Proceedings of MultilingualBIO 2020, pages 33–36
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automated Processing of Multilingual Online News for the Monitoring of Animal Infectious Diseases

Sarah Valentin[1,2,*], **Renaud Lancelot**[1], **Mathieu Roche**[2]
[1] UMR ASTRE, CIRAD, F-34398 Montpellier, France.
ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France.
[2] UMR TETIS, CIRAD, F-34398 Montpellier, France.
TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.
[*] Corresponding author: sarah.valentin@cirad.fr

Abstract

The Platform for Automated extraction of animal Disease Information from the web (PADI-web) is an automated system which monitors the web for detecting and identifying emerging animal infectious diseases. The tool automatically collects news via customised multilingual queries, classifies them and extracts epidemiological information. We detail the processing of multilingual online sources by PADI-web and analyse the translated outputs in a case study.

Keywords: Animal health, Web monitoring, Text mining, Multilingual

1. Introduction

The timely detection of (re)emerging animal infectious diseases worldwide is a keystone for risk assessment and risk management regarding both human and animal health. Traditional surveillance relies on official notifications from intergovernmental organisations such as the World Organisation for Animal Health (OIE) and the Food and Agriculture Organization of the United Nations (FAO). While these systems provide verified and structured information, they are prone to notification delays and are not appropriate to detect new threats. To enhance early detection performances, surveillance activities increasingly integrate unstructured data from informal sources such as online news (Bahk et al., 2015). The daily curation and analysis of web-based information are time-consuming. Thus, several systems were designed to automatize the monitoring of online sources regarding a wide range of health threats, such as MedISys (Mantero et al., 2011), HealthMap (Freifeld et al., 2008), GPHIN (Blench, 2008), ProMED (Madoff, 2004) or PADI-web (Valentin et al., 2020). PADI-web[1] (Platform for Automated extraction of Disease Information from the web) is an automated system dedicated to the monitoring of online news sources for the detection of animal health infectious diseases. PADI-web was developed to suit the need of the French Epidemic Intelligence System (FEIS, or Veille sanitaire internationale in French), which is part of the animal health epidemiological surveillance Platform (ESA Platform). The tool automatically collects news with customised multilingual queries, classifies them and extracts epidemiological information. In this paper, we describe how the PADI-web pipeline processes multilingual textual data. We also provide a case study to highlight the added-value of integrating multiple languages for web-based surveillance.

2. Multilingual news processing

PADI-web pipeline includes four consecutive steps (Figure 1), extensively detailed elsewhere (Valentin et al., 2020): data collection, data processing, data classification and information extraction.

2.1. Data collection

PADI-web collects news articles from Google News on a daily basis, through two types of customised really simple syndication (RSS) feeds (Arsevska et al., 2016). Disease-based feeds target specific monitored diseases, thus they contain disease terms such as *avian flu* or *African swine fever*. To be able to detect emerging threats or undiagnosed diseases, PADI-web also relies on symptom-based RSS feeds. These feeds consist of combinations of symptoms and species (hosts), for instance, *abortions AND cows*. To retrieve non-English sources, we also implemented non-English feeds by translating existing ones in other languages. The languages were selected to target risk areas regarding specific diseases (e.g. we integrated RSS feeds in Arabic for monitoring foot-and-mouth disease in endemic countries). To translate the disease terms, we used Agrovoc[2], a controlled vocabulary developed by the Food and Agriculture Organization (FAO).

2.2. Data processing

PADI-web fetches all the news webpages retrieved by the RSS feeds. The title and text of each news article are cleaned to remove irrelevant elements (pictures, ads, hyperlinks, etc.). The language of the source is detected using the *langdetect* python library. All non-English news articles are translated into English using the Translator API of the Microsoft Azure system [3].

2.3. Data classification

To select the relevant news (i.e. the news describing a current outbreak as well as prevention and control measures,

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

[2] http://aims.fao.org/vest-registry/vocabularies/agrovoc

[3] https://azure.microsoft.com/en-gb/services/cognitive-services/translator-text-api/

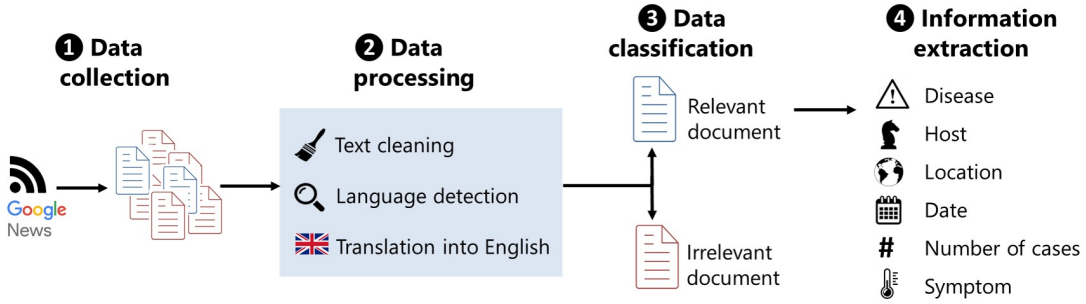

Figure 1: PADI-web pipeline

preparedness, socioeconomic impacts, etc.), PADI-web relies on an automated classifier developed with a supervised machine learning approach. The training dataset consists in a corpus of 600 annotated pieces of news labelled by an epidemiology expert (200 relevant news articles and 400 irrelevant news articles). Using the scikitlearn python library, several models from different families are trained:

- Linear classifiers: Logistic Regression, Gaussian and Multinomial Naive Bayes
- Support vector machines: Linear Support Vector Machine (Linear SVM)
- Decision tree: Random Forest
- Quadratic classifiers: Quadratic Discriminant Analysis
- Instance-based learning models: K-nearest neighbor learner
- Neural networks: Multilayer Perceptron

The model obtaining the highest mean accuracy score along the 5-fold cross-validation scheme is subsequently used to classify each new retrieved article. Currently, Random Forest (composed of 50 trees with a maximum depth of 12) and the Multilayer Perceptron are the best classifiers, obtaining an average accuracy score of 0.944 ± 0.01 (Table 1).

Classifier	Average accuracy score	Standard deviation
Multilayer Perceptron	**0.944**	0.01
Random Forest	**0.944**	0.01
Linear SVM	0.935	0.02
Quadratic Discriminant Analysis	0.905	0.01
Gaussian Naive Bayes	0.896	0.01
K-nearest neighbor learner, K=2	0.896	0.04
Logistic Regression	0.881	0.04
Multinomial Naive Bayes	0.867	0.04

Table 1: Results of the relevance classification in terms of average accuracy score, for different classifiers.

2.4. Information extraction

The extraction of epidemiological information relies on a combined method founded on rule-based systems and data mining techniques (Arsevska et al., 2018). Diseases, hosts and symptoms are extracted using a list of terms of disease names, hosts and clinical signs. To obtain our list of terms, we use BioTex (Lossio-Ventura et al., 2014), a tool for automatic extraction of biomedical terms from free text, as detailed elsewhere (Arsevska et al., 2016). Locations are identified by matching the text with location names from the GeoNames gazetteer (Ahlers, 2013) and dates with the rule-based HeidelTime system (Strotgen and Gertz, 2010). The number of cases is extracted from a list of regular expressions matching numbers in numerical or textual form. A confidence index is automatically assigned to the extracted entities to reflect the probability that they correspond to the desired piece of epidemiological information.

The models for classification (Section 2.3.) and information extraction (Section 2.4.) tasks have been learnt with labeled data in English. English is a "bridge-language" (or "pivot language") for PADI-web. In this context, a translation method has been applied for non-English news before using the classification and information extraction algorithms of the PADI-web pipeline.

3. Case study

We conducted a preliminary case study to evaluate the processing of non-English sources by PADI-web.

3.1. Methods

We extracted the translated news articles from PADI-web database from 01 July 2019 to 31 July 2019 (1 month period). We manually reviewed each news to select the ones containing an animal disease event. An event corresponds to the occurrence of a disease at a specific location and date. Then, we compared the detected events with official events extracted from the FAO Emergency Prevention System for Priority Animal and Plant Pests and Disease (EMPRES-i)[4]. This system receives information from different official data sources, such as governments or OIE, and is a global reference database for animal diseases. We calculated the delay between the official notification and the detection by PADI-web (corresponding to the publication date of the news article). The events present in online news but absent from the official database are considered as unofficial (they cannot be verified). For both official and unofficial events detected by non-English sources, we deter-

[4] http://empres-i.fao.org/eipws3g/.

Disease	Country (no of events)	Source language[a]	Detected in English news	Range of detection delays (days)[b]
African swine fever	Bulgaria (n=1)	TR, IT	yes	-10
	China (n=6)	FR, KO, ZH-CN	yes	- 12 to 1
	Laos (n=1)	ZH-CN	no	5
	Slovakia (n=1)	DE, IT, ZH-CN	yes	0
Avian influenza	Denmark (n=1)	KO	no	4
	Mexico (n=1)	KO	no	1
	Taiwan (n=2)	KO	no	-12 to -2
Foot-and-mouth disease	Morocco (n=1)	AR, FR	no	-5 to 0

Table 2: Official events detected by non-English sources.
[a]Languages: AR: Arabic, DE: German, FR: French, IT: Italian, KO: Korean, TR: Turkish, ZH-CN: Chinese.
[b] Lag between the official notification and the detection by PADI-web.

mined if they were also detected by English news retrieved by PADI-web during the same period.

3.2. Results and discussion

From 01 July 2019 to 31 July 2019, PADI-web retrieved 104 online news, among which 47 online news contained one or several animal disease events. The remaining 57 news were related to control measures (n=34), outbreak follow-up (n=6), human disease outbreak (n=7), disease awareness (n=2), or were irrelevant (n=8). The low number of irrelevant news (8/104) indicates that the classification module was able to perform well on translated news.
The information extraction module extracted 93 disease entities, 218 host entities, 47 dates, 584 locations, 125 symptoms and 45 numbers of cases. PADI-web detected 14 distinct official events from 35 non-English online news (Table 2), involving 3 diseases and 8 countries. English-news did not detect six out of 14 events. The events were detected up to 12 days before their official notification. Besides, PADI-web discovered 5 unofficial events from 12 non-English online news (Table 3.), among which 4 were not detected by English-news.

Disease	Country (no of events)	Source (language[a])	Detected in English news
Anthrax	Guinea (n=1)	FR	no
Eastern equine encephalitis	USA (n=1)	IT	yes
Foot-and-mouth disease	Morocco (n=1)	AR, FR	no
Lumpy skin disease	Kazakhstan (n=1)	RU	no
Peste des petits ruminants	Algeria (n=1)	FR	no

Table 3: Unofficial events detected by non-English sources.
[a]Languages: AR: Arabic, FR: French, IT: Italian, RU: Russian.

During one-month, the non-English sources increased both the sensitivity and the timeliness of PADI-web in detecting official events. This is consistent with the fact that local sources are more reactive in reporting outbreaks from their area or country. The added value of integrating multilingual sources was also highlighted by an in-depth comparison of event-based tools in the human health domain (Barboza et al., 2014).

During the manual analysis, we found out that two diseases were wrongly translated. The most frequent errors occurred when translating *African swine fever* from several languages. We found the following wrong expressions: *African swine plague*, *African pig plague*, *plague of pig*, *African wild boar plague*. In one piece of Chinese news, the translated form was *swine fever in Africa*, which led to the detection of a false location (*Africa*). Errors also occurred in its acronyms translation (*ASP* instead of *ASF*). From Russian news, lumpy skin disease was translated as *nodular dermatitis*. Many animal disease names consist in a combination of host, symptom and location terms. Thus, they are prone to translation errors which should be taken into account to avoid impacting the performances of monitoring tools. Translated texts underline the limits of relying on vocabulary matching for entity recognition. Existing NER models based on machine learning do not include domain-specific entities such as diseases and hosts. However, the python package spaCy allows adding new classes to the default entities, by training its existing model with new labelled examples. Such an approach could enhance the detection of out of vocabulary terms produced by translation.

4. Conclusion

We described how we integrated multilingual sources in the existing PADI-web system. The preliminary evaluation yielded promising results regarding the added-value of integrating non-English news to web-based surveillance. In future work, we will conduct a more in-depth analysis of the translated outputs in terms of sensitivity and timeliness, and we will evaluate the quality of the geographical entities after applying the translation task. Besides, we aim to improve the detection of named entities such as disease names by training NER models.

5. Acknowledgements

This work was 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 MOOD project. This work was supported by the French National Research

Agency (ANR) under the Investments for the Future Program (ANR-16-CONV-0004). This work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement MOOD 874850. The authors thank J. Rabatel, E. Arsevska, S. Falala and the members of the French Epidemic Intelligence in Animal Health (A. Mercier, J. Cauchard and C. Dupuy) for their contribution and expertise in developing PADI-web.

6. Bibliographical References

Ahlers, D. (2013). Assessment of the Accuracy of GeoNames Gazetteer Data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pages 74–81, New York, NY, USA. ACM.

Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., and Dufour, B. (2016). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, 123:104–115.

Arsevska, E., Valentin, S., Rabatel, J., de Goër de Hervé, J., Falala, S., Lancelot, R., and Roche, M. (2018). Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLOS ONE*, 13(8):e0199960, August.

Bahk, C. Y., Scales, D. A., Mekaru, S. R., Brownstein, J. S., and Freifeld, C. C. (2015). Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. *BMC Infectious Diseases*, 15(1), December.

Barboza, P., Vaillant, L., Le Strat, Y., Hartley, D. M., Nelson, N. P., Mawudeku, A., Madoff, L. C., Linge, J. P., Collier, N., Brownstein, J. S., and Astagneau, P. (2014). Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PLoS ONE*, 9(3):e90536, March.

Blench, M. (2008). Global public health intelligence network (GPHIN). In *8th Conference of the Association for Machine Translation in the Americas*, pages 8–12.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, 15(2):150–157, March.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014). BIOTEX: A system for Biomedical Terminology Extraction, Ranking, and Validation. In *International Semantic Web Conference*.

Madoff, L. C. (2004). ProMED-mail: an early warning system for emerging diseases. *Clinical infectious diseases*, 39(2):227–232.

Mantero, J., Belyaeva, J., Linge, J., European Commission, Joint Research Centre, and Institute for the Protection and the Security of the Citizen. (2011). *How to maximise event-based surveillance web-systems: the example of ECDC/JRC collaboration to improve the performance of MedISys.* Publications Office, Luxembourg. OCLC: 870614547.

Strotgen, J. and Gertz, M. (2010). HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, July.

Valentin, S., Arsevska, E., Falala, S., de Goër, J., Lancelot, R., Mercier, A., Rabatel, J., and Roche, M. (2020). PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169:105163, February.