








# PADI-web: An Event-Based Surveillance System for Detecting, Classifying and Processing Online News

Sarah Valentin<sup>1,2,3</sup> , Elena Arsevska<sup>2,3</sup> , Alize Mercier<sup>2,3</sup> ,  
Sylvain Falala<sup>2</sup>, Julien Rabatel<sup>3</sup>, Renaud Lancelot<sup>2,3</sup> ,  
and Mathieu Roche<sup>1,3</sup> 

<sup>1</sup> UMR TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, 34398 Montpellier, France  
sarah.valentin@cirad.fr

<sup>2</sup> UMR ASTRE, Univ Montpellier, CIRAD, INRAE, 34398 Montpellier, France  
<sup>3</sup> CIRAD, Montpellier, France

**Abstract.** The Platform for Automated Extraction of Animal Disease Information from the Web (PADI-web) is a multilingual text mining tool for automatic detection, classification, and extraction of disease outbreak information from online news articles. PADI-web currently monitors the Web for nine animal infectious diseases and eight syndromes in five animal hosts. The classification module is based on a supervised machine learning approach to filter the relevant news with an overall accuracy of 0.94. The classification of relevant news between 5 topic categories (confirmed, suspected or unknown outbreak, preparedness and impact) obtained an overall accuracy of 0.75. In the first six months of its implementation (January–June 2016), PADI-web detected 73% of the outbreaks of African swine fever; 20% of foot-and-mouth disease; 13% of bluetongue, and 62% of highly pathogenic avian influenza. The information extraction module of PADI-web obtained F-scores of 0.80 for locations, 0.85 for dates, 0.95 for diseases, 0.95 for hosts, and 0.85 for case numbers.

PADI-web allows complementary disease surveillance in the domain of animal health.

**Keywords:** Epidemic intelligence · Animal health · Web monitoring · Text mining · Classification · Information extraction

## 1 Introduction

Until the early 2000s, the surveillance of diseases has been essentially dependent on the principles of traditional, indicator-based surveillance (IBS). The IBS mainly involves reporting of known diseases based on sets of rules for verification and confirmation of cases, from clinicians to laboratories and health officials [14].

However, the rapid growth of the Internet and the connectivity of the users to the World Wide Web addressed the need of a supplementary disease surveillance,

i.e. event-based surveillance (EBS). Compared to the IBS, the EBS has been proven to be more flexible to detect both known and new diseases through the use of multiple different sources, languages and geographic coverage [8, 10, 18].

ProMED-mail, one of the first EBS systems, is based on the sharing of sanitary information between users and manual Web searches implemented by human analysts [11]. In contrast, the HealthMap [6] and MedISys [16] systems automatically detect disease-related contents on the Web, extract, and visualise events on interactive maps to monitor trends. The current EBS systems mainly focus on diseases of public health interest [7]. As a consequence, these systems have a limited value for animal health authorities due to the inconsistent coverage of animal health topics.

This paper describes the contributions towards the development and implementation of an EBS system dedicated to the detection and monitoring of new and emerging animal infectious diseases occurring worldwide. We present the Platform for Automated Extraction of Animal Disease Information from the Web (PADI-web), a text mining platform for automatic detection, translation, classification and extraction of disease (outbreak) information from news articles published on the Web (further on referred to as “news”).

PADI-web was primarily developed for the French Epidemic Intelligence System (FEIS, or *Veille sanitaire internationale* in French), which is part of the French animal health epidemiological surveillance Platform (ESA Platform<sup>1</sup>). It is now publicly available<sup>2</sup> through an online platform, in both English and French. Currently, PADI-web scans the Web for nine animal infectious diseases, i.e. African swine fever, classical swine fever, avian influenza, foot-and-mouth disease, bluetongue, Schmallenberg virus infection, West Nile, lumpy skin disease and Rift Valley fever. In order to detect the emergence of potentially new or unknown infectious diseases, PADI-web monitors the Web for eight syndromes, i.e. general, respiratory, digestive, locomotion/neurologic, skin/mucous, haemorrhagic, reproductive and postnatal/congenital in five hosts, i.e. avian, bovine, ovine, caprine and porcine animals.

## 2 PADI-web Approaches

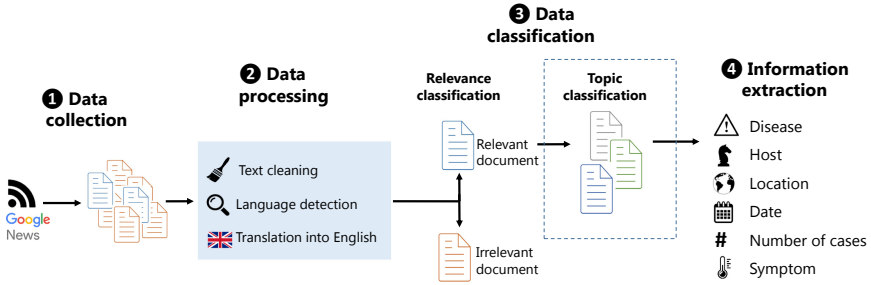
Figure 1 shows the main steps implemented in PADI-web, i.e. data collection, data processing, data classification, and extraction of epidemiological information (features of an outbreak).

### 2.1 Data Collection

PADI-web collects news articles in near-real time through Really Simple Syndication (RSS) feeds. Once a day, 7 days a week, RSS feeds from Google News are processed. Similarly to other EBS systems, we have chosen Google News for

<sup>1</sup> <https://www.plateforme-esa.fr/>.

<sup>2</sup> <https://padi-web.cirad.fr/en/>.



**Fig. 1.** Pipeline implemented in PADI-web. PADI-web scans the Web through customised RSS feeds (step 1). Once collected and stored in a database, the news contents are cleaned and translated (step 2). Then, the news are classified as relevant or irrelevant (step 3). The topic classification, presented in this paper, is not yet integrated into the PADI-web pipeline. Using a combined method for information extraction, epidemiological indicators are extracted from the news content (step 4).

to its international coverage and flexible RSS feeds. In order to detect relevant disease-related news articles, the RSS feeds use specific terminology based on: i) disease names - to detect news which describe outbreaks of known diseases, and ii) associations of terms on hosts and clinical signs - to detect news related to the occurrences of unknown diseases and syndromes in animals.

The terminology is automatically extracted using text mining techniques from a corpus of relevant news articles. We apply additional text mining techniques to automatically obtain associations between terms describing the clinical signs and hosts. Finally, using a Delphi approach and through a consensus, a group of animal health experts validates the extracted terms and associations and, if necessary, complements the list with additional terms. Detailed description of our methods for extraction of disease-related terminology is described elsewhere [2].

All the RSS feeds have an English version. In addition, several RSS feeds are adapted into additional languages, including Chinese, Turkish, French or Arabic. The RSS feed keywords (disease, hosts and symptoms) are translated using two vocabularies, i.e. UMLS [4] and Agrovoc<sup>3</sup>, a controlled vocabulary developed by the Food and Agriculture Organization (FAO). The choice of languages is based on epidemiological expertise, in order to target specific high-risk areas. For instance, we are currently monitoring foot-and-mouth disease through RSS feeds in English, French and Arabic language, as it is endemic in Northern Africa.

## 2.2 Data Processing

Duplicated news (i.e. news for which the url already exists in the database) are filtered out. PADI-web retrieves the news content from its webpage. The textual content (title and text) is cleaned to remove irrelevant elements (e.g.

<sup>3</sup> <http://aims.fao.org/vest-registry/vocabularies/agrovoc>.

pictures, ads, hyperlinks) and the language is detected, using respectively the *BeautifulSoup* [17] and *langdetect* python libraries. All non-English news articles are translated using the Translator API of the Microsoft Azure system<sup>4</sup>. We use English as a bridge-language because the models for classification (Sect. 2.3) and information extraction (Sect. 2.4) modules have been trained with labeled data in this language.

## 2.3 Data Classification

First, news are classified as “relevant” or “irrelevant”. Then, the relevant news are classified according to their topic.

**Relevance Classification.** We define a news article as relevant if it explicitly refers to a recent or current infectious animal health event. This definition includes several topics to capture all the available online information about an on-going event. It excludes topics such as research or general information about a disease.

In its first version, PADI-web categorized the collected news articles by using a list of 32 outbreak-related keywords, i.e. “positive keywords”. More precisely, news articles were classified as relevant if they contained in the title or the body one of the text positive keywords related to an outbreak event (e.g. outbreak, cases, spread). This approach is called *keyword-based classification*.

To improve the accuracy of the classification, we integrated a classifier based on a supervised machine learning approach [22]. To create a learning dataset, a corpus of 800 annotated news labelled by an epidemiology expert (400 relevant news articles and 400 irrelevant news articles). To obtain a feature representation of the corpus, each document from the corpus was converted into a bag-of-words representation using the Term Frequency - Inverse Term Frequency (TF-IDF) as term weighing method [19]. The meaningless terms (stop-words) and the punctuation are removed, all the remaining terms are lowercased. Using the *scikit-learn* python library [15], a selection of model families is trained on the corpus (random forests, linear support vector classifier, neural networks, etc.). The model obtaining the highest mean accuracy score along the 5-fold cross-validation scheme is subsequently used to classify each newly retrieved article.

The PADI-web interface allows users to manually label the relevance of the retrieved news articles. Each manually labelled article is added to the initial training corpus. Thus, each training step is enriched with the user contribution.

**Topic Classification.** To go beyond the binary relevance classification, we defined more fine-grained categories for the relevant news. These categories were created in collaboration with the FEIS team, and aim at improving the news classification regarding two points. First, a part of the relevant news does not

<sup>4</sup> <https://azure.microsoft.com/en-gb/services/cognitive-services/translator-text-api/>.

directly refer to a disease outbreak, but rather describes a disease-free country in alert or the economic impacts on an affected area several days after an outbreak. Therefore, automatically extracting epidemiological information from their content can generate a number of false positive alerts, which has been identified as a significant limitation of PADI-web performances [3]. Secondly, the news declaring or suspecting an outbreak have a higher priority level than those describing outbreak consequences, for instance. In the context of daily monitoring of a continuous stream of news, it is therefore crucial to correctly identify the topic and prioritize the retrieved news. We present the topic categories as follows, in decreasing priority order:

- *Confirmed outbreak*: the news declares or provides updates about a current or a recent confirmed outbreak<sup>5</sup>;
- *Suspected outbreak*: the news refers to current or a recent cases not yet diagnosed, associated with an explicit suspicion of an infectious disease;
- *Unknown outbreak*: the news refers to current or a recent cases not yet diagnosed, not associated with any suspicion;
- *Preparedness*: the news refers to the alert status of a country at risk of being affected by a disease spreading in a neighbouring area;
- *Impact*: the news refers to the economic, political or social consequences of an outbreak in an affected country or area.

## 2.4 Information Extraction

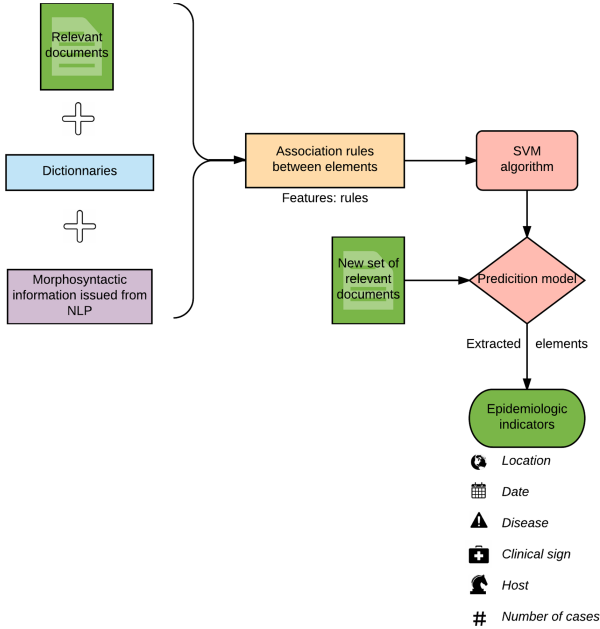
Once a new article is categorized as relevant, PADI-web uses a combined method for information extraction (IE). The combined IE method uses dictionaries and machine learning algorithms (Fig. 2). It allows the identification of key pieces of epidemiological information in the free text (epidemiological events), i.e. location and date of an outbreak, affected hosts, their numbers and encountered clinical signs.

Firstly, our method uses the previously defined dictionaries (Sect. 2.1) to identify relevant candidates for extraction of disease names, hosts, and clinical signs in a given free text. External resources, such as GeoNames [1] and Heidelberg-Time [20] allow the detection of the location and date of a given outbreak. The number of infected cases is recognized using regular expressions.

Secondly, as some of the candidates might be incorrect (e.g. not every date mentioned in the news is the date of an outbreak), each candidate is tested against a set of rules that distinguish correct from incorrect candidates. Such rules, which are at the core of the IE, are automatically extracted as association rules [21] from a corpus of 352 news articles where correct/incorrect candidates were manually annotated by two domain experts in epidemiology and health informatics (EA and JR). Finally, these rules are used as features feeding a Support Vector Machine to predict the relevance of a given candidate [13].

---

<sup>5</sup> This definition is only based in the news semantic, and do not take into account the official confirmation by a formal source.



**Fig. 2.** Event extraction method implemented in PADI-web, based on dictionaries and SVM classification.

Finally, once each candidate has been processed, the interface of PADI-web permits users to visualise all extracted elements from a given news article and the location of a given event are associated with a link to Google Maps (Fig. 3). The interface of PADI-web offers users additional features such as trend analysis, i.e. monthly number of relevant news articles for a given topic. Users can also filter outbreak events by disease, hosts, clinical signs, date interval, and source of information. These events can then be downloaded in a structured format.

### 3 Experiments

We evaluated PADI-web in its integrity and for each step of its pipeline. The results of the first two steps of the method are detailed elsewhere [2]. In this work, we present the results from the evaluation of the classification step, the information extraction step and the overall performance of PADI-web.

#### 3.1 Performance of the Classification

**Relevance Classification.** To evaluate the improvement of integrating a supervised classifier, we compared the performances of the *keyword-based classification* to the performances of three classifiers from different model families, i.e. Random Forest [5], Linear Support Vector Machine (Linear SVM) [9] and Multilayer Perceptron [12] on the learning dataset described in Sect. 2.3.

The screenshot displays the PADI-web interface for a news article titled "Another dead pig found on Kinmen beach confirmed infected with ASF". The article is dated April 10, 2019. The interface is annotated with five numbered circles (1-5) highlighting key features:

- 1**: The article title and date.
- 2**: The "KEYWORDS" panel, which lists categories like "disease", "host", "symptom", "various", and "location". Under "disease", "AFRICAN SWINE FEVER" is highlighted in green. Other terms like "PORCINE", "FEVER", "MORTALITY", "OUTBREAKS", "CASE", "CASES", "ASIA", "PEOPLE'S REPUBLIC OF CHINA", and "REPUBLIC OF CHINA (TAIWAN)" are also listed.
- 3**: The "CLASS LABELS" section, where the article is classified as "relevant".
- 4**: The main text of the article, where several words and phrases are highlighted in green, indicating they are estimated to be correct by the algorithm.
- 5**: The "LOCATION" panel, which shows "Country: TW" and "Zone: Fukien", along with a "Go to Map" link and a "Machine label" with a "CONFIDENCE" score of 0.00%.

**Fig. 3.** Print screen of a user interface in PADI-web. The example shows a news article classified as relevant and related to African swine fever (1, 3). The automatically annotated candidates are coloured in green when they are estimated to be correct by the algorithm (4) and summarized in a keywords panel (2). The locations candidates are associated with complementary information: country, administrative zone, and a link to Google Maps (5) (Color figure online).

**Topic Classification.** To evaluate the performances of supervised classifiers for the topic classification, an epidemiologist first annotated the initial set of 400 relevant news from the learning dataset (Sect. 2.3). This first annotation phase led to a very imbalanced dataset, the class *confirmed outbreak* being over-represented among the other classes. Thus, excluding the *confirmed outbreak* class, we further increased the dataset by annotating additional relevant news extracted from PADI-web database until reaching balanced classes. The final dataset contained 631 news, distributed as follows: *confirmed outbreak*: 308 news, *suspected outbreak*: 86 news, *unknown outbreak*: 77 news, *preparedness*: 80 news, and *impact*: 80 news.

We evaluated two textual representations (Fig. 4):

- $R_1$ : Bag-of-words representation
- $R_2$ :  $R_1$  enriched with a terms-count matrix

$R_1$  corresponds to the bag-of-words matrix obtained after the pre-processing steps described in Sect. 2.3. We further enriched this matricial representation with 5 features, i.e. the counts of diseases, hosts, symptoms, outbreak-related terms, and mystery-related terms in the news. More precisely, in each news, we counted the total number of terms belonging to each category listed here-above (using 5 lists of terms), thus obtaining a term-count matrix. This matrix was

concatenated with the initial bag-of-words matrix, thus increasing the representation space by 5 columns.

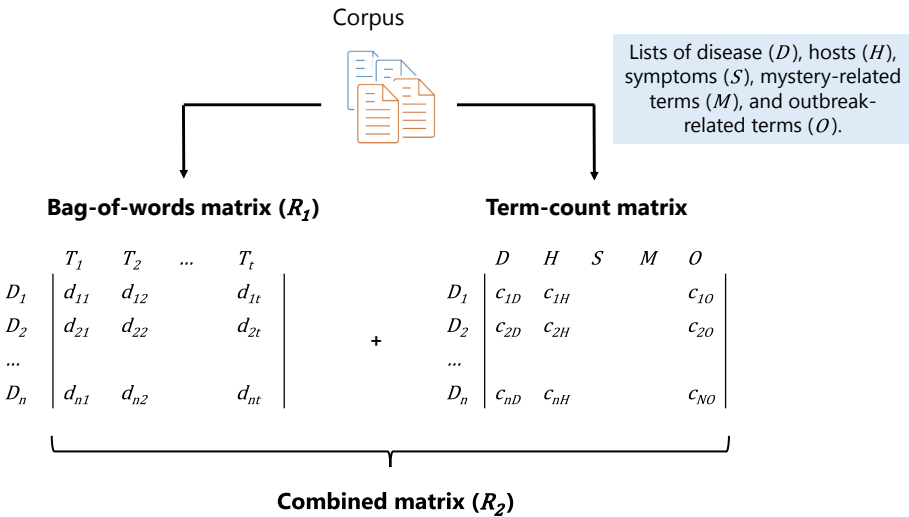
As lists of terms, for diseases, hosts and symptoms, we used the previously defined dictionaries (Sect. 2.1). We further enriched the disease and host lists with synonyms from UMLS and Agrovoc. The disease list and host list contain 1,967 terms and 265 terms respectively. For symptoms, we enriched the vocabulary with the list of disorders from Agrovoc and we manually retrieved the clinical signs from the technical disease cards of the World Organisation for Animal Health (OIE)<sup>6</sup>, obtaining a final list of 798 terms. For mystery-related terms, we manually created a list of terms related to the mysterious and unknown aspect of an event. Outbreak-related terms consists in the list of 32 keywords described in Sect. 2.3.

We compared the same classifiers as described previously (i.e. Random Forest, Linear Support Vector Machine and Multilayer Perceptron).

Both relevance classification and topic classification were evaluated through a 5-fold cross-validation scheme.

### 3.2 Performance of the Information Extraction

The IE step was evaluated on a set of 352 manually labelled news articles. These articles were acquired from Google News and covered content related to



**Fig. 4.** Two types of textual representations used for the topic classification, where  $d_{ij}$  is the weight of the term  $j$  in the document  $i$  (bag-of-words matrix), and  $c_{iX}$  in the count of terms  $X$  in the document  $i$  (term-count matrix).  $R_2$  is the contatnation of both representations.

<sup>6</sup> <https://www.oie.int/en/animal-health-in-the-world/technical-disease-cards/>.



a reporting of at least one disease outbreak in animals from 2014 to 2015. For each relevant document, the information about the candidates was automatically found, i.e. using the dictionaries (Sect. 2.1) for each type of information (disease, host, clinical sign, etc.). Next, the information about each candidate was annotated by two independent annotators (EA and JR) as: i) correct, when the corresponding candidate correctly described the desired piece of information, and ii) incorrect, when the candidate had no link to the corresponding event. We validated the machine learning approach, i.e. Support Vector Machine (SVM), by using a ten-fold cross-validation technique.

Both Data classification and Information extraction were evaluated using the following metrics:

- **Accuracy** measures the overall correctness of classification, i.e. the fraction of documents (resp. candidates) correctly classified from the total number of documents (resp. candidates).
- **Precision** indicates the correctness of classification, i.e. the fraction of documents (resp. candidates) correctly classified in a class  $i$  from the total number of documents (resp. candidates) classified in class  $i$ .
- **Recall** indicates the completeness of classification, i.e. the fraction of documents (resp. candidates) correctly classified in a class  $i$  from the total number of documents (resp. candidates) from class  $i$ .
- **F-score** is the harmonic mean of precision and recall.

### 3.3 Overall Performance of PADI-web

The general performance of PADI-web was evaluated during a six-month period, from 1st January to 28th June 2016. We evaluated the performance in terms of precision, sensitivity and timeliness of PADI-web to alert signals of emergence of African swine fever (ASF), foot-and-mouth disease (FMD), bluetongue (BTV), and highly pathogenic avian influenza (HPAI) at international level.

We considered as a gold standard all official reports for these diseases, freely available from the Empres-i database of the Food and Agriculture Organization (FAO). This database consists of verified information about new epidemiological events (immediate notifications) and ongoing outbreaks (follow-up reports) from a list of obligatory notifiable animal infectious diseases.

**Precision** is the capacity of PADI-web to alert about an event which corresponded to a verified outbreak (TP) from the Empres-i database (immediate notification or follow-up report) or an event that was judged as relevant by a veterinary epidemiologist (one of the authors of this work, EA). A False Positive (FP) was an event not found in the Empres-i database or an event which was annotated by the veterinary epidemiologist as irrelevant.

**Sensitivity** is the capacity of PADI-web to alert about an exceptional epidemiologic event (immediate notification) reported in the Empres-i database (TP). From January to June 2016 a total of 11 outbreaks for ASF, 15 for FMD, 8 for BTV, and 26 for HPAI, respectively, were reported in the Empres-i database.

False Negatives (FN) were all outbreaks from the immediate notifications that were not detected by PADI-web.

## 4 Results

### 4.1 Performance of the Classification

**Relevance Classification.** The *keyword-based classification* obtained imbalanced performances: the precision for the relevant class was 0.76 while its recall reached 0.97 (Table 1). The three classifiers included in this study outperformed *keyword-based classification*, increasing the precision of the relevant class up to 0.96 (linear SVM classifier). Among all the classifiers, the linear SVM classifier obtained the highest accuracy (0.94) and F-scores (0.95 and 0.92 for the relevant and irrelevant classes respectively).

**Table 1.** Comparison of the keyword-based method and three supervised classifiers (RF: Random Forest, MLP: Multilayer Perceptron, linear SVM: linear Support Vector Machine) for the classification of the relevance, in terms of precision, recall, F-score per class, and overall accuracy. For each class, the best performances are shown in bold.

Classification	Class	Precision	Recall	F-score	Accuracy
Keyword-based	Relevant	0.76	0.97	0.85	0.80
	Irrelevant	0.91	0.53	0.67	
RF	Relevant	0.86	<b>0.98</b>	0.92	0.86 ± 0.03
	Irrelevant	<b>0.96</b>	0.75	0.84	
MLP	Relevant	0.93	0.95	0.94	0.93 ± 0.03
	Irrelevant	0.92	0.90	0.91	
Linear SVM	Relevant	<b>0.96</b>	0.94	<b>0.95</b>	<b>0.94 ± 0.02</b>
	Irrelevant	0.91	<b>0.93</b>	<b>0.92</b>	

**Topic Classification.** The overall performances of the topic classification were lower than the relevance classification (Table 2). MLP and RF classifiers obtained comparatively equal performances, obtaining the highest accuracy (0.75) and F-score (0.73). The enriched representation  $R_2$  outperformed the bag-of-words representation  $R_1$  regarding all the scores evaluated.

The recall was heterogeneous between the different classes (Table 3), varying from 0.34 (*preparedness*) to 0.90 (*confirmed outbreak*). The best F-scores were obtained for *confirmed outbreak* (F-score = 0.82) and *impact* (F-score = 0.82).

News wrongly classified as *confirmed outbreak* represented respectively 56%, 31% and 18% of the categories *preparedness*, *suspected outbreak* and *impact* (Table 4). Only 6% of the *confirmed outbreak*, 1% *suspected outbreak* and none of the *unknown outbreak* news were classified as *preparedness* or *impact* categories.

**Table 2.** Comparison of three supervised classifiers (RF: Random Forest, MLP: Multilayer Perceptron, linear SVM: linear Support Vector Machine) and two textual representation ( $R_1$ : bag-of-words representation,  $R_2$ :  $R_1$  enriched which term-count representation) for the topic classification, in terms of weighted mean precision, recall, F-score, and overall accuracy. The best performances are shown in bold.

Classifier	Representation	Precision	Recall	F-score	Accuracy
RF	$R_1$	0.64	0.62	0.57	$0.62 \pm 0.04$
	$R_2$	0.66	0.65	0.60	$0.65 \pm 0.04$
MLP	$R_1$	0.68	0.68	0.65	$0.70 \pm 0.05$
	$R_2$	<b>0.76</b>	<b>0.75</b>	<b>0.73</b>	<b><math>0.75 \pm 0.05</math></b>
Linear SVM	$R_1$	0.70	0.70	0.68	$0.70 \pm 0.07$
	$R_2$	0.75	<b>0.75</b>	<b>0.73</b>	<b><math>0.75 \pm 0.04</math></b>

**Table 3.** Comparison of performances scores in the different classes with linear SVM classifier.

Class	Precision	Recall	F-score
Confirmed outbreak	0.75	0.90	0.82
Suspected outbreak	0.69	0.51	0.59
Unknown outbreak	0.76	0.79	0.78
Preparedness	0.68	0.34	0.45
Impact	0.82	0.78	0.79

**Table 4.** Normalized confusion matrix obtained from the topic classification step with linear SVM classifier. The figures correspond to the percentage of news in each actual class (rows) classified in each predicted class (columns).

		Predicted				
		Confirmed outbreak	Suspected outbreak	Unknown outbreak	Preparedness	Impact
Actual	Confirmed outbreak	90	3	1	4	2
	Suspected outbreak	31	51	16	0	1
	Unknown outbreak	10	10	79	0	0
	Preparedness	56	0	1	34	9
	Impact	18	13	1	1	78

## 4.2 Performance of the Information Extraction

The accuracy of the IE step was higher than 0.80, with the lowest accuracy occurring when detecting locations and the highest for names of diseases and host species (Table 5). Similarly, in terms of F-score, the IE obtained a score of 0.80 for locations, 0.85 for dates, 0.95 for diseases, 0.85 for numbers of cases, and 0.95 for hosts in the free text news articles.

**Table 5.** Performance of the information extraction step of the text-mining approach.

Entity	Precision	Recall	F-score	Accuracy
Location	0.81	0.80	0.80	0.80
Date	0.82	0.88	0.85	0.88
Disease	0.95	0.96	0.95	0.96
Number of cases	0.86	0.85	0.85	0.85
Host	0.94	0.96	0.95	0.96

## 4.3 Overall Performance of PADI-web

In the first six months of its implementation, from January to June 2016, PADI-web alerted on 123 outbreaks for ASF, 191 for FMD, 71 for BTV, and 632 for AI. The precision of PADI-web to adequately alert for true outbreaks was 30% (37/123 events) for ASF, 27% (51/191 events) for FMD, 45% (32/71) for BTV, and 54% (342/632) for AI. The sensitivity of PADI-web to alert for exceptional epidemiological events was 73% (8/11 events) for ASF, 20% (3/15 events) for FMD, 13% (1/8 events) for BTV, and 62% (16/26 events) for HPAI.

## 5 Conclusion

PADI-web, which is operational since January 2016, is a text mining tool specialized in monitoring the Web for the emergence of new animal infectious diseases.

The implementations of a supervised classifier to automatically filter out irrelevant news significantly improved the accuracy of the relevance classification in PADI-web, which is crucial to control the amount of daily news presented to the FEIS experts. The evaluation of the integration of a topic classification step for the relevant news obtained promising results. However, numbers of *preparedness* and *suspected outbreak* news were classified as *confirmed outbreak*. The alert status in a country is generally consecutive to an on-going outbreak in another area, thus the topics *preparedness* and *confirmed outbreak* can overlap in a same news. In addition, the distinction between a confirmed and a suspected outbreak often solely relies on the use of a different adjective (use of “detected” or “confirmed” instead of “suspected” or “investigated”). In practice, misclassification

between *confirmed outbreak*, *suspected outbreak* and *unknown outbreak* is not a strong limitation, since all these categories have a high priority level. Increasing the learning dataset and enriching the feature representation may improve the accuracy of topic classification.

The results from the evaluation of IE show that this method is suitable for the extraction of epidemiological indicators. However, the evaluation of the IE step on the data stream from PADI-web showed a lower precision to correctly alert on events of epidemiological relevance. For example, a number of alerts produced by PADI-web were actually false positives as they corresponded to a location irrelevant to a disease outbreak. This suggests that further improvements, possibly training the SVM algorithm on a larger dataset, can improve the outcomes of the predicted model from the IE step. Our first evaluations showed, however, a higher sensitivity of PADI-web, especially for HPAI and ASF. These results suggest that new epidemiological events are reported in the media, particularly zoonotic events or outbreaks of significant economic impact.

In its first version, PADI-web used Google News in English as the main source of information. This might have influenced the results of the performance evaluation, especially the sensitivity for BTV and FMD which mostly occurred in Central Europe, Southern Africa and Asia. In order to overcome this limitation, the new version of PADI-web integrates a multilingual module to increase the local and regional coverage.

We believe these updates significantly improved the operational use of PADI-web by animal health authorities interested in integrating an event-based surveillance component to their epidemic intelligence activities. Furthermore, in December 2019, PADI-web detected COVID-19 related news through RSS feeds designed for animal health, highlighting its genericity and its ability to detect public health emergence signals.

**Acknowledgements.** We thank J. de Goër, B. Belot, C. Hemeury, M. Devaud, and T. Filiol for their contribution in the development of PADI-web. We also thank the members of the French Epidemic Intelligence Team in Animal Health for their constructive comments during the development of PADI-web. This work has been supported 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 as ANR-16-CONV-0004 (#DigitAg). This work has also been funded by the “Monitoring outbreak events for disease surveillance in a data science context” (MOOD) project from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 874850 (<https://mood-h2020.eu/>).

## References

1. Ahlers, D.: Assessment of the accuracy of GeoNames gazetteer data. In: Proceedings of the 7th Workshop on Geographic Information Retrieval, pp. 74–81. ACM, New York (2013)

2. Arsevska, E.: Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Comput. Electron. Agric.* **123**, 104–115 (2016). <https://doi.org/10.1016/j.compag.2016.02.010>
3. Arsevska, E., et al.: Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLoS ONE* **13**(8), e0199960 (2018). <https://doi.org/10.1371/journal.pone.0199960>
4. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**(Database issue), D267–D270 (2004). <https://doi.org/10.1093/nar/gkh061>
5. Breiman, L.: Random Forests. *Mach. Learn.* **45**(1), 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
6. Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D.: Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the healthmap project. *PLOS Med.* **5**(7), 1–6 (2008). <https://doi.org/10.1371/journal.pmed.0050151>
7. Collier, N., Doan, S.: GENI-DB: a database of global events for epidemic intelligence. *Bioinformatics* **28**(8), 1186–1188 (2012). <https://doi.org/10.1093/bioinformatics/bts099>
8. Collier, N., et al.: BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* **24**(24), 2940–2941 (2008). <https://doi.org/10.1093/bioinformatics/btn534>
9. Joachims, T.: Text categorization with Support Vector Machines: learning with many relevant features. In: Nédellec, Nédellec, Rouveirol, C. (eds.) *ECML 1998*. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0026683>
10. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. *Artif. Intell. Med.* **65**(2), 131–143 (2015)
11. Madoff, L.C.: ProMED-Mail: an early warning system for emerging diseases. *Clin. Infect. Dis.* **39**(2), 227–232 (2004). <https://doi.org/10.1086/422003>
12. Murtagh, F.: Multilayer perceptrons for classification and regression. *Neurocomputing* **2**(5), 183–197 (1991). [https://doi.org/10.1016/0925-2312\(91\)90023-5](https://doi.org/10.1016/0925-2312(91)90023-5)
13. Nahm, U.Y., Mooney, R.J.: Using information extraction to aid the discovery of prediction rules from text. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining, KDD-2000 Workshop on Text Mining*, pp. 51–58 (2000)
14. Paquet, C., Coulombier, D., Kaiser, R., Ciotti, M.: Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro. Surveill.* **11**(12), 212–214 (2006). 665 [pii]
15. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
16. Steinberger, R., Fuart, F., van der Goot, E., Best, C., von Etter, P., Yangarbe, R.: Text Mining from the Web for Medical Intelligence. *NATO Science for Peace and Security Series, D: Information and Communication Security*, pp. 295–310 (2008)
17. Richardson, L.: Beautiful soup documentation (April 2007)
18. Robertson, C., Yee, L.: Avian influenza risk surveillance in North America with online media. *PLoS ONE* **11**(11), 1–21 (2016). <https://doi.org/10.1371/journal.pone.0165688>
19. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* **24**(5), 513–523 (1988). [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)

20. 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, pp. 321–324 (July 2010)
21. Uno, T., Asai, T., Uchida, Y., Arimura, H.: LCM: an efficient algorithm for enumerating frequent closed item sets. In: Proceedings of Workshop on Frequent Itemset Mining Implementations, FIMI 2003 (2003)
22. Valentin, S., et al.: PADI-web: a multilingual event-based surveillance system for monitoring animal infectious diseases. *Comput. Electron. Agric.* **169**, 105163 (2020). <https://doi.org/10.1016/j.compag.2019.105163>