总体实现方案

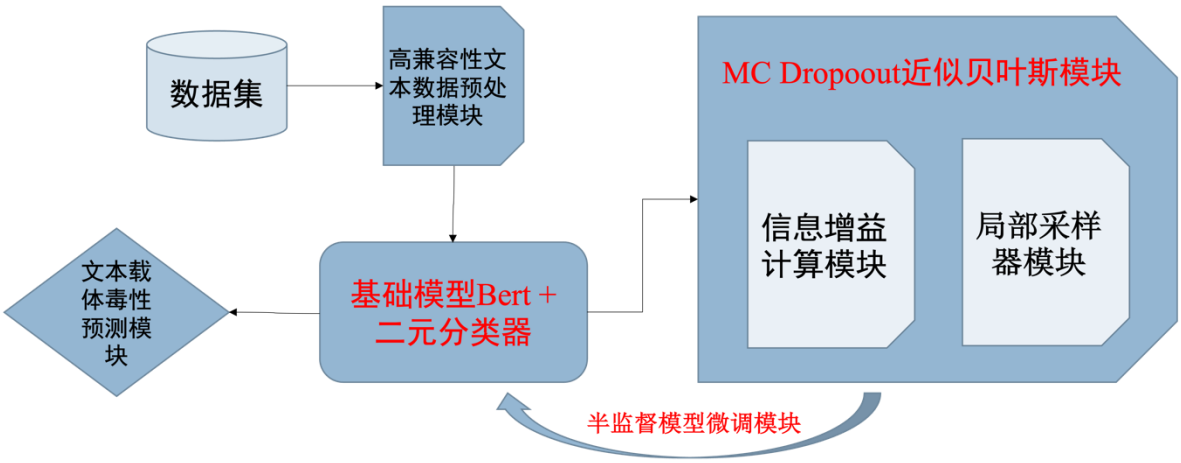

图 1 总体实现方案框架图

在本系统的实现过程中，我们将贝叶斯神经网络理论与预训练模型相结合，以维持系统在样本标记数量受限场景下的性能，设计了一整套高效且可靠的算法和模块，具有极强的现实意义与实践意义。总体实现方案框架图如图 1 所示。

系统设计涵盖了从基础模型搭建、文本数据预处理模块、MC Droopoout 贝叶斯近似模块等各个环节，确保了系统可以高效完成对恶意隐写载体的检测，同时大大降低系统实现成本，特别是数据标注成本。

- **基础模型搭建**：基础模型搭建是初始微调过程中的关键步骤，我选择了使用 BERT 预训练模型加上密集层分类器进行构建。在这一过程中，首先利用预训练模型 BERT 强大的特征提取能力处理文本数据，然后通过一个全连接的密集层对输出特征进行分类。这样，基础模型既能捕捉文本的深层语义信息，又能进行准确的分类，为后续的伪标签生成和自训练提供坚实的基础。

- **高兼容性文本数据预处理模块**：该模块的目的是从 CSV 文件中读取文本数据，并保证代码在 Python2 和 Python3 这两种常见环境中的兼容性，通过选择合适的编码方法，确保数据能够正确读取并处理，进而为后续的模型训练和分析提供高质量的输入数据。这种兼容性的处理方式确保在不同的应用场景下，数据预处理步骤能够稳定、可靠地运行。

- **MC Dropout 近似贝叶斯模块**：MC Dropout 近似贝叶斯模块通过在模型预测时持续启用 dropout 层，从而在前向传播过程中生成多个不同的模型预测。这种方法等效于采样多个模型，并以此近似贝叶斯推断，捕捉模型预测的不确定性，在这个过

程中 MC Dropout 不仅提升了模型的泛化能力，还提供了关于预测结果的置信度评估。

- **信息增益计算模块**：该模块用于评估特征在划分数据集时带来的信息纯度的提升，通过计算某一特征在当前节点分裂后的信息熵减少量，从而衡量其对目标变量的不确定性降低效果。此模块在采样器构建过程中扮演关键角色，帮助选择最优的分裂特征，提高模型的分类性能，优化系统的检测效果。

- **局部采样器模块**：局部采样器模块用于从无标签数据集中选取有代表性的小批量样本组成伪标签数据集，以便在训练过程中进行二次参数更新和模型优化。该模块确保所选样本能够覆盖数据的多样性和特征分布，从而提升模型在训练阶段的效果和收敛速度。

- **半监督模型微调模块**：半监督模型微调模块通过利用有限的有标签数据和大量的无标签数据来逐步优化模型性能。该模块首先使用标签数据对基础模型进行初步训练，然后使用模型对无标签数据进行预测并生成伪标签。再通过多次迭代，不断将高置信度的伪标签样本与有标签样本进行合并训练，使模型在每次迭代中渐进式提高对数据的全面理解和分类能力，从而在标签数据不足的情况下显著提升模型的泛化性能。

- **文本载体毒性预测模块**：该模块主要包含一个高效的分类器，使系统能够精准区分恶意隐写载体和正常载体。分类器使用特殊的 loss 损失函数进行损失度量，之后通过训练更新其内部参数。经过训练的分类器能够在实际应用中快速识别和标记潜在的恶意隐写文本，提高系统的安全性和可靠性。