Dual Consistency-Enhanced Semi-Supervised Sentiment Analysis Towards COVID-19 Tweets

Teng Sun, Liqiang Jing, Yinwei Wei, *Member, IEEE*, Xuemeng Song, *Senior Member, IEEE*, Zhiyong Cheng, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—In the context of COVID-19, numerous people present their opinions through social networks. It is thus highly desired to conduct sentiment analysis towards COVID-19 tweets to learn the public's attitudes, and facilitate the government to make proper guidelines for avoiding the social unrest. Although many efforts have studied the text-based sentiment classification from various domains (e.g., delivery and shopping reviews), it is hard to directly use these classifiers for the sentiment analysis towards COVID-19 tweets due to the domain gap. In fact, developing the sentiment classifier for COVID-19 tweets is mainly challenged by the limited annotated training dataset, as well as the diverse and informal expressions of user-generated posts. To address these challenges, we construct a large-scale COVID-19 dataset from Weibo and propose a dual COnsistency-enhanced semi-superVIseD network for Sentiment Anlaysis (COVID-SA). In particular, we first introduce a knowledge-based augmentation method to augment data and enhance the model's robustness. We then employ BERT as the text encoder backbone for both labeled data, unlabeled data, and augmented data. Moreover, we propose a dual consistency (i.e., label-oriented consistency and instance-oriented consistency) regularization to promote the model performance. Extensive experiments on our self-constructed dataset and three public datasets show the superiority of COVID-SA over state-of-the-art baselines on various applications.**

***Index Terms*—Semi-supervised text classification, sentiment analysis, social media dataset on COVID-19.**

I. Introduction

In MODERN societies, social networks, such as Twitter,[1] and Weibo[2] have been important venues for people to express their opinions via social posts. In the context of Novel Coronavirus Pneumonia (COVID-19), numerous users have been actively participating in the online discussion about this hot topic on social platforms, and there has been a large number of user-generated posts conveying their sentiments. In a sense, by analyzing the content of their generated posts on social networks, we are able to discover the users' attitudes towards COVID-19, which is essential for the government to perceive the potential crisis and hence take proactive measures. In light of this, in this work, we focus on sentiment analysis towards COVID-19 based upon the textual posts of users on social media, to help experts make proper guidelines for avoiding social unrest.

Although many researches have studied the text-based sentiment classification for various domains, including but not limited to hotel [1], restaurant [2], delivery [3], and shopping [4], relatively sparse efforts have been dedicated to the emerging COVID-19 domain. Moreover, we argue that directly using existing models of other domains One possible solution is to utilize the domain adaptation method [5], [6] to transfer the knowledge of existing models that are learned from other domains to our domain. However, these domain adaptation methods can only perform well when the domain gap is not large, such as transferring knowledge from the handwritten digits domain to the house numbers domain [6]. Unfortunately, the word distribution gap between COVID-19-related posts and other domain datasets is significant. To illustrate the difference, we collect a dataset of COVID-19 related posts from the largest Chinese social platform, i.e., Weibo, and compare the keyword distribution of the collected posts with other three existing datasets of other domains, namely ChnSentiCorp,[3] Waimai,[4] and OnlineShopping.[5] Fig. 1 shows the top-10 keywords in terms of TF-IDF of all the four datasets. As can be seen, the distribution of the top keywords across these datasets is distinct. Specifically, as shown in Figure 1(a), users in ChnSentiCorp mainly talk about the *Room* and *Service*; whereas users in Waimai frequently express their opinions on *Delivered* and *Flavour* (see Fig. 1(b)). By contrast, the top-10 keywords in our COVID-19 dataset, including *Epidemic*, *Mask*, and *Coronavirus*, are far different from those of the other three datasets. Inspired by this, we propose to learn an effective and specific sentiment classifier for the user-generated posts on COVID-19.

Manuscript received 3 January 2022; revised 27 January 2023; accepted 20 April 2023. Date of publication 27 April 2023; date of current version 8 November 2023. This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2022YQ59, in part by the National Natural Science Foundation of China under Grants U1936203 and 62236003, and in part by Shenzhen College Stability Support Plan under Grant GXWD20220817144428005. Recommended for acceptance by J. Tang. *(Corresponding authors: Xuemeng Song; Liqiang Nie.)*

Teng Sun, Liqiang Jing, and Xuemeng Song are with the Department of Computer Science and Technology, Shandong University, Qingdao, Shandong 250316, China (e-mail: stbestforever@gmail.com; jingliqiang6@gmail.com; sxmustc@gmail.com).

Yinwei Wei is with the School of Computing, National University of Singapore, Singapore 119077 (e-mail: weiyinwei@hotmail.com).

Zhiyong Cheng is with the Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong 250316, China (e-mail: jason.zy.cheng@gmail.com).

Liqiang Nie is with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: nieliqiang@gmail.com).

Digital Object Identifier 10.1109/TKDE.2023.3270940

[1]https://twitter.com/

[2]https://weibo.com/

[3]https://tinyurl.com/3y2sffn6/

[4]https://tinyurl.com/3yz8dppp/

[5]https://tinyurl.com/48p97mvj/

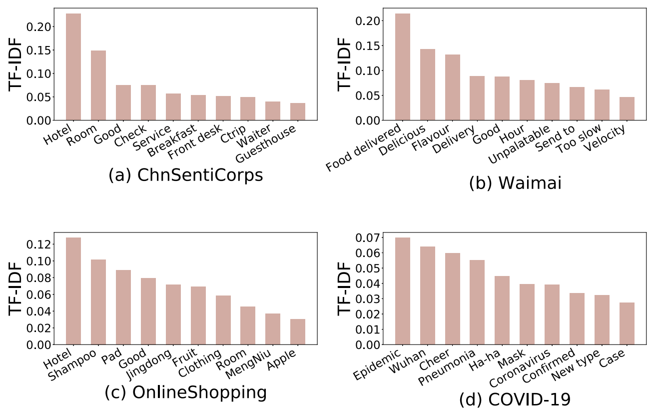


Fig. 1. Top-10 keywords *w.r.t.* TF-IDF on four datasets designed for different domains.

However, it is non-trivial to develop such a sentiment classifier due to the following challenges:

- Although we can obtain a large-scale dataset of user-generated posts on COVID-19 via the social platform API, annotating a huge amount of posts with their corresponding sentiments is rather time-and-labor-consuming. In other words, we can only obtain limited labeled posts as well as a large number of unlabeled ones. Therefore, how to efficiently utilize the limited annotated and ample unlabeled posts to optimize the model is the first challenge we are facing.
- Considering the fact that COVID-19 is a global situation, whereby users are easy to convey the same sentiment via diverse expressions due to the language variants. Data augmentation is one promising solution to deal with the diverse expressions, and hence enhance the model's robustness. However, due to the informality of the user-generated posts, directly adopting the conventional data augmentation methods, like the back translation, may hurt the performance. Accordingly, how to devise an effective data augmentation method to cope with informal user-generated posts on social media is another crucial challenge.

To address these challenges, we present a dual COnsistency-enhanced semi-superVIseD network for Sentiment Analysis (COVID-SA), where both labeled and unlabeled data are jointly exploited. As can be seen in Fig. 2, COVID-SA consists of three key components: knowledge-based data augmentation, text encoding, and dual consistency-enhanced text classification. The first component works on augmenting each unlabeled post to mimic the user's diverse posts. Considering that the user-generated posts are usually informal, instead of using the conventional back translation augmentation method, we introduce a new knowledge-based augmentation method, which is able to use the external knowledge to enhance the representations of the informal posts. The second one is designed to encode both the labeled and unlabeled text, where we adopt the Bidirectional Encoder Representations from Transformers (BERT) [7] as the backbone. Notably, the BERT backbone can be replaced by any other common text encoding framework, like the Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM). The third component is a Multi-Layer Perceptron (MLP)-based sentiment classifier, which is not only supervised by the cross-entropy loss over the limited labeled posts, but also regularized by the dual consistency (namely, label-oriented consistency and instance-oriented consistency) modeling over the unlabeled and augmented data. Specifically, the label-oriented consistency is devised to model the semantic consistency between each augmented post and its original unlabeled post, while the instance-oriented one targets at modeling the semantic consistency among the augmented instances of the same unlabeled post.

To evaluate the generalization of our proposed model, we conduct extensive experiments on not only our constructed dataset from Weibo, consisting of 20 million user posts whereby 87,361 are labeled, but also three existing benchmark datasets from other domains, namely topic classification, news classification, and sentiment analysis of product reviews. Extensive experiments demonstrate the superiority of our COVID-SA over existing methods on various applications.

In a nutshell, the contributions of this work can be summarised as follows:

- To the best of our knowledge, we are the first to construct a large-scale dataset from the largest Chinese social platform Weibo, for the sentiment classification of COVID-19 related posts. This dataset comprises 20 million user posts, including 87,361 labeled ones.
- We devise a novel COVID-SA model, where a knowledge-based augmentation scheme is presented for informal context. In addition, both the label-oriented consistency between the augmented instance and the original sample as well as the instance-oriented consistency among augmented instances of the same sample are jointly modeled. As far as we know, we are the first to combine instance-oriented loss and label-oriented loss to encourage the mode to learn the latent semantic representation for semi-supervised learning. In addition, we also introduce a confidence hyperparameter to reduce the impact of false pseudo-label samples.
- To justify our proposed method, we have conducted extensive experiments on both the self-constructed dataset and three additional benchmark datasets from various domains. The results demonstrate the universality and effectiveness of our proposed model. As a side contribution, we also released the dataset, source code, and trained parameters to facilitate the reproduction.[6]

II. Related Work

Our work is related to semi-supervised text classification, contrastive learning, sentiment analysis toward COVID-19, and domain adaptation.

[6]https://github.com/LiqiangJing/DCSTC

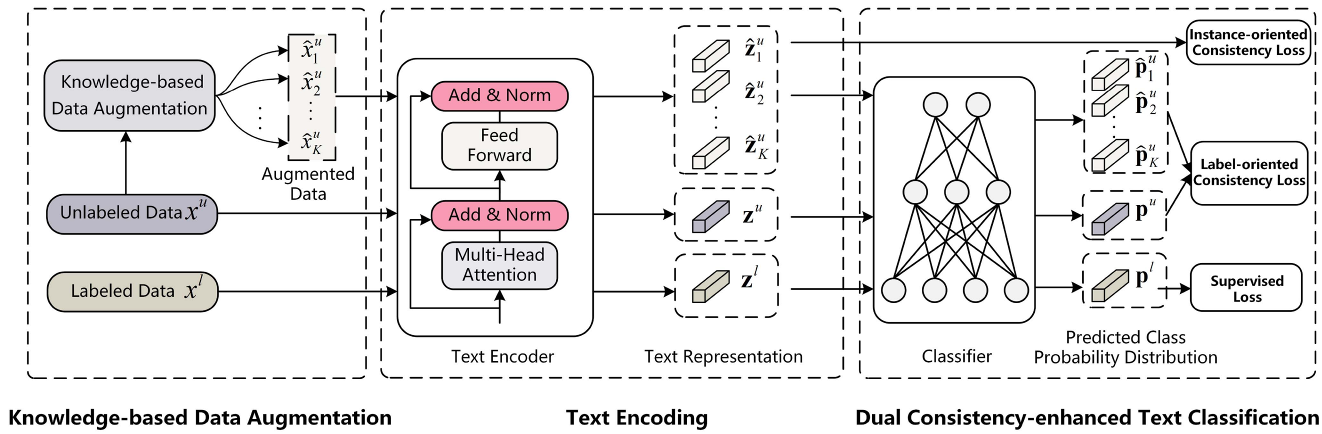


Fig. 2. Overall Architecture of COVID-SA. It consists of three key components: knowledge-based data augmentation, text encoding, and dual consistency-enhanced text classification.

A. Semi-Supervised Text Classification

Because training data annotation is time-and-labor-consuming, semi-supervised learning has received much attention in the natural language processing community [8], [9], [10]. To leverage the unlabeled data in the training phase, Chen et al. [11] adopted variational autoencoders to generate the pseudo labels for unlabeled instances, and fed them into the sequential variational labelers to predict sentence labels. In addition, Grandvalet et al. [12], Lee et al. [13], and Menget al. [14] incorporated the self-training strategy with the semi-supervised learning setting to infer the pseudo labeled data for the unlabeled ones and treated them as the labeled data to optimize the predictor. Despite their remarkable performance, the pseudo label hardly provides the supervision with a high confidence.

To address the problem, the consistency regularization is utilized in semi-supervised learning models based on the assumption that the decision boundary is unlikely to go through the high density areas [15]. Therefore, it aims to augment some unlabeled data by perturbing the original one and enforce their outcomes closer, providing more confident cues to optimize the classification model. Chen et al. [16], [17] introduced the consistency regularization into semi-supervised text classification, which encourages the pseudo labels of unlabeled data and their perturbed ones to be consistent. Moreover, Xie et al. [18] utilized the back translation techniques to obtain the perturbed instances and achieved state-of-the-art performance supervised by the consistency signal between the labels of the original sample and its perturbed instances. Although these methods have achieved compelling progress, they only consider the label-oriented consistency between the perturbed instance and the original sample, overlooking the instance-oriented consistency among perturbed instances of the same original sample, which is one major concern of our work.

B. Contrastive Learning

Contrastive learning works on extracting the discriminative features from the input information by identifying the semantically similar sample pairs from the multiple dissimilar ones. This technique has been widely-used in multiple fields, including computer vision [19], [20], information retrieval [21], and recommendation [22]. In a sense, the core of the contrastive learning is to construct the contrastive pairs for the model optimization by data augmentation. For example, SimCLR [23] augments the images by blurring operations and pairs the blurred images with the original ones as the positive and negative pairs. In the recommendation domain, Wei et al. [22] put the collaborative embeddings and content features of items together to build the contrastive pairs.

Recently, several studies have been dedicated to introducing the contrastive learning into natural language processing tasks [24], [25]. However, different from the spatial augmentations (e.g., rotation and flipping) in computer visions, the augmentation in natural language processing takes the complex syntax and semantic structure of textual data into consideration. Zhang et al. [24] applied contrastive learning to text clustering, and constructed contrastive pairs by back translation and word replacement. Lee et al. [25] added a perturbation to the hidden representation of the target sequence to augment data for conditional text generation. Gao et al. [26] utilized dropout as data augmentation. Despite their compelling success, they are designed for high-quality (formal) text. When it comes to the informal text, they probably cause the semantically inconsistency between the original and augmented instances. In light of this, we design a new knowledge-based data augmentation method for informal text.

C. Sentiment Analysis Toward COVID-19

During the COVID-19 epidemic, people tend to express opinions on social media because of the epidemic prevention regulation. Hence, recent work began to focus on analyzing people's emotions through posts written by people. Some studies trained neural networks with limited labeled data for sentiment analysis toward COVID-19. For example, Lyu et al. employed ERNIE [27], BERT and LSTM to conduct sentiment analysis on posts from Weibo [28], respectively. Furthermore, to compare the human behavior in two geographical locations affected by

TABLE I
SUMMARY OF THE MAIN NOTATIONS

Symbol	Explanation
$\mathcal{X}^l$	The set of labeled training text.
$\mathcal{X}^u$	The set of unlabeled training text.
B	The batch size of labeled data.
μB	The batch size of unlabeled data. μ is the hyperparameter for controlling the batchsize.
K	The number of augmented data for each text.
$f(\cdot)$	The BERT-based text encoder.
$g(\cdot)$	The MLP-based text classifier.
w_j	The j-th word in a given text.
$\mathbf{z}_i^l$	The representation of the i-th labeled text in the batch.
$\hat{x}_{i,k}^u$	The k-th augmented data for the i-th unlabeled text in the batch.
$\hat{\mathbf{z}}_{i,k}^u$	The representation of the k-th augmented data for the i-th unlabeled text in the batch.
τ	The threshold hyperparameter.
$\mathbf{p}_i^u$	The predicted class probability distribution of the i-th unlabeled text in the batch.
$\hat{\mathbf{p}}_{i,k}^u$	The predicted class probability distribution of $\hat{x}_{i,k}^u$.

COVID-19, Garcar et al. analyzed tweets in Portuguese and English by BERT [29]. In addition, Yu et al. directly utilized the dictionary-based sentiment analysis method for news and social media when the annotated data are not available [30]. Existing work has either trained neural models with small amounts of labeled data or used dictionary-based approaches without utilizing the labeled data, overlooking the potential of unlabeled data. Different from the existing work, we design a dual consistency-enhanced semi-supervised network for sentiment analysis where a new knowledge-based augmentation method is devised to augment the informal posts on COVID-19, to maximize the value of labeled and unlabeled data.

D. Domain Adaptation

Domain adaptation aims to apply the model trained in the plentiful labeled source domain to the target domain [31] which has attracted enormous research attention as it can save laborious tasks for collecting labeled data in the target domain. The existing domain adaptation methods can be categorized into two classes: unsupervised domain adaptation methods and semi-supervised domain adaptation methods. The unsupervised adaptation methods just utilize the labeled source domain data and unlabeled target domain data [6], [32]. For example, Huang et al. proposed a deep multi-representation adversarial learning method, which can mitigate the inconsistency between feature transferability and discriminability in unsupervised domain adaptation task [33]. Differently, the semi-supervised domain methods not only utilize a sufficiently labeled source domain but also use the partially labeled target domain to achieve domain adaptation [5], [34]. Although these domain adaptation methods have achieved great success, their performance is largely limited by the domain gap between the source domain and the target domain. These methods usually perform well when the domain gap is small, such as the gap between the handwritten digits domain and the house numbers domain [6]. However, as we mentioned before, the distribution of the keywords between COVID-19-related posts and other domain datasets is quite distinct, which demonstrates that the corresponding domain gap is relatively large and the domain adaptation methods could yield suboptimal performance.

III. Method

In this section, we first formulate the research problem, and then detail the three key components of our proposed dual consistency-enhanced semi-supervised network for sentiment analysis.

A. Problem Formulation

Let us first declare some notations. In particular, we use bold uppercase letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to denote matrices and vectors, respectively. We employ nonbold letters (e.g., x) to represent scalars, and Greek letters (e.g., γ) as parameters. If not clarified, all vectors are in column forms. The main notations used in this paper are summarized in Table I.

Suppose that we have a set of labeled text, denoted as $\mathcal{X}^l$. Let x^l be a labeled text, and y be the corresponding one-hot label vector $\mathbf{y} \in \{0,1\}^C$. C is the total number of classes. In addition, we have a set of unlabeled text, termed as $\mathcal{X}^u$. Our goal is to derive an accurate semi-supervised text classifier based on both labeled and unlabeled text, i.e., $\mathcal{X}^l \bigcap \mathcal{X}^u$.

B. Knowledge-Based Data Augmentation

A fact is that the robust model should make an invariant prediction for the perturbed sample whose nature is not changed [18]. Hence, we resort to the data augmentation strategy to scale up the original training data set, make data as diverse as possible, and hence improve the performance of the model.

One potential data augmentation method is the widely used back translation method [35], [36]. Given a text in some language (e.g., English), back translation first translates it into another intermediate language (e.g., French), and then translates it back into the original language. For example, given the English text "economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending", back translation first translates it into the French text "La croissance

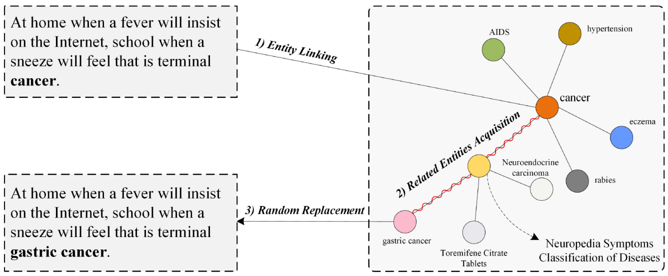


Fig. 3. Illustration of knowledge-based data augmentation.

économique au Japon ralentit á mesure que le pays subit une baisse des dépenses intérieures et des dépenses des entreprises", and then translates the French sentence back to the English sentence "Japan's economic growth slows as country suffers decline in domestic and business spending". In this manner, the new English sentence, even slightly different, is deemed semantically similar to the original English sentence and can be treated as one augmented data of the original sentence. Although the back translation may work well for the formal text (e.g., the press released documents) that seldom has mistakes in grammar and spelling, its performance can be largely degraded for COVID-19 tweets due to the following two facts. 1) The user-generated posts on social media are informal text that usually contains typos, grammatical errors, and abbreviations. 2) COVID-19 is an emerging topic with many new phrases or expressions (e.g., COVID-19) being created. As a result, the semantics of the original text can be hardly retained in the back-translation augmented text, which hinders the following semi-supervised text classification.

Another potential data augmentation method is the unsupervised text style transfer method [37], [38] which can translate one style text into another style text under the unsupervised setting. However, these methods rely on the domain gap between the source domain and the target domain. As we mentioned before, the keyword distribution of COVID-19-related posts is significantly different from the other domain datasets. Meanwhile, under the unsupervised text style transfer setting, the quality of augmented texts is hard to guarantee. In light of this, to ensure the quality of the augmented data for informal tweets, we propose a knowledge-based data augmentation method, where we focus on replacing the local entities of the text without changing the structure of the whole text to generate the augmented data. The underlying philosophy is that although the informal tweet is somehow noisy, its reference entities are still relatively reliable. Similar to Unsupervised Data Augmentation (UDA) [18], we focus on the augmentation of the unlabeled data. Specifically, we replace the entities of a given unlabeled text with the help of a public concept knowledge graph, which comprises a large number of common entities, concepts, and relationships. As shown in Fig. 3, our knowledge-based data augmentation process can be summarized as follows:

1) *Entity Linking*. Given an unlabeled text x^u, we first identify the set of entities involved in the text, denoted as $\mathcal{E} = \{e_1, e_2, \ldots, e_T\}$, according to a concept knowledge graph, whereby T represents the total number of entities identified in the given text.
2) *Related Entities Acquisition*. For each entity $e_i \in \mathcal{E}, i = 1, \ldots, T$, we retrieve its related entities by looking up entities that belong to the same concept of the original one according to the concept knowledge graph. These related entities are deemed as the entity candidates for replacing the original one e_i in the text x^u for data augmentation. We denote the set of entity candidates for entity e_i as $\mathcal{E}_i^r$
3) *Random Replacement*. We generate the augmented data for the text x by replacing the entity e_i, which is randomly selected from the entity set $\mathcal{E}$, with a randomly selected related entity from the candidate set $\mathcal{E}_i^r$.

Ultimately, for a given unlabeled text x^u, we generate K augmented text, i.e., $\{\hat{x}_1^u, \hat{x}_2^u, \ldots, \hat{x}_K^u\}$, with the aforementioned knowledge-based data augmentation method.

C. Text Encoding

After the data augmentation, we have three kinds of text: the labeled, unlabeled, and the augmented text. In this work, we use the same text encoding network to process all these texts. For simplicity, we temporally omit the subscripts and superscripts, and denote an arbitrary text as x. Specifically, we denote each text as $x = \{w_1, w_2, w_3, \ldots, w_m\}$, where w_j refers to the jth word in the text, and m is the total number of words in the text. In accordance with the previous work [18], we adopt BERT as the text encoder, due to its great success on many natural language processing tasks, such as Text Generation [39], [40], Sentiment Analysis [41], [42], and Named Entity Recognition [43], [44]. Mathematically, each text x can be encoded as follows,

$$\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_m} = f(w_1, w_2, \ldots, w_m), \tag{1}$$

where $f(\cdot)$ denotes the BERT based text encoder, $\mathbf{h_j} \in \mathbb{R}^{d_1}$ refers to the hidden state vector for the jth word w_j, and d_1 is the corresponding dimension. We then derive the final embedding for text x by averaging the hidden state vectors for all words in the text as follows,

$$\mathbf{z} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{h_j}, \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^{d_1}$ is the embedding for the text x. For clarity, we denote the embedding for each labeled text x^l as $\mathbf{z}^l$, each unlabeled text x^u as $\mathbf{z}^u$, and each augmented text $\hat{x}_k^u$ of the unlabeled text x^u as $\hat{\mathbf{z}}_k^u$, $k = 1, 2, \ldots, K$.

D. Dual Consistency-Enhanced Text Classification

Regarding the labeled text, based on their embeddings, we resort to the multi-layer perceptron (MLP) with cross-entropy loss function, which has shown remarkable success in the text classification, to classify the text. Similar to the work [45], we adopt the batch-based text classification and hence have the following objective function:

$$\begin{cases}\mathcal{L}^l = \frac{1}{B}\sum_{i=1}^{B} CE(\mathbf{y}_i, \mathbf{p}_i), \\ \mathbf{p}_i = g(\mathbf{z}_i^l), \\ \quad = softmax(\mathbf{W}_2 tanh(\mathbf{W}_1\mathbf{z}_i^l + \mathbf{b}_1) + \mathbf{b}_2).\end{cases} \tag{3}$$

where $g(\cdot)$ denotes the MLP-based text classifier, and $\mathbf{p}_i \in \mathbb{R}^C$ is the predicted class probability distribution for the ith labeled text in the batch. C is the total number of classes. $\mathbf{W}_1 \in \mathbb{R}^{d_2\times d_1}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times d_2}$ are weight matrices. $\mathbf{b}_1 \in \mathbb{R}^{d_2}$ and $\mathbf{b}_2 \in \mathbb{R}^C$ are the bias vectors. d_2 refers to the dimension of the latent output of the first layer. $tanh(\cdot)$ stands for the hyperbolic tangent activation function and $softmax(\cdot)$ refers to the softmax function. $CE(\cdot)$ denotes the cross entropy loss function, and B is the batch size of labeled data in the training process.

To make full use of the unlabeled data, we devise the comprehensive dual consistency based regularization, where both the label-oriented consistency between the augmented instance and the original sample, and the instance-oriented consistency among augmented instances of the same sample are jointly explored.

Label-Oriented Consistency Modeling. Similar to previous studies, we argue that the semantic of the augmented text and that of the original text should maintain certain latent consistency. Thus, the augmented instance of one sample should share the same label with the original sample. Towards this end, we first sample a set of μB unlabeled text from $\mathcal{X}^u$ as a batch, where μ refers to the hyperparameter, controlling the relative size of the labeled and unlabeled dataset. We then obtain the predicted class probability distribution of each sampled unlabeled text x_i^u, termed as $\mathbf{p}_i^u$, and that of the kth augmented text for it, termed as $\hat{\mathbf{p}}_{i,k}^u$, according to (3). Moreover, to encourage the consistency between the original unlabeled text and the augmented text, we treat the predicted label for the unlabeled text x_i^u as the pseudo-label for its augmented instance and employ the cross entropy loss function below,

$$\mathcal{L}_{label}^u = \frac{1}{\mu BK}\sum_{i=1}^{\mu B}\sum_{k=1}^{K}\mathbb{1}(\max(\mathbf{p}_i^u) > \tau)CE(\mathbf{p}_i^u, \hat{\mathbf{p}}_{i,k}^u), \tag{4}$$

where $\mathbb{1}(\cdot)$ is an indicator function, and $\max(\mathbf{p}_i^u)$ denotes the maximum entry of the predicted distribution vector $\mathbf{p}_i^u$. A fact is that the pseudo labels may be inaccurate. Hence, we introduce a threshold to select accurate pseudo labels as soon as possible. It is worth noting that, similar to the work [45] in the computer vision field, we introduce the threshold τ to reduce the error caused by false pseudo-labels. Intuitively, we assume that the pseudo-label should be reliable, only if the predicted probability of the most possible class for the given text is larger than the predefined threshold τ. In a sense, only the reliable pseudo-labels are meaningful to regularize the latent consistency.

Instance-Oriented Consistency Modeling. As a matter of fact, the literature in semi-supervised learning only takes into account the label-oriented consistency between each augmented instance and the original sample, but overlooks the instance-oriented consistency between each pair of augmented instances of the same sample. In a sense, the semantics of the augmented instances derived from the same sample should be consistent. To achieve the instance-oriented consistency modeling, we resort to the contrastive learning, where we expect the semantics of the augmented instances derived from the same sample should be more similar than that of augmented instances derived from different samples. In this way, the feature representation of the augmented instances derived from the same sample is similar.

In this part, we adopt the same set of μB unlabeled text used for the label-oriented consistency modeling, and for each unlabeled text x_i^u, we have a set of K augmented data, denoted as $\{\hat{x}_{i,1}^u, \hat{x}_{i,2}^u, \ldots, \hat{x}_{i,K}^u\}$. The text encoding of these augmented data can be represented as $\hat{\mathbf{z}}_{i,1}^u, \hat{\mathbf{z}}_{i,2}^u, \ldots, \hat{\mathbf{z}}_{i,K}^u$, respectively. In our context, we adopt the InfoNCE loss that has shown compelling success in the field of contrastive learning [46], [47]. In particular, we have the following objective function:

$$\begin{cases}\mathcal{L}_{insta}^u = \frac{1}{\mu BK(K-1)}\sum_{i=1}^{\mu B}\sum_{k_1=1}^{K}\sum_{k_2\neq k_1} l_i(k_1,k_2), \\ l_i(k_1,k_2) = -\log\frac{exp(s(\hat{x}_{i,k_1}^u, \hat{x}_{i,k_2}^u))}{\sum_{j\neq i}^{\mu B}\sum_{k_3=1}^{K} exp(s(\hat{x}_{i,k_1}^u, \hat{x}_{j,k_3}^u))}, \\ s(\hat{x}_{i,k}^u, \hat{x}_{j,k_3}^u) = \cos(\hat{\mathbf{z}}_{i,k}^u, \hat{\mathbf{z}}_{j,k_3}^u).\end{cases} \tag{5}$$

where $\hat{x}_{i,k}^u$ represents the kth augmented data of the ith unlabeled text. $l_i(k_1,k_2)$ denotes the pairwise loss between augmented data $(\hat{x}_{i,k_1}^u, \hat{x}_{i,k_2}^u)$ for the ith unlabeled text. $s(\hat{x}_{i,k}^u, \hat{x}_{j,k_3}^u)$ calculates the cosine similarity between the encoding of the kth augmented instance of the ith unlabeled text, i.e., $\hat{\mathbf{z}}_{i,k}^u$, and that of the k_3th augmented instance of the jth unlabeled text, i.e., $\hat{\mathbf{z}}_{j,k_3}^u$. We treat each pair of augmented instances from the same original sample, e.g., $(\hat{x}_{i,k_1}^u, \hat{x}_{i,k_2}^u)$ for the ith unlabeled text, as the positive pair. Meanwhile, for each positive pair $(\hat{x}_{i,k_1}^u, \hat{x}_{i,k_2}^u)$, we regard all the text pairs $(\hat{x}_{i,k_1}^u, \hat{x}_{j,k_3}^u)$, where $j \neq i, k_3 = 1, 2, \ldots, K$, as the corresponding negative ones. We expect that the similarity between the positive pair is larger than that between negative ones.

Combining the supervised text classification and the dual consistency modeling, we reach the final objective function:

$$\mathcal{L} = \mathcal{L}^l + \alpha\mathcal{L}_{label}^u + \beta\mathcal{L}_{insta}^u, \tag{6}$$

where α and β are hyperparameters, balancing the effect of each component. Algorithm 1 summarizes the entire training process.

IV. Experiment

In this section, we elaborate the experiment details and analyze the results by answering the following research questions:

RQ1. Does the proposed method outperform state-of-the-art baselines towards text classification?

RQ2. How does the proposed method perform with varying amounts of unlabeled data?

RQ3. How does each component of the proposed method affect the performance?

A. Datasets

To evaluate our proposed sentiment classification model in practice, we collected a large corpus of tweets related to COVID-19 from Sina Weibo, one of the largest Chinese social media platforms, to analyze netizen's sentiments to breaking events on COVID-19. It is worth mentioning that considering sentiment

Algorithm 1: Training Process.

Input: Labeled examples $\mathcal{X}^l$, unlabeled examples $\mathcal{X}^u$, batch size B of labeled examples, batch size μB of unlabeled examples, number of augmentations K, Epoch E, iteration $Iter$.
Require: text encoder $f(\cdot)$, text classifier $g(\cdot)$.
Load the pre-trained text encoder $f(\cdot)$.
Initialize the text classifier $g(\cdot)$ with random parameters.
for $e = 1$ to E **do**
  for $t = 1$ to $Iter$ **do**
    Randomly sample B examples from $\mathcal{X}^l$ to construct a labeled batch.
    Randomly sample μB examples from $\mathcal{X}^u$ to construct an unlabeled batch.
    for $i = 1$ to B **do**
      Predict the label of sample in the labeled mini-batch by (3).
    end for
    for $i = 1$ to μB **do**
      for $k = 1$ to K **do**
        Data augmentation for the unlabeled data x_i^u.
      end for
      Predict the labels of unlabeled data and the corresponding augmented data by (3).
    end for
    Compute the overall loss by (6).
    Update the parameters of $f(\cdot)$ and $g(\cdot)$.
  end for
end for
returntext encoder $f(\cdot)$ and text classifier $g(\cdot)$.

classification can be cast as a text classification task in our context, we also justified our model on three benchmark datasets on text classification from other domains. In this subsection, we will describe our collected dataset and the benchmark datasets, respectively.

1) Our Collected Dataset: Data Collection. We built the COVID-19 dataset by collecting tweets posted between January 1, 2020 and April 1, 2020, when COVID-19 was at its peak and attracting numerous people expressing their personal opinions at Sina Weibo. In particular, to ensure the crawled tweets are related to the COVID-19 outbreak, we artificially constructed a set of 230 keywords, including but not limited to, Zhong Nanshan, and Jinyintan Hospital.[7] We then crawled the tweets by the keyword-based search function provided via the support of Weibo API. Overall, we crawled 20 million posts from Sina Weibo. Table II shows some example pieces of data.[8] In order to ensure the dataset quality, we removed the duplicated tweets as well as the retweets.

To get the intuitive insights of the collected data, we show the distribution of posts per day in Fig. 4. As can be seen, the posts

TABLE II
SOME INSTANCES IN ALL CRAWLING DATA

keyword	time	content
fever	01/01 23:38	I'm sick and have a fever, and I haven't been well. Take this as a sign that 2020 will be a year of good luck. Meat will be eaten on the first day of the new year, and meat will be eaten every day thereafter.
epidemic	01/19 19:03	I hope the sick will recover soon.

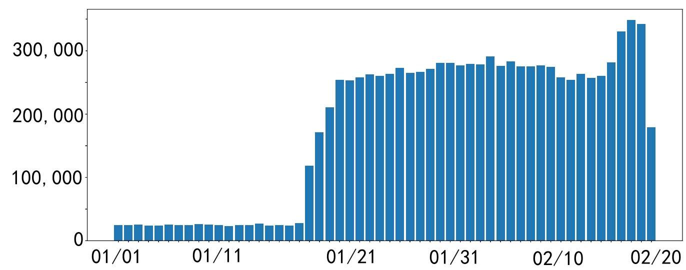


Fig. 4. The daily post distribution on COVID-19.

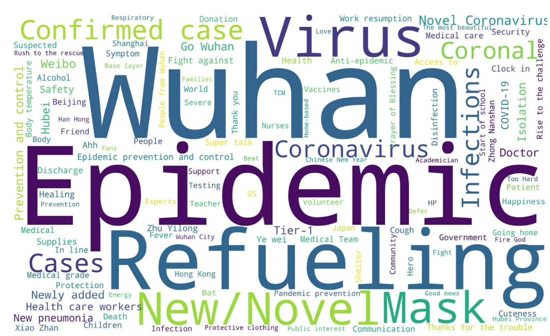


Fig. 5. Word cloud on COVID-19 dataset.

about COVID-19 before January 20, 2020 are rather limited (less than 25 k), while the number of posts related COVID-19 is growing rapidly since then. A possible explanation is that the Chinese Government issued an important instruction on COVID-19 on January 20, 2020. Meanwhile, to intuitively understand what people discussed on COVID-19, we draw the word cloud of our dataset. As can be seen from Fig. 5, the most frequent words appear in our dataset are "Wuhan", "Epidemic", "Case" and "Virus", which are indeed related to COVID-19. Meanwhile, we also found that the frequency of word "Refueling" is also much high, which reflects that many people have a positive attitude towards the COVID-19.

Data Annotation. Before annotation, we excluded the tweets with length less than 10 or greater than 150 characters. In addition, we filtered out the non-Chinese characters, hashtag, urls, and emogis from each tweet before data annotation. Due to the huge amount of the dataset, we randomly selected 100,000 tweets for annotation. Specifically, we employed 15 volunteers

[7] All the keywords can be accessed by here https://github.com/LiqiangJing/DCSTC

[8] Tables II, III, and IX are translated by us. You can see the original chinese text on https://github.com/LiqiangJing/DCSTC

TABLE III
SOME LABELED INSTANCES IN COVID-19

Content	Label
donate less, go on Wuhan, go on China, and hope that tomorrow's China will be even better	positive
I think time flies so fast.	neural
Today, my limbs are weak and I feel my strength is drained. Sigh.	negative

TABLE IV
STATISTICS OF THE FOUR DATASET. $\#C$: THE TOTAL NUMBER OF CLASSES

Dataset	Label Type	#C	#Unlabel	#Dev	#Test
Yahoo! Answers	QA Topic	10	50,000	50,000	60,000
AG News	News Topic	4	20,000	8,000	7,600
Product Review	Sentiment	2	6,000	1,000	2,500
COVID-19	Sentiment	3	15,000	6,000	6,000

#Unlabeled, #Dev, and #Test Refer to the number of samples of the corresponding type per class.

to label each selected tweet. Each volunteer was asked to label the given tweet with one of the four categories: "positive", "neutral", "negative" and "unknown", based on the text content of the tweet. The former three categories correspond to the typical sentiments, while the "unknown" category refers to the cases whose sentiments are hard to be clearly determined. For instance, it is uneasy to identify the sentiment for the tweet "I didn't expect to be lying in bed with a fever on the first day of 2020, hoping to be better in 2020", which seems to contain multiple sentiments. Each tweet is labeled by at least three volunteers, and the final label for each tweet is determined by the majority voting scheme. In case that a tweet is labeled with totally different labels by three volunteers, an extra volunteer would be invited for further annotating. Eventually, we only kept the tweets that are finalized into the former three categories (i.e., positive, neutral, and negative) to build our dataset. Finally, we obtained 87,361 tweets as well as their corresponding sentiment labels. Table III shows some annotated valid tweet examples. The average number of words per post in our annotated dataset is 64. For experiments, we randomly sampled 6,000 tweets as the test set.

2) Existing Datasets: To demonstrate the universality of our method COVID-SA, we also choose three existing datasets from other domains: Yahoo! Answers [48], AG News [49], and Product Review[9] dataset from NLPCC2014 [50].

Yahoo! Answers is an English topic classification dataset with 10 topic categories. Each category contains 140,000 training samples (i.e., question-answer pairs) and 5,000 test samples respectively.

AG News is an English news classification dataset with four categories. Each category has 30,000 training samples and 1,900 test samples, where each sample is composed of the news topic and description.

Product Review is a Chinese dataset for sentiment analysis of product reviews, which has two categories: positive and negative. Each category has 5,000 training samples (i.e., product reviews) and 1,250 test samples.

Overall, we have two Chineses datasets (i.e., COVID-19 and Product Review) and two English datasets (i.e., Yahoo! Answers and AG News). The statistics of these four datasets are shown in Table IV.

[9]http://tcci.ccf.org.cn/conference/2014/pages/page04_dg.html

B. Experiment Settings

Similar to previous study [16], in the experiments, the labeled set and unlabeled set for training, and the development set for validation of each dataset were randomly sampled from the corresponding training set. Notably, the data distribution of the labeled set and unlabeled set for each dataset were kept the same.

We used bert-base-chinese tokenizer[10] to tokenize text in both Chinese datasets (i.e., COVID-19 and Product Review), and bert-base-uncased tokenizer[10] for both English datasets (i.e., Yahoo! Answers and AG News). To facilitate researchers and developers to analyze the sentiment towards COVID-19, we pre-trained a BERT [7] model, named Co-BERT, with 20 million unlabeled data of our COVID-19 dataset, and used that as the text encoder.[11] Regarding the other three existing datasets, we directly used existing pretrained models. Specifically, we used the bert-base-chinese model[10] for the Chinese dataset Product Review, and bert-base-uncased[10] for the two English datasets. For text that has more than 256 English words or Chinese characters, we truncated the excess.

According to the characteristics of the text in different datasets, we adopted different data augmentation methods. For COVID-19 dataset with informal text, we augmented the unlabeled text based on the knowledge graph. In particular, we selected the knowledge graph CN-Probase[12] [51], a public large-scale structured encyclopedia in general domain covering tens of millions of entities and hundreds of millions of relationships, to augment the unlabeled data. For the other three datasets with relatively formal text, we used the back translation-based method to augment their unlabeled data. To be specific, for each English dataset, we selected two kinds of intermediate languages (Russian and German) for back translation with the help of fairseq,[13] where the random sampling temperature was set to 0.9. For the Chinese dataset Product Review, we selected English and French as intermediate languages for back translation using Baidu Translation.[14]

The learning rate of our BERT text encoder was set to 1e-5, and the learning rate of the MLP-based classifier was set to 1e-3. We set the number of augmentation K for unlabeled data to 2

[10]https://huggingface.co/transformers/
[11]We released Co-BERT via https://577279815.wixsite.com/website
[12]http://kw.fudan.edu.cn/cnprobase/
[13]https://github.com/pytorch/fairseq
[14]https://fanyi.baidu.com

and d_1 to 768. Regarding the two-layer MLP based classifier, we set the dimension of the latent output of the first layer as 128. The batch size for labeled data was set to 8 and that for unlabeled data was 4. The hyperparameter α was set to 1. Due to the concern that the loss $\mathcal{L}_{inter}^{u}$ cannot explicitly filter out predictions with low confidence like the loss $\mathcal{L}_{intra}^{u}$, and may hence introduce noise to the whole framework, we dynamically increase the trade-off weight β with the iteration gets larger as follows,

$$\beta = 1.0 \times \frac{Epoch_c}{Epoch_T}, \tag{7}$$

where $Epoch_T$ is total number of training epochs and $Epoch_c, c = 1, 2, \ldots, T$, is the current epoch.

C. On Model Comparison (RQ1)

In order to verify the effectiveness of our method COVID-SA, we compare COVID-SA with several state-of-the-art baselines, including both supervised methods and semi-supervised methods.

- *BERT* [7] is a supervised model, pre-trained on large-scale datasets, which has achieved compelling success in many text classification tasks. In details, we used the average pooling over the output of the BERT-based text encoder and two-layer MLP as the text classifier, which is the same as our method. The following baselines also used the same classifier architecture.
- *RoBERTa* [52] is a supervised model, devised based on BERT by making the following improvements: 1) training the model longer, with bigger batch size, over more data; 2) removing the next sentence prediction objective; 3) training on longer sequences; and 4) dynamically changing the masking pattern applied to the training data. Specifically, we used the public chinese-roberta-wwm-ext model[15] for the two Chinese datasets, and roberta-base model[10] for the two English datasets.
- *XLNet* [53] is also a supervised model, which uses Transformer-XL as the backbone for pre-training, and is pre-trained on a larger corpus than BERT.
- *UDA* [18] is also a semi-supervised method, which only considers the label-oriented consistency between the augmented instance and the original unlabeled sample. Similar to the work [18], we selected a subset of the original dataset to train the model, due to the lack of a TPU to train the whole model. We used the same data augmentation method of our model for the fair comparison.
- *SMDA* [17] is a semi-supervised model, which introduces another loss term to minimize the entropy of model's output on the basis of UDA. This loss term is based on one common assumption in many semi-supervised learning methods that a classifier's decision boundary should not pass through high-density regions of the marginal data distribution [12].

  For the fair comparison, we used the same data augmentation method of our model for SMDA, i.e., back translation-based augmentation for Yahoo! Answers, AG News, and Product Review, respectively, and the knowledge-based augmentation for COVID-19.

[15]https://github.com/ymcui/Chinese-BERT-wwm

TABLE V
PERFORMANCE (TEST ACCURACY) COMPARISON WITH BASELINES ON CHINESE DATASETS

Dataset	COVID-19			Product Review		
Labeled	**60**	**150**	**3000**	**20**	**400**	**1000**
BERT	65.78%	70.50%	74.93%	54.21%	75.59%	77.12%
RoBERTa	60.41%	68.96%	75.05%	55.00%	78.39%	79.73%
XLNet	54.17%	65.07%	74.7%	53.25%	78.18%	79.55%
UDA	66.58%	70.95%	75.27%	55.97%	78.75%	79.69%
SMDA	63.26%	63.83%	75.12%	44.63%	77.48%	79.32%
COVID-SA	**67.72%**	**71.28%**	**75.56%**	**58.47%**	**78.81%**	**79.82%**
Improv.	+1.71%	+0.47%	+0.39%	+4.50%	+0.08%	+0.11%

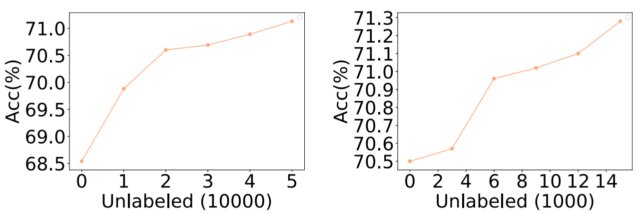
Similar to the work [18] and [45], we evaluated the performance of baselines and our COVID-SA with different amounts of labeled data for a comprehensive comparison. Tables V and VI show the performance comparison in terms of accuracy on the Chinese datasets and English datasets, respectively. Notably, for each baseline, we conducted 5 runs and reported the average performance. The last row 'Improv.' represents the relative improvement of our COVID-SA over the best baseline method. First, as can be seen, COVID-SA outperforms all the baseline methods consistently across different datasets, which demonstrates the superiority of COVID-SA over existing methods. Specially, we noticed that the performance improvement of COVID-SA is more significant when the amount of the labeled data is small. For example, as can be seen from Table VI, the accuracy is improved from 62.49% to 66.79% on Yahoo! Answers with only 100 labeled data. This indicates that COVID-SA is more superior to existing methods in cases when we have limited labeled data. Second, with the increase of the amount of labeled data, the performance of both supervised (i.e., BERT, RoBERTa and XLNet) and semi-supervised methods (i.e., SMDA, UDA and COVID-SA) get improved. This is reasonable as the more labeled data we have, the stronger the supervision can be achieved. Third, in most cases, semi-supervised models UDA and COVID-SA outperform the supervised ones (i.e., BERT, RoBERTa and XLNet), which indicates the advantage of taking into account the unlabeled data. However, with the increase of the amount of labeled data, the accuracy gap between semi-supervised and supervised methods becomes smaller. Last but not least, the performance improvement of COVID-SA is significant when the amount of labeled data is small. One possible reason is that labeled data play a more important role in model training and when the labeled data is small, the more gain unlabeled data brings.

D. On Varying Amounts of Unlabeled Data (RQ2)

To justify the power of unlabeled data, we explored the effect of the amount of unlabeled data on COVID-SA with two datasets: the English dataset Yahoo! Answers and the Chinese

TABLE VI
PERFORMANCE (TEST ACCURACY) COMPARISON WITH BASELINES ON ENGLISH DATASETS

Dataset	**Yahoo! Answers**					**AG News**		
Labeled	**100**	**500**	**1000**	**2000**	**5000**	**40**	**800**	**2000**
BERT	56.33%	66.13%	66.35%	68.54%	70.21%	66.63%	87.85%	88.66%
RoBERTa	58.52%	65.67%	66.80%	68.97%	70.78%	73.91%	87.18%	87.79%
XLNet	60.40%	66.48%	67.47%	69.20%	70.41%	77.15%	86.95%	87.86%
UDA	62.49%	67.71%	68.67%	69.85%	71.66%	83.71%	88.79%	89.50%
SMDA	20.45%	64.29%	69.07%	70.58%	71.59%	44.63%	88.70%	89.59%
COVID-SA	**66.79%**	**69.12%**	**70.21%**	**71.13%**	**71.72%**	**86.18%**	**89.16%**	**89.74%**
Improv.	+6.88%	+2.08%	+1.65%	+0.78%	+0.08%	+2.95%	+0.42%	+0.17%



(a) Yahoo! Answers (b) COVID-19

Fig. 6. Performance of our COVID-SA with different number of unlabeled data on two datasets.

dataset COVID-19. Specifically, we fixed the number of labeled data of each dataset. According to the scale of each dataset, we randomly selected 2000 labeled data for Yahoo!Answers, and 150 labeled data for COVID-19. Fig. 6 shows the performance of our COVID-SA with respect to different amounts of unlabeled data on the two datasets. From Fig. 6, we can see that the performance of COVID-SA keeps growing with the increase of the amount of unlabeled data from zero. On one hand, this confirms the advantage of incorporating the unlabeled data in text classification. On the other hand, this suggests that the more the unlabeled data, the better performance COVID-SA can achieve.

TABLE VII
PERFORMANCE (TEST ACCURACY) ON YAHOO! ANSWER WITH 2000 LABELED DATA AND COVID-19 WITH 150 LABELED DATA AFTER REMOVING DIFFERENT PARTS OF OURS

Dataset	**Model**	**Acc**
Yahoo! Answers	Ours	**71.13%**
	w/o Label-Oriented	68.89%
	w/o Instance-Oriented	70.52%
	w/o Label-Instance	68.54%
COVID-19	Ours	**71.28%**
	w/o Label-Oriented	71.01%
	w/o Instance-Oriented	70.86%
	w/o Label-Instance	70.50%

TABLE VIII
PERFORMANCE (TEST ACCURACY) OF OUR MODEL WITH DIFFERENT DATA AUGMENTATION METHODS ON COVID-19, WHERE 15000 UNLABELED DATA ARE USED

Labeled	**60**	**150**	**3000**
Ours+BT	67.15%	70.69%	74.22%
Ours+KG	67.72%	71.28%	75.56%

'BT' represents the back translation-based data augmentation and 'KG' stands for the knowledge-based data augmentation.

E. On Ablation Study (RQ3)

To gain deep insights of our COVID-SA, we comprehensively perform the ablation study from the following three aspects: 1) the effect of each component; 2) the validity of the proposed knowledge-based data augmentation method; and 3) the model universality over different text encoder backbones.

1) On Component-Wise Evaluation: To prove that each component of our proposed COVID-SA is valid, we compared COVID-SA with its following variants.

- *w/o Label-Oriented*. To explore the effect of the label-oriented consistency regularization, we removed the loss $\mathcal{L}^u_{label}$ by setting $\alpha = 0$.
- *w/o Instance-Oriented*. To study the effect of the instance-oriented consistency regularization, we removed the loss $\mathcal{L}^u_{insta}$ by setting $\beta = 0$.
- *w/o Label-Instance*. To investigate the effect of the dual consistency regularization on COVID-SA, we removed both losses $\mathcal{L}^u_{label}$ and $\mathcal{L}^u_{insta}$ by setting $\alpha = 0$ and $\beta = 0$. Notably, in this manner, this variant totally ignores the unlabeled data.

Table VII shows the ablation study results of COVID-SA. From the results, we gained the following observations. 1) w/o Label-Oriented performs worse than our COVID-SA, indicating that the label-oriented consistency regularization contributes to the performance of COVID-SA. 2) Our method surpasses w/o Instance-Oriented, which proves that contrastive learning based consistency regularization is indeed helpful for semi-supervised text classification. 3) w/o Label-Instance yields the worst performance. This confirms the necessity of incorporating the dual consistency regularization. Meanwhile, this reflects the advantage of exploiting the unlabeled data. Overall, these results show that both the label-oriented consistency and instance-oriented consistency regularizations are necessary to strengthen the performance of the model.

2) On Knowledge-Based Data Augmentation: To verify the utility of the proposed knowledge-based data augmentation

TABLE IX
SOME EXAMPLES FOR DATA AUGMENTATION

Origin Text	Back Translation-based Augmentation	Knowledge-based Augmentation
I just returned to my hometown Xianning a few days ago. . . I want to go back to Wuhan and isolate myself...	A few days ago I just returned to my hometown Xianning, now I want to go back to Wuhan to isolate myself...	I just returned to my hometown of Ningbo (a municipality under the jurisdiction of Zhejiang Province) a few days ago... I now want to go back to Ordos and isolate myself...
Compilation of the microbiology part of the real test questions of the examiner over the years @Dingxiangyuan	Compilation of the microbiology part of the real problems for the technicians in Dingxiangyuan laboratory for many years	Compilation of the microbiology part of the real test questions of the examiner over the years @Youku

TABLE X
PERFORMANCE (TEST ACCURACY) OF DIFFERENT MODELS ON THE YAHOO! ANSWERS DATASET WITH 5000 LABELED DATA

Model	Unlabeled	Accuracy
CNN	0	63.81%
Ours+CNN	50000	64.76%
LSTM	0	40.82%
Ours+LSTM	50000	56.03%

method, we also conducted experiments on COVID-19 dataset with different augmentation methods, and results are shown in Table VIII. From Table VIII, we can observe that the model with our proposed knowledge-based data augmentation method consistently outperforms that with the back translation-based data augmentation method under different settings regarding the amount of labeled data. This suggests that our proposed knowledge-based data augmentation method can maintain the semantic meaning of social media tweets better than the back translation-based augmentation method, and hence be more powerful to deal with social media text. For intuitive comparison, we also presented some augmented instances with different augmentation methods in Table IX. From Table IX, we can observe that the augmented text by back translation has semantic changes with respect to the original text, and is less readable than that obtained by our proposed augmentation method.

3) On Model Universality Over Text Encoders: In order to verify the universality of our proposed framework on different text encoders, we replaced the text encoder in our model, i.e., BERT, with the two common text encoders: CNN and LSTM, respectively. Accordingly, we have two new methods, named as Ours+CNN, and Ours+LSTM. To demonstrate the effectiveness of our method, we also introduce the purely supervised baselines CNN and LSTM, similar to the baseline BERT, for comparison. Table X shows the performance of different models on the Yahoo! Answers dataset. As can be seen, our method significantly improves the performance of the two pure supervised models, i.e., CNN and LSTM, respectively. The relative improvement on LSTM is even as high as 37.11%. This verifies that our framework can be effectively applied to different text encoders.

V. CONCLUSION AND FUTURE WORK

In this work, to study people's sentiments toward COVID-19 by user-generated posts on social media, we present a dual COnsistency-enhanced semi-superVIseD network for Sentiment Analysis (COVID-SA), which jointly explores both labeled and unlabeled text. In particular, COVID-SA consists of three key components: knowledge-based data augmentation, text encoding, and dual consistency-enhanced text classification. As one major novelty, we introduce a new knowledge-based augmentation method for informal text, as well as the dual consistency modeling to enhance the text classification performance. For evaluation, we created a large-scale dataset on COVID-19 from the largest Chinese social platform Weibo. In addition, to demonstrate the effectiveness of our model on text classification, we also conducted experiments on three existing benchmark datasets from other domains. Extensive experiment results demonstrate the superiority of our framework over existing methods under different settings. Moreover, ablation studies validate the validity of each component of our method, the effectiveness of the knowledge-based data augmentation method as well as the universality of our model over different text encoders.

Currently, we only focus on the textual modality for sentiment analysis. In the future, we plan to investigate the task of multimodal sentiment analysis.

REFERENCES

[1] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel arabic-reviews dataset construction for sentiment analysis applications," in *Proc. Intell. Natural Lang. Process.: Trends Appl.*, 2018, pp. 35–52.

[2] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, "A challenge dataset and effective models for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 6279–6284.

[3] H. M. Ahmed, M. Javed Awan, N. S. Khan, A. Yasin, and H. M. Faisal Shehzad, "Sentiment analysis of online food reviews using Big Data analytics," *Elementary Educ. Online*, vol. 20, pp. 827–836, 2021.

[4] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surveys*, vol. 54, pp. 62:1–62:40, 2021.

[5] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8049–8057.

[6] N. Xiao and L. Zhang, “Dynamic weighted learning for unsupervised domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15242–15251.

[7] J. Devlin, M. 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*, 2019, pp. 4171–4186.

[8] D. Yang, M. Wen, and C. P. Rosé, “Weakly supervised role identification in teamwork interactions,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process. Asian Federation Natural Lang. Process.*, 2015, pp. 1671–1680.

[9] F. Wang and C. Zhang, “Label propagation through linear neighborhoods,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 55–67, Jan. 2008.

[10] G. Li, K. Chang, and S. C. H. Hoi, “Multiview semi-supervised learning with consensus,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 11, pp. 2040–2051, Nov. 2012.

[11] M. Chen, Q. Tang, K. Livescu, and K. Gimpel, “Variational sequential labelers for semi-supervised learning,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 215–226.

[12] Y. Grandvalet and Y. Bengio, “Semi-supervised learning by entropy minimization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 529–536.

[13] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1–6.

[14] Y. Meng, J. Shen, C. Zhang, and J. Han, “Weakly-supervised neural text classification,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 983–992.

[15] A. Abuduweili, X. Li, H. Shi, C.-Z. Xu, and D. Dou, “Adaptive consistency regularization for semi-supervised transfer learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6923–6932.

[16] J. Chen, Z. Yang, and D. Yang, “Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2147–2157.

[17] J. Chen, Y. Wu, and D. Yang, “Semi-supervised models via data augmentation for classifying interactive affective responses,” in *Proc. Workshop Affect. Content Anal. Co-Located Conf. Artif. Intell.*, 2020, pp. 151–160.

[18] Q. Xie, Z. Dai, E. H. Hovy, T. Luong, and Q. Le, “Unsupervised data augmentation for consistency training,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6256–6268.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[20] J. Grill et al., “Bootstrap your own latent - a new approach to self-supervised learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[21] L. X. C. Xiong et al., “Approximate nearest neighbor negative contrastive learning for dense text retrieval,” in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–16.

[22] Y. Wei et al., “Contrastive learning for cold-start recommendation,” in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 5382–5390.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[24] D. Zhang et al., “Supporting clustering with contrastive learning,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 5419–5430.

[25] S. Lee, D. B. Lee, and S. J. Hwang, “Contrastive learning with adversarial perturbations for conditional text generation,” in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–25.

[26] T. Gao, X. Yao, and D. Chen, “SimCSE: Simple contrastive learning of sentence embeddings,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, M. X. Moens Huang, L. Specia, and S. W. Yih, Eds., 2021, pp. 6894–6910.

[27] Y. Sun et al., “ERNIE: Enhanced representation through knowledge integration,” 2019, *arXiv: 1904.09223*.

[28] X. Lyu, Z. Chen, D. Wu, and W. Wang, “Sentiment analysis on chinese weibo regarding COVID-19,” in *Natural Language Processing and Chinese Computing*, ser. Lecture Notes in Computer Science, vol. 12430. Berlin, Germany: Springer, 2020, pp. 710–721.

[29] K. Garcia and L. Berton, “Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA,” *Appl. Soft Comput.*, vol. 101, 2021, Art. no. 107057.

[30] X. Yu, C. Zhong, D. Li, and W. Xu, “Sentiment analysis for news and social media in Covid-19,” in *Proc. ACM SIGSPATIAL Int. Workshop Emerg. Manage. Using*, 2020, pp. 1–4.

[31] L. Zhang, “Transfer adaptation learning: A decade survey,” 2019, *arXiv: 1903.04687*.

[32] Y. Ganin et al., “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, pp. 59:1–59:35, 2016.

[33] J. Huang, N. Xiao, and L. Zhang, “Balancing transferability and discriminability for unsupervised domain adaptation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, pp. 1–8, 2022.

[34] P. Jiang, A. Wu, Y. Han, Y. Shao, M. Qi, and B. Li, “Bidirectional adversarial training for semi-supervised domain adaptation,” in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 934–940.

[35] S. Edunov, M. Ott, M. Ranzato, and M. Auli, “On the evaluation of machine translation systems trained with back-translation,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2836–2846.

[36] Z. Dou, A. Anastasopoulos, and G. Neubig, “Dynamic data selection and weighting for iterative back-translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 5894–5904.

[37] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, “Style transfer through back-translation,” in *Proc. Annu. Meeting Assoc. Comput. Assoc. Comput. Linguistics*, 2018, pp. 866–876.

[38] A. R. Kashyap, D. Hazarika, M. Kan, R. Zimmermann, and S. Poria, “So different yet so alike! constrained unsupervised text style transfer,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 416–431.

[39] Q. Chen, J. Lin, Y. Zhang, H. Yang, J. Zhou, and J. Tang, “Towards knowledge-based personalized product description generation in E-commerce,” in *Proc. ACM Int. Conf. Knowl. Discov.*, 2019, pp. 3040–3050.

[40] N. Dai, J. Liang, X. Qiu, and X. Huang, “Style transformer: Unpaired text style transfer without disentangled latent representation,” in *Proc. Conf. Assoc. Comput. Linguistics*, A. Korhonen, D. R. Traum, and L. M´rquez, Eds., 2019, pp. 5997–6007.

[41] K. Zhang et al., “EATN: An efficient adaptive transfer network for aspect-level sentiment analysis,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 377–389, Jan. 2023.

[42] E. öhman, M. P´mies, K. Kajava, and J. Tiedemann, “XED: A multilingual dataset for sentiment analysis and emotion detection,” in *Proc. Int. Conf. Comput. Linguistics*, 2020, pp. 6542–6552.

[43] J. Li, B. Chiu, S. Feng, and H. Wang, “Few-shot named entity recognition via meta-learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4245–4256, Sep. 2022.

[44] Y. Nie, Y. Tian, Y. Song, X. Ao, and X. Wan, “Improving named entity recognition with attentive ensemble of syntactic information,” in *Proc. Conf. Empirical Methods Natural Lang. Process.: Findings*, 2020, pp. 4231–4245.

[45] K. Sohn et al., “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[46] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[47] K. Sohn, “Improved deep metric learning with multi-class N-pair loss objective,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1849–1857.

[48] M. Chang, L. Ratinov, D. Roth, and V. Srikumar, “Importance of semantic representation: Dataless classification,” in *Proc. Conf. Artif. Intell.*, 2008, pp. 830–835.

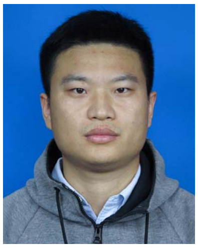
[49] X. Zhang, J. J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 649–657.

[50] Y. Zhou, R. Xu, and L. Gui, “A sequence level latent topic modeling method for sentiment analysis via CNN based diversified restrict boltzmann machine,” in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, 2016, pp. 356–361.

[51] J. Chen et al., “CN-Probase: A data-driven approach for large-scale chinese taxonomy construction,” in *Proc. IEEE Int. Conf. Data Eng.*, 2019, pp. 1706–1709.

[52] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, *arXiv: 1907.11692*.

[53] Z. Yang et al., “XLNet: Generalized autoregressive pretraining for language understanding,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5754–5764.



Teng Sun received the master's degree in School of Computer Science and Technology from Shandong University, Shandong, in 2020. He is currently working toward the PhD degree with the School of Computer Science and Technology, Shandong University, under the supervision of L. Nie. His research interests include multimedia computing and information retrieval.

Liqiang Jing received the BE degree in School of Computer Science and Technology from the Hefei University of Technology, Anhui, in 2020. He is a graduate student with the Department of Computer Science and Technology, Shandong University. His research interests include semi-supervised learning and natural language processing.

Yinwei Wei (Member, IEEE) received the MS degree from Tianjin University and the PhD degree from Shandong University, respectively. Currently, he is a research fellow with National University of Singapore. His research interests include multimedia computing and recommendation. Several works have been published in top forums, such as ACM MM, TMM, and TIP. He has served as the PC member for several conferences, such as MM, AAAI, and IJCAI, and the reviewer for TMM, TKDE, and TIP.

Xuemeng Song (Senior Member, IEEE) received the BE degree from the University of Science and Technology of China, in 2012, and the PhD degree from the School of Computing, National University of Singapore, in 2016. She is currently an associate professor of Shandong University, Jinan, China. Her research interests include the information retrieval and social network analysis. She has published several papers in the top venues, such as ACM SIGIR, MM, and TOIS. In addition, she has served as reviewers for many top conferences and journals.

Zhiyong Cheng received the PhD degree in computer science from Singapore Management University, in 2016, and then worked as a research fellow in National University of Singapore. He is currently a professor with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences). His research interests mainly focus on large-scale multimedia content analysis and retrieval. His work has been published in a set of top forums, including ACM SIGIR, MM, TMM, TIP, TOIS, TKDE. He has served as the PC member for several top conferences such as MM, SIGIR, WSDM, and the regular reviewer for journals including TKDE, TIP, TMM.

Liqiang Nie (Senior Member, IEEE) received the BEng and PhD degree from Xi'an Jiaotong University and National University of Singapore (NUS), respectively. He is currently the dean with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen campus). His research interests lie primarily in multimedia content analysis and information retrieval. He has co-/authored more than 100 CCF-A papers and 5 books, with 16 k plus Google Scholar citations. He is an AE of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Information Science*. Meanwhile, he is the regular area chair or SPC of ACM MM, NeurIPS, IJCAI and AAAI. He is a member of ICME steering committee. He has received many awards over the past three years, like ACM MM and SIGIR best paper honorable mention, in 2019, the AI 2000 most influential scholars 2020, SIGMM rising star, in 2020, MIT TR35 China 2020, DAMO academy young fellow, in 2020, SIGIR best student paper, in 2021, first price of the provincial science and technology progress award, in 2021 (rank 1), and provincial youth science and technology award, in 2022. Some of his research outputs have been integrated into the products of Alibaba, Kwai, and other listed companies.