

An Effective Deployment of Diffusion LM for Data Augmentation in Low-Resource Sentiment Classification


Zhuowei Chen¹, Lianxi Wang^{1,2*}, Yuben Wu¹,
Xinfeng Liao¹, Yujia Tian¹, Junyang Zhong¹


¹Guangdong University of Foreign Studies

²Guangzhou Key Laboratory of Multilingual Intelligent Processing

INTRODUCTION

Sentiment classification models often overfit and lack generalization in low-resource scenario.

It's sunny today,
I am not feeling good.  positive

It's rainy today,
I am feeling amazing.  negative

Failure Cases

Most existing Data Augmentation methods:

- Replace minor context, keep crucial tokens.
- Simple logical modifications.
- Solely rely on pretrained knowledge.
- Introduce inconsistent samples.

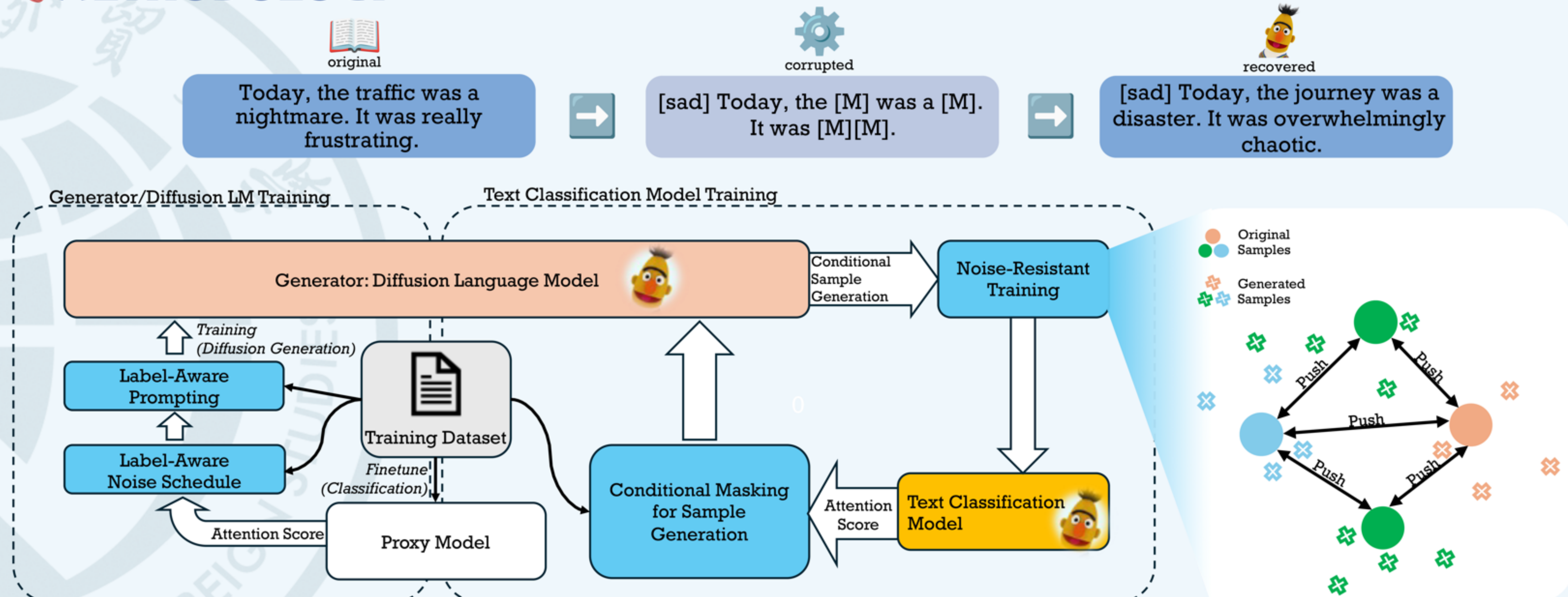
DiffusionCLS(ours):

- Reconstruct highly attentioned context.
- Applied Diffusion LM for DA.
- Build generators on specific task corpus.
- Balance diversity and consistency.

✦ Main Contributions:

- Proposed DiffusionCLS, a diffusion LM-based DA method for SC, generating diverse but consistent pseudo samples.
- Designed a noise-resistant training method, boosting model performance with pseudo samples.
- Experiments validate DiffusionCLS's superior performance, with ablation studies highlighting its module effectiveness.
- Visualization study discusses the diversity-consistency trade-off, further validating the effectiveness of DiffusionCLS.

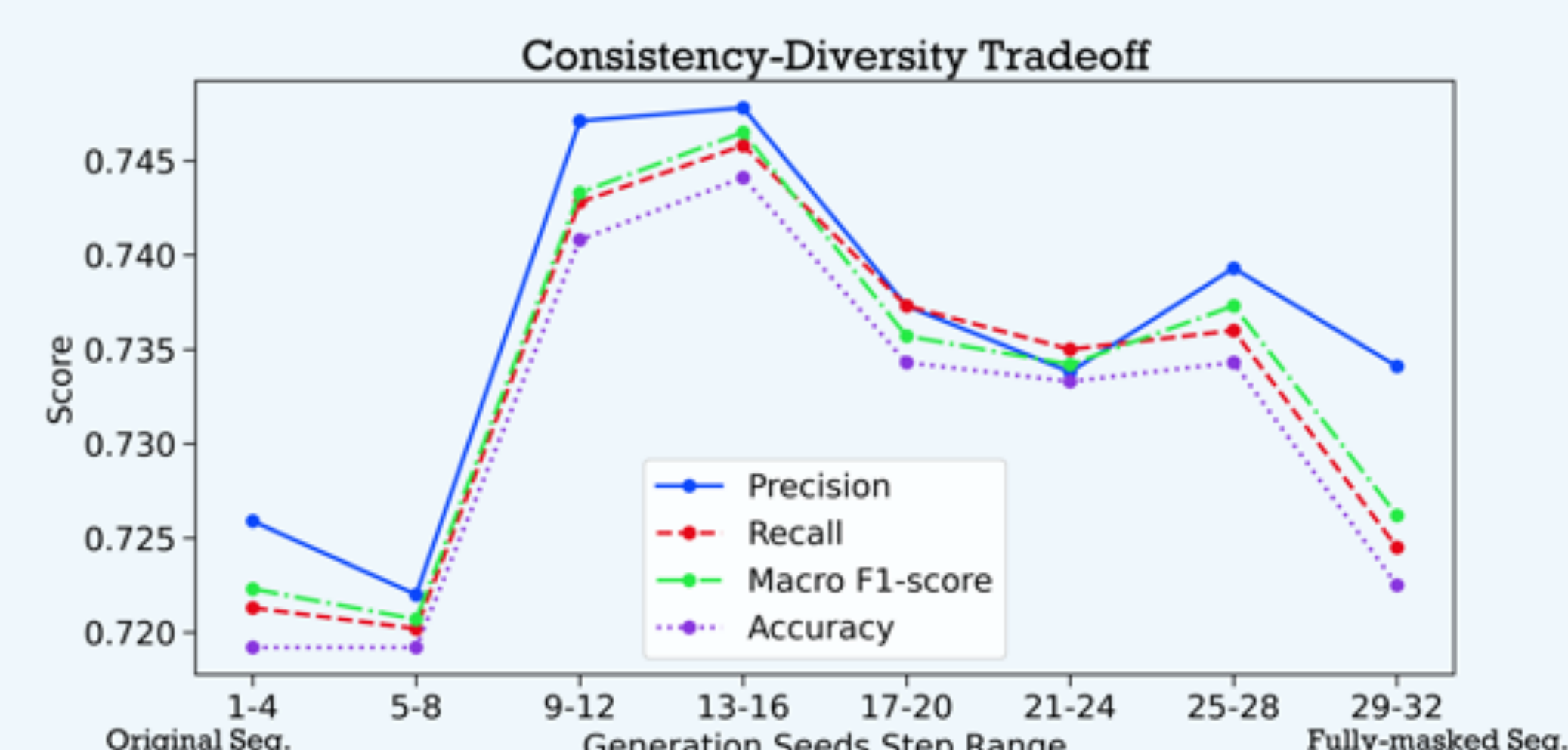
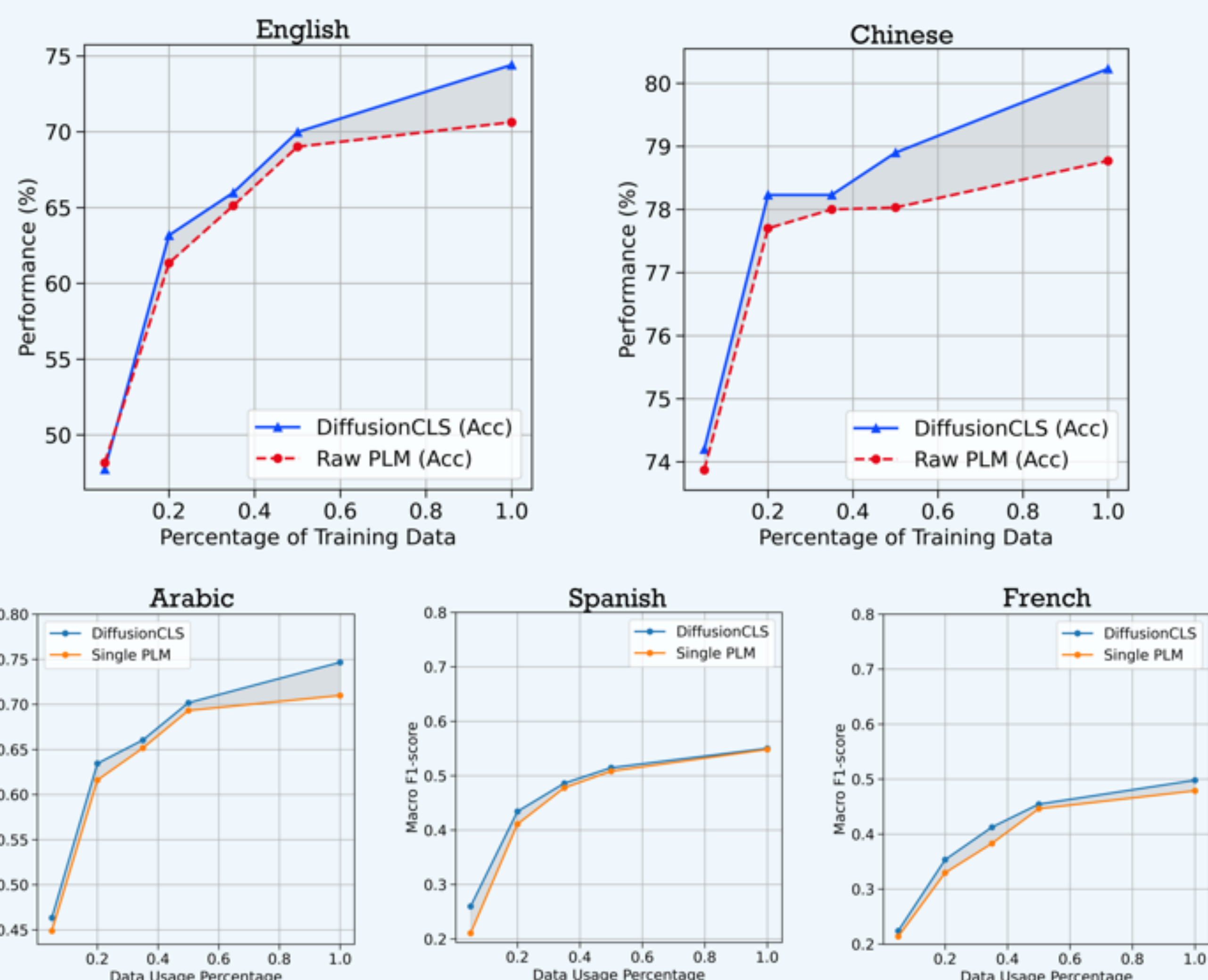
METHODOLOGY



RESULTS

Methods	Policy	SMP2020-EWECT				India-COVID-X			
		Macro-F	Acc	ΔF	ΔAcc	Macro-F	Acc	ΔF	ΔAcc
Raw PLM	N/A	65.87%	79.17%	-	-	70.99%	70.63%	-	-
+ Resample	B/D	64.84%	78.17%	-1.03%	-1.00%	72.74%	72.57%	1.75%	1.94%
+ BT (2019)	B/D	64.03%	77.93%	-1.84%	-1.24%	72.93%	72.79%	1.94%	2.16%
+ EDA (2019)	B/D	65.88%	78.87%	0.01%	-0.30%	66.83%	66.41%	-4.16%	-4.22%
+ AEDA (2021)	B/D	66.58%	79.50%	0.71%	0.33%	72.90%	72.89%	1.91%	2.26%
+ GENIUS (2022)	B/D	64.27%	78.23%	-1.60%	-0.94%	72.84%	72.46%	1.85%	1.83%
+ DiffusionCLS (ours)	B/D	66.47%	79.43%	0.60%	0.26%	72.80%	72.57%	1.81%	1.94%
+ BT (2019)	G/E	65.15%	77.93%	-0.72%	-1.24%	74.40%	74.30%	3.41%	3.67%
+ EDA (2019)	G/E	50.12%	71.87%	-15.75%	-7.30%	74.15%	73.87%	3.16%	3.24%
+ GPT-2 (2020)	G/E	65.06%	77.80%	-0.81%	-1.37%	69.55%	69.58%	-1.44%	-1.05%
+ AEDA (2021)	G/E	65.81%	78.93%	-0.06%	-0.24%	75.49%	75.27%	4.50%	4.64%
+ GENIUS (2022)	G/E	64.30%	78.07%	-1.57%	-1.10%	74.28%	74.08%	3.29%	3.45%
+ DiffusionCLS (ours)	G/E	67.98%	80.23%	2.11%	1.06%	74.65%	74.41%	3.66%	3.78%

Dataset	#Shot	Data Augmentation Methods									
		N/A	+EDA [†] (2019)	+BT [†] (2019)	+SSMBA [†] (2020)	+ALP [†] (2022)	+SE [†] (2023)	+GPT-2 (2020)	+mixup (2017)	+AWD (2023a)	+DiffusionCLS
SST-2	5	54.38	56.22	55.77	56.34	63.40	-	52.18	61.81	58.86	65.30
	10	61.82	53.96	62.05	59.05	69.72	57.56	54.17	61.55	64.62	68.29



FINDINGS

- The framework mitigates spurious pattern problems in the SC model.
- DiffusionCLS has a good balance of the diversity-consistency trade-off.
- It works across various languages and domains under different low-resource scenarios.

LIMITATIONS

Like most model-based data augmentation methods, the performance of data generators is also limited in extreme low-resource scenarios, due to limited samples for finetuning.