arXiv:2301.03127v2 [cs.CL] 3 Feb 2023

Logically 参加 Factify 2：基于证据检索技术与 Transformer 编码器架构的多模态事实核查系统

Pim Jordi Verschuuren，Jie Gao，AdelizevanEeden，Stylianos Oikonomou 和Anil Bandhakavi

Brookfoot Mills，Brookfoot Industrial Estate，布里豪斯，*HD6 2RW*，英国

摘要

本文介绍了 Logically 参加 De-Factify 2 挑战赛（DE-FACTIFY 2023）多模态事实核查任务 1 的方案。我们详细介绍了该挑战赛的参赛方案，包括探索的证据检索与选择技术、预训练跨模态与单模态模型，以及基于成熟的 Transformer 编码器（TE）架构的跨模态真实性模型，该架构高度依赖自注意力机制。我们还对 Factify 2 数据集进行了探索性分析，揭示了其中显著的多模态模式，并提出了推动本研究所提架构设计的假设。我们开展了一系列初步实验，以调查并评估不同的预训练嵌入模型、证据检索设置和阈值。最终系统采用标准的两阶段证据驱动真实性检测系统，在任务 1 的验证集和最终盲测集上均取得了 0.79 的加权平均 F1 分数；在排行榜的 9 个参赛团队中排名第 3，与表现最佳的系统仅有微小差距。

关键词

事实核验、多模态表示学习、多模态蕴涵、文本蕴涵、多头注意力

1. 引言

错误信息和假新闻可能会迅速传播，并在各个层面造成危害。保护自己免受这些负面影响的一种方式，是通过事实核查，并依据证据进行报道来辟谣。然而，这一过程可能需要大量资源和时间。为解决这一问题，研究人员利用深度学习技术开发了自动化事实核查系统，这些系统可以运用自然语言处理技术处理文本内容，完成主张检测、主张匹配、证据检索和真实性预测等任务。尽管该领域已取得进展，但仍需要能够同时处理文本与图像输入的多模态方法。为弥补这一空白，本文提出了一种用于自动化事实核查的多模态真实性预测系统，该系统

De-Factify：多模态事实核查与仇恨言论检测研讨会，与 *AAAI 2023* 同期举行。*2023*

华盛顿特区，美国 Pim.jv@logically.co.uk (P.J. Verschuuren); jie@logically.co.uk (J. Gao); adelize.ve@logically.co.uk (A. v. Eeden); stylianos@logically.co.uk (S. Oikonomou); anil@logically.co.uk (A. Bandhakavi) https://www.logically.ai/team/leadership/anil-bandhakavi (A. Bandhakavi) 0000-0002-3610-8748 (J. Gao)© 2023 年本文作者版权所有。允许依据知识共享署名 4.0 国际许可协议（CC BY 4.0）使用。CEUR Workshop Proceedings (CEUR-WS.org)

作为由 De-Factify@AAAI 2023 组织的 Factify 2 竞赛的一部分而开发。

本文其余部分的结构如下：第2节简要概述相关工作，第3节介绍我们的总体框架和模型架构。第4节讨论 Factify 2 竞赛提供的数据集，随后第5节概述我们的实验。第6节和第7节分别介绍最终结果和结论。

2. 相关工作

作为自动化事实核验的重要组成部分，针对真实性预测对主张与证据进行建模的有效技术，一直是现有事实核查方法中的热门话题和关键研究问题。近期大多数工作着眼于在真实性预测中使用文本证据，主要有两条研究路线。一种方向 [1, 2, 3] 是使用单篇文档（例如 Factify 任务数据集所提供的文档），其中包含长文本证据，并利用为长序列构建的模型。BigBird [4], Longformer[5]等模型，以及卷积网络架构的最新进展（如 Long Range Arena 排行榜中的 Mega [6], S5[7]），在各类任务和其他排行榜上均取得了领先成绩。在文档级使用长序列模型的优势包括：a）整体架构简单；b）能够将整篇文章的更多上下文纳入建模和自然语言推理。通常需要为主张（或查询）和文档序列设置合适的最大长度，并提供文档级真实性标签 [8, 1, 3]。将大量上下文纳入推理的优势，在问答（QA）任务的建模中也有所体现 [4, 5], ，而此类任务的文档级真实性标签相对“廉价”。在文档级使用简单长文本模型技术的缺点是缺乏可解释性（就证据选择而言）、计算成本高、难以处理某些复杂的（多跳）主张 [9], ，以及在真实应用中面对大量不同文档时缺乏多样性和可扩展性。在利用商业搜索引擎提取的网络数据作为事实核查系统构建模块、以纳入更多样化来源的开放领域事实核查任务中，这些限制尤为明显。值得注意的是，长序列模型可以适配证据选择任务，例如将其设定为词元级预测任务。比如，作为 SciFact 排行榜上的顶尖系统之一[1]，LongChecker [10] 使用 LongFormer [5] 进行科学主张核验，并进行段落级证据选择。在他们的方法中，每个句子都会插入一个带有全局注意力的[CLS] 词元，使模型能够将该句子级词元作为证据进行预测。这些工作大多聚焦于有限的上下文，例如少数维基百科文档、单篇文章和摘要，或来自研究文献或小型合成语料库的文本片段。

另一条广泛采用的研究路线，也是 FEVER [11, 12] 中的关键任务之一，是纳入证据检索与选择。该框架利用更大的文档上下文，首先提取证据（或推理依据）段落，随后根据主张和选定的推理依据来构建真实性预测模型。证据段落报告了

[1] https://leaderboard.allenai.org/scifact/submissions/public

与主张相关的研究发现，可用于论证每个真实性标签，并可在句子级或段落级进行选择。尽管大规模语言模型（大规模语言模型），例如 GPT-3[13] 和 ChatGPT[2]，，取得了革命性的突破并展现出强大的生成能力，但这些大型模型仍缺乏事实核查任务所需的关键零样本或少样本学习能力。这主要是因为其权重中存储的知识可能检索错误、不完整或过时，导致这些技术容易产生幻觉 [14, 15],，这与事实核查任务相冲突，因为事实准确性是建模中的必要要素。此外，如何高效地使大规模语言模型跟上持续增长的事实信息和新信息，并以这些信息为依据，至今仍是亟待解决的问题。近期研究 [15, 16] 表明，在包括自然语言推理（NLI）、文本蕴含识别（RTE）、阅读理解（RC）、问答等在内的一系列知识密集型自然语言处理任务中，采用微调和较小模型的轻量级方法表现优于这些大型模型。Sentence-BERT（SBERT） [17] 是一种基于 BERT 语言模型 [18] 、用于证据选择 [19, 20] 的热门技术，可将该任务构建为句对回归任务。SBERT 模型用于编码每个证据段落的上下文化表示，随后根据其与相应主张的上下文化表示之间的语义相似度进行排序。最后一步是选取排名靠前的 k 个证据段落，用于真实性预测。这种多阶段验证框架面临的挑战包括：1）脱离上下文提取的推理依据可能缺少作出预测所需的信息（例如，首字母缩略词、未解决的共指）；2）证据提取（通过段落排序）需要高质量训练数据，而从封闭领域和开放领域任务中获取这类数据需要领域专家参与，成本高昂 [21]。为应对这些限制，研究者开展了多项探索：1）使用科学文献中的段落级训练数据，并将论文标题作为主张、摘要作为证据，采用高精度启发式方法（例如，SciFact [1]）；2）将问题和答案分别视为主张和证据的问答数据集 [22]；3）将主张视为假设、证据视为证据的自然语言推理数据集 [23]。我们沿用第二条研究路线，并在系统中按照当前最先进方法实现证据检索组件。

与基于文本的技术相比，自动化多模态或跨模态事实核查仍是一个发展不足的领域。近期进展表明，跨模态预训练模型（例如 VideoBERT [24], VisualBERT [25], UNITER [26], CLIP [27]）在下游跨模态任务中取得了显著成果 [28, 29, 30] ，并且在零样本或少样本场景中具有很强的迁移能力。我们的工作受到 [31], 的启发，该研究是对多模态事实核查任务的早期探索之一。在其提出的方法中，采用对比语言-图像预训练（CLIP）模型 [27]）作为编码器，学习每张图像与输入主张文本之间的语言-图像联合嵌入。该方法将排名前五的候选图像证据与多模态主张一同输入多模态主张验证模型，并使用简单的交叉注意力网络。值得注意的是，CLIP 模型能够在粗粒度的上下文（全局）层面建模图文上下文对齐，但忽略了解耦概念的组合匹配（即区域-词语层面更细粒度的跨模态对齐）[30, 32, 32]。

[2]https://openai.com/blog/chatgpt/

3. 方法论

3.1. 问题陈述

与前一版提交相同，我们将 Factify 2 问题定义为多模态蕴涵任务[3], ，其中多模态主张 $c = c_{text} + c_{image}$ 作为假设，多模态文档 $d = d_{text} + d_{image}$ 作为前提。目标是学习一个函数 $f(c,\ d)$，推断五种蕴涵类别之一，包括“支持_多模态”、“支持_文本”、“驳斥”、“多模态信息不足”和“文本信息不足”。有关该任务的更多详情，请参阅
_ _[33]。

3.2. 总体架构

我们的系统架构遵循近年来各类共享任务所确立的标准两阶段主张核验方法，典型示例包括 FEVER[34], FEVER 2.0 [35], FEVEROUS[36] 和 SCIVER [37]。首先，文本证据检索组件从给定文档中识别出与相应主张文本最相关的证据段落。然后，利用所有输入数据训练一个基于 Transformer 的跨模态模型，这些数据涵盖多种模态，包括所选证据段落文本、主张文本、主张图像、文档图像、主张 OCR 文本和文档 OCR 文本，以针对多模态主张预测五种多模态蕴涵类别。嵌入层同时采用预训练跨模态模型（即 CLIP）和预训练文本嵌入模型，利用统一的多模态表示和单模态表示来学习跨模态匹配模型。总体而言，所实现的架构采用列表式拼接策略 [38] ，这是近期大多数最先进的序列到序列真实性预测模型所采用的常见策略之一。

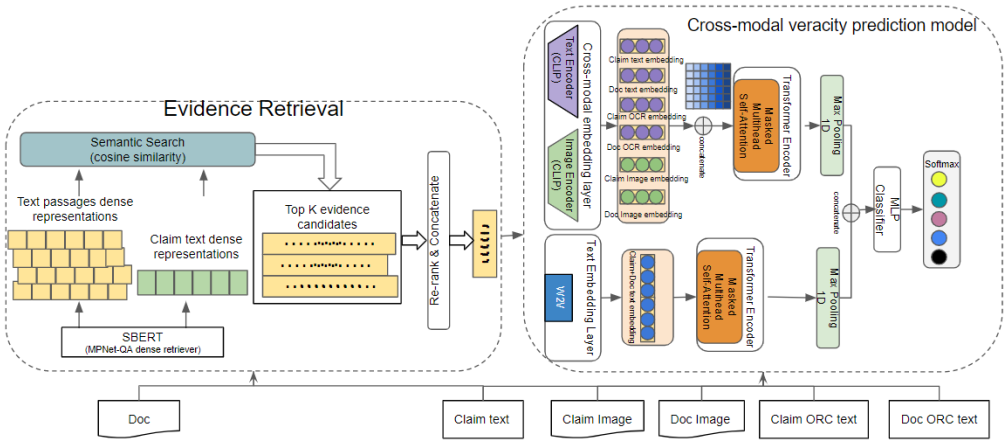


图 1： Logically 通用系统架构

3.3. 证据检索

在证据检索中，‘multi-qa-mpnet-base-dot-v1‘ [3] 用于在段落级别计算主张文本和文档文本的嵌入。关于段落粒度，我们对段落级和句子级检索都进行了实验（见第 5节）。这是一个基于 MPNet 架构的 SBERT 模型 [39]，使用来自不同来源的 2.15 亿个问答对进行训练，所用数据集为问答（QA）数据集。该模型针对语义搜索进行了调优，采用点积评分函数来查找与给定查询相对应的相关段落。该模型将文本编码为 768 维向量，最多支持 512 个词元。

在相似度计算和语义搜索方面，我们使用归一化的 SBERT 嵌入进行简单的点积运算（作为余弦相似度的近似），从而能够快速、高效地进行段落排序，并支持扩展至约 100 万条记录。

通过语义搜索获得的前 K 个段落随后会根据其与主张文本的相关性重新排序，并拼接成一段较长的文本片段，然后输入跨模态真实性预测模型。

3.4. 嵌入层

我们的嵌入层由一个跨模态编码器和一个单模态文本编码器组成。我们假设，仅对文本间交互（即文本前提与假设）进行建模，可以补充仅对跨模态前提与假设交互进行建模，反之亦然。通过将更多文本对齐信号映射到后续语义空间，该架构有助于在这项多模态事实核查任务中衡量多模态语义相关性。这是因为文本专用模型能够捕捉更准确且在语义上更有意义的词级或句子级对齐。

跨模态编码器采用预训练的 CLIP 模型实现，旨在将视觉和文本嵌入映射到同一空间中。本研究选择 ViT-B/32 变体（采用图像块大小为 32 的 ViT-Base），因为它的参数量较少、浮点运算次数较少，推理速度也更快。ViT-B/32 由文本编码器和图像编码器组成，分别用于编码文本输入（包括主张文本、证据段落以及两张图像中的 OCR 文本）和图像输入（包括主张图像和文档图像），随后将这些输入拼接为一个 6×512 矩阵，作为后续 Transformer 编码器的单一输入。CLIP 架构允许输入的文本长度最多为 77 个词元。预训练的 Word2vec 模型（“Word2vec Google News 300”）[40] 被用作单模态文本编码器。它对主张文本和文档证据段落文本拼接而成的序列进行编码，并为每个词元生成一个 300 维特征向量。通过零填充，使序列长度与训练集中最长的句子匹配。预训练的 CLIP 和 Word2Vec 嵌入模型均未进行微调。

3.5. 跨模态真实性预测

真实性预测的第二个组成部分基于成熟的 Transformer 编码器（TE）架构，该架构高度依赖自注意力 [41]的概念，以有效地

[3]该模型可在 Hugging Face Hub 上获取，访问地址为 https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

对输入中的高阶交互和上下文进行建模。近期研究表明，多头自注意力机制和 Transformer 架构在这方面计算效率高且准确。TE编码器的自注意力机制能够进行简单而强大的推理，从而识别向量实体之间隐藏的关系，无论这些实体是视觉的还是文本的。因此，我们的跨模态真实性预测模型基于自注意力机制实现，用于学习声明-文档文本对的文本表示与主张和文档中所有模态的跨模态特征表示的联合分布。

具体而言，CLIP 对联合输入中的主张和文档进行嵌入，文本嵌入层则对文本输入进行嵌入，随后将两者分别送入两个独立的 Transformer 编码器 [41]，每个编码器 N 由多个相同的连续模块构成，每个模块包含多头注意力（MHA）和全连接前馈网络（FFN）。每个 Transformer 编码器中的多个模块能够更深入地理解输入。对于每个模块，输入 x 会经过多头注意力层，注意力层的输出与初始输入相加，从而避免丢失初始序列中的信息。随后，对输出进行层归一化，以加快训练并实现轻度正则化，即 $x = \text{LayerNorm}(x + \text{MHA}(x))$. 接着，将输出传入前馈网络，以提高模型的复杂度。然后再次将输出与原始输入相加，并进行层归一化，即 $x = \text{LayerNorm}(x + \text{FFN}(x))$. 最终模块的输出（即图中每个 Transformer 编码器的输出）会经过自适应最大池化层，以降低输出维度。随后，将两个独立 Transformer 编码器的输出拼接，再输入多层感知机分类器，对五个类别进行预测。五个类别的概率由最终 Softmax 层的输出得出。

4. Factify数据集

4.1. 数据集描述

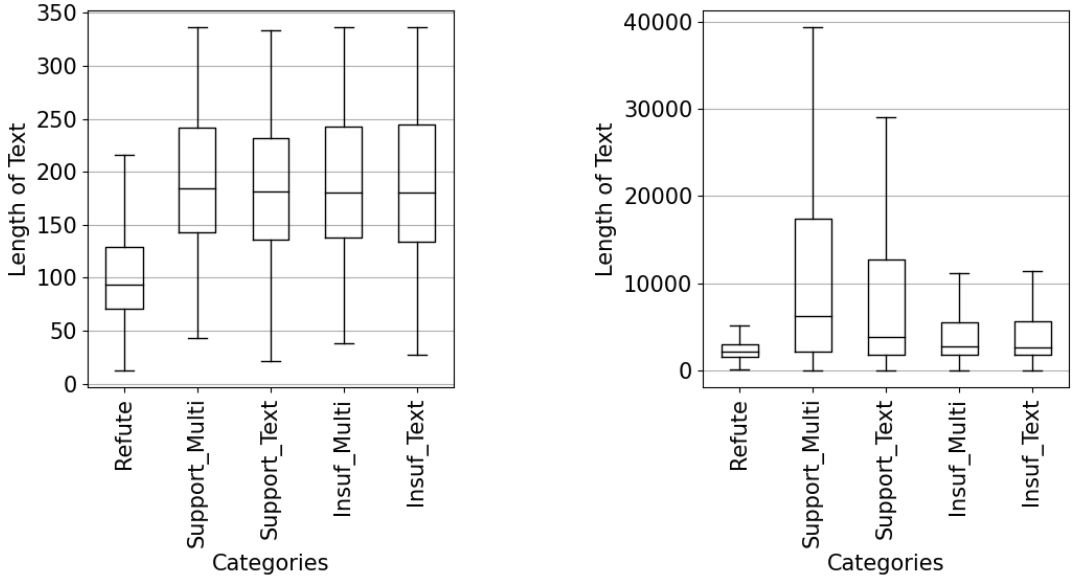
组织者创建并提供的 Factify 2 数据集包含训练集、验证集和测试集。训练集包含 35000 对数据，验证集和测试集则各包含 7500 对数据。每对数据由一个主张和一个文档组成，两者均包含一张图像、一段文本，以及从图像中提取的 OCR 文本。每对数据均标注为 5 个类别之一：支持_多模态、支持_文本、反驳、信息不足_多模态或文本信息不足。_

4.2. 文本长度分布

训练集文本和 OCR 文本长度分布如图 2和 3所示。主张文本与文档文本的文本长度分布有所不同，文档文本往往要长得多。这是预期的，因为文档文本用于验证主张。从图 2（a）中可以看出，与其他类别相比，反驳类别的主张文本明显更短且变化更小，而其余类别的主张文本长度分布较为相似。图 2（b）显示，支持_多模态和支持_文本类别的文档文本长度分布范围更广，且文档文本长度也最长。两个不足

类别的平均文档文本长度较短，而反驳类别的文档文本长度方差和最大长度均最小。

从图中可以看出主张 OCR 长度3，反驳类别的主张 OCR 长度分布和最大长度都远高于其他类别。第二大的主张 OCR 长度分布属于支持_文本和不足_文本类别，因此两个多模态类别的主张 OCR 文本长度最短。文档 OCR 长度分布与主张 OCR 长度分布非常相似。从图 3b 可以看出，唯一明显的差异是，两个文本类别的文档 OCR 长度分布比主张 OCR 长度分布更小。

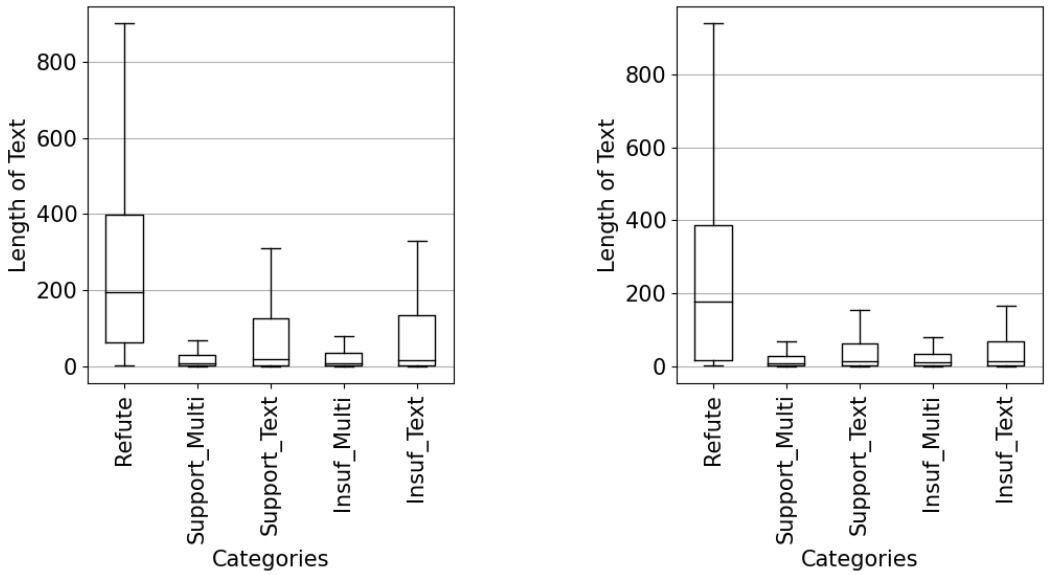


（a）主张文本长度　　（b）文档文本长度

图 2： 所有类别的文本长度分布箱线图

4.3. 图像相似度分布

为了直观了解各类别中声明图像与文档图像之间的相似度，我们开展了图像相似度研究。我们利用图像对的 CLIP 嵌入计算相似度分数，并按类别进行分析。图4a和4b 显示，在支持_多模态和文本信息不足_多模态类别中，声明图像与文档图像之间的相似度明显高于其他类别。与去年的 factity 1 数据集相比，标签与图像对相似度之间的相关性大幅提高。这进一步表明，多模态类别内部存在明确的相关性，可用于学习和验证多模态蕴涵类别。



(a) 主张 OCR 文本长度

(b) 文档 OCR 文本长度

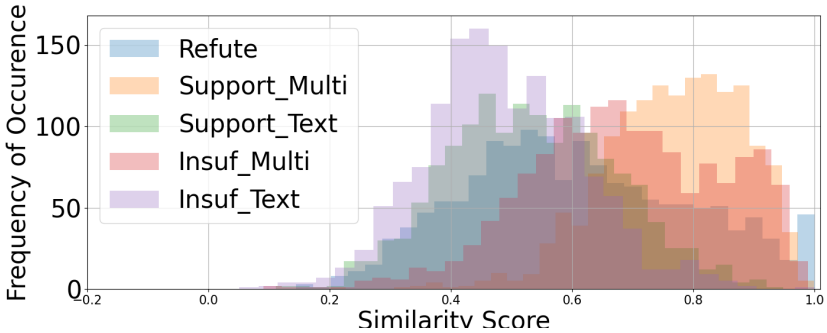
图 3： 所有类别的 OCR 文本长度分布箱线图

4.4. 多模态相似度分布

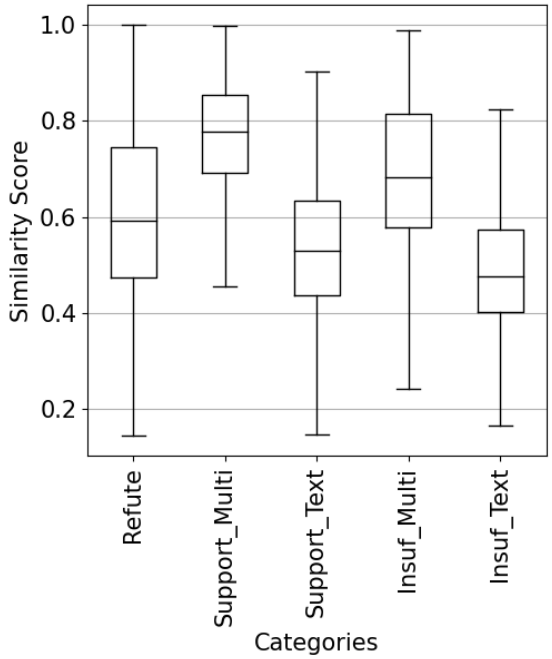
我们研究了多模态主张与文档图像对之间的多模态 CLIP 相似度，以检验我们的假设：文档图像应包含与主张相关的内容，才能得出支持或反驳的判定结果。图5a和5b展示了主张文本与文档图像之间的余弦相似度分数。值得注意的是，“支持_多模态”类别的标签与主张-证据对之间，成对相似度相关性最高。“文本信息不足”类别的成对相似度相关性最低，尽管我们最初假设“文本信息不足_多模态”应具有最低的数值。这项分析表明，要根据主张文本与文档图像之间的相关性区分不同类别，可能颇具挑战。

就主张图像与文档文本之间的相关性而言，由于 CLIP 的最大文本序列限制，文本输入的最大长度会被截断。因此，文档文本中较长的上下文未纳入本分析。如图 6a 和 6c所示，五个类别之间的相似度相关性整体较低，其中“反驳”类别显示出最高的相似度相关性。

最后，图6b 和图6d 展示了主张图像与主张文本之间的相似度相关性；将主张图像与主张文本相互比较时，不同类别的相似度分数没有显著差异。针对这项任务和这个数据集，我们假设主张图像应为主张文本提供补充信息。



(a) 主张图像与文档图像相似度分数直方图



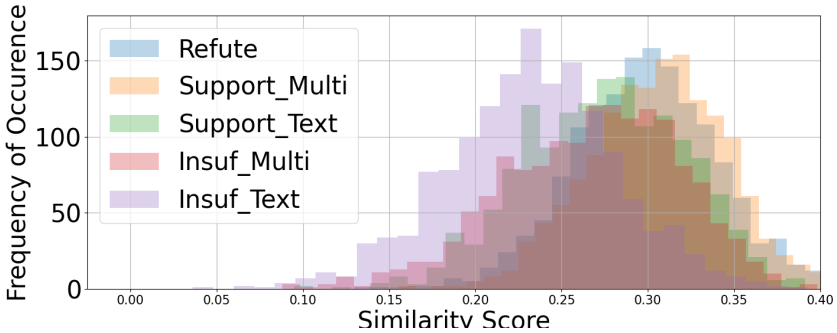
(b) 主张图像与文档图像相似度箱线图

图 4： 主张图像与文档图像相似度分布

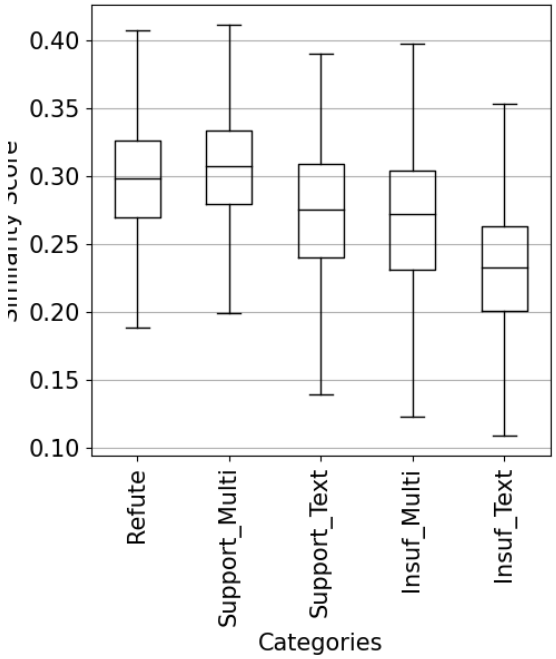
5. 实验

5.1. 模型设置

为验证并优化证据检索的效果，我们尝试通过以下方式对模型进行实验：1）纳入或排除证据选择；2）改变证据检索器排序后的证据文档文本长度；3）比较段落级别与句子级别的段落排序；4）比较使用 SBERT 进行文本到文本对齐与使用 CLIP 进行跨模态对齐。SBERT 和 CLIP 均用于对段落级别和句子级别的证据文档进行排序；5）在问答数据集上训练的 SBERT 模型是否优于通用SBERT模型。请注意，排序



(a) 主张文本与文档图像相似度分数直方图



(b) 主张文本与文档图像相似度箱线图

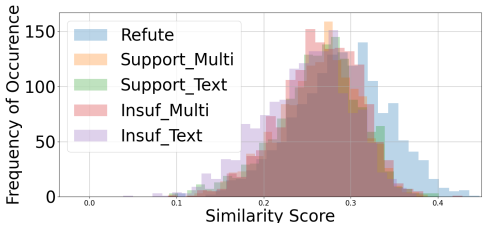
图 5： 主张文本与文档图像相似度分数

由于最大允许长度限制，段落级别置于顶部 <5 或句子级别置于顶部 <5仅适用于 CLIP。

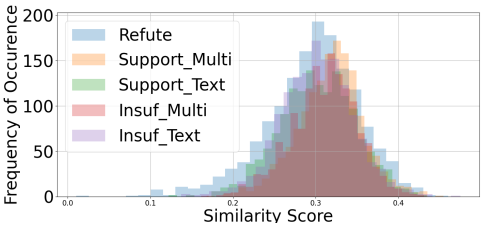
对于两个 Transformer 编码器，我们采用经验设置，在两个多头注意力中各使用四个注意力头。每个嵌入输入对应的连续多头注意力和前馈网络块数量为$N_{blocks} = 2$。所有实验均使用一个三层多层感知机进行训练，各层节点数分别设为 3072、1024 和 5。多层感知机各层之间采用丢弃率 0.5，并应用 ReLU 激活。

本研究开展的初步实验详述如下：

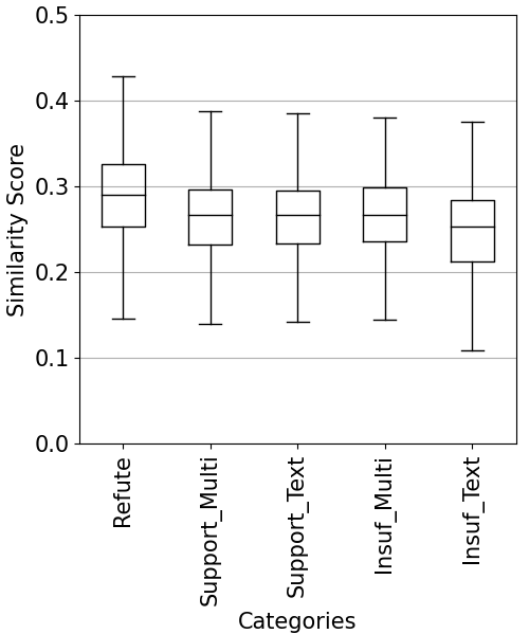
- “模型_不使用_证据检索”：为验证证据检索的有效性，我们从系统中移除证据



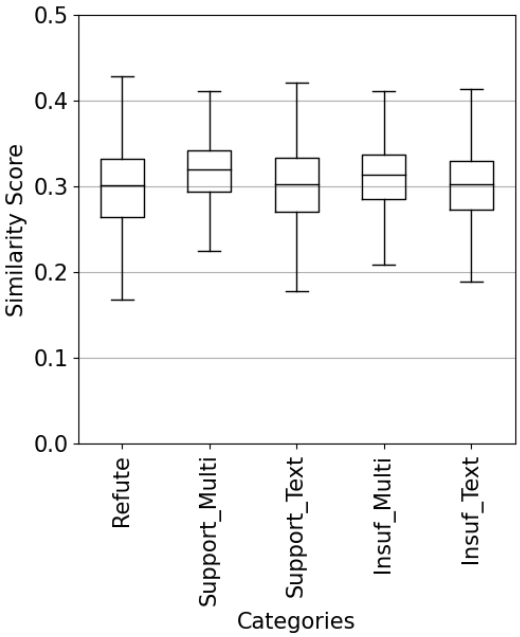
(a) 主张图像与文档文本相似度分数直方图图



(b) 主张图像与主张文本相似度分数直方图图



(c) 主张图像与文档文本相似度分数箱线图



(d) 主张图像与主张文本相似度分数箱线图-plot

图 6： 多模态主张与文档之间的图像和文本相似度分布

检索，并将原始文档文本提供给“跨模态真实性预测模型”。

- "SBERT_sentence_ER_top5”：在我们的实验中，选用了表现“顶尖”的通用SBERT模型（“all-MiniLM-L6-v2”）。这是一个针对多种使用场景调优的全能型模型，速度快5倍，同时与最佳全能型模型“all-mpnet-base-v2”相比，也能提供良好的质量。该模型基于一个庞大且多样化的数据集进行训练，其中包含超过10亿个训练对；此外，还针对适用于余弦相似度的点积评分函数进行了微调。使用全能型模型，使我们能够评估采用问答微调对应模型的价值；我们推测后者是最优方案。Top

[4]根据 SBERT 提供的模型性能排序列表及推荐使用场景，选取表现最佳的通用模型。相关信息可通过 https://www.sbert.net/docs/pretrained_models.html 获取

在此设置中，按全能型SBERT模型排序的前5个句子已配置完成。

- “SBERT_sentence_ER_top10”：此设置采用全能型SBERT模型排序后的前 10 个句子。
- “SBERT_sentence_ER_top15”：此设置采用全能型SBERT模型排序后的前 15 个句子。
- “SBERT-QA_段落_ER_top5”：此设置采用经过问答数据集微调的模型（如3.3节所述），以获取前 5 个段落，作为用于真实性推断的证据段落。
- “SBERT-QA_sentence_ER_top5”：此设置采用 SBERT 问答模型排序后的前 5 个句子，并将其选作证据段落。
- “BigBird_w/o_ER”：为评估证据选择相较于长上下文建模方案的价值，此设置使用去年在 Factity 数据集上微调的谷歌 BigBird 预训练模型 [3]，替换“文本嵌入层”中的 Word2Vec 模型。该 BigBird 模型最多支持 1396 个词元，并在此设置中采用文本的上下文表示。

5.2. 训练与验证

在本次实验中，我们使用自适应 AdamW 优化器 [42]，以初始学习率 $\gamma = 1e{-}4$ 和 epsilon $\epsilon = 1e{-}8$、批次大小 $N_{batch} = 16$，通过最小化交叉熵损失函数训练模型，最多训练 80 轮次，并在验证损失达到最小值时提前停止。提前停止的耐心值设为 5。我们使用了线性递减的学习率调度器，其中包括 $N_{steps} = 438$ 个预热训练步，在此期间，学习率线性增加至设定值。

我们发现，组织者提供的开发数据集中存在数据抓取错误，导致文档文本内容无效：训练集和验证集中分别有 463 个和 114 个无效样本。测试集中也有 112 个无效样本。这些样本的文档文本仅包含“我们检测到此浏览器中已禁用 JavaScript……” 。这些无效样本已从训练数据中移除。

6. 结果与讨论

第 5 节所述初步实验中表现最佳的模型结果分别列于表1、表2和表3 中。

首先，表1显示，不使用ER的真实性判定模型表现相当不错；使用长序列模型（BigBird）生成文本嵌入，使基础模型的表现略有提升，除“反驳”类别外，所有类别均提升了1%。作为对比，我们进一步开展了使用ER的实验，结果见表2和表3。表2 的结果表明，根据目前涵盖三种 top K 句子设置（K=5、10、15）的初步探索，全能型SBERT模型的证据选择并未带来明显的性能提升。相比之下，基于SERT-QA的模型在段落级别和句子级别均取得了显著提升。我们的实验涵盖了 top 5 段落和句子的设置，使表现最佳的基础模型（不使用ER）分别提升了

表 1验证集上未使用 ER 的实验五分类结果

类别	未使用 ER 的模型_ _			BigBird_未使用_ER		
	P	R	F1	P	R	F1
支持_多模态	0.73	0.79	0.76	0.73	0.81	**0.77**
支持_文本	0.71	0.61	0.66	0.77	0.59	**0.67**
不足 _多模态	0.66	0.66	0.66	0.64	0.70	**0.67**
不足 _Text	0.71	0.75	0.73	0.73	0.75	**0.74**
反驳	0.99	0.98	0.98	0.98	0.98	0.98
加权平均	0.76	0.76	0.76	0.77	0.77	**0.77**

表2使用全方位 SBERT + ER 在验证集上的实验五分类结果

类别	SBERT_句子_ER_前5			SBERT_句子_ER_前10			SBERT_句子_ER_前15		
	P	R	F1	P	R	F1	P	R	F1
支持_多模态	0.72	0.85	0.78	0.74	0.78	0.76	0.75	0.77	0.76
支持_文本	0.63	0.73	0.68	0.71	0.61	0.66	0.71	0.62	0.66
不足 _多模态	0.70	0.64	0.67	0.66	0.67	0.66	0.65	0.67	0.66
不足 _Text	0.80	0.58	0.67	0.70	0.77	0.74	0.71	0.76	0.73
反驳	0.96	0.99	0.97	0.96	0.99	0.97	0.98	0.98	0.98
加权平均	0.76	0.76	0.75	0.76	0.76	0.76	0.76	0.76	0.76

表 3在验证集上使用 SBERT-QA +ER 进行实验的五分类结果

类别	SBERT-QA_段落_ER_前5			SBERT-QA_句子_ER_前5		
	P	R	F1	P	R	F1
支持_多模态	0.80	0.77	0.78	0.79	0.83	**0.81**
支持_文本	0.70	0.68	0.69	0.70	0.69	**0.70**
不足 _多模态	0.66	0.72	0.69	0.71	0.72	**0.73**
不足 _Text	0.76	0.72	**0.74**	0.74	0.72	0.73
反驳	0.96	1.00	0.98	0.99	0.98	0.98
加权平均	0.78	0.78	0.78	0.79	0.79	**0.79**

1%和2%。在7种不同实验设置下取得的最终结果显示，与不使用ER的基础模型以及使用全能型SBERT模型相比，结合SBERT-QA进行 top K 句子级证据段落检索可获得最佳性能。最佳模型“SBERT-QA_sentence_ER_top5”在第20个训练轮次时取得了0.79的加权平均F1。

表 4 Factify 官方排行榜

Rank	Team	支持_文本	支持_多模态。	文本信息不足_	信息不足，多模态。_	反驳	最终
1	三重核查	**0.828**	**0.914**	0.852	**0.892**	**1.0**	**0.818**
2	INO	0.812	0.9	**0.888**	0.852	0.999	0.808
3	Logically	0.804	0.905	0.844	0.856	0.985	0.79
4	张	0.766	0.879	0.816	0.879	0.999	0.774
5	gzw	0.785	0.863	0.814	0.833	1.0	0.761
6	coco	0.773	0.865	0.815	0.83	1.0	0.757
7	Noir	0.771	0.873	0.785	0.816	0.997	0.745
8	Yet	0.707	0.826	0.786	0.719	1.0	0.691
9	TeamX	0.582	0.709	0.537	0.556	0.698	0.456
-	基线	0.5	0.827	0.802	0.759	0.988	0.65

6.1. 竞赛结果

最终测试集结果和竞赛排行榜列于表格 4。结果显示，排名前 3 的参赛系统表现相近，我们的系统以微小差距（0.028）位列第 3，与表现最佳的系统仅有细微差距。有关竞赛详情，请参阅 [43]。

7. 结论

在本研究中，我们介绍了提交至 De-Factify 2023 竞赛的多模态事实核查系统。该系统由多个组件构成，包括一个多模态事实核查数据集、一个问答增强的证据段落检索组件，以及一个基于 Transformer 的跨模态序列到序列真实性预测模型。我们在 De-Factify 2023 竞赛中的发现表明，近期预训练跨模态模型（如 CLIP）的进展，使这类模型具备强大的零样本或少样本能力，并能有效迁移至多种下游任务，包括多模态事实核查。然而，我们仍需要更有效的多模态建模与可解释性技术，特别是通过联合建模模态内和模态间关系，学习更细粒度的跨模态表征，并将视觉区域与句子词语或实体对齐。此外，应更加关注真实场景中的挑战，例如处理来自多个来源和领域的大量文本及多模态信息，以核验主张。我们还需要能够有效应对更复杂、细微的真实场景的技术，例如涉及讽刺、反语和误导性语境的场景。正如先前研究 [44, 3], 所指出的，创建能够准确反映真实场景（例如，证据不足或证据泄露）的高质量大规模多模态事实核查数据集仍是一项重大挑战。

参考文献

[1] D. Wadden, K. Lo, L. Wang, A. Cohan, I. Beltagy, H. Hajishirzi, Multivers：通过弱监督和全文档上下文改进科学主张核验，载于：计算语言学协会论文集：NAACL 2022，2022，第 61–76 页。[2] D. Stammbach，将证据选择作为词元级预测任务，载于：第四届事实抽取与验证研讨会论文集（FEVER），EMNLP，计算语言学协会（ACL），2021。[3] J. Gao, H.-F. Hoffmann, S. Oikonomou, D. Kiskovski, A. Bandhakavi, Logically 在 factify 2022：多模态事实验证，arXiv（2021）。网址：https://arxiv.org/abs/2112.09253。doi：10.48550/ARXIV.2112.09253.[4] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang 等，Big bird：面向更长序列的 Transformer，《神经信息处理系统进展》33（2020）17283–17297。[5] I. Beltagy, M. E. Peters, A. Cohan, Longformer：长文档 Transformer，arXiv 预印本 arXiv:2004.05150（2020）。[6] X. Ma, C. Zhou, X. Kong, J. He, L. Gui, G. Neubig, J. May, L. Zettlemoyer，Mega：配备移动平均的门控注意力，arXiv 预印本 arXiv:2209.10655（2022）。[7] J. T. Smith, A. Warrington, S. W. Linderman，面向序列建模的简化状态空间层，arXiv 预印本 arXiv:2208.04933（2022）。[8] J. Zhao, J. Bao, Y. Wang, Y. Zhou, Y. Wu, X. He, B. Zhou，Ror：用于长文档机器阅读理解的逐层阅读，arXiv 预印本 arXiv:2109.04780（2021）。[9] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders 等，WebGPT：结合人工反馈的浏览器辅助问答，arXiv 预印本 arXiv:2112.09332（2021）。[10] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi，事实还是虚构：核验科学主张，arXiv 预印本 arXiv:2004.14974（2020）。[11] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal，FEVER：用于事实抽取与验证的大规模数据集，载于：2018 年计算语言学协会北美分会会议论文集：人类语言技术，第1卷（长篇论文），计算语言学协会，新奥尔良，路易斯安那州，2018，第 809–819 页。网址：https://aclanthology.org/N18-1074。doi：10.18653/v1/N18-1074.[12] R. Aly, Z. Guo, M. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal，FEVEROUS：在非结构化和结构化信息中进行事实抽取与验证，arXiv 预印本 arXiv:2106.05707（2021）。[13] L. Floridi, M. Chiriatti，GPT-3：其本质、范围、局限与影响，《心智与机器》30（2020）681–694。[14] J. Maynez, S. Narayan, B. Bohnet, R. McDonald，论抽象式摘要的忠实性与事实准确性，arXiv 预印本 arXiv:2005.00661（2020）。[15] A. Lazaridou, E. Gribovskaya, W. Stokowiec, N. Grigorev，通过少样本提示实现互联网增强型语言模型，用于开放域问答，arXiv 预印本 arXiv:2203.05115（2022）。[16] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le,

微调语言模型是零样本学习者，arXiv 预印本 arXiv:2109.01652 (2021)。[17] N. Reimers, I. Gurevych, Sentence-bert：使用孪生 BERT 网络的句子嵌入，arXiv 预印本 arXiv:1908.10084 (2019)。[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert：用于语言理解的深度双向 Transformer 预训练，arXiv 预印本 arXiv:1810.04805 (2018)。[19] A. Saakyan, T. Chakrabarty, S. Muresan, Covid-fact：新冠疫情期间现实世界主张的事实抽取与验证，arXiv 预印本 arXiv:2106.03794 (2021)。[20] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, L. Huang, 端到端多模态事实核查与解释生成：一项具有挑战性的数据集与模型，arXiv 预印本 arXiv:2205.12487 (2022)。[21] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov 等，CheckThat! 2020 概述：社交媒体中主张的自动识别与核验，载于：欧洲语言跨语言评测论坛国际会议，Springer，2020，第 215–236 页。[22] N. Lee, C.-S. Wu, P. Fung，使用可分解注意力模型和词汇标注改进大规模事实核查，载于：2018年自然语言处理实证方法会议论文集，2018，第 1133–1138 页。[23] T. Schuster, A. Fisch, R. Barzilay，补充维生素 C！运用对比证据进行稳健事实核验，arXiv 预印本 arXiv:2103.08541 (2021)。[24] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, Videobert：用于视频与语言表征学习的联合模型，arXiv (2019)。URL: https://arxiv.org/abs/1904.01766。doi:10.48550/ARXIV.1904.01766。[25] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert：一种简洁且高效的视觉与语言基线，arXiv (2019)。URL: https://arxiv.org/abs/1908.03557。doi:10.48550/ARXIV.1908.03557。[26] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter：通用图文表征学习，arXiv (2019)。URL: https://arxiv.org/abs/1909.11740。doi:10.48550/ARXIV.1909.11740。[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark 等，利用自然语言监督学习可迁移视觉模型，载于：国际机器学习大会，PMLR，2021，第 8748–8763 页。[28] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, Y. Qiao，CLIP-Adapter：通过特征适配器改进视觉语言模型，arXiv 预印本 arXiv:2110.04544 (2021)。[29] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, B. Cui, CALIP：通过无参数注意力实现 CLIP 的零样本增强，arXiv 预印本 arXiv:2209.14169 (2022)。[30] K. Jiang, X. He, R. Xu, X. E. Wang, ComCLIP：无需训练的组合式图文匹配，arXiv 预印本 arXiv:2211.13854 (2022)。[31] W.-Y. Wang, W.-C. Peng，Factify 2022 的 Yao 团队：利用预训练模型和协同注意力网络进行多模态事实核验，arXiv (2022)。URL: https://arxiv.org/abs/2201.11664。doi:10.48550/ARXIV.2201.11664。[32] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet，使用 Transformer 编码器进行跨模态检索的细粒度视觉文本对齐，ACM

多媒体计算、通信与应用汇刊 (TOMM) 17 (2021) 1–23。[33] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar，Factify 2：多模态假新闻与讽刺新闻数据集，载于：Defactify 2 论文集：第二届多模态事实核查与仇恨言论检测研讨会，CEUR，2023。[34] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal，事实抽取与验证（FEVER）共享任务，arXiv 预印本 arXiv:1811.10971 (2018)。[35] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal，FEVER 2.0 共享任务，载于：第二届事实抽取与验证研讨会论文集（FEVER），2019，第 1–6 页。[36] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal，非结构化与结构化信息上的事实抽取与验证（FEVEROUS）共享任务，载于：第四届事实抽取与验证研讨会论文集（FEVER），2021，第 1–13 页。[37] D. Wadden, K. Lo，SCIVER 科学主张核验共享任务的概述与见解，arXiv 预印本 arXiv:2107.08188 (2021)。[38] K. Jiang, R. Pradeep, J. Lin，探索使用 T5 进行列表式证据推理以实现事实核验，载于：计算语言学协会第 59 届年会暨自然语言处理国际联合会议第 11 届会议论文集（第 2 卷：短论文），2021，第 402–410 页。[39] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu，MPNet：用于语言理解的掩码与置换预训练，《神经信息处理系统进展》33 (2020) 16857–16867。[40] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean，词语和短语的分布式表征及其组合性，《神经信息处理系统进展》26 (2013)。[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polo-sukhin，注意力机制即一切所需，arXiv (2017)。URL：https://arxiv.org/abs/1706.03762。doi：10.48550/ARXIV.1706.03762。[42] I. Loshchilov, F. Hutter，解耦权重衰减正则化 (2017)。URL：https://arxiv.org/abs/1711.05101。doi：10.48550/ARXIV.1711.05101。[43] S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A. Reganti, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar，Factify 2 研究成果：多模态假新闻检测，载于：Defactify 2 论文集：第二届多模态事实核查与仇恨言论检测研讨会，CEUR，2023。[44] M. Glockner, Y. Hou, I. Gurevych，缺失反证使 NLP 事实核查无法真实反映错误信息，arXiv 预印本 arXiv:2210.13865 (2022)。