Multi-label Infectious Disease News Event Corpus

Jakub Piskorski[1], **Nicolas Stefanovitch**[2], Brian Doherty[2], Jens P. Linge[2], Sopho Kharazi[3], Jas Mantero[4], Guillaume Jacquet[2], Alessio Spadaro[2], Giulia Teodori[2]

[2]**European Commission Joint Research Centre**, Ispra, Italy
[1]Polish Academy of Sciences, Warsaw, Poland

[3]Piksel SRL, Milan, Italy

[4]Ending Pandemics

European
Commission

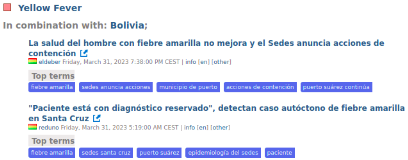
Text2Story 2023, 2 April 2023

Content

- Motivation
- Taxonomy
- Corpus Creation
- Statistics
- Benchmark Models
- Conclusions & Future Work

Background

MEDISYS (health-related news aggregator)



Yellow Fever

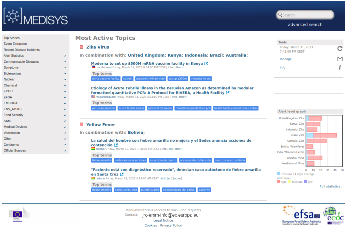
In combination with: **Bolivia;**

La salud del hombre con fiebre amarilla no mejora y el Sedes anuncia acciones de contención
eldeber Friday, March 31, 2023 7:38:00 PM CEST | info [en] [other]

Top terms
fiebre amarilla | sedes anuncia acciones | municipio de puerto | acciones de contención | puerto suárez continúa

"Paciente está con diagnóstico reservado", detectan caso autóctono de fiebre amarilla en Santa Cruz
reduno Friday, March 31, 2023 5:19:00 AM CEST | info [en] [other]

Top terms
fiebre amarilla | sedes santa cruz | puerto suárez | epidemiología del sedes | paciente



Friday, March 31, 2023
7:42:00 PM CEST

Most Active Topics

Zika Virus

In combination with: **United Kingdom; Kenya; Indonesia; Brazil; Australia;**

Moderna to set up $500M mRNA vaccine facility in Kenya
manilatimes Friday, March 31, 2023 6:04:00 PM CEST | info [other]

Top terms
mrna vaccine facility | nairobi | president william ruto | set up $500m | moderna to set

Etiology of Acute Febrile Illness in the Peruvian Amazon as determined by modular formatted quantitative PCR: A Protocol for RIVERA, a Health Facility
researchsquare Friday, March 31, 2023 5:28:00 PM CEST | info [other]

Top terms
peruvian amazon | acute febrile illness | protocol for rivera | formatted quantitative pcr | health facility-based case-control

Yellow Fever

In combination with: **Bolivia;**

La salud del hombre con fiebre amarilla no mejora y el Sedes anuncia acciones de contención
eldeber Friday, March 31, 2023 7:38:00 PM CEST | info [en] [other]

Top terms
fiebre amarilla | sedes anuncia acciones | municipio de puerto | acciones de contención | puerto suárez continúa

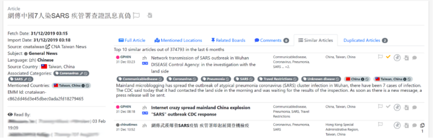
"Paciente está con diagnóstico reservado", detectan caso autóctono de fiebre amarilla en Santa Cruz
reduno Friday, March 31, 2023 5:19:00 AM CEST | info [en] [other]

Top terms
fiebre amarilla | sedes santa cruz | puerto suárez | epidemiología del sedes | paciente

Manuals/Tutorials (access to wiki upon request)
Contact: jrc-emm-info@ec.europa.eu
Legal Notice
Cookies - Privacy Policy

EIOS (Epidemics Intelligence from Open Sources)

- First signal picked up at 03:15 UTC on the 31st of December 2019



Article
網傳中國7人染SARS 疾管署查證訊息真偽

Fetch Date: **31/12/2019 03:15**
Import Date: **31/12/2019 03:18**
Source: cnataiwan CNA Taiwan News
Subject: **General News**
Language: (zh) **Chinese**
Source Country : Taiwan, China
Associated Categories: Coronavirus SARS
Mentioned Countries: Taiwan, China
EMM Id: cnataiwan-c862dd46d3e45dbec0ada2fd18279465

Read By:
[redacted] 03 Feb 19:09
[redacted]

Top 10 similar articles out of 374793 in the last 6 months

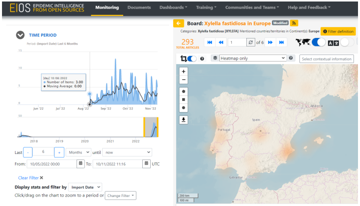
GPHIN 31 Dec 03:23 zh Network transmission of SARS outbreak in Wuhan DISEASE Control Agency: in the investigation with the land side — Communicabledisease, Coronavirus, Pneumonia, SARS ... +2. — China, Taiwan, China

Communicabledisease | Coronavirus | Pneumonia | SARS | Travel Restrictions | Unknown disease | China | Taiwan, China

Mainland microblogging has spread the outbreak of atypical pneumonia coronavirus (SARS) cluster infection in Wuhan, there have been 7 cases of infection. The CDC said today that it had contacted the land side in the morning and was waiting for the results of the inspection. As soon as there is a new message, a press release will be sent.

GPHIN 31 Dec 08:18 zh **Internet crazy spread mainland China explosion "SARS" outbreak CDC response** — Communicabledisease, Pneumonia, SARS, Travel Restrictions — China, Taiwan, China

chinatimes 31 Dec 13:52 zh 網傳武漢爆發SARS疫情 疾管署即起展開登機檢疫 — Coronavirus, Pneumonia, SARS — Hong Kong Special Administrative Region, Taiwan, China — 1



Period: (Import Date) Last 6 Months

[day] 10/08/2022
Number of Items: 3.00
Moving Average: 0.00

Last - 6 + Months until now
From: 10/05/2022 00:00 To: 10/11/2022 11:16 UTC
Clear Filter
Display stats and filter by Import Date
Click/drag on the chart to zoom to a period or Change Filter

Board: Xylella fastidiosa in Europe Modified
Categories: **Xylella fastidiosa [XYLEFA]** Mentioned countries/territories in Continent(s): **Europe**
293 TOTAL ARTICLES
1 of 6

Motivation Example

DUBAI, United Arab Emirates Dubai's Expo 2020 world's fair will be postponed to Oct. 1, 2021, over the new coronavirus pandemic, a Paris-based body behind the events said Monday. The announcement by the Bureau International des Expositions came just hours after police in Kuwait dispersed what they described as a riot by stranded Egyptians unable to return home amid the coronavirus pandemic . The riot was the first reported sign of unrest from the region's vast population of foreign workers who have lost their jobs over the crisis

Motivation Example

DUBAI, United Arab Emirates Dubai's Expo 2020 world's fair will be postponed to Oct. 1, 2021, over the new coronavirus pandemic, a Paris-based body behind the events said Monday. The announcement by the Bureau International des Expositions came just hours after police in Kuwait dispersed what they described as a riot by stranded Egyptians unable to return home amid the coronavirus pandemic. The riot was the first reported sign of unrest from the region's vast population of foreign workers who have lost their jobs over the crisis

Impact: Events **Impact: Displacement of people**
Violation: restrictions and unrest

Motivation

- capture pandemic-related events for large array of use cases
- automatic processing of large amount of news
- enhancement of Medisys/EIOS platforms
- lack of resources for this application

⇒

- created a new 2-tiers taxonomy:
  - 8 coarse grained labels
  - 25 fine grained labels
- created a new multi-label dataset annotated for it
  - 7 annotators
  - 4.4k items
- trained baselines classifiers

Taxonomy: Coarse-Grained

- **REPORTING:** reporting cases and general situation overview
- **IMPACT:** events impacted by the outbreak of the disease
- **MEASURE:** new legislation, restrictions and recommendations of preventive nature, roll-out of vaccines, medicines, equipment
- **VIOLATION:** illegal activities, fraud, fake product discovery, unrests related to measures, spread of misinformation.
- **RESEARCH & DEVELOPMENT:** phenomena observed, science progress (vaccines, medicine, equipment), research support
- **COMMUNICATION:** high-level meetings and launch of new information sharing/collection instruments
- **SUPPORT:** provision of financial support, goods, services, and mentions of the need or lack of such support
- **MISCELLANEOUS:** any other events related or unrelated to infectious diseases and non-events (e.g. descriptions of processes)

Taxonomy: Fine-grained

Reporting
- cases
- situation

Impact
- displacement of people
- health system
- economy
- events
- other

Measure
- authority regulation
- facilities
- travel
- vaccine/medicine roll-out
- other

Violation
- restrictions and unrest
- fake product or fraud
- misinformation

R&D
- medicine progress
- phenomena
- funding

Communication
- meeting
- launch instrument

Miscellaneous
- other (related)
- unrelated
- non events

Support
- financial
- goods

Data examples

NEW YORK (AP) U.S. health officials are telling people to avoid eating because of an E. coli outbreak that has sickened 32 people in 11 states.

Reporting cases Measure authority regulation

We analyzed demographic, clinical, laboratory and exposure information of symptomatic laboratory-confirmed brucellosis cases notified 2006-18. Using official population data, we calculated incidences and risk ratios (RR).

Research phenomena

Prion proteins are small bits of proteins that are responsible for causing sporadic Creutzfeldt-Jakob disease (sCJD) that is rare and affects one in a million individuals worldwide annually. CJD is related to mad cow disease.

Miscellaneous: non event

Annotation Guidelines

Event definition: situation (past/present/future) which has a place or precise span in space and time or ongoing event that has not yet ended; or statements made about events[1]

The **Miscellaneous** category aims at capturing a) events of interest not captured by the other types in the taxonomy, b) unrelated events, (c) non-events

The **Taxonomy** was refined after a first pass with 5 annotators over 400 items. The rest of the data was annotated by 7 annotators (with at least 2 annotators per snippet) with a following curation step

[1] Our notion of events is based on the TimeML standard specifications

Corpus Creation

Procedure

- sampling news from https://medisys.newsbrief.eu/ 2016-2021
  - prefetching of article using list of 800 infectious disesases
  - prefetching data using keyword combination
- extraction of text snippets
- simultaneous manual review and annotation

Bias

- 70% of the items are about Covid

Corpus Creation

sourcing query for `Impact Economy`:

```
10n(OR(economy, economic, economies, financial, unemployment,
bankrupt, bankruptcy, unemployed), OR(pandemic, lockdown,
disease, diseases, infection, infections, infectious, virus,
viruses))
```

retrieved article (first hit):

> *PHILADELPHIA, Pa. (WHTM) Three years after the enactment of the Coronavirus Aid, Relief and Economic Security (CARES) Act, The Internal Revenue Service Criminal Investigation (IRS-CI) has investigated 975 tax and money laundering cases related to COVID fraud with alleged fraud in these cases totaling $3.2B*

actually an instance of `Violation: Fake product or Fraud`

Corpus Statistics

- 4441 text snippets
- average length: 412 chars
- 77.1% are monolabel
- "relatively" balanced
- Average Cohen's Kappa: 0.59 (strict) 0.63 (loose)

Event Type	Number	Fraction	Co-occurrence
Reporting	1089	24.5%	3.31%
Impact	853	19.2%	4.34%
Measure	987	22.2%	3.24%
Violation	378	8.51%	3.87%
R&D	532	12.0%	1.5%
Communication	358	8.06%	6.15%
Support	293	6.6%	5.80%
Miscellaneous	779	17.5%	0.0%

Corpus Statistics

Event Type	Number	Fraction	Co-occurrence
Reporting	1089	24.5%	3.31%
Reporting cases	614	13.83%	10.75%
Reporting situation	641	14.43%	11.23%
Impact	853	19.2%	4.34%
Impact: displacement of people	107	2.41%	20.56%
Impact: health system	117	2.63%	13.68%
Impact: economy	346	7.79%	7.23%
Impact: events	157	3.54%	6.37%
Impact: other	178	4.01%	8.99%
Measure	987	22.2%	3.24%
Measure: authority regulation	322	7.25%	16.77%
Measure: facilities	116	2.61%	11.21%
Measure: travel	137	3.08%	16.79%
Measure: vaccine/medicine roll-out	387	8.71%	6.46%
Measure: other	100	2.25%	7.00%

Statistics

Event Type	Number	Fraction	Co-occurrence
Violation	378	8.51%	3.87%
Violation: restrictions and unrest	127	2.86%	11.81%
Violation: fake product or fraud	121	2.72%	7.44%
Violation: misinformation	149	3.36%	5.37%
R&D	532	12.0%	1.5%
R&D: medicine progress	187	4.21%	2.67%
R&D: phenomena	272	6.12%	2.21%
R&D: funding	97	2.18%	2.06%
Communication	358	8.06%	6.15%
Communication: meeting	158	5.81%	9.30%
Communication: launch instrument	101	2.27%	4.95%
Support	293	6.6%	5.80%
Support: financial	189	4.26%	8.47%
Support: goods	113	2.54%	7.08%
Miscellaneous	779	17.5%	0.0%
Miscellaneous: other	158	3.56%	0.0%
Miscellaneous: unrelated	508	11.44%	0.0%
Miscellaneous: non events	115	2.59%	0.0%

Analysis of co-occurrences

Event Type 1	Event Type 2	Count	Fraction 1	Fraction 2
Reporting cases	Reporting situation	166.0	27.0	25.9
Measure: Authority Regulation	Reporting situation	67.0	20.8	10.5
Measure: Authority Regulation	Reporting cases	48.0	14.9	7.8
Measure: Vaccine/Medicine Roll-out	Reporting situation	32.0	8.3	5.0
Communication: Meeting	Reporting situation	31.0	12.0	4.8
Impact: Economy	Support: Financial	28.0	8.1	14.8
Impact: Health system	Reporting situation	24.0	20.5	3.7
Measure: Authority Regulation	Measure: Travel	21.0	6.5	15.3
Impact: Economy	Impact: Other	20.0	5.8	11.2
Measure: Authority Regulation	Measure: Facilities	20.0	6.2	17.2

Table: Top co-occurring pairs of fine-grained event labels: (a) Count stands for absolute number of co-occurrences, (b) Fraction 1 stands for the count normalised by the total number of co-occurrences of event Type 1; (c) Fraction 2 stands for the count normalised by the total number of co-occurrences of event Type 2.

Benchmark models

- L2-regularized linear SVM using the One-vs-the-Rest strategy, with log TFIDF-weighted 3-5 character n-grams as features
- RoBERTA base, a transformer-based model, batch size of 32, learning rate of 2^{-5} and 100 warming steps with 5 training epochs
- 5-fold cross-validation

Benchmark models: Overall

model	micro	macro	weighted	samples
SVM	0.60	0.56	0.59	0.55
RoBERTa	0.69	0.65	0.68	0.68

Table: **Fine-grained Event Types**: F_1 scores for benchmark models.

model	micro	macro	weighted	samples
SVM	0.69	0.68	0.69	0.67
RoBERTa	0.76	0.76	0.76	0.76

Table: **Coarse-grained Event Types**: F_1 scores for benchmark models.

Benchmark models: Coarse-grained

Event Type	SVM	RoBERTa
Reporting	0.80	0.85
Impact	0.65	0.73
Measure	0.66	0.68
Violation	0.72	0.79
R&D	0.69	0.71
Communication	0.63	0.78
Support	0.58	0.70
Miscellaneous	0.68	0.73

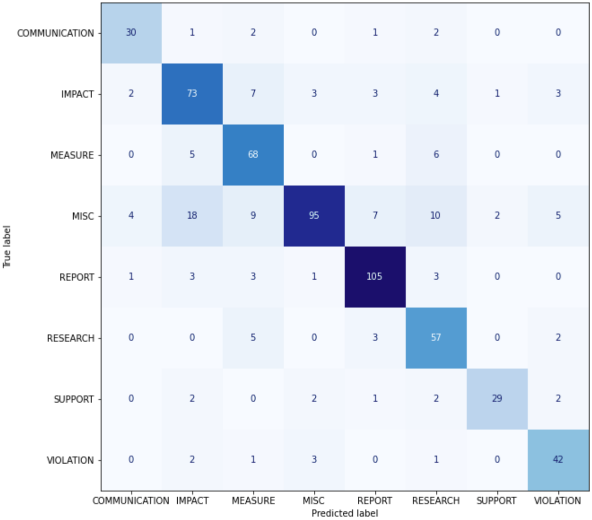
Table: F_1 scores for coarse-grained types

Benchmark models: Fine-grained

Event Type	SVM	RoBERTa
Reporting cases	0.74	0.85
Reporting situation	0.64	0.75
Impact: displacement of people	0.75	0.81
Impact: health system	0.40	0.55
Impact: economy	0.63	0.71
Impact: events	0.64	0.83
Impact: other	0.24	0.42
Measure: authority regulation	0.43	0.45
Measure: facilities	0.55	0.70
Measure: travel	0.60	0.79
Measure: vaccine/medicine roll-out	0.67	0.64
Measure: other	0.21	0.24
Violation: restrictions and unrest	0.54	0.71
Violation: fake product or fraud	0.75	0.80
Violation: misinformation	0.71	0.64
R&D: medicine progress	0.55	0.58
R&D: phenomena	0.59	0.72
R&D: funding	0.64	0.81
Communication: meeting	0.68	0.76
Communication: launch instrument	0.59	0.70
Support: financial	0.55	0.76
Support: goods	0.49	0.63
Miscellaneous: other	0.11	0.34
Miscellaneous: unrelated	0.70	0.50
Miscellaneous: non events	0.48	0.78

Table: **Fine-grained**: F_1 scores for benchmark models per class.

Confusion matrix



True label \ Predicted label	COMMUNICATION	IMPACT	MEASURE	MISC	REPORT	RESEARCH	SUPPORT	VIOLATION
COMMUNICATION	30	1	2	0	1	2	0	0
IMPACT	2	73	7	3	3	4	1	3
MEASURE	0	5	68	0	1	6	0	0
MISC	4	18	9	95	7	10	2	5
REPORT	1	3	3	1	105	3	0	0
RESEARCH	0	0	5	0	3	57	0	2
SUPPORT	0	2	0	2	1	2	29	2
VIOLATION	0	2	1	3	0	1	0	42

Figure: Confusion matrix for coarse-grained event types

Future Work

- data augmentation
- add more samples with diseases different from Covid-19 and test with named entities substitution
- actual event detection versus news item flagging
- event metadata extraction
- evaluate on historical data