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] and Giulia Teodori[2]

[1] *Polish Academy of Sciences, Warsaw, Poland*

[2] *European Commission Joint Research Centre, Ispra, Italy*

[3] *Piksel SRL*

[4] *Ending Pandemics*

Abstract

This paper describes a new corpus consisting of circa 4.5K news snippets (multi-)labelled with fine-grained infectious disease-related event types. The paper presents the underlying event taxonomy consisting of 25 fine-grained event types grouped into 8 main categories, the process of creating the corpus, related statistics and reports on the performance of SVM- and RoBERTa transformer-based baseline models for multi-label event classification. The former model obtains macro F_1 score of 0.56 and 0.68 for fine- and coarse-grained classification, respectively, whereas the corresponding macro F_1 scores for the latter model are 0.65 and 0.76, respectively.

Keywords

multi-label event classification, infectious diseases, machine learning, linguistic resources

1. Introduction

Surveillance and quick response to situations emerging from outbreaks of infectious diseases, e.g. Covid-19, relies on comprehension of all related events, which, among others, are reported in large amounts in news articles being published every day. Automated solutions that facilitate extraction and classification of such events is crucial in order to leverage such sources of information, especially for early-warning systems.

In this paper, we describe a new corpus consisting of news snippets multi-labelled with fine-grained infectious disease-related event types reported therein. The main drive behind this endeavour was to create material for training and building respective ML-based models for event detection/classification in epidemics-related online news gathered by a large-scale news aggregation and analysis engine, and to share such a resource with the scientific community, since, to the best of our knowledge, no similar publicly accessible event-centred corpus exists for this specific domain. Event detection and classification constitutes a key enabling technique

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'23 Workshop, Dublin (Republic of Ireland), 2-April-2023

[*] Corresponding author.

jpiskorski@gmail.com (J. Piskorski); nicolas.stefanovitch@ec.europa.eu (N. Stefanovitch); brian.doherty@ec.europa.eu (B. Doherty); jens.linge@ec.europa.eu (J. P. Linge); sopho.kharazi@ext.ec.europa.eu (S. Kharazi); jasmantero@gmail.com (J. Mantero); guillaume.jacquet@ec.europa.eu (G. Jacquet); alessio.spadaro@ec.europa.eu (A. Spadaro); giulia.teodori@ec.europa.eu (G. Teodori)

CEUR Workshop Proceedings (CEUR-WS.org)

to build higher-level applications, e.g. event extraction, news summarization, and narrative understanding.

Since the beginning of the Covid-19 pandemic, a vast amount of work on studying Covid-19-related online media and automated analysis thereof has been reported, which mainly focused on exploiting topic detection [1], fake news/misinformation narrative analysis [2], entity and demographic-based analysis [3], and sentiment detection [4, 1], whereas relatively little work on automated event detection and extraction has been published in this context.

A corpus of 10K tweets containing public reports of Covid-19 events centered around reporting cases, deaths, prevention measures, and cures was presented in [5]. A large hand-coded dataset of over 13K policy measures introduced worldwide related to Covid-19, gathered among others from online news, is presented in [6]. Other online media resources related to Covid-19 have been listed on the CLARIN Covid-19 response web page. [1]

[7] presented a BERT-based system that extracts and classifies Covid-19 related events and relations between them, using a semi-automatically created event taxonomy consisting of 76 event types. The event taxonomy in the aforementioned work exhibits, to some degree, similarity with the one presented in this paper; however, no event-labelled corpora have been released by the authors. Furthermore, our event taxonomy was not created automatically, but emerged from a business requirement analysis by public health experts and has been designed upfront to cover any infectious diseases, going beyond the Covid-19 pandemic. Finally, the news snippets in our corpus are multi-label annotated.

Related to our work, some short news text classification datasets have been published, e.g. [8] introduce a corpus of ca. 200k news headlines labelled with 40 general news categories, and work related to exploring ML-based models (accompanied with datasets) for the detection and classification of natural disasters [9], financial [10] and socio-political events [11] reported in the news, covering domains that, however, have little in common with pandemics and infectious diseases.

2. News Snippet Event Corpus

This section describes the event taxonomy, creation of the corpus of news snippets with labels corresponding to events referred to in these snippets, and provides some corpus statistics. We consider an event[2], a situation (or a group thereof) that has either: occurred, is currently taking place, or is planned or considered to happen in the future, in some place and at a certain point in time (`punctual events`) or spanned/spans a time period with a start date and potential end date. Furthermore, references to a `state` (of play) of a situation (an ongoing event) that has not yet ended, statements and opinions made about it are also considered events.

2.1. Infectious Disease-related Event Taxonomy

The events are grouped into 8 main categories that revolve around: `reporting` on the disease outbreak development, `impact`, `measures`, `violations`, `research`, `support`, `communication`, which

[1] https://www.clarin.eu/content/clarin-responses-covid-19

[2] Our notion of events is based on the *TimeML* standard specifications [12]

REPORTING: reporting single/multiple infection cases and deaths that occurred within a short period of time and provision of general situation overview (in terms of people affected) spanning a longer time period.

IMPACT: all events that are impacted by the outbreak of the infectious disease/pandemic, e.g. cancellation of events

MEASURE: introduction and changes to legislation, restrictions and recommendations of preventive nature necessary to combat the disease, i.e. the number of infected/affected people and spread of the disease, roll-out of related vaccines, medicines and equipment.

VIOLATION: any illegal activity, fraud, fake product discovery, unrest related to the introduced measures, and spread of misinformation.

RESEARCH & DEVELOPMENT: reporting on the phenomena observed during the spread of the disease, progress on vaccines, medicine and relevant equipment development, and support to research and development related to diagnose or treat the disease.

COMMUNICATION: high-level meetings to discuss the situation, impacts and/or introduce measures, and launch of new information sharing/collection instruments concerning the disease and related phenomena.

SUPPORT: provision of financial and other type of support to the affected entities, community, economy, etc., and mentions of the need or lack of such support.

MISCELLANEOUS: any other events related (not covered above) or unrelated to infectious diseases, and non-events, i.e. texts not referring to any actual event nor a state of an event, e.g. descriptions of processes.

Figure 1: Coarse-grained Infectious Disease-related Event Categories.

are all further subdivided into 25 fine-grained event types that refer to specific aspects of the main categories, e.g. `Reporting` is subdivided into `Reporting cases` and `Reporting situation`. The brief description of the main event categories is provided in Figure 1, whereas the one for the fine-grained types is provided in Annex A in Figure 6. The event definitions are to a large extent ‘inclusive’, e.g., the `Support: goods` category covers not only the factual provision of goods to the affected people, but also plans and intents to do so, and expression of the needs of those in need to receive such support.

The `Miscellaneous` category is envisaged to capture everything that does not fit anywhere else, and is subdivided into: (a) `other` events that are related to the domain, but do not fall under any other type, (b) events that are `unrelated`, and (c) `non events`, e.g., descriptions of certain generic processes and phenomena that are neither tailored in time nor refer to any specific event instances, although relevant for the domain though. It is important to emphasize at this stage that, in a practical set-up, a different merging and subdivision of `Miscellaneous` might be more beneficial for ML modelling purposes; however, the main drive behind this subdivision was to explore how well the 3 different fine-grained classes can be distinguished. Furthermore, the `Miscellaneous: Other` category was deemed as relevant from end-user perspective, i.e. constituting a source of providing ‘interesting’ information.

2.2. Data Sampling

The input data for annotation was randomly sampled from news articles gathered by *MEDISYS*[3], a large-scale health-related news aggregation engine [13] from a period that spans 2016-2021. Apart from conventional media sources, *MEDISYS* also monitors news on hundreds of official public health websites such as ministry of health and public health agency websites.

[3] https://medisys.newsbrief.eu/

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

Figure 2: An example of a Solr query used to target articles with specific categories, in this case the category *Impact: Economy*. This query specifies that the word *economy* and its synonyms should be at max. 10 tokens away from the word *pandemic* and related terms.

News articles were sourced using keywords, and snippets were further extracted from them by selecting up to max. first 4 sentences comprised within the first 500 characters of the article[4]. The rationale behind considering the initial part of news articles was the assumption of *inverted-pyramid* style [14] of writing news articles, i.e. the most relevant events are placed in the beginning and the least important ones are left toward the end. First, news articles were randomly sourced using a list of circa 800 infectious disease names[5], e.g. *Covid-19*, *ebola*, *zika*, *malaria*, etc., and relevant name variants and acronyms. Given that a large fraction of text snippets acquired in this way fell under `Miscellaneous` in order to populate proportionally the other classes in the taxonomy, an additional document sampling for each category was carried out through the use of a more 'focused' combination of keywords (including synonyms) which were required to be found within a specific text window anywhere in the body of a news article. An example of such a keyword query is provided in Figure 2. This allowed to improve the precision (i.e. ca. 50% of the fetched articles were reporting on events in the taxonomy that fall into non-`Miscellaneous` categories). The potential bias that might have been introduced by the use of specific keywords is mitigated by extracting text snippets only from the beginning of articles, which do not necessarily contain any of the keywords of the query and instead use a different wording to report on an event.

In addition, circa 10% of text snippets were further manually selected from the news articles to ensure the corpus is even more balanced.

2.3. Data Annotation

From the sampled text snippets described above, circa 4.5K were randomly selected for annotation. 7 annotators were involved in this process, all of which had prior experience of annotating news texts. 2 of the annotators have a background in NLP and computational linguistics, whereas 5 others were news analysts. Initially, circa 400 randomly selected snippets were annotated by 5 annotators, who subsequently jointly resolved the conflicts. The main motivation behind this part of the annotation process was to revise the event codebook comprising the event definitions, which turned to be overlapping or incomplete to some degree. Next, the remainder of the snippets was annotated, each by at least 2 annotators. Given the fact that annotations are sets of labels, we have computed strict and loose Cohen's κ, where for the former an agreement is considered only for identical label sets, whereas in the latter case, a non-empty overlap of the label sets is considered an agreement. The average strict and loose κ for a pair of annotators are

[4] Some snippets are longer than 500 characters in order to respect sentence boundaries.

[5] This list contains diseases considered as the most common public health threats created for the *MEDISYS* platform for the purpose of retrieving relevant news articles.

> *Amid rising vaccination rates across the European Union, the 27 EU leaders on Tuesday committed to collectively donate at least 100 million doses of Covid-19 vaccine to countries in need by the end of 2021. The bloc, which described itself in a joint statement signed off at summit.*

Figure 3: An example of text snippet annotated with two labels, namely, *Support: goods* and *Communication: meeting*, with the corresponding phrases referring to the respective events underlined.

0.59 and 0.63 resp. The conflict resolution in the annotations was jointly carried out by 2 to 4 annotators.

An example of a news snippet annotated with two event labels is provided in Figure 3. Further examples are provided in Figure 7 in Annex A.

2.4. Data Statistics

The corpus consists of 4441 text snippets, whose average length is 412 characters. The average number of fine- and coarse-grained labels per snippet is 1.26 and 1.19, respectively. Concerning fine-grained labels, circa 77.1% of the snippets have only one such label assigned to them, whereas the percentage of the snippets with 2, 3 and 4 labels are 19.65%, 2.9% and 0.34%, respectively. The corpus is relatively well-balanced. The statistics for the coarse- and fine-grained labels are provided in Table 1. The columns labelled with 'Co-occurrence' provide the percentage of instances of the given class that are labelled with at least one other label. While this figure is maximum 6.15% (`Communication` class) for the coarse-grained types, it can reach up to 20.56% (`Impact: displacement of people` class) for the fine-grained types. The snippets labelled with `Miscellaneous` do not co-occur with other labels by definition of the former. Figure 4 presents the text snippet length histogram.

Table 2 provides a list of most frequently co-occuring pairs of fine-grained event types, while the complete event co-occurrence matrix is shown in Figure 8 in Annex A. Interestingly, the two Reporting classes are the most co-occurring ones and co-occur most frequently together; `Measure: Authority Regulation` and `Impact: Health System` tend to frequently co-occur with them as well. The other `Measure` classes tend to co-occur with `Measure: Authority Regulation`. Given the fact that Covid-19 has triggered a vast amount of news articles over the last 3 years, a large part (more than 70%) of the snippets in the corpus are related to the Covid-19 pandemic.

3. Benchmark Models

We have evaluated two benchmark models, namely: (a) L2-regularized linear SVM [6] using the One-vs-the-Rest strategy, with log TFIDF-weighted 3-5 character n-grams as features, using vector normalization and $c = 0.2$ resulting from parameter optimization, and (b) RoBERTA base [15], a transformer-based model, using a batch size of 32, learning rate of 2^{-5} and 100 warming steps with 5 training epochs.

[6] We used the LibLinear implementation provided in Scikit-learn library: https://scikit-learn.org/

Table 1
Corpus statistic: fine- and coarse-grained event labels. The co-occurrence statistics for the coarse-grained types refer to the co-occurrence with other coarse-grained types.

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%
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%

For the purpose of evaluation of these models, we use *micro*, *macro*, *weighted* and *samples*

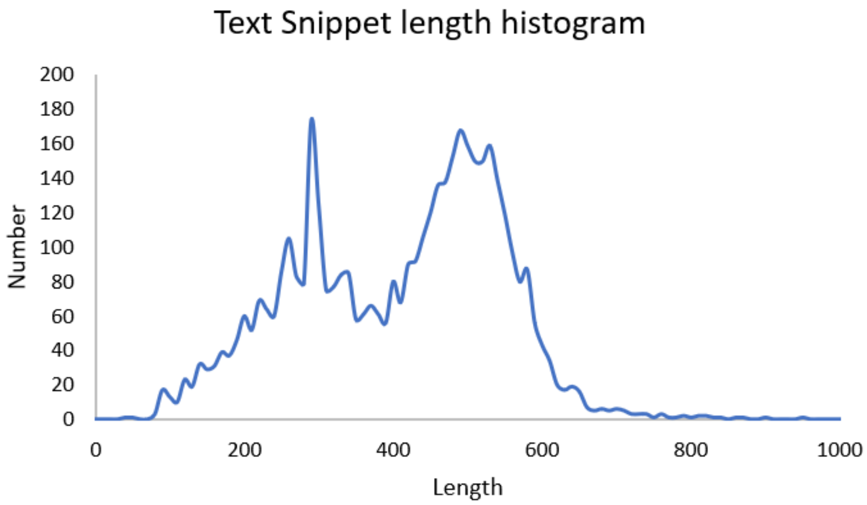

Figure 4: Text Snippet Length histogram.

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

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

F_1 scores, where the latter is computed as an average of F_1 scores computed for each pair of sets of ground-truth and system-response labels for each instance in the training data. 5-fold cross-validation was used.

The overall results for both fine- and coarse-grained classification are provided in Table 3, whereas the per-class performance of the benchmark models for the fine- and coarse-grained scenarios is provided in Table 4 and 5, respectively.

For both models, the overall performance shows little variation between the F_1 measures. The performance of RoBERTa vis-à-vis SVM is better in both the coarse- and the fine-grained classification scenario, with improvements of up to 9 and 13 points in F_1 measures, respectively.

Table 3
F_1 scores for benchmark models for fine- and coarse-grained event classification.

	Fine-grained Event Types				Coarse-grained Event Types			
Approach	Micro	Macro	Weighted	Samples	Micro	Macro	Weighted	Samples
SVM	0.60	0.56	0.59	0.55	0.69	0.68	0.69	0.67
RoBERTa	0.69	0.65	0.68	0.68	0.76	0.76	0.76	0.76

Table 4
F_1 scores for benchmark models per class for the fine-grained event types.

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 5
F_1 scores for benchmark models for coarse-grained event types.

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

As regards the models' performance on individual classes, one can observe that, for both SVM and RoBERTa, the three worst performing classes, namely, `Impact: Other`, `Measure: Other`, `Miscellaneous: Other`, have almost all an $F_1 < 0.45$, and reduce the global F_1 scores. The performance behaviour might be linked to the more open and less-focused nature of the definition of the `Other` classes.

Studying the most common confusion between labels, when both classifier and ground truth have only one label, shows (see Figure 5) that `Miscellaneous` has the most false positives and that the classes `Impact`, `Measure` and `Research` have more false positives than all the other classes.

4. Conclusions

This paper briefly described the creation of a new corpus consisting of circa 4.5K news snippets (multi-)labelled with fine-grained infectious disease-related event types and reported on the

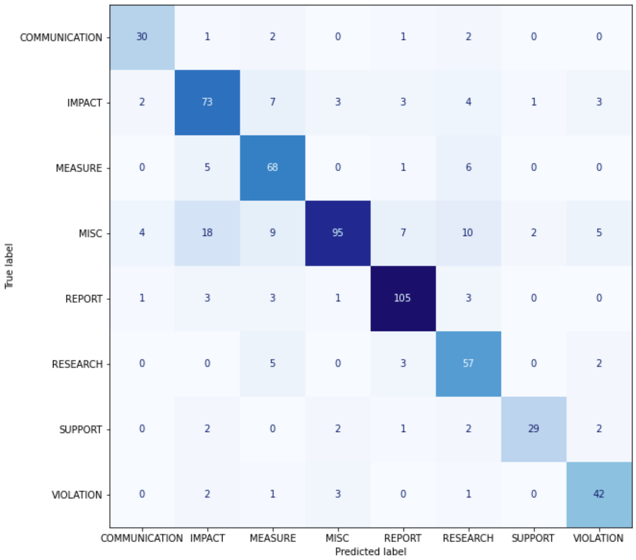

Figure 5: Confusion matrix for coarse-grained event types, considering only the snippets that have a single label both in the prediction and in the ground truth.

performance of SVM- and transformer-based baseline models trained using the corpus. We intend to enlarge the corpus in the future, in particular using snippets that cover a wider range of diseases.

The news event corpus, accompanied by the full-fledged Codebook and annotation guidelines is publicly available at https://github.com/jpiskorski/infectious-diseases-events to the scientific community for research purposes. All future extensions and updates of the corpus will be made available under the same link.

References

[1] P. Ghasiya, K. Okamura, Investigating covid-19 news across four nations: A topic modeling and sentiment analysis approach, IEEE Access 9 (2021) 36645–36656. doi:10.1109/ACCESS.2021.3062875, publisher Copyright: © 2013 IEEE.

[2] T. Marcoux, N. Agarwal, Narrative Trends of COVID-19 Misinformation, in: Proceedings of the 4th Workshop on Narrative Extraction From Texts (Text2Story 2021), held in conjunction with the 43rd European Conference on Information Retrieval (ECIR 2021), Association for Computational Linguistics, 2021, pp. 77–80.

[3] A. E. Varol, V. Kocaman, H. U. Haq, D. Talby, Understanding covid-19 news coverage using medical nlp, 2022.

[4] R. Chandra, A. Krishna, Covid-19 sentiment analysis via deep learning during the rise of novel cases, PLOS ONE 16 (2021) 1–26.

[5] S. Zong, A. Baheti, W. Xu, A. Ritter, Extracting a knowledge base of covid-19 events from social media, 2020. URL: https://arxiv.org/abs/2006.02567. doi:10.48550/ARXIV.2006.02567.

[6] C. Cheng, J. Barceló, A. Hartnett, R. Kubinec, L. Messerschmidt, COVID-19 Government Response Event Dataset (CoronaNet v.1.0), Nat Hum Behav. 4 (2020) 756–768.

[7] B. Min, B. Rozonoyer, H. Qiu, A. Zamanian, N. Xue, J. MacBride, ExcavatorCovid: Extracting events and relations from text corpora for temporal and causal analysis for COVID-19, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 63–71.

[8] R. Misra, News category dataset, 2018. URL: https://www.kaggle.com/datasets/rmisra/news-category-dataset. doi:10.13140/RG.2.2.20331.18729.

[9] T. Nugent, F. Petroni, N. Raman, L. Carstens, J. L. Leidner, A comparison of classification models for natural disaster and critical event detection from news, in: 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3750–3759.

[10] E. Lefever, V. Hoste, A classification-based approach to economic event detection in Dutch news text, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 330–335. URL: https://aclanthology.org/L16-1051.

[11] J. Haneczok, G. Jacquet, J. Piskorski, N. Stefanovitch, Fine-grained event classification in news-like text snippets - shared task 2, CASE 2021, in: Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021), Association for Computational Linguistics, Online, 2021, pp. 179–192.

[12] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, J. Pustejovsky, TimeML annotation guidelines, https://www.researchgate.net/publication/248737128_TimeML_Annotation_Guidelines_Version_121, 2006.

[13] J. Linge, R. Steinberger, T. Weber, Internet surveillance systems for early alerting of health threats, Euro Surveill 14 (2009) 1–2.

[14] J. Canavilhas, Web journalism : from the inverted pyramid to the tumbled pyramid, https://www.bocc.ubi.pt/pag/canavilhas-joao-inverted-pyramid.pdf, 2007.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

A. Supplementary corpus information

The definition (in a simplified form) of the fine-grained event types related to infectious diseases is provided in Figure 6. Some examples of news snippets annotated with these labels are provided in Figure 7. Figure 8 presents event type co-occurrence matrix.

Reporting cases: reporting on cases of infections, hospitalizations, deaths, recoveries of single persons and groups, provision of updates thereon, which covers a short time span and specific location.

Reporting situation: provision of updates on the overall situation of the outbreak, current total figures, observed trends, forecast, which spans longer period of time, and also covers cross-regional and cross-country comparisons.

Impact: Displacement of people: reporting on movement of persons/groups that were either forced, obliged or voluntarily fled or left their homes of places of habitual residence as a consequence of the spread of the infectious disease and/or introduction of measures to combat the disease. Bringing back displaced people to the place of origin falls under this category as well.

Impact: Health system: covers events related to the impact the disease has on the health-care system, e.g. deployment of additional staff, shortage of medical equipment, high bed occupancy rate, establishment of new facilities, etc.

Impact: Economy: covers events related to impact on the economy, e.g., decline/growth of certain sectors, reducing/increasing production, gains/losses, unveiling studies on the analysis and prognosis of the economic situation.

Impact: Events: reporting on cancellation, postponement, and changing of modi operandi in the context of political, sport, cultural and other mass events, etc.

Impact: Other: reporting on other impacts of the disease, e.g., societal phenomena, political situation, future predictions, etc.

Measure: Authority Regulation/Recommendation: covers events related to the introduction of measures like, e.g. law, formal regulations, restrictions, and recommendations by competent government authorities and international bodies which are specifically put in place to decrease the number of infected/affected people and thwart further spread of the disease.

Measure: Facilities: covers closures of facilities (e.g. schools, universities, museums, parks) resulting from regulations and/or situations, re-openings, changing related modi operandi, e.g. the introduction of teleworking, etc.

Measure: Travel: introduction of travel restrictions, recommendations, closure of borders, cancellation of flights, closure of airports, provision of specific transportation means to facilitate travel, etc.

Measure: Vaccine/Medicine Roll-out: covers events revolving around the roll-out of vaccines, medicines, equipment to combat the disease or mitigate the consequences, and includes also events related to sharing experience, measure hesitancy, anti-vax movements, etc.

Measure: Other: covers any other events related to measures, resulting from non-governmental organization decisions, private sector, e.g. linked to introduced laws and regulations.

Violation: Restrictions and Unrest: covers violations against introduced laws, regulations, measures and potential lockdowns, and protests against the introduced laws and measures.

Violation: Fake product or Fraud: covers events related to unveiling or warning on fake medicine or any counterfeits, falsified or substandard disease-related material/equipment being sold and/or distributed, and infectious disease-related fraud.

Violation: Misinformation: embraces events related to revealing misinformation incidents and attempts, and issuing warnings about disease-related misinformation.

Research & Development: Medicine Progress: dissemination of information and updates on the progress of research and development of medicines, vaccines and equipment to combat and/or protect against infectious diseases.

Research & Development: Phenomena: reporting on research on specific phenomena observed in the context of infectious diseases and findings which might potentially contribute to the development of medicines, vaccines, etc.

Research & Development: Funding: raising funding, launching programmes and resources for R&D of technologies and materials related to fight infectious diseases.

Communication: Meeting: covers official meetings, conferences and meetings, press conferences of authorities, states, international organizations, task forces, experts, etc., to discuss topics related to the (outbreak of) infectious diseases and related topics

Communication: Launch Instrument: reporting on new communication, information sharing and gathering instruments and methods related to infectious diseases, e.g. online platforms, databases, smartphone apps, etc.

Support: Financial: launching, proposing and elaborating financial instruments to support affected people, organizations, economy, etc., e.g. the introduction of changes in tax regulations to relieve the most vulnerable groups.

Support: Goods: providing affected people with goods, materials, and services to help and alleviate the problems resulting from the outbreak of the disease.

Miscellaneous: Other: is a placeholder to capture other events related to infectious diseases, which do not fall under any of the above categories, e.g. recruitment of new experts by a company that develops infectious disease-related vaccines.

Miscellaneous: Unrelated: covers events that are not related to infectious diseases in any way.

Miscellaneous: Non Events: covers texts that do not refer to any event that could be tailored to a particular point in time, e.g. general descriptions of processes, etc.

Figure 6: Infectious Disease-related Event Taxonomy.

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

The World Health Organization (WHO) has confirmed the first three cases of Zika virus disease in India. Health Ministry officials said Sunday that the three patients in western Gujarat state had recovered. ”There is no need to panic,” Dr. Soumya Swaminathan, a top health ministry official, told reporters. The World Health Organization said in a statement released Friday that the three cases that India reported to the WHO on May 15 were detected through routine blood surveillance in a hospital in Ahmadabad, Gujarat's capital”
EVENTS: ***Reporting: cases***

The Gates Foundation will give Rotary $255 million, with Rotary pledging to raise $100 million, and the UK and Germany contributing $150 million and $130 million respectively to the global initiative. It is the second such grant from the foundation to Rotary International — in 2007, it gave Rotary a $100 million grant for a polio eradication programme, which Rotary matched dollar for dollar. The new money will go to vaccination programmes, better disease surveillance and research on new vaccines.
EVENTS: ***Research: funding Support: financial***

Warsaw (dpa) - Czech Prime Minister Andrej Babis said on Sunday that he would like residents over the age of 60 to be able to register for a Covid-19 vaccination from March. The move would see the offer of vaccinations extended beyond the current priority groups of health care workers, nursing home residents and staff and all citizens aged over 80.
EVENTS: ***Measure: vaccine/medicine roll-out***

Little air cleansers are digital gadgets that are utilized to tidy up the air by decreasing or removing interior toxins such as germs, odours, smoke and chemicals that could be hazardous to the wellness. These small air purifier cleansers have different types such as the HEPA air cleanser, ozone air cleanser, or the ionic air cleanser.
EVENTS: ***Miscellaneous: non event***

RAI News 24 reports that, as of January 7, Italy will go back to the colour-coded system sub-dividing Regions on the basis of Covid-19 restrictions. The government will decide on the colour zone for each Region on the basis of 21 Covid-19-related criteria. However, the Regions are calling on the government to revise these criteria. Meanwhile, up until January 6, all of Italy will be in a red zone, meaning that bars and restaurants will stay closed
EVENTS: ***Measure: authority regulation, Measure: facilities***

Figure 7: Examples of text snippets annotated with event labels. The text fragments triggering the respective events are underlined in blue.

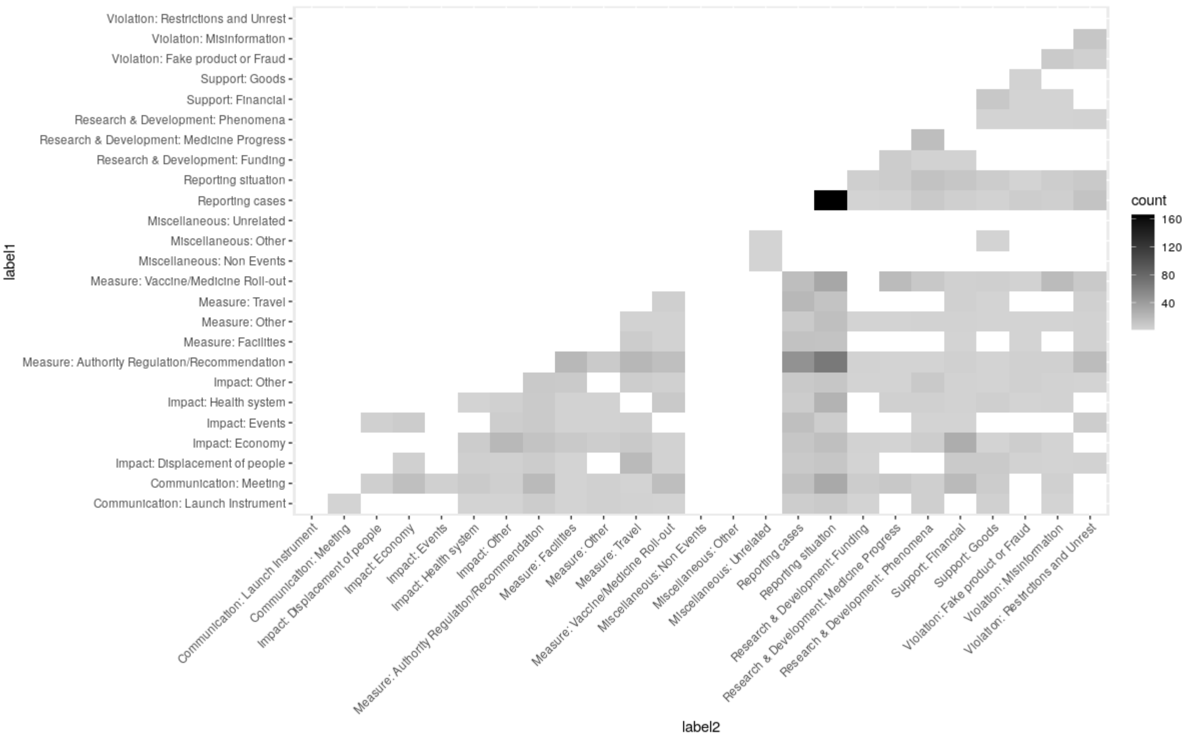

Figure 8: Event type co-occurrence matrix.