIGURE 2 – Résultats TF-IDF + XGBoost : (a) courbe d'apprentissage, (b) matrice de confusion, (c) synthèse des métriques.



(a) PR (classe VS).



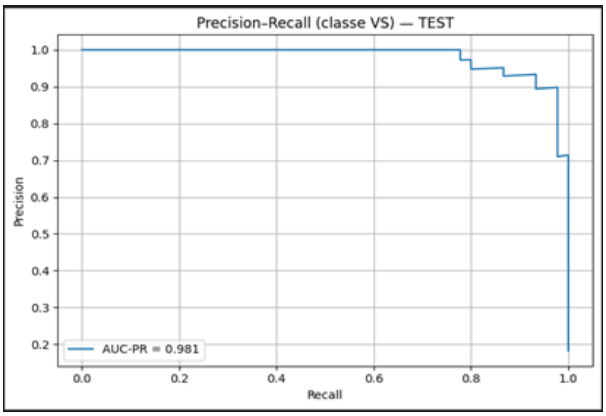
(b) Confusion (test).

Classe	Prec.	Rapp.	F1	Supp.
NVS	0.99	1.00	0.99	202
VS	0.95	0.87	0.91	23
Acc.		0.98		
Macro	0.97	0.93	0.95	225
Weighted	0.98	0.98	0.98	225

(c) Synthèse métriques.

FIGURE 3 – Résultats de SBERT fine-tuné.

et ont des difficultés à capturer des formulations implicites ou synonymiques. À l'inverse, les approches sémantiques reposant sur le fine-tuning de modèles de langage pré-entraînés, comme SBERT, montrent une meilleure capacité à saisir le sens global des textes et à détecter des signaux faibles, caractéristique essentielle dans un contexte



(a) Courbe Precision–Recall (classe VS).

Classe	Prec.	Rapp.	F1	Supp.
NVS	0.98	0.99	0.98	202
VS	0.93	0.89	0.91	45
Acc.		0.97		
Macro	0.95	0.94	0.94	247
Weighted	0.97	0.97	0.97	247

(b) Synthèse des métriques (test).

FIGURE 4 – Résultats de la phase 3 : (a) courbe Precision–Recall sur la classe VS et (b) synthèse des métriques sur l'ensemble de test, après augmentation par rétro-traduction (allemand comme langue pivot). L'augmentation conduit à un test set contenant 45 articles VS (contre 23 en Phase 2 sans augmentation).

de veille syndromique.

Néanmoins, la détection des articles de veille syndromique demeure une tâche complexe en raison du fort déséquilibre entre les classes. Les résultats obtenus confirment que ce déséquilibre constitue un frein majeur à l'apprentissage de modèles robustes, en particulier pour la classe minoritaire. Bien que les représentations sémantiques améliorent la cohérence et la stabilité des prédictions, elles restent sensibles à la quantité limitée de données disponibles en santé végétale.

Les premiers résultats issus de la phase exploratoire d'augmentation de données textuelles suggèrent que l'enrichissement artificiel de la classe minoritaire constitue une piste prometteuse. L'augmentation par rétro-traduction introduit une variabilité lexicale supplémentaire qui se traduit par une amélioration modérée des performances et une meilleure assurance du modèle dans ses prédictions. Toutefois, cette approche reste limitée par la simplicité des transformations appliquées et devra être approfondie afin de générer des textes plus diversifiés et plus représentatifs des situations réelles de veille.

Enfin, des expérimentations complémentaires ont été réalisées pour tester les capacités de transfert des modèles en apprenant sur des données en SA (resp. SV) et testant en SV (resp. SA). Le Tableau 4 montre des résultats satisfaisants quant aux capacités de transfert des modèles. Ceci peut s'expliquer par un vocabulaire utilisé qui peut être relativement commun pour décrire des évenements syndromiques. Ceci met en évidence l'intérêt de renforcer les liens entre la veille en santé animale, santé publique et la veille en santé végétale. Les similarités observées dans les descriptions de symptômes ouvrent des perspectives intéressantes pour le transfert de connaissances entre domaines, notamment dans le contexte de maladies émergentes ou de phénomènes de type « maladie X » (maladies inconnues), dont les manifestations peuvent évoluer rapidement sous l'effet du changement climatique.

Train : SV / Test : SA	F1
NVS	1
VS	0.96
Train : SA / Test : SV	**F1**
NVS	0.98
VS	0.81

TABLE 4 – Résultats de transfert fondés sur la méthode TF-IDF+XGBOOST (notons que les noms de maladie ont été supprimés comme pré-traitement pour éviter de potentiels bias).

7 Conclusion et perspectives

Ce travail mené dans le cadre du projet SURSY [6] a exploré l'apport des modèles de langue pour la détection automatique d'articles de veille syndromique en santé végétale, dans un contexte marqué par des données rares, hétérogènes et fortement déséquilibrées. Les résultats obtenus montrent que les approches sémantiques, et en particulier les modèles SBERT fine-tunés, sont mieux adaptées que les méthodes lexicales classiques pour capturer des signaux faibles et implicites caractéristiques de la veille syndromique.
L'étude met également en évidence l'intérêt de l'augmentation de données textuelles par rétro-traduction pour renforcer la robustesse des modèles, notamment vis-à-vis de la classe minoritaire. Bien que les gains observés restent modérés, cette approche contribue à améliorer la stabilité des prédictions et la confiance du modèle.
Les perspectives de ce travail portent sur l'exploration de techniques d'augmentation plus avancées, incluant l'usage de plusieurs langues pivots ainsi que l'intégration de modèles génératifs récents, tels que les approches basées sur le RAG. À plus long terme, l'extension de ces méthodes à des scénarios de veille multi-domaines, en lien avec la santé animale et publique (contexte One Health), constitue une voie prometteuse pour renforcer la détection précoce de maladies émergentes dans un contexte de changement

6. https ://clapas.umontpellier.fr/projets/projet-sursy/

climatique et de mobilité humaine.

Remerciements

Les auteurs tiennent à remercier les bailleurs de fonds CLAPAS [7] (Collaborative Local Research Actions on Plant health and AgroSystems) – Université de Montpellier, la DGAL (Direction générale de l'alimentation) ainsi que le projet ANR BEYOND (20-PCPA-0002) dont le soutien a contribué au bon déroulement de ce travail. Enfin, les auteurs remercient Sarah Valentin (Cirad, TETIS) qui a mis à disposition un corpus dédié la veille syndromique.

Références

[1] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining Text Data*, pages 163–222. Springer, 2012.

[2] Elena Arsevska, Mathieu Roche, Pascal Hendrikx, and Renaud Lancelot. From syndromic surveillance to event-based surveillance : monitoring animal health events using online news. *Preventive Veterinary Medicine*, 160 :135–144, 2018.

[3] John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. Healthmap : global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2) :150–157, 2008.

[4] Martin Carrion and Lawrence C Madoff. Promedmail : 22 years of digital disease detection. *Online Journal of Public Health Informatics*, 9(1), 2017.

[5] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[7] Marc Dion, Peter AbdelMalik, and Abla Mawudeku. Gphin : Global public health intelligence network. *Online Journal of Public Health Informatics*, 7(1), 2015.

[8] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of EMNLP*, pages 489–500, 2018.

[9] Haibo He and Edwardo Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9) :1263–1284, 2009.

[10] Arslan Jamil, Muhammad Kashif Hanif, Muhammad Umer Sarwar, and Muhammad Irfan Khan. An empirical evaluation of dimensionality reduction and class balancing for medical text classification. *Scientific Reports*, 16(815), 2026.

7. https ://clapas.umontpellier.fr/

[11] Thorsten Joachims. Text categorization with support vector machines : Learning with many relevant features. In *Proceedings of ECML*, pages 137–142, 1998.

[12] Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on K-Means and SMOTE. *arXiv preprint arXiv :1711.00837*, 2017.

[13] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[14] Yinhan Liu et al. Roberta : A robustly optimized bert pretraining approach. In *arXiv preprint arXiv :1907.11692*, 2019.

[15] Nkounghawe Rosalie ; Meltatagia Paulin ; Pieretti Isabelle ; Trevenec Carlène ; Roche Mathieu. Sursy data : Textual data for syndromic surveillance. https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/MYMPSO, 2025. Dataverse.

[16] Leland McInnes, John Healy, and James Melville. Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*, 2018.

[17] Rosalie NKOUNGHAWE. Sursy : Projet_sursy. https://github.com/rosalie0309/Projet_SURSY, 2025.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, 2019.

[19] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. In *PLOS ONE*, volume 10, page e0118432. Public Library of Science, 2015.

[20] Sarah Valentin, Elena Arsevska, Alize Mercier, Sylvain Falala, Julien Rabatel, Renaud Lancelot, and Mathieu Roche. Padi-web : An event-based surveillance system for detecting, classifying and processing online news. In Zygmunt Vetulani, Patrick Paroubek, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 87–101, Cham, 2020. Springer International Publishing.

[21] Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, and Mathieu Roche. Padi-web 3.0 : A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, 13 :100357, 2021.

[22] Jason Wei and Kai Zou. Eda : Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP*, 2019.