# Adversarial Learning-based Stance Classifier for COVID-19-related Health Policies

Feng Xie*, Zhong Zhang*, Xuechen Zhao, Haiyang Wang, Jiaying Zou,
Lei Tian, Bin Zhou✉, and Yusong Tan
College of Computer, National University of Defense Technology, Changsha, China
{xiefeng, binzhou}@nudt.edu.cn

## Introduction

**Background**

The COVID-19 pandemic has caused immeasurable losses for people worldwide. To contain the spread of the virus and further alleviate the crisis, various health policies have been issued which spark heated discussions as users turn to share their attitudes on social media (e.g., Weibo, Twitter). Stance detection is of great practical value as an effective tool for Internet public opinion monitoring, which detects the attitude (i.e., *in favor of*, *against*, or *neutral*) of an opinionated text toward a pre-defined topic automatically.

**Challenges**

- Emerging policies. Policymakers will present new policies based on the dynamics of the epidemic situation in response to complex virus challenges. However, for these newly proposed health policies, the available labeled data is limited.
- Lack of background knowledge. Due to the specific scope of the training corpus, it is hard for language model to master the background knowledge of COVID-19 pandemic.

## Methodology (Fig. 1)

**➢ Encoder Module**

We apply BERT [1] to encode the texts and condition them with external knowledge about the topic (i.e., policy description) to enhance the model's deeper understanding.

**➢ Feature Separation Module**

The contextual representation generated by BERT contains both *topic-specific* and *topic-invariant* information. To allow the model to generalize to unseen topics, it is effective to learn and utilize transferable topic knowledge (i.e., topic-invariant information). We apply a linear transformation to distill the topic-invariant information for topic adaptation.

**➢ GeoEncoder Module**

Geographic signals can reflect potential characteristics and profiles of groups, e.g., cultural backgrounds and epidemiological contexts. We leverage GCN [2] to integrate geographic information as learnable region-specific features for model learning.

**➢ Stance Classifier**

We apply a linear layer with a softmax as the stance classifier to predict the stance labels.

**➢ Topic Discriminator**

We also apply a linear network with a softmax as the topic discriminator to classify the corresponding topic label based on topic-invariant features. Following [3], we add a GRL layer and the adversarial training process is essentially a min-max game:

$$\min_{\Theta_M} \max_{\mathbf{W}_{td}, \mathbf{b}_{td}} \mathcal{L}_{sc} - \alpha \mathcal{L}_{td}$$

## Contributions

**Approach.** We put forward an adversarial learning-based stance classifier to detect the public's attitudes toward COVID-19-related health policies, which can be applied to emerging policies with no labeled data.

**Enhancements.** To further enable the model to get a deeper understanding of the topics, we infuse policy descriptions as the external knowledge into the model. Meanwhile, a GeoEncoder is designed which encourages the model to learn unobserved background factors specified by each region and then represent them as non-text features.

**Experiments.** Experimental results demonstrate that our proposed method achieves state-of-the-art performance in both cross-target and zero-shot settings.

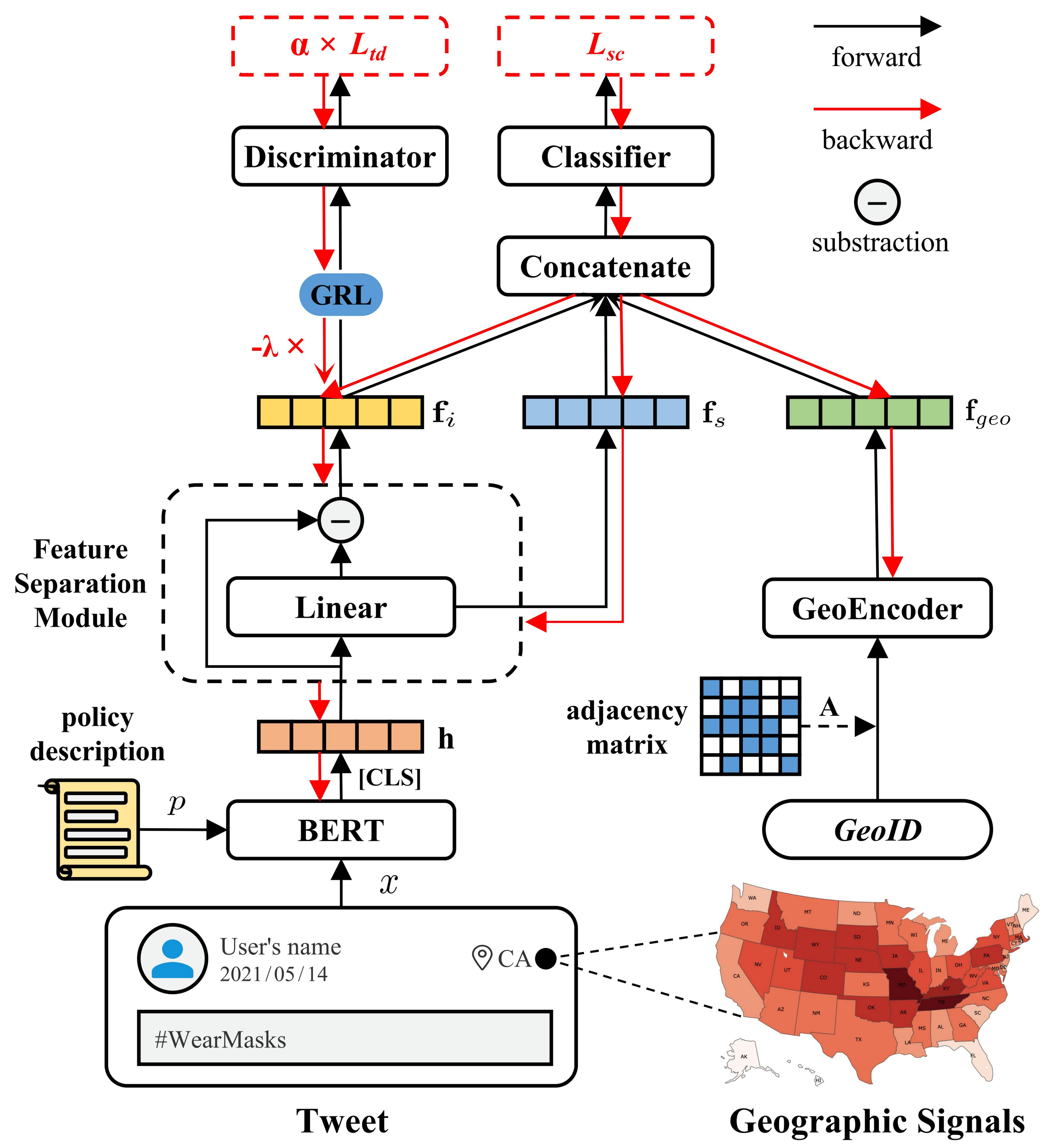


**Figure 1: The proposed method.**

## Experiments

- We select three health policies: (1) Stay at Home Order, (2) Wear Masks, and (3) Vaccination. There are a total of 1702 labeled texts and 3343 unlabeled texts. We choose a broad range of baselines including (1) neural network-based methods, (2) attention-based methods, and (3) BERT-based methods.
- We evaluate all models both in cross-target and zero-shot settings, and the results are reported in Table 1 and Table 2, respectively. Our proposed method outperforms comparison baselines on most tasks and improves the average $F_{avg}$ and $F_m$ by **3.5%** and **2.7%**, respectively.

| Models | Cross-target settings (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SH→WM | | SH→VA | | WM→SH | | WM→VA | | VA→SH | | VA→WM | |
| | $F_{avg}$ | $F_m$ | $F_{avg}$ | $F_m$ | $F_{avg}$ | $F_m$ | $F_{avg}$ | $F_m$ | $F_{avg}$ | $F_m$ | $F_{avg}$ | $F_m$ |
| BiLSTM | 25.4 | 30.6 | 25.6 | 30.5 | 39.1 | 45.1 | 40.8 | 47.5 | 31.5 | 38.1 | 33.0 | 38.6 |
| BiCond | 29.0 | 33.1 | 30.1 | 34.5 | 37.3 | 42.1 | 37.5 | 44.4 | 33.8 | 40.2 | 35.5 | 40.9 |
| TextCNN | 34.6 | 37.8 | 31.5 | 36.6 | 39.4 | 43.9 | 37.6 | 42.5 | 30.7 | 33.3 | 35.8 | 38.5 |
| TAN | 44.3 | 46.2 | 34.5 | 39.0 | 45.5 | 47.4 | 45.1 | 48.5 | 37.7 | 38.2 | 42.6 | 44.1 |
| CrossNet | 45.7 | 49.9 | 39.4 | 43.6 | 43.4 | 47.3 | 47.7 | 50.7 | 37.7 | 38.3 | 46.7 | 48.1 |
| BERT | 44.7 | 49.3 | 34.9 | 41.2 | 44.3 | 49.7 | 52.6 | 55.3 | 44.4 | 45.6 | 53.7 | 55.1 |
| WS-BERT-S | 45.4 | 49.1 | **40.3** | **44.9** | 41.9 | 48.0 | 51.0 | 54.8 | 39.9 | 41.4 | 47.2 | 49.9 |
| WS-BERT-D | 40.1 | 47.1 | 30.5 | 38.9 | 48.2 | 52.5 | 55.4 | 57.5 | 43.5 | 44.9 | 49.5 | 51.1 |
| Ours | **47.6** | **51.9** | 39.4 | 44.4 | **50.9** | **54.1** | **57.6** | **59.3** | **46.1** | **47.4** | **54.5** | **56.3** |
| Improve (%) | 4.2% | 4.0% | - | - | 5.6% | 3.0% | 3.9% | 3.1% | 3.8% | 3.9% | 1.5% | 2.1% |

**Table 1: Performance comparison of cross-target stance detection.**

| Models | Zero-shot settings (%) | | | | | |
|---|---|---|---|---|---|---|
| | SH | | WM | | VA | |
| | $F_{avg}$ | $F_m$ | $F_{avg}$ | $F_m$ | $F_{avg}$ | $F_m$ |
| BiLSTM | 45.6 | 49.6 | 25.7 | 31.7 | 36.7 | 42.4 |
| BiCond | 45.7 | 50.1 | 29.8 | 34.6 | 29.6 | 35.2 |
| TextCNN | 41.0 | 41.5 | 35.7 | 37.8 | 34.8 | 39.2 |
| TAN | 45.8 | 47.7 | 50.2 | 51.7 | 46.5 | 49.3 |
| CrossNet | 45.9 | 49.3 | 55.6 | 56.8 | 45.1 | 48.2 |
| BERT | 49.6 | 53.8 | 63.4 | 64.6 | 57.5 | 59.4 |
| WS-BERT-S | 48.6 | 53.0 | 61.0 | 62.4 | 55.6 | 57.9 |
| WS-BERT-D | 51.6 | 55.2 | 61.6 | 63.3 | 55.3 | 57.6 |
| Ours | **53.3** | **56.2** | **65.1** | **66.4** | **58.9** | **59.9** |
| Improve (%) | 3.3% | 1.8% | 2.7% | 2.8% | 2.4% | 0.8% |

**Table 2: Performance comparison of zero-shot setting.**

## Misc

**References**

[1] Kenton, Jacob Devlin Ming-Wei Chang and Toutanova, Lee Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL 2019*.
[2] Kipf, Thomas N and Welling, Max. 2017. Semi-supervised classification with graph convolutional networks. In *Proc. of ICLR 2017*.
[3] Ganin, Yaroslav and Lempitsky, Victor. 2015. Unsupervised domain adaptation by back-propagation. In *Proc. of ICML 2015*. 1180--1189.

**codes**: https://github.com/Xiefeng69/stance-detection-for-covid19-related-health-policies

Website: