World Wide Web (2023) 26:55–70
https://doi.org/10.1007/s11280-022-01034-1

Identifying informative tweets during a pandemic via a topic-aware neural language model

Wang Gao[1] · Lin Li[2] · Xiaohui Tao[3] · Jing Zhou[1] · Jun Tao[1]

Received: 18 October 2021 / Revised: 20 January 2022 / Accepted: 28 February 2022 /
Published online: 16 March 2022

Abstract
Every epidemic affects the real lives of many people around the world and leads to terrible consequences. Recently, many tweets about the COVID-19 pandemic have been shared publicly on social media platforms. The analysis of these tweets is helpful for emergency response organizations to prioritize their tasks and make better decisions. However, most of these tweets are non-informative, which is a challenge for establishing an automated system to detect useful information in social media. Furthermore, existing methods ignore unlabeled data and topic background knowledge, which can provide additional semantic information. In this paper, we propose a novel Topic-Aware BERT (TABERT) model to solve the above challenges. TABERT first leverages a topic model to extract the latent topics of tweets. Secondly, a flexible framework is used to combine topic information with the output of BERT. Finally, we adopt adversarial training to achieve semi-supervised learning, and a large amount of unlabeled data can be used to improve inner representations of the model. Experimental results on the dataset of COVID-19 English tweets show that our model outperforms classic and state-of-the-art baselines.

Keywords Informative tweet identification · Social media · Topic model · Adversarial training

Wang Gao and Lin Li contributed equally to this work.

This article belongs to the Topical Collection: *Special Issue on Decision Making in Heterogeneous Network Data Scenarios and Applications*
Guest Editors: Jianxin Li, Chengfei Liu, Ziyu Guan, and Yinghui Wu

✉ Wang Gao
gaow@jhun.edu.cn

✉ Lin Li
cathylilin@whut.edu.cn

[1] School of Artificial Intelligence, Jianghan University, 430056 Wuhan, China

[2] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, 430070 Wuhan, China

[3] School of Sciences, University of Southern Queensland, 4072, Queensland, Toowoomba, Australia

1 Introduction

The popularity of social networks has generated a large amount of social interaction between users, which in turn produces massive unstructured textual data [15, 22]. In recent years, social media platforms such as Twitter and Facebook have received widespread attention as a possible tool for tracking a pandemic [1, 24, 26]. The main reason is that these platforms can provide real-time monitoring at a lower cost than conventional monitoring systems. For example, by the end of September 2021, the COVID-19 pandemic has caused approximately 4.6 million deaths and 228.4 million infected cases worldwide[1]. Millions of people are using social media platforms to share information related to COVID-19 such as testing or travel history.

However, although there are massive COVID-19 related posts on Twitter, only a few of them can provide useful information for monitoring systems. Manual detection of informative tweets is costly and ineffective for large amounts of data. Therefore, there is an urgent need to develop automated systems that can identify informative tweets. These tweets contain geographic location and information about confirmed, suspected and death cases. Many studies regard this problem as a binary classification task that classifies a related tweet as informative or non-informative [9, 25, 35].

Recently, the pre-trained language model Bidirectional Encoder Representations from Transformers (BERT) has achieved impressive performance improvements in various Natural Language Processing (NLP) tasks such as text classification and event detection [7]. The BERT model utilizes a large amount of textual data to pre-train encoders, and then makes effective fine-tuning for a certain target task. Although great progress has been made, identifying informative tweets during a pandemic remains a challenging issue. This is due to the short length and high noise of tweets, and high levels of text overlap between the two categories. Probabilistic topic models may provide additional topic information for semantic differences, and the latest research on neural networks has shown that topic integration can improve the performance of NLP tasks such as summarization and question answering [8, 11, 38]. However, there is no standard method to integrate topic information with pre-trained language models such as BERT. Furthermore, when there are huge amounts of labeled data in a classification task, BERT can achieve state-of-the-art results. Unfortunately, obtaining annotated instances is time-consuming and requires expensive human labor.

To address the above challenges, we design a novel model based on BERT to detect informative tweets during a pandemic. The main idea of the proposed model is derived from the answers to the following questions: (1) How to combine topic knowledge and BERT to learn distinguishable representations of short texts for informative tweet detection? (2) When there is little labeled data, how to extend the BERT model to improve its generalization ability?

Specifically, we propose a Topic-Aware BERT (TABERT) model that combines topic modeling with BERT using a simple architecture. TABERT leverages a three-stage framework to solve both topic integration and generalization capability issues in informative tweet detection. In the first stage, TABERT utilizes a Conditional Random Field regularized Topic Model (CRFTM) [10] to extract the topic information of tweets. CRFTM first merges short texts into long pseudo-documents using an embedding-based distance metric.

[1] https://coronavirus.jhu.edu/map.html

Semantic correlations are then integrated into the topic model to increase the probability that semantically related words belong to the same topic. Secondly, TABERT concatenates the topic distribution extracted by CRFTM with the last layer of BERT as the representation of a tweet. In the third stage, the proposed model extends the fine-tuning process of BERT from the perspective of the Generative Adversarial Network (GAN) [14], which conducts adversarial training in a zero-sum game. TABERT is used as a discriminator to classify tweets as informative or non-informative, and a generator generates "false" tweets similar to the distribution of real data. In this way, we can employ unlabeled tweets to improve the performance and generalization ability of the proposed model. To evaluate the performance of TABERT, we conduct extensive experiments on a COVID-19 related dataset. Experimental results demonstrate that the performance of the proposed model is significantly better than state-of-the-art baselines. This paper makes three main contributions as follows:

1. We propose a new Topic-Aware BERT (TABERT) model to identify informative tweets during a pandemic. TABERT adopts CRFTM to discover hidden topics of tweets. We combine topic information with BERT, which helps to enrich the semantics of short texts. To the best of our knowledge, this is the first work to integrate topic information captured by CRFTM into BERT.
2. TABERT exploits adversarial training to achieve semi-supervised learning in a GAN structure. The proposed model trains a generator to produce new tweets, and a discriminator is used to classify tweets as generated or real. Therefore, we train the discriminator with labeled tweets, while unlabeled tweets improve the output representation of the model.
3. Experimental results on a COVID-19 related dataset show that the proposed model achieves improvements against state-of-the-art baseline methods. Furthermore, TABERT can still achieve comparable performance in identifying informative tweets, even though the number of annotated tweets is drastically reduced.

The rest of the paper is arranged as follows. Section 2 reviews the work related to TABERT. Section 3 describes the details of the proposed model. Section 4 contains experiments with evaluations and comparisons. Finally, we conclude the paper in Section 5.

2 Related work

There have been many research reports on how to effectively use disaster-related tweets for situational awareness during emergencies and disasters [5, 6]. Discovering informative content from social media platforms is an important task for government agencies and rescue organizations [2, 41]. In this section, we give a brief overview of the work related to TABERT.

Deep neural networks have made impressive progress in various artificial intelligence tasks over recent years [32, 36, 40]. These techniques are widely used in NLP to extract textual features [12, 23, 39]. Neppalli et al. proposed various neural network models based on Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to identify informative tweets [28]. They found that in different natural disasters, deep neural networks have better generalization ability than traditional machine learning techniques. Kumar et al. proposed a multi-modal neural network that uses multi-modal features to

detect disaster-related information from social media [19]. They employed Long Short-Term Memory (LSTM) and VGG-16 to extract textual and visual features respectively, which proved to be better than using text or images alone. Gao et al. also proposed a multi-modal neural network that captures the transferable features shared between different disaster events [13]. The model utilizes adversarial training to evaluate the similarity between different events and improve its performance in emerging disaster events.

Furthermore, Roy et al. proposed a summarization and classification method to detect informative social media data during a disaster [30]. In their model, a Support Vector Machine (SVM) is trained to capture Parts of Speech (POS) tags and other features to identify informative short texts. After that, they used an abstractive summarization algorithm to obtain a real-time informative summary from these short texts. Zahera et al. added several stacked layers on top of BERT, and then applied the model to the multi-label classification of short texts [42]. They first preprocessed short texts in social media and then input them into BERT, leading to better results.

Due to the increasing amount of social media data associated with COVID-19, there have been many studies to explore the intent and content of these data. Singh et al. analyzed COVID-19 related social media data based on location, content and error message propagation [34]. The results show that despite the existence of a large amount of noise information, there is a significant spatio-temporal relationship between social media data and new cases of COVID-19. Shahi et al. presented a comprehensive study of social media analysis techniques applicable to the COVID-19 epidemic [33]. They analyzed the differences between error messages and other COVID-19 related tweets, and how they spread. However, the above methods ignore topic information or unlabeled data, which can improve the performance of informative tweet identification.

Topic knowledge is able to provide additional information for short text classification [17]. Zeng et al. proposed a topic memory network that encodes topic representations and classifies documents by memory networks [43]. Chaudhary et al. proposed a text classification model that combines topic modeling and a neural language model [3]. Their approach consists of two components: Neural Variational Document Model (NVDM) and BERT for complementary and efficient document understanding. However, due to the short length of each tweet and lack of context, NVDM suffers from a severe feature sparsity problem. In contrast, our model utilizes CRFTM to discover underlying topic knowledge in tweets, which has been shown to reveal more coherent topics from short texts. Furthermore, in Computer Vision (CV), GAN has been shown to be effective for semi-supervised learning. For instance, Salimans et al. proposed a semi-supervised learning method based on GAN, which achieves competitive performance with less labeled data and a large number of unlabeled data [31]. To the best of our knowledge, our model is the first to combine topic information and GAN into BERT to identify informative tweets.

3 Methodology

TABERT divides the process of identifying informative tweets during a pandemic into three stages. In the first stage, we train the CRFTM model on all tweets, and extract the document-topic probability distribution of each tweet. In the second stage, we develop an efficient approach to combine topic information with BERT. Finally, TABERT utilizes unlabeled tweets in an adversarial training setting to extend the training process.

3.1 Topic extraction

As a common information extraction method, topic models have been widely applied in sentiment analysis, event detection and other NLP tasks. Traditional topic models infer topics based on word co-occurrence patterns in documents. However, the features of short texts are sparse, and it is difficult to provide sufficient co-occurrence information for topic modeling. To alleviate the sparsity problem, we utilize CRFTM to extract topic information of tweets. CRFTM first merges tweets into regular-sized pseudo-documents, and then combines word embeddings with word correlation knowledge to enhance the coherence of extracted topics.

Specifically, Embedding-based Minimum Average Distance (EMAD) is used to measure the distance between tweets [10]. EMAD is able to find semantically related words in two different tweets, which may be assigned to the same topic label. Based on a clustering algorithm, CRFTM then aggregates tweets into long pseudo-documents. Next, CRFTM draws a topic distribution $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$ for the whole corpus. For each topic k, the topic model draws a word distribution $\phi_k \sim Dir(\boldsymbol{\beta})$. α and β are Dirichlet priors. For each pseudo-document d, the topic model draws each word $w_{di} \sim Mult(\phi_{z_{di}})$ and a topic assignment $\mathbf{z}_d$ can be written as follows:

$$p(\mathbf{z}_d \mid \boldsymbol{\theta}_d, \mathbf{x}_d) = \prod_{i=1}^{N_d} p(z_{di} \mid \boldsymbol{\theta}_d) \boldsymbol{\Psi}(z_{di}, x_{di}), \tag{1}$$

where x_{di} represents the contextual words of the i_{th} word, $\mathbf{x}_d$ denotes the set of contextual words for each word, $\boldsymbol{\Psi}$ is a potential function that considers the impact of semantic relevance, and N_d denotes the number of words in d.

In CRFTM, collapsed Gibbs sampling can be adopted to do posterior inference. Therefore, the topic z_{di} of word w_{di} in pseudo-document d is derived as:

$$p(z_{di} = k \mid \mathbf{z}_{d,-di}, \mathbf{w}) \propto (n_{d,-di}^{(w_{di})} + \alpha) \frac{n_{k,-di}^{(w_{di})} + \beta}{n_{k,-di} + V\beta} \boldsymbol{\Psi}(z_{di} = k, x_{di}), \tag{2}$$

where $n_{\cdot,-di}^{(\cdot)}$ represents the number of times the word is assigned to topic k, when word w_{di} is excluded from topic k or pseudo-document d. V denotes the dimension of the vocabulary.

Accordingly, the document-topic distribution $\boldsymbol{\theta}$ and the topic-word distribution $\boldsymbol{\phi}$ can be calculated as follows:

$$\theta_{d,k} = \frac{n_d^{(k)} + \alpha}{\sum_{k=1}^{K} n_d^{(k)} + K\alpha} \tag{3}$$

$$\phi_{k,w} = \frac{n_k^{(w)} + \beta}{\sum_{w=1}^{V} n_k^{(w)} + V\beta}, \tag{4}$$

where K is the number of topics. In the proposed model, two distributions representing the topic information of tweets are integrated into BERT in different ways.

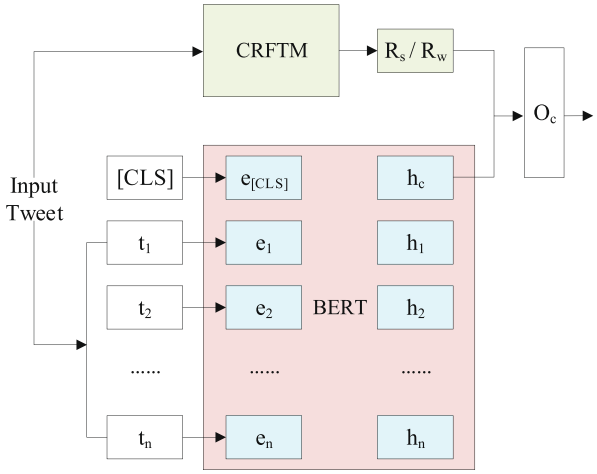

Fig. 1 Architecture of TABERT combining topic information and BERT

3.2 Topic information fusion

In this section, we study how to fuse topic information and BERT to improve the performance of informative tweet detection. During pre-training, since the length of sentences input to BERT is limited, it is more suitable for extracting the semantics of short texts. The architecture of TABERT that combines topic information and BERT is shown in Figure 1. Let $T = \{t_1, t_2, ..., t_j, ..., t_n\}$ be the input tweet, where n denotes the length of the tweet and t_j represents the j-th token of the tweet. The input of BERT is a concatenation of token embeddings, segment embeddings and positional embeddings. Token embeddings are used to convert tokens into vector representations, while segment embeddings separate different tweets. BERT is based on the transformer structure that cannot encode the ordering information of input tweets [37]. Therefore, the BERT model takes advantage of the sequence of tweets by adding positional embeddings to the input representation.

We add these three embeddings element-wise to form a single vector $E = \{e_1, e_2, ..., e_j, ..., e_n\}$, and then employ it as the input of an encoding layer. Subsequently, BERT maps E into a sequence of hidden representations $H = \{h_1, h_2, ..., h_j, ..., h_n\}$ by applying self-attention and multi-head attention mechanisms. An additional token [*CLS*] is appended to each input tweet as the first token, and its hidden state h_c is the output representation of BERT:

$$h_c = BERT(T) \in \mathbb{R}^{d_{BERT}}, \tag{5}$$

where d_{BERT} is the hidden dimension size of BERT. Following [21, 27], we combine BERT with topic information at word and sentence levels respectively. For sentence-level topic information R_s, all words in a tweet are input to CRFTM to infer a document-topic distribution $p(z \mid d)$ for each tweet:

$$R_s = CRFTM([t_1, t_2, ..., t_j, ..., t_n]) \in \mathbb{R}^K. \tag{6}$$

For word-level topic information R_w, TABERT leverages Summation over Words (SW) representations [21] to infer $p(z \mid d)$, which has proved to be an ideal approach for tweets:

$$R_w = \sum_w p(z = k \mid w)p(w \mid d) \in \mathbb{R}^K, \tag{7}$$

where $p(w \mid d)$ is the number of times w occurs in d. Since topic features and textual features are structurally consistent, and semantically continuous and related, the proposed model directly combines sentence-level topic information with the output of BERT as:

$$O_c = h_c \oplus R_s, \tag{8}$$

where $\oplus$ denotes the concatenation operator. The way of combining word-level topic information is as follows:

$$O_c = h_c \oplus R_w. \tag{9}$$

3.3 Adversarial training

By training with additional unlabeled data in an adversarial training setting, Semi-supervised GAN (SGAN) proposed by [31] can significantly improve the effectiveness of a supervised task. In the SGAN model, a discriminator module divides the data into $(c + 1)$ classes. Real data is classified as one of the target $(1, ..., c)$ categories, while the data generated by a generator is classified as a new "generated" class $(c + 1)$.

Specifically, $\mathcal{G}$ and $\mathcal{D}$ represent the generator module and discriminator module respectively, while $p_{\mathcal{G}}$ and $p_{\mathcal{D}}$ are the generator distribution and real example distribution. We then define $p_m(y = c + 1 \mid x)$ to provide the probability that a data point x belongs to a "generated" class. $p_m(y \in (1, ..., c) \mid x)$ denotes the probability that x is a real data, which is associated with an original category. To train a semi-supervised c-class classifier, the objective function $\mathcal{L}_{\mathcal{D}}$ of $\mathcal{D}$ can be defined as:

$$\mathcal{L}_{\mathcal{D}} = \mathcal{L}_{supervised} + \mathcal{L}_{unsupervised}. \tag{10}$$

The total cross-entropy loss can be decomposed into a supervised loss $\mathcal{L}_{supervised}$ and an unsupervised loss function $\mathcal{L}_{unsupervised}$:

$$\begin{aligned} \mathcal{L}_{supervised} &= -\mathbb{E}_{x,y\sim p_{\mathcal{D}}(x,y)} log[p_m(y \in (1, ..., c) \mid x)] \\ \mathcal{L}_{unsupervised} &= -\mathbb{E}_{x\sim p_{\mathcal{D}}(x)} log[1 - p_m(y = c + 1 \mid x)] \\ &\quad -\mathbb{E}_{x\sim p_{\mathcal{G}}(x)} log[p_m(y = c + 1 \mid x)], \end{aligned} \tag{11}$$

where $\mathcal{L}_{supervised}$ measures the cumulative loss of classifying real data into a wrong category among the target c classes. $\mathcal{L}_{unsupervised}$ measures the cumulative loss of identifying real unlabeled data as the generated class $(c + 1)$ and identifying generated data as real.

Meanwhile, the generator module $\mathcal{G}$ needs to generate data close to the data sampled from the real example distribution $p_{\mathcal{D}}$. Therefore, $\mathcal{G}$ should generate examples that match the statistics of real examples as much as possible. The goal of training the generator is to learn the expected values of features on the middle layer of the discriminator. By training $\mathcal{D}$, SGAN captures the features that can best distinguish the real data from the data generated by $\mathcal{G}$. The model defines the feature matching objective function of the generator as:

$$\mathcal{L}_{fm} = || \mathbb{E}_x \sim p_{\mathcal{D}}(x)f(x) - \mathbb{E}_x \sim p_{\mathcal{G}}(x)f(x) ||_2^2, \tag{12}$$

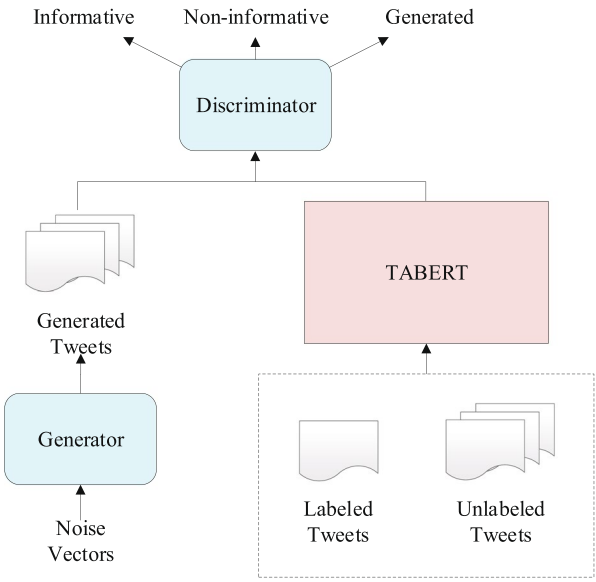

Fig. 2 Architecture of TABERT with adversarial training

where $f(x)$ denotes an activation function. When the examples generated by $\mathcal{G}$ are input to $\mathcal{D}$, their feature representations are very similar to that of real data. SGAN also needs to consider the error $\mathcal{L}_{generated}$ that the discriminator classifies the generated data as real:

$$\mathcal{L}_{generated} = \mathbb{E}_x \sim p_{\mathcal{G}}(x) log[1 - p_m(y = c + 1 \mid x)] \tag{13}$$

The objective function of $\mathcal{G}$ is $\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{fm} + \mathcal{L}_{generated}$. Although SGAN is usually applied in CV, we extend TABERT by using it to improve the performance of informative tweet detection.

In this paper, SGAN and TABERT are combined in the fine-tuning phase. The proposed method adjusts the fine-tuning process of the TABERT model by adding an SGAN layer containing a discriminator and a generator. We employ the discriminator for classification and the generator for semi-supervised learning. Figure 2 illustrates the architecture of the TABERT model with adversarial training.

As shown in the figure, we add the SGAN framework to the top of TABERT by integrating a discriminator $\mathcal{D}$ and a generator $\mathcal{G}$. $\mathcal{D}$ classifies the input tweets as informative, non-informative or generated, while $\mathcal{G}$ produces generated data for adversarial training. More formally, $\mathcal{D}$ is a multi-layer perceptron whose input is vector $I_{\mathcal{D}}$. $I_{\mathcal{D}}$ can be either O_c that is the output of TABERT for real unlabeled or labeled tweets, or $O_{\mathcal{G}}$ generated by $\mathcal{G}$. The generator is also composed of a multi-layer perceptron, which receives noise vectors sampled from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and outputs a generated vector $O_{\mathcal{G}}$. Similar to SGAN, a softmax layer is the last layer of $\mathcal{D}$ to classify tweets.

In the training process, when the input is a real tweet (*i.e.*, $I_{\mathcal{D}} = O_c$), the discriminator should identify whether it contains useful information. When generated data is used as input (*i.e.*, $I_{\mathcal{D}} = O_{\mathcal{G}}$), $\mathcal{D}$ should identify whether it is a real tweet. Two competitive losses, $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{G}}$, can be optimized during the adversarial training process.

Unlabeled tweets only contribute to the unsupervised loss of $\mathcal{D}$ (*i.e.*, $\mathcal{L}_{unsupervised}$) during back-propagation. In other words, only if they are incorrectly identified as generated

tweets, they are considered in the loss calculation. Furthermore, their contribution to the loss is not considered in other cases. Correspondingly, annotated tweets contribute to the supervised loss of the discriminator (*i.e.*, $\mathcal{L}_{supervised}$). The generated data produced by the generator has an impact on both $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{G}}$. If $\mathcal{D}$ recognizes tweets generated by $\mathcal{G}$, then $\mathcal{G}$ will be penalized, and vice versa. When training the discriminator, we also update TABERT to fine-tune its network weights, which requires consideration of both labeled and unlabeled tweets.

The generator can be discarded after the model training is completed, while the rest of the model can be used for inference. As a result, in actual use, there is no additional time consumption compared to the TABERT model.

4 Experiments

In this section, we conduct extensive experiments to validate the effectiveness of TABERT. The performance is reported over a real-world tweet dataset, *i.e.*, a COVID-19 related corpus. The experimental results illustrate the effectiveness of the proposed model for informative tweet detection.

4.1 Dataset

The dataset used in the experiment is provided by [29], which contains informative English COVID-19 tweets. In the dataset, informative tweets contain various information about COVID-19 cases, as well as their location and travel history. To collect unlabeled data, we first employ the Twitter API to crawl English tweets containing keywords such as "covid19", "covid2019", or "coronavirus". These tweets usually contain lots of noisy text. Hence, we perform the following data pre-processing techniques on tweets, which help TABERT achieve better performance: (1) convert all letters to lowercase; (2) utilize the emoji library[2] to replace emojis with short text descriptions; (3) replace all hyperlinks in the corpus with "URL"; (4) remove tweets with less than five words; (5) remove all non-alphabetic characters as well as unnecessary newlines, spaces and tabs.

To balance the dataset, we ask three annotators to label a portion of the collected tweets. The annotators first divide tweets into two categories: "Informative" and "Non-Informative". A post is considered "Informative" if it contains information useful to emergency response organizations (*e.g.*, location of suspected cases). If there is a disagreement between the annotators, the post is removed. Finally, we select 15,935 labeled tweets, consisting of 7,983 informative tweets and 7,952 non-informative tweets as well as 158,341 unlabeled tweets.

4.2 Baseline methods

In the experiment, the following classic and state-of-the-art baseline methods are compared by precision, recall and F1-score:

[2] https://pypi.org/project/emoji/

- **CNN:** CNN uses convolution filters to learn local features of documents. CNN not only shows encouraging results in CV, but also has been widely used in various NLP tasks.
- **BiLSTM:** LSTM shows excellent performance in classification problems like short text classification by extracting contextual features. BiLSTM further improves performance by comprehensively considering the bi-directional context information of words.
- **BERT:** The architecture of BERT is based on a multi-layer bi-directional transformer model [7]. The model is pre-trained on a large-scale corpus such as Wikipedia, and replaces RNN with a self-attention mechanism.
- **ALBERT:** This is a variant of BERT that utilizes parameter reduction methods to reduce memory consumption and speed up BERT training [20]. To further boost the performance, a self-supervised loss is introduced.
- **TABERT-G:** TABERT-G is a variant of TABERT, but it does not use GAN to train the model. Specifically, TABERT-G directly adds a softmax layer on the top of the TABERT model with word-level topic information to identify informative tweets.

For the topic model, we run 1,000 Gibbs sampling iterations and set the Dirichlet priors $\alpha = 50/K$, $\beta = 0.01$. The number of topics is 100, and other parameters are set according to the original paper. For adversarial training, $\mathcal{D}$ is a multi-layer perceptron with a hidden layer, and the top softmax layer is used for classification. The generator is also implemented by a multi-layer perceptron with a hidden layer. A noise representation sampled from a Gaussian distribution $\mathcal{N}(0, 1)$ is used as the input of $\mathcal{G}$. These noise representations are converted by the multi-layer perceptron into vectors with the same size as the output of TABERT, which are used as generated data in adversarial training. For CNN and BiLSTM, the word embedding we use is freely-available Glove[3], and the dimension size is equal to 300. We exploit these word embeddings to build a vector matrix that converts the words of input tweets into corresponding vector representations. Binary cross-entropy loss is used to train these models, and the dropout rate is set to 0.2.

According to the ratio of 7:2:1, the COVID-19 dataset is randomly divided into a training set, a validation set and a testing set. When the loss of the validation set for three consecutive epochs does not decrease, the training process of the model will stop. For BERT, we employ a pre-trained 12-layer BERT-base[4] architecture with 12 self-attention heads and the hidden size of 768. The adam optimizer with a learning rate of 2e-5 is utilized to train the model. For the ALBERT model, we use albert-base-v2[5] with 12 repeating layers and 11M parameters.

4.3 Classification results

Figure 3 illustrates the experimental results of our models and five baseline methods. It can be intuitively seen from the classification results that TABERT with word-level topic information achieves the best performance. The F1-score of the proposed model is 4.1% higher than BERT, and the precision and recall are 2.3% and 5.8% higher, respectively. This experimental result validates that our model utilizes topic information and unlabeled

[3] http://nlp.stanford.edu/projects/glove/
[4] https://github.com/google-research/bert
[5] https://github.com/google-research/ALBERT

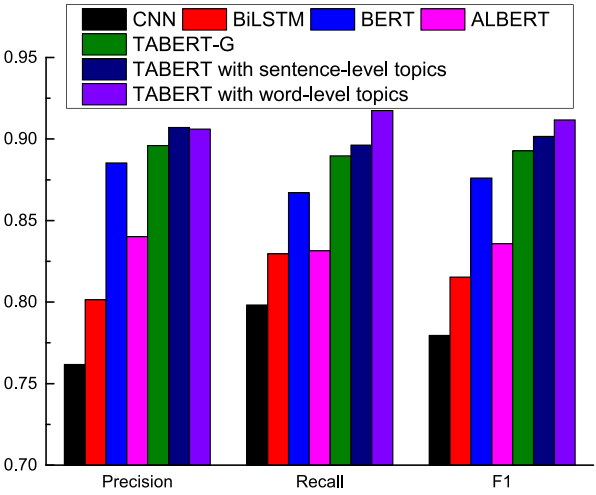

Fig. 3 Experimental results of seven methods

data as an additional source of dataset-specific features, which is beneficial for achieving better classification performance.

As shown in the figure, the combination of TABERT and word-level topic information achieves better results than sentence-level topic information. As discussed in [21], regardless of topic models or the number of topics, using the SW approach is more effective than other approaches for the topic representations of short texts. The reason may be that the direct use of a document-topic distribution to infer $p(z \mid d)$, which results in extremely sparse sentence-level topic information. Therefore, this is not an ideal approach for short texts such as tweets. In the rest of the experiment, we only use TABERT with word-level topic information to classify tweets.

Additionally, the performance of BiLSTM is better than CNN because Bi-LSTM is capable of learning long-term dependencies without retaining repetitive contextual information. The experimental results of classical methods such as CNN and BiLSTM are worse than transformer-based models such as BERT and ALBERT. This may be because the transformer structure depends on an attention mechanism to encode the interdependence between input and output for better parallelization. Another reason is that BERT can learn high-quality vector representations of tweets by pre-training on a large-scale unlabeled corpus. The performance of ALBERT is slightly worse than BERT may be due to its fewer parameters.

TABERT-G adds CRFTM topics to BERT, and experimental results show that this mechanism leads to its performance better than other neural systems. The reason is that TABERT introduces topic information to extend the features of tweets, which can alleviate data sparsity issues. By using unlabeled data, we show that TABERT-G is inferior to TABERT in this tweet detection task. The result indicates that adversarial training not only enhances the performance of the model, but also helps the model to be generalized better, which can also be observed in [13].

4.4 Influence of topic models

To investigate whether adding topic information improves the performance of BERT, the effects of four different short text topic models on informative tweet identification

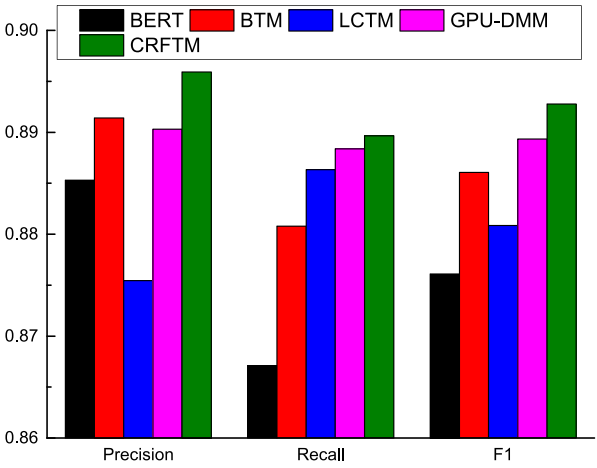

Fig. 4 Experimental results of different topic models

are compared. Biterm Topic Model (BTM) directly models word co-occurrence information from short text datasets, and then extracts hidden topics [4]. In the BTM model, unordered word pairs that co-occur in a sliding window are called biterms. Latent Concept Topic Model (LCTM) treats each latent concept as a local Gaussian distribution in the word embedding space, and each topic is a probability distribution over latent concepts [16]. Generalized Pólya Urn Dirichlet Multinomial Mixture (GPU-DMM) utilizes the GPU model to promote semantically correlated words belonging to the same topic during training [21].

Figure 4 illustrates the classification results of BERT, TABERT-G with BTM, TABERT-G with LCTM, TABERT-G with GPU-DMM and TABERT-G with CRFTM. From the figure, we can find that adding any kind of topic information to BERT improves the performance of the model. The CRFTM model outperforms other models, which validates that the combination of word semantic relations and CRF is beneficial for extracting discriminative topic information. GPU-DMM achieves the second best performance in terms of F1-score. Furthermore, the performance of BTM is worse than GPU-DMM. The result suggests that biterms may only bring little additional word co-occurrence patterns for short text topic modeling. LCTM achieves the worst performance among all models. This may be because tweets in the COVID-19 dataset are much shorter, and adding a latent concept layer would cause more serious sparsity problems, resulting in worse topic inference results.

4.5 Analysis of semi-supervised learning

To assess the impact of semi-supervised learning on TABERT, different scale labeled tweets are employed to identify informative tweets. We also report the performance of the BERT model on the same scale of training corpus as a comparison. We first randomly sample 1% (150 instances) of the entire labeled dataset. Secondly, we repeated the training of TABERT and BERT with gradually increasing labeled data. The proposed model also provides 50 unlabeled examples for each labeled tweet to enable semi-supervised learning from a GAN perspective. We randomly sample the COVID-19 corpus five times and report the average performance.

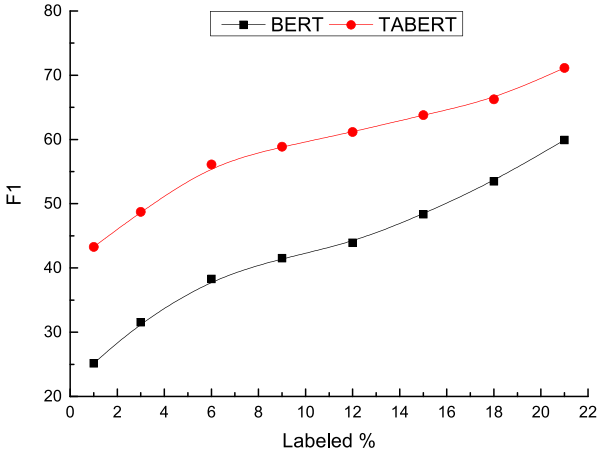

Fig. 5 F1 scores for different ratios of annotated data

Figure 5 shows F1 scores of TABERT and BERT, with the ratio of annotated data used ranging from 1 to 21. When 1% of data is available, the performance of BERT is poor, and the proposed model achieves a score of more than 40%. This trend continues until 21% of labeled tweets are used. Using more annotated tweets would result in closer F1 scores, but TABERT is always better than BERT. Experimental results demonstrate that the proposed model can improve the robustness of transformer-based architectures without incurring additional costs. In the inference process, our model only needs the discriminator module, while the generator module is only utilized in the training phase.

4.6 Actionable information mining

To help emergency response organizations make better decisions and formulate corresponding strategies, after filtering out non-informative tweets, we can further classify the informative content. This helps send a tweet with actionable information to a specific response agency to better understand the situation related to the epidemic, and deploy targeted epidemic prevention and control work.

Following [18], actionable information can be defined as information that can alert emergency response organizations of a certain type (*e.g.*, injured people or infrastructure damage). To avoid affecting the judgment of responding organizations, we divide the data according to the following sufficient granularity:

- **Caution and advice:** These tweets provide the public with some advice on the epidemic (*e.g.*, "You can be a hero to children and elderly by simply staying at home").
- **Emerging threats:** These tweets report information that may lead to the spread of the epidemic (*e.g.*, "A goat at a zoo tested positive for COVID-19").
- **Render services:** These tweets indicate that some people are providing services (*e.g.*, "A hotel provides free rooms for medical staff fighting against COVID-19").
- **Volunteer:** These tweets ask people to volunteer in response to special events (*e.g.*, "Hospital staff need you to help them make food").
- **New event:** These tweets report new incidents that relevant agencies need to respond to in a timely manner (*e.g.*, "The subway station should be closed tonight because a staff member tested positive for covid-19").

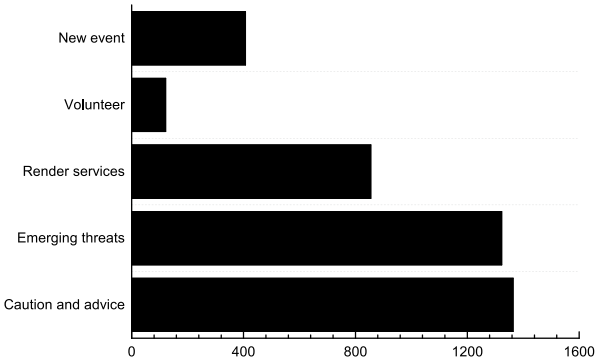

Fig. 6 Number of tweets for five actionable information types

Table 1 Performance of TABERT and BERT in actionable information mining

Model	Precision	Recall	F1
BERT	0.6114	0.6037	0.5976
TABERT	0.6530	0.6306	0.6373

According to the above actionable information categories, Figure 6 shows the number of informative tweets for each category (posts that do not belong to any category are discarded). Table 1 reports the performance of TABERT and BERT in actionable information mining. As seen from the table, in the face of serious class imbalance, the proposed model is better than the BERT model under all metrics. As a result, TABERT is a model built on data collected during past epidemics, and can detect and track new events to strengthen the decision-making process of government agencies.

5 Conclusion

In this paper, we propose a new Topic-Aware BERT (TABERT) model to detect informative posts on social media platforms such as Twitter. In the proposed model, CRFTM is first used to discover the topic knowledge of each tweet. Next, we design a simple architecture to combine topic information with BERT. TABERT finally extends the training process with unlabeled tweets in a GAN framework. Experimental results show that our model is not only better than baseline methods, but also reduces the requirement for annotated data. Since TABERT does not exploit the domain-specific features of the dataset, the model can be generalized for identifying informative tweets in different domains. In the future, we will study how to introduce topic knowledge into BERT without corrupting pre-trained contextual information, and evaluate the model on large-scale datasets. Moreover, we will also explore how to directly apply adversarial training in the pre-training phase to further improve performance.

Acknowledgements This work is supported by National Science Foundation of China (NSFC, No. 62106086), Scientific Research Program of Hubei Provincial Department of Education (No. B2021063) and Research Projects of Jianghan University (No. 2021yb062).

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Al-garadi, M.A., Khan, M.S., Varathan, K.D., Mujtaba, G., Al-Kabsi, A.M.: Using online social networks to track a pandemic: A systematic review. Journal of Biomedical Informatics **62**, 1–11 (2016)
2. Cai, T., Li, J., Mian, A.S., Sellis, T., Yu, J.X., et al.: Target-aware holistic influence maximization in spatial social networks. IEEE Transactions on Knowledge and Data Engineering **34**(4), 1993–2007 (2022)
3. Chaudhary, Y., Gupta, P., Saxena, K., Kulkarni, V., Runkler, T.A., Schütze, H.: Topicbert for energy efficient document classification. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pp. 1682–1690 (2020)
4. Cheng, X., Yan, X., Lan, Y., Guo, J.: Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering **26**(12), 2928–2941 (2014)
5. Chowdhury, J.R., Caragea, C., Caragea, D.: Cross-lingual disaster-related multi-label tweet classification with manifold mixup. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 292–298 (2020)
6. Chowdhury, J.R., Caragea, C., Caragea, D.: On identifying hashtags in disaster twitter data. In: Proceedings of Conference on Artificial Intelligence (AAAI), pp. 498–506 (2020)
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186 (2019)
8. Feng, J., Rao, Y., Xie, H., Wang, F.L., Li, Q.: User group based emotion detection and topic discovery over short text. World Wide Web **23**(3), 1553–1587 (2020)
9. Gao, W., Fang, Y., Li, L., Tao, X.: Event detection in social media via graph neural network. In: Web Information Systems Engineering (WISE), pp. 370–384 (2021)
10. Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., Tian, G.: Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems **61**, 1123–1145 (2019)
11. Gao, W., Peng, M., Wang, H., Zhang, Y., Han, W., Hu, G., Xie, Q.: Generation of topic evolution graphs from short text streams. Neurocomputing **383**, 282–294 (2020)
12. Gao, W., Fang, Y., Zhang, F., Yang, Z.: Representation learning of knowledge graphs using convolutional neural networks. Neural Network World **30**, 145–160 (2020)
13. Gao, W., Li, L., Zhu, X., Wang, Y.: Detecting disaster-related tweets via multimodal adversarial neural network. IEEE MultiMedia **27**(4), 28–37 (2020)
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680 (2014)
15. Haldar, N.A.H., Reynolds, M., Shao, Q., Paris, C., Li, J., Chen, Y.: Activity location inference of users based on social relationship. World Wide Web **24**(4), 1165–1183 (2021)
16. Hu, W., Tsujii, J.: A latent concept topic model for robust topic inference using word embeddings. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 380–386 (2016)
17. Huang, J., Peng, M., Li, P., Hu, Z., Xu, C.: Improving biterm topic model with word embeddings. World Wide Web **23**(6), 3099–3124 (2020)
18. Imran, M., Mitra, P., Castillo, C.: Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In: Proceedings of International Conference on Language Resources and Evaluation (LREC), pp. 1–6 (2016)
19. Kumar, A., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: A deep multi-modal neural network for informative twitter content classification during emergencies. Annals of Operations Research **7**, 1–32 (2020)
20. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: Proceedings of International Conference on Learning Representations (ICLR), pp. 1–17 (2020)
21. Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Transactions on Information Systems **36**(2), 1–30 (2017)

22. Li, J., Cai, T., Deng, K., Wang, X., Sellis, T., Xia, F.: Community-diversified influence maximization in social networks. Information Systems **92**, 1–12 (2020)
23. Li, Z., Wang, X., Li, J., Zhang, Q.: Deep attributed network representation learning of complex coupling and interaction. Knowledge-Based Systems **212**, 1–15 (2021)
24. Long, Z., Alharthi, R., El Saddik, A.: Needfull-a tweet analysis platform to study human needs during the covid-19 pandemic in new york state. IEEE Access **8**, 136046–136055 (2020)
25. Mahata, D., Talburt, J.R., Singh, V.K.: From chirps to whistles: Discovering event-specific informative content from twitter. In: Proceedings of the ACM Web Science Conference (WebSci), pp. 1–10 (2015)
26. Mukherjee, S., Kumar, R., Bala, P.K.: Managing a natural disaster: actionable insights from microblog data. Journal of Decision Systems **31**, 134–149 (2022)
27. Narayan, S., Cohen, S.B., Lapata, M.: Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pp. 1797–1807 (2018)
28. Neppalli, V.K., Caragea, C., Caragea, D.: Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In: Proceedings of Information Systems for Crisis Response and Management (ISCRAM), pp. 1–10 (2018)
29. Nguyen, D.Q., Vu, T., Rahimi, A., Dao, M.H., Nguyen, L.T., Doan, L.: WNUT-2020 task 2: identification of informative COVID-19 english tweets. In: Proceedings of the Workshop on Noisy User-generated Text (WNUT), pp. 314–318 (2020)
30. Roy, S., Mishra, S., Matam, R.: Classification and summarization for informative tweets. In: Proceedings of IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), pp. 1–4 (2020)
31. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 2226–2234 (2016)
32. Sarki, R., Ahmed, K., Wang, H., Zhang, Y.: Automated detection of mild and multi-class diabetic eye diseases using deep learning. Health Information Science and Systems **8**(1), 1–9 (2020)
33. Shahi, G.K., Dirkson, A., Majchrzak, T.A.: An exploratory study of COVID-19 misinformation on twitter. Online Social Networks and Media **22**, 1–16 (2021)
34. Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E.K., Wang, Y.: A first look at COVID-19 information and misinformation sharing on twitter. 1–24 arxiv:2003.13907 (2020)
35. Sreenivasulu, M., Sridevi, M.: Classifying informative and non-informative tweets from the twitter by adapting image features during disaster. Multimedia Tools and Applications **79**(3), 28901–28923 (2020)
36. Supriya, S., Siuly, S., Wang, H., Zhang, Y.: Automated epilepsy detection techniques from electroencephalogram signals: a review study. Health Information Science and Systems **8**(1), 1–15 (2020)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 5998–6008 (2017)
38. Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., Du, Q.: A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 4453–4460 (2018)
39. Yang, Y., Guan, Z., Li, J., Zhao, W., Cui, J., Wang, Q.: Interpretable and efficient heterogeneous graph convolutional network. IEEE Transactions on Knowledge and Data Engineering, 1–14 (2021)
40. Yin, J., Tang, M., Cao, J., Wang, H., You, M., Lin, Y.: Vulnerability exploitation time prediction: an integrated framework for dynamic imbalanced learning. World Wide Web **25**, 401–423 (2022)
41. Yin, H., Yang, S., Song, X., Liu, W., Li, J.: Deep fusion of multimodal features for social media retweet time prediction. World Wide Web **24**(4), 1027–1044 (2021)
42. Zahera, H.M., Elgendy, I.A., Jalota, R., Sherif, M.A.: Fine-tuned BERT model for multi-label tweets classification. In: Proceedings of the Text Retrieval Conference (TREC), pp. 1–7 (2019)
43. Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M.R., King, I.: Topic memory networks for short text classification. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pp. 3120–3131 (2018)