A Risk Communication Event Detection Model via Contrastive Learning

Mingi Shin[1], **Sungwon Han**[1], **Sungkyu Park**[2] and **Meeyoung Cha**[2,1]
[1] Korea Advanced Institute of Science and Technology (KAIST)
[2] Institute for Basic Science (IBS)
Daejeon, South Korea
{mingi.shin, lion4151, shaun.park, meeyoungcha}@kaist.ac.kr

Abstract

This paper presents a time-topic cohesive model describing the communication patterns on the coronavirus pandemic from three Asian countries. The strength of our model is two-fold. First, it detects contextualized events based on topical and temporal information via contrastive learning. Second, it can be applied to multiple languages, enabling a comparison of risk communication across cultures. We present a case study and discuss future implications of the proposed model.

1 Introduction

The novel coronavirus disease (COVID-19) is affecting public health and the economy worldwide. The surge in social media usage during the pandemic led the online content to an excellent tool to examine risk communication (Lazer et al., 2018; Beaunoyer et al., 2020). As more people seek and share information online, NGOs and especially the WHO have warned of the danger of the increasing misinformation on the pandemic. A new term, *infodemic*, was coined to describe this phenomenon.

On the other hand, the natural language processing (NLP) community's recent developments enable in-depth analysis of topical changes from online resources. Latent Dirichlet Allocation (LDA) can detect major topics from unstructured text data (e.g., extract topics in the context of a global pandemic (Park et al., 2020)). Advanced language models like BERT can be used to learn representations (e.g., the pandemic discourse on Twitter (Müller et al., 2020)). A language-agnostic version of BERT further extends its capability to handle multiple languages (Gencoglu, 2020).

Despite social media's potential to understand the risk communication pattern and infodemic during the pandemic, several challenges remain. One of them is the temporal aspect. The existing topic models which exclude temporal information cannot represent how the conversations/themes developed over time (Blei and Lafferty, 2006). While some works propose the refinements of such models by incorporating the time or metadata information (Blei and Lafferty, 2006; Roberts et al., 2013), they still need to dissect the data into arbitrarily chosen time chunks. Particularly in risk communication, where the public attention evolves quickly in a short period, the contextual information intertwining topic and time becomes a critical component (Atefeh and Khreich, 2015). Considering the time aspect also allows identifying topics that are of significant influence yet over a short span of time.

Addressing the limitation above, we present a time-topic cohesive model to detect contextualized events and topics over time. Our model marries key ideas from contrastive learning (hereafter CL) and multilingual BERT (hereafter mBERT). CL is a machine learning and computer vision technique to classify similar objects by devising a triplet loss function among one anchor and two targets (Dai and Lin, 2017). By designing a triplet loss such as computing the difference of topic and time between anchor and target tweets, our model can jointly consider temporal and topical characteristics when detecting major events about the pandemic. mBERT allows us to apply the model to multiple countries for comparison.

The final model is applied to a collection of Twitter messages gathered from three Asian countries: South Korea, Vietnam, and Iran. Based on the determined events, we can understand what information (or misinformation) was mainly talked about at what stage of the pandemic in each country. Unlike

Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom, pages 39–43
Barcelona, Spain (Online), December 12, 2020

existing topic models, this new method also captures the temporal coherence of topics. We present a case study that shows how the risk communication on COVID-19 starts with several key events initially and then expands to diverse domains in South Korea.

2 Method

2.1 Data

We analyze GeoCoV19 (Qazi et al., 2020), a multilingual dataset of tweets about COVID-19 with location information. The data comprise tweets that contain COVID-19-related keywords in multiple languages. The coverage of data is for 90 days, from February 1 to May 1, 2020. We only utilize tweets written in Korean, Farsi, and Vietnamese, all local languages corresponding to South Korea, Iran, and Vietnam. This was done by constraining the language to the `lang` attribute, the auto-detected language of the tweet text. The total number of tweets were 43,347 (% of retweets: 79), 19,174 (34), and 4,359 (16) for each language.

The studied Asian countries had different epidemic situations. South Korea was one of the first countries affected by the virus, recording a surge in the increase of confirmed cases in February and March. By May, the country saw a flattening trend in the confirmed cases. Meanwhile, Iran is one of the most severely affected countries. Within our target period, the number of confirmed cases rapidly increased and maintained over several hundreds of cases. On the other hand, in Vietnam, the numbers have consistently stayed below a hundred throughout the data period.

Preprocessing Data. We excluded retweets, URLs, language-dependent stopwords, and redundant whitespace. We also replaced the mentioned names with a `UNK` token prior to analysis.

LDA Topic Modeling. We utilized the LDA model to extract topical information from the text. We first tokenized the tweet text by standard Python libraries for each specific language. We then trained the LDA model for 50 topics. Each tweet was labeled with the most probable topic.

2.2 Contextualized Event Detection over Time

We propose an event detection algorithm that considers word occurrence patterns and time concurrently. We regard tweets with similar word patterns but large time discrepancy as entries from different events. This approach for training is inspired by the CL approach. By constructing triplets that reflect the time and topical distance among tweets, optimizing triplet loss can directly lead the embedding to gather similar events within a short period. We then perform a clustering algorithm over the trained embeddings to extract the topic clusters. The structure of mBERT is utilized, and its pre-trained weights are used as initialization to distill contextualized information during training the embeddings.

Our model adds a linear layer on the concatenation of the pooled output from mBERT and the normalized timestamp (min-max scaling) to embed tweets within the fixed sized vector. Then, we project the concatenated vector into L2-normalized space. As a pooling strategy, we used the output from the `CLS`-token which is trained for the next sentence prediction. The result model is mBERT which has a spherical embedding head on the top. To fine-tune this model, we construct two kinds of triplets: (1) *LDA-dependent triplet*, in which anchor and positive tweets shows the same topic from the trained LDA model while negative does not; (2) *time-dependent triplet*, in which timestamp from positive tweet is nearer to the anchor than negative. We used a combined loss between the two triplet losses as an objective function. Given a as an anchor Tweet, p as a positive sample from the dataset and n as a negative, two kinds of triplet loss (L_{tri}) can be defined below:

$$L_{tri\_topic} = \max\{s(w_a, w_n) - s(w_a, w_p) + \tau_{topic}, 0\} \tag{1}$$

$$L_{tri\_time} = \max\{s(w_a, w_n) - s(w_a, w_p) + \tau_{time}, 0\} \tag{2}$$

where s is a similarity function, τ_{topic} and τ_{time} are thresholds, and w_a, w_p, w_n are the embeddings from the model using a, p, n as an input respectively. Since our embedding is L2-normalized, s is defined as a dot-product function which represents the cosine similarity. Two triplets can be denoted as L_{tri_topic}

and L_{tri_time}, so the eventual loss (L_{total}) to be minimized has become $L_{total} = L_{tri\_topic} + L_{tri\_time}$. Through grid search, we could find that 0.1 and 0.1 are optimized values for τ_{topic} and τ_{time}, respectively. Finally, we perform spherical k-means clustering to identify topic clusters. The Silhouette value is measured to determine the number of clusters.

3 Results

3.1 Embedding Results

Figure 1 illustrates two embedding examples from the various combinations on choosing τ in the case of South Korea. The figures are plotted via t-SNE by reducing dimensionality by 2. We could confirm that our hyperparameter setting (i.e., τ_{topic} = 0.1, τ_{time} = 0.1) results more distinctive clusters than other conditions. The same values for two τ may mean that the topic and time information equally contributes to the embeddings.

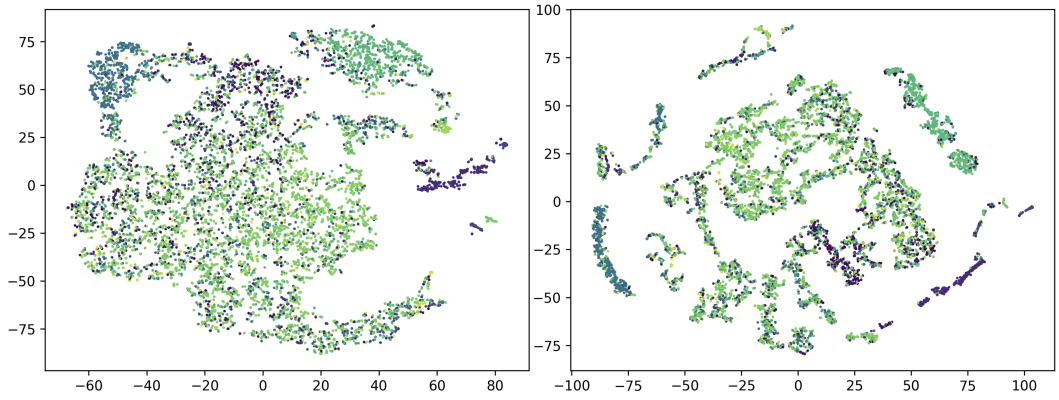

Figure 1: Embedding comparison via t-SNE between two different τ settings in the case of South Korea. τ_{topic} = 0.1, τ_{time} = 0.01 (left) whereas 0.1, 0.1 (our case, right).

3.2 Clustering Results with Evaluation

We then compare the clustering results within the South Korean case based on ablation studies, which are utilizing 1) LDA; 2) LDA+mBERT (concatenating two normalized outputs, then run k-means clustering); 3) LDA+mBERT+CL (our model). The statistics of the clustering results are presented in Table 1). The number of tweets per event of our model varies less than those of LDA, although our model has fewer events. We used the average standard deviation (SD) and standard error (SE) in timestamp as a metric to evaluate whether detected events share the same temporal information. Concerning the timestamp metric, our model shows the smallest time dissimilarity per event on average, meaning our model well reflects the temporal information in practice. By combining these two observations, we conclude our model smoothed LDA by successfully contemplating the time information together as planned.

Metric	Statistics	LDA	LDA+mBERT	**LDA+mBERT+CL (ours)**
# of Events	Count	50 (fixed)	64	20
# of Tweets	Mean (SD)	182.38 (489.43)	142.48 (136.63)	455.95 (307.46)
Timestamp	Average SD	0.25	0.24	**0.16**
	Average SE	0.03	0.02	**0.01**

Table 1: Statistics of the clustering results from three models in the case of South Korea.

3.3 Detected Events across Time by Country

For Iran and Vietnam, Figure 2 shows the trends of the detected events, respectively. When determining those countries' events, we have used the same hyperparameter values from the South Korean case. Particularly with South Korea, we have further qualitatively interpreted the events and merged similar topics, as depicted in Figure 3. For instance, total 20 topics detected from the South Korea data are labeled and merged into 13 discriminative topics. Tweet volumes became larger from mid-February as the number of confirmed cases abruptly rose due to a regional church outbreak, and it lasted until the end of March. During this period, the number of events was relatively small as people focused on a few events like news about the global confirmed cases. As the situation eased from April, the events became more diverse. Also, we could see that the rumor events were relatively steady across the whole period.

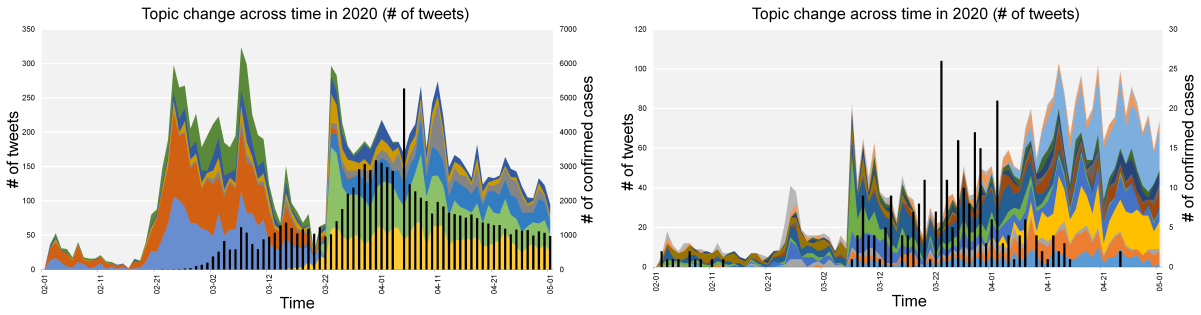

Figure 2: Daily event trends in Iran (left) and Vietnam (right). Black bar graphs represent the number of daily confirmed cases.

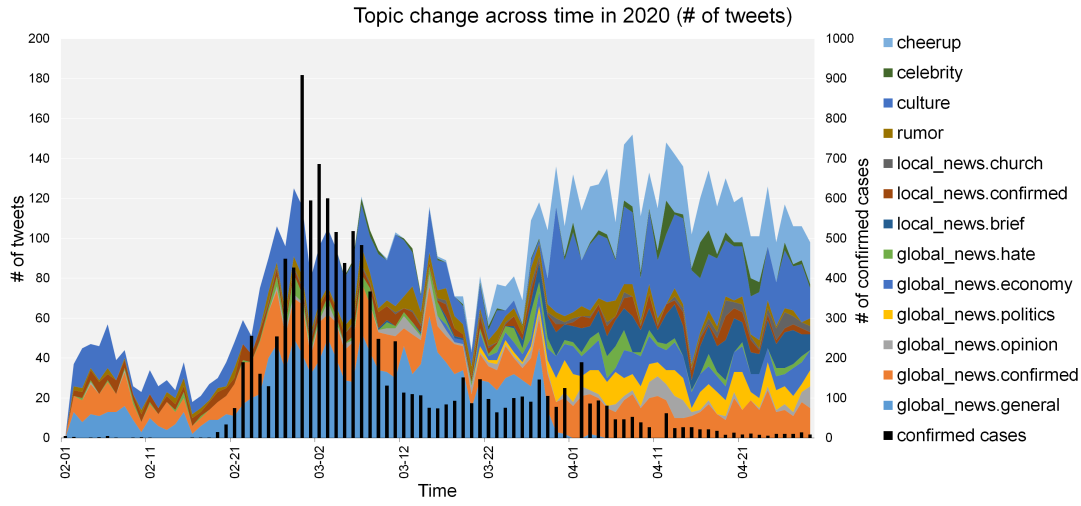

Figure 3: Daily event trends after qualitatively assessed in South Korea.

4 Discussion

We have extracted event trends based on the contextualized event detection model with CL in the paper. By introducing two kinds of the triplets: LDA-dependent and time-dependent triplets, our model efficiently trains the embedding to gather similar events within a short period. We also present the qualitative interpretation of the trends in South Korea as a possible application. As the proposed method can be executed by country, if we detect misinformation or disinformation prevalent in only one country first, we can quickly alarm other countries to deal with this issue preemptively.

For future studies, we can think of utilizing other topic modeling algorithms beyond LDA, like BTM (Yan et al., 2013) or DocNADE (Larochelle and Lauly, 2012) by considering temporal traits together, then compare the performance to the current model. Some research has shown the efficacy of

variational autoencoder on modeling topics (Miao et al., 2016) in terms of both topic coherence and perplexity. Besides, the framework of the current work is retrospective, and we can also try to build a monitoring/analyzing framework for detecting events in real-time. By investigating how the risk communication on COVID-19 proceeds in real-time, it would be easier to react to misinformation immediately.

Acknowledgements

This work was supported by the Institute for Basic Science (IBS-R029-C2).

References

Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

Elisabeth Beaunoyer, Sophie Dupéré, and Matthieu J Guitton. 2020. Covid-19 and digital inequalities: Reciprocal impacts and mitigation strategies. *Computers in Human Behavior*, page 106424.

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

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907.

Oguzhan Gencoglu. 2020. Large-scale, language-agnostic discourse classification of tweets during covid-19. *arXiv preprint arXiv:2008.00461*.

Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. In *Advances in Neural Information Processing Systems*, pages 2708–2716.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Sungkyu Park, Sungwon Han, Jeongwook Kim, Mir Majid Molaie, Hoang Dieu Vu, Karandeep Singh, Jiyoung Han, Wonjae Lee, and Meeyoung Cha. 2020. Risk communication in asian countries: Covid-19 discourse on twitter. *arXiv preprint arXiv:2006.12218*.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: A dataset of hundreds of millions of multilingual covid-19 tweets with location information.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4. Harrahs and Harveys, Lake Tahoe.

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.