基于媒体新闻的缺陷汽车的早期舆情分析与风险指标体系构建

张林逍

中国科学技术大学 管理学院

2025 年 12 月 7 日

摘要

随着汽车产业规模的扩大和车辆电子化程度的提升，汽车质量问题呈现出更高的复杂性和隐蔽性。大量潜在缺陷信息最早往往出现在媒体新闻和舆情报道中，而对海量新闻进行人工筛查和标注不仅成本高昂，也难以及时抽取可能引发召回的风险线索。如何从多源新闻文本中自动识别潜在缺陷并评估其召回风险并建立可能的指标集，已成为汽车安全监管、车企质量管理以及风险预警系统的重要研究方向。

本研究结合国家市场监管总局的召回分级标准及相关文献，构建了适用于新闻场景的 0–3 四级汽车召回风险体系，并提出了一种融合召回烟雾词挖掘（Smoke Words Mining）–弱监督伪标签生成–自训练迭代–BERT 深度语义模型的召回风险预测方法。首先基于之前文章提出的烟雾词理论，设计了多级风险加权的词频统计指标，自动挖掘新闻文本中与不同召回等级显著相关的 “召回烟雾词”。随后基于烟雾词构建弱监督伪标签，并通过多轮自训练不断提升稀少类别（尤其是中高风险 2、3 档）的识别能力。最终构建了以 BERT 为核心的四分类召回风险预测模型。

在测试集上，模型在多数类（0 档，1 档与 3 档）取得了很高的 F1 分数。尽管中间风险类别（2 档）仍呈现数据稀缺导致的性能不足，但整体 Macro-F1 也达到了很高的分数，表明模型在不平衡数据环境下具备较强的召回风险识别能力。混淆矩阵分析亦显示模型能够有效区分无风险与最高风险文本，为后续提升中风险类的识别能力提供了明确方向。

关键词： 汽车召回；风险预测；烟雾词挖掘；多级风险加权；弱监督；BERT 模型

1 引言

1.1 研究背景

缺陷汽车召回是指汽车制造商在发现其生产的车辆由于设计、制造、标识等原因导致在同一批次、型号或者类别的汽车产品中普遍存在的不符合保障人身、财产安全的国

家标准、行业标准的情形或者其他危及人身、财产安全的不合理的危险时，主动或应监管机构的要求，将存在缺陷的车辆从消费者处召回进行维修或更换的产品机制。这一机制旨在保障消费者权益和道路交通安全。

缺陷汽车召回同时受到世界各国政府的重视。美国于 1966 年出台《国家交通与机动车安全法》，宣布正式实行汽车召回制度。于 1970 年成立美国国家公路交通安全管理局（NHTSA），隶属于美国交通部，下设政策运营、交通伤害控制、车辆安全三大部门，负责交通安全研究（人、车、环境），制定机动车安全标准、燃油经济性标准，开展机动车缺陷调查，法规符合性监管和调查里程表欺诈等。其中车辆安全部门下设执法局，专门负责缺陷调查、机动车安全符合性监管等汽车召回工作。欧盟也于 2005 年 1 月 15 日启动 RAPEX 系统，该系统是一个在欧洲范围内的警报系统，依据欧洲一般产品安全指令 2001/95/EC 创建，允许在欧盟内自由交换各种产品的合规状态和违规信息（不包括医疗器械、食品和药品）。该系统目前更名为 “Safety Gate System”，是欧洲共同体针对危险产品的信息交换系统，欧盟委员会和成员国利用该系统交换有关危险产品及相关措施的信息。其中缺陷汽车的危险信息是重要的子模块之一。

2004 年，我国由发改委、商务部和海关总署首次颁布实施《缺陷汽车产品召回管理规定》。由国家质量监督检验检疫总局负责组织和管理工作。2012 年 10 月 30 日，《缺陷汽车产品召回管理条例》（以下简称《条例》）正式发布，并于 2013 年 1 月 1 日取代《规定》正式实施。经过 2018 年的机构重组，目前我国由市场监督管理总局总管市场综合监督管理工作，机构下设缺陷产品召回技术中心，负责缺陷产品的预警与召回工作。其中汽车召回是重要子模块。根据新华网报道，截至 2023 年底，我国已累计实施汽车召回 2842 次、涉及车辆达 1.03 亿辆。而仅 2023 年，我国共实施汽车召回 214 次、涉及车辆 672.8 万辆，分别比上年增长 4.9% 和 49.9%。近年来，有非常多的趋势使得缺陷汽车的早期问题发现和召回这个跨学科研究问题越来越重要。

1.2 研究问题与目标

新闻是 “缺陷早期预警” 的关键来源。很多召回事件在官方公告前，往往先在媒体报道中出现 “隐患迹象”，如 “起火” “自燃” “制动失灵”。然而，在海量的新闻中，含有高风险信息的新闻文本极少，人工标注费时费力，标注成本大且容易出错，传统监督学习也难以奏效。同时也缺少相关的风险预测指标。

本研究问题，聚焦于探讨如何利用自然语言处理，结合特征工程以及深度学习的方法，利用媒体新闻数据库中的大规模语料库进行缺陷汽车的早期舆情分析，试图通过舆情分析来识别含有汽车召回信息和潜在缺陷的文本，以及预测潜在缺陷汽车的早期风险表现，为汽车安全审查召回决策提供决策支持。

本文的研究目标包括：

- 对大量新闻文本数据集进行合理有效的数据清洗与早期处理
- 从新闻中自动挖掘和构建 “召回烟雾词表”
- 建立弱监督 + 自训练流程，构建基于 BERT 的召回风险多分类模型

- 结合相关标准构建评估体系，分析模型的预测表现

1.3 研究意义

一方面，现有工作多聚焦于已确认的缺陷案例或社交媒体投诉，缺乏面向新闻场景、能够输出细粒度多级风险（0–3 档）的系统方法，本研究通过引入“召回烟雾词 + 多级风险加权统计 +BERT 深度模型 + 弱监督自训练”的整体框架，丰富了车辆缺陷挖掘与文本风险评估的研究体系，为有序多分类、数据不平衡条件下的风险建模提供了可借鉴的技术路线；另一方面，在应用层面，所提出的模型可以从海量新闻中自动筛选出潜在高风险事件，显著降低监管部门和车企在缺陷线索研判上的人工成本与时间延迟，有助于更早发现严重质量隐患并触发调查或召回决策，从而提升汽车产品全生命周期的安全监管能力和企业质量管理水平，对构建智能化的汽车召回预警与舆情监测系统具有现实价值。

2 文献综述

现有研究中诸多研究者主要关注从自然语言处理与文本分析的技术从社交媒体、在线论坛、用户评论等用户生成内容中识别产品缺陷。这部分的语料场景主要基于消费者的投诉数据。

在传统的文本分析与特征构建工作上，Abrahams 等（2012）和 Abrahams 等（2013）提出了 VDDS（Vehicle Defect Discovery System）系统和烟雾词（vehicle smoke words）概念，通过专家标注和统计检验挖掘缺陷烟雾词、自动识别缺陷帖和讨论组件；在此基础上，Zhang 等（2015）进一步将论坛帖子按车型与时间窗口聚合，用词频或烟雾词频特征训练分类器预测未来召回事件，把社交媒体信号正式引入召回预测链条；

在较为新颖的神经网络与深度学习的工作上，Xindong You 等的 *VDRF: Sensing the Defect Information to Risk Level of Vehicle Recall based on BERT Communication Model*（2020），在缺陷案例数据库上结合静态风险因子、故障标签与文本语义，引入 BERT “communication model” 实现高/中/低/无四级召回风险多分类，并在不平衡样本条件下取得较高 Macro-F1，首次将预训练语言模型与监管风险分级制度直接对接；Yang 与 Deng（2021）提出 MetaQNL/MetaInduce 框架从数据中学习可读符号规则，这类“文本抽取 + 规则推理 + 规则学习”的方法为今后在汽车召回场景下构建结合烟雾词、事件特征和监管经验的可解释召回风险规则系统提供了重要的方法储备。Titus Hei Yeung Fong（2021）等人基于客户评论，结合 RNN 递归神经网络和 LDA 狄利克雷主题模型，从 OCR 中提取产品缺陷信息，用于为保险公司和投保人提供产品缺陷的早期评审，定位客户在 OCR 中提到的产品缺陷和问题的关键词，为保险公司和投保人提供决策支持，以主动阻止缺陷产品的扩散，从而降低保险成本。Jingjin Tian 等人（2022）利用 Bert 语料预训练模型，根据历史缺陷案例的分类特征集训练了一个分类器来分类汽车缺陷信息。Costa Silva 等的 *Predictive quality model for customer defects*（2024），基于汽车企

业内部的质保记录和客户投诉等结构化数据，通过特征工程与决策树、XGBoost 等机器学习算法构建预测性质量模型和责任归属模型，展示了在 Quality 4.0 背景下利用 AI 提升投诉处理效率和质量管理水平的可行性，其数据来源多局限于企业内部系统。

3 数据分析与模型构建

3.1 数据预处理与特征工程

本文选取 2022 年某财经与汽车资讯平台抓取的汽车相关新闻作为研究对象，原始数据以 Excel 形式存储，有 67233 条新闻文本数据，其字段包括：

- 媒体名称：报道新闻的媒体名称；
- 内容：新闻正文文本，是后续 NLP 建模的核心字段；
- 标题、日期、字数、版面、作者、链接、城市等辅助字段；
- risk：部分样本包含人工标注的召回风险等级。

其中，人工标注的风险等级是后期处理的，之后会提及，原数据集中并无这一字段。

在数据预处理阶段，我依次执行了以下步骤：

媒体类型与权威度映射：在 ‘NLP-car.py‘ 与 ‘context-cleaning.py‘ 中，我们首先基于专家知识构建了 “媒体类型映射表” 和 “媒体权威度分级字典”。其中，媒体类型映射列表将所有媒体划分为官方机构、垂直汽车媒体、综合新闻、财经媒体、专业资讯、社交平台、内容聚合、券商平台等类别，并建立了一一对应关系（见表 1）。

之后在此基础上建立了媒体权威度分级词典，该词典将媒体权威度分为 1-7 级，代表最高权威度到最低权威度。划分权威度的原则是参考相关规定，对与汽车缺陷和召回高度相关的官方机构与专业媒体赋予更高权威度，例如：国家市场监管总局、中国质量报、专业召回网站、中国汽车召回网等。对地方小型网站赋予更低权威度。具体划分标准见表 2.

品牌识别与多品牌展开：

在 `NLP_car.py` 中建立了一个覆盖主流与新势力品牌的 CAR-BRANDS 列表，利用字符串匹配在 “内容” 字段中识别出现的汽车品牌，生成：

- 汽车品牌：新闻中提到的所有汽车品牌列表；
- 主要品牌：取首个汽车品牌作为主品牌（在多品牌新闻中后续再做细分）。

在 `copy_brand.py` 中，我们进一步将 “汽车品牌” 列从列表展开，即：

- 将每条新闻的每个品牌拆为独立样本，得到新闻-品牌对；
- 为避免多品牌新闻权重过高，按每条新闻的品牌数计算 $sampleweight = 1/numberofcartypes$，用于后续训练加权。

文本清洗与分词编码：

在观察分析原始数据集之后，发现出现了很多格式错误，以及大量文本重复的问题，即多家媒体报道完全相同的新闻或者只是稍加修改，内容基本一致。这大大减少了数据

媒体名称	媒体类型
国家市场监督管理总局	官方机构
国务院发展研究中心信息网	官方机构
东方财富 APP	财经媒体
同花顺财经	财经媒体
腾讯新闻 APP	综合新闻
网易新闻 APP	综合新闻
懂车帝 APP	垂直汽车
汽车之家	垂直汽车
上海证券报 APP	传统媒体转型
证券时报 APP	传统媒体转型
中信建投证券 APP	券商平台
国泰君安证券 APP	券商平台
新浪微博	社交平台
知乎	社交平台
虎嗅网 APP	专业资讯
36 氪	专业资讯
周到上海 APP	地方媒体
齐鲁壹点 APP	地方媒体
我的钢铁网	行业垂直
长江有色金属网	行业垂直
微信公众号	内容聚合
看点快报 APP	内容聚合

表 1: 媒体名称与媒体类型映射表（部分）

权威度等级	典型媒体角色/类型
最高权威媒体	汽车召回官方发布机构；国家级主流媒体；专业质量监管机构
高权威媒体	专业财经媒体；证券权威媒体；专业汽车媒体
中等权威媒体	主流综合新闻平台；专业资讯平台；垂直汽车媒体
基础权威媒体	财经数据服务；地方主流媒体；券商研究机构
一般权威媒体	财经媒体；行业垂直类媒体；专业资讯网站
较低权威媒体	内容聚合平台；专业社交平台；汽车资讯网站
最低权威媒体	社交娱乐平台；论坛社区；自媒体与小型网站；其他杂项网站

表 2: 媒体权威度等级与典型媒体类型对应表

集的信息密度，也会增加后面模型训练的时间并。并且，大量重复文本会大幅增加重要新闻的权重，进而增大训练后模型的误差。于是在 NLP_car.py 中，我们进行了统一的文本预处理：

- 去除非中文字符、URL、特殊符号等噪声；
- 使用“jieba”包进行中文分词，基于中文停词表（chinese-stopwords.txt）去停用词，并过滤单字或无意义符号；
- 将分词结果存入 tokenized_text（以空格分隔）和 tokens（列表形式）；
- 基于所有样本分词结果统计词频，构建词汇表 vocab（过滤低频词），并将新闻编码为整数序列，保存为“编码序列”和“processed-tensors.pt”字段。

上述分词过程在后续所有统计与建模脚本中通过 smokewords_stats.tokenize 统一复用，保证口径一致。对于文本重复的问题，在 context_cleaning.py 中，我们基于 SimHash 的 Hamming 距离计算新闻内容近似度：

SimHash 把每条新闻编码成一个固定长度的二进制指纹（如 64 位），Hamming 距离（汉明距离）就是衡量两段等长二进制串“有多少位不同”的指标：

把两条新闻的 SimHash 值看成两个 0/1 串；对应位逐位比较，相同记 0，不同记 1；Hamming 距离 = 不同位的个数；距离越小，说明两条文本越相似（在当前特征与 SimHash 构造下）。

设文本 d 的 SimHash 指纹为 $h(d) \in \{0,1\}^n$。对于两条文本 $d^{(1)}, d^{(2)}$，其汉明距离定义为

$$d_H\big(h(d^{(1)}), h(d^{(2)})\big) = \sum_{i=1}^{n} \delta\big(h_i(d^{(1)}), h_i(d^{(2)})\big),$$

其中

$$\delta(a,b) = \begin{cases} 1, & \text{若} a \neq b, \\ 0, & \text{若} a = b. \end{cases}$$

当 d_H 小于给定阈值（如 3 位）时，两条新闻被视为相似或重复。

对不同媒体发表的高度相似新闻用 Hamming 距离聚类后，对每个相似组按照：

1. 媒体权威度（越高越优）；
2. 发布时间（越早越优）；
3. 字数（越多越优）；
4. 相同权威度下媒体类型优先级（官方机构 > 垂直汽车 > 综合新闻 > 社交平台…）

筛选保留一条“代表性新闻”，其余视为重复样本删除，从而降低噪声和冗余。

3.2 模型构建与优化

3.2.1 多档风险下的文档频数统计

本文引入了多级风险加权的烟雾词挖掘框架（在 risk_smokewords_v0.py 中）。首先仅使用人工已标注风险（人工标注的标准等见后一节）的样本，对每个 term 计算在

不同风险档位中出现的文档频数：

$$df_r(t) = \text{包含词}t \text{ 且人工风险为}r \text{ 的文档数}, \ r \in \{0,1,2,3\}. \tag{1}$$

在此基础上定义“非缺陷文档频数”与“缺陷文档频数”：

$$df_{\text{nondefect}}(t) = df_0(t), \tag{2}$$

$$df_{\text{defect}}(t) = 1 \cdot df_1(t) + 2 \cdot df_2(t) + 3 \cdot df_3(t). \tag{3}$$

3.2.2 相对权重 rtw 与平均风险值 $risk\_avg$

其中系数 $1,2,3$ 为不同风险档位的权重，体现高风险样本对烟雾词的重要性贡献更大。

Abrahams 等人提出的 relative term weight（rtw）主要用于二分类场景，通过比较词项在缺陷类与非缺陷类文本中的文档频数来度量其“缺陷相关性”，其基本形式可写为：

$$rtw_{\text{Abrahams}}(t) = \frac{df_{\text{defect}}(t) + \alpha}{df_{\text{nondefect}}(t) + \alpha}, \tag{4}$$

其中 $df_{\text{defect}}(t)$ 与 $df_{\text{nondefect}}(t)$ 分别表示词项 t 在缺陷/非缺陷文本中出现的文档数，其中 $\alpha = 1.0$ 为平滑系数。$rtw(t)$ 越大，说明该词越倾向于出现在高风险新闻中。

本文面向召回风险分级任务，在 Abrahams 的二分类基础上做出改进，在将“缺陷类”进一步细分为多档风险等级 $r \in \{0,1,2,3\}$（或扩展到 0–4），并对不同风险档位赋予不同权重 w_r，使高风险样本对烟雾词挖掘贡献更大。定义：

该设计使得在高风险新闻中具有代表性的词项（如起火、制动失灵等）更容易获得较高权重，从而避免低风险投诉词对词表的稀释。

借鉴 Abrahams 等人提出的 relative term weight 思路，同时为了刻画“该词出现时平均对应的风险等级”，定义平均风险值：

$$df_{\text{defect}}(t) = \sum_{r=1}^{R} w_r \cdot df_r(t), \tag{5}$$

优化一下即

$$risk\_avg(t) = \frac{0 \cdot df_0(t) + 1 \cdot df_1(t) + 2 \cdot df_2(t) + 3 \cdot df_3(t)}{df_0(t) + df_1(t) + df_2(t) + df_3(t) + \varepsilon}, \tag{6}$$

其中 $\varepsilon = 10^{-9}$ 用于避免除零。$risk\_avg(t)$ 介于 0–3 之间，越大说明该词更集中出现在高风险样本中。

3.2.3 初版烟雾词表 v0 的构建

在得到词级统计特征后，通过以下门槛筛选得到初版烟雾词表 v0：

- 缺陷文档频数：$df_{\text{defect}}(t) \geq \texttt{MIN\_DF\_DEFECT} = 5$；

- 相对权重：$rtw(t) \geq 1.5$；
- 平均风险：$risk\_avg(t) \geq 1.0$，并记 $risk\_avg(t) \geq 2.5$ 的词为高风险烟雾词。

筛选后得到的 v0 按 rtw 与 $risk\_avg$ 降序排序，包含了一批在高风险新闻中显著高频的“召回烟雾词”，例如起火、自燃、刹车失灵等，从而为后续弱监督标注提供基础词典。具体的烟雾词表会在下一节中给出。

3.2.4 基于烟雾词的连续风险与伪标签生成

在 `risk_smokewords_v0.py` 与 `risk_smokewords_v1_pl1prime.py` 中，我们利用词表 v0/v1 的 `risk_avg` 指标，为每条新闻定义**连续风险得分**。

设文档 d 的分词集合为 $\{t_i\}$，其词频为 $tf(t_i, d)$，则连续风险计算公式为：

$$raw\_risk(d) = \frac{\sum_{t \in d} tf(t,d) \cdot risk\_avg(t)}{\sum_{t \in d} tf(t,d)}$$

对该文档里的每个词 t，用该词的 $risk\_avg$（由人工标注样本统计出来）做权重。并按词频 $tf(t,d)$ 做加权平均，将 raw_risk 离散映射至 0–3 档：

- $raw\_risk < 0.5 \to 0$
- $0.5 \leq raw\_risk < 1.5 \to 1$
- $1.5 \leq raw\_risk < 2.5 \to 2$
- $raw\_risk \geq 2.5 \to 3$

为保证伪标签可靠性，引入高置信筛选规则：

- $raw\_risk < 0.3 \to$ 高置信 0 档；
- $raw\_risk > 2.7 \to$ 高置信 3 档；
- 中间区域可作为普通伪标签或暂不使用。

高置信样本构成 PL_1 或 PL_1'，用于后续模型训练。

3.3 模型训练

以下是完整的弱监督 + 自训练的 BERT 多分类模型的训练流程：

3.3.1 训练数据构建：人工标签 + PL1

在 risk_bert_riskmodel.py 中，首先读取原始数据 data_combined_4types.xlsx 与 PL1 伪标签 pseudo_labels_pl1.csv。根据如下策略构建训练样本：

1. 对于存在人工标注 risk 的样本，直接使用人工标签；
2. 对于未标注样本，若其在 PL1 中存在伪标签，则使用对应的 pseudo_label_pl1；
3. 过滤非法标签，保留区间 $0-3$ 之间的整数标签。

最终构造 TextLabelExample 列表，每个样本由文本与四分类标签组成。

3.3.2 基于 BERT 的四分类风险模型 M1

采用 bert-base-chinese 作为预训练语言模型，在其之上接入线性分类头构建四分类模型 $M1$。输入为新闻正文文本，输出为四维风险概率分布。训练细节如下：

- 最大序列长度：256；
- 损失函数：交叉熵；
- 优化器参数：学习率 2×10^{-5}，权重衰减 0.01；
- Batch 大小：8（训练与验证）；
- 训练轮数：3；
- 验证集划分：按标签分层抽样，10% 作为验证集；
- 使用 eval_strategy="steps"，每 200 step 在验证集上评估一次并保存最优模型；

训练完成后，模型与分词器保存至目录 ./bert-risk-model-M1。

3.3.3 全量推断与第二轮伪标签 PL2

使用训练好的 $M1$ 对全量新闻进行推断，得到每条新闻的预测标签 pred_label 和最大 softmax 概率 confidence。将预测结果写入 risk_predictions_M1.xlsx。

为进一步扩充伪标签，定义第二轮伪标签集合 PL2：

- 从全量预测结果中筛选 confidence ≥ 0.85 的样本；
- 将其预测标签视为可靠伪标签 pseudo_label；
- 输出字段 doc_id, pseudo_label, confidence 至 pseudo_labels_pl2.csv。

3.3.4 基于 PL2 更新烟雾词表 v1

在 risk_smokewords_v1.py 中，将人工标签与 PL2 伪标签合并，形成最终标签列 final_risk：若某样本存在人工 risk，则优先使用，否则采用 PL2 的 pseudo_label。

在此新标注基础上，重新执行如下步骤：

1. 对带 final_risk 的样本重新统计 $df_r(t)$、$df_{\text{defect}}(t)$、$df_{\text{nondefect}}(t)$；
2. 按前文公式重新计算 $rtw(t)$ 与 $risk\_avg(t)$；
3. 使用与 v0 相同的阈值筛选，得到更新后的烟雾词表 v1；
4. 将词统计与新词表分别保存为 riskword_term_stats_v1.csv 与 smokewords_v1.csv。

通过这一轮自训练，烟雾词表从依赖少量人工标签的 v0 升级为融合大规模 $M1$ 预测信息的 v1，使得词表更加贴合整体数据分布，尤其能够提升中高风险领域在少样本条件下的表示质量。

4 实验结果与分析

4.1 实验设置

在确定召回风险等级划分标准时，我们一开始采用了 0-4 标签，将风险划分成 5 个标准，但中途出现了 2、3 等级不好划分以及各等级标签数量严重不均衡的问题。之后的训练效果也不理想（结果见后面小节）。

于是之后参考 VDRF 论文在汽车缺陷案例数据集里用的是 4 个召回风险等级：High / Medium / Low / None，并用 BERT 做多分类预测。且国家层面的《汽车产品召回预警规则》及其引用的 GB/T 35253、GB/T 34402 也基本采用 4 级预警 / 风险分级（高、较高、中、较低/低，对应红、橙、黄、蓝）的思路。因此，为了和监管逻辑、论文实践统一，新闻文本的召回风险改为 4 档（0-3）。人工标注风险的判断标准见表 4.

综合以上模块，整体数据处理与训练流程如下：

1. 使用 NLP_car.py 对原始新闻进行品牌抽取、媒体类型归一、文本清洗、分词和基础缺陷类别识别，得到 cleaned_normalized_data_full.xlsx；
2. 使用 context_cleaning.py 执行按月份的 SimHash 去重，得到 deduplicated_news_full.xlsx，同时为每条新闻附带媒体权威度特征；
3. 将去重后的数据与人工风险标签合并，形成 data_combined_4types.xlsx，在为人工标注的数据中随机抽出 100 条数据作为测试集，并进行人工风险标注；
4. 运行 risk_smokewords_v0.py，在人工已标注样本上统计词频并获得烟雾词表 v0（见表 4）和第一轮伪标签 PL1；

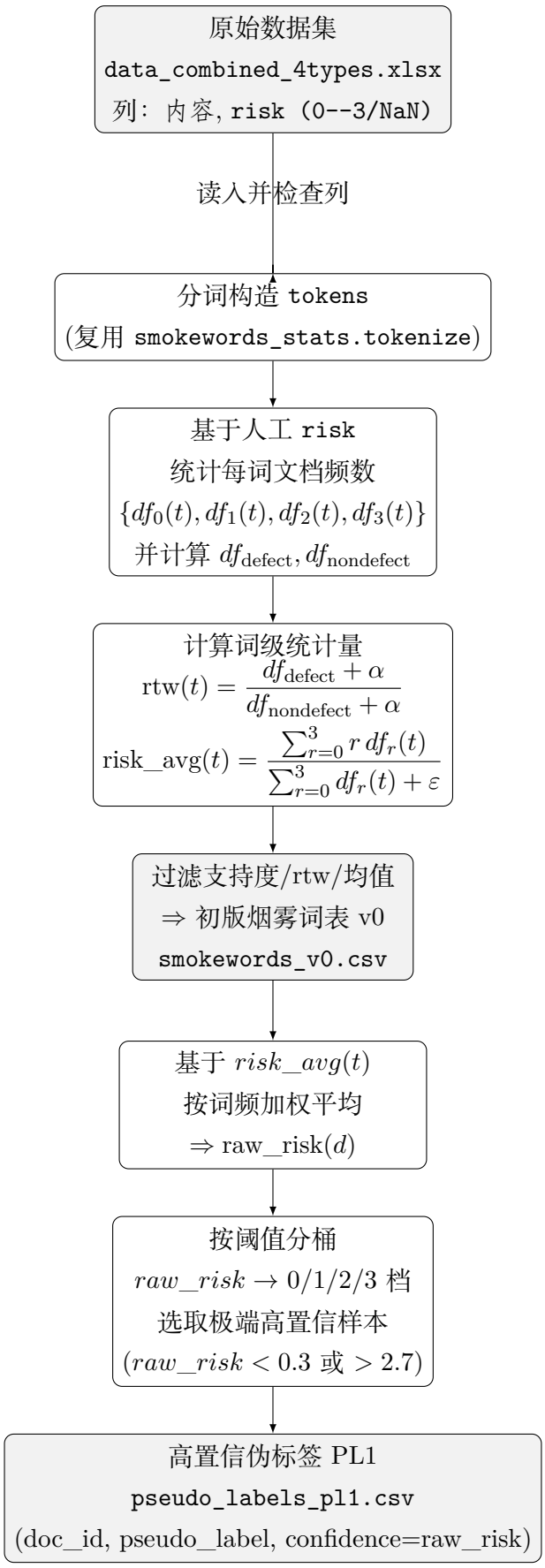


图 1: 阶段 1：从原始数据构建 v0 烟雾词表并生成 PL1 伪标签

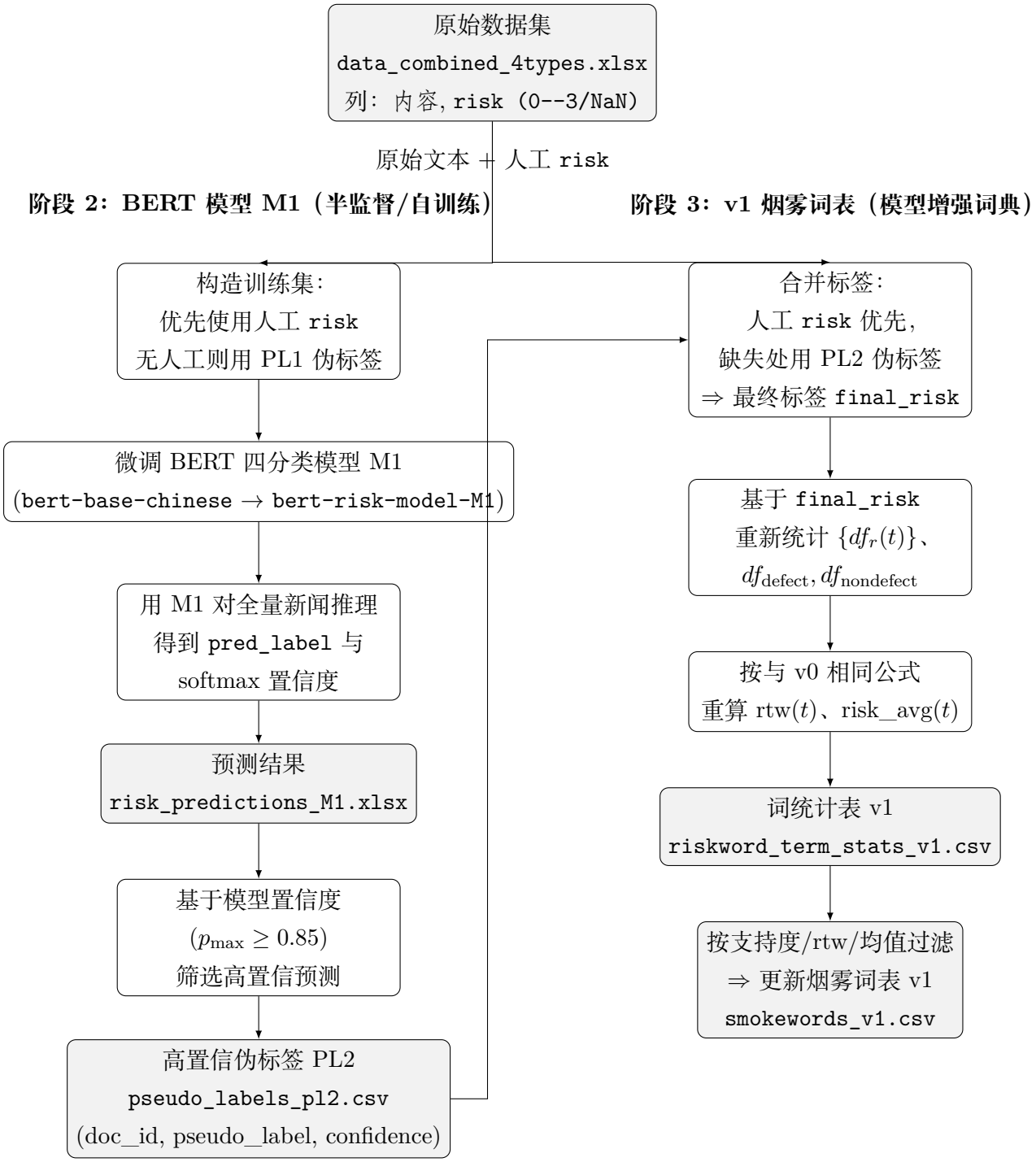


图 2: 阶段 2（左）：训练 BERT 模型 M1 并生成 PL2；阶段 3（右）：融合标签更新 v1 烟雾词表

风险等级	判定标准
高风险（High）	报道直接涉及安全相关缺陷并已造成严重后果或具有高度风险；官方机构已介入调查或发布预警；车企已承认缺陷或启动召回；由高权威媒体发布或多家媒体一致报道；若情况持续发展，高度怀疑一定会触发正式召回。
中风险（Medium）	涉及疑似安全隐患或严重性能缺陷但证据不足；影响范围不明确，仅为少量投诉或爆料；尚无召回决定但监管部门已约谈、调查或排查；国外已有召回而国内尚未实施；由中等权威媒体或多家媒体引用报道；具有较高可能触发召回但仍停留在调查或曝光阶段。
低风险（Low）	主要为一般性能问题或非安全类缺陷（如异响、轻微渗油等），短期不太可能触发召回；为个别车主案例且缺乏规模投诉；媒体权威度较低、缺乏官方回应；多涉及主观抱怨、口碑或售后纠纷；或仅为已结束召回的历史回顾，无新增风险。
无风险（None）	报道内容与缺陷、安全无关（如上市、营销、销量、人事变动等）；若有负面仅涉及价格、服务、金融政策等非质量因素；或为历史召回总结且无新增隐患；缺乏任何指向“车辆存在缺陷或安全隐患”的描述，即便内容属实也不影响召回判断。

表 3: 新闻文本的四级召回风险划分标准

5. 运行 `risk_bert_riskmodel.py`，在“人工标签 + PL1”上微调 BERT 得到模型 M1，使用 M1 对全量数据预测并生成高置信伪标签 PL2；

6. 运行 `risk_smokewords_v1.py`，基于人工标签 + PL2 重估词统计并构建烟雾词表 v1（训练流程见图 1，图 2）。

在模型效果评估部分，选取约 100 条人工标注且未参与训练的样本作为测试集，对最终 BERT 模型进行评价，并计算准确率、各类 Precision / Recall / F1、Macro-F1 等指标。

4.2 多方法效果对比

本实验中衡量模型预测效果的指标分别为：精确率 (accuracy)、Macro-F1、Per-class F1、混淆矩阵。

精确率

$$Precision_c = \frac{TP_c}{TP_c + FP_c}$$

Per-class F1 值：可以单独衡量模型在某个类别上的性能

$$F1_c = \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c}$$

宏平均 F1（$Macro\text{-}F1$）：定义为每个类别的 F1 值的平均，特别适用于衡量类别不平衡任务的效果

$$Macro\text{-}F1 = \frac{1}{C}\sum_{c=1}^{C} F1_c$$

其中 C 为类别总数。

混淆矩阵是一张真实标签 vs 模型预测标签的矩阵，用于展示模型在哪些类别上预测正确、在哪些类别间容易混淆。能直观显示模型的错误类型。

本文首先对比了三种模型 / 策略的效果：

1. **规则基线 (Smoke-Score)**：仅使用烟雾词表 v0，按照 $raw\_risk(d)$ 的连续得分与预设离散化规则，直接输出 0–3（或映射回 0–4）的风险类别预测，不使用 BERT。

2. **监督 BERT（BERT-supervised)**：仅使用人工标注样本训练 BERT 多分类模型。

3. **弱监督 + 自训练 BERT（BERT-semi)**：即本文提出的完整流程，使用“人工标签 + PL1”训练模型 M_1，之后借助 PL2 与词表 v1 进行迭代式更新。

在 100 条测试集上的一组实验结果表明：

BERT-semi 的整体准确率约为 0.80，Macro-F1 约为 0.71；在样本较多的 0 类、1 类与 3 类上，F1 分别约为 0.80 、0.73 与 0.85，说明该模型已能较好地区分“无风险”，“低风险”与“高风险”（见表 5）；对于样本极少的中间类别（2 类），由于训练样本与伪标签均极度不足，其 F1 值较低 0.46，混淆矩阵中大部分样本被模型吸收到 1 或 3 类。这体现出中间风险类别的识别能力依然是当前系统的主要短板。

与之相比：**Smoke-Score** 在高风险类别的召回率上表现尚可，但整体精度与 Macro-F1 显著低于 BERT 模型；**BERT-supervised** 由于仅依赖有限的人工标注样本，在少数类上的表现与 BERT-semi 相差不大，但在多数类上略逊；采用弱监督 + 自训练策略的 **BERT-semi** 通过引入 PL1 与 PL2 显著扩大了有效训练样本规模，使得模型在 0 类、1 类与 3 类上都获得了更高的稳定性和概率校准能力。

之后对比按 0-4 分类训练的模型 **BERT-5types**、**BERT-semi** 和用烟雾词 v1 再次训练得到的模型 **BERT-semi-M2** 在测试集上的效果，并比较了它们的 **Per-class F1** 和混淆矩阵。（见图 3）结果显示 **BERT-5types** 对 2 类和 3 类风险几乎没有预测能力，在最低和最高风险上的预测能力也不如 **BERT-semi**，而同样使用 4 分类的 **BERT-semi-M2** 在大部分指标上的表现都不及 **BERT-semi**。

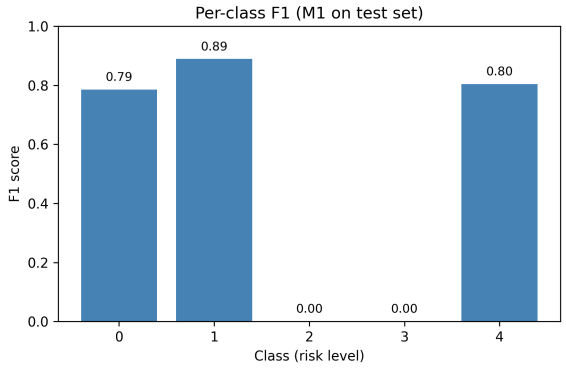
对于 **BERT-semi**，其混淆矩阵在对角线上的集中程度增强，尤其是“高风险 (3) → 低风险 (0/1)”的严重误判比例较低，这表明其与 **BERT-supervised**、**BERT-5types** 和 **BERT-semi-M2** 相比有助于缓解“高风险样本过少”的数据不均衡问题。

表 4: Term 与统计得分（score_stat）示例

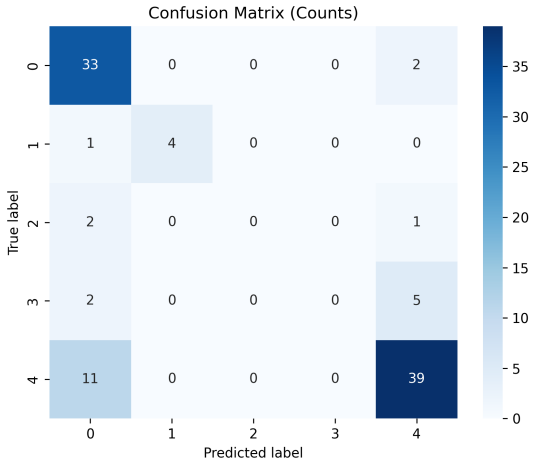
term	score_stat
生产日期	8.12
自即日起	7.63
松动	7.29
日至	7.27
挂号信	7.24
涉嫌	7.21
财政部	7.19
拨打	7.18
东风汽车	7.11
发布公告	7.10
安全隐患	7.04
造型	7.02
退坡	7.02
实施办法	7.00
下车	7.00
锁扣	6.95
风神	6.94
爆料	6.94
梅赛德斯	6.94
下线	6.92

Class	Precision	Recall	F1-score	Support
0	0.7500	0.8684	0.8049	38
1	0.6667	0.8000	0.7273	5
2	0.7500	0.3333	0.4615	9
3	0.8696	0.8333	0.8511	48
Accuracy			0.8000	100
Macro avg	0.7591	0.7088	0.7112	100
Weighted avg	0.8032	0.8000	0.7923	100

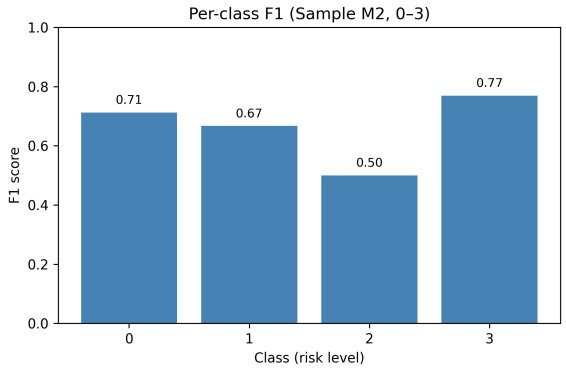
表 5: **BERT-semi** 的分类各指标结果 (0–3)



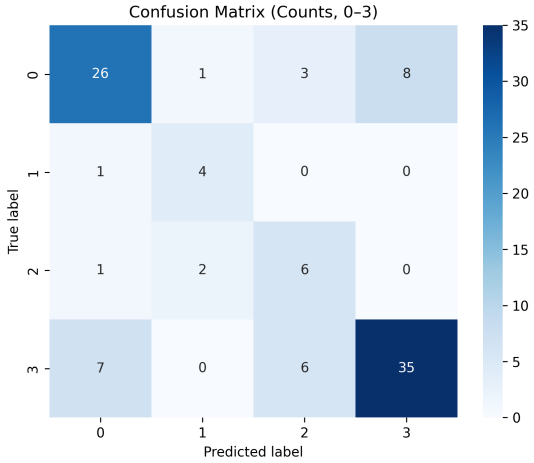
(a) **BERT-5types** 的 F1



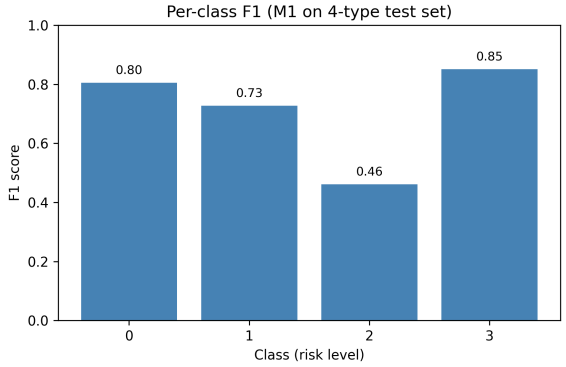
(b) **BERT-5types** 的混淆矩阵



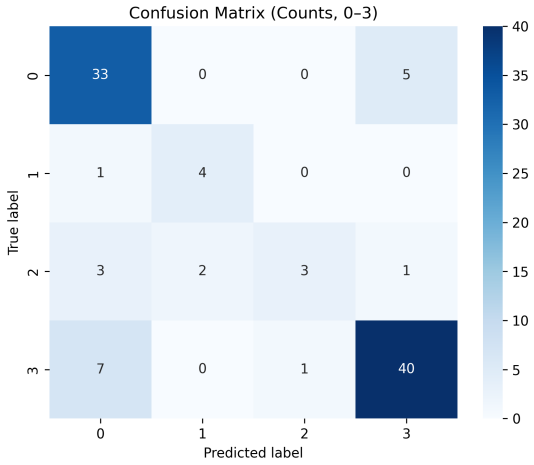
(c) **BERT-semi-M2** 的 F1



(d) **BERT-semi-M2** 的混淆矩阵



(e) **BERT-semi** 的 F1



(f) **BERT-semi** 的混淆矩阵

图 3: 三种模型的预测效果对比

4.3 收敛性分析

本实验基于 `bert-base-chinese` 初始化四分类风险识别模型 M1，在由 299 条人工标注样本 + 147 条高置信 PL1 伪标签样本组成的训练集（共 446 条样本）上进行微调，其中 10% 划为验证集。训练配置为：*batch size* = 8，训练轮数固定为 3 个 epoch，对应约 153 个 global step；每经过 30 个 step 在验证集上评估一次性能（记录 `eval_loss` 与 Macro-F1），并保存一次 checkpoint（step = 30, 60, 90, 120, 150 等），最终根据验证损失最小的 checkpoint 作为最终模型参数（`load_best_model_at_end=True`，以 `eval_loss` 为优先指标）。

如图 X 所示，训练损失（Train loss）在整个训练过程中呈现单调下降趋势：在 global step = 30 时约为 0.914，之后依次降低到 step = 60: 0.565，step = 90: 0.376，step = 120: 0.298，step = 150: 0.252。这一变化表明，在当前学习率和 batch 设置下，优化过程稳定，未出现震荡或发散现象，模型较快完成对训练集的拟合。

验证损失（Val loss）与 Macro-F1 的变化也体现出良好的收敛性。具体而言，在验证集上 step = 30 时 `eval_loss` ≈ 0.639，Macro-F1 约为 0.582；到 step = 60 时 `eval_loss` 降至 0.549（为全程最低值），Macro-F1 略有回落至 0.575；随后在 step = 90 时 `eval_loss` ≈ 0.551，Macro-F1 提升至 0.681，在 step = 120 时两者分别为 0.585 和 0.667，而在 step = 150 时 `eval_loss` ≈ 0.571、Macro-F1 进一步上升至 0.737。可以看出，验证损失在前两个 epoch 内快速下降并在 60–90 step 左右达到稳定区间，而验证 Macro-F1 则在后期持续小幅提升，在最终评估点（step = 150）达到最高值，说明模型在有限轮数内已经学到较为稳定且具有泛化能力的判别边界。

综合上述结果，可以认为：在 3 个 epoch、约 153 个 step 的训练设置下，M1 模型的训练损失实现了平滑收敛，验证损失在中后期保持在 0.55–0.59 的相对稳定区间，验证 Macro-F1 则从约 0.58 起步逐步提升至约 0.74，未观察到明显的过拟合拐点或性能骤降。因此，当前的训练轮数与停止策略（固定 3 个 epoch，并在训练结束后选取验证损失最小的 checkpoint）在收敛性与泛化性能之间取得了较为合理的折中。

在我们的实验设置下，第二轮（M + v2）后上述指标已经接近收敛门槛，说明继续增加迭代轮数的边际收益有限。未来在扩充更多人工标注样本并调整伪标签置信阈值后，可进一步绘制不同置信阈值设置下的收敛曲线，从而分析噪声标签比例对模型收敛速度与最终性能的影响。

5 结论与展望

5.1 研究结论

本文面向“媒体新闻文本视角下的汽车召回风险识别”任务，构建了一条从数据预处理、烟雾词挖掘到弱监督 BERT 风险模型的完整技术路线：

1. 在数据预处理阶段，通过媒体类型与权威度建模、品牌与缺陷类别识别、SimHash

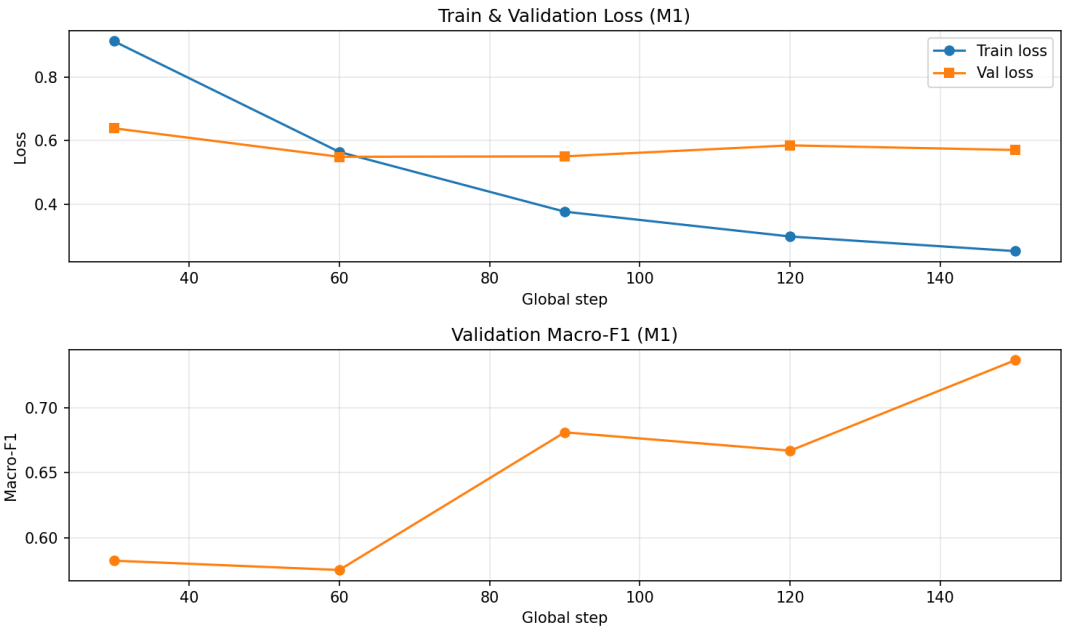


图 4: M1（**BERT-semi**）模型训练 loss/Macro-F1

去重等步骤，构建了结构化、去冗余的汽车新闻语料库；

2. 在二分类烟雾词框架基础上提出了基于多档风险加权的烟雾词挖掘框架，优化了 $df_r(t)$、$df_{\text{defect}}(t)$、$df_{\text{nondefect}}(t)$、$rtw(t)$ 与 $risk\_avg(t)$ 等指标，并基于此生成初版烟雾词表 v_0 与第一轮伪标签 PL1；

3. 在“人工标签 + PL1”上微调 BERT 多分类模型，利用模型预测结果构造高置信伪标签 PL2，再次更新词统计得到更贴合整体数据分布的烟雾词表 v_1，实现了一轮完整的“弱监督 + 自训练”闭环训练方法。

实验结果表明，在样本量有限且类别极度不平衡的条件下，所提出的方法能够在较好保持高风险识别能力的同时，大幅减少人工标注成本，实现对 0–3 召回风险的自动化预测。

之后在对比了分类方式（4 类或 5 类），训练轮数（M1 或 M2）和训练流程（是否引入 PL1 标签；是否加入原始 BERT 模型）后，用多种指标证明了 M1（**BERT-semi**）模型是在各方面表现最好的（见图 4）。

整体而言，本文验证了“烟雾词 + 媒体权威度 + 弱监督 BERT”这一组合在汽车召回风险识别问题上的可行性与潜力，为在真实监管环境中部署自动化召回预警系统提供了方法基础。

5.2 研究不足与未来工作

由于人力原因，人工标注的数据和测试集的数据数量还较少，模型训练轮数和原始数据集的数据量由于算力原因也没有继续增加。后续研究可以进一步增加全部数据量以

训练更高质量的模型

目前模型对中间风险档位（2 类）的识别能力仍然不足，主要原因在于这两类样本数量较少且语义分布多样。后续研究可以对原始数据进行扩充与精细化标注，针对 2、3 档中高风险新闻开展有目的的样本采集与精标，提高少数类的覆盖度；也可以引入有序分类与成本敏感学习，将 0–3 风险视为有序标签，引入 ordinal regression 或 cost-sensitive loss，减轻中间类被极端化吸附的问题。

同时，媒体权威度的指标因为模型复杂的原因未被加入最后的训练流程中，未来可以考虑将媒体权威度与烟雾词特征融合，在 BERT 输入端显式编码媒体权威度和烟雾词命中特征，例如通过额外的 embedding 或前置特殊 token 方式，增强模型对信息可信度的敏感性。

致谢

本研究工作受到指导教师和学长学姐们在数据集和思路上的支持与帮助，在此一并表示感谢。

参考文献

[1] A. S. Abrahams, J. Jiao, G. A. Wang, W. Fan. Vehicle Defect Discovery from Social Media[J]. *Decision Support Systems*, 2012, 54(1): 87–97.

[2] A. S. Abrahams, J. Jiao, W. Fan, G. A. Wang, Z. Zhang. What' s Buzzing in the Blizzard of Buzz? Automotive Component Isolation in Social Media Postings[J]. *Decision Support Systems*, 2013, 55(4): 871–882.

[3] X. Zhang, S. Niu, D. Zhang, G. A. Wang, W. Fan. Predicting Vehicle Recalls with User-Generated Contents: A Text Mining Approach[C]. In: *PAISI 2015, LNCS 9074*. Springer, 2015: 41–50.

[4] A. S. Abrahams, J. Jiao, G. A. Wang, W. Fan. Vehicle defect discovery from social media[J]. *Decision Support Systems*, 2012, 54(1): 87–97.

[5] X. You, J. Ma, Y. Zhang, X. Lv, J. Han. VDRF: Sensing the Defect Information to Risk Level of Vehicle Recall based on Bert Communication Model[J]. *Computer Science and Information Systems*, 2020, 17(3): 795–817. DOI:10.2298/CSIS190903021Y.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. In: *Proceedings of NAACL-HLT*, 2019.

[7] I. H. Kim, M. H. Lee, H. H. Park. Classification Scheme for Root Cause and Failure Modes and Effects Analysis (FMEA) of Passenger Vehicle Recalls[J]. *Reliability Engineering & System Safety*, 2020, 200: 630-638.

[8] J. H. Lim, B. K. Kim. Predicting Vehicle Recalls with User-Generated Contents: A Text Mining Approach[C]. *Business Information Systems*, 2015: 69-80.

[9] G. Liu, X. Wu, J. Zhu, H. Cai. What ' s Buzzing in the Blizzard of Buzz? Automotive Component Isolation in Social Media Postings[J]. *Decision Support Systems*, 2013, 55(4): 871-882.

[10] A. Costa Silva, J. Machado, P. Sampaio. Predictive Quality Model for Customer Defects[J]. *The TQM Journal*, 2024, 36(9): 155-174.

[11] H. Xia, C. Yao, N. Ding, M. Chen, J. Gu, Z. Yang, S. Chen, Z. Zhou. Abductive Symbolic Solver on Abstraction and Reasoning Corpus[J]. *arXiv preprint*, 2024.

[12] Y. Wang, A. Yates, E. Cambria. Learning Symbolic Rules for Reasoning in Quasi-Natural Language[J]. *arXiv preprint*, 2021.

A 代码与数据说明

由于篇幅限制，本文不在正文中逐一给出所有源代码。完整的项目文件已经上传到 github 上：https://github.com/Linxiao077/Construction-of-an-Early-Public-Opinion-Analysis-and-Risk-Indicator-System。本研究的主要 py 代码如下：

- `NLP_car.py`：第一个数据集处理文件，完成了汽车品牌的提取、新闻发布媒体的分类和提取、删除了没数据的 5 行。最后完成了文本清洗、分词与词向量提取；
- `context_cleaning.py`：处理数据的第二个文件，定义媒体权威度映射字典（共 7 个权威度档位）。利用 Hamming 距离找出相似新闻，并根据相关策略保留权威度更高、发布时间更早的新闻，最后进行批量处理；
- `risk_smokewords_v0.py`：基于少量人工 risk 标签统计词–风险共现关系，生成初版烟雾词表 v_0，并据此为全量新闻打出第一轮词典伪标签 PL1；
- `risk_bert_riskmodel.py`：在“人工标签 + 高置信 PL1”上训练四分类 BERT 模型 M1，并用 M1 对全量新闻推理生成第二轮高置信伪标签 PL2；
- `risk_bert_riskmodel_M2.py`：在 v_1 词表生成的高置信伪标签 PL1 基础上训练第二轮 BERT 风险模型 M2，并用 M2 对全量新闻推理生成新的高置信伪标签 PL2；
- `risk_smokewords_v1.py`：将“人工标签 + PL2”合并为最终标签集，再次统计词–风险关系，生成更新后的烟雾词表 v_1。

读者可结合上述目录结构与源代码，复现实验结果并在此基础上进行扩展研究。