arXiv:2304.02945v1 [stat.AP] 6 Apr 2023

使用 BERT 对开放式问题进行多标签分类

Matthias Schonlau*[1]、Julia Weiß[2],和 Jan Marquardt[2]

[1]滑铁卢大学统计与精算科学系[2]曼海姆莱布尼茨社会科学研究所 GESIS

2023年4月7日

摘要

调查中的开放式问题很有价值，因为它们不会限制受访者的回答，从而避免产生偏差。然而，开放式问题的答案是文本数据，较难分析。传统上，答案会按照编码手册中的规定进行人工分类。在过去10年里，研究人员一直在尝试实现编码自动化。大部分工作都集中在较简单的单标签预测问题上，即将答案归入单个代码。然而，需要进行多标签分类（即分配多个代码）的开放式回答十分常见。在社会科学调查中，这类开放式回答也常常只涉及少量标签。在轻度多标签分类中，每段答案文本所含标签的平均数量相对较少（例如 < 1.5）。例如，我们分析的数据集提出了这样一个问题："您认为目前德国最重要的政治问题是什么？"尽管问题只询问一个问题，但有些答案包含多个问题。当然，每段答案所涉及问题（或标签）的平均数量仍然较少。

本文聚焦于社会科学调查中开放式调查问题的文本答案的多标签分类。我们在一项德国社会科学调查——GLES面板调查（N=17,584，55 个标签）中，针对德语，将基于 Transformer 的架构 BERT 的性能与传统多标签算法（二元关联、标签幂集、ECC）进行比较。由于我们的数据集要求每个文本答案至少对应一个标签，因此我们还提出了一种修改方案，以应对分类方法无法预测任何标签的情况。我们使用0-1 损失评估这些算法：仅当所有标签都被正确预测时，损失才为零；某一个标签上的错误会导致完整损失（1）。这种损失与人工文本分类的实际情况相符：即使只需复核一个可疑标签，也要使用所有标签对整个文本答案进行编码。

我们发现，在所考察的方法中，使用 BERT 进行分类（强制至少预测一个标签）的 0-1 损失最低（13.1%；其他方法为 18.9%–21.6%）。正如预期，正确预测对应单个标签的答案文本（损失为 7.1%），比预测对应多个标签的答案文本（损失为 ∼50%）容易得多。由于 BERT 仅对 1.5% 的答案预测零个标签，强制至少预测一个标签虽然有效且值得推荐，但最终并不会大幅降低 0-1 损失。

我们的研究对社会科学家具有重要意义：1) 我们已证明，使用BERT 进行多标签分类可以处理德语开放式问题。2) 对于轻度多标签分类任务而言，如今损失似乎已经小到足以实现完全自动分类。此前，损失更为显著，通常需要采用半自动方法。3) 使用 BERT 进行多标签分类只需要一个模型。主要竞争者 ECC 是一种迭代方法，会逐一迭代单标签预测。

1 引言

开放式问题很有用，因为它们避免了限制受访者的答案选项可能带来的偏差。然而，由此产生的文本答案不便于分析。因此，开放式问题在调查研究中可能未得到充分利用。当开放式问题不可或缺时，通常会依据预先制定的编码方案进行人工分类。

*电子邮箱：schonlau@uwaterloo.ca

过去 10 年间，研究逐渐转向自动分类开放式问题的文本答案（Schonlau 和 Couper，2016；Schierholz 和 Schonlau，2021；He 和 Schonlau，2020）。通常会对部分文本答案（例如 300—500 条）进行人工编码，并将其作为训练数据。然后，按照词袋法计算 n 元语法变量（Schonlau 等人，2017）。n 元语法变量是指某个词或词序列在答案文本中出现的次数。最后，利用监督学习，在训练数据上训练学习器，并用其对文本答案进行分类。由于准确率往往不足以满足研究需求，因此有人提出了半自动算法（Schonlau 和 Couper，2016），自动分类容易分类的答案，并对难以分类的答案进行人工分类。

大多数工作都投入到将开放式问题的文本答案分类为恰好一个类别的研究中。这称为单标签分类。文本答案也可以标记多个标签；例如，当受访者在答案中提到多个主题时。这类似于全选题，因为这类问题也允许选择多个答案。这种多标签分类要困难得多：分类器不仅必须正确地对多个标签进行分类（而不是仅对一个标签进行分类），同时还必须确定对于给定的文本答案，多少个标签是合适的。

尽管许多开放式问题属于多标签问题，但针对社会科学领域多标签开放式问题的研究相对较少。尤其是，尚缺乏将 BERT 这样的基于 Transformer的架构与标准多标签算法的性能进行比较的研究。使用BERT 分析英语以外语言的社会科学开放式问题的研究也同样缺乏。

本文其余部分的结构如下。第 2 节回顾了算法、BERT，以及社会科学调查中开放式问题的多标签分类。第 3节介绍了本文使用的德语数据集和实验设置。第 4 节给出了 BERT 与多标签分类算法的比较结果。第 5 节以讨论作结。

2 背景

2.1 多标签分类算法

对多标签数据进行分类最简单、也可能是最常用的方法是二元关联（BR）（Tsoumakas 和 Katakis，2007）。BR 将多标签问题视为 L 个独立的单标签回归问题，其中 L 是标签的数量。这种方法易于实现，但忽略了标签之间的相关性。

标签幂集（LP）（Tsoumakas 和 Katakis，2007）方法将多标签问题转换为单标签问题：它将训练数据中出现的每个标签集（标签组合）视为一个单独的标签。例如，对于 10 个标签，可能的两两标签组合有 10 选 2 种。尽管这些标签集中有许多不会出现在训练数据中，但标签集的数量仍可能很大。LP 是一种单模型方法，它只允许使用训练数据中出现过的标签集，从而在模型中考虑标签之间的相关性。虽然 LP 确实考虑了标签之间的相关性，但类别数量过多会使这种方法在实践中无法使用。一个可能的例外是轻度多标签数据，其平均标签数量较少，因此产生的标签组合也较少。

分类器链（CC）是一种迭代方法，通过将其他标签的指示变量作为额外的 x 变量纳入模型，从而考虑标签间的相关性（Read 等人，2011年）。与二元相关（BR）一样，标签是逐个预测的：首先按常规方式预测第一个标签。第二个标签的回归模型将预测得到的第一个标签作为额外的 x 变量。第三个标签的回归模型则加入两个额外的 x 变量，即预测得到的第一个和第二个标签，依此类推。为减少对标签顺序的依赖，作者提出了分类器链的集成版本 ECC。ECC 使用随机标签顺序，在自助采样上反复运行分类器链。由于 ECC 优于分类器链，通常只报告 ECC 的结果。ECC 往往具有较高的预测性能（Madjarov 等人，2012；Schonlau 等人，2021）。

多标签分类的其他方法例如见Tsoumakas 和 Katakis(2007）和Zhang 和 Zhou (2013）。

2.2 BERT

双向变换器表示（BERT）（Devlin 等人，2019年）是一种基于变换器架构（Vaswani 等人，2017）的语言模型。BERT 能够同时考虑掩码（或隐藏）词的左侧和右侧上下文，这被称为双向注意力。BERT 将输入词转换为数值向量。在预训练期间，词语的数值表示以及超过 100 万个系数都会被优化，以尽可能准确地预测掩码（隐藏）词。（还有第二项任务，即下一句预测，但其重要性较低。）预训练的目的是在海量的无关文本数据上学习语言模型——即良好地预测掩码词。随后，微调会使用感兴趣的特定数据，优化与分类相关的准则。对于单标签分类，BERT 为每个类别设置一个输出变量（y 变量）。适当的损失函数和激活函数确保每个类别的估计概率之和为 1。预测概率最大的类别。通过为输出变量指定不同的损失函数和激活函数，BERT 也可用于多标签分类。对于多标签分类，BERT 同样为每个类别设置一个输出变量。然而，估计概率之和不再必须为 1。预测的标签集合包括所有对应概率大于阈值概率的标签，该阈值通常为 0。5。

总而言之，多类别分类和多标签分类惊人地相似：它们使用相同的语言模型，输出变量的数量也相同，每个类别对应一个变量。它们的区别仅在于输出概率的约束和损失函数。由于多标签预测来自单个模型，而不是多个彼此独立的模型，因此无需额外处理标签之间的相关性。

2.3 社会科学调查中开放式问题的多标签分类

开放式问题的多标签分类相对而言研究较少。由于多标签算法是准确率较低的难题（子集准确率与 0-1 损失等价；二者都要求正确预测 所有 文本答案的标签），Gweon 等人（2020）提出了一种用于编码多标签开放式问题的半自动算法。其思路是：当算法相当确定分类正确时，自动地 进行分类，否则手动地 进行分类。这种半自动分类确保了总体上的高准确率。

使用n元语法变量对文本进行编码时，用于开放式问题的多标签分类算法（见第 2.1节）在使用 0-1损失时表现优于二元相关（Schonlau 等人，2021）。标签之间的相关性越强，表现似乎越好。

Meidinger 和Aßenmacher (2021)提供了 19 个英语多标签基准数据集，并使用BERT和其他基于 Transformer 的模型对其进行评估。令人惊讶的是，他们发现，就0-1 损失而言，基线模型逻辑回归（一对其余）在 10 次比较中有 3 次优于 BERT。

Ford (2023) 使用 BERT对文本答案进行编码（而不是用于分类）。随后将这些编码作为逻辑回归的输入；每个标签对应一个逻辑回归。各个逻辑回归彼此独立，但共享相同的 x 变量。文本数据为英文。

标签的平均数量（也称为 基数）是刻画多标签分类任务的一种方式。在本文中，如果一个分类任务平均每个文本回答包含少于 1.5 个标签，我们称其为 轻度多标签任务。相当大一部分多标签开放式问题实际上只是轻度多标签。例如，如果要求受访者只提供一个理由，即使部分受访者不遵守这一要求，数据中的理由或标签平均数量也会较少。其他研究报告称，三个数据集中的标签平均数量分别为 1.15、1.44 和 2.77 个（Schonlau 等人，2021）。在Meidinger 和 Aßenmacher（2021）所报告的数据集中，10 个数据集中的 4 个仅属于轻度多标签。

3 Methods

我们在一个大型数据集上将多标签 BERT 与多种多标签算法进行比较。首先介绍该数据集（3.1），然后详细说明模型的架构和调优方式（3.2），

我们还考虑模型预测零个标签这一特殊情况（3.3），并讨论评估标准（3.4）。

3.1 数据

我们使用德国纵向选举研究（GLES）（GLES，2021）中2016年选举竞选面板第1波的数据。这项网络调查以德语提出了以下问题："您认为目前德国最重要的政治问题是什么？请只说出一个问题。"调查对象来自一个访问面板，并按性别（2类）、教育程度（3类）和年龄（5类）设置配额。共有18,128名受访者参与。其中，97.0%回答了开放式问题，共获得17,584条文本答案。这里，"回答"是指答案至少包含一个字符。所有非空答案都纳入预测，但任何无意义的答案都被归类为"未回答"（示例见表1）。

平均答案长度为3.57个词。61.8%的答案仅包含一个词；第三四分位数为3个词。单词的最大长度为333个单词。

每个答案都被人工编码为 55 个不同答案标签中的一个或多个。标签数量最多为 5 个，平均为 1.17 个。12。6% 的答案有多个标签。表 1 列出了最常见的 10 个标签及其占比（占所有标签的比例）。在这些数据中

表1：排名前10的标签、百分比及示例。百分比指标签总数，而非回答文本总数。

标签名称	code	%	示例
移民、庇护与融入	3750	47.93	"关闭边境"，"新移民太多"
社会正义	3720	5.38	"不平等"、"平等机会"
价值观、政治文化与社会批评	2400	4.40	"向右倾向"
激进化、极端主义	2411	3.35	"打击极端主义"、"伊斯兰威胁"
贫困	3721	3.03	"不公正的社会制度"、"社会问题"
未作答	-99	2.84	"..."，"xxx"
人口结构变化和退休	3740	2.78	"社会保障缴费"
政治结构与进程	2000	2.60	联邦制, 总理
恐怖主义	3412	2.57	"恐怖主义"，"攻击"
对政党的批评（具体）	2439	1.61	为政党命名

集中，移民标签（3750）占标签总数的近一半。只有百分之几的标签是"没有答案"，而属于"不了解"或统称类别"其他类别"的标签则更少（不在前10个标签之列）。当分析单位是单个标签时，评定者间信度的 kappa为 $\kappa = 0.88$。当分析单位是文本答案时，即所有标签必须同时正确，评定者间信度的 kappa 为 $\kappa = 0.82$。

3.2 模型

我们评估了几种用于多标签预测的算法：多标签 BERT、二元关联、标签幂集和分类器链。下面详细介绍所使用的算法。数据被随机划分为训练数据（60%）、评估数据（20%）和测试数据（20%）。所有算法均使用相同的数据划分。

BERT

我们的语言模型是 bert-base-german-cased（Chan 等人，2019年）。"base"指的是 12 个编码器层逐层堆叠、12 个注意力头和 768 个隐藏单元。该模型约有 1.1 亿个参数。我们通过 Thilina Rajapakse 的 Python 软件包"Simple Transformers"调用该模型，该软件包基于 Pytorch 构建。

我们使用了 55 个输出变量，分别对应数据集中的 55 个类别。每个输出变量都是一个指示变量。我们在评估数据上对训练轮数（5、10、20）和学习率（1e-3、1e-4、1e-5）进行了网格搜索。最终运行使用了 20 个训练轮数和一个学习

率 $lr = 1e-5$。最终运行的损失函数与使用默认学习率 $lr = 1e-4$ 并训练 10 或 20 个训练轮数的运行结果之间差异极小。

预处理

预处理适用于使用 n 元语法变量的多标签算法。我们删除了仅由单个字母组成的“单词”。德语中不存在单字母词（与英语不同，例如“a”和“I”）。其次，受访者书写德语变音字母的方式不一致（¨a 或 ae、¨u 或 ue、¨o 或 oe，以及 ß 或 ss）。因此，我们将所有变音字母都转换为相应的替代形式（ae、ue、oe、ss）。第三，两个德国政党的名称可能会与对应的常见德语单词混淆。绿党“die Gr¨unen”也是一种颜色，而“die Linke”也有“左边”之意。为将政党与对应的单词区分开来，我们使用正则表达式，在创建 n 元语法变量之前，将所有提及该政党的内容替换为（die_gr¨unen 和 die_linke）。我们还将所有单词转换为小写、删除标点符号并移除了所有数字。为避免意外删除重要单词，我们没有移除停用词。

接着，我们采用词袋法将文本转换为数值变量。具体来说，我们使用了 HanTa 词形还原器（Wartena，2019年）。这是一种基于隐马尔可夫模型、针对德语训练的简单词形还原方法。HanTa 词形还原器生成了一元语法（即单个词的计数）。最后，我们应用了 TF-IDF 转换。这种转换会为较不常见的词和较大的计数赋予更高权重。

支持向量机与二元关联

我们使用了 Python 的“scikit-learn”库中支持向量机的实现（Pedregosa 等人，2011年）。对于二元关联，我们采用“一对其余”的分类方法，以支持向量机作为基学习器，并使用常见的 0.5 阈值。对于线性核和径向核，我们都进行了网格搜索，以确定合适的超参数值 C 和 gamma（gamma 仅适用于径向核），从而最小化 0-1 损失。x 变量是 TF-IDF 变量。模型在训练数据上训练，并在验证数据上评估。对于二元关联，网格搜索确定使用线性核，且 $C = 100$。

集成分类器链

我们使用“sklearn.multioutput”中的“ClassifierChain”函数创建了 10 个分类器链，并以支持向量机作为基学习器。我们采用了常用的 0.5 分类阈值。支持向量机的参数与二元关联方法相同，即使用线性核和 $C = 100$。

标签幂集

训练数据中，55 个标签共有 813 种独特组合。标签幂集只需要一个单标签分类器。与之前一样，我们使用网格搜索来优化验证数据上的 0-1 损失。对于标签幂集方法，网格搜索确定使用径向核以及 $C = 100$ 和$gamma = 0.5$。

3.3 预测出零个标签时

数据集中的所有答案文本都至少有一个标签。如果答案不符合某个主题类别，则根据情况为其指定“无答案”“不知道”或“其他类别”中的一个标签。虽然文本始终有一个标签，但算法可能预测出零个标签。（标签幂集算法是一个例外：由于标签幂集会将问题转换为单标签预测，因此它必须至少预测一个标签。）我们知道，零个标签的预测必然是不正确的。因此，当多标签算法预测出零个标签时，我们用单标签预测所预测的标签来替代。在表格中，我们将这种方法称为“BERT最少一个标签”。同样的方法也可以用于其他算法，但为清晰起见，我们只将其应用于获胜算法BERT。

3.4 评估标准

对于单标签预测，计算损失或等价地计算准确率只有一种方法。存在多个标签时，常用的损失函数至少有两种：0-1 损失和汉明损失。当所有标签都预测正确时，0-1 损失为 0（无损失），否则为 1。这是一个非常严格的标准，因为在我们预测的 55 个标签中，即使只错一个，也会产生完整的损失。汉明损失计算预测错误的标签数的平均值。这种损失函数更宽容；当标签的平均数量较多或很多时，可能会很有用。

我们有55个标签，但标签的平均数量为1.17。由于标签的平均数量较少，0-1 损失准则在此处是合适的。这里的汉明损失极小，因为 55 个标签中的大多数始终会被正确预测为 0。

计算机科学论文中常见的其他评估标准包括宏平均 F、微平均 F 和 LRAP。（LRAP会考虑真实标签在预测标签中的排名位置。）我们认为，0-1 损失是社会科学中最相关的标准。如果怀疑某个回答文本被错误分类，编码员就必须手动为该回答文本编码。无论错误的是一个标签还是多个标签，这项工作所需的投入都是相同的。

4 结果

在第 4.1节中，我们评估 BERT、二元关联、标签幂集和集成分类器链（ECC）的 0-1损失。对于 BERT，我们还评估强制至少输出一个标签后的0-1 损失。第 4.2节考察错误预测，以及我们能从中学到什么。鉴于 BERT 在单标签预测方面的特殊优势，第 4.3 节提出了一种半自动方法。

4.1 主要结果

表2 显示了各方法总体上以及按真实标签数量划分的 0-1 损失。BERT 和至少有一个标签的 BERT 的 0-1 损失远低于其他方法。根据 0-1 损失，它们仅将略高于 13% 的文本回答分类错误。正如预期，二元关联的表现最差。

表2：测试数据中各方法按标签数量和真实标签数量划分的 0-1 损失。算法按总体损失从低到高排序。

	总体	1 个标签	2 个标签	3 标签	4 个标签	5 个标签
BERT最少一个标签	0.1308	0.0705	0.4893	0.8225	0.9130	1.0000
BERT	0.1347	0.0749	0.4893	0.8225	0.9130	1.0000
ECC	0.1891	0.1315	0.5319	0.8387	0.9565	1.0000
标签幂集	0.1959	0.1063	0.8116	1.0000	0.9565	1.0000
二元关联	0.2161	0.1464	0.6657	0.9032	1.0000	1.0000

按真实标签数量考察损失后，我们发现，带有多个标签的回答远比带有单个标签的回答更难预测。对于 BERT，标签数为 1 时，0-1 损失约为 7%；标签数为 2 时，损失则跃升至约 49%。事实上，标签越多，0-1 损失就越高。

算法的标签数量分布如表3所示。与之前一样

Table 3：测试数据中按预测方法划分的标签数量分布（%）

	0 个标签	1 个标签	2 个标签	3 个标签	4 个标签	5 个标签
true	0.0	88.0	9.4	1.8	0.7	0.3
BERT	1.5	89.0	8.1	1.1	0.2	0.0
ECC	2.8	88.4	8.0	0.8	0.0	0.0
标签幂集	0.0	96.0	3.6	0.3	0.0	0.0
二元关联	8.7	84.9	5.9	0.5	0.0	0.0

如前所述，该数据集仅具有轻度多标签特性：88%的数据为单标签。与

BERT和其他算法较少预测多个标签。二元关联的表现尤其不佳，有 8.7% 的情况下预测零个标签。虽然可以通过强制至少预测一个标签来改善性能，但显然其表现不会像BERT 那么好，因为当标签数量较多时，二元相关的表现也更差（见表 2）。标签幂集无法预测零个标签，因为多标签问题被转换为单标签问题。

强制 BERT 预测至少包含一个标签（“BERT最少一个标签”）时，总体损失的降幅绝对值较小，约为 0.4%（表 2）。然而，BERT 的预测中只有 1.5% 为零个标签（表 3）。考虑到这些答案对 BERT 来说很难分类，正确分类 0.4 个百分点（总计 1.5 个百分点）已经相当不错。

4.2 错误预测

BERT 将少量答案预测为没有标签，尽管所有答案至少都有一个标签。这些答案的词数往往高于整个数据集的平均词数。被预测为零个标签的答案包括例如“ich bininpolitischen dingenneutral!”（“我在政治问题上保持中立”），以及只有一个词的答案“boeh”（含义不明确）和“Menschlichkeit”（“人性”）。这些答案的真实标签非常多样。

接下来，我们考察那些本身为单标签、且 BERT和 ECC也预测为单标签但标签错误的答案。ECC 经常错误地分配 -99（“无答案”）或主导类别（出现频率最高的类别）。当 ECC 错误地分配“无答案”时，该答案通常包含拼写错误。在预处理过程中，我们没有尝试纠正拼写错误，而这似乎对 ECC 的影响远大于对 BERT 的影响。

当 ECC和 BERT 都预测出错误标签时，正确代码最常见的是 2400（“价值观、政治文化与社会批评”）、2000（“政治结构与进程”）和 5000（“其他”）。这三个类别的共同点是，它们涵盖了广泛的主题。表 4 给出了一些真实代码为 2400 时的示例；附录中的表 6 展示了相应的德语原始文本答案。值得注意的是，ECC 和 BERT 经常预测出相同的错误标签，而这些错误标签通常是合理的。例如，“不断违反我们的基本法，以及默克尔女士毫无替代方案的政策”被 ECC 和 BERT都归入代码2441“对政治家的具体批评”，但人工编码时归入了更一般的代码2400（见表 4）。代码 2441也是正确的，因为当时的德国总理安格拉·默克尔受到了具体批评。这种标签差异或许更适合被描述为编码手册中一般代码与更具体代码之间存在的歧义。

表 4：正确代码为 2400 时，ECC 和 BERT 预测错误的代码。常见代码的说明见表 1。代码 5000 表示“其他”。

回答	ECC	BERT
“他们不照顾德语”	3100	5000
“利己主义”	-99	5000
“带来满足感”	5000	5000
“目标不明确”	-99	5000
“更好地解释政治”	2000	2000
“未投票者太多”	-99	-98
“贯彻落实一致”	-99	2000
“一致性”	-99	5000
“对我们的基本法和默克尔夫人政策的持续违背	2441	2441
没有替代方案	2441	2441
“沟通”	5000	5000
“说话太少了”	2000	2000
“人人都想自己煮自己的汤”	5000	2000
“通常在面对和处理问题时缺乏诚实		
直截了当的方式，以及有时错误的理解		
的政治正确。”	3750	2420
“无法设定优先级”	-99	2000

当答案是单个标签，但 BERT 或 ECC 预测出多个标签（82 个案例）时，该

真实标签几乎总是出现在预测标签中。对于 BERT，唯一一次真实标签未包含在预测标签中时，真实标签（人工编码）可以说是个错误：“对德国公民的社会支持”被人工编码为标签 3700（“社会政策”）。BERT 则预测了代码 2400（“价值观、政治文化与社会批评”）和 3720（“社会正义”）。由于模型的多标签预测几乎总是包含实际的单个标签，因此问题在于额外的标签是否恰当。我们发现，BERT 预测的额外标签始终合理，而 ECC 则会预测一些不恰当的标签。

表 5 中列出了 ECC 错误预测多个标签时的一些糟糕预测结果。（附录中的表 7 展示了原始德语文本答案。）例如，对于“公平分配税收”，ECC 预测为 2400（“价值观、政治文化与社会批评”）和 -99（“无答案”），而正确答案是 4320（“税收政策”）。ECC 多次难以正确处理“无答案”标签。

表 5：正确代码为单个标签时，ECC 多标签预测不佳的示例

	True	ECC 标签 1	ECC 标签 2	ECC 标签 3
“公平分配税收”	4320	2400	-99	
“细微的动作声响很大，让人觉得动作很大”	2400	3412	3750	
“穆斯林”	3750	2411	-99	
“无论如何，不是寻求庇护者，那只是夸张的”	-99	3720	3750	
“养老金或退休年龄时的贫困”	3740	3721	3740	4100

最后，考虑那些属于多标签的问题，而 ECC 或 BERT 预测出了多个标签，但这些标签并不正确的情况。ECC 和 BERT往往比正确结果预测出更少的多个标签。在极少数情况下，它们会预测出完全错误的标签。其中一个罕见的例子是“难民融入与德国人口老龄化”，其真实标签为 3750（“移民、庇护与融入”）和 3740（“人口结构变化与养老金”），而BERT预测为 3750（“移民、庇护与融入”）和 2400（“价值观、政治文化与社会批评”）。

4.3 一种替代性的半自动方法

半自动算法会自动编码部分答案，并手动编码其他答案，以提高整体准确率或其他指标。纯粹采用多标签策略的一种半自动替代方案如下：如果 BERT 预测出一个标签，就采用该标签；如果 BERT 预测出多个或零个标签，则手动编码。根据表 3 ，我们知道 BERT 将 89.0% 的文本答案预测为单标签。对于 BERT 预测为 1 个标签的文本答案，0-1 损失为 0.086。（表 2中未列出这一结果，该表显示的是基于真实标签分布计算的损失。）

对于这些数据，半自动方法建议手动编码 11% 的答案。其余答案的损失为 0.086%。这种方法需要在降低损失（8.6% 对比 13.1%）与手动编码 11% 的答案之间权衡。是否值得这样权衡，取决于分析人员的判断。此外，这一策略或许只适用于多标签程度较低的数据集，否则被预测为非单标签的数据比例会增加。

5 讨论

多标签 BERT 具有几个决定性优势：a) 在0-1 损失方面（是否正确预测所有标签），多标签 BERT有力地胜过通常的多标签算法。预训练在训练数据较少时可能尤其有利。我们的训练数据规模更大（N=17,584），超过了社会科学中开放式问题通常使用的训练数据规模，然而 BERT 仍明显胜出。b) 由于多标签问题较为困难，多标签方法一直难以做到足够准确，因而无法用于对质量要求较高的场景。例如，在对开放式问题进行分类时，ECC 的 0-1 损失介于 35.8% 和 44.9% 之间（Schonlau 等人，2021）。对于轻度多标签问题，我们首次达到了足够低的 0-1 损失，我们认为这使得（完全）自动分类具有吸引力。也就是说，半自动分类是可选的，而非必需的。c) 多标签BERT 从一个单一模型中获得预测结果。对于其他多标签

引入的算法需要为每个标签建立一个模型。使用多标签 BERT 无需通过逐一迭代各个模型来改进预测结果。d) 对错误预测的分析显示，根据编码手册，BERT 的一些错误预测实际上是合理的。此外，与领先的基于 n-gram 的方法 ECC 相比，BERT 对拼写错误的鲁棒性强得多。

我们发现了以下局限性：a) 需要使用 Python。社会科学家更熟悉 Stata 和 R 等程序。使用 BERT 需要用 Python 进行编程。我们在补充材料中提供了代码，以在一定程度上减轻这一负担。b) 我们使用了德语 BERT，这是一种特定语言模型。由于德语 BERT 不如原始 BERT 成熟，并且单标签分类取得了成功（Gweon 和 Schonlau，2023），因此没有理由认为该方法无法用于英语。c) 我们展示的多标签算法使用支持向量机作为基学习器。然而，我们知道，在对开放式问题的回答进行分类时，梯度提升和随机森林等其他算法的表现也相近（He 和 Schonlau，2022；Gweon 和 Schonlau，2023）。

我们使用的数据集每篇文本平均包含 1.17 个标签，因此属于轻度多标签数据集。如果我们拥有一个平均标签数量较高的数据集，情况会发生什么变化？0-1 损失可能会增加，因为标签越多，出现错误预测的机会就越多。此外，另一种半自动方法将不可行，因为必须手动编码的文本数量过多。

对于英语，Meidinger 和 Aßenmacher （2021）发现，他们的基线模型——采用逻辑回归的二元相关性模型——有时在0-1 损失方面胜过BERT。然而，在 4 个具有轻度多标签特征的数据集（数据集 7-10）中，BERT 和其他基于 Transformer 的模型胜过基线模型。

基于 N-gram的模型往往比处理较长的文本答案更适合处理较短的文本答案，因为在较短的文本答案中，单个词往往能够表明其所属标签。由于我们数据中超过一半的答案是单词答案，即使是在通常被认为是其优势的任务上，基于 N-gram的模型也败给了BERT（基于 Transformer 的分类器）。我们的结果与近期关于单标签分类中短文本的研究一致，该研究得出结论：“Transformer 是短文本分类器”（Karl 和 Scherp，2022），这意味着诸如 BERT这样的基于 Transformer 的文本分类器在短文本上表现极为出色。

致谢

我们衷心感谢加拿大社会科学与人文研究理事会（SSHRC）提供的资助，资助项目编号为 435-2021-0287（首席研究员：Schonlau）。我们感谢 Stefan Eschenwecker 和 Jan-Eric Meurer 的支持。

数据可用性与补充材料

数据可从 https://search.gesis.org/research_data/ZA6838?doi=10.4232/1.13783(GLES，2021年）。我们仅使用面板调查的第一轮数据。第一轮的参与者可通过变量 *kp1 participation*识别。开放式回答的文本条目可在第一轮的单独 CSV 文件中找到，并可通过标识符“lfdn”与其他数据合并。我们用于实现不同模型的代码作为补充材料提供。

参考文献

Chan, B., T. M¨oller, M. Pietsch, and T. Soni (2019年)。德语 BERT。https://huggingface.co/bert-base-german-cased。

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019年)。BERT：用于语言理解的深度双向变换器预训练。arXiv 预印本 arXiv:1810.04805。

Ford, S. S.（2023）。开放式调查回答的自然语言处理方法：以“最重要的国家问题”为例。2023年2月15日在范德堡大学及线上发表的演讲。

GLES（2021）。GLES Panel 2016–2021，第1–15轮。数据档案，科隆：ZA6838。数据文件版本 5.0.0。DOI=10.4232/1.13783。Gweon, H. 和 M. Schonlau（2023）。使用 BERT 对开放式问题进行自动分类。《调查统计与方法学期刊》。即将发表，arXiv 预印本 arXiv:2209.06178。Gweon, H.、M. Schonlau 和 M. Wenemark（2020）。多标签开放式问题的半自动分类。《调查方法学》46（2），265–282。He, Z. 和 M. Schonlau（2020）。将开放式问题自动编码为多个类别：是否以及如何使用双重编码数据。《调查研究方法》14（3），267–287。He, Z. 和 M. Schonlau（2022）。一种模型辅助方法，用于发现开放式问题人工编码中的错误。《调查统计与方法学期刊》10（2），365–376。Karl, F. 和 A. Scherp（2022）。Transformer 是短文本分类器：基于基准数据集和真实世界数据集研究归纳式短文本分类器。arXiv 预印本 arXiv:2211.16878。Madjarov, G.、D. Kocev、D. Gjorgjevikj 和 S. Dˇzeroski（2012）。多标签学习方法的广泛实验比较。《模式识别》45（9），3084–3104。Meidinger, M. 和 M. Aßenmacher（2021）。社会科学中的自然语言处理新基准：评估预训练语言模型在开放式调查回答分类中的实用性。载于第13届国际智能体与人工智能会议，第866–873页。Pedregosa, F.、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Pret- tenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay（2011）。scikit-learn：Python 中的机器学习。《机器学习研究期刊》12，2825–2830。Read, J.、B. Pfahringer、G. Holmes 和 E. Frank（2011）。用于多标签分类的分类器链。《机器学习》85，333–359。Schierholz, M. 和 M. Schonlau（2021）。职业编码的机器学习——一项比较研究。《调查统计与方法学期刊》9（5），1013–1034。Schonlau, M. 和 M. P. Couper（2016）。开放式问题的半自动分类。《调查研究方法》10（2），143–152。Schonlau, M.、N. Guenther 和 I. Sucholutsky（2017）。使用 n 元语法变量进行文本挖掘。《Stata 期刊》17（4），866–881。Schonlau, M.、H. Gweon 和 M. Wenemark（2021）。开放式问题的自动分类：全选题。《社会科学计算机评论》39（4），562–572。Tsoumakas, G. 和 I. Katakis（2007）。多标签分类：综述。《国际数据仓储与挖掘期刊》（IJDWM）3（3），1–13。Vaswani, A.、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin（2017）。注意力就是你所需要的一切。《神经信息处理系统进展》（NIPS）30。Wartena, C.（2019）。用于德语词形还原的概率形态模型。载于第15届自然语言处理会议论文集（KONVENS 2019），第40–49页。Zhang, M.-L. 和 Z.-H. Zhou（2013）。多标签学习算法综述。《IEEE知识与数据工程汇刊》26（8），1819–1837。

附录：含有德语文本答案的表格

本附录中的表格包含原始德语文本答案，并在论文主体中有对应表格。

表6：与表4相对应的原始德语文本答案：当正确代码为2400时，ECC和BERT错误预测的代码。非文本列与此表相同。

答案	ECC	BERT
“他们不去看德国人”	3100	5000
“自私”	-99	5000
创造满意感	5000	5000
“不明确的目标”	-99	5000
“更好地解释政策”	2000	2000
“太多未投票者。”	-99	-98
“持续落实”	-99	2000
“统一性”	-99	5000
“持续不断地违反我们的基本法以及默克尔女士那无可替代的政策”	2441	2441
“沟通”	5000	5000
“交流得太少了”	2000	2000
“每个人都想各自另起炉灶。”	5000	2000
“往往缺乏坦诚，不愿坦然地谈论问题并应对这些问题，以及有时对政治正确”的错误理解	3750	2420
“无法设定优先¨级”	-99	2000

表 7：表 5 中的德语原始文本答案：当正确代码为单标签时，ECC 给出的多标签预测不佳示例。非文本列与本表相同。

	True	ECC 1	ECC 2	ECC 3
“公平的税费分配”	4320	2400	-99	
“小动作的声音如此之大，以至于因此看起来很大”	2400	3412	3750	
“穆斯林”	3750	2411	-99	
“绝对不是那些寻求庇护者，这只会被夸大的	-99	3720	3750	
“养老金或老年贫困”	3740	3721	3740	4100