Global Responses to the COVID-19 Pandemic: A Case Study of Spatiotemporal Evidence Finding and Verification

Anonymous ACL submission

Abstract

This paper explores methods for adapting fact verification models to real-world scenarios that require spatial and temporal inference. As a case study, we search for evidence on governments' responses to the COVID-19 pandemic. We demonstrate that existing fact verification models perform poorly when the verification requires reasoning about spatiotemporal information. The suggested techniques lead to great improvements and we recommend implementing them for such uses.

"EDMONTON – The province of Alberta said on Sunday that there are another 69 cases of COVID-19, bringing the provincial total to 1,250. There were also three more COVID-19 deaths reported, bringing the total to 23. The government did not hold a press conference to update the numbers on Sunday. Press conferences will resume Monday."

Figure 1: Example of an article that reports the number of deaths and new cases of COVID-19. The spatial (Canada) and temporal (April 5th-6th, 2020) information cannot be inferred from the highlighted text.

1 Introduction

During the COVID-19 pandemic, it became imperative to follow the progress of the disease simultaneously in multiple locations and to compare the responses of different authorities in a variety of settings and conditions (Alam et al., 2021; Jin et al., 2021). However, since the pandemic was extensively covered in the media, following and gathering proof of any decisions or actions made by governments became extremely difficult.

In this paper, we aim to find evidence of occurrences of events in extremely large textual corpora for scenarios where the information being sought is timely and localized. We use the AYLIEN Coronavirus Dataset[1] as the extremely large text corpus that constitutes our search space and the information we seek is evidence of actions taken by governments in their particular jurisdiction (thus localized) at a particular time (thus timely). For example we may want to verify the following claim: *The government of Germany decided to restrict gatherings of 10 people or less from 2020-03-21 to 2020-07-06*. The events are extracted from the Oxford COVID-19 Government Response Tracker (Hale et al., 2020).

The task of evidence finding and verification (Thorne et al., 2018) focuses on verifying a statement using retrieved potential evidence from a large collection of texts. It differs from the tasks of fact checking (Vlachos and Riedel, 2014), textual entailment, and natural language inference (Dagan et al., 2010; Bowman et al., 2015; Williams et al., 2018) where the goal is to label a certain statement as true or entailed with respect to a *given* text.

In this study, we show that conventional methods for retrieving documents and identifying textual entailment used in fact verification are ineffective when applied to the challenging and highly relevant setting described above. See for example the article in Figure 1 where the country and the dates are not mentioned specifically in the text, hence cannot be inferred. We propose improvements to these processes in order to identify specific details in the text that may otherwise be overlooked.

As a first step, all location-named-entities and time expressions are automatically extracted to provide explicit spatial and temporal information to each document, as described in §4. Then, we filter out documents that are irrelevant either temporally or spatially for each claim and continue with a smaller collection of more relevant documents for retrieval. This filtering is equivalent to setting hard constraints for the retrieval algorithm.

Next, we choose the top-k ranked documents for each claim (see details in §5) to form the input for the entailment identification step. We argue

[1] https://aylien.com/blog/free-coronavirus-news-dataset

that if A entails B then this could mean that A contains evidence for claim B. However, textual entailment methods in recent years are mostly trained on datasets where both the premise and the hypothesis are single sentences (Bowman et al., 2015; Williams et al., 2018; Khot et al., 2018; Eisenschlos et al., 2020). We adapt an entailment model that works and trained on sentence level to aggregate the outputs from each sentence to output a document level label and demonstrate that it performs similarly to models trained on long texts (See §6).

The contribution of our work is in integrating temporally and spatially relevant signals to enhance the performance of retrieval and entailment methods for evidence-finding and verification of claims that are time and location-specific. Although we perform relatively simple manipulations to existing methods, the improvements are substantial for this case study. We demonstrate the effectiveness of our proposed methods by comparing the responses of governments to the pandemic (§7).

2 Related Work

As a key task aimed at detecting false information and fake news, fact verification has received much attention from the NLP community (Nie et al., 2019a; Zhou et al., 2019; Liu et al., 2019b; Zhang et al., 2020). Recent fact verification shared tasks use Wikipedia as the large corpus to extract the evidence from since the claims are general in nature (Thorne et al., 2018; Jiang et al., 2020; Aly et al., 2021; Eisenschlos et al., 2021). However, we are interested in finding evidence for occurrences of recent global events. To this end, we use the AYLIEN dataset, which contains content of world news articles, better reflecting the purposes of this research. Furthermore, a key difference between common fact verification tasks and the one we study in this paper is that our claims include both spatial and temporal information that must be addressed in order to find evidence of their validity even if the information is not explicitly mentioned in the text.

3 Datasets

This paper uses two datasets to demonstrate how to seek evidence and verify it in the context of global policy responses to COVID-19. The first dataset, from which we extract the facts to be validated is the Oxford COVID-19 government response tracker (OxCGRT, Hale et al., 2020). This tool enables rigorous and consistent tracking and comparison of policies around the world.

The OxCGRT tool collects publicly available information on 20 indicators of government responses. The indicators cover three topics: containment and closure policies, economic policies, and health system policies. The dataset is organized in a table where for each country appears a number indicating the level of severity of each of the indicators by date. See example in Appendix B.

We formulate a list of claims containing the policies of 20 countries/states[2] that represent diverse countries of the world during the year of 2020. Taking into account all 20 indicators, this template is used to create the claims: The government of [country/state name] decides to [indicator details] on [date range].

The second dataset, which is used as the corpus for finding evidence, is the AYLIEN Coronavirus Dataset. More than 1.5 Million news articles in English related to the pandemic were included in the dataset since the outbreak began in November 2019 to July 2021. For the 20 countries/states selected for this research we have made sure that there are at least a few dozen articles to make up the search space. The next section outlines the steps taken to process the AYLIEN documents in order to identify and verify the claims derived from OxCGRT.

4 Temporal and Spatial Filtering

We seek evidence to support claims on global government actions for the COVID-19 pandemic during 2020. The actions are formulated as claims that include spatial (name of country/state) and temporal (range of dates) information.

AYLIEN articles are annotated with publication time and publication source location (e.g., The New York Times is published in New York). However, we argue that this temporal and spatial information is insufficient to achieve our goal of evidence-finding and verification as the text can describe events that happened in locations other than the publication site as well as events that did not take place at the date of publication, but rather in the past (or in the future).

Temporal Annotations: Every document is annotated with a time frame that describes the range

[2]North America: New York, California, Florida, Canada. South America: Mexico, Chile. Europe: Italy, France, Russia, England, Germany. Asia: China, India, Oman, Israel, Iran, Japan. Africa: South Africa, Nigeria. Australia and Oceania: Australia, New Zealand.

of dates it refers to. To begin, we create a list of time expressions (e.g., yesterday, tomorrow, etc. See Appendix A for a complete list) and then identify which of them appear in the document. Next, we annotate these time expressions with the date they refer to using the publication date of the article as an anchor. For instance, if the text refers to an event that will occur the day after tomorrow and the publication date is October 24th, 2020, then the time phrase "day after tomorrow" will be annotated with October 26th, 2020. Finally, we assign each document the relevant date range based on the dates mentioned in the article. Time expressions appeared in 1503405 out of 1673353 articles in the dataset, i.e., in 89.84% of the articles.

Spatial annotations: The documents are grouped by country or state based on the location entities mentioned in them (a document may appear in more than one cluster). We first identify all of the LOCATION entities using the NER tool of Guo and Roth (2021). We then check if the location entity appears in a list of countries and states derived from the OxCGRT table. If it does not, meaning that it might be a settlement such as a city, we associate it with a country/state based on a list of cities and towns that we have extracted from Wikipedia for each country/state.

Filtering: Each document in the original corpus is annotated with both temporal and spatial information, including the range of dates and locations (to the state/country level) that are discussed in the document. By filtering out documents that do not pertain to the time and the geographical entity of each claim, we create a search space for the retrieval step. This filtering process is equivalent to forcing the retrieval algorithm to only return documents that are spatiotemporally accurate. Despite this effort, there may still be irrelevant documents. E.g., Georgia is both a state in the USA and a country located at the intersection of eastern Europe and western Asia. In this case, we would add the document discussing Georgia to both search spaces and would have to rely on the retrieval and entailment mechanisms to resolve the ambiguity.

5 Retrieval

Generally, fact verification systems consist of three components: document retrieval, sentence selection, and textual entailment. We next compare the performance of existing document retrieval methods when utilizing the original corpus and the filtered corpus.

	Emb		BM25	
	Filtered	Unfiltered	Filtered	Unfiltered
k=1	0.25	0	0.16	0
k=5	0.5	0.16	0.5	0.16
k=10	0.67	0.16	0.5	0.33

Table 1: Retrieval results. The values in the table are the percentage of HITS@k for *Emb* and BM25. The results are for the cases where the corpus was filtered and unfiltered.

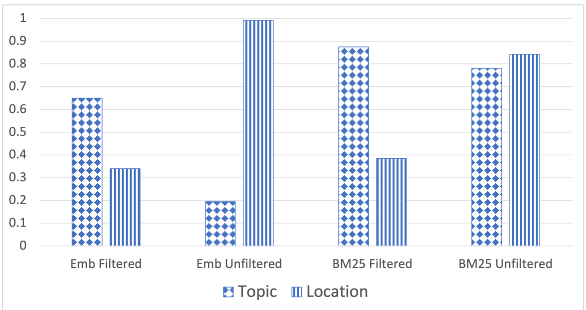

Figure 2: Error analysis for retrieval. The dotted/striped columns are the percentage of topic/location **irrelevant** retrieved documents.

Methods: We experiment with two retrieval methods. The first is Okapi-BM25, which is a bag-of-words method that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document[3]. The second is an embedding-based method for retrieval (denoted as *Emb*) in which the documents and the claims are embedded and then the ranking is held by solving a nearest neighbor search problem in the embedding space. We apply the method of Wu et al. (2018)[4].

Results: To evaluate the retrieval methods, we manually annotated the top 10 documents retrieved from each method in the filtered and unfiltered case for 12 claims that were randomly sampled (overall 480 documents were annotated with entailed/not-entailed labels). Table 1 presents the results. Additionally, we conducted an error analysis (see Figure 2) that classified the documents retrieved according to the types of errors – topic, location, or both. Temporal errors were not annotated since the dates are mostly not mentioned in the text, but

[3]We use the Python implementation of rank_bm25 (Robertson et al., 1995) imported from BM25Okapi package.

[4]`https://github.com/facebookresearch/StarSpace`

Model	Accuracy	Precision	Recall	F1
DocNLI	0.31	0.1	0.73	0.18
BERT	0.82	0.12	0.12	0.12
ANLI	0.88	0.33	0.14	0.20

Table 2: Entailment results. The performance of DocNLI, BERT, and ANLI entailment models for the top-10 retrieved documents from all retrieval models in both filtered and unfiltered cases.

are mentioned or inferred based on the meta-data (publication date).

For both retrieval methods, the results are substantially better in the filtered scenario with a gap ranging from 0.16 to 0.51 in the percentage of hits. Due to the fact that the filtered corpus contains a higher percentage of relevant documents than the unfiltered one, finding relevant content becomes easier. In the filtered scenario, *Emb* outperforms BM25, but not in the unfiltered scenario. The error analysis indicates that *Emb* made fewer mistakes with regard to the topic of the claims than BM25 in both the filtered and unfiltered cases. The error analysis also reveals that the filtering process prevents most errors concerning spatial information.

6 Entailment

We compare three textual entailment models for predicting a binary label (entailed/not entailed). One model is trained on single-sentence inputs and the other two are trained on inputs of varied lengths.

Methods: The first model is BERT-Based (Devlin et al., 2019) that is trained on an argument mining dataset from IBM debater (Ein-Dor et al., 2020) (denoted as BERT). This model's input is limited in length, hence we only send it single sentences as premise and hypothesis at a time. To determine if the entire text entails the claim we look for at least one positive response. The Second system is a RoBERTa-based architecture (Liu et al., 2019a) that is trained on the DocNLI corpus (Yin et al., 2021). This corpus consists of multiple genres and multiple ranges of length documents in both premises and hypotheses. The third model is another RoBERTa-based model trained on Adversarial NLI (ANLI, Nie et al., 2019b).

Results: Based on our annotations for the retrieval part, we calculate accuracy, precision, recall, and F1 scores for each of the entailment models. The results are shown in Table 2.

The best performing method is ANLI with 0.2 F1 score and 0.88 accuracy. Since the labels are very unbalanced (49 entailments out of 480 documents), the precision is critical to determine which method performs best. In this case it is also ANLI with 0.33 precision score.

The next section demonstrates how the best performing retrieval and entailment methods, together with the filtering adjustments can be used to finding evidence in a real-world scenario.

7 Case Study: Comparison between Germany and Nigeria

We compare the responses of developing and developed countries to the outbreak of COVID-19 in the first three months of 2020. As representatives of developing and developed countries, we selected Nigeria and Germany at random.

We were able to extract from OxCGRT 52 claims of government actions for Nigeria and 68 claims for Germany for the relevant time period. One possible reason for the difference in the number of actions is Germany's extensive global media coverage. Another explanation is Nigerian government being less proactive during that period of time. By applying our methods on the claims from both countries we can determine which explanation is more plausible.

Using the best methods for retrieval and entailment in the filtered case (*Emb* for retrieval and ANLI for entailment) we were able to verify 8 claims for Nigeria and 7 claims for Germany. That is, 36.3%/22.5% of the claims were verified for Nigeria/Germany, respectively. According to this, there is no significant difference between government response times and the number of actions taken. This finding supports the explanation that the difference in the number of claims originates from the global report bias toward Germany and not from Nigeria being less proactive. More results and comparisons to the unfiltered case appear in Appendix C.

8 Conclusion

We present methods for enhancing fact verification methods to be applicable for finding evidence for claims requiring temporal and spatial inferences. We demonstrate the benefits of these adjustments with a case study comparing global government responses to the COVID-19 pandemic.

9 Ethical Consideration

Manual annotations were made by the first and second authors in order to evaluate the proposed methods. Both authors independently annotated the examples, and then discussed each example for which they disagreed until agreement was reached (as well as explaining why the final label is correct). We believe the annotation level is high, and there are no ethical issues associated with this process since the authors are NLP researchers, working independently, and all discrepancies were resolved. Labels for annotated data will be released upon acceptance of the paper.

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, et al. 2021. Fighting the covid-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2020. Corpus wide argument mining—a working solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7683–7691.

Julian Martin Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from wikipedia gamification. *arXiv preprint arXiv:2104.04725*.

Julian Martin Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. *arXiv preprint arXiv:2010.00571*.

Ruohao Guo and Dan Roth. 2021. Constrained labeled data generation for low-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4519–4533.

Thomas Hale, S Webster, A Petherick, T Phillips, and B Kira. 2020. Oxford covid-19 government response tracker (oxcgrt). *last updated*, 8:30.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Zhijing Jin, Zeyu Peng, Tejas Vaidhya, Bernhard Schoelkopf, and Rada Mihalcea. 2021. Mining the cause of political decision-making from social media: A case study of covid-19 policies across the us states. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 288–301.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2019b. Fine-grained fact verification with kernel graph attention network. *arXiv preprint arXiv:1910.09796*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019b. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. Starspace: Embed all the things! In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449*.

Yi Zhang, Zachary Ives, and Dan Roth. 2020. "who said it, and why?" provenance for natural language claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4416–4426.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*.

A Temporal Expressions

A list of temporal expressions and connectives used to annotate articles with the relevant time frame:

Time expressions: "today", "tomorrow", "week", "month", "year", "days", "weeks", "months", "years", "Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November", "December".

Time connectives: "before", "after", "ago", "before the", "after the", "start of the", "end of the", "earlier in the", "later in the", "earlier this", "later this", "earlier", "later", "following", "previous", "next", "last".

Any combination of a time expression and time connective was detected and annotated based on simple mathematical operations using the publication date as an anchor point.

B OxCGRT Example

See Figure 3.

C Comparison between Nigeria and Germany Responses

In this section, we present more results from our case study comparing Nigeria and Germany with regards to government responses to the COVID-19 pandemic during the first three months of 2020. Figures 4 and 5 present timelines of the government's responses that we have been able to validate. Both governments appear to have begun responding actively to the epidemic around the end of February 2020, and the Nigerian government appears to have acted more broadly than the German government.

We also utilized the BM25 retrieval and ANLI entailment methods in the unfiltered case in order to demonstrate the benefits of filtering. We managed to verify 8 claims for Nigeria and 9 claims for Germany. However, after further review, we found that only one claim for Germany was correctly labeled, and no claim for Nigeria, since the majority of articles discussed other countries (i.e., were about countries other than Germany and Nigeria).

Country Name	Date	School closing	Workplace closing	Cancel public events	Restrictions on gatherings	Close public transport	Stay at home requirements	Internal movement restriction	International travel controls	Income support	Debt/contract relief	Fiscal measures	International support	Public information campaigns	Testing policy	Contact tracing	Emergency investment in healthcare	Investment in vaccines	Facial Coverings	Vaccination policy	Protection of elderly people	Confirmed Cases	Confirmed Deaths
Aruba	20200327	3	2	0	0	0	1	0	4	0	0	0	0	2	1	1	0	0	0	0	1	33	0
Aruba	20200328	3	2	0	0	0	1	0	4	0	0	0	0	2	1	1	0	0	0	0	1	46	0
Aruba	20200329	3	3	2	4	0	2	2	4	0	0	0	0	2	1	1	0	0	0	0	1	50	0
Aruba	20200330	3	3	2	4	0	2	2	4	0	2	0	0.56	2	1	1	0	0	0	0	1	50	0

Figure 3: Example of the OxCGRT table. The numbers in the table indicate the level of severity for which the action is being enforced. For example, on March 29th the government of Aruba changed its policy from having no restrictions on public events to canceling all public events.

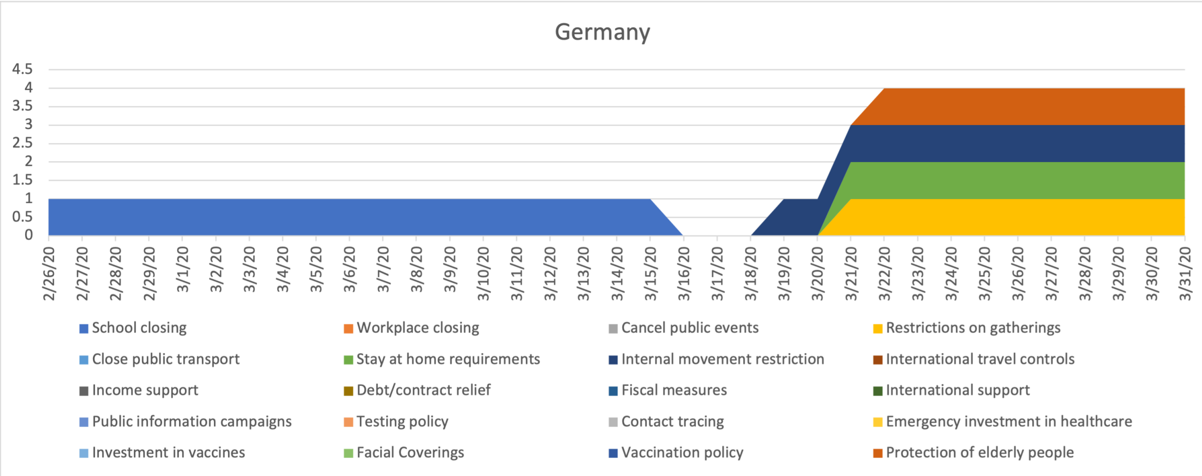

Figure 4: The government of Germany validated responses.

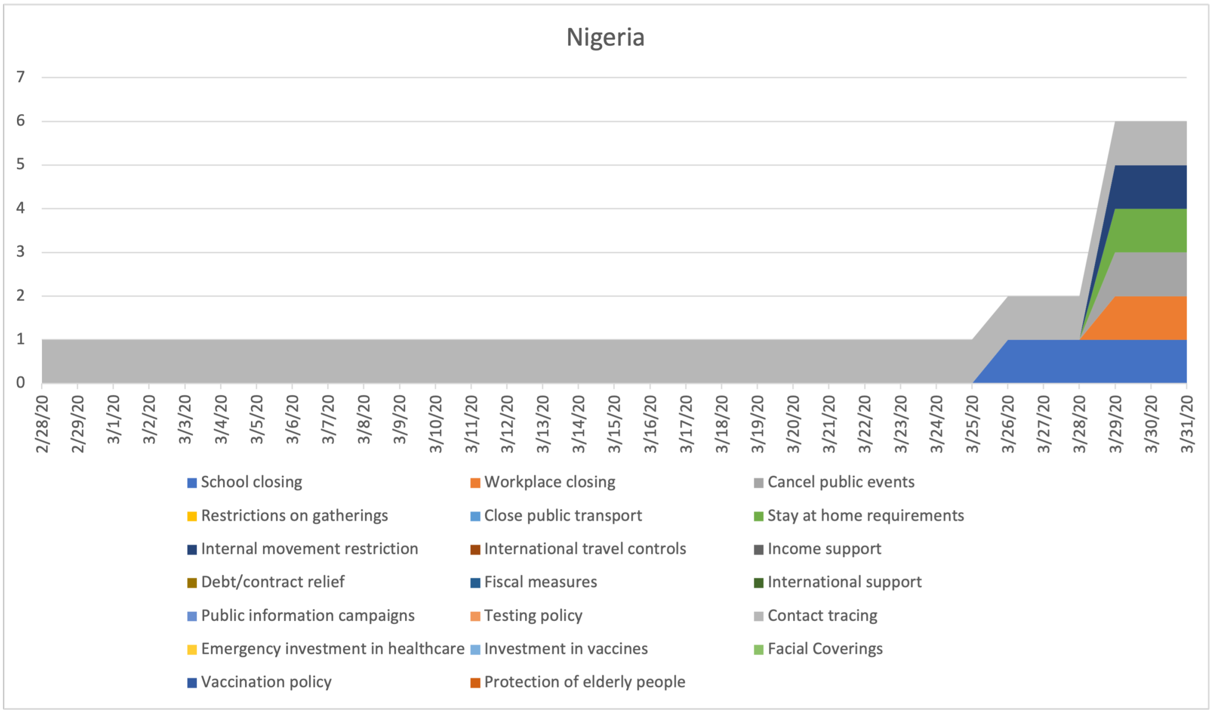

Figure 5: The government of Nigeria validated responses.