GASP: Graph-Augmented Soft Prompt Tuning for Few-Shot Text Classification

Sen Zhang, Mieradilijiang Maimaiti*, Wushour Silamu

School of Computer Science and Technology, Xinjiang University

IEEE SMC 2026.10
Washington, USA

CATALOG

01 **Background and Motivation**

Why the problem matters

02 **Research Goal and Contributions**

What we introduce

03 **Methodology**

How GraphPrompt works

04 **Experimental Setup**

How we evaluate it

05 **Results and Analysis**

What the evidence shows

06 **Conclusion and Q&A**

Key takeaways

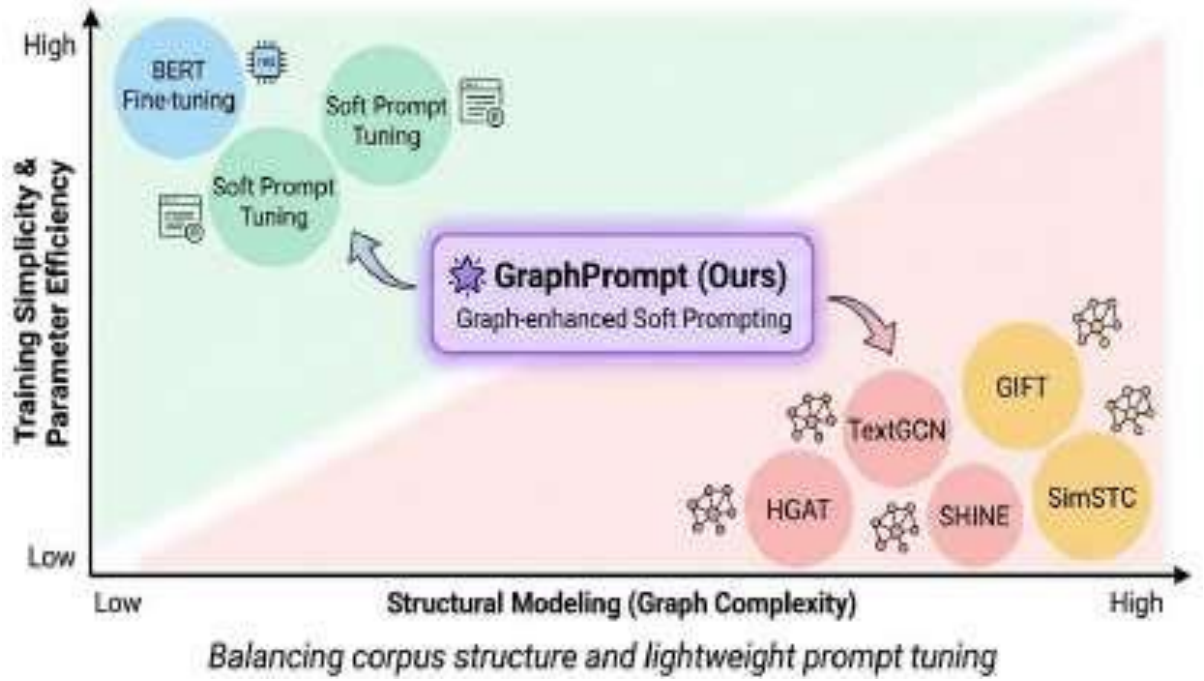
Background and Motivation

Research gap
Few-shot short-text classification combines semantic sparsity with very limited labeled supervision.

Graph-based STC captures corpus structure, but often adds document graphs, extra objectives, or multi-stage training.

Soft prompt tuning is lightweight, but standard prompts do not explicitly exploit word co-occurrence structure.

Our goal
Generate instance-specific continuous prompts from a PMI word graph, then prepend them to a standard BERT classifier.



GraphPrompt occupies the middle ground: structured enough to address sparse semantics, yet much lighter than complex graph pipelines.

Research Goal and Contributions

Goal **Improve few-shot short-text classification by injecting corpus-level word structure directly into continuous prompt embeddings.**

01 **Graph-guided prompts**

Encode a PMI word graph with WordGCN, pool document words by TF-IDF, and generate L continuous prompt tokens.

02 **Lightweight integration**

Prepend the generated prompt after [CLS] and retain a standard BERT classifier with single-stage cross-entropy training.

03 **Controlled evaluation**

Evaluate six public datasets under one 20-train + 20-validation protocol across five independently sampled splits.

04 **Evidence beyond accuracy**

Analyze BERT tuning, prompt length, trainable parameters, peak memory, runtime, and representative behavior.

Methodology

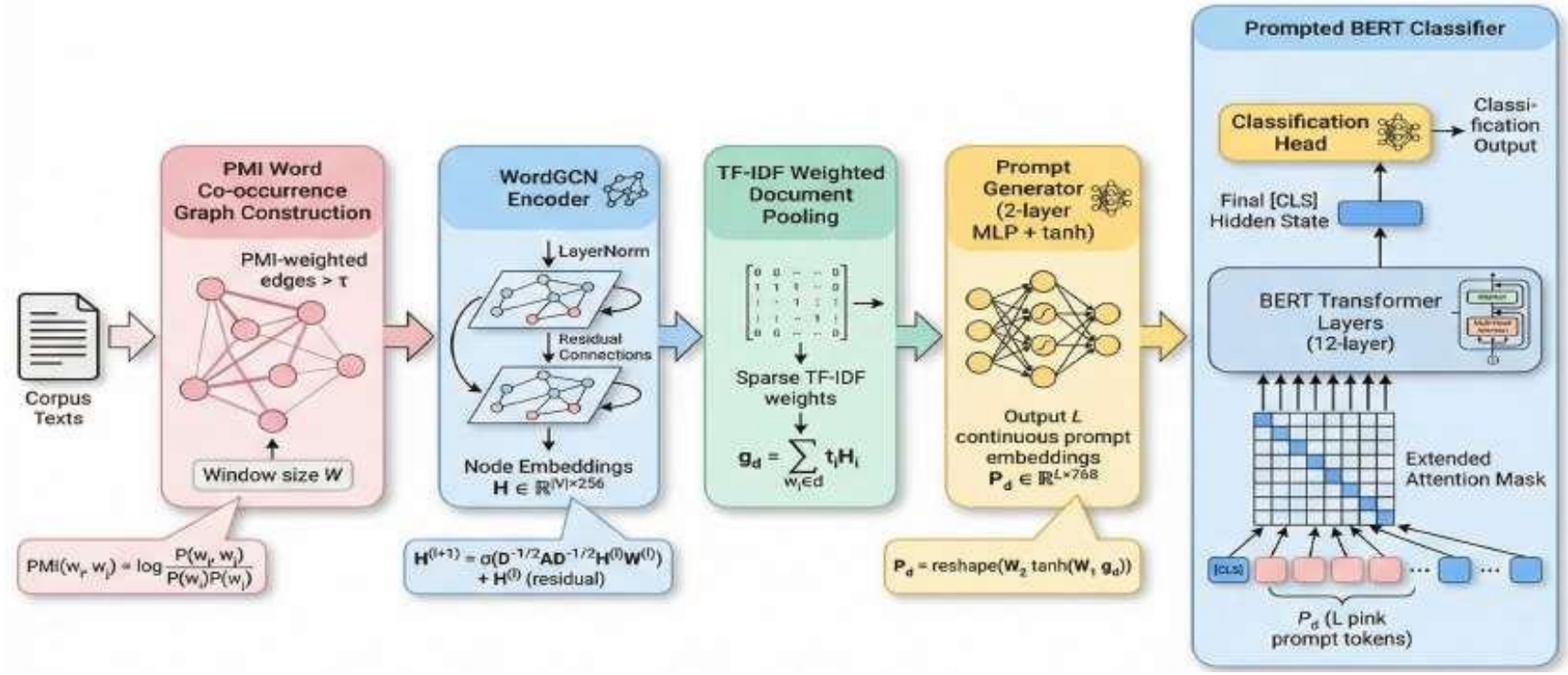


Fig. 2. GraphPrompt converts corpus-level co-occurrence evidence into instance-specific prompt tokens for BERT.

Four-step workflow

1. **PMI graph**
   Build a sparse word graph from unlabeled corpus co-occurrences.
2. **WordGCN**
   Encode structural neighborhoods with residual GCN layers.
3. **TF-IDF pooling**
   Aggregate word nodes into a document-specific graph vector.
4. **Prompted BERT**
   Generate L prompt embeddings and classify from [CLS].

Experimental Setup

Six benchmarks

Dataset	Classes	Documents	Avg. length
MR	2	10,662	19.3
Twitter	2	9,143	11.3
Snippets	8	12,331	18.1
TagMyNews	7	32,600	25.9
Ohsumed	23	7,400	12.4
AG News	4	127,600	32.5

Domains span movie reviews, user posts, web snippets, news, and biomedical text.

Protocol
20 train + 20 validation examples per class; all remaining instances for test; five splits (seeds 42–46).

Metrics
Accuracy and Macro-F1, reported as mean ± standard deviation.

Default model
bert-base-uncased; prompt length L=5; max length 64; fine-tune the last Transformer layer.

Optimization
AdamW, learning rate 2×10^{-4}, batch size 8, dropout 0.3, early stopping patience 20.

Transductive graph packs use raw text only — no test labels are accessed.

Experiments: Main Results I

Method	MR		AG News		Twitter	
	Acc.	F1	Acc.	F1	Acc.	F1
TF-IDF+SVM (1988)*	56.93 ± 1.56	56.68 ± 1.57	77.31 ± 0.83	76.79 ± 0.84	60.47 ± 2.98	60.39 ± 3.00
Graph-only WordGCN *	54.44 ± 0.86	54.14 ± 0.80	71.24 ± 2.43	71.42 ± 2.18	58.50 ± 2.93	58.01 ± 3.05
BERT-only (2019)*	69.83 ± 3.76	69.62 ± 3.81	83.36 ± 2.34	83.25 ± 2.31	97.81 ± 1.42	97.81 ± 1.42
Soft Prompt (2021)*	68.01 ± 5.62	66.87 ± 6.92	83.16 ± 1.73	83.17 ± 1.67	97.36 ± 2.14	97.34 ± 2.15
Token-Based Graphs (2024)†	70.20	70.20	–	–	83.70	83.70
SHINE (2021)†	64.58	63.89	–	–	72.54	72.19
GIFT (2024)†	65.21	65.21	–	–	73.16	73.16
SimSTC (2025)†	66.52	66.36	–	–	73.51	73.49
MI-DELIGHT (2025)†	66.49	66.47	–	–	75.11	75.06
GraphPrompt	**69.99 ± 4.09**	**69.85 ± 4.05**	**84.13 ± 1.30**	**84.17 ± 1.24**	**99.00 ± 0.49**	**99.00 ± 0.49**

Results on MR, AG News, and Twitter under the unified 20-shot protocol

Key findings

GraphPrompt achieves the best results on AG News and Twitter.

Compared with BERT-only:

- **AG News: +0.77 Acc. / +0.92 F1**
- Twitter: +1.19 Acc. / +1.19 F1

On MR, GraphPrompt remains competitive with the strongest reported method.

Experiments: Main Results II

Method	Snippets		TagMyNews		Ohsumed	
	Acc.	F1	Acc.	F1	Acc.	F1
TF-IDF+SVM (1988)⋆	71.11 ± 0.89	69.50 ± 1.06	51.12 ± 2.07	45.18 ± 2.28	25.46 ± 2.62	19.22 ± 1.78
Graph-only WordGCN ⋆	67.30 ± 1.36	67.01 ± 1.42	51.25 ± 2.66	47.85 ± 2.49	26.26 ± 1.68	21.52 ± 0.78
BERT-only (2019)⋆	84.76 ± 0.48	84.00 ± 0.71	78.58 ± 1.44	75.12 ± 1.68	41.67 ± 0.52	32.06 ± 0.55
Soft Prompt (2021)⋆	84.71 ± 0.24	83.64 ± 0.40	78.41 ± 1.29	74.55 ± 1.42	42.27 ± 1.75	32.72 ± 1.68
Token-Based Graphs (2024)†	80.40	79.70	69.90	66.80	–	–
SHINE (2021)†	82.39	81.62	62.50	56.21	45.57	30.98
GIFT (2024)†	83.73	82.35	63.26	56.92	45.62	31.25
SimSTC (2025)†	85.14	83.94	–	–	–	–
MI-DELIGHT (2025)†	**87.90**	**86.84**	69.72	65.94	**48.56**	**33.20**
GraphPrompt	85.06 ± 0.60	84.35 ± 0.55	**79.37 ± 1.41**	**75.99 ± 1.44**	44.35 ± 1.83	34.46 ± 1.15

Results on Snippets, TagMyNews, and Ohsumed under the unified 20-shot protocol

Key findings

GraphPrompt achieves the best results on TagMyNews and the highest Macro-F1 on Ohsumed.

Against Soft Prompt, it improves both Accuracy and Macro-F1 on Snippets, TagMyNews, and Ohsumed.

The gains are clearest on sparse or domain-specific corpora.

Experiments: BERT Tuning Ablation & Prompt Length Ablation

Dataset	Mode	BERT Tune	Acc.	F1
Ohsumed	none	freeze	15.97±1.20	11.85±1.24
	none	last1	41.75±0.34	32.17±0.70
	learnable	freeze	28.92±3.03	19.55±2.22
	learnable	last1	42.80±0.56	33.74±0.51
	graph	freeze	17.75±1.27	15.45±1.06
	graph	last1	44.35±1.83	34.46±1.15
Snippets	none	freeze	69.38±0.41	68.14±0.39
	none	last1	84.75±0.20	84.32±0.58
	learnable	freeze	76.91±1.66	76.07±2.13
	learnable	last1	84.65±0.63	83.88±0.60
	graph	freeze	71.39±2.95	70.99±2.86
	graph	last1	85.06±0.60	84.35±0.55

Key findings

Fine-tuning only the last BERT layer is essential.

GraphPrompt + last1 gives the strongest result on both Ohsumed and Snippets.

Freezing BERT sharply reduces performance, especially on Ohsumed.

Dataset	L	Acc.	F1
MR	5	69.99±4.09	69.85±4.05
	10	66.74±9.19	63.47±11.09
	20	62.31±1.59	59.64±1.38
AG News	5	84.13±1.30	84.17±1.24
	10	83.67±1.84	83.68±1.90
	20	82.48±1.56	82.54±1.50
Twitter	5	99.00±0.49	99.00±0.49
	10	97.96±2.24	97.96±2.24
	20	97.93±1.15	97.93±1.15
Snippets	5	85.06±0.60	84.35±0.55
	10	85.34±0.57	84.41±0.47
	20	85.39±0.61	84.44±0.42
TagMyNews	5	79.37±1.41	75.99±1.44
	10	79.19±0.66	75.90±0.67
	20	79.39±0.46	75.78±0.62
Ohsumed	5	44.35±1.83	34.46±1.15
	10	43.12±0.73	33.75±0.73
	20	44.01±2.61	34.49±0.89

Key findings

Performance is broadly stable for L = 5, 10, and 20.

L = 5 performs best on MR, AG News, and Twitter.

Longer prompts provide only small, inconsistent changes, so L = 5 is the compact default.

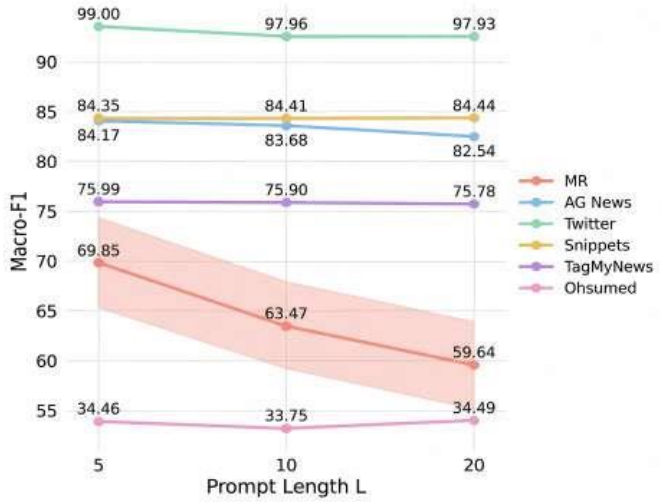
Analysis and Evidence

BERT tuning

Setting	Ohsumed F1	Snippets F1
Graph / freeze	15.45	70.99
Soft / last1	33.74	83.88
Graph / last1	**34.46**	**84.35**

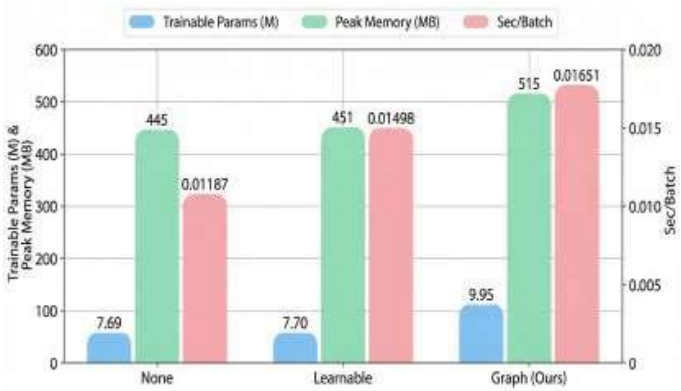
Fine-tuning only the last Transformer layer preserves the best accuracy–efficiency trade-off; freezing BERT is insufficient.

Prompt length sensitivity



Performance is broadly robust for L∈{5,10,20}. The default L=5 offers a concise, efficient choice.

Efficiency



9.95M trainable parameters
515 MB peak memory
0.01651 s per batch

Conclusion and Q&A

Takeaways

GraphPrompt transforms PMI word-graph evidence into instance-specific continuous prompts.
A single-stage BERT classifier gains structural information without document-level graphs or extra contrastive objectives.
Controlled 20-shot experiments show improvements over vanilla soft prompts on all six datasets in both Accuracy and Macro-F1.
Last-layer tuning with L=5 gives a strong quality–efficiency balance with moderate memory and runtime overhead.

Thank you!

Questions are welcome.

Paper & Contact

Corresponding author
Mieradilijiang Maimaiti
miradeljan51@xju.edu.cn