Extraction and combination of epidemiological information from informal sources for animal infectious diseases surveillance

Sarah Valentin

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale Information Structures Systèmes (I2S)

Unité de recherche UMR TETIS et ASTRE

Extraction et combinaison d'informations épidémiologiques à partir de sources informelles pour la veille des maladies infectieuses animales

Présentée par Sarah Valentin
Le 24 septembre 2020

Sous la direction de Mathieu Roche et Renaud Lancelot

Devant le jury composé de

Carmen Gervet, Professeur, Université de Montpellier (France)	Présidente
Diana Inkpen, Professeur, Université d'Ottawa (Canada)	Rapporteur
Bruno Martins, Professeur Associé, Université de Lisbonne (Portugal)	Examinateur
Gaël Dias, Professeur, Université de Caen (France)	Examinatrice
Daniela Paolotti, Chercheuse, ISI, Turin (Italie)	Examinatrice
Renaud Lancelot, Chercheur, CIRAD, Montpellier (France)	Co-directeur
Mathieu Roche, Chercheur HDR, CIRAD, Montpellier (France)	Directeur
Elena Arsevska, Chercheuse, CIRAD, Montpellier (France)	Invitée

UNIVERSITÉ DE MONTPELLIER

A THESIS SUBMITTED IN CONFORMITY WITH THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
UNIVERSITY OF MONTPELLIER

In Computer Science

Doctoral School Information Structures Systèmes (I2S)

Research Unit UMR TETIS

Extraction and combination of epidemiological information from informal sources for animal infectious diseases surveillance

Sarah Valentin
Septembre 24, 2020

Under the supervision of Mathieu Roche and Renaud Lancelot

Presented to a jury composed of:

Diana Inkpen, Professor, University of Ottawa (Canada)	Rapporteur
Bruno Martins, Associate Professor, University of Lisbon (Portugal)	Rapporteur
Gaël Dias, Professor, University of Caen (France)	Examiner
Daniela Paolotti, Research scientist, ISI, Turin (Italy)	Examiner
Carmen Gervet, Professor, University of Montpellier (France)	Examiner
Renaud Lancelot, Research scientist, CIRAD, Montpellier (France)	Co- supervisor
Mathieu Roche, Research scientist, CIRAD, Montpellier (France)	Supervisor
Elena Arsevska, Research scientist, CIRAD, Montpellier (France)	Guest

REMERCIEMENTS

Je crains que malgré tous les efforts et les heures de travail apportés à ce manuscrit, il reste quelque peu incomplet. En effet, s'il tente de développer le plus précisément possible mes travaux, il ne reflète que très partiellement la richesse de tout ce que j'ai appris et l'ouverture que m'a apportée la thèse. Cette ouverture, si elle est bien sûr en partie scientifique, est avant tout personnelle. Je la dois à toutes les personnes avec qui j'ai pu échanger et partager du temps au cours de ces trois ans. Toutes ne sont pas citées dans ces remerciements non exhaustifs, mais ont contribué à ce que je ressorte grandie de cette aventure (malheureusement, pas physiquement).

Dans le parcours parfois fastidieux de la thèse, j'ai eu la chance et l'honneur d'avoir été accompagnée par une équipe d'encadrants hors pair. Mathieu, Renaud, Elena, c'était un plaisir de travailler avec vous durant ces trois ans. Merci de m'avoir transmis, chacun à votre manière, votre passion de la recherche et de vos domaines respectifs. J'espère pouvoir continuer à échanger avec vous longtemps, que ce soit sur le plan professionnel comme personnel.

Je tiens à exprimer ma profonde gratitude à mon directeur de thèse, Mathieu Roche, pour avoir partagé avec moi son expertise de l'informatique et de la fouille de texte, toujours dans la bonne humeur. Son implication dans mon travail de thèse à toutes heures, son exigence scientifique, son soutien, et par-dessus tout sa bienveillance, m'ont été essentiels pour aller au bout de l'aventure. Mes sincères remerciements à mon co-directeur, Renaud Lancelot, pour les échanges extrêmement stimulants que nous avons pu avoir, que ce soit au sujet de la thèse ou sur d'autres sujets scientifiques. Ces discussions m'ont poussée à prendre du recul sur mon travail, tâche ardue dans le brouillard de la thèse.

Elena, de collègue, tu es devenue une amie. Je me souviens de nos premiers échanges téléphoniques, lorsque tu m'encourageais avec ton énergie déjà communicative à postuler à la thèse ; j'ai évidemment été convaincue :). Un immense merci pour tout l'enthousiasme et l'aide que tu as consacrés à mon travail, ces trois ans n'auraient pas été les mêmes sans toi. Tes valeurs humaines et scientifiques et ta vision de la recherche, ont été, et sont toujours, une grande source d'admiration et d'inspiration pour moi. Je suis très heureuse que nous puissions continuer à travailler ensemble, et je te souhaite le meilleur pour l'aventure MOOD et la suite.

Au-delà de mes encadrants, je remercie chaleureusement les membres de l'équipe de PADI-web : Sylvain, ambassadeur de choc de PADI-web, pour avoir été un super co-bureau et avoir supporté mes débats houleux avec Geoffrey, Julien, pour avoir construit, transformé et rendu PADI-web si beau (quel chemin depuis la v1 !) et pour m'avoir aidée à comprendre ses rouages parfois obscurs, Alizé, merci de m'avoir initiée à la veille en santé animale, bientôt la libération pour toi :), Jocelyn, et toutes les personnes ayant apporté leur brique à l'édifice. Merci aux membres de la plateforme ESA, Cécile, Julien et Didier, pour leur collaboration dans les différents développements de PADI-web.

I am grateful to the members of my thesis jury: Prof. Carmen Gervet, Prof. Gaël Dias, Dr. Daniela Paolotti. A special thanks to my *rapporteurs*, Prof. Diana Inkpen and Prof. Bruno Martins, for their insigthful comments which allowed me to significantly improve my manuscript. Je remercie également sincèrement les membres de mes comités de thèse : Didier Calavas, Pascal Hendricks, Annie Château et Maguelonne Teisseire pour vos commentaires constructifs et encourageants lors de nos réunions annuelles.

Dans le cadre de la collaboration Moriskin/PADI-web, je remercie Valérie et Aline, pour nos échanges enrichissants et toujours très agréables. Merci de vous être prêtées avec enthousiasme au jeu parfois fastidieux de l'annotation ! Je remercie également les membres du LIRMM pour nos travaux autour d'EpidNews : Arnaud, Rohan, Samiha, Alexis, et Pascal.

I thank Goran Nenadic and Mercedes Arguello Casteleiro from the informatics department from the University of Manchester. Thank you for your welcoming me and guiding me in the discovery of word embedding models.

Je remercie également Cédric, pour m'avoir permis d'utiliser les fonctionnalités d'EMVISTA dans le cadre de mes travaux.

S'il a parfois été ardu dans le cadre de la thèse, le grand écart entre informatique et épidémiologie a été avant tout source de supers rencontres avec des personnes d'horizon et de parcours différents. Je remercie donc tous les doctorants et autres membres des équipes d'ASTRE et TETIS, pour les événements organisés ensemble, les pauses café, et les échanges, scientifiques ou non, que nous avons eus.

Un merci particulier à Jacques, pour ton aide et tes explications toujours patientes dans mon apprentissage laborieux de Python et de la fouille de texte, et ce même dans tes moments de rush. Tu auras contribué jusqu'aux dernières heures, désolée pour ce samedi matin à débuguer du LaTeX. Fred, je suis super heureuse de t'avoir rencontrée, merci pour les bouffées d'air confinés, et pleins de courage pour la fin de la thèse. Merci à mes co-bureaux TETISiens, Milo, pour ta bonne humeur de tous les instants, Karun, good luck for the future ! Merci à deux astres parmi les ASTRE, Mariline et Claire, hâte de trinquer à votre libération ! Gab, merci d'avoir apporté un peu de folie dans les couloirs du bâtiment E, c'est devenu un peu trop calme après ton départ.

Merci à Laure pour avoir été une super partenaire d'escalade et d'escapades en montagne.

Namrata, merci de m'avoir permis de me lancer dans le grand bain de l'amphithéâtre.

Arthur, Maxime, merci de m'avoir supportée en temps de rédaction ET de confinement. Maxime, j'espère que tu me pardonneras un jour d'avoir travaillé le 1er mai... Promis c'était une exception !

Léa-chou, merci pour ta joie de vivre et les fous rires qui aident à traverser ces temps un peu moroses, et d'être à l'instar de ta cuisine, douce, pétillante et exotique !

Un immense merci aux colocs et amis de la rue des Pâquerettes : Mathilde et Mathilde, Elsa, Eliott, Lucas, Jo, Delphine, Yves, Antoine, Steph, Léa, et tous les autres. Je me suis installée dans

cette maison en commençant ma thèse pour quelques mois, et j'y ai finalement passé 3 ans. Votre folie collective a été une super bouffée d'air frais.

Mes amis Jungle toucheurs et auzevillois, vous savez à quel point vous êtes des amis précieux. C'est un plaisir d'avancer dans la vie avec vous et d'évoluer, chacun à notre manière. Comme vous avez eu droit aux honneurs il y a quelques années de ça, place aux nouveaux venus : Nino, Enoha, Alicia, et Roméo, vivement qu'on puisse danser le Kuduro avec vous tous !

Merci à ma famille, vous me supportez et m'encouragez quels que soient mes choix, c'est surement vous les plus méritants. Marge, ma sœur, je suis fière de toi. Binto, petit bout du Gabon, merci de ta fidélité féline.

En recherche comme dans la vie, il arrive que l'on trouve ce que l'on ne cherchait pas. Arthur, merci d'avoir toujours su comment me redonner confiance et d'avoir supporté mes montées de stress intense. Merci d'avoir passé des heures à relire, corriger, et mettre en forme ce manuscrit, tu le connais mieux que moi maintenant. Mais surtout, parce que l'essentiel est ailleurs, merci de me faire rire, rêver, réfléchir autrement, et pour tout le reste. Il me tarde que nous tournions cette dernière page ensemble.

Merci encore à tous, Sarah.

PUBLICATIONS, DATA AND SOFTWARE

International Journals

Valentin S, Mercier A, Roche M, Lancelot R, Arsevska E. Monitoring online media reports for early detection of unknown diseases: insight from a retrospective study of COVID-19 emergence. *Transboundary and Emerging Diseases*. 2020, to appear.
https://onlinelibrary.wiley.com/doi/10.1111/tbed.13738 (*Available on Wiley Online Library*).

Valentin S, Arsevska E, Falala S, de Goër J, Lancelot R, Mercier A, Rabatel J, Roche M. PADI-web: a multilingual web-based biosurveillance system for the monitoring of animal infectious diseases. *Computer and Electronics for Agriculture*, Elsevier, 169: 105163. 2020.
http://agritrop.cirad.fr/594604/.

Goel R, **Valentin S**, Delaforge A, Fadloun S, Sallaberry A, Roche M, Poncelet P. EpidNews: An epidemiological news explorer for monitoring animal diseases. *Journal of Computer Languages*, Elsevier, 56: 100936. 2020.
http://agritrop.cirad.fr/594528/.

Arsevska E, **Valentin S**, Rabatel J, de Herve JG, Falala S, Lancelot R, Roche M. Web Monitoring of Emerging Animal Infectious Diseases Integrated in the French Epidemic Intelligence System in Animal Health. *PLOS One*; 13: e0199960. 2018.
Selected as one of the "Best paper" of IMIA Yearbook of Medical Informatics 2019: https://www.thieme-connect.com/products/ejournals/pdf/10.1055/s-0039-1677939.pdf/.
http://agritrop.cirad.fr/588533/.

Conference Proceedings

Valentin S, Arsevska E., Mercier A., Falala S., Rabatel J., Lancelot R., Roche M. PADI-web: an event-based surveillance system for detecting, classifying and processing online news. In: *Post-Proceedings of 8th Language and Technology Conference, LTC 2017, Lecture Notes in Computer Science - Lecture Notes in Artificial Intelligence - Springer*. 2020, to appear.

Valentin S, Lancelot R, Roche M. Automated Processing of Multilingual Online News for the Monitoring of Animal Infectious Diseases.In: *2nd MultilingualBIO: Multilingual Biomedical Text Processing Workshop - LREC, p33-36*. 2020.
http://agritrop.cirad.fr/595745/.

Valentin S, Lancelot R, Roche M, Arsevska E, Mercier A, Rabatel J, Falala S, de Goër J. PADI-web : un système automatique multilingue pour la veille sanitaire internationale en santé animale. In: *PFIA2019, Plate-Forme Intelligence Artificielle*. Toulouse (France), 2019.

Valentin S, Roche M, Lancelot R, Arsevska E, Mercier A, Rabatel J, Falala S, de Goër J. PADI-web: un système automatique multilingue pour la veille sanitaire internationale en santé animale. In: 30èmes Journées Francophones d'Ingénierie des Connaissances (IC), pages 235-239, AFIA. Toulouse (France), 2019.
https://agritrop.cirad.fr/596217/.

Goel R, Fadloun S, **Valentin S**, Sallaberry A, Roche M, Poncelet P. EpidNews: An epidemiological news explorer for monitoring animal diseases. In: *VINCI 18, 11th International Symposium on Visual Information Communication and Interaction; p1-8*. Vaxjo (Sweden). 2018.
http://agritrop.cirad.fr/588534/.

Valentin S, Roche M, Lancelot R. How to combine spatio-temporal and thematic features in online news for enhanced animal disease surveillance? In: *KES2018, International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Belgrade (Serbia), 2018.
http://agritrop.cirad.fr/588729/.

Conference Posters

Valentin S, Arsevska E, Mercier A, Lancelot R, Roche M. *Analysis of the dissemination of the information between sources used by event-based surveillance systems.* In: 4th International Conference on Animal Health Surveillance, ICAHS 2020. November 11-13, Copenhagen (Denmark), 2020.

Valentin S, Arsevska E, Mercier A, Lancelot R, Roche M. *An automatic animal diseases surveillance system based on textual media analysis.* In: InnovSur 2018: Innovation in Health Surveillance. International Forum. May 16, Montpellier (France), 2018.

Under Review

Valentin S, de Waele V, Vilan A, Arsevska E, Lancelot R, Roche M. Elaboration and validation of a framework for fine-grained annotation of news articles for event-based surveillance in animal health. *Language Resources and Evaluation.*

Data

Valentin S, Mercier A, Lancelot R; Roche M; Arsevska E, "PADI-web COVID-19 corpus: news articles manually labelled", doi:10.18167/DVN1/MSLEFC, CIRAD Dataverse, 2020.
https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/MSLEFC.

Valentin S, De Waele V, Vilain A., Arsevska E., Lancelot R, Roche M, "Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus", doi:10.18167/DVN1/YGAKNB, CIRAD Dataverse, 2019.
https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/YGAKNB.

Arsevska E, **Valentin S**, Rabatel J, de Goër de Hervé J, Falala S, Lancelot R, Roche M, "PADI-web dataset manually evaluated (January 1 - June 28 2016)", doi:10.18167/DVN1/JZM34U, CIRAD Dataverse, 2017.
https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/JZM34U.

Software

PADI-Web. Rabatel J, Falala S, Arsevska E, **Valentin S**, Mercier A, De Goër J., Lancelot R, Roche M. IDDN1.FR2.0013.1300234.0005.S6 .C7 .20198.00, APP (Agence pour la protection des programmes). 22/03/2019.

I did my PhD thesis at the French Agricultural Research Centre for International Development (Cirad), in the TETIS and ASTRE units. PADI-web project is part of the French Platform for Animal Health Surveillance (ESA Platform) activities, led by Cirad and French Agency for Food, Environmental and Occupational Health and Safety (Anses), also including the French Ministry of Agriculture as member.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

LIST OF ABBREVIATIONS

ADNS	Animal Disease Notification System
AI	Avian Influenza
ANN	Artificial Neural Network
ASF	African Swine Fever
AUC	Area Under Curve
BOW	Bag-Of-Words
CBOW	Continuous Bag-Of-Words
COVID-19	Coronavirus Disease 2019
CRF	Conditional Random Field
EBS	Event-Based Surveillance
EC	European Commission
ECDC	European Centre for Disease Prevention and Control
EI	Epidemic intelligence
EU	European Union
FAO	Food and Agriculture Organisation
FEIS	French Epidemic Intelligence System
FMD	Foot-and-Mouth Disease
GAUL	Global Administrative Unit Layers
GPHIN	Global Public Health Intelligence Network
IBS	Indicator-Based surveillance
IDF	Inverse Document Frequency
IE	Information Extraction
IHR	International Health Regulations
ILI	Influenza-Like Illness
IR	Information Retrieval
JRC	Joint Research Center
MedISys	Medical Information System
MERS	Middle East respiratory syndrome
MI	Mutual Information
NB	Naive Bayes
NER	Named Entity Recognition
NGO	Non-Governmental Organization
NLP	Natural Language Processing
OIE	World Organisation for Animal Health for animal diseases
PADI-web	Platform for Automated extraction of animal Disease Information from the web
POS	Part-Of-Speech
ProMED	Program for Monitoring Emerging Diseases

ROC	Receiver Operating Characteristic curve
SARS	Severe Acute Respiratory Syndrome
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SNOMED	Systematized Nomenclature of Medicine Clinical Terms
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
UMLS	Unified Medical Language System
WAHIS	World Animal Health Information Database Interface
WHO	World Health Organization

RÉSUMÉ

La libéralisation des mouvements d'animaux et des produits d'origine animale, la mobilité accrue des personnes et l'introduction délibérée ou accidentelle d'espèces sont à l'origine de la dissémination d'agents pathogènes à travers les pays et les continents (Brugere et al., 2017; Tatem et al., 2006). Au-delà de leur impact direct sur la santé et le bien-être des espèces, les maladies animales ont un coût sociétal majeur, en raison des pertes de production et des contraintes économiques qu'elles engendrent (Rich and Perry, 2011). De plus, parmi les maladies connues, 60% sont des maladies dites zoonotiques, c'est-à-dire des maladies dont les agents pathogènes se transmettent naturellement d'hôtes animaux à l'homme et vice-versa (Taylor et al., 2001). Les zoonoses émergentes représentent une menace croissante en terme de santé publique (Karesh et al., 2012). Un exemple tristement frappant est le coronavirus 2 du syndrome respiratoire aigu sévère (SRAS-CoV-2), qui a émergé en décembre 2019 dans la ville chinoise de Wuhan. A la date du 1er juillet 2020, le virus a tué plus de 500 000 personnes dans le monde WHO (2020). L'existence d'un hôte animal intermédiaire du SRAS-CoV-2 entre un réservoir probable de chauves-souris et l'être humain fait actuellement toujours l'objet de recherches (Li et al., 2020).

La capacité à détecter rapidement les maladies (ré)émergentes est une priorité en termes de santé publique et vétérinaire. L'alerte précoce, définie par l'OMS comme « le mécanisme organisé pour détecter le plus tôt possible toute occurrence anormale ou toute divergence par rapport à la fréquence habituelle ou normalement observée des phénomènes », est indispensable à la mise en œuvre de stratégies de lutte efficaces aux niveaux mondial et local (Heymann and Rodier, 2001). Au cours des dernières décennies, plusieurs épidémies ont mis en évidence les limites de la surveillance traditionnelle, sujette à des notifications retardées et à une latence des canaux de communication (Ben Jebara and Shimshony, 2006). En outre, les maladies émergentes posent des problèmes spécifiques en matière de détection et de notification car, par nature, il n'existe pas de tests de diagnostic ni de systèmes de notification officiels spécifiques de ces maladies.

La disponibilité croissante des données numériques représente une source sans précédent d'informations sur les maladies en temps réel. Les articles de presse en ligne, les réseaux sociaux, ou encore les dossiers médicaux digitalisés se sont révélés être de précieuses sources d'informations sur les maladies animales et humaines (Soto et al., 2008; Wilson et Brownstein, 2009). Leur intégration dans les systèmes de surveillance, à travers le processus d'intelligence épidémiologique (IE), a changé le paradigme de la surveillance et du contrôle des maladies. Sous l'impulsion du Règlement Sanitaire International (RSI) (WHO, 2005), l'IE intègre deux composantes dans un système de surveillance unique : la surveillance basée sur des indicateurs (collecte de données structurées par le biais des systèmes de surveillance traditionnels) et la surveillance basée sur les événements (collecte de données non structurées à partir de sources informelles) (Paquet et al., 2006). La combinaison de ces deux composantes permet d'améliorer la performance des systèmes de surveillance en terme de sensibilité et de rapidité de détection, tout en fournissant des informations épidémi-

ologiques supplémentaires telles que la sensibilisation ou l'état de préparation face à une menace (Arsevska et al., 2018; Bahk et al., 2015; Barboza et al., 2013; Dion et al., 2015).

La quantité d'information produite quotidiennement par les sources informelles peut très rapidement submerger les systèmes de veille, qui reposent toujours fortement sur la modération d'experts. Si les sources informelles sont diverses par leur nature, elles ont en commun d'être toutes le résultat d'une communication écrite, d'un journaliste par exemple, et donc disponibles au format textuel (un article, un tweet, un post de blog, etc.). Les textes sont généralement qualifiés de données non-structurées. Ce terme illustre le fait qu'un texte est la transcription d'une activité verbale produite non pas pour stocker une information, mais pour communiquer une idée, des faits, des instructions, etc. Ainsi, l'information n'est pas directement utilisable au sens d'une donnée, mais est structurée par les règles du langage. L'analyse de texte repose donc sur une étape d'abstraction, classiquement, une conversion dans un format numérique lisible par ordinateur (Hearst, 1999). Ce format intermédiaire, structuré et quantitatif, permet l'utilisation de méthodes de fouille de texte (text-mining), combinant le traitement automatique du langage naturel (TALN) et les techniques de fouille de données. La fouille de textes vise in fine à découvrir de nouvelles informations à partir de grandes quantités de textes, en développant des algorithmes de hautes performances et en sélectionnant les informations en fonction des besoins des utilisateurs humains (Collier, 2012).

PADI-web (Platform for Automated extraction of animal Disease Information from the Web) est un outil de veille automatique d'articles médias publiés sur le Web, dédié à la surveillance des maladies animales. Il est utilisé depuis 2014 par la cellule de Veille sanitaire internationale (VSI) de la Plateforme d'Epidémiosurveillance en santé animale (ESA) (Arsevska et al., 2017; Valentin et al., 2020a). Cette thèse se déroule dans le contexte de l'amélioration du système PADI-web. Aussi nos contributions visent à être aussi génériques que possible, en comblant certaines des limites rencontrées par les systèmes EBS. Notre objectif principal est de proposer des méthodes fondées sur l'utilisation de descripteurs épidémiologiques pour répondre à des tâches spécifiques de fouille de texte dans le contexte de la surveillance de la santé animale. Plus précisément, nous cherchons à répondre aux questions suivantes :

1. Comment détecter des informations épidémiologiques fines à partir d'articles en ligne ?
2. Comment améliorer l'extraction des événements à partir de descripteurs épidémiologiques d'une part, et de descripteurs lexico-sémantiques d'autre part ?
3. Comment utiliser les descripteurs épidémiologiques pour la recherche d'articles qui sont similaires du point de vue épidémiologique ?
4. Quelles sont les perspectives et les enjeux en termes de (i) surveillance des menaces sanitaires inconnues et (ii) compréhension de la diffusion de l'information épidémiologique par le biais de sources informelles ?

Cette thèse est structurée en cinq chapitres. Le chapitre 1 détaille les enjeux et les limites actuelles

des approches de fouille de texte mises en œuvre dans les systèmes de veille fondées sur les événements dans le cadre de la surveillance de la santé animale.

Le chapitre 2 décrit une nouvelle approche pour l'identification d'informations épidémiologiques fines à partir des articles en ligne. Notre approche est fondée sur une annotation au niveau de la phrase. Elle étend le champ de l'extraction d'entités épidémiologiques classiques (telles que les maladies, ou les hôtes) ou de l'extraction d'événements. L'approche d'annotation inclut deux niveaux : le type d'évènement, qui permet de filtrer des phrases sur la base de leur pertinence, et le type d'information épidémiologique, tels que les facteurs de risque ou les voies de transmission. Notre schéma d'annotation a été co-construit en collaboration avec des experts en épidémiologie, et a conduit à la publication d'un corpus annoté. Sur la base de ce corpus, nous comparons deux approches de recherche d'information.

La première est une approche de classification supervisée classique. Nous comparons trois classifieurs : Naive Bayes (NB), Support Vector Machine (SVM) et Multilayer Perceptron (MLP), sur la base d'une représentation textuelle « sac-de mots » (bag-of-words). Les deux derniers classifieurs obtiennent des résultats comparables, atteignant des exactitudes (accuracy) de l'ordre de 0.70 (type d'évènement) et de 0.65 (type d'information épidémiologique). Nous étudions ensuite une représentation textuelle fondée sur les plongements lexicaux (word embedding), en utilisant l'algorithme word2vec. Nous comparons différents paramètres (type d'algorithme, taille des vecteurs) et de pre-processing du texte (lemmatisation, suppression des stop-words, etc.). Le modèle sélectionné est celui obtenu par l'algorithme CBOW (continuous bag-of-words) et des vecteurs de dimension 300. La seule étape de prétraitement textuel est la conversion de tous les caractères en lettres minuscules. Sur la base de cette représentation, le classifieur MLP obtient les meilleures performances atteignant des exactitudes de 0.76 (type d'évènement) et de 0.72 (type d'information épidémiologique). Nous constatons en particulier une amélioration de la classification pour des classes minoritaires. La deuxième approche consiste à recherches les phrases d'une catégorie donnée sur la base de patterns (motifs) lexico-syntaxiques. Pour générer ces patterns, nous mettons en place une approche incrémentale et semi-automatique afin d'intégrer la connaissance d'experts. Les patterns sont tout d'abord identifiés manuellement à partir d'un nombre réduit de phrases, sous la forme d'une séquence de termes. Une première étape d'extension lexicale permet de créer des ensembles de termes similaires sur la base de leur proximité dans l'espace vectoriel du modèle de plongements lexicaux. Une deuxième étape consiste à la validation manuelle des ensembles de termes générés automatiquement, et de l'ajout éventuel de nouveaux termes. Enfin, les patrons sont transformés en relations syntaxiques afin de s'affranchir de la rigidité des séquences fixes de termes. Cette approche est évaluée sur deux classes véhiculant des informations fines et peu représentées.

L'approche par phrase nous apparait comme un complément important à la classification à l'échelle du document. Un module intégrant la classification supervisée va être implémenté à court terme dans le système PADI-web. Ainsi, sur la base d'un nouveau corpus de données externes qui va être constitué, de futurs travaux pourront consister à évaluer la précision de la classification, et la pertinence des informations détectées d'un point de vue de la veille. Dans le chapitre 3,

nous nous intéressons à deux approches pour extraire des événements. La première est fondée sur la cooccurrence de paires d'entités (localisation – maladie et localisation – hôtes). La force d'association des paires de mots est calculée par deux mesures classiquement utilisées en fouille de texte, l'Information Mutuelle et l'indice de Dice. Nous évaluons l'impact du contexte, défini comme la distance à considérer entre deux entités pour déterminer la cooccurrence. L'approche combinant mesure statistique et contexte est évaluée comme une fonction de rang, c'est-à-dire sa capacité à mieux classer des paires pertinentes. Nous montrons que fixer une fenêtre de mots, plutôt que considérer les co-occurrences à l'échelle du document en entier, améliore la détection de paires pertinentes. De plus, nous améliorons la qualité des paires localisation – maladie en généralisant les entités spatiales au niveau national. Dans ces travaux, l'indice de Dice obtient de meilleurs résultats que l'information mutuelle. L'approche statistique est classiquement utilisée sur de gros volumes de données, par exemple pour détecter des événements à partir de corpus de tweets. Elle permet de s'affranchir de méthodes de traitement du langage naturel sophistiquées qui sont plus sensibles à la qualité du texte, par exemple après une traduction. Ainsi, nous pourrions adapter notre approche à l'ensemble des articles collectés quotidiennement par PADI-web. La généralisation des entités spatiales à l'échelle nationale nous parait une approche pertinente pour réduire la quantité de paires générées. De plus, d'un point de vue de la veille, l'échelle nationale est un niveau correct pour la détection d'alerte. Cela pourrait être le cas pour les foyers de peste porcine africaine à la frontière française.

Dans un deuxième temps, nous évaluons une méthode lexico-syntaxique applicable à un échantillon réduit d'articles en français. Dans cette approche, une première étape consiste à identifier des verbes d'action particulièrement pertinents dans un contexte de veille épidémiologique. Les verbes servent de pivots pour l'extraction d'attributs : selon sa catégorie (définie sur la base d'une ressource lexicale de verbes en français), chaque verbe est associé à un cadre sémantique comprenant plusieurs attributs. Les résultats suggèrent que ce type d'approche n'est pas adapté à la détection d'attributs thématiques spécifiques telles que les maladies ou les localisations. Cependant, elle semble particulièrement performante pour la détection du nombre de cas, des hôtes et des unités épidémiologiques, qui sont typiquement les sujets ou objets des verbes liés aux événements. Nous suggérons d'étendre l'évaluation de cette approche à un plus grand nombre de textes. De plus, des approches similaires adaptées sur l'anglais pourraient être également étudiées pour permettre leur intégration dans les systèmes EBS.

Ensuite, nous proposons et évaluons des méthodes de fusion matricielle de descripteurs épidémiologiques (entités spatio-temporelles et thématiques) pour l'association d'articles sur la base de leur similarité (chapitre 4). La similarité est ici définie dans le contexte épidémiologique : deux articles sont similaires s'ils font référence au(x) même événement(s) sanitaire(s). Nous comparons les représentations issues de fusion à des représentations textuelles classiquement utilisées pour le calcul de similarité entre textes : sac-de-mots (BOW), lemmatisation, sélection de descripteurs sur la base de leur rôle grammatical. Les différentes méthodes sont évaluées de la même manière que les cooccurrences dans le cadre de la détection d'évènement, c'est-à-dire comme des fonctions de rang. La méthode de fusion de descripteurs améliore ou égale les résultats de toutes les

représentations classiques testées, tout en reposant sur un vocabulaire beaucoup plus réduit (1 400 descripteurs, contre 13 800 descripteurs pour le BOW). L'information spatiale est celle qui obtient, seule, les meilleurs résultats. L'information temporelle ressort comme la moins déterminante. La combinaison des 4 types d'entités obtient de meilleurs résultats par rapport à l'utilisation des entités seules ou combinées deux à deux. Nous proposons des poids à associer à chaque entité.

Dans le chapitre 5, nous abordons deux perspectives de ce travail, à partir de deux études de cas. Dans un premier temps, nous cherchons à comprendre la manière dont l'information sanitaire (définie ici comme la déclaration d'un événement) se diffuse entre les sources avant d'être détectée par un système de veille fondé sur les événements. Nous utilisons la méthode de l'analyse de réseau, représentant le système de sources et de communication par un graphe dans lequel les nœuds sont les sources, et les arrêtes représentent la transmission d'une information. Les résultats préliminaires indiquent que la diffusion de l'information se fait de manière très accélérée, et implique majoritairement un à deux intermédiaires entre la première détection et la détection par un système EBS. La source primaire d'information (sources qui communiquent les premières l'informations sanitaire) sont les autorités vétérinaires nationales. Les journaux en ligne et agences de presse apparaissent comme des sources intermédiaires majeures entre ces sources primaires et les systèmes EBS.

Enfin, nous avons mené une étude rétrospective pour évaluer la capacité de trois systèmes de surveillance événementielle (ProMED, HealthMap et PADI-web) à détecter les premiers signaux d'émergence de COVID-19. Nous nous sommes concentrés sur les changements dans le vocabulaire utilisé dans les articles avant et après l'identification de COVID-19. ProMED a été le système le plus rapide, détectant un article lié au COVID-19 un jour avant la notification officielle du cluster de cas de pneumonie par la Chine. À ce stade précoce, le vocabulaire spécifique utilisé était lié aux symptômes de pneumonie et à celui du « mystère ». Nos résultats suggèrent que les méthodes d'EBS doivent être adaptées aux différents stades de l'émergence de la maladie afin d'améliorer la détection précoce des futures émergences.

Cette thèse nous a permis d'utiliser différents types d'approches, statistiques, lexicales, sémantiques, dans le cadre de différentes tâches de fouille de texte appliquées à la veille en santé animale. Nous avons de plus contribué à la production de ressources annotées disponibles pour la communauté scientifique et permettant la comparaison avec de futures méthodes. Parmi les approches présentées, la classification par phrases sera implantée à court terme dans le système PADI-web. Cela nous permettra, à moyen terme, d'évaluer la pertinence de l'approche sur un jeu de données plus large. Ce travail ouvre de nombreuses perspectives de recherche. Si les systèmes de veille fondés sur les événements sont fréquemment évalués en termes de sensibilité, spécificité ou réactivité par rapport à des maladies spécifiques, nous pensons que leur capacité à détecter d'autres type d'informations, telles que l'origine d'un foyer, ou encore les conséquences économiques locales d'une épidémie, doivent être davantage valorisées. A ce titre, des méthodes d'évaluation spécifiques pourraient être proposées. De plus, la détection des signaux faibles nous semble une voie à privilégier. Nous proposons deux définitions, et donc deux ouvertures méthodologiques relatives aux signaux faibles. La première consiste en l'apparition, dans un ensemble d'articles,

d'un terme ou d'une paire de termes rares, potentiel signal d'un événement inhabituel. Cette définition implique des approches de détections statistiques, fondées sur l'analyse dans le temps de la fréquence et/ou discriminance d'un ou plusieurs termes. Plusieurs métriques existantes, telles que le TF-IDF temporel ou le degré de visibilité pourraient être comparées dans ce contexte (Preoţiuc-Pietro et al., 2016; Yoon, 2012). La deuxième définition désigne les articles décrivant un événement inconnu, c'est-à-dire un ou plusieurs cas symptomatiques dont la cause n'est pas identifiée. Cette définition pourrait dans un premier temps mener à la mise en place d'une approche de classification supervisée adaptée à cette problématique.

INTRODUCTION

Background

In 1920, a shipment of Zebu cattle coming from India transited through the port of Antwerp (Belgium) on its way to Brazil. The cattle were infected with the rinderpest virus — causing one of the deadliest livestock diseases. This virus introduction triggered a devastating rinderpest outbreak in Belgium (Vallat et al., 2013). Indeed, the history of zoonotic and animal diseases that have spread through international livestock trade is abound with examples (Seimenis, 2008).

The globalisation of animal and animal product movements, the increased mobility of people, and the deliberate or accidental introduction of non-native pathogen agents, as well as their possible vectors, are major drivers of pathogen dissemination across the countries and continents (Brugere et al., 2017; Tatem et al., 2006). Animal diseases have detrimental impacts on animal health and the economy in terms of lost revenues and societal costs (Rich and Perry, 2011). Some diseases have the potential of rapidly killing large numbers of animals (e.g., Newcastle disease, African swine fever), while others also prompt significant drops in demand of animal products, through consumer fears of getting infected with zoonotic diseases (e.g. avian influenza) (Rushton et al., 2005). In recent decades, concern has been growing with regard to so-called zoonotic diseases which are caused by pathogen agents shared by animals and humans - a common situation (Taylor et al., 2001). A striking and recent example is the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which hwas identified in December 2019 in the Chinese city of Wuhan. Up to early July 2020, the virus had killed more the 500,000 persons worldwide (WHO, 2020). Investigations are ongoing to determine if there is an intermediate animal host of SARS-CoV-2 between a probable bat reservoir and humans (Li et al., 2020).

The ability to rapidly recognise emerging and re-emerging diseases is a critical global health priority. Early warning, defined by WHO as "the organised mechanism to detect as early as possible any abnormal occurrence or any divergence from the usual or normally observed frequency of phenomena", is crucial for the implementation of effective control strategies and their quick implementation at global and local levels (Heymann and Rodier, 2001). In recent decades, several outbreaks have highlighted the limitations of conventional disease surveillance, which is hampered by delayed detection and latency of the communication channels (Ben Jebara and Shimshony, 2006). Besides, by their nature, emerging diseases raise specific issues with regard to their detection and reporting because in many cases, there are no diagnostic tests or formal reporting systems when the emerging event starts.

The growing availability of digital data represents an unprecedented source of real-time disease information. Online news, social media and electronic health records are among the so-called informal sources that have proven to be valuable sources of disease information (Soto et al., 2008; Wilson and Brownstein, 2009). Their mainstreaming into surveillance systems, via the epidemic

intelligence (EI) concept, has been a game-changer for disease surveillance and control. Driven by the International Health Regulations (IHR) (WHO, 2005), EI integrates two components in a single surveillance system: indicator-based surveillance (collection of structured data through traditional surveillance systems) and event-based surveillance (collection of unstructured data from informal sources) (Paquet et al., 2006). Combining these two components has proven to enhance the performance of surveillance systems by increasing the outbreak detection timeliness and number (Arsevska et al., 2018; Bahk et al., 2015; Dion et al., 2015; Barboza et al., 2013).

Informal information sources are diverse in their spectrum, but they all share the information in textual format. Peculiarities of textual data include linguistic ambiguities, redundant and noisy information, a lack of normalisation, etc. Besides, daily amounts of such information can rapidly overwhelm surveillance systems, which include a step of moderation by experts. Event-based surveillance (EBS) systems thus increasingly marshal text-mining methods to alleviate the amount of manual curation of the continuous flow of free text (Hartley et al., 2010). Text-mining — which combines natural language processing (NLP) and data-mining techniques—enables the conversion of free text into a computer-readable format (Hearst, 1999). Text mining aims at discovering new information in a timely manner from large text collections through the development of high-performance algorithms, along with information filtering according to the needs of human users (Collier, 2012).

In this setting, the Platform for Automated extraction of Animal Disease from the Web (PADI-web[1]) is an open-access EBS system dedicated to the detection of new and emerging animal infectious disease events (Arsevska et al., 2017; Valentin et al., 2020c). It was developed to meet the needs of the French Epidemic Intelligence System (FEIS, or *Veille sanitaire internationale* in French) via online news monitoring. FEIS has been involved in activities of the French Platform for Animal Health Surveillance (ESA Platform) since 2013. FEIS aims to identify, monitor and analyse reports of animal health hazards (including zoonotic diseases) threatening France as a whole, by monitoring official and unofficial information sources. PADI-web monitors Google News in real-time and automatically retrieves animal disease related news articles, classifies them and extracts epidemiological entities (Arsevska et al., 2018).

Motivation

Research for this thesis has been under way at a time when PADI-web pipeline is being enhanced to better address FEIS needs. However, we are also striving to ensure that our contributions are as generic as possible, by filling some of the gaps inherent to the EBS systems. Our main objective is to enhance the use of epidemiological features regarding specific text-mining based tasks in the animal health surveillance context. More precisely, we address the following questions:

1. How could fine-grained epidemiological information be retrieved from online news?
2. How could event extraction be improved based on epidemiological feature co-occurrence and lexico-semantic cues?

[1] https://padi-web.cirad.fr/en/

3. How could epidemiological features be used to detect news articles related to the epidemiological rationale?

4. What are the perspectives and stakes, in terms of: (i) unknown threat surveillance, and (ii) understanding the dissemination of epidemiological information through informal sources before being detected by event-based surveillance systems?

Outline

This thesis is structured in five chapters. **Chapter 1** details the stakes and current limits of text-mining approaches implemented in event-based surveillance systems in the animal health surveillance context. **Chapter 2** describes a new framework to retrieve epidemiological information from the news. Our approach is based on sentence-level annotation and broadens the scope of conventional extraction of epidemiological entities (e.g. disease, host) or event mentions with new types of epidemiological topics such as risk factors or transmission pathways. **Chapter 3** evaluates two approaches to extract events based on the co-occurrence of pairs of entities, and on a graph-based and linguistic approach. **Chapter 4** describes how epidemiological entities are used to retrieve related news. **Chapter 5** showcases two perspectives of this work based on two case studies.

Chapter 1

Text mining for animal health epidemic intelligence: stakes and limits

Table of contents

In this section, we describe the stakes and limits of current text-mining methods for the analysis of informal sources of sanitary information in the animal health surveillance context. To our knowledge, the earliest surveys on text mining methods applied to epidemic intelligence were published in 2009 and 2012 (Keller et al., 2009; Collier, 2012). Since then, the landscape of event-based surveillance systems has been extensively described (Hartley et al., 2010; Barboza et al., 2013; Velasco et al., 2014). First, we introduce epidemic intelligence (EI) concepts as a novel framework for disease surveillance. Next, we describe the main characteristics of formal EI systems in animal health, including their limitations for early warning. We further describe informal EI systems and focus on EI using online news sources. We detail the text-mining methods that have been developed to date to address the challenge of processing online news articles, while highlighting their current limitations.

1 Animal disease surveillance context

In this Section, we present the traditional surveillance approach based on formal sources. We introduce alternative sources of information for disease surveillance, i.e. unofficial sources. We further explain how both types of sources are integrated into the EI process through two main information channels: **indicator-based surveillance** and **event-based surveillance**.

1.1 Stakes and limits of traditional surveillance

1.1.1 Traditional data sources for surveillance

The collection of data from formal sources is the conventional way of monitoring diseases—this is referred to as **indicator-based surveillance** in the EI framework. Formal sources are typically **official sources** that include governmental institutions and organisations present at different geographical levels:

- Subnational level: local, district or regional public health and veterinary health agencies or networks;
- National level: national public health and veterinary health agencies, which are generally under the direction of ministries of health or agriculture,
- Supranational level: institutions, such as the European Centre for Disease Prevention and Control (ECDC), or regional networks such as the East Africa Integrated Disease Surveillance Network or EpiSouth;
- International level: the World Organisation for Animal Health (OIE) for animal diseases and the World Health Organization (WHO) for human diseases (including human cases of zoonotic diseases).

Formal sources collect outbreak data through a routine formalised process that relies on well-established case definitions. They produce structured and reliable data, while being highly comprehensive regarding the pathogen, outbreak source, species, clinical signs, etc. As they first detect and report diseases to international authorities, national public health and veterinary health agencies are the key actors of conventional disease surveillance. International institutions (i.e. OIE, WHO) play a critical role in collecting and communicating of notifiable disease events at an international scale. Besides, at the subnational and national levels, sentinel, syndromic and participatory-based surveillance complement mandatory disease-specific notifications (Brugere et al., 2017; Paolotti et al., 2014). Contrary to surveillance based on individual notifiable cases, such approaches are used to collect and analyse more timely aggregated data for the purpose of syndrome detection. In addition to official sources, they usually collect information from the so-called **authorised sources**, i.e. sources that do not depend on government authority, such as NGOs (e.g. Doctors Without Borders), clinician or laboratories.

1.1.2 Limits

Although formal sources are still the keystone of disease surveillance, they also have limitations for early warning. Underreporting and delayed notifications significantly alter their reactivity (Chandrasekar, 2012). From 2015 to 2018, countries reported confirmed animal outbreaks to OIE within a median of 3 days and a maximum reporting delay of 280 days. During this period, only 29% of immediate notification reports were submitted within the recommended time limit of 24 h after confirmation of the event (Kuribreña et al., 2019).

The reasons for underreporting include the inability and unwillingness to report (Halliday et al., 2012). A surveillance system may be unable to detect or report a disease outbreak due to a lack of training or insufficient personnel resources, and to a poor diagnosis capacity. The administrative process for official notification creates a non-compressible delay between disease observation, laboratory confirmation and disease reporting to an international health authority (Bahk et al., 2015). Besides, farmers and competent animal health authorities may be reluctant to notify diseases to avoid negative political or economic consequences (McNabb, 2010).

These factors may be responsible for disease introductions and spreading into unaffected territories, with major animal health and economic impacts. An eloquent example is the emergence of African swine fever (ASF) in Eastern Europe in 2007 (Rowlands et al., 2008). The earliest cases were observed in Georgia in April 2007. Georgia notified the OIE of several outbreaks on May 17, 2007. The event was first suspected to be caused by the porcine circovirus. Tests carried out on samples finally confirmed the presence of the African swine fever virus (OIE, 2007). The disease is now endemic in Eastern Europe.

1.2 Novel sources of health surveillance information

1.2.1 Motivation

In 2005, World Health Organization (WHO) members adopted the revised International Health Regulations 2005 (WHO, 2005). This decision followed a series of international infectious disease outbreaks, including the severe acute respiratory syndrome (SARS) outbreak in 2003, thus prompting the need for early detection and response to health threats worldwide. By complying to IHR, the WHO Member States agreed to strengthen their capacity to prevent, detect, and respond to infectious disease outbreaks that have a high potential for global spreading (WHO, 2005). In addition to reinforcing existing national surveillance systems, IHR also encouraged the use of new technologies to broaden the range of data sources used for surveillance. These are called **informal sources** since they are not official or formal (WHO, 2014).

Informal sources are typically digital and freely available online sources. Their monitoring is hence referred to as “web-based” or “internet-based” surveillance. The first sources used were **online news articles** and blogs (Brownstein et al., 2008). From July 1998 to August 2001, 65% of the 578 human disease outbreaks verified by WHO came from online media (Heymann and

Rodier, 2001). Authors used Internet reports for timely detection of human and animal disease outbreaks (Arsevska et al., 2018; Carrion and Madoff, 2017; Choi et al., 2016). From 2003 to 2009, informal sources reported 50% of avian influenza and H1N1 outbreaks at least 4.96 and 2.26 days before WHO, respectively (Tsai et al., 2013). News articles served as an alternative data source in the absence of detailed epidemiological information rapidly available from formal sources during epidemic emergencies. Detailed accounts of epidemiological clusters were extracted during the 2014–2015 Ebola epidemic in West Africa and the 2015 Middle East respiratory syndrome (MERS) outbreak in South Korea (Chowell et al., 2016). Trends in the volume of cholera-related news articles were significantly temporally correlated with official cholera cases during the 2010 Haitian cholera outbreak (Chunara et al., 2012). In the SENTINEL system, online news articles are used as an indicator of situational awareness. Events are first detected from a stream of tweets, and each detected event is further linked with related news (Șerban et al., 2019).

The volume of **web searches** has further been shown to be an effective estimator of public health disease events. A well-known example is Google Flu Trends, which is a model that was developed by Google.org to estimate influenza-like illnesses (ILI) in the United States from internet searches (Ortiz et al., 2011) . The logistic regression model fit actual ILI numbers with mean correlations of over 0.9. More recently, (Lu and Reis, 2020) found that increases in COVID-19-symptom-related searches predict increases in reported COVID-19 cases and deaths 18.53 days and 22.16 days in advance, respectively. In recent years, the use of data from **social media**, especially Twitter, has been tremendously studied for event detection, including disease surveillance and forecasting. Self-reported influenza cases on Twitter were compared to ILI cases in the US (Broniatowski et al., 2013; Li and Cardie, 2013). Daily counts of tweets related to enterohemorrhagic Escherichia coli (EHEC) in Germany were effective in triggering EHEC outbreak alerts before official outbreak detection (Diaz-Aviles and Stewart, 2012). To our knowledge, the use of social media to monitor animal disease outbreaks has been poorly studied. When comparing avian influenza (AI)-related tweets to the OIE official cases, the authors noted that tweet contents tend to be article titles (Robertson and Yee, 2016)). Indeed, over 85% of tweets topics are headline news or persistent news (Kwak et al., 2010), thus suggesting that Twitter is an efficient tool to access AI event reports in news articles, blog posts, and other sources of online media.

1.2.2 Limits

Importantly, although they belong to the same source family, informal sources differ markedly in terms of nature, relation to official disease events and interpretability. Online news typically reports one or several disease events at the outbreak level rather than symptom onset at the individual level. They can thus be used as direct indicators of disease events and awareness at the cost of manual curation (Arsevska et al., 2018; Chowell et al., 2016). Current efforts are hence aimed at reducing the number of noisy data and organising information to facilitate the manual validation. As they rely on journalistic decisions, the reporting of disease events in online news is biased in several ways (Section 3.1). Among other limitations, the editorial process creates a time lag that sharply contrasts with real-time communication on social media platforms. That is

why the DEFENDER surveillance system uses news articles as a secondary source of information: outbreak alarms are generated based on streams of symptom-related tweets, and news articles are subsequently retrieved to provide a larger information context (Thapen et al., 2016b).

Yet both social media and internet searches generate a huge amount of daily data that serve as indirect indicators of an outbreak. They are analysed on the basis of a sydromic surveillance paradigm whereby statistical models help detect volume trend changes (Thapen et al., 2016b; Sharpe et al., 2016). A key challenge is to fit models on historical data while preserving their ability to forecast future outbreaks. The first version of Google Flu Trends has been persistently overestimating flu prevalence for several years and missed the non-seasonal 2009 influenza A–H1N1 pandemic. An overfitting problem was identified as the cause of the lack of forecasting robustness, as 50 million search terms were used to fit 1152 ILI data points. The algorithm was described as "part flu detector, part winter detector" (Lazer et al., 2014). Moreover, the frequency of both internet searches and social media posts may be driven by personal perceptions or media interest. An increase in searches for "bird flu" that occurred in USA between 2005 and 2006 was attributed to media attention to an influenza outbreak affecting Asia at the time (Carneiro and Mylonakis, 2009).

Informal sources have common structural limitations, as they all produce textual data. Several aspects are particularly critical: data is unstructured (i.e. epidemiological entities are not directly available), noisy (i.e. extracted epidemiological entities are not necessarily linked to an event), and not validated (i.e. the trustfulness is unknown). Thus, before being used as input in statistical models or presented to experts, natural language processing (NLP) methods are used to clean free texts and detect relevant elements. Such approaches are detailed in the online news processing context in Section 4.

The value of informal sources to increase the timeliness of disease outbreak detection and provide detailed epidemiological information in the early warning and preparedness context is recognised (Arsevska et al., 2018; Bahk et al., 2015; Barboza et al., 2013; Dion et al., 2015). Their integration is formalised through the so-called epidemic intelligence (EI) process.

1.3 Epidemic intelligence process

1.3.1 Concept and definitions

The Epidemic Intelligence Service of the Center for Disease Control is considered to be the earliest public health system dedicated to early warning (Langmuir, 1980). It was created to enhance the surveillance and eradication of both infectious and non-infectious human diseases, such as poliomyelitis and leukaemia, and bioterrorism. The epidemic intelligence (EI) concept as it is used today was developed in the early 2000s. The French *Institut de Veille Sanitaire* and the European Centre for Disease Prevention and Control (ECDC) proposed an EI framework to enhance disease surveillance in Europe in 2006 (Kaiser et al., 2006; Paquet et al., 2006). Eight years later, the World Health Organization (WHO) published a comprehensive guide providing key definitions and de-

tailing the implementation of early warning activities (WHO, 2014). Epidemic intelligence corresponds to a formalised surveillance process that encompasses "all activities related to the early identification of potential health hazards that may represent a risk to health, and their verification, assessment and investigation" (WHO, 2014). It relies on two main channels of information: **indicator-based surveillance** (IBS) and **event-based surveillance** (EBS).

Indicator-based surveillance is defined as "*the systematic collection, monitoring, analysis and interpretation of structured data (i.e. indicators)*"(WHO, 2014). It corresponds to conventional surveillance of formal sources and is based on established case definitions (Section 1.2).

Event-based surveillance is defined by WHO as "*the organised collection, monitoring, assessment and interpretation of mainly unstructured ad hoc information regarding health events or risks, which may represent an acute risk to human [or animal] health*" (WHO, 2014). EBS involves the use of data streams from informal sources (Section 1.2.1).

The definitions and concepts from both ECDC and WHO were elaborated for public health. However, they have been successfully transferred to other domains, such as plant health (Alomar et al., 2016) and both terrestrial and aquatic animal health (Arsevska et al., 2018; Lyon et al., 2013b,a).

1.3.2 Workflow

Both EBS and IBS can be formally represented as consecutive steps, corresponding to the flow of epidemiological information from its detection to its communication to relevant authorities (Barboza et al., 2013; Kaiser et al., 2006; Rotureau et al., 2007) (Figure 1).

The first stages of the EI process consist of identifying and extracting relevant and accurate information from raw (unverified) data. They include: (i) **data detection**, (ii) **data triage**, and (iii) **signal verification**.

Data detection (or collection consists of defining modalities (e.g. format, frequency) through which **raw data** is detected, and implementing the collection process. The strategy differs according to the type of source. For instance, IBS usually relies on well-established notification procedures submitted through the electronic platform by a country, for instance. Otherwise EBS systems use specific queries implemented through RSS feeds.

Data triage is a crucial step to avoid overwhelming the EI system with irrelevant data. Raw data is filtered (triaged) based on predefined selection criteria. In IBS, for instance, selection criteria vary according to the institution mandates, such as its geographical coverage (the whole world - such as OIE -, or regional - such as ECDC), or its thematic mandate (e.g. animal health, public health, or infectious diseases). In EBS, **data triage** includes: (i) **data filtering**, which consists of removing duplicates and irrelevant data (e.g. a review about a disease), and (ii) **data selection**, which consists of sorting information according to the EBS system priority criteria. Data retained as relevant regarding early warning activities is a **signal**.

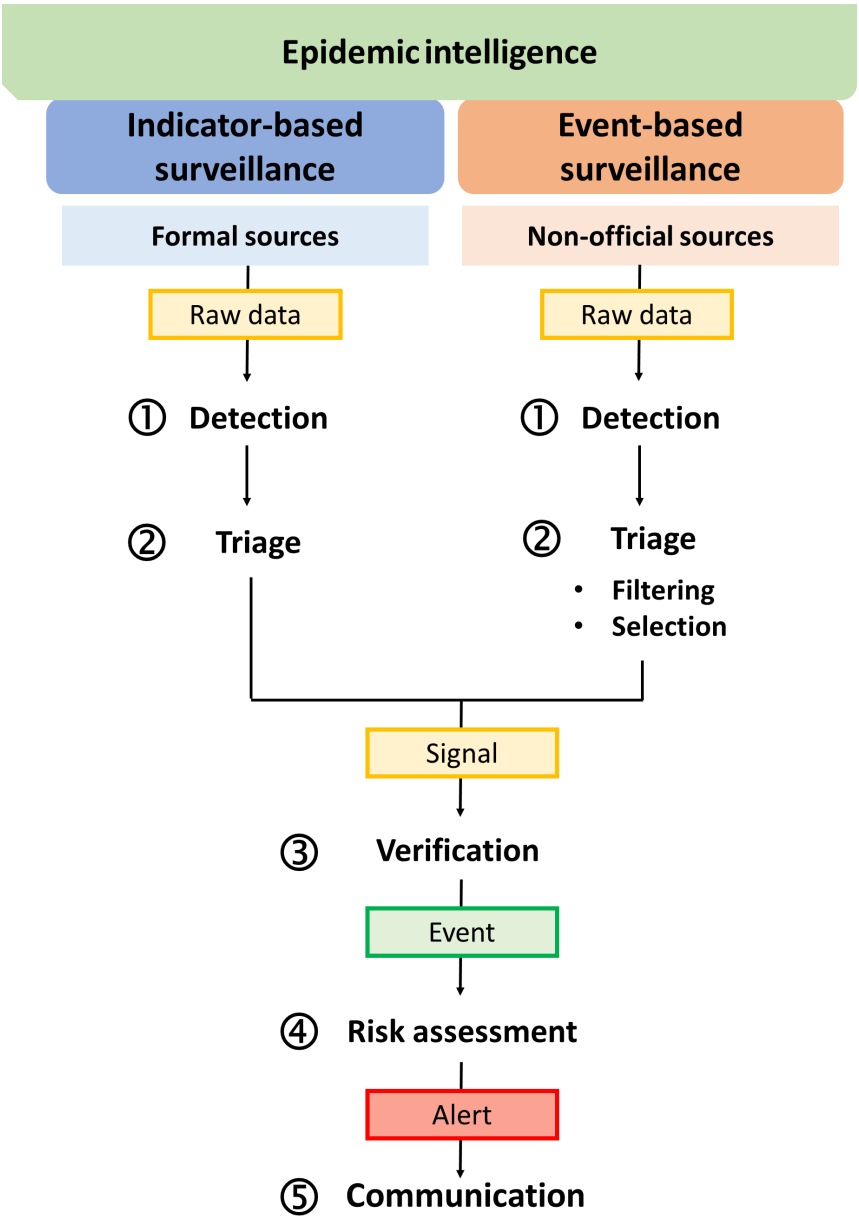


Figure 1 Epidemic intelligence workflow (adapted from WHO (2014)).

Signal verification consists of validating the truthfulness of a signal. This step is especially important in the EBS workflow since data sources are informal. This step is usually automatic in IBS. Once validated, a signal becomes an **event**, i.e. "*a manifestation of a disease or an occurrence that creates a potential for disease*" (Baker and Forsyth, 2007). This definition encompasses events from all possible known origins (e.g. infectious, zoonotic, food safety, chemical, radiological or nuclear in origin), as well as signals of unknown origin in the EBS context.

The final stages consist of **risk assessment** and **alert communication**. **Risk assessment** aims at determining the level of risk associated with a detected signal. The signal becomes an **alert** if the risk is considered significant to public or animal health. Alerts are communicated through adapted channels to relevant authorities (e.g. ministries of health, an international organization such as OIE, public health national networks, etc.) or the larger network (e.g. users of an EBS system).

2 Indicator-based and event-based surveillance systems in animal health

2.1 International indicator-based surveillance systems

In Section 1.1.1, we present the different stakeholders involved in IBS surveillance. In this section, we focus on the supra-national level, as international organizations are the major actors of the development of online IBS systems. Besides, **international IBS systems** are usually considered as gold standard data sources for evaluating EBS systems (Arsevska et al., 2018; Robertson and Yee, 2016).

Three IBS systems currently focus on infectious animal diseases: the OIE World Animal Health Information System (WAHIS) Interface, the FAO Emergency Prevention System for Priority Animal and Plant Pests and Disease (EMPRES-i) and the EC Animal Disease Notification System (ADNS). Their features are outlined in Appendix A.

WAHIS is an updated version of the OIE notification system created in 1996 (OIE, 2017). The WAHIS Interface provides data from all outbreak reports (for 117 notifiable animal infectious diseases) reported by member countries[1] as well as additional information on surveillance and control strategies, animal population numbers, etc. (Caceres et al., 2017).

At the European level, the Animal Disease Notification System (ADNS) has been centralising information on 45 animal infectious diseases since 1998 (Council, 2016). The main objective of the ADNS system is to ensure rapid exchange of outbreak information between EU countries. The system is based on European regulation, which makes it compulsory for the EU countries to notify disease outbreaks.

[1] 182 Member Countries in 2020 (https://www.oie.int/en/about-us/our-members/member-countries)

In 1994, the Food and Agriculture Organization of the United Nations (FAO) created EMPRES-i, an animal disease information system. This system was initially focused on transboundary livestock diseases with significant food security or economic impacts, such as avian influenza (Martin et al., 2007), and it currently covers 32 animal infectious diseases.

The information workflow is summarized in Figure 2. The three IBS systems are exhaustive in terms of the epidemiological information they collect and share. WAHIS and ADNS are more comprehensive than EMPRES-i and record the type of event (immediate notification versus follow-up, primary versus secondary outbreaks). This feature is important for further risk assessment by EI experts since solely immediate notifications and primary outbreaks are of interest for early warning purposes. Moreover, WAHIS and ADNS provide complementary information such as the diagnostic method, the source (origin) of the disease and the control measures implemented.

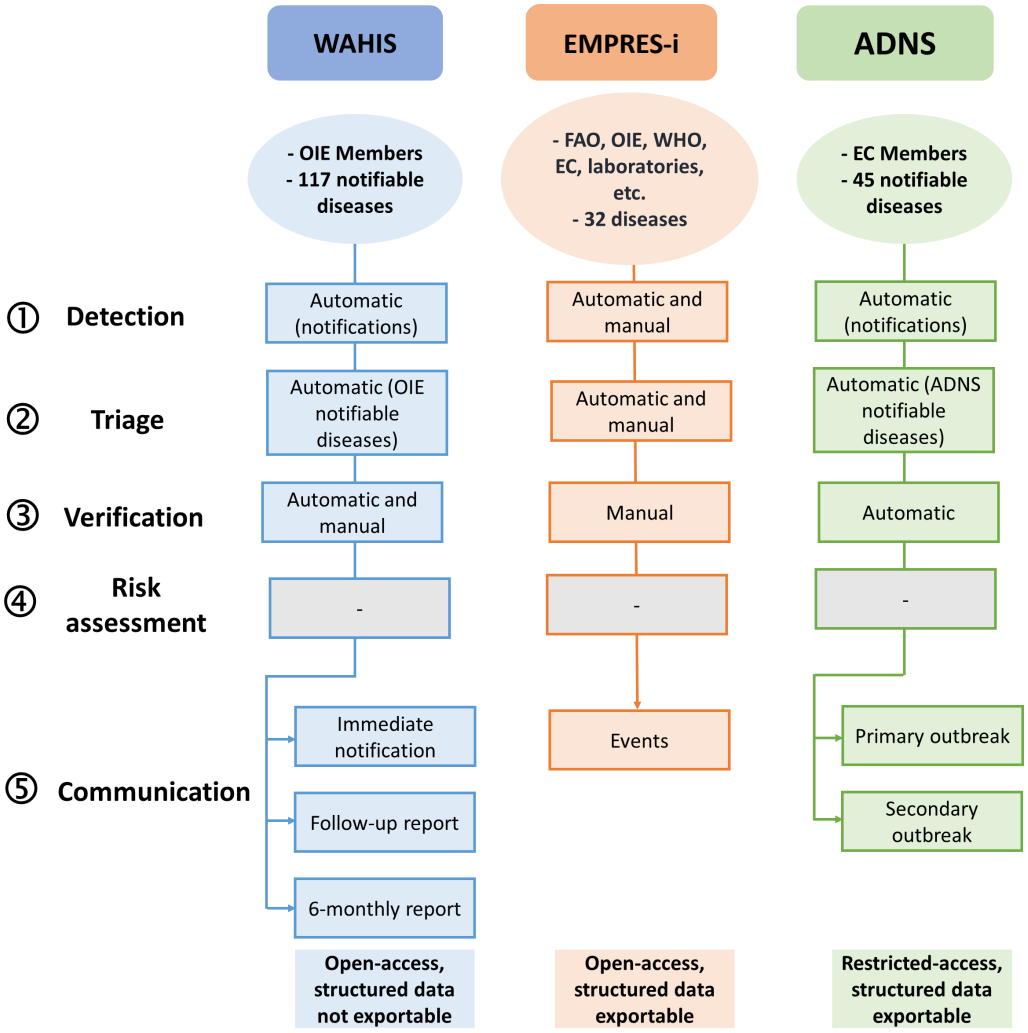


Figure 2 The workflow of three IBS surveillance systems dedicated to animal health.

2.2 Event-based surveillance systems

The development of EBS systems aims at meeting the challenges posed by the integration of unstructured data in the formalised EI process. Below we present the EBS systems that encompass the animal health threat in their scope.

EBS systems were pioneered in 1994 by the International Society for Infectious Diseases (ISID), through the Program for Monitoring Emerging Diseases (ProMED) (Madoff, 2004; Woodall, 2001). ProMED is a **human-curated** system that relies on an extensive network of experts worldwide who detect and share reports on disease outbreaks in a common platform (Carrion and Madoff, 2017). Moderators validate the information.

BioCaster and the Platform for Automated extraction of animal Disease Information from the web (PADI-web) rely on fully **automated** pipelines. BioCaster was a public health surveillance system supported by the University of Tokyo from 2006, with a priority focus on the Asia-Pacific region (Kawazoe et al., 2008). BioCaster is no longer operational, but it is included in our review because it relied on a unique and well-documented ontology-based approach. PADI-web was created in 2016 to monitor online animal health-related news for the French Epidemic Intelligence System (FEIS) (Arsevska et al., 2017; Valentin et al., 2020c).

Between these two extremes of pure automation and pure manual data collection and analysis, other prominent systems combine automated text-mining based steps and a dedicated team of curators to assess and verify the outputs. **Semi-automated** systems include HealthMap, founded by the Boston Children's Hospital in 2006, the Canadian Public Health Agency Global Public Health Intelligence Network (GPHIN), the European Union MediSys, Argus and AquaticHealth.net.

2.3 Epidemiological information within the EI process

Epidemiological information is at the core of IBS and EBS inputs and outputs (i.e. signals, events and alerts). In practice, epidemiological information available to assess an event varies according to the type of surveillance. Data used in IBS systems come from official sources, so a high level of accuracy and completeness is expected. Conversely, epidemiological information provided by EBS systems highly varies, depending on: (i) the information available in the data source (e.g. online news mentioning the disease and location solely), and (ii) the information extracted by the EBS system (e.g. the system was able to detect the disease and location mentioned in the content of online news).

Essential information to describe an event may include at least the disease and spatiotemporal information. In animal health surveillance, this also includes the host (i.e. the species affected). In the computational field, such atomic pieces of information are referred to as **event attributes** (Chambers and Jurafsky, 2011). When used in classification models and information retrieval tasks, they can also be called **features**. In this thesis, we refer to them as **epidemiological entities**, but both **event attributes** and **features** are used in Chapter 3 and Chapter 4 for consistency

with the computer science domain. Each entity is defined by its type: temporal (e.g. date of start of an event), spatial (e.g. event location), numeric (e.g. number of cases), or thematic, including the disease, host and symptoms. Some type of entities can be provided at different granularity levels (e.g. city, state, or country for spatial entities).

3 Online news sources for event-based surveillance

In this section, we highlight key traits of online news as a data source that might influence the final detection of information by EBS systems. We distinguish reporting bias, which depends on the **newsworthiness** of animal health events, and **structural characteristics** of the news (Figure 3).

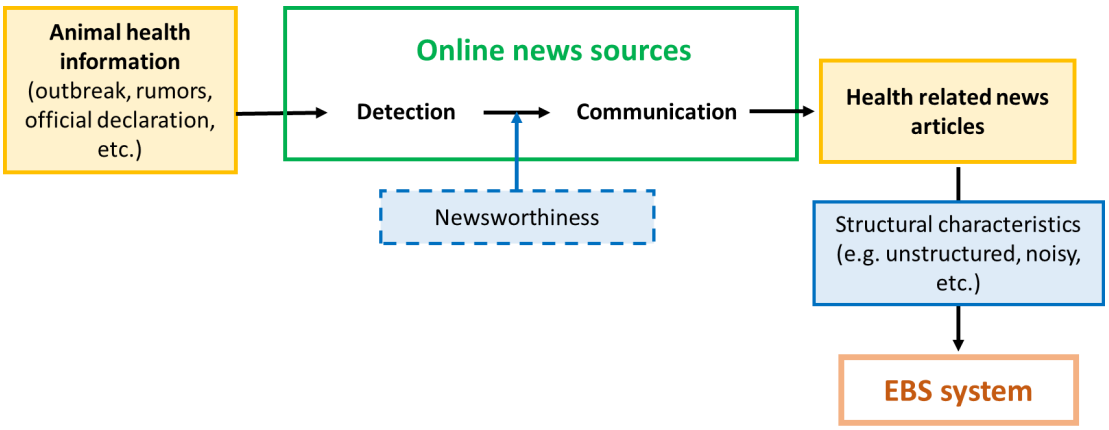


Figure 3 Bias and limits of online news sources in the flow of health information. EBS systems pipelines address structural characteristics of online news.

Both aspects are described hereafter. We further introduce the challenges that may arise when analysing EBS systems that deal with online news sources.

3.1 Animal health information newsworthiness

EBS systems that rely on online news as a data source closely depend on the journalistic willingness to share the information. In journalism, the criteria used to assess the newsworthiness of an information item and select what is published or not is called "newsworthiness" or "news value". Journalism practices consider criteria such as recency, conflict, unexpectedness, relevance, proximity, and social impact (Harcup and O'Neill, 2017). Little is known about criteria underlying journalistic selection criteria in the animal health context. Factors influencing the publication of animal disease outbreaks are mostly determined empirically, by examining online news content. Natural disasters and contagious diseases of public health importance have proven to be sensitive to sensationalism (Drache et al., 2003; Rezza et al., 2004). Some diseases are showcased by media

which overestimate their significance and risk to public or animal health. This bias is particularly marked with regard to zoonotic disease outbreaks (Poglayen et al., 2008). A striking example was the media coverage of the outbreak of bovine spongiform encephalopathy (BSE), also known as "mad cow disease". Isolated cases of BSE, as well as cases of another transmissible spongiform encephalopathy affecting deer (chronic wasting disease), are still highly covered by the media, yet their public health risk is minor.

On the other hand, some diseases receive little attention either because they are endemic in poor areas, or there is little interest from health authorities and stakeholders, or official reporting is lacking. A study evaluating the reporting of health events in Nepal's national media showed severe underreporting of animal diseases and environmental threats compared to human diseases (Schwind et al., 2017). Lack of media interest is also noted in highly specific domains, such as aquaculture (Thrush et al., 2012).

When published, animal health-related information is not directly available and is provided in a journalistic way.

3.2 Structural features of online news

In Section 1.2.2, we introduced some limits of informal textual data, i.e. unstructured, noisy and not validated data. Figure 4 shows an example of a news article published by an online source [2], illustrating these features.

First, contrary to formal data, epidemiological information is not directly available in online news content. In Figure 4, disease, host, and spatiotemporal entities related to the recent event (shown in orange) are disseminated in the content. They can be present in various lexical forms (e.g. disease names as acronyms) and at different hierarchy levels (e.g. in Figure 4, locations are described at village, region and country levels). The expression "Wednesday" cannot be directly interpreted as a locatable temporal entity without contextual information (here, the news article publication date). Text-mining and machine learning methods are thus needed to extract, identify and normalise epidemiological entities from free texts.

Secondly, while official sources release event information at a known epidemiological scale, a single news article may report a single outbreak, but also provide a case count of one or several diseases over a specific time range, or even refer to older outbreaks (entities in green). Thus, in addition to the identification of epidemiological entities, newly reported events have to be distinguished from historical and hypothetical events. Besides, numerous entities may be present while not having any link with the target event (here, the location "China", and entities used to provide general information are noted at the end of the news article).

Another major challenge is the evaluation of the trustfulness of the information. Different types of clues can be taken into account, such as the type of online news source, primary sources cited in the news, or the completeness of the information. Several sources are specialized in animal

[2] https://www.porkbusiness.com/article/african-swine-fever-found-two-farms-russia

health or economics, thus are more likely to share validated information. When an international authority is cited in the news, there is a higher probability that the information is correct.

However, no procedures or guidelines provide a precise way to assess the reliability of an online news item or its source. As detailed further, several EBS systems compute indirect trustfulness indicators, based for example on users' ratings.

Figure 4 News article published by *PorkBusiness.com* on August 21, 2019.

Such aspects highlight the need for developing and implementing NLP methods to detect, extract and analyse information from free texts. Indeed, NLP methods shape the pipelines implemented in EBS systems.

3.3 Challenges in describing and comparing EBS pipelines

For several reasons, EBS system pipelines and the methods they rely on are heterogeneously described in the literature, thereby making it difficult to precisely and formally compare them.

First, research devoted to text-mining methods for web-based health surveillance and the implementation of such methods in EBS systems are two parallel yet non-synchronous processes. Hence, it is not easy to know which method is actually integrated into an EBS system. Similarly, some implemented algorithms are not documented or fully described.

Second, there is a lack of consistency in the vocabulary used to describe the steps of the pipeline and the data flow. Indeed, EBS systems vocabulary are at the crossroads between recent EI definitions and concepts and the well-established informatics jargon. We noted several discrepancies between the vocabulary describing EBS pipelines and WHO formal definitions of computational terms. For instance, the "data triage" step is referred to as "classification" in most EBS systems, as it usually relies on classification algorithms to categorise news articles. In HealthMap, however, the

term "classification" is used in reference to the extraction of the location and disease describing an outbreak. In computational terms, this task corresponds to event extraction. In both HealthMap and PADI-web, the term "signal" refers to unverified data detected by the system (Arsevska et al., 2018; Brownstein et al., 2008), while in the formal EBS process, a signal corresponds to data considered relevant after the triage step (Section 1.3.2). As detailed later, the term "event" is particularly prone to multiple definitions, whether it is used in the epidemiological or informatical context.

Secondly, the EBS process is formally described as a unidirectional process since it was modelled on the IBS scheme (Figure 2). However, the flow of data within EBS systems is not necessarily linear. For instance, filtering and validation steps can be done at different levels: a news article can be categorised as relevant, while some events extracted from its content will be considered irrelevant (e.g. reference to an old outbreak).

Thus, in the following section we do not categorize text-mining methods based on steps of the EBS process. Instead, we distinguish approaches dedicated to analysing news articles as a whole (document-based) and more fine-grained methods based on epidemiological entities (entity-based).

4 Approaches, stakes and limits of text-mining methods applied to online news monitoring

In this section, we describe how EBS systems rely on text-mining methods to automate or enhance part or all of their process. We excluded the ProMED system in the remainder of this section since it involves a fully manual process. However, we refer to ProMED in several parts of the discussion to compare automated approaches to the totally expert-based paradigm. Besides, a comprehensive description of the ProMED pipeline is detailed elsewhere (Carrion and Madoff, 2017).

Importantly, all of the studied EBS systems, except for AquaticHealth.net and PADI-web, were primarily designed for public health surveillance. The methods and limits discussed hereafter are therefore not specific to animal health.

We divided the approaches into two levels, i.e. document and entity levels. The term document refers to a news article, while the term entity refers to an epidemiological entity, as described in Section 2.3. The main features of each EBS systems are summarized in Table 3, at the end of the section. The current pipeline of PADI-Web is based on 4 steps (i.e. data collection, data processing, data classification, information extraction) (detailed in Appendix B and C). The contributions of this thesis enable to improve the last steps of PADI-Web by proposing new information retrieval and event extraction methods.

4.1 Document level

4.1.1 Document retrieval

Compared to IBS systems, which mostly passively receive information, EBS systems implement an active retrieval process. When automated, data retrieval is a critical step to ensure the sensitivity of the EBS system since it channels what will be further analysed and potentially detected as signals.

A commonly shared strategy to retrieve news articles from online sources involves the implementation of Really Simple Syndication (RSS) feeds. RSS feeds are easily customisable by the user and they can be used to automatically retrieve data from a broad range of sources:

- A specific online news source, e.g. AquaticHealth.net monitors websites dedicated to aquaculture, such as thefishsite.com;
- A range of online news sources through news aggregators, e.g. Google News for HealthMap and PADI-web; Al Bawaba and Factiva for GPHIN;
- Another EBS system, e.g. HealthMap, BioCaster and AquaticHealth, collects ProMED alerts;
- An IBS system, e.g. HealthMap, Argus and AquaticHealth, retrieves official OIE notifications through the WAHIS platform. BioCaster and Argus also collect WHO alerts.

EBS systems are more heterogeneous than IBS regarding the range of sources they monitor, with a number of them retrieving validated events from official sources or another EBS system. In such cases, the EBS systems act as aggregators, and the detected signals are systematically validated and displayed. Web aggregators gather material from a variety of sources.

An RSS feed consists of a query, i.e. a combination of keywords and boolean operators (e.g. “and”, “or”). There are two types of query, corresponding to two distinct strategies. Disease-based (specific) queries contain names of diseases (including scientific names) and pathogens and aim at detecting timely information of a known disease. Moreover, non-specific queries do not contain any disease-specific terms, but instead host, symptom and outbreak-related keywords (e.g. “cases”, “spread”, etc.). These queries aim at detecting early signals of an unknown disease or a rare condition that is not yet known. They are implemented in HealthMap, Medisys, and PADI-web (Arsevska et al., 2017; Blench, 2008; Mantero et al., 2011).

The terms used in the queries are either proposed by domain experts or trained analysts (Medisys and GPHIN), by the system users (AquaticHealth.net) or based on a medical ontology (BioCaster) (Collier et al., 2007; Mantero et al., 2011; Mykhalovskiy and Weir, 2006). A mixed approach was proposed to create PADI-web queries, which combines automatic extraction of terms from a corpus of relevant documents and validation with a consensus of domain expert (Arsevska et al., 2016). The data retrieval frequency sharply differs, i.e. from every 15 min for the most reactive system (GPHIN) to every 24 h (PADI-web).

4.1.2 Document translation

Most EBS systems have implemented RSS feeds in several languages. As their pipelines are designed for texts in standard English, news are translated from their source language into English. This approach assumes that even lossy translation can be sufficient for machine learning components (e.g. classifiers) trained on English data.

The HealthMap system has a translation scheme that uses the article link (i.e. URL) to detect the original language automatically. For instance, if "trto=enandtrf=zh" appears in the link, the source language is identified as Chinese. Non-English articles are displayed in their language source in the HealthMap interface, and users are provided a link to Google Translate to enable them to access to the translated version.

Both PADI-web and GPHIN rely on the Microsoft Azure Translate API that was selected to detect the source language and translate the news content to English (Carter et al., 2020). Microsoft Azure is neural machine translation system[3], whereby each word in a sentence is coded along a 500-dimensional vector representing its unique characteristics within a specific language (Research, 2018).

Expert feedbacks suggest that automatic translations are somewhat noisy (Carter et al., 2020; Valentin et al., 2020c). Domain-specific terms, especially compound expressions, can result in erroneous translation. For instance, in PADI-web, the African swine fever disease was translated into various incorrect forms. Translation errors can negatively impact the extraction of epidemiological information, e.g. by creating spurious matches with geographical names. To overcome these shortcomings and limit the time lag needed by automatic translation, (Lejeune et al., 2015) proposed a language-independent approach that directly extracts information from the news in the native language. This sharply different approach, implemented in the so-called Data Analysis for Information Extraction in any Language (DAnIEL) system, handles language variations by processing text at the character level, rather than at the word level (Lejeune et al., 2012). This allows the detection of subpatterns of location and diseases names, such as "deng" in denge and dengi, two Polish variants of dengue.

4.1.3 Document triage: de-duplication and related documents

De-duplication is the fact of eliminating redundant information. In the NLP field, this task is linked with the text similarity concept: two documents having a similarity score above a predefined threshold are considered redundant. Text similarity encompasses two notions: (i) **semantic similarity** and (ii) **lexical similarity**. The lexical similarity between two documents relies on characters or lexical units that occur in both texts (Corley and Mihalcea, 2005). It is usually calculated in terms of edit distance (Guégan and Hernandez, 2006), lexical overlap (Jijkoun and de Rijke, 2005) or largest common substring. Another common approach derived from the lexical overlap involves representing two documents as dimensional vectors, where each dimension corresponds

[3]The principle of artifical neural network (ANN) models is briefly discussed in Chapter 2 Section 2.2.3

to a term (Gudivada et al., 2018). This representation relies on the segmentation of texts into words, referred to as "bag-of-words" (described in Section 2.2.1). The similarity between two documents can be further calculated by different metrics, such as the cosine of the angle (Cosine similarity) or the Jaccard coefficient (Gomaa and Fahmy, 2013).

To a certain extent, lexical similarity methods are often used to achieve semantic similarity. However, they fail to capture all semantic similarity in trivial cases, e.g. the use of different terms to describe the same concept. More sophisticated methods have thus been proposed: topological or knowledge-based methods, which rely on semantic and ontological relationships between the words (e.g. polysemy, synonym, etc.), corpus-based methods, which learn statistical similarity from data (e.g. latent semantic analysis) and recent word embedding representation (Billah Nagoudi et al., 2017; Majumder et al., 2016; Nguyen et al., 2019).

In EBS systems, de-duplication is often referred to as a single task. However, it is essential to define cases in which redundant information is retrieved, to highlight the different objectives that the de-duplication step addresses and the method involved.

First, the **same document** can erroneously be retrieved several times, either because it matches several queries integrated into the EBS system, or it is detected several times by the same query. This case is not a real de-duplication problem since it is based on the document's metadata rather than its textual content. PADI-web and AquaticHealth.net check the URL, while MedISys compares the title with those of already retrieved documents (Steinberger et al., 2008; Valentin et al., 2020c).

Second, an online source can reproduce the content of another source identically, thus generating **real duplicates**, also referred to as "exact duplicates" (Brownstein et al., 2008). This usually occurs with regard to news contents released by press agencies. It is also frequent that a source reproduces a document with slight modifications, by slightly rewording and adding or removing extra paragraphs of context (Lyon et al., 2011). Real duplicates are filtered using lexical similarity since they aim at detecting identical or highly similar documents. The GPHIN similarity score corresponds to the percentage of word triplets in common between two documents. As the similarity threshold is flexible, this method allows detection of both identical and highly similar documents (Carter et al., 2020). MedISys uses cosine similarity, calculated from the first 200 words of each article. The publication date is also taken into account—the similarity is computed only if the news articles were published within 8 h of each other at most (Ralf et al., 2008). To evaluate the number of single news articles retrieved by different EBS systems, (Lyon et al., 2011) proposed the $2 * M/T$ similarity ratio, where M is the number of common terms and T is the total number of terms. Documents whose pairwise ratio is higher than 0.9 are considered as duplicates. AquaticHealth.net uses the Python SequenceMatcher class, whose algorithm is based on the largest common subsequence (Wagner and Fischer, 1974).

Two documents could eventually relate to the same event without being real duplicates, i.e. having profoundly different textual contents. These documents are hereafter referred to as **related documents** (also called similar documents). This step is usually addressed after the event

extraction step by using the document epidemiological information to aggregate associated events Section 4.2.3. In the HealthMap system, this step is referred to as "news clustering", whereby news articles are grouped based on their lexical similarity and the specific epidemiological features they contain (same disease and same location).

When identified by the EBS system, **real duplicates** are usually considered as noisy and redundant data. They are thus filtered out by the system (HealthMap, PADI-web, AquaticHealth.net) or stored in a cluster, with the earliest document being used for exemplar identification (GPHIN, AquaticHealht.net).

4.1.4 Document triage: relevance classification

The classification steps of the EBS system differs regarding: (i) the number of categories used to label the documents, (ii) the type of classification method, and (iii) the kind of moderation Table 1.

Most EBS systems (GPHIN, Argus, Medisys, HealthMap, AquaticHealth.net and PADI-web) involve binary classification: news articles are either relevant or irrelevant. Interestingly, no formal definition of what relevance is, which poses a significant limitation when comparing EBS system performances. Besides the lack of shared gold standard and annotated resources hampers knowledge and experience sharing.

Two types of classification method, i.e. PADI-web (first version, Arsevska et al. (2018)), Mediys and AquaticHealth use a keyword-based approach; while GPHIN, Argus and HealhtMap rely on machine learning-based classifiers.

Keyword-based classification

In its first version, PADI-web categorised collected news articles by using a list of 32 outbreak-related keywords, i.e. "positive keywords". More precisely, news articles are classified as relevant if they contain—in the title or the body—one of the text positive keywords related to an outbreak event (e.g. "outbreak", "cases", "spread") (Arsevska et al., 2017).

MediSys classification relies on a more sophisticated approach involving boolean combinations and keyword weighting. A document is considered relevant if it matches one of a predefined set of alerts (Mantero et al., 2011; Steinberger et al., 2008). Two types of alerts are implemented: (i) single keyword alerts and (ii) combination alerts. A single keyword alert consists of attributing positive and negative weights to relevant and irrelevant keywords, respectively. An article is kept if the sum of the keyword weights it contains is above a specific threshold. A combination alert consists of relevant keywords combined by boolean expressions (i.e. « AND » and « AND NOT »). Documents are selected if they contain at least two relevant keywords (« AND » combination) and do not include any irrelevant keyword (« AND NOT » combination).

In AquaticHealth.net, news articles are tagged by the users if it contains specific key terms, usually scientific names for diseases. This strategy is based on the assumption that the authors using correct scientific terminology are more likely to disseminate accurate and relevant information.

Machine learning-based classification

HealthMap, BioCaster, Argus and GPHIN rely on a supervised machine learning classifiers, namely three Bayesian algorithms and support vector machines (SVM), respectively. The classifiers are trained on manually-labelled datasets and automatically learn rules to label the unclassified retrieved news articles (so-called supervised classification).

HealthMap's Bayesian algorithm automatically assigns a label out of the five described in Table, but the size of its training set is unknown. GPHIN automatically computes a relevance score for each retrieved report. This score corresponds to the confidence estimate of the SVM classifier, trained on approximately 1,400 documents (Carter et al., 2020). In Argus, relevant articles are identified by keyword-matching (with a set of concepts and keywords relevant to infectious disease surveillance) combined with Bayesian software tools. The performances of HealthMap, GHPIN and Argus classifiers are unknown.

Experts further evaluate the automatic classification in HealthMap, GPHIN and Argus systems. In GPHIN, articles with a high relevance score are published immediately, while the system discards low-scoring reports automatically. The remaining medium-relevance reports are triaged by analysts. Analysts also review automatically discarded articles to verify that relevant information has not been erroneously filtered out by the automated system (Carter et al., 2020).

BioCaster classification is totally automated. A naive Bayes classifier was trained on a gold standard corpus of 1,000 news articles. Each annotated article was manually assigned one of four relevance categories: alert, publish, check, and reject. However, a binary classification was implemented with the alert, publish and check classes being merged into a single category (relevant). The classification obtained an F-measure of 0.93 (Conway et al., 2009).

Table 1 Categories used to classify news in EBS systems. "N/A" indicates that the no definition was provided.

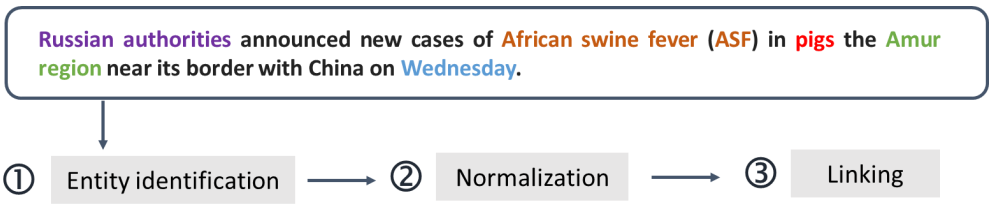
	Category	Definition (expert-based)	Classification method
HealthMap	Relevant	Newly discovered outbreak	News article classified as relevant (automatic classification and subsequent manual validation)
	Irrelevant	N/A	News article classified as relevant (automatic classification)
GPHIN	Relevant	News connected to the possibility of future international disease outbreaks, whether natural or the result of suspected bioterrorism.	New article with a high relevance score
	Irrelevant news	N/A	New article with a low relevance score (automatic classification) News article with a medium relevance score (automatic classification and subsequent manual validation)
PADI-web	Relevant	News mentioning outbreak(s) or suspicion(s) thereof.	News article containing a positive keyword
	Irrelevant news	N/A	News article not containing a positive keyword
MedISys	Relevant	N/A	News article verifying an alert definition
	Irrelevant	N/A	News article not verifying an alert definition
AquaticHealth.net	Relevant	Information that is relevant to aquatic animal health, often not about disease outbreaks, not even indirectly.	News article containing a positive keyword
	Irrelevant	N/A	News article not containing a positive keyword
Argus	Relevant	N/A	News article classified as relevant (automatic classification and subsequent manual validation)
	Irrelevant	N/A	News article classified as irrelevant (automatic classification)

In its second version, PADI-web integrates a supervised classifier, as described in Appendix C.

In addition to relevance classification, several systems attribute a meta-category to news articles based on the keywords they contain, thus further enhancing information retrieval by users. In several systems which monitor several health domains, it allows classification of a news article as plant in AquaticHealth; news articles are automatically tagged according to the search terms which retrieved it. For instance, an article found by using a specific disease term is tagged with the disease and its pathogen. The user controls the list of tags to adapt to changing contexts and information needs.

4.2 Entity level

Three main tasks are applied at the entity level, i.e. (1) their recognition in the text (entity extraction), (2) their normalisation and linking with external knowledge resources, and (3) their interpretation (event extraction). The first two steps are summarized in Figure 5, and the steps are detailed in the following sub-sections.



Value	Type	Normalized form	External ressource identifier
African swine fever, ASF	Disease	African swine fever	UMLS id: C0001752
pigs	Host	pig	UMLS id: C1135183
Amur region	Location	*50.27961, 127.5405*	Geonames id: 2026609
Wednesday	Date	2019-06-27	_
Russian authorities	Organisation	_	_

Figure 5 Illustration of entity level processing subtasks, including entity extraction, normalization and linking.

4.2.1 Entity extraction

Information extraction (IE) aims at locating specific pieces of data in natural-language documents, thereby extracting structured information from unstructured text (Mooney and Bunescu, 2005). Entity extraction, also called named entity recognition (NER), is an IE subtask that seeks to locate and classify textual elements into predefined categories, such as:

- locations (e.g. “Lagos”, “China”),
- temporal expressions (e.g. “last month”, “July 28, 1990”),

- organisations (e.g. “Ministry of Health“),
- person names,
- quantities (e.g. “2”), etc.

This list of predefined categories can be extended to include domain-specific entities (thematic entities). In the animal health domain, this involves:

- disease names (e.g. “avian influenza”, “AI”),
- causal agents (e.g. “H5N1 virus”),
- animal species (e.g. “chicken”),
- symptoms (e.g. “appetite loss”), etc.

Regarding geographical entities in online news, it is important to distinguish: (i) geographic entity extraction and resolution from (ii) identification of the event-related location. Geographic entity extraction and resolution aim at correctly extracting and identifying all locations from a text.

Broadly, two types of approaches are used to extract entities from texts : (i) **dictionary-based** approaches and (ii) **classifier-based** approaches. Both methods can be combined with expert-built rules.

The **dictionary-based** approach simply involves matching terms from a document with a list of words. Geographical dictionaries are usually called gazetteers. Some dictionaries can have an ontological structure rather than a simple list of terms. Ontologies aim at modelling the relations between entities (Guarino et al., 2009). In the Geonames ontology, for instance, spatial entities are structured into different hierarchical classes identified by a letter, with each of the letter corresponding to a precise category (e.g. A for administrative borders). In the health domain, an ontology can represent the causality relationships between a disease and a pathogen (Chanlekha et al., 2010).

The coverage of the word list is a substantial limiting factor of dictionary-based methods. Dictionary and ontologies need regular updates to include new terms, which requires time-consuming manual work. Besides, the type (i.e. class) of a named entities can be ambiguous. For instance, the term “May” can refer either to a date, a location or a person’s name. In the sentence “The virus can be transmitted between pigs by their body fluids”, the term “body fluids” refers to the transmission route, yet in another context it may relate only to an anatomy concept. In location extraction, this level of ambiguity is referred to as geo/non-geo ambiguity (Amitay et al., 2004).

To overcome the rigidity of the dictionary-based approach, another approach consists of considering NER as a classification task, where the type of entity is the label to assign. Extraction rules can be generated by hand or automatically. The latter case method relies on machine learning trained on manually annotated data. Conditional random fields (CRF) is amongst the most

prominent classifier used for NER (Lafferty et al., 2001), at the core of well-established pretrained NER tools, including StanfordNER (Manning et al., 2014) and NLTK (Bird and Loper, 2004). This approach is designed for sequential data: CRFs predict the probability of output sequence by giving an input sequence (Song et al., 2019). Classification approach is particularly suitable for misspelled locations or texts short in terms of length such as tweets, for which gazetteer lookup suffers from low precision due to irrelevant matches (Inkpen et al., 2017). While **classifier-based** approaches achieve good results, they are limited to the predefined categories upon which they are trained, i.e. unspecific domain entities (e.g. dates, locations, etc.). Recent tools increasingly allow users to add new types of entities to NER algorithms by training its model on annotated datasets, such as the neural network-based NER algorithm from SpaCy package (Honnibal and Montani, 2018). Locations are also prone to another level of ambiguity that occurs when several distinct places have the same name, i.e. the *geo/geo ambiguity*, or *referent ambiguity*. Several methods are described in the literature to address geo/geo ambiguities, based on the spatial entity itself (e.g. selecting the entity with the largest population (Amitay et al., 2004)), the spatial proximity between the spatial entities from the text (Li et al., 2003) or the frequency of a spatial entity. Heuristics can combined several of these aspects (Martins et al., 2008).

We further detail the NER approaches integrated into EBS systems by distinguishing the extraction of domain-specific (thematic) and domain-unspecific entities.

Domain non-specific entities

HealthMap extracts locations using a dictionary of 2,300 location place patterns. The extraction is consolidated with heuristics to infer people's job titles and full names and acronyms of organizations. Redundant locations are further filtered out based on container relationships by keeping the highest granular level of information. For instance, if "Boston" and "Massachusetts" are identified as locations, "Massachusetts" is eliminated (Freifeld et al., 2008).

PADI-web extracts locations by matching the text with the GeoNames ontology (Ahlers, 2013), and identifies dates using the HeidelTime rule-based system (Strotgen and Gertz, 2010). GPHIN extracts several domain-unspecific entities (e.g. person names, organisations and locations) with the classifier-based Stanford CoreNLP NER.

AquaticHealth only extracts locations using the Alchemy Location Extraction application programming interface (API) developed by IBM. Users can further manually add or refine locations from a report (Lyon et al., 2013a).

Thematic entities

In all EBS systems, thematic entity extraction is **dictionary-based**. The lists used are either external knowledge resources (GPHIN) or manually built by domain experts (AquaticHealth.net, PADI-web, HealthMap, MedISys, BioCaster). GPHIN extracts medical entities (i.e. syndromes and disease vectors) by combining matching with UMLS and expert heuristics. A hand-curated list of

frequent false positive terms is applied to filter out irrelevant terms. All EBS systems extract at least the disease name. GPHIN, MedISys, BioCaster and PADI-web extract symptoms. PADI-web also detects the host species and the number of cases using regular expressions. Both MedISys and BioCaster use their own ontologies to extract both thematic and domain-unspecific entities (Collier et al., 2007; Ralf et al., 2008). The multilingual BioCaster ontology (BCO) contains 18 classes encompassing both epidemiological concepts (e.g. virus, symptom) and generic concepts (e.g. locations) (Kawazoe et al., 2008, 2006).

4.2.2 Entity normalisation and linking

Domain non-specific entities

Mapping locations to an external gazetteer has several advantages, as it allows: (i) geocoding to map the detected location via latitude-longitude coordinates, (ii) inferring parent-child relationships between different granularity levels to group synonymous mentions, or to synthesise very local information from a global perspective (e.g. at the country level), (iii) geotagging a document to improve information retrieval from EBS databases.

Both PADI-web and GPHIN map geographical entities with the GeoNames gazetteer. In PADI-web, location mapping is merged with the entity extraction step as described above, while these two phases are separate in GPHIN. Using a classification-based approach prior to matching with an external knowledge resource reduces geographical and non-geographical ambiguities that occur when a noun is the same as an existing location name. For instance, the term "More" may erroneously match the city of More in England in the PADI-web pipeline. However, in both cases, a place name has multiple entries in the gazetteer, thus creating geographical-geographical ambiguities. In GPHIN, such cases are resolved through heuristic rules that take where an article was published into account, but further details on the procedure are not available. PADI-web does not address this problem, and all entries are kept.

In AquaticHealth.net, locations are geocoded using the Google Maps API so that reports can be presented on a Google Map on the system's website.

Thematic entities

Thematic entities are usually normalised to their canonical form (e.g. disease acronyms are converted into the full disease name). GPHIN provides a link between the detected entities and UMLS terminology and definitions. BioCaster Ontology provides access to term definitions, synonyms and translations in eight languages, along with a link to medical ontologies (including ICD-10, MedDRA, MeSH and SNOMED-CT) (Collier et al., 2008).

4.2.3 Event extraction

The term "event" is used differently in the medical and computational communities. In epidemiology, an event denotes the entire course of an epidemic. In health informatics, an event is a single "factoid" belonging to a group of factoids that jointly describe the entire epidemic. In the rest of this chapter, we adopt the computational definition for consistency with the event extraction terms. Event extraction encompasses several subtasks defined by a specific terminology (Ahn, 2006; Pyysalo et al., 2015; Xiang and Wang, 2019):

- Detecting an event from the text, i.e. recognising the event triggers or event mentions, and the event type,
- Identifying the event arguments (or "attributes"),
- Determining the roles of the arguments,
- Resolving event co-reference.

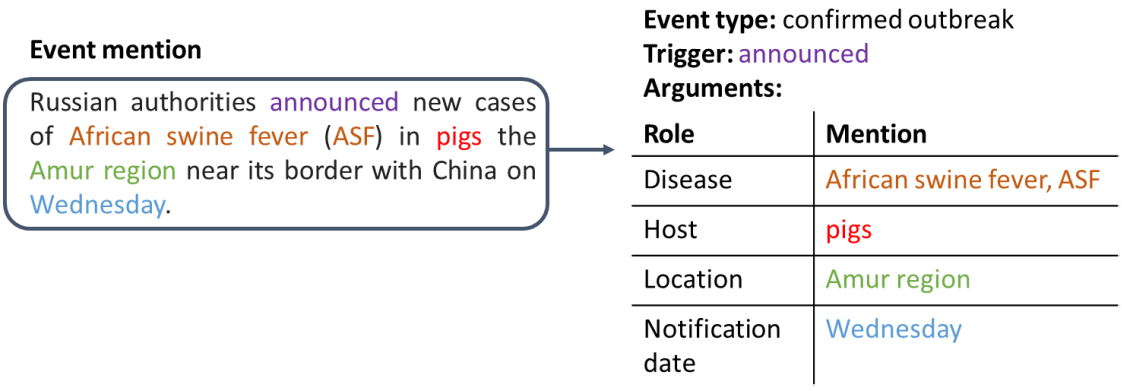


Figure 6 Illustration of the extraction results of three event extraction subtasks, i.e. event mention/trigger detection and event type identification, argument detection and argument role identification.

An event mention is a phrase or sentence that describes the event, while containing an event trigger and its arguments (Figure 6). A trigger (or "anchor") is a semantic clue indicating the presence of an event (e.g. the presence of the verb "detected"). The trigger is specific to the event type (e.g. attack, outbreak, etc.). Arguments (or "attributes") are entities having a specific role in the event, e.g. the location where an outbreak occurred. The role of arguments concerns the relationship between an argument and the event in which it participates. Event co-reference consists of identifying the multiple event mentions of the same event.

Event extraction methods have been extensively studied in many domains such as business and financial (Du et al., 2016), biomedical (Zhu and Zheng, 2020), and outbreak-event detection (Piskorski et al., 2011) domains. (Xiang and Wang, 2019) propose a comprehensive and synthetic

survey of event extraction methods. Briefly, they include pattern-based methods, machine learning methods (supervised or semi-supervised), deep learning methods, and unsupervised methods.

In the studied EBS systems, HealthMap, PADI-web, BioCaster and MedISys include an event extraction step, all of which rely on a different approach.

Unsupervised approach

HealthMap event extraction is unsupervised, i.e. it does not train any event extraction models. Typically, in the unsupervised approach, the detection of triggers and arguments is based on word distributional representations. In HealthMap, the event mention and trigger detection steps are ignored. Instead, the approach relies on the document structure, based on the hypothesis that the most relevant information (i.e. event attributes) appears at the beginning of a news report. Diseases and locations are first searched in the title, then in the document headlines, and finally in the full content. If the algorithm is unable to extract relevant elements from these three levels, the name of the online news source is used instead. This last step relies on the assumption that news articles which do not contain any specific location refer to a place near the publication source. Erroneous extractions are further corrected by analysts when necessary (Brownstein et al., 2008). This approach decreases the risk of false-positive extraction (i.e. extraction of locations which are not en event attribute). Two shortcomings should be noted: the spatial granularity is decreased since the attribute extraction stops at the first entity detected, and also this approach cannot address cases of news articles containing several events.

Pattern-based approach

Both Medisys and BioCaster rely on pattern-based methods, which were the earliest approaches proposed for event extraction. They consist of matching text with specific event templates. Patterns are constructed manually or automatically. Manual event construction typically relies on domain expert proposals, thereby achieving high accuracy. However, manual construction is time-consuming, and expert bias can lead to a lack of recall. A weakly supervised method or bootstrapping can generate patterns automatically from a pre-classified training corpus or from seed patterns. In MediSys, event extraction is performed by the Pattern-based Understanding and Learning System (PULS) developed at the University of Helsinki. PULS relies on a cascade of patterns applied to each sentence of the news article content to extract the event attributes. Patterns use both syntactical and semantic information of the sentence, such as:

NP(disease) VP(kill) NP(victim) ['in' NP(location)]

This pattern matches a noun phrase (NP) of semantic type, i.e. "disease"; a verb phrase (VP) headed by the verb "kill" (or its synonyms in the ontology) and has the adverbial phrase "so far", etc. The square brackets indicate an optional match. If the location is omitted in the sentence, it is inferred from the surrounding context. Verb phrases are not rigid and allow the presence of

Table 2 Advantages and limitations of event extraction methods implemented in EBS systems.

EBS	Method	Advantage	Limitation
HealthMap	Unsupervised	High precision No manual annotation required Easy implementation	Need high document classification precision Decreased spatial granularity. Not suitable for multi-events. Event status unknown
Medisys (PULS)	Pattern-based (manual)	High precision High spatial granularity Tolerant of document classification errors Suitable for multi-events	Time-consuming Lack of recall
BioCaster	Pattern-based (semi-supervised)	Reduced manual work for patterns creation High spatial granularity Suitable for multi-events	Manual curation of a spatial entity list
PADI-web	Machine learning based	High spatial granularity High precision	Needs high document classification precision Event status unknown Not suitable for multi-events.

modifier elements, such as the auxiliary verb (e.g. "has") or adverb (e.g. "so far") (Steinberger et al., 2008). PULS implements weakly-supervised learning to reduce the amount of manual labour as far as possible by automatically learning new patterns via bootstrapping (Grishman et al., 2002).

BioCaster event extraction uses a simple rule language (SRL), inspired by the so-called Declarative Information Analysis Language (DIAL) (Feldman et al., 2001). SRL creates sophisticated matching patterns combining entity classes, string literals, regular expressions, entity types such as verbs of infection, common victim expressions, occupation names, etc.

Machine learning-based approach

While several machine-learning based methods have been proposed for event extraction (Margineantu et al., 2010), they are not yet implemented in any operational EBS system for animal health surveillance.

The PADI-web event extraction module aims at identifying dates and locations which are the attributes of an event. By default, diseases and hosts that appear in the text are considered as event attributes. The approach relies on rules that were discovered automatically by a data mining technique known as frequent itemset discovery. This technique can reveal correlations in a large volume of data by providing elements that frequently occur together in a document. Each

candidate (i.e. dates and location identified) is transformed into a set of features describing both the candidate and its context:

- Elements related to the candidate entity. Each word positioned around a candidate is encoded as an element which describes both the word itself and its position relative to the candidate. For example, the element (infected, -5) means that the word "infected" is four words before an entity candidate. The position of an element is also expressed according to the relative position in a sentence.
- Elements related to the candidate's part-of-speech. Each entity is associated with its grammatical function (e.g. verb) or its lemma, both produced by TreeTagger, a tool for annotating text with part-of-speech (Schmid, 1994).
- Elements related to the position. This position is expressed with respect to the paragraph in which the candidate is located. For example, the element (position, 0%-10%) means that the current candidate occurs in the first 10% of the document, while (position, PAR1) means that it is in the first paragraph.

These features allow attribute candidate identification. An SVM classifier is further applied to predict if the candidate is correct of not (Arsevska et al., 2018).

Advantages and limitations of the different event extraction approaches are summarized in Table 2. Both PADI-web and HealthMap do not rely on the detection of an event trigger (or mention) in the news content. Rather, they hypothetise that the documents classified as relevant describe a new event (event trigger detection is replaced by document classification). This hypothesis requires highly efficient classification of news articles upstream of the event extraction. HealthMap automatic classification is manually verified to avoid potential errors. In PADI-web, the extraction of spatial entities from a document wrongly classified as relevant led to a high number of false alerts. Besides, neither PADI-web nor HealthMap are designed for multi-events: HealthMap extracts a single disease–location pair, while PADI-web is able to extract all possible locations and diseases, but without computing the link between them.

4.2.4 Event relevance

In HealthMap, two strategies are used to compute the relevance of an event. The noteworthiness is a user-based indicator. Each HealthMap user is allowed to rate the significance of an alert (scale from 0 to 5). If the alert has not yet been manually rated, a composite score is calculated based on the news volume associated with the latter and the disease importance. This score (called **heat index**) is calculated based on: (i) the reliability of the data source (e.g. increased weight is given to official reports and reduced weight to local media); and (ii) the number of single data sources reporting the event, with an increased weight if different types of sources are involved (i.e. formal and informal sources). This latter component is based on the idea that multiple sources of information about an event provide greater confidence in the reliability of the report

than any single source alone (Brownstein et al., 2008). Explicit details about this algorithm have not been published.

In GPHIN, a proprietary algorithm is applied to assign a relevancy score to each news item, which is derived based on the values attributed to the keywords and terms it contains (Blench, 2008).

EBS systems which involve a dedicated team of analysts and experts, the relevance of the news is determined manually. In Argus, analysts write an event report for each relevant news item and assign a stage based on a detailed heuristic model devoted to the assessment of biological events (Hartley et al., 2010). The model includes six stages, ranging from preparatory (e.g. prevention activities and conditions favourable for disease emergence) to recovery (Wilson and Brownstein, 2009). In AquaticHealth.net, all registered users can create summaries of the most relevant information about an event, and they are allowed to add broader context (Lyon et al., 2013b).

4.2.5 Spatio-temporal abberation detection

Beyond event extraction, GPHIN and MedIsys integrate statistical methods to detect spatiotemporal abberations.

In GPHIN, statistically unusual pairs of the form (keyword, location) are detected (referred to as "narratives"). This approach produces between 20 to 50 alerts daily, which are further manually analysed. The authors have highlighted that the stream of news articles retrieved by the EBS systems is too noisy to be able to use highly sensitive aberration detection algorithms (Carter et al., 2020). Besides, GPHIN applies topic modelling methods to estimate when the language of the narrative is changing, e.g. from uncertain to certain. However, none of these unusual pair detection methods nor the topic modelling are described in further detail. MediSys counts the number of documents mentioning a given disease and country over a 2-week time window. Sudden increases in the last 24 h compared to a two-week rolling average create an alert (Mantero et al., 2011).

In the light of the multiple challenges of automating the processing of online news, the main objective of this thesis is to evaluate methods for enhancing several tasks in EBS pipelines. We propose and compare approaches based on the use of epidemiological entities and involving NLP and machine learning techniques. More precisely, in the following chapters, we describe the retrieval of new kinds of epidemiological information and propose how to use epidemiological features to enhance both event extraction and related news retrieval.

Table 3 Characteristics of EBS systems encompassing animal health threats.

Systems	GPHIN	MedISys	BioCaster	HealthMap	Argus	PADI-web	AquaticHealth.net
General characteristics							
Year launched	1997	2004	2006	2006	2004	2016	2013
Institution	Public Health Agency (Canada)	Joint research centre (European Union)	Tokyo University	Boston Children's Hospital (USA)	Georgetown University Medical Center (USA)	ESA Platform (France)	University of Melbourne (Australia)
Geographical coverage	World	World	World	World	World	World	World
Type of threats	A, H, P, E	A, H, P, E	A, H, P, E	A, H, P, E	A, H, P	A	A
Interface (access)	Restricted	Public, restricted	Public	Public	Restricted	Public, restricted	Public
Interface (languages)	5	9	50	2	1	2	1
Export of structured data	No	Yes	No	Yes	No	Yes	No
Automatic alerts	E-mail	E-mail	Email, RSS, Twitter	E-mail, social media, RSS feeds and smartphone	E-mail	No	E-mail
Document level							
Document retrieval							
Official data sources	No	No	Yes	Yes	Yes	No	Yes
Non-official data sources	Web	Web	Web, EBS systems	Web, EBS systems, Twitter.	Web, blogs	Web	Web, Twitter, EBS systems, scientific literature, users.
Frequency	15 min	Hourly	Hourly	Hourly	Six times daily	Daily	Hourly
Number of languages	10	43	13	7	40	8	Not available
Moderation	Automatic and manual	Automatic	Automatic	Automatic and manual	Automatic and manual	Automatic	Automatic

Table 3 (continuation)

Systems	GPHIN	MedISys	BioCaster	HealthMap	Argus	PADI-web	AquaticHealth.net
Document translation							
Method	Automatic	No translation	Only for NE	Automatic	Manual	Automatic	Not available
Available in the system interface	Yes	No	No	No	Yes	Yes	Not available
Document filtering: de-duplication							
Duplicates removal	Automatic	Automatic	Automatic	Automatic and manual	Automatic	Automatic	Automatic
Document clustering	Yes	Yes	No	Yes	No	No	Yes
Document filtering: relevance classification							
Classification	Automatic and manual	Automatic	Automatic	Automatic and manual	Automatic and manual	Automatic	Automatic
Entity level							
Entities extraction	External resources and heuristics	Ontology	Ontology	Dictionary	Dictionary	Dictionary and machine-learning rules	Dictionary, users tags
Entity linking	Yes	No	No	No	No	Yes	Yes
Event extraction	No	Pattern-based	Pattern-based	Unsupervised	No	Rule-based	No
Event relevance	Automatic	No	No	Automatic	Manual	No	Manual

Chapter 2

Epidemic intelligence and information retrieval

Table of contents

In this chapter, we describe how retrieving fine-grained epidemiological information from animal-health news improves the performance of current EBS, while presenting the two following contributions:

1. First, we outline the sentence-based annotation framework developed during this thesis—it enhances the identification of new types of epidemiological information of relevance to EI.
2. Second, we evaluate and compare the performances of two methods (i.e. supervised classification and pattern-based classification) to retrieve fine-grained epidemiological information.

1 Elaboration of a new framework for fine-grained epidemiological annotation

In this section, we first describe the needs for developing a new annotation framework by highlighting the limitations of current approaches and available resources. We further describe our global protocol for guideline elaboration, followed by a detailed description of the final annotation guidelines. We discuss how we addressed the annotation challenges of the global process, and we highlight the contributions and limitations of our framework.

1.1 Motivation and context of framework elaboration

1.1.1 Collaboration context

This work was conducted as part of a collaboration with the Belgian Moriskin (Method for threat analysis in the context Of the RISK of emergence or re-emergence of INfectious animal diseases) research project. Moriskin aims at developing an EI system for horizon scanning of animal infectious diseases and assessment of the risk of outbreaks in Belgium. This project is funded by the Belgian Federal Public Service (Health, Food Chain Safety and Environment) and coordinated by Sciensano (Institute of Public and Animal Health). The system—currently in development—will pool semi-automatically outbreak data from multiple sources: (i) formal sources (structured data), such as infectious diseases events reported by international agencies (e.g. WAHIS, Animal Disease Notification System), and (ii) informal sources through PADI-web.

1.1.2 Objectives of the new framework

Classification in text mining usually assigns a single topic (category) per news piece (document-based classification). However, animal-health news is rich in different types of epidemiological information. For instance, news articles that report an outbreak often also describe outbreak control measures or economic impacts, share information about the outbreak source or draw attention to a given area at risk (Figure 7). Those elements may be of relevance to EI teams to assess risks associated with the occurrence of an event.

(1) “The official said he had heard reports that some smallholders had been dumping the corpses of infected pigs into the Danube, suggesting the highly contagious virus might have been spread by river water”.
(2) “The concern for Australia is the closeness of East Timor to Australia; the distance from the two East Timor municipalities and Darwin is only 650 km.”

Figure 7 Example of sentences extracted from two disease-related news.
The first sentence provides suspicion about the spread of African swine fever virus in Romania through water (1). The second sentence indicates Australia’s concern for risk of ASF introduction from East Timor (2).

When a news piece contains several topics, a single-label classifier has to decide on a topic (i.e. a label) among the other ones, which usually decreases the classification performance (Zhang et al., 2009). Besides, as noted in Chapter 1 Section 4.1.4, most classification approaches in EBS systems focus on the binary news relevance. Little attention has been focused on the retrieval of other types of epidemiological information.

In this context, we propose to split news content into sentences that are annotated into different categories according to their epidemiological topic, which we refer to as fine-grained information. Empirically, sentence-level classification seems more homogeneous in terms of the topic than document-level classification. We therefore believe that sentence-level classification can more accurately identify specific types of information.

To create annotated data able to be integrated into a machine learning pipeline, we first need to elaborate and evaluate a generic annotation framework that should be as reproducible as possible. Besides, the list of classes for the sentence-based annotation should allow us to identify new types of epidemiological information in animal disease-related news.

1.1.3 Limitation of existing resources

Supervised learning algorithms implemented in EBS systems must be trained on labelled datasets to further classify unknown data. Several annotated textual resources thus have been created to support classifier training tasks in animal health. Table 4 presents examples of labelled datasets of news in the animal disease surveillance context. Datasets are compared based on their aim, the characteristics of the annotated data and their reproducibility in terms of availability (indicating whether the corpus and guidelines are freely available for download) and reliability (corresponding to the evaluation of inter-annotator agreement).

Depending on the context in which it was created (typically the scope of the EBS system), the labelled corpus is either generalist, i.e. encompassing both human and animal disease events (Chanlekha et al., 2010; Conway et al., 2010), or specific, i.e. targeting one or several animal diseases (Zhang et al., 2009). The annotation unit and labels (categories) closely depend on the aim of the text-mining tasks in the animal disease domain, i.e. i) classification, ii) named entity recognition and iii) event extraction.

- For classification tasks, annotation is usually at the document level. The labels are often related to the news relevance so as to filter out irrelevant ones (Conway et al., 2009; Doan et al., 2007; Torii et al., 2011; Valentin et al., 2020a). Other classification frames assign a broad thematic label to the news, such as “outbreak-related” or “socioeconomic” (Zhang et al., 2009). To our knowledge, all document-based annotation approaches allow a single label per news piece.
- For named entity recognition tasks, the corpus is annotated at the word level (including multi-word expressions). A typical example is the annotation framework of the BioCaster Ontology (Chanlekha et al., 2010) described in Chapter 1 Section Thematic entities.

- For event extraction tasks, the annotation unit depends on the definition adopted for the event. Some authors opt for a linguistic definition, i.e. a verb (called predicate) and a subject or object (called argument). Some sophisticated event annotation schemes allow extraction of fine-grained temporal information such as the beginning and end of an event (Chanlekha et al., 2010), or thematic attributes such as the transmission mode (Conway et al., 2010).

No currently available annotated data and frameworks can fulfil the needs of our current objective to detect fine-grained epidemiological information (i.e. topics). Document-based approaches are not precise enough to detect the variety of information contained in a single news piece. Word-based annotation frameworks provide accurate information at the word level, yet they are task-oriented (extraction of events or named entities) and partly address the potential of other types of epidemiological information. Midway between these two approaches, (Zhang and Liu, 2007) proposed a sentence-based annotation to detect outbreak-related sentences, while recognising that a news piece contains many sentences with different semantic meanings. However, as the primary goal was outbreak detection, outbreak-unrelated sentences (e.g. describing treatment or prevention) were all merged into one negative category.

In addition to the shortcomings mentioned above, the availability and reproducibility of the annotated data and guidelines vary between the studies. Several corpora were not published or are no longer available because of unstable storage. For instance, the BioCaster disease event corpus has to be retrieved through a Perl script that downloads documents from their web source. As some sources become unavailable, the corpus size inevitably decreases over time (only 102 source web pages among 200 were still available online in 2015 (Lejeune et al., 2015)). The availability of EBS tools also hampers data access—two EBS tools from Table 3 were no longer operational (Argus, BioCaster).

Most of the proposed approaches lack reproducibility. First, annotation guidelines usually consist of brief label descriptions rather than detailed schemes. Second, in the provided examples, only three annotation frameworks were evaluated in terms of inter-annotator agreement. According to biomedical text annotation recommendations (Wilbur et al., 2006), the BioCaster disease event corpus authors used percentage scores (pairwise agreement) rather than the kappa statistic (Conway et al., 2010). Chanlekha et al. (2010) included both metrics to evaluate the event annotation scheme. Torii et al. (2011) measured agreement using multi-kappa statistics (Artstein and Poesio, 2008) to take the five annotators involved in their scheme into account.

Similarly to the approach of (Zhang and Liu, 2007), we aim at sentence-level annotation to enrich the binary outbreak-related/unrelated classification with thematic categories. Our objective is to make effective use of the epidemiological information contained in the news, especially when the information is relevant for assessing an epidemiological situation.

Table 4 Example of annotated data used for online news processing in event-based surveillance applications.
C: classification, En. E: entity extraction, Ev. E: event extraction, ST: Spatiotemporal

Data source	Tasks	Annotated data			Annotation guidelines			Limitations
		Content	Annotation unit	Labels	Availability	Availability	Agreement evaluation	
BioCaster	C	Corpus of 1,000 news articles, generalist	Document	Alert, publish, check, and reject	No	Yes (brief)	No	Document-based Corpus unavailable Reproducibility
FMD BioPortal	C	Corpus of 1,674 news articles, specific	Document	Outbreak-related, control program-related, and general information	No	Yes (brief)	No	Document-based Corpus unavailable Reproducibility
Argus	C	Corpus of news articles, generalist	Document	Relevant, irrelevant	No	Yes (brief)	Yes	Corpus unavailable
ProMED	C	Corpus of 2,342 sentences, generalist	Sentence	Disease reports, not disease reports	No	No	No	Corpus unavailable Reproducibility
PADI-web	C	Corpus of 600 news articles, specific	Document	Relevant, irrelevant	No	Yes (brief)	No	Document-based Corpus unavailable Reproducibility
	En. E	Corpus of 532 news articles, specific	Entity	Location, date, disease, host, and number of cases	Yes	Yes (brief)	No	Word-based Reproducibility
BioCaster	En. E	Multilingual ontology, generalist	Entity	Location, organisation, time, etc.	Yes	Yes (detailed)	No	Word-based Reproducibility
	Ev. E	Corpus of 200 news articles, generalist	Event (disease – location pairs) and attributes	ST and thematic attribute categories: disease, host, food contamination, etc.	Partial	Yes (brief)	Yes	Event-based Corpus partially unavailable
	Ev. E	Corpus of 100 news articles, generalist	Event (a verb and a subject), ST attributes	Event categories: temporally-locatable, generic, and hypothetical ST attribute classes: starting time, etc.	No	Yes (detailed)	Yes	Event-based Corpus unavailable

1.2 Methods

In this section, we describe our approach for building annotated resources for the extraction of fine-grained epidemiological information. We first describe the global process we adopted to develop the annotation guidelines. Then we present the final annotation framework and describe the proposed categories (labels). The annotation guidelines and annotated corpus are publicly available (Valentin et al., 2019).

1.2.1 Global annotation process

We extracted candidate news from the PADI-web database, while focusing on those classified as relevant (Valentin et al., 2020a). The content of each news piece was segmented by sentence using the NLTK library (Bird et al., 2009). The annotation dataset provided to the annotators consisted of: (i) a list of segmented news articles (each news article corresponds to a set of sentences), (ii) the news article metadata (i.e. title, source, and publication date).

The four annotators were veterinarians working in epidemic intelligence. Two of them had previous experience with annotation tasks. During the process, we followed four consecutive steps (Figure 8). After each annotation step, we calculated the agreement metrics, as outlined in Section 1.3. Annotators discussed the main disagreement results and modified the guidelines to improve the annotation process. We describe the main modification choices that led to the final guidelines in Section 1.4.2. We stopped the process when satisfactory agreement measures were attained (step 3).

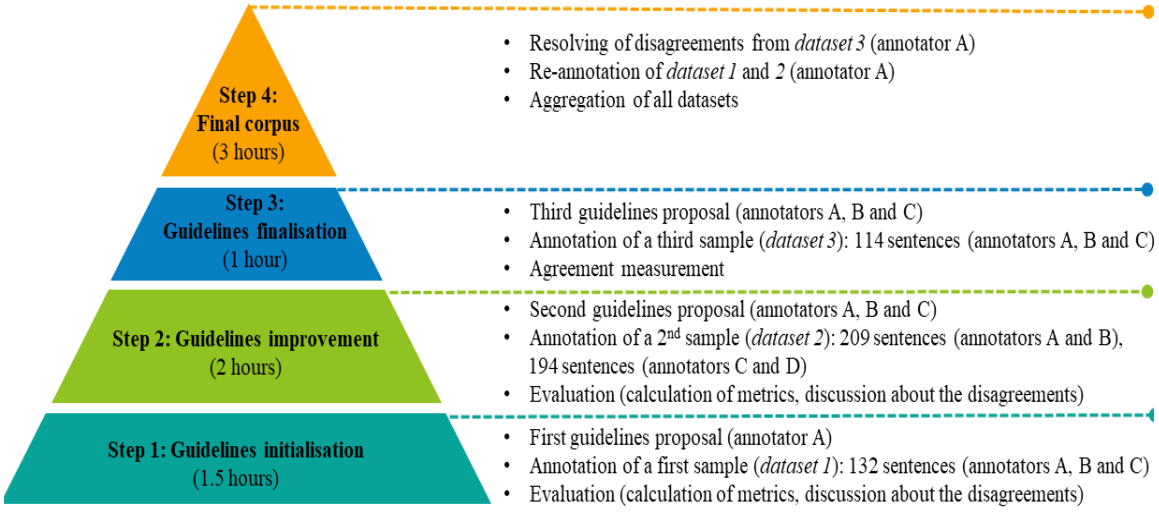


Figure 8 Pipeline of the annotation guideline elaboration process.
Indicative time per annotator is given

To build the final corpus (step 4), we aggregated all previously labelled datasets (datasets 1, 2 and 3). To choose one label per sentence in case of disagreement, we adopted the following procedure:

1. For dataset 3 (labelled with final guidelines):
   (a) If at least two out of three annotators assigned the same label, we selected the majority label;
   (b) If each of the three annotators assigned a different label, annotator A chose a final label among those proposed.
2. For datasets 1 and 2 (labelled with previous guidelines):
   (a) For unchanged labels, we followed the same guidelines as for dataset 3;
   (b) For new/modified labels, annotator A chose a label consistent with the final guidelines.

1.2.2 Annotation guidelines

In this subsection, we present the final annotation framework and definitions for each label from the guidelines. A detailed version of the guidelines and the labelled corpus are publicly available in a Dataverse repository (Valentin et al., 2019).

Our annotation framework relies on the attribution of two levels per sentence: Event type and Information type, as illustrated in Figure 9. Each category is further described. The Event type level indicates the relation between the sentence and the epidemiological situation. At this step, irrelevant sentences are discarded; these sentences are labelled "irrelevant" for Event type classification. The Information type level qualifies the type of epidemiological information, i.e. the fine-grained topics. The meaning of a sentence depends on the entire news content, as well as its epidemiological context. Therefore, for each set of sentences (from a single news piece), the annotator first reads the news metadata (i.e. title, source, and publication date). The annotator chooses a single label per level and per sentence. As some sentences may contain information belonging to several Information type categories, the annotator must pinpoint the primary information (label).

Event type

While focusing on sentence epidemiological topics, the relation between the sentence and the current epidemiological situation must be taken into account: sentences in news pieces may describe an outbreak that happened several years before or provide general information about a disease. More precisely, from the EBS standpoint, only sentences referring to current events or events at risk are of interest.

In our context, we define an event as the occurrence of a disease within a specific area and time range . The Event type label aims to differentiate sentences referring to the current/recent outbreak ("**Current event**" and "**Risk event**") from sentences referring to old outbreaks ("**Old event**") or general information ("**General**"). Sentences which do not contain any epidemiological information are considered irrelevant ("**Irrelevant**").

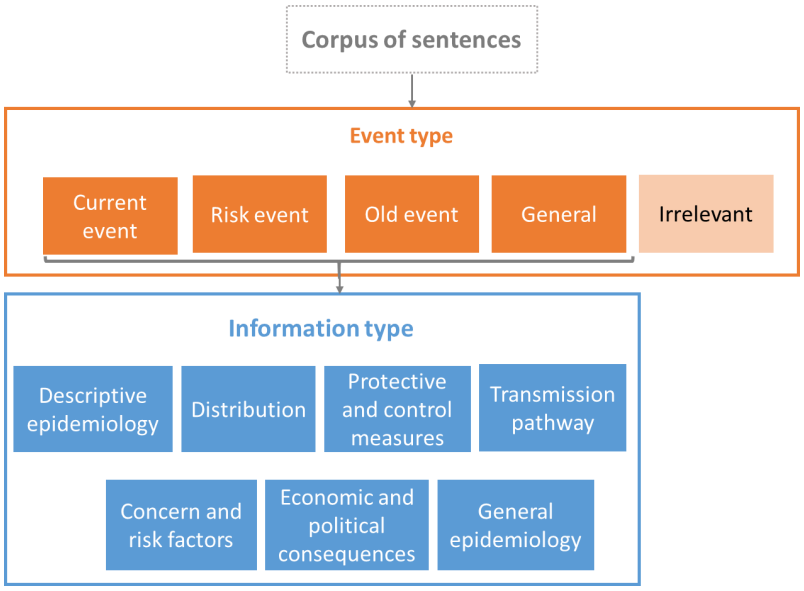


Figure 9 Two-level annotation framework.

- **Current event**: this class includes sentences related to the current situation. There are five major groups of sentences that are considered "current":

  1. **Recent event, relative to the main event**. This includes events occurring at a nearby location and/or within a short-time window around the main event. For instance, "On Saturday, similar infections were found in 30 pigs on a farm in the Huangpu district of Guangzhou."
  2. **Aggregation of events between a prior date and a recent/current date**. For instance, "According to data from the Council of Agriculture, 94 poultry farms in Taiwan have been infected by avian flu so far this year." The temporal expression "so far this year" indicates a relationship between the start of the outbreak and the publication date.
  3. **Recent/current epidemiological status of a disease within an area**. For instance, "In recent months, the disease has spread more rapidly and further west, affecting countries that were previously unscathed."
  4. **Events that will definitively occur in the future**. In general, this category includes the direct consequences of an event, such as control measures that will be taken. For instance, "All pigs in the complex will be killed, and 3 km and 10 km protection and surveillance zones will be installed."

- **Old event**: This class includes sentences about events that provide a historical context for the main event. Those sentences contain explicit references to old dates, either absolute ("In 2007") or relative ("Back in days"). This category includes two groups of sentences:

  1. **Old event**. For instance, "The most recent case of the disease in the UK came in 2007."

2. **Aggregation of events between two past dates**. For instance, "During last year, 132 cases were recorded across the country".
3. **Past epidemiological status of a disease within an area**. For instance, "Between 2006 and 2010, BTV serotype 8 reached parts of north-western Europe that had never experienced bluetongue outbreaks previously."

- **Risk event**: This class includes all sentences referring to hypothetical events. These sentences are generally about an area at risk of introduction or dissemination of a pathogen. This category includes two groups of sentences:

1. **An unaffected area expressing concern and/or preparedness**. For instance, "Additional outbreaks of African swine fever are likely to occur in China, despite nationwide disease control and prevention efforts."
2. **An area with unknown disease status**. For instance, "If the outbreak is verified, all pigs at the feeding station will have to be culled, Miratorg said."

- **General**: This class includes general information about a disease or pathogen. Conventionally, the sentences describe the disease hosts, its clinical presentation and pathogenicity. For instance, "Bluetongue is a viral disease of ruminants (e. g. cattle, sheep goats, and deer)."

- **Irrelevant**: This class includes sentences that do not contain any epidemiological information. This group includes, for instance, disease-unrelated general facts ("Pig imports from Hungary only represented about 0. 64 per cent of all pork products to the UK in 2017.") or article news artefacts ("Comments will be moderated.").

Information type

The Information type level describes the sentence epidemiological topic. As epidemiological topic, we include the notification of a suspected or confirmed event, the description of a disease in an area ("**Descriptive epidemiology**" and "**Distribution**"), preventive or control measures against a disease outbreak ("**Preventive and control measures**"), an event's economic and/or political impacts ("**Economic and political consequences**"), it's suspected or confirmed transmission mode ("Transmission pathway"), the expression of concern and/or facts about risk factors ("**Concern and risk factors**") and general information about the epidemiology of a pathogen or a disease ("**General epidemiology**").

- **Descriptive epidemiology**. This class includes sentences containing the standard epidemiological indicators (e.g. disease, location, hosts, and dates) that describe an event. It includes:

1. **Epidemiological description of the event**. For instance, "Cases of African swine fever (ASF) have been recorded in Odesa and Mykolaiv regions."

2. **Information about the pathogenic agent cause of the event**. For instance, "Results indicated that the birds were infected with a new variety of H5N1 influenza."
3. **Clinical signs of the suspected event**. For instance, "The remaining buck appears healthy at this time and is showing no clinical signs associated with the disease."

- **Distribution**. This class contains sentences giving indications on the presence of a disease in a specific area (i.e. a region, a country). It includes:

1. **Description of the epidemiological status**. For instance, "In recent months, the disease has spread more rapidly and further west, affecting countries that were previously unscathed."
2. **Aggregation of events between a past date and a recent/current date**. For instance, "According to data from the Council of Agriculture, 94 poultry farms in Taiwan have been infected by avian flu so far this year."

- **Preventive and control measures**. This class includes sentences describing:

1. **Preventive measures**, i.e. all sanitary and physical actions taken to avoid the introduction of a disease into an unaffected area. For instance, "ASF: France about to end the fencing in the borderland with Belgium."
2. **Control measures**, i.e. all sanitary and physical actions taken to eradicate a pathogen once introduced into an area (e.g. vaccination, slaughtering, disinfection, zoning, etc.). For instance, "All the infected animals have been killed, and the area has been disinfected."
3. **Instructions/recommendations**, i.e. actions for both preventive and control measures, we include recommendations in this class. For instance, "Hunters, travellers, and transporters are asked to take extra care concerning hygiene."

- **Transmission pathway**. This class includes the sentences indicating the origin (source) of the disease or the transmission route. For instance, "The authorities suggest that the highly contagious virus might have been spread by a river".

- **Concern and risk factors**. This class includes sentences indicating a risk of introduction or spread of disease in an area. We include two types of sentences in this group:

1. **Confirmation of suspicion of one or several risk factors**, i.e. an individual, behavioural and environmental characteristics associated with an increased disease occurrence. For example, "A recent wave of inspections revealed 4,000 different biosecurity violations on farms and Gosvetfitosluzhba warned that this could result in further outbreaks soon."
2. Semantic **expression of fears or concerns** regarding: (i) The hypothetical intrusion of a pathogen into an unaffected area. For instance, "ASF is a real threat to the UK," she said." (ii) The worrying development of a situation. For instance, "Several countries are

affected, alarming governments and pig farmers due to the pace at which the disease has spread."

- **Economic and political consequences**. This category includes all references to direct or indirect economic or political impacts of an outbreak on an area. It includes the consequences of preventive and control measures. For instance, "Gorod estimated that financial losses due to ASF could amount to €17 million to Latvia's industry in 2017."
- **General epidemiology**. This category is only used for the sentences labelled "General" as the Event type level. It merges the classes "Event description" and "Transmission pathway" described above. In this particular Event type level, those two categories include the description of a disease's hosts, pathogenicity and transmission route. For instance, "The virus is transmitted by midge bites, and it does not affect humans."

Multi-topic sentences

To handle multi-topic sentences, we provide two rules to help annotators make choices:

- If one category (label) is the consequence of another one, the annotator should select the first one. For instance, if a sentence describes both a control measure and its economic effects, the sentence should be labelled as "**Protective and control measures**".
- Both "**Concern and risk factors**" and "**Transmission pathway**" provide highly valuable information to assess the risk of emergence or spread of a disease. The annotator should therefore prioritise them against other labels into a multi-topic sentence.

Table 5 provides examples of frequently encountered multi-topic cases and the choice of the main label according to the two rules shown above.

Table 5 Resolution of multi-topic sentences in typical cases. DE: Descriptive epidemiology, PCM: Protective and control measures, EPC: Economic and political consequences, TP: Transmission pathway

Sentence topics	Example	Possible labels →main label	Rationale
Description of an event and its control measures.	*The Philippines confirms African swine fever, culls 7000 pigs.*	DE, PCM →DE	Control measures are consequences of the event.
Sanitary bans.	*Russia's agriculture authorities introduced temporary restrictions on pig and pork imports from Hungary due to an outbreak of the disease.*	PCM, EPC →PCM	Economic consequences of the ban.
Description of an event and its source.	*The strain detected in China is similar to the one that infected pigs in eastern Russia last year, but there is no conclusive evidence of the outbreak's source, it said.*	DE, TP →TP	Transmission pathway category prevails over the other types.

1.3 Results

In this section, we describe changes in the agreement metrics during the framework elaboration. As quantitative agreement measures, we calculated the inter-annotation agreement and Cohen's kappa coefficient. For inter-annotation agreement, we defined three different levels, i.e. total agreement (all annotators reached a consensus), partial agreement (two annotators agreed), and complete disagreement (all annotators disagreed). In case of multi-labels, we defined the agreement as strict, i.e. there is agreement between two annotators if they give precisely the same labels.

Cohen's kappa coefficient (κ) is a widely used statistical measure of inter-annotator agreement, which takes into account the extent of agreement expected by chance (Cohen, 1960). κ was calculated as follows:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{1}$$

Where $Pr(a)$ is the observed agreement among two annotators, $Pr(e)$ is the hypothetical probability of reaching an agreement.

Table 6 compares the agreement results obtained in step 1 (initial version of the guidelines) and step 3 (final version of the guidelines). We calculated the kappa by pairs of annotators separately

and then computed the average. At step 1, we obtained poor agreement for Event type annotations (κ= 0.30), while we obtained fair agreement for Information type (κ= = 0.53). Annotators totally agreed on Event type labels for only 29% of sentences, while almost 49% of the sentences obtained a total agreement for Information type.

Table 6 Agreement statistics at step 1 (initial guidelines, N=132 sentences) and step 3 (final guidelines, N=114 sentences). The inter-annotator agreement was computed in terms of relative agreements (total and partial), disagreements and Cohen's kappa (κ).

		Inter-annotator agreement			
		Total agreements	Partial agreements	Disagreements	κ
Step 1	Event type	29%	48%	23%	0.30
	Information type	49%	43%	8%	0.53
Step 3	Event type	87%	19%	4%	0.71
	Information type	75%	22%	3%	0.78

Statistics at step 3 (final guidelines) indicate a substantial improvement in the agreement for both classes. The Event type kappa was still lower than the Information type kappa (0.71 and 0.78, respectively).

1.4 Discussion

In this Section, we present critical issues that emerged during the framework elaboration process, while outlining our choices to improve the inter-annotator agreement. We first discuss two characteristics of our global framework and then explain two different strategies adopted to modify the annotation guidelines.

1.4.1 Global framework

Double-level annotation

Similar to event annotation approaches in which the annotator labels the event type and its attributes separately (Chanlekha et al., 2010), our final annotation framework relies on the attribution of two labels per sentence: Event type and Information type. We chose this approach because the thematic labels (Information type) encompass different temporal and event levels. Their relevance from an event-based surveillance viewpoint differ. For instance, a sentence describing an outbreak that occurred 2 years before the publication date ("**Old event**") is obviously less relevant than a sentence describing a current one. However, the type of information provided (description of an outbreak) remains the same. Therefore, the double-level approach is geared towards assigning consistent Information type labels among different event statuses. This choice increases the annotation time and complexity, but we believe that it substantially enhances the value of assigned labels by allowing us to consider spatiotemporal and topic labels separately.

Single-label annotation

We chose the sentence-based approach to address the lack of granularity in document-level approaches. However, a single sentence may also contain distinct topics. Therefore, until step 3, we allowed multi-labelling (the annotator could allocate as many labels as wished to a sentence, for both Event type and Information type). For Event type, only two sentences from the third dataset had multi-labels, both of them with "**Current event**" and "**Old event**". In both sentences, the reference to historical outbreaks was provided as context, e.g. "It has not been confirmed what caused the outbreak, but there have been other incidents in the region during the 20th century."

Multi-labelled sentences were more frequent for Information type, representing from 11% (12/114) to 25% (28/114) of the sentences according to the annotator. The most frequent associations were:

- "**Protective and control measures**" with either "**Descriptive epidemiology**" or "**Economic and political consequences**". In these sentences, there was a causal relationship between two labels. For instance, in the following sentence, a ban was decided in response to a related outbreak: "The Polish news agency reported that the ban was in relation to two cases of African swine fever found in dead wild boar on the Polish border with Belarus."

  These cases were resolved by providing rules to choose the main label in case of a causal relationship. We prioritised the causal label, claiming that it usually contains the main information. In the previous sentence, the outbreak occurrence prevails over the ban. Therefore, the sentence should be labelled as "**Descriptive epidemiology**".

- "**Descriptive epidemiology**" and "**Clinical presentation**", used in 12 sentences by annotator C.

1.4.2 Strategies to improve inter-annotator agreement

Creation of new classes

During this process, we created the "**Distribution**". In the first guidelines, sentences such as "In recent months, the disease has spread more rapidly" were labelled by annotators as either "**Descriptive epidemiology**" or "**General epidemiology**". Such sentences describe the current situation but they do not inform on a specific event. On the other hand, they describe an epidemiological situation that depends on a specific context (spatiotemporally locatable). Therefore, they cannot be considered as "**General epidemiology**".

Merging of classes

We merged the following categories in the annotation process:

(1) Current event and related event

Initially we had divided Event type labels into three groups for current and past events:

- Current event, i.e. the main event notified in the news article and which recently occurred,
- Related events, i.e. events that happened in the past but are related to the current one,
- Old events, i.e. events that occurred in the past without any link with the current situation (same definition as in the final guidelines)

This distinction between present and related events was the leading cause of disagreements in step 1. Deciding whether an event was a present or a related one was not trivial because it depended on a spatiotemporal cutoff which differed between annotators. Therefore, we decided to gather current and recent outbreaks in the same category ("**Current event**"). Some authors have proposed to use a temporally fixed window. For instance, events occurring within a 3 month window are related (Lejeune et al., 2012). This threshold was also used to label events as historical (occurred more than 3 months ago), in addition to recent events (occurred between 2 weeks and 3 months ago), and present ones (occurred within the last 2 weeks) as described by Chanlekha et al. (2010).

We believe that setting a rigid time window is not consistent with the epidemiological specificity of each disease. We instead decided to aggregate these two categories and distinguish only current/related events from old events. This distinction improved the agreement for the Event type level—all six sentences labelled as "**Old event**" obtained total agreement. In these sentences, typical semantic clues (e.g. the use of temporal expressions such as "back in days" or "in 2006") explicitly indicated the absence of epidemiological link.

(2) Clinical presentation

The "**Clinical presentation**" category was present in the first version of the guidelines. The label was mainly used by one annotator in association with the "**Descriptive epidemiology**" label (section 6.1). It appeared that in these sentences, all symptom-like terms were related to "deaths" or "died", e.g. "So far, six adult cattle and two calves have died from the disease". Rather than providing a clinical picture, these expressions were used to indicate the number of cases. We therefore decided to merge it with "**Descriptive epidemiology**" in the final frame.

(3) Protective and control measures

In the intermediate guidelines, we divided preventive and control measures into two distinct categories. This choice increased the number of disagreements in this class because several types of measures could be considered as both protective or control according to the context. For instance, the slaughtering of infected animals is a control measure for the concerned affected area but is a preventive measure from the unaffected area standpoint (limiting the disease spread). The ban of the animal movements, as well as vaccination, can also be control measures (avoid disease spread from the affected area) as well as a preventive measure (prevent disease introduction in an unaffected area). In the BioCaster ontology scheme, this context-dependency made the "control" category the most challenging class in terms of agreement (Kawazoe et al., 2006).

1.4.3 Limitations

Several limitations in the proposed annotation framework should be noted, as they may influence the performance of further classification tasks.

First, we adopted a single-label approach for each level. Not allowing multiple labels per sentence was questionable since several sentences belonged to several classes, and the annotator may have had difficulty in determining which category should take precedence. This may lead to misclassification errors and information loss during the supervised approach. However, the use of multi-labelling raises the issue of finding suitable agreement metrics while adding a major complication in finding proper classification methods (Koyejo et al., 2015). As some typical cases occurred, we tried to harmonise the annotators' choices by resolving multi-label cases in the guidelines.

Besides, we did not include polarity or sentiment analysis in our labelling scheme. For instance, sentences indicating the absence of outbreaks or a negative result for a test should be labelled as "**Descriptive epidemiology**". In practice, sentences claiming a negative event are quite rare in online news narratives. The current frame could be enhanced by adding a polarity label to each sentence as it is necessary to include negation detection to avoid false alarms.

In this Section, we proposed a sentence-based annotation scheme with the aim of going beyond conventional document-based classification and entity recognition. We built the framework by heavily relying on domain expert opinions, while intending to find a trade-off between fair inter-annotator agreement and class granularity. The final inter-annotator scores were 0.71 Kappa on average for Event type labels and 0.78 Kappa on average for Information type labels. While some classes of interest from an epidemiological viewpoint (e.g. **Concern and risk factors**, **transmission pathway**) are under-represented, we believe that the proposed framework helps increase the number of instances quickly and reproducibly. In the following section, we evaluate two approaches to automatically retrieve sentence-level epidemiological information.

2 Retrieval of fine-grained epidemiological information

In this section, we aim at automatically retrieving fine-grained information in online news articles, i.e. identifying event and information types at the sentence level as defined in the previous section. We compare two approaches, i.e. a supervised classification method and a pattern-based approach. Beyond the global performance of the models, we focus particularly on their ability to retrieve two categories, i.e. transmission pathway and concern and risk factors. Indeed, these two categories have two key features: (i) they are of critical interest for risk assessment by EI teams, (ii) contrary to event extraction, their retrieval has not yet been addressed in the literature, and (iii) they are under-represented in the corpus.

2.1 Corpus

To evaluate retrieval methods on sufficient class sizes, we increased the annotated corpus based on the final guidelines described in the previous section (32 news articles, 486 sentences) with 56 additional news articles. We obtained a final corpus containing 1,245 sentences (from 87 news articles). From this initial corpus (1,245 sentences), 161 sentences were labelled as irrelevant. The subset of sentences for Information type classification hence consisted of 1,084 sentences (Figure 10).

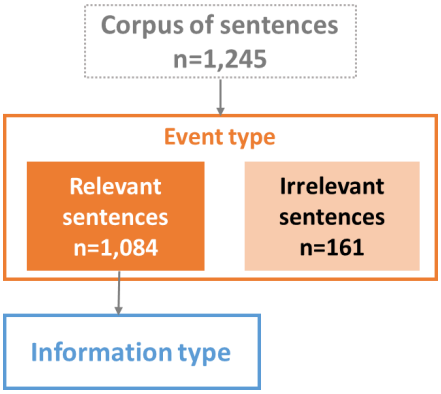


Figure 10 Distribution of labelled sentences in the corpus. Relevant sentences include current event, risk event, old event and general.

Figures 11 and 12 show the distribution of sentence labels at the Event type and Information type levels, respectively.

For the Event type-level, 64% of the sentences (799/1245) were labelled as **Current event**, 11% (136/1245) as **General**, 8% (105/1245) as **Risk event**, and 4% (44/1245) as **Old event**. **Irrelevant** sentences represented 13% of the corpus (161/1245). The Information type-level contained 1,084 annotated sentences. Among these sentences, 37% of the sentences (401/1084) were labelled as **Descriptive epidemiology**, 29% (310/1084) as **Preventive and control measures**, 10% (110/1084) as **Concern and risk factors**, 10% (109/1084) as **General epidemiology**, 6% (69/1084) as **Transmission pathway**, 5% (58/1084) as **Economic and political consequences**, and 2% (27/1084) as **Distribution**.

The distribution of sentences at the Event type level was highly imbalanced, indicating that disease-related news articles primarily provide information about the current situation (**Current event**).

The Information type level was more balanced, with two classes (**Descriptive epidemiology** and **Protective and control measures**) representing 67% of the sentences (711/1084).

The classes of interest were under-represented in the corpus (**Transmission pathway**, n= 69; **Concern and risk factors**, n=110). Thus, we evaluate the retrieval methods regarding their ability to detect minority classes that are usually "drowned" in the news content.

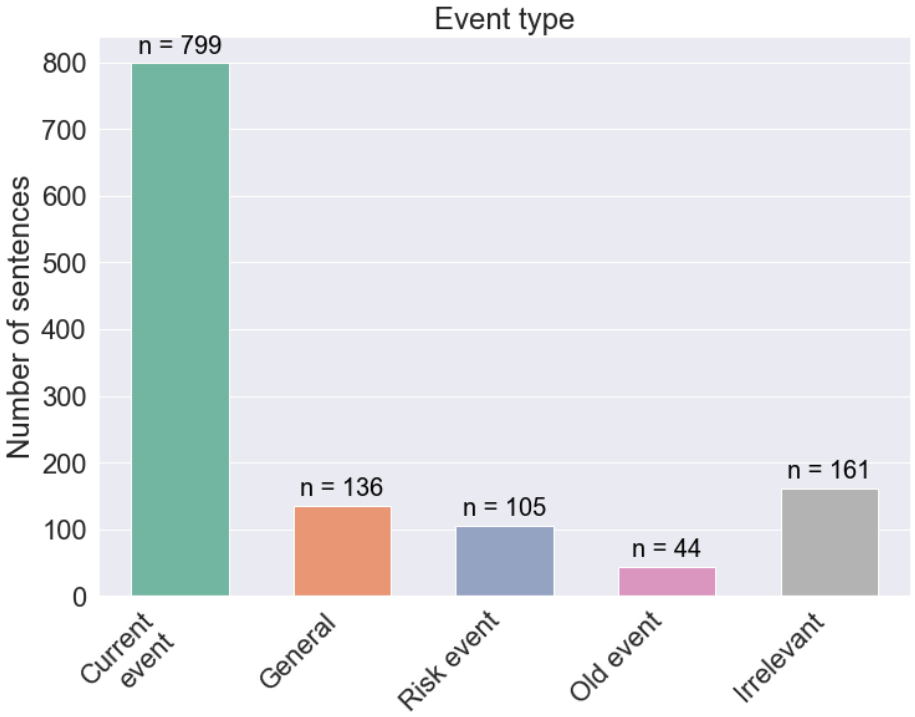


Figure 11 **Number of sentences per Event type level.**

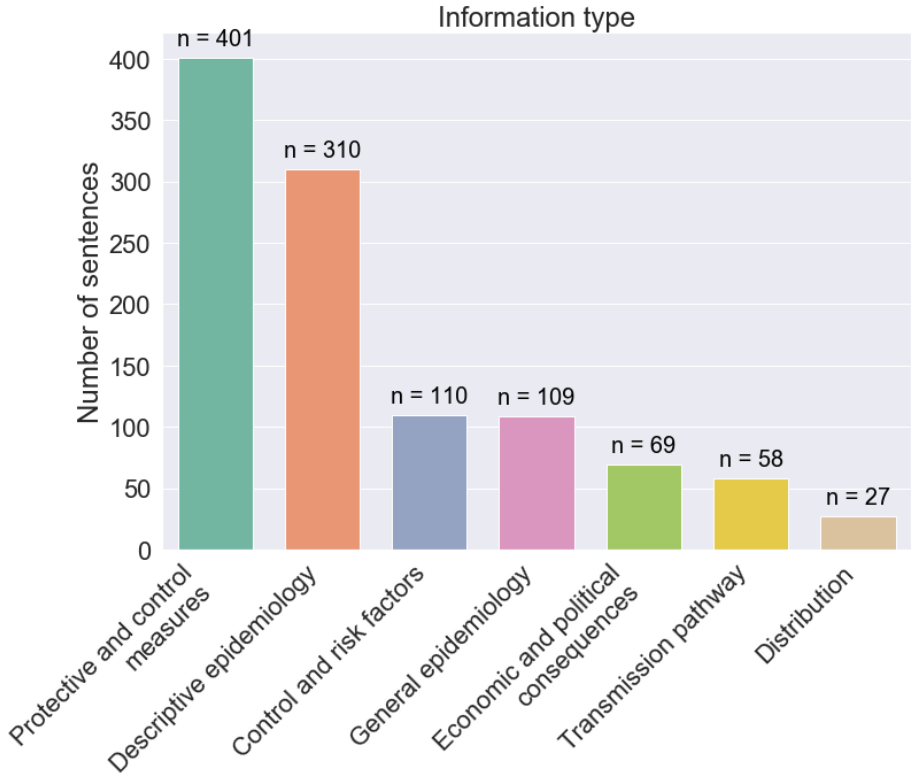


Figure 12 **Number of sentences per Information type level.**

2.2 Supervised classification

Textual classification involves using statistical learning models to classify text (e.g. a whole document, a sentence, etc.) into specific sets of categories. The classification is called supervised when models are trained on instances whose labels are known (i.e. annotated by domain experts) (Witten et al., 2016). When labels are unknown, clusters of texts are generated automatically, and the task is referred to as unsupervised classification.

In this study, we adopt the supervised classification paradigm, using the previously described corpus of sentences as training dataset. Our objective is to fit a global classification model able to correctly identify both the Event type and Information type of an unlabelled sentence, as illustrated in Figure 13.

Models (i.e. classifiers) are fitted on annotated instances during the training step, which includes two steps (Witten et al., 2016):

- Textual vectorisation (Sections 2.2.1 and 2.2.2), which converts textual data into a machine-readable format;
- Training of different models (Section 2.2.3), whose performances are calculated through a 5-fold cross-validation process (Section 2.2.3).

We evaluated two types of textual representations in this framework, i.e. the traditional bag-of-words and word embedding representation.

2.2.1 Bag-of-words representation

In this section, we outline the foundations of the model used to represent the sentences, i.e. the vector-space model.

The vector-space model

The vector-space model is an algebric model for converting textual data into a machine-readable format, which was introduced by Salton in 1971 for information retrieval tasks (Salton, 1971). This model is based on the assumption that a document from a corpus can be represented as a numeric vector, derived from the terms it contains. The aim is to achieve a consistent representation of the meaning of a document, i.e. its semantics. *In fine*, the closeness of two document vectors in the vector space model should reflect their semantic similarity.

The transformation of a corpus of documents in the vector-space model involves two steps:

1. Representation of each document in a vector of selected features,
2. Computation of the vector's numerical values (feature weighting),

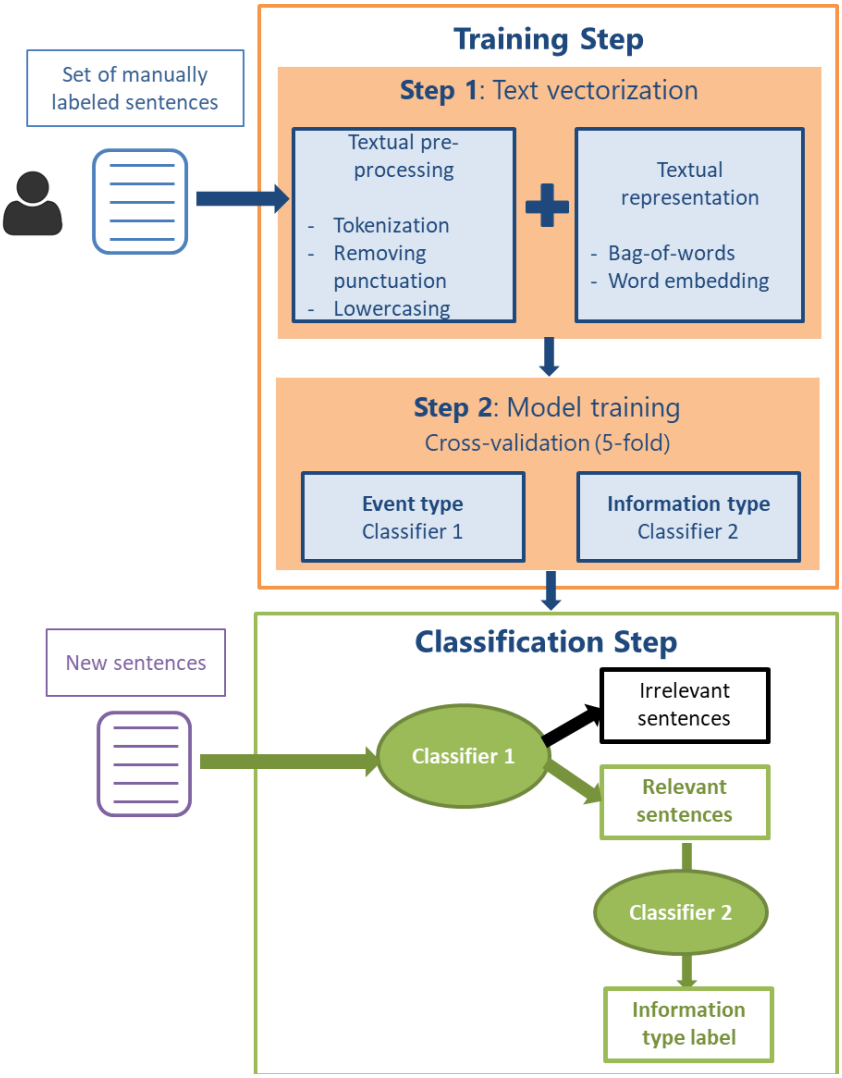


Figure 13 Supervised learning framework for sentence classification.

Feature selection.

The first step aims at determining the set of features (i.e. "vocabulary") that provides the most relevant representation of a document (or a sentence), i.e. which best reflects its content. Bag-of-words (BOW) is one of the most popular models used to convert textual documents into vectors. In this model, the vocabulary corresponds to all of the terms present in the whole corpus (Zhang et al., 2010). Each document d is encoded in an n-dimensional vector where each component w_{td} represents the absence or presence of a feature (term) t in the document (where n is the length of the vocabulary). If the feature t occurs in the document, the feature weight w_{td} has a non-zero value. BOW is an easy to understand but effective model to convert a document into fixed-length vectors.

However, it has several limitations (Brownlee, 2017; Zhao and Mao, 2018):

1. **Vocabulary**: the size of the vocabulary can result in high dimensional feature vectors,
2. **Sparsity**: the BOW model leads to highly sparse vectors (i.e. most of the vector elements have zero value, since a document only contains a very small portion of all of the vocabulary). This may result in computational complexity while drowning out information,
3. **Meaning**: the BOW model overlooks the grammar and word order in a document, as reflected by the "bag" concept. The context of the terms is discarded even though it provides meaningful information regarding the semantics of terms, such as synonymy. For instance, the BOW model may not effectively capture the closeness of semantically similar documents with different term usages, as they are converted to very different vectors.

Several approaches can be used to enhance BOW models, such as bag-of-n-grams (contiguous sequence of n terms) (Aizawa, 2001), using noun phrases as terms (Lewis, 1992), or bag-of-concepts (Grootendorst and Vanschoren, 2020). Lexical (based on terminology) and ontology-based methods (relying on knowledge representation in the studied domain) have also been proposed to semantically enrich textual representation (Ranwez et al., 2013; Sbattella and Tedesco, 2013).

Feature weighting

The second step corresponds to the numerical transformation of a document, whereby a weight is assigned to each feature in the document. The most basic representation is to compute the presence of a feature as a boolean value, i.e. 0 for absent and 1 for present. In 1957, Luhn proposed the first term weighting scheme, based on the assumption that the weight of a term in a document is simply proportional to the number of times the term occurs in the document (Luhn, 1957). This weighting scheme is referred to as term frequency (Equation 3). The boolean weight (B_{ij}) and term frequency (TF_{ij}) of a term i in a document d are given by Equations 2 and 3:

$$B_{ij} = \begin{cases} 1 & if \quad TF_{ij} > 0 \\ 0 & if \quad TF_{ij} = 0 \end{cases} \tag{2}$$

$$TF_{id} \quad = \quad \frac{n_{id}}{n_d} \tag{3}$$

where n_{id} is the number of times the term i appears in document d,
n_d is the total number of terms in d,

A major shortcoming of term frequency weighting methods is their sensitivity to highly frequent terms, while rarer and more specific terms have very low weights in the feature space (Manning et al., 2009). In 1972, Jones scaled down the importance of highly frequent terms by introducing a new term-weighting method called *inverse document frequency* (*IDF*) (Jones, 1972).

In this approach, the specificity of a term is quantified as an inverse function of the number of documents in which it occurs (Equation 4).

$$IDF_i = log(\frac{N}{DF_i}) \tag{4}$$

where N is the total number of documents in the corpus,
and DF_i is the frequency of the term i in the whole corpus (the number of documents which contain i).

Term Frequency-Inverse Document Frequency ($TF - IDF$) is the product of term frequency and inverse document frequency (Equation 5). Terms with the highest $TF - IDF$ values are distinctively frequent in a document in comparison to the collection of documents (Salton and Buckley, 1988).

$$TF - IDF_{id} = TF_{id} \times IDF_i \tag{5}$$

We transformed all the sentences into the bag-of-words model, using the $TF - IDF$ weight. Several normalizing (or "pre-processing") steps which can be applied prior to the creation of the vocabulary to account for the noisy aspect of textual data are further presented, in Chapter 4 Section 1.1. For sentence classification, we solely removed the punctuation and normalized the words to lowercase.

2.2.2 Word embedding representation

Word embedding methods produce word representations corresponding to dense real-valued vectors in a predefined vector space, typically leveraging the distributional principle which states that words that occur in similar contents should also be close in the vector space. Vector values are learned according to the context in which the word appears, based on the assumption that words that frequently appear in the same context (i.e. surrounded by the same words) tend to have the same meaning (Goldberg, 2017). In most models, the context corresponds to the window of neighbouring words—which is a configurable parameter.

For instance, in health-related news, the verbs “declared” and “reported” are typically used in the same types of sentences (e.g. “France declared/reported an outbreak of foot-and-mouth disease”). While the traditional bag-of-words model will encode the verbs “declared” and “reported” as two distinct features, a word embedding model can capture their semantic closeness.

Word embedding models have been shown to perform with higher accuracy than bag-of-words representations. They have been applied to a variety of linguistic tasks in the disease surveillance domain, such as veterinary necropsy report classification (Bollig et al., 2020), disease taxonomy development (Ghosh et al., 2016), or epidemiological feature extraction from WHO reports (Ghosh et al., 2017). An interesting feature of word embedding is the ability to compute the similarity between two words by computing the cosine similarity between their two vectors. As a counterpart,

word embedding models must be trained on large datasets.

Model training

Word2Vec, developed by Tomas Mikolov in 2013, is one of the most popular techniques to learn word embedding (Mikolov et al., 2013b). The Word2Vec algorithm is based on an artificial neural network (ANN). ANNs are modelled on neurons in a biological brain. They consist of at least two layers of nodes, or so-called artificial neurons, i.e. an input layer and one or more hidden layers. During the learning phase, each node computes a linear combination of input data, adding a value that is referred to as bias. A non-linear function is applied to the output data and transmitted to the following node. The network automatically adjusts node parameters to minimise the average error between the expected and observed outputs (e.g. a document label).
Word2Vec uses the shortest possible ANN structure, i.e. a 2-layer neural network. Two models are implemented:

- Continuous Bag-Of-Words (CBOW) that predicts a target word based on its context (surrounding context words);
- Continuous Skip-gram model that predicts a target context using a word.

Figure 14 illustrates the CBOW architecture. The input layer corresponds to the vector representation of the context of the word "chickens", using a 2-word window. In this example, the stop-word "by" is discarded before the model training. The hidden layer analyses the vector representation to predict the word "chickens".

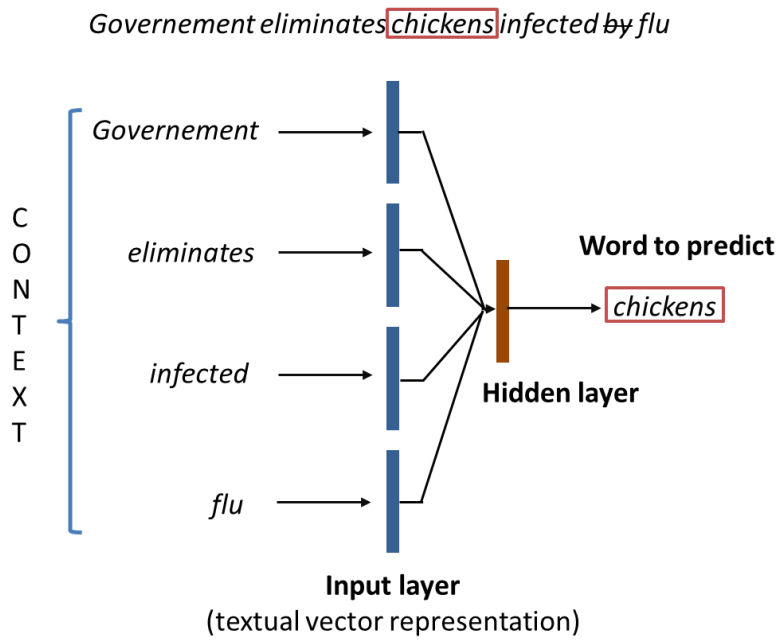


Figure 14 CBOW architecture, inspired by (Mikolov et al., 2013a).

According to Mikolov et al. (2013a), each of these models has its own advantages depending on the training data. The CBOW model is faster to train compared to Skip-gram, but the latter works better with small corpora and rare word representation. Thus in this thesis, we investigate the performance of each model. Several pre-trained word embeddings are publicly available, but training a word embedding model on text specific to the target domain has been shown to improve performances (Pyysalo et al., 2015). Moreover, the pre-processing techniques applied to the training dataset pre-trained word embeddings are not always straightforward. We thus decided to train a word embedding model on a dataset of news articles extracted from the PADI-web database using the gensim library (Řehůřek and Sojka, 2010). We extracted the whole PADI-web database (all news in order to have a large dataset) on December 14, 2018, obtaining a training set of 35,577 news articles. The training set length was 33,417,501 words. We compared this model with the Word2Vec pre-trained Google News model[1], trained on a 3 billion words corpus with the CBOW algorithm.

Two types of parameters can influence the quality of word embedding vectors: (i) the pre-processing steps of the training dataset, and (ii) the model parameters. Here we evaluated the influence of lowercasing, lemmatisation, and stop-word removal.

Textual pre-processing was done using the NLTK library (Bird and Loper, 2004). We opted to focus on the CBOW model. We evaluated two dimensions for the trained vectors: 100, which is the default length implemented in gensim, and 300, which is the most commonly used dimensionality in various studies (Mikolov et al., 2017, 2013b; Pennington et al., 2014). We used the default parameter for the window size (5 words)— while setting the minimum word frequency at 10. We evaluated the impact of the different parameters based on the Information type classification as we wanted to optimise the classification of this level.

From word to sentence embeddings

The representation of sentence-vectors through different algebraic combinations of a sentence's word vectors have been widely explored (Dilawar et al., 2018; Mitchell and Lapata, 2010). One of the most popular and simple methods is to compute the unweighted average of embeddings of a sentence's words (Wieting et al., 2016). Similar to the boolean version of bag-of-words model, unweighted approaches tend to bias the importance of frequently occurring but not informative words. Thus, as proposed by (De Boom et al., 2016), we leveraged words based on their $TF-IDF$ values in the corpus (these $TF-IDF$s are identical to those used for bag-of-words representations presented in the previous section). Finally, each sentence is represented by a vector that pools the information of all of the words.

While the weighted average is simple to implement, it has the same shortcoming as the bag-of-words representation regarding the loss of the word order. More sophisticated approaches propose to jointly learn embeddings for both words and paragraphs using models similar to Word2Vec

[1]https://github.com/mmihaltz/word2vec-GoogleNews-vectors

(Peters et al., 2018; Le and Mikolov, 2014), or to compute the distance between two documents (or sentence) based on a new metric called Word Mover's Distance (WMD) (Kusner et al., 2015).

2.2.3 Classifiers

We compared three classifiers that are widely used for text classification, i.e. Naive Bayes, Support Vector Machines and Multilayer Perceptron. As detailed in Chapter 1 Section 4.1.4, Naive Bayes and Support Vector Machines are currently used for document classification in several EBS systems.

1. Naive Bayes (NB), is a family of probabilistic classifiers based on the Bayes' theorem. These classifiers are based on the assumption that there is high independence between features. We used a multinomial Naive Bayes classifier, which assumes that features have a multinomial distribution that is suitable for fractional counts such as $TF-IDF$ (Kibriya et al., 2005). Naive Bayes classifiers have several limitations, i.e. they are sensitive to highly correlated features, and they cannot handle negative values generated by word embedding vectorisation (Langley and Sage, 1994). Thus, while it is possible to scale all vectors uniformly to avoid negatives values, we did not include NB in the word embedding model evaluation.

2. Support Vector Machines (SVM) is a non-probabilistic and linear classification technique. SVM has been widely used for text classification, including small sized texts such as sentences (Khoo et al., 2006; Zhang and Liu, 2007) and tweets (Go et al., 2009). It achieves robust performance regarding important textual data vector properties, which are sparse and dense (containing few relevant features) (Joachims, 1998). We used a linear kernel parameter (linear SVM) classifier, as linear kernels perform well with textual data (Uysal and Gunal, 2014; Kumar and Gopal, 2010).

3. Multilayer Perceptron (MLP) is an Artificial neural network-type classifier. ANN classifiers were shown to perform well when combined with word embedding representations (Agibetov et al., 2018; Mandelbaum and Shalev, 2016).

The whole classification and evaluation pipeline was performed using the scikit-learn library (Python) (Pedregosa et al., 2011). We used the default parameters implemented in scikit-learn, except for class weighting. Indeed, class sizes are highly imbalanced in both Event type and Information type classifications. The overall accuracy may be artificially maximised by assigning all examples to the majority class. To avoid this bias, both SVM and NB can take class weights during the classifier fitting into account (instances are assigned a weight inversely proportional to the frequency of their class). MLP does not allow class weighting and was therefore used with all of its default parameters.

2.2.4 Evaluation

We estimated the performances of the trained models via the widely used of cross-validation method. During cross-validation, the dataset is split into K equal-sized subsets, referred to as "folds". The classifier is trained on K-1 folds, and its performances are evaluated when predicting the remaining fold, as illustrated in Figure 15.

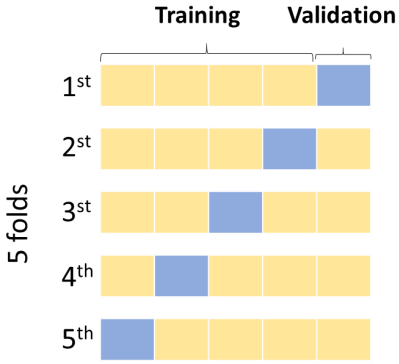


Figure 15 Illustration of 5-fold cross-validation.

We used a fold number of 5, as this value was empirically shown to yield test error rate estimates with low variance, while not being hampered by excessively high bias (Hastie et al., 2009).

At each fold, we computed the traditional metrics used in supervised classification, i.e. precision, recall, accuracy and F-measure. At the class A level, precision corresponds to the proportion of correct sentences classified in class A (equation 6), and recall corresponds to the proportion of sentences belonging to class A that are correctly identified (equation 7):

$$Precision(A) = \frac{number\ of\ sentences\ correctly\ attributed\ to\ class\ A}{number\ of\ sentences\ attributed\ to\ class\ A} \tag{6}$$

$$Recall(A) = \frac{number\ of\ sentences\ correctly\ attributed\ to\ class\ A}{total\ number\ of\ sentences\ belonging\ to\ class\ A} \tag{7}$$

F-measure is the harmonic mean of precision and recall (equation 8).

$$F - measure(A) = \frac{2 \times Precision(A) * Recall(A)}{Precision(A) + Recall(A)} \tag{8}$$

To calculate the performances over all classes to account for class imbalance, we computed the weighted precision, recall, and F-measure. The metrics (i.e. recall, precision, F-measure) are calculated for each label, and the average is weighted by the support (the number of true instances per label). The weighted method can result in an F-measure that is not between precision and recall.

For instance, considering a binary classification between a class A (frequency= N_a) and a class B (frequency = N_b), the weighted precision $Precision_w$ is:

$$Precision_w = \frac{N_a}{N_a + N_b} \times Precision(A) + \frac{N_b}{N_a + N_b} \times Precision(B) \quad (9)$$

The overall accuracy corresponds to the proportion of correctly classified sentences among all classes.

2.2.5 Results and discussion

We conducted two experiments. The first aimed at identifying the best pre-processing steps and parameters for training of the word embedding model. The second compared the results of the classifiers and representations detailed above.

Word embedding parameters and pre-processing

We evaluated the word embedding models for the classification of the Information type level in terms of accuracy, with both MLP and SVM classifiers (Table 7). Standard deviations correspond to the variation in the accuracy among the 5-fold cross-validations. Lowercasing and lemmatisation did not increase the accuracy. However, with the CBOW algorithm, removing stop-words decreased the classification performances for both SVM and MLP classifiers.

Table 7 Impact of pre-processing steps and the vector dimension for classification of the Information type level in terms of accuracy based on SVM and MLP classifiers.

Textual pre-processing	Vocabulary length	Classifier	Model	Vector dimension 100	Vector dimension 300
None	82433	SVM	CBOW	0.65 (+/-0.03)	0.66 (+/-0.02)
			Skip-gram	0.66 (+/-0.03)	0.65 (+/-0.03)
		MLP	CBOW	0.72 (+/-0.03)	0.73 (+/-0.03)
			Skip-gram	0.67 (+/-0.03)	0.67 (+/-0.03)
E1: lowercase	67524	SVM	CBOW	0.64 (+/-0.02)	0.65 (+/-0.02)
			Skip-gram	0.65 (+/-0.02)	0.66 (+/-0.03)
		MLP	CBOW	0.72 (+/-0.02)	**0.74 (+/-0.03)**
			Skip-gram	0.68 (+/-0.03)	0.67 (+/-0.03)
E1 + lemmatisation	57928	SVM	CBOW	0.62 (+/-0.03)	0.66 (+/-0.01)
			Skip-gram	0.69 (+/-0.02)	0.69 (+/-0.03)
		MLP	CBOW	0.71 (+/-0.03)	0.73 (+/-0.02)
			Skip-gram	0.71 (+/-0.02)	0.71 (+/-0.01)
E1 + stop-words removal	57785	SVM	CBOW	0.59 (+/-0.03)	0.57 (+/-0.05)
			Skip-gram	0.65 (+/-0.03)	0.65 (+/-0.03)
		MLP	CBOW	0.64 (+/-0.02)	0.66 (+/-0.02)
			Skip-gram	0.68 (+/-0.02)	0.69 (+/-0.03)

The Word2Vec pre-trained model obtained accuracies of 0.66 +/- 0.02 (SVM classifier) and 0.69 +/- 0.02 (MLP classifier).

In several studies, removing stop-words improved the output quality of word embedding models (Lison and Kutuzov, 2017). In our experiments, we used the stop-word list implemented in NLTK. This list includes, among others, conjugated auxiliary verbs (e.g. "am", "is") as well as adverbs such as "because" or "during". Indeed, some of these words are important for the sentence meaning. For instance, Transmission pathway sentences rely heavily on causality semantics expressed by specific adverbs (e.g. "The infection occurred **during** transportation"). We thus hypothesise that removing stop-words based on the NLTK list deleted important cues. Further work could focus on editing a specific stop-word list.

The Skip-gram algorithm yielded lower performances than CBOW but was less sensitive to stop-word removal. Skip-gram is known to be more efficient with infrequent words and small training datasets (Naili et al., 2017). Hence, we expected this model to perform better than CBOW.

In our results, the 300-dimension vectors achieved higher accuracy than the 100-dimension vectors, even though the difference was slight. The choice of embedding dimension is still an open issue in the literature— it relies on a trade-off between small dimensionality, that may not capture all possible word relations, and large dimensionality which suffers from overfitting (Yin and Shen, 2018). In the following analysis, we used the 300-dimension vectors trained on the lowercased text and the CBOW algorithm. However, our results suggest that 100-length vectors trained without any pre-processing could also be used without impacting the overall accuracy of the classification. This simpler model could be preferred in terms of reduced computing time and complexity (for both textual pre-processing and model training steps), which are constraints that can hinder model applicability of large embedding vectors (Wu et al., 2016).

The quality of word embedding models is sensitive to other parameters that were not evaluated in this study, such as the size of the training dataset and the window size. The size of the sliding window has a marked effect on the vector similarities—small windows tend to produce functional and syntactic similarities, while larger windows tend to produce more topical similarities (Goldberg, 2017). Moreover, larger training datasets tend to increase the quality of embedding models. Besides, we ignore the multi-word terms in our model. Several techniques could be used to take them into account in word embedding representation, such as averaging the component word vectors or directly creating multi-word term vectors after merging them (Henry et al., 2018).

Considering the classification performances obtained with minimal tuning of the word embedding model, we believe that further evaluation of the training parameters could enhance its quality. Moreover, other word embedding algorithms such as FastTex or Glove (Pennington et al., 2014) could be compared to the Word2Vec model performances. Besides, new word embedding architectures have been proposed, such as the model BERT (Bidirectional Encoder Representations from Transformers), released in late 2018. This model achieved new state-of-the-art results on several NLP tasks, including sentence classification (Devlin et al., 2019). Compared to the classic word embedding models, where each word has a fixed representation regardless of the context

within which it appears, BERT produces word representations that are dynamically informed by the words around them. Such an approach could be evaluated in the context of animal disease surveillance.

Classifiers and representations

In this second experiment, we compared the different classifiers for both Event type and Information type classification, depending on the textual representation. The performances are summarised in Table 8.

Table 8 Performances of classifiers trained on bag-of-words (BOW) and word embedding (Emb) representations, in terms of weighted precision, recall, F-measure and accuracy over 5-fold cross-validation. The best performances are shown in bold for each level.

Level	Textual representation	Classifier	Precision	Recall	F-measure	Accuracy
Event type	**BOW**	SVM	0.71 (+/-0.02)	0.71 (+/-0.01)	0.70 (+/-0.01)	0.71 (+/-0.01)
		NB	0.50 (+/-0.02)	0.72 (+/-0.02)	0.55 (+/-0.03)	0.50 (+/-0.02)
		MLP	0.72 (+/-0.02)	0.70 (+/-0.02)	0.69 (+/-0.02)	0.72 (+/-0.02)
	Emb	SVM	0.63 (+/-0.02)	0.67 (+/-0.03)	0.65 (+/-0.04)	0.64 (+/-0.02)
		MLP	**0.76 (+/-0.02)**	**0.75 (+/-0.04)**	**0.75 (+/-0.03)**	**0.76 (+/-0.03)**
Information type	**BOW**	SVM	0.67 (+/-0.04)	0.66 (+/-0.03)	0.66 (+/-0.03)	0.66 (+/-0.03)
		NB	0.55 (+/-0.06)	0.67 (+/-0.04)	0.58 (+/-0.05)	0.55 (+/-0.06)
		MLP	0.66 (+/-0.03)	0.66 (+/-0.04)	0.64 (+/-0.03)	0.65 (+/-0.03)
	Emb	SVM	0.65 (+/-0.03)	0.68 (+/-0.03)	0.66 (+/-0.03)	0.65 (+/-0.03)
		MLP	**0.73 (+/-0.03)**	**0.72 (+/-0.04)**	**0.72 (+/-0.03)**	**0.72 (+/-0.03)**

At both levels, the MLP classifier combined with the word embedding representation outperformed the other approaches in terms of recall, precision, F-measure and accuracy. MLP and SVM achieved comparatively equal performances when trained on the BOW representation and clearly

outperformed the NB classifier. These behaviours were identical for Event type and Information type classification. Classification performances were lower on average for the Information type level than for the Event type level. Indeed, the number of classes was higher at the Information type level (7 versus 5). Besides, the Information type level contained three classes with very small sizes (**DI**, n=27; **EPC**, n= 58 and **TP**, n=69).

Similar to the global performances, the word embedding representation outperformed the BOW representation. For both types of classification, MLP obtained higher precision than SVM, which contributed to achieving higher F-measures. To evaluate the influence of word-embedding representation in intra-class classification, we compared the performances obtained by MLP for both Event type and Information type classification (Tables 9 and 10). Within each level, classification performances with the BOW representation were highly heterogeneous between classes. For Event type classification, F-measures ranged from 0.14 (**Old event**) to 0.81 (**Current event**). For the Information type level, F-measures ranged from 0.24 (**Distribution**) to 0.76 (**General epidemiology**). Based on the word embedding representation, F-measures ranged from 0.32 (**Old event**) to 0.84 (**Current event**) for Event type classification. For Information type classification, F-measures ranged from 0.29 (**Distribution**) to 0.82 (**Descriptive epidemiology**, **General epidemiology**).

With the word embedding representation, four classes obtained an F-measure higher than 0.8, i.e. **Current event** and **General** (Event type level); **Descriptive** and **General epidemiology** (Information type level). Preventive and control measures obtained an F-measure of 0.78. Lowest recall and precision tended to be correlated with the class having the lowest number of instances, such as Old event (F-measure=0.32) and Distribution (F-measure=0.29).

Table 9 Performances of MLP for Event type classification. The best performances are shown in bold for each level.

Label (n)	Bag-of-words			Word embedding		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Current event (n=799)	0.74	0.98	0.81	0.80	0.88	**0.84**
Risk event (n=105)	0.39	0.29	0.33	0.42	0.36	**0.39**
Old event (n=44)	0.33	0.09	0.14	0.40	0.27	**0.32**
General (n=136)	0.79	0.58	0.67	0.84	0.79	**0.82**
Irrelevant (n=161)	0.69	0.41	0.52	0.72	0.57	**0.64**
Weighted average	0.72 (+/-0.02)	0.70 (+/-0.02)	0.69 (+/-0.02)	0.76 (+/-0.02)	0.75 (+/-0.03)	**0.75 (+/-0.03)**

Classification based on word embedding outperformed the BOW classification in terms of weighted precision, recall and F-measure. This improvement was observed in all classes, including under-represented ones. This is in line with the fact that word embedding methods are able

Table 10 Performances of MLP for Information type classification. The best performances are shown in bold for each level.

Label (n)	Bag-of-words			Word embedding		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Descriptive epidemiology (n=401)	0.70	0.78	0.73	0.80	0.84	**0.82**
Distribution (n=27)	0.67	0.15	0.24	0.32	0.26	**0.29**
Preventive and control measures (n=310)	0.57	0.75	0.65	0.75	0.81	**0.78**
Concern and risk factors (n=110)	0.53	0.35	0.42	0.64	0.55	**0.59**
Transmission pathway (n=69)	0.56	0.28	0.37	0.57	0.39	**0.47**
Economic and political consequences (n=58)	0.68	0.26	0.38	0.62	0.59	**0.60**
General epidemiology (n=109)	0.83	0.70	0.76	0.83	0.82	**0.82**
Weighted average	0.66 (+/-0.03)	0.66 (+/-0.04)	0.66 (+/-0.03)	0.74 (+/-0.01)	0.75 (+/-0.03)	**0.73 (+/-0.02)**

to address the sparsity of short texts (Dai et al., 2017; Mandelbaum and Shalev, 2016). Embedding models are trained on external datasets contrary to traditional bag-of-words representations. They thus allow the classifiers to generalise more effectively beyond their limited number of training examples (Thapen et al., 2016a). This feature enables them to override the constraints inherent to low-sized training datasets.

The MLP classifier was not fitted to take class imbalance into account, which may explain the higher recall rate obtained for over-represented classes when using the BOW representation, especially at the Event level. Our results suggest that this bias was reduced with the word embedding representation.

The Information type classification results, based on both BOW and word embedding representations, showed that the supervised approach performed well in distinguishing sentences describing the ongoing event (**Descriptive epidemiology**) from sentences providing general information (**General epidemiology**). During the annotation phase, many instances of General epidemiology involved the same sentence structure, e.g. "The virus causes a hemorrhagic fever with high mortality rates in pigs", thus partly explaining why it can be easily detected by supervised learning approaches. Being able to differentiate Descriptive from General epidemiology is critical for the detection of relevant symptoms from the news. Indeed, we found that most

Table 11 Confusion matrix for classification of the Information type by MLP classifiers and word embedding representation.

		Predicted label						
		DE	**DI**	**PCM**	**CRF**	**TP**	**EPC**	**GE**
True label	**DE**	333	6	38	10	6	2	6
	DI	11	10	1	4	0	0	1
	PCM	28	1	252	10	6	7	6
	CRF	10	0	19	65	6	9	1
	TP	17	0	14	8	27	0	3
	EPC	3	0	7	10	0	37	1
	GE	9	1	1	7	5	0	86

symptom occurrences in the corpus were not related to an event, i.e. 16 out of 21 sentences belonged to the General epidemiology class. Noisy symptom detection may therefore occur when blindly extracting symptom keywords from the text without taking the keyword context into account, i.e. the sentence. Culotta (2013) showed that Twitter messages containing specific keywords (i.e. variants of avian influenza or symptoms) correlated with influenza-like illness official reports. However, in that study, the authors highlighted the issue of "spurious matches", which are text fragments containing one of the targeted keywords but which are not related to a flu-event (e.g. tweets mentioning the recall of a flu vaccine or an official policy announcement). A semi-supervised classification model was used to automatically detect such matches and reduce the number of false alarms (Culotta, 2013; Edo-Osagie et al., 2019). Our results showed that a similar approach could be implemented in the EBS pipeline to distinguish spurious from real symptom occurrence.

We hypothesize that this behaviour could partly be explained by the class imbalance, with under-represented class instances being classified in majority classes (i.e. **DE** and **PCM**). This hypothesis is confirmed by the confusion matrix (Table 11). This is a major limitation in practice when the aim is to retrieve under-represented classes. However, our results suggested that training on word embedding models could overcome this limitation by substantially improving both MLP recall and precision.

As highlighted in Table 8, the SVM performance for the Event type dropped when using embeddings. These results suggested that using word embedding with traditional classifiers (here, SVM) had no additional value for our task, as confirmed in previous studies (d'Amato et al., 2017).

In this first experiment, we showed that classic machine learning approaches were able, with promising results, to detect epidemiological information at the sentence level. However, a drawback of learning-based methods is that their decision rules are not directly interpretable by humans, and the use of neural networks further amplifies this aspect. Human input solely occurs during the annotation step, and experts cannot tune the final model. Besides, learning algorithms are sensitive to class imbalances, which may limit their performances regarding the retrieval of rare categories.

For specific fine-grained Information type classes, such as Transmission pathway or Concern and risk factors, symbolic approaches could prove to be more relevant. In the following section, we propose and evaluate a pattern-based approach relying on both automated vocabulary expansion and expert input.

2.3 Lexicosyntactic pattern-based approach

In this section, we aim at identifying sentences from specific classes based on the patterns they contain. We opted for the pattern definition of Du and Yangarber (2015), i.e. a pattern consists in "a place-holder for specific tokens (terms) and their surrounding context". The surrounding context may be fixed, and the token may be the variable. For instance, "**X** was detected in **Y**", where X is a disease and Y a location, is a sample pattern for detecting sentences from the **Descriptive epidemiology** category.

Among current pattern extraction approaches, unsupervised methods have gained some popularity due to the marked reduction in the amount of manual curation they require (Ghosh et al., 2017; Ibekwe-Sanjuan et al., 2011; Yangarber, 2003). An intuitive approach is to rely on experts to provide a list of patterns. Such methods have two major shortcoming regarding the pattern generalisation: (i) the vocabulary is limited to the expert knowledge, and (ii) the syntactic structure may be rigid. Thus, even if expert-based patterns may achieve high precision, the problem of recall, or coverage, is critical.

Bootstrapping methods have been proposed to automatically discover patterns from an initial set of patterns (hereafter referred to as seed patterns) (Jones et al., 1999; Thelen and Riloff, 2002). Ibekwe-Sanjuan et al. (2011) used the local context of seed patterns, i.e. their surrounding words, to generate variants. This method relied on the assumption that patterns occurring in the same context tend to have the same semantic meaning, i.e. the paradigm of word embedding models. In line with (Ghosh et al., 2017; Ibekwe-Sanjuan et al., 2011), we propose an incremental approach integrating both word embedding and expert knowledge to expand patterns, in the sentence retrieval context. We propose two types of pattern expansion: (i) lexical and (ii) syntactical.

2.3.1 Method

We evaluated a semi-automated and incremental process. In this approach, an expert is at the core of the process (Figure 16, steps 1 and 3). Indeed, we believe that expert knowledge is particularly suitable for the retrieval of fine-grained classes. Our objective is to use a minimal set of sentences (referred to as seed sentences) to identify patterns specific to the class (seed patterns) and expand the patterns at lexical and syntactical levels. All steps are detailed in the following subsections.

The pattern extracted and expanded after steps 1, 2, 3 and 4 are hereafter referred to as P_S (seed patterns), P_{E1} (first expansion), P_{E2} (second expansion), and P_{E3} (third expansion).

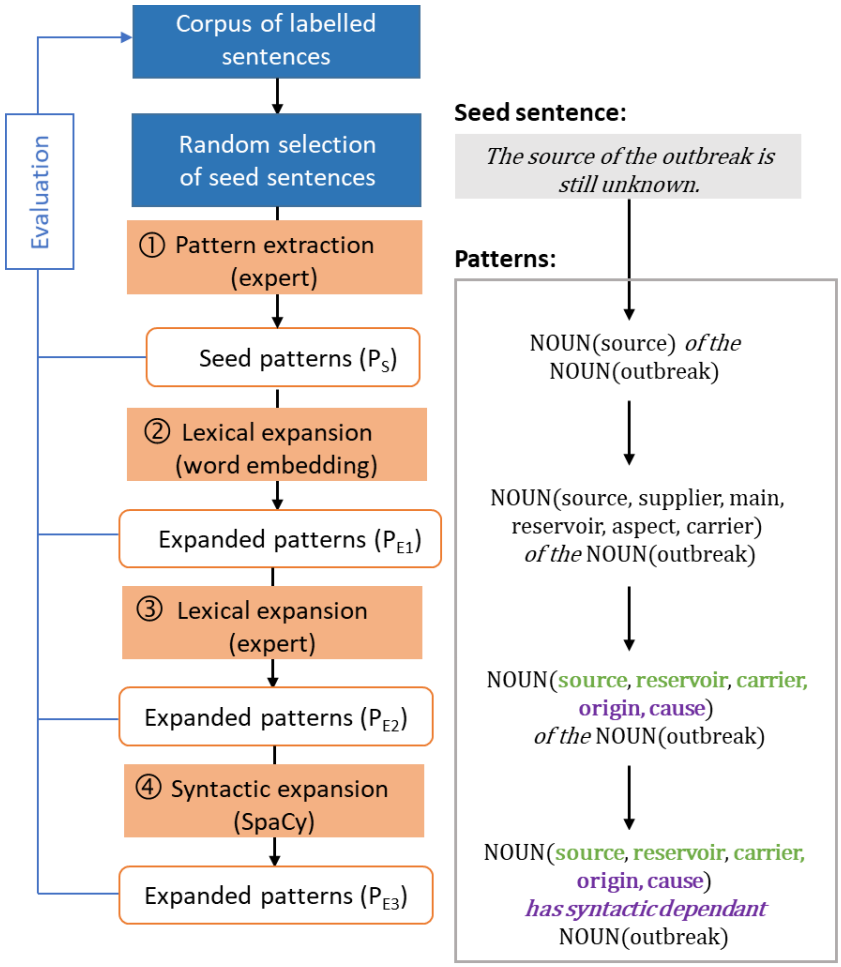
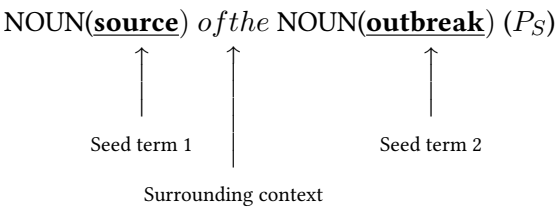


Figure 16 Incremental pattern expansion.
In the pattern box, seed terms are indicated between parentheses, and preceded by their POS. Terms validated by the expert are shown in green, and violet terms correspond to terms added by the expert. For readability, only the expansion of seed term 'source' is represented. The expansion steps (in orange) are detailed hereafter. Evaluation is done after each step.

Manual extraction of patterns (step 1)

This first step aims at extracting an initial set of patterns based on a minimal subset of sentences (seed sentences). We relied on the expert to read each seed sentence and identified one or several patterns specific to the sentence class (seed patterns). For instance, based on the seed sentence from Figure 16, the identified pattern is:



Where **source** and **outbreak** are the **seed terms** with *of* and *the* being linking words. Linking words usually consists of prepositions (e.g. "of", "through"), adjectives, adverbs or auxiliary verbs. Seed terms include all terms present in the **seed pattern**. In our approach, seed terms include nouns, verbs (and their preposition), and adjectives.To match the patterns with sentences, we used the seed term lemmas labelled with their part-of-speech (POS) to avoid possible ambiguities between nouns and verbs. For instance, the seed term NOUN(cause) matches both nouns "cause" and "causes", but does not match conjugated forms of the verb "to cause", such as "causes" or "causes".

Several patterns appeared to involve disease or host as seed term, e.g. in "Investigators look for **swine fever links**". All disease names (including acronyms) and host were thus replaced in the text by the word "disease" and "host" respectively, so that they could be represented by NOUN(disease) and NOUN(host).

Automatic lexical expansion (step 2)

At this step, we aim at expanding extracted patterns (P_s) at the lexical level. We focus on the **seed terms**, by automatically generating closed terms. We used a property of word embedding models, whereby words are represented as vectors. Words with common contexts (thus considered as similar) have close vectors in the produced vector space. In word2vec, the metric used to calculate the distance between two vectors is the standard cosine similarity (Leeuwenberg et al., 2016). The closer the cosine similarity between two word vectors is to one, the more similar the words are according to the model. Thus, for each word, cosine similarity can be used to rank terms in decreasing order of similarity.

For each **seed term**, we retrieved the K closest terms based on word embedding cosine similarity. For instance, the five closest terms for "source" are "supplier", "main", "reservoir", "aspect" and "carrier". The pattern is expanded based on the list of seed term synonyms: in (P_{E1}), the first term will match either the seed term "source" or one of its synonyms.

NOUN(**source, supplier, main, reservoir, aspect, carrier**) $of the$ NOUN(**outbreak**) (P_{E1})

Seed term 1 — Surrounding context — Seed term 2

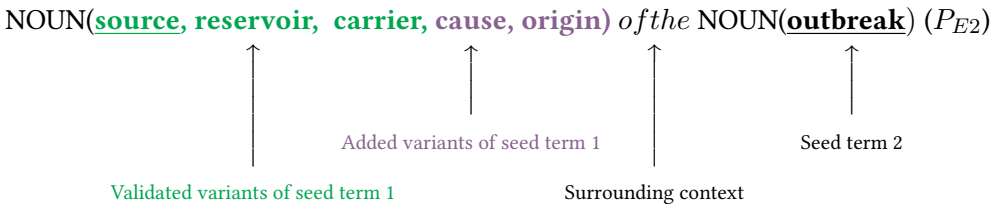
Finally, each **seed term** generates a set of K+1 variants (the seed term plus its close terms). Thus, if a pattern contains two seed terms generating K+1 variants each, we obtain $(K+1) \times (K+1)$ combinations.

As word embedding model, we used the 300-length CBOW model trained on the lemmatised text. The lemmatized version avoids the retrieval of derived forms of the same word (e.g. "detection" and "detects" for the term "detect"). We set K at 15, as a trade-off between the amount of input information and limited manual curation.

Manual lexical expansion (step 3)

At the previous step, the K closest terms provided by the word embedding model were considered as seed term synonyms by default. At this step, we use expert knowledge to validate this assumption and enhance the lexical expansion at different steps:

1. Manual validation of the list of variants (K closest terms) generated automatically; terms judged as irrelevant or not specific enough regarding the sentence category are removed.
2. Adding a new term, if not present in the initial list of variants;
3. Merging seed terms (and their corresponding variants), when they are considered as synonyms.



In (P_{E2}), the variants "supplier", "main", and "aspect" were considered as irrelevant by the expert and removed. The expert further added the variants "origin" and "cause".

Structure expansion (step 4)

The final step of manual curation consists of modifying the rigid contextual surrounding to improve the generalisation of patterns. A common approach is to use wildcards, i.e. symbols representing optional or specific characters and words in pattern matching. For instance, the pattern NOUN(source) of the NOUN(infection) could be replaced by:

NOUN(**source**) (W)? NOUN(**infection**)

The symbol (W)? indicates that NOUN(**source**) and NOUN(**infection**) can be separated by zero or more words. A major shortcoming of using wildcards is that syntactic information is not taken into account: the previous pattern only matches the source (or its variants) followed by the term infection (or its variants) but is not able to detect "the infection's source", for instance.

Thus, we proposed to expand the pattern structure based on the syntactic dependence between terms. More precisely, we modified the pattern structure when two (or more) seed terms were immediately syntactic dependent, i.e. connected by a single arc in the dependency tree (e.g. the subject of a verb, the adjective of a noun, etc.). In the following example, the new pattern is:

NOUN(**source**) *has immediate syntactic dependant* NOUN(**infection**) *(P_{E3})*

This new pattern is now able to match both "the source of the infection" and "the infection's source".

This approach has two advantages regarding pattern generalisation. First, as shown in the previous example, it increases the recall, as it does not rely on a specific sequence of terms such as wildcards. Second, the immediate syntactic relation a more fine-grained understanding of the meaning, thus avoiding irrelevant matches. We illustrate this feature with the following example. Sentence (1) and (2) belongs to the **Transmission pathway** and **Concern and risk factors**, respectively. Both sentences could match a pattern from the **Transmission pathway** class, such as NOUN(disease) (W)? VERB(spread).

"The disease has spread through water." (1)
"Fear of the disease spreads through the country" (2)

Figures 17 and 18 represent the dependency relations (i.e. dependency tree) of sentence (1) and (2):

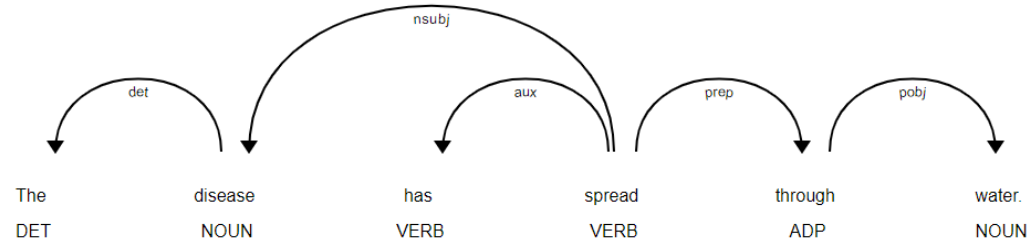


Figure 17 Dependency tree of sentence (1), generated with spaCy dependency parser.

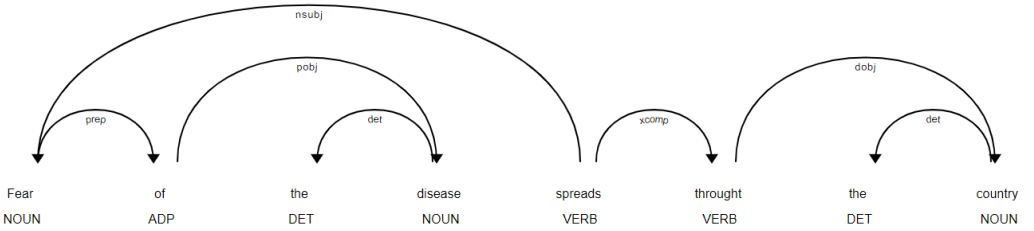


Figure 18 Dependency tree of sentence (2), generated with spaCy dependency parser.

The arcs represent the syntactic relations between the terms. In Figure 17, "spread" and "disease" are immediate syntactic dependents (connected by a single arc), while in Figure 18, "spread" is headed by "fear".

The analysis, including pattern-matching and generation of syntactic dependencies, was done using spaCy, i.e. a free open-source library for NLP in Python (Honnibal and Montani, 2018). We chose spaCy because it provides a pattern-matching function that readily allows pattern creation and enrichment. Syntactic relations were based on the spaCy dependency parser.

2.3.2 Evaluation

The pattern-based approach is particularly relevant for under-represented classes containing highly precise information, which are hard to identify by supervised methods. In this context, we evaluated the pattern-based approach on two Information type classes, i.e. **Concern and risk factors** (CRF) and **Transmission pathway** (TP). As seed sentences, we extracted 15% of an initial set of sentences belonging to the same class. The number of seed sentences was 16 (among 110) and 10 (among 69), respectively.

As shown in Figure 16, we evaluated the pattern performances after each expansion step in terms of recall, precision and F-measure. The supervised classification approach presented in Section 2.2 was multiclass (i.e. the sentences were classified into one out of three or more classes). Here, as we focused on two classes, we evaluated the pattern approach as a binary classification—a sentence is classified as **TP** (resp. **CRF**) if it matches at least one **TP** (resp. **CRF**) patterns (positive sentence). Otherwise it is considered a negative sentence. We excluded seed sentences from the evaluation to avoid artificial positive matches. The testing dataset hence respectively consisted of 59 positive and 1,025 negative sentences for **TP**, and 94 positive and 990 negative sentences for **CRF** categories.

2.3.3 Results and discussion

At step 1, the expert extracted 9 and 12 patterns from **TP** and **CRF** seed sentences, respectively (Table 12). The sentences and their corresponding patterns are shown in Appendix D and E. No identifiable pattern could be found in two sentences (one in each category). In the **CRF** category, three sentences had the same pattern. After step 3 (manual lexical expansion), the number of terms represented 27% (65/240) and 68% (113/165) of the terms generated automatically in **TP** and **CRF** classes, respectively. For the same final number of patterns (7), the number of term variants in the **CRF** class was twofold higher in the **CRF** class than in the **TP** class.

Table 12 Numbers of patterns and terms at the different pattern expansion steps for both Transmission pathway and Concern and risk factor categories. P_S, P_{E1}, P_{E2} and P_{E3} correspond to the seed patterns, and the 1st, 2nd and 3rd expansions, respectively.

	Transmission pathway (TP)				**Concern and risk factors (CRF)**			
	P_S	P_{E1}	P_{E2}	P_{E3}	P_S	P_{E1}	P_{E2}	P_{E3}
Number of patterns	9	9	7	7	12	12	8	7
Number of terms (seeds and variants)	16	240	65	65	11	165	113	113

Tables 13 and 14 show the performances of the pattern-based approach for the retrieval of **TP** and **CRF** sentences. The first extracted patterns (P_S) retrieved only 7% (4/59) of κ= sentences and 47% (44/94) of **CRF** sentences.

Table 13 Performances of the patterns for TP sentence retrieval in terms of precision, recall, and F-measure. The variation in the percentage corresponds to the change from the previous step.

	P_S	P_{E1}	P_{E2}	P_{E3}
Sentences retrieved (nb)	7	32 (+357%)	28 (-12,5%)	78 (+179%)
Precision	0.57 (4/7)	0.25 (10/32)	0.75 (21/28)	0.50 (39/78)
Recall	0.07 (4/59)	0.17 (10/59)	0.36 (21/59)	0.68 (40/59)
F-measure	0.12	0.20	0.49	0.58

Table 14 Performances of the patterns for CRF sentence retrieval in terms of precision, recall, and F-measure. The variation in the percentage corresponds to the change from the previous step.

	$\mathbf{P}_S$	P_{E1}	P_{E2}	P_{E3}
Sentences retrieved (nb)	70	227 (+224%)	138 (-39%)	139 (+0.7%)
Precision	0.63 (44/70)	0.33 (75/227)	0.52 (72/138)	0.53 (74/139)
Recall	0.47 (44/94)	0.80 (75/94)	0.77 (72/94)	0.79 (74/94)
F-measure	0.53	0.47	0.62	0.63

In both classes, the precision decreased after the automatic lexical expansion (step 2) and increased after the manual expansion (step 3).

Manual and automatic pattern expansion did not impact the **TP** and **CRF** recall in similar ways. In the **TP** class, lexical expansions obtained mitigate improvement in recall, reaching a maximum value of 0.36 after step 3. The syntactic expansion increased the number of retrieved sentences by 179% (28 to 78 sentences), thus obtaining the highest recall of all steps (0.68).

In the **CRF** class, the automatic lexical expansion reached a recall of 0.80. Manual lexical and syntactic expansion did not improve the recall but contributed to increasing the precision from 0.33 to 0.53. The syntactic expansion did not impact the number sentences retrieved (+0.7%).

These results were consistent with the characteristics of the patterns extracted from both **TP** and **CRF** classes at the first step. In the **CRF** category, the seed terms mostly consisted of a noun such as threat, risk, or fear (14 out of 16 seed sentences). The recall stability among the lexical and syntactic expansion steps confirmed that **CRF** sentences were homogeneous regarding their syntactic structure and vocabulary. On the contrary, the extracted **TP** patterns were more complex. Seed terms mostly consisted of a verbal-linguistic structure such as “could have been brought by”. Such syntactic structures were not generalizable, as highlighted by the poor recall at the first step.

Our results indicated that both lexical and syntactic information were crucial for improving the retrieval quality by the pattern-based approach. Yet the relative importance of each expansion step depended on semantic and syntactic specificity of each category. The manual curation by an

expert allowed filtering out of irrelevant terms automatically generated by the word embedding model, thus improving the precision. However, the final precision (after step 4) did not exceed 0.53, indicating that some patterns were ambiguous and not sufficiently class-specific. This constraint was offset by the fact that we aimed to minimize the number of false-negative instances.

To understand what impacted the final recall, we manually evaluated the sentences not detected by the pattern-based approach (false negative sentences).

We found that 13/19 and 11/20 sentences were based on identifiable patterns that were not captured as seed patterns in step 1, in **TP** and **CRF** classes, respectively. For instance, four **TP** sentences referred to ongoing investigations about the outbreak's cause. Indeed, the term "investigators" was present in one of the seed sentences but was not identified by the expert as specific to the **TP** category. In our approach, we only tagged disease and host with their entity type in the text. However, the "vector" category appeared particularly important for the **TP** category, as highlighted by the two following sentences:

1. "Migratory birds behind South Korea disease outbreak."
2. "The intense monsoon this year has caused a bloom of the midges exposing native animals and leading to disease."

Both terms "midges" and "migratory birds" are disease vectors. Tagging such terms with their entity type could generate new patterns such as VERB(cause) has syntactic dependent ENTITY(vector).

In the second group of sentences (5/19 and 9/20 sentences), no specific patterns could be identified in, for instance:
"The minister said in one case a man bought meat from Ukraine and gave the water he washed the meat in to his pigs, which got sick and died."

This type of sentence underlines the limitations of the pattern-based approach. The classification decision is based only on matching with pre-defined terms (pattern seed terms), which are sometimes not sufficient to capture the whole semantics of the sentences. We hypothesise that supervised classification that takes all the sentence terms into account may perform better in such cases.

Eventually, one last **TP** sentence was not detected because it contained the pronoun "it" in reference to the disease, which did not match the list of expanded terms:
"It has cooperated with public security departments to trace its origin."

This classic NLP problem is known as a noun phrase coreference resolution (Ng and Cardie, 2002). It highlights one limitation of the sentence-based approach in which references to entities are not inferred from other sentences. Further work could be focused on coreference resolution at the corpus level.

As a lexical expansion, we relied on terms automatically generated by the word embedding models. While a lot of retrieved terms were highly relevant, i.e. semantically close to the seed

term, a substantial number of them were irrelevant, as shown by the drop in the number of variants after expert validation (step 3). A well-known alternative to retrieve term variants is the use of an external lexical database such as case WordNet (Luhn, 1958). However, as highlighted by (Ibekwe-Sanjuan et al., 2011), WordNet is not domain-specific and may fail to provide appropriate words. For instance, the seed term "source" has "beginning", "root" and "informant" among its synonyms in WordNet. Word embedding models are also prone to generate irrelevant terms, i.e. not only retrieving synonyms but also antonyms, derived forms, hyper and hyponyms, etc. (Nooralahzadeh et al., 2018). Besides, there is no consensus regarding the number of terms to retrieve (K). (Ghosh et al., 2017) used the top-5 terms to expand medical expression patterns. Relying on the same K for all seed terms unavoidably overlooks some relevant terms, or retrieves irrelevant ones, as the number of variants varies between terms. An alternative could be to set a minimum threshold for the cosine similarity value (Rekabsaz et al. 2017). But determining whether the similarity score obtained from word embedding is indicative of term synonymy is still an open question. Leeuwenberg et al. (2016) showed that cosine similarity alone is a bad indicator to determine if two words are synonymous. They proposed a new measure, i.e. relative cosine similarity, which calculates similarity relative to other cosine-similar words in the corpus.

3 General discussion

In this chapter, we compared two approaches to retrieve epidemiological information at the sentence level. We evaluated different textual representations (bag-of-words and word embedding model) and classifiers for the supervised classification. We further described an incremental approach to create and expand patterns at both lexical and syntactic levels, with expert knowledge input.

The supervised classification results showed that the method classified several categories efficiently, such as **Descriptive** and **General epidemiology**. It would be interesting to evaluate how the pattern-based method would perform in such categories compared to the supervised method. The limited size of the learning dataset may have limited the performances of the supervised method, especially in under-represented classes. In the literature, extensive work has been devoted to feature enrichment of bag-of-words models. For instance, Morid et al. (2016) included UMLS concepts, semantic groups and cue words, among other features, to classify sentences as clinically useful or not. The enriched-sentences obtained an F-measure of 74% versus 37% for the feature count method. Another important feature is the tense: event-related sentences are typically written in the past tense while general sentences are written in the present tense. We believe that incorporating additional features, such sentiment score (Wu et al., 2018), named entities (Doan et al., 2007) or sentence tense (Dias et al., 2014), could enhance the performance of our supervised classification models .

With minimal computational and manual investment, the pattern-based approach outperformed the supervised approach in terms of recall (0.68 versus 0.55 for **TP** class, and 0.79 versus 0.39 for

CRF class). In line with systems that learn patterns automatically, automatic lexical expansion patterns increase the recall to the detriment of the precision (Nguyen et al., 2019). The role of the expert was therefore crucial to pinpoint the irrelevant terms generated by the word embedding model. Besides, the time cost of manual curation is minimal, as it mostly consists of validating or adding terms. This time depends directly on the threshold chosen for top K retrieved terms. The use of syntactic dependency was easy to implement, thus alleviating the cost and bias of wildcard fine-tuning.

Contrary to the supervised approach, the pattern-based method is not hampered by so-called "black box" problems and can be easily enhanced by expert knowledge. Besides, even though it was not evaluated in this study, it is suitable for multi-label sentences as a sentence can match patterns from several classes. On the other hand, as the precision of the patterns is poor to moderate, the multi-label approach may decrease the intra-class precision by boosting the number of misclassified sentences.

We also noted that the advantages of each method could be synergistic. Indeed, only 16/59 and 17/94 were not detected by pattern-based or supervised approaches . A promising perspective could thus be to evaluate how to jointly take full advantage the strength of both methods. (Cui et al., 2019) proposed an interesting approach to combine both supervised learning and manually built patterns and rules. They applied heuristic-based regular expression when the prediction confidence of the supervised classifier confidence was less than 0.6. The pattern-based classifier was used for the top 5 predictions (categories) with the highest confidence scores.

Chapter 3

Event extraction from news articles

Table of contents

In this chapter, we evaluate two different approaches to extract events from news article content: (i) a statistical approach based on the cooccurrence of epidemiological features, and (ii) a morphosyntactical approach. We first outline the current operationnal mode of the PADI-web information extraction module (Section 1), which allows users to identify and extract epidemiological entities from news article content. Then we present and evaluate both approaches. The event definition was tailored to each method and is detailed in the headlines of Sections 2 and 3.

1 Information extraction in PADI-web pipeline

The PADI-web Information extraction module is described in Chapter 1, Section 4.2.3. Briefly:

1. **Thematic entities** (i.e. diseases and hosts) are extracted by matching with a list of disease and host keywords.
2. **Spatio-temporal entities** (i.e. dates and locations): locations are identified by matching the text with the GeoNames gazetteer, and the dates with the HeidelTime rule-based system (Strotgen and Gertz, 2010). The extraction module automatically calculates the confidence score with which the locations and dates are related to an event. Relevant entities can thus be filtered based on their relevance according to a pre-determined confidence threshold.

The Information Extraction output consists of a structured list of epidemiological entities, but their relation to the event in the document is unknown. For instance, if two diseases d_1 and d_2 and a location l are identified, it is not possible to determine whether $\{d_1 \ ; \ L\}$ or $\{d_2 \ ; \ L\}$ is the attribute of event. Based on the event extraction terminology defined in Chapter 1, Section 4.2.3, two tasks are not addressed:

1. **The trigger identification**, i.e. the identification of main words(s) which most clearly express an event, is not performed.
2. **The attributes identification**, i.e. the identification of the relation between the attributes and the event.

In Section 2, we evaluate a statistical approach which ignores the event trigger identification. An event is only represented by its attributes, and we aim at detecting and ranking relevant sets of attributes (attributes describing the same event). In Section 3, we evaluate a morphosyntactical approach to identify the event attributes based on the event trigger identification.

2 Event extraction: a statistical approach

In this section, we define the event extraction task as the detection of pairs of epidemiological entities from different types (e.g. a disease name and location) describing the same event. Two entities form an event-related pair if they are attributes of the same event. Event-related pairs of attributes are hereafter referred to as "relevant pairs". Relevant pairs can only contain entities labelled as relevant following the Information Extraction step (Section 1). On the other hand, a pair of entities individually labelled as relevant can be irrelevant if the entities are not attributes of a same event. Thus, blindly grouping the extracted entities may generate wrong associations.

To address this issue, we evaluate a statistical method based on entity cooccurrence in news articles. More precisely, our approach involves two steps: (i) the detection of pairs of entities based

on their relative position in the news article content, and (ii) their ranking based on two state-of-the-art term association measures (*Pointwise Mutual Information* and $Dice$). Our contribution addresses the following questions:

1. What are the best cooccurrence parameters to select relevant pairs of entities from a corpus of news articles?
2. What is the impact of two association measures for the ranking of relevant pairs of entities?
3. How can contextual aspects be integrated for the ranking measures?
4. Does the generalisation of spatial entities improve the retrieval of relevant pairs?

Below we outline the proposed statistical approach and further describe the protocol and corpus used for the evaluation.

2.1 Statistical approach

2.1.1 Detection and ranking of pairs of entities

The computation of the association strength between two or more words (i.e. cooccurrence) is applied in several tasks, such as the discovery of association rules (Blanchard et al., 2005), feature extraction (Torkkola, 2003) and document summarization (Aji, 2012). Our objective is to identify the best parameters regarding entity cooccurrence and spatial hierarchy to improve the retrieval of relevant pairs, using $Dice$ and *Pointwise Mutual Information*. In the following, *Pointwise Mutual Information* will be referred to as *Mutual Information* (MI) for reason of simplification. MI has been used to discover and cluster words specific to events in a stream of tweets (Preoţiuc-Pietro et al., 2016). Our approach is based on the same rationale, but rather than taking all the words into account we compute the association measure only between predefined epidemiological entities (i.e. disease, host and location). Several other text-mining association metrics could be applied to our task, such as Jaccard, Cubic MI (Niwattanakul et al., 2013) or other measures such Bayes Factor, as applied in the data mining domain (Lallich et al., 2007). However, we opted to focus on $Dice$ and MI due to their simplicity, interpretability, and highly different behaviour regarding cooccurrence counts (Roche and Prince, 2010).

Mutual Information (MI) measures the relative difference between observed word cooccurrences, and their expected cooccurrence assuming independence (Church and Hanks, 1989). MI is defined as the probability that two words cooccur in the same context (the context concept is discussed below), divided by the product of the probabilities of each word occurrence in a corpus:

$$MI \quad = \quad log_2 \times \frac{P_{xy}}{P_x \times P_y} \tag{10}$$

where P_x is the probability of occurence of x, P_y is the probability of occurence of y, and P_{xy} is the probability of cooccurence of x and y (joint probability). *Mutual Information* is sensitive to rare and specific cooccurences (Roche et al., 2004).

$Dice$ coefficient (Equation 11) is also based on the joint probability, divided by the sum of the individual occurence probabilities. $Dice$ is less sensitive to low-count cooccurences (Smadja et al., 1996).

$$Dice = 2 \times \frac{P_{xy}}{P_x + P_y} \tag{11}$$

In both Equations 11 and 10, $P_x = \frac{N_x}{N}$, $P_y = \frac{N_y}{N}$ and $P_{xy} = \frac{N_{xy}}{N}$, where N_x is the number of occurences of x, N_y is the number of occurences of y and N_{xy} is the number of cooccurences of x and y. Moreover, as both metrics are used in a ranking purpose while the log function is a strictly increasing function, we can simplify Equations 10 and 11 as:

$$MI = \frac{N_{xy}}{N_x \times N_y} \tag{12}$$

$$Dice = \frac{N_{xy}}{N_x + N_y} \tag{13}$$

The results of both metrics heavily depend on the context chosen to compute the cooccurrence between two words. In our approach, this context controls the detection of pairs of features. In this thesis, we propose three definitions of cooccurrence contexts, hereafter referred to as "levels":

1. At the document-level: N_{xy} is the number of documents in which x and y cooccur;
2. At the sentence-level: N_{xy} is the number of sentences in which x and y cooccur;
3. At the word-level: N_{xy} is the number of times that x and y cooccur in a w word window.

Word and sentence levels rely on two parameters, i.e. the window size and the window side. The window size corresponds to the number of words (or sentences) separating two entities. The window side can be positive (y appears after x), negative (y appears before x) or bi-directional (y appears before or after x). For both disease-location and disease-host pairs, disease entities are considered as "pivot". Thus, a positive (resp. negative) window of w words corresponds to searching for another entity within the w words on the right (resp. on the left) of the disease feature. A bilateral window consists of searching for an entity on the right or left of a disease feature in a sliding window of w words.
We illustrate the influence of the window parameters on pair detection with an example extracted from an news article[1]. Location features are in bold while disease features are in italic (Figure 19).

[1] https://www.theguardian.com/environment/2020/apr/08/african-swine-fever-outbreak-reported-in-western-poland

An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Więckowice** near **Poznań** in western **Poland**, less than 150km (93 miles) from the border with **Germany**.

Figure 19 A news article content extract (The Guardian, 8 April 2020).

When setting the word window size at 15 words and the sentence window size at 1 sentence[2], the disease - location pairs are:

- At the document level: {*African swine fever*, Więckowice}, {*African swine fever*, **Poznań**}, {*African swine fever*, Poland}, {*African swine fever*, **Germany**};
- At a word level, rigth side, window of 15 words: {*African swine fever*, **Więckowice**}, {*African swine fever*, **Poznań}**;
- At a word level, left side, window of 15 words: no cooccurence;
- At a word level, both sides, window of 15 words: {*African swine fever*, **Więckowice**}, {*African swine fever*, **Poznań**};
- At the sentence level: {*African swine fever*, **Więckowice**}, {African swine fever, Poznań}, {*African swine fever*, **Poland**}, {*African swine fever*,**Germany**}.

2.1.2 Spatial generalisation

As illustrated in the previous example, spatial information can be provided at different granularity levels (e.g. city and administrative level). They generate different pairs of entities while representing the same location. Thus, we evaluated the impact of generalising the spatial entities to different granularity levels. More precisely, based on the GeoNames hierarchy, we converted the location entities into lower granular levels (e.g. converting "Allier" into "France"), hereafter referred to as "generalisation". We evaluated three generalisation levels:

- Level 0: No generalisation. This level corresponds to raw location values, without applying any generalisation. It includes spatial entities with heterogeneous granularity levels (e.g. cities, villages, as well as countries, etc.)
- Level 1: Administrative generalisation. This level corresponds to the conversion of spatial features into their first administrative level. This conversion is applied only if the initial spatial granularity is higher than the first administrative level. This level thus still includes heterogeneous granularity levels, such as administrative regions and countries.
- Level 2: Country generalisation. This level corresponds to the conversion of spatial features into their country. This last level only contains countries and supra-national entities (e.g. Asia, European Union).

[2]These parameters were chosen as an example, but a range of values are evaluated in Section 2.2.3.

We illustrate the impact of generalisation on cooccurrence weights with the previous example (Figure 20):

An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Więckowice**$_{LEVEL0}$ near **Poznań**$_{LEVEL0}$ in western **Poland**$_{LEVEL2}$, less than 150km (93 miles) from the border with **Germany**$_{LEVEL2}$.

Level 0

An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Greater Poland**$_{LEVEL1}$ near **Greater Poland**$_{LEVEL1}$ in western **Poland**$_{LEVEL2}$, less than 150km (93 miles) from the border with **Germany**$_{LEVEL2}$.

Level 1

An outbreak of *African swine fever* was confirmed on Monday on a farm near the village of **Poland**$_{LEVEL2}$ near **Poland**$_{LEVEL2}$ in western **Poland**$_{LEVEL2}$, less than 150km (93 miles) from the border with **Germany**$_{LEVEL2}$.

Level 2

Figure 20 Generalisation levels of spatial entities. The level of each location (based on the GeoNames hierarchy) is shown.

At level 1, all locations with a lower granularity than the first administrative level (i.e. **Więckowice** and **Poznań**) are converted into their administrative level (**Greater Poland**). At level 2, all locations are converted into their country level, which increases the joint probability of the pair {*African swine fever*, **Poland**}:

- Level 0: {*African swine fever*, **Więckowice**}: $N_{xy} = 1$, {*African swine fever*, **Poznań**}: $N_{xy} = 1$, {*African swine fever*, **Poland**}: $N_{xy} = 1$, {*African swine fever*, **Germany**}: $N_{xy} = 1$;
- Level 1: {*African swine fever*, **Greater Poland**}: $N_{xy} = 2$, {*African swine fever*, **Poland**}: $N_{xy} = 1$, {*African swine fever*, **Germany**}: $N_{xy} = 1$;
- Level 2: {*African swine fever*, **Poland**}: $N_{xy} = 3$, {*African swine fever*, **Germany**}: $N_{xy} = 1$.

The combination of association measures (Equations 12 and 13), cooccurrence contexts and spatial generalisation provides a mixed measure to evaluate both the detection quality and the ranking of relevant pairs:

- The window parameters control pair detection;
- The association measure (MI or $Dice$) controls the ranking of the detected pairs;
- For disease-location pairs, the spatial generalisation level jointly contributes to the detection of a pair and its ranking.

In the following section, we describe the evaluation protocol and the corpus used for the experiments.

2.2 Corpus and evaluation

To evaluate the proposed approach, we first annotated a corpus of news articles with events (Section 2.2.1). We further used the list of annotated events as a gold standard to automatically determine the relevance of the retrieved pairs of entities (Section 2.2.2). The quality of the ranked lists of pairs was evaluated using specific ranking evaluation metrics (Section 2.2.3).

2.2.1 Event corpus

We used a publicly available annotated corpus of 438 documents (i.e. news articles) related to animal disease events (either describing a recent outbreak or providing complementary insight regarding control measures, economic impacts, etc.) (Rabatel et al., 2019). This corpus was initially designed for training and evaluating the PADI-web information extraction module. The corpus contains information about the news article itself (publication date, title, content, URL, etc.), as well as epidemiological features (locations, diseases, hosts, dates and symptoms), which were first automatically identified by data mining and rule-based approaches. A veterinary epidemiologist and a computer scientist subsequently labeled each candidate as correct or incorrect. For each document and type of feature (i.e. disease, host, date and location), only candidates manually labelled as correct in the corpus were retained for analysis (including the geographical-geographical disambiguation of locations).

An epidemiologist read each of the 438 documents to detect all disease events they contained. To ensure a consistent and reproducible annotation, events found in the documents were compared to a gold standard database, i.e. the EMPRES-i database (Section 2.1). Each detected event was labelled using the unique EMPRES-i identifier. When the epidemiologist could not link an event to an official one, she created a new event identifier and manually recorded the epidemiological features (location, date, disease and host). The final corpus annotated with the event identifiers is hereafter referred to as the *event corpus*.

The number of news articles containing at least one event represented 53% of the corpus (n= 229/438). Among them, 52% (n= 127/229) reported several events, with a median number of 3 events (Table 15). One news article contained a maximum number of 208 events due to the reporting of 200 avian influenza outbreaks in Taiwan on 28 January 2015.

Overall, 771 events were detected in the corpus. Among them, 70% (n= 541/771) were reported in a single news article. The events present in several news articles were reported in up to 11 news articles (median number of 3 news articles).

In the following experiments, we selected only news articles containing at least one event (corpus of 229 documents).

Table 15 Descriptive statistics of the number of articles ($N_{article}$) per event and number of events (N_{event}) per articles in the event corpus.

	Min	Median	Mean	Max
n_{event} per article:				
Articles with $N_{event} >= 1$ (n=229)	1	2.0	5.1	208
Articles with $N_{event} >= 2$ (n=127)	2	3.0	8.4	208
$N_{article}$ per event:				
Events with $N_{article} >= 1$ (n=771)	1	1.0	1.5	11
Events with $N_{article} >= 2$ (n=230)	2	3	2.8	11

2.2.2 Relevant pairs

From the annotated event corpus, we can map each publication date $date_i$ with the set of its corresponding events (i.e. the event annotated in the news articles published at $date_i$). Sets including the publication dates and their epidemiological attributes are used as gold standard lists to evaluate the relevance of extracted pairs of features. We created a gold standard list specific to each type of pairs, as follows:

1. We aggregated news articles from the event corpus by publication date;
2. For each distinct date $date_i$, we extracted all events $event_i$ labelled in the set of news articles published at $date_i$ (gold standard list);
3. For each event $event_j$, we retrieved its disease ($disease_j$), host ($host_j$) and country ($country_j$);
4. The gold standard lists include all of the formed $date_i, disease_j, country_j$ and $date_i, disease_j$ and $host_j$ sets .

Identifying the extracted event per publication date was required to avoid false positive matches between retrieved pairs and the gold standard lists. As the event corpus covers a 2-year period, a pair of features extracted at $date_j$ could erroneously correspond to a pair corresponding to a different event.

The disease-location and disease-host gold standard lists contained 248 and 228 sets, respectively. Each retrieved pair extracted from a set of articles at date $date_i$ was relevant if it matched at least one pair from the gold standard list corresponding to date $date_i$ (Figure 21). To match the gold standard terms (disease names, species names and country codes from the EMPRES-i database), diseases and hosts were normalized to their canonical form using a manually built dictionary, and locations were normalized to their country code.

Note that the normalization of locations differs from the spatial generalisation described in Section 2.1.2. Normalization aims at matching a pair with the gold standard list features. Locations from a same country are not aggregated and considered as two distinct values at the pair extraction step. In the example from Figure 21, "Henan" and "China" are considered as two distinct values, even though they are normalized to the same country code.

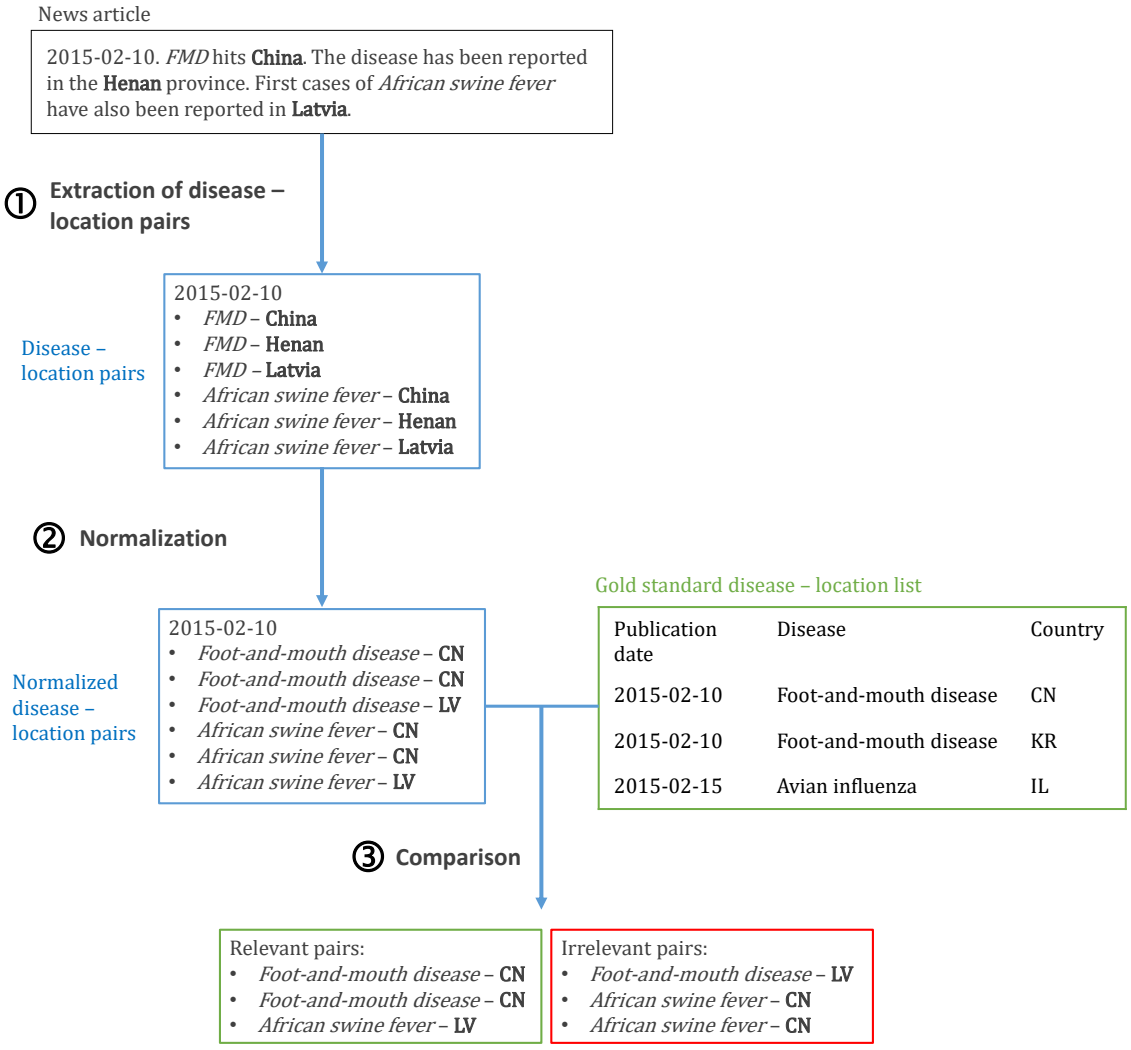


Figure 21 Steps to evaluate the relevance of the disease-location pairs extracted from a news article.
After extraction (1), disease and location features are normalized (2). Pairs matching a pair from the gold standard list is considered relevant.

2.2.3 Evaluation

Pairs extraction and ranking

We extracted all the disease-host and disease-location pairs using the cooccurrence parameters described in Section 2.1.1. The word window size ranged from 1 to 200 words on each side (left, right, and both). This window was chosen by (Piskorski et al., 2011) for an event extraction task, assuming that most relevant information would be present in the first 200 words. The sentence window size ranged from 0 to 20 sentences per side. We ranked the retrieved pair in decreasing order based on their *Mutual Information* or *Dice* values. We evaluated the quality of the ranked list according to the ability of the parameters and association measures to give a better rank to

relevant pairs than to irrelevant ones. The ranking was evaluated in terms of normalized precision (P_{norm}), normalized recall (R_{norm}) and F-measure (F_{norm}). R_{norm} and P_{norm} are based on the difference between the sum of ranks of R relevant pairs obtained by a ranking function, and the sum of ranks of an ideal list, where all relevant pairs are retrieved before all the irrelevant ones (Kishida, 2005; Salton and Lesk, 1968):

$$R_{norm} = 1 - \frac{1}{R * (N - R)} \times \sum_{i=1}^{R} r_i - \sum_{i=1}^{R} i \tag{14}$$

$$P_{norm} = 1 - \frac{1}{log(C(N, R))} \times \sum_{i=1}^{R} log(r_i) - \sum_{i=1}^{R} log(i) \tag{15}$$

where N is the total number of pairs,
r_i is the rank of the i^{th} relevant pair in the ordered list,
and $C(N, R)$=$\frac{N!}{R! \times (N-R)!}$.

Graphically, R_{norm} corresponds to the area under curve (AUC) of the receiver operating characteristics (ROC) curve, or AUC. Figure 22 provides an example of how ROC curves work regarding ranking evaluation.
Let R_1 and R_2 being the two ranked lists of pairs P_i :

- R_1 = P_2, P_1, P_4, P_6, P_5, P_3
- R_2 = P_3, P_4, P_6, P_5, P_1, P_2

For each relevant pair (in bold, the curve increases one unit in the Y-axis direction. For each irrelevant pair, the curve increases one unit in the X-axis direction. Consequently, the AUC of the best ranking function (here, R_1) is greater than that of a function giving a poorer ranking (here, R_2).

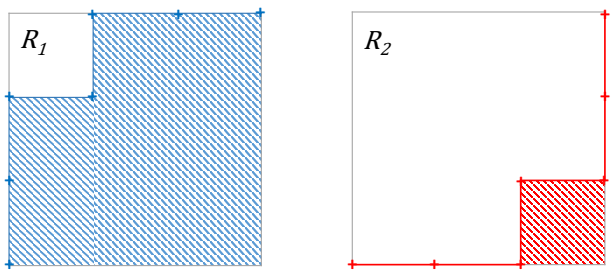


Figure 22 ROC curves obtained by two different rankings, R_1 and R_2. The dashed areas correspond to the AUC.

The normalized F-measure F_{norm} is the harmonic mean of R_{norm} and P_{norm} (Equation 16).

$$F_{norm} = 2 \times \frac{R_{norm} \times P_{norm}}{R_{norm} + P_{norm}} \quad (16)$$

We also evaluated the quality the 5 first pairs retrieved by calculating the precision at k ($P@k$), recall at k ($R@k$), and the F-measure ($F@k$), with k=5. We chose this threshold because it provides of the local ranking quality, and 95% of the sets of relevant pairs had 1 to 5 elements (pairs).

2.3 Results

2.3.1 Disease - host pairs

Table 16 summarizes the best results obtained among all the window parameters evaluated, and the performances at the document-level. The word-level window outperformed document-level and sentence-level windows in terms of normalized precision and recall. The highest precision and recall values were obtained with *Dice* using a window of 26 words on the right side (R_{norm}=0.90, P_{norm}=0.92). The performances obtained with *Dice* values were higher than the *MI* values.

Table 16 Performances of MI and $Dice$ to retrieve and rank relevant disease - host pairs at document level, based on P_{norm}, R_{norm} and F_{norm}.
For sentence-level and word-level, the performances correspond to the best values among the range of window sizes and sides.

	Mutual Information			**Dice**		
	R_{norm}	P_{norm}	F_{norm}	R_{norm}	P_{norm}	F_{norm}
Document level	0.79	0.80	0.80	0.83	0.85	0.84
Sentence level	0.78	0.81	0.80	0.84	0.88	0.86
Word level	0.82	0.85	0.87	**0.90**	**0.92**	**0.91**

The maximum recall at 5 (R@5) ranged from 0.89 to 0.92, while the precision at 5 ($P@5$) reached a maximum value of 0.88 (Table 17). The word level obtained the best recall-precision balance ($F@5 = 0.88$).

Figure 23 shows the normalized F-measure (F_{norm}) among the word window sizes and sides. The horizontal lines correspond to the F_{norm} values obtained at the document level. At a given window size and side, *Dice* systematically outperformed *MI*. For both metrics, we achieved better F_{norm} values using a right or bilateral window, clearly outperforming the left side windows. For all curves, the slope rapidly increased when the word distances increased from 1 to 100. *MI* performances decreased with window sizes of more than 100 words. Dice had a different behaviour, where the performances stayed stable among all the window sizes when the values peaked (100 to 200 words).

Contrary to the global ranking, $Dice$ and MI obtained similar performances for the retrieval of the first 5 pairs (Figure 24). The F-measure behaviour was similar to the global ranking, with right and bilateral sides obtaining the best results while remaining stable among the window sizes.

Table 17 Performances of MI and $Dice$ to retrieve and rank relevant disease-host pairs, based on $P@5$, $R@5$ and $F@5$.
For sentence-level and word-level windows, the performances correspond to the best values among the range of window sizes and sides.

	Mutual Information			Dice		
	$R@5$	$P@5$	$F@5$	$R@5$	$P@5$	$F@5$
Document level	0.91	0.81	0.86	**0.92**	0.81	0.86
Sentence level	0.89	0.85	0.87	0.90	**0.85**	0.87
Word level	0.90	0.84	0.87	0.91	0.84	**0.88**

2.3.2 Disease-location pairs

The maximal normalized F-measure values for disease-host pairs ranged from 0.62 (document-level, MI) to 0.88 (word-level, $Dice$) (Table 18). At the document level, the generalisation at the administrative level (level 1) slightly improved the performances. The second level (country level), improved the recall and precision of the Dice ranking at the word level (improving the F-measure from 0.81 to 0.88). However, it decreased the MI ranking performances (at the word level, the F-measure decreased by 0.11).

Table 18 Performances of MI and $Dice$ based on P_{norm}, R_{norm} and F_{norm} to retrieve and rank relevant disease-location pairs at the document level according to the spatial generalisation level.
Level 0: no generalisation, level 1: first generalisation level, level 2: second generalisation level.

		Mutual Information			Dice		
	Generalisation	R_{norm}	P_{norm}	F_{norm}	R_{norm}	P_{norm}	F_{norm}
Document-level	Level 0	0.73	0.73	0.73	0.75	0.76	0.76
	Level 1	0.76	0.76	0.76	0.77	0.80	0.78
	Level 2	0.64	0.60	0.62	0.81	0.82	0.81
Sentence-level	Level 0	0.73	0.74	0.74	0.75	0.75	0.75
	Level 1	0.73	0.74	0.74	0.75	0.76	0.75
	Level 2	0.63	0.66	0.64	0.62	0.66	0.64
Word-level	Level 0	0.72	0.78	0.75	0.78	0.84	0.81
	Level 1	0.71	0.74	0.73	0.77	0.82	0.80
	Level 2	0.68	0.73	0.71	**0.88**	**0.88**	**0.88**

Figure 25 highlights the different behaviours of *Dice* and *MI* regarding the generalisation level. Without generalisation (level 0), the best F-measures were obtained for both metrics with a window of 25 words (on both sides), and the scores slightly decreased with the highest window sizes. At the country level, the MI F-measure stayed below 0.70 while that of Dice ranged from 0.80 to 0.88. The $Dice$ ranking reached maximum values between 100 and 125 words (both sides) and remained elevated for all window sizes.

The ranking quality at 5 was sensitive to the word windows regarding both the level and gen-

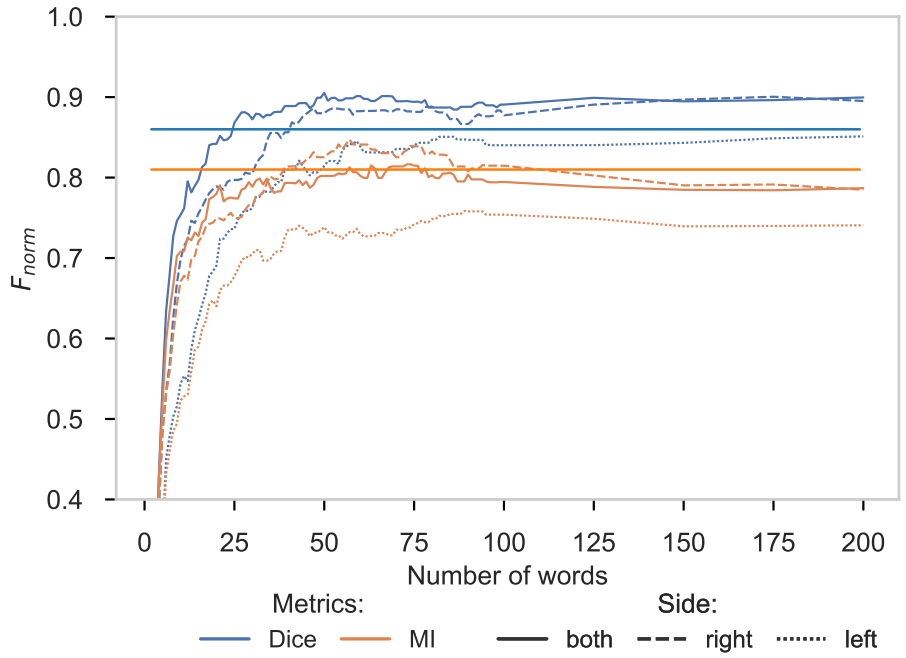


Figure 23 Performances of MI and $Dice$ to retrieve and rank relevant disease-host pairs in terms of F_{norm}, depending on the window parameters used for the co-occurence count. For the left side, distances were converted into their positive values. Horizontal lines correspond to the F_{norm} values obtained at the document-level for *MI* (orange line) and *Dice* (blue line).

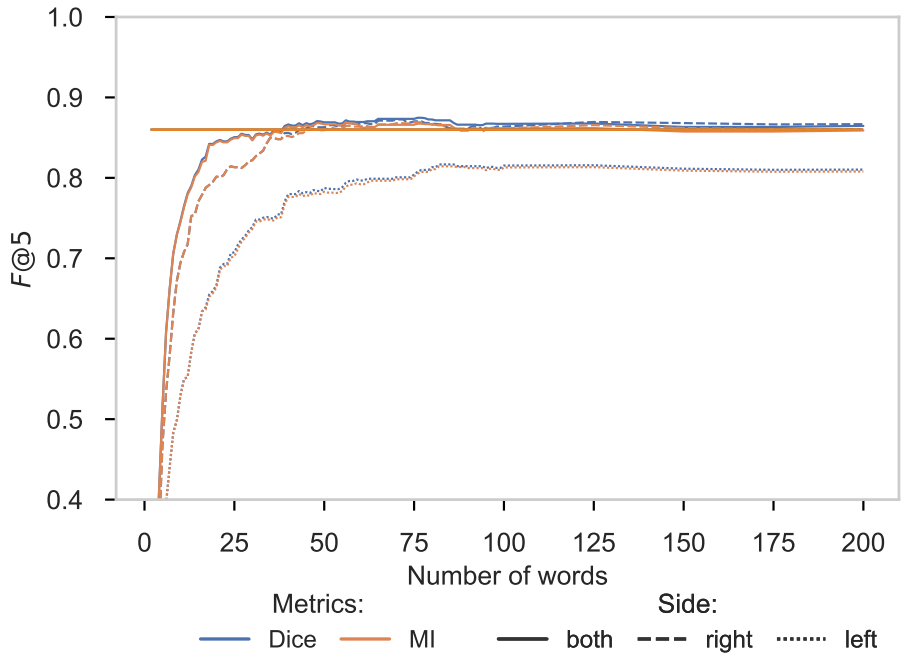
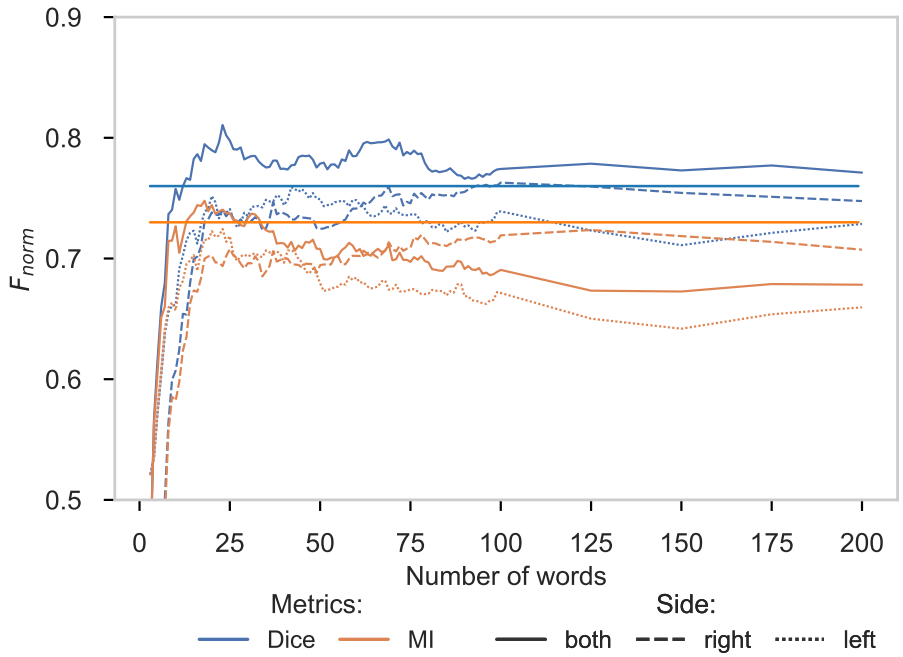
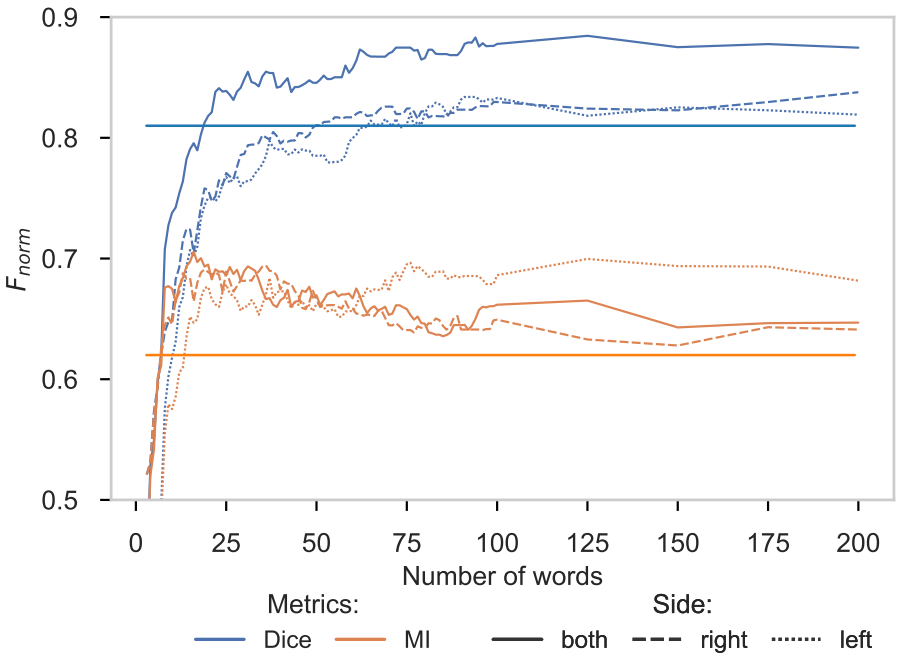


Figure 24 Performances of MI and $Dice$ to retrieve and rank relevant disease-host pairs in terms of $F@5$, depending on the window parameters used for the co-occurence count. For the left side, distances were converted into their positive values. Horizontal lines correspond to the P@5 values obtained at the document level for MI (orange line) and Dice (blue line)



a) Level 0



b) Level 2

Figure 25 Performances of MI and $Dice$ to retrieve and rank relevant disease-location pairs in terms of F_{norm} depending on the window parameters used for the cooccurrence count and different spatial generalisation levels.
For left side, distances are converted to their positive value. Horizontal lines correspond to the normalized F-measure values at document-level. Level 0: no generalisation, level 2: generalisation at the country level.

eralisation (Figure 26, Table 19), with the best F-measures reached within a 50 word window.

Table 19 Performances of MI and $Dice$ based on $P@5$, $R@5$ and $F@5$ to retrieve and rank relevant disease-location pairs at the document level according to the level of spatial generalisation.
Level 0: no generalisation, level 1: first generalisation level, level 2: second generalisation level.

		Mutual Information			Dice		
	Generalisation	$R@5$	$P@5$	$F@5$	$R@5$	$P@5$	$F@5$
Document-level	Level 0	0.89	0.78	0.83	0.95	0.77	0.85
	Level 1	0.90	0.77	0.83	0.95	0.77	0.85
	Level 2	0.95	0.67	0.79	**0.98**	0.68	0.80
Sentence-level	Level 0	0.90	0.85	0.87	0.89	0.86	0.87
	Level 1	0.90	0.85	**0.88**	0.90	0.85	**0.88**
	Level 2	0.93	0.76	0.84	0.93	0.76	0.84
Word-level	Level 0	0.91	0.85	0.88	0.91	0.85	**0.88**
	Level 1	0.91	0.80	0.84	0.92	0.80	0.85
	Level 2	0.84	**0.87**	0.85	0.85	0.86	0.86

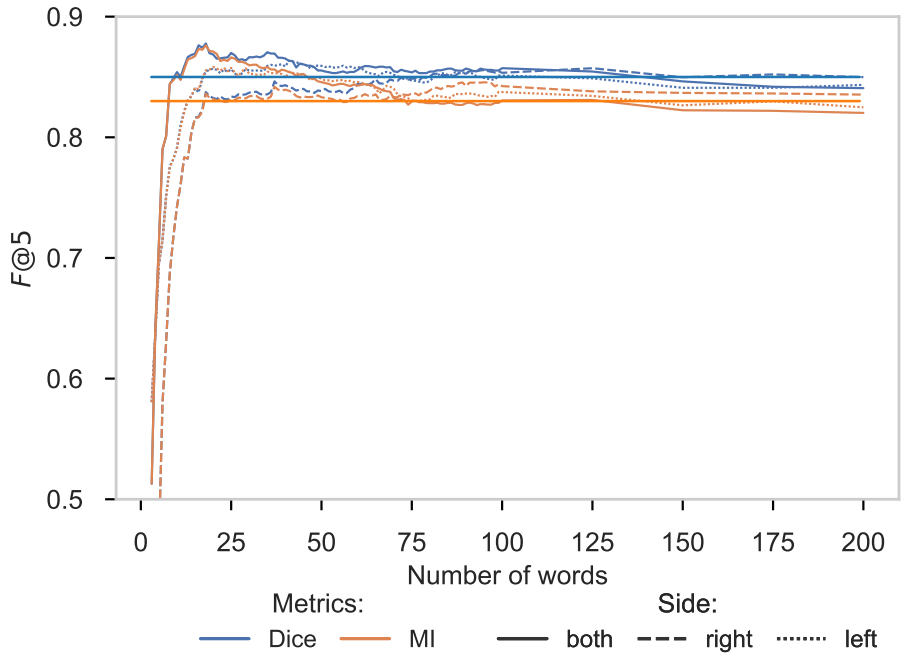
2.4 Discussion

The statistical approach was able to detect event-related pairs of epidemiological features with a good trade-off between precision and recall. Our results showed that using a window of words outperformed document-based and sentence-based approaches, while reducing the probability of detecting false pairs.

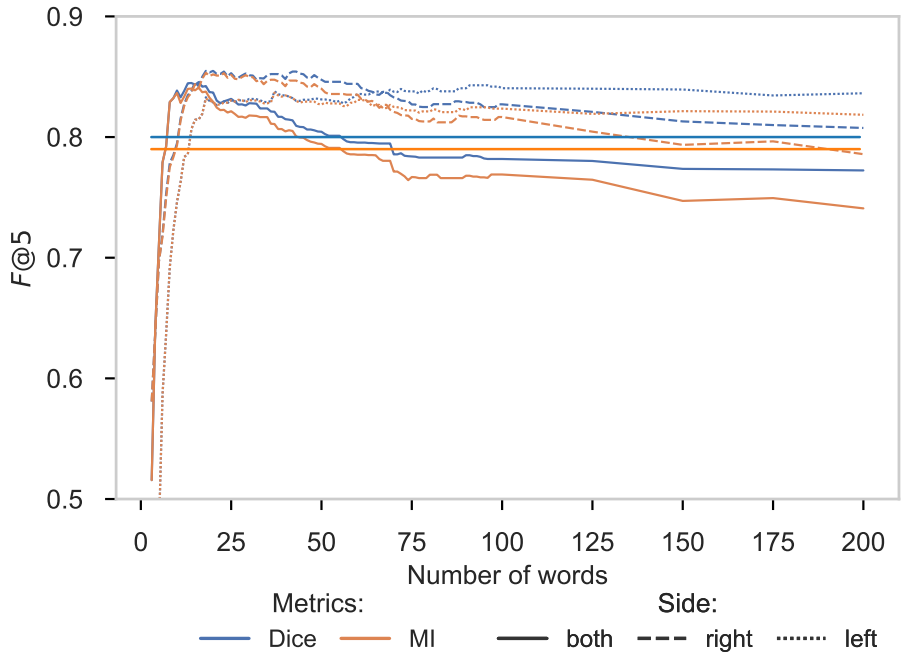
Our results indicated that Mutual Information was less adapted than the Dice coefficient for the ranking of pairs of features in the event extraction framework. This was especially true when generalising spatial features and increased occurrence counts. Besides, Dice ranking was found more resistant to larger word windows, in line with the findings of (Bouma, 2009), who proposed to add a normalization factor to the Mutual Information formula to address low-count issues. We believe that Mutual Information would be more relevant for rare pair detection (i.e. weak signals), but requires higher manual curation to avoid false positive extraction pairs.

Disease-location detection and retrieval

The results obtained for the retrieval of disease-location pairs without any generalisation suggested that relevant spatial features tended to occur within a small window around the disease feature (25 words, bilateral window). Beyond this window range, the global ranking performance decreased. However, global ranking with the Dice coefficient after country-level generalisation had a different behaviour, i.e. staying stable and close to its maximum value throughout the window size range. Event-related spatial features are provided at different granularity levels. Spatial generalisation allowed us to aggregate related locations in single features, thus increasing the



a) Level 0



b) Level 2

Figure 26 Performances of MI and $Dice$ to retrieve and rank relevant disease-location pairs in terms of $F@5$ depending on the window parameters used for the cooccurrence count and different spatial generalisation levels.
For the left side, distances are converted into their positive value. Horizontal lines represent the $F@5$ values at the document level. Level 0: no generalisation, level 2: generalisation at the country level

weights of event-related pairs. Moreover, spatial generalisation overcomes possible location extraction and disambiguation errors. For instance, news articles often refer to the place where the sample analyses are done, thus citing a laboratory place. If the laboratory-based city is extracted as a candidate for being event-related, using the city feature itself would generate a false alarm. This issue may be overcome by converting the location value into its country, yet this would lead to lower spatial precision. In the epidemic intelligence framework, the country is an acceptable level for signal analysis (note that the signal concept is outlined in Chapter 1 Section 1.3.1), but this approach may not relevant if fine-grained location extraction is needed.

Several gold standard disease-location pairs were not detected due to a problem of linkage between GeoNames and the Global Administrative Unit Layers (GAUL) used by the EMPRES-i database. In the latter, Taiwan and Hong Kong are considered as distinct countries. On the contrary, the GeoNames hierarchy considers them as administrative units from China. (Claes et al., 2014) used a manual procedure to link both databases.

Manual analysis of irrelevant retrieved pairs showed that most of them were due to the presence of multiple events, which the statistical approach did not succeed in separating.

Disease-host detection and retrieval

The detection and ranking of disease-host pairs achieved better results than the retrieval of disease-location pairs. The variation in the word window had less impact than for the disease-location pairs. This indicates that when the highest recall value was achieved, the precision also remained high. This was expected since thematic features are much less prone to ambiguity than spatial or temporal features. When additional events were present in a news article, they were often summarized in a few sentences containing only the disease and location, thus reducing the probability of creating false disease-host pairs. Several pairs were due to the fact that irrelevant host terms were extracted from two disease variants, i.e. "small ruminant plague" and "cattle plague". As these two expressions were not in the dictionary, they were not recognized as diseases, which led to erroneous extraction of "small ruminant" and "cattle" as hosts. Both cases were corrected by adding the new variants to the dictionary and reconducting the experiments. Animal disease expressions are often composed of several terms containing epidemiological entities such as hosts and symptoms, so the information extraction performance (and, in this particular case, the coverage of dictionary-based methods) is critical for event extraction. We observed that using the broader cooccurrence detection level (i.e. document level), we did not achieve a recall of 1. After manual investigation, we discovered that 5 news articles did not contain any reference to a host. Four out of 5 news articles were about foot-and-mouth disease in endemic areas and the last one referred to a suspected case of African swine fever. Such cases were rare in the studied corpus, but it should be noted that the host attribute is not necessarily communicated in news content, either because the disease is well known yet (e.g. endemic disease in an area) and the host is thus implied, or because this information is secondarily compared to spatial features. Models that determine the relevance of news articles based on the presence or absence of a host name may be

biased by this behaviour.

The proposed method relies on basic statistical features derived from word occurrences and their relative position in the text. Controlling the distance between two features is a simple intuitive and approach to decide whether or not they form a relevant pair. However, chosing a window threshold implies achieving a balance between recall and precision, which is particularly important for disease-location detection. Using measures to compute the cooccurrence strength, such as the Dice coefficient, allows the user to rank numerous pairs and prioritize the most relevant ones, rather than relying on binary classification. Besides, this approach overlooks the semantic structure of the text. This can be an asset, for instance to overcome possible errors or approximations after automatic translation. The current approach only requires efficient entity extraction.

Yet ignoring the semantic structure also has limitations. This method should first of all be used within an efficient classification pipeline able to precisely identify news describing events. Otherwise many false positive pairs may be extracted from news articles about other disease-related topics. Moreover, controlling the word distance cannot prevent the detection of irrelevant pairs, thus limiting the overall precision. For instance, a number of irrelevant disease-host and disease-location pairs were due to a general description of a disease and its epidemiology:

Brucellosis is caused by different Brucella species, which are named for their primary hosts: Brucella melitensis is found mostly is goats, sheep and camels, B. abortus is a pathogen of cattle, B. suis is found primarily in swine and B. canis is found in dogs. In such cases, semantic and temporal clues (use of a state verb, present tense) are essential (see "Sentence classification" section).

As further work, a comprehensive comparison of the state-of-the-art association metrics could provide an extensive overview of their performances.

3 Event extraction: lexicosyntactical approach

In this section, we evaluate a turnkey approach to extract event attributes. Contrary to the previously described statistical method, the lexical and semantic aspects of the text here are essential in the attribute identification process—with verbs having a pivotal role. This seems particularly relevant to our work since the major role of verbs in event extraction, especially event mention, has been previously highlighted (Chapter 1, Section 4.2.3). The approach is already implemented in an existing software package developed for French textual data. The aim of this study was to investigate a generic and fine-grained event extraction method for comparison with statistical-based methods. As we used French data, we could not perform an exhaustive and comparative study with the approach presented in Section 2, so we conducted a preliminary and mostly qualitative evaluation.

3.1 Lexicosemantic representation approach

The fine-grained event extraction method is implemented in a software package developed by Emvista[3] , as described in (Mekaoui et al., 2020). This method is based on the lexicosemantic representation approach with two modules: (i) a module based on lexicosemantic resources, and (ii) a module independent of any lexicosemantic resources. The resource-based module uses a French version of VerbNet, which is an English verb lexicon that relies on Levin verb classes to construct lexical entries systematically (Schuler, 2006). In VerbNet, all verbs members for a specific class have common semantic and syntactic properties. More precisely, each verb (hereafter referred to as **predicate**) is assigned a syntactic frame containing a list of specific attributes. An attribute is defined by its role and has specific restrictions. For instance, the verb class $transfer-mesg-37.1.1$ has, among others, an "Agent" attribute which is restricted to animated objects. In the sentence *"Margot sent me a letter yesterday"*, the verb "sent" has four attributes:

- **Agent**, which is "Margot";
- **Recipient**, which is "the author" (inferred by the use of the pronoun "me");
- **Theme**, which is "a letter";
- **TimeExact**, which is "yesterday".

Besides, the system detects if the predication is negated or not.

The French version of VerbNet, called "VerbeNet", was recently developed by (Danlos et al., 2015). As the coverage of VerbeNet is limited, the resource was manually enriched with new verbs, and a second module was created. This module is based on manually built rules and focuses on verbs that were not extracted by the first module to increase the coverage.

[3] https://pss.prevyo.com/

It involves two steps:

1. The generation of candidates (for both predicates and attributes), using rules based on syntactic dependencies and morphosyntactic patterns.

2. The selection of a single candidate from the set generated at the previous step. A candidate is selected if its semantic corresponds to the most likely restrictions of each role according to VerbeNet. For instance, an Agent is most likely a person or an organisation. If several candidates are relevant, the system returns several solutions.

The output of the semantic representation is available in two formats, i.e. a graph view and a json file, as illustrated in Figures 27 and 28. These figures represent the output of the semantic analysis of the sentence in French (1) and its translation in English (2):

"La France a déclaré un foyer de grippe aviaire le 9 janvier 2019." *(1)*

"France declared an outbreak of avian influenza on January 9th, 2019." *(2)*

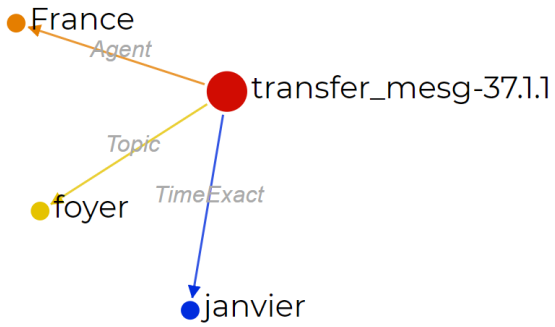


Figure 27 Example of graph-based semantic representation.
The main nodes (in red) represent the predicate (i.e. verbs), while the secondary nodes represent the attributes. The edges represent the relation between the predicates and attributes

The VerbeNet class for the verb "declared", i.e. $transfer - mesg - 37.1.1$ and the negation value is False, which means that the verb is used in its affirmative form (Figure 28). Attributes corresponding to the verb frame and which were found in the sentences are listed, i.e. Agent, TimeExact, and Topic. Each attribute is tagged to indicate whether it is a named entity (here, both "France" and "January" are identified as named entities).

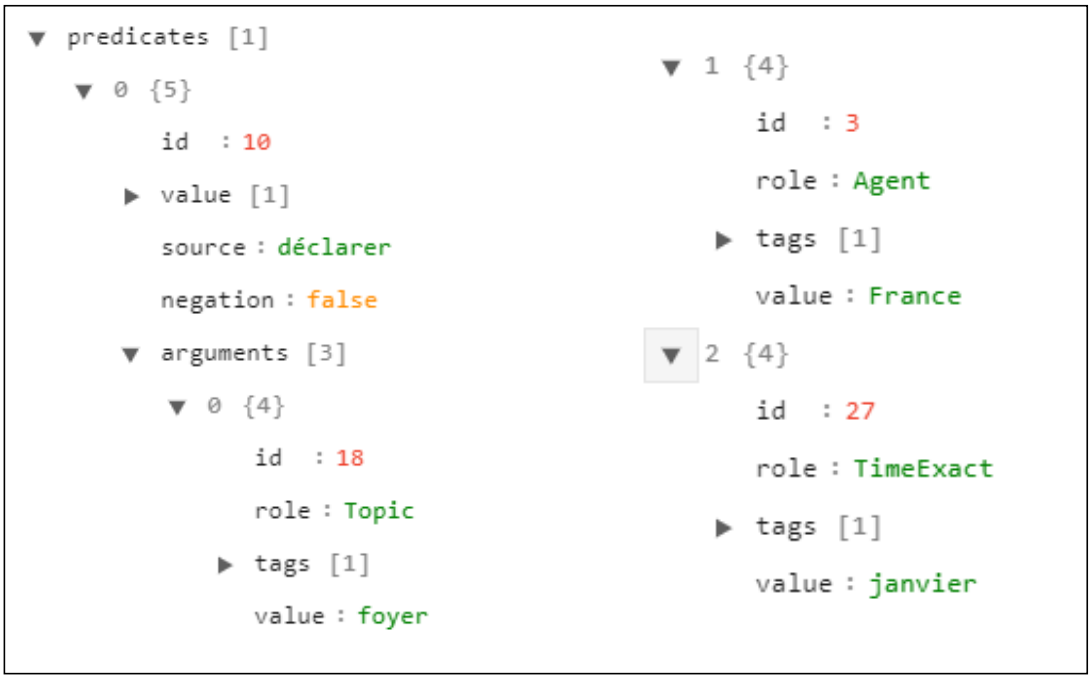


Figure 28 Example of semantic representation, focusing on the "declared" predicate.

3.2 Evaluation

To date, the method has only been developed to deal with textual data in French. As PADI-web retrieves news in several native languages, we could select new articles from the PADI-web database that were published in French. We further manually selected 11 news articles declaring a new event. Once the semantic representation is generated, filters can be used to retain only the desired information. Filters can be based on the predicates (selecting only to predicates belonging to a specific VerbNet class) or on the roles (selecting a predicate whose attribute is a location). Such an evaluation would consist of predefining classes of verbs typically mentioning an event, such as “declared” or “detected”, and selecting the predicates on this basis. We manually identified the relevant sentences since we did not have any a priori knowledge about the verb classes. As relevant sentences, we refer to sentences containing at least one epidemiological entity related to an event (i.e. a disease name, host, location or date), considering them as event mentions. We obtained 44 relevant sentences which will be used as an evaluation corpus.

For wider analysis and prospective use, the filtering approach based on predicates and attributes allows automation of this step, as long as the filter coverage is sufficient.

We evaluated the approach regarding the retrieval of spatiotemporal and thematic attributes (i.e. disease, host, location and date). All of these attributes were first manually identified, and then we calculated the recall, precision and F-measure after the semantic representation. As the predicate frames were generic, they did not include “host” or “disease”. These entities were thus

searched in the "Theme" or "Topic" attributes. Besides, after the initial tests, we included the epidemiological unit (i.e. "case", "outbreak"), the number and type of cases. This latter attribute corresponds to the adjective used to qualify the case, which is relevant from an epidemiological standpoint, e.g. "first", "new".

3.3 Results and discussion

Relevant predicates were found in 38 out of 44 sentences. In these 38 sentences, we found 43 occurrences of predicates (some sentences had several occurrences). Table 20 shows the predicate values ranked by decreasing number of occurrences. Among the verbs, seven were related to the disease appearance or spread, five to the disease detection, four to the disease declaration, and one to the host's death. In 6 out of 44 sentences, no relevant predicates were extracted. These sentences were:

- nominal sentences, e.g. "Two cases of African swine fever detected in Belgium";
- sentences in which the event mention is identified through a noun or an adjective instead of a verb, e.g. "The detection of two positive cases is worrying.", "After the first detected case […]".

These results could be explained by the fact that the verbs (predicates) are at the core of the approach, and only terms which fit the verb frame can be captured.

Table 20 Predicates (verbs) extracted from event-related sentences.

Verb	Class of verb (VerbNet)	Broad category	Number of occurrences
confirm	accept-77-1	Declaration	9
die	die	Death	5
detect	unknown	Detection	3
spread	appear-48.1.1	Appearance, spread	3
announce	transfer_mesg-37.1.1	Declaration	3
appear	appear-48.1.1	Appearance, spread	3
notice	notice	Declaration	2
contaminate	unknown	Appearance, spread	1
decimate	unknown	Appearance, spread	1
reach	unknown	Appearance, spread	1
discover	discover-84	Detection	1
strike	hurt-1	Appearance, spread	1
affect	give-13.1	Appearance, spread	1
register	estimate-34.2	Detection	1
find	discover-84	Detection	1
inform	transfer_mesg-37.1.1-1	Declaration	1
prove to be	state	Detection	1
Total			43

Table 21 shows the performance of the approach for attribute retrieval.

Table 21 Retrieval of event attributes in terms of recall, precision and F-measure.

	Disease	Host	Locations	Temporal	Number of cases	Epid.unit	Type of cases
Recall	0.04 (1/24)	0.58 (7/12)	0.25 (13/53)	0.33 (7/21)	0.91 (32/35)	0.79 (27/34)	1 (21/21)
Precision	1 (1/1)	1 (7/7)	0.81 (43/53)	1 (7/7)	1 (32/32)	1 (27/27)	0.95 (21/22)
F-measure	0.08	0.73	0.38	0.50	0.95	0.88	0.97

Considering the low number of evaluated sentences, we formulated the following hypotheses:

- The approach yields excellent precision, regardless of the type of attribute;
- The approach yields poor results in terms of recall for the retrieval of spatiotemporal and disease entities;
- The approach is efficient in identifying the hosts, number and type of cases, and the epidemiological unit.

Several features of the approach could explain its poor performance regarding the retrieval spatiotemporal and disease entities. We hypothesise that several syntactic frames did not include a spatial or temporal attribute. The approach failed to identified disease names, which most of them are multiword expressions, as a whole entity. Besides, each attribute could have a single value. This hampered the retrieval of locations when there were several spatial granularity levels (e.g. a city and an administrative region) in the sentence. False-positive locations were due to multiword entities, such as *Corée du Sud*, which is the French expression for South Korea. In such cases, the preposition "du" was identified as the location. We do not know what caused this behaviour.

Conversely, the approach correctly identified hosts and epidemiological units. Indeed, such terms were typically the theme (or topic) of a declarative verb, thus easily captured by the frame, e.g. *"An outbreak of foot-and-mouth disease was detected"*, *"China declared a case of African swine fever"*. The ability to identify both the numbers and types of cases is an interesting feature of the semantic approach. Besides, numbers are provided at a fine-grained level, thus indicating if the quantity is exact, e.g. *"45 pigs died"*, or if the exact quantity is known, e.g. "several cases". For the latter, the number is called "minimal measure" and set at two. Beyond the number of cases, retrieving the type of case is a major advantage. Indeed, it allows us to automatically determine if the case (or outbreak) is new in a given area (e.g. "the first case"), or if it is part of an ongoing situation (e.g. "a ninth case", "an additional case").

Importantly, the method correctly identified the affirmative or negative version for all of the 43 predicates, including three negated predicates. Even if negated events are rare in disease-related news, this feature avoids false alerts in sentences such as *"no additional cases were detected"*.

This preliminary evaluation suggests that the semantic-based approach could be used to enhance the extraction of specific event attributes in the animal health surveillance context. The evaluated methods appear as complementary to domain-specific approaches as they are able to detect quantities (number of cases) at a higher semantic level than simple entity extraction, indicating for instance if a measure or a temporal entity is exact or fuzzy. However, a broader evaluation is needed to validate the results outlined in this work.

Chapter 4

Using epidemiological features to improve information retrieval

Table of contents

In this chapter, we address the task of linking news articles which are related in epidemiological terms. This task corresponds to the identification of related documents, as described in Chapter 1, Section 4.1.3. More precisely, we aim to identify the best textual representation, i.e. the features used to represent news articles. We compare two types of textual features: (i) morphosyntactic features, and (ii) lexicosemantic features (or "entities"). The latter correspond to the epidemiological entities described in Chapter 3, Section 1. Feature selection is essential for information retrieval tasks in animal health event-based systems, as well as in other fields such as genomics (Nadkarni, 2002). In this section, we thus provide a generic methodological framework for feature selection, partly relying on the specificity of the animal epidemiology domain. We use the fusion method to evaluate the importance of the different types of lexicosemantic features.

1 Proposed approach

In our approach, we used the vector-space model to represent the documents and compute their similarity. We further describe and illustrate the features used for the morphosyntactic and the lexicosemantic representations.

1.1 Morpho-syntactic features

1.1.1 Preprocessing and feature selection

The morphosyntactic representation is based on the bag-of-words (BOW) representation, i.e. the terms present in all documents in the corpus, as detailed in Section 2.2.1. The text is first segmented into words (tokenisation). Prior to feature selection and representation, several steps known as "text preprocessing" or "text normalisation" can be applied to remove noisy elements from the vocabulary (Figure 29).

Common preprocessing steps include the removal of stop-words, part-of-speech (POS) tagging, normalization to lowercase and lemmatisation (Uysal and Gunal, 2014; HaCohen-Kerner et al., 2020). Stop-words are frequently used terms that are not dependent on a particular topic, such as conjunctions, prepositions or articles. They are usually assumed to be irrelevant and removed from the vocabulary. POS tagging allows users to tag and select the words according to their syntactic functions, i.e. verbs, nouns and proper nouns (Chua, 2008). Lemmatisation consists of transforming the different inflected forms of a word into its canonical form (e.g. singular form for nouns), so they can be analysed as a single item. Lowercase conversion is another widely used preprocessing step. Both lowercasing and lemmatisation aim at computing the occurrence of derived forms of a term (which are semantically similar) as a single item. In this work, all of the preprocessing steps were conducted using the NLTK python library (Bird et al., 2009). We used the list of 318 English stop words provided by the library. The above-mentioned methods are illustrated in Table 22.

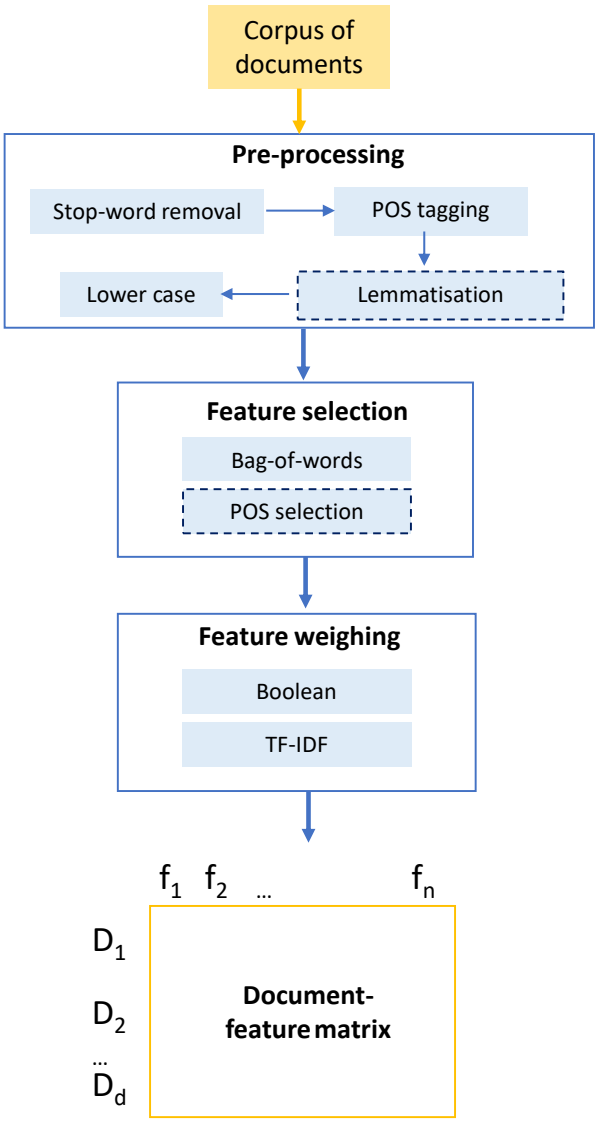


Figure 29 Steps to create the document-features matrix based on morphosyntactic features. Dashed steps corresponds to optional preprocessing steps.

Table 22 Document pre-processed methods evaluated. Bag-of-words (BOW), stop-words removal (SW), BOW lemmatized (BOWlem), part-of-speech selection (POS): verbs (V), nouns (N), proper nouns (PN)

Pre-processing method	Description	Transformed text
No transformation	—	*Bluetongue cases were declared in France*
BOW	Breaking the text into words (tokenization)	*Bluetongue, cases, were, declared, in, France*
BOW, SW	Removing stop-words	*Bluetongue, cases, declared, France*
BOWlem	Transforming words in their canonical form	*Bluetongue, case, be, declare, France*
POS selection:		
Verbs (V)	Selecting only the verbs	*were, declared*
Nouns (N)	Selecting only the common nouns	*cases*
Proper nouns (PN)	Selecting only the proper nouns	*Bluetongue, France*

There is no *gold standard* procedure for preprocessing steps which should be applied to a text corpus. The choice of including or not a normalization depends on the task and the corpus nature (e.g. its language), and different prepreprocessing combinations should be tested (Uysal and Gunal, 2014).

1.1.2 Feature weighting

We used the boolean and $TF - IDF$ weights defined in Chapter 2 Section 2.2.1. Other term weighting methods have been proposed and successfully applied. For instance, OKAPI measures take the document length into account (Robertson and Jones, 1976). Our corpus is homogeneous in terms of length, which justified the choice of $TF - IDF$ weighting.

1.2 Lexicosemantic features

In the lexicosemantic approach, features used to represent the text are selected according to their lexical type. In our context, this includes four epidemiological types of lexical features, or "entities": diseases, hosts, dates and locations. We used the fusion method to evaluate the importance of each type of entity with regard to our task (i.e. linking of related documents). Fusion methods were initially used in multimedia analysis to address the problem of combining multimodal data (i.e. data of different types). For instance, fusion methods are used to combine textual and visual data features to improve multimedia retrieval (Clinchant et al., 2011). Soriano-Morales (2018) describes how these methods can be applied to the fusion of textual features at different linguistic levels (e.g. lexical, syntactic, semantic). Four types of fusion methods are described in the literature, but in this thesis we focus on the two most commonly used methods, i.e. early fusion and late fusion (Figure 30).

As input, both early fusion and late fusion take two unimodal matrices in which rows represent elements (e.g. pictures, texts) and columns represent features. Early fusion involves preliminary combination of the input matrices into a single multimodal representation (combining, for instance, textual and visual features). This combination step consists of column-wise concatenation of the unimodal matrices, thus creating a new matrix with the same number of rows but with an increased number of columns. The second step ("learning step"), consists of computing the similarity between the elements. It can be simple and just involve similarity matrix calculation, or otherwise involve more sophisticated approaches such as the manual attribution of scores by experts (Clinchant et al., 2011) or the use of machine learning methods (Seeland et al., 2017).

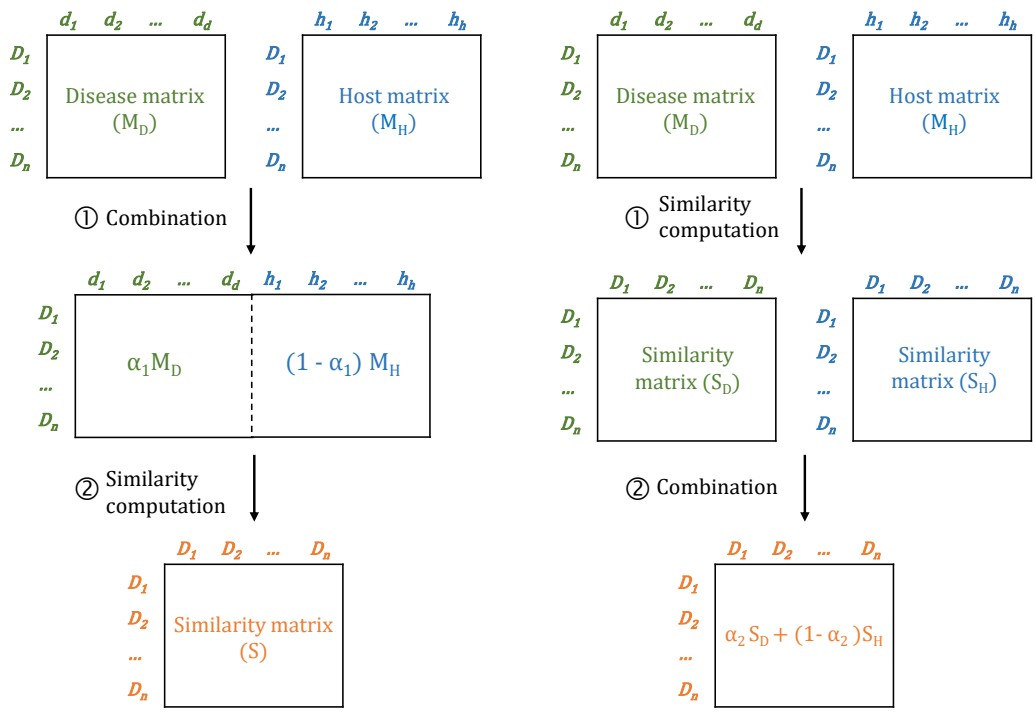
In late fusion, the similarity computation is first performed on each unimodal matrix separately. Then the output matrices are combined into a single representation. For both approaches, a weight can be applied at the combination step to control the relative importance of each modality. With early fusion, one single matrix goes through the learning step, which reduces the computing time and leverages the correlation between the concatenated features. However, the representation space is increased and it may be difficult to combine features into a common representation

(Snoek et al., 2005). In late fusion, features are combined at the same representation level (after the decision step). As the decision step is performed on two matrices, it can increase the computing time and result in a loss of correlation (Soriano-Morales, 2018).

For each type of epidemiological feature (i.e. disease, host, date and location), we first converted the corpus into a document-term matrix in which the rows represent the documents (i.e. news articles) and the columns represent the distinct values of each type of feature. We used either boolean or $TF - IDF$ values as term weights (Equations 2 and 5). We fused the spatiotemporal features (i.e. locations and dates), in addition to the thematic features (i.e. disease and host), and further combined all the features.

1.2.1 Fusion of thematic and spatiotemporal features

To fuse thematic and spatiotemporal features, we applied early and late fusion method as illustrated in Figure 30.



(a) Early fusion **(b) Late fusion**

Figure 30 Fusion of the disease matrix (M_D) and host matrix (M_H) representing n documents D, with d and h features, respectively.
S, S_D, and S_H are the similarity matrices. α_1 and α_2 are the weights to take into account in linear combinations.

We used weighted concatenation as a combination step. With M_D, M_H, M_S and M_T respectively being the disease, host, spatial and temporal feature matrix, and with sim being the cosine similarity function, the fused disease-host matrix (M_{DH}) and the spatiotemporal fused matrix

(M_{ST}) obtained by early fusion are defined by Equations 17 and 18:

$$M_{DH} = sim(\alpha_1 \times M_D + (1 - \alpha_1) \times M_H) \tag{17}$$

$$M_{ST} = sim(\alpha_2 \times M_S + (1 - \alpha_2) \times M_T) \tag{18}$$

For each fusion, α_1 and α_2 ranged from 0 to 1 with a step of 0.1.

1.2.2 Fusion of all features

This step consisted of combining the fused matrices from the previous step into a final matrix (M_F). We selected the best M_{DH} and M_{ST} on the basis of their F-measure (Section 2.2), and concatenated them by using a weight β ranging from 0 to 1 with a step of 0.1 (Equation 19):

$$M_F = \beta \times M_{DH} + (1 - \beta) \times M_{ST} \tag{19}$$

1.2.3 Feature generalisation

The lexicosemantic representation does not take the similarity between related features into account, i.e. the terms "pig" and "boar" are considered as different as "pig" and "bird". To overcome this shortcoming, we evaluated the influence of converting the features into lower granular features. We defined one generalisation level for thematic features (i.e. disease and host), and two granularity levels for spatiotemporal features (Table 23).

For disease and host entities, generalisation aims to take the synonyms used to refer to the same disease or host into account. To convert each feature into its generalized form, we manually built a dictionary with each disease and host variant being mapped to its canonical form and species, respectively. For spatiotemporal features, we aimed to account for the fact that several news articles describing the same outbreak may have different levels of detail when describing the location and occurrence date. We used the GeoNames gazeteer to transform each spatial feature into its first administrative level (level 1) or into its country (level 2), which corresponds to the same methodology used in Chapter 3 Section 2.1.2. For temporal features, we used the normalized values given by HeidelTime to transform features into their week number (level 1) or month (level 2) (Strotgen and Gertz, 2010). The generalisation step can only decrease the granularity of features having higher granularity than the chosen thresholds. Features having a lower granularity were not modified. For instance, the names of continent such as "Africa" were not transformed.

The generalisation reduced the number of distinct features in each category (Table 24), especially for spatial and temporal features whose vocabulary decreased by 83% and 73%, respectively, from level 0 to level 2.

We further describe the corpus which supported the experiments and the evaluation protocol.

Table 23 Generalisation levels of the different types of entities in the event corpus.

Type of feature	Level	Description	Example
Disease	1	Canonical name	– *ASF⟶African swine fever*
Host	1	Species name	– *boars⟶pig* – *ewe⟶sheep*
Location	1	Administrative	– *Toulouse⟶Occitanie*
	2	Country	– *Toulouse⟶France*
Date	1	Week	– *14-02-2015⟶2015-W07*
	2	Month	– *14-02-2015⟶2015-02*

Table 24 Number of distinct features of each type according to the level of generalisation in the Event Corpus. The evolution compared to the number of features at level 0 is indicated between parenthesis.

	Level 0	Level 1	Level 2
Disease	55	29 (-47%)	29 (-47%)
Host	82	36 (-56%)	36 (-56%)
Spatial	761	591 (-22%)	127 (-83%)
Temporal	561	272 (-52%)	152 (-73%)

2 Corpus and evaluation

2.1 Corpus

We used the *event corpus* described in Chapter 3, Section 2.2.1. In this corpus (438 news articles), 229 news articles are labelled with one or more events (the remaining news articles were about general disease information, economic impacts of outbreaks, etc.). Among the news articles labelled with an event, 52% (n= 127/229) reported several events, with a median number of 3 events per news article (Table 15). We consider two (or more) news articles as related if they reported at least one event in common. We created sets of related news articles by linking each document with those having at least one event (i.e. an event identifier) in common. A set consists of a document D_i and its related documents $\{D_k\}$, where k in [1,229] and $i \neq k$. We obtained 157 sets of related documents.

Contrary to the event extraction evaluation (Chapter 3), we used the entire *event corpus* (438 news articles) for the evaluation of related news retrievals. News articles which did not contain any event generate data noise.

2.2 Evaluation protocol

The comparison protocol of morphosyntactic (Section 1.1) and lexicosemantic representation (section 1.2) involved three steps:

1. Representation of each document in a vector of features for each representation,
2. Calculation of pair-wise document vector similarity (similarity computation),
3. Evaluation of the ranking quality.

2.2.1 Similarity computation

As for weights, several measures can be used to compute the similarity between two documents (Huang, 2008), e.g. Euclidian distance, cosine similarity, or the Jaccard coefficient. In our work, we used cosine similarity, calculated as follows (Turney and Pantel, 2010):

$$sim(D_a, D_b) = \frac{\sum_{i=1}^{V} w_{ia} \times w_{ib}}{\sqrt{\sum_{i=1}^{V} w_{ia}^2 \times w_{ib}^2}} \tag{20}$$

where w_{ia} is the weight of term i in document D_a,
w_{ib} is the weight of term i in document D_b,
and V is the total number of terms.

2.2.2 Ranking evaluation

We evaluated matrices obtained by as ranking functions, i.e. regarding their ability to give higher similarity scores to relevant elements than to irrelevant ones. In our process, the related documents of each set D_k were sorted in decreasing order of their similarity values according to the different ranking functions. Figure 31 shows an example of two rankings from a fictive corpus containing 7 documents D_i, where

$$i \in [1:7]$$

and D_1, D_2, D_6, D_7 are related. The displayed rankings are obtained by sorting the documents by decreasing similarity scores with D_1.

Cases of equality in the similarity values could lead to randomly ranked lists (a relevant element may artificially have a higher rank than an irrelevant one, even if they have the same similarity score). We opted to assign the lowest rank to the relevant elements in such cases. We thus favored ranking functions that were able to discriminate relevant from irrelevant elements.

We evaluated the ranking quality in terms of R_{norm}, P_{norm} and F_{norm}, as described in Chapter 3, Section 2.2.3. We also evaluated the R-precision ($R - pre$), corresponding to the precision after R documents have been retrieved, where R is the number of relevant documents for a set (Buckley and Voorhees, 2017). This is calculated as follows:

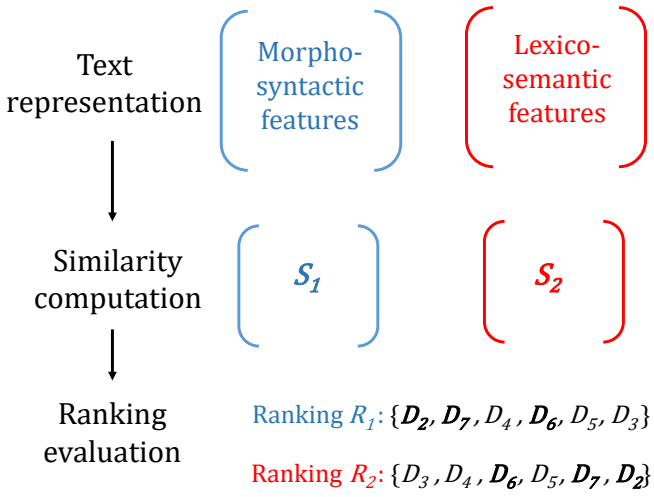


Figure 31 Approaches evaluated for information retrieval. R_1 and R_2 represent the two rankings obtained by the morphosyntactic and lexicosemantic representations, respectively, relative to a document D_1, by decreasing order of similarity. Documents related to D_1 are shown in bold.

$$R_{pre} = \frac{1}{R} \times \sum_{i=1}^{R} x_i \tag{21}$$

where

$$x_i = \begin{cases} 1 & \text{if the } i^{th} \text{ element is relevant} \\ 0 & \text{if the } i^{th} \text{ element is irrelevant} \end{cases} \tag{22}$$

R_{norm}, P_{norm} and F_{norm} evaluate the ranking in terms of the whole set of documents. R-precision considers only a single precision point for each set, thus it is a more stringent measure. Hereafter, we used the first three indicators to evaluate the global ranking, and the R-precision to evaluate the local ranking. For each evaluated model, we calculated the average performances of these 4 measures over the 157 sets of related documents.

3 Results

3.1 Morphosyntactic features

Table 25 shows the ranking performance obtained for different baseline representations. The vocabulary length indicates the number of distinct features used to represent the document (number of columns in the document-term matrix). Among all of the evaluated models, the normalized recall (R_{norm}) was better than the normalized precision (P_{norm}), i.e. reaching up to 0.98. The BOW representation with stopword removal outperformed the other types of representations (F_{norm} =

0.89, $R - pre$ = 0.44), with a vocabulary length of 14,996. In comparison, the lemmatized BOW with the selection of proper nouns obtained very close results (F_{norm} = 0.87, $R - pre$ = 0.42), with a representation space of 4,185 features. The lowest performances were obtained by lemmatization and verb selection only (F_{norm} = 0.53, $R - pre$ = 0.09), which corresponds to the lowest space dimensionality (vocabulary length of 2,151).

Table 25 Ranking performances of morphosyntactic features.
Bag-of-words (BOW), stop-words removal (SW), BOW lemmatized (BOWlem), part-of-speech selection (POS): verbs (V), nouns (N), proper nouns (PN)

	R_{norm}	P_{norm}	F_{norm}	R-pre	Vocabulary length
BOW	0.96	0.77	0.85	0.39	15,278
BOW (SW)	**0.98**	**0.82**	0.89	**0.45**	14,996
BOWlem (SW)	**0.98**	**0.82**	**0.89**	0.44	13,794
BOWlem (SW, V)	0.74	0.41	0.53	0.09	2,151
BOWlem (SW, N)	0.89	0.61	0.72	0.25	4,185
BOWlem (SW, PN)	0.96	0.80	0.87	0.42	5,129

3.2 Lexicosemantic features

3.2.1 Fusion of thematic and spatiotemporal features

First we evaluated the ranking performances of the output matrices from Step 1, without applying any feature generalisation. Feature fusion improved the ranking of both disease-host and spatiotemporal features (Figures 32 and 33). For the disease-host fusion, the lowest F_{norm} values were obtained with the unimodal matrices (corresponding to α_1 = 0 and α_1 = 1 in Equation 17).

For spatiotemporal fusion, the lowest F_{norm} values were obtained with the temporal matrix alone (α_1 = 0), while the spatial matrix alone (α_1 = 1) performed better than fused matrices, with α_2 < 0.6. In both fusion models, the $TF - IDF$ representation outperformed the boolean representation. Early and late fusion obtained very close results, especially when the F_{norm} values peaked. The disease-host fusion performance slightly differed with α_1, ranging from 0.2 to 0.8. For spatiotemporal fusion, the performances significantly increased when α_2 ranged from 0 to 0.5.

Tables 26 and 27 show the best results obtained by the different studied models (based on the best F_{norm} values) for the disease-host and spatiotemporal fusion. The best disease-host fusion models were obtained with α_1 = 0.4 (level 0) and α_1 = 0.7 (level 1). The best disease-host fusion models were obtained with α_2 = 0.7 (level 0), α_2 = 0.8 (level 1) and α_2 = 0.7 (level 2).

As in the baseline approaches, the normalized recall was better than the normalized precision among all of the studied models. Disease-host fusion obtained a maximal F_{norm} of 0.77 and a poor R-precision value (0.28). Spatiotemporal fusion performed better than disease-host fusion, especially regarding the R-precision results (F_{norm} = 0.85 and $R - pre$ = 0.47).

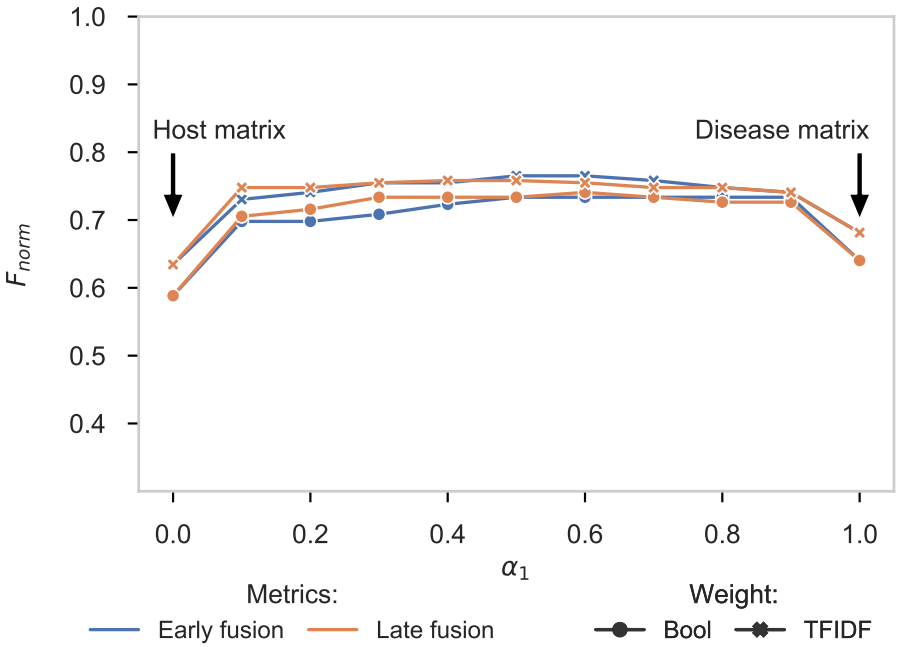


Figure 32 Comparison of fusion methods to combine disease and host features. The arrows correspond to the use of unimodal matrices.

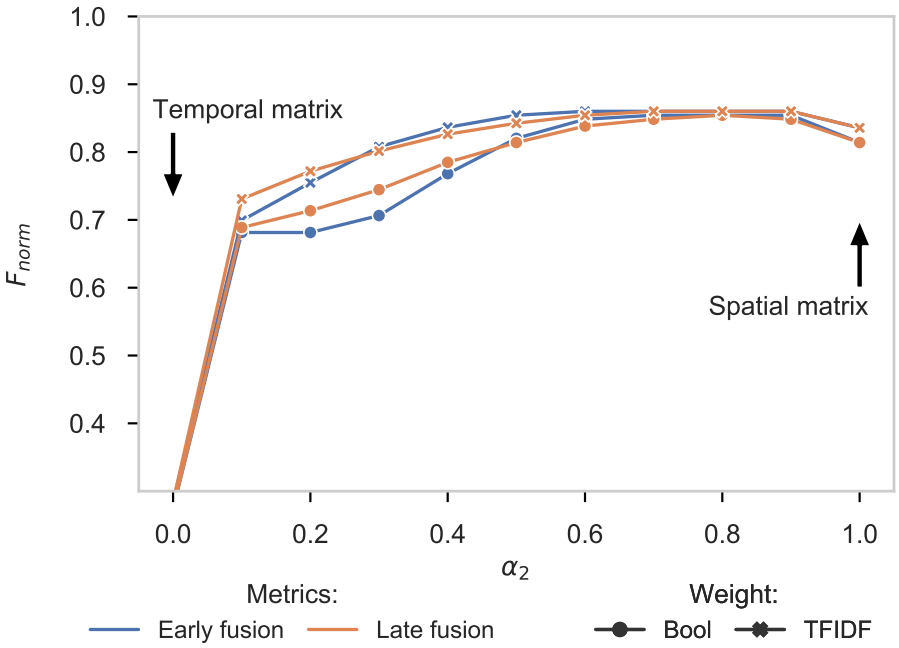


Figure 33 Comparison of fusion methods to combine spatial and temporal features. The arrows correspond to the use of unimodal matrices.

Feature generalisation had different impacts on the performance metrics. In the $TF-IDF$ models, the normalized recall and precision increased when applying the generalisation steps. When applied to boolean representations, it decreased R_{norm} for the first and second disease-host and spatiotemporal fusion levels, respectively. For disease-host fusion, generalisation decreased all of the R-precision values, except for $TF-IDF$ early fusion. For spatiotemporal features,

generalisation at level 1 (administrative level) increased all the R-precision values (up to 0.47). However, generalisation of the country (level 2) decreased these values sharply (from 0.47 to 0.28 for boolean early fusion).

Table 26 Ranking performances of disease and hosts features, using different types of fusion and term representations.
level 0: no generalisation, level 1: first generalisation level.

	Early fusion				Late fusion			
	R_{norm}	P_{norm}	F_{norm}	$R-pre$	R_{norm}	P_{norm}	F_{norm}	$R-pre$
Boolean								
M_{DH} (level 0)	0.91	0.60	0.72	0.23	0.91	0.60	0.72	0.23
M_{DH} (level 1)	0.92	0.59	0.72	0.19	0.91	0.58	0.71	0.19
$TF-IDF$								
M_{DH} (level 0)	0.91	0.63	0.74	0.27	0.92	0.63	0.75	0.27
M_{DH} (level 1)	**0.93**	**0.65**	**0.77**	**0.28**	0.93	0.63	0.75	0.25

Table 27 Ranking performances of spatial and temporal features, using different types of fusion and term representations.
level 0: no generalisation, level 1: first generalisation level, level 2: second generalisation level

	Early fusion				Late fusion			
	R_{norm}	P_{norm}	F_{norm}	$R-pre$	R_{norm}	P_{norm}	F_{norm}	$R-pre$
Boolean								
M_{ST} (level 0)	0.88	0.75	0.81	0.44	0.88	0.74	0.80	0.42
M_{ST} (level 1)	0.89	0.77	0.83	**0.47**	0.89	0.76	0.82	0.46
M_{ST} (level 2)	0.91	0.70	0.79	0.28	0.92	0.73	0.81	0.32
$TF-IDF$								
M_{ST} (level 0)	0.89	0.76	0.82	0.46	0.89	0.76	0.82	0.46
M_{ST} (level 1)	0.89	0.77	0.83	**0.47**	0.89	0.77	0.83	0.46
M_{ST} (level 2)	**0.93**	0.77	0.84	0.38	**0.93**	**0.79**	**0.85**	0.43

3.2.2 Fusion of all features

The second fusion step improved the global ranking of the bimodal matrices (from Step 1), as illustrated in Figure 34. For each distinct β value, the model based on the highest generalisation level (level 1 for M_{DH} and level 2 for M_{ST}) outperformed the models with lower generalisation levels. In the three models, the highest F_{norm} were obtained when more weight was given to the spatiotemporal matrix (β ranging from 0.5 to 0.8).

Table 28 compares the performances obtained with the best fusion and baseline models. For the fusion models, the vocabulary length corresponds to the number of features used, i.e. the sum of vocabulary lengths of each type of feature. The best final fusion model (M_{DH} (level 1) + M_{ST} (level 2)$_{max}$) slightly outperformed the best F_{norm} obtained with baseline models (from 0.89 to 0.92), and improved the R-precision (from 0.45 to 0.58). Compared to baseline models, the

representation space was reduced to 344 features (sum of all distinct disease-host and thematic feature values).

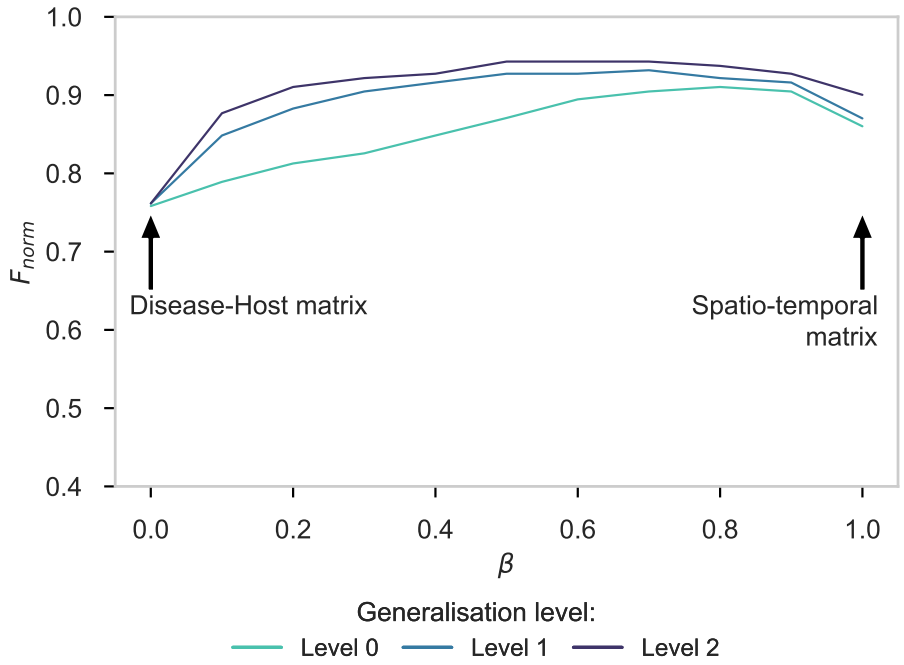


Figure 34 Performance of all lexicosemantic features to retrieve relevant documents, according to varying fusion weight β.
The arrows correspond to the bimodal matrices (thematic matrix M_{DH} and spatiotemporal matrix M_{ST}).

Ranking performance and vocabulary length of morphosyntactic and lexicosemantic representations, and the best fused models at step 1 (disease-host and spatiotemporal fusion) and step 2 (fusion of all features)

Table 28 Ranking performance and vocabulary length of morphosyntactic and lexicosemantic representations, and the best fused models at step 1 (disease-host and spatiotemporal fusion) and step 2 (fusion of all features).
Bag-of-words (BOW), stop-words removal (SW), BOW lemmatized (BOWlem),
part-of-speech selection (POS): verbs (V), common nouns (CN), proper nouns (PN)

Morpho-syntactic features	R_{norm}	P_{norm}	F_{norm}	$R-pre$	**Vocabulary length**
BOW (SW)	**0.98**	0.82	0.89	0.45	14,996
BOWlem (SW, V)	0.74	0.41	0.53	0.09	2,151
Lexico-semantic features (bimodal):					
M_{DHmax}	0.93	0.65	0.77	0.28	65
M_{STmax}	0.93	0.79	0.85	0.43	279
Lexico-semantic features (all):					
(M_{DH} (gen0)+ M_{ST} (gen0))$_{max}$, β=0.8	0.97	0.83	0.89	0.54	1,459
(M_{DH} (gen1) + M_{ST} (gen1))$_{max}$, β=0.6	0.97	0.85	0.91	**0.58**	928
(M_{DH} (gen1) + M_{ST} (gen2))$_{max}$, β=0.6	**0.98**	**0.87**	**0.92**	**0.58**	344

4 Discussion

In this section, we evaluated the performance of morphosyntactic and lexicosemantic representations for the retrieval of related documents. In both representations, the $TF - IDF$ term weights outperformed the boolean ones. Regarding the lexicosemantic features, the spatial features were most efficient when considered separately. Besides, the best retrieval results were achieved when more weight was given to the spatiotemporal matrix, in the bimodal representation as well as in the final fusion. These results were consistent with the task performed—the retrieval of related documents corresponded to the retrieval of the related events contained in the documents. These events were characterized by their thematic attributes (disease and host), but were truly identified by their spatiotemporal attributes. Contrary to the spatial features, the temporal features obtained poor performances when used alone. News articles often refer to the date of official notification of an event, but the true occurrence date may be unknown. Hence, several systems such as HealthMap only use the publication date as a proxy for the occurrence date rather than extracting temporal information from the news article content.

Our final best model combined both early fusion matrix and a late fusion matrix. However, during our experiments, we did not find any clear trend favouring one representation over another. We believe that the impact of the different types of fusion may be reduced when combining features of the same type (here, textual). Thus, from a practical viewpoint, it would be reasonable to chose the early fusion method due to its reduced computational time.

The ranking performances were significantly impacted by the weights used. Rather than specific weights, which could obtain comparable results, we provide ranges of recommended values: $0.2 \leq \alpha_1 \leq 0.8$, $0.6 \leq \alpha_2 \leq 0.8$ and $0.5 \leq \beta \leq 0.8$.

Feature generalisation allowed us to increase the global model performances (in terms of F_{norm}) while reducing the representation space. However, increasing generalisation at the country level for spatial features reduced the local precision. Thus, a balance has to be found between the global ranking and the precision over a set of retrieved documents. This balance depends on the user's needs, as well as the size of the corpus—if plenty of news articles contain events from the same country, mapping each entity with its country would certainly decrease the retrieval performance. Our generalisation approach is similar to an ontological approach—we used predefined conceptual structures to enrich the documents with "meta" entities (e.g. a country instead of a city name). The core of the BioCaster system relies on a complex ontology which maps each named entity to a canonical form (Collier et al., 2008). However, the event extraction performances with and without the use of the ontology were not compared in that study.

Regarding the morphosyntactic features, all lemmatized representations which included the proper nouns gave very good results, i.e. outperforming the other representations. We assumed that the vocabulary in terms of verbs and common nouns was homogeneous between news articles reporting outbreaks (e.g. "reported", "declared", "cases", etc.). Including these features to compare the epidemiological content would not be very informative. Conversely, as proper names include

locations and disease names, they contain much accurate and rich information. This is consistent with previous results which showed that the spatiotemporal matrices should be assigned a higher weight. Moreover, proper names include other types of named entities, such as organization names, which we did not take into account in the fusion models. Such types of features could be of interest to link two related news articles when, for instance, referring to a local official source. The best baseline and fusion representations obtained comparable performances in terms of global ranking (F_{norm}). However, the fusion approach has the advantage of being based on a reduced number of features, thus limiting the computing time when dealing with larger corpora. Besides, the fused models significantly increased the local precision (R-precision) compared to the baseline approach.

Chapter 5

Integration of event-based surveillance systems into epidemic intelligence activities – Case studies

Table of contents

As introduced in Chapter 1 Section 3, the detection of an event by EBS systems depends on: (i) the detection and reporting of the event by an informal source (e.g. an online news outlet), and (ii) the extraction of the event information from the source content. In the previous chapters of this thesis, we focused on the second point: evaluate statistical and NLP methods to retrieve and combine information from free texts. Our work highlighted several gaps. First, little is known about how disease outbreak information circulates between sources before being detected by EBS systems. Knowledge on which sources communicate and disseminate disease information is crucial for understanding their role regarding EBS system performance. Second, we only evaluated our approaches through the lens of specific animal diseases surveillance, so EBS system performance regarding the emergence of unknown threats has not been evaluated. We believe that these two aspects are crucial for enhancing the mainstreaming of EBS systems into the global epidemic intelligence process. In this Chapter, we provide preliminary results regarding:

1. The understanding of the dissemination of sanitary information through sources used by EBS systems;
2. The reaction of EBS systems to the COVID-19 outbreak.

1 Dissemination of information in event-based surveillance

In this Section, we assess how sanitary information (i.e. events) propagates between sources before being detected by an EBS system. This study was done in collaboration with the HealthMap project, which kindly provided us the ProMED and HealthMap data. Both PADI-web and HealthMap data were used for this analysis. More precisely, we aimed to answer two main questions:

- How does outbreak-related information propagate between sources?
- What is the role of sources regarding the detection and propagation of outbreak-related information?

To address these questions, we assessed an approach involving the following steps:

1. We manually extracted the structure regarding the dissemination of information through sources, relying on the source citations in disease-related online news;
2. We formalised this dissemination using network theory, i.e. a well-established approach for analysing the dissemination of content through online sources (Weber and Monge, 2011).
3. We described the network based on qualitative and quantitative attributes of its components to highlight the value of sources in information dissemination.

1.1 Methods

1.1.1 Network construction

To extract and analyse information dissemination and the roles of sources, we first collected the report data, stored them in a database, and used the latter to generate the network.

Data collection

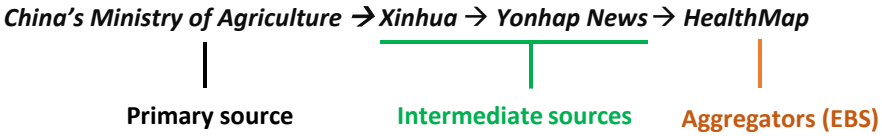
We extracted all English reports related to African swine fever (ASF) published between 1 August 2018 and 31 July 2019 from PADI-web and HealthMap databases, and containing one or several events (i.e. unverified sets of epidemiological information for an ASF case or outbreak). PADI-web relevant reports consisted of news classified as relevant (Appendix B). HealthMap reports included reports from different types of informal sources, e.g. online news, ProMED, Twitter, etc. We obtained 136 ASF-related reports from HealthMap and 594 ASF-related reports from PADI-web (total of 730 final reports).

Database creation

At this step, the objective was to trace back the origin of the event information. We assumed that this pathway could be deducted from the sources cited in the final reports. For instance, based on a simplified report retrieved by the HealthMap system:

> ***Yonhap News***. According to ***Xinhua***, ***China's Ministry of Agriculture*** on Friday said it has detected more outbreaks of African swine fever.

The information pathway is:



The arrows indicate the direction of the information (here, African swine fever outbreaks) flow. As shown hereabove, we distinguish:

- Primary sources, i.e. the earliest emitter source for a given event;
- Intermediate sources, i.e. all sources involved in a path, except the primary source and final aggregator;
- Aggregators, i.e. the EBS system used to retrieve the final reports (i.e. HealthMap and PADI-web).

When a source was cited with a hyperlink, we followed the link to follow the information pathway as far as possible.

Each source was characterised by its type, as defined in Table 29, and its publication date.

We retrieved the information pathways of all events present in the reports, representing a total of 359 events. All pathways and events were stored in a specially designed database. In the database, each path was segmented in pairs of emitter SE and receptor sources SR. Thus, the previous example is represented as:

Emitter source SE	Receptor source SR
China's Ministry of Agriculture	Xinhua
Xinhua	Yonhap News
Yonhap News	HealthMap

Network construction

We use a graph structure to represent the desired network. A graph is composed of nodes and edges. Formally, a **graph** $G = (V, E)$ is a mathematical structure consisting of a set V of **vertices**

Table 29 Definition of the types of sources.

Type of source	Definition	Example
International veterinary authority	International source for animal disease notification	FAO, OIE
National veterinary authority	National source for animal disease notification	French Ministry of Agriculture
Local veterinary authority	Local source for animal disease notification (typically, veterinary authority at a province-level in large countries such as China or USA)	Department of Health of Gansu province
Control authority	National control authority, such as customs	General Administration of Customs
Public organisation	Public organisms with a veterinary scope, such as laboratories	
Private company	Private companies, typically animal industry companies	Canadian Pork Council
Online news	Online news	PigProgess
Press agency	Agency that collects news reports for newspapers and distributes it electronically.	AFP, Reuters
Radio, television	Radio and television channels	Bulgarian National Radio
Social platform	Blogs and social media	FluTrackers.com, Twitter
A person	A person cited as a source, who cannot be related to a source of this list	A farmer
EBS system	EBS system	HealthMap, ProMED, PADI-web or other

(also commonly called **nodes**) and a set E of **edges** (also commonly called **links**) (Kolaczyk, 2009). The network nodes represent the sources and final aggregators (HealthMap and PADI-web). The edges represent the transmission of event information between two nodes (an emitter, which sends the event, and a receptor, which receives the event). We thus created an edge between an emitter SE and receptor sources S_R if S_R cited S_E at least once. The first node of a path is called the primary node; the last node is called the final node. The primary and final nodes can be separated by intermediate nodes. The combined edges from a primary to a final node correspond to a path.

Figure 35 represents the construction of a subset of the graph based on the previous example. The graph is directed, as the information is transmitted from an emitter source S_E to receptor sources S_R. A directed graph is formally defined as a graph G with each edge in E having an ordering to its vertices (i.e. such that $e1 = (u, v)$ distinct from $e2 = (v, u)$, for $e1, e2 \in E$). Directed edges are also called **arcs**. In our approach, the edges are not weighed.

Based on this method, we generated two types of graphs. The **event graph** is specific to an event and contains the path(s) through which the event propagated. The graph shown in Figure 35 is an example of an event graph. There is only one **total graph**, which consists of all paths extracted from the reports. Formally, it was generated as follows:

1. We initialize a graph containing all nodes (i.e. sources extracted during the database creation process), without any edges;

2. For each event P_i: for each pair of nodes (v, u) connected in the event graph of P_i, we created an edge $e = (v, u)$ in the total graph.

Importantly, each node has an identifier specific to its source. This allowed us to further aggregate all metrics obtained by a specific node in all the event graphs.

1.1.2 Network analysis

To describe the network, we relied on qualitative and quantitative attributes of nodes. All nodes Si have the same attributes, as described in the next insert. The node quantitative attributes correspond to centrality measures used to describe the importance of a node in the network (Kolaczyk, 2009). These attributes are essential to understand the role of the sources. Sources with a high in-degree collect information from a broad range of sources. Sources with a high out-degree are often cited sources, thus which are able to communicate information with high visibility. Sources with a high degree combine both capacities (also referred to as "hubs" (Weber and Monge, 2011)).

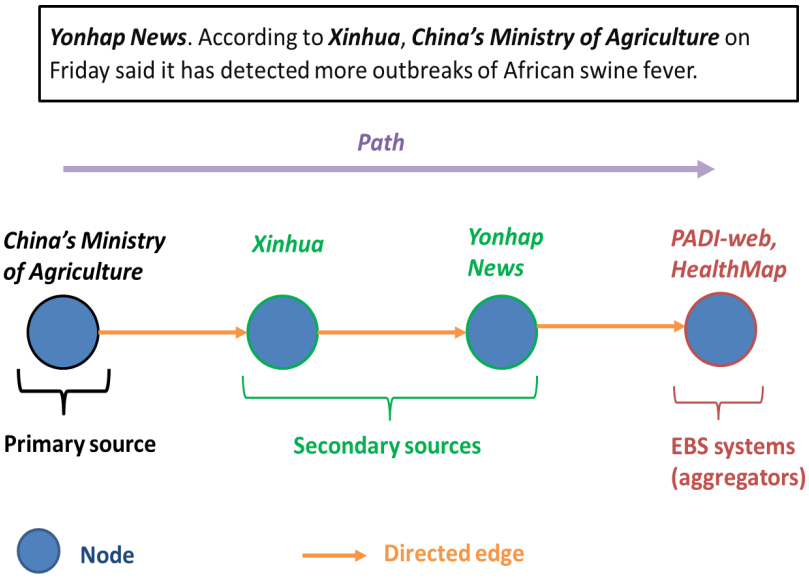


Figure 35 Example of network representation of information dissemination between sources.
The inset (above) contains a simplified extract from a news article containing the citation of the primary and secondary sources.

Node attributes

- Qualitative attribute:
  - Name of the source (i.e. identifier)
  - Type of source
  - Date of publication
- Quantitative attributes:
  - In-degree is the number of incoming edges to a node.
  - Out-degree is the number of outcoming edges from a node. It measures the degree to which the node connects outwards to other nodes;
  - Degree is the sum of in-degree and out-degree;
  - Time lag. Given a node S_R connected with an upstream node S_E and, with respective publication dates $S_E : d$ and $S_R : d$, where d is the time lag attribute of S, the time lag is $S_R : d$ - $S_E : d$, corresponding to the transmission time lag between S_E and S_R. S_R have as many time lag values as upstream nodes connected to S_R, with each one belonging to a specific path.

Path analysis

First, to evaluate the network global performances regarding event dissemination, we calculated the following aggregated measures on the total graph:

- The **path length**, which is the number of edges in the paths. The path length measures the number of intermediate sources between the primary source and the aggregator.
- The **path reactivity**, which is the sum of the time lag of all the nodes composing the paths. The path reactivity measures the number of days between the communication by the primary source and the detection by the aggregator.

Then we compared the primary and secondary nodes of each path based on their type of source. The secondary node corresponds to the node immediately following the primary node, in paths whose length is strictly higher than one edge (i.e. the first intermediate nodes).

Node centrality measures

Node centrality measure (i.e. in-degree, out-degree and degree) were first calculated on each **event graph**. For each node, we calculated the mean and standard deviation of each metric. These values are indicators of the node performance for a given event.

Next, we calculated the total in-degree, total out-degree and total degree of the **total graph**. These metrics measure the centrality degree of nodes based on all events that circulated through

the graph. For instance, the total degree of a node corresponds to the total number of sources from which it has been connected at least once.

The analysis was done using the *igraph* package available in R version 3.6 (Csardi and Nepusz, 2006).

1.2 Preliminary results and discussion

Path analysis

First we analysed the paths (defined as the sequence of edges followed by an event from a primary node to the final node) in terms of length, reactivity and qualitative attribute "type of sources" of the nodes involved, based on the **total graph**.

The **total graph** contains 295 nodes and 477 edges, corresponding to 813 distinct paths. Among the paths, 47.4% (385/813) had a length of two, and 39.1% (318/813) had a length of three (Figure 36). Thus, 86.5% of the paths were transmitted from a primary node to the final node through one or two intermediate nodes. Marginally, 1.7% (14/813) of the paths had a length of one. These paths were extracted from final reports which did not cite any source. The remaining 11.8% (96/813) of the paths had a length of four or more.

The network was highly reactive, with 85.7% (687/813) of the paths propagating events in less than a day (Figure 36).

National veterinary authorities were the primary source of 74% of the paths (601/813) (Figure 37). In the remaining paths (26%, 212/813), the primary source was an online news outlet, a local veterinary authority, a private organisation, a control authority, a press agency, a public organisation, a television channel or radio, or a person. Press agencies and online news outlets are major secondary sources, representing 49% (390/799) and 32% (252/799) of the secondary sources, respectively. International veterinary authorities represent only 7% (55/799) of the secondary nodes.

These results indicated that 74% of the paths of the graph propagated events which were acknowledged at the national level. However, we noted that international veterinary authorities (the main source of official information for international monitoring) were involved as secondary sources in a few paths. Importantly, this result does not reflect the intrinsic value of international veterinary authorities in collecting information from national authorities, as we only took events present in the report detected by the EBS systems into account. However, the result indicates that online news and press agencies detect information directly from primary sources, thus bypassing the international official notification process.

Node centrality measures

We further analysed the nodes in terms of connection with other nodes.

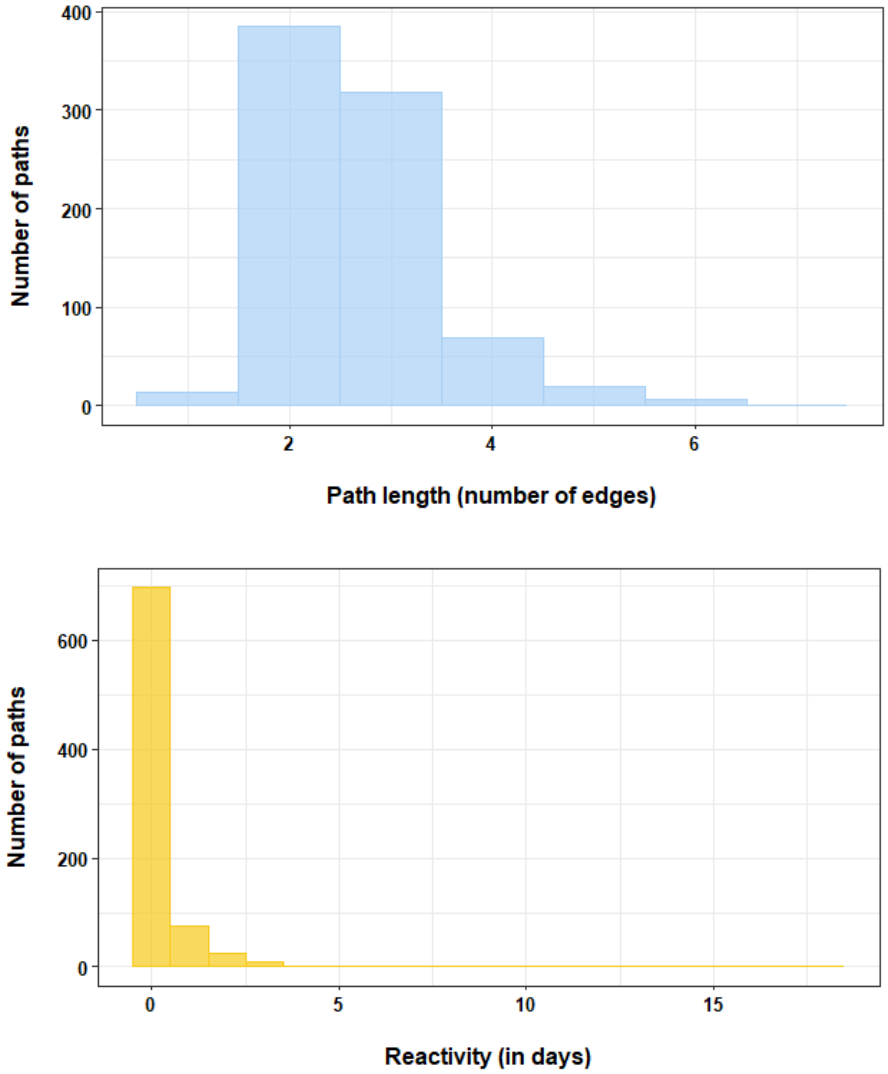


Figure 36 Path length (above) and reactivity (below).

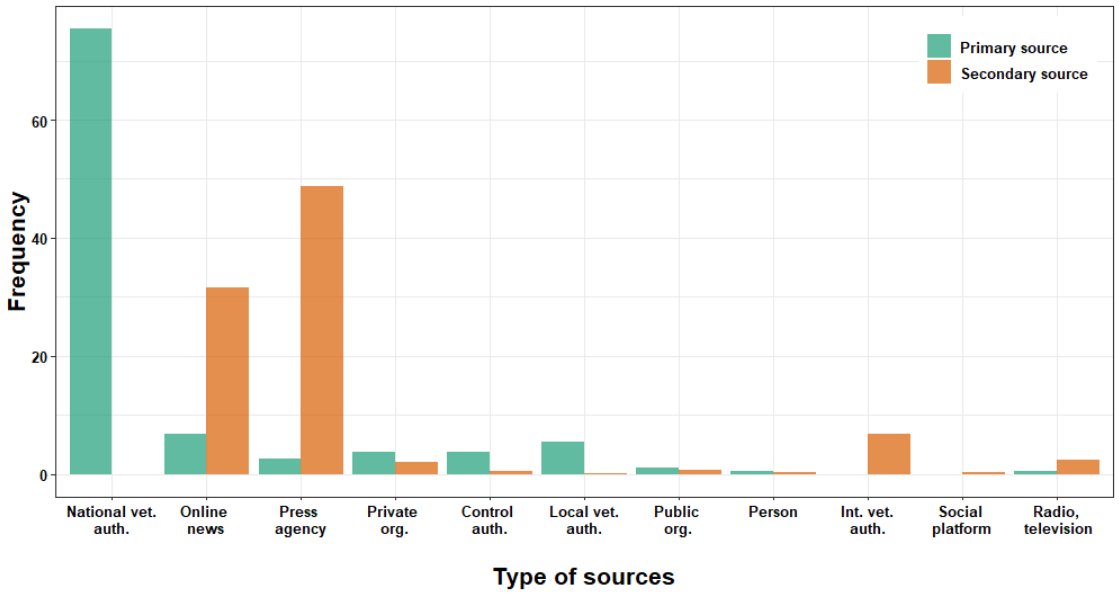


Figure 37 Type of source of the primary and secondary nodes among all the paths of the graph.

The total degree ranged from 1 to 677. The node degree distribution followed a power law, indicating the presence of major hub sources, i.e. highly connected nodes within a network (Berlingerio et al., 2011; Fornito et al., 2016).

Table 30 shows the node degree centrality measures (in-degree, out-degree and degree), aggregated by the type of sources. The first set of metrics corresponds to the average values of each node over the **event graphs**.

Private companies and international health authorities had the highest average degree. National health authorities and control authorities were pure emitter sources, with an in-degree of 0. EBS systems had the highest average in-degree, consistent with the fact that they aggregate a broad range of different types of sources. Hence they are more prone to retrieving information through different edges. Their out-degree was not zero as ProMED was used as a source by HealthMap.

Private companies had the highest out-degree. This result suggests that, for a given event, event information communicated by private companies is highly picked up by other sources. Indeed, sanitary information communicated by a private company typically involves the detection of a case in a slaughterhouse or a feed maker, thus having a potentially important economic impact or posing a food security threat. The analysis of the top 5 sources confirmed this hypothesis (Table 31). For instance, the Tangrensten Group is a major Chinese animal feed maker which reported on 9 November 2018 that the feed produced by one of its units was suspected to be contaminated by the ASF virus. This contamination was further confirmed. The media widely communicated the suspicion and confirmation of this event (22 reports in our corpus), as it raised the fear of ASF virus transmission to pig farms through the contaminated feed. The second set of metrics corresponds to the degree centrality measures of each node in the total graph, also aggregated by

Table 30 Node performances in terms of mean and total in-degree, out-degree and degree, aggregated by type of sources.

	Average per node over all events			Sum per node		
	In-degree	Out-degree	Degree	In-degree	Out-degree	Degree
Online news	1.2 (0.9)	1.3 (0.8)	2.6 (1.5)	1.9 (1.9)	1.5 (1.1)	3.4 (2.5)
National vet auth.	0	2.9 (2.5)	2.9 (2.5)	0.2 (0.5)	5.3 (9.1)	5.6 (9.5)
Radio/TV	1.1 (0.7)	1.1 (0.3)	2.2 (0.9)	1 (0.6)	1.2 (0.4)	2.3 (0.8)
Press agency	1.9 (1.5)	2 (1.5)	4.0 (3)	4.8 (7.0)	6.4 (12.1)	11.2 (18.9)
Local vet auth.	0 (0.2)	1.6 (1.6)	1.6 (1.6)	0.1 (0.2)	1.9 (2.1)	2.0 (2.1)
Private company	2.1 (3.6)	**4.3 (5.5)**	**6.4 (8.4)**	0.5 (0.7)	2.1 (1.9)	2.5 (2.2)
Control auth.	0.9 (0.3)	1.2 (0.6)	2.1 (0.6)	0.2 (0.4)	2.2 (1.6)	2.4 (1.7)
Other health assoc.	0.5 (0.8)	1.5 (0.6)	2.0 (1.3)	1.0 (1.2)	2.2 (1.6)	3.2 (2.7)
Social platform	1.2 (0.7)	1.2 (0.6)	2.4 (1.2)	3.4 (2.3)	1.0 (0)	4.4 (2.3)
Local person	0.5 (0.9)	1.3 (0.5)	1.75 (1.0)	2.3 (0.5)	1.3 (0.5)	1.5 (1.0)
EBS system	**2.7 (2.5)**	0 (0.1)	2.7 (2.5)	**68.0 (80.0)**	0.3 (0.6)	**68.3 (79.6)**
Int health auth.	2.5 (2.4)	2.8 (2.6)	5.3 (5.0)	6.0 (7.1)	**10.0 (12.7)**	16.0 (19.8)

the type of sources.

EBS systems had the highest in-degree values, which was in line with the fact that HealthMap and PADI-web were the aggregators. Excluding EBS systems, international health authorities had the highest degree. This indicates that they were connected with a large number of incoming and outcoming sources. Incoming sources were national veterinary authorities, and outcoming sources were typically online news and press agencies. This result should be tempered by the fact that international authorities were involved in a few paths as a secondary source.

International authorities were cited as a source for only 9% of the events (31/359). We hypothesise that these were major events in terms of epidemiological and/or economic impact, such as the first emergence in a country or large-scale outbreaks (e.g. on large pig farms, feed producers, etc.). In such cases, the event is highly reconveyed by a broad range of sources. Conversely, online news were major secondary sources, but to a much lower degree. Online news sources, including local and regional sources, were more prone to communicate follow-up events such as an additional and local outbreak in an already affected region, thus not necessarily capturing much media

Table 31 Top 5 sources (nodes) with the highest average in-degree, out-degree, and degree for each generated graph in which they are involved.

In-degree	Out-degree	Degree
Shenzhen Stock Exchange platform	Tangrensten Group	Tangrensten Group
015.BY	North Korean Vet Authority	Shenzhen Stock Exchange platform
Belsat.tv	Sanquan Food (Henan Province, China)	North Korean Vet Authority
Flanders News	Northern Ireland Vet Authority	Sanquan Food (Henan Province, China)
Tangrensten Group	Shenzhen Stock Exchange platform	015.BY

interest.

Press agencies had performances close to those obtained for international authorities. Compared to online news, they had access to a broader range of information sources and, by nature, distributed their information to a vast network of sources.

All the metrics had high standard deviations, thus indicating that the groups of sources were not homogeneous in terms of event dissemination, confirmed by the presence of hubs, i.e. sources with high in-degree and out-degree. In Table 32, we show the top 5 of the nodes based on the highest in-degree, out-degree, and degree. We indicate the EBS systems in parenthesis are the final aggregators, thus biasing their in-degree. The top 5 includes two press agencies (Reuters and Xinhua), the OIE international health authority, the Chinese Ministry of Agriculture, and the online news outlet Outbreak News Today. The presence of Xinhua (a Chinese press agency) and the Chinese veterinary authority is consistent with the corpus coverage in terms of events. Indeed, ASF has emerged in China for the first time and propagated through the country during the study period. Outbreak News Today is an online news outlet that provides a summary of all on-going outbreaks. Thus, its behaviour is close to that of press agencies, i.e. aggregating information from a large number of sources.

Table 32 Top 5 sources (nodes) with the highest sum of in-degree, out-degree, and degree over all events. Aggregators are shown in parenthesis.

In-degree	Out-degree	Degree
(PADI-web)	Chinese Vet Authority	(PADI-web)
Reuters	Reuters	Reuters
(HealthMap)	Xinhua	Chinese Vet Authority
Xinhua	Outbreak News Today	Xinhua
Outbreak News Today	OIE	(HealthMap)

1.3 Prospects

In this preliminary research, we show that EBS systems rely on a network of sources able to rapidly communicate event information. Besides, our results show the global performance of some types of sources was driven by a small number of sources that served as hubs, such as the Reuters and Xinhua press agencies. As highlighted by the standard deviations of the centrality measures, the type of source alone was not sufficient to determine the intrinsic value of the source. Thus, other source attributes should be evaluated, e.g. the language and geographical focus. Such attributes were recorded during the corpus annotation, and we aim to include them in further analyses. Besides, other centrality measures can characterise nodes as hubs or authorities, so sources could be compared on this basis (Weber and Monge, 2011; Berlingerio et al., 2011).

A key result was the fact that national health authorities appeared to be a major primary source of events. This shows that a significant number of events detected by the EBS systems were national officials. Press agencies and online news sources (which belong to the informal source category) played a major role in retrieving event information from national authorities and in disseminating it to inform the general public. Our results suggest that in bypassing the international notification process, online news and press agencies are crucial actors for early warning. Importantly, these results do not reflect a total lack of transparency at the international level: 84% of the events extracted from the corpus were officially notified to OIE. However, they suggest that notifications to the international authority were probably delayed, thus explaining why another primary source was instead used by the intermediate sources. Besides, in addition to being crucial links between national authorities and EBS systems, intermediate sources facilitate information transmission by translating into English, for instance, national reports that were initially released in native languages.

In a follow-up to this study, we aim to evaluate another network property, i.e. the ability to disseminate complete and truthful information. We extracted the presence or absence of a set of specific epidemiological features (i.e. host, location, dates, etc.) as well as their trustworthiness (correct or not based on official data). We aim to evaluate the ability of the source to provide and communicate complete and truthful information.

EBS systems had two roles in this analysis. PADI-web and HealthMap, which were used to retrieve the corpus of reports, were final aggregators. ProMED is also an EBS system, but it was not evaluated as an aggregator, so this study did not reflect its intrinsic value. Thanks to an ongoing collaboration, we aim to integrate ProMED data in order to evaluate the system with PADI-web and HealthMap. The evaluation of the source roles individually was biased by the focus of the corpus (African swine fever events). Thus, a second annotation process is currently under way to compare the ASF results with another disease, i.e. avian influenza. More precisely, we aim to determine the similarity and difference in terms of sources involved and network characteristics (e.g. if the same press agencies are involved).

2 Early detection of unknown diseases by event-based surveillance systems

2.1 Context of the study

This thesis has focused on the use of epidemiological entities specific to animal disease events in online news, such as disease name, host, location, etc. We also investigated how semantic and lexical information can be used to retrieve fine-grained information about a disease event. However, the monitoring of unknown diseases, i.e. clusters of cases for which the diagnosis is not known, has not been evaluated in terms of appropriate methods. Unknown diseases might be reported in informal sources in cases of delayed diagnosis of common diseases, if access to laboratory testing is limited, for instance. Less frequently, unknown disease reporting might be an early warning of an emerging disease (Rolland et al., 2020), which monitoring is the mandate of EBS systems.

In the following Sections, we conducted a retrospective analysis to compare three systems, i.e. ProMED, HealthMap and PADI-web regarding the detection of COVID-19 emergence. We investigated how PADI-web, which was designed for monitoring animal disease, detected COVID-19 events. We further analysed the specificity of the vocabulary used in the news (Valentin et al., 2020d).

On 31 December 2019, local health officials of the Chinese city of Wuhan reported a cluster of 27 cases of "pneumonia of unknown cause". These cases were linked to a wholesale live animal and seafood market in the city. The first death was reported in January 2020 and the causative agent was identified as a new coronavirus, i.e. SARS-Cov-2, and the disease was named COVID-19. The first epidemiological study on patients with laboratory-confirmed COVID-19 infection reported the onset of illness as early as 1 December 2019 (Huang et al., 2020).

This retrospective study aimed first to evaluate three EBS systems (ProMED, HealthMap and PADI-web) and their capacity for timely detection of the COVID-19 emergence in China. Secondly, we focused on PADI-web to understand how an animal health EBS system contributed to the detection of a human EID. We analysed the RSS feeds from PADI-web that detected COVID-19-related news articles (hereafter referred to as "news"). Thirdly, we assessed the vocabulary in the news detected by PADI-web and its change in relation to identification of the pathogen and the EID spread.

2.2 Material and methods

2.2.1 COVID-19-related news detection

News from 1 to 31 December 2019 was mined to assess the timeliness of the three EBS. We compared the first news regarding the publication date, language and source.

To gain insight into how PADI-web detected the COVID-19 emergence, we further filtered a second corpus of news published from 31 December 2019 to 26 January 2020 containing at least one of the following words in the title and body of the news: “pneumonia”, “respiratory illness”, “coronavirus”, “nCoV” (an early name for COVID-19), and “Wuhan”. After manual verification of their relevance, we retained 275 out of 333 news items for analysis (Valentin et al., 2020a).

We assessed the link between the detected news items and the animal health RSS feeds from PADI-web that served to retrieve those news items. To this end, we read each news item and categorised it into: i) disease-specific RSS feeds (containing specific disease names), and ii) syndromic RSS feeds (containing combinations of symptoms and animal hosts).

2.2.2 News vocabulary

We analysed the vocabulary change spanning the period from the initial discovery of the COVID-19 outbreak to its spread outside China by extracting terms from the whole corpus. A word frequency-based method was first implemented to highlight important keywords according periods (Figure 38).

Secondly, we used a ranking function based on the frequency and discriminance of terms (i.e. words and multi-word terms) extracted with BioTex, a text-mining tool tailored for biomedical terminology (Lossio-Ventura et al., 2016). BioTex is based on the use of: (i) a relevant combination of information retrieval techniques and statistical methods, and (ii) a list of syntactic structures of terms that have been learnt via relevant sources (e.g. UMLS, MeSH). BioTex-extracted terms can be lowercase words (e.g. influenza), or phrases (e.g. avian influenza). We used the F-TFIDF-C measure as a ranking function (Lossio-Ventura et al., 2014). We further identified terms referring to COVID-19, such as “new virus” and “mystery pneumonia”. We manually categorised the terms as: “mystery” (referring to the unknown threat), “pneumonia” (referring to the clinical signs), “coronavirus” (referring to the virus taxonomy) and “technical” (technical acronyms specifically pertaining to the virus). One news item could contain terms from different categories. We calculated the daily proportion of each category, expressed as the sum of occurrences of the category divided by the total number of occurrences.

2.3 Results and discussion

2.3.1 News detection

ProMED was the first EBS system to detect and report a news item from a Chinese online source . The ProMED report dated back to 30 December 2019—a day before the first official notification of pneumonia-like cases in Wuhan (Wuhan Municipal Health Commision, 2020). PADI-web and HealthMap respectively detected three and one COVID-19-related news items on 31 December 2019—the same day as the first official notification of pneumonia-like cases in Wuhan (one HealthMap news item from an English source, three PADI-web news items from two English sources and one Chinese source). The news detected by the three EBS originated from five different media outlets.

Among the three EBS systems compared, only ProMED relies on local expert information to alert on health threats. This result suggest that the network of local field experts is crucial for the detection of EID events and their reporting. Otherwise, HealthMap and PADI-web detected news on the same day as the official reporting. It is therefore essential to understand their current limitations and promote the key role of experts in EBS systems. Further studies should also focus on assessing whether the timeliness of automated systems depends on the communication strategies of online media, as well as on determining their health event reporting threshold, and how these features impact the sensitivity of EBS systems.

The three EBS systems included in this study monitor media in multiple languages, thus facilitating detection of local media news. A further increase in the number of available languages should enhance the sensitivity of EBS systems (Barboza et al., 2014). Our study also showed that the three EBS systems were complementary regarding scope (animal and public health), moderation (manual, semi-automated, automated), and number of covered languages . PADI-web could retrieve COVID-19 related news through animal-health related RSS feeds, thus proving its usefulness for the detection of information of relevance for public health risk assessors. From 275 COVID-19-related news items retrieved by PADI-web, 54.5% (n=150) were retrieved via syndromic RSS feeds, while the remaining 45.5% (n=125) were retrieved via disease-specific RSS (Table 33).

Content-wise, 31.7% (n=87) of the news items compared COVID-19 to five animal diseases (avian influenza, African swine fever, classical swine fever, West Nile virus, and Rift Valley fever), 24.4% (n=67) of the news items described the broad range of animal species sensible to coronaviruses, 18.2% (n=50) reported that avian influenza was ruled out from possible causes of the outbreak, 7.7% (n=21) described other on-going outbreaks in addition to COVID-19 (avian influenza, African swine fever, classical swine fever, and foot-and-mouth disease), 2.5% (n=7) referred to animal species present in Chinese markets as potential COVID-19 sources, and 0.7% (n=2) advised people to avoid contact with animals. Irrelevant keyword matches were found in 12 news items (e.g. finding a host keyword in the name of a source), and no link could be established between the RSS feed and the article for 29 other news items (10.5%).

The fact that disease-specific RSS feeds contributed as much as syndromic RSS feeds to the

detection of COVID-19 news by PADI-web was unexpected, thus highlighting the importance of combining both disease-specific and syndromic feeds. Many of the news items detected by PADI-web compared the magnitude and economic impact of COVID-19 with regard to avian influenza and African swine fever outbreaks in China. Indeed, prior to COVID-19 identification, the reported pneumonia-like illness was compared to avian influenza zoonotic infections. Some news also presented a summary of several recent disease outbreaks in China, including African swine fever (which is not a zoonotic disease), thus explaining why they were detected by PADI-web.

Table 33 Percentage (%) and number (n) of COVID-19-related news items retrieved by PADI-web from 31 December 2019 to 26 January 2020. Each article is categorized by type of feed (disease-related or syndromic) according to the link between the feed and COVID-19.

Link with COVID-2019	Type of RSS feed		
	Disease-specific	Syndromic	Total
Comparison with another disease	20.4% (n=56)	11.3% (n=31)	31.7% (n=87)
Disease ruled out	17.8% (n=49)	0.4% (n=1)	18.2% (n=50)
Aggregation with other disease outbreaks	6.2% (n=17)	1.5% (n=4)	7.7% (n=21)
Coronaviruses in animals	-	24.4% (n=67)	24.4% (n=67)
Market animals	-	2.5% (n=7)	2.5% (n=7)
Avoid contact with animals	-	0.7% (n=2)	0.7% (n=2)
Irrelevant keyword matches	0.4% (n=1)	4.0% (n=11)	4.4% (n=12)
Unknown	0.7% (n=2)	9.8% (n=27)	10.5% (n=29)
Total	45.5% (n=125)	54.5% (n=150)	100% (275)

The ability of EBS tools to encompass a broad scope of health-related topics through a limited number of queries (RSS feeds) is a major asset compared to formal sources. This capacity largely depends on the intrinsic features of online news in which outbreak-related content is often bulked up with additional information, such as comparisons with previous disease outbreaks, thus increasing the probability of being detected by EBS tools. However, the probability of detection of an EID event might be higher for (actual or assumed) zoonotic diseases and countries with on-going animal disease outbreaks. This is not a major shortcoming in practice.

2.3.2 News vocabulary

From the terms referring to either the virus or the disease, 18 terms were in the “pneumonia” category, eight terms in the “mystery” category, three terms in the “coronavirus” category (one of them, “coronovirus” being a misspelt form of “coronavirus”), and seven terms in the “technical” category (Table 34).

The word clouds generated from the overall news contents mined over three consecutive periods are shown in Figure 38.

Before identification of the virus (31 December 2019 – 8 January 2020), 58.1% (n=317) of the COVID-19 terms were in the “pneumonia” category, 29.1% (n=159) in the “mystery” category and 12.8% (n=70) in the “coronavirus” category. From the official identification of the virus to the

Table 34 Terms used to describe SARS-CoV-2 and COVID-19 in the corpus and their corresponding category after manual classification.

Category	Terms
pneumonia	pneumonia, respiratory outbreak, lung disease, respiratory tract illness, respiratory illness, respiratory infection, pneumonia-like disease, upper-respiratory illness, respiratory condition, lung infection, pneumonia-like cases, pneumonia-like illness, respiratory virus, lung virus, pneumonia-like virus, pneumonia-causing virus, pneumonialike virus
mystery	mystery, mysterious, unidentified, undocumented, disease x, unknown, abnormal, unexplained
technical	2019-ncov, ncov, 2019 novel coronavirus, n-cov2019, novel coronavirus 2019, ncov2019, cov2019
coronavirus	coronavirus, betacoronavirus, coronovirus

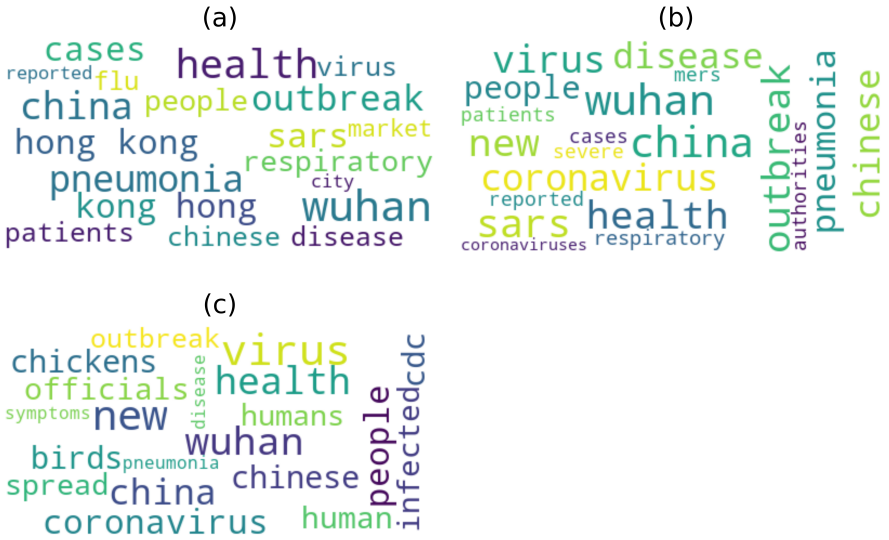


Figure 38 Wordclouds generated from COVID-19 related news articles during three consecutive periods: (a) 31 December 2019 – 08 January 2020, (b) 09 – 19 January 2020, (c) 20 – 26 January 2020.

first report of a case outside China (09 – 12 January 2020), 48.5% (n=127) of the terms were in the "coronavirus" category, 34.7% (n=91) in the "pneumonia" category, and 16.8% (n=44) in the "mystery" category. From this first report to the confirmation of human-to-human transmission (13 – 19 January 2020), 58.3% (n =196) of the terms were in the "coronavirus" category, 27.4% (n=92) in the "pneumonia" category, 11.3% (n=38) in the "mystery", category and 3.0% (n=10) were in the "technical" category. From the confirmation of human-to-human transmission to the end of the studied period 62.9% (n=906) of the terms were in the "coronavirus" category, 17.4% (n=250) in the "technical" category, 14.1% (n=203) in the "pneumonia" category, and 5.6% (n=81) in the "mystery" category (Figure 39).

The incorporation of terms semantically related to "unknown" and "mysterious" events into existing RSS feeds could enhance the detection and retrieval of relevant news. We suggest that

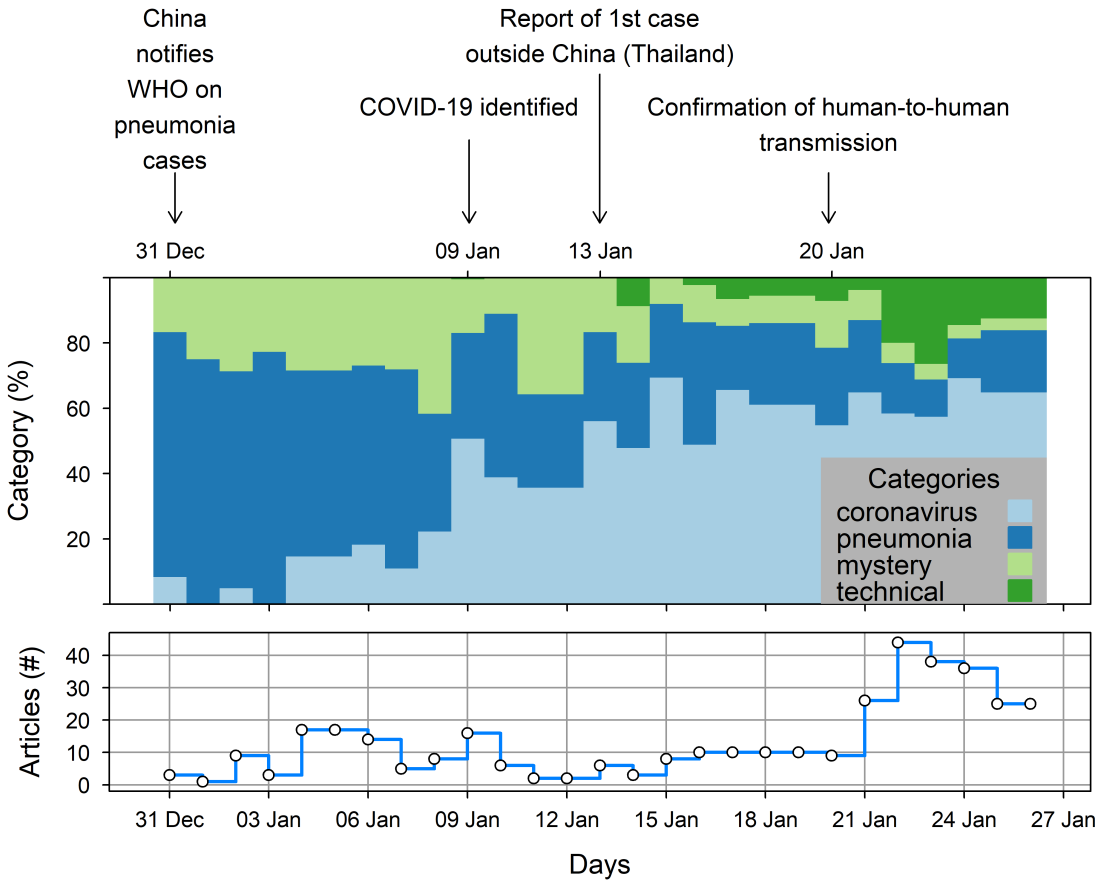


Figure 39 Frequency of the different categories used to describe COVID-19 outbreaks (above), and stepped curve of the daily number of COVID-19 news articles retrieved by PADI-web (below), from 31 December 2019 to 26 January 2020. Daily counts for Saturday and Sunday are merged to account for weekday/weekend trends.

this category of terms could boost the identification of classic epidemiological entities (e.g. disease, hosts, locations, dates) in news feeds.

Our results revealed that the vocabulary changed as the disease spread. EBS methods used to mine and analyse news from the web should thus be tailored to the different disease epidemiology stages.

With MERS-CoV in 2014 and SARS in 2003, COVID-19 is the third coronavirus outbreak emergence that has occurred over the past two decades, thereby highlighting the need to closely monitor the emergence of pneumonia-like illnesses using existing EBS systems. Our results revealed the complementarity of the existing systems and underlined the need for collaborative development. Pooling veterinary and public health information resources seems crucial to improve early detection of unknown diseases in a One Health context. Our future work will focus on identifying the most relevant keywords for rapid detection of unknown threats, in collaboration with experts of other EBS systems. Moreover, EBS tools may be used in a broader setting, such as monitoring

the implementation of protective and control measures.

Efforts invested in improving the timeliness and sensitivity of EBS systems make sense if their outputs (EID event alerts) are formatted, supervised, and interpreted by epidemiologists in collaboration with disease experts and reference laboratories. Most importantly, EID event alerts should feed the risk assessment process to ensure early mitigation of EID events by the health managers and decision-makers.

Chapter 6

Major contributions and perspectives

Table of contents

In this chapter we sum-up the main contributions of the PhD research and discuss future prospects and developments of the work.

1 Summary of the main contributions

The major contributions of this thesis are:

- For the **retrieval of fine-grained epidemiological information** (Chapter 2), we proposed a sentence-based annotation framework involving two annotation levels— (i) the type of event and (ii) the type of epidemiological information. The sentence level provides better insight into the document topic, especially for specific fine-grained types of information. Our contribution stresses the under-exploited value of online news as a source of epidemiological information of interest for risk assessment. We provide the annotation framework and a manually labelled corpus. Based on this corpus, we evaluated two different approaches for automated retrieval of epidemiological information. We showed that the word embedding representation associated with the Multilayer Perceptron classifier outperformed the Support Vector Machines and Naive Bayes classifiers with bag-of-words representations (obtaining overall accuracies of 0.76 and 0.72 for Event type and Information type, respectively). We further proposed an incremental pattern-based approach based on automatic expansion of terms and expert inputs. In this approach, expert validation and proposition of terms are crucial for improving the retrieval quality.
- For **event extraction** (Chapter 3), we showed that the detection of relevant pairs depends on the context, i.e. the window used to retrieve the features. Restraining the context to a fixed window of words achieved better results than retrieving all pairs occurring in a document. Association measures, such as Dice and Mutual Information, could also be used to compute the co-occurrence strength in a simple and interpretable way. Besides, generalizing the spatial features to the country-level enabled better discrimination (i.e. ranking) of relevant disease–location pairs.
- We implemented two fusion methods to evaluate the importance of the different types of lexicosemantic and epidemiological features for the **retrieval of related documents** (Chapter 4). We showed that the use of only four types of epidemiological features, i.e. disease, host, location and dates with appropriate weights, outperformed bag-of-words representations. Spatial features were the most discriminant features, thus highlighting the need for robust methods for spatial entity extraction, disambiguation and representation. Conversely, temporal features performed poorly on their own.

We further considered two case studies to assess the integration of online news (Chapter 5):

- We retrospectively evaluated the ability of three event-based surveillance systems to detect early signs of unknown diseases based on the COVID-19 emergence. Animal health-specific RSS feeds were able to retrieve COVID-19 related news, thus highlighting the porosity between veterinary and public health event-based surveillance. The emergence of COVID-19 was characterized on the basis of symptoms and mystery-related terms, which is consistent

with the emergence of an unknown disease. We produced and shared an annotated corpus of COVID-19-related-news (Valentin et al., 2020e).

- We investigated the dissemination of information between sources. Our results showed that online news sources and press agencies quickly disseminated information of official origin (national veterinary services) at the national level, before their notification to the World animal health organisation (OIE). The reactivity of the information spread was high (i.e., less than one day), thus confirming the interest of event-based systems for the early warning of emerging diseases.

2 Perspectives of text mining in epidemic intelligence

2.1 Retrieval of fine-grained epidemiological information

A sentence-based classifier will shortly be integrated as a new module into the PADI-web information system. Online news will be tagged with their own fine-grained types of information, i.e. the sentences categories. We intend to assess the performances of this module on the daily corpus of retrieved online news.

A further challenge is to identify and implement methods integrating lexicosemantic knowledge for this specific task, as well as for event and information extraction.

Special attention was paid to the genericity of tools and reproducibility of results when building the annotation framework. It would therefore be interesting to assess the proposed annotation framework and classification approaches for other types of threats (i.e. human or plant diseases, antimicrobial resistance, etc.). In this context, the MOOD[1] project -coordinated by CIRAD, aims to improve epidemic intelligence in public-health agencies, providing them with new, innovative, co-designed and co-constructed tools such as PADI-web.

2.2 Event detection

The use of the statistical approaches for event detection is particularly suitable for large volumes of information. We applied this approach in the context of event extraction from a corpus of online news, aligned with methods designed for identifying disease events from streams of tweets (Preoţiuc-Pietro et al., 2016; Thapen et al., 2016b). Our evaluation corpus was relatively small in comparison to those described in the literature, containing several thousand labelled units. Thus, in our future work, we intend to evaluate the statistical approach on a higher volume of data. The statistical approach avoids the need for sophisticated natural language processing methods which are more sensitive to the text quality, for example, after a translation. We could hence tailor our

[1] Monitoring outbreaks for disease surveillance in a data science context", European Commission, Horizon 2020, Health work programme 2020-2023

approach to all articles retrieved daily by PADI-web, including translated articles. The generalization of spatial entities at a national scale looks like a consistent way to reduce the number of generated pairs. Besides, in an event-based perspective, the national level is relevant for the initial early warning. If necessary, the granularity of spatial entities can be increased to monitor a restricted geographical area for example.

2.3 Retrieval of related documents

The retrieval of related documents might be mainstreamed into an EBS system allowing to identify clusters of related documents. Besides, it would be interesting to assess the same approach to link online news with other types of informal textual data, such as tweets, or with structured official data from indicator-based systems. This feature would be helpful for suggesting official events that are potentially related to each news article based on the epidemiological features it contains.

In our work, we did not address extraction and disambiguation tasks for spatial entities. However, location features were crucial for related document retrieval, as well as for event extraction. Thus, it would be interesting to compare the location extraction algorithm form PADI-web to text-based document geocoding approaches, that aim at predicting the geospatial coordinates that best correspond to an entire document, based on its textual contents (Melo and Martins, 2017).

2.4 Role of the expert

In this thesis research, experts were solicited at different levels: annotation guideline creation, corpus annotation, qualitative evaluation of method outputs and consistent feedback with the proposed methods. (Velasco et al., 2014) noted that health agencies have been reluctant to incorporate event-based surveillance system outputs into their systems because many technical issues had not yet been addressed. Health experts are inclined to understand automatic processes beyond event-based surveillance outputs, as they directly support the decision-making process (Cui et al., 2019). The frontier between informatics methods and epidemiological expertise therefor must be porous.

In this context, it would be interesting to assess semi-automated approaches, combining machine-learning methods with expert heuristics and knowledge. For instance, active learning might achieve precise more accurate classification using as few labelled samples as possible. Active learning relies on an iterative procedure for selecting the most informative unlabeled examples in terms of how much further information they would bring once manually labelled (Liere and Tadepalli, 1997). It is efficient in decreasing the number of training samplings used in several different contexts, such as text or image classification (Esuli and Sebastiani, 2009; Figueroa et al., 2012). In the health domain, active learning effectively reduces the sample size of training sets and outperforms the passive selection approach (Figueroa et al., 2012). We hence believe that this could be an interesting alternative to the supervised approaches currently implemented in event-based systems.

2.5 Detection of weak signals

The concept of weak signal should be precisely defined in the context of event-based surveillance so as to formalize specific research questions and assess adapted methods. Our preliminary insight suggests that the weak signal applies to two different level: (i) the feature level, and (ii) the document level (i.e. online news):

- Rare occurrences of event attributes, such as disease–location pairs are potential weak signals, as they can also be the sign of an event that has not yet been confirmed at the national or international level. In this context, a weak signal could be defined as a temporal anomaly of the frequency of a term or an association of terms compared to a baseline. Detection of weak signals by event-based systems could consist of implementing alerts based on terms-weighted metrics which take the temporal dimension into account. For instance, the Peakiness Score is a normalized word frequency metric, similar to the TF-IDF metric, for identifying words that are specific to a fixed-length time window (Shamma et al., 2011). Trending Score, a version adapted to n-gram, was proposed by (Benhardus and Kalita, 2013). For a given n-gram and time window, it involves computing the normalized frequency of that n-gram with regard to the frequency of other n-grams in this window. The advantage of such approaches is the unsupervised aspect. However, blindly creating alerts may lead to an increase of false-positive alerts. It would be interesting to evaluate the impact of selecting types of entities or terms, such as spatial features, symptoms or mystery-related terms. The two latter types of terms were specifically used in reference to COVID-19 before its official identification (Chapter 5).

- Some news articles themselves represent a weak signal, e.g. news articles reporting an occurrence of an unknown disease. In such cases, the detection of a weak signal could be addressed as a conventional document classification task. During the thesis research, a binary supervised classification module was implemented in PADI-web to automatically assess the relevance of news, as fully described in Appendix C. This methods might be further refined to encompass different themes at the document level. On this basis, our results suggest unknown disease news could be detected with fair accuracy (Valentin et al., 2020b). Such an approach could be the first step towards weak signal detection, while being easily integrable in EBS pipelines.

3 Perspectives of the event-based surveillance systems

In this Section, we discuss the medium- and long-term prospects for event-based surveillance systems.

3.1 Information dissemination between sources

The fact that online news can be a source of timely epidemiological information should be mitigated by the risk of false information, or so-called *fake news*. On the one hand, domain experts can easily assess the truthfulness of information based on their knowledge and solicitation of a network of local sources of information. On the other hand, automatic event-based surveillance systems are more at risk of releasing wrong information. Misinformation has a major impact, not only causing delays or hampering effective care but also interfering with official health knowledge dissemination (Wang et al., 2019). Brainard and Hunter (2020) suggest that controlling the spread of misinformation or susceptibility to it could reduce disease burdens. Research has focused on developing algorithms able to detect health misinformation at an early stage. For instance, supervised classification, based on syntactic and semantic features of news, has proven effective in differentiating fake from correct online news with an accuracy of 76% (Pérez-Rosas et al., 2018). Such an approach could be assessed with regard to misinformation in animal health-related news, especially in event-based surveillance automated systems. In that context, another challenge is to assess methods to evaluate the reliability the sources used.

3.2 Enhancing One Health surveillance through cross-sectorial collaboration

Our results revealed the complementarity of existing event-based surveillance systems and underlined the need for collaborative development to improve early detection of unknown diseases. Beyond the need for experience sharing in terms of text-mining and informatics methods, pooling veterinary and public health information resources seems crucial. The recognition that human and veterinary medicine could be developed in a joint and complementary manner is not new (Majok et al., 1996): in the last two decades, the One Health approach has seen an unprecedented revival in scientific debates, research programs and epidemiological studies . This has been further confirmed since the COVID-19 emergence, with recognition that the long-term experience gained by veterinary medicine with animal coronaviruses could drive future research in human medicine towards the development of vaccines and effective antiviral drugs (Decaro et al., 2020). However, applying the One Health concept in practical methods in the integrated disease surveillance field is still hindered by the fact that a major share of human and animal health thinking, training and actions remain in separate disciplinary silos (Zinsstag et al., 2011). Data from conventional animal and human health surveillance systems are obtained at different resolutions and scales, and for various purposes. In this context, several event-based surveillance systems have been re-

trieving public and animal health data for many years. Thus, event-based surveillance systems have a crucial role to play in establishing links between animal and human health data. Efforts to build bridges between health domains by taking their specificity into account are still hampered by methodological challenges that should be addressed in further research.

In this thesis research, we trained models based on traditional supervised learning methods where the trained models are specific to a task and a training dataset. An alternative possibility would be to focus research on transfer learning, where the knowledge arising from previously trained models (e.g. features, weights, etc.) is used to leverage the training of new models, addressing problems such as the lack of learning data (Pan and Yang, 2010). This approach might be assessed in the context, for instance, of transferring a classifier trained on animal diseases-related news to a classifier for plant disease-related-news or even to new topics such as antimicrobial resistance.

3.3 Event-based surveillance systems in developing countries

By recognizing the weaknesses of conventional indicator-based systems for detecting and reporting disease outbreaks efficiently, the EBS paradigm has concentrated effort on retrieving publicly available sources. EBS is hence an inexpensive and effective alternative to formal surveillance. Several authors thus encourage implementation of EBS systems in areas that are most prone to outbreaks, with limited technological infrastructure, and where indicator-based surveillance is suboptimal (Kuehne et al., 2019; Wilburn et al., 2019). The majority of EBS systems are based in North America and Europe, and fewer systems focus on monitoring epidemic threats in developing countries (Velasco et al., 2014). Nevertheless, apart from restricted systems such as GPHIN, most of EBS systems are open-access and allow free and rapid access to health information worldwide (pending internet access). Besides, we noted that increasing reliance on private stakeholders has accompanied the development of EBS pipelines. Private solutions allow EBS systems to rapidly mainstream efficient cutting-edge technological advances, such as deep-learning-based translation among other NLP tasks. We believe that special attention should be paid to the role of private solutions within EBS pipelines. It is crucial that they do not threaten the sustainability of open source systems. Besides, in the broader open science context, the sharing of resources such as annotated corpora and gold standards should be encouraged to enable accurate comparison of methods between the event-based surveillance systems.

3.4 Monitoring of online news through news aggregators

When schematizing factors influencing the retrieval of online news by event-based surveillance systems in Chapter 1, we omitted an important intermediate layer, i.e. news aggregators. Currently, six out of seven event-based surveillance systems that collect online news from the Web rely on the Google News aggregator. Efforts that have focused on designing relevant health-relevant RSS feeds are based on the assumption that the retrieval of new by Google News is unbiased. However, the Google News algorithm is not publicly available and little is known about the factors that drive the selection and curation of the information sources. A recent study showed that a bias emerged in the way Google News retrieved and ranked online news in the political field (Trielli and Diakopoulos, 2019). Besides, in this study, the algorithm provided a high degree of source concentration, to the detriment of source diversity and retrieval of local news sources. Although the existence of political bias is not necessarily an indication of parallel bias in the health domain, attention should be paid to the ability of Google News to access local media. Indeed, multilingual queries are implemented in event-based surveillance systems to enhance the detection of health information at the local level. Besides, it would be interesting to compare the results of queries obtained via Google News and alternative news aggregators that are currently in use.

4 Conclusion

Increases in the emergence or re-emergence of animal and human infectious diseases have been evident in many parts of the world for several years. Beyond the well-known role of human and animal mobility in the spread of pathogens, climate change and biodiversity loss are likely to exacerbate the global disease burden (Keesing et al., 2010; Ostfeld, 2009). National and international institutions are currently experimenting a global paradox—conciliating trade extension with the control of the risk to public and animal health.

In this context, event-based surveillance can identify events faster than indicator-based surveillance reporting procedures. In this thesis, we have addressed several research questions with the aiming of enhancing the use of a specific informal source, i.e. online news, for the detection of animal health events and epidemiological information. Methods implemented in event-based surveillance systems, as well as their assessment, are still maturing. The involvement of domain experts in developing new methods increases the interpretability and performance of event-based surveillance systems and may participate in ensuring their long-term acceptability. Whether the detection of an event through an event-based surveillance system is capable of triggering response to enhance early mitigation of emerging diseases events has yet to be assessed. Efforts made to improve the timeliness and sensitivity of event-based surveillance systems only make sense if their outputs (event alerts) are taken into account by health managers and decision makers, while fuelling the risk assessment process. Without a targeted response, event-based surveillance is limited to monitoring, and the use of the term “surveillance” instead of “monitoring” could be questioned (Anholt et al., 2014).

Event-based surveillance systems are at the crossroads of human and animal (and plant and ecosystem) health, epidemiology, statistics, and informatics. Their deployment thus faces many challenges specific to each domain and their intersection, such as the relation between automation, artificial intelligence and expertise, the link between surveillance and response, the status of open access tools, or the ethical implications related to using the Web as a source for health surveillance. To overcome these challenges and obstacles, upstream research on event-based surveillance systems – fed itself by interdisciplinary research and exchanges, is crucially needed to better anticipate, quickly detect and early control the ever-growing risk of emerging diseases.

Bibliography

Bibliography

Agibetov, A., Blagec, K., Xu, H., and Samwald, M. (2018). Fast and scalable neural embedding models for biomedical sentence classification. *BMC Bioinformatics*, 19(1):541. 61

Ahlers, D. (2013). Assessment of the Accuracy of GeoNames Gazetteer Data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pages 74–81, New York, NY, USA. ACM. 27, 178

Ahn, D. (2006). The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, page 8. Association for Computational Linguistics. 29

Aizawa, A. (2001). Linguistic Techniques to Improve the Performance of Automatic Text Categorization. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001)*, page 8. 57

Aji, S. (2012). Document Summarization Using Positive Pointwise Mutual Information. *International Journal of Computer Science and Information Technology*, 4(2):47–55. 81

Alomar, O., Batlle, A., Brunetti, J. M., García, R., Gil, R., Granollers, T., Jiménez, S., Laviña, A., Reverté, C., Riudavets, J., and Virgili-Gomà, J. (2016). Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats. *EFSA Supporting Publications*, 13(12). 10

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 273–280, Sheffield, United Kingdom. Association for Computing Machinery. 26, 27

Anholt, R., Berezowski, J., Jamal, I., Ribble, C., and Stephen, C. (2014). Mining free-text medical records for companion animal enteric syndrome surveillance. *Preventive Veterinary Medicine*, 113(4):417–422. 146

Arsevska, E., Falala, S., De Goer, J., Lancelot, R., Rabatel, J., and Roche, M. (2017). PADI-web: platform for automated extraction of animal disease information from the web. In *Proceedings of LTC - Language and Technology Conference*, pages 241–245. xxiv, 2, 14, 19, 22

Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., and Dufour, B. (2016). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, 123:104–115. 19, 178

Arsevska, E., Valentin, S., Rabatel, J., de Goër de Hervé, J., Falala, S., Lancelot, R., and Roche, M. (2018). Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLOS ONE*, 13(8):e0199960. xxiv, 2, 8, 9, 10, 12, 18, 22, 32, 177

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596. 40

Bahk, C. Y., Scales, D. A., Mekaru, S. R., Brownstein, J. S., and Freifeld, C. C. (2015). Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. *BMC Infectious Diseases*, 15(1). xxiv, 2, 7, 9

Baker, M. G. and Forsyth, A. M. (2007). The new International Health Regulations: a revolutionary change in global health security. *The New Zealand Medical Journal*, 120(1267):U2872. 12

Barboza, P., Vaillant, L., Le Strat, Y., Hartley, D. M., Nelson, N. P., Mawudeku, A., Madoff, L. C., Linge, J. P., Collier, N., Brownstein, J. S., and Astagneau, P. (2014). Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PLoS ONE*, 9(3):e90536. 133

Barboza, P., Vaillant, L., Mawudeku, A., Nelson, N. P., Hartley, D. M., Madoff, L. C., Linge, J. P., Collier, N., Brownstein, J. S., Yangarber, R., Astagneau, P., and on behalf of the Early Alerting, Reporting Project of the Global Health Security Initiative (2013). Evaluation of epidemic intelligence systems integrated in the Early Alerting and Reporting project for the detection of A/H5N1 influenza events. *PLoS ONE*, 8(3):e57252. xxiv, 2, 5, 9, 10

Ben Jebara, K. and Shimshony, A. (2006). International monitoring and surveillance of animal diseases using official and unofficial sources. *Veterinaria Italiana*, 42(4):431–441. xxiii, 1

Benhardus, J. and Kalita, J. (2013). Streaming trend detection in Twitter. *International Journal of Web Based Communities*, 9(1):122. 143

Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., and Pedreschi, D. (2011). The pursuit of hubbiness: Analysis of hubs in large multidimensional networks. *Journal of Computational Science*, 2(3):223–237. 127, 130

Billah Nagoudi, E. M., Ferrero, J., Schwab, D., and Cherroun, H. (2017). Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences. In *6th International Conference on Arabic Language Processing*, Fez, Morocco. 21

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. 42, 104

Bird, S. and Loper, E. (2004). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics. 27, 60

Blanchard, J., Guillet, F., Gras, R., and Briand, H. (2005). Using information-theoretic measures to assess association rule interestingness. In *5th IEEE International Conference on Data Mining ICDM'05*, pages 66–73, United States. IEEE Computer Society. 81

Blench, M. (2008). Global public health intelligence network (GPHIN). In *8th Conference of the Association for Machine Translation in the Americas*, pages 8–12. 19, 33

Bollig, N., Clarke, L., Elsmo, E., and Craven, M. (2020). Machine learning for syndromic surveillance using veterinary necropsy reports. *PLOS ONE*, 15(2):e0228105. Publisher: Public Library of Science. 58

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of German Society for Computational Linguistics and Language Technology Conference,*, pages 31 – 40. 93

Brainard, J. and Hunter, P. R. (2020). Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus. *SIMULATION*, 96(4):365–374. Publisher: SAGE Publications Ltd STM. 144

Broniatowski, D. A., Paul, M. J., and Dredze, M. (2013). National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS ONE*, 8(12):e83672. 8

Brownlee, J. (2017). *Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems*. Machine Learning Mastery. Google-Books-ID: _pmoDwAAQBAJ. 57

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS medicine*, 5(7):e151. 7, 18, 21, 30, 33

Brugere, C., Onuigbo, D. M., and Morgan, K. L. (2017). People matter in animal disease surveillance: Challenges and opportunities for the aquaculture sector. *Aquaculture*, 467:158–169. xxiii, 1, 6

Buckley, C. and Voorhees, E. M. (2017). Evaluating Evaluation Measure Stability. *ACM SIGIR Forum*, 51(2):8. 111

Caceres, P., Awada, L., Barboza, P., Lopez-Gatell, H., and Tizzani, P. (2017). The World Organisation for Animal Health and the World Health Organization: intergovernmental disease information and reporting systems and their role in early warning. *Revue Scientifique et Technique de l'OIE*, 36(2):539–548. 12

Carneiro, H. and Mylonakis, E. (2009). Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564. 9

Carrion, M. and Madoff, L. C. (2017). ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. *International Health*, 9(3):177–183. 8, 14, 18

Carter, D., Stojanovic, M., Hachey, P., Fournier, K., Rodier, S., Wang, Y., and de Bruijn, B. (2020). Global Public Health Surveillance using Media Reports: Redesigning GPHIN. *arXiv e-prints*, page arXiv:2004.04596. _eprint: 2004.04596. 20, 21, 23, 33

Chambers, N. and Jurafsky, D. (2011). Template-based Information Extraction Without the Templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 976–986, Stroudsburg, PA, USA. Association for Computational Linguistics. 14

Chandrasekar, K. (2012). Use of Information Communication Technology (ICT) in communicable disease surveillance – a review of literature. *Sri Lanka Journal of Bio-Medical Informatics*, 2(2):41–52. Number: 2 Publisher: HISSL. 7

Chanlekha, H., Kawazoe, A., and Collier, N. (2010). A framework for enhancing spatial and temporal granularity in report-based health surveillance systems. *BMC medical informatics and decision making*, 10(1):1. 26, 39, 40, 49, 51

Choi, J., Cho, Y., Shim, E., and Woo, H. (2016). Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health*, 16(1). 8

Chowell, G., Cleaton, J. M., and Viboud, C. (2016). Elucidating Transmission Patterns From Internet Reports: Ebola and Middle East Respiratory Syndrome as Case Studies. *The Journal of Infectious Diseases*, 214(suppl_4):S421–S426. Publisher: Oxford Academic. 8

Chua, S. (2008). The Role of Parts-of-Speech in Feature Selection. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, Hong Kong. 104

Chunara, R., Andrews, J. R., and Brownstein, J. S. (2012). Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45. 8

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics* -, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics. 81

Claes, F., Kuznetsov, D., Liechti, R., Von Dobschuetz, S., Dinh Truong, B., Gleizes, A., Conversa, D., Colonna, A., Demaio, E., Ramazzotto, S., Larfaoui, F., Pinto, J., Le Mercier, P., Xenarios, I., and Dauphin, G. (2014). The EMPRES-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. *Database*, 2014(0):bau008–bau008. 95

Clinchant, S., Ah-Pine, J., and Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, page 44. ACM. 107

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46. 48

Collier, N. (2012). Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health*, 7(7):19. xxiv, 2, 5

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., and Taniguchi, K. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941. 28, 117

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R. A., Takeuchi, K., and Kawtrakul, A. (2007). A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language resources and evaluation*, 40(3-4):405. 19, 28

Conway, M., Doan, S., Kawazoe, A., and Collier, N. (2009). Classifying Disease Outbreak Reports Using N-grams and Semantic. *International Journal of Medical Informatics*, 78(12). 23, 39

Conway, M., Kawazoe, A., Chanlekha, H., and Collier, N. (2010). Developing a Disease Outbreak Event Corpus. *Journal of Medical Internet Research*, 12(3):e43. 39, 40

Corley, C. and Mihalcea, R. (2005). Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Ann Arbor, Michigan. Association for Computational Linguistics. 20

Council, E. (2016). Council Directive 82/894/EEC on the notification of animal diseases within the Community. 12

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695. 125

Cui, M., Bai, R., Lu, Z., Li, X., Aickelin, U., and Ge, P. (2019). Regular Expression Based Medical Text Classification Using Constructive Heuristic Approach. *IEEE Access*, 7:147892–147904. Conference Name: IEEE Access. 78, 142

Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*, 47(1):217–238. 68

Dai, X., Bikdash, M., and Meyer, B. (2017). From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon 2017*, pages 1–7, Concord, NC, USA. IEEE. 67

d'Amato, C., Fernandez, M., Tamma, V., Lecue, F., Cudré-Mauroux, P., Sequeda, J., Lange, C., and Heflin, J. (2017). *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I.* Springer. Google-Books-ID: qHg5DwAAQBAJ. 68

Danlos, L., Nakamura, T., and Pradet, Q. (2015). Traduction de VerbNet vers le français. In *Congrès ACFAS*, page 1, Rimouski, Canada. 97

De Boom, C., Van Canneyt, S., Demeester, T., and Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156. arXiv: 1607.00570. 60

Decaro, N., Martella, V., Saif, L. J., and Buonavoglia, C. (2020). COVID-19 from veterinary medicine and one health perspectives: What animal coronaviruses have taught us. *Research in Veterinary Science*, 131:21–23. 144

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. arXiv: 1810.04805. 64

Dias, G., Hasanuzzaman, M., Ferrari, S., and Mathet, Y. (2014). TempoWordNet for Sentence Time Tagging. In *23rd international conference on World wide web companion*, pages Pages 833–838, Seoul, South Korea. 77

Diaz-Aviles, E. and Stewart, A. (2012). Tracking Twitter for epidemic intelligence: case study: EHEC/HUS outbreak in Germany, 2011. In *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*, pages 82–85, Evanston, Illinois. ACM Press. 8

Dilawar, N., Majeed, H., Beg, M. O., Ejaz, N., Muhammad, K., Mehmood, I., and Nam, Y. (2018). Understanding Citizen Issues through Reviews: A Step towards Data Informed Planning in Smart Cities. *Applied Sciences*, 8(9):1589. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute. 60

Dion, M., AbdelMalik, P., and Mawudeku, A. (2015). Big Data and the Global Public Health Intelligence Network (GPHIN). *Canada Communicable Disease Report*, 41(9):209–214. xxiv, 2, 9

Doan, S., Kawazoe, A., and Collier, N. (2007). The Role of Roles in Classifying Annotated Biomedical Text. In *Biological, translational, and clinical language processing*, pages 17–24, Prague, Czech Republic. Association for Computational Linguistics. 39, 77

Drache, D., Feldman, S., and Clifton, D. (2003). Media coverage of the 2003 Toronto SARS outbreak. Technical report, Robarts Centre for Canadian Studies, York University, Toronto, Robarts Centre for Canadian Studies, York University;. 15

Du, M., Pivovarova, L., and Yangarber, R. (2016). PULS: natural language processing for business intelligence. In *Proceedings of the 2016 Workshop on Human Language Technology and Intelligent Applications*, New York, United States. 29

Du, M. and Yangarber, R. (2015). Acquisition of domain-specific patterns for single document summarization and information extraction. In *Proceedings of the The Second International Conference on Artificial Intelligence and Pattern Recognition*, Shenzhen, China. 69

Edo-Osagie, O., Smith, G., Lake, I., Edeghere, O., and Iglesia, B. D. L. (2019). Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. *PLOS ONE*, 14(7):e0210689. 68

Esuli, A. and Sebastiani, F. (2009). Active Learning Strategies for Multi-Label Text Classification. In Boughanem, M., Berrut, C., Mothe, J., and Soule-Dupuy, C., editors, *Advances in Information Retrieval*, pages 102–113, Berlin, Heidelberg. Springer Berlin Heidelberg. 142

Feldman, R., Aumann, Y., Liberzon, Y., Ankori, K., Schler, J., and Rosenfeld, B. (2001). A domain independent environment for creating information extraction modules. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 586–588, New York, NY, USA. Association for Computing Machinery. 31

Figueroa, R. L., Zeng-Treitler, Q., Ngo, L. H., Goryachev, S., and Wiechmann, E. P. (2012). Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association : JAMIA*, 19(5):809–816. 142

Fornito, A., Zalesky, A., and Bullmore, E. T. (2016). Chapter 4 - Node Degree and Strength. In *Fundamentals of Brain Network Analysis*, pages 115–136. Academic Press, San Diego. 127

Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). 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. 27

Ghosh, S., Chakraborty, P., Cohn, E., Brownstein, J. S., and Ramakrishnan, N. (2016). Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach. *arXiv:1603.00106 [cs, stat]*. arXiv: 1603.00106. 58

Ghosh, S., Chakraborty, P., Lewis, B. L., Majumder, M. S., Cohn, E., Brownstein, J. S., Marathe, M. V., and Ramakrishnan, N. (2017). Guided Deep List: Automating the Generation of Epidemiological Line Lists from Open Sources. *arXiv:1702.06663 [cs]*. arXiv: 1702.06663. 58, 69, 77

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, 150. 61

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309. 58, 64

Gomaa, W. H. and Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13). 21

Grishman, R., Huttunen, S., and Yangarber, R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4):236–246. 31

Grootendorst, M. and Vanschoren, J. (2020). Beyond Bag-of-Concepts: Vectors of Locally Aggregated Concepts. In Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., and Robardet, C., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 11907, pages 681–696. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science. 57

Guarino, N., Oberle, D., and Staab, S. (2009). What Is an Ontology? In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 1–17. Springer, Berlin, Heidelberg. 26

Gudivada, V. N., Rao, D. L., and Gudivada, A. R. (2018). Information Retrieval: Concepts, Models, and Systems. In Gudivada, V. N. and Rao, C. R., editors, *Handbook of Statistics*, volume 38 of *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*, pages 331–401. Elsevier. 21

Guégan, M. and Hernandez, N. (2006). Recognizing Textual Parallelisms with Edit Distance and Similarity Degree. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics. 20

HaCohen-Kerner, Y., Miller, D., and Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE*, 15(5):e0232525. Publisher: Public Library of Science. 104

Halliday, J., Daborn, C., Auty, H., Mtema, Z., Lembo, T., Bronsvoort, B. M. d., Handel, I., Knobel, D., Hampson, K., and Cleaveland, S. (2012). Bringing together emerging and endemic zoonoses surveillance: shared challenges and a common solution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1604):2872–2880. 7

Harcup, T. and O'Neill, D. (2017). What is News? *Journalism Studies*, 18(12):1470–1488. Publisher: Routledge _eprint: https://doi.org/10.1080/1461670X.2016.1150193. 15

Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., Thinus, G., and Lightfoot, N. (2010). The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, 3(0). 2, 5, 33

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science and Business Media. Google-Books-ID: tVIjmNS3Ob8C. 62

Hearst, M. A. (1999). Untangling Text Data Mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10, College Park, Maryland, USA. Association for Computational Linguistics. xxiv, 2

Henry, S., Cuffy, C., and McInnes, B. T. (2018). Vector representations of multi-word terms for semantic relatedness. *Journal of Biomedical Informatics*, 77:111–119. 64

Heymann, D. and Rodier, G. (2001). Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *Lancet Infectious Diseases*, 1(5):345–353. xxiii, 1, 7

Honnibal, M. and Montani, I. (2018). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. In *To appear*. 27, 73

Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the 6th New Zealand Computer Science Research Student Conference*. 111

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506. Publisher: Elsevier. 131

Ibekwe-Sanjuan, F., Silvia, F., Eric, S., and Eric, C. (2011). Annotation of Scientific Summaries for Information Retrieval. *arXiv:1110.5722 [cs]*. arXiv: 1110.5722. 69, 77

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., and Ghazi, D. (2017). Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2):237–253. 27

Jijkoun, V. and de Rijke, M. (2005). Recognizing Textual Entailment Using Lexical Similarity. *Proceedings Pascal 2005 Textual Entailment Challenge Workshop*, page 4. 20

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In Carbonell, J. G., Siekmann, J., Goos, G., Hartmanis, J., van Leeuwen, J., Nédellec, C., and Rouveirol, C., editors, *Machine Learning: ECML-98*, volume 1398, pages 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg. 61

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21. 57

Jones, R., Mccallum, A., Nigam, K., and Riloff, E. (1999). Bootstrapping for Text Learning Tasks. In *In IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, pages 52–63. 69

Kaiser, R., Coulombier, D., Baldari, M., Morgan, D., and Paquet, C. (2006). What is epidemic intelligence, and how is it being improved in Europe? *Weekly releases (1997–2007)*, 11(5):2892. 9, 10

Karesh, W. B., Dobson, A., Lloyd-Smith, J. O., Lubroth, J., Dixon, M. A., Bennett, M., Aldrich, S., Harrington, T., Formenty, P., Loh, E. H., Machalaba, C. C., Thomas, M. J., and Heymann, D. L. (2012). Ecology of zoonoses: natural and unnatural histories. *The Lancet*, 380(9857):1936–1945. Publisher: Elsevier. xxiii

Kawazoe, A., Chanlekha, H., Shigematsu, M., and Collier, N. (2008). Structuring an event ontology for disease outbreak detection. *BMC Bioinformatics*, 9(Suppl 3):S8. 14, 28

Kawazoe, A., Jin, L., Shigematsu, M., Barrero, R., Taniguchi, K., and Collier, N. (2006). The Development of a Schema for the Annotation of Terms in the Biocaster Disease Detecting/Tracking System. In *KR-MED*. 28, 51

Keesing, F., Belden, L. K., Daszak, P., Dobson, A., Harvell, C. D., Holt, R. D., Hudson, P., Jolles, A., Jones, K. E., Mitchell, C. E., Myers, S. S., Bogich, T., and Ostfeld, R. S. (2010). Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, 468(7324):647–652. Number: 7324 Publisher: Nature Publishing Group. 146

Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Mandl, K. D., Mawudeku, A., Eysenbach, G., and Brownstein, J. S. (2009). Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. *Emerging Infectious Diseases*, 15(5):689–695. 5

Khoo, A., Marom, Y., and Albrecht, D. (2006). Experiments with Sentence Classification. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 18–25, Sydney, Australia. 61

Kibriya, A. M., Frank, E., Pfahringer, B., and Holmes, G. (2005). Multinomial Naive Bayes for Text Categorization Revisited. In Webb, G. I. and Yu, X., editors, *AI 2004: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 488–499, Berlin, Heidelberg. Springer. 61

Kishida, K. (2005). Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. *NII Technical Reports*, 2005(14):1–19. 88

Kolaczyk, E. D. (2009). Descriptive Analysis of Network Graph Characteristics. In Kolaczyk, E. D., editor, *Statistical Analysis of Network Data: Methods and Models*, Springer Series in Statistics, pages 1–44. Springer, New York, NY. 122, 123

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. (2015). Consistent Multilabel Classification. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3321–3329. Curran Associates, Inc. 52

Kuehne, A., Keating, P., Polonsky, J., Haskew, C., Schenkel, K., Waroux, O. L. P. d., and Ratnayake, R. (2019). Event-based surveillance at health facility and community level in low-income and

middle-income countries: a systematic review. *BMJ Global Health*, 4(6):e001878. Publisher: BMJ Specialist Journals Section: Research. 145

Kumar, M. A. and Gopal, M. (2010). A Comparison Study on Multiple Binary-Class SVM Methods for Unilabel Text Categorization. *Pattern Recogn. Lett.*, 31(11):1437–1444. 61

Kuribreña, M. A., Awada, L., Mur, L., and Tizzani, P. (2019). Current animal health situation worldwide: analysis of events and trends. In *87th General Session*, page 34, Paris. 7

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From Word Embeddings To Document Distances. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR. 61

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 591, Raleigh, North Carolina, USA. ACM Press. 8

Lafferty, J., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. 27

Lallich, S., Teytaud, O., and Prudhomme, E. (2007). Association Rule Interestingness: Measure and Statistical Validation. In Kacprzyk, J., Guillet, F. J., and Hamilton, H. J., editors, *Quality Measures in Data Mining*, volume 43, pages 251–275. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Studies in Computational Intelligence. 81

Langley, P. and Sage, S. (1994). Induction of Selective Bayesian Classifiers. In *Uncertainty Proceedings 1994*, pages 399–406. Elsevier. 61

Langmuir, A. D. (1980). The Epidemic Intelligence Service of the Center for Disease Control. *Public Health Reports*, 95(5):470–477. 9

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205. 9

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196, Beijing, China. JMLR.org. 61

Leeuwenberg, A., Vela, M., Dehdari, J., and van Genabith, J. (2016). A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. *The Prague Bulletin of Mathematical Linguistics*, 105(1):111–142. 71, 77

Lejeune, G., Brixtel, R., Doucet, A., and Lucas, N. (2012). DAnIEL: Language Independent Character-Based News Surveillance. In Isahara, H. and Kanzaki, K., editors, *Advances in Natural Language Processing*, Lecture Notes in Computer Science, pages 64–75, Berlin, Heidelberg. Springer. 20, 51

Lejeune, G., Brixtel, R., Doucet, A., and Lucas, N. (2015). Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, 65(2):131–143. 20, 40

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 37–50, Copenhagen, Denmark. Association for Computing Machinery. 57

Li, H., Srihari, K. R., Niu, C., and Li, W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 39–44. 27

Li, J. and Cardie, C. (2013). Early Stage Influenza Detection from Twitter. *arXiv:1309.7340 [cs]*. arXiv: 1309.7340. 8

Li, X., Zai, J., Zhao, Q., Nie, Q., Li, Y., Foley, B. T., and Chaillon, A. (2020). Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *Journal of Medical Virology*, 92(6):602–611. xxiii, 1

Liere, R. and Tadepalli, P. (1997). Active Learning with Committees for Text Categorization. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, pages 591–596. AAAI Press. event-place: Providence, Rhode Island. 142

Lison, P. and Kutuzov, A. (2017). Redefining Context Windows for Word Embedding Models: An Experimental Study. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 284–288, Gothenburg, Sweden. Association for Computational Linguistics. 64

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014). BIOTEX: A system for Biomedical Terminology Extraction, Ranking, and Validation. In *International Semantic Web Conference*. 132

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, 19(1):59–99. 132

Lu, T. and Reis, B. Y. (2020). Internet Search Patterns Reveal Clinical Course of Disease Progression for COVID-19 and Predict Pandemic Spread in 32 Countries. preprint, Epidemiology. 8

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4):309–317. 57

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165. Conference Name: IBM Journal of Research and Development. 77

Lyon, A., Grossel, G., Burgman, M., and Nunn, M. (2013a). Using internet intelligence to manage biosecurity risks: a case study for aquatic animal health. *Diversity and Distributions*, 19(5-6):640–650. 10, 27

Lyon, A., Mooney, A., and Grossel, G. (2013b). Using AquaticHealth.net to Detect Emerging Trends in Aquatic Animal Health. *Agriculture*, 3(2):299–309. 10, 33

Lyon, A., Nunn, M., Grossel, G., and Burgman, M. (2011). Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap: Comparison of Web-Based Biosecurity Intelligence Systems. *Transboundary and Emerging Diseases*, 59(3):223–232. 21

Majok, A. A., Schwabe, C. W., and Majok, A. (1996). *Development Among Africa's Migratory Pastoralists*. Greenwood Publishing Group. Google-Books-ID: gt6elh3jaAEC. 144

Majumder, G., Pakray, P., Gelbukh, A., and Pinto, D. (2016). Semantic Textual Similarity Methods, Tools, and Applications: A Survey. *Computación y Sistemas*, 20(4). 21

Mandelbaum, A. and Shalev, A. (2016). Word Embeddings and Their Use In Sentence Classification Tasks. *arXiv:1610.08229 [cs]*. arXiv: 1610.08229. 61, 67

Manning, C., Raghavan, P., and Schuetze, H. (2009). Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*, page 581. Cambridge University Press. 57

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. 27

Mantero, J., Belyaeva, J., Linge, J., European Commission, Joint Research Centre, and Institute for the Protection and the Security of the Citizen (2011). *How to maximise event-based surveillance web-systems: the example of ECDC/JRC collaboration to improve the performance of MedISys*. Publications Office, Luxembourg. OCLC: 870614547. 19, 22, 33

Margineantu, D., Wong, W.-K., and Dash, D. (2010). Machine learning algorithms for event detection: A special issue of Machine Learning. *Machine Learning*, 79(3):257–259. 31

Martin, V., Von Dobschuetz, S., Lemenach, A., Rass, N., Schoustra, W., and DeSimone, L. (2007). Early warning, database, and information systems for avian influenza surveillance. *Journal of wildlife diseases*, 43(3_Supplement):S71. 13

Martins, B., Manguinhas, H., and Borbinha, J. (2008). Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. In *IEEE International Conference on Semantic Computing*, pages 1–9. 27

McNabb, S. J. (2010). Comprehensive effective and efficient global public health surveillance. *BMC Public Health*, 10(1):S3. 7

Mekaoui, M., Tisserant, G., Dodard, M., and Cédric, L. (2020). Extraction de tâches dans les e-mails : une approche fondée sur les rôles sémantiques. In *Extraction et Gestion des Connaissances (EGC'2020)*. 97

Melo, F. and Martins, B. (2017). Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, 21(1):3–38. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12212. 142

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. arXiv: 1301.3781. 59, 60

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*. 60

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc. 59, 60

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2010.01106.x. 60

Mooney, R. J. and Bunescu, R. (2005). Mining knowledge from text using information extraction. *ACM SIGKDD*, 7(1):3–10. 25

Morid, M. A., Fiszman, M., Raja, K., Jonnalagadda, S. R., and Del Fiol, G. (2016). Classification of clinically useful sentences in clinical evidence resources. *Journal of Biomedical Informatics*, 60:14–22. 77

Mykhalovskiy, E. and Weir, L. (2006). The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. *Canadian journal of public health = Revue canadienne de sante publique*, 97(1):42–44. 19

Nadkarni, P. M. (2002). An introduction to information retrieval: applications in genomics. *The pharmacogenomics journal*, 2(2):96–102. 103

Naili, M., Chaibi, A. H., and Ben Ghezala, H. H. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, 112:340–349. 64

Ng, V. and Cardie, C. (2002). Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. 76

Nguyen, H. T., Duong, P. H., and Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, 182:104842. 21, 78

Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of Jaccard Coefficient for Keywords Similarity. *Hong Kong*, page 5. 81

Nooralahzadeh, F., Øvrelid, L., and Lønning, J. T. (2018). Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). 77

OIE (2007). African swine fever (ASF) confirmed in Georgia : OIE - World Organisation for Animal Health. Library Catalog: www.oie.int. 7

OIE (2017). World Animal Health Information Database (WAHIS) Interface. 12

Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., and Goss, C. H. (2011). Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. *PloS One*, 6(4):e18687. 8

Ostfeld, R. S. (2009). Biodiversity loss and the rise of zoonotic pathogens. *Clinical Microbiology and Infection*, 15:40–43. 146

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. 145

Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., Koppeschaar, C., Rehn, M., Smallenburg, R., Turbelin, C., Van Noort, S., and Vespignani, A. (2014). Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, 20(1):17–21. 6

Paquet, C., Coulombier, D., Kaiser, R., and Ciotti, M. (2006). Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Eurosurveillance*, 11(12):5–6. xxiii, 2, 9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830. 61, 179

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. 60, 64

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365 [cs]*. arXiv: 1802.05365. 61

Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E., and Zavarella, V. (2011). Online news event extraction for global crisis surveillance. In Nguyen, N. T., editor, *Transactions on Computational Collective Intelligence V*, volume 6910, pages 182–212. Springer Berlin Heidelberg, Berlin, Heidelberg. 29, 87

Poglayen, G., Baldelli, R., and Battelli, G. (2008). Zoonoses and information of the public: the role of media, with special reference to Italy. *Vet Ital*, 44:6. 16

Preoţiuc-Pietro, D., Srijith, P. K., Hepple, M., and Cohn, T. (2016). Studying the Temporal Dynamics of Word Co-occurrences: An Application to Event Detection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4380–4387, Portorož, Slovenia. European Language Resources Association (ELRA). xxviii, 81, 141

Pyysalo, S., Ohta, T., Rak, R., Rowley, A., Chun, H.-W., Jung, S.-J., Choi, S.-P., Tsujii, J., and Ananiadou, S. (2015). Overview of the Cancer Genetics and Pathway Curation tasks of BioNLP Shared Task 2013. *BMC Bioinformatics*, 16(10):S2. 29, 60

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics. 144

Rabatel, J., Arsevska, E., and Roche, M. (2019). PADI-web corpus: Labeled textual data in animal health domain. *Data in Brief*, 22:643–646. 85

Ralf, S., Flavio, F., Erik, v. d. G., Clive, B., Peter, v. E., and Roman, Y. (2008). Text Mining from the Web for Medical Intelligence. *NATO Science for Peace and Security Series, D: Information and Communication Security*, pages 295–310. 21, 28

Ranwez, S., Duthil, B., Sy, M.-F., Montmain, J., Augereau, P., and Ranwez, V. (2013). How ontology based information retrieval systems may benefit from lexical text analysis. In Oltramari, Vossen, A., Qin, P., Hovy, L., and Eduard, editors, *New Trends of Research in Ontologies and Lexical Resources*, Theory and Applications of Natural Language Processing, pages 209–230. Springer. 57

Research, M. (2018). Customized neural machine translation with Microsoft Translator. Library Catalog: www.microsoft.com. 20, 178

Rezza, G., Marino, R., Farchi, F., Taranto, M., and Superiore di Sanità, I. (2004). SARS Epidemic in the Press. *Emerging Infectious Diseases*, 10(2):381–382. 15

Rich, K. M. and Perry, B. D. (2011). The economic and poverty impacts of animal diseases in developing countries: New roles, new demands for economics and epidemiology. *Preventive Veterinary Medicine*, 101(3):133–147. xxiii, 1

Richardson, L. (2007). Beautiful soup documentation. *April*. 178

Robertson, C. and Yee, L. (2016). Avian Influenza Risk Surveillance in North America with Online Media. *PLOS ONE*, 11(11):e0165688. 8, 12

Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146. Publisher: John Wiley and Sons, Ltd. 107

Roche, M., Azé, J., Kodratoff, Y., and Sebag, M. (2004). Learning interestingness measures in terminology extraction. A ROC-based approach. In *ROCAI*, pages 81–88. 82

Roche, M. and Prince, V. (2010). A Web-Mining Approach to Disambiguate Biomedical Acronym Expansions. *Informatica*, 34:12. 81

Rolland, C., Lazarus, C., Giese, C., Monate, B., Travert, A.-S., and Salomon, J. (2020). Early Detection of Public Health Emergencies of International Concern through Undiagnosed Disease Reports in ProMED-Mail. *Emerging Infectious Disease journal*, 26(2):336. 131

Rotureau, B., Barboza, P., Tarantola, A., and Paquet, C. (2007). International Epidemic Intelligence at the Institut de Veille Sanitaire, France. *Emerging Infectious Diseases*, 13(10):1590–1592. 10

Rowlands, R. J., Michaud, V., Heath, L., Hutchings, G., Oura, C., Vosloo, W., Dwarka, R., Onashvili, T., Albina, E., and Dixon, L. K. (2008). African Swine Fever Virus Isolate, Georgia, 2007. *Emerging Infectious Diseases journal*, 14(12). 7

Rushton, J., Viscarra, R., Guerne Bleich, E., and McLeod, A. (2005). Impact of avian influenza outbreaks in the poultry sectors of five South East Asian countries (Cambodia, Indonesia, Lao PDR, Thailand, Viet Nam) outbreak costs, responses and potential long term control. *World's Poultry Science Journal*, 61(3):491–514. Publisher: Cambridge University Press on behalf of World's Poultry Science Association. 1

Salton, G. (1971). *The SMART Retrieval System-Experiments in Automatic Document Processing*. Prentice-Hall, Inc., USA. 55

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523. 58

Salton, G. and Lesk, M. E. (1968). Computer Evaluation of Indexing and Text Processing. *Journal of the Association for Computing Machinery*, 15(1):8–36. 88

Sbattella, L. and Tedesco, R. (2013). A novel semantic information retrieval system based on a three-level domain model. *Journal of Systems and Software*, 86(5):1426–1452. 57

Schmid, H. (1994). Probabilistic Part-Of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49. 32

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania. 97

Schwind, J. S., Norman, S. A., Karmacharya, D., Wolking, D. J., Dixit, S. M., Rajbhandari, R. M., Mekaru, S. R., and Brownstein, J. S. (2017). Online surveillance of media health event reporting in Nepal: digital disease detection from a One Health perspective. *BMC International Health and Human Rights*, 17(1). 16

Seeland, M., Rzanny, M., Alaqraa, N., Wäldchen, J., and Mäder, P. (2017). Plant species classification using flower images—A comparative study of local feature representations. *PLOS ONE*, 12(2):e0170629. 107

Seimenis, A. M. (2008). The spread of zoonoses and other infectious diseases through the international trade of animals and animal products. *Veterinaria Italiana*, 44(4):591–599. 1

Shamma, D. A., Kennedy, L., and Churchill, E. F. (2011). Peaks and Persistence: Modeling the Shape of Microblog Conversations. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 355–358, New York, NY, USA. Association for Computing Machinery. event-place: Hangzhou, China. 143

Sharpe, J. D., Hopkins, R. S., Cook, R. L., and Striley, C. W. (2016). Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis. *JMIR Public Health and Surveillance*, 2(2). 9

Smadja, F., Hatzivassiloglou, V., and McKeown, K. R. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 2(1). 82

Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM. 108

Song, S., Zhang, N., and Huang, H. (2019). Named entity recognition based on conditional random fields. *Cluster Computing*, 22:1–12. 27

Soriano-Morales, E.-P. (2018). *Hypergraphs and information fusion for term representation enrichment. Applications to named entity recognition and word sense disambiguation.* PhD thesis, Université de Lyon - Lumière Lyon 2. 107, 108

Soto, G., Araujo-Castillo, R. V., Neyra, J., Fernandez, M., Leturia, C., Mundaca, C. C., and Blazes, D. L. (2008). Challenges in the implementation of an electronic surveillance system in a resource-limited setting: Alerta, in Peru. In *BMC proceedings*, volume 2, page S4. BioMed Central. 1

Steinberger, R., Fuart, F., Goot, E., Best, C., Etter, P., and Yangarber, R. (2008). Text Mining from the Web for Medical Intelligence. In *Mining Massive Data Sets for Security*. IOS Press. 21, 22, 31

Strotgen, J. and Gertz, M. (2010). HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. 27, 80, 109, 178

Tatem, A. J., Rogers, D. J., and Hay, S. I. (2006). Global Transport Networks and Infectious Disease Spread. In Hay, S. I., Graham, A., and Rogers, D. J., editors, *Advances in Parasitology*, volume 62 of *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*, pages 293–343. Academic Press. xxiii, 1

Taylor, L. H., Latham, S. M., and Woolhouse, M. E. (2001). Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1411):983–989. xxiii, 1

Thapen, N., Simmie, D., and Hankin, C. (2016a). The early bird catches the term: combining twitter and news data for event detection and situational awareness. *Journal of Biomedical Semantics*, 7(1):61. 67

Thapen, N., Simmie, D., Hankin, C., and Gillard, J. (2016b). DEFENDER: Detecting and Forecasting Epidemics Using Novel Data-Analytics for Enhanced Response. *PLOS ONE*, 11(5):e0155417. 9, 141

Thelen, M. and Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, pages 214–221, Not Known. Association for Computational Linguistics. 69

Thrush, M. A., Dunn, P. L., and Peeler, E. J. (2012). Monitoring Emerging Diseases of Fish and Shellfish Using Electronic Sources: Monitoring Emerging Fish and Shellfish Diseases. *Transboundary and Emerging Diseases*, 59(5):385–394. 16

Torii, M., Yin, L., Nguyen, T., Mazumdar, C. T., Liu, H., Hartley, D. M., and Nelson, N. P. (2011). An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. *International Journal of Medical Informatics*, 80(1):56–66. 39, 40

Torkkola, K. (2003). Feature Extraction by Non Parametric Mutual Information Maximization. *J. Mach. Learn. Res.*, 3:1415–1438. Publisher: JMLR.org. 81

Trielli, D. and Diakopoulos, N. (2019). Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–15, Glasgow, Scotland Uk. ACM Press. 146

Tsai, F.-J., Tseng, E., Chan, C.-C., Tamashiro, H., Motamed, S., and Rougemont, A. C. (2013). Is the reporting timeliness gap for avian flu and H1N1 outbreaks in global health surveillance systems associated with country transparency? *Globalization and health*, 9(1):14. 8

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188. 111

Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, 50(1):104–112. 61, 104, 107

Valentin, S., Arsevska, E., Falala, S., de Goër, J., Lancelot, R., Mercier, A., Rabatel, J., and Roche, M. (2020a). PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169:105163. xxiv, 39, 42, 132, 177

Valentin, S., Arsevska, E., Mercier, A., Falala, S., Rabatel, J., Lancelot, R., and Roche, M. (2020b). PADI-web: an event-based surveillance system for detecting, classifying and processing online news. In *Post-Proceedings of 8th Language and Technology Conference, LTC 2017, November 17-19, 2017*, Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence - Springer, Poznań, Poland. 143

Valentin, S., De Waele, V., Vilain, A., Arsevska, E., Lancelot, R., and Roche, M. (2019). Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus. *Dataverse Cirad*. type: dataset. 42, 43

Valentin, S., Lancelot, R., and Roche, M. (2020c). Automated Processing of Multilingual Online News for the Monitoring of Animal Infectious Diseases. In *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, pages 33–36, Marseille, France. European Language Resources Association. 2, 14, 20, 21

Valentin, S., Mercier, A., Lancelot, R., Roche, M., and Arsevska, E. (2020d). Monitoring online media reports for early detection of unknown diseases: insight from a retrospective study of COVID-19 emergence. *Transboundary and Emerging Diseases*, to appear. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tbed.13738. 131

Valentin, S., Mercier, A., Lancelot, R., Roche, M., and Arsevska, E. (2020e). PADI-web COVID-19 corpus: news articles manually labelled. *Dataverse Cirad*. type: dataset. 141

Vallat, B., Thiermann, A., Ben Jebara, K., and Dehove, A. (2013). Notification of animal and human diseases: the global legal basis. *Rev. sci. tech. Off. int. Epiz*, 32(2):331–335. 1

Velasco, E., Agheneza, T., Denecke, K., Kirchner, G., and Eckmanns, T. (2014). Social media and Internet-Based data in global systems for public health surveillance: A systematic review. *The Milbank Quarterly*, 92(1):7–33. 5, 142, 145

Wagner, R. A. and Fischer, M. J. (1974). The String-to-String Correction Problem. *Journal of the ACM*, 21(1):168–173. 21

Wang, Y., McKee, M., Torbica, A., and Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science and Medicine*, 240:112552. 144

Weber, M. S. and Monge, P. (2011). The Flow of Digital News in a Network of Sources, Authorities, and Hubs. *Journal of Communication*, 61(6):1062–1081. 120, 123, 130

WHO (2005). *International Health Regulation (2005)*. WHO Press, Geneva, 3rd edition. xxiii, 2, 7

WHO (2014). *Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance*. WHO Press, Geneva: The Organization, interim version edition. 7, 10, 11

WHO (2020). WHO Coronavirus Disease (COVID-19) Dashboard. Library Catalog: covid19.who.int. xxiii, 1

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Towards Universal Paraphrastic Sentence Embeddings. *arXiv:1511.08198 [cs]*. arXiv: 1511.08198. 60

Wilbur, W. J., Rzhetsky, A., and Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7(1):356. 40

Wilburn, J., O'Connor, C., Walsh, A. L., and Morgan, D. (2019). Identifying potential emerging threats through epidemic intelligence activities—looking for the needle in the haystack? *International Journal of Infectious Diseases*, 89:146–153. 145

Wilson, K. and Brownstein, J. S. (2009). Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal*, 180(8):829–831. 1, 33

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. 55

Wu, L., Morstatter, F., and Liu, H. (2018). SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52(3):839–852. 77

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Łukasz., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*. arXiv: 1609.08144. 64

Wuhan Municipal Health Commision (2020). Information session of the Wuhan Municipal Health and Health Commission on the current situation of the pneumonia epidemic in our city. 133

Xiang, W. and Wang, B. (2019). A Survey of Event Extraction From Text. *IEEE Access*, 7:173111–173137. 29

Yangarber, R. (2003). Counter-training in discovery of semantic patterns. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 343–350. Association for Computational Linguistics. 69

Yin, Z. and Shen, Y. (2018). On the Dimensionality of Word Embedding. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 887–898. Curran Associates, Inc. 64

Yoon, J. (2012). Detecting Weak Signals for Long-term Business Opportunities Using Text Mining of Web News. *Expert Syst. Appl.*, 39(16):12543–12550. xxviii

Zhang, Y., Dang, Y., Chen, H., Thurmond, M., and Larson, C. (2009). Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems*, 47(4):508–517. 39

Zhang, Y., Jin, R., and Zhou, Z. (2010). Understanding bag-of-words model: A statistical framework. *International journal of machine learning and cybernetics*, 1(1-4):43–52. 56

Zhang, Y. and Liu, B. (2007). Semantic text classification of emergent disease reports. In *Proceedings of the 11th European Conference on Principles and Pratice of Knockledge Discovery in Databases (PKDD)*, Warsaw, Poland. Springer. 40, 61

Zhao, R. and Mao, K. (2018). Fuzzy Bag-of-Words Model for Document Representation. *IEEE Transactions on Fuzzy Systems*, 26(2):794–804. Conference Name: IEEE Transactions on Fuzzy Systems. 57

Zhu, L. and Zheng, H. (2020). Biomedical event extraction with a novel combination strategy based on hybrid deep neural networks. *BMC Bioinformatics*, 21(1):47. 29

Zinsstag, J., Schelling, E., Waltner-Toews, D., and Tanner, M. (2011). From "one medicine" to "one health" and systemic approaches to health and well-being. *Preventive Veterinary Medicine*, 101(3-4):148–156. 144

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. 60

Șerban, O., Thapen, N., Maginnis, B., Hankin, C., and Foot, V. (2019). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing and Management*, 56(3):1166–1184. 8

APPENDIX

Appendix A

Indicator-based surveillance systems

Table 35 Characteristics of IBS systems encompassing animal health threats.

Systems	WAHIS	Empres-i	ADNS
General caracteristics			
Year launched	1996	2004	1998
Institution	OIE	FAO	EC
Geographical coverage	World	World	Europe
Type of cases	Animal	Animal, human (zoonoses)	Animal
IBS process			
Data detection			
Data sources	OIE Members countries.	OIE (WAHIS), WHO, FAO, etc.	European countries.
Moderation	Automatic and manual	Automatic and manual	Automatic
Data triage			
Selection criteria	Diseases notifiable to OIE by OIE Members.	Events detected by a specific range of sources.	Diseases notifiable to ADNS by European countries.
Moderation	Automatic	Automatic	Automatic
Signal validation			
Moderation	Manual	Manual	No
Risk assessment			
Spatial representation	Static map	Dynamic map	Static map
Data analysis	Static tables, qualitative status	Dynamic charts and figures	No
Additional data	Yes	Geographical layers, genetic information	
Communication			
Interface (access)	Open-access	Open-access	Restricted
Interface (langages)	3 (EN, FR, SP)	1 (EN)	1 (EN)
Export of structured data	No	Yes (csv, pdf, maps)	Yes
Automatic alerts	Yes	No	Yes
Alerts format	E-mail, smartphone app	No	E-mail

Appendix B

PADI-web pipeline

The PADI-web pipeline involves four steps ranging from online news collection to the extraction of epidemiological features: (1) data collection, (2) data processing, (3) data classification and (4) information extraction (Figure 40). All these steps are detailed in (Arsevska et al., 2018; Valentin et al., 2020a).

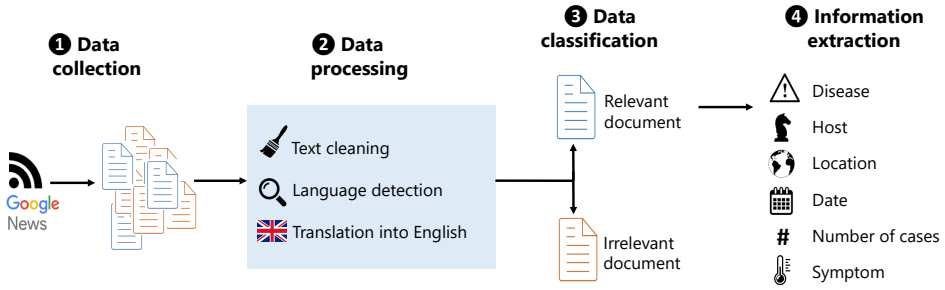


Figure 40 PADI-web pipeline.

1. Data collection

PADI-web collects news articles from Google News through customised really simple syndication (RSS) feeds daily. Each RSS feed consists of an association of terms (disease names, clinical signs or hosts) extracted by an integrated approach combining text mining and domain experts. PADI-web currently uses two types of RSS feed:

- Disease-based RSS feeds include disease terms (e.g. avian flu OR avian influenza OR bird flu) and target seven animal diseases
- Symptom-based RSS feeds consist of combinations of clinical signs and hosts (e.g. abortions AND cows) and do not include any disease term. Those feeds enable detection of diseases

that are not explicitly monitored or not confirmed at the time of article release, as well as unknown hazards (Arsevska et al., 2016).

RSS feeds are implemented in 7 langages (English, Chinese, Arabic, Italian, French, Russian, Turkish).

2. Data processing

To avoid duplicates, PADI-web checks if retrieved articles already exist in the database based on the URL. For each news article that is not a duplicate, the corresponding webpage is visited to fetch its content. The title and text are cleaned to remove irrelevant elements (pictures, ads, hyperlinks, etc.), using the BeautifulSoup python library (Richardson, 2007). The *langdetect* python library is used to detect the language of the article. Then all news articles that are not in English are translated using the Translator API of the Microsoft Azure system (Research, 2018).

3. Data classification

The data classification step aims to select relevant news, which is subsequently processed by the information extraction module. By relevant news, we mean a news report that is related to a disease event (describing a current outbreak as well as prevention and control measures, preparedness, socioeconomic impacts, etc.). The new classification module uses a supervised machine learning approach described in Appendix C. The pipeline is generic, allowing users to easily create new classification tasks that are independent of each other.

4. Information extraction

The final step aims to extract epidemiological entities from the relevant news content. The information extraction process relies on a combined method founded on rule-based systems and data mining techniques. Briefly, to extract names of diseases, hosts and symptoms, PADI-web relies on a vocabulary which was created using text mining methods and validated by domain experts (Arsevska et al., 2016). Locations are identified by matching the text with location names from the GeoNames gazetteer (Ahlers, 2013) and dates with the rule-based HeidelTime system (Strotgen and Gertz, 2010). The number of cases is extracted from a list of regular expressions matching numbers in numerical or textual form. A confidence index is automatically assigned to the extracted entities to reflect the probability that they correspond to the desired piece of epidemiological information.

Appendix C

PADI-web classification module

The new classification module of PADI-web uses a supervised machine learning approach and heavily relies on the scikit-learn python library (Pedregosa et al., 2011). It relies on two steps, outlined in Figure 41.

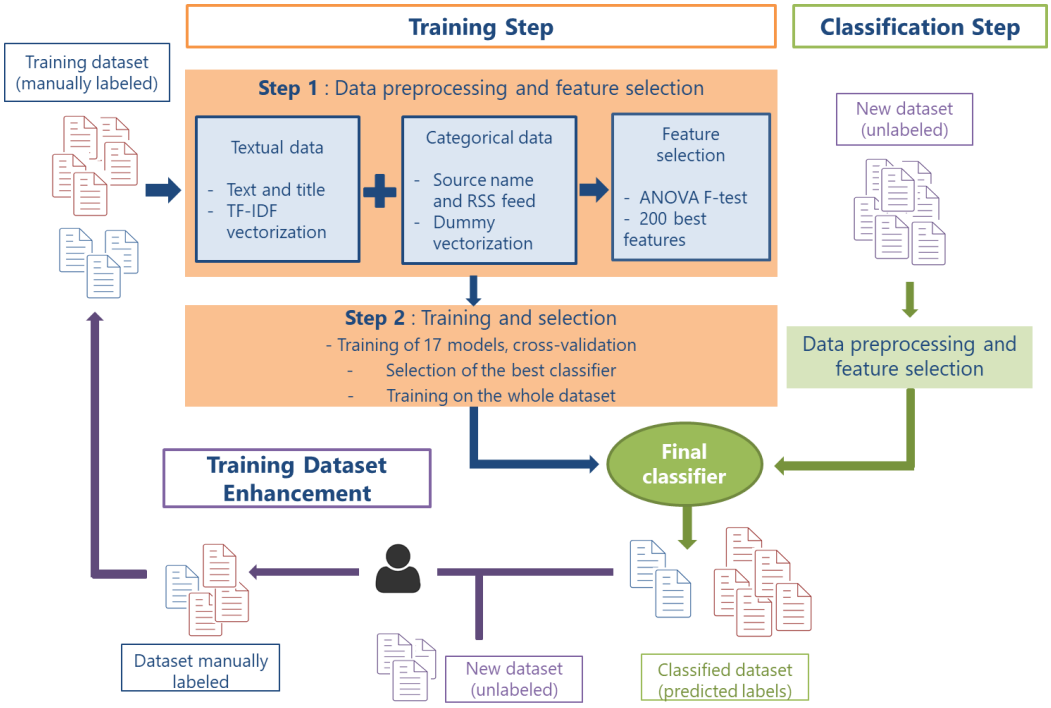


Figure 41 PADI-web classification module.

1. Training and classification

The training dataset is a corpus of 600 news items randomly retrieved from the PADI-web database and manually labelled by an epidemiology expert (200 relevant news articles and 400 irrelevant news articles). Data preprocessing involves the vectorization of both textual (text and title) and categorical data (source and RSS feed), followed by a feature selection step (based on a one-way ANOVA F-test). A selection of model families is trained on the dataset (random forests with various parameters, linear support vector classification, neural networks, Gaussian-based models, etc.), using a 5-fold cross-validation scheme. The model obtaining the highest mean accuracy score along the folds is trained on the whole dataset and is subsequently used to classify each new retrieved article. Currently Random Forest is the best classifier built with this methodology (composed of 50 trees with a maximum depth of 12), which obtains an average accuracy score of 94.2% through the 5-fold cross-validation process. We evaluated the classifier performance in a new set of 100 articles (external validation) that was preannotated by two epidemiologists (gold standard). The classifier obtained a 92% accuracy score.

2. Training dataset enhancement

The PADI-web interface allows users to label the relevance of the retrieved news articles manually. If different from the classifier label, the user label prevails over the classifier label. Each manually labelled article is added to the initial learning dataset. Therefore, each new Training Step relies on the initial dataset enriched with the user contribution. It allows a quick increase in the learning dataset size and adaptation of the classifier if users detect significant classification bias or errors. This feature is of primary importance since we are dealing with textual data that are prone to rapid change (e.g. onset of a new disease).

Table 36 Relevance classification results in terms of accuracy.

Classifier	Parameters*	Accuracy
LR	-	0.90 (+/-0.02)
NB	distribution: gaussian	0.895 (+/-0.03)
NB	distribution: multinomial	0.885 (+/-0.03)
SVM	kernel: linear	0.940 (+/-0.02)
RF	trees: 50, max. depth:12	0.942 (+/-0.01)
RF	trees: 50, max. depth:15	0.933 (+/-0.01)
RF	trees: 200, max. depth:3	0.927 (+/-0.01)
RF	trees: 200, max. depth:5	0.936 (+/-0.02)
MLP	-	0.93 (+/-0.02)

*: if not specified, scikitlearn library default parameters.

LR: Logistic Regression, NB: Naive Bayes, RF: Random Forest, SVM: Support Vector Machine, MLP: MultiLayer Perceptron

Appendix D

Seed sentences and corresponding patterns for the category Transmission pathway

Seed-sentence	Pattern
Zhu Zengyong [...] said the three cases may be related, with cross-infection occurring during transportation of live pigs.	NOUN(infection) VERB(occured) ADV(during)
The most likely route of infection is said to be discarded infected meat products brought in from ASF-infected areas.	NOUN(route) of NOUN(infection)
Salvage slaughter of infected pigs and selling of the pork at discount prices and movement of pigs from one village to the nearby villages to save them have been the major routes of spread.	NOUN(routes) of NOUN(spread)
While the main source of the outbreaks remains inconclusive contact with wild species has been determined as the primary cause of the outbreaks.	NOUN(source) of the NOUN(outbreak) NOUN(cause) of the NOUN(outbreak)
Investigations into how the animal became infected will focus on the feed supplied during the first year of its life.	ADV(how) ENTITY(host) VERB (became infected)
The Norwegian Veterinary Institute use genotyping to determine whether ISA outbreaks are an occurrence or a reoccurrence.	-
"The movement of pig products can spread disease quickly and, as in these recent cases, it's likely that the virus was spread to other parts of China by the movement of such products, rather than live pigs, "[...].	NOUN(virus) VERB(was spread) ADV(by)
However, the possibility that it could have been brought here by soldiers is now being discussed.	VERB(have been brought) ADV(by)
The source of the outbreak is still unknown.	NOUN(source) of the NOUN(outbreak)
Agriculture deputy minister livestock Paddy Zhanda yesterday said government had stopped issuing out new licences for abattoirs accusing them of spreading foot-and-mouth disease .	VERB(accusing) ADV(of) VERB(spreading) the NOUN(disease)

Appendix E

Seed sentences and corresponding patterns for the category Concern and risk factors

Seed-sentence	Pattern
Pig producers have been warned about an increased chance of African Swine Fever (ASF) descending on the UK.	VERB(have been warned)
[...], the spread of the disease in eastern Europe has been causing immense concern in Germany.	NOUN(concern)
It said Heves county was the most at risk from the disease at the moment.	NOUN(risk)
[...] it is essential that farmers, vets and government agencies remain vigilant to the threat of disease spread, given the mild weather we have been experiencing in the UK.	ADJ(vigilant) NOUN(threat)
Until the origin of disease is understood and the extent of spread it is difficult to assess whether this outbreak signifies an increase in our risk level from low to medium .	NOUN(risk)
Concerns were originally raised about the existence of the virus in October, when the past exposure to bluetongue virus was detected in several 12-month-old dairy heifers on a property near Bamawm.	NOUN(concern)
Nearby Busan and Ulsan also emergency alert rush	NOUN(alert)
Zhanda warned others were on surveillance with a view to containing the disease.	VERB(warned)
But experts have warned further outbreaks could emerge in the coming days.	VERB(warned) MD(could) VERB(emerge)
Hungary, Russia, Poland, Ukraine and Romania are among the countries affected, alarming governments and pig farmers due the pace at which it has spread.	VERB(alarming)
The recent detection of bluetongue [...] highlights to farmers the risks that come with bringing animals from disease-affected areas into their flocks and herds.	NOUN(risk)
We have also informed the EFSA and the European Centre for Disease Prevention and Control; atypical cases do not threaten food safety.	VERB(threaten)
Bluetongue virus (BTV) is prevalent in Europe and is a threat to Irish ruminants, in an update on the virus this week.	NOUN(threat)
The tough new measures risk incentivizing the illegal movement of pigs as farmers and traders struggle to maintain their livelihoods, said an animal health expert at a large pig producer who was not authorized to talk to the media.	NOUN(risk)

RÉSUMÉ

Extraction et combinaison d'informations épidémiologiques à partir de sources informelles pour la veille des maladies infectieuses animales

L'intelligence épidémiologique a pour but de détecter, d'analyser et de surveiller au cours du temps les potentielles menaces sanitaires. Ce processus de surveillance repose sur des sources dites formelles, tels que les organismes de santé officiels, et des sources dites informelles, comme les médias. La veille des sources informelles est réalisée au travers de la surveillance basée sur les événements (event-based surveillance en anglais). Ce type de veille requiert le développement d'outils dédiés à la collecte et au traitement de données textuelles non structurées publiées sur le Web. Cette thèse se concentre sur l'extraction et la combinaison d'informations épidémiologiques extraites d'articles de presse en ligne, dans le cadre de la veille des maladies infectieuses animales. Le premier objectif de cette thèse est de proposer et de comparer des approches pour améliorer l'identification et l'extraction d'informations épidémiologiques pertinentes à partir du contenu d'articles. Le second objectif est d'étudier l'utilisation de descripteurs épidémiologiques (i.e. maladies, hôtes, localisations et dates) dans le contexte de l'extraction d'événements et de la mise en relation d'articles similaires au regard de leur contenu épidémiologique. Dans ce manuscrit, nous proposons de nouvelles représentations textuelles fondées sur la sélection, l'expansion et la combinaison de descripteurs épidémiologiques. Nous montrons que l'adaptation et l'extension de méthodes de fouille de texte et de classification permet d'améliorer l'utilisation des articles en ligne tant que source de données sanitaires. Nous mettons en évidence le rôle de l'expertise quant à la pertinence et l'interprétabilité de certaines des approches proposées. Bien que nos travaux soient menés dans le contexte de la surveillance de maladies en santé animale, nous discutons des aspects génériques des méthodes proposées, vis-à-vis de de maladies inconnues et dans un contexte One Health (« une seule santé »).

Keyword : intelligence épidémiologique, fouille de texte, articles en ligne, santé animale, données textuelles non-structurées.

ABSTRACT

Extraction and combination of epidemiological information from informal sources for animal infectious diseases surveillance.

Epidemic intelligence aims to detect, investigate and monitor potential health threats while relying on formal (e.g. official health authorities) and informal (e.g. media) information sources. Monitoring of unofficial sources, or so-called event-based surveillance (EBS), requires the development of systems designed to retrieve and process unstructured textual data published online. This manuscript focuses on the extraction and combination of epidemiological information from informal sources (i.e. online news), in the context of the international surveillance of animal infectious diseases. The first objective of this thesis is to propose and compare approaches to enhance the identification and extraction of relevant epidemiological information from the content of online news. The second objective is to study the use of epidemiological entities extracted from the news articles (i.e. diseases, hosts, locations and dates) in the context of event extraction and retrieval of related online news. This manuscript proposes new textual representation approaches by selecting, expanding, and combining relevant epidemiological features. We show that adapting and extending text mining and classification methods improve the added value of online news sources for event-based surveillance. We stress the role of domain expert knowledge regarding the relevance and the interpretability of methods proposed in this thesis. While our researches are conducted in the context of animal disease surveillance, we discuss the generic aspects of our approaches regarding unknown threats and One Health surveillance.

Keyword : epidemic intelligence, text mining, online news, animal health, unstructured textual data.