TaCL-CoMoE: Task-adaptive Contrastive Learning with Cooperative Mixture of Experts for Multi-task Social Media Analysis

Anonymous ACL submission

Abstract

Social media has become a crucial platform for information dissemination and opinion expression. The massive and continuous generation of user content has given rise to various natural language processing tasks, such as sentiment analysis and topic classification. However, existing mainstream approaches typically focus on modeling individual tasks in isolation, lacking systematic exploration of collaborative modeling across multiple tasks. This neglects the inherent correlations among social media tasks, thereby limiting the model's ability to fully comprehend and exploit the rich, multi-dimensional semantic information embedded in text. To address this challenge, we propose **Ta**sk-adaptive **C**ontrastive **L**earning with **Co**operative **M**ixture **o**f **E**xperts (**TaCL-CoMoE**), a unified framework for social media multi-task learning. Specifically, we improve the gating mechanism by replacing the traditional softmax routing with sigmoid activation, enabling cooperative selection among multiple experts and mitigating the "expert monopoly" phenomenon. In addition, we introduce a task-adaptive contrastive learning strategy to further enhance the model's ability to capture and distinguish semantic structures across different tasks. Experimental results on multiple public social media datasets demonstrate that TaCL-CoMoE consistently achieves state-of-the-art (SOTA) performance. The code is available at https://anonymous.4open.science/r/TaCL-CoMoE.

1 Introduction

In recent years, social media has become a major platform for information acquisition, opinion expression, and social interaction (Islam, 2025). The continuous and rapid growth of user-generated content has given rise to a variety of natural language processing tasks, including sentiment analysis, topic classification, and misinformation detection (Zhou et al., 2025; Antypas et al., 2024; Wang et al., 2025). Although these tasks differ in objectives and characteristics, they often exhibit underlying semantic correlations, which suggests the potential for joint modeling. However, most existing approaches focus on single-task learning as shown in Figure 1 (a), overlooking the latent relationships and shared knowledge across tasks. Therefore, how to effectively leverage the semantic correlations among tasks for joint multi-task modeling remains an open and important research question.

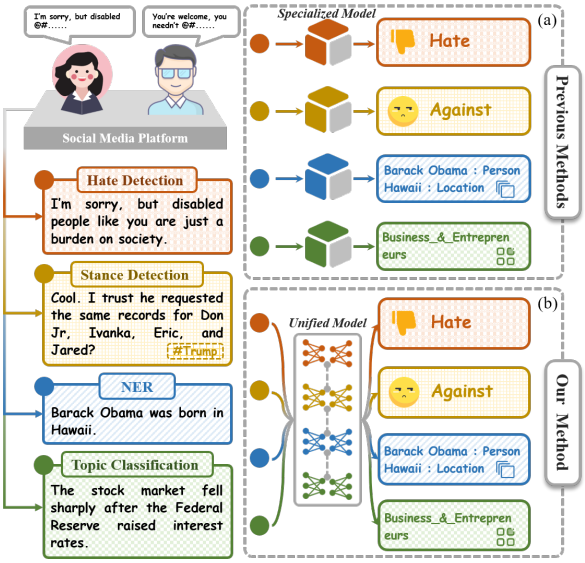


Figure 1: Comparison Between Traditional Methods and TaCL-CoMoE on the Social Media Multi-task Analysis. Unlike traditional approaches that train separate models for each task, TaCL-CoMoE adopts a unified architecture capable of capturing underlying inter-task relationships and enabling knowledge sharing.

To address this challenge, researchers have increasingly turned their attention to large language models (LLMs), which have demonstrated remarkable performance across a wide range of natural language processing tasks due to their powerful semantic understanding and generative capabilities (Touvron et al., 2023; Hurst et al., 2024; Zeng

et al., 2024). However, as model sizes continue to grow, the computational and storage costs associated with fine-tuning have escalated significantly, posing serious limitations to the efficient deployment of such models in real-world applications. Low-Rank Adaptation (LoRA), a representative parameter-efficient fine-tuning approach, mitigates these issues by freezing the backbone model and only training a small number of low-rank parameters (Hu et al., 2022). This strategy substantially reduces training costs while maintaining competitive performance. Nevertheless, in practical multi-task scenarios, methods like LoRA often require frequent switching between specialized models and lack flexible gating and routing mechanisms, making it challenging to efficiently adapt to the dynamic variability across tasks (Zhao et al., 2024; Xia et al., 2024).

To address the aforementioned issue, Liu et al. (2024) and Dou et al. (2024) propose MOELoRA, which integrates the Mixture-of-Experts (MoE) (Shazeer et al., 2017; Fedus et al., 2022) architecture with the LoRA efficient fine-tuning technique. A flexible gating network in MOE is leveraged to dynamically select the most appropriate expert subnetwork across different tasks or input samples, effectively alleviating the problem of frequent switching between specialized models in multi-task scenarios. However, existing MoE primarily rely on softmax-based gating mechanisms, which tend to make overly confident expert selections, i.e., "expert monopoly" phenomenon (Nguyen et al., 2024).

In this paper, we propose a unified multi-task learning framework for social media scenarios, called **Ta**sk-adaptive **C**ontrastive **L**earning with **Co**operative **M**ixture **o**f **E**xperts (**TaCL-CoMoE**), aiming to enhance both collaborative modeling capabilities and semantic discrimination in multi-task learning. Firstly, we introduce a sigmoid-based cooperative expert routing mechanism that allows multiple experts to be activated simultaneously, alleviating the common issue of "expert monopoly" observed in softmax gating, and promoting balanced cooperation among experts. Secondly, we integrate a contrastive learning mechanism into TaCL-CoMoE and flexibly adopt supervised or unsupervised strategies depending on the availability of task label information, guiding the model to learn task-specific semantic representations and improving its fine-grained semantic discrimination. Finally, we conduct extensive experiments on multiple public social media datasets to validate the effectiveness of the proposed approach. The results demonstrate that TaCL-CoMoE outperforms existing SOTA methods across all tasks. The main contributions of this paper are as follows:

- We propose TaCL-CoMoE, a unified multi-task modeling framework for social media.
- We introduce a sigmoid-based cooperative expert routing mechanism to alleviate expert selection polarization and task interference issues commonly found in traditional MoE architecture.
- We introduce a task-aware contrastive learning strategy that flexibly selects supervision methods based on the nature of the task's label information, enhancing the model's ability to capture semantic structures across different tasks.
- Extensive experimental results demonstrate that our method outperforms existing SOTA approaches.

2 Related Work

The related work is provided in Appendix A

3 Method

In this section, we elaborate on the methodological details of TaCL-CoMoE. The overall architecture of TaCL-CoMoE is illustrated in Figure 2.

3.1 Task Definition

This study aims to address various text analysis tasks in the context of social media from a multi-task learning perspective. Specifically, it focuses on four representative tasks: Stance Detection, Hate Detection, Named Entity Recognition (NER), and Topic Classification. A unified language modeling paradigm is adopted by formulating all tasks as text-to-text generation problems. Specifically, given a target task instruction $\mu \in G$ and a social media post $X = \{x_1, \ldots, x_i, \ldots, x_n\}$, where G is a set of multi-task instructions, x_i represents the i^{th} token in the sequence, the model is required to learn a conditional generation function $F_t : X \rightarrow \hat{Y}_t$, where $\hat{Y}_t$ represents the predicted output. The set G is shown in Appendix C.

3.2 Cooperative Mixture of Experts

LoRA has demonstrated remarkable advantages in the parameter-efficient fine-tuning of LLMs. The

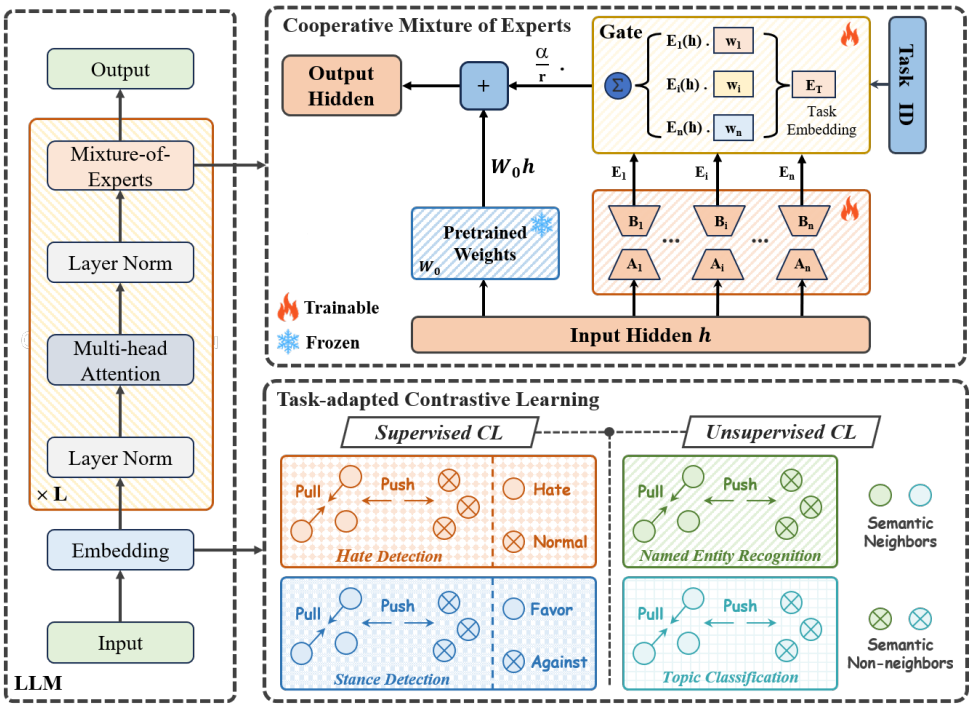


Figure 2: Illustration of the overall framework of TaCL-CoMoE, which consists of two essential components: Cooperative Mixture of Experts, and Task-adaptive Contrastive Learning.

core idea is to replace full-parameter updates with the learning of a pair of low-rank matrices, thereby significantly reducing the number of trainable parameters and improving convergence efficiency. Specifically, LoRA represents the update to linear weights as $W_0 + \Delta W = W_0 + BA$. Here, $W_0 \in \mathbb{R}^{d_{in} \times d_{out}}$ denotes the fixed weight matrix from the pre-trained LLMs, while $B \in \mathbb{R}^{d_{in} \times r}$ and $A \in \mathbb{R}^{r \times d_{out}}$ are the trainable low-rank matrices. The forward pass is defined as:

$$h = W_0 x + \frac{\alpha}{r} \cdot BAx \tag{1}$$

Here, α is a scaling factor that controls the influence of the low-rank update on the output, and r is a rank hyperparameter that determines the number of trainable parameters. The input vector x has a dimensionality of d_{in}, and the output vector h has a dimensionality of d_{out}.

Existing research demonstrates that the widely used softmax gating mechanism in MoE models may induce unnecessary competition among experts, leading to issues such as expert monopolization and representation collapse (Nguyen et al., 2024; Csordás et al., 2023). To address this problem, we introduce a cooperative MoE layer to replace each dense layer in LLM.

In the cooperative MoE layer, each expert $\{E_i\}_{i=1}^{N}$ consists of a pair of low-rank matrices $\mathbf{B}_i \in \mathbb{R}^{d_{in} \times \frac{r}{N}}$ and $\mathbf{A}_i \in \mathbb{R}^{\frac{r}{N} \times d_{out}}$, where N denotes the number of experts. Collectively, these experts form a trainable module for modeling the parameter update ΔW. For each task $\mathcal{T}_j \in \mathrm{T}$ we assign a unique task identifier, which is mapped to a task vector $\mathbf{e}_j \in \mathbb{R}^{d_T}$ via a task embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathrm{T}| \times d_T}$, where d_T denotes the task embedding dimension and T denotes the set of all tasks. The task vector is then fed into a task-aware gating network, which generates task-specific expert weights through a linear transformation followed by a sigmoid activation:

$$\omega_j = \mathrm{Sigmoid}(\mathbf{W}_T \mathbf{e}_j) \tag{2}$$

where, $\mathbf{W}_T \in \mathbb{R}^{N \times d_T}$ is a learnable gating matrix, and $\omega_j \in \mathbb{R}^N$ denotes the contribution of each expert to task $\mathcal{T}_j$. Based on this structure, the forward process of the cooperative MoE layer for task $\mathcal{T}_j$ can be represented as:

$$\begin{aligned} \mathbf{h}_j &= \mathbf{W}_0 \mathbf{x}_j + \frac{\alpha}{r} \cdot \sum_{i=1}^{N} \omega_{ji} \cdot E_i(\mathbf{x}_j) \\ &= \mathbf{W}_0 \mathbf{x}_j + \frac{\alpha}{r} \cdot \sum_{i=1}^{N} \omega_{ji} \cdot \mathbf{B}_i \mathbf{A}_i \mathbf{x}_j \end{aligned} \tag{3}$$

where $\mathbf{h}_j$ and $\mathbf{x}_j$ represent the input and output of intermediate LLM layers for samples from $\mathcal{T}_j$.

3.3 Task-adaptive Contrastive Learning

In the task-adaptive contrastive learning module, we dynamically select between unsupervised and supervised contrastive strategies based on the nature of the task's label information. All tasks share a unified contrastive loss framework.

3.3.1 Contrastive Loss

For each sample in a batch, two augmented views are generated, resulting in a total of $2N$ representation vectors $\{z_k\}_{k=1}^{2N}$, where each $z_k \in \mathbb{R}^d$ is a normalized embedding. In unsupervised contrastive learning, if z_i and z_j are two different augmented views derived from the same original sample, they are regarded as a positive pair, while all other samples are treated as negative examples. In supervised contrastive learning, class labels are utilized to construct positive and negative sample pairs. Samples belonging to the same class are regarded as positive samples, while those from different classes are treated as negative samples. The unsupervised and supervised contrastive losses are defined as follows:

$$\mathcal{L}_x = \frac{1}{|I|} \sum_{i \in I} l_x(z_i), x \in \{\text{unsup}, \text{sup}\} \quad (4)$$

$$l_{\text{unsup}}(z_i) = -\log \frac{\sum_{j \in I} \mathbb{1}_{(i,j)} \cdot \exp \mathcal{S}(i,j)}{\sum_{k \in I} \mathbb{1}_{(i,k)} \cdot \exp \mathcal{S}(i,k)} \quad (5)$$

$$l_{\text{sup}}(z_i) = -\log \frac{\sum_{j \in I} \mathbb{1}'_{(i,j)} \cdot \exp \mathcal{S}(i,j)}{\sum_{k \in I} \mathbb{1}'_{(i,k)} \cdot \exp \mathcal{S}(i,k)} \quad (6)$$

Here, $\mathbb{1}$ and $\mathbb{1}'$ denote the indicator function, which $\mathbb{1}$ returns 1 if the two input elements originate from the same view, and $\mathbb{1}'$ returns 1 only if the two input elements belong to the same label class. The function $\mathcal{S}(i,j) = \text{sim}(z_i, z_j)/\tau$ computes the cosine similarity and scales it under the control of a temperature parameter τ. I denotes the index set of all augmented samples in the batch.

3.3.2 Multi-task Contrastive Loss Integration

We assign an independent contrastive loss balancing parameter λ_t to each sub-task $t \in \mathcal{T}$. The final contrastive loss is formulated as a weighted sum of the individual task-specific losses:

$$\mathcal{L}_{\text{contrastive}} = \sum_{t \in \mathcal{T}} \lambda_t \cdot \mathcal{L}^{(t)}_{\text{sup/unsup}} \quad (7)$$

The final training objective combines the contrastive learning loss with the primary task loss from multi-task fine-tuning:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{contrastive}} \quad (8)$$

where $\mathcal{L}_{\text{main}}$ denotes the sum of the primary loss functions across all tasks, formulated as a cross-entropy loss.

4 Experimental Settings

4.1 Dataset

Four representative datasets from social media are utilized, each corresponding to a distinct task: Hate Detection, Stance Detection, NER, and Topic Classification. The **IHC** dataset (ElSherief et al., 2021) consists of annotated tweets labeled for hate speech, distinguishing between hateful and normal content. The **PStance** dataset (Li et al., 2021) contains stance annotations toward political figures, with labels of favor or against. The **TweetNER7** dataset (Ushio et al., 2022) includes tweets annotated with seven types of named entities, covering categories such as person, corporation, location, etc. The **Tweet Topic Multi** dataset (Antypas et al., 2022) comprises multi-labeled tweets annotated with various topics, including family, gaming, sports, etc. More details are given in Appendix B

4.2 Baselines

LLMs-based (Zero-shot) To evaluate the zero-shot performance of LLMs in multi-task social media scenarios, we conducted zero-shot experiments on six LLMs: GLM-4-32B (GLMTeam et al., 2024), DeepSeek-V3 (Guo et al., 2025), InternLM2.5-20B-chat (Zang et al., 2025), Qwen2.5-72B-Instruct (Hui et al., 2024), GPT-3.5-turbo, and GPT-4o (Hurst et al., 2024).

LLMs-based (Fine-tuning) To investigate the fine-tuning capabilities of LLMs in multi-task social media scenarios, we selected four representative open-source models: Baichuan2-7B (Yang et al., 2023)[1], DeepSeek-7B (Bi et al., 2024)[2], Llama2-7B (Touvron et al., 2023)[3], and Qwen2-7B (QwenTeam, 2024)[4]. All fine-tuning baselines adopted LoRA as an efficient parameter-efficient

[1] https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat

[2] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[3] https://huggingface.co/meta-llama/Llama-2-7b

[4] https://huggingface.co/Qwen/Qwen2-7B-Instruct

Methods	Hate Detection	Stance Detection	NER	Topic Classification	Avg
LLMs-based (Zero-shot)					
GLM-4-32B	68.00	82.05	33.11	52.41	58.89
DeepSeek-V3	60.70	83.27	57.43	56.74	64.53
Internlm2.5-20B-chat	62.92	76.60	13.23	41.79	48.63
Qwen2.5-72B-Instruct	62.12	78.62	57.36	53.91	63.00
GPT-3.5-turbo	66.85	79.35	41.80	55.89	60.97
GPT-4o	70.83	82.98	52.85	55.16	65.45
LLMs-based (Fine-tuning)					
Baichuan2-7B	74.80	84.12	73.27	62.99	73.79
DeepSeek-7B	74.28	84.55	<u>73.52</u>	62.21	73.64
Llama2-7B	76.72	84.23	72.66	62.60	74.05
Qwen2-7B	77.15	84.25	72.81	<u>63.24</u>	<u>74.36</u>
LLMs-based (KTO)					
Baichuan2-7B	67.48	83.54	51.59	54.91	64.38
DeepSeek-7B	65.87	84.26	66.00	58.40	68.63
Llama2-7B	63.52	<u>84.94</u>	71.07	57.89	69.35
Qwen2-7B	63.76	84.53	70.92	60.05	69.81
SOTA	<u>78.19</u>	83.43	63.10	62.02	71.68
TaCL-CoMoE (Ours)	**78.21**	**85.90**	**78.04**	**64.37**	**76.63**

Table 1: The overall results(%) of the competing baselines and TaCL-CoMoE on the social media multi-task datasets. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>. The results of all LLM-based methods are derived from experiments conducted using self-constructed instruction data.

tuning strategy.

LLMs-based (KTO) Knowledge Transfer Optimization (KTO) (Ethayarajh et al., 2024) is a reinforcement learning method based on human preference, which aims to optimize the behavior of language models through human feedback. All baseline models are trained with the KTO strategy.

SOTA To validate the superiority of our approach, we conducted a comparative study against current SOTA methods on four selected social media tasks. Hate Speech Detection: Hoang et al. (2024) propose ToXCL, a unified framework for detecting and explaining implicit harmful speech. ToXCL integrates a target group generator, an encoder-decoder architecture, and a teacher classifier, leveraging knowledge distillation to enhance detection performance. Stance Detection: Lan et al. (2024) introduce COLA, a three-stage framework composed of collaborative large language model agents. The framework includes multidimensional text analysis, reasoning-enhanced debate, and stance inference stages. NER and Topic Classification: Due to inconsistent dataset splits in prior work, fair comparisons are infeasible. Therefore, we follow the original dataset splits for our experiments and report the best performance reported in the respective papers as baselines (Ushio et al., 2022; Antypas et al., 2022).

Evaluation Metrics All tasks are evaluated using the Macro F1 score as a unified metric.

4.3 Implementation Details

ChatGLM3-6B (GLMTeam et al., 2024)[5] is employed as the base model in TaCL-CoMoE, which is built upon the transformer architecture and consists of 28 transformer layers, exhibiting strong capabilities in language understanding and generation. The model is fine-tuned using LoRA for parameter-efficient adaptation, with a rank of 16 and a dropout rate of $\alpha = 0.1$. The number of experts is set to 8. All training stages use the AdamW optimizer (Loshchilov and Hutter, 2017) with an initial learning rate of 2e-4. The maximum input and output lengths are set to 2048 and 512, respectively. All experiments are conducted on two NVIDIA RTX 4090 GPUs, each with 24GB of memory. For all experiments, we report the results as the average over three runs with different random seeds.

[5] https://huggingface.co/THUDM/chatglm3-6b

5 Results and Discussions

5.1 Main Results

The performance of TaCL-CoMoE compared to all baselines is presented in Table 1. The detailed analyses are as follows.

Hate Detection TaCL-CoMoE achieves an F1 score of 78.21%, slightly outperforming the current SOTA result of 78.19%. Compared to GPT-4o and the best result under the KTO paradigm, TaCL-CoMoE yields absolute improvements of 7.38% and 10.73%, respectively. The result suggests that both zero-shot and KTO methods exhibit certain limitations in recognizing offensive intent within social contexts.

Stance Detection TaCL-CoMoE achieves an F1 score of 85.90%, outperforming GPT-4o, the best KTO-based model, and the best supervised fine-tuned model by 2.92%, 0.96%, and 1.35% , respectively. This result demonstrates the effectiveness of TaCL-CoMoE in Stance Detection.

NER TaCL-CoMoE achieves an F1 score of 78.04%, substantially surpassing existing best-performing methods. Compared to the current SOTA, the best supervised fine-tuned model, and the best KTO-based approach, it yields improvements of 14.94%, 4.52%, and 6.97%, respectively. The generally poor performance of zero-shot methods indicates that pretrained language models exhibit clear limitations in structured information extraction tasks.

Topic Classification TaCL-CoMoE achieves an F1 score of 64.37%, outperforming the current SOTA by 2.35 %. Compared to GPT-4o, the best KTO-based model, and the best supervised fine-tuned model, TaCL-CoMoE yields improvements of 9.21%, 4.32%, and 1.13%, respectively.

Overall, TaCL-CoMoE achieves SOTA performance across four social media tasks, demonstrating notable advantages particularly in NER and Stance Detection. These results validate the effectiveness of TaCL-CoMoE in multi-task modeling within the context of social media.

5.2 Ablation Study

In this section, we perform ablation studies to analyze the effects of critical modules in our TaCL-CoMoE, detailed in Table 2.

Impact of the MoE Architecture To evaluate the impact of the MoE architecture on model performance, we remove the MoE module and directly fine-tune the base model, ChatGLM3-6B, on the four tasks (θ). Experimental results show that removing the MoE leads to performance degradation across all four tasks, with the most notable drop of 4.38% observed in the NER task.

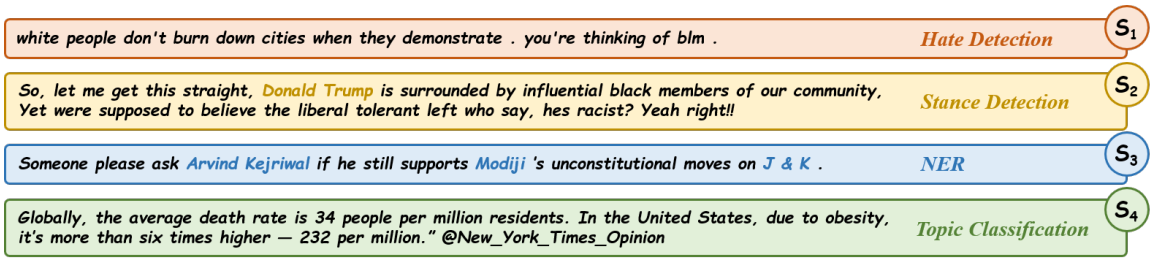
Impact of Contrastive Learning To investigate the impact of contrastive learning on model performance, we remove the $\mathcal{L}_{\text{con}}$ (ρ) and replace the task-adaptive contrastive learning (ϕ) with a unified unsupervised contrastive learning approach. Experimental results show that both modifications lead to performance degradation, with average F1 scores decreasing by 1.86% and 0.69%, respectively. These findings demonstrate the effectiveness of task-adaptive contrastive learning.

Methods	Hate	Stance	NER	Topic
TaCL-CoMoE	78.21	85.90	78.04	64.37
(θ) w/o MoE	$76.18_{\downarrow 2.03}$	$84.49_{\downarrow 1.41}$	$73.66_{\downarrow 4.38}$	$62.60_{\downarrow 1.77}$
(ρ) w/o $\mathcal{L}_{\text{con}}$	$76.37_{\downarrow 1.84}$	$84.87_{\downarrow 1.03}$	$75.38_{\downarrow 2.66}$	$62.47_{\downarrow 1.90}$
(ϕ) w CL_{unsup}	$77.19_{\downarrow 1.02}$	$84.89_{\downarrow 1.01}$	$77.58_{\downarrow 0.46}$	$64.12_{\downarrow 0.25}$
(κ) w Softmax	$77.04_{\downarrow 1.17}$	$84.55_{\downarrow 1.35}$	$76.95_{\downarrow 1.09}$	$63.20_{\downarrow 1.17}$
(Ω) w Random	$76.15_{\downarrow 2.06}$	$84.10_{\downarrow 1.80}$	$74.68_{\downarrow 3.36}$	$62.53_{\downarrow 1.84}$
(η) w/o N.+T.	$77.50_{\downarrow 0.71}$	$84.80_{\downarrow 1.10}$	/	/
(ζ) w/o H.+S.	/	$77.02_{\downarrow 1.02}$	$63.74_{\downarrow 0.63}$	/
(ψ) w/o MTL	$75.48_{\downarrow 2.73}$	$84.27_{\downarrow 1.63}$	$72.61_{\downarrow 5.43}$	$61.02_{\downarrow 3.35}$

Table 2: The experimental results(%) of the ablation study for TaCL-CoMoE (F1 score).

Impact of the Gating Mechanism To evaluate the impact of the gating mechanism design on model performance, we conduct comparative experiments by replacing the Sigmoid gating function with Softmax (κ) and random (Ω) gating mechanisms, respectively. The experimental results show that both alternative designs lead to a decline in overall performance, with the average F1 score dropping by 1.2% and 2.27%, respectively.

Impact of Task Interactions To investigate the influence of different tasks on each other, we designed three task ablation settings: (1) removing NER and Topic Classification while training the model only on Hate Detection and Stance Detection (η); (2) removing Hate Detection and Stance Detection while retaining only NER and Topic Classification (ζ). (3) removing the multi-task learning (MTL) mechanism and fine-tuning the base model ChatGLM3-6B separately on each of the four tasks (ψ). Experimental results show that all three ablation settings lead to a degradation in overall model performance, which verifies the importance of cross-task joint training.



ID	TaCL-CoMoE	w/o MoE	Ground Truth
S_1	Hate ✓	Normal ✗	Hate
S_2	Favor ✓	Against ✗	Favor
S_3	Arvind Kejriwa : Person Modiji : Person J & K : Location ✓	Arvind Kejriwa : Person J & K : Location ✗	Arvind Kejriwa : Person Modiji : Person J & K : Location
S_4	News_&_Social_Concern Fitness_&_Health ✓	News_&_Social_Concern ✗	News_&_Social_Concern Fitness_&_Health

Figure 3: Case Study of TaCL-CoMoE and w/o MoE on Social Media Multi-task Datasets

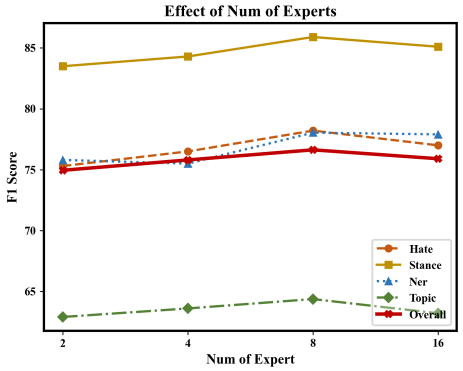


Figure 4: The Impact of the Number of Experts

5.3 Case Study

To better understand the impact of the MoE module on multi-task learning, we selected four representative examples from four distinct tasks. Figure 3 presents a comparative analysis of the performance of TaCL-CoMoE and w/o MoE on these examples.

In both the hate speech detection and stance detection tasks, TaCL-CoMoE accurately captures the emotional tone and stance expressed in the inputs, successfully identifying hateful content and correctly determining the supportive stance. In contrast, w/o MoE fails in both cases, misclassifying the inputs. This suggests that TaCL-CoMoE exhibits a stronger capacity for fine-grained opinion understanding.

For the NER task, the input text includes three entities: two person names and one location. TaCL-CoMoE successfully identifies and categorizes all entities, whereas w/o MoE identifies only a subset, missing a key person name. A similar pattern is observed in the topic classification task, where w/o MoE identifies only part of the topic information, while TaCL-CoMoE correctly captures the complete set of topic categories.

5.4 The Impact of the Number of Experts

Figure 4 illustrates the impact of varying the number of experts on the performance across different tasks. As the number of experts increases from 2 to 8, the F1 scores generally exhibit an upward trend, particularly in Hate Detection and Stance Detection, indicating that appropriately increasing the number of experts can effectively enhance the model's performance in multi-task learning. However, when the number of experts further increases to 16, the performance of all tasks declines to varying degrees. This phenomenon suggests that more experts do not necessarily lead to better results. An excessive number of experts may introduce redundancy or noise, thereby undermining the model's performance.

5.5 Feature Representation Visualization

To more intuitively illustrate the role of contrastive learning, we visualize the learned representations on four tasks using dimensionality reduction, as shown in Figure 5. The figure presents the distribution of samples in the feature space at both the Initial State and Trained State.

For the hate speech detection and stance detection tasks, we adopt supervised contrastive learning. As observed, at the initial state, samples from different classes are mixed and poorly separated, with

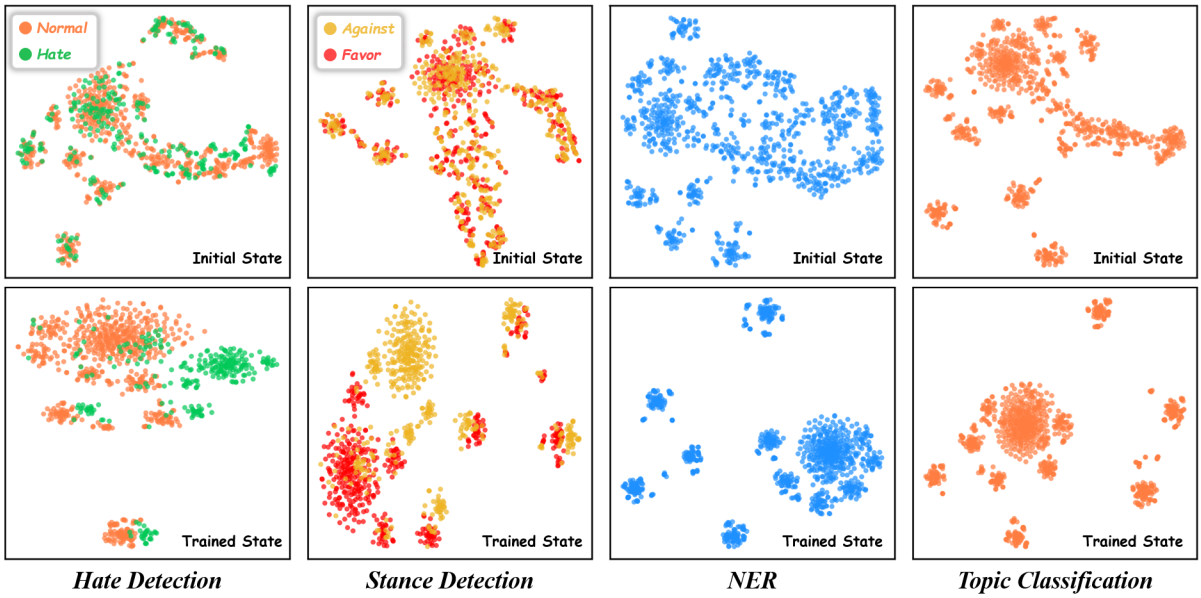


Figure 5: Visualization of Feature Representations on Social Media Multi-task Datasets

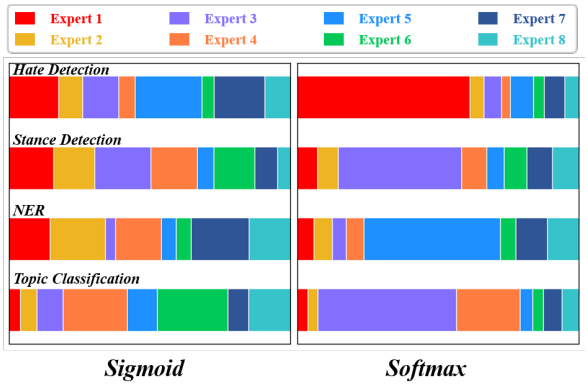


Figure 6: The Visualization of Expert Weights for Various Tasks. In each task, the length of the bar in different colors represents the weights for the corresponding expert.

fuzzy class boundaries. After training, the samples exhibit a clearer clustering structure, with improved intra-class compactness and increased inter-class separation.

In the named entity recognition and topic classification tasks, we adopt unsupervised contrastive learning. Initially, the sample distribution appears scattered and lacks discernible semantic structure. After training, the samples progressively form dense clusters in the feature space, revealing more coherent and semantically meaningful groupings.

Overall, both supervised and unsupervised contrastive learning effectively facilitate the emergence of semantic structures in the feature space across different tasks, enabling the model to learn more organized and discriminative semantic representations.

5.6 Expert Weights Visualization

Figure 6 presents the visualization of expert weights across four tasks under two gating mechanisms: Sigmoid and Softmax. It is observed that the Softmax gating tends to concentrate weights on a small subset of experts, exhibiting the issue of "expert monopoly". In contrast, the Sigmoid gating more evenly activates multiple experts in each task, resulting in a more balanced distribution of expert weights. These results suggest that the Sigmoid gating, by independently computing the activation probability of each expert, effectively mitigates the over-reliance on a few experts seen in Softmax gating and promotes better collaboration among experts.

6 Conclusion

In this paper, we propose TaCL-CoMoE, a multi-task learning framework for the social media domain that incorporates task-adaptive contrastive learning into an MoE architecture. To mitigate the expert domination issue inherent in traditional MoE models, we design a sigmoid-based expert routing mechanism that facilitates cooperative expert selection and reduces task interference. Experimental results and analyses demonstrate the effectiveness of the proposed TaCL-CoMoE.

7 Limitations

Despite achieving state-of-the-art results on multiple social media tasks, the proposed TaCL-CoMoE still has certain limitations. Firstly, due to constraints in computational resources and time, experiments are conducted on only four social media tasks. Future work will aim to extend the evaluation to a broader range of tasks to further verify the model's generalizability and effectiveness across diverse social media scenarios. Secondly, the proposed task-adaptive contrastive learning relies on task labels or semantic similarity to construct positive and negative sample pairs, which introduces additional training overhead to some extent.

8 Ethical Considerations

For the Hate Detection task, the examples provided in this paper are solely for research purposes and do not reflect the authors' personal values or viewpoints. The goal of this task is to identify and prevent the spread of harmful content on social media, thereby fostering a healthy and positive online environment. All data used in this study are derived from publicly available datasets, with no additional data collection, annotation, or external dissemination involved.

References

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, and Jose Camacho-Collados. 2024. Multilingual topic classification in x: Dataset and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20136–20152.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vítor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. 2023. Approximating two-layer feedforward networks for efficient transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 674–692.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, and 1 others. 2024. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, pages 12634–12651. PMLR.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

GLMTeam, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Nhat Hoang, Do Long, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. Toxcl: A unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1

others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tunazzina Islam. 2025. Understanding microtargeting pattern on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29269–29270.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 891–903.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Bin Liang, Wangda Luo, Xiang Li, Lin Gui, Min Yang, Xiaoqi Yu, and Ruifeng Xu. 2021. Enhancing aspect-based sentiment analysis with supervised contrastive learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3242–3247.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. 2024. Sigmoid gating is more sample efficient than softmax gating in mixture of experts. *Advances in Neural Information Processing Systems*, 37:118357–118388.

QwenTeam. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Asahi Ushio, Francesco Barbieri, Vitor Sousa, Leonardo Neves, and Jose Camacho-Collados. 2022. Named entity recognition in twitter: A dataset and analysis on short-term temporal shifts. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 309–319.

Zhuang Wang, Linna Zhou, Xuekai Chen, Zhili Zhou, and Zhongliang Yang. 2025. Stlc-kg: A social text steganalysis method combining large-scale language models and common-sense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25461–25469.

Yifei Xia, Fangcheng Fu, Wentao Zhang, Jiawei Jiang, and Bin Cui. 2024. Efficient multi-task llm quantization and serving for multiple lora adapters. *Advances in Neural Information Processing Systems*, 37:63686–63714.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, and 1 others. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *CoRR*.

Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. 2024. Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4447–4462.

Ziyi Zhou, Xiaoming Zhang, Shenghan Tan, Litian Zhang, and Chaozhuo Li. 2025. Collaborative evolution: Multi-round learning between large and small language models for emergent fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1210–1218.

A Related Work

Parameter-Efficient Fine-Tuning for LLMs. With the continuous growth in the number of parameters in LLMs, traditional full-parameter fine-tuning faces significant computational and storage overhead, limiting its scalability in real-world applications. To address this challenge, various approaches have been proposed to improve fine-tuning efficiency. Representative techniques include Adapter (Houlsby et al., 2019), BitFit (Zaken et al., 2022), Prefix Tuning (Li and Liang, 2021), and LoRA (Hu et al., 2022). Among them, LoRA introduces trainable low-rank matrices into the linear transformations of pretrained models, enabling effective adaptation to downstream tasks without updating the original model parameters. Owing to its simplicity, low resource overhead, and stable performance, LoRA has been widely adopted in practice.

Mixture-of-Experts. The Mixture of Experts (MoE) is an approach that expands model capacity through sparsely activated expert networks, without significantly increasing computational cost (Shazeer et al., 2017; Fedus et al., 2022). Traditional MoE architectures have been widely adopted in both pretrained language models and vision models (Lepikhin et al., 2021; Riquelme et al., 2021). The core idea is to utilize a router network to dynamically assign input data to different expert networks, thereby enabling specialization and collaboration among experts.

Recently, researchers have begun to explore the integration of MoE with Parameter-Efficient Fine-Tuning approaches. For instance, Liu et al. (2024) introduce a multi-expert architecture in which each expert consists of a pair of low-rank matrices. A task-driven gating function generates task-specific parameters, enabling the model to achieve notable performance gains in multi-task medical applications. Similarly, Dou et al. (2024) impose a local balancing constraint to encourage a subset of experts to focus on leveraging world knowledge for downstream tasks, while the remaining experts concentrate on task-specific objectives. This design enhances multi-task performance while preserving essential world knowledge.

Contrastive Learning. Contrastive Learning aims to learn more discriminative feature representations by constructing positive and negative sample pairs, guiding the model to draw semantically similar samples closer while pushing dissimilar ones apart in the embedding space. Khosla et al. (2020) propose Supervised Contrastive Learning, which extends the self-supervised contrastive learning framework by utilizing label information to cluster embeddings of samples from the same class more tightly. This method significantly improves classification accuracy on the ImageNet benchmark and strengthens the model's robustness to noise and hyperparameter sensitivity. Moreover, Liang et al. (2021) introduce contrastive learning into aspect-based sentiment analysis by designing a multi-task framework that jointly optimizes the supervised contrastive objective and the primary task, thereby enhancing the model's capacity to distinguish aspect-specific sentiment features.

B Dataset Statistics

In this section, we present the dataset statistics for social media multitask learning, as shown in Table 3 and Figure 7. Table 3 lists the sizes of the training, validation, and test sets, the average token lengths, and the number of labels for the four tasks. Figure 7 illustrates the label distribution for each task.

Hate Detection This task aims to determine whether a given text contains hate speech. It includes 13,797 training samples, 1,838 validation samples, and 3,912 test samples, with an average token count of 21.77. It comprises two labels: Normal and Hate. The Normal class contains 13,291 samples, while the Hate class contains 6,256 samples.

Stance Detection This task aims to identify the attitude expressed in a text toward a specific target individual. The targets include three former U.S.

Task Type	# Train	# Validation	# Test	# Avg Tokens	# Label Num
Hate Detection	13797	1838	3912	21.77	2
Stance Detection	17191	2174	2176	43.50	2
NER	7111	886	3383	45.71	7
Topic Classification	5005	708	5536	44.54	19

Table 3: Data Statistics of Social Media Multi-task Datasets

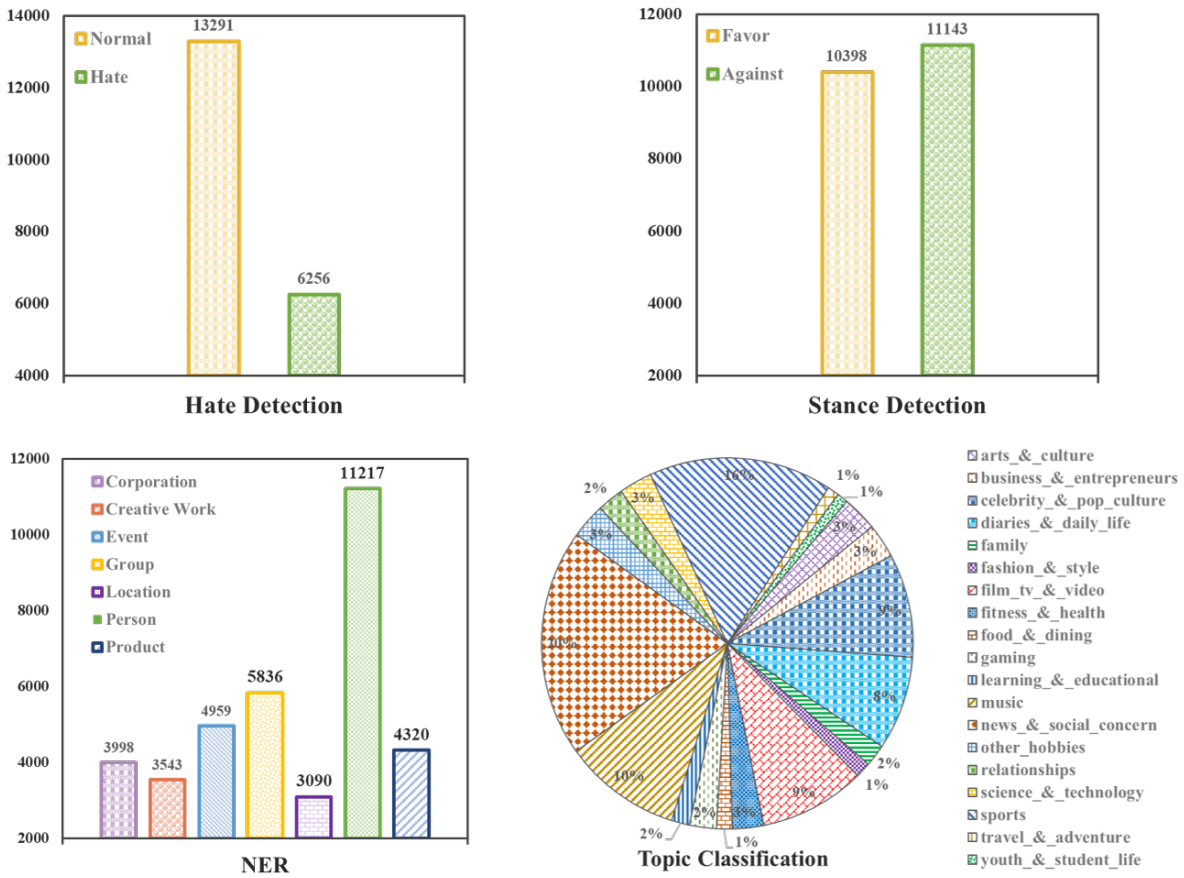


Figure 7: Label Distribution Statistics on Social Media Multi-task Datasets

presidential candidates: Donald Trump, Joe Biden, and Bernie Sanders. The stance labels consist of Favor and Against. The dataset comprises 17,191 training samples, 2,174 validation samples, and 2,176 test samples, with an average token count of 43.50. It includes two labels, with 10,398 samples labeled as Favor and 11,143 samples labeled as Against.

Named Entity Recognition This task aims to identify and classify named entities in text into predefined categories. It includes 7,111 training samples, 886 validation samples, and 3,383 test samples, with an average token count of 45.71. It comprises seven entity categories: Person, Creative Work, Location, Corporation, Group, Product, and Event.

Topic Classification This task is formulated as a multi-label text classification problem, aiming to assign one or more relevant topics to each social media post. The dataset consists of 5,005 training samples, 708 validation samples, and 5,536 test samples. The average number of tokens per sample is 44.54, and the label space includes 19 distinct topic categories.

C Task Instructions

The instruction design for the four tasks is shown in Figure 8.

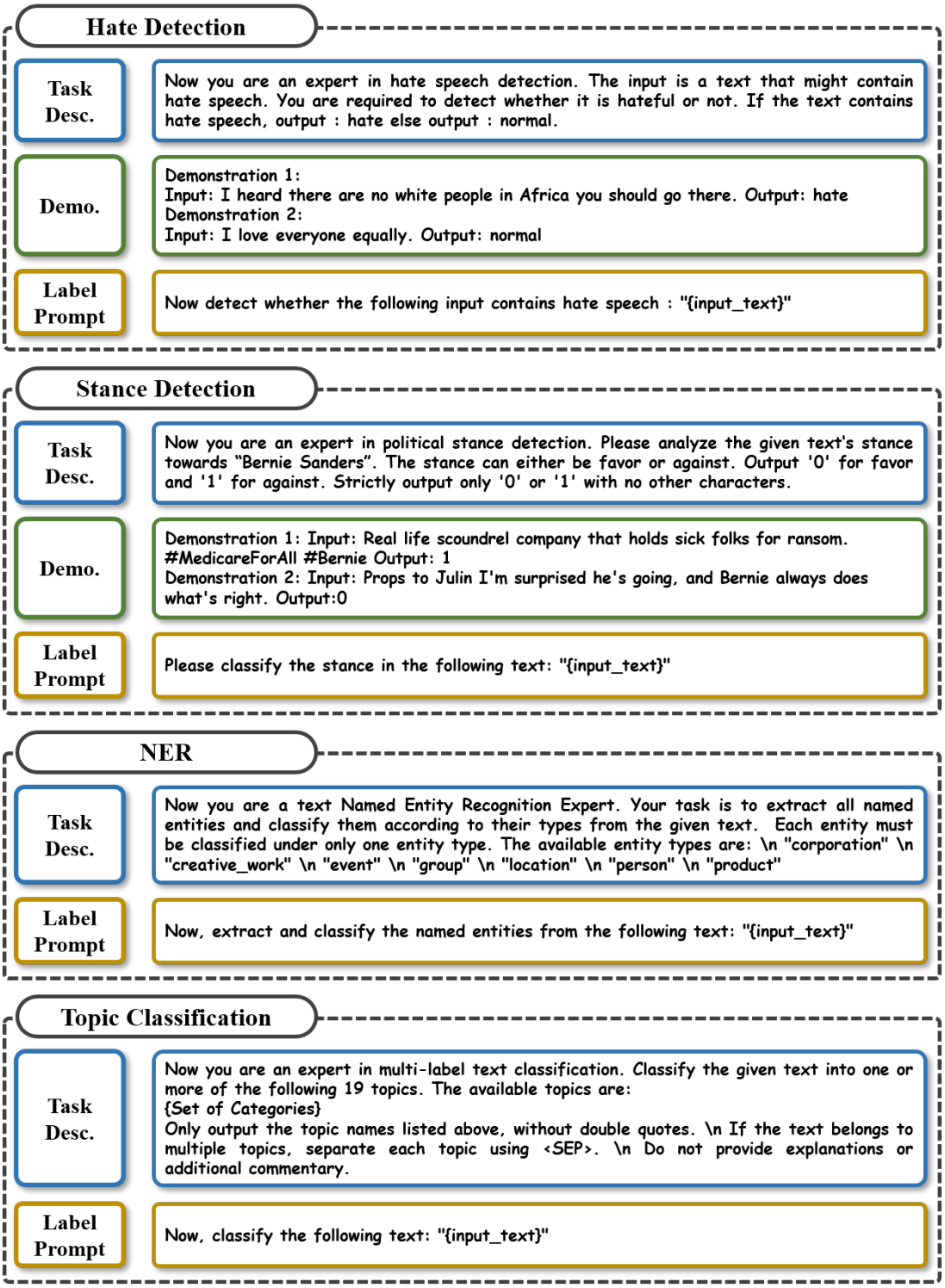


Figure 8: Instruction Design for the Four Tasks