Zero-Shot Rumor Detection with Propagation Structure via Prompt Learning

Hongzhan Lin[1*], Pengyao Yi[2*], Jing Ma[1†], Haiyun Jiang[3†], Ziyang Luo[1], Shuming Shi[4], Ruifang Liu[2]

[1]Hong Kong Baptist University
[2]Beijing University of Posts and Telecommunications
[3]Fudan University [4]Tsinghua University
{cshzlin, majing, cszyluo}@comp.hkbu.edu.hk, jianglu@fudan.edu.cn, {yi.py, lrf}@bupt.edu.cn

Code: https://github.com/PengyaoYi/zeroRumor_AAAI

AAAI'2023

gesis
Leibniz-Institut für Sozialwissenschaften

Forschungszentrum · Research Center L3S

Reported by Yang Jiang

INTRODUCTION

pretrain fine-tuning

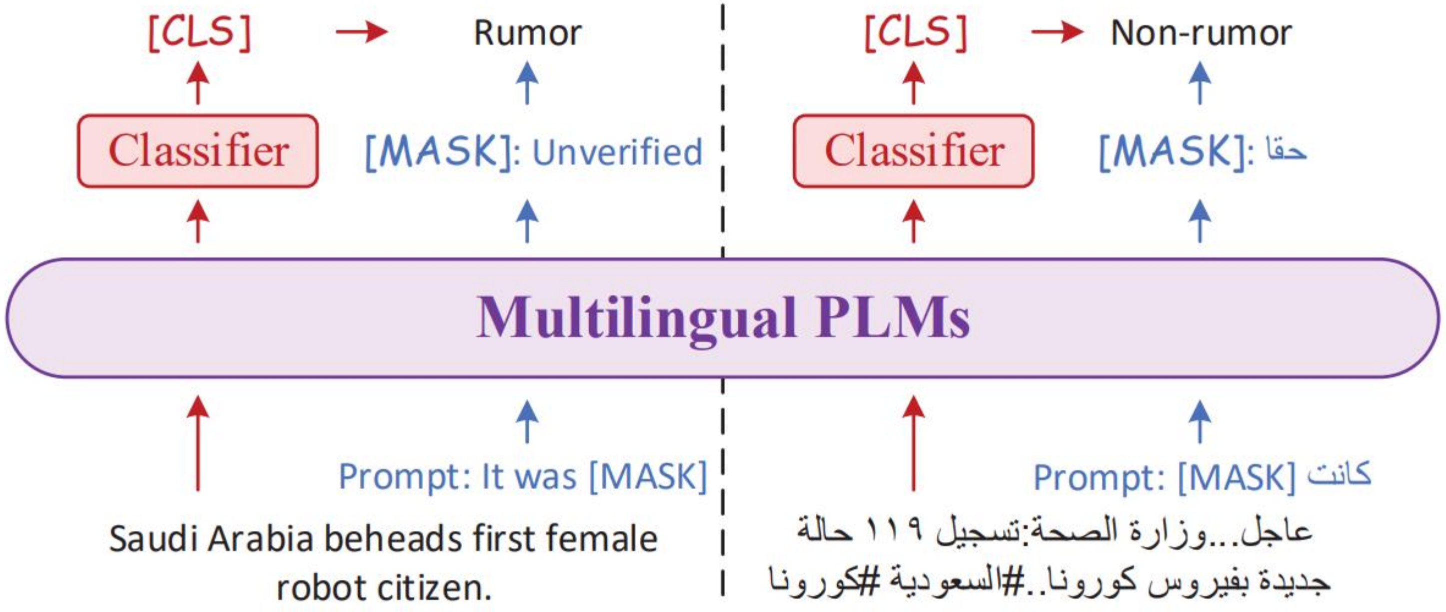
Detecting the single claim post with a heavy task-specific fine-tuning stage, which makes it deviate from the pre-training target on masked language modeling, even ignoring the domain-invariant interaction of user opinions during the diffusion of rumors

contrastive learning

A small number of target annotation is required. However, it is prone to be poor at emerging events propagated in minority languages without any expertise annotation, especially in some underdeveloped countries and regions.

INTRODUCTION

ZRD



a) Training on ***Politics*** domain in ***English***   b) Testing on ***COVID-19*** domain in ***Arabic***

INTRODUCTION

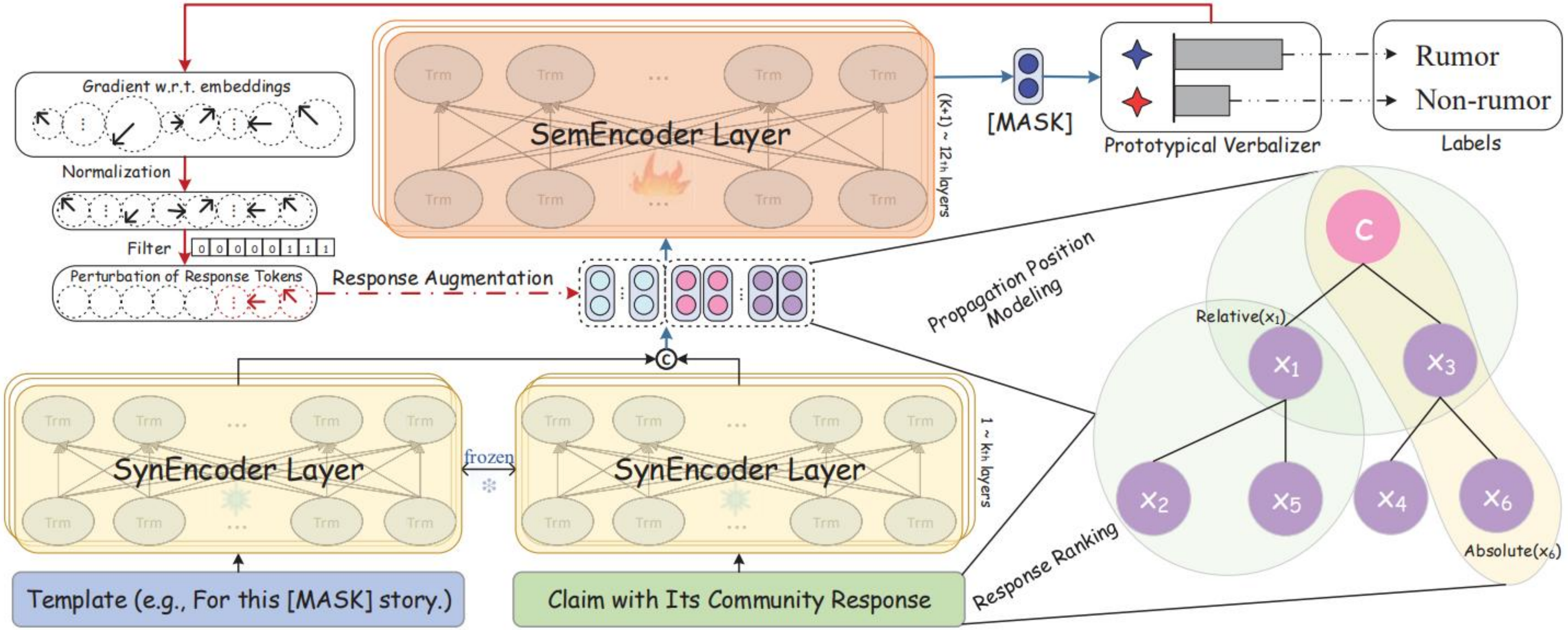
prompt learning

discrete prompt requires experts of native speakers to design rumor-related templates/rules

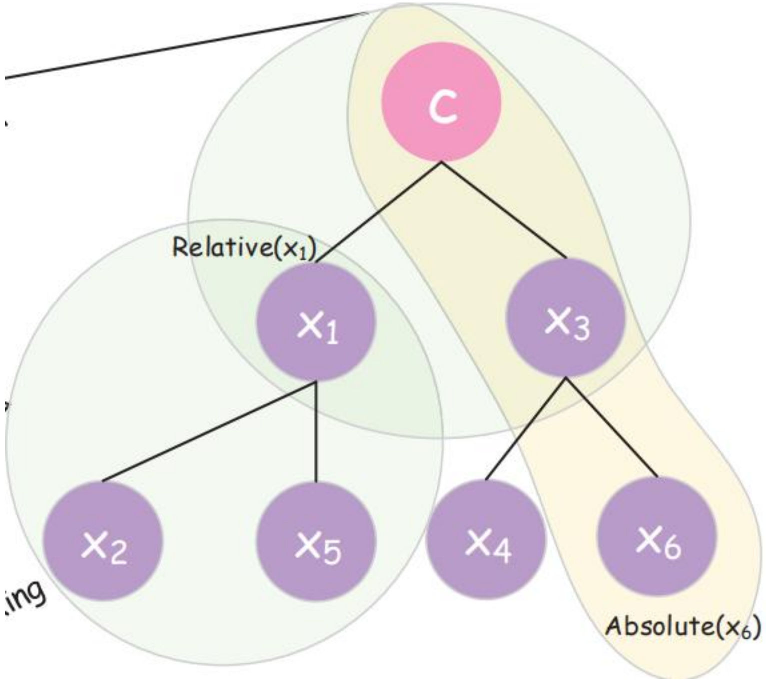
soft prompt soft prompt uses optimized token representations trained on a large dataset

RPL we propose to decouple shared semantic information from the syntactic bias in specific languages based on multilingual PLMs, which could enhance the semantic interaction between the prompt

Overview


Gradient w.r.t. embeddings
Normalization
Filter
0 0 0 0 0 1 1 1
Perturbation of Response Tokens
Response Augmentation
SemEncoder Layer
Trm
(k+1) ~ 12th layers
[MASK]
Prototypical Verbalizer
Rumor
Non-rumor
Labels
Propagation Position Modeling
Relative(x1)
Absolute(x6)
c
x1
x2
x3
x4
x5
x6
Response Ranking
SynEncoder Layer
frozen
1 ~ kth layers
Template (e.g., For this [MASK] story.)
Claim with Its Community Response

Definition



source

$$\mathcal{D}_s = \{C_1^s, C_2^s, \cdots, C_M^s\}$$

$$C^s = (y, c, \mathcal{T}(c))$$

$$\mathcal{T}(c) = [x_1^s, x_2^s, \cdots, x_m^s]$$

target

$$\mathcal{D}_t = \{C_1^t, C_2^t, \cdots, C_N^t\}$$

$$C^t = (c', \mathcal{T}(c'))$$

$$f(C^t|\mathcal{D}_s) \rightarrow y$$

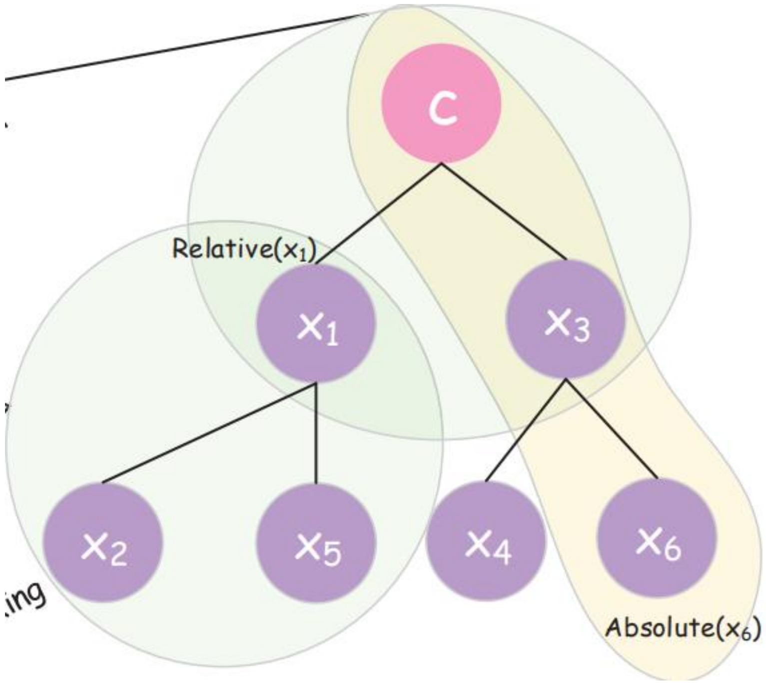
$$\mathcal{P}(y|\hat{c}) = g(\mathcal{P}([MASK] = v|\hat{c})|v \in \mathcal{V}_y) \quad (1)$$

$$\hat{c} = [p, c]$$

"For this [MASK] story."

Method

Response Ranking



time

$$\mathcal{T}(c) = [x_1, x_2, \cdots, x_m]$$

$$\mathcal{T}(c) = [x_m, x_{m-1}, \cdots, x_1]$$

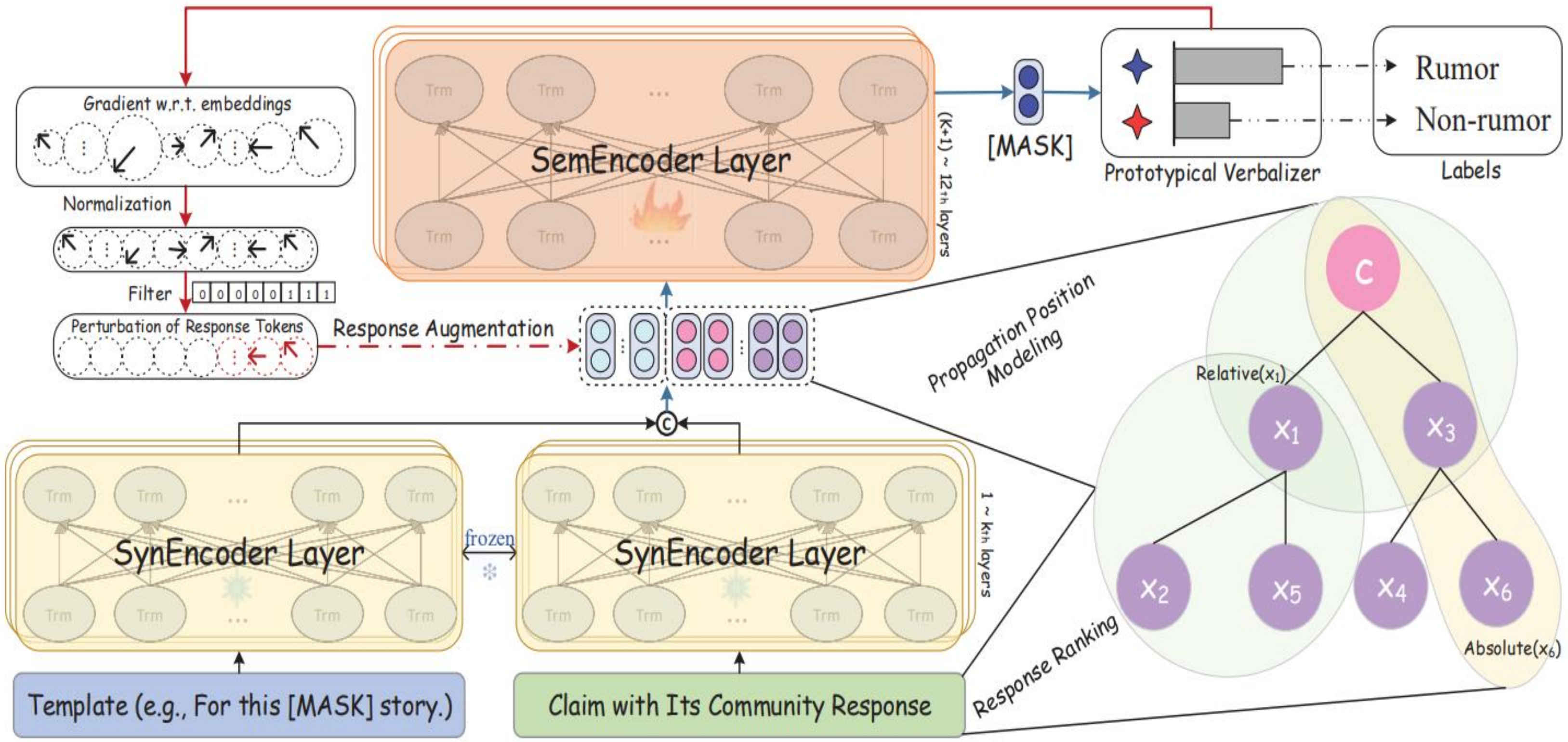
space

$$\mathcal{T}(c) = \langle \mathcal{G}, \vec{\mathcal{E}} \rangle$$

$$[x_1, x_2, x_5, x_3, x_4, x_6]$$

$$[x_1, x_3, x_2, x_4, x_5, x_6]$$

Method



Hierarchical Prompt Encoding

XLM-RBase (12 layers)

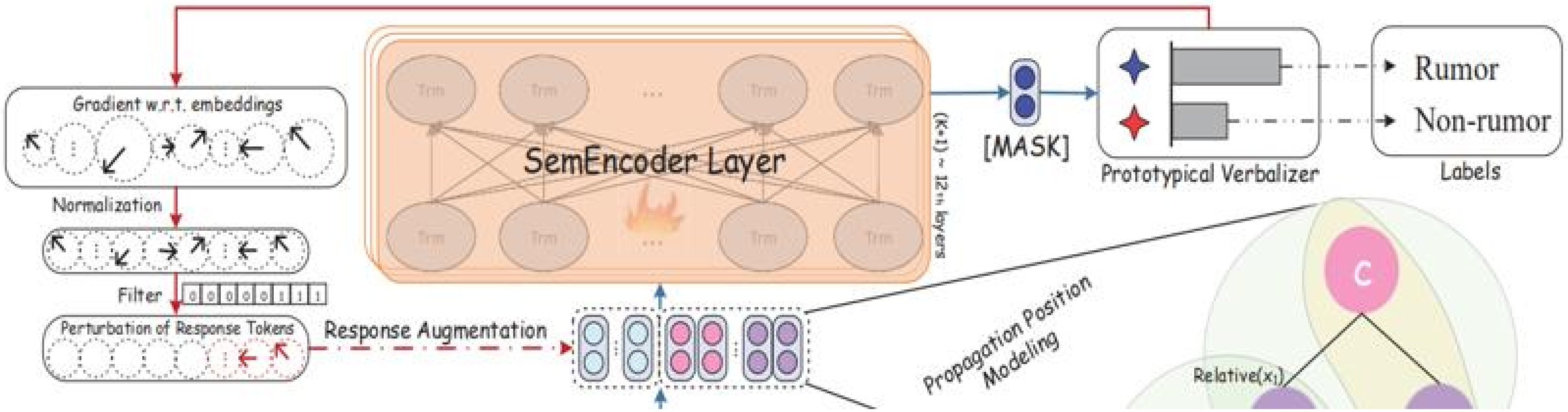
$$X_p = \mathrm{SynEncoder}\,(p) \tag{2}$$

$$X_{cr} = \mathrm{SynEncoder}\,([c, \mathcal{T}(c)]) \tag{3}$$

$$H = \mathrm{SemEncoder}\,([X_p, X_{cr}]) \tag{4}$$

$$abs_{pro}(q) = \mathrm{distance}_{tree}\,(x_i, c)$$

Method



$$\mathcal{L}_{proto} = -log \frac{e^{\mathcal{S}(H_i^m, l_y)}}{\sum_{y'} e^{\mathcal{S}(H_i^m, l_{y'})}} \quad (5)$$

$$\mathcal{L} = \alpha \mathcal{L}_{proto} + (1-\alpha)\hat{\mathcal{L}}_{con}.$$

$$\mathcal{L}_{con} = -\frac{1}{B_{y_i} - 1} \sum_j \mathbb{1}_{[i \neq j]} \mathbb{1}_{[y_i = y_j]} \log \frac{e^{\mathcal{S}(H_i^m, H_j^m)}}{\sum_{j'} \mathbb{1}_{[i \neq j']} e^{\mathcal{S}(H_i^m, H_{j'}^m)}} \quad (6)$$

$$\text{loss } \mathcal{L}_{avg} = \text{mean}(\mathcal{L} + \tilde{\mathcal{L}})$$

Experiments

Dataset	Source		Target		
	TWITTER	WEIBO	Twitter-COVID19	Weibo-COVID19	CatAr-COVID19
# of events	1154	4649	400	399	1699
# of tree nodes	60409	1956449	406185	26687	168276
# of non-rumors	579	2336	148	146	998
# of rumors	575	2313	252	253	701
Avg. time/tree	389 Hours	1007 Hours	2497 Hours	248 Hours	858 Hours
Avg. depth/tree	11.67	49.85	143.03	4.31	13.26
Language	English	Chinese	English	Chinese	Cantonese&Arabic
Domain	Open	Open	COVID-19	COVID-19	COVID-19

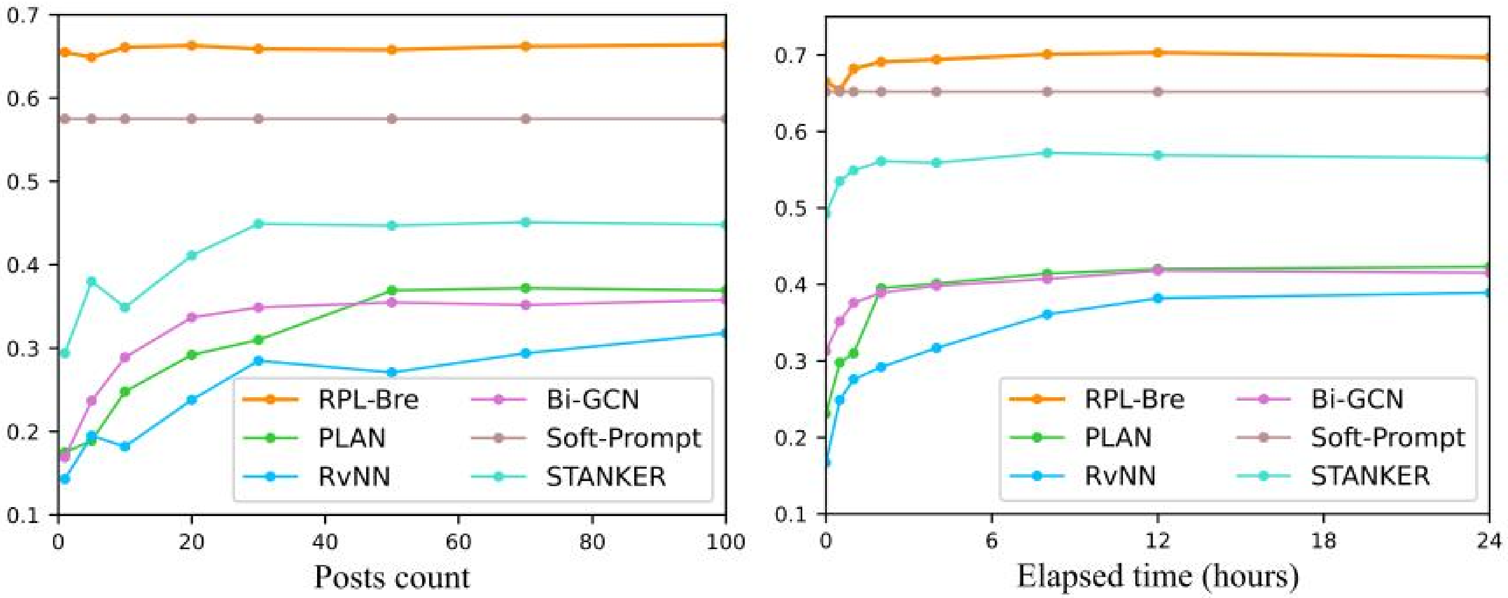
Experiments

Source	TWITTER								WEIBO							
Target	Weibo-COVID19				CatAr-COVID19				Twitter-COVID19				CatAr-COVID19			
Model	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1	Acc.	Mac-F_1	Rumor F_1	Non-rumor F_1
Vanilla-Finetune	0.623	0.585	0.711	0.459	0.518	0.402	0.583	0.220	0.603	0.602	0.619	0.585	0.481	0.481	0.479	0.474
Translate-Finetune	0.639	0.567	0.745	0.388	0.523	0.457	0.637	0.277	0.634	0.574	0.653	0.495	0.505	0.512	0.528	0.496
Contrast-Finetune	0.656	0.582	0.759	0.405	0.584	0.458	0.720	0.196	0.653	0.644	0.699	0.590	0.562	0.561	0.571	0.551
Adapter	0.644	0.600	0.737	0.463	0.558	0.438	0.665	0.211	0.652	0.612	0.736	0.487	0.548	0.556	0.605	0.508
Parallel-Adapter	0.651	0.598	0.730	0.467	0.567	0.450	0.701	0.198	0.667	0.653	0.731	0.574	0.579	0.585	0.636	0.534
Source-Prompt	0.664	0.648	0.722	0.574	0.589	0.564	0.460	0.669	0.670	0.616	0.760	0.472	0.599	0.565	0.688	0.441
Translate-Prompt	0.650	0.489	0.776	0.201	0.573	0.568	0.519	0.617	0.674	0.651	0.740	0.562	0.604	0.542	0.374	0.711
Soft-Prompt	0.652	0.574	0.756	0.392	0.590	0.565	0.446	0.683	0.685	0.652	0.758	0.546	0.609	0.575	0.518	0.633
RPL-Cho	0.713	0.675	0.786	0.563	0.613	0.581	0.455	0.707	0.715	0.689	0.778	0.601	0.634	0.633	**0.616**	0.650
RPL-Inv	0.728	0.666	**0.810**	0.521	0.601	0.592	0.473	**0.711**	**0.733**	0.710	0.788	0.632	0.647	0.640	0.586	0.693
RPL-Dep	0.732	0.689	0.805	0.574	**0.640**	**0.619**	0.530	0.708	0.723	**0.711**	0.771	**0.650**	0.657	0.636	0.547	**0.724**
RPL-Bre	**0.745**	**0.719**	0.804	**0.634**	0.631	0.617	**0.544**	0.689	0.727	0.697	**0.793**	0.601	**0.672**	**0.664**	0.614	0.714

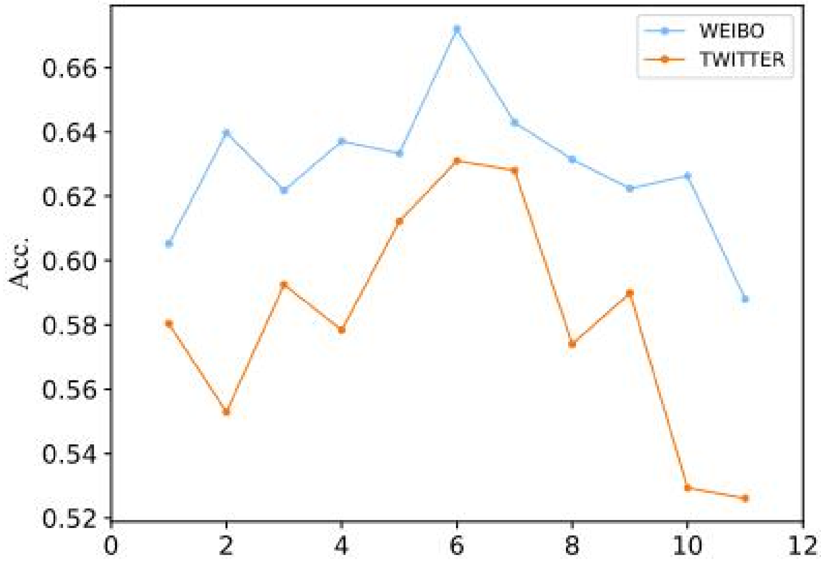
Experiments

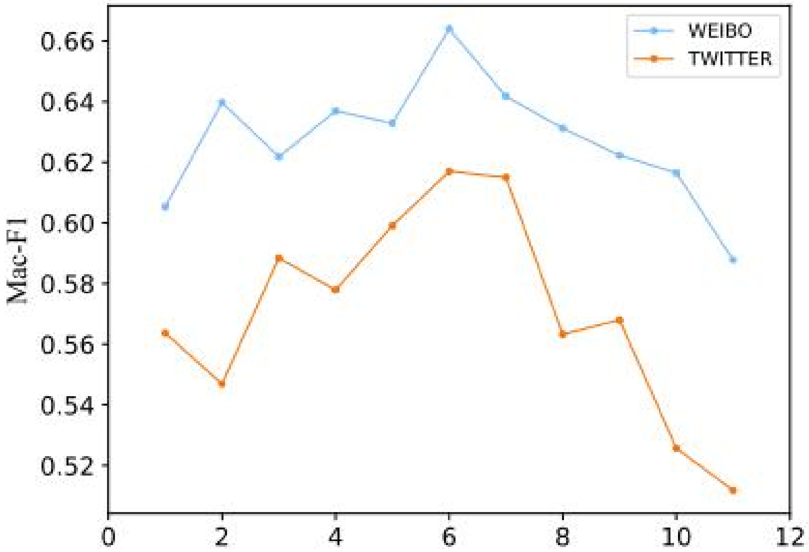
Source	TWITTER		WEIBO	
Model	Acc.	Mac-F_1	Acc.	Mac-F_1
RPL-Bre	0.631	0.617	0.672	0.664
RPL-Bre w/o RR	0.605	0.598	0.613	0.611
RPL-Bre w/o APP	0.622	0.607	0.626	0.624
RPL-Bre w/o RPP	0.610	0.601	0.633	0.632
RPL-Bre w/o ViRA	0.626	0.612	0.644	0.634
RPL-Bre w/o HPE	0.571	0.451	0.581	0.433
RPL-Bre w/o PV	0.592	0.589	0.621	0.617

Experiments

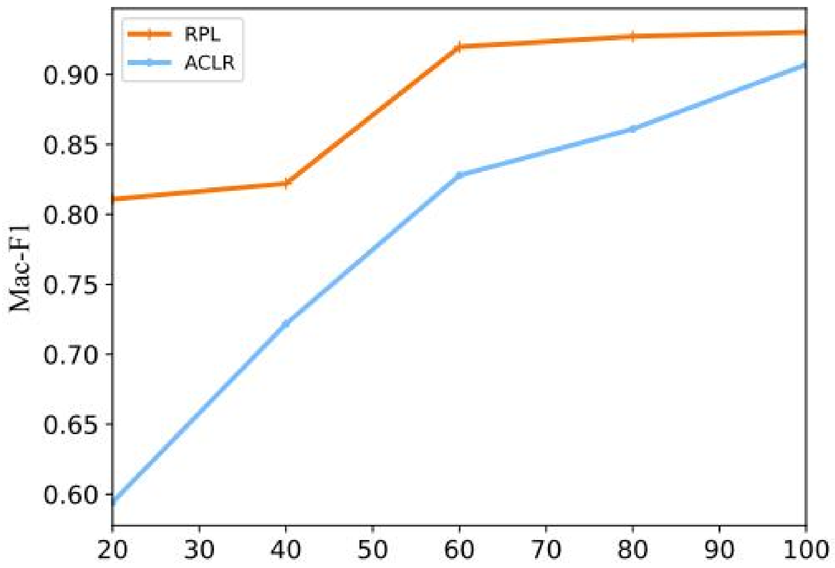


Experiments





Experiments



Thanks!