深呼吸：用于应对网络虚假信息的机器学习浏览器扩展

Marc Kydd 设计与信息学学院 网络安全系 阿伯泰大学 英国邓迪
m.kydd1800@abertay.ac.uk

Lynsay A. Shepherd 设计与信息学学院 网络安全系 阿伯泰大学 英国邓迪
lynsay.shepherd@abertay.ac.uk

arXiv:2301.03301v1 [cs.HC] 9 Jan 2023

摘要——**过去十年间，媒体格局发生了巨大变化。随着越来越多的公众通过在线信息来源了解时事，各家媒体争相吸引关注，竞争也日益激烈。这种竞争促使一些网络媒体发布耸人听闻、危言耸听的内容来吸引读者注意。这类做法可能会歪曲事实，误导读者对事件性质的认知，从而威胁民主。本文提出一种新颖的系统，用于检测和处理网络误导性内容，并向用户发出警告，以应对错误信息带来的威胁。通过在现有的 32,000 条标题党新闻标题数据集上训练机器学习模型，该模型可以预测标题的煽动程度，随后与浏览器扩展交互，依据现有设计原则生成独特的内容警示通知，并纳入模型的预测结果。本研究为机器学习和以人为本的安全领域作出了创新贡献，其研究发现也为未来研究带来了希望。在用户可能正在浏览错误信息时向其发出警告，有助于避免其做出冲动反应，帮助用户深呼吸、保持清醒的头脑来浏览网络媒体。**

索引词—**错误信息、机器学习、以人为中心的安全、网络安全、网络技术。**

I. 引言与背景

互联网迅速成为用户保持联系的主要方式，新冠肺炎疫情暴发后更是如此；然而，这也带来了若干问题 [1]。无论是与亲友联系、评论时事，还是单纯寻求娱乐，互联网在我们获取信息和保持信息畅通的过程中都发挥着至关重要的作用[2]。然而，在一个奖励关注度而非质量的环境中，有些人转而采用更可疑的做法，例如煽情主义、错误信息和令人震惊的图像。

A. 错误信息与标题党的激励机制

随着内容每秒不断涌现，关注重点也从制作优质内容转向攫取注意力。当用户的兴趣不断从一篇文章、一个视频或一张图像转移到下一个内容时，出版商便需要寻找新的方式来吸引点击。因此，这很容易导致煽情主义、标题党，最糟糕的情况下，甚至会出现公然传播错误信息而用户却毫不知情的情况。

信息时代持续不断的新闻周期，让新闻媒体难以及时报道每一条新兴新闻。相反，一些媒体选择聚合内容：要么对现有内容稍作编辑，要么引导用户前往其他来源阅读相关信息。新闻聚合这种做法颇具争议，有些人认为这公然侵犯了内容所有权，另一些人则视其为在快节奏的信息经济中生存的唯一可行之道 [3] [4]。

B. 分享错误信息与诱饵式标题的影响

无论标题党及类似做法是否符合道德规范，一旦错误信息触达更广泛的受众，发布错误信息都可能造成危害。许多互联网用户会使用社交网络，这会通过快速传播虚假信息和半真半假的信息，加剧错误信息的影响。通常，这会形成一种连锁反应：一名用户将故事分享给另一名用户，后者再将其分享给另一名用户，如此循环往复。这一现象的影响并不要求用户在社交媒体平台上拥有大量关注者。Cˆot´e 和 Darling [5]发现，一旦 Twitter 账户的关注者超过约 1000 人，它就更有可能吸引来自各种不同背景的其他用户，包括公众成员、外联团体，甚至具有影响力的决策者。这会大幅扩大一段内容能够触达并被分享的潜在受众范围。这样的情况可能导致用户多次看到相同的信息，尽管这些信息是由不同用户发布的。

Fazio、Rand 和 Pennycook 分析了广泛分享信息的影响 [6]。作者考察了重复接触信息（无论真实与否）如何影响个人对所呈现数据的真相感知。参与者被要求根据陈述的可信度，按照百分比量表为一系列陈述打分。陈述以两种方式呈现：有些只展示一次，另一些则反复展示，并穿插在只展示一次的陈述之间。除了反复看到相同信息这一因素外，没有其他因素影响参与者在准确性量表上的真相感知，而这种感知却有所提升。

参与者将信息评为真实时，该信息确实被评为真实。同样，被强烈认为是虚假的信息，也不足以影响参与者，使其

转而将其归入“认为真实”这一类别。在 1%-30% 区间内，这一点尤为明显：重复接触并未使参与者意见发生显著变化。论文的一项关键发现与被认为模棱两可的信息有关。起初被认为不明确的信息，随着用户接触该陈述的次数增加，逐渐被视为准确且真实可信。这在 41-70% 区间尤为明显，该区间的评分平均比最初提高了 6 分。

现有研究 [3]–[6]表明，在当今高度社交化的时代，用户可能会因误导性陈述、夸张言辞以及重复接触信息而产生认知偏差。这种重复接触不一定仅仅源于反复看到同一篇新闻文章，也可能来自围绕该主题展开的讨论。网上呈现的观点日益分化且充满对抗，使用户更难判断真假。然而，仅仅声称用户“受到错误信息误导”或“错了”，并不会让他们更愿意重新审视自己的理解。相反，我们需要采取更审慎的方法，为用户提供一个起点，让他们得以形成自己的结论。正如 Bronstein 等人 [7] 所指出的，旨在促进“开放和分析性思维”的干预措施，或有助于遏制错误信息的影响。

C. 告知用户网络风险

近年来，错误信息愈发显著且广泛传播，关于在用户受到误导时向其发出警告的研究，因此在学术界和业界都受到越来越多的关注。许多社交媒体网站已采取措施，遏制错误信息对各自平台的影响，通常会向用户展示警告标签。这种做法简单直接，能够迅速告知用户，他们正在查看的内容可能具有误导性；由于包括 Twitter [8] 和 Meta 旗下的 Facebook[9]在内的多个社交媒体平台都采取了类似做法，因此颇受欢迎。

Ross 等人 [10] 分析了大型社交平台采用的警告标签的有效性。作者主要关注用于劝阻用户分享错误信息的方法，并测试了两种不同的标签样式：一种仿照 Meta 使用的样式，另一种则参考当代研究设计。参与者（N = 151）看到的内容包括六篇真实报道和六篇虚假报道。第一组看到的所有报道都没有警告标签。第二组看到的报道中，一半带有警告，另一半没有。随后，研究人员要求每组参与者判断哪些报道经过篡改或属于虚假内容，哪些未被篡改。研究发现，两种警告信息都没有改变用户行为。用户似乎并未对带标签内容更加警惕，与没有标签的内容相比，他们与带标签内容互动的可能性同样高。

D. 设计有效的警告标签

Ross 等人 [10]指出，为用户提供额外的背景信息有助于遏制错误信息的影响。在设计过程中，仔细考虑向用户传达哪些信息以及如何传达，有助于限制错误信息的传播范围。

Shepherd 和 Renaud [11]对如何在浏览器中设计有效的安全警告标签进行了文献综述。作者评估了这一领域已有的研究，发现用于设计警告标签的时间和资源分配不足，导致用户感到沮丧和困惑。Ross 等人 [10]的研究似乎也印证了这一看法。作者发现，现有解决方案的有效性各不相同。有些警告错误地假设用户具备相关主题的背景知识，降低了其有效性；另一些警告的措辞则过于含糊，以至于用户无法理解自身选择所带来的后果。

为应对上述问题，Shepherd 和 Renaud[11] 最后提出了一套浏览器警告设计指南。作者指出，为隐私设计的警告与为安全设计的警告有所不同，这表明必须根据预期用途考虑不同的优先事项。所提出的指南建议使用简单、简洁的语言提醒用户注意潜在问题，并使用中性色，以避免引发过度的情绪反应。此外，指南还建议，如果最初的说明不够充分，应链接至其他资源。尽管相关研究主要针对安全用途的警告标签，但将这些建议应用于应对错误信息仍然具有价值。

E. 设计有效的浏览器警告和标签

此前，可用安全领域在警告标签设计方面已开展了大量研究。早期的警告系统通常采用情境措施，例如在页面上显示一个小型弹出窗口，告知用户潜在风险。然而，Wu、Miller 和 Garfinkel [12] 的研究表明，这些弹出窗口往往会被忽略、误解，或者用户甚至没有意识到它们的存在。

该领域的后续研究采用了不同的方法，即使用插页式警告。用户必须先与警告标签进行交互才能继续。这种方法背后的理论是，让警告成为用户注意力的中心，能够提高用户阅读警告内容并据此做出明智决定的可能性。

直到最近，大多数关于有效警告设计的研究都局限于网络安全主题，例如证书过期或网络钓鱼链接。Kaiser等人 [13] 研究了警告标签的设计，以告知用户网上可能存在的虚假信息。作者评估了几种不同的警告设计风格，从信息密集、色彩极少的设计，到视觉冲击力强但细节有限的警告，并考察了用户在真实环境中对这些设计的反应。

在一项针对238名参与者开展的调查中，作者发现，参与者最青睐包含感知风险提示（“此页面包含错误信息”）

以及建议后续采取的步骤（“请考虑查找替代来源。”）的设计。作者指出，所有受评估的设计在用户之后查阅第二个或替代来源的可能性方面均未显示出显著差异。作者认为，行为变化更可能源于手动点击查看警告所带来的阻力，而非警告内容本身。

用户对警告标签的反应受多种因素影响，其中包括标签所使用的语言。Mozilla开展的一项研究 [14] 旨在了解如何设计更好的警告标签，研究结果强调，采用带有明确立场的设计比提供客观信息更重要。这意味着，传达存在威胁这一理念，比说明威胁具体是什么更重要——先前研究表明，过于专业的警告会令用户困惑。Mozilla通过简化警告标题来落实这一做法，使用抽象但易于理解的语言。此外，对于想了解问题根源细节的用户，警告还以通俗易懂的方式说明了与安全漏洞相关的风险。

为误导性内容添加标签是一项颇具挑战的工作。必须界定何为错误信息，而设计能够促进批判性思维、避免用户条件反射式反应的警告系统，仍是一个持续探索的领域。尽管各大社交平台采用的警告用户方式可能效果有限 [10]，Shepherd和Renaud [11] 指出，警告标签可以适应用户的既有知识和理解水平，而不会引发不必要的恐慌或担忧。同样，Kaiser等人 [13] 认为，此类研究可用于打击错误信息。

F. 使用机器学习检测标题党

每秒都有海量内容发布到网上，人工审核员不可能评估并审查所有可疑上传内容。因此，有必要采用一种自动化系统，能够快速、可靠地分析内容，识别其中可能具有误导性的信息，并整合干预措施。机器学习在发现和预测信息模式方面具备固有优势，非常适合用于应对错误信息。此外，过去十年来，随着计算能力和数据存储技术日趋成熟，机器学习重新受到关注，使其能够在众多应用场景中落地。

Chen、Conroy 和 Rubin [15] 探讨了是否可以使用机器学习方法检测标题党，进而检测错误信息。通过对该领域研究进行全面考察，作者指出了四种独特的标题党检测方法。最初，作者将重点放在标题党文章的文本内容上，发现标题党通常呈现出其形式所特有的词汇和语义特征。作者引用了 Lex、Juffinger 和 Granitzer [16] 的研究，该研究根据词长、选词和术语等因素对标题党进行了分析，并发现机器学习模型可以训练为无论讨论的主题是什么，都能以 77% 的准确率检测标题党。

利用未指明所指的代词，或暗示文章中的内容来迎合用户天生的好奇心，这也符合标题党的特征。Rubin 等人后续发表的一篇论文[17] 指出，自动化事实核查和验证系统可以帮助检测文本中的语言模式，并提醒用户，他们即将阅读的内容可能具有误导性。作者还指出，这类工具也可能对记者有所帮助，在他们混淆主张或无意中造成误导时向其发出提醒。

标题党并不严格局限于相关文章的文本内容；作者还发现，图像以及用户与文章的互动方式等周边因素也发挥着关键作用。关于前者，作者 [17]引用了 Ecker 等人的研究 [18] ，他们发现标题党文章往往会使用与标题不相符的图片。在这类文章中，图片可以通过醒目但无关的图像吸引潜在读者的注意力，或在读者阅读文章之前影响其观点。

作者 [17] 还指出，先前的研究发现，标题党媒体通常旨在吸引用户的注意力，然后将他们引向赞助内容或广告。此外，“阅读文章所花费的时间”与“分享和评论该文章所花费的时间”之间的差异，也可能是标题党的一种标志。在这方面，标题党文章往往使用煽动性或耸人听闻的标题，以便在用户实际阅读其中内容之前，激发其下意识的回应（无论是评论还是分享）。

G. 问题空间

信息时代的迅速兴起，促使一些人采用不道德的做法来提高参与度。无论这些做法是否有意实施，都会对社会构成严重威胁。尽管已有研究探讨了警告标签的使用，但其效果可能取决于设计方式。考虑到每秒发布的内容数量庞大，人工标注内容的准确率是不可能的。因此，机器学习提供了一种颇具吸引力的替代方案。错误信息构成了严重威胁；因此，将机器学习和警告设计方面的进展结合起来，便可提出一种有效的解决方案，保障用户安全。

II. 方法论

所提出的方法由两个部分组成：用于分析和分类内容的机器学习模型，以及用于向用户传达潜在风险的浏览器扩展。图 1 展示了简化后的流程。

使用 TensorFlow [19] ，并采用 Chakraborty 等人 [20]，所用的相同数据集，在 32,000 条新闻文章标题上训练了一个序贯模型，并将其标记为“标题党”或“非标题党”。该模型由四层组成（不包括输入层），会将输入文本分词并转换为一个 64 维稠密向量，再将其输入全局

平均池化滤波器。随后，输出经过一层 ReLU 节点，再经过最后一层 Sigmoid 节点，得出一个介于零（表示中性）和一（表示极具误导性）之间的实数。

随后，模型通过 Native Manifest 与浏览器扩展相连，使浏览器能够将文章的部分内容（例如标题）发送给模型进行分析并生成评分。评分随后会返回浏览器，之后便可向用户显示相关警告。警告旨在为用户提供信息并便于采取行动，清楚说明感知风险及建议采取的后续步骤。

A. 开发机器学习模型

1）数据集：本项目采用了 Chakraborty 等人[20]使用的同一数据集，其中包含 32,000 条新闻文章标题，并标注为“标题党”或“非标题党”。该数据集为模型构建提供了稳健且相关的基础。具体而言，标题党和错误信息都依赖煽动情绪的语言来引发反应，这表明该数据集有助于开发出特别适用于检测此类语言的模型。

2）模型：在实际应用中，本项目的后端以二元分类为核心：这段文本是否属于标题党／错误信息？因此，序贯模型凭借其单一的输入输出结构，被认为是最合适的选择；这种结构非常适合此类分类任务。

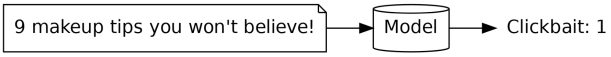

图 1. 数据处理流程示例。

3) 数据预处理：数据集被拆分为 26,666 个训练样本和 5,334 个测试样本，分别占数据集的 83% 和 17%。在训练之前，必须对数据集进行格式化，使模型能够区分标题党和非标题党。为此，我们将数据集中的标题转换为词嵌入词汇表（图 2）。为确保数据集中的数据格式统一，所有序列都填充至 24 个词元；我们认为这是标题的安全最大长度。超过 24 个词的标题会自动截断至最大长度。

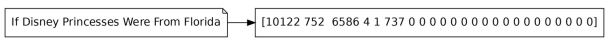

图 2. 词嵌入示例。

4）构建模型：模型如图 3 所示，由输入层、两个隐藏层和输出层组成。第一层接收输入（经过分词的句子），并将其转换为 64 维稠密向量。使用稠密向量可以压缩句子的语义含义，从而确保更好的泛化能力。

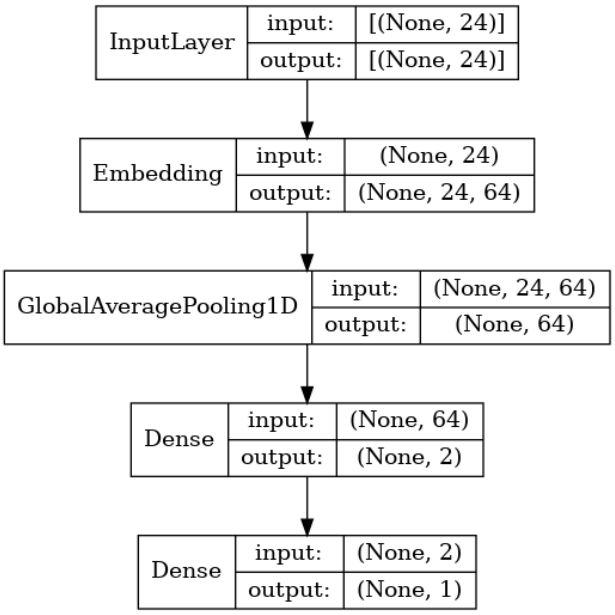

图 3. 模型可视化。

尽管稠密向量能够提取给定句子的潜在含义和联系，但如果它们过于细致，也可能导致过拟合。为了解决这一问题，第二层由全局平均池化组成。该层接收 64 维向量，并计算每个输入通道（24 维词元序列）的均值，从而使模型能够学习嵌入的近似值，而非其精确值（图 4）。

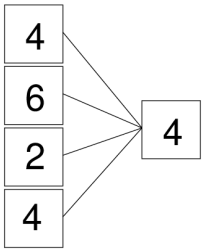

图 4. 一维全局平均池化的可视化。

在此阶段，输入数据已完成格式化并经过近似处理，以限制过拟合，随后即可构建各层，由这些层决定模型的输出。模型的第一个激活函数采用修正线性激活函数（即“ReLU”）。ReLU 可确保网络的下一层接收到正值：对于小于或等于 0 的输入值，ReLU 输出 0；对于大于 0 的输入值，则输出原值。最后，数据经过 Sigmoid 激活层，确保最终输出介于 0 和 1 之间，即：“这段文本是不是标题党？”

模型根据其预测结果的准确率来衡量性能。该任务涉及分类；

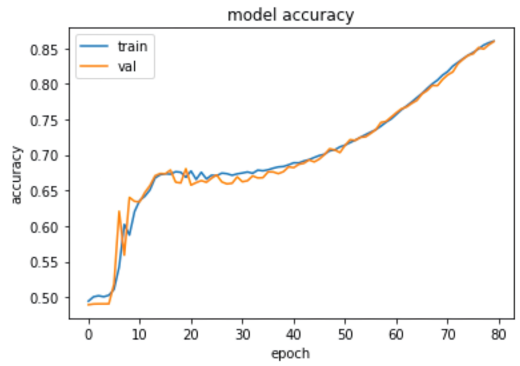

图5. 模型训练期间的准确率图（越高越好）。

因此，使用二元交叉熵损失函数。该函数计算模型的预测结果与数据集标签之间的偏差。通过使用动量优化器（也称为随机梯度下降或 SGD）的梯度下降，进一步最小化损失函数。优化器通过在训练迭代中最小化损失值，帮助提高模型的训练速度。如此一来，模型的预测结果便会逐渐趋近于预期输出。

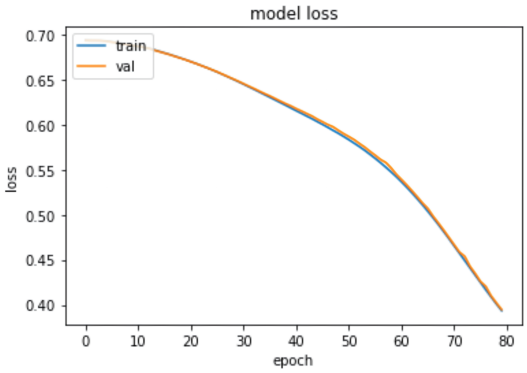

图6. 模型训练期间的损失图（越低越好）。

模型使用批量大小为128的数据训练了80个训练轮次。尽管批量大小大于32可能导致欠拟合，但我们的设想是，较大的训练轮次会逐渐提高准确率并实现收敛，这一点可以从图5和图6中看出。模型经过编译、训练和评估后，被导出以供浏览器扩展使用。

B. 浏览器扩展和警告信息

1) 创建浏览器扩展：我们开发了一个浏览器扩展，以确保模型能够部署于真实环境中。借助该扩展，模型可以在用户浏览新闻文章时对其进行分析，并在发现误导性内容时发出警告。

Native Manifests [21] 用于让网络浏览器与本机应用程序进行交互，在两者之间来回传递数据。这些清单使 TensorFlow 模型能够在本地计算机上进行预测，然后将预测结果发送到浏览器，以便进行进一步分析并输出。

在浏览器中，分析会在获取用户访问页面的标题后开始。首先，会清除空白字符和控制字符。随后对标题进行处理，并返回一个数值，用于表示该标题的煽动性程度。

在用户的设备上，程序会将任务交给本机应用程序，由它负责解析标题并将其转换为适合模型的格式，然后计算评分并将其返回浏览器。浏览器传来的数据封装为 JSON，并通过标准输入（标准输入）发送，程序会从中读取数据。随后，脚本开始导入用于加载模型并相应格式化传入数据的库。模型加载完成后，分词器会被实例化，用于转换传入的标题。至此，初始设置已完成，模型可以投入使用。

脚本的核心部分是一个循环，它会等待接收浏览器发来的消息；收到消息后即可进行解码。现在，文章的标题已经就绪，脚本会对其进行分词并添加填充，以确保与模型兼容。接下来，可以调用标准模型预测函数，对结果进行编码并返回浏览器，供用户查看。

2) 向用户显示警告：本机应用程序生成结果后，会将其返回浏览器。随后，浏览器扩展会生成一条消息并发送到新闻文章所在的页面，同时将本机应用程序的结果存储在一个变量中。内容脚本会接收这条消息；该脚本由扩展注入每个页面。

内容脚本会动态地为用户生成警告标签。具体做法是等待接收来自后台脚本的消息（即结果）。随后，该数值会乘以十，以应对可能出现的浮点数问题（例如将 0.8 转换为 8）。为避免反复向用户发出有关无害内容的警告，只有评分超过十分之五的标题才会生成警告。警告显示在屏幕上时，滚动功能会被禁用，以确保用户必须确认该警告。

就本项目而言，现有文献表明，插页式警告最有可能促使用户行为发生改变。此外，即使大多数用户似乎并不会实际阅读警告标签的内容，他们仍然倾向于保留此类信息。

受第 I-D 节所讨论论文的启发，该浏览器扩展经过设计，确保具备较强的视觉清晰度，从而有效传达与一则误导性内容相关的风险。

该警告采用付费墙式设计，模仿用户可能已在其他新闻网站上熟悉的方式。这有助于确保警告不会被忽略，而情境式警告就可能出现这种情况。为进一步确保警告引起用户注意，系统会使用覆盖层使文章变暗，并阻止用户滚动

页面。警告设计有 5 种变体，根据文章的评分从最轻微到最严重不等。图7展示了示例设计

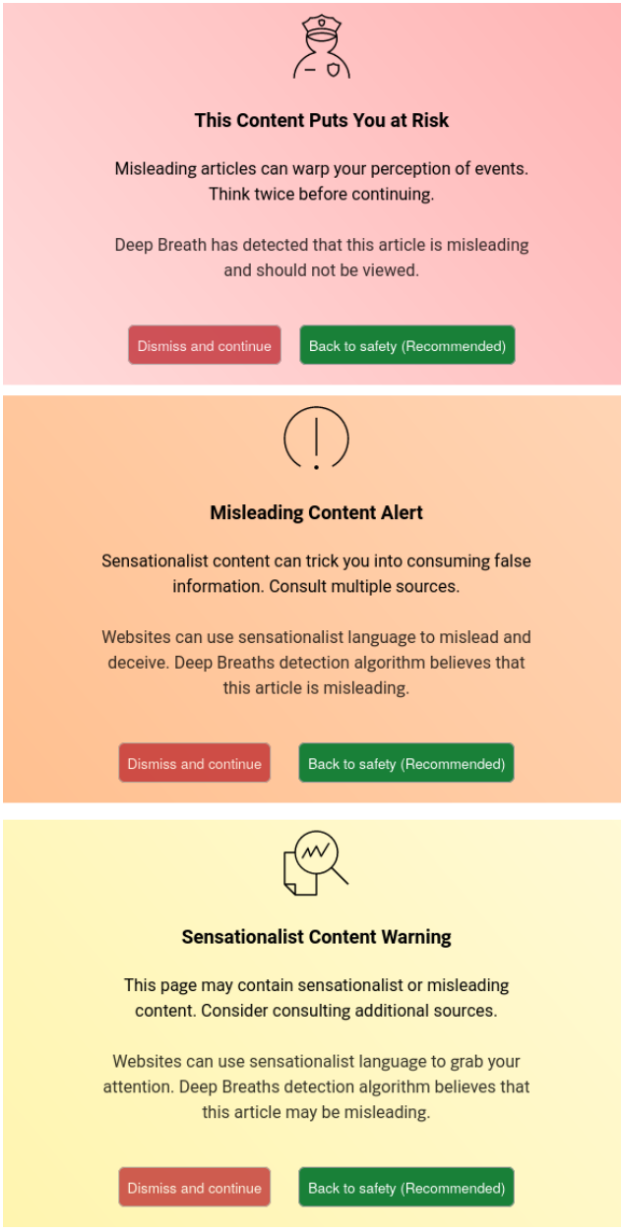

图 7. 警告设计示例。

此前警告设计存在的一个关键问题是风险传达不佳，警告可能会被行业术语 [13]掩盖。该警告设计旨在通过以易于阅读的格式传达尽可能多的相关信息，最大限度地减少现有问题。醒目的符号根据文章的评级表示不断上升的风险等级。风险刻度较低的文章会使用更为通用的“放大镜”符号，以强化人们更批判性地思考文章价值的观念。如果文章包含更严重级别的错误信息，则会使用更加醒目的“警报”导向的符号，例如警告标志、停车标志以及警察形象等权威象征。

此外，警告后方采用了振荡渐变。警告的严重程度不同，所使用的颜色也会从黄色逐渐变为橙色，再变为红色。渐变的细微变化旨在吸引用户注意警告，而每条警告采用的独特颜色也有助于用户理解相应的风险等级。

归根结底，该扩展旨在改变用户行为，并帮助用户了解未来可采取哪些切实有效的措施来保护自己免受错误信息的影响。因此，每个警告标签都采用了独特的措辞，不仅告知用户感知风险是什么，还提供有关后续可采取行动的建议。

警告底部向用户显示两个按钮，用户可以选择关闭警告并继续浏览，或离开当前页面。为表明推荐的行为，离开页面的选项以醒目的绿色显示，并在括号中标注“推荐”。相反，关闭警告的选项以红色显示，并稍微淡化处理，使其刻意与背景融为一体，直到用户将鼠标悬停在按钮上。

三、结果与讨论

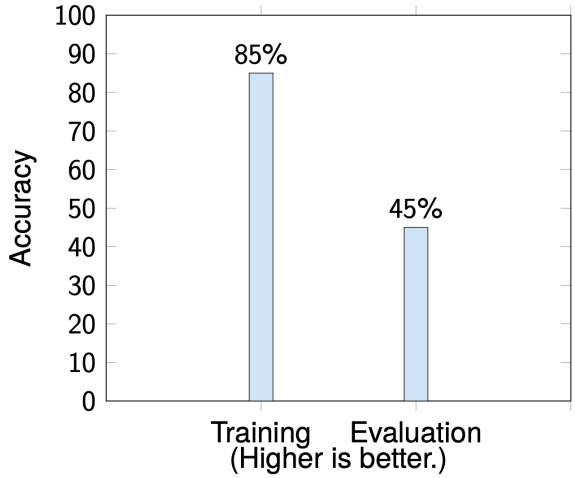

图 8. 训练与评估的准确率比较。

在训练过程中，模型的准确率约为 85%，损失值为 0.39；而在评估数据集上测试时，模型的准确率约为 45%，损失值为 1.15（图8、图9）。准确率下降可能是由于现有评估数据集存在不一致之处，例如数据格式不规范，或部分标题的标签似乎不正确。可以通过进一步的数据清洗来解决这一问题。另一个值得注意的地方是，评估数据集只使用了二元标签（这个标题是“标题党”还是“新闻”？），这也可能导致准确率出现差异，因为模型输出的是 0 到 1 之间的数值，而不是纯粹的二元输出。

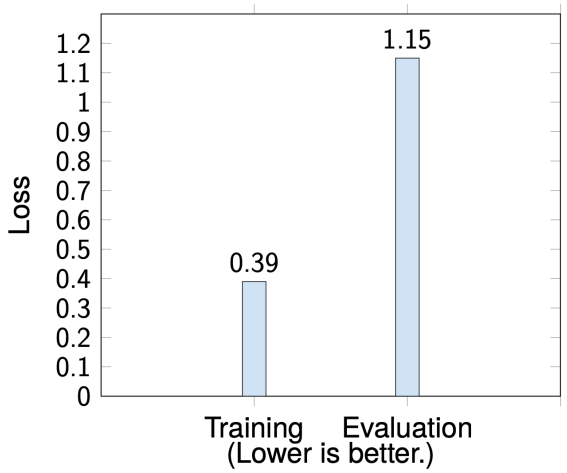

图 9. 训练与评估损失值对比 on.

关于机器学习，该项目证实了 Lex、Juffinger 和 Granitzer（2010）的研究发现：可以根据词汇语义检测标题党和错误信息，具体依据包括选词、词长和词语共现性，即单词 x 经常与单词y一同出现。

IV. 结论与未来工作

本文展示的模型已表现出较为可靠的性能，但仍可进一步改进，以取得更好的结果。模型的准确率仍在上升，损失值仍在下降，这表明在曲线趋于平缓之前，模型性能还有望进一步提升。此外，还可以拓展模型的能力。该模型仅使用标题党风格的标题进行训练，且效果良好。不过，若使用标题党文章的内容进行训练，或许能取得更稳健的结果，使模型更深入地理解文章，并作出更细致的预测。本文采用的模型是一种序贯模型，旨在接收单个输入并生成单个结果。尽管这种模型能够有效地将单个标题分类，但支持多个输入和输出可以实现更多功能，例如评估文章标题以及从文章中选取的若干句子。内容的评级是动态的，但相应的警告仍是固定的。拓展模型的能力后，或许还可以针对内容提供个性化警告。在实际应用中，这意味着可以就文章的特定方面向用户发出警告，例如耸人听闻的作者、误导性句子和错误说明的图像。尽管我们已尽力确保模型作出平衡且准确的预测，但任何系统都并非万无一失。开展用户测试，并允许用户在认为模型出现误判时进行报告，或有助于纠正模型的失误。

本文所 प्रस्तुत的工作通过应对网络错误信息问题，取得了令人鼓舞的进展将机器学习、人机交互研究和网络技术相结合。研究结果验证并拓展了先前的研究，而将机器学习与易用性安全相结合仍是一个相对鲜有探索的研究领域。

参考文献

[1] H. S. Lallie、L. A. Shepherd、J. R. Nurse、A. Erola、G. Epiphaniou、C. Maple 和 X. Bellekens，“COVID-19 时代的网络安全：疫情期间网络犯罪与网络攻击的时间线及分析”，《计算机与安全》，第 105 卷，页码 102248，2021 年。[2] A. Mitchell、E. Shearer 和 G. Stocking，“Twitter 上的新闻：大多数用户都会浏览，许多人信任”，2021 年。[在线]。可访问：https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/ [3] H. I. Chyi、S. C. Lewis 和 N. Zheng，“寄生者还是合作伙伴？新闻聚合时代对谷歌新闻的报道”，《新闻与大众传播季刊》，第 93 卷，第 4 期，第 789–815 页，2016 年。[4] M. Coddington，“收集证据的证据：新闻聚合作为一种认识论实践”，《新闻学》，第 21 卷，第 3 期，第 365–380 页，2020 年。[5] I. M. Cˆot´e 和 E. S. Darling，“Twitter 上的科学家：是在向志同道合者布道，还是在大声宣讲？”《*Facets*》，第 3 卷，第 1 期，第 682–694 页，2018 年。[6] L. K. Fazio、D. G. Rand 和 G. Pennycook，“重复会同等程度地提高人们对可信与不可信陈述的感知真实性”，《心理学通报与评论》，第 26 卷，第 5 期，第 1705–1710 页，2019 年。[7] M. V. Bronstein、G. Pennycook、A. Bear、D. G. Rand 和 T. D. Cannon，“对假新闻的信念与妄想倾向、教条主义、宗教原教旨主义以及较弱的分析性思维有关”，《记忆与认知应用研究杂志》，第 8 卷，第 1 期，第 108–117 页，2019 年。[8]Y. Roth 和 N. Pickles，“更新我们应对误导性信息的方法”，2020 年，https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information（访问日期：2022 年 2 月 8 日）。[9] Facebook，“对屡次分享错误信息的人采取行动”，2021 年，https://about.fb.com/news/2021/05/taking-action-against-people-who-repeatedly-share-misinformation/（访问日期：2022 年 2 月 8 日）。[10] B. Ross、A. Jung、J. Heisel 和 S. Stieglitz，“社交媒体上的假新闻：警告信息的（无）效力”，载于*CIS 2018* 论文集，第 *16* 卷。信息系统协会，2018 年。[11] L. A. Shepherd 和 K. Renaud，“如何设计浏览器安全与隐私警报”，载于*2018 AISB* 大会：网络安全数字行为干预研讨会。人工智能与仿真研究学会，2018 年，第 21–28 页。[12] M. Wu、R. C. Miller 和 S. L. Garfinkel，“安全工具栏真的能防止网络钓鱼攻击吗？”载于*SIGCHI* 计算系统中的人为因素会议论文集，2006 年，第 601–610 页。[13] B. Kaiser、J. Wei、E. Lucherini、K. Lee、J. N. Matias 和 J. Mayer，“调整安全警告以应对网络虚假信息”，载于第 *30* 届 *USENIX* 安全研讨会（*USENIX Security 21*），2021 年，第 1163–1180 页。[14] M. Walkington，“设计更好的安全警告”，2019 年，https://blog.mozilla.org/ux/2019/03/designing-better-security-warnings/（访问日期：2022 年 2 月 8 日）。[15] Y. Chen、N. J. Conroy 和 V. L. Rubin，“误导性网络内容：识别标题党式‘假新闻’”，载于*2015* 年 *ACM* 多模态欺骗检测研讨会论文集，2015 年，第 15–19 页。[16] E. Lex、A. Juffinger 和 M. Granitzer，“在线媒体中的客观性分类”，载于第 *21* 届 *ACM* 超文本与超媒体会议论文集，2010 年，第 293–294 页。[17] V. L. Rubin、N. Conroy、Y. Chen 和 S. Cornwell，“假新闻还是真相？利用讽刺线索检测可能具有误导性的新闻”，载于第二届欺骗检测计算方法研讨会论文集，2016 年，第 7–17 页。[18] U. K. Ecker、S. Lewandowsky、E. P. Chang 和 R. Pillai，“新闻标题中微妙错误信息的影响。”《实验心理学杂志：应用》，第 20 卷，第 4 期，页码 323，2014 年。[19] M. Abadi、A. Agarwal、P. Barham、E. Brevdo、Z. Chen、C. Citro、G. S. Corrado、A. Davis、J. Dean、M. Devin 等，“TensorFlow：面向异构分布式系统的大规模机器学习”，*arXiv* 预印本 *arXiv:1603.04467*，2016 年。

[20] A. Chakraborty、B. Paranjape、S. Kakarla 和 N. Ganguly，“停止标题党：检测并防止在线新闻媒体中的标题党”，载于《社会网络分析与挖掘进展》（*ASONAM*），*2016* 年 *IEEE/ACM* 国际会议。IEEE，2016 年，第 9–16 页。[21] Mozilla，“Native Manifests”，2022 年，https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/Native manifests（访问日期：2022 年 3 月 5 日）。