基于知识增强和对抗训练的立场检测方法

黄莉媛[1,2,3]，张瑾[1,2]，靳小龙[1,2]，徐辉[2]，郭嘉丰[1,2]
1. 中国科学院计算技术研究所网络数据科学与技术重点实验室，北京 100190；
2. 智能算法安全全国重点实验室，北京 100190；
3. 中国科学院大学计算机科学与技术学院，北京 101408

摘要
立场检测是自然语言处理领域的一个重要研究方向，旨在判断作者对特定目标所持有的支持、反对或中立态度。针对社交媒体场景中文本语言复杂、领域知识匮乏和模型泛化能力弱的问题，提出了一种基于知识增强和对抗训练的文本立场检测方法KABERT。方法结合生成式与判别式模型的优势，先使用生成式模型从关键词、隐含情感和修辞手法角度提取文本与目标之间的深层语义关系，生成隐含知识；再使用判别式模型作为主干分类网络进行有监督微调，并引入快速梯度法进行对抗训练。在两个标准立场检测数据集SEM16和P-Stance上的实验结果显示，所提方法的Macro-F1分别为68.85%和79.24%，较目前主流方法均有不同程度的提升，证明了所提方法的有效性。

关键词
立场检测；知识增强；对抗训练；预训练模型；大语言模型

中图分类号：TP391.1 文献标志码：A doi：10.11959/j.issn.2096-0271.25261

Stance detection method based on knowledge augmentation and adversarial training

HUANG Liyuan[1,2,3], ZHANG Jin[1,2], JIN Xiaolong[1,2], XU Hui[2], GUO Jiafeng[1,2]
1. Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China；
2. State Key Laboratory of AI Safety, Chinese Academy of Sciences, Beijing 100190, China；
3. School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China

Abstract
Stance detection is a crucial research direction in natural language processing, aiming to determine the author's supportive, opposing, or neutral attitude towards a specific target. To address challenges such as complex linguistic expressions, insufficient domain-specific knowledge, and weak model generalization in social media contexts, this paper proposed a stance detection method named KABERT, based on knowledge augmentation and adversarial training. The method integrated the advantages of generative and discriminative models. A generative model was first used to extract deep semantic relationships between text and the target from the perspectives of keywords, implicit sentiment, and

rhetorical devices, generating implicit knowledge. Then, a discriminative model was used as the classification backbone network for supervised fine-tuning, and a fast gradient method was introduced for adversarial training. Experiments were conducted on two standard stance detection datasets, SEM16 and P-Stance. Results show that KABERT achieves Macro-F1 scores of 68.85% and 79.24% respectively, outperforming current mainstream approaches by varying margins, demonstrating the effectiveness of the proposed approach.

Key words
stance detection, knowledge augmentation, adversarial training, pre-trained models, large language model

0 引言

立场检测是自然语言处理（natural language processing，NLP）领域中的一个重要研究方向，旨在判断作者对特定话题、事件或实体等所表达的支持、反对或中立的态度。随着互联网的迅速发展，社交媒体平台数量的激增彻底改变了现代交流方式，社交媒体平台成为个人获取新闻和信息的主要来源。越来越多的用户在推特（Twitter）和微博等网络社交平台上发布评论来表达自己的观点，形成了海量的用户生成内容（user-generated content，UGC）[1]。对这些数据进行立场检测能够在网络舆情分析、公共政策制定、社会事件监测以及虚假新闻和谣言识别等应用场景中提供技术支持，具有重要研究意义。

然而，在社交媒体环境中，文本内容具有高度的复杂性，语境依赖性强，语义多样且表达方式偏口语化，同时还涉及情感倾向和隐含态度等因素，因此文本与目标之间的立场关系往往是隐含的。此外，在某些专业领域或特定任务中，模型还可能因领域知识匮乏和数据稀缺而受到限制，难以准确建模复杂的语义关系，无法精准捕捉文本的立场。

针对上述问题，本文结合知识增强和对抗训练策略，提出了一种针对社交媒体场景的立场检测方法KABERT（knowledge augmentation and adversarial training on BERT），通过结合生成式和判别式模型的优势，充分挖掘文本中蕴含的深层语义信息以提高模型整体性能。首先，使用生成式模型进行知识增强，从关键词（keyword）、隐含情感（implied emotion）和修辞手法（rhetorical device）等角度提取文本与目标之间的深层语义关系，生成隐含知识的推理分析内容，实现对原始文本数据的扩展增强。其次，采用判别式模型作为主干分类网络，对增强后的数据进行有监督微调，并通过对抗训练增强模型的鲁棒性和泛化能力。

本文的主要贡献如下：

（1）提出一种利用生成式模型驱动的知识增强策略，从多个角度提取目标与文本之间的关系，丰富语义信息，提高模型理解能力；

（2）在基于判别式模型的微调过程中，引入基于快速梯度法的对抗训练策略，提高模型的鲁棒性和泛化能力；

（3）在以上基础上，提出一种结合生成式与判别式模型的立场检测方法KABERT，引入知识增强和对抗训练策略，能够有效建模隐含语义关系并提取立场信息，提高模型在社交媒体语境下的立场检测性能。

1 相关工作

1.1 立场检测方法

早期的立场检测方法主要依赖于特征工程和传统机器学习算法，主要包括支持向量机（support vector machine，SVM）、逻辑回归、决策树和K近邻算法等[1]。随着深度学习技术的发展，长短期记忆（long short term memory，LSTM）网络、卷积神经网络（convolutional neural network，CNN）和基于注意力（Attention）机制的方法逐渐被用于立场检测。例如，Du等[2]提出一种目标特定网络，利用注意力机制提取文本中与目标相关的特定部分，从而提高立场分类性能。白静等[3]结合双向长短期记忆（BiLSTM）网络和卷积神经网络来提取文本语义特征，并使用注意力机制融合两种特征。杨顺成等[4]使用BiLSTM获取句子特征，构建图卷积网络（graph convolutional network，GCN），并通过建立针对话题的注意力机制进行立场分类。

预训练模型的出现为立场检测任务提供了新的思路。以GPT（generative pre-training transformer）[5]和BERT（bidirectional encoder representations from transformer）[6]为代表的预训练模型能够学习到复杂的语言表示和上下文信息，研究者可以通过微调使其适应特定的立场检测任务。例如，Nguyen等[7]提出了一种专门针对推文（Tweet）数据进行优化的预训练模型BERTweet。Li等[8]则在其基础上进一步针对政治领域的社交媒体数据进行训练，使得模型表现更加优异。

近年来，大语言模型（large language model，LLM）在立场检测任务中也展现出了巨大的潜力。目前主要存在两种应用策略：一种是直接利用LLM作为预测器（LLM-as-predictor），即通过零样本或少样本方式，基于自然语言提示实现立场检测[1]。例如，Huang等[9]提出了一种零样本提示方法，通过直接向LLM提供查询指令来实现高效的立场预测；另一种是将LLM作为增强器（LLM-as-enhancer）[1]，如Li等[10]提出的知识增强立场检测方法（knowledge-augmented stance detection，KASD），结合LLM基于思维链（chain-of-thought，CoT）的推理与BERT的微调机制，通过引入外部的背景知识来捕捉更深层次的语义关系，从而进一步提高模型在复杂语境下的推理能力。

1.2 对抗训练方法

对抗训练（adversarial training）是一种用于提升模型泛化能力与鲁棒性的训练策略，其核心思想是在训练过程中动态生成并引入对抗样本，最小化模型在原始样本与对抗样本上的联合损失，使模型在面对正常输入与扰动输入时能保持一致的预测表现，从而让模型学习更泛化的特征表达。常见的对抗训练方法有快速梯度符号法（fast gradient sign method，FGSM）[11]、快速梯度法（fast gradient method，FGM）[12]和投影梯度下降（projected gradient descent，PGD）法[13]。

（1）快速梯度符号法（FGSM）

FGSM[11]是最早提出的一种对抗训练方法，起初应用于图像领域。其核心思想是在输入空间沿损失函数梯度的符号方向施加小幅扰动，从而生成对抗样本，算法的具体步骤如下。

算法1. FGSM

输入： 模型$f(\cdot)$，原始样本(x,y)，扰动大小ε，损失函数L

输出： 对抗样本x_{adv}

1. 计算损失 $L\left(f(x),y\right)$
2. 计算输入x的梯度：$g=\nabla_x L(f(x),y)$
3. 施加微小扰动：$\delta=\varepsilon\cdot\text{sign}(g)$
4. 生成对抗样本：$x_{\text{adv}}=x+\delta$
5. RETURN 对抗样本x_{adv}

其中，$\text{sign}(\cdot)$表示符号函数，用于提取梯度的方向信息。

（2）快速梯度法（FGM）

FGM[12]是FGSM的改进版，它不仅考虑了符号扰动方向，还对扰动幅度在梯度方向进行了归一化处理，使其对模型的干扰更具稳定性，其生成对抗样本的方法为：

$$x_{\text{adv}}=x+\varepsilon\cdot\left(g/\|g\|_2\right) \tag{1}$$

其中，$g=\nabla_x L(f_\theta(x),y)$同样表示损失函数$L$关于输入$x$的梯度，$\varepsilon$表示扰动大小，$\|g\|_2$表示对$g$使用$L_2$范数进行缩放。该方法与FGSM相比，能够更精确地朝着梯度方向移动一定距离，同时还具有计算开销低、实现简单的优势。由于文本会转化为词嵌入向量，扰动的方向性对模型更敏感，因此该方法更适用于NLP领域。

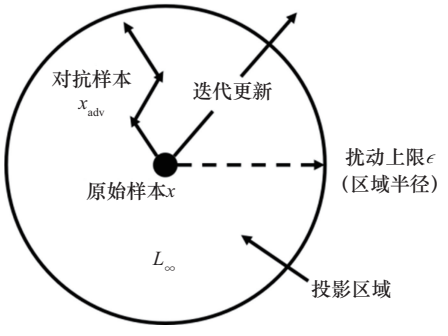

图1 PGD示意图

（3）投影梯度下降（PGD）法

PGD[13]在FGM的基础上引入了多步迭代的扰动生成过程，并且每次迭代都会将扰动投影到指定的半径为ϵ的范围内，这就像是在原始样本周围划一个“安全圈”，以保证扰动不会过大，其示意图如**图1**所示。该方法通过投影和多次迭代操作能够提升对抗样本的攻击强度，但同时也导致了计算成本的增加[14]。

2 研究方法

本文针对社交媒体场景下的立场检测任务，提出了一种基于生成式模型知识增强与判别式模型对抗训练的方法KABERT。方法的整体框架如**图2**所示。

首先，对原始数据集进行预处理，得到目标、文本和标签字段，确保数据质量和数据格式的一致性。然后，使用生成式LLM从多角度提取信息，生成额外的隐含知识，并将生成的隐含知识与原始文本拼接，实现对数据的扩展增强，提高模型对文本立场的理解能力。在模型训练阶段，采用判别式模型作为主干分类网络，并同时引入对抗训练策略进行有监督微调。最后，在测试集上进行推理并全面评估模型的性能。

2.1 知识增强模块

在立场检测任务中，文本长度较短，且上下文缺失，模型仅依赖输入的目标和文本的表面特征难以捕捉更深层次的语义关系。Zhang等[15]的研究工作表明，可以通过引入大模型驱动生成的知识来补充相关背景信息，从而达到提高模型效果的目

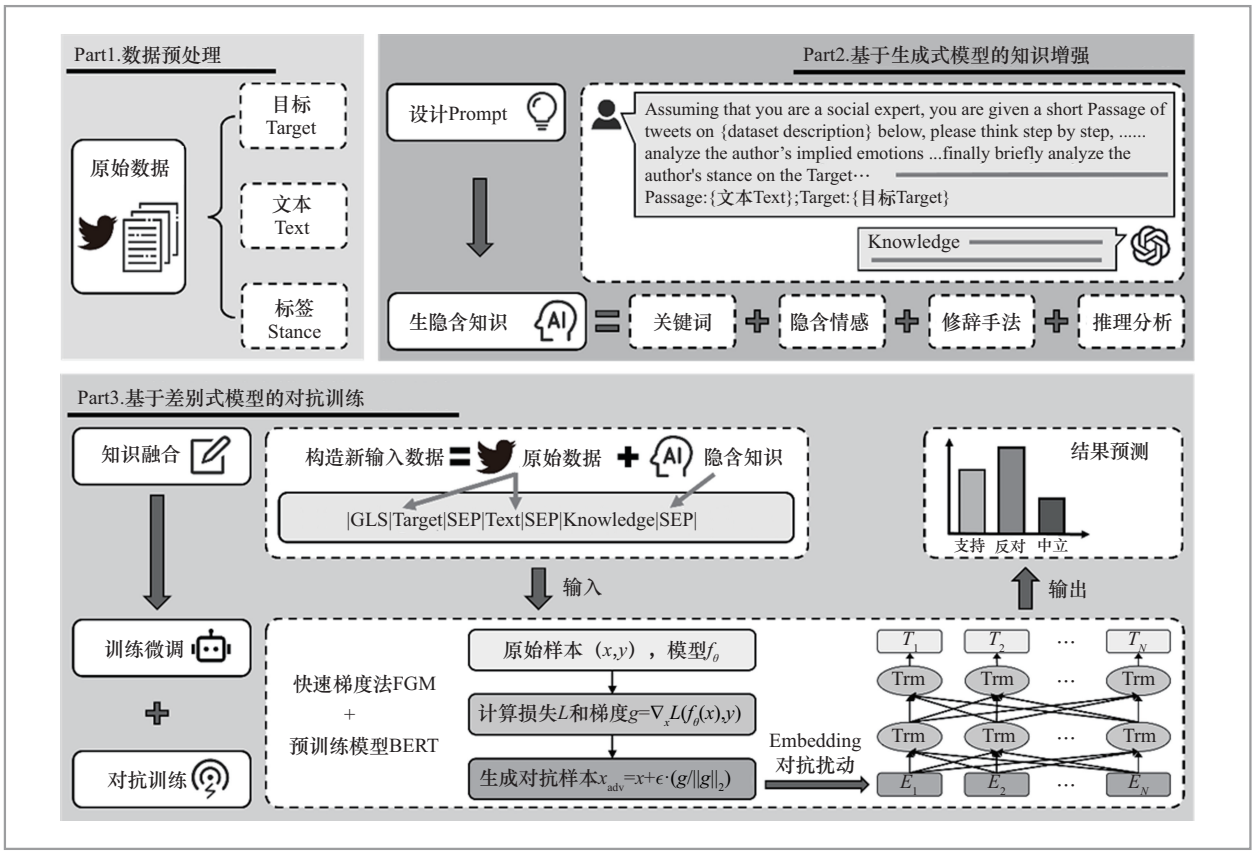

图2 KABERT方法整体框架图

的。因此，本文使用ChatGPT作为知识增强工具，帮助模型捕捉深层语义关系并获取隐含的立场信息。

具体来说，本文对不同的数据集分别设计了部分填充的零样本提示（prompt），并从以下3个方面生成隐含知识。

（1）关键词：文本中的核心词汇能够直接或间接反映作者的观点和立场。例如，高频出现的倾向性词汇或与特定主题相关的词语，能够为立场检测提供重要的语义线索。

（2）隐含情感：某些文本不会直接表达立场，而是通过隐含的情感色彩进行传递。通过情感分析，可以识别文本对目标的潜在态度，如褒贬倾向、情绪强度等。

（3）修辞手法：人们常用一些比喻、讽刺、夸张等修辞手法来表达自己对某个观点的立场，但这不可避免地会在一定程度上增加文本的歧义性。通过识别修辞手法，可以更准确地解析文本的真实意图，避免误判。

为了降低LLM固有的幻觉风险，本文在使用LLM简要分析文本立场时要求避免直接给出结论性的答案，否则可能会导致后续的判别式模型产生错误的预测[15]。此外，在设计prompt时额外添加了对不同数据集的简要描述，以便LLM可以通过检索其内部知识找到更多关于数据集的背景信息，从而充分挖掘文本与目标之间的潜在逻辑关系，补充显式文本未表达的关键信息，完整的prompt设计如**图3**所示。

Prompt设计

Assuming that you are a social expert, you are given a short Passage of {dataset description} below, please think step by step, extract the keywords in the passage, analyze the author's implied emotions, rhetorical devices, etc., finally briefly analyze the author's stance on the Target, paying attention to giving the process of analysis without giving the conclusion.

Passage : {text} Target: {target}

Passage：原始文本数据内容；Target：目标话题；dataset description：数据集描述

图3 prompt设计

2.2 主干分类网络

本文采用基于双向Transformer架构[16]的判别式模型BERT[6]作为主干分类网络，能够同时从左右两个方向学习文本信息，有效捕捉长距离依赖关系，从而更准确地理解词汇语义。BERT的输入以特殊符号“[CLS]”开头，句子间以“[SEP]”分隔，并由词嵌入（token embedding）、段嵌入（segment embedding）和位置嵌入（position embedding）3个向量相加构成，其结构如图4所示。

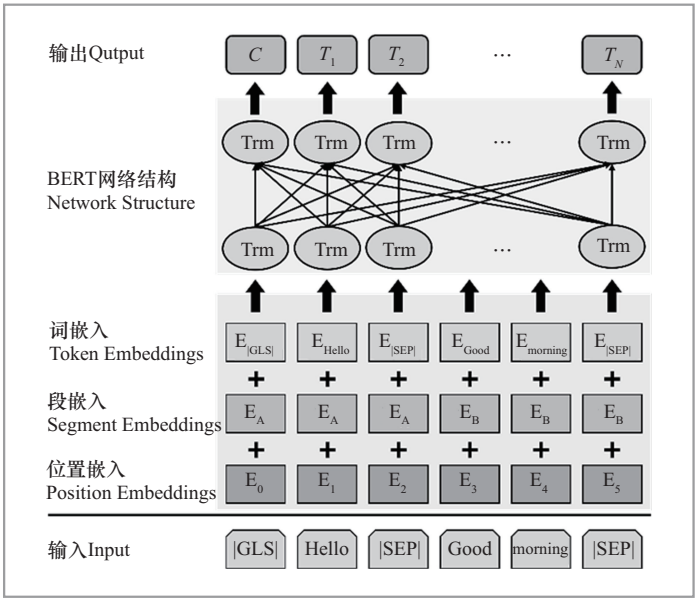

图4 BERT模型结构图

（1）模型输入构建

为了充分利用LLM生成的外部知识，模型将原始样本中的目标（target）、文本（text）以及LLM生成的隐含知识（knowledge）进行拼接，构建成“目标—文本—知识”的三元组结构，以增强模型对目标与文本之间语义关系的理解能力。该过程的示意图如图5所示。

（2）基于自注意力机制的语义编码

输入序列构建完成后，会进入多层Transformer编码器进行语义编码。模型利用其核心的自注意力（self-attention）机制，通过计算元素之间的相关性权重，充分捕捉目标、文本和知识增强信息之间的深层语义关系。自注意力机制的计算依赖查询（query）、键（key）和值（value）3个矩阵的加权关系，并通过缩放点积和归一化来计算，其核心计算式如下：

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{d_k}}\right)V \quad (2)$$

$$Q=X\cdot W_Q, K=X\cdot W_K, V=X\cdot W_V \quad (3)$$

其中，X表示输入的嵌入矩阵，W_Q、W_K和W_V均为可训练的参数矩阵，d_k等于矩阵Q和K的列数，通常与词向量维度相等。除

以$\sqrt{d_k}$的目的是对内积结果进行归一化处理，以避免内积过大。特别地，BERT模型中“[CLS]”位置的隐藏状态向量被视为整个输入序列的全局表示，记作h_{CLS}，并传递至分类层进行预测。

（3）分类与训练

在进入分类层之前，模型引入了Dropout机制以防止过拟合并提升模型的泛化能力。在训练阶段，Dropout以设定的概率随机“屏蔽”部分神经元的输出，迫使模型在学习过程中不依赖特定神经元。具体而言，在h_{CLS}输入分类层之前，以一定的概率p将向量中的部分元素置为零，得到$h_{\mathrm{drop}}=\mathrm{Dropout}(h_{\mathrm{CLS}})$。然后，$h_{\mathrm{drop}}$进入全连接层，通过一个线性变换和Softmax函数计算每个类别的概率，其计算式如下：

$$P\left(y|X\right)=\mathrm{softmax}\left(Wh_{\mathrm{drop}}+b\right) \tag{4}$$

其中，W和b分别为权重矩阵和偏置向量。

模型训练采用交叉熵损失（cross-entropy loss）函数作为优化目标。设真实标签为$\boldsymbol{y}=[y_1,y_2,\cdots,y_K]$，模型预测的概率分布为$\hat{y}=[\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_K]$，则交叉熵损失函数定义如下：

$$\mathrm{L}_{\mathrm{CE}}=-\sum_{k=1}^{K}y_k\log\left(\hat{y}_k\right) \tag{5}$$

其中，K表示类别总数，$y_k\in\{0,1\}$表示真实标签，$\hat{y}_k$表示模型预测的第k类的概率。通过反向传播不断更新模型参数，模型最终能够有效融合原始文本信息与LLM生成的隐含知识，从而增强对立场检测任务的理解能力，提高分类精度。

2.3 对抗训练模块

考虑到用于立场检测的推文数据集规模较小，容易出现过拟合问题，本文使用快速梯度法[12]进行对抗训练。由于文本具有离散的性质，无法直接对原始输入的数据生成扰动，因此Miyato等[12]将对抗训练技术应用于文本分类的单词嵌入空间。本文参考Karimi等[17]的工作，采用标准监督学习方式对BERT进行微调，在嵌入层加入微小扰动，构造对抗样本参与损失函数优化，使用FGM对原始样本和对抗样本分别计算损失，并将常规梯度和对抗梯度累加进行训练，训练过程如图6所示。

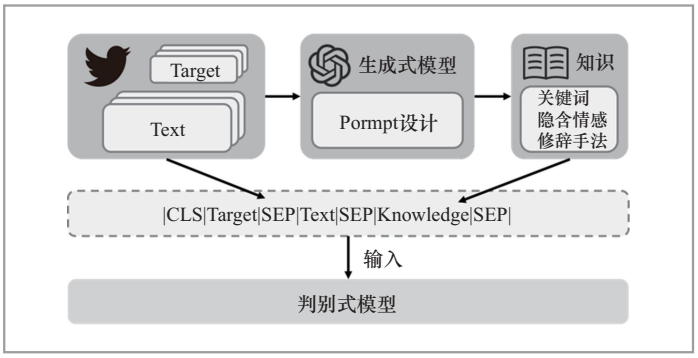

图5 模型输入构造示意图

在常规训练过程中，模型首先对输入进行前向传播，经过编码层、Dropout层和分类层得到预测结果并计算损失，随后通过反向传播计算梯度并更新参数。本文引入对抗训练策略的做法是，在常规反向传播后，对嵌入层沿梯度方向注入微小扰动，生成对抗样本，并基于该对抗样本再次进行前向传播与反向传播，从而在原始梯度的基础上累加对抗梯度。最终，通过恢复嵌入层参数并统一完成一次优化器的更新，保证模型参数的一致性。

在具体实现上，本方法构建了一个对抗训练生成模块，具有扰动的注入与恢复两个功能。首先，在常规的前向传播与反向传播完成后，定位嵌入层参数并保存当前值以便后续恢复。然后，基于参数的梯度计算其L_2范数，并将归一化后的梯度向量乘以扰动强度ε生成扰动向量，与原始

图6 模型训练过程

嵌入参数相加构造对抗样本E_{adv}。接着，模型基于该对抗样本进行新的前向传播与反向传播，从而在原始梯度基础上叠加对抗梯度。最后，对抗训练生成模块会将受扰的嵌入层参数恢复成原始状态，从而确保整体训练过程的稳定性与参数一致性。

3 实验与分析

3.1 数据集

本文使用了两个标准的立场检测数据集进行实验。

（1）SEM16[18]：该数据集来源是推特（Twitter）用户评论，包含无神论（atheism，AT）、气候变化（climatic change，CC）和希拉里·克林顿（hillary clinton，HC）等5个目标话题，含有支持（favor）、反对（against）和中立（none）3类立场标签。

（2）P-Stance[8]：该数据集由2020年美国总统大选期间收集的选民评论数据组成，以唐纳德·特朗普（Trump）、乔·拜登（Biden）和伯尼·桑德斯（Sanders）3位政治人物作为目标，含有支持和反对两类立场标签。

3.2 评价指标与实验设置

为了有效降低类别不均衡对整体评价的影响，本文采用宏平均F1分数（Macro-F1）作为主要评价指标。Macro-F1是基于F1分数的宏平均形式，F1分数是精确率P和召回率R的调和平均，而Macro-F1则是对各类别的F1进行平均，其计算式如下：

$$F1=\frac{2\times P\times R}{P+R} \tag{6}$$

$$\text{Macro}-\text{F1}=\frac{1}{N}\sum_{i=1}^{N}F1_i \tag{7}$$

其中，N为类别总数，在一般的立场检测任务中，使用支持和反对两个类别平均数作为某个目标上的评价结果，中立或不相关不参与该指标的计算。

本文的具体实验参数设置见**表1**。

3.3 对比方法

本文选取了多个主流方法进行对比实

验，以验证所提方法的有效性。

BiLSTM[19]：双向长短期记忆网络。

BiCond[20]：双向条件编码模型。

TextCNN[21]：基于卷积神经网络（CNN）的文本分类模型。

TAN[2]：基于注意力机制的LSTM模型。

BERT[6]：基于Transformer的多层双向编码器模型。

BERTweet[7]：一种基于RoBERTa[22]架构的社交媒体文本立场检测模型，在大规模推特语料上进行预训练。

EZSD-CP[23]：结合对比学习和提示学习策略，通过一个门控多层感知器来动态捕捉并增强实例与提示之间的语义关联。

WS-BERT[24]：基于BERT的立场检测模型，利用维基百科等外部信息来增强性能。

3.4 对比实验

KABERT在两个数据集上与上述基线方法的Macro-F1对见**表2**。

表1 实验参数设置

参数名称	取值	说明
Max Length	512	输入最大长度
Batch Size	64	批处理大小
Learning Rate	1e-5	学习率
Epochs	20	最大训练轮次
Epsilon	0.5	对抗扰动大小
Dropout	0.1	神经元随机失活

实验结果表明，本文提出的KABERT方法在两个数据集上均表现出较好的性能提升。在SEM16数据集上，KABERT的Macro-F1达到68.85%，整体上取得了优于全部基线模型的结果；在P-Stance数据集上，KABERT的Macro-F1达到79.24%，较大部分基线模型均有提升。针对BERTweet模型在P-Stance数据集上的优异表现，我们分析很大程度上源于其对特定政治领域语料的深度学习。相比之下，本文方法通过轻量级的立场提示微调，并未对政治主题做特殊优化，它通过融合外部知识来提升泛化能力，使其在处理不

表2 Macro-F1对比

模型	SEM16						P-Stance			
	AT	CC	FM	HC	LA	ALL	Trump	Biden	Sanders	ALL
BiLSTM	42.11%	6.11%	39.10%	36.83%	40.30%	36.41%	77.18%	75.47%	67.43%	73.36%
BiCond	55.32%	35.64%	52.87%	56.12%	61.17%	60.52%	73.03%	69.39%	64.58%	72.77%
TextCNN	56.52%	**53.79%**	53.77%	61.72%	60.98%	64.98%	74.79%	74.11%	66.68%	71.86%
TAN	**59.33%**	53.59%	55.77%	65.38%	63.72%	68.79%	78.30%	75.26%	70.67%	74.74%
BERT	59.24%	45.77%	55.00%	55.88%	64.04%	68.45%	79.19%	76.02%	73.59%	76.27%
BERTweet	56.40%	45.35%	53.56%	57.49%	58.44%	67.89%	**83.81%**	**79.08%**	77.75%	**80.21%**
EZSD-CP	54.48%	42.50%	56.07%	**75.20%**	58.50%	68.80%	–	–	–	–
WS-BERT	57.43%	50.39%	53.29%	65.46%	59.41%	57.20%	77.69%	77.82%	77.78%	78.20%
KABERT*	58.09%	53.63%	**56.24%**	63.19%	**64.17%**	**68.85%**	80.94%	78.94%	**77.83%**	79.24%

注*表示本文提出的方法，粗体表示最优，下划线表示次优。

同领域、不同任务时都具有稳健的表现。因此，本文方法在通用性和可迁移性上展现出了更强的潜力。

3.5 消融实验

为验证KABERT中知识增强（KA）和对抗训练（AT）两个关键模块的有效性，我们设计了如下3组消融实验，结果见**表3**。

（1）去除KA：在原始数据集上使用对抗训练策略训练BERT模型。

（2）去除AT：在知识增强后的数据集上训练BERT模型，不使用对抗训练策略。

（3）去除KA & AT：在原始数据集上训练BERT模型，不使用对抗训练策略。

消融实验结果表明，KA和AT两个模块都能在一定程度上提高模型的性能。对于KA模块，当未使用LLM生成的隐含知识作为补充时，模型的性能明显下降，两个数据集上的Macro-F1分别下降3.63%和4.56%，平均下降约4.10%，说明通过知识增强给模型提供额外的上下文信息和隐含知识，能显著提升立场检测的准确性。对于AT模块，在未使用对抗训练时，模型在两个数据集上的表现均略有下降但并不显著，Macro-F1平均下降0.86%，表明对抗训练能够在一定程度上提升模型的鲁棒性和泛化能力，对模型准确性提升也起到了一定的积极作用。同时，可以发现在两个模块都加入的情况下，模型的性能的提升最大。

表3 消融实验结果

实验设置	SEM16	P-Stance
KABERT	68.58%	79.24%
w/o KA	64.95%(↓3.63)	74.68%(↓4.56)
w/o AT	67.68%(↓0.90)	78.42%(↓0.82)
w/o KA & AT	64.77%(↓3.81)	74.31%(↓4.93)

注："w/o X"符号表示去除模块X

此外，本文进一步探究了知识增强模块内部各子成分的贡献。如2.1节所述，KABERT使用LLM从关键词、修辞手法和隐含情感3个角度生成隐含知识。为了验证各部分的有效性，本文进一步设计了模块内部的消融实验，即分别移除某一子成分，以观察模型性能的变化。实验结果表明，当任意一部分知识被去除时，模型的性能均出现不同程度的下降，说明3类知识均对立场检测具有积极作用。实验结果见**表4**。

表4 知识增强模块内部消融实验结果

实验设置	SEM16	P-Stance
KABERT	68.58%	79.24%
w/o keyword	67.42%(↓1.16)	77.92%(↓1.32)
w/o implied emotion	67.59%(↓0.99)	78.34%(↓0.90)
w/o rhetorical device	68.40%(↓0.18)	79.01%(↓0.23)

注："w/o X"符号表示去除子成分X

综上所述，本文提出的KABERT方法在立场检测任务中表现出良好的性能。该方法通过结合知识增强和对抗训练策略，有效提升了模型对复杂语义的理解能力及其在实际应用中的鲁棒性。

4 结束语

本文面向社交媒体语境下的立场检测任务，针对现有方法在目标与文本之间隐含关系建模不足，以及模型泛化能力与鲁棒性较弱等问题，提出了一种基于知识增强与对抗训练的立场检测方法KABERT。首先，利用生成式模型ChatGPT进行知识增强，从关键词、隐含情感和修辞手法

等角度挖掘目标与文本之间的深层语义关联，生成可显式利用的隐含知识。其次，采用判别式模型 BERT 作为主干分类网络，并结合基于 FGM 的对抗训练策略，对增强后的数据进行鲁棒性微调。最后，在两个标准的公开数据集上与多种主流方法进行对比实验，结果表明KABERT方法较这些方法均有不同程度的提升，并且通过系列消融实验进一步验证了各个模块的有效性。

参考文献：

[1] Zhang B, Dai G, Niu F, et al. A Survey of Stance Detection on Social Media: New Directions and Perspectives[J]. arXiv preprint arXiv: 2409.15690, 2024.

[2] Du J, Xu R, He Y, et al. Stance classification with target specific neural attention networks[C]//International Joint Conferences on Artificial Intelligence, 2017. 3988-3994.

[3] 白静，李霏，姬东鸿．基于注意力的BiLSTM-CNN中文微博立场检测模型[J]. 计算机应用与软件, 2018, 35(3): 266-274.
Bai J, Li F, Ji D H. Attention Based BiLSTM-CNN Chinese Microblogging Position Detection Model. Computer Applications and Software, 2018, 35(3): 266-274.

[4] 杨顺成，李彦，赵其峰．基于GCN和Bi-LSTM的微博立场检测方法[J].重庆理工大学学报(自然科学), 2020, 34(6): 167-173.
Yang S C, Li Y, Zhao Q F. Stance Detection Method of Chinese Micro-Blog Based on GCN and Bi-LSTM. Journal of Chongqing University of Technology (Natural Science), 2020, 34(6): 167-173.

[5] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.

[6] Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]//Proceedings of NAACL-HLT. 2019: 4171-4186.

[7] Nguyen D Q, Vu T, Nguyen A T. Bertweet: A pre-trained language model for English Tweets[J]. arXiv preprint arXiv: 2005.10200, 2020.

[8] Li Y J, T S, A S, et al. 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.

[9] Huang H, Zhang B W, Li Y Y, et al. 2023. Knowledge-enhanced prompt-tuning for stance detection. ACM Trans. Asian Low Resour. Lang. Inf. Process., 22(6):159:1 - 159:20.

[10] Li A, Liang B, Zhao J Q, et al. 2023. Stance Detection on Social Media with Background Knowledge. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15703 - 15717, Singapore. Association for Computational Linguistics.

[11] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412.6572, 2014.

[12] Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semi-supervised text classification[J]. arXiv preprint arXiv:1605.07725, 2016.

[13] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.

[14] Zhu D, Lin W, Zhang Y, et al. At-bert: Adversarial training bert for acronym identification winning solution for sdu@ aaai-21[J]. arXiv preprint arXiv:

2101.03700, 2021.

[15] Zhang Z, Li Y, Zhang J, et al. Llm-driven knowledge injection advances zero-shot and cross-target stance detection[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). 2024: 371-378.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[17] Karimi A, Rossi L, Prati A. Adversarial training for aspect-based sentiment analysis with bert[C]//2020 25th International conference on pattern recognition (ICPR). IEEE, 2021: 8797-8803.

[18] Saif M, Svetlana K, Parinaz S, et al. Semeval-2016 Task 6: Detecting Stance in Tweets. In Proceedings of the International Workshop on Semantic Evaluation (SemEval '16). June 2016. San Diego, California.

[19] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural networks, 2005, 18 (5-6): 602-610.

[20] Augenstein I, RocktÄSchel T, Vlachos A, et al. Stance Detection with Bidirectional Conditional Encoding[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 876 - 885.

[21] Kim Y. Convolutional neural networks for sentence classification[J]. Empirical Methods in Natural Language Processing, 2014: 1746-1751.

[22] Liu Y H, Ott M, Goyal N, et al. RoBERTA: a robustly optimized BERT pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[23] Yao Z, Yang W, Wei F. Enhancing zero-shot stance detection with contrastive and prompt learning[J]. Entropy, 2024, 26(4): 325.

[24] Nguyen Q M, Kim T. Is External Information Useful for Stance Detection with LLMs? [J]. arXiv preprint, 2019, arXiv: 2507.01543, 2025.

作者简介

黄莉媛（2003– ），女，中国科学院计算技术研究所硕士生，主要研究方向为自然语言处理。

张瑾（1978– ），男，博士，中国科学院计算技术研究所高级工程师，主要研究领域为自然语言处理、大数据处理、信息检索等。

靳小龙（1976– ），男，博士，中国科学院计算技术研究所研究员，主要研究方向为知识图谱、知识计算、大数据知识工程等。

徐辉（1989– ），男，硕士，中国科学院计算技术研究所工程师，主要研究方向为自然语言处理。

郭嘉丰（1980– ），男，博士，中国科学院计算技术研究所研究员，现任中国科学院计算技术研究所网络数据科学与技术重点实验室主任，大数据分析系统国家工程研究中心常务副主任，主要研究方向为信息检索、自然语言理解、大数据分析系统等。

收稿日期：2025-11-30

通信作者：张瑾，jinzhang@ict.ac.cn

基金项目：国家自然科学基金资助项目(No.62441229)

Foundation Items: The National Natural Science Foundation of China (No.62441229)