Article:
Song, X., Petrak, J., Jiang, Y. et al. (Submitted: 2020) Classification aware neural topic model and its application on a new COVID-19 disinformation corpus. arXiv. (Submitted)

Classification Aware Neural Topic Model and its Application on a New COVID-19 Disinformation Corpus

Xingyi Song[1] **Johann Petrak**[1,2], **Ye Jiang**[1]
Iknoor Singh[1,3], **Diana Maynard**[1], **Kalina Bontcheva**[1]
[1]Department of Computer Science, University of Sheffield, Sheffield, UK
[2]Austrian Research Institute for Artificial Intelligence, Vienna, Austria
[3]Panjab University, Chandigarh, India
{x.song, johann.petrak, yjiang18}@sheffield.ac.uk
{i.singh, d.maynard, k.bontcheva}@sheffield.ac.uk

Abstract

The explosion of disinformation related to the COVID-19 pandemic has overloaded fact-checkers and media worldwide. To help tackle this, we developed computational methods to support COVID-19 disinformation debunking and social impacts research. This paper presents: 1) the currently largest available manually annotated COVID-19 disinformation category dataset; and 2) a classification-aware neural topic model (CANTM) that combines classification and topic modelling under a variational autoencoder framework. We demonstrate that CANTM efficiently improves classification performance with low resources, and is scalable. In addition, the classification-aware topics help researchers and end-users to better understand the classification results.

1 Introduction

COVID-19 is not just a global disease pandemic, but has also led to an 'infodemic'[1] (WHO, 2020) and a 'disinfodemic'[2] (Posetti and Bontcheva, 2020). The increased volume (Brennen et al., 2020) of COVID-19 related disinformation has already caused public mistrust (Clare and Christie, 2020) and even real-life damage to health and 5G masts.[3]

Consequently fact-checkers and media worldwide are having to triage carefully their limited resources in order to uncover and debunk quickly and effectively the most damaging kinds of COVID-19 disinformation. For example, Brennen et al. (2020) found that most disinformation in the early stage of the pandemic made false claims related to actions and statements by public authorities.

Guided by these needs, we developed an automatic COVID-19 disinformation classifier and made this available for testing and use by professionals at AFP and First Draft. [4]

The challenges of this task are that: 1) there is no sufficiently large existing dataset annotated with COVID-19 disinformation categories, which can be used to train and test machine learning models. 2) Due to the time-consuming nature of manual fact-checking and disinformation categorisation, manual corpus annotation is expensive and slow to create. Therefore the classifier should robustly handle training with low resources. 3) COVID-19 disinformation classification is a fast-moving research area, thus the model should provide suggestions to researchers about relevant categories. 4) The classifier and decisions should be self-explanatory, enabling journalists to understand the rationale for the auto-assigned category.

To address the first challenge, we created a new COVID-19 disinformation classification dataset. The corpus contains disinformation (e.g. false information or misleading tweets) debunked by the CoronaVirusFacts Alliance led by the International Fact-checking Network (IFCN) and has been manually annotated with the categories defined in the most recent social science research on COVID-19 disinformation (Brennen et al., 2020).

For the remaining challenges, we propose a Classification Aware Neural Topic Model (CANTM) which combines the benefits of BERT (Devlin et al., 2019) with a Variational Autoencoder (VAE)(Kingma and Welling, 2013; Rezende et al., 2014) based document model (Miao et al., 2016) to provide:

1. Robust classification performance especially on a small training set – instead of training

[1]"an over-abundance of information" (WHO, 2020)

[2]"the disinformation swirling amidst the COVID-19 pandemic" (Posetti and Bontcheva, 2020)

[3]https://www.bbc.co.uk/news/uk-england-52164358,https://news.sky.com/story/coronavirus-church-ordered-to-stop-selling-bleach-based-covid-19-cure-11975002

[4]https://cloud.gate.ac.uk/shopfront/displayItem/covid19-misinfo

the classifier directly on the original feature representation, the classifier is trained based on generated latent variables from the VAE (Kingma et al., 2014). In this case the classifier has never seen the 'real' training data during the training, thus reducing the chance of over-fitting. Our experiment shows that combining BERT with the VAE framework improves the classification results on a small dataset, and is also scalable to larger datasets.

2. Ability to discover the hidden topics related to the pre-defined classes – the success of the VAE as a topic model[5] has already been proven in previous research (Miao et al., 2016, 2017; Card et al., 2018). We further adapt the VAE-based topic modelling to be classification-aware, by proposing a stacked VAE and introducing the classification information directly in the latent topic generation.
3. The classifier is self-explaining[6] – in CANTM the same latent variable (topic) is used in both the classifier and topic modelling. The topic can be seen as an explanation of the classification model. We further introduce 'class-associated topics' that directly map the topic words to classifier classes. This enables the inspection of topics related to a class, thus providing a 'global' explanation of the classifier.

The main contributions of this paper are: 1) A new COVID-19 disinformation corpus with manually annotated categories. 2) A BERT language model with a VAE topic modelling framework, which shows a performance improvement (over using BERT alone) in a low resource classifier training setting. 3) The CANTM model, which takes classification information into account for topic generation. 4) The use of topic modelling to introduce 'class-associated' topics as a global explanation of the classifier. The corpus and source code of this work will be open-source, and the web service and API will be publicly available.

2 COVID-19 Disinformation Category Dataset

The corpus was created in three stages. Firstly, we collected COVID-19 related debunks of disinformation until 13th April, 2020 published on the IFCN Poynter website[7]. The data has the following fields derived from the published html tags:

- 'Claim': claim of the disinformation, rephrased by the IFCN fact-checker;
- 'Explanation': the explanation of why this is a false claim as provided by the fact checkers;
- Source link': link to original page of the debunk, as published on the fact-checking organisation's website;
- 'Date': date of publication on IFC Poynter website.

Due to the language restrictions of our human annotators, we could only focus on debunks in English. Thus we applied a language detector[8] over the source of the debunk and filtered out all non-English debunks automatically. In total, 1,480 debunked claims remained.

Category
Public authority (PubAuthAction)
Community spread and impact (CommSpread)
Medical advice, self-treatments, and virus effects (GenMedAdv)
Prominent actors (PromActs)
Conspiracies (Consp)
Virus transmission (VirTrans)
Virus origins and properties (VirOrgn)
Public Reaction (PubRec)
Vaccines, medical treatments, and tests (Vacc)
Cannot determine (None)

Table 1: Categories for annotation, the abbreviations are in the parentheses

The next stage involved manual annotation, where an adapted version of Label Studio[9] was used as a web-based annotation tool. The claim, explanation and source link were all provided to the annotators, who assigned to each text the most relevant one of 10 COVID-19 disinformation categories (see Table 1) and indicated their confidence (from 0-9) in their decision. Originally, these categories were proposed in a recent social science analysis of a small sample of 225 debunks(Brennen et al., 2020). We adopted them unchanged, except for widening their 'Public preparedness' category to become 'Public Reaction' and to include also disinformation about public protests and other civil disobedience which are a more recent phenomenon. In addition, we added a new category 'Cannot determine (None)' to enable annotators to flag cases of COVID-19 disinformation that did not fit any of the other categories.

[5] Some researchers distinguish 'document model' from 'topic model' (Miao et al., 2017; Korshunova et al., 2019). For simplicity, we consider both as a topic model.

[6] BERT attention weights could also be treated to explain the decision, but this is outside the scope of this paper.

[7] https://www.poynter.org/ifcn-covid-19-disinformation/

[8] https://pypi.org/project/langdetect/

[9] https://github.com/heartexlabs/label-studio

We recruited 27 volunteers for the annotation, and randomly split the data into batches of 20 debunks. In the first round, all annotators worked on unique batches. In the second round, annotators received randomised debunks from the first round, which were then used to measure inter-annotator agreement (IAA) on COVID-19 disinformation classification.

The exercise produced 2,192 classified debunks (see Table 2). Amongst these, 424 samples were double- or multiple-annotated, from which we calculate the IAA. At this stage, vanilla Cohen's Kappa (Cohen, 1960) was only 0.46.

To increase the data quality and provide a good training sample for our ML model, we applied a cleaning step to filter the annotations. We first measured annotator quality by observing agreement change when removing an (anonymous) annotator. This annotator quality was scored based on the magnitude of score variance. Based on this, we then removed annotations from the two annotators with the lowest scores.

We also measured the impact of the annotator confidence score on the annotation agreement and the amount of filtered data, and set a confidence threshold for each annotator, based on the quality check from the first round (for most annotators, this threshold was 6). Any annotation below this threshold was filtered out.

Finally, 1,293 debunks remained with at least one reliable classification, and IAA was boosted to 0.7336 (in percentage) and Cohen's Kappa to 0.7040.

	All	Cleaned
Single Annotated	1056	1038
Double Annotated	213	186
Multiple Annotated	211	69
Annotation Agreement	0.5145	0.7336
Kappa	0.4660	0.7040

Table 2: Label counts and annotation agreements of unfiltered annotation (All) and filtered annotation (Cleaned)

The final dataset was produced by merging the multiple-annotated debunks on the basis of: 1) majority agreement between the annotators where possible; 2) confidence score – if there is no majority agreement, we use the highest confidence score. Table 3 shows the statistics of the merged dataset in each category. The category distribution is consistent with that found in Brennen et al. (2020).

PubAuthAction	CommSpread	PubRec	PromActs
251	225	60	221
GenMedAdv	VirTrans	Vacc	Consp
177	80	76	97
VirOrgn	None		
63	43		

Table 3: Label count after merge in each category

3 Model

In this section, we review some related work, using this to explain the motivation for our model. Then we describe our CANTM model in Section 3.2. Other related work is reviewed in Section 5.

3.1 Background and Preliminaries

Miao et al. (2016) introduce a generative neural variational document model (NVDM) that models the document (x) likelihood $p(x)$ using a variational autoencoder (VAE), which can be described as:

$$\log p(x) = ELBO + D_{KL}(q(z|x)||p(z|x))$$
$$ELBO = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)) \quad (1)$$

Where $p(z)$ is the prior distribution of latent variable z. $q(z|x)$ is the inference network (encoder) used to approximate the posterior distributions $p(z|x)$. $p(x|z)$ is the generation network (decoder) to reconstruct the document based on latent variable (topics) $z \sim q(z|x)$ sampled from the inference network.

According to Equation 1, maximising the ELBO (evidence lower bound) is equivalent to maximising the $p(x)$ and minimising the difference between $q(z|x)$ and $p(z|x)$. Therefore, maximising ELBO will be the objective function in the NVDM or VAE framework, or negative ELBO for gradient descent optimisation. The latent variable z then can be treated as the latent topics of the document.

NVDM is an unsupervised model, hence we have no control on the topic generation. In order to uncover the topics related to the target y (e.g. category, sentiment or coherence) in which we are interested, we can consider several previous approaches. The Topic Coherence Regularization (NTR) (Ding et al., 2018) applies the topic coherence as additional loss (i.e. loss $\mathcal{L} = -ELBO + C$) to regularise the model to generate more coherent topics. SCHOLAR (Card et al., 2018) directly inserts the target information into the encoder (i.e. $q(z|x, y)$), making the latent variable also dependent on the target. However, when target information is missing at application time, SCHOLAR treats the target

input as a missing feature (i.e. all zero vector) or all possible combinations. Hence the latent variable becomes less dependent on the target.

Inspired by the stacked VAE of Kingma et al. (2014), we combined ideas from NTR and SCHOLAR. We stacked a classifier-regularised VAE (M1) and a classifier-aware VAE (M2) enabling the provision of robust latent topic information even at testing time without label information.

3.2 Classifier Aware Neural Topic Model (CANTM)

The training sample $D = (x, x_{bow}, y)$ is a triple of the BERT word-pieces sequence representation of the document (x), a bag-of-words representation of the document (x_{bow}) and its associate target label y.

The general architecture of our model is illustrated in Figure 1. CANTM is a stacked VAE containing 6 sub-modules:

1. M1 encoder (or M1 inference network) $q(z|x)$
2. M1 decoder (or M1 generation network) $p(x_{bow}|z)$
3. M1 Classifier $\hat{y} = f(z)$
4. M1 Classifier decoder $p(x|\hat{y})$
5. M2 encoder (or M2 inference network) $q(z_s|x, \hat{y})$
6. M2 decoder (or M2 generation network) $p(x_{bow}|\hat{y}, z_s)$ and $p(\hat{y}|z_s)$

Sub-modules 1 and 2 implement a VAE similar to NVDM. The modification over original NVDM is that instead of bag-of-words (x_{bow}) input and output to the model, our input is a BERT word-pieces sequence representation of the original document (x). The reason for this modification is that x can be seen as a grammar-enriched x_{bow}, and we could capture better semantic representation in the hidden layers (e.g. though pre-trained BERT) and benefit the classification and topic generation. Also, $q(z|x)$ is an approximation of $p(z|x_{bow})$, and they do not have to follow the same condition (Kingma and Welling, 2013), as our model is still under the VAE framework. Sub-modules 5 and 6 implement another VAE that models the joint probability of document x_{bow} and label $\hat{y}$. Note that the label in M2 is a classifier prediction, hence this label information will always be available for M2 VAE. To apply CANTM to unlabelled test data, we fix the M1 weights that are pre-trained with labelled data, and only train the M2 model. In Sections 3.2.1 to 3.2.5, we will describe the detail of each sub-module.

3.2.1 M1 Encoder

The M1 encoder is illustrated in the yellow part of Figure 1. During the encoding process, the input x is first transformed into a BERT-enriched representation h using a pre-trained BERT model. We use the CLS token output from BERT as h. Then linear transformations $l_1(h)$ and $l_2(h)$ transform the h into parameters of variational distribution that are used to sample latent variable z. The variational distribution is a Gaussian distribution ($\mathcal{N}(\mu, \sigma)$) The M1 Encoder is represented in Equation 2

$$\begin{aligned} q(z|x) &= \mathcal{N}(\mu, \sigma) \\ \mu &= l_1(h), \sigma = l_2(h) \\ h &= BERT(x) \end{aligned} \tag{2}$$

Following previous approaches (Rezende et al., 2014; Kingma and Welling, 2013; Miao et al., 2016), a re-parameterisation trick is applied to allow back-propagation to go though the random node.

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, 1) \tag{3}$$

where ϵ is random noise sampled from a 0 mean and variance 1 Gaussian distribution. In the decoding process (next section), the document is reconstructed from latent variable z, hence z can be considered as the document topic.

3.2.2 M1 Decoder

The decoding process (red part in Figure 1) is to reconstruct x_{bow} from latent variable z. This is modelled by a fully connected feed-forward (FC) layer with softmax activation (sigmoid activation normalised by softmax function. For the rest of the paper we will describe this as softmax activation for simplicity). The likelihood of the reconstruction $p(x_{bow}|z)$ can be calculated by

$$p(x_{bow}|z) = softmax(zR + b) \odot x_{bow}$$

Where $R \in \mathbb{R}^{|z| \times |V|}$, and $|V|$ is the vocabulary size. R is a learnable weight for mapping between topics and words. The topic words for each topic can be extracted according to this weight. $\odot$ is the dot product.

3.2.3 M1 Classifier and Classifier Decoder

The classifier $\hat{y} = softmax(FC(z))$ is a softmax activated FC layer. It is based on the same latent variable z from the M1 encoder. Since the M1 VAE and classifier are jointly trained based on z, it can be seen as a 'class regularized topic' and also serve

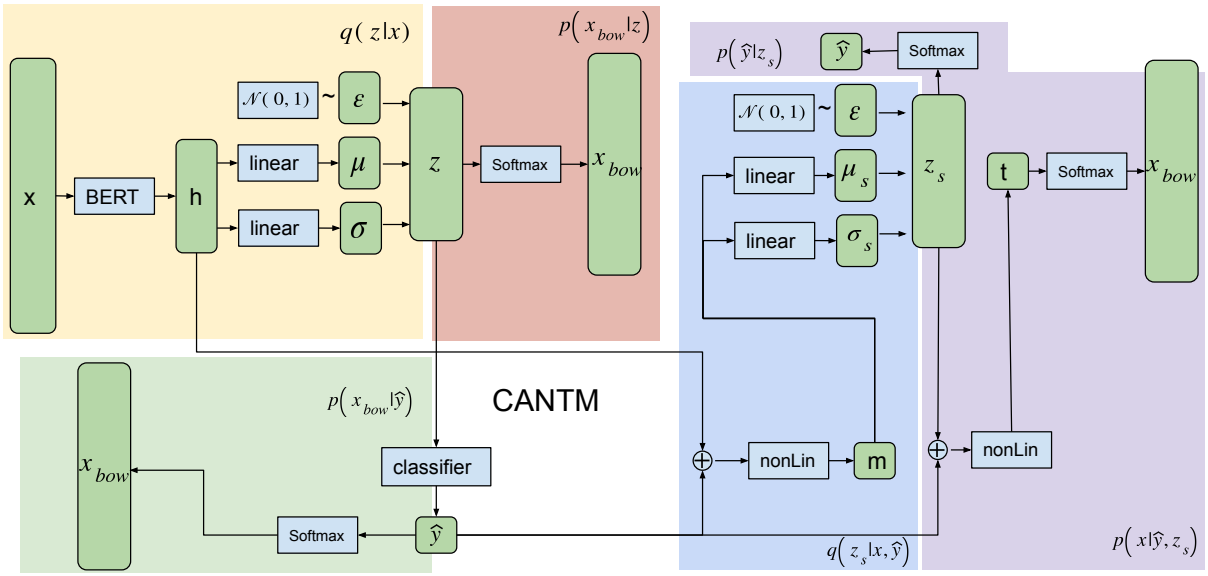

Figure 1: Overview of model architecture, Linear is the linear transformation (i.e. Linear(x)=xW+b), nonLin is linear transformation with non-linear activation function f(Linear(.)), Softmax is Softmax activated linear function

as a 'global explanation' of the classifier. Furthermore, $\hat{y}$ itself can be seen as a compressed topic of z, or 'class-associated topic'. The document can be reconstructed by $\hat{y}$ in the same way as the M1 decoder, and the likelihood of $p(x_{bow}|\hat{y})$ is given by:

$$p(x_{bow}|\hat{y}) = softmax(\hat{y}R_{ct} + b) \odot x_{bow}$$

Where $R_{ct} \in \mathbb{R}^{|y| \times |V|}$ is a learnable weight for 'class-associated topic' word mapping.

3.2.4 M2 Encoder

The encoding process of M2 (blue part in Figure 1) is similar to M1, but instead of only encoding x, M2 encodes both the document and predicted label from the M1 classifier $q(z_s|x, \hat{y})$. In the M2 encoder process, we first concatenate ($\oplus$) the BERT representation h and predicted label $\hat{y}$, then merge them through a leaky rectifier ($LRelu$)(Maas et al., 2013) activated FC layer. We refer to this as $nonLin$ in the remainder of the paper.

$$\begin{aligned} m &= nonLin(h \oplus \hat{y}) \\ &= LRelu(FC(h \oplus \hat{y})) \end{aligned}$$

As for the M1 encoder, a linear transformation then maps the merged feature m to the parameters of the variational distribution represented by the latent variable of M2 model z_s. The variational distribution is a Gaussian $\mathcal{N}(\mu_s, \sigma_s)$:

$$\begin{aligned} q(z_s|x, \hat{y}) &= \mathcal{N}(\mu_s, \sigma_s) \\ \mu_s = l_3(m), \sigma_s &= l_4(m) \end{aligned}$$

3.2.5 M2 Decoder

The decoding process of M2 $p(x_{bow}, \hat{y}|z_s)$ is divided into two decoding steps ($p(x_{bow}|\hat{y}, z_s)$ and $p(\hat{y}|z_s)$) by Bayes Chain Rule.

The step $p(\hat{y}|z_s)$ can be considered as M2 classifier, calculated by softmax FC layer, the likelihood function is modelled as $p(\hat{y}|z_s) = softmax(FC(z_s)) \odot \hat{y}$. The M2 classifier will not be used for classification in this work, only for the loss calculation (see Section 3.2.6).

In step $p(x_{bow}|\hat{y}, z_s)$, we first merge $\hat{y}$ and z_s using $nonLin$ layer $t = nonLin(\hat{y} \oplus z_s)$ where t is a 'classification aware topic'. Then x_{bow} is reconstructed using a softmax layer. The likelihood function is:

$$p(x|\hat{y}, z_s) = softmax(tR_s + b) \odot x_{bow}$$

where $R_s \in \mathbb{R}^{|z_s| \times |V|}$ is a learnable weight for the 'classification aware topic' word mapping.

3.2.6 Loss Function

The objective of CANTM is to: 1) maximise $ELBO_{x_{bow}}$ for M1 VAE; 2) maximise $ELBO_{x_{bow},\hat{y}}$ for M2 VAE; 3) minimise cross-entropy loss $\mathcal{L}_{cls}$ for M1 classifier and 4) maximise the log likelihood of M1 class decoder $\log[p(x_{bow}|\hat{y})]$. Hence the loss function[10] for

[10]For full details of the ELBO term deriving process please see Appendix C

CANTM is

$$\begin{aligned}\mathcal{L} &= \lambda\mathcal{L}_{cls} - ELBO_{x_{bow}} - ELBO_{x_{bow,\hat{y}}} \\ &\quad - \mathbb{E}_{\hat{y}}[\log p(x_{bow}|\hat{y})] \\ &= \lambda\mathcal{L}_{cls} - \mathbb{E}_{z}[\log p(x_{bow}|z)] + D_{KL}(q(z|x)||p(z)) \\ &\quad - \mathbb{E}_{z_s}[\log p(x_{bow}|\hat{y}, z_s)] - \mathbb{E}_{z_s}[\log p(\hat{y}|z_s)] \\ &\quad + D_{KL}(q(z_s|x, \hat{y})||p(z_s)) - \mathbb{E}_{\hat{y}}[\log p(x_{bow}|\hat{y})]\end{aligned}$$

where $p(z)$ and $p(z_s)$ are zero mean diagonal multivariate Gaussian priors ($\mathcal{N}(0, I)$), $\lambda = vocabSize/numclass$ is a hyperparameter controlling the importance classifier loss.

4 Experiments

In this section, we evaluate the performance of CANTM on both COVID-19 disinformation classification and topic modelling (with 50 topics). Three experiments are presented. We first compare the performance of CANTM against baseline approaches on the COVID-19 corpus (Section 4.1); then we apply CANTM to the IMDB sentiment corpus (Maas et al., 2011) to test its compatibility with other tasks with larger data (Section 4.3.1); finally, in Section 4.3 we discuss topic interpretability by visualising the topic words.

We compare to the following: BERT, SCHOLAR, NVDM, and LDA. The settings of CANTM and baselines are:

- BERT (Devlin et al., 2019): We use Huggingface[11] (Wolf et al., 2019) 'BERT-based-uncased' pre-trained model and Pytorch implementation in this experiment. As with CANTM, we use BERT [CLS] output as BERT representation, and an additional 50 dimensional feed-forward hidden layer (with leaky ReLU activation) after that.[12] Only the last transformer encoding layer (layer 11) is unlocked for fine-tuning, the rest of the BERT weights were frozen for this experiment. The Pytorch[13] implementation of the Adam optimiser (Kingma and Ba, 2014) is used in the training with default settings. The batch size for training is 32. All BERT-related (CANTM, NVDMb) implementations in this paper follow the same settings.
- CANTM (our proposed method): We use the same BERT implementation and settings as described above. The sampling size (number of samples z and z_s drawn from the encoder) in training and testing are 10 and 1 respectively, and we only use expected value (μ) of $q(z|x)$ for the classification at testing time. Unless mentioned otherwise, the topics reported from CANTM are 'classification-aware'.
- NVDM (Miao et al., 2016): We re-implement NVDM[14], with two versions: 1) original NVDM as described in (Miao et al., 2016) ("NVDMo" in the results); 2) NVDM with BERT representation ("NVDMb" in the results).
- SCHOLAR (Card et al., 2018): We use the original author implementation[15] with all default settings (except the vocabulary size and number of topics).
- Latent Dirichlet Allocation (LDA) (Blei et al., 2003): the Gensim (Řehůřek and Sojka, 2010) implementation is used.

All bag-of-words inputs are pre-processed using the script publicly available from Card et al. (2018).[15] The vocabulary sizes are 2000 for the COVID-19 set and 5000 for the IMDB set (consistent with (Card et al., 2018) to make a fair comparison) based on word counts from each set.

4.1 COVID-19 Disinformation Classification

In this experiment, the input text for each instance is the combination of the Claim and the Explanation (the average text length is 23 words). The results are reported based on 5-fold cross validation. Since class distribution is imbalanced, we report the macro F-1 measure (F-1)[16] and accuracy (Acc.) for the classification task. For the topic modelling task, the metrics reported are perplexity (Perp.) and non-negative point-wise mutual information (NPMI (Chang et al., 2009; Newman et al., 2010)). As in previous work (Miao et al., 2016; Card et al., 2018), the perplexity is estimated by ELBO, and NPMI scores were calculated based on the top 10 topic words of each topic.

[11] https://github.com/huggingface/transformers

[12] CWNTM contains a sampling layer after the BERT representation, this additional layer is added for fair comparison. Please check Appendix E on impact of the additional hidden layer

[13] https://pytorch.org/

[14] Based on code at https://github.com/YongfeiYan/Neural-Document-Modeling

[15] Using code from https://github.com/dallascard/scholar

[16] The F-1 is calculated as the average F-1 measure of all classes, please refer to Appendix E for the class level F-1 score.

	Acc.	F-1	Perp.	NPMI
Bert	58.78	54.19	n/a	n/a
SCHOLAR	48.17	36.40	2947	**0.25**
NVDMb	n/a	n/a	1084	0.09
NVDMo	n/a	n/a	781	0.08
LDA	n/a	n/a	8518	0.12
CANTM	**63.34**	**55.48**	**749**	0.14

Table 4: COVID-19 disinformation results, n/a stands for not applicable for the model

The COVID-19 evaluation results are shown in Table 13. BERT as a strong baseline outperforms SCHOLAR in accuracy by more than 10% and almost 18% F-1 measure. This is expected, because BERT is a discriminative model pre-trained on large corpora and with a much more complex model structure than SCHOLAR. Our model CANTM shows almost 5% increase in accuracy and more than 1% F-1 further improvement over BERT. Training on latent variables with multi-task loss is thus an efficient way to train on a small dataset even with a pre-trained embedding/language model.

In the topic modelling task, using BERT in NVDM has better topic coherence than the vanilla NVDM, but also increases the perplexity. LDA has high perplexity in the COVID-19 experiment, which may be because of the relatively small dataset and short document length (average 19 words after pre-processing and vocabulary filtering), but LDA still has relatively better topic coherence than both NVDM versions. CANTM has the best perplexity performance, while SCHOLAR has the best coherence score. It is very difficult to draw conclusions from the topic modelling task performance; in Section 4.3 we will discuss the lack of correlation between topic interpretability and topic coherence.

4.2 IMDB Sentiment Experiment

The IMDB sentiment corpus contains 50,000 movie reviews annotated with positive and negative sentiment. The number of positive and negative labels is balanced in this corpus (25,000 positive, 25,000 negative, average document length is 282 words). We use the original train-test split for evaluation, and report the results on the test set only. All settings including vocabulary size and pre-processing steps exactly follow Card et al. (2018). Hence the NVDM, LDA and SCHOLAR results are as reported in Card et al. (2018).

The IMDB results are shown in Table 5. The classification results are consistent with the COVID-19 experiment. Baseline BERT has better accuracy than SCHOLAR, and CANTM further improves over BERT by about 0.5%. The topic modelling performance of CANTM is almost the same as for NVDM, while LDA and SCHOLAR have the best performance in perplexity and coherence, respectively.

Metrics	Accuracy	Perplexity	Coherence
NVDM*	n/a	1748	0.06
LDA*	n/a	**1508**	0.13
SCHOLAR*	87	1905	**0.14**
BERT	89.54	n/a	n/a
CANTM	**90.00**	1786	0.06

Table 5: IMDB results, n/a stands for not applicable for the model (*NVDM, LDA and Scholar results are borrowed from Card et al. (2018))

4.3 Topic Interpretability Discussion

CANTM 0.50	please patents link ecuador patent read click full article guayaquil
CANTM 0.04	cure proven met protection leader pope aajtak within elizabeth developed
SCHOLAR 0.58	article link read please click full ecuador cases guayaquil ecuadorian
SCHOLAR 0.07	coronavirus story lab china created website similar general chinese director

Table 6: Topic words of the best and worst coherence topics NPMI score in parentheses

Table 14 shows the topics from CANTM and SCHOLAR with the best and worst NPMI scores. We found there is no strong correlation between coherence score and topic interpretability in supervised topic models.[17] The SCHOLAR topics are included here to demonstrate this is not just the case in CANTM. With knowledge of the predefined classes (see Table 1), the lowest coherence topic (Row 2, CANTM 0.04) in Table 14 can be easily interpreted as a mixture of the topics 'General Medical advice' and 'Prominent actors', while the highest two topics (CANTM 0.50 and SCHOLAR 0.58) are more general words appearing in the text.

CANTM 0.09	worst waste like boring lousy wasted lame sucks bottom tedious
CANTM 0.03	animation movie enjoy better film disney time acting recommend make

Table 7: The IMDB topics from the two best and worst topic coherence scores

Table 7 shows the CANTM-generated IMDB topics. We select two topics, based on the best and worst topic coherence score. Since IMDB is a sentiment-labelled data set, we can clearly see that the topics generated here are the sentiment

[17] For a full comparison to NVDM and LDA and discussion, please check Appendix E

and aspect words. Row 1 is the topic related to negative sentiment. Row 2 shows the topics related to positive sentiment in animation movies. Again, the lowest coherence CANTM topic is still highly interpretable.

4.3.1 Class-Associated Topics

In Section 3.2.3 we discussed the Class-Associated Topics, which can be used to visualise the word distribution in the training data associated with the pre-defined classes. Table 8 shows an example of topic words of class-associated topics. As the topics are guided by the classifier, the topic words are strongly associated with the pre-defined classes, and can be used to discover concepts related to the classes. For example, temperature (topic word 'hot' in GenMedAdv) is one of the most connected concepts to GenMedAdv. In addition, this feature could be potentially used to check the biases of the trained classifier.

Vacc	cure vaccine new covid developed novel scientists claimed coronavirus claims
VirOrgn	coronavirus shows video bat wuhan novel source outbreak virus taken
GenMedAdv	coronavirus cure experts novel prevent water evidence kill scientific hot
Consp	coronavirus chinese covid virus lab wuhan outbreak new china posts

Table 8: Top 10 class topic words for Vaccines(Vacc), Medical advice(GenMedAdv), General Medical advice (GenMedAdv) and Conspiracies (Consp)

4.3.2 CANTM with Unlabelled COVID-19 Disinformation

To test the CANTM with unlabelled data, we collected further 4587 COVID-19 debunks (until 26th May 2020) from IFCN (the same collection as described in Section 2). In the training, we reuse the pre-trained M1 model (with labelled data), and only train M2 model with $-ELBO_{x_{bow},\hat{y}}$ loss. Table 9 shows the example classification-aware topics trained with newly collected data. We can clearly see these topics are still classification-related even without labels. (Row 1: virus transmission; Row 2:Public authority; Row 3:Medical advice and Row 4:Conspiracies)

police bat cdc spread visit tourists answer data july clip
mention conference professor since april quarantined starting spoke supporting please
health ever salt swat ginger pope uses welfare hands singapore
people vaccine since weapon hospital scientific man group cells working

Table 9: COVID-19 classification-aware topics from unlabelled data

5 Further Related Work

In addition to the work cited in the previous sections, the following research is related to our approach: **VAE based topic/document modelling** e.g. Mnih and Gregor (2014) trained a VAE based document model using the REINFORCE algorithm (Williams, 1992); Miao et al. (2017) introduce Gaussian Softmax distribution, Gaussian Stick Breaking distribution and Recurrent Stick Breaking process for topic distribution construction. Srivastava and Sutton (2017) proposed a ProdLDA that applies a Laplace approximation to re-parameterise Dirichlet distribution in VAE. Zhu et al. (2018) apply a Biterm Topic Model (Cheng et al., 2014; Yan et al., 2013) into the VAE framework for short text topic modelling. **Topic models with additional information (e.g. author, label etc.)**: example work includes Supervised LDA(Mcauliffe and Blei, 2008), Labeled LDA (Ramage et al., 2009), Sparse Additive Generative Model (Eisenstein et al., 2011), Structural Topic Models (Roberts et al., 2014), Author Topic Model (Rosen-Zvi et al., 2004), Time topic model (Wang and McCallum, 2006) and topic model conditional on any arbitrary Features (Mimno and McCallum, 2008; Korshunova et al., 2019). **NVDM in text classification:** Zeng et al. (2018); Gururangan et al. (2019), apply NVDM as additional topics feature in text classification. Compare to these approaches, CANTM is an asymmetric (different encoder input and decoder output) VAE that directly use VAE latent variable as classification feature without external features, hence we can use latent topics as classifier explanation.

6 Conclusion

In this paper, we introduced the COVID-19 disinformation corpus, which has 10 manually annotated categories of debunked COVID-19 disinformation. After quality control and a filtering process, the inter-annotator agreement average Cohen's Kappa is 0.70. We also present a new classification-aware topic model, that combines the BERT language model with the VAE document model framework and demonstrate improved classification accuracy over a vanilla BERT model. In addition, the classification-aware topics provide class related topics, which are: a) an efficient way to discover the class of (pre-defined) related topics, and b) a proxy explanation of classifier decisions.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Scott Brennen, Felix Simon, Philip Howard, and Rasmus Kleis Nielsen. 2020. Types, sources, and claims of covid-19 misinformation. Technical report, Reuters Institute.

Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Clare Clare and Lorna Christie. 2020. Covid-19 misinformation. *UK Parliament Post*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836, Brussels, Belgium. Association for Computational Linguistics.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1041–1048.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. 2019. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5880–5894.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the conference paper at the 3rd International Conference for Learning Representations*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

Iryna Korshunova, Hanchen Xiong, Mateusz Fedoryszak, and Lucas Theis. 2019. Discriminative topic modeling with logistic lda. In *Advances in Neural Information Processing Systems*, pages 6767–6777.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proceeding of International Conference on Machine Learning*, volume 30, page 3.

Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2410–2419. JMLR. org.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 411–418.

Andriy Mnih and Karol Gregor. 2014. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, pages II–1791.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108. Association for Computational Linguistics.

Julie Posetti and Kalina Bontcheva. 2020. Policy brief 1, disinfodemic: Deciphering covid-19 disinformation. Technical report, United Nation Educational, Scientific and Cultural Organization.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of 2017 International Conference on Learning Representations*.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433.

WHO. 2020. Novel coronavirus(2019-ncov) situation report - 13. Technical report, World Health Organization.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. 2018. Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3120–3131.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4663–4672, Brussels, Belgium. Association for Computational Linguistics.

7 Appendix A - COVID-19 Annotation Categories Definition

- Public authority: Claims about policy, action, or communication by a public authority (e.g. government department, police, fire brigade, government officials), including claims about WHO guidelines and recommendations as well as those about governments' action or advice.
- Community spread and impact: Claims about people, groups, or individuals with regard to how the virus is spreading (internationally, regionally, or within more specific communities); impact on people, groups (including religions and ethnic minorities), or individuals; deaths, etc.
- Medical advice, self-treatments, and virus effects: Claims about health remedies, self-treatments, self-diagnosis, signs and symptoms, effects of the virus, etc.
- Prominent actors: Claims about pharmaceutical companies, media organisations, health-care supply businesses, other companies, or famous people (including celebrities and politicians). Note that this does not include claims made bypoliticians or other famous people unless they are about other prominent actors.
- Conspiracies: Claims that the virus was created as a bioweapon, that some organization supposedly created the pandemic, that it was predicted, etc.
- Virus transmission: Claims about how the virus is transmitted and how to prevent transmission. This includes cleaning as well as use of specific lighting, appliances, protective equipment, etc.
- Virus origins and properties: Claims about the origins of the virus (e.g,. in animals) or its properties.
- Public Reaction: Claims that encourage hoarding, buying supplies, practising or avoiding social distancing, compliance or non-compliance with public health measures, protests and civil disobedience against official measures (including government measures). etc.
- Vaccines, medical treatments, and tests: Claims about vaccines, tests, and treatments, including the development and availability of a vaccine or a treatment. (Claims about self-treatment fall under the medical advice category, however.)
- Cannot determine: Use this category if the claim does not fit into any category above, if it does not seem to contain misinformation, or if you cannot read the language or understand the text for any reason.

8 Appendix B - Corpus Structure

The COVID-19 Disinformation corpus is organised in Json format with the following fields.

- Debunk_Date: The date of the disinformation debunked
- Date: The date of the disinformation first posted online
- Country: Country location of the fact-checker
- Claim: The claim of the disinformation
- Explanation: The explanation from the fact-checker of why this is a disinformation
- Source: The link to the disinformation debunk page
- unique_wv_id: hash code based on the first 200 words of 'Claim' and 'Explanation'
- Factcheck_Org: The organisation of the fact-checker from
- annotations: Contains annotations before merge.
- selected_label: The merged annotation label

9 Appendix C - Deriving the ELBO

This section describes the details of $ELBO_{x_{bow}}$ and $ELBO_{x_{bow},\hat{y}}$ derivation and calculation.

$$
\begin{aligned}
& z \sim q(z|x) \\
& \log p(x_{bow}) = \mathbb{E}_z \log p(x_{bow}) \\
& = \mathbb{E}_z[\log p(x_{bow}, z)] - \mathbb{E}_z[\log p(z|x_{bow})] \\
& = \mathbb{E}_z[\log \frac{p(x_{bow}, z)}{q(z|x)}] - \mathbb{E}_z[\log \frac{p(z|x_{bow})}{q(z|x)}] \\
& = EBLO_{x_{bow}} - D_{KL}(p(z|x_{bow})||q(z|x)) \\
& = EBLO_{x_{bow}} + D_{KL}(q(z|x)||p(z|x_{bow}))
\end{aligned}
$$

$$
\begin{aligned}
& ELBO_{x_{bow}} \\
& = \mathbb{E}_z[\log p(x_{bow}, z)] - \mathbb{E}_z[\log q(z|x)] \\
& = \mathbb{E}_z[\log p(x_{bow}|z)] + \mathbb{E}_z[\log p(z)] - \mathbb{E}_z[\log q(z|x)] \\
& = \mathbb{E}_z[\log p(x_{bow}|z)] - D_{KL}(q(z|x)||p(z))
\end{aligned}
$$

$$
\begin{aligned}
& z_s \sim q(z|x, \hat{y}) \\
& \log p(x_{bow}, \hat{y}) = \mathbb{E}_{z_s} \log p(x_{bow}, \hat{y}) \\
& = \mathbb{E}_{z_s}[\log p(x_{bow}, \hat{y}, z_s)] - \mathbb{E}_{z_s}[\log p(z_s|x_{bow}, \hat{y})] \\
& = \mathbb{E}_{z_s}[\log \frac{p(x_{bow}, \hat{y}, z_s)}{q(z_s|x, \hat{y})}] - \mathbb{E}_{z_s}[\log \frac{p(z_s|x_{bow}, \hat{y})}{q(z_s|x, \hat{y})}] \\
& = EBLO_{x_{bow},\hat{y}} - D_{KL}(p(z_s|x_{bow}, \hat{y})||q(z_s|x, \hat{y}))
\end{aligned}
$$

$$
\begin{aligned}
& ELBO_{x_{bow},\hat{y}} \\
& = \mathbb{E}_{z_s}[\log p(x_{bow}, \hat{y}, z_s)] - \mathbb{E}_{z_s}[\log q(z_s|x, \hat{y})] \\
& = \mathbb{E}_{z_s}[\log p(x_{bow}|\hat{y}, z_s)] + \mathbb{E}_{z_s}[\log p(\hat{y}, z_s)] - \mathbb{E}_{z_s}[\log q(z_s)|x, \hat{y})] \\
& = \mathbb{E}_{z_s}[\log p(x_{bow}|\hat{y}, z_s)] + \mathbb{E}_{z_s}[\log p(\hat{y}|z_s)] + \mathbb{E}_{z_s}[p(z_s)] - \mathbb{E}_{z_s}[\log q(z_s)|x, \hat{y})] \\
& = \mathbb{E}_{z_s}[\log p(x_{bow}|\hat{y}, z_s)] + \mathbb{E}_{z_s}[\log p(\hat{y}|z_s)] - D_{KL}(q(z_s|x, \hat{y})||p(z_s))
\end{aligned}
$$

Where $p(z) = p(z_s) = \mathcal{N}(0, I)$ is a zero mean diagonal multivariate Gaussian prior, hence the $D_{KL}(q(z|x)||p(z))$ and $D_{KL}(q(z_s|x, \hat{y})||p(z_s))$ will be

$$
\begin{aligned}
& p(z) = p(z_s) = \mathcal{N}(0, I) \\
& D_{KL}(q(z|x)||p(z)) = 0.5(\sigma^2 + \mu^2 - \log(\sigma^2) - 1) \\
& D_{KL}(q(z_s|x, \hat{y})||p(z_s)) = 0.5(\sigma_s^2 + \mu_s^2 - \log(\sigma_s^2) - 1)
\end{aligned}
$$

10 Appendix D – Experimental Details

The bag-of-words pre-processing step is the same as (Card et al., 2018): All characters are transformed to lower case; stopwords[18], punctuation, all tokens less than 3 characters and all tokens that include numbers are removed.

The pre-processing step for BERT representation is different from bag-of-words pre-processing. For the COVID-19 corpus, all characters are lowercased, and tokenised by the BERT tokeniser from Huggingface[19] (Wolf et al., 2019) Library. The IMDB corpus has a longer average document length, and some of the documents are longer than the pre-trained BERT length limitation (510 + CLS and SEP). Therefore, we only keep the first 510 tokens.

The ADAM optimiser parameters are default from the Pytorch Library: Learning Rate = 0.001, betas=(0.9, 0.999). The number of training epochs are 200 as in Card et al. (2018), with early stopping when no training loss (classification loss for CANTM) decrease after 4 epochs.

[18] snowball.tartarus.org/algorithms/ english/stop.txt

[19] https://github.com/huggingface/transformers

The fine tuning layers for BERT (Huggingface BERT-base implementation) are:

- encoder.layer.11.attention.self.query.weight,
- encoder.layer.11.attention.self.query.bias,
- encoder.layer.11.attention.self.key.weight,
- encoder.layer.11.attention.self.key.bias,
- encoder.layer.11.attention.self.value.weight,
- encoder.layer.11.attention.self.value.bias,
- encoder.layer.11.attention.output.dense.weight,
- encoder.layer.11.attention.output.dense.bias,
- encoder.layer.11.intermediate.dense.weight,
- encoder.layer.11.intermediate.dense.bias,
- encoder.layer.11.output.dense.weight,
- encoder.layer.11.output.dense.bias

The number of parameters in CANTM (include BERT) are 110,464,382 and number of trainable parameters are 8,066,942. The experiment hardware environment are: Intel(R) Xeon(R) Bronze 3204 CPU, TITAN RTX GPU, average epoch run time for COVID corpus is 41 seconds. The full list parameters number and epoch time shown in Table 10. Please note Gensim LDA does not have GPU support, hence it running on single core CPU.

Model	num. params	epoch time (sec.)
CANTM	110,464,382	41
BERTraw	109,489,930	36
BERT	109,521,200	37
SCHOLAR	740,360	0.05
NVDMb	109,661,140	37
NVDMo	1,152,600	20
LDA	151,750	0.6

Table 10: Number of parameters and epoch training time. Gensim LDA does not have GPU support

11 Appendix E – Additional Experiment Results

To ensure fair comparison between CANTM with the BERT classifier, we first compared: 1) BERT with additional hidden layer that matches the dimension of latent variables (denoted BERT in the result); 2) BERT without additional hidden layer, i.e. applying BERT [CLS] token output directly for classification (denoted BERTraw in the result). The COVID corpus results are shown in Table 11; the BERT with additional hidden layer has better performance in both accuracy and F-measure. Therefore, we report the BERT result in the paper.

Metrics	Acc.	F-1
BERT	58.78 (3.36)	54.19 (6.85)
BERTraw	58.77(3.56)	49.74 (7.62)

Table 11: BERT setting comparison on COVID-19 disinformation standard deviation in parentheses

Table 12 shows the class level F1 score of the COVID-19 disinformation corpus. CANTM has the best F1 score over most of the classes (CommSpread, MedAdv, PromActs, Consp, Vacc,None), also with better

	PubAuth	CommSpread	MedAdv	PromActs	Consp
BERT	61.17(4.50)	62.27(5.83)	75.03(6.54)	60.12(3.25)	49.92(12.04)
BERTraw	**65.64(2.91)**	59.35(4.77)	75.82(5.53)	65.51(4.34)	41.90 (10.46)
SCHOLAR	47.92(9.77)	48.84(11.56)	71.11(6.99)	46.93(8.66)	31.30(13.78)
CANTM	64.35(1.44)	**66.50(3.87)**	**79.68(2.12)**	**67.21(3.72)**	**60.06(6.80)**
	VirTrans	VirOrgn	PubRec	Vacc	None
BERT	**42.67(8.70)**	**57.62(6.72)**	23.68(10.01)	64.62(9.66)	12.59(11.35)
BERTraw	41.42(5.36)	53.20(15.92)	**27.19(13.55)**	65.48(9.62)	1.90 (3.8)
SCHOLAR	11.71(10.06)	45.15(20.49)	5.71(11.42)	55.37(15.78)	0.0(0.0)
CANTM	40.21(8.56)	55.19(3.43)	25.04(9.87)	**72.28(8.40)**	**15.52 (15.0)**

Table 12: COVID-19 disinformation class level F1 score, standard deviation in parentheses

	Acc.	F-1	Perp.	NPMI
Bert	58.78(3.36)	54.19(6.85)	n/a	n/a
Scholar	48.17(6.78)	36.40(10.85)	2947(353)	**0.25(0.015)**
NVDMb	n/a	n/a	1084(88)	0.09(0.004)
NVDMo	n/a	n/a	781(35)	0.08(0.001)
LDA	n/a	n/a	8518(1132)	0.12(0.005)
CANTM	**63.34(1.43)**	**55.48(6.32)**	**749(63)**	0.14(0.012)

Table 13: COVID-19 disinformation results, n/a stands for not applicable for the model

standard deviations. Except for the None class, standard deviations for CANTM are below 10. From the results, the most difficult class to classify is 'None'. This class represents anything that the annotators could not decide on, and therefore it could be anything that does not belong to the other 9 classes. In future work, we might need a better algorithm to handle this problem.

Table 13 shows the results of the COVID-19 performance with different baselines. The scores reported are the same as Table 4 in the paper, but standard deviation is added (standard deviation here is the average standard deviation from all classes). According to the results, CANTM not only improves the accuracy and F1 measure over the BERT baseline, but also improves standard deviation.

Table 14 shows the topics of the best and worst NPMI scores from CANTM and the baselines. We already discussed the fact that topic interpretability is not strongly associated with the NMPI score in supervised topic models (CANTM and SCHOLAR) in the paper. However, we found additionally that the NPMI score may have a better connection to the topic interpretability with the unsupervised topic modelling (LDA and NVDM). The best NMPI LDA topic (LDA0.149) can be interpreted as a mixed topic of Russian public authority and medical advice. However, the lowest NMPI LDA topic (LDA0.013) is difficult to interpret.

12 Appendix E – CANTM Topics

In this section we demonstrate the topics generated from CANTM (Table 15 to Table 18 are the COVID-19 topics from CANTM.) Table 15 is Classification-Aware topics Table 16 is Classification-Regularised topics Table 17 is Classification-Associate topics Table 18 is Classification-Aware topics updated from unlabelled data.

CANTM 0.50	please patents link ecuador patent read click full article guayaquil
CANTM 0.04	cure proven met protection leader pope aajtak within elizabeth developed
SCHOLAR 0.50	claim posts facebook false novel times shared multiple twitter thousands
SCHOLAR 0.05	people china doctors conditions barack obama masks pre existing containing
LDA 0.149	putin keep implemented contrary days code president avoiding drinking talk
LDA 0.013	cross breath anything empty generator broadcast hanks indicates external apparently
NVDM 0.192	scientifically context reached notification decided vitamin carrying alternative hair preventing
NVDM 0.037	publish quarantine corporation dna listed staff described restricted popular platform

Table 14: Topic words of the best and worst coherence topics

pope vatican francis filipinos region bat mosque seeks giuseppe daniel
strains animal reddit new visited tourist soup original suspected scene
production couple alongside lankan concentration photos image airport unleashed testing
article day click full johannesburg positive read tested italian due
please patents link ecuador patent read click full article guayaquil
korean robredo chloroquine request remedy existing zoology approved vice end
prime lockdown trupti modi police curb wake detained commissioner minister
lions committed drowned protests earthquake gandhi police detained indoors image
cure proven met protection leader pope aajtak within elizabeth developed
bodies originally show seconds london setting stopped rahul libya people
art islamic victims washed tribute lying bodies doctors hong suicide
times quoted robredo novel philippine graphic saddam contracting facebook activist
conspiracy russian movie update anything unleashed disaster trump doctored guard
case patients lockdown wake complete mock medical announcement government streets
died girl jan january kills ecuadorian link first please ecuadorians
strains drug traditional along replies comfort kit focus institute cure
developed treatment check development tea alcohol breath medicine warm study
tweet biden quote leni giuseppe rodrigo paid rappler urged visiting
juice via solution steam chicken lace coronavirus dettol kills effect
palau source spreading via soup says circulated animal claim online
coronavirus cure patents garlic study water sputum ecuador election vinegar
warned times facebook written letter claims transmission advisory data posts
kill ramesh intermediate related viral virus research patented negated book
philippines positive confirmed advisory task patient mask pakistan italian remains
cases case died sars handling chickens ebola reported dead thousand
contaminated steroids advice ministry purported gargling issued issuing colored red
warm dry avoiding remedy practices transmission treatments vinegar ways bakery
abdullah desai badawi swat lockdown minister force extension region police
langowan vatican market indonesia seconds palau wuhan prime roof alkaline
trump president donald quote approval forward roche friday felt intermediate
breath cause patented seconds garlic vinegar scientific deaths studies bat
ministry advisory case bakery aap wash pib notice positive dismissed
bat joe match origin palau flu federal bloggers source former
click humans viruses full indonesia cattle bat ago two let
wife justin multiple illegally bed hospital prayers photo migrants trudeau
pib government considering edited disaster issued forward act offense affairs
restaurant ahmad uploaded saddam former china nazi pretty mosque xia
vinegar india clapping ronaldo salt getting mask wear suggesting message
affairs salt vinegar drinking method test lemon fda ministry media
woman video hospital patients thanksgiving barcelona drill tribute newly gandhi
biden obama bill joe allah china narendra funding giuseppe evers
dead lives tribute night circulated photo left italian croatia picture
covid said kindly virus known use clearly gargling cure suggest
bicarbonate drinking cured dryer effective still eliminates temperature ice lemon
dry maximum masks voted outbreaks doctors happens sars temperatures mask
click full tested read confirmed vatican dettol please seeks positive
positive delhi jan doh january tea chart pope negative strains
posts youtube viewed multiple television issue cases based incidents prevention
barcelona couple development image photograph happens croatia broiler picture virus
link ecuador please article read full lions russian biden saves

Table 15: Full list of Classification-Aware topics for COVID-19 corpus, each line is a topic

lace lay manufactured imports treated demonstration strewn pedestrian gas fronts
chance indoors supermarket bodies citizens streets items frequently thrown fronts
media coronavirus viral taken world social shows video novel man
health coronavirus pandemic cases one covid organization italy countries outbreak
khan living foundation patents london dinner camilla fictional cornwall actor
therefore tony recommendations airborne mouth facial generally copd hair chance
shows actually chinese victims photo art august biden italy protesters
seconds karnataka enter went husband breath converted bjp colored cuban
local click newspaper passengers actor employee corporation link article bank
internet went america mumbai related recent worked extended offense june
deaths covid number health account cases take state confirmed obama
washed wake sea video wife ministry recorded case trudeau department
coronavirus new people cases china kill evidence virus article kills
world doctors health people medical according food sri misleading facebook
found pictures images march viral image old disease along chicken
kong hong video wuhan shows clips police suspected bodies seconds
often buy allow masks buying lockdown important australian mask march
please link alongside full click jair kit crying read article
coronavirus claim evidence said twitter novel facebook times posts multiple
disease using prevention says whether election official cure study research
available president south buy priyanka bill trump testing test vaccine
click please full read labs cuban guayaquil ministry ecuador treating
number aap severe transmission mers likely centre posts worked philippines
china reported internet link francis sars read pope manufactured contain
stated scientific various name oil cure made barack israel clarified
ahmad weave seek jinping badawi abdullah ultraviolet xia additionally soup
place viruses new yet animal products strain sanitizer strains get
considering federal ireland announce patrolling passengers wife anti presidential rubbished
flu sars inaccurate studies important mers around type runny suggests
full medicine ashore items sea wash migrant click ecuadorian organization
link covid two full read doctor confirmed created please click
purportedly notice doses supermarkets try promotes abortion experts levels earthquake
coronavirus ministry novel home caused related video traditional viral youtube
indian claim minister india also false misleading claims prime sri
italy bodies pictures china prayer mosque plates video allah coffins
tested buckingham tourists hand lysol ronaldo advises washington met manufacturer
two wuhan china denied chinese center report confirmed reports officials
desai visuals dung saddam cow july kills sold older lips
satellite accurate sulfur sun dioxide kingdom forecasts translated maps camilla
tested positive doctor coronavirus help recent person getting india novel
stopped company sanitizer label desai trupti saddam theories entering buying
coronavirus different million new family covid say allegedly respiratory death
video man covid first circulated claiming false woman least taken
shows covid hospital video wuhan doctor photo cure woman test
twitter facebook photo times italy claim victims curfew health image
key copd recognized tourist antibiotics rabies rumour short emphasized medication
year modi india old positive narendra curfew muslim wife announced
military full code daniel roche speech event archive please actor
spread show patients new covid virus viral also outbreak response
movie picture couple italian lions lion photograph volleyball barcelona tom

Table 16: Full list of Classification-Regularised topics for COVID-19 corpus, each line is a topic

PubAuthAction	china government people claim india march facebook coronavirus outbreak covid
CommSpread	coronavirus photo covid people shows video claim novel confirmed shared
GenMedAdv	coronavirus cure experts novel prevent water evidence kill scientific hot
PromActs	coronavirus said novel trump video shared president media covid hospital
Consp	coronavirus chinese covid virus lab wuhan outbreak new china posts
VirTrans	coronavirus covid evidence virus claim spread video said people surfaces
VirOrgn	coronavirus shows video bat wuhan novel source outbreak virus taken
PubPrep	video shows facebook image shared show times outbreak false circulated
Vacc	cure vaccine new covid developed novel scientists claimed coronavirus claims
None	video image taken shared covid due streets old india reports

Table 17: Full list of Classification-Associate topics for COVID-19 corpus, each line is a topic

say carry meeting come weed china publicly regularly director consumption
mention conference professor since april quarantined starting spoke supporting please
fake ministry close study refused february attempt video beaten administration
photo post french smoking circulating bank side account eating image
police bat cdc spread visit tourists answer data july clip
announced lockdown stay school office deadly arrested ground degrees always
lockdown every end afp common true islamic concerns rapid undergoing
health ever salt swat ginger pope uses welfare hands singapore
news president victims use minutes cases day continue laboratory developed
corporation denied present force official palau show give post pneumonia
says reports elderly infection claimed beijing speech generally reporting experts
video french time patient includes place victims threatened forward close
leave taken investigation recommendations ventilators exposed organisms people deaths put
italy aired patent election malaria pepper working contrary five growing
suicide included indicates kansas temperatures staying jamaat communities italy two
prayer effective discovered led herbal cov patient article china takes
people vaccine since weapon hospital scientific man group cells working
prime vaccination worldwide due zone created planning airlines producing ultraviolet
evidence human gives even temperature science end claimed across conte
video south covid emergency response never chinese seconds changed images
leave research conspiracy indian starting individuals though text tanker germany
correlation visible sometimes lay produce outright super district initiation six
newly hanks definitively last six hence lack barack elizabeth subway
modular considering stories gotabaya strewn abdullah vibration ramesh miami commission
match preventions relationship indonesia eliminate herbal obama diagnosed bjp japan
key screening dung worldwide try teacher carbon thai decisions spokesman
key advising physically solutions restriction doh camilla promotes vibration scattered
sunday ready netflix production pib telecast cause nazi emergency stopped
reports known study findings hospitals movement vaccine garlic family onion
stating died gas kill best strain announcing chain mass acid
outright experts warm respiratory helps announcements damage factual instead crisis
false ago video president cases undergoing clarification chinese bulletin project
hours cause jamaat make romania xia stock effort isolation kenyan
new medical police city media china social agency also back
coronavirus dangerous medicine advise intermediate graphics brazil ebola imposed generator
quarantine warm sent claims cells members bill conte soup company
government face available outbreak exist proof head service found cruise
found suspended man joe congress whatsapp trump weed claim sauna
case facebook north used production many protocol context citizens text
cures issued onion spokesperson failure even spread actually returned moment
virus published confirmed phishing link cure station extension taking spread
continues widespread stands visible patrolling mandate strewn violating absolutely sophie
trump items blood guard technology told home amid suspended intermediate
coffins announcement october protect coast capacity company carry supplies committed
shows please war paulo nose road flu refer runny post
spain published people south taken showing offering reviewed anything mock
police july infected coffins protocol experts receive say world website
outbreak social hospital virus bats eligible latest hoax taken risk
video north social lemon whole report sources rahul ground ahmad
cases lost official found students manipulated support patients thousand forward

Table 18: Full list of Classification-Aware topics for unlabelled COVID-19 corpus, each line is a topic