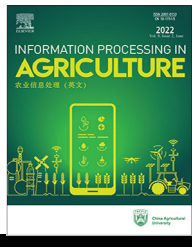
Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE 10 (2023) 347–360

journal homepage: www.keaipublishing.com/en/journals/information-processing-in-agriculture/



Fusion of spatiotemporal and thematic features of textual data for animal disease surveillance

Sarah Valentin [a,b,c,d], *Renaud Lancelot* [a,b], *Mathieu Roche* [a,c,*]

[a] *CIRAD, F-34398 Montpellier, France*
[b] *ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France*
[c] *TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France*
[d] *Département de Biologie, Université de Sherbrooke, Sherbrooke, Quebec, Canada*

ARTICLE INFO

Article history:
Received 9 March 2021
Received in revised form
14 March 2022
Accepted 24 March 2022
Available online 28 March 2022

Keywords:
Animal disease surveillance
Text mining
Ranking
Fusion

ABSTRACT

Several internet-based surveillance systems have been created to monitor the web for animal health surveillance. These systems collect a large amount of news dealing with outbreaks related to animal diseases. Automatically identifying news articles that describe the same outbreak event is a key step to quickly detect relevant epidemiological information while alleviating manual curation of news content. This paper addresses the task of retrieving news articles that are related in epidemiological terms. We tackle this issue using text mining and feature fusion methods. The main objective of this paper is to identify a textual representation in which two articles that share the same epidemiological content are close. We compared two types of representations (i.e., features) to represent the documents: (i) morphosyntactic features (i.e., selection and transformation of all terms from the news, based on classical textual processing steps) and (ii) lexicosemantic features (i.e., selection, transformation and fusion of epidemiological terms including diseases, hosts, locations and dates). We compared two types of term weighing (i.e., Boolean and TF-IDF) for both representations. To combine and transform lexicosemantic features, we compared two data fusion techniques (i.e., early fusion and late fusion) and the effect of features generalisation, while evaluating the relative importance of each type of feature. We conducted our analysis using a corpus composed of a subset of news articles in English related to animal disease outbreaks. Our results showed that the combination of relevant lexicosemantic (epidemiological) features using fusion methods improves classical morphosyntactic representation in the context of disease-related news retrieval. The lexicosemantic representation based on TF-IDF and feature generalisation (F-measure = 0.92, r-precision = 0.58) outperformed the morphosyntactic representation (F-measure = 0.89, r-precision = 0.45), while reducing the features space. Converting the features into lower granular features (i.e., generalisation) contributed to improving the results of the lexicosemantic representation. Our results showed no difference between the early and late fusion approaches. Temporal features performed poorly on their own. Conversely, spatial features were the most discriminative features, highlighting the need for robust methods for spatial entity extraction, disambiguation and representation in internet-based surveillance systems.

* Corresponding author at: CIRAD, Campus de Lavalette, Avenue Agropolis, 34398 Montpellier Cedex 5, France.
E-mail address: mathieu.roche@cirad.fr (M. Roche).

Peer review under responsibility of China Agricultural University.
https://doi.org/10.1016/j.inpa.2022.03.004

1. Introduction

The globalisation of 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 countries and continents [1,2]. Animal diseases have detrimental impacts on animal health and the economy in terms of lost revenues and societal costs [3]. Some diseases have the potential to rapidly kill large numbers of animals (e.g., avian influenza and African swine fever). Furthermore, other animal diseases also prompt significant drops in the demand for animal products through consumer fears of becoming infected with zoonotic diseases (e.g., avian influenza) [4]. In recent decades, concern regarding so-called zoonotic diseases, which are caused by pathogen agents shared by animals and humans - a common situation - has been growing [5]. In this context, management of data dealing with animal health is challenging for improving health surveillance systems.

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" [6]. 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)" [6]. It corresponds to conventional surveillance of formal sources and is based on established case definitions. Event-based surveillance is defined by the WHO (World Health Organization) 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" [6]. EBS involves the use of data streams from informal sources. In the medical domain, the extraction and use of epidemiological indicators from new data sources have become a hot research topic over the last few years. A large range of sources can be used in human medicine, such as chief complaints [7], electronic medical records [8] or more informal sources, such as social media [9]. Symptoms are extracted from unstructured textual data and gathered into syndromes, manually or through text mining methods. A syndrome can be defined as "a combination of clinical signs that repeatedly occurs in different observations, indicating a possible presence of disease" [10]. Monitoring those syndromes should allow the detection of an outbreak before an outbreak has been diagnosed. However, optimal syndrome definitions adapted to each specific data source have not yet been determined [11]. The syndromic surveillance approaches are mostly cumulative: an alert is created when a deviation from a baseline level appears (for example, an increase in cattle mortality). Other initiatives focus on noncumulative approaches that can detect weak signals. For instance, weighted vectors were created from health-related tweets to gather these tweets into clusters sharing the same content [8]. This approach aimed to detect "latent infectious disease": when a tweet, represented by its vector, could not be associated with an existing cluster, it was a candidate for a potential emerging health event. The association of several indicators has been studied to improve surveillance [12,13]. In veterinary medicine, data sources are not as numerous and can be more challenging to obtain. Moreover, due to the specificity of social media users (i.e., the patients themselves), there is no interest in using those sources to extract animal symptoms. Therefore, clinical data from practitioners and laboratory data are currently the main sources used in animal syndromic surveillance [14]. Nevertheless, an increasing number of publications evaluate the potential of other data sources, ranging from production data [15] to disease outbreak news [16].

Several internet-based surveillance systems were created to monitor disease outbreak news for potential public health threats [17]: Argus, BioCaster, GPHIN (Global Public Health Intelligence Network), IBIS (International Biosurveillance System) and MedISys. However, none of them is specifically dedicated to animal disease surveillance. In order to assist the French epidemic intelligence team in the international monitoring of animal health, a platform dedicated to automatic surveillance of electronic media, PADI-web (Platform for Automated extraction of animal Disease Information from the web), has been implemented [18]. This EBS tool automatically detects, classifies, and extracts epidemiological clues from news (i.e., disease, hosts, symptoms, dates, and locations) [19]. Since its implementation, a large volume of articles has been retrieved (i.e., more than 380 000 articles).

In these different systems, one of the major challenges in the news article processing pipeline is to identify articles that are related to each other, i.e., that describe the same outbreak. In this paper that is an extension of a preliminary study [20], we address the task of linking news articles, which are related in epidemiological terms, collected by PADI-web. Finding the most relevant documents regarding a chosen document (related documents) can be viewed from the perspective of the document ranking. Document ranking first requires the transformation of texts into vectors, where each component of a vector represents the weight of a feature in a document. Selecting the features used to represent the texts is essential for information retrieval tasks in animal health event-based systems and other fields such as genomics [21]. Therefore, this paper aims to identify the best textual representation (i.e., features) to associate articles that share epidemiological content.

Section 2 presents our process to combine the epidemiological features and our evaluation approach. Sections 4 and 5 discuss the results obtained with different fusion methods.

2. Material and methods

2.1. Morphosyntactic features

In the information retrieval domain, document ranking generally consists of computing numeric scores between a query and documents to retrieve the more relevant documents [22]. A common approach is to convert the text into a structured representation such as the vector space model [23]. This model relies on creating a link between textual content and linear algebra area analysis tools. This representation encodes a document in a k-dimensional space where each component w_{ij} represents the weight of term j in document i. The text's grammatical structure is neglected; this is also referred to as a bag-of-words representation [24]. The most basic weight method uses a binary value (i.e., 1 or 0) to represent whether the term is present in the document. Another approach is the term frequency–inverse document frequency (TF-IDF) function, which calculates the weight w_{ij} (Eq. (1)).

$$w_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\left(\frac{N}{df_j}\right) \quad (1)$$

where tf_{ij} is the frequency of term j in document i, N is the total number of documents in the corpus, and df_j is the number of documents containing the term j.

A large range of measures can then be applied to compute the similarity score between the vectors, e.g., the Euclidian distance, Jaccard coefficient, or cosine similarity [25,26]. The cosine similarity used in this work (Eq. (2)) between two documents D_1 and D_2 is calculated as follows [27]:

$$sim(D_1, D_2) = \frac{\sum_{i=1}^{V} w_{i1} \times w_{i2}}{\sqrt{\sum_{i=1}^{V} w_{i1}^2 \times w_{i2}^2}} \quad (2)$$

where w_{i1} is the weight of term i in document D_1, w_{i2} is the weight of term i in document D_2, and V is the total number of terms (features).

In our work, the morphosyntactic representation is based on the bag-of-words (BOW) model. The text is first segmented into words (tokenisation). Then, several steps known as "preprocessing" are applied to remove noisy elements from the vocabulary (Fig. 1). Preprocessing steps include removing stop words, part-of-speech (POS) tagging, lemmatisation and normalisation to lowercase [28,29]. 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 consists of tagging the words according to their syntactic functions, i.e., verbs, nouns and proper nouns [30]. Lemmatisation consists of transforming the different inflected forms of a word into its canonical form (e.g., singular form for nouns) to be analysed as a single item. Both lowercasing and lemmatisation aim to compute 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 [31]. To remove stop words, we used the list of 318 English stop words provided by the library. After preprocessing, the features can be selected according to their POS tag (Fig. 1). For instance, only the terms that are verbs are retained. The methods mentioned above are illustrated in Table 1. There is no gold standard procedure for preprocessing steps that should be applied to a text corpus. The choice of whether to include normalisation depends on the task and the nature of the corpus (e.g., its language) [28]. Thus, we compared different preprocessing and feature selection combinations.

As term weighting (Fig. 1), we used the Boolean and TF-IDF weights (Eq. (1)). Other term weighting methods have been proposed and successfully applied. For instance, OKAPI measures consider the document length [32]. Our corpus is homogeneous in terms of length, which justified the choice of TF-IDF weighting.

The morphosyntactic output matrix, M_{MS}, corresponds to the cosine similarity matrix of the morphosyntactic document-features matrix. Suppose a document-term matrix containing N rows (representing the documents) and V columns (representing the terms, or features), the similarity matrix is an $N \times N$ matrix where each component $x_{i,j}$ corresponds to the cosine similarity between the documents i and j, as defined in Eq. (2).

2.2. 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 for our task (i.e., retrieving related documents). Four types of fusion methods are described in the literature [33–38], but in this paper, we focus on the two most commonly used methods, i.e., early fusion and late fusion (Fig. 2(a) and Fig. 2(b)).

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

2.2.1. Fusion of thematic and spatiotemporal features

Data fusion methods are increasingly being used in content analysis and retrieval, especially when handling diverse and complementary data sources. Fusion methods were initially used in multimedia analysis to address the problem of combining multimodal data (i.e., different types of data, also referred to as heterogeneous data). For instance, fusion methods combine textual and visual data features to improve multimedia retrieval or user recommendation systems [33–35]. These methods can also be applied to the fusion of homogeneous data, such as textual features at different linguistic levels (e.g., lexical, syntactic, and semantic) [36]. Several types and levels of fusion strategies exist, among which early and late fusion are the most commonly used [37]. Early and late fusion functions differ in how they integrate the results from feature extraction.

Early fusion consists of combining the features into a unique multimodal representation (e.g., textual and visual

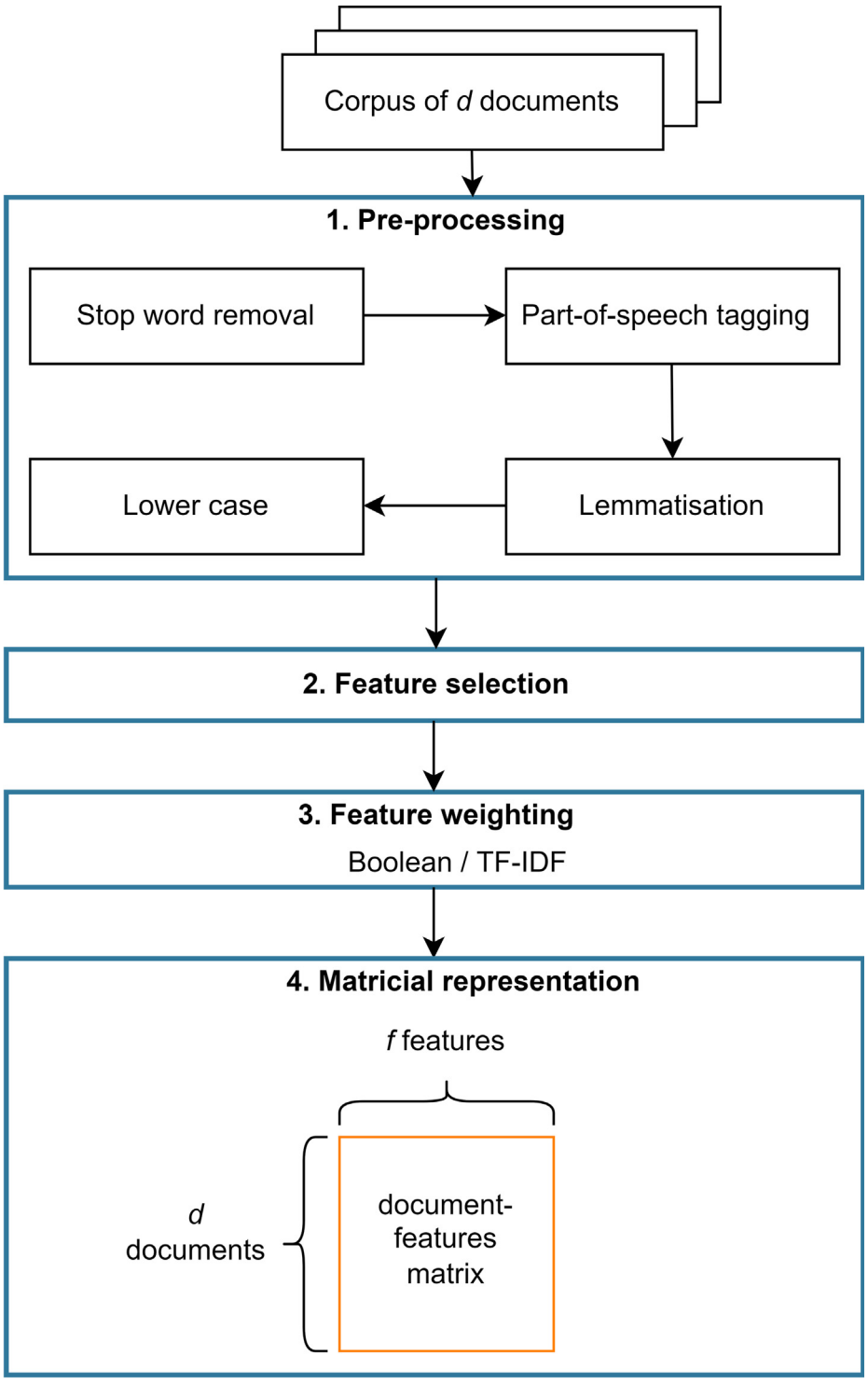


Fig. 1 – Steps to create the document-features matrix based on morphosyntactic features.

features [38]). This representation is used as an input for the "decision step". This step, also called the "learning phase", can be as simple as the calculus of a similarity matrix [37]. However, this step can also involve more sophisticated approaches, such as the manual attribution of scores by experts [33] or machine learning methods [38].

The main advantage of early fusion is that one unique matrix goes through the learning phase, which reduces the computing time and leverages the correlation between the concatenated features. The main disadvantages are increasing the representation space and the difficulty of combining features into a common representation [37]. The decision step is first performed using unimodal features in the context of late fusion. Then, the outputs of the decision step are combined into a single final dataset. The advantage is that the features are combined at the same representation level (e.g., similarity

Table 1 – Textual preprocessing methods evaluated. BOW: bag-of-words, SW: stop word removal, BOWlem: BOW lemmatisation, POS: part-of-speech, V: verbs, N: nouns, and PN: proper nouns.

Method name	Description	Processed text
No transformation		*Bluetongue cases were declared in France*
BOW	Breaking the text into words (tokenisation)	*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, V	Selecting only the verbs	*were, declared*
POS selection, N	Selecting only the common nouns	*cases*
POS selection, PN	Selecting only the proper nouns	*Bluetongue, France*

matrices). The main limitation is the increase in computing time and the potential loss of correlation [37]. A weight can be applied to control the impact of early and late fusion.

We applied early and late fusion methods to fuse thematic and spatiotemporal features, as illustrated in Fig. 2(a) and Fig. 2(b). The fusion methods involve three matrices operations: the addition, hereafter symbolized by "+", the similarity calculation, hereafter symbolized by the function sim, and the concatenation, hereafter symbolized by "|". The addition consists in adding the corresponding elements together (the matrices must have the same number of rows and columns). The similarity calculation is based on the cosine similarity function. The concatenation consists in joining the two matrices along rows. The number of columns in the output matrix corresponds to the sum of the columns from the concatenated matrices, thus increasing the features space (i.e., the number of columns). In weighted combination, the elements of the matrices are multiplied by two different coefficients before being combined. Here, M_D, M_H, M_S and M_T are the disease, host, spatial and temporal feature matrices, respectively; and sim is the cosine similarity function.

The combination step for early fusion is based on the concatenation. The fused disease-host matrix (M_{DH}) and the spatiotemporal fused matrix (M_{ST}) obtained by early fusion are defined by Eqs. (3) and (4), respectively:

$$M_{DH} = \text{sim}(\alpha_{DH} \times M_D | (1 - \alpha_{DH}) \times M_H \quad (3)$$

$$M_{ST} = \text{sim}(\alpha_{ST} \times M_S | (1 - \alpha_{ST}) \times M_T \quad (4)$$

The combination step for late fusion is based on addition. The disease-host matrix (M_{DH}) and the spatiotemporal fused matrix (M_{ST}) obtained by late fusion are defined by Eqs. (5) and (6), respectively:

$$M_{DH} = \alpha_{DH} \times S_D + (1 - \alpha_{DH}) \times S_H \quad (5)$$

$$M_{ST} = \alpha_{ST} \times S_S + (1 - \alpha_{ST}) \times S_T \quad (6)$$

Where S_D = sim(M_D), S_H= sim(M_H), S_S= sim(M_S) and S_T = sim (M_T).

For each fusion, α_{DH} and α_{ST} ranged from 0 to 1 with a step size of 0.1.

2.2.2. *Fusion of all features*

This step consisted of combining the fused matrices from the previous step into a final matrix. We selected the best M_{DH} and M_{ST} based on their F-measures and concatenated them by using a weight β ranging from 0 to 1 with a step size of 0.1, obtaining the lexicosemantic matrix M_{LS} (Eq.(7)):

$$M_{LS} = \beta \times M_{DH} | (1 - \beta) \times M_{ST} \quad (7)$$

2.2.3. *Feature generalisation*

The lexicosemantic representation does not consider the similarity between related features, 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 2).

For disease and host entities, generalisation aims to consider the synonyms used to refer to the same disease or host. To convert each feature into its generalised form, we manually built a dictionary with each disease and host variant mapped to its canonical form and species, respectively. For spatiotemporal features, we aimed to consider 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[1] gazetteer to transform each spatial feature into its first administrative level (level 1) or into its country (level 2). For temporal features, we used the normalised values given by HeidelTime[2] to transform features into their week number (level 1) or month (level 2) [39]. 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 continents such as "Africa" were not transformed.

The generalisation reduced the number of distinct features in each category (Table 3), 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 that supported the experiments and the evaluation protocol.

[1] https://www.geonames.org/.

[2] https://dbs.ifi.uni-heidelberg.de/research/heideltime/.

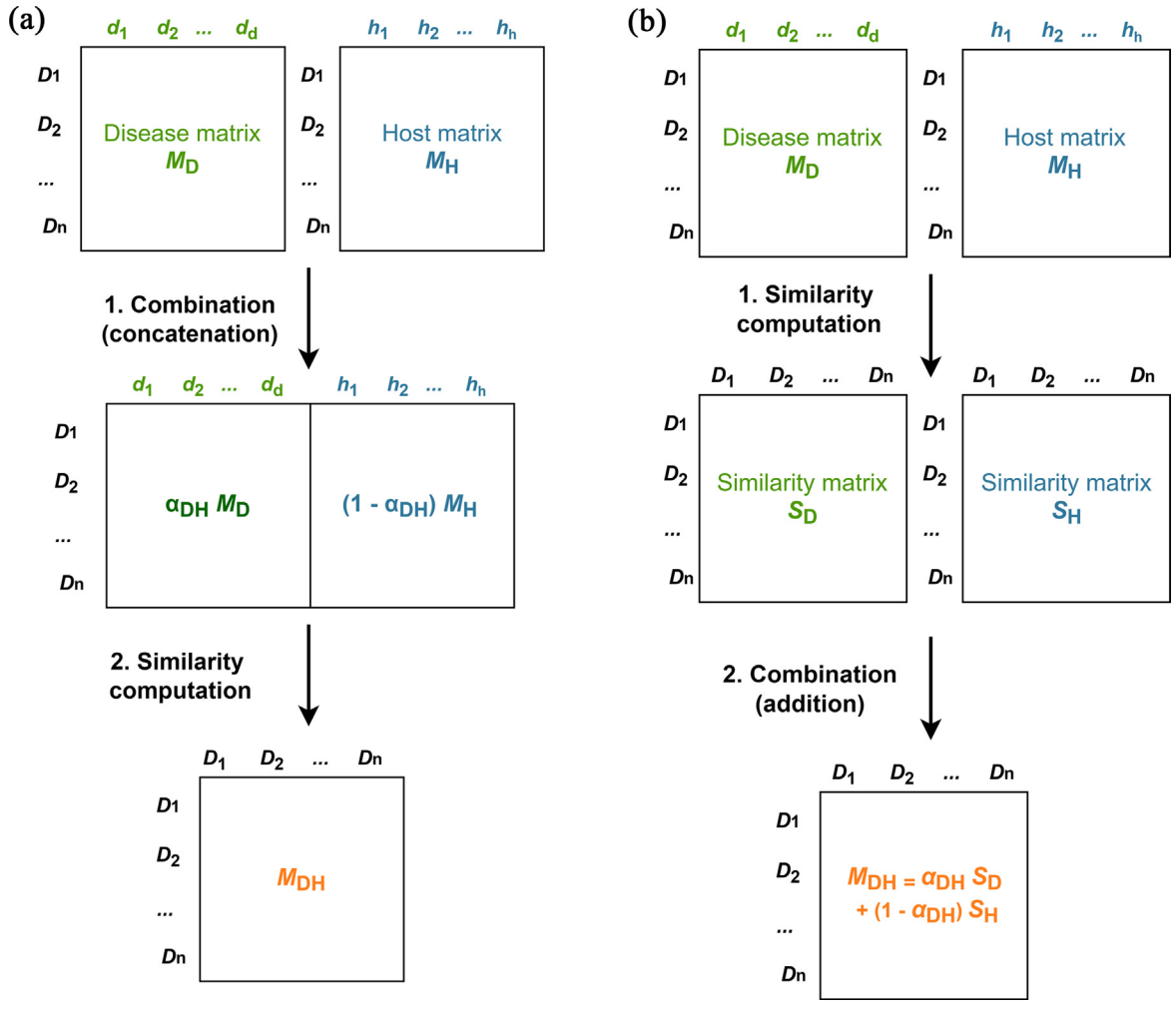


Fig. 2 – Early fusion (a) and late fusion (b) of the disease matrix (M_D) and host matrix (M_H) representing *n* documents D, with *d* and *h* features, respectively. S_D, and S_H are the similarity matrices. α_{DH} is the weight considered in linear combinations. Fusion of the spatial matrix and temporal matrix is based on the same process, by replacing the disease and the host matrices by the spatial and temporal matrices, respectively.

2.3. *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.) [40]. This corpus was initially designed for training and evaluating the PADI-web information extraction module [18]. The corpus contains information about the news article itself (publication date, title, content, URL, etc.) and epidemiological features (locations, diseases, hosts, dates and symptoms), which were automatically identified by data mining and rule-based approaches. A veterinary epidemiologist and a computer scientist subsequently labelled each candidate as correct or incorrect. For each document and type of feature (i.e., disease, host, date and location), only candidates manually labelled correct in the corpus were retained for analysis (including the 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. Each detected event was labelled using the unique EMPRES-i identifier. When the epidemiologist could not link an event to an official event, she created a new event identifier and manually recorded the epidemiological features (location, date, disease and host). News articles containing at least one event represented 53% of the corpus (n = 229/438). Among these news articles, 52% (n = 127/229) reported several events, with a median number of 3 events (Table 4).

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 these events, 70% (n = 541/771) were reported in one single news article. The events present in multiple news articles were reported in up to 11 news articles (median number of 3 news articles). In

Table 2 – 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-W015
	2	Month	14-02-2015 → 2015-02

Table 3 – 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 parentheses.

	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%)

the following experiments, we selected only news articles containing at least one event (corpus of 229 documents). We considered 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 is in [1, 229] and $i \neq k$. We obtained 157 sets of related documents. Even if it still has a modest size, our corpus is highly specialised regarding both its domain (i.e., animal health) and its nature (i.e., online news articles).

3. Results

3.1. Evaluation protocol

The matrices obtained by the morphosyntactic and lexicosemantic representations (M_{MS} and M_{LS}, respectively) were evaluated using ranking functions, i.e., regarding their ability to give higher similarity scores to relevant elements than to irrelevant documents. 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. Fig. 3 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 and D_7 are related. The displayed rankings are obtained by sorting the documents by decreasing similarity scores with D_7.

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 favoured ranking functions that were able to discriminate relevant from irrelevant elements.

We evaluated the ranking quality according to the ability of the fusion to give a better rank to relevant pairs (i.e., related documents) than to irrelevant pairs. The ranking was evaluated in terms of the normalised precision (P), normalised recall (R) and F-measure (F). R and P are based on the difference between the sum of the ranks of V relevant pairs obtained by a ranking function and the sum of the ranks of an ideal list, respectively, where all relevant pairs are retrieved before all the irrelevant pairs [41,42]:

$$R = 1 - \frac{1}{V \times (N-V)} \times \sum_{i=1}^{V} r_i - \sum_{i=1}^{V} i \tag{8}$$

$$P = 1 - \frac{1}{\log \binom{N}{V}} \times \sum_{i=1}^{V} \log(r_i) - \sum_{i=1}^{V} \log(i) \tag{9}$$

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

Graphically, R corresponds to the area under the curve (AUC) of the receiver operating characteristic (ROC) curve or AUC. F-measure is the harmonic mean between the recall and the precision:

$$F = 2 \times \frac{P \times R}{P + R} \tag{10}$$

We also evaluated the r-precision (rP), corresponding to the precision after r documents have been retrieved, where r is the number of relevant documents for a set [43]. This is calculated as follows:

$$rP = \frac{1}{r} \times \sum_{i=1}^{r} x_i \tag{11}$$

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

R, P and F evaluate the ranking in terms of the entire set of documents. The 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.

Table 4 – Descriptive statistics of the number of articles ($N_{article}$) per event and number of events (N_{event}) per article in the event corpus.

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

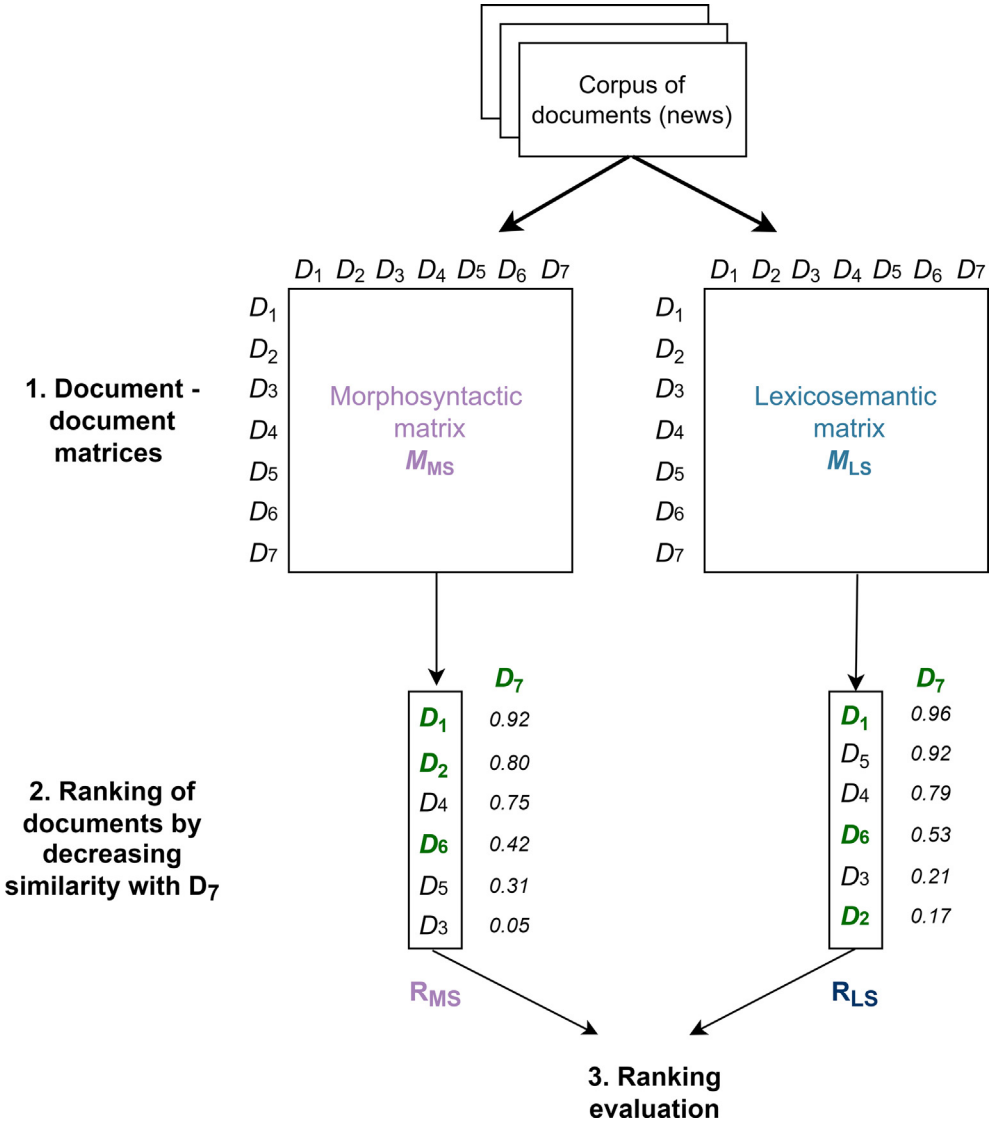


Fig. 3 – Evaluation pipeline for information retrieval. M_{MS} and M_{LS} represent the two output matrices obtained by the morphosyntactic and lexicosemantic representations, respectively. R_{MS} and R_{LS} represent the two rankings obtained by the morphosyntactic and lexicosemantic representations, respectively, relative to a document D_7 by decreasing order of similarity. Documents related to D_7 are shown in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Morphosyntactic features

Table 5 shows the ranking performances 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 normalised recall (R) was better than the normalised precision (P), i.e., reaching up to 0.98. The BOW representation with stop word removal outperformed the other types of representations (F = 0.89, rP = 0.45), with a vocabulary length of 14 996. In comparison, the lemmatised BOW with the selection of proper nouns obtained very close results (F = 0.87, rP = 0.42), with a representation space of 5 129 features. The lowest performances were obtained by lemmatisation and verb selection only (F = 0.53, rP = 0.09), which corresponds to the lowest space dimensionality (vocabulary length of 2 151).

3.3. Lexicosemantic features

3.3.1. 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 (Fig. 4) and spatiotemporal features (Fig. 5). For the disease-host fusion, the lowest F values were obtained with the unimodal matrices (corresponding to α_{DH} equals 0 or 1 in Eq. (3)).

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

Tables 6 and 7 show the best results obtained by the different studied models (based on the best F values) for the disease-host and spatiotemporal fusion. The best disease-host fusion models were obtained with $\alpha_{DH} = 0.4$ (level 0) and $\alpha_{DH} = 0.7$ (level 1). The best spatiotemporal fusion models were obtained with $\alpha_{ST} = 0.7$ (level 0), $\alpha_{ST} = 0.8$ (level 1) and $\alpha_{ST} = 0.7$ (level 2).

As in the baseline approaches, the normalised recall was better than the normalised precision among all of the studied models. Disease-host fusion obtained a maximal F 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 (rP = 0.47).

Feature generalisation had different impacts on the performance metrics. In the TF-IDF models, the normalised recall and precision increased when applying the generalisation steps. When applied to Boolean representations, it decreased the recall R 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).

3.3.2. Fusion of all features

The second fusion step improved the global ranking of the bimodal matrices (from step 1), as illustrated in Fig. 6. 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-measure was obtained when more weight was given to the spatiotemporal matrix (β ranging from 0.5 to 0.8). Table 8 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) slightly outperformed the best F-measure 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). The 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) are shown in Tables 6 and 7.

4. Discussion

In this paper, we evaluated the performance of morphosyntactic and lexicosemantic representations for the retrieval of

Table 5 – Ranking performances of morphosyntactic features. BOW: bag-of-words, SW: stop word removal, BOWlem: BOW lemmatisation, V: verbs, N: nouns, and PN: proper nouns), in terms of recall (R), precision (P), F-measure (F) and r-precision (rP).

	R	P	F	rP	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, V	0.74	0.41	0.53	0.09	2 151
BOWlem, N	0.89	0.61	0.72	0.25	4 185
BOWlem, PN	0.96	0.80	0.87	0.42	5 129

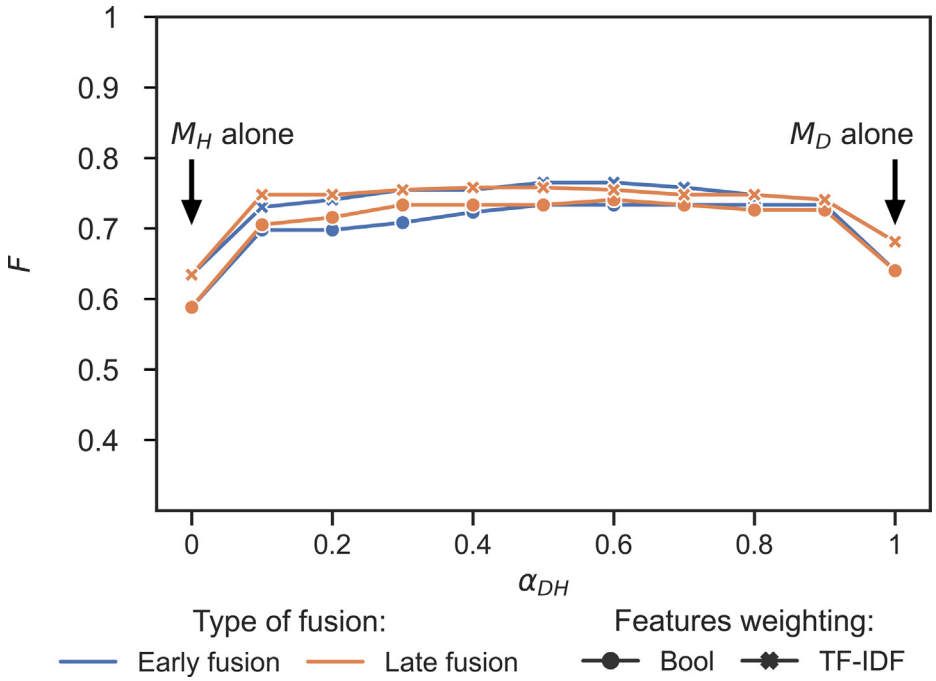


Fig. 4 – Comparison of fusion methods to combine disease and host features in terms of F-measure. The arrows correspond to the use of unimodal matrices (disease matrix M_D and host matrix M_H).

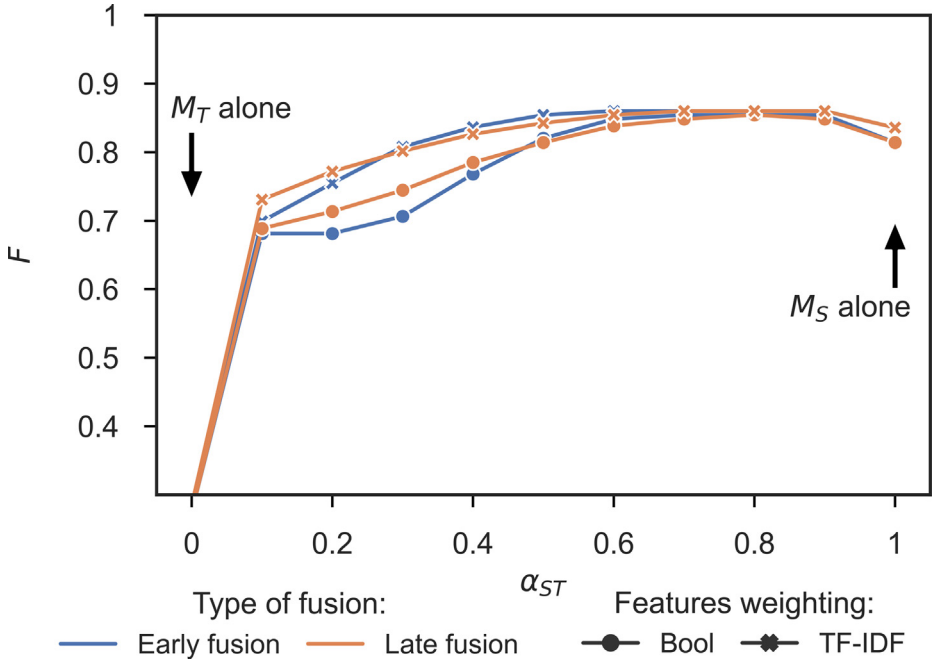


Fig. 5 – Performance of fusion methods to combine spatiotemporal features in terms of F-measure. The arrows correspond to the use of unimodal matrices (spatial matrix M_S and temporal matrix M_T).

related documents. In both representations, the TF-IDF term weights outperformed the Boolean ones. The best morphosyntactic and lexicosemantic representations obtained comparable performances in terms of the global ranking (F-measure). However, the lexicosemantic models significantly increased the local precision (r-precision) compared to the morphosyntactic approach.

Regarding the lexicosemantic features, the spatial features were the most efficient features when considered separately. Besides, the best retrieval results were achieved when more weight was given to the spatiotemporal matrix in the bimodal representation and the final fusion. These results were consistent with the task performed, i.e., the retrieval of related documents corresponds to the retrieval of the related events

Table 6 – Ranking performances of disease and host features (matrix M_{DH}) using different types of fusion, term weighting (Boolean or TF-IDF) and generalisation levels (level 0: no generalisation and level 1: first generalisation level), in terms of recall (R), precision (P), F-measure (F) and r-precision (rP).

	Early fusion				Late fusion			
	R	P	F	rP	R	P	F	rP
Boolean								
level 0	0.91	0.60	0.72	0.23	0.91	0.60	0.72	0.23
level 1	0.92	0.59	0.72	0.19	0.91	0.58	0.71	0.19
TF-IDF								
level 0	0.91	0.63	0.74	0.27	0.92	0.63	0.75	0.27
level 1	**0.93**	**0.65**	**0.77**	**0.28**	0.93	0.63	0.75	0.25

Table 7 – Ranking performances of spatial and temporal features (matrix M_{ST}) using different types of fusion, term weighting (Boolean or TF–IDF) and generalisation levels (level 0: no generalisation, level 1: first generalisation level, and level 2: second generalisation level), in terms of recall (R), precision (P), F-measure (F) and r-precision (rP).

	Early fusion				Late fusion			
	R	P	F	rP	R	P	F	rP
Boolean								
level 0	0.88	0.75	0.81	0.44	0.88	0.74	0.80	0.42
level 1	0.89	0.77	0.83	**0.47**	0.89	0.76	0.82	0.46
level 2	0.91	0.70	0.79	0.28	0.92	0.73	0.81	0.32
TF–IDF								
level 0	0.89	0.76	0.82	0.46	0.89	0.76	0.82	0.46
level 1	0.89	0.77	0.83	**0.47**	0.89	0.77	0.83	0.46
level 2	**0.93**	0.77	0.84	0.38	**0.93**	**0.79**	**0.85**	0.43

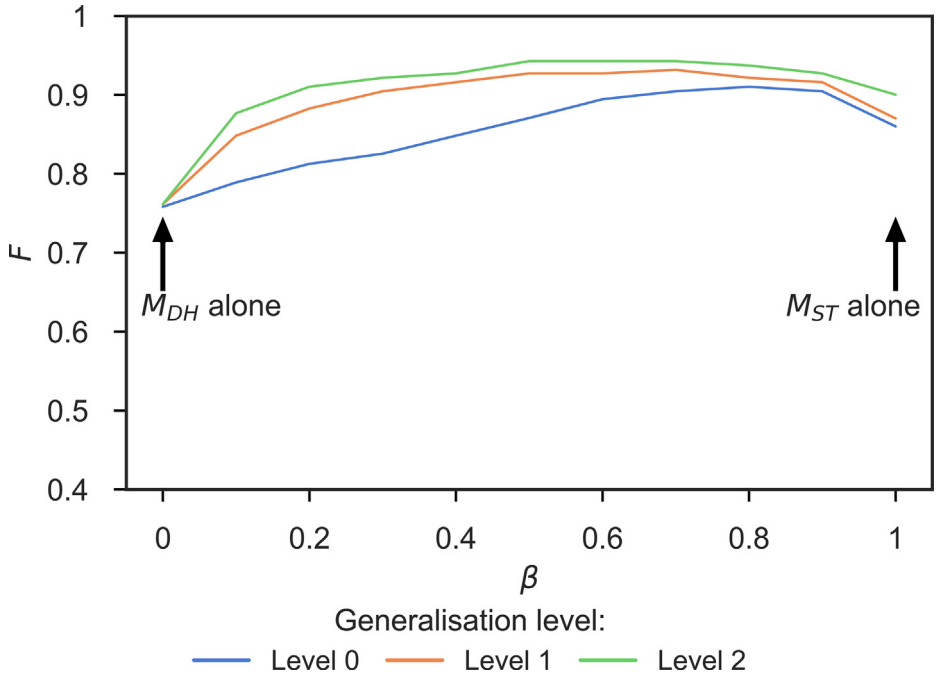


Fig. 6 – Performance of all lexicosemantic features to retrieve relevant documents according to different fusion weights β, in terms of F-measure. The arrows correspond to the bimodal matrices (thematic matrix M_{DH} and spatiotemporal matrix M_{ST}).

Table 8 – Comparison of ranking performances of morphosyntactic and lexicosemantic representations in terms of recall (R), precision (P), F-measure (F), r-precision (rP) and vocabulary length. The morphosyntactic representations include the bag-of-words with stop-word removal (BOW, SW) and the lemmatized bag-of-words with selection of verbs (BOWlem, V). The lexicosemantic representations include the best bimodal disease-host and spatio-temporal fusions (M_{DH} and M_{ST}) from Tables 6 and 7, and the combinations of the best bimodal fusions at different generalisation levels (level 0: no generalisation, level 1: first generalisation level, and level 2: second generalisation level). We show the best combinations in terms of F-measure and the corresponding weight β.

Morphosyntactic features:	R	P	F	rP	Vocabulary length
BOW, SW	0.98	0.82	0.89	0.45	14 996
BOWlem, V	0.74	0.41	0.53	0.09	2 151
Lexicosemantic features (bimodal):					
M_{DH}	0.93	0.65	0.77	0.28	65
M_{ST}	0.93	0.79	0.85	0.43	279
Lexicosemantic features (all):					
(M_{DH}, level 0 + M_{ST}, level 0), $\beta = 0.8$	0.97	0.83	0.89	0.54	1 459
(M_{DH}, level 1 + M_{ST}, level 1), $\beta = 0.6$	0.97	0.85	0.91	**0.58**	928
(M_{DH}, level 1 + M_{ST}, level 2), $\beta = 0.6$	**0.98**	**0.87**	**0.92**	**0.58**	344

contained in the documents. These events were characterised 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 the 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 an early fusion matrix and a late fusion matrix. During our tests, we did not find any clear trend favouring one type of fusion over another. We believe that the impact of the different types of fusion may be reduced when combining the same type of features (here, textual). Thus, from a practical viewpoint, it would be reasonable to choose the early fusion method due to its reduced computational time.

The ranking performances were significantly impacted by the weights used. While we recommend giving equal weight to disease and host features, we advise attributing a higher weight to spatial features than to temporal features ($0.6 \leq \alpha_{ST} \leq 0.8$).

Feature generalisation allowed us to increase the global model performances (in terms of the F-measure) 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 many 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 [44], i.e., 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 that maps each named entity to a canonical form [45]. However, the event extraction performances with and without the use of the ontology were not compared in that study.

Regarding the morphosyntactic features, all lemmatised representations that 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 that showed that the spatiotemporal matrices should be assigned a higher weight.

Moreover, proper names include other types of named entities, such as organisation names, which we did not consider 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.

5. Conclusion

In this paper, we propose a lexicosemantic representation for the retrieval of disease-related news articles. We show that the fusion of two groups of features, i.e., thematic (disease and host) and spatio-temporal, combined with appropriate weights and generalisation steps, outperformed the classical morphosyntactic (bag-of-words) representation. In the lexicosemantic context, spatial features were the most discriminative features, thus highlighting the need for robust methods for spatial entity extraction, disambiguation and representation. Conversely, temporal features performed poorly on their own. The lexicosemantic approach provides a thematic and comprehensible representation of the disease-related news articles. Besides, the approach has the advantage of being based on a reduced number of features, thus limiting the computing time when handling larger corpora.

In this paper, we focused on animal disease monitoring. Even if veterinary surveillance has some specificities, such as detecting the host, it has similarities with human disease surveillance. Indeed, we used PADI-web to detect early signals of COVID-19 [46]. Thus, we believe that our approach could be easily applied to public health surveillance.

Expansions of text content before applying fusion approaches by using word embedding architectures such as the bidirectional encoder representations from transformers (BERT) model could be proposed. This model achieved new state-of-the-art results on several NLP tasks [47,48]. BERT produces word representations that are dynamically informed by the words around them (i.e., context-dependent). As discussed in [47], considering the time complexity of BERT, simple representations could be adapted, especially when the proportion of the training data is low. These general conclusions could be further evaluated in our specialised context dedicated to animal disease surveillance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD009. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. This work has also been funded by the French General Directorate for Food (DGAL), the French Agricultural Research Centre for International Development (CIRAD), the SONGES Project (FEDER and Occitanie), and the French National Research Agency under the Investments for the Future Program, referred to as ANR-16-CONV-0004 (#DigitAg).

REFERENCES

[1] Brugere C, Onuigbo DM, Morgan KL. People matter in animal disease surveillance: Challenges and opportunities for the aquaculture sector. Aquaculture 2017;467:158–69.
[2] Tatem AJ, Rogers DJ, Hay SI. Global Transport Networks and Infectious Disease Spread. In: Hay SI, Graham A, Rogers DJ, editors. Advances in Parasitology. Academic Press; 2006. p. 293–343.
[3] Rich KM, Niemi JK. The economic impact of a new animal disease: same effects in developed and developing countries? Revue Sci Tech de l'OIE 2017;36:115–24.
[4] Rushton J, Viscarra R, Guerne Bleich E, McLeod A. 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 Sci J 2005;61:491–514.
[5] Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci 2001;356:983–9.
[6] World Health Organization. Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version. Link: https://apps.who.int/iris/bitstream/handle/10665/112667/WHO_HSE_GCR_LYO_2014.4_eng.pdf; 2014.
[7] Conway M, Dowling JN, Chapman WW. Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America. J Biomed Inform 2013;46:734–43.
[8] Hazewinkel MC, de Winter RFP, van Est RW, van Hyfte D, Wijnschenk D, Miedema N, Hoencamp E. Text Analysis of Electronic Medical Records to Predict Seclusion in Psychiatric Wards: Proof of Concept. Front Psychiatry 2019;10:188.
[9] Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. J Biomed Inform 2017;66:82–94.
[10] Fricker RD, Hegler BL, Dunfee DA. Comparing syndromic surveillance detection methods: EARS' versus a CUSUM-based methodology. Stat Med 2008;27:3407–29.
[11] Hennings KJ. What is syndromic surveillance. Syndromic surveillance: reports from a national conference. Morbidity and mortality weekly report 2003;53(supplemental):7–11.
[12] Vial F, Wei W, Held L. Methodological challenges to multivariate syndromic surveillance: a case study using Swiss animal health data. BMC Veterinary Res 2016;12(1).
[13] Faverjon C, Andersson MG, Decors A, Tapprest J, Tritz P, Sandoz A, Kutasi O, Sala C, Leblond A. Evaluation of a multivariate syndromic surveillance system for West Nile Virus. Vector-Borne Zoonotic Diseases 2016;16:382–90.
[14] Dórea FC, Vial F. Animal health syndromic surveillance: a systematic literature review of the progress in the last 5 years (2011–2016). Veterinary Med: Res Reports 2016;7:157–70.
[15] Madouasse A, Marceau A, Lehébel A, Brouwer-Middelesch H, van Schaik G, Van der Stede Y, Fourichon C. Use of monthly collected milk yields for the detection of the emergence of the 2007 French BTV epizootic. Preventive Veterinary Med 2014;113:484–91.
[16] Aiello AE, Renson A, Zivich PN. Social Media- and Internet-Based Disease Surveillance for Public Health. Annu Rev Public Health 2020;41:101–18.
[17] Choi J, Cho Y, Shim E, Woo H. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. BMC Public Health 2016;16:1238.
[18] Arsevska E, Valentin S, Rabatel J, de Goër de Hervé J, Falala S, Lancelot R, et al. Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. PLOS One 2018;13:e0199960.
[19] Valentin S, Arsevska E, Falala S, de Goër J, Lancelot R, Mercier A, Rabatel J, Roche M. PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. Comput Electron Agric 2020;169.
[20] Valentin S, Lancelot R, Roche M. How to combine spatio-temporal and thematic features in online news for enhanced animal disease surveillance? Proc Comput Sci 2018;126:490–7.
[21] Nadkarni PM. An introduction to information retrieval: applications in genomics. Pharmacogenomics J 2002;2:96–102.
[22] Strat ST, Benoit A, Lambert P, Bredin H, Quénot G. Hierarchical late fusion for concept detection in videos. Fusion Computer Vision Springer 2014:53–77.
[23] Lops P, de Gemmis M, Semeraro G. Content-based recommender systems: state of the art and trends. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender Systems Handbook. Springer, US, Boston: MA; 2011. p. 73–105.
[24] Berry MW, Castellanos M. Survey of text mining II. Springer; 2008.
[25] Gomaa WH, Fahmy AA. A survey of text similarity approaches. Int J Comput Appl 2013;68(13).
[26] Huang A. Similarity measures for text document clustering. In: Proceedings of the 6th New Zealand Computer Science Research Student Conference. Christchurch, New Zealand; 2008.
[27] Turney PD, Pantel P. From frequency to meaning: vector space models of semantics. J Artif Int Res 2010;37:141–88.

[28] Uysal AK, Gunal S. The impact of preprocessing on text classification. Inf Process Manage 2014;50:104–12.
[29] HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. PLoS ONE 2020;15.
[30] Chua S. The Role of Parts-of-Speech in Feature Selection. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong; 2008. Vol 1.
[31] Bird S, Loper E. NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions; Association for Computational Linguistics, Barcelona, Spain; 2004. p. 214–217.
[32] Robertson SE, Jones KS. Relevance weighting of search terms. J Am Soc Inform Sci 1976;27:129–46.
[33] Clinchant S, Ah-Pine J, Csurka G. Semantic combination of textual and visual information in multimedia retrieval. In: Proceedings of the 1st ACM international conference on multimedia retrieval. ACM; 2011. p. 44.
[34] Wang J-H, Wu Y-T, Wang L. Predicting Implicit User Preferences with Multimodal Feature Fusion for Similar User Recommendation in Social Media. Appl Sci 2021;11:1064.
[35] Unar S, Wang X, Wang C, Wang Y. A decisive content based image retrieval approach for feature fusion in visual and textual images. Knowl-Based Syst 2019;179:8–20.
[36] Eke CI, Norman AA, Shuib L. Multi-feature fusion framework for sarcasm identification on twitter data: A machine learning based approach. PLoS ONE 2021;16.
[37] Soriano-Morales E-P. Hypergraphs and information fusion for term representation enrichment. Applications to named entity recognition and word sense disambiguation. Doctor thesis. Univ. Lumière Lyon 2: France; 2018.
[38] Snoek CG, Worring M, Smeulders AW. Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on Multimedia; ACM; 2005. p. 399–402.
[39] Strotgen J, Gertz M. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation; 2010, p. 321–324.
[40] Rabatel J, Arsevska E, Roche M. PADI-web corpus: Labeled textual data in animal health domain. Data in Brief 2019;22:643–6.
[41] Kishida K. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. National Institute of Informatics. NII Technical Report; 2005.
[42] Salton G, Lesk ME. Computer Evaluation of Indexing and Text Processing. J Assoc Comput Mach 1968;15:8–36.
[43] Buckley C, Voorhees EM. Evaluating Evaluation Measure Stability. ACM SIGIR Forum 2017;51(2):235–42.
[44] Drury B, Fernandes R, Moura M-F, de Andrade Lopes A. A survey of semantic web technology for agriculture. Inform Process Agric 2019;6:487–501.
[45] Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, et al. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 2008;24:2940–1.
[46] Valentin S, Mercier A, Lancelot R, Roche M, Arsevska E. Monitoring online media reports for early detection of unknown diseases: Insight from a retrospective study of COVID-19 emergence. Transbound Emerg Dis 2021;68:981–6.
[47] Piskorski J, Haneczok J, Jacquet G. New Benchmark Corpus and Models for Fine-grained Event Classification: To BERT or not to BERT? In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain; 2020. p. 6663–78.
[48] Torregrossa F, Allesiardo R, Claveau V, Kooli N, Gravier G. A survey on training and evaluation of word embeddings. Int J Data Sci Anal 2021;11:85–103.