Global Surveillance of COVID-19 by mining news media using a multi-source dynamic embedded topic model

Yue Li
yueli@cs.mcgill.ca
School of Computer Science and McGill Centre for Bioinformatics, McGill University
Montreal, Quebec, Canada

Pratheeksha Nair
School of Computer Science, McGill University
Montreal, Quebec, Canada

Zhi Wen
School of Computer Science, McGill University
Montreal, Quebec, Canada

Imane Chafi
School of Computer Science, McGill University
Montreal, Quebec, Canada

Anya Okhmatovskaia
School of Population and Global Health, McGill University
Montreal, Quebec, Canada

Guido Powell
School of Population and Global Health, McGill University
Montreal, Quebec, Canada

Yannan Shen
yannan.shen@mail.mcgill.ca
School of Population and Global Health, McGill University
Montreal, Quebec, Canada

David Buckeridge
david.buckeridge@mcgill.ca
School of Population and Global Health, McGill University
Montreal, Quebec, Canada

Abstract

As the COVID-19 pandemic continues to unfold, understanding the global impact of non-pharmacological interventions (NPI) is important for formulating effective intervention strategies, particularly as many countries prepare for future waves. We used a machine learning approach to distill latent topics related to NPI from large-scale international news media. We hypothesize that these topics are informative about the timing and nature of implemented NPI, dependent on the source of the information (e.g., local news versus official government announcements) and the target countries. Given a set of latent topics associated with NPI (e.g., self-quarantine, social distancing, online education, etc), we assume that countries and media sources have different prior distributions over these topics, which are sampled to generate the news articles. To model the source-specific topic priors, we developed a semi-supervised, multi-source, dynamic, embedded topic model. Our model is able to simultaneously infer latent topics and learn a linear classifier to predict NPI labels using the topic mixtures as input for each news article. To learn these models, we developed an efficient end-to-end amortized variational inference algorithm. We applied our models to news data collected and labelled by the World Health Organization (WHO) and the Global Public Health Intelligence Network (GPHIN). Through comprehensive experiments, we observed superior topic quality and intervention prediction accuracy, compared to the baseline embedded topic models, which ignore information on media source and intervention labels. The inferred latent topics reveal distinct policies and media framing in different countries and media sources, and also characterize reaction to COVID-19 and NPI in a semantically meaningful manner. Our PyTorch code is available on Github (https://github.com/li-lab-mcgill/covid19_media).

CCS Concepts: • **Text mining and classification**; • **Infectious Disease Networks and Computational Epidemiology**; • **Knowledge Representation Applications**; • **Advancing Algorithms and Methods**;

Keywords: Topic models, Bayesian inference, coronavirus, media news, text mining

ACM Reference Format:
Yue Li, Pratheeksha Nair, Zhi Wen, Imane Chafi, Anya Okhmatovskaia, Guido Powell, Yannan Shen, and David Buckeridge. 2020. Global Surveillance of COVID-19 by mining news media using a

ACM-BCB'20, Sep 21–24, 2020, Virtual due to COVID-19

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
DOI: 10.1145/3388440.3412418

multi-source dynamic embedded topic model. In *11th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics Sep 21–24, 2020, Virtual due to COVID-19*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

The current pandemic of coronavirus disease (COVID-19), caused by the SARS-CoV-2 virus [20], is an unprecedented global public health emergency. It has spread rapidly, affecting people on all inhabited continents within six months, with 16.8 million confirmed cases and over 660,000 deaths as of July 29, 2020.[1] The severe clinical outcomes and high case fatality rate in some populations have prompted global efforts to control COVID-19 so that healthcare systems will not be overwhelmed. Given the current absence of effective vaccines and drugs, governments around the world have so far relied upon non-pharmacological interventions (NPI) to curb the spread and impact of the virus. NPI include measures to limit mobility (e.g., quarantining highly infected cities, banning international and domestic travel) and social distancing measures (e.g., cancellation of public events, closure of schools and work places). The nature, timing, and stringency of NPI implementation has varied across countries and there is little evidence to guide their use as we prepare for future waves of COVID-19.

To monitor and evaluate the global use of NPI across countries, multiple groups are manually reviewing official documents and news media. These efforts recently coalesced into a project supported by the World Health Organization[2] (WHO), but given that more than 100,000 news articles about COVID-19 were published daily over the last three months, manual review of news media has limited timeliness, recall, and ability to extract complex information, such as the reaction of communities to implemented interventions.

Public health agencies are aware of the potential value of online news surveillance and have developed systems to semi-automate the review of news media [8]. However, these systems currently use natural language processing methods only to detect entities such as geographical location, disease, and symptom. Even prior to COVID-19, it was clear that novel methods to automatically extract richer information from online media were needed [2]. Now, this need is even more pressing, as timely and comprehensive information on the global use and impacts of NPI are required urgently to guide disease control efforts [1]. In this paper, we present the Multi-Source Dynamic Embedded Topic Model (MSDETM) and the MixMedia model, which extend existing embedded topic methods [6, 7], and demonstrate the application of these methods for the surveillance of global efforts to control COVID-19. We explore the benefits of modeling news articles in a source-specific fashion and compare topic dynamics across countries and media outlets.

The structure of the remaining paper is as follows. Section 2 reviews the related work on media text mining. Section 3 briefly reviews the Embedded Topic Model (ETM) and the Dynamic Embedded Topic Model (DETM). The details of the proposed MSDETM and MixMedia models are then presented in Section 4. Section 5 describes the inference algorithm for learning these models. Section 6 introduces the COVID-19 media news data used in our experiments. Section 7 outlines the experimental evaluations and Section 8 presents our quantitative and qualitative results in comparison with the baseline methods. Section 9 concludes and identifies future directions.

2 Related work

While separate from surveillance purposes, analyses of policy reporting in news media and similar sources often focus on the identification of *media frames*. Framing refers to the ability of the news media to influence an audience regarding 'how to think' about an issue, by presenting selective and limited accounts of issues and events[19]. A traditional method for examining media framing is content analysis, which entails manual classification of articles using a set of, often predefined, keywords. Machine learning approaches have been developed to study media framing of policies while avoiding biases due to the knowledge and expectations of researchers. Among these methods, latent topic models, such as Latent Dirichlet Allocation (LDA)[5], are popular due to their rigorous Bayesian formalism and interpretability.

Recently, Walter et al. (2019) [19] used LDA to study media framing of epidemics. However, their works did not have a specific focus on surveillance of interventions. Ophir et al. (2018) [16] assessed the evolution of latent topic themes over the course of three different epidemics (H1N1, Ebola, Zika) and compared how closely articles within each topic followed risk factors pre-defined by the Centers for Disease Control (CDC) . The authors identified emphases and omissions in media coverage, such as the scarcity of information on individual responses to interventions to promote healthy behaviours. Poirier et al. (2020) [17] used topic modelling to examine the framing of COVID-19 in Canadian media. The authors derived six topics using LDA and interpreted them as common media frames (e.g. Chinese Outbreak, Social Impact, Health Crisis, etc). Although they found differences in framings between the anglophone and francophone media, this finding may have been due to the wider geographic coverage of the anglophone media.

An important aspect of this type of research is ensuring that the inferred topics capture semantic themes as opposed to syntactic context or superficial word co-occurrences. A main cause of low topic quality is the failure to account for variation in the data due to factors such as different media

[1] https://coronavirus.jhu.edu/

[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm

news outlets and biases in coverage across different countries. For instance, if United States had more news outlets than South Africa, then one might expect the inferred topics to be dominated by the news content related to United States.

However, most existing studies use naïve topic models without modeling the distinct patterns of dynamic topic evolution tailored to the coverage of each media outlet or country. This limitation is due mainly to the lack of a principled topic modeling approach that can model dynamic topics from multiple sources. Furthermore, there is no end-to-end model that can leverage the topic mixture for each document to predict document labels such as the COVID-19 intervention tags, which are often manually curated by human experts. To the best of our knowledge, our proposed models are the first to address these issues.

3 Background

The original LDA model [5] defines a fixed set of $V \times K$ topic distributions β as K independent Dirichlet distributions over a word vocabulary of size V. The Embedded Topic Model (ETM) [7] decomposes the unnormalized topic distribution β^* into the word embedding $\rho^\top \in \mathbb{R}^{V \times L}$ and the topic embedding $\alpha \in \mathbb{R}^{L \times K}$, where L denotes the size of the embedding space. The data generating process is as follows:

1. Draw a topic proportion θ_d for a document d from logistic normal $\theta_d \sim \mathcal{LN}(0, I)$:
   $\delta_d \sim \mathcal{N}(0, I)$; $\quad \theta_d = \text{softmax}(\delta_d) = \frac{\exp(\delta_d)}{\sum_k \exp(\delta_{kd})}$
2. For each token n in document d,
   a. Draw a topic assignment $z_{dn} \sim \text{Cat}(\theta_d) = \prod_k \theta_{dk}^{[z_{dn}=k]}$
   b. Draw a word $w_{dn} \sim Cat(\beta_{z_{dn}}) = \prod_v \beta_{vz_{dn}}^{[x_{dn}=v]}$, where $\beta_k = \text{softmax}(\rho^T \alpha_k) = \frac{\exp(\rho^\top \alpha_k)}{\sum_v \exp(\rho_v^\top \alpha_k)}$

The matrix of word embeddings ρ can either be initialized with previously fitted embeddings such as skip-gram [14] or learned along with the topic embeddings. Both ρ and α are fixed point estimates rather than time-varying variables.

Dynamic Embedded Topic Model (DETM) [6] extends ETM by learning the evolution of topics over time. The model inherits the generative process of the Dynamic LDA (D-LDA) [4] but operates in the embedding space:

1. Draw a topic proportion θ_d for a document d from logistic normal $\theta_d \sim \mathcal{LN}(\eta^{(t_d)}, \delta^2 I)$
2. For each token n in the document,
   a. Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$
   b. Draw word $w_{dn} \sim \text{softmax}(\rho^T \alpha_{z_{dn}}^{(t_d)})$, where t_d indexes the time at which document d is published

Here the topic prior η_t and the topic embeddings $\alpha^{(t)}$ are dynamic Gaussian variables at time point t, which depends only on the previous time point $t-1$ in the first-order Markov chain:

$$p(\eta^{(t)}|\eta^{(t-1)}) = \mathcal{N}(\eta^{(t-1)}, \zeta^2 I), \quad p(\alpha^{(t)}|\alpha^{(t-1)}) = \mathcal{N}(\alpha^{(t-1)}, \gamma^2 I) \tag{1}$$

Therefore, DETM models the transition of topics over time using a Markov chain where the transition probabilities are drawn from a normal distribution with the variances γ^2 and δ^2 controlling the smoothness of the Markov transition.

4 Multi-Source Dynamic Embedded Topic Models

Both models we propose capture the source-specific temporal evolution of the mean topic proportion as a properly defined dynamic topic prior (e.g., country-specific media preference over a fixed set of topics at any given time).

Inspired by the supervised topic model [13], we introduce a supervised component as an option into both frameworks. This enables our models to jointly learn topic embeddings and the classification of labeled documents (e.g., a manually applied intervention category label for a news article).

4.1 Proposed model 1: MSDETM

We extend DETM by incorporating a source-specific dynamic topic prior mean:

$$p(\eta_s^{(t)}|\eta_s^{(t-1)}) = \mathcal{N}(\eta_s^{(t-1)}, \zeta^2 I) \tag{2}$$

Our rationale is that different media sources may follow different prior distributions over the topics. For example, given a set of 3 topics at time t (say quarantine, lockdown, vaccine research), the media news may have a topic prior probabilities equal to $\eta_{media}^{(t)} = (0.1, 0.5, 0.4)$, whereas the official announcements may have a topic prior equal to $\eta_{official}^{(t)} = (0.2, 0.6, 0.2)$. Therefore, we seek to capture such source-specific topic dynamics.

Additionally, we add a linear classifier to the unsupervised model. The classifier uses the K-topic mixture θ_d for each document d to predict the multi-class C document labels:

$$p(\mathbf{y}_d|\theta_d) = \text{softmax}(\mathbf{W}^\top \theta_d) = \frac{\exp(\mathbf{W}^\top \theta_d)}{\sum_c \exp(\mathbf{W}_c^\top \theta_d)} \tag{3}$$

where $\mathbf{W}$ is a $K \times C$ matrix of weights that indicate the association of each topic with each class label c. The generative process of MSDETM is displayed in Figure 1a and described in the Appendix.

4.2 Proposed model 2: MixMedia

In the original D-LDA [4], D-ETM [6] and the above proposed MSDETM, both the topic mean $\eta^{(t)}$ and the topic embedding $\alpha^{(t)}$ are dynamic variables (i.e., lower-left portion in Figure 1a). While the dynamic topic embedding may capture low frequency topic evolution over many years, its benefit for modeling the topic evolution at higher frequency (e.g., daily or weekly) over months (e.g., COVID-19 media news collected within less than a year) is unclear.

From a computational perspective, modelling many time points with a large vocabulary size and many topics is expensive because the model complexity is quadratic in the T

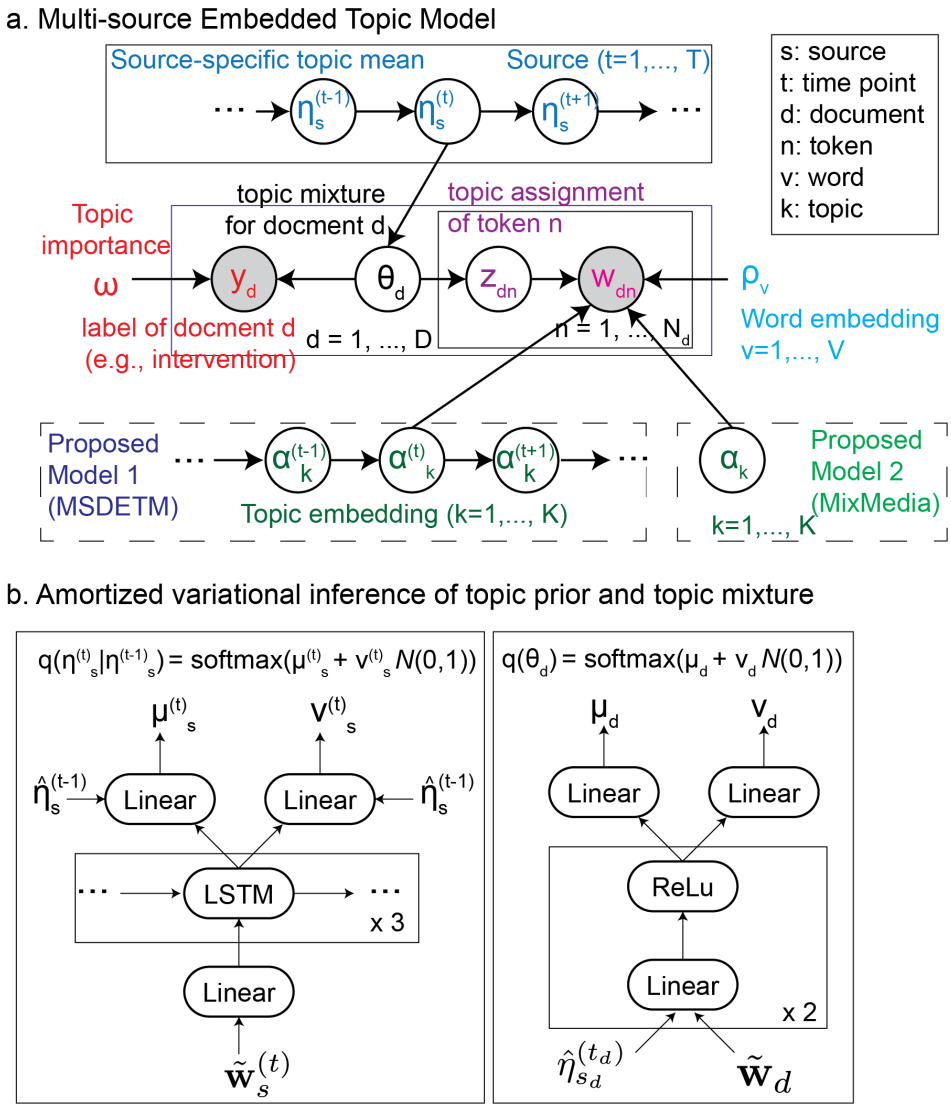

Figure 1. Proposed multi-source semi-supervised embedded topic model (MSDETM). **a**. Graphical model of MSDETM and MixMedia. **b** Amortized variational inference of topic mean η and topic mixture θ.

time points, vocabulary size of V, and K topics (i.e., $O(TVK)$). Interpretation of the dynamic topics is also not trivial as one needs to keep track of the changes of the same topic over time while trying to extract meaningful contents from many different topics.

While there is prior work on improving the scalability of the dynamic topic model (e.g., [3]) and visualization of the dynamic topics (e.g., [9]), we sought a simplified model that is more tailored towards mining media news over a short time interval. To this end, we have proposes the second model, called *MixMedia*. MixMedia is very similar to MSDETM except that it treats the topic embedding α_k for each topic as a *time-independent* Gaussian variable: $p(\alpha_k) = \mathcal{N}(0, \gamma^2)$ (see lower-right portion in Figure 1a). As in MSDETM, we keep the dynamic source-specific topic prior mean $\eta_s^{(t)}$.

To generate a document $\mathbf{w}_d$, we first sample the topic mixture θ_d from the source-specific topic prior $\eta_{s_d}^{(t_d)}$. We then sample the topic assignment for each word token z_{dn} from θ_d, and then sample the word w_{dn} from the fixed set of static topics $\rho^\top \alpha_{z_{dn}}$. As a result, MixMedia has the same model complexity as the ETM (i.e., $O(VK)$) and still takes into account the source-dependent temporal changes of documents via the topic priors $\eta_s^{(t)} \sim \mathcal{N}(\eta_s^{(t-1)}, \delta^2)$.

5 Model inference

In MSDETM, we have four latent variables, namely the dynamic topic prior η_s per source s, the topic mixture per document θ_d, the topic assignment per word per document token z_{dn}, and the dynamic topic embedding α_k per topic k. We

treat the word embedding ρ and topic importance ω as fixed point estimates and numerically optimize them via empirical Bayes.

We denote $\mathcal{D}$ as the data (i.e., the bag of words $\mathbf{w}$ over D documents). To ease the inference, we first marginalize discrete latent topic assignments z_{dn} in the conditional data multinomial likelihood: $p(\mathcal{D}|\theta,\beta) = \prod_{d,n}\sum_k p(w_{dn}|z_{dn} = k, \beta_k^{(t_d)})p(z_{dn} = k|\theta_d) = \prod_{d,n}\sum_k \theta_{dk}^{\top}\beta_{w_{dn}k}^{(t_d)}$, where $\beta_{vk}^{(t_d)} = \text{softmax}(\rho_v^{\top}\alpha_k^{(t_d)})$ for word v and topic k.

The posterior distribution of the rest of the latent variables $p(\eta,\alpha,\theta|\mathcal{D})$ are intractable. Hence, we took an amortized variational inference approach using a family of proposed distributions $q(\eta,\alpha,\theta)$ to approximate the true posterior [6] (Figure 1b). Specifically, we define the following proposed distribution:

$$q(\eta,\alpha,\theta) = \prod_d q(\theta_d|\eta_{s_d}^{(t_d)}, \mathbf{w}_d) \prod_s\prod_t q(\eta_s^{(t)}|\eta_s^{(t-1)}\tilde{\mathbf{w}}_s^{(t)}) \prod_k\prod_t q(\alpha_k^{(t)})$$

where

$$q(\theta_d|\eta_{s_d}^{(t_d)}, \mathbf{w}_d) = \text{softmax}(\delta_d); \quad (4)$$

$$\delta_d \sim \mathcal{N}(\mu_d, \text{diag}(\sigma_d^2)) = \mu_d + \text{diag}(\sigma_d)\mathcal{N}(0, I)$$

$$[\mu_d, \log\sigma_d^2] = NNET([\eta_{s_d}^{(t_d)}, \tilde{\mathbf{w}}_d]; \mathbf{W}_\theta) \quad (5)$$

$$q(\eta_s^{(t)}|\eta_s^{(t-1)}, \tilde{\mathbf{w}}^{(t)}) = \text{softmax}(\lambda_s^{(t)}) \quad (6)$$

$$\lambda_s^{(t)} \sim \mathcal{N}(\mu_s^{(t)}, \text{diag}(\nu_s^{(t)})) = \mu_s^{(t)} + \text{diag}(\nu_s^{(t)})\mathcal{N}(0, I)$$

$$[\mu_s^{(t)}, \nu_s^{(t)}] = LSTM([\eta_s^{(t-1)}, \tilde{\mathbf{w}}_s^{(t)}]; \mathbf{W}_\eta) \quad (7)$$

$$q(\alpha_k^{(t)}) = m_k^{(t)} + \text{diag}(\nu_k^{(t)})\mathcal{N}(0, I) \quad (8)$$

Here $\tilde{\mathbf{w}}_d$ is the word frequency for document d divided by the document length and $\tilde{\mathbf{w}}_s^{(t)}$ denotes average word frequency at time t for source s. The function $NNET(\mathbf{x};\mathbf{W})$ is a feed-forward neural network and $LSTM(\mathbf{x};\mathbf{W})$ is a recurrent neural network with Long Short Term Memory (LSTM) unit. We use NNET to estimate the sufficient statistics of the proposed distribution for the topic mixture for each document θ_d, and we use $LSTM$ to estimate the sufficient statistics of the proposed distribution for dynamic source-specific topic prior mean $\eta_s^{(t)}$. For computational efficiency, we assume a mean-field distribution for the topic embedding variables $\alpha^{(t)}$'s and use black-box variational inference to estimate its mean and variance.

Taking advantage of properties of Normal distribution, we use the re-parameterization trick [12] to stochastically sample the latent variable θ_d, $\eta_s^{(t)}$, and $\alpha_k^{(t)}$ from the their means with added Gaussian noise weighted by their variances as shown in Equations (5), (7), and (8), respectively.

To learn the above variational parameters $\Theta_q = \{\mathbf{W}_\theta, \mathbf{W}_\eta, \mathbf{m}_\alpha, \nu_\alpha\}$, we optimize the evidence lower bound (ELBO), which is equivalent to minimizing the Kullback-Leibler (KL) divergence between the true posterior and the proposed distribution $KL(q(\Theta)|p(\Theta|\mathcal{D}))$:

$$ELBO = \mathbb{E}_q[\log p(\mathcal{D}|\eta,\alpha,\theta)] + KL[p(\eta,\alpha,\theta)|q(\eta,\alpha,\theta)] \quad (9)$$

Here, the first term defines the data likelihood (or the negative loss function) and second term in defines the KL divergence between the proposed distribution and the prior distribution. Therefore, the Bayesian model is learned by maximizing the data likelihood while being regularized by its deviation from the prior.

We optimize ELBO with respect to the variational parameters by black-box stochastic variational inference [6, 10, 12, 18]. Specifically, we draw a sample of the latent variables from $q(\eta)$ (7), $q(\alpha)$ (8), $q(\theta)$ (5) based on a minibatch of data. We then use those draws as the noisy estimates of the variational expectation for the ELBO (9). We up-weight the ELBO by a factor that is equal to the ratio of the full training data over the batch size. The optimization is then carried out by back-propagating the gradients of the weighted ELBO into variational parameters.

Inference for MixMedia is the same as MSDETM except the topic embedding prior α_k's are static Gaussian variables ($p(\alpha_k) = \mathcal{N}(0, I)$), which is actually closer to the proposed mean-field distribution of $q(\alpha_k) = \mathcal{N}(m_k, v_k^2)$ in Eq 8.

6 Datasets

We used two corpora of news articles, one from the Global Public Health Intelligence Network (GPHIN) and the other from the World Health Organization (WHO). GPHIN is operated by the Public Health Agency of Canada (PHAC) and relies on an automated web-based system to scan newspapers and other communications worldwide for potential indicators of outbreaks, which are reviewed and rapidly assessed by multilingual, multidisciplinary analysts [8]. The GPHIN system processes over 20,000 online reports daily in nine languages, and maintains an annotated database of reports focusing primarily on communicable diseases. Since the beginning of the epidemic, GPHIN has maintained a table recording all reports that refer to an NPI and manually assigned labels to each report indicating the one or more types of NPI referred to in the report. From this table, we selected articles from an official or news media source that were assigned at least one label for a NPI. The resulting corpus included 8872 articles about NPI and the current COVID-19 outbreak dating from January 2020 to May 2020. Of these articles, 7334 were from media sources (news reports, social media postings), and 1538 were from official sources (government notices, articles published by global health institutions).

The global database of public health and social measures is a collaboration supported by WHO to harmonize the results of several projects tracking the global implementation of NPI

using a common taxonomy and structure for classifying NPI. [3] Each entry in the WHO database records an implemented NPI along with its geographical location and timing, and provides a link to a source. We extracted articles by limiting the source type to media and official sources, and excluded articles where no source was provided. Intervention labels were coded according to the WHO taxonomy. The resulting corpus contained 12 185 articles with 10 882 media articles (i.e., news reports) and 1 303 official articles (i.e., governmental notices). These articles are dated from January to May 2020.

Both datasets were cleaned to remove empty articles, misspelled country names were corrected and article dates were represented as week numbers of the year. The summaries of the articles were converted to lower case and the vocabulary of words was constructed after removing stop words, rarely occurring words (in less than 10 articles) and frequently occurring words (in more than 70% of the articles). The data was then split into train (85%), test (10%) and validation (5%) sets. All words not occurring in the train set were removed from the vocabulary. These steps are consistent with the data processing used in DETM [6].

7 Experiments

We used the training set for the model inference, the validation set to monitor overfitting and to choose the best model at a certain epoch, and the testing set to evaluate the model performance. For model selection, we used the validation perplexity [5]. We randomly split each validation document into two halves (i.e., two bags of words). We used the first half to infer the topic mixture θ_d for each validation document and the second half to compute the perplexity as the negative log multinomial likelihood on the held-out document ($\mathcal{L} = -\sum_d \sum_k \log \theta_d^\top \beta_{\mathbf{w}_d k}$), which can be thought of as the reconstruction loss that is commonly used in the matrix factorization literature.

To assess topic quality, we used the same metrics in [7] and [6]. Specifically, we first computed two topic scores, topic coherence (TC) and topic diversity (TD). TC is defined by the average point-wise mutual information of the top-10 most likely words under each topic k, that are drawn randomly from the same document [15]:

$$TC = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{45}\sum_{i=1}^{10}\sum_{j=i+1}^{10} f(w_i, w_j)$$

where

$$f(w_i, w_j) = \frac{\log P(w_i, w_j) - \log P(w_i) - \log P(w_j)}{-\log P(w_i, w_j)} = -1 + \frac{\log \frac{D_{w_i}}{D} + \log \frac{D_{w_j}}{D}}{\log \frac{D_{w_i, w_j}}{D}}$$

Here we estimated $P(w_i)$ and $P(w_i, w_j)$ as $\frac{D_{w_i}}{D}$ and $\frac{D_{w_i, w_j}}{D}$, where D_{w_i} denotes the number of documents containing word w_i and D_{w_i, w_j} the number of documents containing both word w_i and word w_j. Therefore, the top-10 most likely words under a coherent topic should co-occur frequently in the same document. TD is defined as the percentage of the unique words in the top 25 words across all of the K topics. Therefore, a diverse set of topics will have TD close to 1 whereas a set of redundant topics will have TD close to 0. Finally, the overall topic quality (TQ) is simply the product of TC and TD: $TQ = TC \times TD$.

Each media news article can be labeled with zero, one, or more intervention labels. Accordingly, we use the Top-K recall to assess the classification performance in predicting public health interventions. Top-K recall is defined as the number of true interventions among the top $K = \{3, 5, 10\}$ predicted interventions. As a comparison, we trained the baseline ETM and D-ETM followed by a multinomial linear regression model with a L2 norm penalty that used the topic mixture from the unsupervised models to predict intervention labels on the training set (i.e., ETM+LR and DETM+LR). We then applied ETM+LR and D-ETM+LR to the testing set to evaluate their predictions.

For all of the models, we set the training batch size to 200 documents per batch and the initial learning rate to 0.001 with Adam [11] to adjust it throughout our experiments. To infer topic mixture $q(\theta)$, we used 2 hidden layers with 800 hidden units each layer for the feedforward network. To infer topic mean $q(\eta)$, we used 3 LSTM layers with 200 hidden units each. For both neural networks, to avoid over-fitting, we used 0.1 dropout to regularize co-adaptation among the hidden units and weight decay rate 1.2e-6 to penalize large network weights.

For numerical stability, we clipped the gradient at 2.0. During the training of each model, we divided the learning rate by a factor of 4 if the validation perplexity at the current epoch was greater than the lowest validation perplexity obtained in the last 10 epochs. For dynamic topic models namely D-ETM, MSDETM, MixMedia, we set the variance $\zeta^2 = \gamma^2 = 0.005$ for both the topic prior mean variables $\eta_{sk}^{(t)} \sim \mathcal{N}(\eta_{sk}^{(t-1)}, \zeta^2)$ and the topic embedding variable $\alpha_k^{(t)} \sim \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2)$. For each model, we experimented with the number of topics $K \in \{5, 10, 15, 20\}$ based on the metrics above. We ran each model for 1000 epochs and took the best model at a certain epoch to be the one with the lowest validation perplexity.

[3] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm

Dataset	# Total Docs	# Official	# Media	# Weeks	# Sources†	# Interventions	Vocab
WHO	12185	1303	10882	20	211	42	4330
GPHIN	8872	1538	7334	19	289	17	3895

Table 1. The COVID-19 news datasets used in this study. †Sources include countries and organizations.

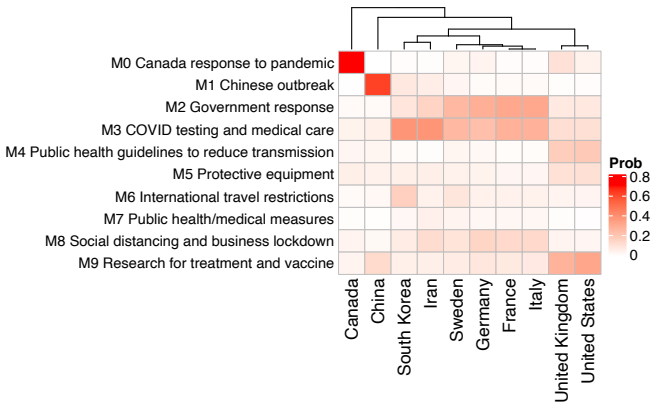

Figure 2. Country-specific topics mean η_s learned by the MixMedia model on the GPHIN data. MixMedia was trained with **countries** as source information **but without using time and label information**.

To obtain robust performance estimates, we repeated 10 random 85%-train/10%-validation/5%-test splits and recorded the mean and standard deviation of the performance scores (i.e., topic quality and top-K recall) for each model on the testing set.

8 Results

8.1 Quantitative analysis

We observed that our proposed MSDETM and MixMedia models resulted in superior topic quality compared to the baseline ETM and DETM models (Table 2). This improvement in topic quality suggests that it is beneficial to systematically integrate the source (i.e., 1. countries or organizations; 2. media news and official news types) and the label information (i.e., interventions). Among the two best performing methods, MixMedia compares well with the more sophisticated MSDETM model. This result suggests the benefits of modeling the dynamics of topic embeddings (α) is smaller than the benefits of modeling the dynamics of the topic prior (η). Accordingly, we focused our qualitative analysis on results from the MixMedia model. We also obtained superior performance over the baseline models in predicting the intervention labels in the GPHIN and WHO datasets (Table 3). This result highlights the advantage of an end-to-end approach of simultaneously inferring topics and predicting labels using the topic mixture as compared to a pipeline approach.

8.2 Qualitative analysis

We compared countries over their related topic coverage using the topic prior mean η_s learned from our best-performing MixMedia models in terms of the topic quality scores. We first labeled the topics based on the top 20 most likely words under each topic (Supplementary Table S1 and Supplementary Table S3).

Here we focused our analysis on the GPHIN news data as the 10 topics capture more interesting contents than the 5 topics learned from the WHO data. As we expected, Canada exhibits high prior over Topic M0 *Canada response to pandemic*, and China on Topic M1 *Chinese outbreak* (Figure 2). Iran and South Korea are high on Topic M3 *COVID testing and medical care*. Germany, France, Italy and Sweden exhibit similarly high preference over Topic M2 *Government response*. United States and UK are high on Topic M9 *Research for treatment*.

Additionally, we also assessed the difference between official news and media news in their preference over the topics. Interestingly, our results suggest that official news announcements focused on *health systems* during January and March and switched to *protective equipment* in April whereas media news outlets focused on *travel restriction* earlier in the pandemic and then switched to focus on *social distancing* in March (Supplementary Fig. 3).

Finally, we examined the learned topic weights associated with the intervention labels in the GPHIN data. These weights were learned by the supervised component in the MixMedia model. We observed meaningful topic correlation with the interventions (Supplementary Fig. 4). In particular, the predicted M0 *social distancing* topic was associated with interventions in mass gathering cancellation, lockdown or curfew, and easing restrictions. Topic M1 *protective equipment* was associated with PPE (personal protective equipment). Topic M3 *travel restriction* was associated with travel advisory, travel ban, and quarantine. Topic M4 *research* was associated with vaccines.

9 Discussion and Conclusion

We developed novel methods for mining media news data about COVID-19 by integrating calendar time, source information (i.e., media source, country), and label information (i.e., NPI) into a unified Bayesian framework. Our goal is to develop a method that can be used in event-based surveillance systems to learn topics indicative of COVID-19 interventions, thereby supporting surveillance of NPI and their effects. Our approach can support the intelligent screening

Dataset	ETM	DETM	MSDETM				MixMedia			
	n/a	n/a	1	2	1†	2†	1	2	1†	2†
WHO (TC)	0.0417	0.0290	0.0014	0.0348	0.0014	0.0325	0.0373	**0.1571**	0.0375	0.1519
	(0.008)	(0.0072)	(0.0005)	(0.0045)	(0.0003)	(0.0087)	(0.0172)	(0.0068)	(0.0068)	(0.0050)
WHO (TD)	0.2192	0.7824	0.8714	**0.9434**	0.8718	0.9162	0.4592	0.7784	0.6120	0.7856
	(0.0096)	(0.0098)	(0.0045)	(0.0089)	(0.0040)	(0.0065)	(0.1307)	(0.0193)	(0.2224)	(0.0147)
WHO (TQ)	0.0091	0.0208	0.0012	0.0318	0.0012	0.0287	0.0174	0.1120	0.0238	**0.1157**
	(0.0018)	(0.0053)	(0.0005)	(0.0041)	(0.0003)	(0.0074)	(0.0103)	(0.0101)	(0.0118)	(0.0036)
GPHIN (TC)	0.1273	0.0759	0.0335	0.0358	0.0334	0.0369	**0.1377**	0.1352	0.1362	0.1329
	(0.0049)	(0.0091)	(0.0032)	(0.0071)	(0.0032)	(0.0053)	(0.0057)	(0.0066)	(0.0063)	(0.0076)
GPHIN (TD)	0.7712	0.7687	**0.9493**	0.9338	0.9186	0.9363	0.7856	0.7760	0.7792	0.7704
	(0.0119)	(0.0327)	(0.0054)	(0.0063)	(0.0031)	(0.0098)	(0.0182)	(0.0113)	(0.0212)	(0.0119)
GPHIN (TQ)	0.0770	0.0748	0.0309	0.0330	0.0307	0.0340	**0.0967**	0.0946	0.0947	0.0910
	(0.0043)	(0.0070)	(0.0029)	(0.0065)	(0.0029)	(0.0049)	(0.0048)	(0.0059)	(0.0059)	(0.0059)

Table 2. Topic quality. TC: topic coherence; TD: topic diversity; TQ: topic quality. Source 1: countries or organizations. Source 2: media versus official announcements. †Using intervention labels.

Dataset/Source	ETM+LR	DETM+LR	MSETM		MixMedia	
	n/a	n/a	1	2	1	2
Top-3 Recall						
WHO (mean)	0.1247	0.2557	0.0965	0.0577	**0.5328**	0.4135
WHO (std)	0.0327	0.0527	0.1094	0.0310	0.0335	0.0698
GPHIN (mean)	0.4801	0.4485	0.1906	0.1508	0.5502	**0.6073**
GPHIN (std)	0.1018	0.0831	0.0532	0.0250	0.0822	0.0143
Top-5 Recall						
WHO (mean)	0.2460	0.3872	0.1465	0.1098	**0.6947**	0.5707
WHO (std)	0.0400	0.0482	0.0884	0.0277	0.0275	0.0755
GPHIN (mean)	0.6279	0.6067	0.3470	0.2602	0.7038	**0.7506**
GPHIN (std)	0.1143	0.0862	0.0798	0.0369	0.0764	0.0130
Top-10 Recall						
WHO (mean)	0.6103	0.6587	0.2888	0.2722	**0.9316**	0.8295
WHO (std)	0.0564	0.0670	0.1618	0.0409	0.0109	0.0416
GPHIN (mean)	0.8676	0.8455	0.6587	0.6389	0.9089	**0.9434**
GPHIN (std)	0.0887	0.0695	0.0516	0.0600	0.0489	0.0091

Table 3. Intervention prediction performance measured by top-k recall (k=3, 5, 10). LR: linear regression; Source 1: countries or organizations; Source 2: media news and official news types. The top-k recall is presented as mean and standard deviation in two rows.

and analysis of media reports for monitoring COVID-19 interventions to inform policy evaluation and assist in understanding the progression of the epidemic.

Basic topic models were not capable of capturing the granularity of region-specific and time-dependent topical events, such as NPI in response to COVID-19. By modeling the source-specific topic prior, our method demonstrated superior performance over the state-of-the-art topic modeling approaches in terms of topic quality. Our model was also interpretable and gave insights into the interventions to control COVID-19 implemented by different countries and organizations. To the best our knowledge, previous studies on mining media news data using topic models have conducted a post-hoc analysis by first running LDA on the documents without the source or time information and subsequently analyzed the identified topics with the other information in an ad hoc manner. In contrast, our model automated this process by incorporating the source and time information into the model training in the form of a Bayesian prior. Furthermore, by training an end-to-end semi-supervised model, we observed superior accuracy in predicting NPIs, which were meaningfully characterized by the inferred topics.

This article is focused on the novel machine learning aspects of our research. One caveat for our studies is that for our model to be properly trained and useful in practice, we need good timely measurement of interventions in place. In

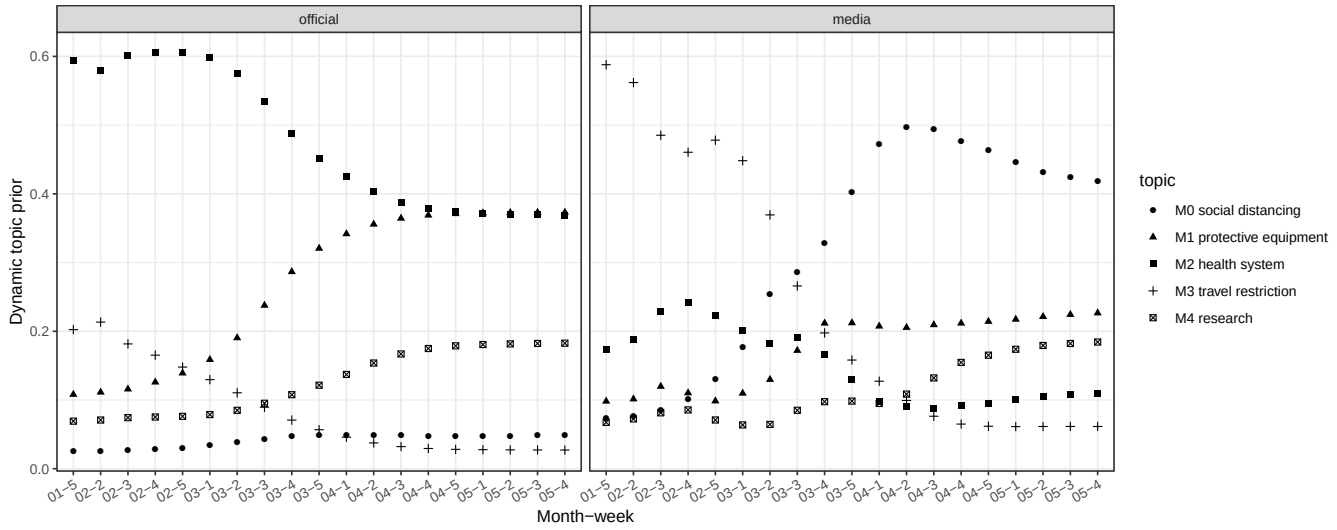

Figure 3. Source-specific dynamic topic priors on the GPHIN data. The dynamic topic priors are displayed over weeks for official news and media news. The 5 topics are labeled based on Supplementary Table S2.

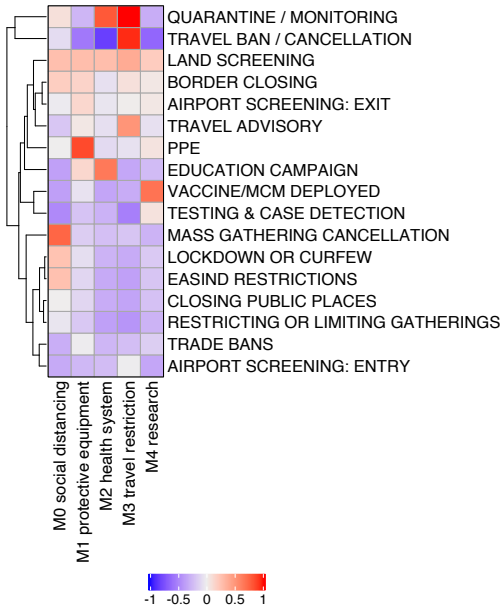

Figure 4. Predicted topic weights of interventions on the GPHIN data. MixMedia was trained with prediction component with the source information set to official and media types. The linear weights of the 5 topics associated with the intervention labels in the GPHIN data were displayed as a heatmap. The 5 topics were labeled based on the top 20 words in Supplementary Table S2.

future work, we will conduct a deeper analysis of the epidemiological aspects of interventions and responses. We will also explore the benefits of incorporating into our model other surveillance data such as the number of confirmed infected cases and deaths and analyze the impacts of region-specific interventions. We believe that our approach is a significant step towards automated intelligent systems for global disease surveillance.

In conclusion, global variation in the use of NPI has likely contributed to the variation in the burden of the pandemic between countries. Clear evidence regarding the effectiveness and optimal use of NPI to control COVID-19, would allow NPI to be selected and implemented for maximum effect, potentially preventing thousands of infections and deaths as the pandemic continues. It is paramount to learn from our early experience around the globe to devise effective strategies for deploying public health interventions to control COVID-19. A methodological foundation for such learning is provided by our results and we intend to continue refining and applying this approach to support the public health response to COVID-19.

Acknowledgments

This work is supported by CIHR through the Canadian 2019 Novel Coronavirus (COVID-19) Rapid Research Funding Opportunity (Round 1), (Application number: 440236).

References

[1] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B Munroe, Bina Joe, and Xi Cheng. 2020. Artificial intelligence and machine learning to fight COVID-19.

[2] Philippe Barboza, Laetitia Vaillant, Yann Le Strat, David M Hartley, Noele P Nelson, Abla Mawudeku, Lawrence C Madoff, Jens P Linge, Nigel Collier, John S Brownstein, et al. 2014. Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PloS one* 9, 3 (2014).

[3] Arnab Bhadury, Jianfei Chen, Jun Zhu, and Shixia Liu. 2016. Scaling up dynamic topic models. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 381–390.

[4] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. 113–120.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3 (March 2003), 993–1022.

[6] Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545* (2019).

[7] Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907* (2019).

[8] M Dion, P AbdelMalik, and A Mawudeku. 2015. Big Data: Big Data and the Global Public Health Intelligence Network (GPHIN). *Canada Communicable Disease Report* 41, 9 (2015), 209.

[9] Nikou Günnemann, Michael Derntl, Ralf Klamma, and Matthias Jarke. 2013. An interactive system for visual analytics of dynamic topic models. *Datenbank-Spektrum* 13, 3 (2013), 213–223.

[10] M D Hoffman, D M Blei, C Wang, and J W Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research (JMLR)* (2013).

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *arXiv.org* (Dec. 2013). arXiv:1312.6114v10 [stat.ML]

[13] Jon D McAuliffe and David M Blei. 2008. Supervised Topic Models. In *Advances in Neural Information Processing Systems 20*, J C Platt, D Koller, Y Singer, and S T Roweis (Eds.). Curran Associates, Inc., 121–128.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[15] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262–272.

[16] Yotam Ophir. 2018. Coverage of epidemics in American newspapers through the lens of the crisis and emergency risk communication framework. *Health security* 16, 3 (2018), 147–157.

[17] William Poirier, Catherine Ouellet, Marc-Antoine Rancourt, Justine Béchard, and Yannick Dufresne. 2020. (Un)Covering the COVID-19 Pandemic: Framing Analysis of the Crisis in Canada. *Canadian Journal of Political Science/Revue canadienne de science politique* (April 2020), 1–7.

[18] R Ranganath, S Gerrish, D Blei Artificial Intelligence Statistics, , and 2014. [n.d.]. Black box variational inference. *jmlr.org* ([n. d.]).

[19] Dror Walter and Yotam Ophir. 2019. News Frame Analysis: An Inductive Mixed-method Computational Approach. *Communication Methods and Measures* 13, 4 (2019), 248–266.

[20] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature* 579, 7798 (2020), 270–273.

A Supplementary Information

The generative process of the MSDETM model described in the main text (Section 4; Fig. 1a) is as follows:

1. Draw the initial topic proportion mean $\eta_{sk}^{(0)} \sim \mathcal{N}(0, I)$ for $s = 1, \ldots, S$ and $k = 1, \ldots, K$
2. Draw the initial topic embedding $\alpha_k^{(0)} \sim \mathcal{N}(0, I)$ for $k = 1, \ldots, K$
3. For time step $t = 1, \ldots, T$,
   a. Draw topic proportion mean $\eta_{sk}^{(t)} \sim \mathcal{N}(\eta_{sk}^{(t-1)}, \delta^2)$ for $s = 1, \ldots, S$ and $k = 1, \ldots, K$
   b. Draw topic embedding $\alpha_k^{(t)} \sim \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2)$ for $k = 1, \ldots, K$
4. For each document $d = 1, \ldots, D$ in the corpus,
   a. Draw a topic proportion for document d from logistic normal $\theta_d \sim \mathcal{LN}(\eta_{s_d}^{(t_d)}, \delta^2 I)$
   b. For each token n in document d,
      i. Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$
      ii. Draw word $w_{dn} \sim \text{softmax}(\rho^T \alpha_{z_{dn}}^{(t_d)})$
   c. Draw document label $\mathbf{y}_d \sim \text{softmax}(\mathbf{W}^\top \theta_d)$

B Supplementary Tables

B.1 GPHIN topics

B.2 WHO topics

Topic Label	Top 20 words
M0 Canada response to pandemic	canada, health, covid, care, government, canadian, province, pandemic, public, ontario, workers, canadians, minister, people, federal, long, april, term, staff
M1 Chinese outbreak	china, chinese, epidemic, coronavirus, medical, control, february, wuhan, prevention, province, outbreak, city, central, beijing, hubei, commission, health, national, measures
M2 Government response	coronavirus, march, outbreak, due, minister, government, april, spread, week, announced, prime, state, emergency, president, scheduled, cancelled, amid, world, events
M3 COVID testing and medical care	coronavirus, cases, people, virus, confirmed, health, tested, positive, symptoms, reported, city, number, country, covid, authorities, infected, home, february, quarantine
M4 Public health guidelines to reduce transmission	health, covid, public, cdc, guidance, risk, disease, transmission, countries, response, information, measures, care, states, cases, community, pandemic, control, report
M5 Protective equipment	masks, million, medical, covid, equipment, pandemic, supplies, protective, countries, support, food, response, face, global, supply, world, products, health, united
M6 International travel restrictions	china, passengers, travel, flights, days, quarantine, february, nationals, korea, countries, italy, iran, japan, south, foreign, hong, kong, apply, enter
M7 Public health/medical measures	health, ministry, coronavirus, country, cases, public, measures, covid, national, hospital, case, isolation, suspected, medical, hospitals, screening, care, minister, virus
M8 Social distancing & business lockdown	april, announced, measures, public, restrictions, lockdown, authorities, government, march, remain, closed, covid, place, country, essential, state, spread, extended, allowed
M9 Research for treatment & vaccine	covid, patients, coronavirus, research, vaccine, clinical, study, test, drug, virus, researchers, sars, disease, testing, cov, published, university, treatment, tests

Table S1. Top 20 words from the 10 topics learned from the **GPHIN** data using the best-performing MixMedia model. MixMedia was trained with **countries** as source information η_s but **without using time and label** information.

Topic Label	Top 20 words
M0 social distancing	april, announced, march, coronavirus, government, measures, public, restrictions, authorities, lockdown, closed, spread, state, remain, people, country, covid, place, schools
M1 protective equipment	covid, health, pandemic, coronavirus, medical, masks, government, million, support, canada, response, outbreak, workers, equipment, countries, protective, supplies, food, world
M2 healthcare system	health, covid, cases, coronavirus, public, ministry, people, confirmed, control, risk, prevention, case, measures, disease, care, february, contact, virus, hospital
M3 travel restriction	china, passengers, coronavirus, february, travel, days, quarantine, flights, march, italy, south, iran, korea, countries, nationals, foreign, citizens, japan, hong
M4 research	covid, coronavirus, patients, virus, test, tests, study, research, testing, clinical, vaccine, drug, disease, researchers, published, data, sars, found, people

Table S2. Top 20 words from the 5 topics learned from the **GPHIN** data using the best-performing MixMedia model. MixMedia was trained with source information as **media or official** using both **time and label** information.

Topic Label	Top 20 words
M0 COVID testing and medical care	health, covid, coronavirus, cases, patients, testing, virus, symptoms, hospital, people, ministry, medical, confirmed, disease, contact, test, case, hospitals, february
M1 International travel restrictions	china, days, quarantine, travel, passengers, countries, flights, italy, iran, korea, south, entry, nationals, international, citizens, foreign, country, march, allowed
M2 Social distancing & business lockdown	april, public, closed, people, march, essential, schools, lockdown, home, announced, services, curfew, closure, gatherings, remain, extended, restrictions, government, measures
M3 Protective equipment	covid, government, masks, support, million, medical, health, workers, pandemic, care, equipment, emergency, companies, provide, supplies, protective, work, additional, canada
M4 Government response	coronavirus, spread, march, minister, government, health, measures, outbreak, covid, state, ministry, emergency, country, national, february, due, announced, authorities, city

Table S3. Top 20 words from the 5 topics learned from the **WHO** data using the best-performing MixMedia model. MixMedia was trained with **countries as source information** but without using time and label information.

Topic Label	Top 20 words
M0	coronavirus, people, government, health, april, measures, minister, country, public, ministry, support, million, gatherings, work, closed, state, order, days, sri
M1	covid, emergency, government, national, march, state, health, essential, days, services, ministry, lockdown, allowed, coronavirus, measures, medical, million, border, home
M2	covid, days, people, health, public, closure, border, march, quarantine, government, closed, essential, measures, state, country, lockdown, pandemic, services, home
M3	coronavirus, covid, country, government, april, health, schools, curfew, pandemic, announced, people, million, public, essential, border, spread, ministry, services, state
M4	covid, people, government, public, april, march, curfew, health, allowed, quarantine, movement, gatherings, lockdown, businesses, day, essential, food, home, citizens

Table S4. Top 20 words from the 5 topics learned from the **WHO data** using the best-performing MixMedia model. MixMedia was trained with the source information as **media or official** using **both time and label** information.

Topic Label	Top 20 words
M0	emergency, government, march, april, quarantine, state, covid, including, public, order, minister, schools, measures, home, closure, country, days, support, bank
M1	covid, health, government, state, coronavirus, people, border, emergency, ministry, lockdown, national, country, pandemic, suspended, measures, public, day, allowed, days
M2	people, coronavirus, government, public, covid, health, police, measures, country, march, schools, state, april, million, spread, emergency, essential, pandemic, closed
M3	covid, government, people, public, coronavirus, days, allowed, country, april, quarantine, work, sri, announced, home, lockdown, curfew, period, ministry, essential
M4	health, covid, april, march, days, government, quarantine, public, allowed, services, country, people, curfew, state, essential, food, national, day, closed

Table S5. Top 20 words from the 5 topics learned from the **WHO data** using the best-performing MixMedia model. MixMedia was trained with the source information as **countries** using **both time and label** information.