Data in Brief 46 (2023) 108870

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

An annotated dataset for event-based surveillance of antimicrobial resistance

Nejat Arınık[a,c], Wim Van Bortel[d], Bahdja Boudoua[a,c], Luca Busani[e], Rémy Decoupes[a,c], Roberto Interdonato[b,c], Rodrique Kafando[a,c], Esther van Kleef[f], Mathieu Roche[b,c], Mehtab Alam Syed[b,c], Maguelonne Teisseire[a,c,*]

[a] INRAE, Montpellier F-34398, France
[b] CIRAD, Montpellier F-34398, France
[c] TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier 34090, France
[d] ITM, Institute of Tropical Medicine, Department of Biomedical Sciences, Antwerp, Belgium
[e] Center for Gender-Specific Medicine, Istituto Superiore di Sanitá Viale Regina Elena 299, 00161 Rome, Italy
[f] ITM, Institute of Tropical Medicine, Department of Public Health, Outbreak Research Team, Antwerp, Belgium

ARTICLE INFO

Article history:
Received 16 November 2022
Revised 15 December 2022
Accepted 27 December 2022
Available online 1 January 2023

Dataset link: MOOD - News AMR dataset - Hackathon 2022 (Original data)

Keywords:
Antimicrobial resistance (AMR)
Text mining
Annotation
Epidemiology

ABSTRACT

This paper presents an annotated dataset used in the MOOD Antimicrobial Resistance (AMR) hackathon, hosted in Montpellier, June 2022. The collected data concerns unstructured data from news items, scientific publications and national or international reports, collected from four event-based surveillance (EBS) Systems, i.e. ProMED, PADI-web, HealthMap and MedISys. Data was annotated by relevance for epidemic intelligence (EI) purposes with the help of AMR experts and an annotation guideline. Extracted data were intended to include relevant events on the emergence and spread of AMR such as reports on AMR trends, discovery of new drug-bug resistances, or new AMR genes in human, animal or environmental reservoirs. This dataset can be used to train or evaluate classification approaches to automatically identify written text on AMR events across the different reservoirs and sectors of One Health (i.e. human, animal,

* Corresponding author at: TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier 34090, France.
E-mail addresses: nejat.arinik@inrae.fr (N. Arınık), wvanbortel@itg.be (W. Van Bortel), el-bahdja.boudoua@inrae.fr (B. Boudoua), luca.busani@iss.it (L. Busani), remy.decoupes@inrae.fr (R. Decoupes), roberto.interdonato@cirad.fr (R. Interdonato), rodrique.kafando@inrae.fr (R. Kafando), evankleef@itg.be (E. van Kleef), mathieu.roche@cirad.fr (M. Roche), mehtab-alam.syed@cirad.fr (M. Alam Syed), maguelonne.teisseire@inrae.fr (M. Teisseire).

https://doi.org/10.1016/j.dib.2022.108870

food, environmental sources, such as soil and waste water) in unstructured data (e.g. news, tweets) and classify these events by relevance for EI purposes.

Specifications Table

Subject	Data Science: Data Mining and Statistical Analysis
Specific subject area	Identification and Classification of AMR information in news data
Type of data	Tables for corpora (*.csv) and pdf for guidelines.
How data were acquired	Manually with different platforms (ProMED, PADI-web, HealthMap and MedISys)
Data format	Raw and Standardized
Parameters for data collection	Keywords dedicated to AMR for each platform.
Description of data collection	The dataset constitutes (i) four table files (one per EBS system) for thematic classification, (ii) four table files (one per EBS system) for host classification, (iii) one pdf document for the annotation guidelines.
Data source location	The data are hosted on the INRAE Dataverse. The data were manually collected within the UMR TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France in the context of the MOOD (MOnitoring Outbreaks for Disease surveillance in a data science context) project.[1]
Data accessibility	Repository name: Data INRAE (Dataverse) Data identification number: doi:10.57745/MPNSPH [1] Direct URL to data: https://doi.org/10.57745/MPNSPH

Value of the Data

- This dataset contributes to the available resources for Natural Language Processing (NLP) on specialized domains and more precisely in the field of AMR surveillance and epidemic intelligence.
- It is useful for computer scientists for NLP and data mining tasks.
- It can be used for evaluation or training purposes for classification tasks.
- It has a degree of heterogeneity capable to cover different disciplines of the AMR domain
- It is useful for the detection of information across the One Health domains (humans, animals and environment) and interactions between these domains (e.g. transmission or exchange of AMR bacteria or AMR genes within and between domain reservoirs)
- It can support the development of methodology that can automatically classify EBS AMR data, hence facilitate relevant epidemic intelligence activities for AMR across the One Health domains

1. Objective

The emergence and spread of drug-resistant pathogens has led to antimicrobial resistance (AMR) being considered a major public health problem. AMR surveillance in Europe and elsewhere helps monitor trends and detect potential emerging threats. These surveillance activities - commonly referred to as epidemic intelligence - largely concern indicator-based surveillance

[1] https://mood-h2020.eu/

(IBS), i.e. the monitoring of pre-defined indicators based on clear case definitions and structured data collection by public, animal, and environmental health institutes nationally and globally (e.g. the yearly number of hospital-based bloodstream infections caused by third-generation cephalosporin resistant Enterobacterales).

Unofficial and unstructured data sources such as news articles and data from social media, may provide alternative and complementary data to the official surveillance to address some of the limitations of IBS. The latter is commonly referred to as event-based surveillance (EBS) [2].

It is in this context that the H2020 MOOD project (MOnitoring Outbreaks for Disease surveillance in a data science context) organized a hackathon challenge led by TETIS.[2] The MOOD project aims to explore the potential of data science for epidemic intelligence purposes. This includes data mining, analysis and visualization techniques of health, environmental and other data to enhance the utility of EBS.

2. Data Description

In the course of the MOOD AMR hackathon, interdisciplinary teams were formed that collectively worked on new technical solutions to mine and visualize unstructured data from four EBS systems. The main objective of the hackathon task was to develop and test classification approaches that automatically identified AMR events and reservoirs of concern (e.g. animal, food, environment, etc.) and classified these events by their relevance for epidemic intelligence purposes. Extracted data were intended to include important events on the emergence and spread of AMR such as reports on AMR trends, discovery of new drug-bug resistances, or new AMR genes in human, animal or environmental reservoir.

The corpus dataset was curated and mined by the TETIS team from four EBS-systems: ProMED, PADI-web, HealthMap and MedISys. The four sub-datasets from the respective EBS sources were manually annotated according to three main classes (New Information, General Information, Not Relevant). Data observation labeled as New Information or General Information were subsequently annotated according to a host classification system referring to host and/or transmission reservoirs (e.g. Humans, animals, Environment, etc). An annotation guideline was provided to facilitate a unified manual annotation across three annotators (AMR experts). Each EBS dataset contains 100 articles and two levels of annotation are provided. The datasets composed by both labeled corpora and annotation guidelines are publicly available online.[3]

We have two corpora: D_1 for thematic classification with three 3 main classes (*New Information, General Information, Not Relevant + Don't Know* class) and D_2 for host classification with 7 classes (*Humans, Human-animal, Animals, Human-food, Food, Environment, and All*). Corpus D_2 is a subset of D_1, it concerns only *New Information* and *General Information* data from D_1.

Tables 1 and 2 respectively detail the distribution per class for D_1 and D_2 corpora. Dominant and minor classes are highlighted in order to illustrate the class imbalance for each dataset. The D_1 corpora datasets, as well as D_2 corpora datasets, are named according to their sources as follows:

- $D_{1\cdot1}$-ProMED_url_thematic_classification.csv
- $D_{1\cdot2}$-PADI-web_thematic_classification.csv
- $D_{1\cdot3}$-HealthMap_url_thematic_classification.csv
- $D_{1\cdot4}$-MedISys_url_thematic_classification.csv

- $D_{2\cdot1}$-ProMED_url_host_classification.csv
- $D_{2\cdot2}$-PADI-web_host_classification.csv
- $D_{2\cdot3}$-HealthMap_url_host_classification.csv
- $D_{2\cdot4}$-MedISys_url_host_classification.csv

[2] https://mood-h2020.eu/event/mood-summer-school-2022/
[3] https://doi.org/10.57745/MPNSPH

Table 1
Statistics on D_1 corpus. The columns represent the four considered EBS surveillance systems, and the first four rows correspond to the classes established for thematic classification. The last row simply indicates the total number of documents in each EBS surveillance system. Finally, each entry in the table indicates as a percentage how frequent a specific class in a given surveillance system. These percentage values are highlighted with different gray scales.

Class	$D_{1\cdot1}$-ProMED	$D_{1\cdot2}$-PADI-web	$D_{1\cdot3}$-HealthMap	$D_{1\cdot4}$-MedISys
New Information	50.00%	48.00%	27.00%	08.00%
General Information	48.00%	47.00%	08.00%	10.00%
Not Relevant	02.00%	03.00%	64.00%	81.00%
Don't Know	00.00%	02.00%	01.00%	01.00%
Total	**100 (100%)**	**100 (100%)**	**100 (100%)**	**100 (100%)**

Table 2
Statistics on D_2 corpus. The columns represent the four considered EBS surveillance systems, and the first seven rows correspond to the classes established for host classification. The last row simply indicates the total number of documents in each EBS surveillance system. Finally, each entry in the table indicates as a percentage how frequent a specific class in a given surveillance system. These percentage values are highlighted with different gray scales.

Class	$D_{2\cdot1}$-ProMED	$D_{2\cdot2}$-PADI-web	$D_{2\cdot3}$-HealthMap	$D_{2\cdot4}$-MedISys
Humans	54 (54.54%)	52 (54.73%)	29 (85.29%)	12 (66.66%)
Human-animal	13 (13.13%)	14 (14.73%)	02 (5.88%)	02 (11.11%)
Animals	12 (12.12%)	10 (10.52%)	01 (2.94%)	01 (5.55%)
Human-food	04 (4.04%)	06 (6.31%)	00 (00.00%)	00 (00.00%)
Food	05 (5.05%)	05 (5.26%)	00 (00.00%)	01 (5.55%)
Environment	04 (4.04%)	05 (5.26%)	02 (5.88%)	02 (11.11%)
All	07 (7.07%)	03 (3.15%)	00 (00.00%)	00 (00.00%)
Total	**99 (100%)**	**95 (100%)**	**34 (100%)**	**18 (100%)**

3. Experimental Design, Materials and Methods

The sub-datasets coming from each of different EBS-systems annotated according to the guidelines describing in the following sub-sections.

3.1. Guidelines

The guidelines were designed according to an iterative process made of multiple annotation rounds. The first annotation round was done by three AMR experts.[4] The experts were asked to annotate independently a sample made of 25 AMR related articles extracted from PADI-web [3]. These AMR related articles were annotated to classify their relevance for epidemic intelligence purposes. The annotator had to choose one single label for each of the two classification types:

1. Thematic classification: This class aims to describe the main content of the article.
2. Host resistance and pathogen classification: This class aims to outline which reported host or hosts the AMR event concerned (reservoir or transmission source).

A first guideline version was established with thematic and host classification classification. After the first round, the experts discussed the annotation disagreements, and decided on a reformulation and refinement of the definitions and on the merging of labels. The goal was to make the guideline as generic as possible and as precise as necessary so that non-expert annotators could annotate the AMR articles without running into ambiguity issues. After this fist round of annotations the modifications were integrated in order to established a final guideline version described in Table 3.

[4] Authors 2, 4, and 7

Table 3
Guideline definitions and examples.

Classification	Class	Class description	Example text
Thematic Classification	New information	Describing an AMR event: information of suspicion or confirmation of an AMR outbreak, a new scientific discovery published in reports or scientific articles	A group of Peruvian researchers have detected increased microbial resistance in samples of chicken and beef in markets of Lima city, the capital of the Latin-American country, 21 samples of food presented 36 strains of Escherichia coli of which 26 were resistant to colistin. This antibiotic is a growth promoter still used for different animal chains around the globe, including in the poultry industry.
	General information	A broad description of an AMR situation in a defined location/region, a description of a project dealing with AMR, a description of measures (e.g. economic/political measures, control measures, vaccination) and impact (e.g. direct or indirect financial or political impacts) that are subsequent to AMR event(s).	Each year, the United States sees around 2.8 million antibiotic-resistant infections and more than 35,000 deaths as a result, according to the Centers for Disease Control. Public health experts warn that the overuse of antibiotics in livestock is contributing to this rising threat.
	Irrelevant	Not related to AMR	Vietnam has reported three outbreaks of highly pathogenic H5N6 bird flu among backyard birds
Host Classification	Animals	AMR associated with animals	Veterinary labs across the country submit animal isolates for antibiotic resistance surveillance and whole-genome sequencing
	Humans	AMR associated with humans	Researchers from the Hudson Institute have revealed how AMR spreads inside the human gut, with their results published in the journal Nature Communications.
	Food	AMR associated with food	The salmonella infections were linked with beef obtained in the United States and soft Cheese obtained in Mexico
	Environment	AMR associated with the environment	A new study performed by Newcastle University and the Indian Institute of Technology, Delhi, measured metal and antibiotic resistance in sediments from the Ganges and Yamuna Rivers.
	Human-animal	AMR transmitted to humans through animals	Pet store puppies spread antibiotic-resistant infection, CDC says a national pet store chain, has spread campylobacter infections to 55 people.
	Human-Food	AMR transmitted to humans via food products	AMR continues to remain a serious threat to public health according to the CDC. Through interviews, the affected individuals reported consumption of beef and soft cheeses.
	All	AMR associated with multiple hosts	A total of 4514 isolates from humans, animals, foods, and the environment were reported

3.2. Corpus

This section outlines how the corpus was acquired. We built our corpus from four EBS surveillance systems. Next, we describe these surveillance systems, as well as their data acquisition strategies.

- *HealthMap:* It is operating since 2006. It is an automated and curated aggregator of a broad range of data sources, such as Twitter, Google News, Baidu and SoSo news aggregators and ProMED in nine languages. The automated part of the tool extracts information on the disease, location of the event and the host. The curated outbreak-related information is displayed on a freely available, visual interface. HealthMap addresses a wide range of health threats (human, animal and plant infectious diseases, as well as environmental risks) [4].
- *ProMED:* It is a program of the International Society for Infectious Diseases (ISID) [5]. It was launched in 1994 as an Internet service to identify unusual health events related to emerging and re-emerging infectious diseases and toxins affecting humans, animals and plants. Information is selected for publication by subject matter expert *moderators* who provide written commentary, giving the reader the necessary historical context and/or clinical background to understand the importance of the information being reported. ProMED posts also supply references to previous reports and to the scientific literature. Reports are simultaneously posted to ProMED's website and sent to ProMED subscribers by email.
- *MedISys:* It is an EBS system developed by the European Commission Directorate General for Health and Consumers (DG SANCO) and the Joint Research Centre (JRC) of the European Commission. The system monitors web-based information (media articles and open-source public health reports) about human, animal and plant infectious diseases, chemical, biological, radiological and nuclear (CBRN) threats, and food contaminations. MedISys retrieves approximately 100,000 news articles from more than 2500 selected media sources per day. In the context of this work, *100* articles are selected from MedISys alerts from communicable disease "Antimicrobial resistance". The data is curated from the RSS feeds of AMR alerts URL.[5] The data is extracted from RSS feeds using automated approach with the help of python libraries *feedparser* and *newspaper3k* [6].
- *PADI-web:* It is operating since 2016. It is an automated EBS tool that monitors the Google News aggregator in sixteen languages. The tool gathers outbreak-related online news and extracts epidemiological information on the disease, location of the event, the host, clinical signs and number of cases. PADI-web monitors animal health threats for known emerging infectious diseases and new infectious diseases via syndromic surveillance. This tool is currently in use by the French epidemic intelligence team in animal health [3].

3.3. Experiments

To illustrate the usefulness of our two annotated corpora D_1 and D_2, we present a text classification task, publicly available online.[3] For each corpus, we perform this task through generic state-of-the-art NLP techniques in five steps.

First, we transform the raw text documents by removing irrelevant elements (pictures, ads, hyperlinks, etc.) and by applying classical preprocessing techniques (removing special characters, lowercasing, lemmatization, tokenization) [7]. Second, we convert the preprocessed texts into numerical feature vectors with two well-known feature extraction methods: 1) Tf-Idf (Term Frequency - Inverse Document Frequency weighting) [8] and 2) Doc2Vec [9]. These two methods are tested separately and in combination.

Third, we rely on Support Vector Machines (SVM) [10] for prediction, which is a widely used machine learning method for text classification. During the prediction process, we exhaustively consider all parameter combinations of SVM to keep the best model. Fourth, we train SVM based on the extracted features using 5-cross-validation. In the cross-validation we split the data into five same-sized parts containing the same ratio of output classes. In each same-sized part, we use a 80%-train / 20%-test split, which means, for each run of the cross-validation, the train set is composed of 8 of those parts while the test set is composed of the remaining two.

[5] https://medisys.newsbrief.eu/rss/?type=category&id=Antimicrobialresist&language=en&duplicates=false

Table 4
Classification results.

	Corpus D_1 (Thematic classification)			Corpus D_2 (Host classification)		
Score (macro-averaged)	Tf-Idf	Doc2vec	Tf-Idf & Doc2vec	Tf-Idf	Doc2vec	Tf-Idf & Doc2vec
Precision	0.753	0.767	**0.769**	0.395	0.445	**0.452**
Recall	0.755	0.747	**0.761**	0.372	**0.509**	0.484
F-measure	0.754	0.753	**0.765**	0.379	**0.465**	**0.465**

It is worth noticing that both corpora D_1 and D_2 have an unequal distribution of classes (i.e. class imbalance), as described in Section 2. This can severely affect the prediction algorithms, depending on the skewness degree of the class distributions. With a greater imbalanced ratio (e.g. three classes with the ratio of 900:300:10 texts), such a prediction algorithm favors the class with the larger number of texts. To deal with this problem, we use a multi-class resampling method called *SMOTETomek* [11] during the training, which combines over and under-sampling techniques employed in the literature to alter the class distributions toward a more balanced distribution. This resampling method is only applied onto the training set in the 5-cross-validation.

Finally, we estimate the performance of the trained SVM model using three evaluation metrics: Precision, recall, *F*-measure. Since each output class is of equal importance, we calculate these metrics for each output class and find their unweighted mean, so-called the *macro-average* calculation. The evaluation is presented in Table 4. We can summarize it in three parts.

First, we see that Tf-Idf and Doc2vec features give very similar results for the corpus D_1 (3 classes), whereas Doc2vec features contribute more to the prediction for the corpus D_2 (7 classes). Since all documents constituting seven output classes in D_2 (Humans, Animals, Environment, Food, Human-animal, Human-food, All) are AMR-related, it is possible that the same relevant words are frequently used in any output class of D_2. This seems to reduce the effect of Tf-Idf, hence favors Doc2vec over Tf-Idf in D_2. Second, Tf-Idf and Doc2vec features seem to be complementary in both corpora D_1 and D_2, since SVM better performs with their combination. Finally, the prediction results for Corpus D_2 are approaching a baseline value of 0.5. However, they are worse than those for Corpus D_1. This is expected, since there are more output classes, and this makes the task more challenging.

Overall, for both corpora we obtain a baseline prediction performance based on Tf-Idf and Doc2vec features. Nonetheless, these results reveal the necessity of considering more features, particularly AMR-related ones (e.g. AMR symptoms, antibiotic names), in order to obtain better prediction results.

Ethics Statement

The content of media data is anonymized by removing the user names and the names of person in transcripts (using SpaCy for name recognition).

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.

Data Availability

MOOD - News AMR dataset - Hackathon 2022 (Original data).

CRediT Author Statement

Nejat Arınık: Data curation, Methodology, Software, Writing – review & editing; **Wim Van Bortel:** Data curation, Writing – review & editing; **Bahdja Boudoua:** Data curation, Methodology, Writing – review & editing; **Luca Busani:** Data curation, Writing – review & editing; **Rémy Decoupes:** Methodology, Writing – review & editing; **Roberto Interdonato:** Data curation, Methodology, Software, Writing – review & editing; **Rodrique Kafando:** Data curation, Methodology, Software, Writing – review & editing; **Esther van Kleef:** Data curation, Writing – review & editing; **Mathieu Roche:** Data curation, Methodology, Writing – review & editing; **Mehtab Alam Syed:** Methodology, Software, Data curation, Writing – review & editing; **Maguelonne Teisseire:** Data curation, Methodology, Writing – review & editing.

Acknowledgments

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD052. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

References

[1] N. Arinik, W. Van Bortel, B. Boudoua, L. Busani, R. Découpes, R. Interdonato, E. Van Kleef, R. Kafando, M. Roche, M. Syed, M. Teisseire, MOOD - news AMR dataset - Hackathon 2022, Recherche Data Gouv (2022), doi:10.57745/MPNSPH.
[2] C. Paquet, D. Coulombier, R. Kaiser, M. Ciotti, Epidemic intelligence: a new framework for strengthening disease surveillance in Europe, Euro Surveillance 11 (12) (2006) 665, doi:10.2807/esm.11.12.00665-en.
[3] S. Valentin, E. Arsevska, S. Falala, J. De Goër, R. Lancelot, A. Mercier, J. Rabatel, M. Roche, PADI-web: a multilingual event-based surveillance system for monitoring animal infectious diseases, Comput. Electron. Agric. 169 (2020) 105163.
[4] C.C. Freifeld, K.D. Mandl, B.Y. Reis, J.S. Brownstein, HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports, J. Am. Med. Inform. Assoc. 15 (2) (2008) 150–157.
[5] M. Carrion, L.C. Madoff, ProMED-mail: 22 years of digital surveillance of emerging infectious diseases, Int. Health 9 (3) (2017) 177–183.
[6] A. Rortais, J. Belyaeva, M. Gemo, E. Van der Goot, J.P. Linge, Medisys: an early-warning system for the detection of (re-) emerging food-and feed-borne hazards, Food Res. Int. 43 (5) (2010) 1553–1556.
[7] D. Jurafsky, J.H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition, second ed., Prentice-Hall, Inc., 2000.
[8] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. - Manag. 24 (5) (1988) 513–523, doi:10.1016/0306-4573(88)90021-0.
[9] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on International Conference on Machine Learning, in: ICML'14, vol. 32, JMLR, 2014, pp. 1188–1196.
[10] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 137–142.
[11] G.E.A.P.A. Batista, A.L.C. Bazzan, M.C. Monard, Balancing training data for automated annotation of keywords: a case study, in: S. Lifschitz, N.F. Almeida Jr., G.J. Pappas Jr., R. Linden (Eds.), WOB, 2003, pp. 10–18.