What is Discussed about COVID-19: A Multi-Modal Framework for Analyzing Microblogs from Sina Weibo without Human Labeling

Hengyang Lu[1, *], Yutong Lou[2], Bin Jin[2] and Ming Xu[2]

Abstract: Starting from late 2019, the new coronavirus disease (COVID-19) has become a global crisis. With the development of online social media, people prefer to express their opinions and discuss the latest news online. We have witnessed the positive influence of online social media, which helped citizens and governments track the development of this pandemic in time. It is necessary to apply artificial intelligence (AI) techniques to online social media and automatically discover and track public opinions posted online. In this paper, we take Sina Weibo, the most widely used online social media in China, for analysis and experiments. We collect multi-modal microblogs about COVID-19 from 2020/1/1 to 2020/3/31 with a web crawler, including texts and images posted by users. In order to effectively discover what is being discussed about COVID-19 without human labeling, we propose a unified multi-modal framework, including an unsupervised short-text topic model to discover and track bursty topics, and a self-supervised model to learn image features so that we can retrieve related images about COVID-19. Experimental results have shown the effectiveness and superiority of the proposed models, and also have shown the considerable application prospects for analyzing and tracking public opinions about COVID-19.

Keywords: COVID-19, public opinion, microblog, topic model, self-supervised learning.

1 Introduction

The outbreak of the new coronavirus disease 2019 has caused huge losses all around the world. Until April 2020, the number of infections has already exceeded 1 million. This is far more than that of Severe Acute Respiratory Syndrome (SARS), which is also a coronavirus that caused a pandemic in 2002. Experts and researchers from various research fields, such as healthcare, biology, computer science, and so on have made many efforts to relieve this disaster. Among them, AI and information technologies play a quite important role in fighting COVID-19. For example, Stebbing et al. [Stebbing, Phelan, Griffin et al. (2020)] utilized an AI-driven knowledge graph to facilitate rapid drug

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China.

[2] National Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing, 210023, China.

[*] Corresponding Author: Hengyang Lu. Email: luhengyang@jiangnan.edu.cn.

Received: 29 April 2020; Accepted: 23 May 2020.

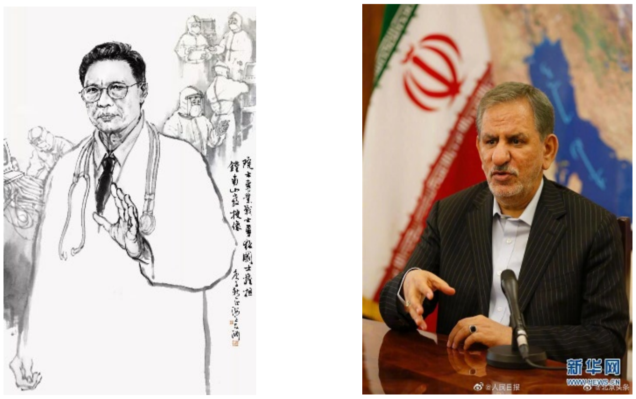
CMC. doi:10.32604/cmc.2020.011270 www.techscience.com/journal/cmc

development. Jin et al. [Jin, Wang, Xu et al. (2020)] built an AI system to automatically analyze Computed Tomography (CT) images by detecting COVID-19 pneumonia features. Jiang et al. [Jiang, Coffee, Bari et al. (2020)] proposed an AI framework for predicting the coronavirus clinical severity.

These latest researches try to beat the pandemic from the perspectives of promoting medical diagnosis and improving medical techniques. At the same time, it is also quite important to track and discover what people think about COVID-19 and find bursty events in a certain period. For example, there exist many rumors online these days. It is impossible to break all the rumors immediately. If we can discover rumors from the most bursty topics as soon as possible, we are able to dispel these rumors in time and minimize the harm caused by rumors. Considering the online social media like Sina Weibo has provided a large amount of valuable public opinions about COVID-19, we have witnessed the positive influence brought by microblogs posted on Sina Weibo. For example, microblogs about “human-to-human infection” have aroused people’s attention to this new virus. Microblogs about “the Thor Mountain (Lei Shen Shan in Chinese) hospital was built” have encouraged people to believe in victory. This kind of valuable information can not only give insights for governments but also keep the public informed and aware of the development of COVID-19 in time. Therefore, it is necessary to apply AI techniques to discover and track online public opinions automatically every day.

Because of the huge amount of data from Sina Weibo, it is almost impossible to analyze and label these microblogs by artificial processing. For texts, the topic model can effectively discover latent thematic information in an unsupervised way and each topic is represented by a set of words. Representative models include probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) [Alghamdi and Alfalqi (2015)]. These conventional models have been successfully applied to various tasks in long-text scenarios. However, there exist two main challenges to discover bursty topics from online social media. Firstly, the average length of texts from online social media is quite short, which causes the performance of conventional models to decrease sharply. We call it the sparsity problem. Secondly, texts from online social media naturally contain timestamps, which have made bursty topics keep changing over time. However conventional models are not able to differentiate the bursty degree between topics. This can be concluded as the burstiness problem.

What’s more, microblogs from online social media usually belong to multiple modalities, including texts, images, and so on. Because an image is worth a thousand words, they can also provide useful insights for public opinions and discussions about COVID-19 in a more vivid way. One of the effective strategies is learning feature representations for each image and then retrieving related images according to their similarities. Recent techniques mainly exploit deep learning to learn deep features based on large image collections such as ImageNet. However, the training images in these collections are annotated with single labels such as “person”, “car” etc., which are quite coarse-grained. In real scenarios like Sina Weibo, a single image often contains multiple semantic information and needs more specific semantic descriptions. Fig. 1 is an example of two microblogs crawled from Sina Weibo. Fig. 1(a) describes the famous Chinese experts Zhong Nanshan and Fig. 1(b) describes the first vice president of Iran, who are both related to news about COVID-19. Common pretrained features can only recognize “person” as objects in these images but fail to differentiate their different semantics.



(a). image in microblog about the Chinese expert “Zhong Nanshan”

(b). image in microblog about the First Vice President of Iran

Figure 1: Examples of images appear in different microblogs

This paper aims to automatically discover bursty topics and related images about COVID-19 from Sina Weibo by exploiting AI techniques. We have collected microblogs related to pneumonia dating from 2020/1/1 to 2020/3/31, which contain over 80,000 texts and over 40,000 images. For texts, we aim to solve the sparsity problem and the burstiness problem at the same time. For images, we aim to learn image features, which can reflect more specific semantic information by utilizing semantics of corresponding texts, and retrieve related images to reveal public opinions during a certain period. The main contributions of our paper are threefold:

1) We have collected a Chinese dataset from Sina Weibo, which contains multi-modal microblogs related to COVID-19, dating from 2020/1/1 to 2020/3/31.
2) We have proposed an improved short text topic model, which generates new phrases for discovering bursty topics in an unsupervised way.
3) We have proposed a self-supervised model to learn image features for topic-related image retrieval. This model can help learn features, which reflect semantic information of images without human labeling.

This paper is organized as follows. Section 2 shows related researches. Section 3 presents the proposed models for dealing with texts and images respectively. Section 4 contains the experiments as well as analysis and Section 5 concludes our paper.

2 Related work

This paper analyzes online data from social media to discover and track public opinions about COVID-19. There exist the latest researches mining Sina Weibo for sentimental analysis on COVID-19 [Chen, Chang, Wang et al. (2020)]. Sina Weibo is a widely used Chinese online social media, which contains rich and real posts from users. Thus, we also use Sina Weibo as the data source in this paper. Microblogs from Sina Weibo are usually multi-modal data, which can provide public opinions in the form of texts and images. This section summarizes related work and techniques from the perspectives of analyzing texts and images respectively.

For texts, the topic model is widely applied to discover public opinions [Ma, Yu, Ji et al. (2019)]. Texts from Sina Weibo are often short in length. This brings the sparsity problem to conventional topic models. Furthermore, each microblog has a timestamp when being posted, so topics keep changing according to different time slices. Bursty Biterm Topic Model (BBTM) [Yan, Guo, Lan et al. (2015)] is a probabilistic model for discovering bursty topics in online social media like Sina Weibo. It extends the famous short-text topic model named Biterm Topic Model (BTM) [Yan, Guo, Lan et al. (2013)] by incorporating prior knowledge on biterms. This makes BBTM can overcome the sparsity problem when learning latent topics. Benefit from the effectiveness of BBTM, there are many further models and applications based on the idea of BBTM. Most of them are designed to deal with data from online social media, such as the Time-User Sentiment/Topic Latent Dirichlet Allocation (TUS-LDA) model [Xu, Qi, Huang et al. (2018)], the User Interaction based Bursty Topic Model (UIBTM) model [Li, Du, Cui et al. (2018)] etc.

For images, it is one of the effective ways to show public opinions with a set of related images. The core solution is representing images with their features and retrieving related images according to their similarities. Researches about image feature learning have experienced the development from hand-crafted features to deep features. Conventional hand-crafted feature extraction methods include Gist, Scale-invariant feature transform (SIFT), Histogram of Oriented Gradient (HoG), and so on [Nanni, Ghidoni and Brahnam (2017)]. These conventional methods are huge in costs because they rely on artificial designs. The emergence of convolutional neural networks (CNNs) has changed this situation. Representative models such as the Alex Krizhevsky Network (AlexNet) [Krizhevsky, Sutskever and Hinton (2012)], the Visual Geometry Group Network (VGGNet) [Simonyan and Zisserman (2014)] and the Residual Network (ResNet) [He, Zhang, Ren et al. (2016)] trained image features with deep neural networks. These deep features have achieved much better improvements than hand-crafted features in various tasks, such as object tracking [Zhang, Jin, Sun et al. (2018)], image super resolution [Wang, Jiang, Luo et al. (2019)], image classification [Wang, Li, Zou et al. (2020)] and so on.

However, these models highly relied on labeled data like ImageNet. They are not suitable for dealing with real data from Sina Weibo, which keep emerging. Recently, self-supervised learning has become a new branch to learn image features without artificial labeling. For example, Gomez et al. [Gomez, Gomez, Gibert et al. (2018, 2019)] made use of the matching relationship of texts and images from Wikipedia, and defined the learning pretext as fitting image features as close as corresponding text's semantics. This novel attempt has achieved success in cross-modal retrieval than the above models, which trained image features based on ImageNet.

This paper takes both advantages of BBTM for texts and the self-supervised learning for images. We propose a unified multi-modal framework to discover public opinions about COVID-19 from Sina Weibo. For texts, we improve the readability of BBTM by generating new phrases from the latest online discussions. For images, we treat the topic distributions of texts as semantic representations of images, which were posted together in the same microblogs. Finally, we can automatically discover and describe public opinions in the form of bursty topics and a set of related images.

3 Proposed models

In this section, we give the problem settings at first. Then we give detailed descriptions of the proposed models in this paper. For discovering bursty topics of texts, we propose a New Phrase based Bursty Biterm Topic Model (NP-BBTM). For retrieving related images about COVID-19, we propose a Topic Distribution based Self-supervised Learning Model (ToDS) to learn image features.

3.1 Problem settings

The general framework is as Fig. 2 shows.

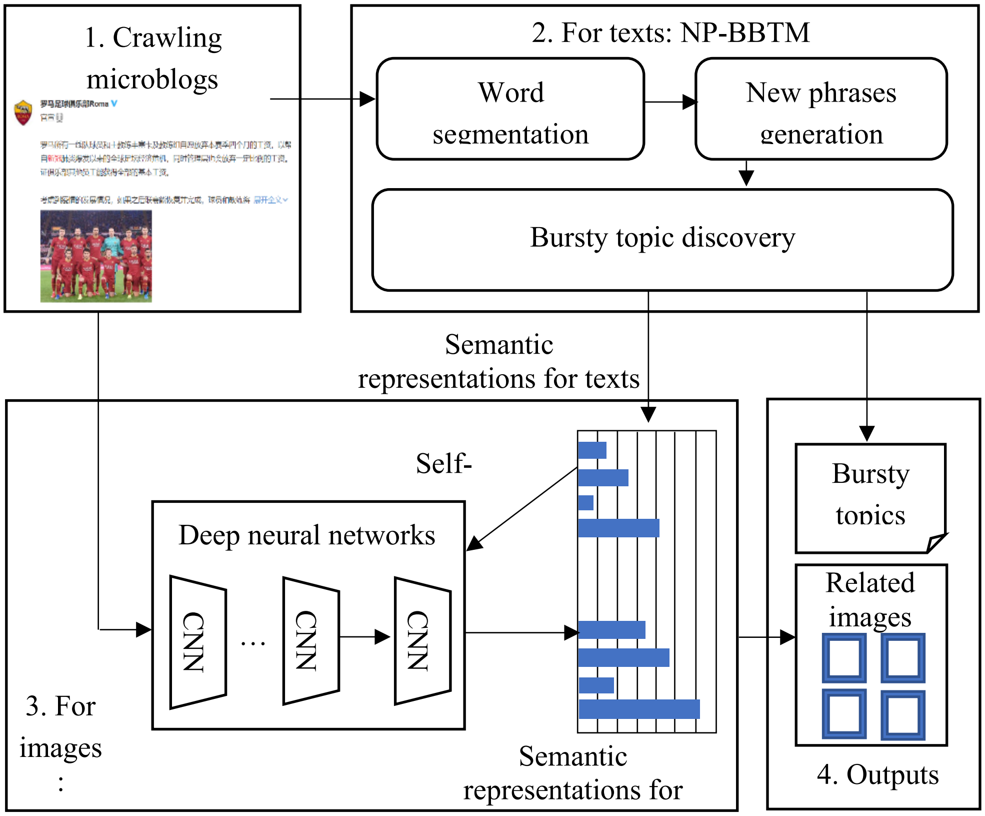


Figure 2: General framework

This paper takes microblogs about COVID-2019 as inputs, which are crawled from Sina Weibo, including texts and images. The outputs are respected to be bursty topics during a certain period and related images that describe public opinions.

For texts, we aim to discover topics to reveal public opinions about COVID-19 in a certain time slice. Each topic is represented as a set of related words or phrases, which is as same as most topic models. In this scenario, topics can be divided into common topics and bursty topics, where common topics are discussed almost every day and bursty topics are widely discussed during a certain period.

For images, we aim to retrieve several semantically related images to represent public opinions about COVID-19. We map the image feature to make semantically related images closer. What's more, we build the connection between images and texts, which can help retrieve related images in a cross-modal way.

3.2 Bursty topic discovery with new phrases

For bursty topics from online social media, there exist many new phrases, which represent important opinions in a certain time slice. Because these new phrases are rarely observed before, most word segmentation tools failed to correctly detect these phrases in the data preprocessing. We propose a new phrase generation algorithm to solve this problem.

For a document $d_i = [w_1, w_2, \dots w_{|d_i|}]$, which consists of $|d_i|$ words. We aim to detect some new phrases, each of which consists of a set of words in d_i. A new phase should satisfy two main constraints. Firstly, all the words should keep their order in the original documents. Secondly, any two adjacent words in the new phrase should share almost the same co-occurrence in the corpus. In this paper, we set this threshold as 3.

To satisfy the first constraint, we can traverse each document d_i according to the order of words' appearance. To satisfy the second constraint, we can compare the co-occurrence of any two adjacent words. We can generate new phrases from the original corpus *D* and achieve the pseudo corpus consisted of new phrases at the same time. Details are summarized in Algorithm 1. We use a 2-dimensional matrix to store the co-occurrence of any two words w_i and w_j, denoted as C.

Algorithm 1. Pseudo corpus generation with new phrases for NP-BBTM

```
Inputs.   D, co-occurrent threshold $\delta$.
Outputs.  PD.
Initialize the co-occurrent matrix $C$ as a zero matrix.
For each d in D:
    For i → 2 to $|d|$:
        $C(w_{i-1}, w_i) = C(w_{i-1}, w_i) + 1$
For each d in D:
    Initialize pd as an empty string.
    Set $startIndex$ as 0, $endIndex$ as 1, Set $newPhrase$ as $w_{startIndex}$.
    While $startIndex < |d|$ and $endIndex < |d|$:
        If $C(t_{startIndex}, t_{endIndex}) > \delta$:
            $newPhrase = newPhrase + w_{endIndex}$
            If $|C(t_{startIndex}, t_{endIndex}) - C(t_{endIndex}, t_{endIndex+1})| < 3$:
                $startIndex = endIndex$
            Else:
                $pd = pd + newPhrase$
                $startIndex = endIndex + 1$
        Else:
                $pd = pd + newPhrase$
                $startIndex = endIndex$
        $endIndex = startIndex + 1$
    Insert $pd$ to $PD$
```

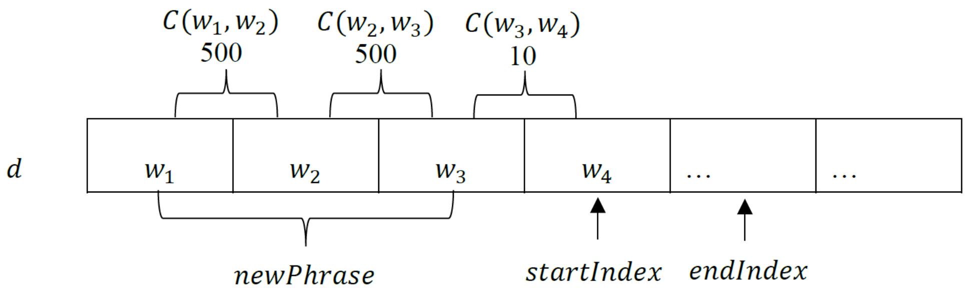


Figure 3: Illustration of the new phrase generation

Fig. 3 is the corresponding illustration for the idea of the new phrase generation. Tab. 1 shows some examples of discovered new phrases by applying Algorithm 1.

Table 1: Examples of new phrases

Id	Original tokens in Chinese along with translations in English	New phrases in Chinese	Translation of new phrases in English
1	新(new)/冠(crown)	新冠	new Coronavirus
2	隔离(isolation)/治疗(therapy)	隔离治疗	isolation therapy
3	疫情(epidemic)/防控(prevention and control)	疫情防控	epidemic prevention and control
4	医务人员(medical staff)/感染 (infection)	医务人员感染	medical staff infection
5	核酸(nucleic acid)/检测(testing)	核酸检测	nucleic acid testing
6	英国(English)/首相(prime minister)	英国首相	British Prime Minister
7	疑似(suspected)/接触者(contacts)/排除 (exclude)	疑似接触者排除	Exclusion of suspected contacts
8	物资(supplies)/缺口(gap)/亟待 (urgently)/支援(support)	物资缺口亟待支援	supplies gap urgently needs support

Take *id 1* as an example, the new phrase “新冠(new Coronavirus)” is unseen in the past, so the word segmentation tool fails to discover this new phrase. If we only model topics with the single word "新(new)" and "冠(crown)" separately, we might lose important public opinions about new Coronavirus. The *id 6* is another case. In common usage, the two words "英国(English)" and "首相(prime minister)" is available for modeling common topics. However, in order to discover bursty topics, it is more reasonable to treat "英国首相(British Prime Minister)" as a phrase, which can promote discover topics about the breaking news about "British Prime Minister was infected with COVID-19".

Then, we aim to discover bursty topics based on PD. Because every pseudo document pd consists of both words and phrases, we use tokens for unified illustration, where $pd_i = [t_1, t_2, \dots t_{|pd_i|}]$. Similar to BBTM, NP-BBTM models topics on co-occurrent token pairs with the assumption that two co-occurrent tokens have a higher probability to share the same topic. **We denote a token pair $\boldsymbol{tp} = (\boldsymbol{t_1}, \boldsymbol{t_2})$, where $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ are two co-occurrent tokens**. Because the number of co-occurrent word pairs is much more adequate than single words in the corpus, which makes NP-BBTM can solve the sparsity problem of short texts from Sina Weibo.

Furthermore, in order to solve the burstiness problem, NP-BBTM involves prior knowledge η for bursty topic discovery, whose definition is similar to BBTM's. Assume that a token pair tp appears $n_{tp}^{\mathcal{T}}$ times in microblogs in the time slice $\mathcal{T}$. It might be used for common chatting or bursty topics. So the count $n_{tp}^{\mathcal{T}}$ can be divided into two parts including $n_{tp,c}^{\mathcal{T}}$ for common usage and $n_{tp,b}^{\mathcal{T}}$ for bursty usage, where $n_{tp}^{\mathcal{T}} = n_{tp,c}^{\mathcal{T}} + n_{tp,b}^{\mathcal{T}}$. Considering the fact that the count for common usage $n_{tp,c}^{\mathcal{T}}$ should keep consistent during several time slices, while the count for bursty usage $n_{tp,b}^{\mathcal{T}}$ should rise steeply in a certain time slice. We can approximately define $n_{tp,c}^{\mathcal{T}}$ by the mean of $n_{tp}^{\mathcal{T}}$ in the last S time slices, denoted as $n_{tp,c}^{\mathcal{T}} = \frac{1}{S}\sum_{s=1}^{S} n_{tp}^{\mathcal{T}-s}$. We use prior knowledge η to reflect the probability of the token pair tp generated from a bursty topic in time slice $\mathcal{T}$, a higher value indicates that a token pair appears more often in this time slice than in others. The calculation is as Eq. (1) shows.

$$\eta_{tp}^{\mathcal{T}} = \frac{\left(n_{tp}^{\mathcal{T}} - n_{tp,c}^{\mathcal{T}}\right)_+}{n_{tp}^{\mathcal{T}}} \qquad (1)$$

where the function $(x)_+ = \max(x, \varepsilon)$ is to avoid zero probability by setting ε as a positive number. By considering the prior knowledge η, the generative procedure of token pairs set $\boldsymbol{TP}$ in a one-time slice of NP-BBTM is as follows.

1) draw a bursty topic distribution $\theta \sim Dir(\alpha)$
2) draw a common usage token distribution $\varphi_0 \sim Dir(\beta)$
3) for each bursty topic $k \in [1, K]$

   draw a token distribution $\varphi_0 \sim Dir(\beta)$
4) for each token pair $tp \in \boldsymbol{TP}$

   draw $e_i \sim Bern(\eta_{tp}^{\mathcal{T}})$

   if $e_i = 0$, draw two tokens $t_1, t_2 \sim Multi(\varphi_0)$

if $e_i = 1$, draw a bursty topic $z \sim Multi(\theta)$ and draw two tokens $t_1, t_2 \sim Multi(\varphi_z)$

where Dir and $Multi$ are the Dirichlet distribution and the Multinomial distribution respectively, which are widely used in topic models. $Bern$ refers to the Bernoulli distribution, which is used to decide whether a token pair belongs to a bursty topic. K is the number of bursty topics, α and β are two hyper-parameters, which should be determined in advance. θ and φ are two parameters to be learned, which refer to the global topic distributions of token pairs and token distributions of topics. We follow the Gibbs Sampling algorithm of BBTM to estimate these two parameters for NP-BBTM, as Eqs. (2) and (3) show.

$$\theta_k = \frac{n_k + \alpha}{n_{\cdot} + K\alpha} \tag{2}$$

$$\varphi_{k,t} = \frac{n_{k,t} + \beta}{n_{k,\cdot} + |V|\beta} \tag{3}$$

where n_k is the number of token pairs assigned to the k-th bursty topic, while $n_{\cdot} = \sum_{k=1}^{K} n_k$, $n_{k,t}$ is the number of token t assigned to the k-th bursty topic, while $n_{k,\cdot} = \sum_{i=1}^{|V|} n_{k,t_i}$. V is the vocabulary of tokens in PD.

In the Gibbs sampling procedure, we model topics on token pairs set $\boldsymbol{TP}$ instead of on documents. In order to infer topic distributions of each document, we can make use of all the tokens, which make up this document. Because PD is generated with new phrases based on D, we can treat the topic distributions of PD as same as that of D. The derivation of the k-th topic's proportion of a document $d \in PD$ is as Eq. (4) shows.

$$p(z_k|d) = p(z_k|pd) = \sum_{t_i} p(z_k|t_i)p(t_i|pd) \tag{4}$$

We can use Bayes formula to calculate $p(z_k|t_i)$, and have $p(z_k|t_i) = \frac{p(z_k)p(t_i|z_k)}{\sum_{k\in K} p(z_{k'})p(t_i|z_{k'})}$, where $p(z_k)$ and $p(t_i|z_k)$ are parameters θ_k and φ_{k,t_i} learned in the Gibbs sampling procedure. We can simply calculate $p(t_i|pd)$ based on the counting statistic, and have $p(t_i|pd) = \frac{n_{t_i}}{|PD|}$, where n_{t_i} is the frequency of token t_i appearing in the pseudo document pd. The outputs of NP-BBTM are two matrices, including a $|D| \times K$ matrix for topic distributions of all documents and a $K \times |V|$ matrix for word distributions of all topics, calculated as Eqs. (5) and (6) show.

$$p(z|D) = p(z|PD) = \left[p(z|pd_1), p(z|pd_2), \dots, p\left(z|pd_{|PD|}\right)\right] \tag{5}$$

$$\varphi = \left[\varphi_{z_1}, \varphi_{z_2}, \dots, \varphi_{z_K}\right] \tag{6}$$

where $p(z|D)$ can be used as feature representations of texts from microblogs, φ can be used to display bursty topics of texts in the form of a set of topic-related tokens.

3.3 Feature learning for images

Images from Sina Weibo can provide auxiliary information about public opinions about COVID-19. However, the number of images is usually of great amount. A common solution is to represent images with features and then retrieve related features according to their similarity. Nowadays, deep learning has been widely used for learning image features. One of the main solutions is pretraining features with labeled datasets such as

ImageNet with CNNs based deep neural networks. This kind of method has two drawbacks in online social media scenarios. Firstly, images from Sina Weibo update very fast and have multiple semantics in each image. However, images in ImageNet are annotated with the single label, which cannot fully reflect the semantics of images. While labeling these images by human kind is expensive both in time and money. Secondly, features pretrained by ImageNet only focus on single modality, which ignores the cross-modal interaction between texts and images in microblogs.

In order to solve these problems, we consider the characteristic that texts and images posted in the same microblogs usually share the same topic, we can make full use of this observation to design new schemes for learning image features. We construct corresponding (text, image) pairs for model training, each pair of which appears in the same microblog. We can apply NP-BBTM proposed in Section 3.2 to get semantic information of texts in the form of topic distributions. Then we set the training target as mapping image features to fit topic distributions of corresponding texts. We call this model as ToDS, as Fig. 4 shows.

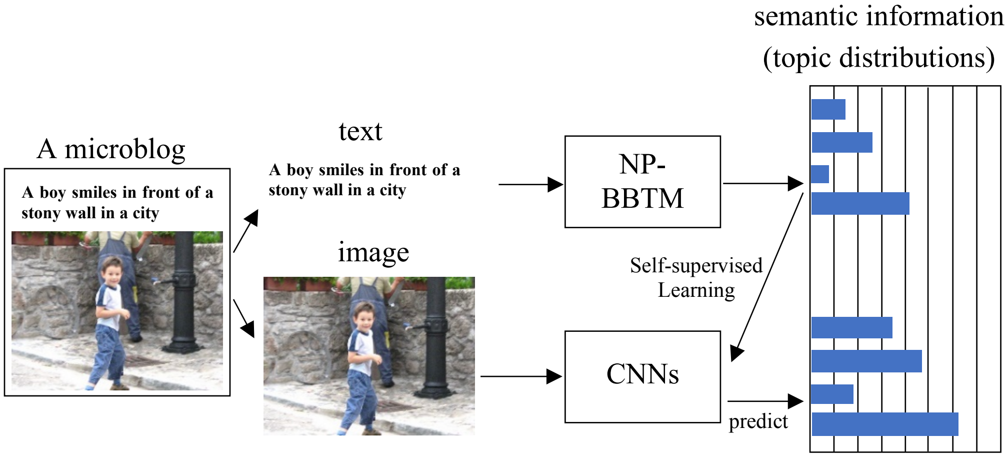


Figure 4: Illustration of self-supervised learning for images from microblogs in ToDS

ToDS does not need labeled data in advance, which belongs to the self-supervised learning. Furthermore, the mapping scheme makes image features and topic distributions of text come to consistent after training, which makes cross-modal interaction possible. In ToDS, we choose ResNet as the neural network structure for learning image features. The inputs of ToDS are original RGB vectors, which can represent their low-level features. Given image-text pairs $X=\{(x_i^v,x_i^t)\}_{i=1}^N$, for each microblog (x_i^v,x_i^t), the topic distribution of x_i^t learned by NP-BBTM is denoted as $\hat{x}_i^t$. We denote the vectors of image x_i^v by mapping with ResNet as $\hat{x}_i^v$, the objective function can be described as Eq. (7) shows.

$$Obj=\arg\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}L(\hat{x}_i^v,\hat{x}_i^t) \tag{7}$$

4 Experiments and analysis

4.1 Experimental settings

We have implemented a web crawler to collect microblogs from Sina Weibo. Because the official API has limitations to pages, we use "pneumonia" as the searching keyword and collect around 100 pages every day. The microblogs for analysis and experiments cover from 2020/1/1 to 2020/3/31, including texts and images. We conduct data preprocessing according to the following steps:

1) Word segmentation: Because microblogs from Sina Weibo are mostly posted in Chinese, we use Jieba[3] for word segmentation.
2) Stop words: We construct a Chinese stop words list based on the widely used stop words list[4] and remove all the stop words from the corpus.
3) Data cleaning: We delete those low-frequency words, which appear in the whole corpus less than 10 times, and delete those documents, which contain less than 3 words.
4) Image resize: We resize all the images into 128×128 for the following experiments.

After data preprocessing, the statistics of our dataset are as Tab. 2 shows.

Table 2: Statistics of Sina Weibo dataset

	2020/1/1 to 2020/1/31	2020/2/1 to 2020/2/29	2020/3/1 to 2020/3/31	Total
number of texts	26,459	29,839	27,838	84,136
number of images	13,258	17,273	15,742	46,273

In order to analyze texts with NP-BBTM, we set parameters as follows. We set the topic model's hyper-parameters $\alpha = 50/K$ and $\beta = 0.01$, which are as same as most topic models. We set the iteration time as 300 for topical inference. We set the co-occurrence threshold $\delta = 300$ for the new phrase generation. We set the positive number $\varepsilon = 0.01$ for calculating the prior knowledge η to reflect the probability of a burstiness token pair.

In order to analyze images with ToDS, we set parameters as follows. We use ResNet as the training networks with learning rate as 0.001, batch size as 16, and dropout rate as 50% to avoid overfitting. We choose Adam as the optimizer and Kullback-Leibler (KL) divergence as the loss function.

4.2 Public opinions with texts

We can use NP-BBTM to discover bursty topics to reveal public opinions about COVID-19. In this experiment, we show the improvement of NP-BBTM compared with BBTM at first, and then show the effectiveness of tracking public opinions about COVID-19 with NP-BBTM.

Topic coherence [Mimno, Wallach, Talley et al. (2011)] is one of the widely used evaluation metrics for the topic model. It measures the extent that the most probable tokens of a topic tend to co-occur within the same documents, which does not rely on

[3] https://pypi.org/project/jieba/

[4] https://github.com/goto456/stopwords

additional corpus for evaluation. Topic coherence can be calculated as follows.

$$Coh = \frac{1}{K}\sum_{z=1}^{K}\sum_{m=2}^{M}\sum_{l}^{m-1} log\,\frac{n_D(t_m^z,t_l^z)+\epsilon'}{n_D(t_l^z)} \qquad (8)$$

where $[t_1^z, t_2^z, \ldots, t_M^z]$ denotes the M most representative tokens of topic z. $n_D(t_l)$ denotes the frequency of token t_l occurs in corpus D. $n_D(t_m, t_l)$ denotes the co-occurrence of token t_m and t_l. Coh is a negative number, where a higher value indicates a better performance. We perform this experiment with $M = 10$ and the number of bursty topics $K = 5$, comparison between BBTM and NP-BBTM is as Fig. 5 shows.

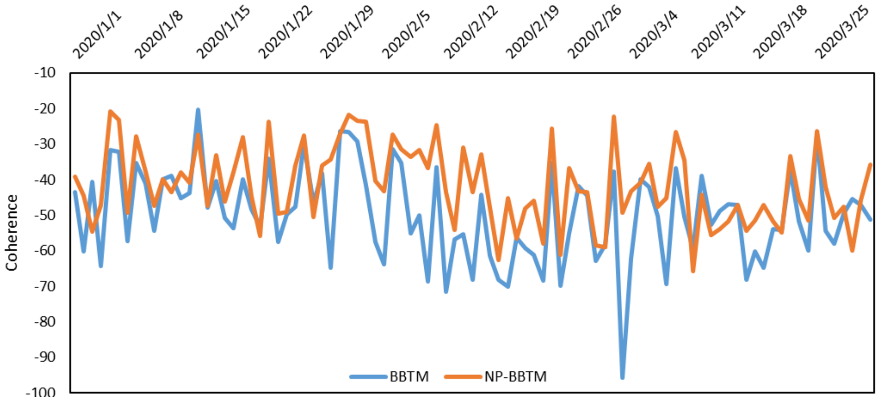


Figure 5: Comparisons of topic coherence between BBTM and NP-BBTM

Fig. 5 shows both models' coherence values of every day. The *x*-axis represents the date (from 2020/01/01 to 2020/03/31), the *y*-axis represents the value of coherence. We can find NP-BBTM performs better than BBTM on most days. Furthermore, the mean average of NP-BBTM's topic coherence is -41.83, which is better than BBTM's -49.94.

We also invite volunteers to artificially judge the performance of topics discovered by two models. They are required to judge whether a topic discovered by a topic model belongs to bursty topics, if the answer is yes, then this topic can achieve one vote from volunteers. The average score of voting is as Tab. 3 shows. We can find that NP-BBTM has a better ability to discover more bursty topics than BBTM.

Table 3: The number of bursty topics detected by both models

	1/1 to 1/15	1/16 to 1/31	2/1 to 2/15	2/16 to 2/29	3/1 to 3/15	3/16 to 3/31	**Total**
BBTM	47	54	51	57	63	61	**333**
NP-BBTM	52	57	61	60	67	64	**361**

Tab. 4 shows some examples of bursty topics discovered by NP-BBTM. For example, Topic 1 in 2020-3-27 discovered by NP-BBTM is about the fact that "British Prime Minister Boris Johnson tests positive for coronavirus", which is consistent with the breaking news in real life[5].

[5] https://www.pressherald.com/2020/03/27/british-prime-minister-boris-johnson-tests-positive-for-virus/#

Table 4: Examples of bursty topics discovered by NP-BBTM

Date	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
2020/3/3	新冠 (COVID)	医生 (Doctor)	美联储 (Federal Reserve)	市场 (Market)	乘客 (Passenger)
	纽约州 (New York state)	去世 (Death)	基点 (Basis point)	消杀 (Germicidal)	游轮 (Cruise)
	肺炎病例 (Pneumonia cases)	李文亮 (Li Wenliang)	降息 (Cut interest rates)	华南海鲜 (Huanan Seafood)	德国 (Germany)
	纽约 (New York)	梅仲明 (Mei Zhongming)	下调 (Come down)	消毒 (Disinfect)	1200 名 (1200 people)
	州长 (Governor of states)	武汉市中心医院 (Wuhan Central Hospital)	基准 (Benchmark)	疾控 (Disease control)	检测 (Testing)
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
2020/3/27	新冠 (COVID)	开学 (Start of school)	人员 (People)	确诊病例 (Confirmed cases)	西班牙 (Spain)
	英国首相 (British Prime Minister)	年级 (Grade)	入境 (Immigration)	美国 (USA)	逝世 (Death)
	肺炎检测 (Pneumonia testing)	错峰 (Avoid peaks)	费用 (Fee)	新增 (New)	皇室 (Royal)
	呈阳性 (Positive)	中小学 (Elementary and secondary schools)	核酸检测 (Nucleic acid testing)	国家 (Country)	成员 (Member)
	英国 (England)	湖南 (Hunan Province)	陕西省 (Shanxi Province)	全球 (Global)	新冠肺炎 (COVID)

Furthermore, NP-BBTM can be used to conduct evolution comparisons between bursty topics and common topics. We determine whether two topics in different days describe the same public opinion based on the word distributions of topics learned by NP-BBTM. We use KL divergence to measure the distance between the probability distributions. A smaller value indicates a higher probability that two topics describe the same public opinion. We take topics discovered in 2020/2/19 as an example. We observe the bursty topic about "tourists in the Diamond Princess Cruise started to disembark", whose evolution trend is the blue line in Fig. 6, and observe the common topic about "COVID-19 in Wuhan", whose evolution trend is the orange line in Fig. 6. We can find that the bursty topic suddenly appears in 2020/2/19 because topics discovered in other days have large KL divergence with it. However, the change of the common topic appears far more slowly and smoothly. These observations not only show the effectiveness of discovering bursty topics by NP-BBTM but also indicate that NP-BBTM can be applied to automatically discover and track public opinions about COVID-19.

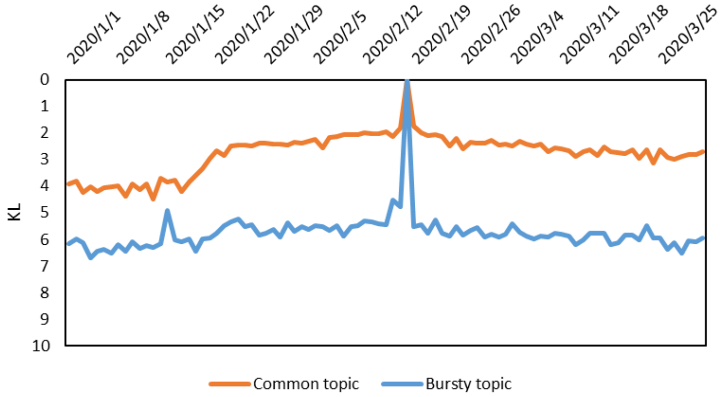


Figure 6: Evolution comparison between bursty topic and common topic

4.3 Public opinions with images

Images can be used as supplements for revealing public opinions in a more vivid way. We can use ToDS to learn feature representations for images and then retrieve related images to represent public opinions about COVID-19 by using given images or texts. We apply ResNet50 as the basic deep neural network structure for training ToDS, which aims to fit the image features similar to semantic topic distributions of texts posted in the same microblog. We randomly choose 70% images of each day as training data, 15% as validation data, and the rest 15% as testing data. We use accuracy to measure how the learned image features fit the corresponding topic distributions. Parameter settings are introduced in Section 4.1. According to the performance of training and validation data, we find when the iteration time is around 400, ToDS comes to convergence, as Fig. 7 shows.

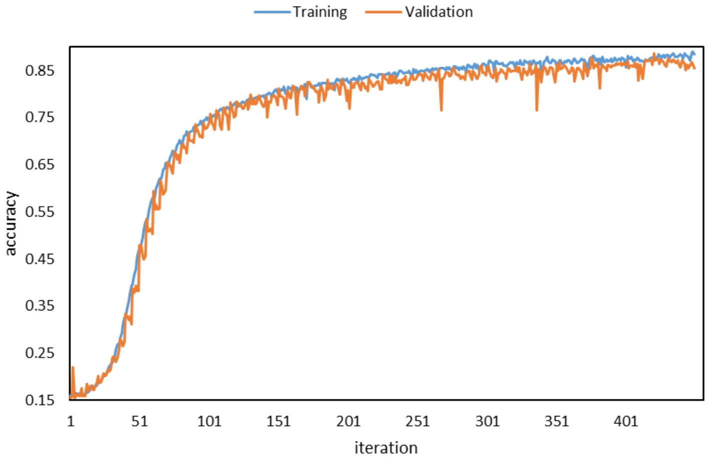


Figure 7: Performance of training and validation with the change of iteration times

We use the ToDS model trained with these settings to predict the testing data for each day separately. Experimental results cover from 2020/1/1 to 2020/3/31, as Fig. 8 shows. The average accuracy of 91 days is 87.87%, which shows that the ToDS model has a good ability to learn features for unseen images.

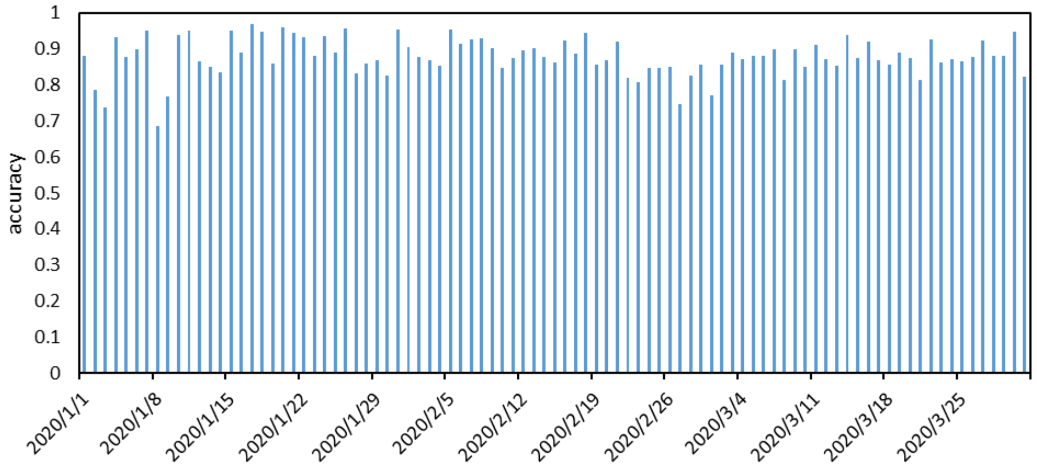
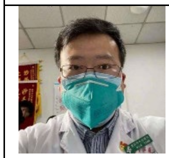
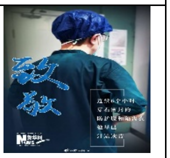


Figure 8: Performance of testing data from 2020/1/1 to 2020/3/31

We can use cosine similarity between two vectors to measure whether they are related for retrieval, which can improve the readability and information of public opinion about COVID-19. Because the features learned by ToDS naturally have connections with texts, we can retrieve related images either with given images or with given texts. Fig. 9 and Fig. 10 have shown several retrieval examples of related images. The filename of each image is organized according to the following rule: the first number refers to post data (0206 means posted on 6th Feb), the second and the third numbers refer to the unique id that records which microblog the current image is posted with.

Given Image	1	2	3	4	5
					
0206_942_0	0206_943_0	0206_986_0	0206_988_0	0206_940_0	0130_506_0

(a)

Given Image	1	2	3	4	5
加藤厚労相 「前の状況とは異なっている」 ⇒事実上の「流行」を初めて認める	加藤厚労相 「前の状況とは異なっている」 ⇒事実上の「流行」を初めて認める	加藤厚労相 「前の状況とは異なっている」 ⇒事実上の「流行」を初めて認める		加藤厚労相 「前の状況とは異なっている」 ⇒事実上の「流行」を初めて認める	加藤厚労相 「前の状況とは異なっている」 ⇒事実上の「流行」を初めて認める
0215_1_0	0215_326_0	0215_33_0	0215_557_0	0215_557_4	0130_746_0

(b)

Given Image	1	2	3	4	5
0303_293_0	0210_821_1	0303_358_0	0321_89_0	0307_37_0	0321_210_2

(c)

Figure 9: Examples of related images retrieved by given images

Fig. 9 shows three examples retrieved by given images and we display the top five similar images. Fig. 9(a) shows the bursty topic about the death of doctor Wenliang Li who was infected by COVID-19, which is a widely discussed event in 2020/02/06. Sugfigure 9(b) shows the breakup news that the Japanese minister of Health Kato held a press conference to announce the virus testing for passengers on the Diamond Princess Cruise. The results of these two examples are quite satisfying while the third example in Fig. 9(c) has several irrelevant images in it. We use the image about doctor Wenhong's Zhang interview for retrieval and only get one most similar image in the top 5 candidates. One of the possible reasons is that the topic of this given image is not quite bursty and important on this day.

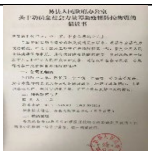
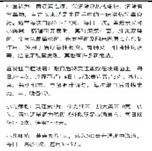
We also conduct experiments to show the performance of retrieving related images with given texts. Firstly, we use the trained NP-BBTM model to infer topic distributions of given texts, which can be regarded as the features of texts. Then we retrieve related images, which are similar to this text. Fig. 10 shows three examples with the top five similar images.

Given Texts	1	2	3	4	5
新冠肺炎疫苗研发在武汉北京等地紧锣密鼓推进 The research of COVID-19's vaccine is in full progress in Wuhan and Beijing					
	0218_315_0	0218_393_0	0218_178_0	0218_708_0	0218_664_0

(a)

Given Texts	1	2	3	4	5
中国又同 19 个国家开防疫视频会议… China has hold a video conference on epidemic prevention with another 19 countries					
	0320_629_1	0320_761_3	0320_761_1	0320_695_1	0320_695_3

(b)

Given Texts	1	2	3	4	5
河北新闻聚焦: 为河北医疗队点赞 … Hebei News Spotlight: Praise the Hebei Medical Team …					
	0303_379_4	0303_379_5	0128_509_0	0205_815_0	0328_52_0

(c)

Figure 10: Examples of related images retrieved by given texts

We observe that Fig. 10(a) shows five same images about the research of COVID-19's vaccine from different microblogs. We check the original microblogs and find that these related microblogs are widely reposted by different users, which can reveal public opinions about today's bursty event. Fig. 10(b) is another widely discussed topic, which is about China sharing experience of preventing COVID-19 with 19 countries online, the retrieved images can have a better auxiliary description for this public opinion. Fig. 10(c) is a less effective example. We use text about "Praise Hebei Medical Team" for retrieval, the first and second images are most related, the last three images are also about fighting with COVID-19 but are less relevant to the given text. This might because the training data for this topic is not enough, and this is also our target to improve in future work.

5 Conclusions

This paper applies AI techniques to analyze and discover public opinions about COVID-19 from the famous Chinese online social media Sina Weibo, in order to help citizens and governments track the development of the pandemic in time. Considering the time costs and expenses are rather huge by artificial processing, we propose a unified framework to deal with multi-modal data in microblogs without human labeling, including the NP-BBTM model for discovering bursty topics for texts and the ToDS model for retrieving related images with both given images and texts.

The main contributions include: 1) we have collected a dataset in Chinese about COVID-19 from Sina Weibo, which can be shared for academic researches. 2) we have improved the BBTM model by generating new phrases and propose an NP-BBTM model, which is suitable to deal with texts produced by breaking events. 3) we have proposed a self-supervised image feature learning model named ToDS, which makes full use of the semantic relationship between texts and images in microblogs, and is suitable for retrieving related images in the online social media scenarios.

For future work, we would like to implement a prototype based on the proposed framework and contribute to tracking online public opinions about COVID-19.

Acknowledgement: We would like to sincerely thank our colleagues, Bojian Hou, Yi Zhang, Kai Zhu, Professor Chongjun Wang, and Professor Junyuan Xie from Nanjing University, for their kindly support and guidance. We would also like to thank all the medical staff working on the front line around the world.

Funding Statement: This paper is supported by the Fundamental Research Funds for the Central Universities [No. JUSRP12021].

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Alghamdi, R.; Alfalqi, K. (2015): A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147-153.

Chen, X.; Chang, T.; Wang, H.; Zhao, Z.; Zhang J. (2020): Spatial and temporal analysis on public opinion evolution of epidemic situation about novel coronavirus pneumonia based on micro-blog data. *Journal of Sichuan University (Natural Science Edition)*, vol. 57, no. 2, pp. 409-416.

Gomez, R.; Gomez, L.; Gibert, J. (2018): Learning to learn from web data through deep semantic embeddings. *Proceedings of ECCV-2018*, pp. 514-529.

Gomez, R.; Gomez, L.; Gibert, J. (2019): Self-supervised learning from web data for multimodal retrieval. *Multimodal Scene Understanding*, pp. 279-306.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Jiang, X.; Coffee, M.; Bari, A.; Wang, J.; Jiang, X. et al. (2020): Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials & Continua*, vol. 63, no. 1, pp. 537-551.

Jin, S.; Wang, B.; Xu, H.; Luo, C.; Wei, L. et al. (2020): AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system in four weeks. *MedRxiv*.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.

Li, Z.; Du, J.; Cui, W.; Zhu, P. (2018): User interaction based bursty topic model for emergency detection. *Proceedings of Chinese Intelligent Systems Conference*, pp. 11-21.

Ma, K.; Yu, Z.; Ji, K.; Yang, B. (2019): Stream-based live public opinion monitoring approach with adaptive probabilistic topic model. *Soft Computing*, vol. 23, no. 16, pp. 7451-7470.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; McCallum, A. (2011): Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262-272.

Nanni, L.; Ghidoni, S.; Brahnam, S. (2017): Handcrafted *vs*. non-handcrafted features for computer vision classification. *Pattern Recognition*, vol. 71, pp. 158-172.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Stebbing, J.; Phelan, A.; Griffin, I.; Tucker, C.; Oechsle, O. et al. (2020): COVID-19: combining antiviral and anti-inflammatory treatments. *The Lancet Infectious Diseases*, vol. 24, no. 4, pp. 400-402.

Wang, W.; Jiang, Y. B.; Luo, Y. H.; Li, J.; Wang, X. et al. (2019): An advanced deep residual dense network (DRDN) approach for image super-resolution. *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1592-1601.

Wang, W.; Li, Y. T.; Zou, T.; Wang, X.; You, J. Y. et al. (2020): A novel image classification approach via Dense-MobileNet models. *Mobile Information Systems*, https://doi.org/10.1155/2020/7602384.

Xu, K.; Qi, G.; Huang, J.; Wu, T.; Fu, X. (2018): Detecting bursts in sentiment-aware topics from social media. *Knowledge-Based Systems*, vol. 141, pp. 44-54.

Yan, X.; Guo, J.; Lan, Y.; Xu, J.; Cheng, X. (2015): A probabilistic model for bursty topic discovery in microblogs. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 353-359.

Yan, X.; Guo, J.; Lan, Y.; Cheng, X. (2013): A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1445-1456.

Zhang, J. M.; Jin, X. K.; Sun, J.; Wang, J.; Sangaiah, A. K. (2018): Spatial and semantic convolutional features for robust visual object tracking. *Multimedia Tools and Applications*, https://doi.org/10.1007/s11042-018-6562-8.