IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 29, 2020, accepted January 14, 2021, date of publication January 20, 2021, date of current version January 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052566

An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model

XU TAN[1,4], MUNI ZHUANG[2,4], XIN LU[3], AND TAITIAN MAO[4]
[1]School of Software Engineering, Shenzhen Institute of Information Technology, Shenzhen 518172, China
[2]Business School, Southern University of Science and Technology, Shenzhen 518172, China
[3]College of Systems Engineering, National University of Defense Technology, Changsha 410022, China
[4]School of Public Management, Xiangtan University, Xiangtan 411005, China

Corresponding author: Muni Zhuang (201821100562@smail.xtu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 71771213, Grant 91846301, and Grant 71790615; in part by the Ministry of Education in China (MOE) Liberal Arts and Social Sciences Foundation under Grant 17YJCZH157; and in part by the Pengcheng Scholar Funded Scheme, and the Basic Research Project of Science and Technology Plan of Shenzhen under Grant JCYJ20200109141218676.

ABSTRACT The purpose of this article is to analyse the emotional evolution of the netizens in reaction to the events of the Anti-ELAB (Anti-Extradition Law Amendment Bill) movement in Hong Kong. We attempt to investigate evolving laws of large-scale Internet public opinion events and provide relevant agencies with a theoretical basis for a public opinion response mechanism. On the basis of improving the Bidirectional Encoder Representations from Transformers (BERT) pre-training task, we add in-depth pre-training tasks, and based on the optimisation results of the LDA topic embedding, we integrate deeply with the LDA model to dynamically present the fine-grained public sentiment of the event. Through the collection of large-scale text data related to the Anti-ELAB Movement from a well-known forum in Hong Kong, a BERT-LDA hybrid model for large-scale network public opinion analysis is constructed in a complex context. Through empirical analysis, we calculate and reveal the emotional change process of netizens and opinion leaders in the three transition stages of the Anti-ELAB Movement with the evolution of the topic word as the core by visualisation. We also analyse the emotional distribution and evolution trend of public opinion under the 'text topic', and deeply analyse the character and role of opinion leaders in Anti-ELAB public opinion events. The improved BERT-LDA model or sentiment classification AUC value exceeds 99.6% in the sentiment classification task for the Anti-ELAB Movement.

INDEX TERMS BERT-LDA hybrid model, large-scale Internet public opinion, emotional evolution, the anti-ELAB movement.

I. INTRODUCTION

In June 2020, on the first anniversary of the birth of the Anti-ELAB Movement, the National Security Law was officially promulgated and implemented in Hong Kong. Since the inception of the Hong Kong Anti-ELAB Movement in June 2019, and under the guise of peaceful gatherings, the opposition and some radical forces have deliberately attacked the police, smashed the Legislative Council building, defaced the national emblem, and torn down the national flag. The haze of illegal incidents continues to spread. Although the Hong Kong SAR government has repeatedly expressed its sincerity and goodwill, violent activities by the opposition have been escalating, the social impact has continued to increase, and the rule of law in Hong Kong society has been openly disrupted. These activities have attracted great attention both from the central and Hong Kong governments. The movement is regarded as being the largest social unrest in Hong Kong for 52 years, since the 1967 riots, and the most serious political crisis since Hong Kong was returned to China in 1997. The evolution of the Anti-ELAB Movement in Hong Kong is closely related to the development of local public opinion. As a large-scale Internet public opinion event, massive amounts of text comment data have been generated on the Internet. How to analyse valuable public opinion and sentiment information from a massive amount of

The associate editor coordinating the review of this manuscript and approving it for publication was Sara Dadras.

public comments on the Internet is in the research category of big data affective computing and sentiment analysis [1], [2] which can be used in order to strengthen the monitoring of the Anti-ELAB Movement, and provide a public opinion response mechanism.

Sentiment analysis is becoming increasingly widely used to process data from online communities, blogs, Wikis, Weibo platforms and other forms of social media [3]. Sentiment analysis is also beginning to be applied to financial forecasting [4], political forecasting [5], e-tourism [6], and user profiles [7], among other fields. Although most authors approach it as a simple categorisation problem, sentiment analysis is actually a suitcase research problem [8], and many NLP tasks remain to be solved, including aspect extraction [9], word polarity disambiguation [10], sentiment and sarcasm classification with multitask learning [11].

In recent years, deep learning has gradually replaced traditional emotion analysis methods based on machine learning [1]. Konate and Du [12] proved that a deep learning model based on a single-layer convolutional neural network (CNN) performed better than machine learning, by applying sentiment analysis to multilingual text on Facebook. In addition, the difficulty of manual label acquisition in supervised learning means that experts are needed to formulate rules and perform manual labelling. Unsupervised learning lacks label guidance, and the effect is often unsatisfactory. Semi-supervised learning is therefore an effective solution. Hussain and Cambria [13] constructed a semi-supervised learning model for social data analysis, and showed that emotion recognition and polarity monitoring could be significantly improved. Akhtar *et al.* [14] proposed a stacked ensemble method for intensity prediction of sentiment and emotion, which proposed model performs significantly better than the state-of-the-art systems on the tasks of sentiment prediction of financial microblog messages and news headlines.

However, most of the above sentiment analysis methods use pre-trained word vectors as the basis of the model. These methods often use a unique word embedding representation for each word. Since this approach obviously cannot correctly represent polysemous words, we need to be able to accurately understand the semantics of each word in combination with the specific context. Although Perter *et al.* [15] proposed ELMo (Embeddings from Language Models) in an attempt to solve the above problems through a pre-trained bidirectional language model based on a stacked bidirectional LSTM, this approach was still limited by the coding ability of LSTM and the computational speeds of complex models. Devlin *et al.* [16] designed a BERT model with a multi-layered transformer encoder structure that was based on the transformer encoder constructed by Vaswani *et al.* [17], an approach which effectively solved the difficulties with parallel computing and depth with a LSTM pre-training model. The stacking model is difficult to converge and other issues, which broke the highest record of 11 NLP test tasks in the GLUE benchmark.

Although BERT has performed well in many NLP tasks, some scholars have questioned the perfect performance of BERT [18], [19]. Some scholars also proposed an improved model based on the BERT pre-training model. Liu presents a replication study of BERT, and he found that hyperparameter choices have a significant impact on the final results [20]. Joshi introduces a span-level pretaining approach that consistently outperforms BERT, with the largest gains on span selection tasks such as question answering and coreference [21]. Liu aims to combine Multi-task with BERT so that the model can be trained on more data and get better transfer ability at the same time [22]. The experimental corpus of the paper contains expressions in complex contexts such as Chinese and English, Cantonese, traditional characters, emojis, and so on. However, BERT lacks the input of a large-scale emotional corpus with a mixed context in the pre-training stage, causing it to hit a bottleneck when performing sentiment analysis tasks. In addition, the sentiment analysis method based on BERT has difficulties mining the semantics hidden behind the massive texts. Therefore, on this basis, we have made relevant improvements to the pre-training task of the BERT model and added a deep pre-training task [23]. Meanwhile, this is deeply integrated with the LDA topic model to increase the fine-grained sentiment topic analysis. We tried to solve two problems in our research: the first was how to improve and design the pre-training and in-depth pre-training tasks of the BERT model to achieve public sentiment classification in complex contexts. The second is how to deeply integrate the advantages of the BERT and LDA models so that the improved BERT-LDA model cannot only make up for the deficiencies of the LDA bag-of-words model, but also provide topic-dimensional semantic information for sentiment analysis tasks of large-scale complex text. In this way, we can fully grasp the netizen's emotional evolutionary context of the Hong Kong Anti-ELAB Movement, explore the evolutionary laws of such large-scale public opinion incidents, and provide a theoretical basis for government departments to formulate effective measures.

The main contributions of our proposed work are highlighted below: a) we effectively integrate the LDA topic model and the BERT word embedding model to optimize the topic vector to realize topic clustering; b) We improved the BERT pre-training task and superimposed the deep pre-training task; c) We propose a large-scale network public opinion event sentiment evolution analysis model. Simulation experiments show that the model has the advantages of high accuracy and high topic relevance.

The remainder of the paper is organized as follows: Section II introduces the LDA topic optimization model and the BERT deep learning improvement model, and describes our proposed model and algorithm in detail; empirical analysis of the Internet public opinion and sentiment in Hong Kong Anti-ELAB movement are presented in Section III; finally, Section IV concludes this paper.

II. CONSTRUCTION OF A LARGE-SCALE INTERNET PUBLIC OPINION EMOTION EVOLUTION MODEL BASED ON BERT-LDA

A. LDA TOPIC EXTRACTION OPTIMISATION MODEL

The LDA model gives the topic of each document in the document set in the form of a probability distribution and then performs topic clustering or text classification according to the topic distribution [24]. The modelling process is shown in Fig.1. We first define the document set $D = \{d_i | i \in \{1, 2, \ldots, M\}\}$ which is composed of M documents; The document $d_i = \{d_{is} | s \in \{1, 2, \ldots, S\}\}$ contains S sentences, and $d_i = \left\{w_{ij} | j \in \{1, 2, \ldots, N_i\}\right\}$ consists of N_i words. Each word corresponds to a potential topic. The topic set corresponding to d_i is $z_i = \{z_{ij} | j \in \{1, 2, \ldots, N_i$; Therefore, the total number of topics in document set D is $l = \sum_{i=1}^{M} count(z_i)$, and the total number of words is $N = \sum_{i=1}^{M} N_i$. The LDA modelling process can be analysed as follows:

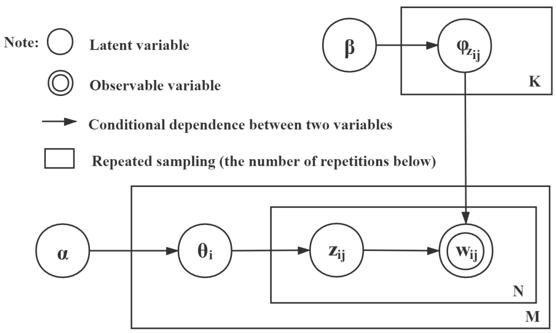


FIGURE 1. LDA topic analysis model.

Step 1: Set the initial probability of each document d_i to $P(d_i)$;

Step 2: Given the parameter α, which obeys the Dirichlet distribution, sample from the distribution to generate the topic distribution θ_i for document d_i;

Step 3: Sample from the topic distribution θ_i to generate the theme z_{ij} of the j$^{\text{th}}$ word in the document d_i;

Step 4: Given the parameter β which obeys the Dirichlet prior distribution, sample from this distribution to generate the word distribution $\varphi_{z_{ij}}$ of topic z_{ij};

Step 5: Generate word w_{ij} from the word distribution $\varphi_{z_{ij}}$.

Set the joint distribution of all variables in the LDA model as shown in (1):

$$P(w_i, z_i, \theta_i, \Phi | \alpha, \beta) = \prod_{j=1}^{N_i} P(w_{ij} | \varphi_{Z_{ij}}) P(z_{ij} | \theta_i) \cdot P(\theta_i | \alpha) \cdot P(\Phi | \beta) \quad (1)$$

where α and β obey the prior Dirichlet distribution; θ_i is the probability distribution of the text topic; $P(\theta_i | \alpha)$ is the 'text topic' distribution probability of the document d_i generated by the Dirichlet prior parameter α;$P(z_{ij} | \theta_i)$ is the topic probability corresponding to the j$^{\text{th}}$ word of the sampled document d_i in the topic distribution θ_i;$P(\Phi | \beta)$ is the Dirichlet prior parameter β to generate the 'topic word' distribution matrix of topic z_{ij}; $P(w_{ij} | \varphi_{Z_{ij}})$ is the probability of generating word w_{ij} in word distribution $\varphi_{z_{ij}}$.

The next step is to use the Gibbs sampling algorithm to estimate the parameters, iterative sampling until convergence and obtain the 'topic word' distribution matrix φ and 'text topic' distribution θ. In the calculation of the original model, the topic vector $\boldsymbol{\mu}$ is calculated from the cosine distance between the high-frequency words of each topic and the document. Subject words are limited by the traditional structure of the LDA bag-of-words model, which cannot effectively combine the semantic and contextual information of the text, and the quality of word segmentation technology has a greater impact on the calculation results of $\boldsymbol{\mu}$. In order to obtain a better topic vector, we fused the word embedding in the BERT model with the topic representation of the LDA model and obtained the optimised topic embedding $\boldsymbol{\mu}$' through iterative calculation of word weights. This approach is expected to provide more accurate topic semantic information for emotion simulation for large-scale complex texts.

B. IMPROVED BERT DEEP LEARNING MODEL

Bidirectional Encoder Representations from Transformers (BERT) [16] is a language pre-training model based on the bidirectional Transformer structure released by Google in 2018. By masking language model (MLM) and next sentence prediction (NSP) tasks, the semantic representation ability of the model is enhanced. It also relies on Transformer's powerful feature extraction and fine-tuning transfer learning capabilities to make it stand out in a number of NLP tasks. In order to improve accuracy of sentiment classification tasks in the analysis of fine-grained, large-scale complex text, we retain the BERT basic model, and add the deep pre-training task on the basis of improving the pre-training task. At the same time, we embed the topic optimisation results of the LDA model in Part A of Section II into the BERT pre-training and fine-tuning stages, so that the model can simultaneously learn text features such as syntax, semantics, and topics when performing sentiment classification tasks. The improved BERT model is shown in Fig. 2.

We set $\boldsymbol{\omega}$, $\boldsymbol{\delta}$, and $\boldsymbol{\rho}$ as the word embedding, segment embedding, and position embedding of the text; *Trm* stands for Transformer encoder unit;$d_i' = \left\{w_{ij}' | j \in \{1, 2, \ldots, N_i\}\right\}$ is a vector set of words in the document d_i that incorporates an improved full-text semantic information. For the special characters inserted in the text, we have made the following definitions (see Table 1).

According to Fig.2, after the word segmentation document $d_i = \left\{w_{ij} | j \in \{1, 2, \ldots, N_i\}\right\}$ is input into the model, each word is mapped to the sum of three vectors $\mathbf{w_{ij}}$ ($\boldsymbol{\omega}\ \mathcal{C}\ \boldsymbol{\delta}\ \mathcal{C}\ \boldsymbol{\rho}$). They are collectively referred to as word vectors. In the process of improvement, the word vector $\mathbf{w_{ij}}$ ($\boldsymbol{\omega}\ \mathcal{C}\ \boldsymbol{\delta}\ \mathcal{C}\ \boldsymbol{\rho}$) is combined with the topic vector optimised by LDA to obtain the word vector $\mathbf{w_{ij}}$ ($\boldsymbol{\omega}\ \mathcal{C}\ \boldsymbol{\delta}\ \mathcal{C}\ \boldsymbol{\rho}\ \mathcal{C}\ \boldsymbol{\mu}'$) that integrates the better topic vector. These are then passed into the

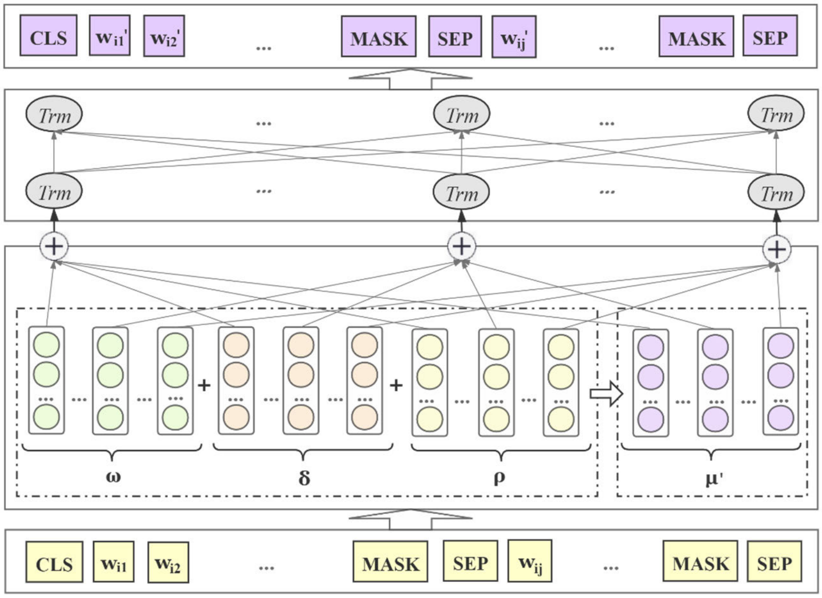


FIGURE 2. Structure of Improved BERT Model.

TABLE 1. Cross-references of text special characters.

Character	Meaning
[CLS]	Start-of-text character
[SEP]	Text spacer and terminator
[MASK]	Cover word
[MUN]	Number
[DEL]	Deleted characters
[PAD]	Fill in the blanks

bidirectional Transformer encoder. In order to learn more information from the model, the Transformer encoder connects the multi-headed attention and feed-forward [25] through a residual network structure. The multi-headed mechanism performs multiple linear transformations on the input vector to obtain different linear values, and then calculates the attention weight. The calculation process is shown in (2) and (3):

$$MultiHead\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = Concat(head_1, head_2, \ldots, head_h)\boldsymbol{W}^{\boldsymbol{O}} \tag{2}$$

$$head_f = Attention\left(QW_f^Q, KW_f^K, VW_f^V\right) \tag{3}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are the input word vector matrix. We calculate Attention after $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are mapped through the parameter matrix, and the calculation results are spliced after repeating h times. $head_f = \{head_f | f \in \{1, 2, \ldots, h$ represents the f$^{\text{th}}$ hyperparameter header. $\boldsymbol{W}^{\boldsymbol{O}}$ is the weight matrix. W_f^Q, W_f^K, W_f^V represents the weight matrix of $\boldsymbol{W}^{\boldsymbol{Q}}$, $\boldsymbol{W}^{\boldsymbol{K}}$, $\boldsymbol{W}^{\boldsymbol{V}}$ corresponding to the f$^{\text{th}}$ hyperparameter head.

Thus, the Transformer encoder learns and stores the semantic relationship and grammatical structure information of the document $\boldsymbol{d_i}$, since the document $\boldsymbol{d_i}$ incorporates better topic feature vectors after the improvement of the BERT model. Therefore, the improved document $\boldsymbol{d_i'} = \left\{\boldsymbol{w_{ij}}' | \boldsymbol{j} \in \{1, 2, \ldots, \boldsymbol{N_i}\}\right\}$ is connected to the output layer of softmax through special characters [CLS] to adapt to Transfer Learning under multi-tasking. The combination of the feature vector of the topic optimisation information and the BERT word embedding is expected to improve the accuracy and give a finer-grained model for large-scale complex text emotion evolution analysis.

C. SENTIMENT EVOLUTION ANALYSIS MODEL OF LARGE-SCALE INTERNET PUBLIC OPINION BASED ON IMPROVED BERT-LDA

The sentiment analysis model aims to analyse accurately the process of emotional evolution of large-scale Internet public opinion, grasp the characteristics and rules of group emotional evolution, and assist government departments to quickly form an effective public opinion response mechanism. We propose the improved BERT-LDA simulation model for the public opinion sentiment evolution (Fig.3). First of all, the corpus TB is pre-trained in the BERT general domain, after data pre-processing and complex corpus dictionary construction, the text word vectors were mapped, and then the train the LDA model and construct BERT sentiment classifier. Secondly, the corpus TW is used to perform in-depth pre-training of the BERT target field on the constructed sentiment classifier. Finally, the Epoch training results are transferred to the BERT fine-tuning model, and the LDA topic emotion calculation is performed.

Because the BERT model lacks emotional corpus training in the pre-training stage, it performs poorly when carrying out sentiment classification tasks. Therefore, we formulated

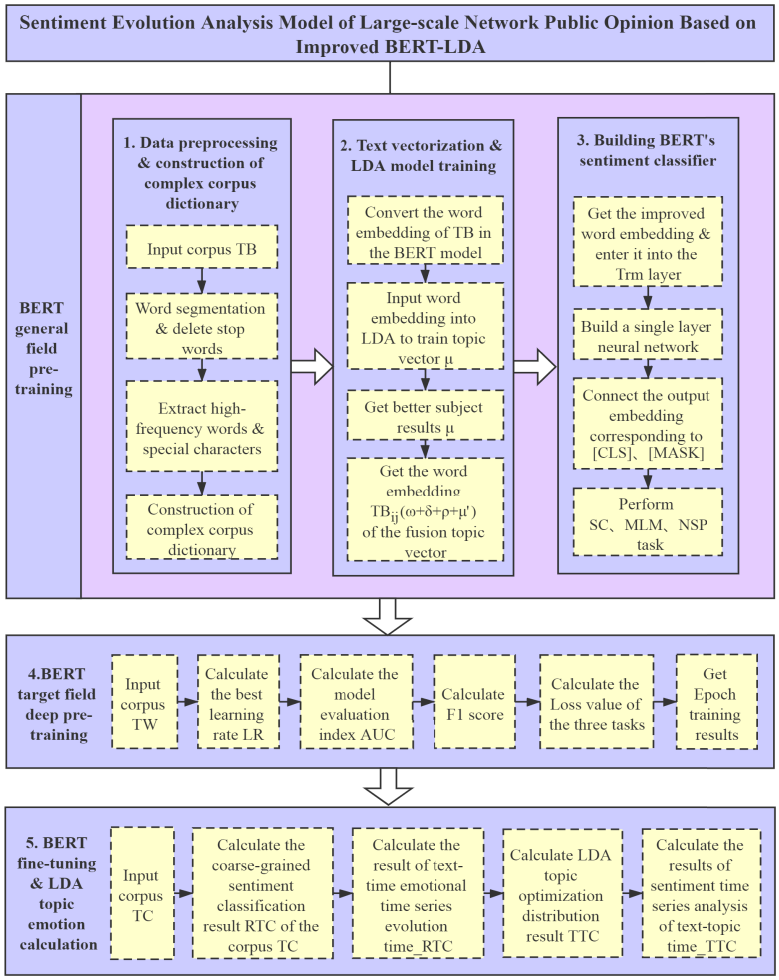


FIGURE 3. Improved BERT-LDA Large Scale Network Public Opinion Sentiment Classification Model.

a new pre-training task for BERT and introduced an improved pre-training corpus; $TB = \{\boldsymbol{TB_i} | \boldsymbol{i} \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{M}\}\}$. That is to say, on the basis of the original Chinese Wikipedia corpus, the emotional corpus of Sina Weibo and Baidu Tieba is added, so that the model can learn more emotional information. At the same time, the public Sina Weibo annotation set and a small amount of specific event emotional annotation set TW are introduced as the BERT deep pre-training corpus. Finally, we define $TC = \left\{\boldsymbol{TC_i} | \boldsymbol{i} \in \{\mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{M}\}\right\}$ as the sentiment classification task corpus of a specific event.

The specific improved model algorithm is described as follows:

Step 1: Data pre-processing and construction of a complex corpus dictionary. We input the corpus *TB* and perform data pre-processing such as word segmentation and delete stop word. Then, we count word frequency information, extract high-frequency words and add special characters to form a complex corpus dictionary; each word corresponds to a unique index.

Step 2: Text vectorisation and LDA model training. After the corpus *TB* is input to the BERT pre-training model, each word is mapped into a word vector $\boldsymbol{TB_{ij}}$ $(\boldsymbol{\omega}\ \boldsymbol{\mathcal{C}}\ \boldsymbol{\delta}\ \boldsymbol{\mathcal{C}}\ \boldsymbol{\rho})$. Next, this is input into the LDA model to assist in training the topic vector μ. After iterative calculation, a better topic vector $\boldsymbol{\mu}'$ is obtained. In other words, the probability distribution of one optimal topics and different 'topic words' is calculated, and the word vector fused with the topic vector is obtained followed by the $\boldsymbol{TB_{ij}}$ $(\boldsymbol{\omega} + \delta + \rho + \mu')$.

Step 3: Building the BERT sentiment classifier. We pass the feature vector $\boldsymbol{TB_{ij}}$ $(\boldsymbol{\omega}\ \boldsymbol{\mathcal{C}}\ \boldsymbol{\delta}\ \boldsymbol{\mathcal{C}}\ \boldsymbol{\rho}\ \boldsymbol{\mathcal{C}}\ \boldsymbol{\mu'})$ output by Step 2 into the bidirectional Transformer encoder. Then we construct a single-layer neural network to connect the output vector corresponding to [CLS] in Transformer as a classifier to perform sentiment classification (SC) tasks. At the same time, MLM and NSP tasks are retained, which are connected to the output vectors corresponding to [MASK] and [CLS] respectively.

Step 4: Deep pre-training of BERT target field. In the neural network constructed by Step 3, we input the corpus set TW, set the initial learning rate, batch size, dropout rate, and epoch training times. In order to avoid over-fitting in the training process, the Adam optimisation algorithm [26] is used to find the best learning rate. Also, we use AUC [27] as an index to test the evaluation effect of the epoch model. The F1 score [28] is used to find the best threshold for positive and negative emotion classification, and the sum of the losses of the three tasks [29] of MLM, NSP, and SC is used as the loss function of the deep pre-training model:

$$Loss = \lambda_1 Loss_{MLM} + \lambda_2 Loss_{NSP} + \lambda_3 Loss_{SC} \tag{4}$$

where λ is the weight distribution of the three tasks. Finally, the results of epoch training are output.

Step 5: BERT fine-tuning and LDA topic emotion calculation. The epoch training results of step 4's deep pre-training model are transferred to the fine-tuning model, and the *TC* of corpus is input for data pre-processing. Then the sentiment value $E\,(TC_i)$ is calculated of each document TC_i to obtain the coarse-grained sentiment classification result *RTC*:

$$RTC = (MnE_{pos}(TC_i), MnE_{neg}(TC_i), VarE_{pos}(TC_i), VarE_{neg}(TC_i)) \tag{5}$$

where $E\,(TC_i)$ is conditionally judged according to Boolean expression. When $E\,(TC_i) > 0$, it is judged as positive emotion; when $E\,(TC_i) < 0$, it is judged as negative emotion. According to the conditional judgment results, we calculate the positive emotion mean value $MnE_{pos}(TC_i)$, negative emotion mean value $MnE_{neg}(TC_i)$, positive emotion variance value $VarE_{pos}(TC_i)$ and negative emotion variance value $VarE_{neg}(TC_i)$. Then the *RTC* is integrated into the time series text set *TC_time*:

$$TC\_time = \{(TC_1, time_1), (TC_2, time_2), \ldots, (TC_M, time_q)\} \tag{6}$$

where time contains q time slices, the coarse-grained 'text time' emotional temporal evolution result *time_RTC* is obtained

$$time\_RTC = \{time\_RTC_1, time\_RTC_2, \ldots, time\_RTC_q\} \tag{7}$$

Next, we input the corpus *TC* into the optimised LDA topic model for the iterative calculation to obtain the document distribution *TTC* corresponding to the 'topic word' under L optimal number of topics. It is integrated into the time series text set to get the time series evolution results under different LDA topics.

Finally, the document distribution *TTC* corresponding to the 'topic word' is combined with the 'text time' sentiment sequence evolution result *time_RTC* to obtain the multidimensional and fine-grained 'text topic' sentiment analysis result:

$$time\_TTC = \{(time\_TTC_1, l), (time\_TTC_2, l), \ldots, \left(time\_TTC_q, l\right)\} \tag{8}$$

By constructing an improved BERT-LDA evolution analysis model, we can not only fully combine contextual semantic information (thus making up for the lack of an LDA bag-of-words model) and train better topic vectors, but also deeply integrate the optimised topic vectors with BERT word embedding. Finally, we obtain finer-grained and more accurate results for text topic clustering and public opinion evolution analysis.

III. AN EMPIRICAL ANALYSIS OF INTERNET PUBLIC OPINION AND SENTIMENT IN HONG KONG ANTI-ELAB MOVEMENT

A. DATA SOURCES

In order to analyse deeply the public opinion and the process of emotional evolution of netizens behind large-scale public opinion incidents which are part of the Hong Kong Anti-ELAB Movement, we used 'Xiu Li', 'Fan Song Zhong', 'Tao Fan Tiao Li', and 'Yin Du Tiao Li' as search keywords to build a Python crawler framework. We captured the posts and comment data regarding the Anti-ELAB Movement from a well-known forum in Hong Kong from 0:00 on 25 April, 2019, to 24:00 on 20 August, 2019, ultimately collecting 125,963 posts and comment data (i.e. corpus *TC*) in complex contexts such as Cantonese, Chinese, and English, as well as emoticons. First, we performed text data pre-processing on the corpus *TC*. Specifically, we used the Translate function of Baidu API to process the comment text data from traditional to simplified and Cantonese to Chinese. Secondly, we used Python regular expression findall and subn functions to extract the English words in the comment text, and then backfilled them after machine translation, having performed regular expression matching and Sinicization on the emojis in the comment text. Third, batch Chinese word segmentation of the text was realised by calling on the Segmentor function in the LTP word segmentation tool of the Harbin Institute of Technology. Finally, we eliminated invalid data and meaningless URLs. Through iterative processing, 105,349 valid data (including 6,066 posts and 99,283 comments) were obtained. The data set was classified into a training set and a test set according to the ratio 7:3; 73,744 and 31,605 respectively.

B. ITERATIVE OPTIMISATION OF PRE-TRAINING AND DEEP PRE-TRAINING MODELS

In accordance with Steps 1 to 4 in the algorithm of part C of Section II, first, we input the corpus *TB* including Chinese Wikipedia, Sina Weibo, and Baidu Tieba, yielding a total of 9,352,561 entries. We filtered 26,523 words and special characters with high frequency to form a complex corpus dictionary. Then, the corpus TB was vectorised and LDA subject training was performed. The document TB_i after the word segmentation of the corpus *TB* was transformed into a document TB_i composed of N_i word vectors $\boldsymbol{TB_{ij}}(\boldsymbol{\omega\, \mathcal{C}\, \delta\, \mathcal{C}\, \rho})$ and input into the LDA model. The Gibbs sampling algorithm was used to obtain the parameters θ and φ, and then the normalised pointwise mutual information (NPMI) [30] was used as the evaluation index of topic relevance. The optimal topic number l and the co-occurrence frequency matrix of 'topic

TABLE 2. Average topic relevance.

Model	LDA	BERT-LDA
NPMI value	0.346	0.703

word', as well as the word vector and topic vector of each topic were obtained. We drew on the word vector training method of the character-word joint of Chen *et al.* [31] and integrated the word vector of the BERT model in the training process of the LDA topic vector. In order to verify the accuracy of the model, a comparative experiment was performed with the original LDA subject vector training method, and the results are shown in Table 2. The results show that subject vector training through a variety of vector fusion can get better results in subject relevance.

Then, the optimised topic vector was embedded in the BERT model, and the word vector TB_{ij} (ω $\mathcal{C}$ δ $\mathcal{C}$ ρ $\mathcal{C}$ μ') of the fused topic vector was obtained and passed into the bidirectional Transformer encoder. On the basis of the BERT-base model, we trained five epochs on a GTX1080ti and built a BERT sentiment classifier to pass the epoch training results into Step 4. In the in-depth pre-training stage of the BERT target field, we input the corpus *TW*, which contained the publicly annotated data set of Sina Weibo (41,768 positive samples and 59,518 negative samples) and 5% of the manual annotation data set of the Anti-ELAB Movement (1,041 positive samples and 1,521 negative samples), a total of 103,848. We randomly shuffled through the shuffle function to obtain 73,316 training sets and 31,421 test sets of the corpus *TW*. We set the initial learning rate as 1e-6, the batch size was 24, and the dropout rate was 0.4. The training process is shown in Fig.4. When the epoch reached 400, the AUC value did not increase, so we set the epoch training times to 400.

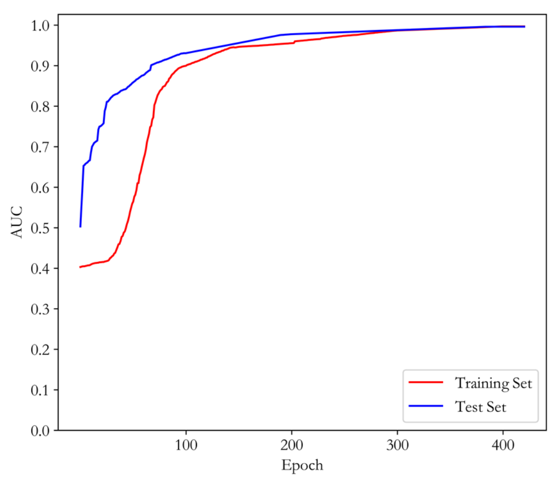


FIGURE 4. Iterative Process of BERT-LDA Deep Pre-training Model.

The optimal loss value of the test set was calculated to be 0.065, and the optimal AUC value was 0.9962. As a result, the trained pre-training model was transferred to the fine-tuning model of the sentiment classification task for the Anti-ELAB Movement, in order to analyse the emotional evolution trends of the Anti-ELAB Movement in-depth.

C. AN EMPIRICAL ANALYSIS OF THE EMOTIONAL EVOLUTION OF HONG KONG's ANTI-ELAB MOVEMENT

In order to analyse the emotional evolution process of posting and comment data in different time periods, we performed a statistical analysis and data visualisation on 105,349 posts and comment data (corpus *TC*) after pre-processing. According to the time nodes of important events and the number of posts and comments, we divided the Hong Kong Anti-ELAB Movement into three-time segments ($q = 3$), where q_1 was the incubation period (25 April, 2019–31 May, 2019), q_2 was the formation period (1 June, 2019–9 July, 2019), and q_3 was the period of volatility (10 July, 2019–20 August, 2019). The analysis results are shown in Fig.5 (the figure above shows the analysis results of comment data, and the figure below shows the analysis results of posting data). It is not difficult to see that the trend of the comment volume curve and the posting volume curve in the figure is basically consistent. During the latent period, there were no major waves, but in the formation period, there were two peaks of public opinion in the volume of comments and posts. Through tracing the source, we found that the two peaks featured intense discussions around the parade on the day. During the period of fluctuation, the volume of public opinion varied, and the trend of public opinion volume was highly consistent with the time axis of the demonstration. The high peak value and demonstration activities generally occurred at the weekend.

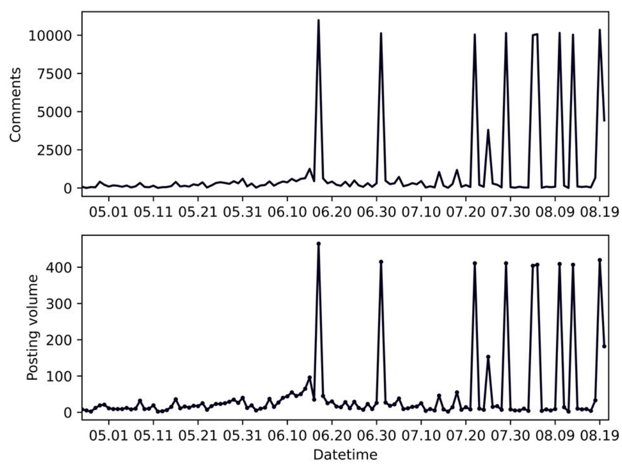


FIGURE 5. The evolution of the number of comments and posts in the Hong Kong Anti-ELAB Movement.

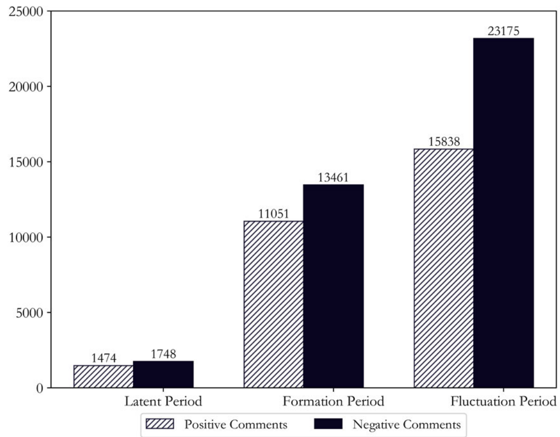
For the sentiment classification task of the Anti-ELAB Movement, we transferred the epoch training results output of Step 4 to Step 5 for fine-tuning. In accordance with the BERT fine-tuning model, the sentiment value of each document in the *TC* was calculated, and then converted into the corresponding emotional polarity according to the Boolean condition judgment, and the number of positive and negative texts and the emotional mean value were calculated to obtain the $RTC = (MnE_{pos}(TC_i), MnE_{neg}(TC_i), VarE_{pos}(TC_i), VarE_{neg}(TC_i))$ of public opinion sentiment classification of the Anti-ELAB Movement. We integrated RTC into the time series set $TC\_time(q = 3)$ to obtain the coarse-grained 'text time' emotional time series evolution result $time\_RTC = \{time\_RTC_1, time\_RTC_2, time\_RTC_3$. It can be visualised as the classification map of emotional tendency (a), the mean value and variance of positive emotion (b), and the mean value and variance of negative emotion (c) as shown in Fig. 6.

According to Fig.6(a), the number of positive and negative comments during the q_2 period both increased significantly, and public opinion began to shift from confrontation to a situation dominated by negative comments. During the q_3 period, the number of negative comments far exceeded the positive comments. With the continuous evolution of the Anti-ELAB Movement, netizens paid increasing amounts of attention to it, and there were more and more critical and negative emotions. According to the emotional transmission law of public opinion events, netizens are more likely to express negative emotions such as anger and anxiety after public opinion events occur [21]. When negative emotions lack the guidance of positive opinion leaders, netizens are more likely to go to extremes. From Fig.6(b) and Fig.6(c), with the evolution of the Anti-ELAB Movement, the value of $MnE_{pos}(TC_i)$, $MnE_{neg}(TC_i)$, $VarE_{pos}(TC_i)$, $VarE_{neg}(TC_i)$ all decreased. In the q_2 and q_3 periods, as the number of netizens participating in the discussion increased, netizens' moods became more stable, and their attitudes became clearer, although public opinion tended to be negative overall.

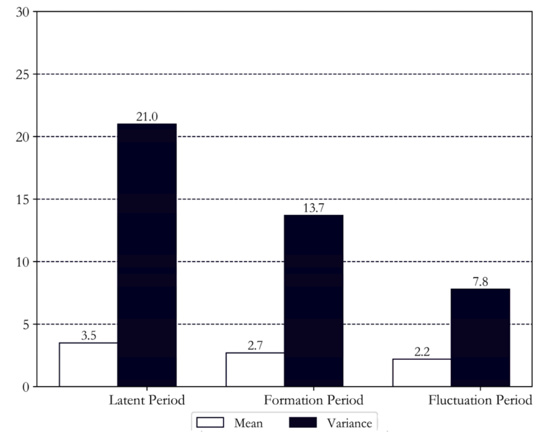
In order to corroborate the above point of view and analyse deeply the key role of opinion leaders in the process of netizens' emotional evolution, we also conducted a topic analysis and calculated the emotional value of the posts and comments of the other 10 opinion leaders in the q_2 and q_3 periods. It turns out that these opinion leaders are not only the guides of public opinion but also the producers of public opinion. The public opinion comments made by these opinion leaders tended to be negative. Therefore, the more opinion leaders commenting on the incident, the greater the negative impact on netizens. At the same time, according to the silent spiral theory, the number of netizens who started to hold positive attitudes gradually decreased as negative attitudes became increasingly dominant. These factors together caused the phenomenon of negative emotions dominating during q_2 and q_3.

D. A FINE-GRAINED ANALYSIS OF THE EMOTIONAL EVOLUTION OF THE TOPIC OF THE HONG KONG ANTI-ELAB MOVEMENT

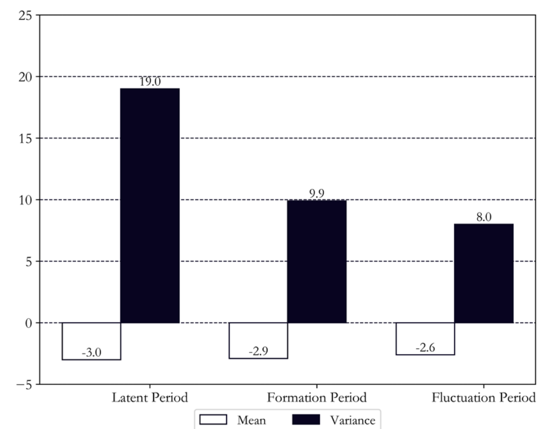
In order to further analyse the topic emotion evolution trend of netizens in the Anti-ELAB Movement, and provide a reasonable basis for the emotion evolution distribution results,



(a) Classification map of emotional tendency



(b)The mean value and variance of positive emotion



(c) the mean value and variance of negative emotion

FIGURE 6. The overall emotional evolution distribution of the Anti-ELAB Movement in different periods.

we introduced the improved BERT-LDA model to analyse the fine-grained topic emotion in accordance with Step 5 of the third part of the algorithm in Section II. After many experiments, we finally chose the a priori parameter combination of $\alpha = 50/L$ and $\beta = 0.01$ and use 5-fold cross-validation to iterate 2,000 times on the training set and 1,000 times on the test set. As shown in Table 3, after iterative calculation, when $l = 3$, the topic similarity is 0.7102. At this time, the topic

TABLE 3. The topic iterative optimisation results of the anti-elab movement after the improvement of the bert-lda algorithm.

Number of topics	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Average similarity
3	Hong Kong	Escapee	Regulations					0.744
4	Demonstration	Hong Kong	Government	Mainland				0.831
5	Government	Demonstration	Hong Kong	Legislation	Hong Kong			0.77
7	Hong Kong	Government	Government	Government	Regulations	Anti-ELAB	Parade	0.793

TABLE 4. The naming results of public opinion topic of the anti-elab movement in different periods.

	X_1		X_2				X_3	
d_1	$(TC_1, q_1)_1$	Amendment	$(TC_1, q_2)_1$	Fan Song Zhong			$(TC_1, q_3)_1$	Fan Xiu Li
	$(TC_1, q_1)_2$	Transfer to the Legislative Council	$(TC_1, q_2)_2$	Fugitive Offenders Ordinance			$(TC_1, q_3)_2$	Fan Song Zhong
	$(TC_1, q_1)_3$	Fugitive Offenders Ordinance	$(TC_1, q_2)_3$	Fan Xiu Li			$(TC_1, q_3)_3$	Anti-ELAB
d_2	$(TC_2', q_1)_1$	Hong Kong middle school	$(TC_2', q_2)_1$	Hong Kong	$(TC_2'', q_2)_1$	Hong Kong	$(TC_2'', q_3)_1$	Hong Kong Demonstration
	$(TC_2', q_1)_2$	Hong Kong Fugitive Offenders Ordinance	$(TC_2', q_2)_2$	Legislation	$(TC_2'', q_2)_2$	Lackey	$(TC_2'', q_3)_2$	Neighborhood
	$(TC_2', q_1)_3$	Draft legislation	$(TC_2', q_2)_3$	Fugitive Offenders Ordinance	$(TC_2'', q_2)_3$	Legislator	$(TC_2'', q_3)_3$	Mainland government
d_3	$(TC_3, q_1)_1$	Hong Kong middle school parade	$(TC_3, q_2)_1$	Revival Hong Kong			$(TC_3, q_3)_1$	Violence
	$(TC_3, q_1)_2$	Government Fugitive Offenders Ordinance	$(TC_3, q_2)_2$	Demonstration			$(TC_3, q_3)_2$	Hong Kong Police
	$(TC_3, q_1)_3$	Hong Kong	$(TC_3, q_2)_3$	Hong Kong government			$(TC_3, q_3)_3$	Yuen Long reporter

similarity reaches the minimum, and the number of topics is optimal.

The document distribution *TTC* corresponding to the optimal number of topics ($l = 3$) after iteration was integrated into the time series text set *TC_time (q=3)*. We set the document composed of the title of the post as TC_1, the document composed of the content of the opinion leader's post, and the comment content of the post as TC_2; and the document composed of the comment content of all posts was TC_3, divided into three-time segments q_1, q_2, and q_3 for modelling. We selected the netizen 'Yiyu∗∗ shentan' with the most posts (TC_2', 153 posts and 1,645 comments on the posts) and the netizen 'Ji∗∗he' with the most comments, (TC_2'', 12 posts, 83,037 comments on the posts) as the representatives of opinion leaders in the Anti-ELAB Movement. A total of 10 LDA topic results were generated, which were (TC_1, q_1), (TC_1, q_2), (TC_1, q_3), (TC_2', q_1), (TC_2', q_2), (TC_2', q_2), (TC_2'', q_3), (TC_3, q_1), (TC_3, q_2), (TC_3, q_3). The respective subject words are presented in word cloud form according to probability, as shown in Fig.7.

In Fig.7, since each LDA topic model is assigned 3 topic classifications, 30 different topic results are calculated by evolution. Each topic is set to contain 9 keywords, and the topic is named according to the probability distribution of keywords. The topic naming results are shown in Table 4. Taking $(\boldsymbol{TC_3}, \boldsymbol{q_1})$ as an example, in its three topic classifications, the keywords with the highest occurrence probability of topic 1 are: P(Hong Kong) = 0.0159, P (Middle School) = 0.0136, P(Parade) = 0.0116, the subject 1 $(\boldsymbol{TC_3}, \boldsymbol{q_1})_{\boldsymbol{1}}$ was named 'Hong Kong Middle School Parade'. In the same way, topic 2 $(\boldsymbol{TC_3}, \boldsymbol{q_1})_{\boldsymbol{2}}$ was named 'Government Fugitive Offenders Regulations'; topic 3 $(\boldsymbol{TC_3}, \boldsymbol{q_1})_{\boldsymbol{3}}$ was named 'Hong Kong'.

According to the theme naming results in Table 4 and the evolutionary distribution in Fig.3, from the time dimension analysed, the topic of the q_1 period was an objective discussion of the main body of the Anti-ELAB Movement. Netizens were in a rational state, without radical speech and offline activities, and in the period of latent risk and public opinion origination. The theme of the q_2 period was more distinctive. Both 'Fan Song Zhong' and 'Fan Xiu Li' clearly indicated the subjective attitudes of netizens to the Anti-ELAB Movement, and their speeches were gradually becoming extreme. During this period, public opinion was boiling, and rumours were spreading. q_3 was an expansion of public opinion which developed during the period of q_2. There have been seven peaks in public opinion. During this period, netizens began to move from online condemnation to offline activities. Public opinion became stronger and more consistent.

In order to further study, the coupling degree between opinion leaders and netizens' opinions in the process of the Anti-ELAB Movement, we made a horizontal comparison of the thematic characteristics among different research subjects. It was not surprising to find that the theme features of TC_2 and TC_3 were highly consistent, which to a certain extent reflects the role of opinion leaders in the event as leaders of public opinion; their ability to guide public opinion overall cannot be underestimated. This conclusion is consistent with the point of view of Chen Yixin *et al.* [32]. He believes that the more consistent the group views become after stability,

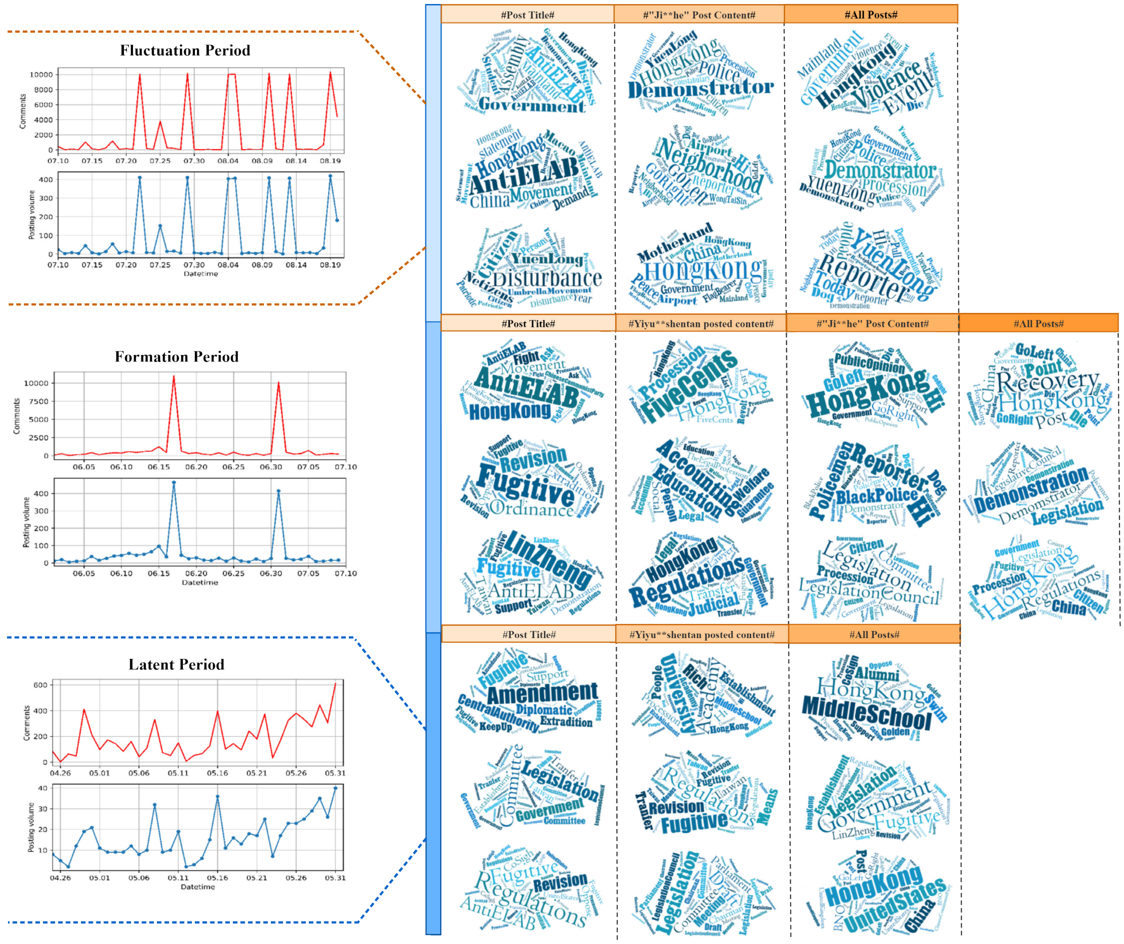


FIGURE 7. The word cloud of LDA Topic Evolution of the Anti-ELAB Movement in different periods.

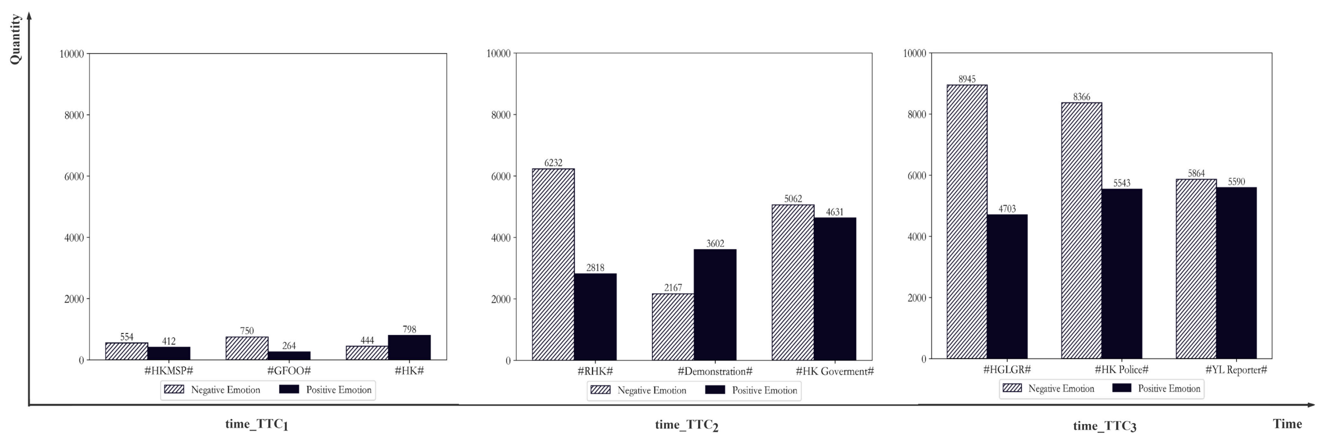


FIGURE 8. Emotional evolution analysis of the Anti-ELAB Movement under 'text topic' in different periods.

the closer the average views are to the opinions of opinion leaders.

Based on the emotional evolution distribution analysed, in order to further explore the emotional evolution trend under the perspective of multi-dimensional topics, and in accordance with Step 5 in the algorithm of part C of Section II, we performed sentiment analysis under multiple dimensions of 'text topic' on the full sample data. The sentiment analysis result $time\_TTC = \{(time\_TTC_1, 3), (time\_TTC_2, 3), (time\_TTC_3, 3)\}$ is shown in Fig. 8.

According to Fig.8, 'HKMSP' is the abbreviation of Hong Kong middle school parade; 'GFOO' is the abbreviation of the government Fugitive Offenders Ordinance; 'HK' is the abbreviation of the Hong Kong; 'RHK' is the abbreviation of the revival Hong Kong; 'HGLGR' is the abbreviation of the hi, go left, go right; 'YL Reporter' is the abbreviation of the Yuen Long reporter. Obviously, netizens had the strongest negative emotions on the topic of 'revival Hong Kong' during the q_2 period, and the negative comments on the topics of 'violent incidents' and 'Hong Kong police' during the q_3 period far exceeded the positive comments. This shows that netizens were most concerned about three themes, and these triggered a fierce manifestation of public opinion. The theme of 'revival Hong Kong' in the forum was closely related to the extreme slogan of 'revival Hong Kong, the revolution of the times' which appeared in the demonstration activities. The themes of 'violent incidents' and 'Hong Kong police' in the forum are also related to the violent clashes between radicals and the Hong Kong police. This proves that in the Hong Kong Anti-ELAB Movement, Internet public opinion deeply reflected the influence of demonstrations, and demonstrations also contributed to the spread of Internet negative public opinion. Therefore, the control of such large-scale public opinion events could start with the identification of emotional themes with the greatest negative emotions, combining the Internet and real life, while paying attention to the specific events reflected by such themes, and by finding targeted public opinion solutions.

IV. CONCLUSION

We focussed on the hot Internet public opinion incident of the Anti-ELAB Movement. An improved BERT-LDA hybrid model was constructed in a complex Cantonese context, involving the mixture of Chinese and English, as well as traditional characters and emoticons. Through the analysis and prediction of sentiment evolution of public opinion data, we attempted to investigate the laws of emotional evolution for such large-scale public opinion events, and we proposed an emotion value measurement algorithm based on machine learning. We analysed the theme characteristics of different research subjects from multi-dimensional perspectives through the analysis of 'theme' evolution, and deeply analysed the netizens' public opinion attitudes towards different topics through the emotional analysis of 'text theme'. We use the BERT-LDA hybrid model to predict the evolution of public opinion and emotional trends in complex contexts, with an average error rate of not more than 9.95%.

The empirical study shows that the event is still in a fluctuation period from the perspective of the recent evolution and prediction of public opinion, and there is no sign of public opinion dissipation; it may even foment to a certain extent. From the perspective of the emotional evolution law of such large-scale public opinion events, opinion leaders are the guides of public opinion and they are also the producers of public opinion. When public opinion is one-sided, the emotional tendency of netizens is close to that of opinion leaders. From the perspective of public opinion control measures, the intervention and control effect of the Hong Kong government can be strengthened. When dealing with relevant public opinion incidents, the government should actively seize the position of opinion leaders, increase the cost of opinion leaders' malicious participation, restrict the malicious incitement of negative opinion leaders, and promote the early settlement of the Anti-ELAB Movement as soon as possible.

REFERENCES

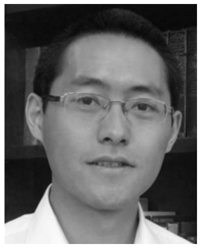
[1] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016.
[2] A. S. M. Alharbi and E. de Doncker, "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information," *Cognit. Syst. Res.*, vol. 54, pp. 50–61, May 2019.
[3] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, *A Practical Guide to Sentiment Analysis*. Cham, Switzerland: Springer, 2017.
[4] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: A survey," *Artif. Intell. Rev.*, vol. 50, no. 1, pp. 49–73, Jun. 2018.
[5] M. Ebrahimi, A. H. Yazdavar, and A. Sheth, "Challenges of sentiment analysis for dynamic events," *IEEE Intell. Syst.*, vol. 32, no. 5, pp. 70–75, Sep. 2017.
[6] A. Valdivia, M. V. Luzon, and F. Herrera, "Sentiment analysis in tripadvisor," *IEEE Intell. Syst.*, vol. 32, no. 4, pp. 72–77, Aug. 2017.
[7] R. Mihalcea and A. Garimella, "What men say, what women hear: Finding gender-specific meaning shades," *IEEE Intell. Syst.*, vol. 31, no. 4, pp. 62–67, Jul. 2016.
[8] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov. 2017.
[9] S. Poria, I. Chaturvedi, E. Cambria, and F. Bisio, "Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 4465–4473.
[10] Y. Xia, E. Cambria, A. Hussain, and H. Zhao, "Word polarity disambiguation using Bayesian model and opinion-level features," *Cogn. Comput.*, vol. 7, no. 3, pp. 369–380, 2015.
[11] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, and E. Cambria, "Sentiment and sarcasm classification with multitask learning," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 38–43, May 2019.
[12] A. Konate and R. Du, "Sentiment analysis of code-mixed bambara-French social media text using deep learning techniques," *Wuhan Univ. J. Natural Sci.*, vol. 23, no. 3, pp. 237–243, Jun. 2018.
[13] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, Jan. 2018.
[14] M. S. Akhtar, A. Ekbal, and E. Cambria, "How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes]," *IEEE Comput. Intell. Mag.*, vol. 15, no. 1, pp. 64–75, Feb. 2020.
[15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Vol. 1, 2018, pp. 2227–2237.
[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, vol. 1, pp. 4171–4186, 2019.
[17] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[18] T. Niven and H.-Y. Kao, "Probing neural network comprehension of natural language arguments," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4658–4664.
[19] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3428–3448.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*. [Online]. Available: http://arxiv.org/abs/1907.11692
[21] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 64–77, Jul. 2020.

[22] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4487–4496.
[23] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. Cham, Switzerland: Springer, 2019, pp. 194–206.
[24] D. Blei, A. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[27] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[28] C. J. van Rijsbergen, F. Crestani, and M. Lalmas, *Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information*. Norwell, MA, USA: Kluwer, 1998.
[29] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8778–8788.
[30] T. Mikolov, I. Sutskever, K. Chen, and G. S. Corrado, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[31] X. Chen, L. Xu, Z. Liu, M. Sun, and H. Luan, "Joint learning of character and word embeddings," in *Proc. 24th Int. Joint Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2015, pp. 1236–1242.
[32] C. Yixin, "Research on Weibo public opinion reversal based on Improved hegselmann-Krause model," *Inf. Theory Pract.*, vol. 43, no. 1, pp. 82–89, 2020.

XU TAN was born in 1981. He received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2009.

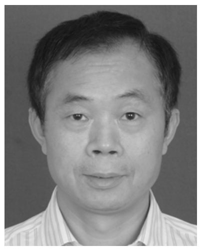
He is currently a Professor with the Shenzhen Institute of Information Technology, and a Pengcheng Distinguished Professor, Shenzhen, China. He has been committed to research on intelligent decision analysis and network public opinion analysis. In recent years, he has published more than 60 academic articles, among which 35 were published in the journals indexed by EI and SCI. He has obtained three Chinese national invention patents, six Chinese utility model patents, four Chinese software copyrights, and edited two textbooks. He has presided over 14 teaching and scientific research projects, including those of the Natural Science Foundation of China (NSFC) and the Ministry of Education in China (MOE) Liberal Arts and Social Sciences Foundation.

MUNI ZHUANG was born in 1996. She is currently pursuing the Ph.D. degree with the Business School, Southern University of Science and Technology, Shenzhen, China. She has completed a few projects in machine learning and data analytics with the guidance of Prof. Tan. Her research interests include distributed public opinion analysis, data mining, and deep learning.



XIN LU was born in 1982. He received the Ph.D. degree from the Department of Public Health Sciences, Karolinska Institutet, in 2009.

He is currently a Professor with the College of Systems Engineering, National University of Defense Technology, Changsha, China. He is a Co-founder of Flowminder Foundation. His main research interests include big data mining, human behavior dynamics, complex network theory, statistical sampling, emergency management, public health, and other related subjects. He is involved in long-term research topics around big data mining, the core technology of complex network algorithms, and human behavior analysis.



TAITIAN MAO was born in 1971. He received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2008.

He is currently a Professor and a Ph.D. Supervisor with the School of Public Administration, Xiangtan University, Xiangtan, China. His research interests include information analysis and evaluation, government public opinion analysis, and information resource management.