Stories and personal experiences in the COVID-19 Discourse

Neele Falk[1], **Gabriella Lapesa**[2]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]GESIS - Leibniz Institute for Social Sciences and Heinrich-Heine University of Düsseldorf, Germany
neele.falk@ims.uni-stuttgart.de, gabriella.lapesa@gesis.org

Abstract

Storytelling, i.e., the use of of anecdotes and personal experiences, plays a crucial role in everyday argumentation. This is particularly true for the highly controversial debates that spark in times of crisis – where the focus of the discussion is on heterogeneous aspects of everyday life. For individuals, stories can have a strong persuasive power; for a larger collective, stories can help decision-makers to develop strategies for addressing the challenges people are facing, especially in times of crisis. In this paper, we analyse the use of storytelling in the COVID-19 discourse. We carry out our analysis on three publicly available Reddit datasets, for a total of 367K comments. We automatically annotate the Reddit datasets by *detecting* spans containing storytelling and *classifying* them into: a) personal vs. general: is the story experienced by the speaker? b) argumentative function: does the story clarify a problem, potentially consisting in harm to a specific group? Does it exemplify a solution to a problem, or does it establish the credibility of the speaker?), and c) topic. We then carry out an analysis which establishes the relevance of storytelling in the COVID discourse and further uncovers interactions between topics and types of stories associated to them.

1. Introduction

While having a discussion about a controversial topic that can impact (or has impacted) the life of a large community, it is a very human strategy to resort to personal experiences or anecdotes to back-up one's position. Not only this is natural: it can be very effective, too. For example, a storyteller can highlight their personal experiences as an expert and appear more credible. In other cases, storytellers share experiences of harm that can arouse empathy among other discourse participants and facilitate perspective-taking (Polletta and Lee, 2006; Hoeken and Fikkers, 2014). Especially in times of crisis, where there is great uncertainty, the exchange of personal experiences plays a major role, both as an alternative to the overwhelming and constantly changing facts and to establish a collective identity.

'Storytelling', the phenomenon at the core of this paper, has increasingly gained importance in linguistic analysis and social science research on public discourse (Polletta and Lee, 2006; Black, 2008; Esau, 2018; Gerber et al., 2018; Dillon and Craig, 2021). Research in deliberative theory (Habermas, 1996) explores various facets of public discourse to understand its productivity. Storytelling is frequently recognized as an alternative mode of reasoning that empowers citizens to engage and contribute, irrespective of their backgrounds. Consequently, it has emerged as a valuable tool for enhancing a fundamental aspect of productive public discourse: equality and inclusion.

Storytelling within arguments can manifest in various ways. These include referencing personal experiences and backgrounds, providing a comprehensive and subjective account of a particular event, as well as making fragmentary allusions to recurring occurrences, all of which are grouped under this label (Falk and Lapesa, 2023). Existing research in computational argumentation has focused on identifying anecdotes and investigating personal experiences as specific types of premises (Park and Cardie, 2014; Song et al., 2016; Al-Khatib et al., 2016; Park and Cardie, 2018; Egawa et al., 2019; Wang et al., 2019; Falk and Lapesa, 2022).

This work focuses on the use of storytelling in the online discourse on the COVID-19 pandemic: we talk the reader through the annotation, modeling, and interpretation steps of our analysis. The COVID-19 discourse sample we analyse is constituted by a collection of English Reddit corpora containing 367K comments posted between January 2020 and October 2021 in different subreddits (e.g., change my view, news comments, general subreddits).

The following two examples provide the reader with a first impression of the facets of storytelling in the COVID-19 discourse: *"i'm a post-grad student in melbourne. i can't access any archives within my state, my thesis has just stalled. i can't work because of the lockdown. i can't even go for a walk with friends because i'm trapped within a 5km radius of my house."* compared to *"maybe if we had actually done something as a nation when any of the other black people were murdered by police we wouldn't be having protests during a pandemic. my wife is high risk (autoimmune) and as much as we would like to we are not out protesting (both white btw)."* While the first is a typical example of storytelling aimed at showing the negative consequences the speaker is experiencing due to something she considers wrong (lockdown),

LREC-COLING 2024, pages 15320–15340

the second example exemplifies the interaction between COVID-19 restrictions and other societal movements (black lives matters protests, in this case).

We structure our analysis in four research questions.

RQ1: How pervasive is storytelling in the discourse fragment encoded in our corpus? And how does it modulate in the different subcorpora?

RQ2: Which types of stories do speakers share? Are they personal (about themselves) or more general (represent collectives)?

RQ3: What is the function of the stories within the post they occur? Do they provide clarification, propose a solution, disclose harm, or establish the speaker's credibility?

RQ4: What are these stories about? That is to say, what are their topics?

To address these research questions, we need classification models able to identify storytelling spans in the COVID-19 Reddit corpus and assign them the properties that are relevant for our investigation. We develop such models exploiting storyARG (Falk and Lapesa, 2023), an existing corpus of argumentative texts drawn from a heterogeneous set of domains and topics with storytelling annotation at multiple layers, including function of the story in the argument (clarification, search for a solution, disclosure of harm, establishing speakers' background). Albeit heterogeneous, storyARG does not include the COVID-19 topic: this creates a domain-adaptation challenge that we successfully overcome with instruction based fine-tuning.

The contribution of this work is at multiple levels.

At the methodological level, we illustrate the different steps of an NLP-supported workflow for analysis of online discourse at large-scale. More specifically, we create a collection of 367k posts extracted from different subreddits related to the COVID-19 discourse and enrich it with the following information: a) whether a post contains storytelling and if so b) whether the story is a first-hand experience of the posts' author, c) which argumentative function(s) it takes on and d) which specific topic is discussed the story belongs to.

At the modeling level, we assess the models' ability to detect and classify storytelling spans across both in-domain and out-of-domain contexts using a manually annotated test set representative of our target discourse. We demonstrate the effectiveness of instruction-based fine-tuning in enhancing model performance when dealing with out-of-domain data.[1]

[1]The manually annotated test set and the COVID-19 storytelling discourse corpus are available under https://github.com/Blubberli/storytellingInCOVID19Discourse.git

At the level of the specific discourse we establish the relevance of storytelling in the COVID-19 discourse and further characterize its modulation through the interactions of our annotation layers with topics on one side (e.g., social distancing, conspiracy theories, lockdown, masking, home schooling) and story properties on the other. We find that personal stories prevail with social distancing, disclosure of harm stories highly correlate with environmental and social issues, stories that illustrate a solution dominate when it comes to home schooling, and stories which establish the background of the speaker often occur in the context of conspiracy theories.

At the level of the potential impact beyond NLP, we believe that understanding of personal narratives during crisis could (and should) become a driving force for the political discourse which targets policy-making. By illustrating how political decisions directly impact the daily lives of citizens, this analysis helps reveal shared concerns among various social classes and identifies potential actionable measures.

2. Data

2.1. Gold Data: storyARG

The basis of our models is the storyARG dataset (Falk and Lapesa, 2023) which contains a total of 2,385 storytelling spans, extracted by human annotators from 507 documents. Albeit not large, storyARG covers different domains with a rich annotation schema. To date, this is the only dataset in the Argument Mining community whose focus is on storytelling.

We exploit the span-level annotation from storyARG to identify the storytelling in our target discourse, thus addressing RQ1.

In storyARG, spans identified as containing storytelling are further characterized with annotation layers that cover both the argumentative and narrative properties of the storytelling spans. We exploit this annotation to address RQ2 and RQ3.

Personal vs. General stories Of relevance for this paper are the storyARG annotation layers regarding the *protagonist of the story* (individual, group or other, e.g. institution) and the *proximity of the story protagonist to the speaker*. The latter contrasts first-hand stories (those experienced by the narrator), second-hand stories (happened to someone known to the narrator), and stories for which the perspective cannot be determined as they are reported from an external point of view (e.g., the narrator may invent a hypothetical story to illustrate their opinion, labelled as "other"). We exploit these annotation layers to address RQ2: how personal are the COVID-19 stories?

Argumentative function of stories The storyARG annotation layer targeting the argumentative functions of the stories was developed based on a social science framework for annotating storytelling in deliberative discussions (Maia et al., 2020). Why did the speaker use the story in their post? Each span can take one of the following four functions:

Clarification: the speaker uses an analogy to illustrate their viewpoint. The story is used to derive a general statement from a concrete experience.

Establish background: the speaker uses personal experiences to establish themselves as an expert in a certain area or to make themselves more credible/authentic.

Disclosure of harm: a negative experience about suffering that is shared to evoke empathy and raise awareness about injustices or disadvantages faced by specific groups, arising from, for example discrimination, exploitation, or stigmatization.

Search for solution: a positive experience can be used to promote established policies or specific actions or to seek resolution of a dispute.

2.2. Target discourse: the COVID-19 Reddit Corpus

Our target discourse is a collection of three publicly available datasets extracted from Reddit.

COVID-19 Vaccine News Discussions This dataset contains ∼34k user comments from the *r/Coronavirus* subreddit[2] about COVID-19 vaccination news posted between November 2020 and January 2021. The dataset is publicly available.[3]

Change My View (CMV) from the reddit-covid-dataset We extract the subreddit *r/CMV* from the Reddit covid dataset[4] that contains any covid-related posts until October 2021 and all corresponding comments. The CMV subreddit consists of ∼35k user comments.

COVID-19 Vaccine Perceptions on Reddit The largest of the three datasets contains ∼267k user comments from 8,300 different subreddits that contain both COVID-19 and vaccine-related keywords. The comments have been posted between January and December 2020. The data has been analyzed by Kumar et al. (2022) who investigate how the topics in this discourse change over time.[5]

[2] https://www.reddit.com/r/Coronavirus/

[3] https://www.kaggle.com/datasets/xhlulu/covid19-vaccine-news-reddit-discussions

[4] https://socialgrep.com/datasets/the-reddit-covid-dataset

[5] Available under https://osf.io/urp2a/

3. NLP Pipeline

We use the gold data to train different models to extract and annotate storytelling spans. We train models for three different tasks:

Task 1 – Extraction of storytelling spans: To extract spans of interest we proceed in two steps. In a first step we classify whether the comment as a whole contains storytelling or not. This helps to reduce the number of comments to a smaller set of candidates. We use a publicly available text classification model[6] provided by Falk and Lapesa (2022). Second, we employ a sentence-based classifier to extract meaningful segments of storytelling. To do so, we train the classifier on the storyARG dataset, segmenting each document into individual sentences. During training, we treat each sentence within a storytelling segment as a positive example. During inference, all positively classified consecutive sentences can be merged into a coherent storytelling segment.

Task 2 – Classification: personal vs. general stories. We distinguish between personal and general stories based on the protagonist and the narrative proximity. We assume a story is personal if it is a first-hand experience and the main protagonist is an individual. Other experiences are labelled as general. We use this heuristic to create a label for *personal* for each storytelling span in the gold data.

Task 3 – Classification: argumentative function. Each storytelling span can belong to one or several argumentative functions. Given a storytelling segment as input, we use the annotations from storyARG to train a binary classifier for each of the four functions.

3.1. Experimental setup

For each task, we generate three distinct data partitions, where each partition consists of 80% for training and 20% for validation. We make sure that there is no overlap of documents between training and validation. We employ the classification models defined below.

Fine-tuning for text classification (all tasks): We fine-tune a `roberta-base` transformer model with a binary classification head for each task for a maximum of 10 epochs with a batch size of 16 and a learning rate of 2e-5. To tackle the class imbalance we use class weights during training and take the model with the highest F1 macro on the validation set.

Instruction fine-tuning (task 3, argumentative functions). As the training data for argumentative functions is small and the class distribution highly

[6] `falkne/storytelling-LM-europarl-mixed-en`

Instruction	Input	Target
Does the following personal experience or story express the a **disclosure of harm**? A disclosure of harm is defined as follows: A negative experience is reported that was either made by the discourse participants themselves or that they can testify to and casts the experiencer as a victim. The experience highlights injustice or disadvantage. Answer with yes or no.	Yes, we do have to face negative (even aggressive) reactions on a regular basis.	yes

Table 1: Format for instruction fine-tuning on argumentative functions.

imbalanced (e.g. 8% positive instances for disclosure of harm in the training data), we experiment with two alternative models.

We use the Flan-t5 model (Chung et al., 2022) which is based on the encoder-decoder model T5 (Raffel et al., 2020), fine-tuned on 1.8k tasks using instructions. In instruction fine-tuning, a model is not only provided with input-output pairs (e.g. English and French sentence pairs for translation) but with additional instructions that describe the target task (e.g. *instruction: translate the following English sentence into French*). This improves generalizability of LLMs with relatively small computational cost. Recent advances in instruction-based fine-tuning have focused on providing additional, large instruction datasets of high quality. We further fine-tune `flan-t5-XL` on the synthetic Alpaca dataset[7] which consists of 52k unique instruction-output pairs resulting in a powerful instruction-following LLM. We compare this model in a zero-shot setup to a version which we additionally instruction fine-tune on the argumentative functions. To do so, we convert the storytelling spans from storyARG into instruction-input-output pairs, where the input represents the span and the instruction describes an argumentative function (example in Table 8). We expect a yes / no answer as output (depending on whether the requested function is expressed by the story or not).

Summing up, for task 3, we contrast two instruction-based models (zero-shot, referred to as `flan-XL-alp-zero` in tables and plots in the paper) vs its fine-tuned counterpart trained on the storyARG functions, (`flan-XL-alp-storyARG`) to RoBERTa base fine-tuned on storyARG (`roberta`). Evaluation results are displayed in table 3.

3.2. Results

Creating a test set for the COVID-19 discourse. The COVID-19 Discourse Corpus, the target of our analysis, is completely out-of-domain compared to our gold corpus, storyARG, which instead contains the annotation layers we are interested in. To be able to assess the performance of our model, we need a manually annotated sample.

[7] https://github.com/gururise/AlpacaDataCleaned

	storyARG	Covid19
Task 1: extraction	0.80 ±0.01	0.71
Task 2: personal	0.87 ±0.01	0.81

Table 2: Extraction: *sentence-based f1-macro* score (storyARG) and *percentage of correctly identified stories* (Covid19); personal: *span-based f1-macro score* for personal vs general classification on both datasets. Averaged over three validation sets for storyARG (standard deviation in brackets). Model: full fine-tuning, `roberta-base`.

The starting point for the selection of this sample is the set of all storytelling spans in the COVID-19 Reddit corpus, identified with both pre-filtering and sentence-classification, as discussed in section 3.[8] To reduce the number of false positives we require a span to consist of at least 25 tokens. This results in a corpus that contains a total of 177,225 storytelling spans distributed over 91,589 comments. The majority come from *VaccinePerceptions* (155k). 15k stem from *CMV* and 8k from *VaccineNews*. As more than a quarter of all comments contain storytelling, the phenomenon is confirmed to be relevant in this discourse and further motivates a closer investigation.

Manual Annotation and Agreement We then extract a sample of 700 instances for manual annotation (250 from CMV and VaccinePerceptions, 200 from VaccineNews). Two annotators[9] validated whether a span had been correctly suggested as storytelling (task 1) and further annotated the storytelling span for the annotation layers relevant to task 2 and 3, following the storyARG guidelines for the relevant annotation layers. We asked annotators to mark difficult cases and specify the reasons for considering these as hard examples.

We compute the agreement for the overlapping subset of the data (cf. Table 7 in the appendix). The overall agreement is low to medium which is in line with the results of the agreement in the original storyARG corpus. One of the main difficul-

[8] As we trained three different sentence-based classifiers (one on each data split), we combine the predictions as an ensemble and take the majority vote.

[9] One student of computational linguistics (700 instances), one author of this paper (250 instances).

model	storyARG background	storyARG clarification	storyARG harm	storyARG solution	COVID19 background	COVID19 clarification	COVID19 harm	COVID19 solution
baseline-random	0.45 ±0.03	0.48 ±0.01	0.40 ±0.02	0.36 ±0.01	0.50±0.02	0.43±0.01	0.49±0.03	0.46±0.01
roberta	**0.69 ±0.02**	**0.61 ±0.05**	**0.71 ±0.02**	0.57 ±0.04	0.60±0.03	0.29±0.02	0.51±0.02	0.55±0.07
flan-XL-alp-zero	0.58 ±0.01	0.49 ±0.02	0.50 ±0.01	0.40 ±0.01	0.58±0.00	**0.46±0.02**	**0.72±0.01**	0.56±0.02
flan-XL-alp-storyARG	0.63 ±0.01	0.51 ±0.02	0.69 ±0.00	**0.64 ±0.00**	**0.62±0.03**	**0.46±0.04**	0.49±0.08	**0.65±0.05**

Table 3: F1-Macro score for each argumentative function for the three models. For storyARG averaged over validation sets, with standard deviation.

model	storyARG background	storyARG clarification	storyARG harm	storyARG solution	COVID19 background	COVID19 clarification	COVID19 harm	COVID19 solution
baseline-random	0.31 ±0.04	0.32 ±0.03	0.14 ±0.02	0.08 ±0.02	0.49±0.01	0.63±0.01	0.48±0.03	0.32±0.02
roberta	0.54 ±0.01	0.45 ±0.06	**0.46 ±0.05**	0.21 ±0.04	0.48±0.06	0.35±0.04	0.31±0.04	0.24±0.14
flan-XL-alp-zero	0.61 ±0.03	**0.60 ±0.01**	0.31 ±0.03	0.23 ±0.03	**0.60±0.00**	**0.66±0.02**	**0.73±0.01**	**0.44±0.01**
flan-XL-alp-storyARG	**0.63 ±0.02**	0.57 ±0.04	**0.44 ±0.02**	**0.37 ±0.03**	**0.58±0.06**	**0.66±0.11**	0.28±0.13	**0.43±0.10**

Table 4: F1-positive score for each argumentative function for the three models. For storyARG averaged over validation sets, with standard deviation. For covid19 averaged over seeds (zero-shot, random-baseline) or fine-tuned models (trained on different storyARG splits). In the model names, `alp`=alpaca and `func`=functions.

ties for annotating the functions is whether a story serves as clarification or not (very low agreement for this function). One annotator perceives the clarification function as widely applicable, suggesting that nearly all cases can be understood as drawing analogies between the story and the argument. In contrast, the other annotator is more conservative in attributing this function.

To consolidate the annotations for the subset, we employ a minority vote approach. This is because identifying storytelling in general and its argumentative functions involves subjective interpretations of the speakers' arguing strategies. We opt for the majority vote to assign the label for *personal*, as its annotation is less subjective.

Evaluation: Task 1 and 2 Table 2 shows the results for task 1 (extraction of storytelling spans) and task 2 (classification into personal vs. general stories). While extraction turned out to be a challenging task, the performance (F1-macro) is high for classifying whether a story is personal or more general for both the gold data and the out-of-domain COVID-19 test dataset.

To understand the challenges encountered by the models we resort to the annotators comments about difficult cases. Out of 700 annotated spans in the COVID-19 test data, 71% are examples of storytelling. Reasons for difficulty in annotation are those in which storytelling is scattered a lot across the post, the post stance is not clear or the span is not even argumentative.

Classification errors frequently arise when the model detects the recounting of concrete events as storytelling due to the presence of a sequential plot, despite the absence of personal involvement or subjective description. In these cases, the reported story lacks an expression of a stance or an argumentative position. Distinguishing between storytelling and non-storytelling instances proved challenging for annotators because the implicit nature of the stance or position conveyed through storytelling added complexity to the task.

Evaluation: Task 3 Table 3 and Table 4 show the performance of each model for each argumentative function (f1-macro and f1 for the positive class). For StoryARG all functions, except for clarification, represent the minority class. We report mean and standard deviation across three different test splits (storyARG) or across different models. For both test sets we use three different models (either trained on different splits of storyARG (`roberta` `flan-XL-alp-storyARG`) or based on running with different seeds (`baseline-random`, `flan-XL-alp-zero`).

It becomes evident that the classification of the functions is challenging. However, all models outperform the random baseline in most cases by a large margin.

We compare all models with each other, testing for significant differences using the Almost Stochastic Order test (Del Barrio et al., 2018; Dror et al., 2019) as implemented by Ulmer et al. (2022). The test compares two distributions of results and measures to which extend stochastic order is being violated. If the amount of violation is small enough, one model can be considered as superior (stochastically dominant) over the other. We report the results as heat maps for each function, each metric and both datasets (storyARG and COVID19 test-set) in Figures 4 to 7 in the Appendix.

We find that for in-domain data (storyARG), full fine-tuning of a transformer works well but does not generalize well to the new data (drop in performance for all functions for the COVID-19 dataset). The simpler text classification model also strug-

gles more with correctly classifying the positive classes.

Compared to roberta, the instruction-tuned models are less prone to class imbalance and improve in performance for the positive class in most cases.

We find a mixed picture when comparing the zero-shot performance of the instruction-tuned model (`flan-XL-alp-zero`) and the one fine-tuned additionally on storyARG (`flan-XL-alp-storyARG`). Testing both models in-domain, additional fine-tuning improves the performance significantly for most argumentative functions (not clarification). On the out-domain case, both models perform equally well on the f1 for the positive classes with an exception of disclosure of harms which are significantly better captured by the zero-shot model. We hypothesize that this function benefits most from related tasks the model has been fine-tuned for (e.g. sentiment analysis) and that additional instruction fine-tuning leads to an overfitting on the disclosures of harm in storyARG. The model tuned on storyARG however performs better for the f1-macro score for establish background and search for solution.

Zero-shot classification with an instruction fine-tuned model (`flan-XL-alp-zero`) does not perform well for detection harm and solution storytelling functions in storyARG but outperforms all other models by correctly classifying disclosures of harm in the COVID19 test set. We hypothesize that this function benefits most from related tasks the model has been fine-tuned for (e.g. sentiment analysis).

4. Analysis

We employ the classifiers described in section 3 to extract storytelling spans and classify them into personal or general using an ensemble of the three models and taking the majority vote. Since `flan-XL-alp-storyARG` 'wins' most model comparisons across all scenarios (10 out of 16 runs the model is best or on par with another best), we use this model for the argumentative functions. For better robustness we also employ an ensemble using the majority vote, except for disclosures of harms for which we find that taking the minority vote results in a higher performance (59% f1-macro score on COVID19 test set).

Additionally, as discussed in the introduction, we are interested in analyzing the distribution of different types of storytelling across the corpus with respect to what people talk about (cf. Antoniak et al. (2024) for a comparable approach to storytelling on Reddit). Therefore we apply topic modeling to identify the sub-topics of the COVID-19 Discourse. We consider each storytelling span as a document and apply neural topic modeling using BERTopic (Grootendorst, 2022).[10] We regulate the minimum size of a topic to only allow those that can be associated with multiple stories. We cap the number of topics permitted at 80. If the identified clusters exceed this number, topics will be merged to fit within this limit.

4.1. COVID-19 stories: which topics are the most salient in the storytelling discourse?

The final topic model identified a total of 72 topics. However, 50% of the data could not be assigned to any of the 72 topics, either because the topics were too small or they did not contain distinctive words to allow for effective categorization.

Detailed results and visualizations of the the topic modeling are reported in appendix section A.3: most frequent topics and their n-grams (figure 8), topics mapped to their coarse-grained labels (tables 9 and 10), hierarchical structure of the topics (figure 9).

The largest topic (Topic 0, *vaccine, covid, flu, people, get, one would, like, said, virus,*) (27%) is relatively general and encompasses all possible experiences and narratives related to vaccination and COVID-19. This includes experiences with the vaccine, negative sentiment on vaccination as well as COVID-19 infections and subjective descriptions of the state of the pandemic in a particular region or country. The second largest topic (Topic 1, *mask, masks, wear, wearing, wear mask, people wearing mask, wear masks, wearing masks, face*) deals with measures such as mask mandates and social distancing (3.5%), while the third largest topic (1.6%) focuses on the status of schools and education during the pandemic (*school, students, schools kids, online classes, teachers, teacher, campus, semester*).

In order to provide a better overview of the distribution of topics in the discourse, each of the 72 identified topics is annotated with a more coarse-grained label from a list of 13 topics that were manually created after qualitatively reviewing all the topics. Table 5 shows the proportion of each broader discourse topic in the COVID-19 storytelling discourse. Prominent topics include vaccination in general, measures, as well as the impact of the pandemic on personal, professional, and societal spheres of life. Personal areas and leisure activities are particularly dominant (7.2%). We also find a high percentage of storytelling for top-

[10] BERTopic uses sentence-transformers to create document embeddings, uses the embeddings for clustering in a first step and then generates coherent topic representations using class-based TF-IDF in a second step.

ics related to specific conspiracy theories (4.7%). These include a few scenario descriptions that are part of the conspiracy theory but the the majority of the reports in this topic involve personal experiences with friends and family members who have developed extreme views during the course of the pandemic. Topics related to broader societal issues are less prevalent, and storytelling is particularly common in relation to the social impact of the pandemic, such as the Black Lives Matter movement.

4.2. How does storytelling differ between sub-corpora?

Figure 1 shows the relative amount of each storytelling property (personal, argumentative functions) for each sub-corpus. With respect to more personal stories, we can see that while *CMV* is rather balanced, the *VaccineNews* corpus contains a high amount of personalized storytelling, while the *VaccinePerceptions* corpus contains more general reports. This can be attributed to the fact that participants in the VaccineNews corpus more frequently use personal background information and experiences (higher amount of *establish background*) to present their arguments, which is by definition personal (first-hand and individual). One possible explanation for this could be that credibility is more important in discussions about news articles, or that people discussing there have very diverse backgrounds and political views that need to be established before presenting the argument. In contrast, the VaccinePerceptions corpus includes many group- and politically-specific communities where the background is already defined by the community itself.

Another difference is the occurrence of disclosures of harm, which are much less common in the VaccineNews corpus than in CMV and VaccinePerceptions. This sub-corpus includes fewer affective experiences and comments with negative sentiment, possibly, because a more objective and balanced writing style is preferred in the context of discussing news articles.

4.3. Topics and storytelling properties: Interactions

In what follows, we report the most prevalent patterns identified with respect to each annotation layer (personal, argumentative function) and their relationship to the topics. For this, we look at more general and specific topics that have a high percentage of a certain property and carry out a qualitative analysis of corresponding identified storytelling spans. Figure 2 and 3 visualize the 5 coarse-grained topics with the highest relative amount of a certain property.

topic	perc.
vaccine development and distribution	60.1
public health measures and restrictions	11.5
personal area of life and leisure	7.2
financial and economic impact	5.1
anti-vax sentiment and conspiracy theories	4.7
professional area of life / job	3.8
social issues and consequences	2.4
government policies and decision-making processes	1.8
spread of the virus	1.0
healthcare and hospital systems capacity	1.0
global health and public health response efforts	0.8
measures against the spread of the virus	0.6
enviromental consequences	0.2

Table 5: Distribution of coarse-grained topic across the subset of COVID-19 storytelling discourse that was assigned a clear topic (84,857 storytelling spans).

Social distancing results in highly personal experiences, particularly in areas where social interaction plays an important role. This can be seen in Figure 2 (a) which shows a high relative amount of personal stories in topics related to personal area of life and leisure. People often share more personal experiences of how they cope with the restrictions and risks of the COVID-19 pandemic in specific contexts, such as whether and under what conditions they engage in social interaction (e.g. weddings, restaurant visits) during the pandemic, as in the following example about a postponed wedding: *"[...] in december if theres no vaccine for covid, ill move it again same wedding, just different day no biggie, [...], the peace of mind of not making any of my loved ones sick is absolutely worth it"*In COVID-specific topics that have a significant impact on everyone (restrictions, healthcare, public health measures), the proportion of personal and general storytelling is relatively balanced (around 40 to 50 percent of personal storytelling in those topics, cf. Figure 2 a). In the discourse on measures such as wearing masks, the participants reflect on these measures and their effects from both a personal perspective (*"i go to places that don't require masks"*), and a more distant observer perspective (*"weekly testing and wearing a mask for months but they still do it"*). From this external perspective, observers often draw evaluative conclusions if they perceive these measures as too strict, as clearly evident in the two examples (*"seeing all these people also not wearing facemasks"*; *"the residents themselves weren't wearing masks"*).

Complex issues often trigger the use of 'establish background' in order to enhance the credibility of a statement through a certain level of expertise. This function is commonly observed in discussions that deal with complex matters, such as the healthcare system or the specific active ingredient of a vaccine (as seen in a higher

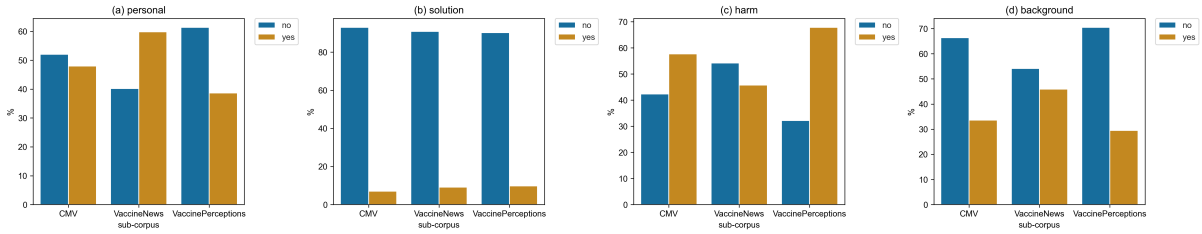
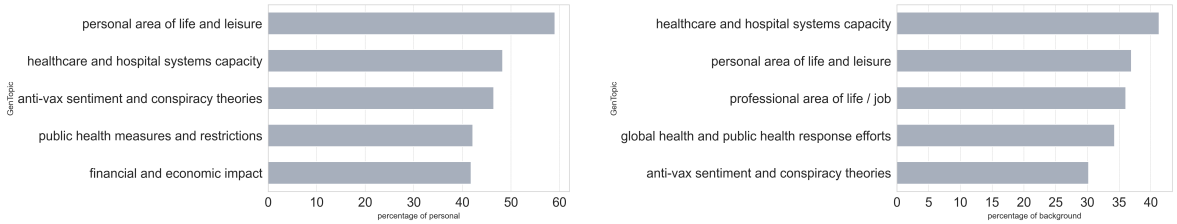


Figure 1: Relative frequency of each annotated storytelling-property in each sub-corpus of the COVID-19 Reddit Corpus.



(a) Highest amount of personal stories.

(b) Highest amount of establish background.

Figure 2: Coarse-grained topics with highest relative amount of storytelling properties.

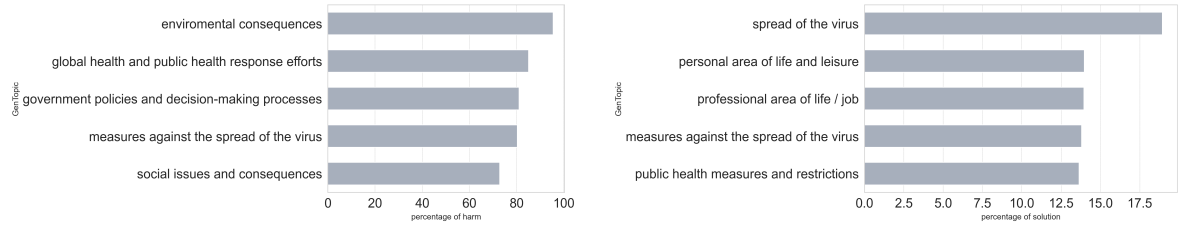
frequency in topics related to healthcare, global health, and vaccine development in Figure 2 (b)).

Participants in a discourse share their personal background to emphasize their expertise in a particular area and strengthen their credibility, especially when their contribution contains stronger claims, such as whether or not one should get vaccinated, as in *"i work in the frontline and has been since march when we first came in contact with covid in my country now, ive seen the effects of it, lived and worked through this shit im still a bit hesitant about a vaccine [...]"*. The speaker emphasizes that they "work on the frontline" and witness the daily events surrounding the COVID-19 pandemic, which enhances their statement "I'm still a bit hesitant about a vaccine," as they imply a closer proximity to the areas that ultimately "know everything" - the "experts". In *"i am a gp practice manager, so i am uniquely qualified to tell you that it's not just you, it is bloody confusing"*, the person mentions their professional background to reassure the other person, implying that if even if you have a professional and relevant background it is difficult to understand due to high complexity.

If it is not their professional or occupational background that gives participants credibility, they use personal experience as irrefutable evidence for something that they, as non-experts, cannot support with facts. These examples are often found when participants share experiences of illnesses and the effects of various vaccinations on themselves or their relatives (e.g. *"I got a flu shot last year for the first time and had zero issues"*).

In controversial topics, "establish background" is often used to distinguish oneself from a group that is being questioned. Especially in discussions that are critical of vaccines, participants often use this function to distance themselves from anti-vaxxers. For example, in *"i ain't no anti-vaxer (i have all my shots) and i know the short-term side effects are basically non-existent in some of the ones we working on but we really can't be that sure about them long term side effects"*, the person sets the stage at the beginning of their statement: "I'm not an anti-vaxxer," but then makes a critical comment about vaccines, attempting to avoid being attacked by those who support vaccinations.

Issues that concern society and social problems often involve a high number of disclosures of harm. There is a significant amount of disclosure of harm found in topics that address environmental consequences, government policies, and social issues (Figure 3 a), such as the COVID-19 pandemic and the Black Lives Matter movement, and the challenges faced by the government in responding to protests during the pandemic (*"[...]ive had to use my body as a shield for black protestors and might have gotten COVID-19 because these right wingers are the same whackos that think masks are the devil. but the cops dont do anything"*). The primary actors in these disclosures of harm are often a social group or a collective, including those of a non-human nature, such as the environment, the country, or the economy. This pattern highlights how the COVID-19 crisis has reinforced social inequality and structural problems, such as the widening gap between the rich



(a) Highest amount of disclosure of harm.

(b) Highest amount of search for solution.

Figure 3: Coarse-grained topics with highest relative amount of storytelling properties.

and poor, and how these issues are particularly evident during the pandemic (*"it's always insane to me how we live in a system where tons of people simply live paycheck to paycheck, can't afford health insurance,[...]"*).

Proposed solutions are mainly made in the work/school sector or in the concrete handling of measures to stop the spread of the virus. These can either be existing positive examples that have already been implemented (e.g., *"[...] they called me monday to tell me some other parents are also sane and said the same thing, and that he could take his tests online or on the flashdrives"*), or they can be suggestions on how to implement something, as in *"[...]have 1/5 of the population attending one day a week they would meet in class rooms with desks spread far enough apart for social distancing, while wearing masks, and have the teachers float from room to room[...]"*, which discusses possible solutions to the topic of home schooling or a hybrid learning setup and how one could implement a partial in-person teaching solution. A model that can automatically extract and categorize such proposals thematically could be useful in such a crisis to enable citizen participation online and to incorporate such citizen-oriented proposals into political decisions.

Other examples include less concrete solutions but rather motivation in the style of "together we can do this" or positive future prospects (e.g., *"I just need to wait till we achieve herd immunity and then I'll be ready to party with the rest of you Until then, I'll continue to be careful, [...]"*). These can be a source of hope in a rather pessimistic discourse and encourage positive thinking.

5. Conclusion

The present study aimed to investigate the use of storytelling in the COVID-19 discourse. To accomplish this, we utilized a pre-existing gold standard dataset containing storytelling spans and annotated properties as training data to develop models for extraction and classification. Our results indicate that while the extraction of storytelling spans works well, the models encounter difficulties in distinguishing between factual event descriptions and those used to support an argumentative position or attitude. While the classification of a storytelling span into personal or general is accurate, classifying the argumentative function is more challenging due to suboptimal class distribution and its inherent subjectivity. However, we find that fine-tuning the models on the argumentative functions using instructions improves over standard text classification and generalizes better on new data.

In the COVID-19 discourse, we identify a high prevalence of storytelling across various topics related to the pandemic, including vaccination, restrictions, and public and global health measures. Moreover, we identify specific trends regarding certain types of stories. Highly personalized stories are shared in the context of leisure activities and social interaction during social distancing, while people established background to increase perceived expertise in complex or domain-specific topics, and whenever they aim to distance themselves from criticized groups. Disclosures of harm are frequently found when discussing social issues in society and concrete solutions are often proposed within the context of home schooling or to motivate each other in times of dispair. These findings highlight the relevance of storytelling as an argumentative device in public discourse encourage further research on this topic.

6. Ethics Statement

The analysis is based on model predictions that are inherently susceptible to errors, particularly due to the models being trained on a separate source dataset that, although covering diverse domains, is limited in size. Consequently, the frequencies extracted for each type of story may not accurately reflect the actual frequencies and must be analyzed only as a proxy. We have assessed the model's performance on a sample of the new discourse to gain a better understanding of its classification accuracy and to identify stories that are miss-classified. Additionally, we have conducted

a manual analysis of several examples across different topics and properties as part of the analysis, and we have found that the model's predictions can be meaningfully interpreted. However, future research should aim to obtain larger training datasets encompassing various types of storytelling for improved generalizability.

Despite using the most recent and advanced topic modeling approach, topic modeling is still subject to limitations. For instance, different topic models can result in variation in the number of type of topics that are found due to the clustering approach. We have trained several topic models and compared their output, and while we have found different numbers of topics, the essential topics are usually identified by all the models. However, the analysis is based on only one topic model and is thus susceptible to bias towards the topic distribution identified by this model. Furthermore, interpreting the latent topics can be challenging as it is based on word lists and manual inspection, which is ultimately subject to the interpreter's bias.

In this study, we employed a state-of-the-art LLM and instruction fine-tuning technique, which has recently shown substantial improvements in performance on various NLP tasks. However, some advanced models, such as chatGPT or GPT3.5, are not open-source and pose significant challenges for researchers to access. Moreover, training such models requires considerable computational and economic resources. To overcome this limitation, we leveraged the publicly available Alpaca dataset and `flan-t5-large` model, which can be easily accessed through the huggingface library. To promote open science, we plan to share our fine-tuned model through the same platform.

Nonetheless, the use of such models requires careful consideration due to the risk of generating harmful or incorrect content and information. In our study, we used the model for text classification, specifically for generating yes/no answers. However, any other application of this model should be approached with caution and ethical considerations.

7. Acknowledgements

This work is supported by Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation). We are thankful to our student annotator whom we do not mention here explicitly for anonymity reasons. We also thank Melanie Andresen for additional feedback on the manuscript.

8. Bibliographical References

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where do people tell stories online? Story detection across online communities. *arXiv preprint 2311.09675 cs.CL*.

L. Black. 2008. Listening to the city: Difference, identity, and storytelling in online deliberative groups. *Journal of Public Deliberation*, 5:4.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

S. Dillon and C. Craig. 2021. *Storylistening: Narrative Evidence and Public Reasoning*. Taylor & Francis.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Katharina Esau. 2018. Capturing citizens' values: On the role of narratives and emotions in digital participation. *Analyse & Kritik*, 40(1):55–72.

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553, Dublin, Ireland. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. Association for Computational Linguistics.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jurgen Habermas. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, Cambridge, MA, USA.

Hans Hoeken and Karin M. Fikkers. 2014. Issue-relevant thinking and identification as mechanisms of narrative persuasion. *Poetics*, 44:84–99.

Navin Kumar, Isabel Corpus, Meher Hans, Nikhil Harle, Nan Yang, Curtis McDonald, Shinpei Nakamura Sakai, Kamila Janmohamed, Keyu Chen, Frederick L. Altice, Weiming Tang, Jason L. Schwartz, S. Mo Jones-Jang, Koustuv Saha, Shahan Ali Memon, Chris T. Bauch, Munmun De Choudhury, Orestis Papakyriakopoulos, Joseph D. Tucker, Abhay Goyal, Aman Tyagi, Kaveh Khoshnood, and Saad Omer. 2022. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. *BMC Public Health*, 22(1).

Rousiley C. M. Maia, Danila Cal, Janine Bargas, and Neylson J. B. Crepalde. 2020. Which types of reason-giving and storytelling are good for deliberation? assessing the discussion dynamics in legislative and citizen forums. *European Political Science Review*, 12(2):113–132.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Francesca Polletta and John Lee. 2006. Is telling stories good for democracy? rhetoric in public deliberation after 9/ii. *American Sociological Review*, 71(5):699–723.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Wei Song, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Anecdote recognition and recommendation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2592–2602, Osaka, Japan. The COLING 2016 Organizing Committee.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

A. Appendix

A.1. Data

A.1.1. StoryARG: preprocessing for training

The storyARG dataset is available in a non-aggregated fashion. That is, every storytelling span has been extracted and annotated by one annotator. Some of the storytelling spans have a token overlap based on which on can create an aggregated version of the dataset with unique storytelling spans. For the extraction of storytelling span (Task 1), all documents are split into sentences and labelled as 1 if they belong to any storytelling span, 0 otherwise.

For classifying a story into personal or general (Task 2), we use the non-aggregated version of the dataset and consider each storytelling span as a single instance which is labelled as personal if the proximity is *first-hand* and the protagonist is an *individual*. This results in having some similar storytelling spans in the training set, which can be seen as a form of augmentation. As there is no overlap in documents between training and validation data, it is made sure that a similar span that is in training, does not occur in validation.

For classifying the argumentative function (Task 3), we aggregate the dataset based on token overlap (70% or more token overlap will merge annotations for one span and we take the maximum span as textual input). All functions that have been annotated by one of the annotator are added as labels to that span.

Table 6 shows the average size and standard deviation of training and validation data for each task. We will release the splits in the paper repository.

task	train	val
Task 1: extraction	4184 (+-50)	888 (+-50)
Task 2: personal	2004 (+-27)	437 (+-27)
Task 3: functions	1384 (+-17)	307 (+-17)

Table 6: Average training and validation size for gold data (*storyARG*) with standard deviation. Averaged over three splits.

A.1.2. COVID-19 test set: annotator agreement

Table 7 displays the Cohen's kappa score for each annotation layer. *Story* depicts the agreement between the two annotators over the subset of 251 instances, the agreement scores for the other annotation layers were computed for the subset for which both annotators agreed that it was storytelling, and therefore also annotated the storytelling-specific properties (n=178).

	cohen's κ
story	0.36
personal	0.23
establish background	0.31
clarification	0.07
disclosure of harm	0.31
search for solution	0.43

Table 7: COVID-19 test set: Cohen's kappa score for each annotation layer.

A.2. Modeling

A.2.1. Implementation Details

We use the transformer library and huggingface (`https://huggingface.co/roberta-base`) to fine-tune `roberta-base` on each task. We train the model on 3 GPUs (NVIDIA RTX A6000, each GPU has 49GB, CUDA Version 11.7). Training the model for 10 epochs on 3 GPUs takes $\sim$ 4 minutes.

To fine-tune the instruction-based models we use `google/flan-t5-xl` (https://huggingface.co/google/flan-t5-xl) and the Alpaca dataset (https://github.com/gururise/AlpacaDataCleaned). This dataset consists of 52k unique instruction-output pairs and is freely available. It has been created to develop an instruction-following open-source alternative, called Alpaca (https://crfm.stanford.edu/2023/03/13/alpaca.html) which uses the Llama LLM (Touvron et al., 2023) as a foundation model. The instruction dataset was automatically generated with GPT 3.5. We use the following repository to fine-tune the model: https://github.com/declare-lab/flan-alpaca.

For fine-tuning the model on additional data from the source domain that contains the task to identify the argumentative function, we convert each instance to the instruction-input-output format shown in Table 8.

We fine-tune the model on the instruction-datasets for 3 epochs with maximum source length of 200, and defaults for all other hyperparameters. During inference we limit the output length to 4 so we are able to map anything containing "yes" or "no" to the respective label. Fine-tuning for 1 epoch takes $\sim$ 3.5 hours on one GPU (NVIDIA RTX A6000, 49GB). To train the topic model we use the library BERTopic (https://maartengr.github.io/BERTopic/index.html#quick-start) and the `sentence-transformers/all-MiniLM-L12-v2` model (https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2) for creating the embeddings.

A.2.2. Significance analysis

Figures 4 to 7 show the model-to-model calculated significance values for the almost stochastic order test for each of the storytelling functions, in StoryARG and in the COVID-19 set, on F1 macro and F1 positive class. They display the Almost Stochastic Order Scores (ϵ) adjusted by using the Bonferroni correction. ϵ = 0.0 means model in row is stochastically dominant over model in column, $\epsilon < 0.5$ denotes almost stochastic dominance.

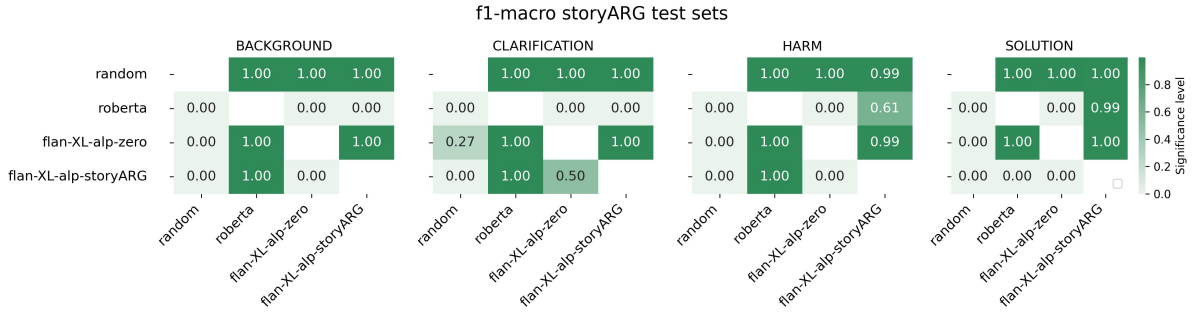


Figure 4: StoryARG: Model-to-model significance test, F1 macro

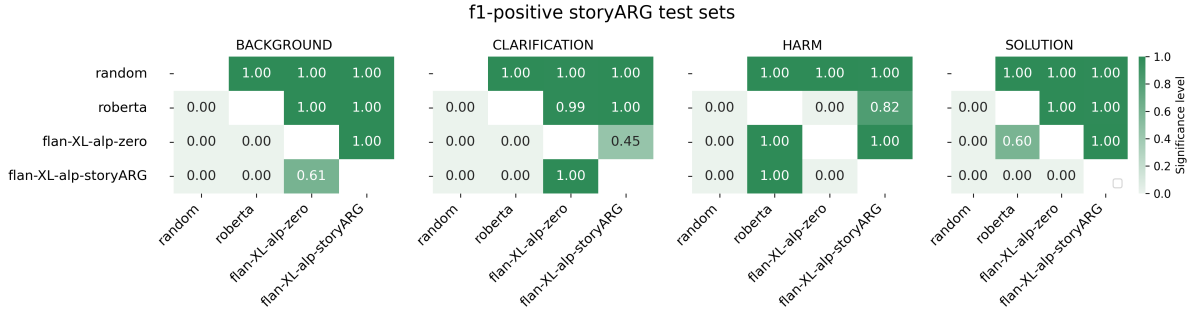


Figure 5: StoryARG: Model-to-model significance test, F1 positive class

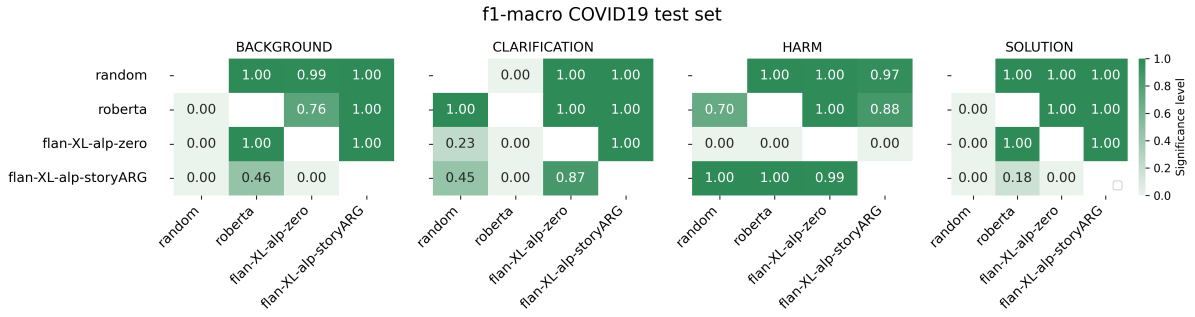


Figure 6: COVID-19: Model-to-model significance test, F1 macro

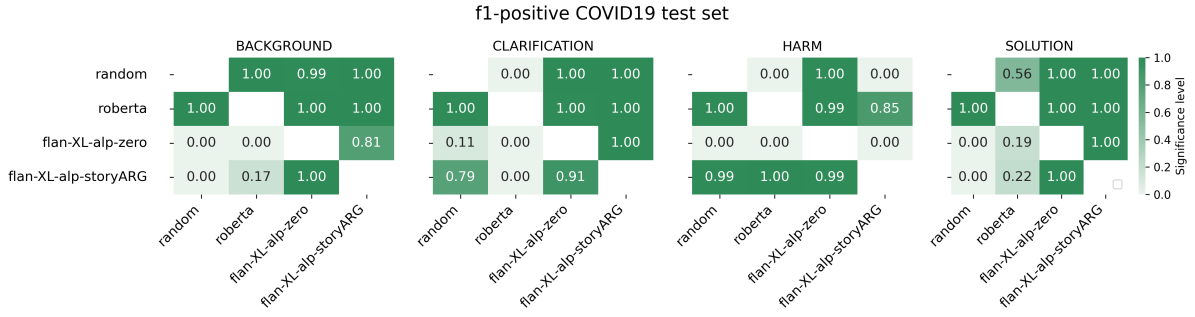


Figure 7: COVID-19: Model-to-model significance test, F1 positive class

A.2.3. Instructions for fine-tuning

Table 8 displays the instructions used for fine-tuning on argumentative functions (bold is only for formatting). The *establish background* example is a multi-label (it is also annotated as *search for solution*). In practice, this means that it has been shown to the model with both instructions (*establish background*) and *search for a solution*, in both cases with a "yes".

Instruction	Input	Target
Does the following personal experience or story express a **clarification**? A clarification is defined as follows: Through the story or personal experience in the argument, the authors clarify what position they take on the topic under discussion. The personal experience clarifies the motivation for an opinion or supports the argument of the discourse participant. The story or personal experience can help the discourse participant to identify with existing groups (pointing out commonalities) or to stand out from them (pointing out differences). The story or personal experience can illustrate how a rule or law or certain aspects of the discourse topic effect everyday life. Answer with yes or no.	I love going into a restaurant and being able to eat everything on the menu.	yes
Does the following personal experience or story **establish background**? Establish background is defined as follows: The participants mention experiential knowledge or share a story to emphasize that they are an 'expert' in the field or that they have the background to be able to reason about a problem. The goal can be to strengthen their credibility. Answer with yes or no.	I've been a vegan for nearly a year now and can tell you that all systems work better!	yes
Does the following personal experience or story express the a **disclosure of harm**? A disclosure of harm is defined as follows: A negative experience is reported that was either made by the discourse participants themselves or that they can testify to and casts the experiencer as a victim. The experience highlights injustice or disadvantage. For example, the negative experience may describe some form of discrimination, oppression, violation of rights, exploitation, or stigmatization. Answer with yes or no.	Yes, we do have to face negative (even aggressive) reactions on a regular basis.	yes
Does the following personal experience or story express the **search for a solution**? A search for solution is defined as follows: A positive experience is reported that can serve as an example of how a particular rule can be implemented or adapted. It may indicate suggestions of what should or should not be done to achieve a solution to the problem. The experience may indicate a compromise. Answer with yes or no.	Mine come from local farmers who treat their animals well.	yes

Table 8: Instructions used for fine-tuning: storytelling functions

A.3. Topic analysis

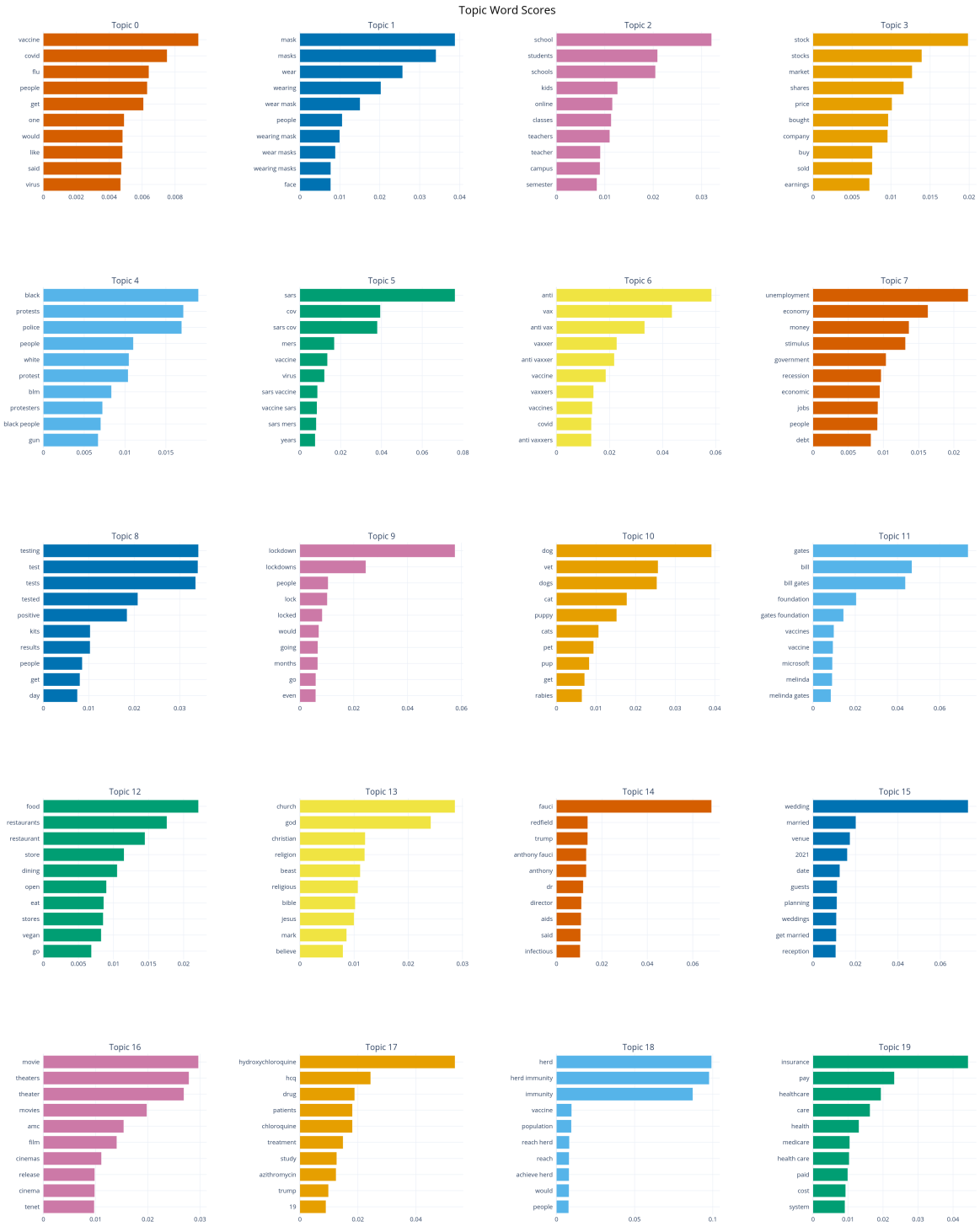


Figure 8: 20 most frequent topics identified by the storytelling BERTopic model and their corresponding most important n-grams.

Topic	BOW	General Topic
0	vaccine covid flu people get one would like said virus	Vaccine development and distribution
1	mask masks wear wearing wear mask people wearing mask wear masks wearing masks face	Public health measures and restrictions
2	school students schools kids online classes teachers teacher campus semester	professional area of life / job
3	stock stocks market shares price bought company buy sold earnings	financial and economic impact
4	black protests police people white protest blm protesters black people gun	social issues and consequences
5	sars cov sars cov mers vaccine virus sars vaccine vaccine sars sars mers years	Vaccine development and distribution
6	anti vax anti vax vaxxer anti vaxxer vaccine vaxxers vaccines covid anti vaxxers	Anti-vax sentiment and conspiracy theories
7	unemployment economy money stimulus government recession economic jobs people debt	financial and economic impact
8	testing test tests tested positive kits results people get day	Public health measures and restrictions
9	lockdown lockdowns people lock locked would going months go even	Public health measures and restrictions
10	dog vet dogs cat puppy cats pet pup get rabies	personal area of life and leisure
11	gates bill bill gates foundation gates foundation vaccines vaccine microsoft melinda melinda gates	Anti-vax sentiment and conspiracy theories
12	food restaurants restaurant store dining open eat stores vegan go	personal area of life and leisure
13	church god christian religion beast religious bible jesus mark believe	personal area of life and leisure
14	fauci redfield trump anthony fauci anthony dr director aids said infectious	Government policies and decision-making processes
15	wedding married venue 2021 date guests planning weddings get married reception	personal farea of life and leisure
16	movie theaters theater movies amc film cinemas release cinema tenet	personal area of life and leisure
17	hydroxychloroquine hcq drug patients chloroquine treatment study azithromycin trump 19	Anti-vax sentiment and conspiracy theories
18	herd herd immunity immunity vaccine population reach herd reach achieve herd would people	Measures against the spread of the virus
19	insurance pay healthcare care health medicare health care paid cost system	Healthcare and hospital systems capacity
20	dr could best tl make original tl dr could make tl bot reduced quot original	None
21	hands wash wash hands touch hand washing gloves sanitizer face clean	Public health measures and restrictions
22	gym gyms training membership going back covid bjj go workout	personal area of life and leisure
23	season football players league fans team game games play player	Personal area of life and leisure
24	quarantine people quarantined reopen reopening 14 going stay quarantining quarantines	Public health measures and restrictions
25	rent property tenants prices month rental house estate housing real estate	financial and economic impact
26	5g towers 5g towers conspiracy gates bill gates covid bill us people	Anti-vax sentiment and conspiracy theories
27	pox chicken pox chicken shingles chickenpox get got vaccine shingles vaccine kid	Healthcare and hospital systems capacity
28	city live lived sf park downtown like place la go	Personal area of life and leisure
29	cuomo twitter reporter 21 ny albany albany based new york york blasio	Government policies and decision-making processes
30	fow fwom aww awe peopwe ouw reeeeeeeeeeeeee wike hew ow	Anti-vax sentiment and conspiracy theories
31	mrna mrna vaccines mrna vaccine vaccines rna vaccine technology protein years mrna technology	Vaccine development and distribution
32	nz new zealand zealand new australia borders country border cases island	Global health and public health response efforts
33	work home office work home working remote working home wfh company job	Professional area of life / job
34	normal back normal back return new normal year 2021 things go back get	Spread of the virus
35	polio polio vaccine vaccine salk children 1955 paralyzed people vaccines paralysis	Global health and public health response efforts
36	left conservative right liberal political wing side conservatives politics right wing	Government policies and decision-making processes

Table 9: Topics 0 to 36 with their top 10 most relevant words and their mapped coarse-grained label.

Topic	BOW	General Topic
37	news media newsmax fox propaganda fox news watch cnn sources truth	Social issues and consequences
38	conspiracy theories conspiracy theories theory conspiracy theory conspiracy theorist theorist conspiracies moon believe	Anti-vax sentiment and conspiracy theories
39	curve wave second wave flatten flatten curve second flattening flattened flattening curve hospitals	Spread of the virus
40	car driving drive cars drunk road accidents safety drivers traffic	Public health measures and restrictions
41	christmas thanksgiving family year holidays santa parents holiday see covid	Personal area of life and leisure
42	ellie joel abby michael dwight fireflies game meredith marlene firefly	Personal area of life and leisure
43	autism autistic cause autism vaccines vaccines cause cause child wakefield mmr license	Anti-vax sentiment and conspiracy theories
44	weight fat obese pounds overweight diet lose weight healthy eating lbs	Personal area of life and leisure
45	arm news armenpress armenia azerbaijan armenpress arm arm shushi armenian artsakh azeri	Government policies and decision-making processes
46	monkeys macaques rhesus animals oxford vaccinated rhesus macaques virus viral sars cov	Vaccine development and distribution
47	ivermectin borody patients study drug 19 covid 19 scabies therapy treatment	Spread of the virus
48	reopening reopen open opening states georgia new cases reopened opened	Public health measures and restrictions
49	freezers storage dry ice pfizer cold ultra cold ice ultra freezer dry	Vaccine development and distribution
50	tb bcg bcg vaccine tuberculosis tb vaccine bcg vaccination vaccine countries vaccination 19	Global health and public health response efforts
51	mink denmark minks farms mink farms mutated danish farm humans mutation	enviromental consequences
52	smoking smokers smoke smoked smoker quit cigarettes nicotine cigarette vape	Personal area of life and leisure
53	concert album tour band concerts music tickets live year see	Personal area of life and leisure
54	needles needle fainted blood faint shot shots get getting vasovagal	Vaccine development and distribution
55	tip tipping tips takeout tipped service food restaurant server orders	Personal area of life and leisure
56	trans gender gay women male rowling female men woman transphobic	social issues and consequences
57	cruise ship cruises cruising passengers ships san diego diego cruise ship san	Spread of the virus
58	contact tracing contact tracing tracers contacts contact tracers health testing close contacts cases	Public health measures and restrictions
59	bright rick bright rick hhs bowen development authority advanced research biomedical advanced whistleblower barda	Government policies and decision-making processes
60	www reddit reddit 2famitheasshole 2fr 2famitheasshole reddit amitheasshole wiki_post_deletion wiki faq compose 2fr faq wiki_post_deletion 2fr	Personal area of life and leisure
61	ship princess diamond princess diamond cruise passengers cruise ship princess cruise infected crew	Spread of the virus
62	coming china coming person coming going fine totally control control one china control 22 totally much shut	Government policies and decision-making processes
63	book books read reading novel fiction author chapters love story	personal area of life and leisure
64	narcolepsy pandemrix swine swine flu vaccine flu flu vaccine h1n1 people sweden	Vaccine development and distribution
65	airline airlines emirates dubai flights travel copeland air westjet said	professional area of life / job
66	translate action performed bot action performed automatically translation translate automatic translation translate sl sl auto auto tl tl en	None
67	microchips microchip covid vaccine covid chip vaccine track chips 19 track us	Anti-vax sentiment and conspiracy theories
68	inovio ino 4800 inovio pharmaceuticals ino 4800 dna vaccine pharmaceuticals company phase	Vaccine development and distribution
69	hair cut haircut beard cut hair salon curly hair cut shave like	Personal area of life and leisure
70	ventilators ventilator people ventilators patients people person hospitals die put treatment	Healthcare and hospital systems capacity

Table 10: Topics 37 to 70 with their top 10 most relevant words and their mapped coarse-grained label.

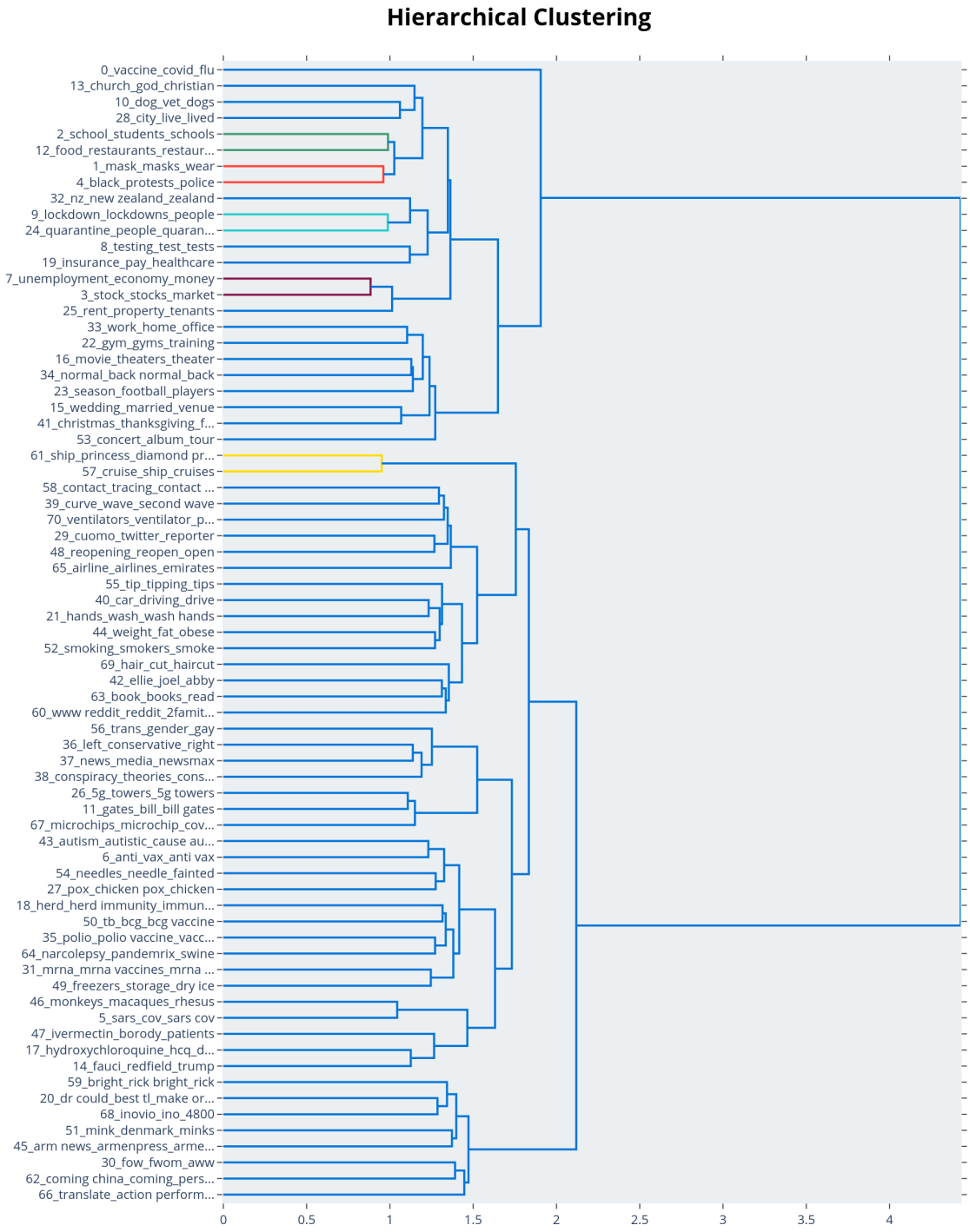


Figure 9: Hierarchical structure of the topics identified by the BERTopic model (storytelling). Topic distance computed based on the topic-term matrix (c-TF-IDF matrix)

.

A.4. Example comments

example	text span	property
(1)	i moved it to march. in december if theres no vaccine for covid, ill move it again same wedding, just different day no biggie, just a lot of phone calls and boom done and while its a hassle, the peace of mind of not making any of my loved ones sick is absolutely worth it	personal
(2)	even if i wear a mask, there's going to be a hundred or more people they'll pass that isn't i go to places that don't require masks, but i did go into a wal-mart once when i just needed to get something another store was out of i wasn't alone, seeing all of these people, also not wearing facemasks	personal
(3)	not a single case in a 150km radius, weekly testing and wearing a mask for months but they still do it the residents themselves weren't wearing masks in your office, not a hospital, not speaking to anyone for hours, distanced from everyone, in summer and sitting down wearing a mask is the height of hygiene theatre	general

Table 11: Storytelling with different perspective and main actor: personal and general examples.

example	text span	property
(1)	i work in the frontline and has been since march when we first came in contact with covid in my country now, ive seen the effects of it, lived and worked through this shit im still a bit hesitant about a vaccine it sounds nice i hope it is everything we are hoping for i will do my research on the vaccine before i decide to take or not take it a close friend of mine developed narcolepsy after the swineflu vaccine and it has ruined her life in so many ways	establish background
(2)	i am a gp practice manager, so i am uniquely qualified to tell you that it's not just you, it is bloody confusing the capitation payment formulae are quite complex, so much so that we don't work out how much we're paid per patient ourselves because it's too difficult	establish background
(3)	i ain't no anti-vaxer (i have all my shots) and i know the short-term side effects are basically non-existent in some of the ones we working on but we really can't be that sure about them long term side effects till the long-term done passed (unless there is reason for long-term effects to be impossible) i honestly haven't looked into pfeizer and moderna as much as should have but i will before i eventually take the well-calculated risk of one (or more?)	establish background

Table 12: Storytelling with establish background.

example		dimension
(1)	[...] its getting crazy out here. right wingers try and hit us with cars. ive seen multiple incidents where they stop get out of their cars and start punching protesters. ive had to use my body as a shield for black protestors and might have gotten COVID-19 because these right wingers are the same whackos that think masks are the devil. but the cops dont do anything [...]	harm
(2)	it's always insane to me how we live in a system where tons of people simply live pay-check to paycheck, can't afford health insurance, or afford education then you have the blatant greed and insanity of the billionaire class, which run the country, get away with not doing shit to help anybody	harm
(3)	two weeks ago i actually had to put my foot down with the schools. so far, they've been surprisingly smart about covid. [...], but we're the only county in the area that has not sent our kids back at all. they called 2 weeks ago to tell me that my eldest would have to go to the school on the bus, with all the other kids, just to take his eocs for the semester. i told them i'd let him go back when everyone was vaccinated and not before that. [...] they called me monday to tell me some other parents are also sane and said the same thing, and that he could take his tests online or on the flashdrives [...]	solution
(4)	and allow students to ask the lecturer question via the phone or online chat then to give them access to teachers again and limited socialization with fellow students, divide the student population and have 1/5 of the population attending one day a week they would meet in class rooms with desks spread far enough apart for social distancing, while wearing masks, and have the teachers float from room to room to teach the different subject[...]	solution
(5)	If that's the case, I just need to wait till we achieve herd immunity and then I'll be ready to party with the rest of you Until then, I'll continue to be careful, wear masks, avoid crowded indoor areas, etc My return to normal life will be more gradual than it will for most people, but I'll get there eventually	solution

Table 13: Storytelling with disclosures of harm and search for solution. Parts of the original comment omitted for space restrictions.