scientific reports

Check for updates

OPEN

结合 Transformer 和 LLM 进行跨语言情感分析的多模态方法

Md Saef Ullah Miah、Md Mohsin Kabir、Talha Bin Sarwar、Mejdl Safran、Sultan Alfarhood 和 M. F. Mridha

情感分析是自然语言处理中的一项重要任务，涉及识别文本的极性，无论它表达积极、消极还是中性情感。随着社交媒体和互联网的发展，情感分析在营销、政治和客户服务等各个领域变得越来越重要。然而，在处理外语时，情感分析变得具有挑战性，特别是在没有用于训练模型的标记数据的情况下。在这项研究中，我们提出了一个 Transformer 的集成模型和一个大型语言模型 (LLM)，该模型通过将外语翻译成基础语言英语来利用外语的情感分析。我们使用了四种语言：阿拉伯语、中文、法语和意大利语，并使用两种神经机器翻译模型进行翻译：LibreTranslate 和 Google Translate。然后使用预先训练的情感分析模型集合来分析句子的情感：Twitter-Roberta-Base-Sentiment-

最新的 bert-base-multilingual-uncased-sentiment 和 GPT-3，它是来自 OpenAI 的 LLM。我们的实验结果表明，使用所提出的模型对翻译句子进行情感分析的准确性超过 86%，这表明通过翻译成英语进行外语情感分析是可能的，并且所提出的集成模型比独立的预训练模型效果更好LLM。

关键词

跨语言交流、情感分析、神经机器翻译、预训练情感分析模型、LLM 的集成

情感分析是确定文本中情绪基调的计算任务，在过去几十年中已发展成为自然语言处理 (NLP) 的一个关键子领域。它系统地分析文本内容，以确定它是否传达积极、消极或中性的情绪。这种能力对于理解公众舆论、客户反馈和社会话语非常重要，使其成为营销、政治和客户服务等跨领域各种应用的基本原则。情感分析的总体领域经历了指数级增长，这主要是由于数字通信平台的扩展和大量日常文本数据的推动。然而，由于广泛的标记数据集的可用性和复杂语言模型的开发，情感分析的有效性主要以英语得到证明。这在分析非英语语言的情绪时留下了巨大的空白，其中标记的数据通常不足或缺失。

尽管情感分析研究不断增长，但仍然存在一个重要的悬而未决的问题：在没有大量标记数据的情况下，我们如何才能有效地将情感分析技术应用于非英语语言？我们的研究通过提出全面的方法论和实证结果来证明通过翻译进行跨语言情感分析的可行性和准确性，寻求对这个问题有说服力的答案。

美国国际大学孟加拉国分校计算机科学系，孟加拉国达卡 1229。罗兰大学信息学院，布达佩斯 1117，匈牙利。沙特国王大学计算机科学系计算机与信息科学学院在线对话与文化传播研究主席，11543 沙特阿拉伯利雅得。沙特国王大学计算机科学系，计算机与信息科学学院，P.O.沙特阿拉伯利雅得 11543 号信箱 51178。电子邮件：mejdl@ksu.edu.sa；firoz.mridha@aiub.edu

本研究探讨将外语翻译成基础语言英语，以分析文本情感。
对于情感分析，将外语翻译成基础语言（例如英语）有几个优点。这些包括：

- 克服语言障碍 语言障碍对分析外语情绪提出了重大挑战。通过将外文文本翻译成英语等基础语言，分析师可以克服这些障碍并更准确地分析情绪。
- 语言标准化将外语翻译成基本语言可以帮助标准化用于情感分析的语言。这可以减少不同语言中使用的语言的变异性，并使比较不同文本之间的情感变得更容易。
- 情感分析工具的可用性 许多情感分析工具都有英文版本，这使得分析翻译文本中的情感变得更加容易。分析师可以使用这些工具更有效、更准确地分析翻译文本中的情绪。
- 提高准确性 与不考虑语言翻译的传统情感分析方法相比，将外语翻译成基础语言可以提高情感分析的准确性。这是因为语言翻译可以捕捉外语的细微差别，并更容易地将其传达给分析师。

这项研究的主要贡献概述如下：

- 情感分析研究的进展：本研究通过提出并实施一种将外语翻译成英语并分析翻译文本的情感的方法，为现有的情感分析研究做出了贡献。这种方法将情感分析的范围扩展到英语文本之外，并提供了分析各种语言情感的框架。
- 深入了解所提出方法的有效性：这项研究提供了对所开发的情感分析和语言翻译方法的有效性的深入了解。通过介绍该方法并讨论其实施，本研究为外语情感分析的准确性和可靠性提供了有价值的信息。
- 方法论贡献：本研究描述了一种将外语翻译成英语并进行情感分析的方法。这项贡献包括翻译过程和情感分析中使用的技术、算法和工具，为其他研究人员复制或进一步改进提供了框架。
- 研究结果和影响：本研究提出了有关外语情感分析的研究结果并讨论了其影响。这些发现揭示了与不同语言的情感分析相关的挑战、局限性和机遇。这些影响对于自然语言处理、机器学习、社交媒体分析和跨文化交流领域的研究人员和从业者来说非常有价值。
- 研究人员和从业人员的实用见解：本研究旨在提供对各个领域的研究人员和从业人员有用的实用见解。本研究中提出的结果和方法可以指导未来的情感分析和语言翻译研究。此外，寻求在多语言环境中实施情感分析的从业者可以从本研究提供的见解和建议中受益。

通过强调这些贡献，本研究展示了本研究的新颖之处及其对情感分析和语言翻译的潜在影响。
本手稿的后续部分结构如下：在“相关作品”部分，我们深入研究与我们的研究相关的现有研究主体。在“问题表述”部分中，制定问题陈述，而“方法论”部分则概述本研究中采用的方法。在“结果和讨论”部分中介绍了实验结果和随附的讨论。“挑战”部分介绍了当前技术的挑战。最后，我们在“结论和未来的工作”部分得出结论。

相关作品

情绪分析是一项重要的自然语言处理任务，涉及自动检测文本中表达的情绪，区分积极、消极或中性情绪。数字时代使得跨不同领域的情感分析成为可能。尽管如此，用外语进行情感分析，特别是在没有注释数据的情况下，提出了复杂的挑战。虽然传统方法依赖于多语言预训练模型进行迁移学习，但有限的研究探索了利用翻译进行外语情感分析的可能性。大多数研究都集中在使用多语言预训练模型应用迁移学习，但其准确性并未取得显着提高。然而，所提出的将外语文本翻译成英语并随后分析翻译文本中的情感的方法仍然相对未经探索。本节概述了该领域的相关工作，重点介绍了主要集中于多语言预训练模型的迁移学习的现有研究，以及测试所提出的基于翻译的方法的有效性方面的差距。

Salameh 等人的工作。提出了一项使用最先进的阿拉伯语和英语情感分析系统以及阿拉伯语到英语翻译系统对阿拉伯语社交媒体帖子进行情感分析的研究。本研究概述了每种方法的优点和缺点，并进行了实验以确定

使用每种技术获得的情感标签的准确性。结果表明，对阿拉伯文本英文翻译的情感分析产生了有竞争力的结果。该研究还回答了与情感预测准确性、将阿拉伯文本翻译成英语时的可预测性损失以及自动情感分析与人工注释相比的准确性相关的几个研究问题。

中的工作系统地研究了英语翻译，并在情感分析的背景下分析了翻译文本的情感。阿拉伯语社交媒体帖子被用作焦点语言文本的代表性示例。该研究表明，与母语阿拉伯语情感分析相比，阿拉伯语文本英文翻译的情感分析产生了有竞争力的结果。此外，这项研究还证明了阿拉伯语情感分析系统可以通过合并自动翻译的英语情感词典获得切实的好处。此外，这项研究还包括人工注释研究，旨在辨别翻译与源词或文本之间情感差异背后的原因。这项研究具有特别重要的意义，因为它有助于自动翻译系统的发展。这项研究有助于开发最先进的阿拉伯语情感分析系统，创建新的阿拉伯语方言情感词典，并建立第一个阿拉伯语-英语平行语料库。值得注意的是，该语料库由阿拉伯语和英语使用者独立注释，从而为情感分析领域添加了宝贵的资源。

中描述的工作重点是通过机器翻译过程仔细检查情感的保存。为此，我们用英语策划了一个情绪黄金标准语料库，其中包含本土金融专家的注释。随后，使用人工翻译人员和三个不同的机器翻译引擎（微软、谷歌和谷歌神经网络）将这个黄金标准语料库翻译成目标语言（德语），并无缝集成到 Geo Fluent 中以促进预处理和后处理程序。本研究进行了两个关键实验。第一个目标是使用 BLEU 算法作为基准来评估整体翻译质量。第二个实验确定了哪些机器翻译引擎最有效地保留了情感。这项调查的结果表明，通过机器翻译成功传递情感可以通过利用 Google 和 Google 神经网络与 Geo Fluent 结合来完成。这一成就标志着在金融领域建立多语言情感平台的关键里程碑。未来的努力将进一步集成特定于语言的处理规则，以提高机器翻译性能，从而推进该项目的总体目标。

中描述的工作介绍了 GLUECoS，这是一个基准测试，旨在评估跨不同任务的代码转换自然语言处理 (NLP) 模型的功效，特别关注情感分析。为了评估情感分析性能，本研究采用英语-西班牙语和英语-印地语数据集，采用一系列跨语言嵌入技术，例如 MUSE、BiCVM 和 BiSkip，以及多语言 BERT (mBERT)。此外，作者提出了 mBERT 模型的改进版本，该模型对综合生成的语码转换数据进行了进一步的微调，以增强其对语码转换设置的适用性。这些发现揭示了情绪分析的显着进步。具体来说，在英语-印地语数据集 (SAIL) 上，最先进 (SOTA) 的 F1 分数为 56.9，而利用修改后的 mBERT 模型产生的最高 F1 分数为 59.35。同样，对于英语-西班牙语数据集（Twitter 情绪），SOTA F1 得分为 64.6，修改后的 mBERT 模型获得了 69.31 的最佳得分。这些结果强调了在合成语码转换数据上微调 mBERT 的有效性，证明了其进一步优化语码转换任务的多语言模型的能力，从而展示了在语码转换环境中增强情感分析的有希望的途径。

机器翻译的最新进展引起了人们对其在情感分析中的应用的极大兴趣。中提到的工作深入研究了机器翻译在跨语言情感分析中的潜在机会和固有局限性。情感分析的关键涉及获取语言特征，通常通过词性标注器和解析器等工具或注释语料库和情感词汇等基本资源来实现。这项研究背后的动机源于为每种语言创建这些工具和资源的艰巨任务，这个过程需要大量的人力。这种限制极大地阻碍了与英语中使用的类似的特定于语言的情感分析技术的开发和实施。情感分析的关键组成部分包括标记语料库和情感词汇。这项研究系统地将这些资源翻译成资源有限的语言。主要目标是提高分类准确性，主要是在处理可用（标记或原始）训练实例时。在训练数据的访问受到限制的情况下，本研究探索了将情感词汇翻译成目标语言的方法，同时努力通过生成额外的上下文信息来提高机器翻译性能。

本研究中进行的实验侧重于英语和土耳其语数据集，包括电影和产品评论。分类任务涉及两类极性检测（正负），不包括中性类。极性检测实验取得了令人鼓舞的成果，特别是通过利用在翻译语料库上训练的通用分类器。然而，需要强调的是，不同语言的语料库之间的差异值得进一步研究，以促进更无缝的资源整合。

此外，定量证据强调了与词汇翻译相关的复杂性，因为语言之间表达情感的固有差异给翻译过程中保留单词和短语的情感带来了挑战。这项研究为跨语言情感分析的不断发展提供了宝贵的见解，揭示了利用机器翻译的潜力和复杂性。

这项工作提出了一种解决方案，通过利用预先训练的多语言转换器模型和数据增强技术来查找用于非英语语言情感分析的大型注释语料库。作者表明，使用机器翻译的数据可以帮助使用带有 Bag-of-N-Grams 的 SVM 模型更好地区分情感分类的相关特征。本研究中使用的数据增强技术涉及机器翻译来增强数据集。具体来说，作者使用了预先训练的多语言

将非英语推文翻译成英语的变压器模型。然后，他们使用这些翻译的推文作为情感分析模型的附加训练数据。这种简单的技术允许利用多语言模型来处理有限大小的非英语推文数据集。

表 1 比较了与情感分析和机器翻译相关的五项最新著作。每项研究都解决了各种语言情感分析的具体方面和挑战，揭示了机器翻译技术的优点和局限性。该表简洁地比较了不同领域的五项最新研究，显示了利用高级语言模型的优点和局限性。这些研究涵盖从加密货币中的情绪分析到网络钓鱼电子邮件检测等主题，强调了与大型语言模型 (LLMs) 在各个领域相关的多样化应用和挑战。我们的研究为跨多种外语的情感分析挑战提供了一种新颖的解决方案。通过引入结合了转换器和大型语言模型的集成模型，我们的研究表明，与单独的预训练模型或单独的 LLMs 相比，情感分析的准确性和可靠性得到了提高。此外，所提出的在进行情感分析之前将外语翻译成英语的方法为跨语言情感分析技术提供了宝贵的见解，对商业、社交媒体分析和政府情报具有实际意义。总体而言，这项研究通过解决外语情感分析的复杂性并为可应用于不同语言环境的跨语言情感分析提供了一个强大的框架，显着推进了情感分析领域。

学习	优点	局限性
瓦希杜尔等人。	本文通过在大型语言模型上实施微调技术，展示了加密货币中情感分析的重大进步。通过在零样本情绪分析性能方面实现 40% 的平均提升，该研究强调了微调策略优化预训练模型有效性的潜力。此外，对基于指令的微调的探索揭示了其功效，特别是在提高大型模型的准确性方面，从而为改进加密货币投资的决策过程提供了宝贵的见解。	尽管做出了显着的贡献，但该论文在研究结果的推广方面面临着局限性，特别是对于较小规模的模型。由于完整的模型容量利用率而观察到的泛化能力下降凸显了在不同模型大小上普遍应用微调技术的潜在挑战。此外，虽然该研究强调了指令清晰度在微调过程中的重要性，但它揭示了更长、更复杂的指令对模型准确性的负面影响，建议进一步研究优化指令设计以减轻此类限制。
Xing	本文通过探索在不进行微调的情况下利用大型语言模型 (LLMs) 的有效性，开创了金融情绪分析 (FSA) 的范式转变。该研究植根于明斯基的心理和情感理论，提出了一种新颖的设计框架，即异构LLM FSA 代理（HAD），其中涉及利用 FSA 错误类型的先验领域知识实例化的专门代理。该框架通过对 FSA 数据集的综合评估（主要是在产生大量讨论时）展示了准确性的提高。这种方法弥补了幼稚提示和微调之间的性能差距。它提供了一种计算效率高的替代方案，为在 FSA 中利用 LLMs 铺平了新的途径，而无需进行大量的模型训练。	首先，可扩展性带来了挑战，因为与LLM代理进行预测或讨论会比传统的统计分析方法产生更高的计算成本。虽然该框架显示出希望，但需要进一步探索来解决可扩展性问题，尤其是在使用更多代理扩展系统时。其次，评估数据集的机密性提出了潜在的问题，特别是在之前的迭代中数据集暴露于LLMs，这可能导致信息泄露和评估偏差。此外，虽然该研究有效地识别和解决了 FSA 中的错误类型，但仍需要进一步调查以了解 LLMs 所犯新错误或剩余错误的原因，并评估 FSA 数据集上的人类水平表现。尽管存在这些局限性，该研究强调了 LLMs 在解决 FSA 等具有挑战性的任务方面的多功能能力，并呼吁在这一重要领域继续进行研究，特别是考虑到先进的通用人工智能 (AGI) 功能。
徐等人。	本研究引入了一种突破性的医学文本语义分析方法，通过利用大型语言模型 (LLMs) 的功能（以 GPT-4 为例）来增强放射学等高度专业化领域的相似性度量。通过一个新颖的框架，LLMs 用于零样本文本识别和标签生成，然后将其用作放射学报告中文本相似性的测量。通过在 MIMIC 数据集上测试该框架，该研究表明与 ROUGE 和 BLEU 等传统 NLP 指标相比，语义相似性评估有了显着改进。所提出的方法超越了词汇比较指标，并展示了人工智能驱动的方法通过促进对临床文档的更深入理解来彻底改变医疗信息学的潜力，最终推进精准医学和循证临床实践。	一个主要限制在于该方法中假设的“人机交互”（HITL）组成部分，其中医疗专业人员直接参与完善人工智能生成的标签并未在当前研究范围内实施。这种缺失可能会影响人工智能生成标签的准确性、相关性和临床实用性，凸显出预期能力和实现结果之间的差距。此外，研究范围仅限于胸部 X 射线放射学报告，限制了研究结果在其他医疗文件类型和专业中的普遍性。未来的研究应侧重于 HITL 框架的实施，并将分析扩展到更广泛的医学文本，确保所提出的方法在各种医疗信息学应用中的相关性和适用性。通过解决这些局限性，后续研究可以以本研究中提出的基础工作为基础，进一步加强人工智能在医学文本分析中的整合，并推进精准医学实践。
乌丁等人。	本文提出了一种使用微调的 DistilBERT 模型进行网络钓鱼电子邮件检测的创新方法，在区分网络钓鱼电子邮件和合法电子邮件方面实现了较高的准确率。通过解决类别不平衡问题并采用 LIME 和 Transformer Interpret 等可解释 AI 技术，模型的决策过程变得透明，从而增强用户的信任和理解。这项研究强调了高级语言模型在显着改进网络安全措施以应对网络钓鱼威胁方面的潜力。	尽管取得了进步，但这项研究缺乏人类专家直接参与完善人工智能生成的标签，这可能会影响模型的临床有效性。 此外，它仅关注网络钓鱼电子邮件检测限制了普遍性。未来的研究应整合“人机交互”方法，并探索跨领域和数据集的更广泛应用，以确保稳健性和有效性。
雷汉等人。	本文介绍了英语和乌尔都语多语言威胁内容检测（MTCD）的开创性框架，解决了有关低资源语言的文献中的重大空白。该研究利用迁移学习和微调技术，使用最先进的 RoBERTa 和 MuRIL 变压器模型探索联合多语言和联合翻译方法。所提出的框架实现了基准性能，以 92% 的准确率和 90% 的宏观 F1 分数超越基准，特别是在联合多语言方法方面表现出色。通过提供对社交媒体内容分类的有效多语言 NLP 框架的见解，该研究大大增强了对威胁性表达的自动检测，促进在线和平与和谐。	首先，评估仅限于两种语言（英语和乌尔都语），忽略了其他资源匮乏的语言，这些语言可能也需要注意威胁内容检测。将框架扩展到包括俄语、中文或印地语等语言可以增强其实用性和普遍性。其次，虽然在半监督语料库上取得了令人印象深刻的结果，但在更大的数据集上进行测试可以提供进一步的见解，并确保框架在不同语言环境中的稳健性。此外，这里采用的二元分类方法可能会过度简化任务；未来的研究可以探索更细致的分类，例如识别威胁内容中的目标个人或社区，以增强该框架在现实世界应用中的相关性和有效性。

表 1. 不同领域情感分析应用研究的比较。

问题表述

本研究中解决的问题可以形式化如下。将情感分析表示为 SA，这是自然语言处理 (NLP) 中的一项任务。 SA 涉及将文本分类为不同的情感极性，即积极 (P)、消极 (N) 或中性 (U)。随着社交媒体和互联网的日益普及，SA 在营销、政治和客户服务等各个领域都变得越来越重要。然而，在处理外语时，情感分析变得具有挑战性，特别是在没有用于训练模型的标记数据的情况下。

考虑到假设 H，通过将文本翻译成英语并分析翻译文本中的情感，外语情感分析是可行的。我们使用四种不同的语言进行了实验来验证这一假设：阿拉伯语 (A)、中文 (C)、法语 (F) 和意大利语 (I)。翻译过程使用 LibreTranslate API (T_libre) 和 Google Translate API (T_google)。然后使用两个预先训练的情感分析模型对每个句子 s 进行情感分析：Twitter-Roberta-Base-Sentiment-Latest (M_Twitter) 和 Bertweet-Base-Sentiment-Analysis (M_Bertweet) 以及由 Twitter-Roberta- BaseSentiment-Latest 、 bert-base-multilingual-uncased-sentiment 和 GPT-3 。

为了衡量翻译句子情感分析的准确性，我们将 Acc 定义为准确性指标。Acc 是正确分类的句子与分析的句子总数的比率。从数学上讲，Acc 由等式给出。 (1):

$$Acc = (C/C) * 100 \quad (1)$$

C_正确 表示正确分类的句子的数量，C_全部的表示句子总数分析了。

本研究的主要目的是评估翻译句子情感分析的可行性，从而深入了解利用翻译文本进行情感分析的潜力，并开发一种新模型以提高准确性。通过使用 Acc 评估情感分析的准确性，我们的目的是验证假设 H，即通过翻译成英语可以进行外语情感分析。

这项研究的结果对跨语言沟通和理解具有重要意义。如果假设 H 得到支持，则意味着外语情感分析的可行性，从而有助于提高对不同语言表达的情感的理解。这项研究的结果对各个领域都很有价值，例如多语言营销活动、跨文化分析和国际客户服务，在这些领域理解外语情绪至关重要。

方法论

在这项研究中，我们采用多步骤方法通过将外语文本翻译成基础语言英语来分析其情感。该方法包括五个阶段：数据收集、数据清理和预处理、翻译成英文、情感分析和结果评估。首先，我们从社交媒体、新闻文章和在线论坛等各种来源收集目标语言的数据。接下来，我们进行数据清理和预处理，以去除数据中的噪声、重复内容和不相关信息。之后，我们使用机器翻译系统将清理和预处理的数据翻译成英语。然后，我们使用专为英语文本设计的情感分析模型来分析翻译数据。最后，我们评估了情感分析的结果，以确定该方法的准确性和有效性。图 1 显示了本研究中采用的方法的概述。在以下部分中，我们详细描述了该方法的每个阶段以及每个阶段中使用的工具和技术。

数据收集

在研究方法的初始阶段，目标语言的数据是从不同且成熟的来源收集的，包括 SemEval-2017 任务 4：Twitter 中的情绪分析、amazon_reviews_multi、DEFT 2017 和 SENTIPOLC 2016。本研究采用四种不同的语言，即阿拉伯语 (ar)、中文 (zh)、法语 (fr) 和意大利语 (it)，这是根据 Twitter 平台上的推文中的频繁使用而选择的。此外，这些语言的选择得到了研究的支持，研究表明它们是 Twitter 上最常用的消息语言之一。此外，这些语言是世界上使用最广泛的语言之一。每个数据集都标有三个情感标签，即“积极”、“消极”和“中性”。这些注释任务是由众包平台中的众包工作人员手动完成的。表 2 显示了本研究中收集和使用的数据概述。该表列出了本研究中使用的数据源、语言和每个数据源的句子数量。图 2 显示了所使用的数据集中不同语言的分布。

数据清洗和预处理

在该方法的第二阶段，收集的数据经过数据清理和预处理的过程，以消除噪声、重复内容和不相关信息。此过程涉及多个步骤，包括标记化、停用词删除以及表情符号和 URL 的删除。标记化是通过将文本划分为单独的单词或短语来执行的。相比之下，停用词删除需要删除常用词，例如“and”、“the”和“in”，这些词对情感分析没有帮助。虽然词干提取和词形还原在某些自然语言处理任务中很有帮助，但在基于 Transformer 的情感分析中通常是不必要的，因为模型旨在处理词形和词形变化的变化。因此，本研究的数据清理和预处理中没有应用词干提取和词形还原

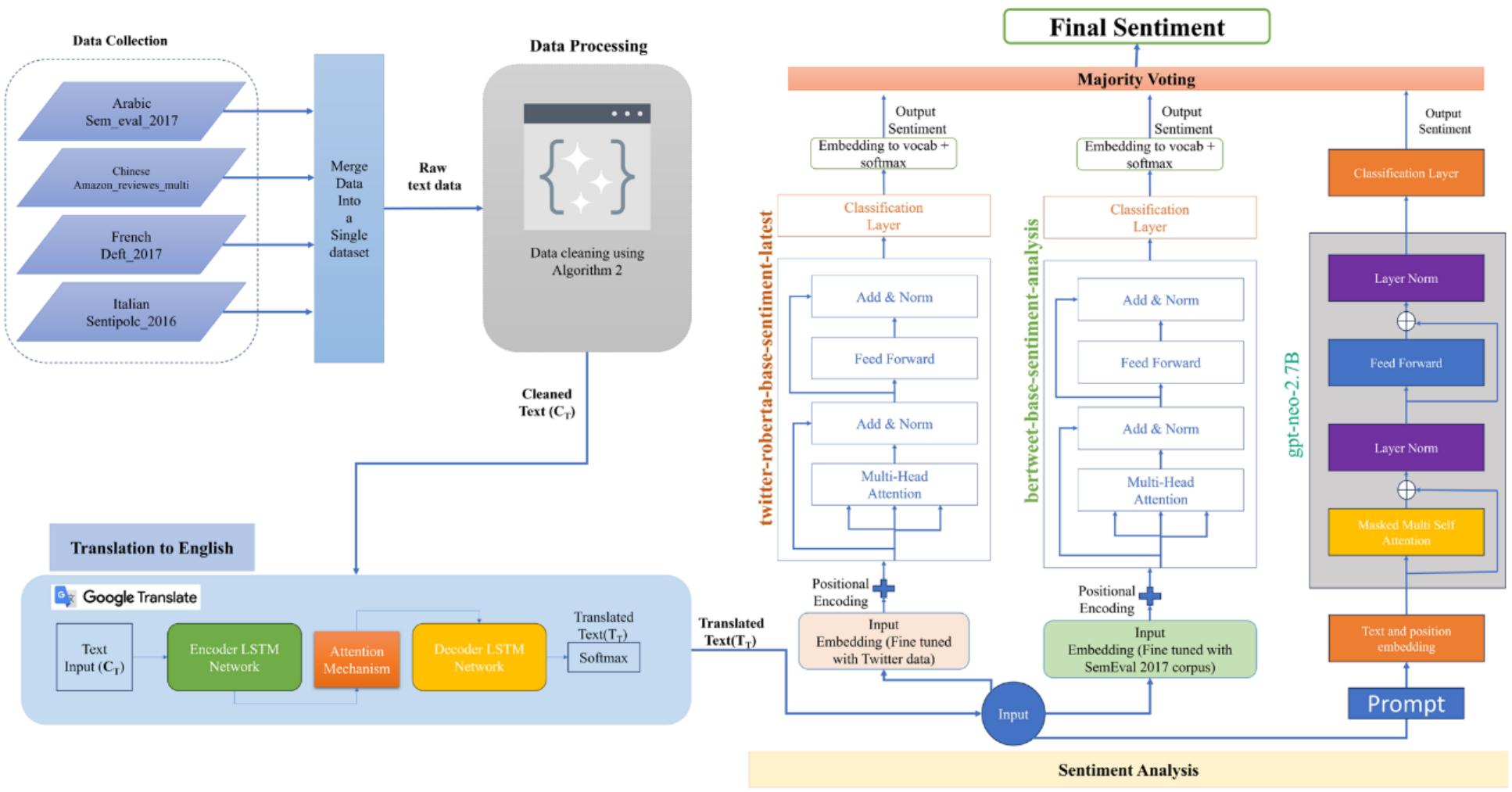

图 1.所提出方法的概述。

数据来源	语言	数据点数量
SemEval-2017 任务 4	阿拉伯语 (ar)	1823
亚马逊评论多	中文（zh）	3000
2017年DEFT	法语（fr）	1715
森蒂波尔 2016	意大利语（它）	1837

表 2. 本研究中使用的数据概述。

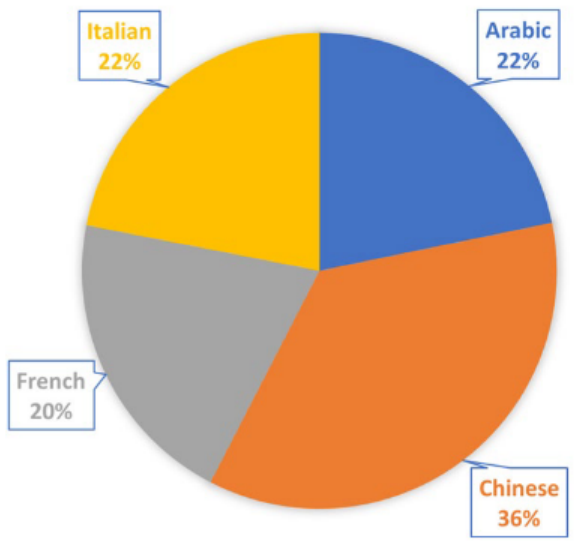

图 2. 本研究使用的数据集中不同语言的分布。

阶段，它利用基于 Transformer 的预训练模型进行情感分析。表情符号删除被认为在情感分析中至关重要，因为它可以传达可能干扰情感分类过程的情感信息。 URL 删除也被认为至关重要，因为 URL 不提供相关信息并且可能占用大量特征空间。完整的数据清理和预处理步骤如算法 1 所示。

```
function TextCleaner(text)
    text ← remove stopwords from text
    text ← remove all URLs from text
    set: emoji_pattern ← compile the regular expression to match and remove emojis from text
    text ← remove user mentions from text
    text ← remove hashtags from text
    text ← remove retweet sign from text
    return text
end function
```

算法1.文本清理算法

翻译为基本语言：英语

在该方法的第三阶段，我们使用自托管机器翻译系统（即 LibreTranslate）和 Google 翻译神经机器翻译（NMT）的云托管服务将清理和预处理的数据翻译成英语。 LibreTranslate 是一个免费的开源机器翻译 API，它使用预先训练的 NMT 模型在不同语言之间翻译文本。输入文本被标记化，然后使用编码器神经网络编码为数字表示。然后，编码表示通过解码器网络，生成目标语言的翻译文本。谷歌翻译 NMT 使用深度学习神经网络将文本从一种语言翻译成另一种语言。神经网络接受大量双语数据的训练，以学习如何有效翻译。在翻译过程中，输入文本首先被标记为单独的单词或短语，并且每个标记都被分配一个唯一的标识符。然后，这些标记被输入神经网络，神经网络在一系列层中处理它们，以生成可能翻译的概率分布。网络的输出是目标语言的一系列标记，然后将其转换回最终翻译文本的单词或短语。神经网络经过训练，可以优化翻译准确性，同时考虑输入文本的含义和上下文。谷歌翻译 NMT 的优势之一是它能够处理复杂的句子结构和语言中的细微差别。

对于预测任务，翻译过程是迭代的。一旦句子翻译完成，就会分析句子的情感并提供输出。然而，首先翻译句子来训练模型，然后执行情感分析任务。情感分析过程将在下一节中讨论。算法 2 介绍了本研究中采用的方法。

```
Input: Cleaned and pre-processed data
Output: Translated text in English
procedure Translate(data)
    Initialize translationResult as an empty string
    for all sentence in data do
        translationResult += LibreTranslate(sentence)          ▷ Translate using LibreTranslate
        translationResult += GoogleTranslate(sentence)         ▷ Translate using Google Translate NMT
    end for
    return translationResult
end procedure
procedure LibreTranslate(sentence)
    Tokenize and encode sentence using an encoder neural network
    Pass encoded representation through a decoder network
    Generate translated text in English
end procedure
procedure GoogleTranslate(sentence)
    Tokenize sentence into individual words or phrases
    Assign a unique identifier to each token
    Feed tokens into neural network for translation
    Process tokens through layers to generate translation probability distribution
    Convert sequence of tokens in target language back into words or phrases
end procedure
Main Procedure:
Call Translate function with cleaned and pre-processed data as input
```

算法2.翻译过程

情感分析

在该方法的第四阶段，我们使用预先训练的情感分析深度学习模型和提出的集成模型对翻译数据进行情感分析。在本研究中，我们利用了 Hugging Face 中两个预先训练的情感分析模型的集合，即 Twitter-Roberta-Base-Sentiment-Latest、bert-base-multilingual-uncased-sentiment 和 GPT-3 LLM 来自 OpenAI。整体情感分析模型分析文本以确定情感极性（积极、消极或中性）。情感分析过程如算法 3 所示。该算法显示了情感分析阶段遵循的逐步过程。

Hugging Face 是一家提供开源软件库以及用于构建和共享自然语言处理 (NLP) 模型的平台的公司。该平台提供对各种预训练模型的访问，包括可用于情感分析的 Twitter-Roberta-Base-Sentiment-Latest 和 Bertweet-Base-Sentiment-Analysis 模型。

使用这些模型的主要优点之一是它们在情感分析任务中的高精度和高性能，特别是对于 Twitter 等社交媒体数据。这些模型经过大量文本数据（包括社交媒体内容）的预先训练，这使它们能够捕捉社交媒体中使用的语言的细微差别和复杂性。使用这些模型的另一个优点是它们能够处理不同的语言和方言。这些模型经过多语言数据的训练，这使得它们适合分析用各种语言编写的文本中的情感。

Require: *dataset_path*, *input_sentence*
1: Load CSV dataset from *dataset_path*
2: Load BERT and RoBERTa sentiment analysis models
3: Load GPT-3 model and tokenizer
4: Define input sentence as *input_sentence*
5: Perform sentiment analysis using BERT and store results as *bert_results*
6: Perform sentiment analysis using RoBERTa and store results as *roberta_results*
7: Generate sentiment using GPT-3
8: Combine predictions using majority voting
9: Print "Ensemble Sentiment:", *final_sentiment*

算法3.情感分析集成模型

所提出的算法概述了一个通过利用三种不同的自然语言处理模型（BERT、RoBERTa 和 GPT-3）的功能来进行情感分析的集成模型。该算法首先导入 CSV 数据集，然后初始化 BERT 和 RoBERTa 的情感分析模型。此外，还加载了 GPT-3 模型和分词器，以方便生成情感相关文本。该算法的核心围绕为情感分析提供的输入句子。使用 BERT 和 RoBERTa 模型分析输入句子，并存储其结果以供进一步处理。接下来，利用 GPT-3 模型根据固定提示“提供给定文本在单个类别中的积极、消极和中立的情绪”来生成给定文本的情绪。

下一步涉及通过称为多数投票的过程将 BERT、RoBERTa 和 GPT-3 模型提供的预测结合起来。这需要统计“积极”、“消极”和“中性”情绪标签的出现次数。根据结果，算法会给出最终的情绪。

评估指标

在该方法的最后阶段，我们评估了情感分析的结果，以确定该方法的准确性和有效性。我们将情感分析结果与真实情感（数据集中标记的文本的原始情感）进行比较，以评估情感分析的准确性。

在评估阶段，我们使用真阳性（TP）、真阴性（TN）、假阳性（FP）和假阴性（FN）作为评估二元分类器性能的指标。这些指标通常用于评估二元分类模型。 TP表示正确识别的正实例数量，TN表示正确识别的负实例数量，FP表示错误分类的正实例数量，FN表示错误分类的负实例数量。算法 4 给出了 TP、TN、FP 和 FN 的定义。该算法需要两个输入：数据集中标记的原始语言情感和将其翻译成英语后的情感。然后它会比较两种情绪并确定它们是积极的、消极的还是中性的。如果原始情绪是积极的并且翻译后的情绪也是积极的，则被认为是真正的积极（TP）。如果原始情绪是积极的，而翻译后的情绪是消极的，则被视为假阴性（FN）。如果原始情绪是积极的，而翻译后的情绪是中性的，则被视为误报 (FP)。如果原始情绪是负面的，而翻译后的情绪也是负面的，则被视为真负面（TN）。如果原始情绪是负面的，而翻译后的情绪是正面的，则被视为误报 (FP)。如果原始情绪是负面的，而翻译后的情绪是中性的，则被视为假阴性（FN）。如果原始情绪和翻译情绪是中性的，则被视为真阳性 (TP)。如果原始情绪是中性的，而翻译后的情绪是积极的，则被视为误报（FP）。 如果原始情绪是中性的，而翻译后的情绪是负面的，则被视为假阴性 (FN)。

Require: Original language sentiment labelled in the dataset, Sentiment after translating to English
Ensure: Output (TP, TN, FP, FN)

```
if original_sentiment = "positive" then
    if translated_text_sentiment = "positive" then
        Output ← "TP";
    else if translated_text_sentiment = "negative" then
        Output ← "FN";
    else if translated_text_sentiment = "neutral" then
        Output ← "FP";
    end if
else if original_sentiment = "negative" then
    if translated_text_sentiment = "negative" then
        Output ← "TN";
    else if translated_text_sentiment = "positive" then
        Output ← "FP";
    else if translated_text_sentiment = "neutral" then
        Output ← "FN";
    end if
else if original_sentiment = "neutral" then
    if translated_text_sentiment = "neutral" then
        Output ← "TP";
    else if translated_text_sentiment = "positive" then
        Output ← "FP";
    else if translated_text_sentiment = "negative" then
        Output ← "FN";
    end if
end if
```

算法 4. 定义 TP、TN、FP 和 FN

我们的研究分为三个不同的类别：积极、消极和中立。因此，我们分别计算了每个类别的指标。精度用于衡量分类器识别为阳性的所有实例中正确分类的实例的比例，而召回率用于衡量数据集中所有阳性实例中正确识别的实例的比例。 F1 分数是精确率和召回率的调和平均值。准确率衡量数据集中所有实例中正确分类的实例的比例。相反，特异性衡量数据集中所有负实例中正确分类的负实例的比例。

评价指标可以用以下等式表示： (2)、(3)、(4)、(5)和(6)。

$$\text{精确} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{记起} = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \cdot \frac{\text{精确} \cdot \text{记起}}{\text{精确} + \text{记起}} \tag{4}$$

$$\text{准确性} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{特异性} = \frac{TN}{TN + FP} \tag{6}$$

结果与讨论

在本节中，我们将介绍并讨论使用机器学习模型进行外语情感分析的实验结果。我们测试了两种不同的翻译服务，即 LibreTranslate 和 Google Translate，将阿拉伯语、中文、法语和意大利语句子翻译成英语。然后使用三种不同的预训练情感分析模型对翻译的句子进行情感分析：Twitter-RoBERTa-BaseSentiment-Latest、BERTweet-Base-Sentiment-Analysis、GPT-3（这是一个 LLM），以及提出的集成模型。我们对每个语言对进行了 8 次实验，总共 32 次实验。我们展示结果

这些实验并讨论翻译服务和情感分析模型的性能。最后，我们总结了利用外语翻译句子进行情感分析的可行性和有效性。

表 3 显示了涉及翻译器和情感分析器模型的情感分析任务的不同组合的结果。所提出的指标包括准确度、精确度、召回率、F1 分数和特异性，共同提供对情感分析任务中各种组合的性能的全面评估。评估涵盖两种主要翻译服务，即 LibreTranslate 和 Google Translate，以及四种不同的情感分析器模型：Twitter-Roberta-Base、Bertweet-Base、GPT-3 和一个新颖的提议集成模型。

关于准确性，值得注意的是，LibreTranslate-Bertweet-Base 组合在所有测试的组合中表现出最低的准确性分数，为 0.5638。相反，Google Translate 与提议的 Ensemble 模型相结合，获得了最高的准确度分数 0.8671，展示了其实现卓越情感分析结果的潜力。

GPT-3 模型的性能值得注意，因为它在与 LibreTranslate 或 Google Translate 服务配合使用时始终表现出强大的情感分析功能。这一发现强调了 GPT-3 模型在不同翻译平台上执行情感分析任务的多功能性和稳健性。

此外，拟议的 Ensemble 模型在多个指标上始终如一地提供了有竞争力的结果，强调了其作为跨各种翻译上下文的情感分析器的有效性。这一观察结果表明，集成方法对于实现准确的情绪预测非常有价值。生成了一系列图表来直观地表示翻译器和情感分析模型的各种组合的实验结果，提供了对这些模型在情感分析中的有效性的全面了解，如图 3 所示。

在准确度图（左上）中，每个图代表翻译器和情感分析器模型的不同组合所达到的准确度。 x 轴标识所使用的转换器，而 y 轴表示

翻译者	情感分析器模型	准确性	精确	记起	F1	特异性
自由翻译	推特-罗伯塔-基地	0.5531	0.6440	0.6069	0.6249	0.4678
	Bertweet-base	0.5638	0.6555	0.6155	0.6349	0.4807
	GPT-3	0.8091	0.6909	0.7081	0.6993	0.4522
	提出的整体模型	0.8491	0.7892	0.7771	0.7831	0.5201
谷歌翻译	Twitter-罗伯塔-基地	0.6182	0.6982	0.6521	0.6744	0.5660
	Bertweet-base	0.6347	0.7066	0.6734	0.6896	0.5761
	GPT-3	0.8571	0.7053	0.7366	0.7199	0.4706
	提出的整体模型	0.8671	0.8091	0.8122	0.8106	0.5713

表 3. 翻译器和情感分析器模型不同组合的实验结果。

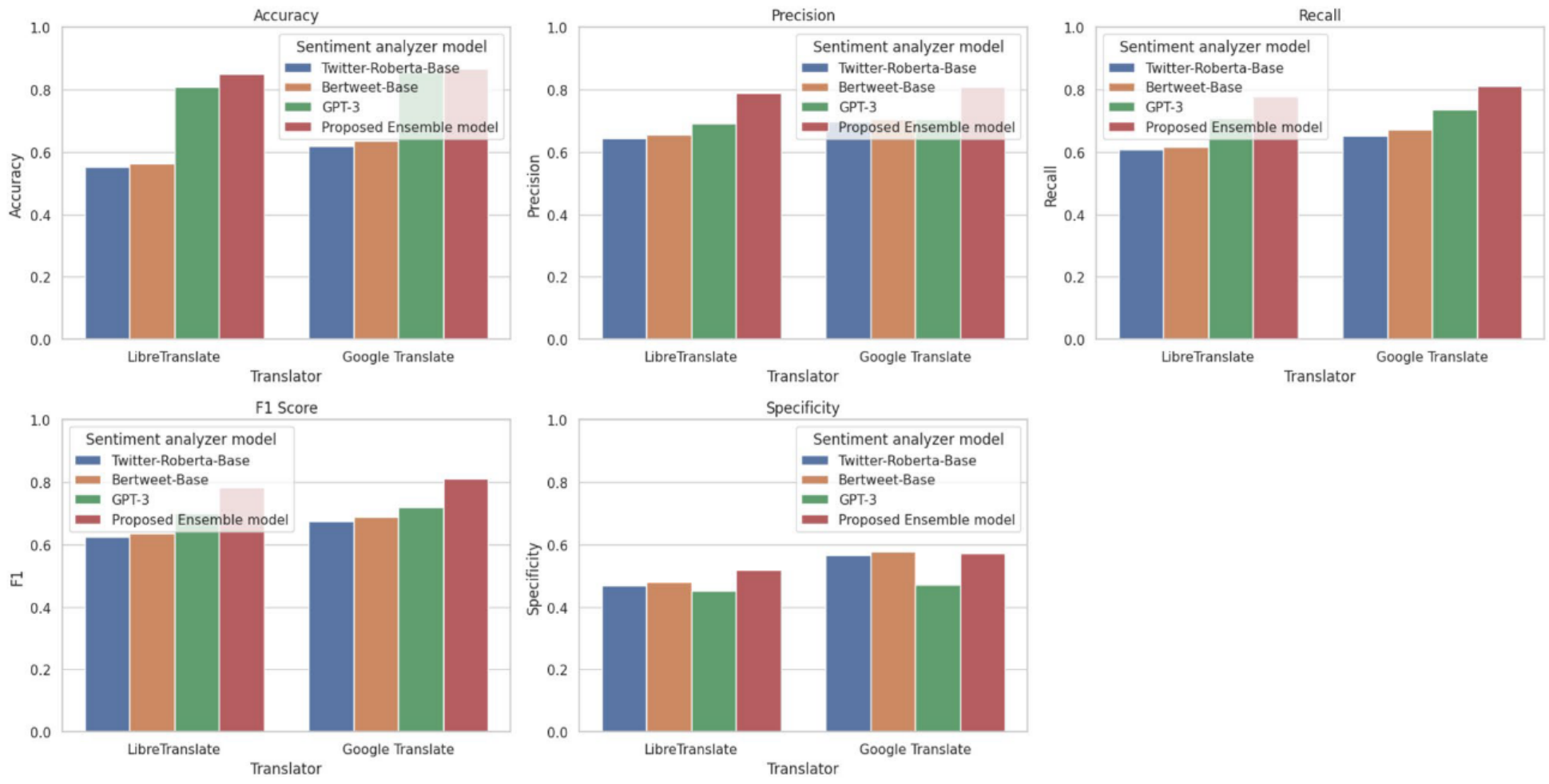

图 3. 实验结果显示了不同评估指标的结果。

卷：.(1234567890)

准确度得分。谷歌翻译与提议的集成模型相结合成为最准确的，产生的准确度分数约为 0.8671。

转向精度图（中上），该可视化重点关注精度，它衡量模型正确识别积极情绪实例的能力。在此图中，谷歌翻译与提出的集成模型相结合，以最高的精度得分脱颖而出，达到 0.8091 左右。

召回图（右上）以召回指标为中心，表明模型在识别所有相关积极情绪实例方面的熟练程度。在这里，谷歌翻译与提议的集成模型相结合，展示了最高的召回率，大约为 0.8122。

F1 分数图（左下）提供了 F1 分数的概述，这是一个平衡精确度和召回率的指标，用于衡量情感分析模型的整体有效性。在此图中，谷歌翻译和提议的集成模型展示了最高的 F1 分数，约为 0.8106。

特异性图（中下）侧重于特异性，这是情绪分析中的一个关键指标，用于评估模型准确识别负面情绪的能力。在这里，谷歌翻译与提议的集成模型配对表现出最高的特异性，大约为 0.5713。

图 4 中所示的混淆矩阵提供了不同翻译器和情感分析器模型组合在情感分类中的性能的详细总结。每个混淆矩阵由四个象限组成，分别代表以下内容：

- 真负面 (TN) 正确分类为负面情绪的实例数量。
- 误报 (FP) 被错误分类为积极情绪的实例数量。
- 假阴性 (FN) 被错误分类为负面情绪的实例数量。
- 真阳性 (TP) 正确分类为积极情绪的实例数量。

例如，考虑 LibreTranslate - Twitter-Roberta-Base 组合的混淆矩阵，它显示在 891 个实例中：

- 312 被正确分类为负面情绪 (TN)。
- 179 种情绪实际上是消极情绪 (FP)，但被错误地归类为积极情绪。
- 222 种情绪实际上是积极的，但被错误地归类为消极情绪 (FN)。
- 346 被正确分类为积极情绪 (TP)。

同样，每个混淆矩阵都可以深入了解不同翻译器和情感分析器模型组合在准确分类情感方面的优缺点。评估这些矩阵中的数字有助于了解模型在情感分析任务中的整体性能和有效性。

这些图形表示可以作为理解翻译器和情感分析器模型的不同组合如何影响情感分析性能的宝贵资源。研究人员和从业者可以利用这些可视化来确定适合其特定应用的最有效组合，无论是社交媒体、客户评论或任何其他环境中的情感分析，确保情感分析模型的最佳性能。在介绍总体实验结果之后，下面详细描述和讨论特定语言的实验结果。

表 4 显示了四种语言的翻译器和情感分析器模型的不同组合的准确度得分：阿拉伯语、中文、法语和意大利语。在 LibreTranslate 和 Google Translate 中，所提出的集成模型始终获得最高的准确度分数，四种语言的值范围为 0.83 到 0.88。对于 LibreTranslate，GPT-3 也表现出相对较高的准确性，特别是对于阿拉伯语，得分为 0.81。与此同时，谷歌翻译与 GPT-3 或所提出的集成模型的配对始终优于其他组合，跨语言的准确度得分在 0.84 到 0.88 之间。值得注意的是，在各种翻译器和情感分析器组合中，中文的准确性始终高于其他语言。这些发现表明，集成模型与 GPT-3 有望提高多语言情感分析任务的准确性，而中文相对更容易准确地分析情感。

表 5 提供了四种语言的翻译器和情感分析器模型的不同组合的精度分数：阿拉伯语、中文、法语和意大利语。在 LibreTranslate 框架内，所提出的集成模型在所有语言中始终达到最高的精度分数，范围从 0.75 到 0.82。值得注意的是，阿拉伯语和中文的精确度分数相对高于法语和意大利语。同样，所提出的 Google Translate 集成模型展示了卓越的精度分数，四种语言的值范围为 0.7 到 0.87。同样，阿拉伯语和中文的精确度得分高于法语和意大利语。此外，GPT-3 与 LibreTranslate 和 Google Translate 配合使用，在所有语言中始终显示出有竞争力的精确度分数。这些发现表明，所提出的集成模型与 GPT-3 有望提高跨不同语言环境的多语言情感分析任务的精度。

表 6 描述了翻译器和情感分析器模型的不同组合的召回分数。在 LibreTranslate 和 Google Translate 框架中，所提出的集成模型始终表现出所有语言的最高召回分数，范围从 0.75 到 0.82。值得注意的是，与意大利语相比，阿拉伯语、中文和法语的召回分数相对较高。同样，GPT-3 与 LibreTranslate 和 Google Translate 配合使用，在所有语言中始终显示出有竞争力的召回分数。对于阿拉伯语，各种组合的召回分数都非常高，表明对该语言进行了有效的情感分析。这些发现表明，所提出的集成模型与 GPT-3 有望提高跨不同语言环境的多语言情感分析任务的召回率。

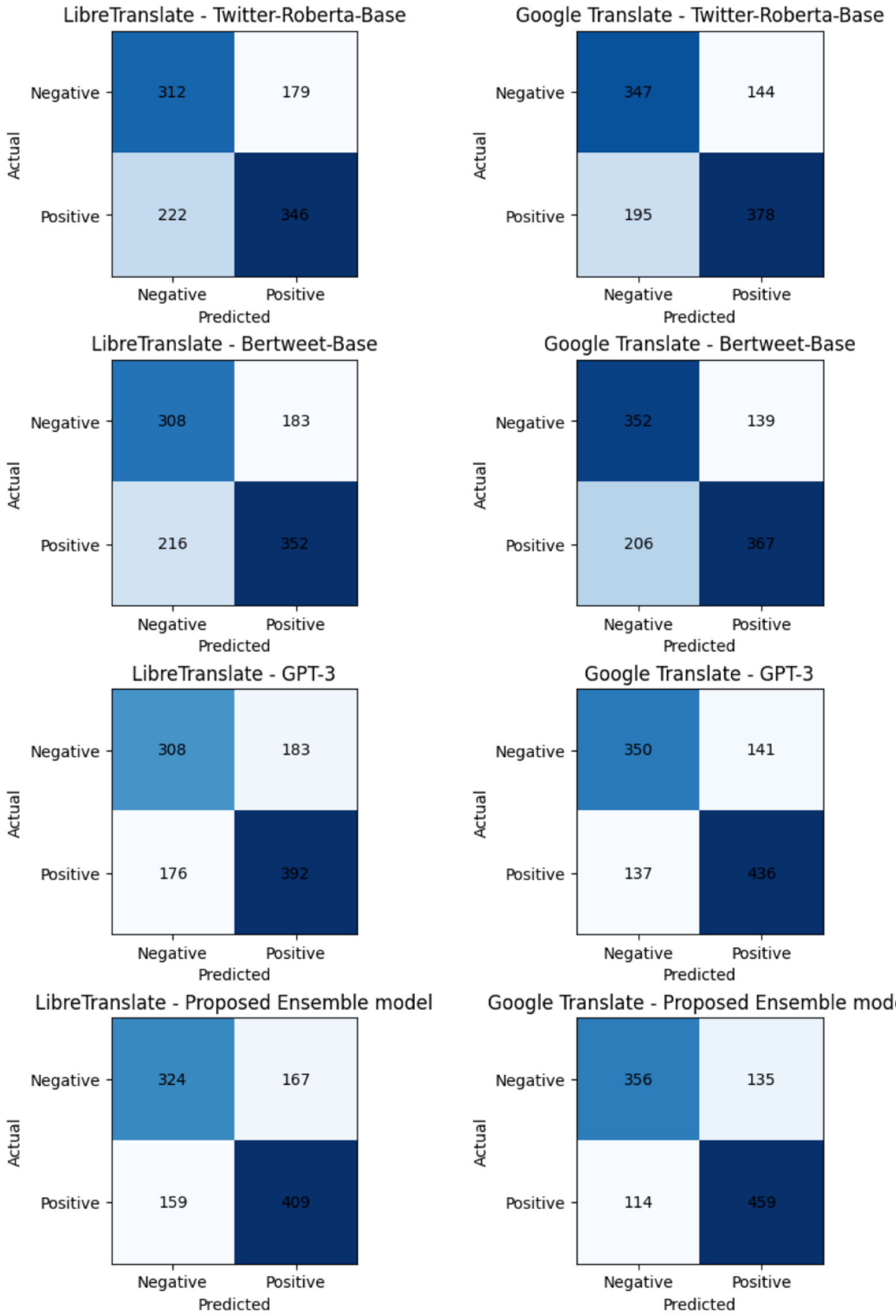

图 4.来自不同实验的混淆矩阵。

表 7 显示了四种语言的翻译器和情感分析器模型的不同组合的 F1 分数：阿拉伯语、中文、法语和意大利语。在 LibreTranslate 和 Google Translate 框架中，所提出的集成模型在所有语言中始终获得最高的 F1 分数，范围从 0.746 到 0.844。值得注意的是，汉语和意大利语的 F1 分数相对高于阿拉伯语和法语。此外，GPT-3 与 LibreTranslate 和 Google Translate 配合使用，在所有语言中始终表现出有竞争力的 F1 分数。特别是对于中文来说，各种组合的 F1 分数都非常高，这表明对该语言的情感分析是有效的。这些发现表明，拟议的

翻译者	情绪分析器模型	阿拉伯	中国人	法语	意大利语
语言准确性					
自由翻译	Twitter-罗伯塔-基地	0.55	0.5	0.56	0.63
自由翻译	Bertweet-base	0.57	0.5	0.6	0.63
自由翻译	GPT-3	0.81	0.78	0.81	0.82
自由翻译	提出的整体模型	0.83	0.85	0.84	0.83
谷歌翻译	Twitter-罗伯塔-基地	0.61	0.61	0.58	0.67
谷歌翻译	Bertweet-base	0.62	0.62	0.62	0.68
谷歌翻译	GPT-3	0.84	0.87	0.84	0.83
谷歌翻译	提出的整体模型	0.86	0.88	0.85	0.84

表 4. 翻译器和情感分析器模型的不同组合之间的语言准确性得分。

翻译者	情绪分析器模型	阿拉伯	中国人	法语	意大利语
语言精度					
自由翻译	Twitter-罗伯塔-基地	0.61	0.61	0.61	0.74
自由翻译	Bertweet-base	0.63	0.62	0.64	0.74
自由翻译	GPT-3	0.64	0.7	0.65	0.77
自由翻译	提出的整体模型	0.79	0.82	0.75	0.78
谷歌翻译	Twitter-罗伯塔-基地	0.67	0.74	0.61	0.75
谷歌翻译	Bertweet-base	0.67	0.74	0.64	0.76
谷歌翻译	GPT-3	0.68	0.78	0.66	0.7
谷歌翻译	提出的整体模型	0.81	0.85	0.7	0.87

表 5. 翻译器和情感分析器模型的不同组合之间的语言精度得分。

翻译者	情绪分析器模型	阿拉伯	中国人	法语	意大利语
语言记忆					
自由翻译	Twitter-罗伯塔-基地	0.62	0.49	0.67	0.7
自由翻译	Bertweet-base	0.65	0.5	0.69	0.68
自由翻译	GPT-3	0.69	0.68	0.75	0.7
自由翻译	提出的整体模型	0.72	0.75	0.78	0.81
谷歌翻译	Twitter-罗伯塔-基地	0.69	0.55	0.68	0.72
谷歌翻译	Bertweet-base	0.71	0.58	0.72	0.73
谷歌翻译	GPT-3	0.73	0.7	0.74	0.78
谷歌翻译	提出的整体模型	0.79	0.81	0.8	0.82

表 6. 翻译器和情感分析器模型的不同组合之间的语言回忆得分。

翻译者	情绪分析器模型	阿拉伯	中国人	法语	意大利语
语言方面的 F1					
自由翻译	Twitter-罗伯塔-基地	0.62	0.54	0.64	0.71
自由翻译	Bertweet-base	0.63	0.553	0.66	0.71
自由翻译	GPT-3	0.66	0.68	0.69	0.73
自由翻译	提出的整体模型	0.75	0.78	0.76	0.79
谷歌翻译	Twitter-罗伯塔-基地	0.68	0.63	0.64	0.74
谷歌翻译	Bertweet-base	0.68	0.65	0.67	0.74
谷歌翻译	GPT-3	0.70	0.73	0.69	0.73
谷歌翻译	提出的整体模型	0.79	0.82	0.74	0.84

表 7. 翻译器和情感分析器模型的不同组合之间的语言方面 F1 分数。

集成模型与 GPT-3 一起有望提高跨不同语言环境的多语言情感分析任务的 F1 分数。

表 8 提供了四种语言的翻译器和情感分析器模型的不同组合的特异性分数：阿拉伯语、中文、法语和意大利语。特异性衡量正确识别的负面实例在所有实际负面实例中的比例。在 LibreTranslate 和 Google Translate 框架中，所提出的集成模型在所有语言中始终获得最高的特异性得分，范围从 0.49 到 0.73。值得注意的是，中文的特异性得分相对高于其他语言。此外，谷歌翻译与所提出的集成模型相结合，显示出中文和意大利语的高特异性得分。然而，值得注意的是，特异性得分通常低于其他评估指标，这表明某些模型可能难以准确识别负面情绪实例。这些发现表明，虽然所提出的集成模型有望提高情感分析的特异性，但可能需要进一步细化以提高所有语言和翻译器的性能。

之后，还使用扩展语言模型 (XLM) XLM-T 对该数据集进行训练和测试。这是一个基于 XLM-R 架构但进行了一些修改的多语言语言模型。与 XLM-R 类似，它可以针对情感分析进行微调，特别是对于包含推文的数据集，因为它专注于非正式语言和社交媒体数据。然而，对于实验来说，该模型是在基线配置中使用的，没有进行任何微调。同样，该数据集也使用名为 mBERT 的多语言 BERT 模型进行了训练和测试。实验结果与所提出的集成模型的比较如表9所示。

实验结果表明，与 XLM-T 和 mBERT 等已建立的情感分析模型相比，所提出的集成模型取得了可喜的性能提升。两个提出的模型分别利用 LibreTranslate 和 Google Translate，表现出更好的准确性和精确度，分别超过 84% 和 80%。与 XLM-T 80.25% 的准确率和 mBERT 78.25% 的准确率相比，这些集成方法明显提高了情感识别能力。谷歌翻译集成模型获得了最高的总体准确率（86.71%）和精确度（80.91%），凸显了其强大的情感分析任务的潜力。而 mBERT 的召回率最高（83.27%）。所有模型的特异性始终较低，凸显了准确区分中性文本与积极或消极情绪的共同挑战，需要进一步探索和完善。与其他多语言模型相比，所提出的模型的性能增益可能是由于在情感分析任务之前对句子进行翻译和清理。

该实验的结果对从事情感分析任务的研究人员和从业者具有重要意义。研究结果强调了翻译器和情感分析器模型选择对情感预测准确性的关键影响。此外，GPT-3 模型和拟议的集成模型的良好性能凸显了改进情感分析技术的潜在途径。它为使用LLMs在这个动态领域的未来研究打开了大门。

在这项研究中，我们比较了两种流行翻译器 LibreTranslate 和 Google Translate 的性能，并结合两种预训练的情感分析模型 Twitter-Roberta-Base 和 Bertweet-Base、一种大型语言模型 GPT-3 和两种多语言模型（XLM-t 和 mBERT）使用我们提出的集成模型以四种不同的语言：阿拉伯语、中文、法语和意大利语。我们的评估基于四个指标：精确度、召回率、F1 分数和特异性。我们的结果表明，谷歌翻译，与提议的

翻译者	情绪分析器模型	阿拉伯	中国人	法语	意大利语
语言方面的特异性					
自由翻译	Twitter-罗伯塔-基地	0.43	0.51	0.4	0.49
自由翻译	Bertweet-base	0.44	0.5	0.45	0.5
自由翻译	GPT-3	0.43	0.43	0.44	0.5
自由翻译	提出的整体模型	0.5	0.54	0.49	0.53
谷歌翻译	Twitter-罗伯塔-基地	0.48	0.7	0.42	0.55
谷歌翻译	Bertweet-base	0.48	0.69	0.48	0.57
谷歌翻译	GPT-3	0.46	0.54	0.45	0.42
谷歌翻译	提出的整体模型	0.54	0.73	0.51	0.52

表 8. 翻译器和情感分析器模型的不同组合之间的语言特异性得分。

型号名称	准确性	精确	记起	F1	特异性
提议的 Ensemble 模型—LibreTranslate	0.8491	0.7892	0.7771	0.7831	0.5201
提出的集成模型——谷歌翻译	0.8671	0.8091	0.8122	0.8106	0.5713
XLM-T	0.8025	0.8089	0.8052	0.807	0.6151
mBERT	0.7825	0.7164	0.8327	0.7701	0.5296

表 9. 多语言模型的实验结果比较。

集成模型，在所有四种语言中取得了最高的 F1 分数。我们的研究结果表明，谷歌翻译更擅长将外语翻译成英语。考虑到不同的语言-翻译器对可能需要不同的模型才能获得最佳性能，所提出的集成模型是这四种语言情感分析的最合适选择。

这项研究中提出的结果提供了强有力的证据，表明可以通过将外语情感翻译成作为基础语言的英语来对其进行分析。这一概念得到了以下事实的进一步支持：使用用英语训练的机器翻译和情感分析模型，我们在预测阿拉伯语、中文、法语和意大利语等非英语语言的情感方面实现了高精度。获得的结果表明，翻译器和情感分析器模型都显着影响情感分析任务的整体性能。它为营销、政治和社交媒体分析等各个领域的情感分析应用开辟了新的可能性。

挑战

尽管将外语翻译成基础语言进行情感分析有很多优点，但这种方法也存在一些挑战，这些挑战在本实验中已经出现，并且不同的研究也面临着这些挑战。
本节更详细地讨论这些挑战并探讨可能的解决方案。

挑战一：翻译准确性

外语情感分析遇到的主要挑战之一是翻译过程的准确性。机器翻译系统通常无法捕捉目标语言错综复杂的细微差别，从而导致翻译错误，从而影响情感分析结果的准确性。

解决翻译不准确的挑战的一种潜在解决方案是利用人工翻译或机器翻译和人工翻译相结合的混合方法。人工翻译通过考虑机器翻译可能忽略的语境因素、惯用表达和文化差异，对源文本进行更细致、更精确的翻译。然而，必须注意的是，这种方法在时间和成本方面可能是资源密集型的。尽管如此，它的采用可以提高准确性，特别是在需要细致语言分析的特定应用中。

或者，机器学习技术可用于训练针对特定语言或领域定制的翻译系统。在广泛的数据集上训练系统并采用专门的机器学习算法和自然语言处理方法可以提高翻译的准确性，从而减少后续情感分析中的错误。尽管它需要访问大量数据集和特定领域的专业知识，但这种方法为外语情感分析提供了可扩展且精确的解决方案。

挑战二：文化敏感性

翻译外语文本进行情感分析的另一个关键考虑因素涉及文化差异对情感表达的影响。不同的文化在传达积极或消极情绪方面表现出不同的惯例，这对翻译工具或人工翻译者准确捕捉情绪提出了挑战。

例如，某些文化可能主要采用间接手段来表达负面情绪，而其他文化可能会采取更直接的方式。因此，如果情绪分析算法或模型无法解释这些文化差异，那么精确识别翻译文本中的负面情绪就会变得困难。

为了减轻这种担忧，将文化知识纳入情感分析过程势在必行，以提高翻译文本中情感识别的准确性。潜在的策略包括使用特定领域的词典、针对特定文化背景策划的训练数据，或应用为适应文化差异而定制的机器学习模型。将文化意识融入情感分析方法中，可以更精细地理解翻译文本中表达的情感，从而能够跨不同语言和文化领域进行全面、准确的分析。

挑战三：惯用表达

翻译外语文本进行情感分析时的另一个挑战是惯用表达和其他特定于语言的属性，这些属性可能无法被翻译工具或翻译人员准确捕获。

习语代表比喻意义偏离组成词字面解释的短语。翻译惯用表达可能具有挑战性，因为比喻含义可能不会立即出现在翻译文本中。

为了熟练地识别翻译文本中的情感，必须全面考虑这些语言特定的特征，从而需要应用专门的技术。例如，采用根据目标语言的大量数据进行训练的情感分析算法可以增强在惯用表达和其他语言特定属性中辨别情感的能力。通过结合专门为解决惯用表达和其他语言特定特征而设计的技术，可以实现更精确和全面的情感分析，从而促进充分的跨语言理解和分析。

挑战四：翻译偏差

翻译外语文本进行情感分析的固有局限性在于翻译过程中可能引入偏见或错误。尽管机器翻译工具通常非常准确，但它们生成的翻译可能会偏离原文的保真度，从而导致翻译失败

捕捉源语言的复杂性和微妙之处。同样，人工翻译通常表现出更高的准确性，但在翻译过程中也难免会引入偏见或误解。

为了最大限度地减少翻译引起的偏见或错误的风险，在情感分析中细致的翻译质量评估变得势在必行。这种评估需要采用多种翻译工具或聘请多名翻译人员来交叉引用翻译，从而有助于识别潜在的不一致或差异。此外，可以采用诸如反向翻译之类的技术，从而将翻译后的文本重新翻译回原始语言，并与初始文本进行比较以辨别任何差异。通过采取严格的质量评估措施，可以有效减少翻译过程中引入的潜在偏差或错误，提高情感分析结果的可靠性和准确性。

挑战五：语言多样性

另一个可能的限制与翻译外语文本的实用性和可行性有关，特别是在涉及大量文本或存在重大挑战的语言的情况下。以大量用于情感分析的语料库或存在极其复杂的语言为特征的情况可能会使传统的翻译方法不切实际或无法实现。在这种情况下，替代方法对于有效进行情绪分析至关重要。

一种可行的途径是开发针对目标语言的复杂性而定制的特定于语言的情感分析算法。这些算法经过优化，可以解决所考虑的语言特有的独特语言特征、文化差异和情感表达模式。通过定制情感分析方法，可以规避与翻译相关的限制，从而促进准确的情感分析结果。

另一种方法涉及利用机器学习技术对来自目标语言的大量数据训练情感分析模型。该方法利用大规模数据可用性来创建强大且有效的情绪分析模型。通过直接在目标语言数据上训练模型，消除了翻译的需要，从而实现更有效的情感分析，特别是在关注翻译可行性或实用性的场景中。

通过采用这些替代方法，例如特定语言的情感分析算法或大规模目标语言数据的训练，可以有效解决外语文本翻译不切实际或不可行所带来的挑战，从而促进改善情感分析结果。

挑战六：处理俚语、口语、讽刺和讽刺

翻译外语文本进行情感分析的一项重大挑战涉及合并俚语或口语，这可能会让翻译工具和人工翻译人员感到困惑。俚语和口语在不同地区和语言之间表现出相当大的差异，这使得将其准确翻译成基础语言（例如英语）具有挑战性。例如，西班牙语评论可能包含许多俚语或口语表达，对于非流利的西班牙语使用者来说可能难以理解。同样，阿拉伯语的社交媒体帖子可能会使用缺乏语言和文化知识的个人不熟悉的俚语或口语。为了准确识别包含俚语或口语的文本中的情感，设计用于处理此类语言特征的特定技术是必不可少的。

翻译外语文本进行情感分析的另一个潜在挑战是反讽或讽刺，即使对于母语人士来说，这在识别和解释方面也很复杂。讽刺和讽刺涉及使用语言来表达与预期含义相反的意思，通常是出于幽默的目的。例如，法国评论可能会使用讽刺或挖苦来传达负面情绪；然而，法语不流利的人可能很难理解这种语气。同样，德语的社交媒体帖子可能会使用讽刺或讽刺来表达积极的情绪，但这对于那些不熟悉语言和文化的人来说可能很难辨别。为了准确识别包含讽刺或讽刺的文本中的情感，专门处理此类语言现象的技术变得必不可少。

值得注意的是，根据目标语言的大量数据进行训练的情感分析算法证明了检测和分析文本中特定特征的能力得到了提高。另一种可能的方法涉及使用经过明确训练的机器学习模型来识别和分类这些特征，并将它们分配为积极、消极或中性情绪。随后，这些模型可以通过结合俚语、口语、反讽或讽刺来对文本中传达的情感进行分类。这有助于更准确地确定所表达的整体情绪。

结论和未来的工作

本研究调查了使用不同的机器翻译和情感分析模型来分析四种外语情感的有效性。我们的结果表明机器翻译和情感分析模型可以准确地分析外语的情感。具体来说，谷歌翻译和提出的集成模型在精度、召回率和 F1 分数方面表现最好。此外，我们的结果表明，在翻译后使用基础语言（在本例中为英语）进行情感分析可以有效地分析外语中的情感。这项研究提供了一种集成模型，通过机器翻译和基础语言分析来进行外语情感分析，该模型在商业、社交媒体分析和政府情报等各个领域都有潜在的应用。该模型可以扩展到本研究中调查的语言以外的语言。我们承认我们的研究存在局限性，例如数据集大小和使用的情绪分析模型。这些限制应该在未来的研究中得到解决。

数据可用性

在当前研究期间生成和/或分析的数据集可根据合理请求从相应作者处获得。

收稿日期：2023 年 12 月 15 日；接受日期：2024 年 4 月 19 日
Published online: 26 April 2024

参考

1. Yadav, A. 和 Vishwakarma, D.K. 使用深度学习架构进行情感分析：综述。阿蒂夫。英特尔。修订版 53, 4335–4385 (2020)。
2. Gandhi, A.、Adhvaryu, K.、Poria, S.、Cambria, E. 和 Hussain, A. 多模态情感分析：对历史、数据集、多模态融合方法、应用、挑战和未来方向的系统回顾。信息。融合 91, 424–444 (2023)。
3. Cambria, E.、Das, D.、Bandyopadhyay, S. 和 Feraco, A. 情感计算和情感分析。情感分析实用指南 1-10 (2017)。
4. Sarker, I. H. 机器学习：算法、实际应用和研究方向。 SN 计算。科学。 2、160（2021）。
5. Das, R. 和 Singh, T. D. 多模式情绪分析：方法、趋势和挑战的调查。 ACM 计算。幸存者。 （2023）。
6. Mercha, E. M. 和 Benbrahim, H. 用于跨语言情感分析的机器学习和深度学习：一项调查。神经计算 531, 195–216 (2023)。
7. Oueslati, O.、Cambria, E.、HajHmida, M. B. 和 Ounelli, H. 阿拉伯语情感分析研究综述。未来一代。计算。系统。 112, 408–430 (2020)。
8. Dewaele, J.-M.、Petrides, K. V. 和 Furnham, A。特质情绪智力和社会传记变量对成人多语言者交际焦虑和外语焦虑的影响：回顾和实证调查。郎.学习。 58, 911–960 (2008)。
9. Chan, J.Y.-L.、Bea, K. T.、Leow, S. M. H.、Phoong, S. W. 和 Cheng, W. K. 最新技术：基于顺序迁移学习的情感分析综述。阿蒂夫。英特尔。修订版 56, 749–780 (2023)。
10. 萨拉梅，M.，穆罕默德，S.M.，基里琴科，S. 等人。翻译后的情绪：阿拉伯社交媒体帖子的案例研究。在 HLT-NAACL 767-777 (2015) 中。
11. Mohammad, S. M.、Salameh, M. 和 Kiritchenko, S. 翻译如何改变情感。 J.阿蒂夫。英特尔。 Res.https://doi.org/10.1613/jair.4787（2016）。
12. 张，C.，卡佩莱蒂，M.，普利斯，A.，斯特曼，T.和内姆科瓦，J。金融情绪分析中机器翻译的案例研究。见：机器翻译峰会（2017）。
13. Khanuja, S.、Dandapat, S.、Srinivasan, A.、Sitaram, S. 和 Choudhury, M. Gluecos：代码转换 NLP 的评估基准（2020）。预印本 arXiv：2004.12376。
14. Wahidur, R. S.、Tashdeed, I.、Kaur, M. 和 Lee, H.-N.通过微调的语言模型和及时的工程来增强零样本加密情绪。 IEEE 访问（2024）。
15. Xing, F. 设计用于金融情绪分析的异构LLM代理（2024）。预印本 arXiv：2401.05799。
16. 徐，S.等人。比较前推理：Llm - 用于领域专业文本分析的增强语义相似性度量（2024）。预印本 arXiv：2402.11398。
17. Uddin, M. A. 和 Sarker, I. H. 用于网络钓鱼电子邮件检测的可解释的基于变压器的模型：一种大型语言模型方法 (2024)。预印本 arXiv：2402.13871。
18. Rehan, M.、Malik, M. S. I. 和 Jamjoom, M. M. 使用迁移学习对多语言威胁文本识别的变压器模型进行微调。 IEEE 访问（2023）。
19. Demirtas, E. 使用机器翻译进行跨语言情感分析。 （埃因霍温科技大学研究门户网站，2013 年）。
20. Barriere, V. 和 Balahur, A. 使用多语言转换器和数据增强自动翻译改进非英语推文的情感分析 (2020)。预印本 arXiv：2010.03486。
21. 卡迪夫nlp/twitter-罗伯塔-基础情绪。拥抱脸 (2023)。
22. nlptown/bert-base-multilingual-uncased-sentiment。拥抱脸。
23. 雷德福，A.等人。语言模型是小样本学习者。副词。神经信息。过程。系统。 33（2020）。
24. Rosenthal, S.、Farra, N. 和 Nakov, P. SemEval-2017 任务 4：Twitter 中的情绪分析。第 11 届国际语义评估研讨会 (SemEval-2017) 502–518（计算语言学协会，2017 年）。 https://doi.org/10.18653/v1/S17-2088。
25. Keung, P.、Lu, Y.、Szarvas, G. 和 Smith, N. A. 多语言亚马逊评论语料库。 2020 年自然语言处理经验方法会议论文集 (2020)。
26. Vinayakumar, R.、SachinKumar, S.、Premjith, B.、Poornachandran, P. 和 Kp, S. Deft 2017 — 文本搜索 @ taln/recital 2017：对法语推文中的观点和比喻语言的深入分析。在《Défi Fouille de Textes》（2017）中。
27. 诺维埃利，N.等人。 SENTIPOLC 2016 数据集。 https://doi.org/10.57771/N279-Q780（2021）。类型：数据集。
28. 阿尔沙比，T.等人。社交媒体的日益放大：测量 2009-2020 年 Twitter 上 150 多种语言的时间和社会传染动态。 EPJ 数据科学。 10、15。https://doi.org/10.1140/epjds/s13688-021-00271-0 (2021)。
29. Semiocast - Twitter 统计上的热门语言 - Semiocast (2023)。
30. 林瓜语。 2022 年（2022 年）世界上使用最多的 20 种语言。
31. 自由翻译。 Libre 翻译 API 2021。（访问日期：2023 年 4 月 26 日）； https://libretranslate.com/。
32. 谷歌翻译。 （2023 年 4 月 27 日访问）； https://translate.google.com/about/intl/en_ALL/。
33. 沃尔夫，T.等人。拥抱脸的变形金刚：2019 年最先进的自然语言处理（2023 年 4 月 27 日访问）； https://huggingface.co/transformers/。
34. Loureiro, D.、Barbieri, F.、Neves, L.、Anke, L. E. 和 Camacho-Collados, J. Timelms：来自 Twitter 的历时语言模型（2022 年）。 arXiv：2202.03829。
35. Wiriyathammabhum, P. Tedb 对 2022 年委婉语检测共享任务的系统描述 (2023)。 arXiv：2301.06602。
36. Schmidt, S.、Zorenböhmer, C.、Arifi, D. 和 Resch, B. 对与 2022 年 Twitter 收购相关的地理参考推文进行基于极性的情感分析。信息https://doi.org/10.3390/info14020071 (2023)。
37. Barbieri, F.、Espinosa Anke, L. 和 Camacho-Collados, J. Xlm-t：Twitter 中用于情感分析等的多语言语言模型。语言资源和评估会议记录 258-266（欧洲语言资源协会，2022 年）。
38. Devlin, J.、Chang, M.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向转换器的预训练。协调委员会（2018）。 arXiv：1810.04805。
39. Klubička, F.、Toral, A. 和 Sánchez-Cartagena, V. M. 机器翻译系统的定量细粒度人类评估：英语到克罗地亚语的案例研究。马赫。译。 32, 195–215 (2018)。
40. Daems, J.、Vandepitte, S.、Hartsuiker, R. J. 和 Macken, L. 识别对译后编辑工作影响最大的机器翻译错误类型。正面。心理。 8、1282（2017）。

41. 基于在线评论情感分析的跨文化学习资源推荐方法及语料库构建。第五届艺术、设计和当代教育国际会议（ICADCE 2019）271-278（亚特兰蒂斯出版社，2019）。

42. Mohammad, S. M. 情感分析：自动检测文本中的效价、情绪和其他情感状态。情绪测量 323-379（爱思唯尔，2021 年）。

43. Singh, M.、Kumar, R. 和 Chana, I。印度语言机器翻译系统：建模技术、挑战、开放问题和未来研究方向回顾。拱。计算。方法工程。 28, 2165–2193 (2021)。
44. Vanroy, B. 翻译中的句法困难。博士论文（根特大学，2021）。
45. 喀什加里，公元。翻译不可译的悖论：从阿拉伯语翻译成英语时的对等与非对等。 J.沙特国王大学。郎.译。 23, 47–57 (2011)。

46. 戈伊米尔·维拉科巴 (Goimil Vilacoba)、V. 詹姆斯·乔伊斯 (V. James Joyce) 翻译：《尤利西斯》西班牙语和加利西亚语版本中的口语、粗俗、惯用语和文化表达。 UDC 存储库 (2014)。
47. Reyes, A.、Rosso, P. 和 Veale, T. 检测 Twitter 中讽刺的多维方法。郎.资源。评估。 47, 239–268 (2013)。

48. Joshi, A.、Bhattacharyya, P. 和 Carman, M. J. 自动讽刺检测：一项调查。 ACM 计算。幸存者。 (CSUR) 50, 1–22 (2017)。

致谢

作者感谢沙特阿拉伯利雅得沙特国王大学在线对话和文化传播研究主席资助这项研究。

作者贡献

作者贡献 概念化和数据管理：M.S.U.M.,T.B.萨瓦尔和 M.M.K.形式分析、调查和方法：M.S.、S.A. 和 M.F.M.监督和可视化：M.F.M.写作——初稿：M.S.U.M.和 M.M.K.写作评论和编辑：M.S.和 S.A.

资金

这项研究由沙特阿拉伯利雅得沙特国王大学在线对话和文化传播研究主席资助。

利益竞争

作者声明没有竞争利益。

附加信息

信件和材料请求应发送至 M.S.或 M.F.M.