计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330967
61(5): 1250−1260, 2024

基于大语言模型隐含语义增强的细粒度虚假新闻检测方法

柯 婧[1] 谢哲勇[2] 徐 童[1] 陈宇豪[3] 廖祥文[3] 陈恩红[1]
[1](中国科学技术大学大数据学院 合肥 230026)
[2](中国科学技术大学计算机科学与技术学院 合肥 230027)
[3](福州大学计算机与大数据学院 福州 350108)
(kejing@mail.ustc.edu.cn)

An Implicit Semantic Enhanced Fine-Grained Fake News Detection Method Based on Large Language Models

Ke Jing[1], Xie Zheyong[2], Xu Tong[1], Chen Yuhao[3], Liao Xiangwen[3], and Chen Enhong[1]
[1](*School of Data Science*, *University of Science and Technology of China*, *Hefei* 230026)
[2](*School of Computer Science and Technology*, *University of Science and Technology of China*, *Hefei* 230027)
[3](*College of Computer and Data Science*, *Fuzhou University*, *Fuzhou* 350108)

Abstract The advancement of generative artificial intelligence technology has significantly contributed to the progress in various fields. However, this technological development has also inadvertently facilitated the creation and widespread dissemination of misinformation. Prior research has concentrated on addressing grammatical issues, inflammatory content, and other pertinent features by employing deep learning models to characterize and model deceptive elements within fake news content. These approaches not only are lack of the capability to assess the content itself, but also fall short in elucidating the reasons behind the model's classification. Based on the above problems, we propose a fine-grained fake news detection method with implicit semantic enhancement. This method fully utilizes the summarization and reasoning capabilities of the existing generative large language model. The method employs inference based on major events, fine-grained minor events, and implicit information to systematically evaluate the authenticity of news content. This method strategically leverages the full potential of the model by decomposing tasks, thereby not only optimizing its proficiency but also significantly enhancing its prowess in capturing instances of fake news. Simultaneously, it is designed to be interpretable, providing a solid foundation for detection. With its inherent ability, this method not only ensures reliable identification but also holds vast potential for diverse applications.

Key words social media; fake news detection; large language models; event extraction; knowledge enhancement

摘 要 随着生成式人工智能技术的发展,许多领域都得到了帮助与发展,但与此同时虚假信息的构建与传播变得更加简单,虚假信息的检测也随之难度增加.先前的工作主要聚焦于语法问题、内容煽动性等方面的特点,利用深度学习模型对虚假新闻内容进行建模.这样的方式不仅缺乏对内容本身的判断,还无法回溯模型的判别原因.针对上述问题提出一种基于大语言模型隐含语义增强的细粒度虚假新闻检测方法.该方法充分挖掘并利用了现有的生成式大语言模型所具有的总结与推理能力,按照主干事件、细粒度次要事件和隐含信息推理的顺序进行层级式推导,逐步判别新闻的真实性.通过分解任务的方式,该方法最

收稿日期: 2023-12-01; **修回日期**: 2024-03-08
基金项目: 国家自然科学基金项目(62222213, U22B2059, 61976054)
This work was supported by the National Natural Science Foundation of China (62222213, U22B2059, 61976054).
通信作者: 徐童(tongxu@ustc.edu.cn)

大程度发挥了模型的能力，提高了对虚假新闻的捕获能力，同时该方法也具有一定的可解释性，能够为检测提供判别依据.

关键词 社交媒体；虚假新闻检测；大语言模型；事件抽取；知识增强

中图法分类号 TP391

在当前社交媒体和生成式人工智能飞速发展的背景下，信息的创建和分享变得异常便捷，同时也导致信息量呈指数级增长. 不仅如此，由于技术的成熟所带来的条件，部分个体开始滥用生成式人工智能的强大创造能力以及社交媒体的快速传播渠道，大量制造和传播虚假新闻. 这种虚假信息在社交媒体上的逐步泛滥[1]，进一步加重了人们的信息负担，甚至引发了误导，激起了恐慌心理，对社会造成了严重的负面影响[2-3]. 这样的现象要求社会对于虚假信息的监管不断加强，通过人工审核、打击追查等方式来降低这些问题带来的负面影响. 因此，为了降低上述过程所带来的人力成本，也为了对虚假信息进行更好地监管，许多研究者开始关注虚假新闻内容检测这一研究方向，并在这个问题上提出各类型的解决方案，试图采用某些方法来替代人工审核，从而规避高额代价.

由于虚假新闻本身作为内容篡改的产物所具有的特质以及其煽动性目的，大量虚假新闻在文本内容上表现出较为明显的脱节与错误，并且通常具有强烈的情感色彩和立场倾向[1]. 因此在先前的研究中，大量研究者主要聚焦于如何依据流畅性或情绪倾向等问题对虚假新闻进行检测. 然而，随着技术水平的进步，内容生成的效果变得更加理想，通过自动化技术生成的语句更加流畅，也具有更加接近于人类的表述风格. 这些变化导致了先前的方案受到了一定程度的影响. 此外，由于先前的虚假新闻检测受限于模型规模，大部分工作更加侧重于对较短内容的新闻进行有效检测，而无法适应真实环境中新闻具有较长篇幅的特点，并且缺乏对于长新闻文本分阶段分析的能力. 在这样的背景下，部分研究者[4-5]开始转向利用大模型所具有的强大推理能力与内容理解能力来对新闻进行直接判别，并进行了相关能力的测评. 在测评内容中可以看到，模型具有一定程度的判别能力，但是仅通过零样本推理的方式利用大模型进行虚假新闻判别仍然具有较大的提升空间.

为了解决上述问题，也为了进一步增强检测过程中的泛用性与解释性，我们提出了一套基于大语言模型隐含语义增强的细粒度虚假新闻检测方法. 具体来说，该方法可以划分为 3 个步骤：

1）通过符合新闻事实[6-7]的模板化方式指导模型对新闻内容进行全局主干事件与细粒度次要事件的提取，得到事件的规范化描述. 并且基于事件描述，再次利用多次询问的方式通过大模型推导出事件背后的隐含假设与倾向性信息.

2）在事件描述的基础上，进一步通过检索的方式引入外部知识，作为新闻事件相关事实的外部信息补充.

3）通过层级递进式的方式进行判别，依照主干事件、若干次要事件与隐含信息的顺序进行判别. 对于每一个事件，模型将会通过当前事件、相关外部知识与先前判别真实事件作为基础对当前内容的合法性与正确性进行判别. 若判断为虚假，则得到一条虚假判别新闻，反之则当前事件将会作为事实基础与下一待判别事件的外部知识一同提供给模型进行细粒度判断. 以此方式对于每个判断成功的事实都会成为后续判别的事实基础，直到最终得出错误或完全正确信息.

通过上述步骤，我们可以对于一个新闻内容进行有效合理的拆解，并给出判断. 以一条声称“某中学补充一类新基础课程”的虚假新闻为例，该新闻中的主干事件为学校在某时间补充课程，即新闻最核心的内容. 虚假新闻在该层级往往逻辑自洽，因而仅使用主干事件难以判断新闻真实性. 次要事件，如学校的具体位置等细节，虽然不构成新闻的核心，但它们在验证新闻真实性时仍具有潜在价值. 为了进一步增强新闻真实性的评估，我们引入了第 3 层次的分析，即隐含信息. 这些信息并未直接在新闻文本中提及，但可以通过结合外部知识进行推理得出. 针对例子中的课程，结合外部知识可以推理出该课程一般是面向高年级同学，不属于基础课程，这就为我们判断虚假新闻提供了一个有力的辅助线索. 通过上述 3 轮不同层次的递进式判断，该方法有效地激发大模型的各项能力，在充足信息条件下对新闻真实性做出合理推断，而无需依赖额外的细粒度信息. 我们的方法在相关数据集上进行了实验，实验结果也表明了我们的方法具有更加优越的性能，能够更好

地对虚假新闻进行召回.

本文的贡献点主要有3个方面：

1)提出了一种基于大语言模型隐含语义增强的细粒度虚假新闻检测方法，可以依靠大语言模型能力通过3个分支设计去对新闻内容中不同的3个角度进行有效建模来捕获虚假新闻，发挥了大语言模型的抽取与推理能力，解决其零样本推理结果不理想与外部信息获取受限的问题，从而提升了整体性能.

2)提出了通过将新闻内容划分为主干事件、若干细粒度次要事件，并且进一步推导得到背后隐含信息的技术框架，为虚假新闻判别提供了更丰富的视角.

3)在真实世界的新闻数据集上对本文的方法进行验证，与当前较好方法相比，本文模型能够有效提高虚假新闻的性能，在FakeNewsAMT数据集上$F1$分数与召回率分别有5个百分点与13个百分点的提升.

1 相关工作

在早期，虚假新闻的构建方式主要基于人为构建以及机器编辑.受限于机器与软件能力，虚假新闻的构建会更容易出现内容上的语法或流畅性等错误.基于上述问题，不少方法主要聚焦于对语法[8]、流畅性以及标点符号[9]等非语义内容的部分进行检查.随着软件技术的不断提升和发展，内容构建更加流畅与合乎逻辑.因此许多研究开始尝试侧重于语义上的检测与发现.其中最直观的做法便是基于虚假新闻与真实新闻在内容上所具有的明显差异去检测，包括煽动性、主观性、特定词语高频出现[10]以及质量更低、冲击力更强等特点.具体而言，可以将这类工作分为2个部分，其中一部分工作基于内容分析[11]，从词典、语法、语义层面和语篇层面进行检测；还有部分工作是基于文本风格[9,12-13]进行检测.在此研究背景下，也有部分工作关注于训练过程中的误差，例如CCD方法[14]针对文本的特征在不同话题领域、不同生成方式下都会有偏差的问题，通过对数学方法的调整来削弱遇到分布外的数据时对模型的性能造成的严重打击.

另一类型的工作则是尝试利用外部知识辅助的方法进行处理.例如有部分工作[15]在新闻内容理解的基础上引入社会背景信息，包括查询新闻创建、发布和传播的来源等内容，间接确定新闻的可信度，或是借助对高阶传播模式[16-18]的分析，确定新闻的真实性.而也有一部分工作[19-20]则尝试通过由大量的人工参与构建的知识库或样例库来辅助真假新闻的分类，以达到更好的分类结果.

为了更进一步提升检测效果，部分工作[21-22]进一步尝试利用深度网络来捕捉虚假新闻的数据特征，加强对于特定语义语法等内容的检测.也有工作[18]利用GRU与CNN模型的能力对虚假新闻的传播路径进行分析和捕捉以达到对虚假新闻检测的效果.同样也有工作结合知识补充与深度学习方法来强化对于虚假检测过程中内容的判别，例如KAN[23]就是通过利用外部补充的实体信息来对新闻内容进行补充，使得模型能够检测新闻的虚假性.

然而上述方法仍存在一些问题.一方面，由于网络的不可解释性以及对数据集的依赖，可能仅能给出虚假内容与否的判别，而无法提供判别依据[24].另一方面，知识库参与的方法具有成本高、耗时长的问题.更严重的是，随着大语言模型的出现，通过人工智能进行内容生成的效果更加理想，先前的方案逐渐失效，无法有效地对内容进行检查.

因此，近期研究开始转向引入大语言模型作为检测流程的一环，尝试使用大语言模型具备的特征抽取能力[25]，或是利用大语言模型进行直接推理[4-5]并给出判别结果.但上述方法仍然具有一定的限制.为了更好地对虚假新闻进行检测，在此基础上，我们的方法考虑将新闻内容进行不同层次的划分，并通过全局、局部与隐含信息3个角度对新闻内容进行建模，以此获得可靠的结果以及解释.

2 隐含语义增强的细粒度虚假新闻检测方法

我们的任务是对一条完整新闻的真假与否进行判别.为了进一步充分发挥大语言模型抽取与推理能力的优势并用于解决虚假新闻检测问题，我们依据新闻的叙述性质，将新闻整体划分为主干事件、细粒度次要事件与隐含信息3个部分，并且在每次处理过程中都依靠上一步的真实事实内容作为依据逐步处理，最终以3部分分支所得到的判别结果为依据，给出新闻是否虚假的最终判定.该方法得到的结果能够更好地捕获虚假新闻中的错误内容、定位错误位置，为虚假新闻判别提供有力支撑.本节将详细介绍基本的定义与各个模块的具体结构，完整结构如图1所示.

2.1 基本定义

对于虚假新闻判别，我们的主要目标是通过内部概念矛盾、外部知识冲突、逻辑错误等方式去判别整个新闻是否存在部分内容与真实情况不符，来

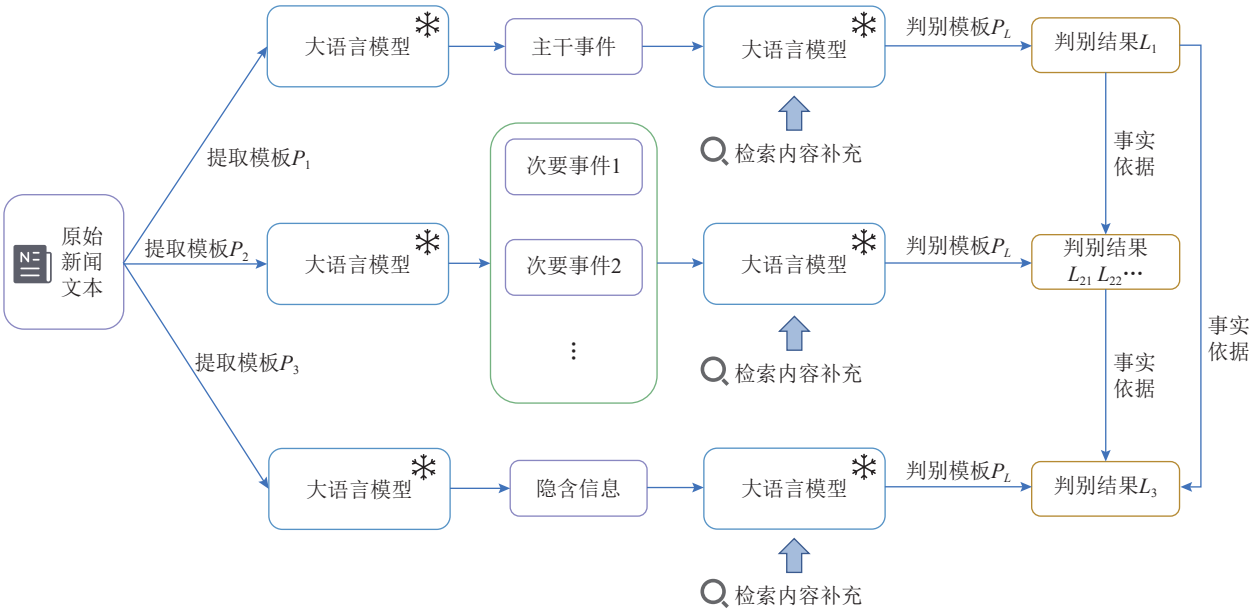


Fig. 1 Illustration of implicit semantic enhanced fine-grained fake news detection method

图 1 隐含语义增强的细粒度虚假新闻检测方法示意图

得到新闻是否捏造的一个判别结果. 具体来说, 我们可以对虚假新闻判别进行如下定义.

定义 1. 虚假新闻判别. 给定一个字符串序列 $T=(t_1, t_2, \cdots, t_n)$, 其中 t_i 代表字符串序列中的某一个字符, 而 n 代表字符序列的整体长度. 该任务的目标是为了得到序列 T 的标签 L, 来指示该新闻是否虚假.

2.2 整体框架介绍

隐含语义增强的细粒度虚假检测方法整体框架如图 1 所示. 首先对于不同分支的内容, 我们基于大语言模型通过新闻六要素的方式提取主干事件、若干个细粒度次要事件和隐含信息的主要内容. 在此基础上, 我们对不同部分的内容进行检索, 以此补充外部知识. 在获得事件的主要内容与外部知识后, 我们依照主干事件、次要事件以及隐含事件的顺序依次进行判别, 并将前序判别为正确的事实内容作为额外补充知识用于后续的判别过程. 最终我们在某一分支得到虚假或全分支真实的结果后, 给出方法的最终判别结果. 通过该方式能够有效捕获新闻虚假内容, 避免错误信息遗漏. 接下来我们将从主干事件判别、细粒度信息判别以及隐含信息推理 3 个部分分别进行介绍.

2.3 主干事件判别

新闻作为一种具有实时性的事实信息传播方式, 一直在各类媒体平台上频繁出现, 并且传播某一些特定的事实. 而也正是因为新闻本身的事实描述的特性, 一般性的新闻都具有一个完整的事件内容贯穿于整个新闻文本中, 作为整个新闻内容的主要脉络与事实基础, 并由这个主干事件来展开及补充较多相关事件与局部细节, 最终有效地描述整个事件发生的细节, 并向读者完整地构建整个事件发生的全过程. 基于上述发现, 本文所提出的虚假新闻检测方法的第 1 步即为主干事件的抽取与判别, 判别正确的主干事件也将作为后续细粒度信息与隐含信息判别的基础, 具有较为重要的作用. 为了更好地抽取新闻内容中的事件, 我们利用大语言模型所具有的强大总结能力作为基础, 设计了一套更加合理的提取方式来针对新闻内容进行提取. 受到新闻事件原本的形式以及新闻信息抽取工作[6-7]的启发, 我们在原先的总结提示模板中额外添加了“5W1H”抽取内容. 具体来说, 我们要求大语言模型在抽取具体信息时, 按照何人(Who)、何时(When)、何事(What)、何地(Where)、为何(Why)及如何(How)这 6 个部分的结构对文本内容进行归纳整理, 并且在此基础上将内容进行合并总结来得到最后的结果.

特别地, 假设输入的初始文本为 $T=(t_1, t_2, \cdots, t_n)$, 大语言模型为 ϕ, 所使用的额外提示词信息模板为 $P_1=(p_1, p_2, \cdots, p_m)$, 抽取获得的主干事件为 $T_1=(t'_1, t'_2, \cdots, t'_{n'})$, 则有

$$T_1 = \phi(P_1, T, \theta),$$

其中 θ 为大语言模型参数, 该参数在使用过程中保持不变. 通过上述变换, 我们最终得到第 1 部分的主干事件抽取.

在获得主干事件之后, 我们需要对主干事件进行细致判别来确定其是否具有因为虚假信息导致的

矛盾. 先前的工作主要集中于语法或情感极性等角度，然而从语义层面分析，虚假新闻的主要问题集中于2个部分：内部矛盾和外部矛盾. 内部矛盾通常为新闻内容的某些实体在特定场景下不能共同出现，或是在逻辑关系上不能匹配，例如在内容中不具备特定背景的相关信息的条件下，北极熊出现在赤道环境中就是一个明显的错误. 因此内部矛盾主要集中于对内部不同部分的推理.

而外部矛盾则将更为复杂，外部矛盾是当前新闻信息与网络中的常识内容出现了矛盾，例如文中出现了某位领导人出现在国内，而网络信息均显示他出发探访他国，那么该新闻为虚假新闻的可能性将大大增加. 为了有效地同时捕获这2种可能的错误类型，我们通过结合大语言模型自身所具有的推理能力与搜索引擎的检索能力，来对抽取得到的主干事件进行处理. 首先使用主干事件中的部分事实以及全部文本对网络信息进行检索，以此来对当前新闻内容的相关事实进行扩展，之后将新闻信息与原先的主干实体合并后共同输入大语言模型，并利用提示词来对内容进行分类，支持模型根据补充信息对当前主干事件进行判别，以获取2个部分可能的错误问题和判别结果. 假设所收集到的检索信息为 $T_{s1}=(t_1'', t_2'', \cdots, t_{n''}'')$，对应的判别提示词为 $P_L=(p_1, p_2, \cdots, p_{m'})$，同时我们会在判别提示词中加入CoT方法的提示，来进一步提升大语言模型的推理能力，则主干事件的判别结果为

$$L_1 = \phi(P_L, T_1, T_{s1}, \theta),$$

这一判别结果将继续支持后续模块的判别过程.

2.4 细粒度信息判别

与主干事件相对应的就是长文本新闻所具有的基于主干事件扩展得到的细粒度信息. 这一部分细粒度信息在很大程度上影响新闻整体的内容、立场、倾向以及相对应的事件细节，也会影响到读者对于事实的判断和理解. 因此大量虚假新闻也会采取在主干事件不变的情况下，通过对细节部分的篡改来使得新闻内容发生较大变化. 因此为了检测这类更细粒度的虚假信息，我们同样需要对每个部分的信息内容进行细致判别. 基于上述动机，我们设计了细粒度信息判别分支，如图1所示，该分支采用与主干事件类似的5W1H信息提取方法对信息内容进行提取，不同的是我们将会提供主干事件的描述，以此来排除原先已经获取到的事实信息，防止信息重复，并使得模型能够有效抽取出其余次要事件. 受到CoT方法[26]的启发，此处我们也采用多轮询问的方式来有效释放大语言模型在内容理解上的潜能，并且在每次给出一个次要事件后，反复使用3次终止条件的询问模板来确定信息抽取完毕，并得到对应的所有次要事件. 具体而言，假设次要事件提示词为 P_2，抽取得到第 i 个次要事件为 T_{2i}，第 i 次询问是否终止的结果为 F_{2i}，F_c 表示非终止信息集合. 因此对于每次抽取有

$$T_{2i} = \phi(P_2, T, T_1, T_{21}, T_{22}, \cdots, T_{2(i-1)}, \theta),\ \text{当}\ F_{2i} \in F_c,$$

其中 $F_{2i} = \phi(P_1, T, T_1, T_{21}, T_{22}, \cdots, T_{2(i-1)}, \theta)$. 通过上述方法，我们可以有效抽取到一组次要信息描述内容. 与主干事件判别部分相同，我们也采用类似的判别提示词 P_L 对每一个子事件进行判别，并且在判别前添加每个部分的检索补充内容 T_{s2i}，得到每一组次要事件的判别结果：

$$L_{2i} = \phi(P_L, T_{2i}, T_{s2i}, \theta),\ 1 \leqslant i \leqslant n_2,$$

其中 n_2 表示最终收集到的次要事件数量.

2.5 隐含信息推理

除了上述主干事件与细粒度次要事件的内容外，许多虚假新闻也会在文字中通过表达上的情感色彩、语言倾向性与隐含假设，使得文本含有对某些观点或事实的假设与暗示. 例如在某一新闻中提及课程普及带来的好处，则这样的新闻明显具有某一课程已经得到推广的内容暗示以及对这一课程的赞许. 这类虚假信息由于其隐蔽性高、逻辑复杂的特点而难以检测，从而导致检测性能受到影响.

为了解决上述问题，我们依据大语言模型在补充事实信息情况下多次询问后能够依据事实给出暗示或假设性隐含信息的特点，构建了用于虚假新闻判别的第3个分支：隐含信息推理. 这个部分将利用大语言模型所具有的隐含信息推理能力，来对新闻背后的倾向与假设进行显示表述，并依靠这些表述来给判别过程提供更多视角，避免虚假信息的遗漏，以此来进一步提升模型的检验能力. 具体而言，我们通过提示词 P_3 与提取得到的主干事件和次要事件作为输入，来迫使大语言模型给出对事件内容背后暗示内容 T_3，并同时加入暗示内容所检索得到的补充内容 T_{s3}，通过这2部分的信息，最终利用判别提示词得到最后的判别结果 L_3，用公式表述为

$$T_3 = \phi(P_3, T, T_1, T_{21}, T_{22}, \cdots, \theta),$$

$$L_3 = \phi(P_L, T_3, T_{s3}, \theta),$$

其中 ϕ 为大语言模型，θ 为大语言模型参数.

得到上述所有分支的结果后，若新闻中出现了1个及以上判别部分为虚假结果，则该新闻将被判定

为虚假内容，并且也将获得对应虚假部分的具体信息. 若所有分支的事实表述均未判定为虚假，则该新闻将被判定为真实合法新闻.

3 实验与分析

3.1 数据集介绍

本文采用了 2 个不同的数据集来进行方法的检验.

1）FakeNewsAMT[27]. 该数据集通过众包收集，涵盖了体育、商业、娱乐、政治、技术和教育 6 个新闻领域. 新闻内容取自主流新闻网站，例如 ABCNews、CNN、美国今日新闻、纽约时报、福克斯新闻、彭博社和 CNET 等. 数据集中包括 240 条真实合法新闻与 240 条虚假新闻，真实合法新闻与虚假新闻数量较为平衡.

2）Snopes[28]. 该数据集是通过收集知名事实核查 Snopes 网站上的经过事实核查的新闻文章实例所获得. Snope 事实核查网站是专家编辑团队运营的核查网站，除了对虚假事实的发现，该网站还提供了详尽的解释. 该网站给出的事实核查内容均包含有细粒度的虚假、真实、混合等标签，此处我们随机选取了其中各 500 条真实合法新闻与虚假新闻作为数据集用于评估各对比模型的能力.

为了方法的一般性考虑，本文所选取的上述数据集均不具有强烈时效性，亦未针对时效性问题进行专门收集，因而适用于一般条件下涉及检索进行外部知识补充的方法.相关数据集指标如表 1 所示.

Table 1 Statistics of Datasets

表 1 数据集统计指标

数据集	虚假新闻	真实新闻	涵盖话题
FakeNewsAMT	240	240	体育、商业、娱乐、政治、技术和教育等
Snopes	500	500	体育、商业、娱乐、政治、技术和教育等

3.2 实验设置与评价指标

本文使用准确率(accuracy)以及假新闻类别上的 $F1$ 值、精确率(precision)及召回率(recall)作为评估指标. gpt-3.5-turbo 模型有较强的语言能力，相较于 LLaMA 模型及其他模型，能更精准地提取和总结事件，因而本文使用 gpt-3.5-turbo 模型提取主干事件、细粒度信息以及隐含信息. 设置生成句子最长 token 数为 2 048.

为了验证模型是否存在信息泄露，我们设计并进行了一个初步实验. 我们尝试选取 FakeNewsAMT 数据集的 240 条真实新闻，首先使用第 1 种方案：gpt-3.5-turbo 模型直接对新闻真实性进行判断；而后使用第 2 种方案：通过将真实新闻中部分名词进行随机替换的方式构造同等数量的假新闻，再使用同样方式判别. 后一种方案相较前一种方案，准确率下降了 75 个百分点. 相对的，采取 LLaMA2-13B 模型就不存在准确率下降的现象，这反映了 gpt-3.5-turbo 模型可能存在标签泄露的问题. 为了更进一步验证上述情况，我们人工采集了数据集外较新的真假新闻各 40 条作为域外数据参与到判别中. 相较于域内数据，域外数据的判别准确率下降了 15 个百分点，是 LLaMA2-13B 结果变动的 2 倍. 该实验结果表明，对于上述 2 个数据集，gpt-3.5-turbo 很可能存在标签泄露问题. 因此，为了更加公平地评测方法，尽量规避可能存在的标签泄露问题，本文选取 LLaMA2-13B 作为判别部分所使用的模型.

对于主干事件与细粒度信息提取部分，我们采用如 2.3 节与 2.4 节所述的方法对信息进行 6 部分要素分别提取再整合. 对于主干事件，我们采用提示模板为“You are a journalist. Given the following news, describe the six major news events elements in a short sentence: Who, What, When, Where, Why, and How.For those not explained in the text, output 'Not stated in the text'. \nnews:”进行 6 要素的划分，同时对于次要事件也采用类似模板，排除原始新闻的主干事件进行划分. 在获得 6 要素后，我们采用“There are six elements to the statement: Who, What, When, Where, Why, and How. Please revert the following statement to a single sentence, ignoring the 'Not stated in the text' message in the six elements.\nStatement:”作为提示词，将主干事件和次要事件各自的 6 要素分别整合为一句话. 而对于隐含事件提取，我们采用的提示模板为“Apart from the related information within the summarization above, is there any message you can from the original news?”以激发模型对新闻事件深层次内涵的发掘. 在检索部分，我们受文献 [5] 的启发，将主干事件、次要事件及隐含信息分别作为搜索引擎的输入进行内容检索. 为了方法的通用性，我们选择的检索引擎不针对时效性问题，目的是通过检索补充外部知识. 此处我们采用谷歌搜索①作为检索部分使用的搜索引擎.

① https://www.google.com/search

同时由于部分对比方法需要进行一定程度的训练，因此对于需要训练的方法，我们参考文献 [27] 统一使用五折交叉验证的方式，将每个数据集划分成 5 份进行交叉验证，从而确保与全量验证的方法设置上的一致性.

3.3 虚假新闻检测性能对比

我们依照文献 [24] 的做法，将本文方法与 3 种传统方法、2 种深度学习方法进行性能对比，如表 2 所示.

Table 2 Performance Comparison of Different Methods

表 2 不同方法的性能比较

数据集	方法	准确率	*F*1 值	精确率	召回率
FakeNewsAMT	LIWC-Summary	0.61	0.64	0.60	0.68
	LIWC-Linguistic processes	0.67	0.66	**0.67**	0.66
	LIWC-Psychological processes	0.56	0.55	0.56	0.56
	Ngrams	0.62	0.62	0.62	0.63
	Syntax	0.65	0.65	0.64	0.67
	BERT	0.53	0.55	0.53	0.58
	RoBERTa	0.51	0.50	0.51	0.50
	LLaMA2-Zero-Shot	0.65	0.68	0.62	0.75
	LLaMA2-Zero-Shot + CoT	0.66	0.70	0.62	0.78
	本文方法	**0.70**	**0.75**	0.64	**0.91**
Sonpes	LIWC-Summary	0.51	0.44	0.54	0.41
	LIWC-Linguistic processes	0.50	0.38	0.55	0.42
	LIWC-Psychological processes	0.49	0.38	0.50	0.42
	Ngrams	0.50	0.19	**0.63**	0.23
	Syntax	0.51	0.27	0.63	0.25
	BERT	0.52	0.55	0.51	0.59
	RoBERTa	0.47	0.49	0.47	0.51
	LLaMA2-Zero-Shot	0.55	0.55	0.55	0.54
	LLaMA2-Zero-Shot + CoT	**0.56**	0.56	0.56	0.56
	本文方法	0.53	**0.63**	0.52	**0.81**

注：黑体数值是最优结果.

1）LIWC[29]. 该方法在先前语言相关的虚假检测中发挥较大的作用. 此处选取了摘要类别（summary）、语言过程（linguistic processes）和心理过程（psychological processes）作为具体的对比方案.

2）Syntax[30]. 使用斯坦福解析器提取了一组基于上下文无关语法的规则，包括所有词组的生产规则（规则包括子节点）以及它们的父节点和祖先节点，这些特征也被编码为 TF-IDF 值.

3）Ngrams.从每篇新闻文章的单词表示包中提取单字组和双字组，并将特征编码为 TF-IDF 值以解决内容长度的偶然差异.

4）BERT.基于 BERT 的模型通常被应用于虚假新闻检测[31]，特别被多次应用于关于 COVID-19 的相关虚假新闻检测[32-33]. BERT 模型基于大量数据的预训练，在多种任务上均取得良好的结果. 此处使用预训练的 BERT-based-uncased 模型①.

5）RoBERTa.作为 BERT 模型的改良版本，其也被应用于虚假新闻检测[33]，并取得了良好的结果. 此处采用标准 RoBERTa-based 模型②进行对比.

同时为了更好地证明本文方法的有效性，我们也与 2 种常用大模型方法进行了比较：

1）LLaMA-Zero-Shot.使用问答模板对 LLaMA 模型进行直接询问以得到判别结果.

2）LLaMA-CoT[26]. 使用模板要求模型首先输出一个思维链，再给出判别结果.

从表 2 可以看到，本文方法在准确率上取得了比较好的结果，高于大部分基础方法，同时在保持准确率领先的前提下，依靠多分支归纳的能力，取得了更加良好的召回率与 *F*1 值. 这也说明了本文方法能够有效地召回虚假信息，避免虚假信息的遗漏，在现实场景中将具有更加良好的应用空间.

同时，相比起直接利用大模型的 Zero-Shot 方式，本文方法能够更有效地对数据进行合理划分与判断，使得整体模型的效果都得到了提升，也能有效地针对虚假新闻进行不同程度的判别，增强了整体的推理性能，得到了更好的检测结果.

3.4 消融实验

为了验证本文方法中不同组件对于实验结果的影响，我们设计了 3 种模型的变体，在 FakeNewsAMT 数据集上进行消融实验，结果如表 3 所示.

Table 3 Ablation Experiment

表 3 消融实验

方法	准确率	*F*1 值	精确率	召回率
本文方法	**0.70**	**0.75**	**0.64**	0.91
去除隐含信息	0.68	0.73	0.64	0.84
去除隐含信息以及细粒度信息	0.7	0.71	0.69	0.71
LLaMA-Zero-Shot	0.65	0.68	0.62	0.75
Ours-LLaMA2	0.61	0.71	0.57	**0.93**

注：黑体数值是最优结果.

① https://huggingface.co/bert-base-uncased

② https://huggingface.co/roberta-base

1)去除隐含信息分析模块. 仅保留主干事件和细粒度次要事件信息的抽取与判别, 以此来验证隐含信息部分对于虚假新闻检测的影响.

2)去除隐含信息以及细粒度次要事件信息. 仅使用从原新闻中提取的主干事件进行检测, 以此来验证细粒度次要事件对最终结果判别的影响.

3)LLaMA-Zero-Shot 方法. 使用 LLaMA2-13B 在没有额外信息提供的前提下, 直接对原新闻文本的真假进行判断. 通过该方法的消融可以进一步对比相对于基础大模型是否能够更充分地利用信息并获得更准确的效果.

4)Ours-LLaMA2 方法. 使用 LLaMA2-13B 模型替换 gpt-3.5-turbo 作为事件提取模型. 由于 LLaMA2-13B 模型对新闻总结能力的限制, 对事件提取的准确率产生了负面影响, 导致最终结果有一定的下降.

从表 3 可以看到, 不同模块的去除对于各个指标都有一定的影响, 这充分证明了各个模块对于整体方法的重要性; 隐含信息与细粒度信息对于召回率的提升有较大的影响, 说明通过 2 个非主干事件的信息补充能够更好地捕获缺失的虚假新闻, 避免缺漏问题的出现.

3.5 归因结果分析

为了更好地说明本文方法不同分支的处理情况, 将本文方法在 FakeNewsAMT 数据集上不同分支中的判别情况进行了统计. 如表 4 所示, 可以发现主干事件能够捕获最多的虚假信息, 而随着模块复杂程度的提升捕获的虚假新闻数量也逐渐变少. 这是由于大部分虚假新闻的构建都是基于主干事件的变化, 并且借由主干事件就可以得到正确判别. 随着信息的粒度与语义层级的提升, 虚假信息的构建也更加困难复杂, 数量也会逐渐减少, 但在该模型框架下仍然能够正确检测出错误的新闻. 这样的实验结果也表明本文方法能够成功对不同粒度和语义层级的虚假内容进行检测, 得到正确的判别结果, 并且通过 3 个不同分支的划分与收集, 能够有效地帮助模型提高虚假新闻召回的数量, 保证模型的有效性.

Table 4 Number of Fake News Detected by Each Branch

表 4 各分支检测出的假新闻数量

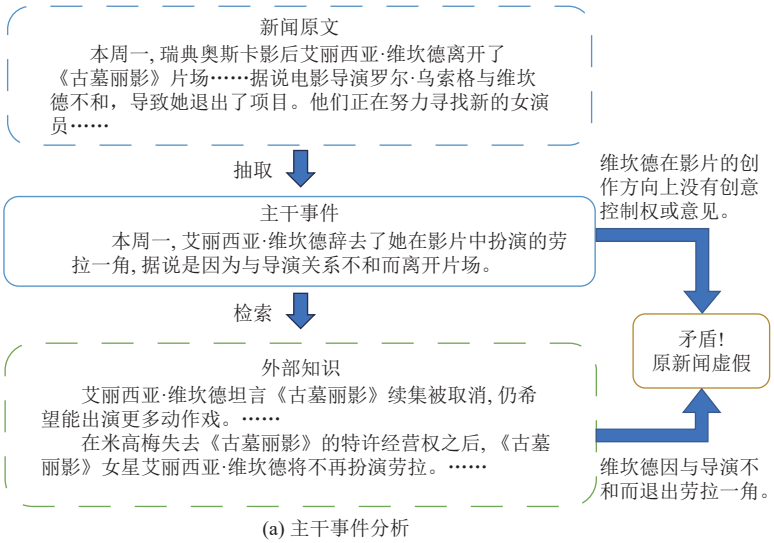
判别分支	主干事件	细粒度信息	隐含信息	未检出
假新闻数量	177	39	2	22

3.6 样例分析

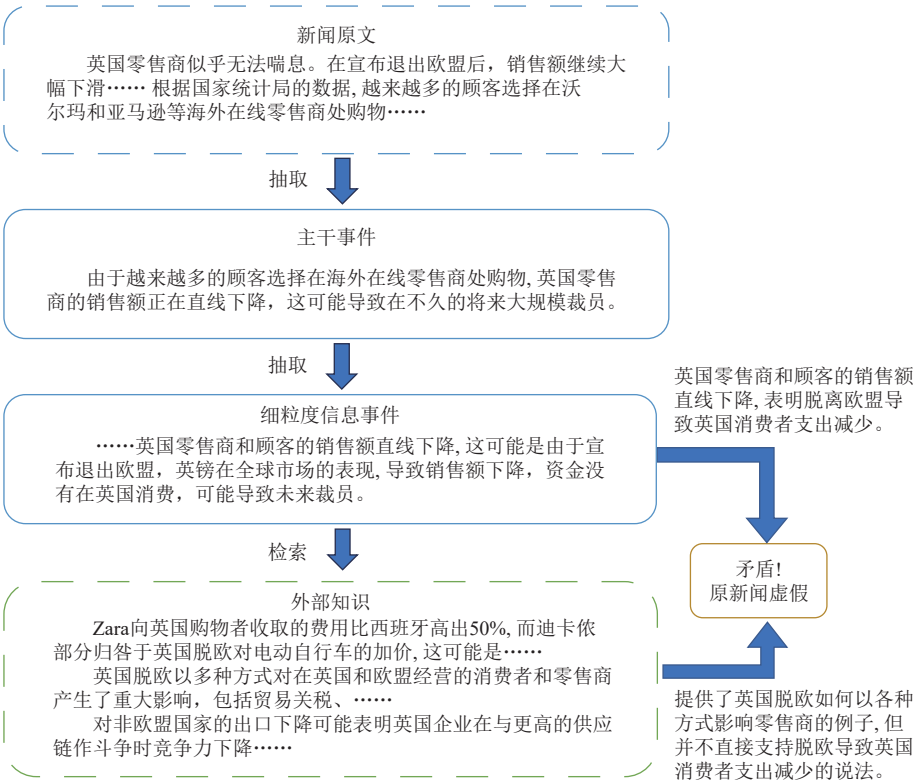
为了更好地说明和展示本文方法, 我们选取具有代表性的样例, 如图 2 所示. 图 2(a)中, 通过主干事件提取, 可以直接清晰地抓取新闻的主体, 即维坎德由于和导演的矛盾不再参与《古墓丽影》. 然而多条检索信息中有关维坎德离开《古墓丽影》原因的解释为其公司失去了《古墓丽影》的特许经营权. 新闻主题与检索信息之间发生冲突, 推断新闻是主体部分虚假的假新闻.

图 2(b)中, 无法仅依靠主干事件判断新闻真假, 故进一步提取细粒度的信息. 细粒度信息提供了“英国脱欧致消费者支出减少”等具体信息, 而检索信息不直接支持这一信息. 本例由新闻细节处判断出其违背逻辑, 是假新闻.

通过图 2 例子的分析可知, 本文方法能够正确地将虚假信息进行捕获, 并给出错误位置与相关信息, 证明本文方法在真实环境下的有效性与可用性.



(a) 主干事件分析



(b) 次要事件分析

Fig. 2 Our case analysis

图 2 本文样例分析

4 总 结

针对大语言模型背景下现有的虚假新闻检测方法的不足，本文提出了一种隐含语义增强的细粒度虚假新闻检测方法，该方法能够对新闻文本以主干事件、次要事件以及隐含语义 3 个角度，从全局到细粒度再到隐藏含义逐步地分析新闻中所存在的虚假信息，充分利用大语言模型所具有的总结能力与推理能力，实现具有一定解释性的虚假新闻检测方法. 实验结果表明，本文方法在准确率和召回率等各项指标上均有明显提升，也证明了基于大语言模型隐含语义增强的细粒度虚假新闻检测方法的有效性. 在数据收集的过程中我们注意到，针对具备时效性的虚假新闻进行检测也是一个值得探究的问题. 未来我们希望能够有机会利用相关的数据针对这一问题进一步研究.

作者贡献声明：柯婧和谢哲勇提出研究思路，设计方案，完成实验数据以及论文撰写；徐童提出研究思路，设计方案，完成最终论文修订；陈宇豪完成实验数据的获取；廖祥文完成最终论文修订；陈恩红完成最终论文修订. 柯婧和谢哲勇具有相同贡献.

参 考 文 献

[1] Maigrot C, Kijak E, Claveau V. Détection de fausses informations dans les réseaux sociaux: l'utilité des fusions de connaissances[C]//COnférence Recherche d'Information et Applications. Narbonne, France: ARIA, 2017: 107−122

[2] Roets A. ‘Fake news’: Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions[J]. Intelligence, 2017, 65: 107−110

[3] Yang Yuzhou, Zhou Yangming, Ying Qichao, et al. Fact-checking based fake news detection: A review[J]. arXiv preprint, arXiv: 2401.01717, 2024(in Chinese)
(杨昱洲，周杨铭，应祺超，等. 基于事实信息核查的虚假新闻检测综述[J]. arXiv preprints, arXiv: 2401.01717, 2024)

[4] Lucas J, Uchendu A, Yamashita M, et al. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 14279−14305

[5] Zhang Xuan, Gao Wei. Towards LLM-based fact verification on news

claims with a hierarchical step-by-step prompting method[J]. arXiv preprint, arXiv: 2310.00305, 2023

[6] Hamborg F, Breitinger C, Schubotz M, et al. Extraction of main event descriptors from news articles by answering the journalistic five W and one H questions[C]//Proc of the 18th ACM/IEEE on Joint Conf on Digital Libraries. New York: ACM, 2018: 339−340

[7] Jin Peiquan, Mu Lin, Zheng Lizhou, et al. News feature extraction for events on social network platforms[C]//Proc of the 26th Int Conf on World Wide Web Companion. New York: ACM, 2017: 69−78

[8] Chen Yimin, Conroy N K, Rubin V L. News in an online world: The need for an “automatic crap detector”[J]. Proceedings of the Association for Information Science and Technology, 2015, 52(1): 1−4

[9] Rubin V L, Conroy N, Chen Yimin, et al. Fake news or truth? using satirical cues to detect potentially misleading news[C]//Proc of the second Workshop on Computational Approaches to Deception Detection. Stroudsburg, PA: ACL, 2016: 7−17

[10] Huang Yue, Sun Lichao. Harnessing the power of ChatGPT in fake news: An in-depth exploration in generation, detection and explanation[J]. arXiv preprint, arXiv: 2310.05046, 2023

[11] Zhou Xinyi, Jain A, Phoha V V, et al. Fake news early detection: An interdisciplinary study[J]. arXiv preprint, arXiv: 1904.11679, 2019

[12] Castelo S, Almeida T, Elghafari A, et al. A topic-agnostic approach for identifying fake news pages[C]//In Companion Proc of the World Wide Web Conf. New York: ACM, 2019: 975−980

[13] Potthast M, Kiesel J, Reinartz K, et al. A stylometric inquiry into hyperpartisan and fake news[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2018: 231−240

[14] Chen Ziwei, Hu Linmei, Li Weixin, et al. Causal intervention and counterfactual reasoning for multi-modal fake news detection[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2023: 627−638

[15] Zhou Xinyi, Zafarani R. Network-based fake news detection: A pattern-driven approach[J]. ACM SIGKDD Explorations Newsletter, 2019, 21(2): 48−60

[16] Gupta M, Zhao Peixiang, Han Jiawei. Evaluating event credibility on Twitter[C]//Proc of the 2012 SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2012: 153−164

[17] Wu Ke, Yang Song, Zhu K Q. False rumors detection on Sina Weibo by propagation structures[C]//Proc of 2015 IEEE 31st Int Conf on Data Engineering. Piscataway, NJ : IEEE, 2015: 651−662

[18] Liu Yang, Wu Y F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018, 32(1): 354−361

[19] Yang Zhiwei, Ma Jing, Chen Hechang, et al. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection[J]. arXiv preprint, arXiv: 2209.14642, 2022

[20] Zhou Xinyi, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities[J]. ACM Computing Surveys, 2020, 53(5): 1−40

[21] Wang Yaqing, Ma Fenglong, Jin Zhiwei, et al. Eann: Event adversarial neural networks for multi-modal fake news detection[C]//Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 849−857

[22] Qi Peng, Cao Juan, Sheng Qiang. Semantics-enhanced multi-modal fake news detection[J]. Journal of Computer Research and Development, 2021, 58(7): 1456−1465 (in Chinese)
(亓鹏, 曹娟, 盛强. 语义增强的多模态虚假新闻检测[J]. 计算机研究与发展, 2021, 58(7): 1456−1465)

[23] Dun Yaqian, Tu Kefei, Chen Chen, et al. KAN: Knowledge-aware attention network for fake news detection[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021, 35(1): 81−89

[24] Yao B M, Shah A, Sun Lichao, et al. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models[C]//Proc of the 46th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2023: 2733−2743

[25] Hu Beizhe, Sheng Qiang, Cao Juan, et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection[J]. arXiv preprint, arXiv: 2309.12247, 2023

[26] Wei J, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 24824−24837

[27] Pérez-Rosas V, Kleinberg B, Lefevre A, et al. Automatic detection of fake news[J]. arXiv preprint, arXiv: 1708.07104, 2017

[28] Asr F T, Mokhtari M, Taboada M. Misinformation Detection in News Text: Automatic Methods and Data Limitations[M]//The Routledge Handbook of Discourse and Disinformation. London: Routledge, 79−102

[29] Pennebaker J W, Boyd R L, Jordan K, et al. The development and psychometric properties of LIWC2015[R]. Austin, Texas: The University of Texas at Austin, 2015

[30] Klein D, Manning C D. Accurate unlexicalized parsing[C]//Proc of the 41st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2003: 423−430

[31] Dementieva D, Kuimov M, Panchenko A. Multiverse: Multilingual evidence for fake news detection[J]. Journal of Imaging, 2023, 9(4): 77

[32] Glazkova A, Glazkov M, Trifonov T. g2tmn at Constraint@ AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection[C]//Proc of Int Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation. Berlin: Springer, 2021: 116−127

[33] Gundapu S, Mamidi R. Transformer based automatic COVID-19 fake news detection system[J]. arXiv preprint, arXiv: 2101.00180, 2021

Ke Jing, born in 2001. Master candidate. Her main research interests include data mining and disinformation detection.

柯　婧, 2001 年生. 硕士研究生. 主要研究方向数据挖掘、虚假信息检测.

Xie Zheyong, born in 1999. Master candidate. His main research interests include multimodal text generation and dialogue system.

谢哲勇, 1999 年生. 硕士研究生. 主要研究方向为多模态文本生成、对话系统.

Xu Tong, born in 1988. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data mining and social media analysis.

徐　童, 1988 年生. 博士，教授，博士生导师. CCF 高级会员. 主要研究方向为数据挖掘、社交媒体分析.

Chen Yuhao, born in 2002. Undergraduate. His main research interests include data mining and social media analysis.

陈宇豪, 2002 年生. 本科生. 主要研究方向为数据挖掘、社交媒体分析.

Liao Xiangwen, born in 1980. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include natural language processing, opinion mining, and sentiment analysis.

廖祥文, 1980 年生. 博士，教授，博士生导师. CCF 高级会员. 主要研究方向为自然语言处理、观点挖掘、情感分析.

Chen Enhong, born in 1968. PhD, professor, PhD supervisor. CCF fellow. His main research interests include data mining, personalized recommendation system, and social media analysis.

陈恩红, 1968 年生. 博士，教授，博士生导师. CCF 会士. 主要研究方向为数据挖掘、个性化推荐系统、社交媒体分析.