

针对中文文本垃圾邮件检测任务的不平衡性、效率和文本伪装问题，提出了一种半监督生成的主动学习（Semi-supervIsed GeNerative Active Learning (SIGNAL)）模型。

本文贡献：

1. 提出了一个“自我多样性”标准来衡量候选注释的“价值”。
2. 提出了一种半监督变分自动编码器和一种特征变化图增强的数据增强方法。
3. 实验表明，该信号模型不仅对垃圾邮件样本选择敏感，而且可以提高一系列传统主动学习模型在中文垃圾邮件检测任务中的性能。
4. 本研究在关注中文垃圾邮件检测任务的同时；理论上，该模型在其他NLP任务中有很大的应用潜力。它可以通过降低标签成本来缓解数据饥饿问题。

创新点：据我们所知，这是首次将主动学习和半监督生成学习结合起来用于文本垃圾邮件检测。

发展过程：

当前模型是基于给定标签下的数据集。为了解决打标签的问题，提出了主动学习的方法。但是在垃圾信息问题中，主动学习的问题有：

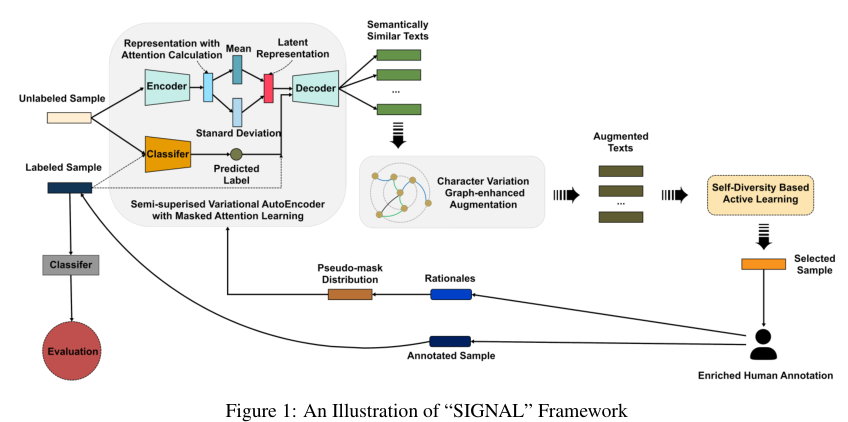
1. 不平衡性：垃圾邮件样本与正常样本的比例非常不平衡。主动学习模型应该对垃圾邮件样本更加敏感。一般的主动学习方法很难解决这个问题。
2. 效率：当与反垃圾邮件模式竞争时，垃圾邮件发送者不断为垃圾邮件文本创建新的形式。未标记的样本数量巨大，并且不断增加。经典的基于多样性的主动学习方法迭代比较每个未标记样本和每个标记样本，以选择最“多样”的样本进行标注，由于其计算复杂度为O(n^2)，因此性能较差。因此，需要一种高效的面向对象的主动学习算法。
3. 伪装性：汉字有字形和语音变体。“账（账户）“和”帐（窗帘）“具有相似的结构和发音。垃圾邮件发送者可以利用这一特征逃避检测算法（相当于谐音字）。

因此，我们提出一种新的主动学习模型来预测标记数据集中未出现的新汉字变异模式。

本文前提假设：对增强文本（augmented text）预测的多样性是预测候选文本样本对分类器性能提升能力的有效指标。

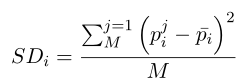
原因：以图像为例，从相同图像生成的patch共享相同的标签，并且自然期望分类器具有相似的预测。因此，面片预测的多样性可以成功地衡量候选图像在提升当前分类器性能方面的“能力”。同样，在本研究中，每个候选样本的一组语义相似的文本通过数据扩充自动生成。

模型：



问题：如何评价多样性指标？

对于样本，有一系列的增强文本= 。多样性定义：

**

是对的预测，是所有的的均值。M是增强文本的数目。

直观的来看，SD表示候选文本的“价值”。SD较大，表明当前分类器对目标候选信息的预测是不稳定的。只要有一点突变，预测就会发生巨大的变化。这样的候选人值得被人工标注。

SD特征：

1. 该准则具有定位重要样本和降低计算复杂度的潜力。
2. 在中文文本垃圾邮件检测的背景下，垃圾邮件候选者获得更大SD的可能性更大。例如，如果垃圾邮件候选者在关键位置发生变异，则增强文本的标签可能会改变。相反，非关键词修改的文本受这种情况影响的可能性较小。

S-VAE和遮挡注意力学习：

作用：生成相似文本

使用带注解的根本标签R（即一组重要的术语，可以理解为一种带标签的训练），一个伪遮挡分布P\_m用于生成候选者样例。对第i候选者样例t\_i，该概率Pr\_i计算公式为：



是一个决定t\_i是否属于R的函数，是用于正则化，用来将重要术语的注意力降低（即关注不重要的部分）。