Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-04551-2 OPEN

A joint-training topic model for social media texts

Simeng Qin[1,2], Mingli Zhang[3], Haiju Hu[3] & Gang Li[1,2]✉

The burgeoning significance of topic mining for social media text has intensified with the proliferation of social media platforms. Nevertheless, the brevity and discreteness of social media text pose significant challenges to conventional topic models, which often struggle to perform well on them. To address this, the paper establishes a more precise Position-Sensitive Word-Embedding Topic Model (PS-WETM) to adeptly capture intricate semantic and lexical relations within social media text. The model enriches the corpus and semantic relations based on word vector similarity, thereby yielding dense word vector representations. Furthermore, it proposes a position-sensitive word vector training model. The model meticulously distinguishes relations between the pivot word and context words positioned differently by assigning different weight matrices to context words in asymmetrical positions. Additionally, the model incorporates self-attention mechanism to globally capture dependencies between each element in the input word vectors, and calculates the contribution of each word to the topic matching performance. The experiment result highlights that the customized topic model outperforms existing short-text topic models, such as PTM, SPTM, DMM, GPU-DMM, GLTM and WETM. Hence, PS-WETM adeptly identifies diverse topics in social media text, demonstrating its outstanding performance in handling short texts with sparse words and discrete semantic relations.

[1] School of Management, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei, China. [2] School of Business Administration, Northeastern University, Shenyang, Liaoning, China. [3] College of Economy and Management, Yanshan University, Qinhuangdao, Hebei, China. ✉email: ligang@neuq.edu.cn

Introduction

The rapid growth of social media has revolutionized the way of communication, generating an immense volume of social media text that reflects diverse perspectives and insights. Platforms such as Twitter, Facebook, and Weibo, in particular, facilitate the widespread dissemination of brief yet information-rich messages, which pose significant analytical challenges due to their brevity and high variability (Qin et al., 2023). Consequently, the analysis of social media text has become an increasingly critical task.

Scholars have primarily approached this analysis from two aspects: sentiment analysis and topic analysis. For sentiment analysis, scholars have conducted extensive research. Faced with short and ambiguous Twitter messages, Bravo-Marquez et al. proposed Annotate-Sample-Average (ASA), a distant supervision approach that utilizes these two resources to generate synthetic training data for sentiment polarity classification. This method yields a classifier that significantly outperforms those trained on tweets annotated with emoticons and those trained without any sampling and averaging, based on the polarity of their words (2016). Wang et al. advanced sentiment analysis by integrating textual information of Twitter messages with sentiment diffusion patterns to achieve superior performance in analyzing on Twitter data (2019). Kumar et al. applied sentiment analysis to short texts using Enhanced Vector Space Model combined with a Hybrid Support Vector Machine classifier. And to further enhance the accuracy of this method when applied to short texts, Kumar extended the sentiment dictionaries through the expansion of Stanford's GloVE tool during the sentiment analysis (2024).

For topic analysis, while scholars have focused on identifying topics within short texts (Xu et al., 2019), there is a paucity of topic models tailored specifically for social media text, which is characterized by brevity and discreteness. Compared to typical short texts, social media text exhibits greater discreteness, due to that contemporary social media text is carried on the social network platforms, which aggregate a wide array of internet slang and buzzwords (Zhang et al., 2021; Qin and Zhang, 2024). Consequently, the opinions shared on these platforms tend to be more informal and unexplainable, making it difficult for conventional models to perform accurate topic extraction and analysis. Thus, there is a pressing need for more sophisticated models capable of addressing the unique characteristics of social media text and providing deeper insights into the underlying topics.

To address these issues and accurately extract topics from social media text, the study introduces the short-text topic model PS-WETM. This model primarily incorporates a novel devised position-sensitive word vector training model and a self-attention mechanism. These innovations aim to capture the unique dynamics and semantic intricacies inherent more effectively in the brief yet complex expressions of social media text. In this short-text topic model, we utilize word vector similarity to enrich the corpus and semantic relations, thereby achieving dense word vector representations. And it employs the proposed position-sensitive word vector training model to differentiate relations between the pivot word and context words in various positions by assigning symmetrical weight matrices to words positioned symmetrically. Additionally, the model incorporates a self-attention mechanism to globally capture dependencies between each element in the input word vectors and calculate the contribution of each word to the topic matching performance accuracy. This method effectively captures the relationships within texts that are more discrete and contain scarce semantic relations. To evaluate the accuracy of short-text topic model PS-WETM, we conduct comparative analyses with six other topic models, utilizing perplexity and topic coherence as metrics. Finally, we select "COVID-19 explosion" as a case study, crawling relevant microblogs as data to verify the accuracy of PS-WETM, and employ Word Cloud visualizations and sentiment analysis of these microblogs to further evaluate the performance of the PS-WETM model and trace the evolving tendencies of their topics.

The contributions of this paper:

- To counterbalance the inherent scarcity of semantic relations in social media text, we implement a short text extension strategy utilizing word vector similarity, thereby enriching the semantic information conveyed by these texts.
- It defined a position-sensitive word vector training model equipped with a customized positional judgment mechanism, which considers the impact of word position on semantic relations, addressing the problem that current word vector training models ignore the position of context words. This innovation allows for a more nuanced understanding of word semantics, considering not just the words themselves but their specific positions within a text, thereby enhancing the model's ability to accurately interpret and process language.
- To further enhance the performance of topic matching for social media text characterized by their conciseness and discreteness features, the model employs self-attention mechanism to post-process the extracted sentence features, which is capable of globally discerning internal semantic relations within a sentence and proactively determines the contribution of each word vector to the accuracy of topic matching. This approach leads to a more precise and representative depiction of word features, enhancing the model's ability to accurately reflect the nuanced dynamics of social media text.

Literature review

Topic models for document. LDA is a basic topic model for identifying latent topics, playing a critical role in text classification, topic detection, evolution tracking, emotion analysis, and other tasks. However, LDA exhibits certain limitations, particularly when applied to document sets rich in metadata, where its performance often proves suboptimal. To overcome the limitations, researchers have proposed the Structural Topic Model (STM). STM extends the basic LDA framework by incorporating document-level covariates into the modeling process (Roberts et al., 2016). These covariates can include any attributes of the documents, such as publication date, author identity, or other contextual information. By integrating these covariates, STM captures not only the latent topic structures within documents but also the influence of covariates on topic prevalence and content. Nevertheless, both LDA and STM demand significant computational resources and and substantial time investment for training document weights when identifying topics from massive corpus (Jiang et al., 2019; Balikas et al., 2016; Huang et al., 2020). Moreover, their probabilistic foundations, adhering to the Bag of Words model and presuming word independence, diverge from the complexities of natural language (Bastani et al., 2019; Meng et al., 2020; Keya et al., 2023). Additionally, it ignores word order, resulting in poor semantic coherence and topic interpretability (Salehi et al., 2015; Ruas et al., 2019).

In response to these challenges, researchers have introduced several enhanced topic models incorporating word or document vectors (Li et al., 2020; Ma et al., 2019), and topic models combined with neural networks, which markedly improve semantic coherence in documents (Brown et al., 2020; Bender et al., 2021). For topic models incorporating word or document vectors (Steuber et al., 2021; Kumar et al., 2022), a notable advancement is the LDA2vec model. This model integrates Word2vec with LDA, where Word2vec utilizes a neural network framework to map input word vectors, and LDA employs probabilistic distributions for training topic weights (Moody Christopher, 2016; Mikolov et al., 2013; Le and Mikolov, 2014).

For topic models combined with neural network, scholars fused neural network with topic models to model topic distributions and learn topic representations (Hofmann, 1999; Blei et al., 2010; O'Callaghan et al., 2015; Belford et al., 2018; Bruni et al., 2014; Shahriare Satu et al., 2021; Chaturvedi et al., 2019). Specifically, Peng et al. introduced the Neural Sparse Topical Coding (NSTC), an advancement of the Sparse Topical Coding (STC) model (2018). NSTC addresses the issue of sparsity and demonstrates the versatility of neural networks with three extended applications without the need for re-deriving inference algorithms. In a further innovation, He et al. proposed a two-phase neural embedding network with redundancy-aware graph-based ranking, which optimizes the identified topics with fewer yet more representative terms. This approach the integrality and fidelity of topics while demonstrating how pre-trained neural embedding can be effectively applied to automatic topic labeling tasks (2021). Gupta et al. introduced a novel neural topic model named Discrete-Variation-Inference-based Topic Model. This model acquires dense topic embeddings that are homomorphic to word embeddings through discrete variational inference, conceptualizing words as mixtures of topics (2023). To further improve semantic relevance, scholars have combined the BERT pre-trained language model with topic models to extract latent topics from short texts, resulting in the BERTopic model (Grootendorst, 2022). The BERTopic model generates document embeddings using pre-trained transformer-based language models, clusters these embeddings, and finally generates topic representations with the class-based TF-IDF. Experimental results indicate that the BERTopic model performs exceptionally well in topic mining.

Topic models for short texts. Whereas social media text on platform like Weibo and Twitter typically manifests as short texts, the lexical characteristics of social media text exhibit greater dispersion compared to regular short texts, presenting significant challenges in the effective extraction of latent topics. These texts often amalgamate a broad range of internet slang and popular expressions, rendering the opinions shared on social network platforms more informal and less interpretable. On the surface, adjacent words may seem unconnected, yet within the context of the entire sentence, they frequently bear profound associations. Additionally, the limited number of words and the insufficiency of semantic information in these texts lead to sparse feature vectors. Consequently, standard document topic models are ill-suited for analyzing such short, semantically sparse texts as those found in social media.

Addressing this issue, several researchers have modeled texts as networks, wherein words are represented as nodes interconnected based on textual similarity. For instance, Machicao et al. characterized text networks to grasp informative spatio-temporal patterns and attributions of texts by considering both the topological and dynamic properties of networks (2018). Furthermore, one commonly method to enhance the performance of short-text topic mining is clustering short text via neural networks. Xu et al. have proposed a clustering technique via convolutional neural networks, which optimizes clustering by imposing constraints on learned features through a self-taught learning framework, eschewing the need for external labels (2015). Moreover, some scholars have attempted to extend or aggregate short text to solve its problem of sparse and unbalanced features. For example, the Bi-term Topic Model (BTM) was introduced to more effectively capture the co-occurrence relationships between words in short texts (Yan et al., 2013), which utilizes word pairs as the fundamental unit for modeling. However, BTM only considers the frequency of bi-term, neglecting the latent semantic relations between them, which may result in semantically similar words being categorized under different topics. Addressing this limitation, Li et al. improved the BTM for short texts and proposed the Latent Semantic Augmented Bi-term Topic Model (LS-BTM) based on BTM, which incorporates semantic information as prior knowledge to infer topics more accurately (2018). Beyond these advancements, researchers have explored incorporating external word-associated knowledge and prior understanding into short texts to improve the coherence of topic modeling. For example, Dieng et al. proposed Embedded Topic Model (ETM) to extract latent topics from short texts, employing a categorical distribution for each word. Within this distribution, the natural parameter is the inner product between word embeddings and specific topic embeddings. ETM has demonstrated the capability to uncover explainable topics within large vocabularies containing rare and stop words, thus outperforming traditional models like LDA in terms of topic quality and predictive performance (2020). Furthermore, another effective topic model for short texts is the fuzzy bag-of-topics model, developed from the fuzzy bag-of-words model (FBoWC). This model utilizes words communities that provide greater coherence than word clusters in FBoWC, replacing word clusters as basis terms in text vector (Jia and Li, 2018).

The above topic models for short texts present better performance compared to traditional topic models for documents. However, their effectiveness diminishes when applied to social media text, primarily due to their inability to specify unique positions of context words, the limited semantic content, and discrete semantic relations. To address these challenges, this paper introduces a joint-training short-text topic model — Position-Sensitive Word-Embedding Topic Model (PS-WETM). The model involves a proposed position-sensitive word vector training model and self-attention mechanism, which together facilitate a more comprehensive capture of global semantic relations within sentences. Additionally, it accurately evaluates the exact contribution of each context words in different positions, thereby significantly enhancing the model's ability to precisely identify topics within social media text.

Methodology

Short text extension. To enhance the accuracy in identifying latent topics from brief contexts on platforms like Weibo and Twitter, we implement text extension techniques based on the semantic similarity of word vectors.

Initially, we generate word vectors using Word2vec. We select a word w from the corpus and verify whether its corresponding word vector is present within the set of generated word vectors. If the word vector for w is identified, we proceed to calculate the similarity between w 's word vector and the vectors of all other words in the corpus. Subsequently, we fill the new text with word w and the first k words with the highest similarity. The semantic similarity between words w_1 and w_2 is calculated as:

$$\text{similarity}\,(w_1, w_2,) = \frac{v_1 \times v_2}{||v_1|| \times ||v_2||} \tag{1}$$

where v_1 and v_2 are their word vectors. $v_1 \times v_2$ denotes the dot product of two words' vectors, and $||v_1||$ represents the length of modulus.

Table 1 The similarities of exampled words.

Example word	epidemic		Wuhan		love	
Number	**words**	**similarity**	**words**	**similarity**	**words**	**similarity**
1	prevention	0.999764	Hunan	0.999746	like	0.999731
2	pandemic	0.999678	Haicang	0.999622	happy	0.999727
3	contagion	0.999671	Market	0.999598	warmth	0.999699
4	pneumonia	0.999653	enterprise	0.999593	help	0.999698
5	infection	0.969667	Yuhua	0.999590	hope	0.999694

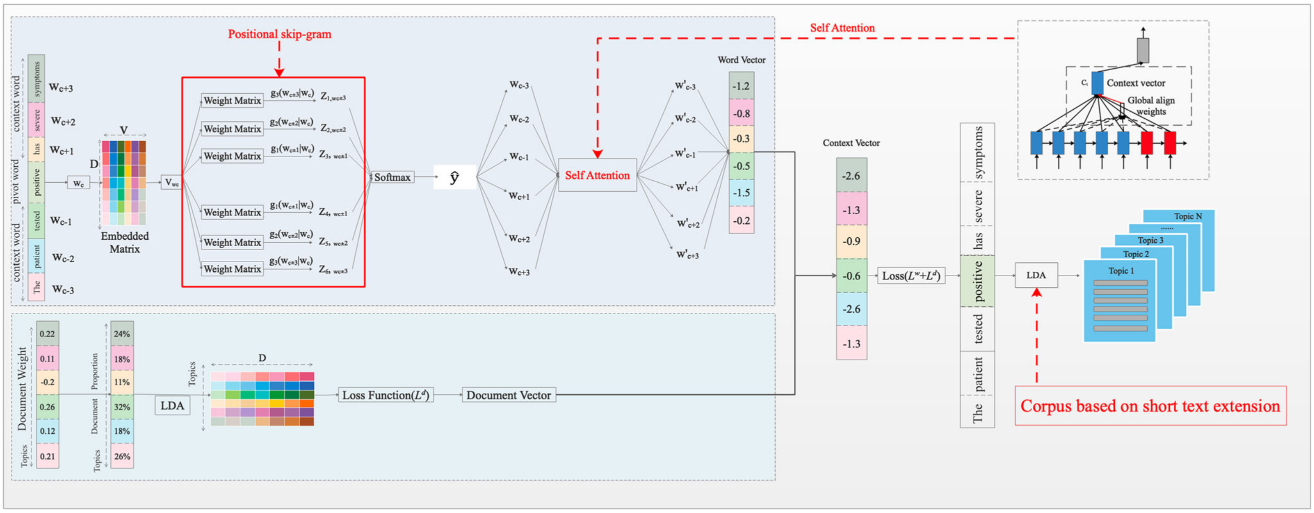

Fig. 1 The framework of the proposed short-text topic model PS-WETM. A framework of the proposed short-text topic model, consisting of the positional-sensitive word training model, the document weight training model, and the self-attention mechanism.

In this paper, we utilize *genism* in *Pycharm* to calculate the similarity of words. And we use microblogs of event COVID-19 in *Sina Weibo* as corpus. The total number of words in the corpus is 188583, the number of unique vocabularies is 5929 and the dimension of feature vectors is 200. It uses negative sampling to get loss function, where negative samples = 5, size of window = 5. The top five words with the highest similarity of the example words are shown in Table 1.

The proposed short-text topic model PS-WETM. In this chapter, we propose a joint-training topic model PS-WETM, which integrates a proposed position-sensitive word vector training model, self-attention mechanism, document proportion and topic matrices obtained by LDA, aimed at enhancing the focus on contextual relationships between words and their associations with topics within documents. The proposed position- sensitive word vector training model in the PS-WETM short-text topic model assigns varying weight matrices to context words based on their relative positions to the pivot word, under the assumption that context words at symmetric positions share a comparable degree of semantic relation with the pivot word. With the proposed position- sensitive word vector training model, we extract sentence features encompassing the feature vectors of each word. In order to better characterize the feature vectors of each word, we employ a self-attention mechanism to obtain the degree to which each word's features in a sentence contribute to the accuracy of topic matching and to adjust the feature representations accordingly, so that not only the semantic information of the word itself is taken into account, but also the internal relations of the sentence in which the word is embedded, so as to achieve the purpose of improving the performance of topic matching. It is worth noting that the self-attention mechanism only updates the training parameters by concentrating on its own information and does not require additional information, thus reducing the computational complexity of the model and making up the defects of the proposed position- sensitive word vector training model. Subsequently, by incorporating the word vector got from the self-attention mechanism with the document vector, we generate the context vector that is used in the topic model. Together, these components form the joint-training short-text topic model PS-WETM. Ultimately, we investigate the perplexity and topic coherence of the PS-WETM, comparing its performance with recent topic models under the data of online social media text. The framework of the proposed short-text topic model PS-WETM is as Fig. 1.

The proposed position-sensitive word vector training model. In this paper, we described a proposed position-sensitive word vector training model to solve the constraint in prototypical word vector training models that it ignores the exact position of each context word and treats all context words equally. The proposed position-sensitive word vector training model operates on the premise that context words in symmetric positions relative to the pivot word share an equivalent degree of semantic relationship with it. It initially assigns different weight matrices to context words in asymmetric positions, and then judges the direction of context words, thereby differentiating the varying impacts of different context words in different positions on the pivot word and distinguishing the semantic relations between context words and the pivot word.

In comparison with the skip-gram word vector training model, the proposed position- sensitive word vector training model

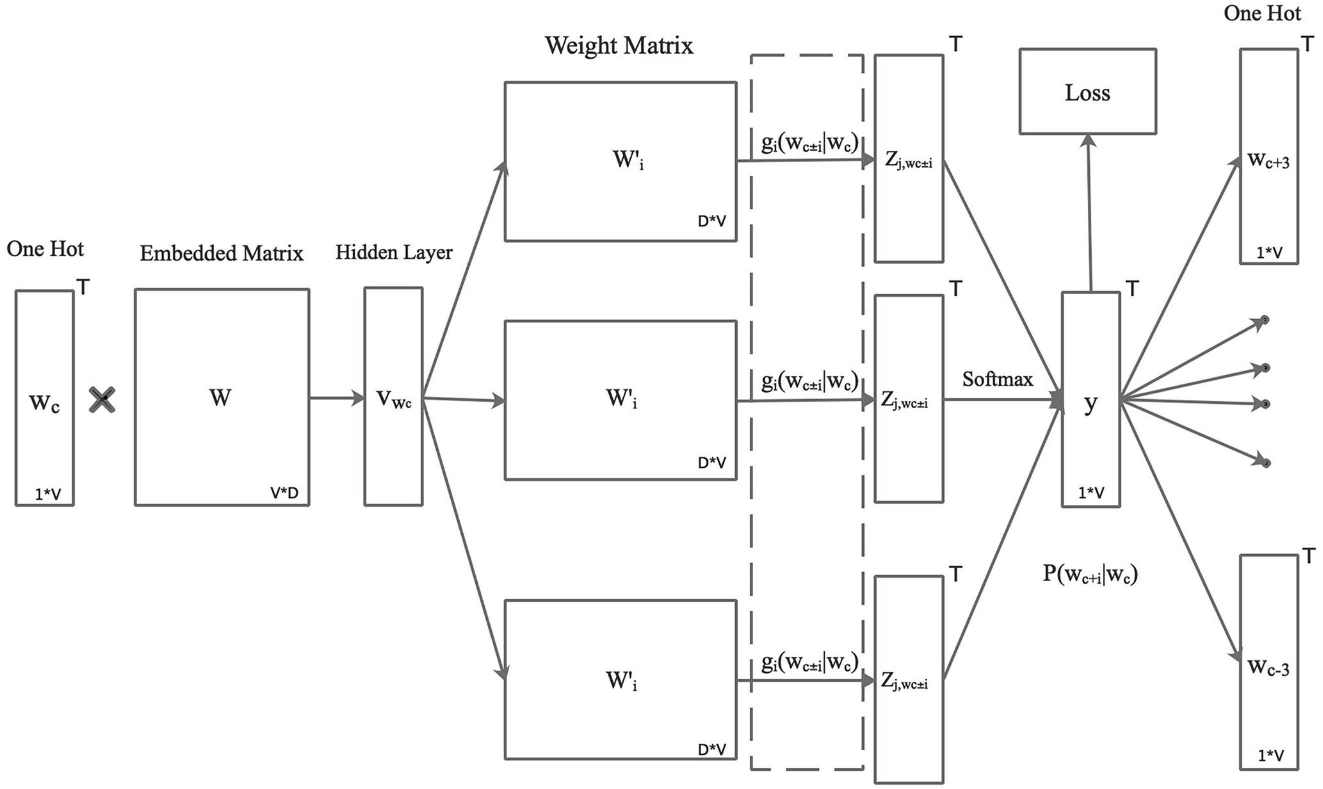

Fig. 2 The framework of the proposed position-sensitive word vector training model. The training process of the position-sensitive word vector training model, depicting how to differentiate the distance between the pivot word and context words positioned differently.

considers the exact positions of different context words, resulting in more concise performance. And compared with the structured skip-gram, it reduces the computation complexity. For instance, if a typical skip-gram has p parameters and the size of window is m, the structured skip-gram will contain $2\,mp$ parameters while the proposed position-sensitive word vector training model contains only mp parameters. The proposed position-sensitive word vector training model is described in Fig. 2.

V is the number of vocabulary and D is the number of features. The embedded matrix has a size of $V \times D$, whereas the weight matrix has a size of $D \times \mathrm{V}$. In the proposed position-sensitive word vector training model, we proposed a new loss function based on the relative position of context words to the pivot word to update the embedded matrices and the weight matrices.

The proposed loss functions. In the proposed PS-WETM model, the dictionary is stored by hash algorithm. We generate an integer array called $\boldsymbol{voc_{hash}}$, with each element initialized to -1. The size of this array is denoted as $\boldsymbol{voc_{hash\_size}}$, which is calculated as:

$$voc_{hash}[h_v(w_j)] = j \tag{2}$$

$h_v\ (w_j)$ signifies the hash value of word w_j. We calculate hash value $h_v\ (w_j)$, and then we find the position of word w_j. If voc_{hash} $[h_v\ (w_j)] = -1$, it indicates that word w_j has not been included in dictionary V. Otherwise, compare w_j with $w_{voc_{hash}[h_v(w_j)]}$, where $w_{voc_{hash}[h_v(w_j)]}$ refers to the word with index $voc_{hash}[h_v(w_j)]$. If w_j and $w_{voc_{hash}[h_v(w_j)]}$ are the same, we set the index of w_j to $voc_{hash}[h_v(w_j)]$.

During the construction of the dictionary, we selectively remove both low-frequency and high-frequency words. If the current scale $|V_{current}|$ of dictionary $V_{current}$ satisfies $|V_{current}| > 0.7voc_{hash\_size}$, we proceed to eliminate words whose frequencies fall below the preset minimum threshold. Additionally, high-frequency words, such as *is, are* will also be deleted. To effectively manage high-frequency words, we adopt the subsampling approach proposed by Mikolov (Mikolov et al., 2013), which involves setting a probability $prob(w)$ for each word, guiding the decision of whether to keep or discard these words. We set the probability $prob(w)$ as:

$$prob(w) = 1 - \left(\sqrt{\frac{t}{f(w)}} + \frac{t}{f(w)}\right) \tag{3}$$

$$f(w) = \frac{[counter(w)]^{\frac{3}{4}}}{\sum_{u \in v}[counter(w)]^{\frac{3}{4}}} \tag{4}$$

where $f\,(w)$ refers to the frequency of word w. Accordingly, word w will be discarded with the probability $prob(w)$. Given the hidden layer v_{w_c}, the proposed position-sensitive word vector training model will generate the predicted word vector $\hat{y}$. We first get the hidden layer through embedded matrix. v_{w_c} is calculated as:

$$v_{w_c} = x_{w_c}\mathrm{W} \tag{5}$$

where x_{w_c} refers to the one-hot vector of the pivot word, W refers to the embedded matrix, which is used as a look up table. Then we import m weight matrices for context words in asymmetric positions and there are sn words in the sentence. Note that if the skip window is 3, there will be 3 weight matrices. After assigning weight matrices, we get positional word vector $z_{n,w_{c+i}}$. It is calculated as:

$$\begin{cases} [v_{w_c} v'_{1,w_{c-\mathrm{i}}}, \dots, v_{w_c} v'_{1,w_{c-2}}, v_{w_c} v'_{1,w_{c-1}}, v_{w_c} v'_{1,w_{c+1}}, v_{w_c} v'_{1,w_{c+2}}, \dots, v_{w_c} v'_{1,w_{c+\mathrm{i}}}] g_{\mathrm{i}}(w_{c-i}|w_c) + b_{c-i} = z_{1,w_{c-i}} \\ [v_{w_c} v'_{2,w_{c-\mathrm{i}}}, \dots, v_{w_c} v'_{2,w_{c-2}}, v_{w_c} v'_{2,w_{c-1}}, v_{w_c} v'_{2,w_{c+1}}, v_{w_c} v'_{2,w_{c+2}}, \dots, v_{w_c} v'_{2,w_{c+\mathrm{i}}}] g_{\mathrm{i-1}}\left(w_{c-(\mathrm{i-1})}|w_c\right) + b_{c-(i-1)} = z_{2,w_{c-(i-1)}} \\ \cdots \\ [v_{w_c} v'_{\mathrm{n-1},w_{c-\mathrm{i}}}, \dots, v_{w_c} v'_{\mathrm{n-1},w_{c-2}}, v_{w_c} v'_{\mathrm{n-1},w_{c-1}}, v_{w_c} v'_{\mathrm{n-1},w_{c+1}}, v_{w_c} v'_{\mathrm{n-1},w_{c+2}}, \dots, v_{w_c} v'_{\mathrm{n-1},w_{c+\mathrm{i}}}] g_{\mathrm{i-1}}\left(w_{c+(i-1)}|w_c\right) + b_{c+(i-1)} = z_{\mathrm{n-1},w_{c+(\mathrm{i}+1)}} \\ [v_{w_c} v'_{\mathrm{n},w_{c-\mathrm{i}}}, \dots, v_{w_c} v'_{\mathrm{n},w_{c-2}}, v_{w_c} v'_{\mathrm{n},w_{c-1}}, v_{w_c} v'_{\mathrm{n},w_{c+1}}, v_{w_c} v'_{\mathrm{n},w_{c+2}}, \dots, v_{w_c} v'_{\mathrm{n},w_{c+\mathrm{i}}}] g_{\mathrm{i}}(w_{c+i}|w_c) + b_{c+i} = z_{\mathrm{n},w_{c+i}} \end{cases} \tag{6}$$

where $i = \{1, 2, 3, \dots m\}$, $j = \{1, 2, 3, \dots n\}$. w_{c+i} refers to the feature vector of context word, and i refers to the relative position of context word.

Here, we regard the judgment of direction as a binary classification and use the sigmoid function to calculate the probability of selecting the direction. It is calculated as:

$$g_i(w_{c+i}|w_c) = \begin{cases} \sigma\left({v'_{j,w_{c+i}}}^T v_{w_c}\right), & L_{w+i} = 1 \\ 1 - \sigma\left({v_{j,w_{c+i}}}^T v_{w_c}\right), & L_{w+i} = 0 \end{cases} \tag{7}$$

Finally, we define a novel softmax function to normalize the probability of context words and get loss L^w. In this softmax function, the probability of each context word is normalized with all probability in $z_{n,w_{c+i}}$. It is shown as:

$$p\left(w_{c+i}|w_c\right) = \frac{\exp\left(\sum_{i=1}^{m} \sum_{j=1}^{n} z_{n,w_{c+i}}\right)}{\sum_{i=-m}^{m} \exp\left(\sum_{i=1}^{m} \sum_{j=1}^{n} z_{n,w_{c+i}}\right)} \tag{8}$$

Hence, the loss function L^w can be defined as:

$$L^w = -\log \prod_{i=-m}^{m} p\left(w_{c+i}|w_c\right) \tag{9}$$

Optimization. The model optimization involves updating the embedded matrix and m weight matrices to optimize vector representation of words. We define u (w, i, m) as the following:

$$u(w,i,n) = {v'_{n,w_{c+i}}}^T v_{w_c} \left[\sigma({v'_{n,w_{c+i}}}^T v_{w_c})\right]^{L_{w+i}} \left[1 - \sigma({v'_{n,w_{c+i}}}^T v_{w_c})\right]^{1-L_{w+i}} \tag{10}$$

In the stochastic gradient descent, we update the embedded matrix and weight matrix as follows:

$$v_{w_c}(\text{new}) = v_{w_c}(\text{old}) + \nabla L^w(v_{w_c}) \tag{11}$$

$${v'_{n,w_{c+i}}}^T(new) = {v'_{n,w_{c+i}}}^T(\text{old}) + \eta \nabla L^w\left({v'_{n,w_{c+i}}}^T\right) \tag{12}$$

where $\nabla L^w(v_{w_c})$ and $\nabla L^w({v'_{n,w_{c+i}}}^T)$ represent the gradient for the embedded matrix and the weight matrix respectively. Then, the update rates of v_{w_c} and ${v'_{n,w_{c+i}}}^T$ in gradient optimization are:

$$\begin{aligned} \nabla L^w(v_{w_c}) &= \frac{\partial \mathrm{u(w,i,n)}}{\partial v_{w_c}} \\ &= \frac{\partial {v'_{n,w_{c+i}}}^T v_{w_c} [\sigma({v'_{n,w_{c+i}}}^T v_{w_c})]^{L_{w+i}} [1 - \sigma({v'_{n,w_{c+i}}}^T v_{w_c})]^{1-L_{w+i}}}{\partial v_{w_c}} \end{aligned} \tag{13}$$

$$\begin{aligned} \nabla L^w({v'_{n,w_{c+i}}}^T) &= \frac{\partial \mathrm{u(w,i,n)}}{\partial {v'_{n,w_{c+i}}}^T} \\ &= \frac{\partial {v'_{n,w_{c+i}}}^T v_{w_c} [\sigma({v'_{n,w_{c+i}}}^T v_{w_c})]^{L_{w+i}} [1 - \sigma({v'_{n,w_{c+i}}}^T v_{w_c})]^{1-L_{w+i}}}{\partial {v'_{n,w_{c+i}}}^T} \end{aligned} \tag{14}$$

Hence, in the proposed position-sensitive word vector training model, v_{w_t} and ${v'_{m,w_{c+i}}}^T$ are updated as follows:

$$v_{w_c}(new) = v_{w_c}(old) + \eta \nabla L^w\left(v_{w_c}\right) \tag{15}$$

$${v'_{n,w_{c+i}}}^T(new) = {v'_{n,w_{c+i}}}^T(\text{old}) + \eta \nabla L^w\left({v'_{n,w_{c+i}}}^T\right) \tag{16}$$

where η refers to the learning rate. The model will adjust the learning rate η based on the following equations after every 10,000 words are processed.

$$\eta = \eta_0\left(1 - \frac{word_n}{word_c + 1}\right) \tag{17}$$

$$word_c = \sum_{\mathrm{w} \in \mathrm{V}} counter(w) \tag{18}$$

$word_c$ represents the count of processed words. η will gradually decrease during the processing. However, the learning rate cannot be too small. Hence, we set a threshold $\eta_{\min}$. Once $\eta < \eta_{\min}$, we initialize it to $\eta_{\min}$, where $\eta_{\min} = 10^{-4} \times \eta_0$ and $\eta_0 = 0.025$.

Self-attention. We get the set of sentence feature $= [\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_j}, \dots . \boldsymbol{w_n}]$ through the above proposed position-sensitive word vector training model. Note that $\boldsymbol{w_1}$ refer to the feature vector of the first word, and n means the number of words in the sentence. As shown in Fig. 3.

Feature vectors we got from the proposed position-sensitive word vector training model will be fed into the self-attention mechanism, yielding a revised set of feature vectors $\widetilde{W} = [\widetilde{w}_1, \widetilde{w}_2, \dots, \widetilde{w}_j, \dots .. \widetilde{w}_n]$, where $\widetilde{w}_i$ is the updated feature vector of the j^{th} word output by the self-attention layer. This self-attention mechanism is adept at learning the structure and contextual relationships within a sentence and updates the training parameters by concentrating on internal information, without the need for external data. Compared with the original feature vector $W = \left[w_1, w_2, \dots, w_j, \dots .. w_n\right]$, the new feature vector $\widetilde{W} = [\widetilde{w}_1, \widetilde{w}_2, \dots, \widetilde{w}_j, \dots .. \widetilde{w}_n]$ got from self-attention reflects the interconnections between different words in the sentence. What's worth to say, the proposed position-sensitive word vector training model compensates for the failure that self-attention didn't consider the influence of word position in the generation of feature vectors.

In this regard, we get three vector sequences by linear transformation based on the given sentence feature $W = [w_1, w_2, w_3, \dots .. w_n]$, which include the query vector sequence Q, the key vector sequence K and the value vector

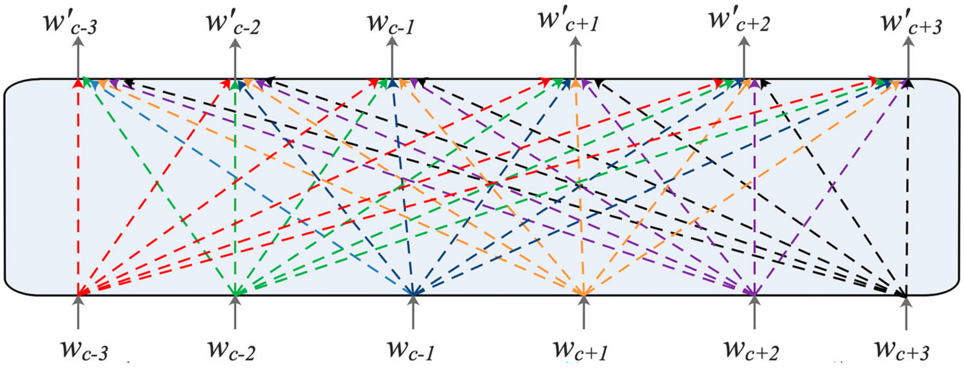

Fig. 3 The abstracted framework of self-attention. The framework presents the process of generating new word vectors that containing richer semantic relations.

sequence V, which is calculated as the followings:

$$Q = H_q * W = [H_q w_1,\ H_q w_2,\ H_q w_3,\ \ldots ..H_q w_n] \tag{19}$$

$$K = H_k * W = [H_k w_1,\ H_k w_2,\ H_k w_3,\ \ldots ..H_k w_n] \tag{20}$$

$$V = H_v * W = [H_v w_1,\ H_v w_2,\ H_v w_3,\ \ldots ..H_v w_n] \tag{21}$$

where H_q, H_k, and H_v represents the initial parameters of the three linear transformations, and Q, K and V are the transformations of the sentence feature $W = [w_1,\ w_2,\ w_3,\ \ldots ..w_n]$. Then, we calculate the attention weight distribution A as followings:

$$A = Q \cdot K^T \tag{22}$$

$$A = [H_q w_1,\ H_q w_2,\ H_q w_3,\ \ldots ..H_q w_n] \cdot [H_k w_1, H_k w_2, H_k w_3, \ldots ..H_k w_n]^T \tag{23}$$

$$A = H_q * H_k * [w_1,\ w_2,\ w_3,\ \ldots ..w_n] \cdot [w_1, w_2, w_3, \ldots ..w_n]^T \tag{24}$$

$$A = H_q * H_k * [\, w_1 * W^T,\ w_2 * W^T, w_3 * W^T,\ \ldots\ \ldots\ w_n * W^T] \tag{25}$$

As depicted in the above, the length n of the attention weight distribution A represents the importance of each word's feature vector. We normalize it using the softmax as followings:

$$\hat{A} = softmax\left(\frac{A}{\sqrt{d_k}}\right) \tag{26}$$

where d_k represents the length of the feature dimension. After that, we multiply the normalized attention weight distribution $\hat{A}$ by the feature vector, and get the new feature vector of the sentence $\widetilde{W}$ as follows:

$$\widetilde{W} = Attention(Q,\ K,\ V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{27}$$

$$\widetilde{W} = \hat{A} \cdot V \tag{28}$$

$$\widetilde{W} = softmax\left(\frac{A}{\sqrt{d_k}}\right) \cdot V \tag{29}$$

$$\widetilde{W} = softmax\left(\frac{A}{\sqrt{d_k}}\right) \cdot H_v * W \tag{30}$$

$$\widetilde{W} = softmax\left(\frac{H_q * H_k * [\, w_1 * W^T,\ w_2 * W^T, w_3 * W^T,\ \ldots\ \ldots\ w_n * W^T]}{\sqrt{d_k}}\right) \cdot H_v * W \tag{31}$$

In this paper, the new feature vector considers both the position of context word and the semantic relations of the whole sentence, which significantly improves the accuracy of the output word vectors.

Document vector. The generation of the document vector is bifurcated into two parts: the document weight and the topic matrix. Document weight encapsulates the significance of each topic in a document. Initialize the document weight vector of each document in the corpus. Subsequently, a set of topic vectors is generated, subject to the constraints imposed by the document vectors $\boldsymbol{d_j}$. Under these constraints, the latent topic vectors are represented as:

$$d_j = p_{j0} \times t_0 + p_{j1} \times t_1 + \ldots\ \ldots\ + p_{jk} \times t_k + \ldots\ \ldots\ p_{jn} \times t_n \tag{32}$$

$t_0, t_1, \ldots, t_k, \ldots, t_n$ are latent topic vectors. p_{jk} is the topic weight transformed by softmax, indicating the proportion of different topics in a document. The total sum of p_{jk} for any given document is 100%. Topic vector t_n is a shared part across all documents, with the document weight being instrumental in modulating its thematic prominence. To enhance the interpretability of topics, we set word vector's dimension as the number of topics and initialize the latent topic vector to get p_{jk}. Then, optimize the Dirichlet(α) likelihood to sparse document weight p_{ij} and the loss function is:

$$L_d = -\lambda \sum_{k=0}^{n} (\alpha - 1) log(p_{jk}) \tag{33}$$

In this model, the parameter α plays a crucial role in determining the sparsity of the document weight vector. Specifically, if $\alpha > 1$, the document weight vector tends to be more concentrated; conversely, if $\alpha \le 1$, the document weight vector exhibits sparsity. Aligning with existing research (Zuo et al., 2023) set $\alpha = \frac{1}{n}$. After continuous iteration and optimization, the document weight vector becomes increasingly concentrated, culminating in the formation of an interpretable topic vector.

The total loss L of PS-WETM is the sum of L_w generated by the proposed position-sensitive word vector training model and L_d generated during optimizing the probability distribution of document weight.

$$L = L_w + L_d \tag{34}$$

Topic vector and document weight vector are synthesized to produce the document vector. This document vector is then added to the word vector, resulting in the formation of the context vector. Finally, minimize the total loss L to train the context vector. Input the well-trained context vector into LDA topic model, the hot topic identification is completed.

Experiment and discussion

In this section, we perform comparative analyses between the newly proposed PS-WETM topic model and six recent proposed topic models, employing word cloud and sentiment analysis as assisted tools to gauge the accuracy of PS-WETM. This methodology aids in evaluating the model's efficacy in capturing the nuanced essence of topics within online social media text.

Dataset. In this paper, we choose COVID-19 as the study case to verify the performance of short-text topic model PS-WETM. We chose to focus on microblogs on *Sina Weibo* from the initial outbreak period of January 2020 through June 2020, when the epidemic first showed signs of remission, for the research data. We crawled 6-month microblogs posted on *Sina Weibo* based on keywords *epidemic, COVID-19, pneumonia, segregate*, and *mask*, and obtained a total of 82,330 microblogs as the analysis data.

Performance analysis. We undertook a comparative analysis of seven topic models, evaluating their performance based on metrics of perplexity and topic coherence. The model assessed include DMM (Yin and Wang, 2014), PTM (Zuo et al., 2016), SPTM (Zuo et al., 2023), GPU-DMM (Li et al., 2017), GLTM (Liang et al., 2018), WETM (Rashid et al., 2023) and the proposed PS-WETM. When given a test set $W = \{w_1, w_2, w_3, \ldots, w_N\}$, the perplexity ($PP$) is calculated as follows:

$$PP(W) = P(w_1 w_2 w_3 \ldots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 w_3 \ldots w_N)}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_1 w_2 w_3 \ldots w_N)}} \quad (35)$$

The perplexity metric inversely relates to the conditional probability of word sequences; a higher conditional probability leads to a lower perplexity. We utilized the crawled microblogs as data to evaluate the performance of these models. The number of topics K is 5,10,15 and 20 respectively. The maximum iteration is 500. The results of this comparison are illustrated in Table 2.

Topic coherence is another crucial metric for assessing the performance of a topic model. It is based on the co-occurrences of words within Wikipedia and has been demonstrated to correlate with human judgment. The calculation of topic coherence (C) is as follows:

$$C(z; S^z) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} \log \frac{D_2(w_n^z, w_l^z) + 1}{D_1(w_l^z)} \quad (36)$$

For a specified topic z, the set of words in top n topics is $S^z = \{w_1^z, w_2^z, \ldots, w_n^z\}$. $D_1(w)$ signifies the frequency of word w in the document. $D_2(w_1, w_2)$ is the co-occurrence of word w_1 and w_2. A higher value of topic coherence indicates better interpretability and clarity of the topics. The comparison of topic coherence among the models is depicted in Table 3.

As shown in Tables 2–3, PTM has the highest perplexity and the lowest topic coherence, presenting a relatively bad performance on topic extraction of short texts, followed by SPTM. And the topic model PS-WETM has the lowest perplexity and the highest topic coherence relatively, which demonstrates that compared with the other six topic models, PS-WETM has the best interpretability of topics.

Table 2 Perplexity score of seven topic models.

Method	$K = 5$	$K = 10$	$K = 15$	$K = 20$
DMM	1.78	1.77	1.74	1.78
PTM	1.84	1.82	1.80	1.85
SPTM	1.82	1.81	1.79	1.83
GPU-DMM	1.76	1.74	1.73	1.74
GLTM	1.74	1.72	1.73	1.74
WETM	1.73	1.68	1.66	1.67
PS-WETM	1.68	1.65	1.64	1.66

Table 3 Topic coherence score of seven topic models.

Method	$K = 5$	$K = 10$	$K = 15$	$K = 20$
DMM	0.60	0.63	0.64	0.63
PTM	0.58	0.62	0.61	0.61
SPTM	0.59	0.61	0.62	0.61
GPU-DMM	0.60	0.64	0.62	0.62
GLTM	0.61	0.65	0.65	0.64
WETM	0.63	0.65	0.65	0.66
PS-WETM	0.64	0.67	0.66	0.66

Word cloud. The high-frequency Chinese words were translated into English and drawn in the Word Cloud. Word Cloud from January to June are shown in Fig. 4.

The Word Cloud visualization reveals that during the initial period of the event, from January to February, there were three prominent topics reflected in the high-frequency words. Some words pertained to the topic of *infection*, such as *case, health, cumulative, medical*, while others are about the topic *prevention*, such as *medical, observation, isolation*, and still others are about the topic *domestic places*, such as *Wuhan, Shenzhen*, highlighting the areas most affected by the initial outbreak of COVID-19 in China.

In March and April, words *rent, family, income, close* arise in the Word Cloud, describing the situation that COVID-19 had a great impact on people's daily lives. The shutdown of businesses led to reduced incomes, making it challenging for many to cover their rent and daily expenses. Additionally, words *global*, *UK*, *overseas* began to appear in the Word Cloud of April, albeit with less prominence compared to the aforementioned words.

As the timeline progressed into May and June, words *work, death, test, close, market* continued to feature in the Word Cloud, but with a declining frequency, indicating that efforts to control COVID-19 and treat patients were yielding positive results, as evidenced by the decreasing number of deaths and infections. In addition, businesses and workplaces in numerous cities started to resume production and work.

The Word Cloud in Fig. 5, encompassing data from 6 months of microblogs, demonstrates that topics *epidemic* and *protection* were persistent and dominant throughout this timeframe. Notably, the prominence of these topics did experience a slight decline during March and April.

Sentiment analysis. In this chapter, we use the Emotion Dictionary of *Dalian University of Technology*, which contains seven emotions, to calculate the emotion value of microblogs and classify microblogs into seven emotions. The results are presented in Fig. 6.

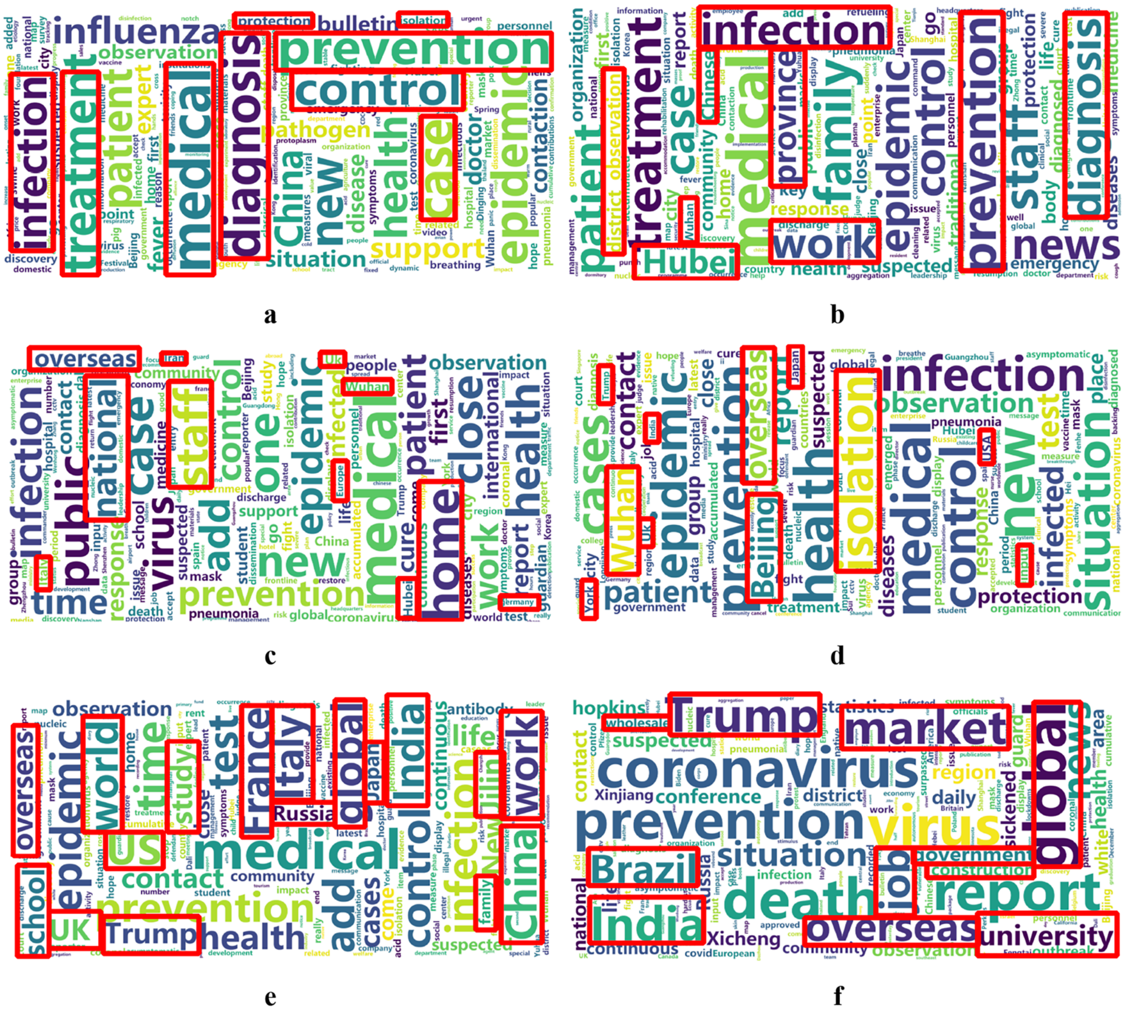

Fig. 4 Word Cloud of microblogs from January to June. **a**, **b**, **c**, **d**, **e**, **f** represent the Word Cloud from January to June respectively.

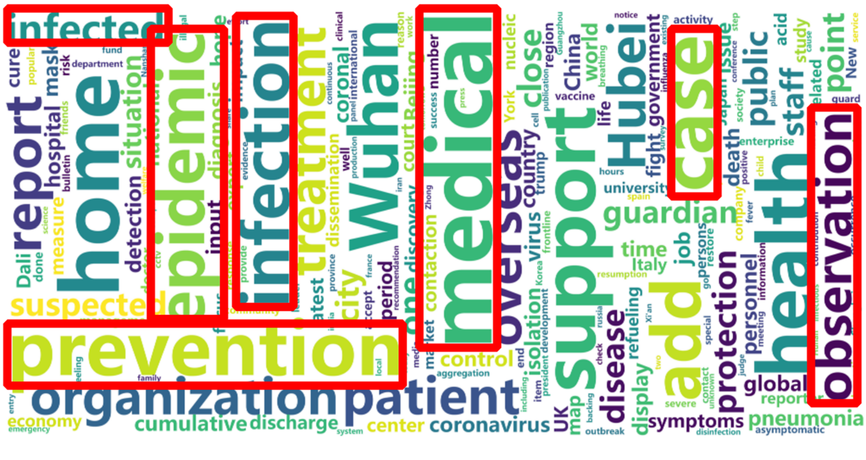

Fig. 5 Word Cloud of 6 months microblogs. It shows the Word Cloud and keywords of all the 6-month microblogs.

The outbreak of COVID-19 led to a marked increase in the emotions of *Love, Depress, Dislike* and *Fear* in February, with *Fear* spiked the fastest. As the gravity of the COVID-19 situation dawned on people, there was a noticeable decrease in microblogs sharing moments of happy daily life and in expressions of *Joy*. However, there were also microblogs expressing encouragement for patients and frontline medical workers, which is why the emotion of *Love* remained predominant.

As the COVID-19 situation evolved, the prevalence of *Anger* diminished, reflecting a reduction in blame or resentment towards the virus's origin. A significant moment of change occurred on May 21 and 22, with the convening of the *NPC&CPPCC* in Beijing, which instilled hope among the Chinese populace. Following this, widespread nucleic acid testing was implemented across China, and business operations began to resume, leading to improvements in the country's situation.

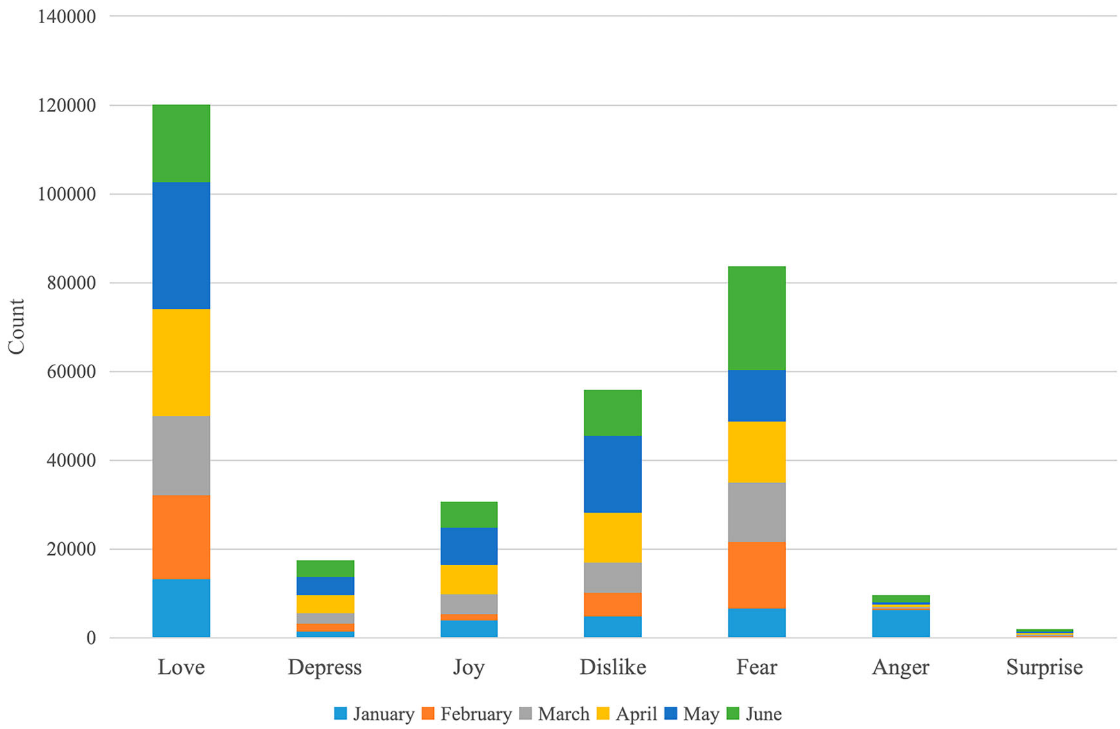

Fig. 6 Count of seven emotions from January to June. The statistical graph of seven emotions in each month from January to June, including emotion *Love*, *Depress*, *Dislike*, *Fear*, *Anger*, and *Surprise*.

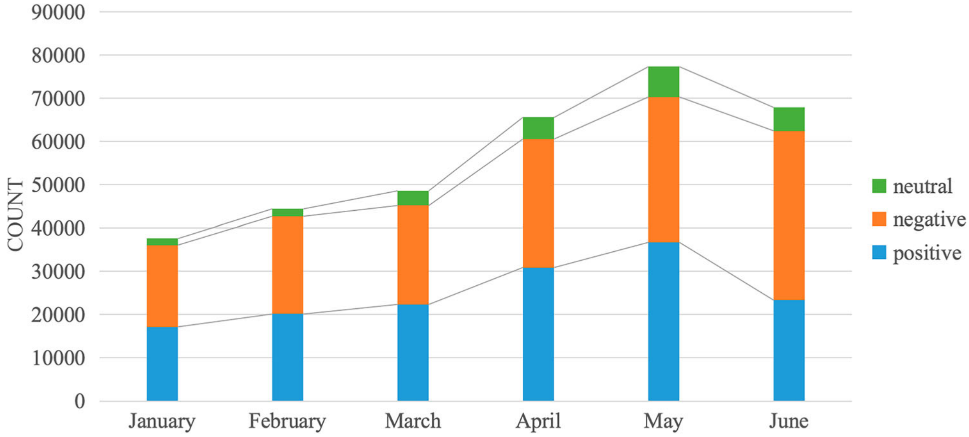

Fig. 7 Count of positive, negative, neutral emotions from January to June. The statistic graph of three emotions, where emotion *Love*, *Depress*, *Dislike*, *Fear*, *Anger*, and *Surprise* are classified into *positive*, *negative*, and *neutral* emotion.

People's lives started to show signs of recovery, and as a result, the emotions of *Love* and *Joy* experienced a resurgence.

The seven emotions identified in our study have been categorized into positive, negative, and neutral groups. *Love* and *Joy* are classified as positive emotions, *Surprise* is considered neutral emotion, while *Depress*, *Dislike*, *Fear*, and *Anger* are categorized as negative emotions. According to Figs. 7, 8, negative emotions slightly outnumber positive emotions. However, with the global spread of COVID-19, the prevalence of negative emotions significantly outweighs that of positive emotions due to events discussed above.

Topic identification by the proposed topic model PS-WETM. In this part, we utilized the PS-WETM short-text topic model to identify topics from the extended texts. After we get the Chinese topics and corresponding keywords, we translate them into English and then visualize them. Topic visualization of January is shown as Fig. 9. (Topic visualization of other 5 months are shown in Supplementary Figs. S1–S5).

The number of topics is set to 6 according to the perplexity. And the number of keywords is set to 20. In Fig. 9, the absence of overlap between any of the circles indicates that the number of topics set here is reasonable. It is also observable that some topics and keywords recur across adjacent months. For better analyzing the identified topics, we combined microblogs of January and February as the first period, microblogs of March and April as the second period, microblogs of May and June as the third period. Then, new topics and keywords of three periods are presented in Figs. 10–12.

When identifying topics and keywords of the three periods, the number of topics is set to 8 according to the perplexity. Topics *epidemic, support, protection and control, affection* identified in the first period are consistent with key words *infection, case,*

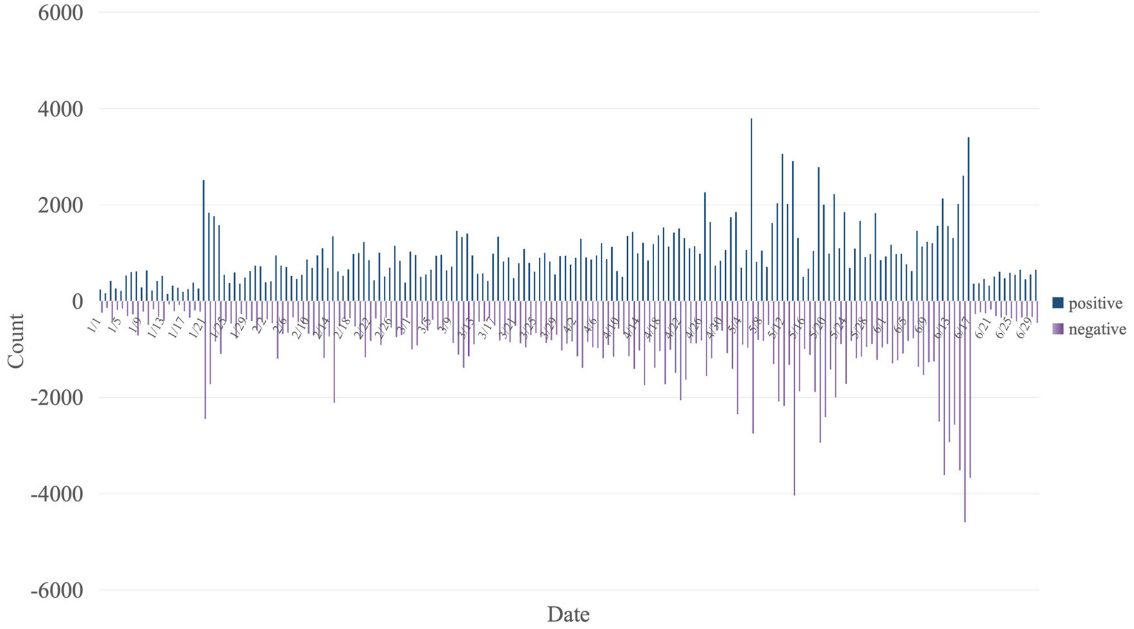

Fig. 8 Daily change of positive and negative emotions in 6 months. The statistical graph of emotions each day, where emotion *Love*, *Depress*, *Dislike*, *Fear*, *Anger*, and *Surprise* are classified into *positive* and *negative* emotion.

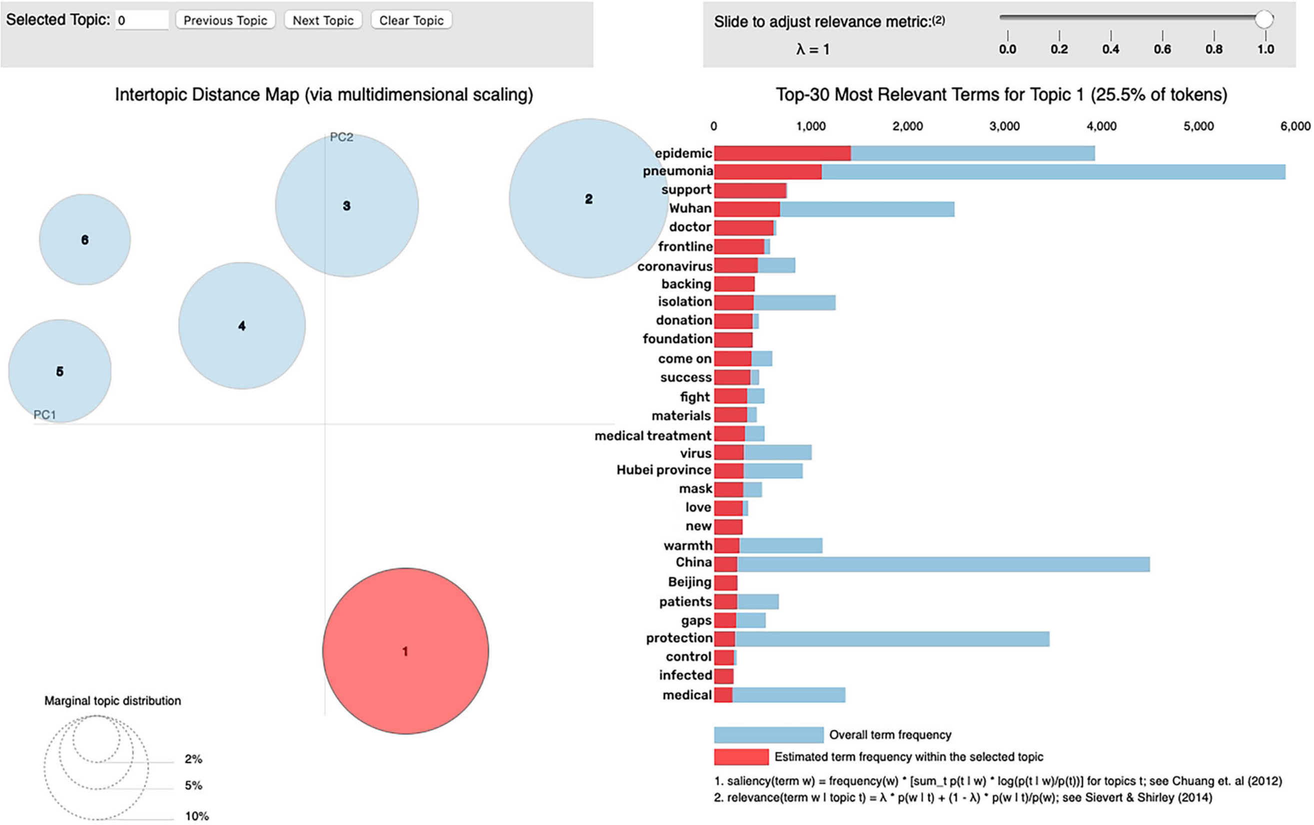
Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (25.5% of tokens)
0
1,000
2,000
3,000
4,000
5,000
6,000
epidemic
pneumonia
support
Wuhan
doctor
frontline
coronavirus
backing
isolation
donation
foundation
come on
success
fight
materials
medical treatment
virus
Hubei province
mask
love
new
warmth
China
Beijing
patients
gaps
protection
control
infected
medical
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 9 Visualization of topics and key words of January. The distribution and key words of topics, where circles represent topics, and words on the right are key words of the corresponding topics.

health, cumulative, medical, prevention, medical, observation, isolation drawn in Word Cloud of January. Topics *international response, domestic places* are dovetailed with key words *Wuhan, hospital, Shenzhen, Guangdong, Beijing* in Word Cloud of February. As described in Figs. 13, 14, in the first period, the number of confirmed cases and deaths in China suddenly climbed, and the places with confirmed cases gradually expanded, resulting in the emergence of domestic regions' names.

During the second period, while China's situation began to stabilize, the global scenario deteriorated significantly. There was

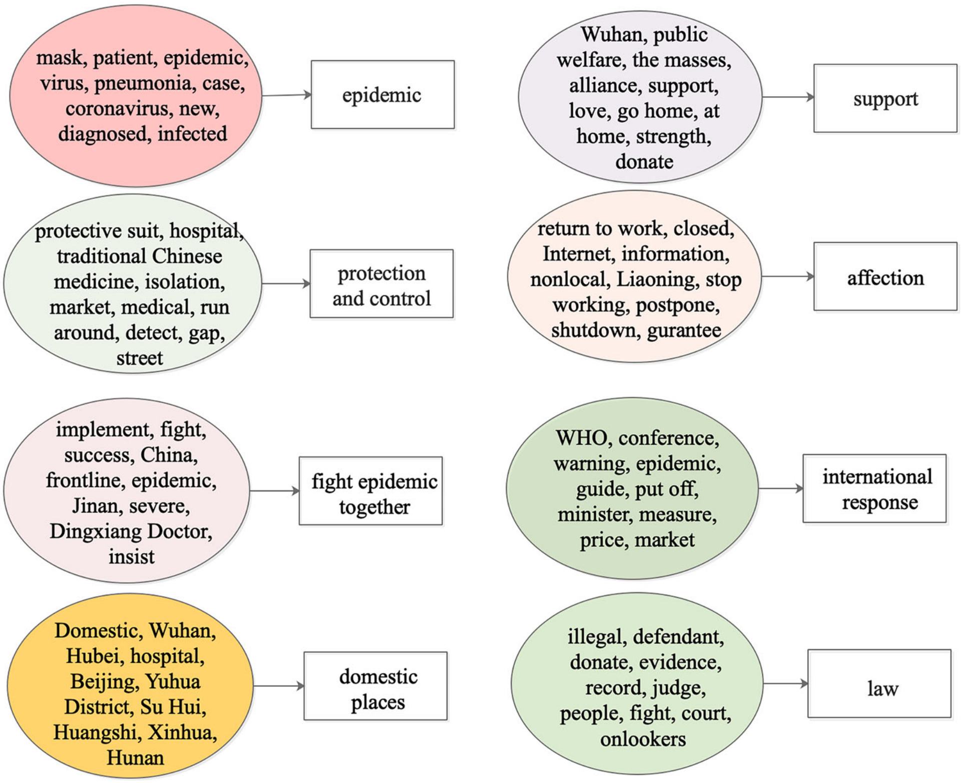
mask, patient, epidemic, virus, pneumonia, case, coronavirus, new, diagnosed, infected
epidemic
Wuhan, public welfare, the masses, alliance, support, love, go home, at home, strength, donate
support
protective suit, hospital, traditional Chinese medicine, isolation, market, medical, run around, detect, gap, street
protection and control
return to work, closed, Internet, information, nonlocal, Liaoning, stop working, postpone, shutdown, gurantee
affection
implement, fight, success, China, frontline, epidemic, Jinan, severe, Dingxiang Doctor, insist
fight epidemic together
WHO, conference, warning, epidemic, guide, put off, minister, measure, price, market
international response
Domestic, Wuhan, Hubei, hospital, Beijing, Yuhua District, Su Hui, Huangshi, Xinhua, Hunan
domestic places
illegal, defendant, donate, evidence, record, judge, people, fight, court, onlookers
law

Fig. 10 Topics and key words of the first period in January and February. The visualization of extracted topics and corresponding key words in the first period.

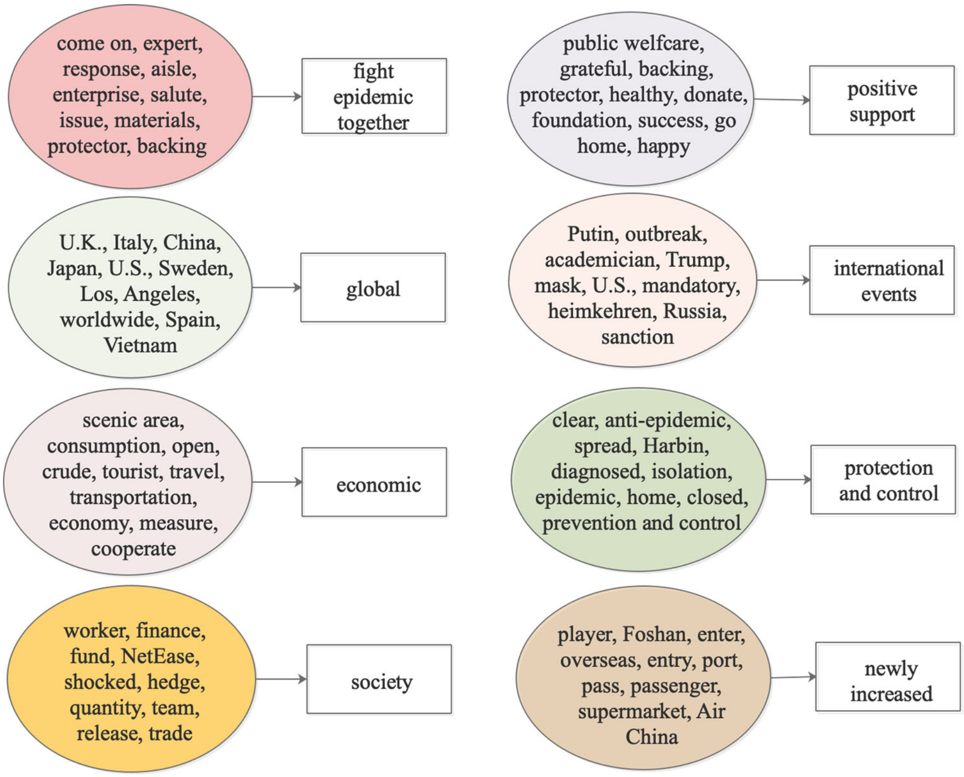
come on, expert, response, aisle, enterprise, salute, issue, materials, protector, backing
fight epidemic together
public welfcare, grateful, backing, protector, healthy, donate, foundation, success, go home, happy
positive support
U.K., Italy, China, Japan, U.S., Sweden, Los, Angeles, worldwide, Spain, Vietnam
global
Putin, outbreak, academician, Trump, mask, U.S., mandatory, heimkehren, Russia, sanction
international events
scenic area, consumption, open, crude, tourist, travel, transportation, economy, measure, cooperate
economic
clear, anti-epidemic, spread, Harbin, diagnosed, isolation, epidemic, home, closed, prevention and control
protection and control
worker, finance, fund, NetEase, shocked, hedge, quantity, team, release, trade
society
player, Foshan, enter, overseas, entry, port, pass, passenger, supermarket, Air China
newly increased

Fig. 11 Topics and key words of the second period in March and April. The visualization of extracted topics and corresponding key words in the second period.

a rapid increment in the number of confirmed cases and deaths worldwide. And as depicted in Fig. 14, the number of countries with confirmed cases also expands gradually. Hence, topics *global* and *international events* appear in the second periods. Furthermore, due to the global outbreak, the number of emotions *Fear* and *Dislike* both rise in the second period, which is also consistent with topics identified in these two periods.

In Fig. 15a, b, the map shows that China is in a severe situation where the number of confirmed cases accounts for the largest. On January 21, the first period, China has 386 confirmed cases, while the United States and Mexico each have only 1 confirmed case. On March 26, the second period, countries such as Italy, Brazil, Canada appear with a large number of confirmed cases, but the situation in China has flattened out, just like the description of topic *global* shown in Fig. 11. In Fig. 15c and Fig. 15d, it can be seen that in the third period, China's COVID-19 has entered a period of remission, while other countries have entered an outbreak period. These all reflect the consistency between identified topics and emotion value, and the high accuracy of the proposed short-text topic model PS-WETM.

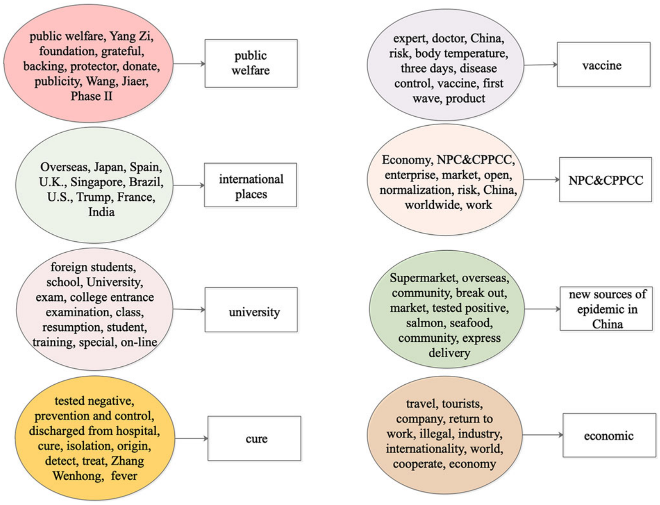
public welfare, Yang Zi, foundation, grateful, backing, protector, donate, publicity, Wang, Jiaer, Phase II
public welfare
expert, doctor, China, risk, body temperature, three days, disease control, vaccine, first wave, product
vaccine
Overseas, Japan, Spain, U.K., Singapore, Brazil, U.S., Trump, France, India
international places
Economy, NPC&CPPCC, enterprise, market, open, normalization, risk, China, worldwide, work
NPC&CPPCC
foreign students, school, University, exam, college entrance examination, class, resumption, student, training, special, on-line
university
Supermarket, overseas, community, break out, market, tested positive, salmon, seafood, community, express delivery
new sources of epidemic in China
tested negative, prevention and control, discharged from hospital, cure, isolation, origin, detect, treat, Zhang Wenhong, fever
cure
travel, tourists, company, return to work, illegal, industry, internationality, world, cooperate, economy
economic

Fig. 12 Topics and key words of the third period in May and June. The visualization of extracted topics and corresponding key words in the third period.

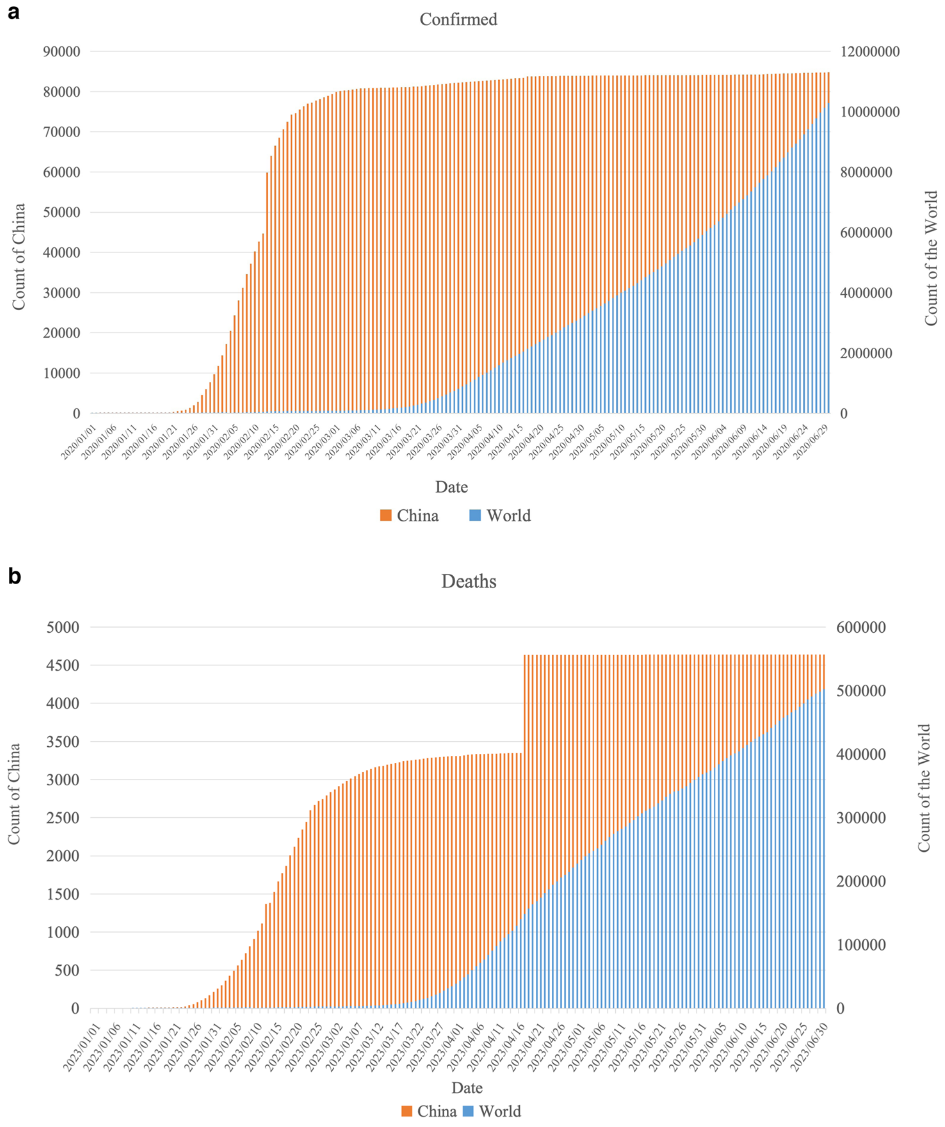

Fig. 13 The total confirmed and deaths in China and the world. In these two subfigures, the orange bar graph represents the count of China, and the blue bar graph represents the county of the World. **a** is the statistical graph of total confirmed cases in China and the World; **b** is the statistical graph of total death in China and the World.

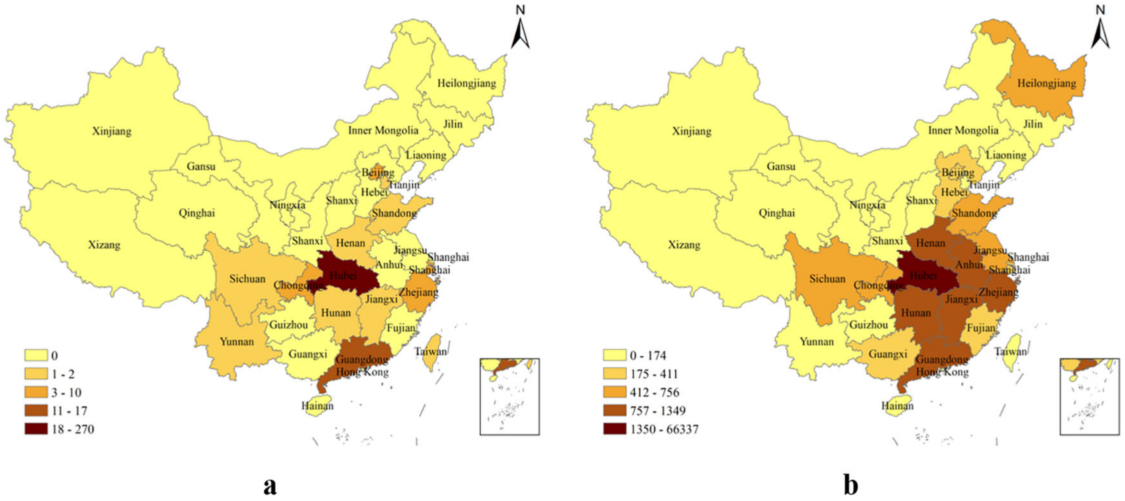

Fig. 14 The map of places containing confirmed cases in China. Subfigure **a** and **b** are respectively situations of China on January 21 and March 15.

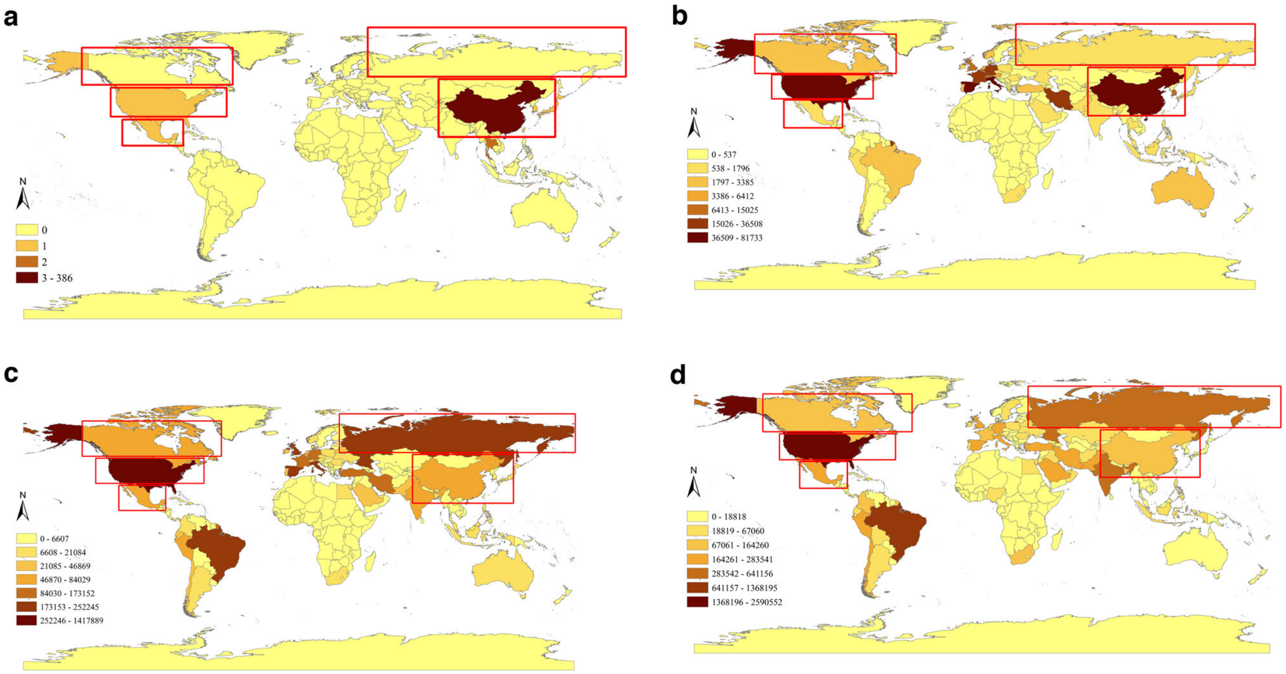

Fig. 15 The map of counties containing confirmed cases in the World. a, **b**, **c**, **d** are respectively situations of the world on January 21, March 26, May 15, June 30.

Conclusions

In this study, we introduce PS-WETM, a joint-training short-text topic model grounded in deep learning, designed to precisely identify latent topics in online social media text of emergencies. Firstly, it extends short texts of online opinions according to the similarities of word vectors, solving the problem of sparse matrix, and improving the accuracy of PS-WETM in generating context vector. Secondly, it combines word vector and document vector to better deal with the semantic relations of words and topic coherence. Thirdly, we define a proposed position-sensitive word vector training model, which assigns different weight matrices to asymmetrically positioned context words and judges the direction of each context word, thereby thoroughly accounting for the precise positions of context words in weight prediction and making up for the deficiency that previous topic models treat context words in different positions equally. Fourthly, in order to enhance the performance of topic matching, we implement a self-attention mechanism that calculates the contribution of each word feature generated by the proposed position-sensitive word vector training model to topic matching accuracy. The mechanism not only adjusts the feature representations of individual words but also considers both their semantic relations and those within the entire sentence. In the short-text topic model PS-WETM, the proposed position-sensitive word vector training model differentiates the relation between context words in different positions, which compensates for the deficiencies that self-attention ignores the position of words. Concurrently, the self-attention mechanism learns the sentence structure solely from internal without the need for external information, which addresses the limitation that the proposed position-sensitive word vector training model relatively large parameters. And the comparison displays that PS-WETM outperforms than other six topic models, evidence by its lowest perplexity and highest topic coherence. Furthermore, the trend of identified topics is considerably consistent with that delineated through the Word Cloud and sentiment analysis.

Data availability

The data is available at: https://github.com/qinsm/data-for-HMM/tree/main.

Received: 7 March 2024; Accepted: 11 February 2025;
Published online: 01 March 2025

References

Balikas G, Amoualian H, Clausel M et al. (2016) Modeling topic dependencies in semantically coherent text spans with copulas. In: Matsumoto YJ, Prasad R (eds) Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The 26th International Conference on Computational Linguistics; 2016 Dec 11–16; Osaka, Japan. Osaka(Japan): The COLING 2016 Organizing Committee, pp 1767–1776

Bastani K, Namavari H, Shaffer J (2019) Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst Appl 127:256–271

Belford M, Mac Namee B, Greene D (2018) Stability of topic modeling via matrix factorization. Expert Syst Appl 91:159–169

Bender EM, Timnit G, Angelina MM et al. (2021) On the dangers of stochastic parrots: can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623. https://doi.org/10.1145/3442188.3445922

Blei DM, Carin L, Dunson D (2010) Probabilistic topic models. IEEE Signal Process Mag 27(6):55–65

Bravo-Marquez F, Frank E, Pfahringer B (2016) Annotate-sample-average (ASA): a new distant supervision approach for Twitter sentiment analysis. ECAI 2016: Proceedings of the 22nd European conference on artificial intelligence, vol 285, pp 498–506. https://doi.org/10.3233/978-1-61499-672-9-498

Brown TB, Mann B, Ryder N et al. (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Bruni E, Tran NK, Baroni M (2014) Multimodal distributional semantics. J Artif Intell Res 49(1):1–47

Chaturvedi I, Satapathy R, Cavallari S et al. (2019) Fuzzy commonsense reasoning for multimodal sentiment analysis. Pattern Recognit Lett 125:264–270

Dieng AB, Ruiz FJR, Blei DM (2020) Topic modeling in embedding spaces. Trans Assoc Comput Linguist 8(2):439–453

Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794. http://arxiv.org/pdf/2203.05794v1

Gupta A, Zhang Z (2023) Neural Topic Modeling via Discrete Variational Inference. ACM Trans Intell Syst Technol 14(2):6904–6912

He D, Ren Y, Khattak AM et al. (2021) Automatic topic labeling using graph-based pre-trained neural embedding. Neurocomputing 463:596–608

Hofmann T (1999). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI'99). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 289–296

Huang JJ, Peng M, Li PW et al. (2020) Improving biterm topic model with word embeddings. World Wide Web 23(6):3099–3124

Jia H, Li Q (2018) Fuzzy bag-of-topics model for short text representation. International conference on neural information processing, Springer, Cham. https://doi.org/10.1007/978-3-030-04221-9_42

Jiang HX, Zhou R, Zhang LM et al. (2019) Sentence level topic models for associated topics extraction. World Wide Web 22(6):2545–2560

Keya KN, Papanikolaou Y, Foulds JR (2023) Neural embedding allocation: distributed representations of topic models. Comput Linguist 48(4):1021–1052

Kumar S, Mallik A, Khetarpal A (2022) Influence maximization in social networks using graph embedding and graph neural network. Inf Sci 607:1617–1636

Kumar KS, Mani ASR, Kumar TA et al. (2024) Sentiment analysis of short texts using SVMs and VSMs-based multiclass semantic classification. Appl Artif Intell 38(1):2321555

Le QV, Mikolov T (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning, 32(2), pp 1188-1196. https://proceedings.mlr.press/v32/le14.html

Li C, Duan Y, Wang H, Zhang Z, Ma Z (2017) Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Trans Inf Syst 36(2):1–30

Li L, Sun Y, Wang C (2018) Semantic augmented topic model over short text. In Proceeding of the 5th IEEE international conference on cloud computing and intelligence systems (CCIS). https://doi.org/10.1109/CCIS.2018.8691313

Li S, Zhang Y, Pan R et al. (2020) Adaptive probabilistic word embedding. WWW'20: Proceedings of the web conference 2020, pp 651–661. https://doi.org/10.1145/3366423.3380147

Liang W, Feng R, Liu X et al. (2018) GLTM: a global and local word embedding-based topic model for short texts. IEEE Access 6:43612–43621

Ma TH, Li J, Liang XN et al. (2019) A time-series based aggregation scheme for topic detection in Weibo short texts. Phys A Stat Mech Appl 536:120972

Machicao J, Correa Jr EA, Miranda GH et al. (2018) Authorship attribution based on life-like network automata. PloS one 13(3):e0193703

Meng Y, Zhang Y, Huang J et al. (2020) Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'20). Association for Computing Machinery, New York, NY, USA, pp 1908–1917

Mikolov T, Sutskever I, Kai C et al. (2013) Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Curran Associates Inc., Red Hook, NY, USA, 2, pp 3111–3119

Moody Christopher E (2016) Mixing dirichlet topic models and word embeddings to make lda2vec. https://doi.org/10.48550/arXiv.1605.02019

O'Callaghan D, Greene D, Carthy J et al. (2015) An analysis of the coherence of descriptors in topic modeling. Expert Syst Appl 42(13):5645–5657

Peng M, Xie QQ, Tian G (2018) Neural sparse topical coding. Proceedings of the 56th annual meeting of the association for computational linguistics, 1(long papers), pp 2332–2340. https://doi.org/10.18653/v1/P18-1217

Qin SM, Zhang ML, Hu HJ (2023) Ternary interaction evolutionary game of rumor and anti-rumor propagation under government reward and punishment mechanism. Nonlinear Dyn 111:21409–21439

Qin SM, Zhang ML (2024) Boosting generalization of fine-tuning bert for fake news detection. Inf Process Manag 61(4):103745

Rashid J, Kimb J, Hussainc A et al. (2023) WETM: a word embedding-based topic model with modified collapsed Gibbs sampling for short text. Pattern Recognit Lett 72:158–164

Roberts ME, Stewart BM, Airoldi EM (2016) A model of text for experimentation in the social sciences. J Am Stat Assoc 111(515):988–1003

Ruas T, Grosky WI, Aizawa A (2019) Multi-sense embeddings through a word sense disambiguation process. Expert Syst Appl 136:288–303

Salehi B, Cook P, Baldwin T (2015) A word embedding approach to predicting the compositionality of multiword expressions", Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado 977-983. https://doi.org/10.3115/v1/N15-1099

Shahriare Satu MD, Imran Khan MD, Mahmud M et al. (2021) TClustVID: a novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets. Knowl -Based Syst 226:107126

Steuber F, Schneider S, Schoenfeld MZ (2021) Embedding semantic anchors to guide topic models on short text corpora. Big Data Res 27:100293. https://doi.org/10.1016/j.bdr.2021.100293

Wang L, Niu J, Yu S (2019) Sentidiff: combining textual information and sentiment diffusion patterns for Twitter sentiment analysis. IEEE Trans Knowl Data Eng 32(10):2026–2039

Xu GX, Wu X, Yao HS et al. (2019) Research on Topic recognition of network sensitive information based on SW-LDA model. IEEE Access 7:21527–21538

Xu JM, Wang P, Tian GH et al. (2015) Short text clustering via convolutional neural networks. Proceedings of the 1st workshop on vector space modeling for natural language processing. https://doi.org/10.3115/v1/W15-1509

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 (WWW '13), pp 1445–1456. https://doi.org/10.1145/2488388.2488514

Yin J, Wang J (2014) A Dirichlet multinomial mixture model-based approach for short text clustering. Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 233–242. https://doi.org/10.1145/2623330.2623715

Zhang ML, Qin SM, Zhu XX (2021) Information diffusion under public crisis in BA scale-free network based on SEIR model—taking COVID-19 as an example. Phys A Stat Mech Appl 571:125848

Zuo Y, Li C, Lin H et al. (2023) Topic modeling of short texts: a pseudo-document view with word embedding enhancement. IEEE Trans Knowl Data Eng 35(1):972–985

Zuo Y, Wu JJ, Zhang H et al. (2016) Topic modeling of short texts: a pseudo-document view. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 2105–2114. https://doi.org/10.1145/2939672.2939880

Acknowledgements

This work was supported by the General Program of National Natural Science Foundation of China (grant number 71971051).

Author contributions

Qin Simeng was responsible for the methodology, data curation, data analysis, original draft writing and review & editing draft. Zhang Mingli was responsible for the resources, conceptualization. Hu Haiju was responsible for the resources, theory development, and supervision. Li Gang was responsible for investigation, supervision, data curation, and funding acquisition.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04551-2.

Correspondence and requests for materials should be addressed to Gang Li.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.