Improving Text Classifier Performance through Human-in-the-Loop Error Correction: Enhancing Learning from Explanations

Author: Xinyang Song Supervisor: Edwin Simpson

Abstract(object: what I hope to say about my thesis)

文本分类模型通过自然语言处理来分析文本并分配标签。在危机场景中例如洪水，地震，使用文本分类器可以识别和转发来自社交媒体的紧急文本报告给相关机构。然而，文本分类器的有效性在很大程度上依赖于大量的标记训练数据，而这些数据可能会出现稀缺以及难以获得的情况[1]。除此以外，训练大量的标记好的数据会延迟模型反应时间，不具有代表性的数据也会影响模型准确率。确定搜索和救援请求等可操作的信息类型仍然具有挑战性。

本项目结合真实以及生成的模拟用户的判断和解释整合到分类器的训练过程中，使用针对多分类问题的主动学习策略to optimize the human-in-the-loop（HITL） process for error correction作为解决上述限制的替代方法。结合以前的一项技术即Representation Engineering with Natural Language Explanations（ExpBERT），此技术增加从解释中生成特征结合原有特征来提高分类器的性能[2]。本项目的主要目标是在分类过程中采用交互系统的多种方式，超过ExpBERT文本分类器的最大可实现精度[3]。在迭代过程中，使用基于不确定性，代表性以及多样性等方面对应的抽样策略来query高信息度的未标记实例。主动学习用少量的有效的数据去训练，可以减少文本分类任务中的标记工作以及处理时间，并将文本的解释精确化。

~~期间探索贝叶斯深度主动学习对系统的有效性，近似神经网络的贝叶斯推理即Monte Carlo dropout (MCDO)可以应用于深度学习模型的训练以及特征提取中，其随机前向传递(SFP)的结果还可以通过不确定性评估来应用于查询策略中。注释者接受并处理用于训练的抽取完成的无标签实例。通过这个过程产生的带标签的有效数据以及最优超参数将被用来微调ExpBERT分类器，反复迭代最终提高其准确性。~~

~~除此以外，传统神经网络分类模型对输入样本的不确定性及其分布无法给出解释，因此无法给出置信度。因此，本项目将应用贝叶斯神经网络（BNN）通过inference方法来降低原本方法中的传统神经网络带来的不确定性。为了评估本方法的性能，此实验将设置对比组，将此方式与未改进模型以及其他算法下的模型例如模块化python库Small-Text的性能对比。其中包括使用不同主动学习策略的ExpBERT分类模型，有无BNN的深度主动学习改进的模型以及其他算法下的模型，在一个 CrisisNLP的数据集上进行比较。最后的结果发现优化后的模型具有高准确率以及低延迟。~~

探索主动学习对系统的有效性。注释者接受并处理用于训练的抽取完成的无标签实例。通过这个过程产生的带标签以及有精确解释的有效数据以及最优超参数将被用来微调ExpBERT分类器的解释生成过程以及训练过程，反复迭代最终提高其准确性。为了评估本方法的性能，此实验将设置对比组，对比使用不同主动学习策略的ExpBERT分类模型，有无贝叶斯主动学习改进的模型以及其他算法下的模型，在一个 CrisisNLP的数据集上进行比较。最后的结果发现优化后的模型具有高准确率以及低延迟。

本论文的主要结论如下：

Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah.

Introduction

Background

社交平台紧急时间响应系统大部分的焦点集中于创造更好的文本分类算法来从数据中学习。然而，获得有用的注释数据集可能被证明是困难的[4]。一般社交平台数据的吞吐量是无法完成批量注释的，因此很多弱监督形式可以控制成本的情况下实现提取信息价值高的数据并标注，已经广泛的应用在了分类项目中[5,6]。为了减少标注的成本，社交网站后台采用主动学习的方法来完成注释。主动学习是解决这些问题的有效方法，它选择信息含量高的未标注样本，由专家进行标注[7, 8]。利用不同的抽样的评估算法查询信息量最大的实例可能是主动学习中最流行的方法。因此，查询策略自然成为主动学习算法的一个研究热点，各种优化算法层出不穷。将注释者放到优化循环中最终达到高准确率。

同时，神经文本分类器通常对早期的不确定性抽样并不适应[9，10]，并对结果的产出通常被认为是过度自信，对NN的复杂性的无效反应依旧是研究的重要领域。因为它没有内在的脆弱性措施。相比之下，将主动学习框架不仅应用于标签的注释中，而且应用在对文本的解释特征的挑选中是一个具有创新性的实验，模型本身的难以解释性也可以采用此主动学习的方式来增加置信度。

Motivation

Against the above background,研究方向主要集中于文本分类系统的准确性，交互性以及健全性和代表性

准确性：

由推特代表的社交平台的统一特征是数据量庞大，网络语言的噪音众多。尤其是在捕捉紧急需求时，需要对紧急信息类别进行分类以及出动相关部门来解决问题。但是，传统的文本分类系统训练中需要完成信息的高吞吐。由于信息文本具有着不同的分布，其所表现的信息量的不同会极大的影响模型的准确度，所以采集并识别重要紧急信息的效果大多差强人意[1]。因此，采用主动学习的半监督模式提高准确率是非常重要的。

交互性：

传统的深度学习模型缺少主动选择样本的能力。在具有进化功能的网络文本中，传统增加了解释功能的文本分类器往往是静态的，因此难以面对新的，变化快速的样本。所以，带有交互性机制的系统能够实现快速的标记样本，并且通过与用户的交互可以进一步提高模型的性能和准确性。尤其是在紧急情况下，深度主动学习的交互性可以使模型更加灵活和适应性强，能够及时识别和分类与公共安全相关的信息，为应急响应提供重要的支持。因此，深度主动学习在处理推特上紧急需求的分类等交互性场景中的重要性不言而喻，它能够通过与用户的交互实现快速准确的标记样本获取，并为应急响应和公共安全提供有力的支持。

健全性：

传统的文本分类模型在处理文本分类任务时常常表现出不确定性和泛化能力不足的问题。传统初训练产出的具有解释的样本通常连接所有实例，并没有为特殊实例去定制。因此将解释生成与主动学习相结合，可以提供更健全的文本分类系统，能够更好地处理模型的不确定性和泛化能力等方面的不足。

代表性：

主动学习的思想能够帮助模型在有限的标记样本下获得更高的性能。通过主动选择样本进行标记，模型可以在每次迭代中选择最有价值的样本，以提高模型的性能。除此以外，能够在大规模未标记数据中提取丰富的特征表示以及解释，这一点在紧急事件多分类主动学习问题中可以减少很大一部分的偏差。

Problem statement

首先，基于上述背景我们可以得出，基于解释的神经文本分类模型需要使用少量的但信息量大的实例来训练，并且可以在训练过程中发挥模型参数的最大潜能。其次，如何选择具有大信息量的实例供给给人类注释者，探测实例的后验表现也是本实验的重点。再其次，由于人工成本的问题，如何模拟人类注释者的处理进程也是减少人工的有效途径。最后，传统的主动学习查询策略基于固定的特征表示，却缺少对主动学习在解释生成中的作用的研究。但是仅仅在主动学习中对文本分类模型进行微调可能出现不一致的问题，所以，对于具有ExpBERT的神经网络文本分类器的优化结构上的改进的探索也将对社交平台上应急时间的检索有着良性的影响。

Object：

因此，结合上述问题有如下的研究目标：

开发一个pool-based human-in-the-loop主动学习框架，并根据不确定性、多样性等评估不同采样查询算法的影响，适当整合提取算法，设计中止标准，研究互动系统是否能提高文本分类模型的性能。

进一步将主动学习应用到ExpBERT文本多分类模型的解释的生成过程。改进主动学习框架，研究其是否能有效提高原有优化架构的性能。

建立注释者评估模拟算法和一个现场注释者模拟程序。

探索先进的ExpBERT文本分类架构在嘈杂环境中的稳健性。

Challenges：

本项目的主要挑战是如何对原有的ExpBERT文本多分类模型的训练框架进行改进来提升性能，是否可以通过交互来提取有用信息，是否可以将解释者的功能不仅应用于标签生成而且附带着解释的精确化是一个对模型性能有着重要作用的挑战。

需要注意的是大多数主动学习是应用于二分类的问题上，因此如何应用多分类主动学习到原有框架中是一个挑战

同时，针对主动学习的文本分类系统有着多种查询策略，查询策略的应用以及对比也极大的影响了主动学习框架的性能。需要在时间复杂度以及用户体验多方面考量查询策略的选择。

最后，是对评估方式的判定，对human-in-the-loop中止标准的制定以及超参数的选择。这个对于框架在测试集上的性能表现有着极大的影响，这也是本项目值得注意的地方。

Thesis Organization

基于研究目标整体的的论文架构一共包括六大模块，分别是介绍，背景，设计，实现，评估以及结论。

在第一章介绍阶段将对项目的背景进行概述，根据背景展现实验动机，然后简述整体交互系统的实现方法，列出研究的目的以及面临的挑战。

在第二章将提供对整体交互系统的技术及其应用进行详细介绍，即交互系统在文本分类系统的应用（第2.1节），2.2节进行主动学习的框架的介绍。由于主动学习中实例查询方式的众多，2.4节将结合伪代码以及公式对抽样查询算法的介绍，以及在后续章节中介绍其他主动学习框架以及结束循环需要的中止标准。

在设计这一章的第一节将对整体框架以及方法的设计细节进行阐述。由于查询策略的重要性，在3.2节将主要介绍本报告采取的查询策略已经对实例代表性的评估设计，最后介绍在ExpBERT中主动学习下解释生成以及训练的框架设计

在实现的这一章首先将对实验环境进行概述，在4.2节对数据库中的元素以及其属性等进行介绍，然后4.3节详细阐述模拟用户操作的实现过程。由于实验中要对模型参数以及抽样策略进行评估，所以需要在4.4节描述评估标准，最后对开发资源进行概述。

在评估这一章首先要阐述超参数的选取过程，对验证集以及测试集的评估结果的展示以及分析，最后完成噪音影响的评估。

在最后一章将对前面的实验进行总结，然后将对整体模型的局限性进行解析，最终探讨进一步的提升以及未来工作。

Literature Background

Human in the loop（HITL）

传统的自然语言处理 pipeline is not designed to take advantage of human feedbac。但是，人类在循环中是作为交互系统中的重要一环，人在回路中的模拟允许识别模型的不足之处，这些不足之处在现实世界的测试之前可能并不明显[11]。Godbole et al. (2004)[12]对基于支持向量机（SVM）主动学习的文本分类器进行了扩展，以自然地吸收人类在特征工程、术语包容/排除以及术语和文档标签方面的输入，做出了统计学上合理的决定。这种新型的人类和机器学习算法之间的交互方式，可以称之为人类循环机器学习（Munro 2020）[13]。Human in the loop根据谁具体控制着学习过程可以确定不同的HITL机器学习方案（Mosqueira-Rey et al. 2023）[14]：

主动学习（AL） (Settles 2009)[15]：系统保持着控制模型学习的过程，人类作为媒介来参与对未标记的数据的注释。但是人类无法根据偏好选择无标签数据。AL将作为优化框架应用在本论文中，在2.2节中将详细介绍主动学习的应用。

交互式机器学习（IML） (Amershi et al. 2014)[16]：人类与系统保持着紧密的交互关系。针对其中AL与IML的不同店Dudley and Kristensson (2018)[17] 认为both AL and IML focus on selecting new points for labeling by the user, but in AL the selection is driven by the model and in IML the selection is driven by the user。由于IML是基于AL所以他们有着共有的缺点，但是，IML本身还有与 Human-Computer Interaction techniques (HCI)混杂的缺点，因此需要更多的独特性的研究（Michael et al.(2020)）[18]。

Machine teaching (MT) (Ramos et al. 2020)[19]: 通过人类教师的指导来训练机器学习模型。即delimiting the knowledge that they intend to transfer to the model.强调人类教师在学习过程中的主动参与和指导。 Devidze et  al. (2020)[20]指出与主动学习相比，MT更依赖于教师的专业知识，并且在样本选择和处理复杂任务方面的灵活性较差。因此，根据具体的需求，选择适当的学习方法是至关重要的。

Active learning（AL）

作为HITL中最热门的学习方案，主动学习系统通过向未标记的实例提出问题并由专家（例如人工标注者）进行标记，试图克服标记瓶颈[21]。简而言之，主动学习是找出信息量最大的未标记来交给注释者，由注释者标记并将已经标记的实例加入到模型的训练过程，通过使用更少的标签数据来实现更好的性能。在此项目中实例庞大噪音较多全部训练任务较重，准确率较低大，因此主动学习适用于此项目去减少无相关实例，增加准确率。

AL process and scenarios：

主动学习的进程如图一所示，图一为基于池的主动学习的进程，机器学习模型初始化后会先从学习已标记的实例集L开始，然后使用模型对未标记的样本集进行特征提取后根据特定的选择策略选择未标记样本提供给人类标注者。并将已经标记的样本从U集中剔除并添加到L集中用来更新模型。通过反复选择、标记和训练来逐渐增加已标记样本的数量。随着机器学习模型性能的提高，它能够更准确地选择最有意义的未标记样本进行标记。最后根据预定义的终止标准（例如达到标记样本数量的限制）来终止主动学习过程。

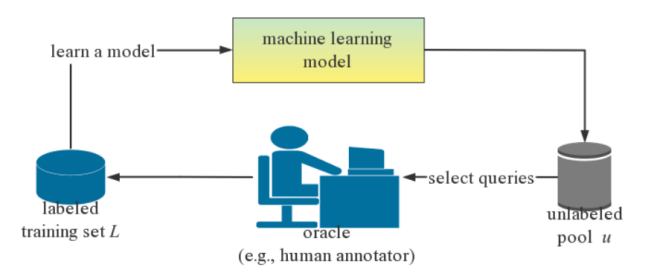


Figure 1: 基于池的主动学习过程

根据数据源， Settles (2009) [15]mentioned three main 场景设置: membership query synthesis [22, 23], stream-based selective sampling [24] and pool-based [25] AL.

Membership query synthesis: 这种设置的查询是由模型生成的，机器学习模型可以请求标记任何未标记的实例，其并未在潜在的自然分布中取样[15]。对于比较绝对的问题领域，例如对绝对坐标采用回归预测的任务中，有效的查询合成通常是有效的。因为它可以解析简单d的数据分布，构造合理数据供人类标注。但是, Baum and Lang (1992)[26]的研究中可以看出对于自然语言处理这类复杂任务中，模型可能会产生不易理解的文本串，导致人类无法对这些混乱文字进行判断。在深度主动学习背景中，可以通过生成对抗网络（GANs）用于数据增强来解决成员查询合成的场景，因为GANs能够生成合理度高的实例[27]。

Stream-Based Selective Sampling：此设置相比于Membership query synthesis而言，可以不受输入分布的影响。基于流的抽样是一次从实际的分布中抽取一个未标记实例，模型通过“信息性度量”或“查询策略“来决定是否请求实例的标签，相当于一种带有偏见的随机抽样[28]。

pool-based AL: 作为普遍用于主动学习的设置，pool-based AL与基于流的采样的区别在于，pool-based AL在选择最佳查询之前采取贪婪机制，对整个数据集进行比较，但是后者单独接收数据进行评估。但是，当数据集非常大时，选择最佳的元素进行标注可能变得困难或耗时。因此此方法适用于人工贴标签的工程量和成本较大的任务。

查询策略：

本节考虑了主动学习中的多种策略，在实现章节将使用基于ExpBERT的文本多分类模型来理解性能的影响。

随机抽样策略：

随机抽样随机选择实例，既不基于预测，也不基于数据以及模型，通常被用作任务的基准。因为，它可能忽略了潜在地信息，从而降低了学习效率在这种情况下，随机抽样将作为底线与下面提到的更复杂的策略对比，特别是当标记池过于庞大时[29]。

Prediction uncertainty sampling：

不确定性抽样，顾名思义，主动学习者（模型）查询最难以判定其分类的实例。在用概率模型做二分类问题中，这种实例的后验正概率最接近0.5[30]。但是对于更复杂多标签分类问题，将使用基于熵的方法。概率分布越均匀，熵越大，随机变量的不确定性越高，信息量越大。当概率集中在几个数据点，表示不确定性越低，信息量小。



表示在分类的概率分布。

在文本分类领域，对于不确定性的测量，常用一种替代方式，即least confident[31]：



表示可能性最大类标签。对于二值文本分类，该方法等价于基于熵的算法有效性。

除了上面两种较为常用的测量方式之外，（Munro 2020）[13]提到了margin of confidence 以及ratio of confidence这两种方式，前者是两个最自信的预测之间的差异，后者则是这两个预测的比率。



是最自信，是次自信。

Semantic-based diversity sampling：

Peng, Hao, et al.（2023）[30]提出了一种基于语义的多样抽样方式，该方式可应用在文本分类中。与采用不确定性抽样的测量置信度的方式的不同之处在于，基于语义的多样抽样的方式是利用欧氏距离在语义上消除文本样本的冗余。确保在后续过程中提供给模型（学习者）更丰富的，重复度更低的样本。此抽象方式使用（Sener和Saveravarese，2017）[31]的贪婪k中心算法进行聚类操作。数据集包含个无标签的文本，并将分成个批次，每个样本集包含m个实例。是数据集编码的结果。首先选择vectors从中去初始化聚类，这里examples将被认为是聚类中心。然后k-centered算法searches from ,这是一个包括不在中的成员的集合。它是所有聚类中距离中心最远的。选择的算法公式如下：



Where：



随后，通过将现有的聚类更新为。经过循环执并将输出结果合并到中。中的所有文本实例会收敛成一个核心集合，在语义空间中最好地体现并概括数据集。

除此之外的多样本抽样

Bayesian active learning by Disagreement：

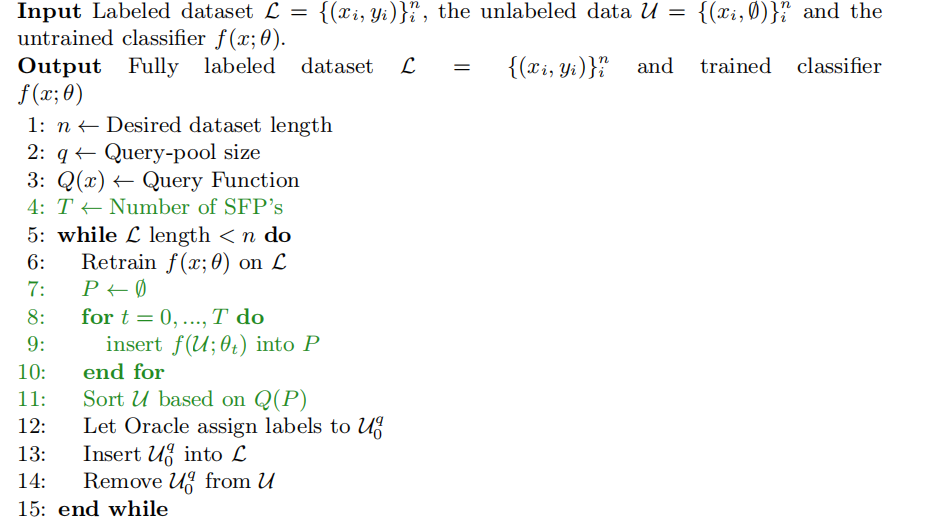
BALD（Bayesian Active Learning by Disagreement）是一种在文本分类主动学习中广泛使用的方法。它基于贝叶斯推断和不确定性度量，旨在选择最具信息量的样本进行标注。在BALD中，首先使用贝叶斯神经网络或其他贝叶斯模型来建模文本分类任务。这些模型能够估计每个样本属于每个类别的概率分布，并提供了关于每个预测的不确定性度量。(Houlsby, N （2011）)[33]提出BALD的关键思想是通过对每个样本的多次预测并计算模型的预测之间的不一致性来选择具有最高不确定性的样本进行标注。学习者（模型）通过输入x最大化模型参数的不确定性，H[y|x,D]表示了目标变量的不确定性，等式第二项表示在参数 θ 服从训练数据集 D 的后验概率分布 p(θ|D) 的条件下，对于 H[y|x, θ] 的不确定性（熵）进行期望值的计算，衡量了平均的不确定性[34]。



通过计算 BALD 的值，可以选择具有最高 BALD 值的样本进行标注，因为这些样本在提供最大信息量的同时也会最大程度地减少模型的不确定性。具体而言，BALD使用信息增益（Information Gain）来衡量每个样本的不确定性，信息增益越大表示样本的信息价值越高。在每个主动学习迭代中，BALD选择具有最高信息增益的样本进行标注，并将其添加到已标注的训练集中。然后，使用更新后的训练集重新训练模型，以提高模型性能和减少不确定性。

AL loop with Monte Carlo Dropout

在传统的dropout中，训练神经网络过程中会以一定的概率将神经元的输出置为零，以减少过拟合。在MCDO中，网络应用了dropout [36] 来使网络具有生成性质。通过执行多个随机前向传递，为相同的输入产生多个输出。这些输出可以用多种方式来总结模型的不确定性。将MCDO方式运用在主动学习循环中的伪代码如下图所示。在MCDO中，通过使用不同的Dropout采样来创建T个稍有差异的模型，从而近似贝叶斯推断。这些所谓的随机前向传播（SFP）的结果可以被查询函数用来计算不确定性，经过排序后将不确定性最高的实例分配给注释者。



Reference

McCreadie, R., Buntain, C., & Soboroff, I. (2019). TREC incident streams: Finding actionable information on social media.

Murty, S., Koh, P. W., & Liang, P. (2020). Expbert: Representation engineering with natural language explanations. arXiv preprint arXiv:2005.01932.

Baram, Y., Yaniv, R.E., Luz, K.: Online choice of active learning algorithms. J. Mach. Learn. Res. 5(Mar), 255–291 (2004)

Prabhu, S., Mohamed, M., & Misra, H. (2021). Multi-class text classification using bert-based active learning. arXiv preprint arXiv:2104.14289.

A. Prest, C. Schmid, and V. Ferrari, “Weakly supervised learning of interactions between humans and objects,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 601–614, 2012.

D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, “A survey of active learning algorithms for supervised remote sensing image classification,” IEEE Journal on Selected Topics in Signal Processing, vol. 5, no. 3, pp. 606–617, 2011.

B. Settles, “Active learning literature survey,” University of Wisconsin, Madison, vol. 52, pp. 55–66, 2010.

Y. Fu, X. Zhu, and B. Li, “A survey on instance selection for active learning,” Knowledge and Information Systems, vol. 35, no. 2, pp. 249–283, 2013.

Schröder, C., & Niekler, A. (2020). A survey of active learning for text classification using deep neural networks. arXiv preprint arXiv:2008.07267.

Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Gupta, B. B., ... & Wang, X. (2021). A survey of deep active learning. ACM computing surveys (CSUR), 54(9), 1-40.

Tomaszewski, J. E. (2021). Overview of the role of artificial intelligence in pathology: the computer as a pathology digital assistant. In Artificial intelligence and deep learning in pathology (pp. 237-262). Elsevier.

Godbole, S., Harpale, A., Sarawagi, S., & Chakrabarti, S. (2004). Document classification through interactive supervision of document and term labels. In Knowledge Discovery in Databases: PKDD 2004: 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, September 20-24, 2004. Proceedings 8 (pp. 185-196). Springer Berlin Heidelberg.

Munro R (2020) Human-in-the-loop machine learning. Manning Publications, Shelter Island.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D. et al. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 56, 3005–3054 (2023). <https://doi.org/10.1007/s10462-022-10246-w>

Settles B (2009) Active learning literature survey. Tech. rep., University of Wisconsin-Madison. Department of Computer Sciences, <https://minds.wisconsin.edu/handle/1793/60660>

Amershi S, Cakmak M, Knox WB et  al (2014) Power to the people: the role of humans in interactive machine learning. AI Magazine 35(4):105–120. <https://doi.org/10.1609/aimag.v35i4.2513>

Dudley JJ, Kristensson PO (2018) A review of user interface design for interactive machine learning. ACM Trans Interact Intell Syst. <https://doi.org/10.1145/3185517>

Michael CJ, Acklin D, Scheuerman J (2020) On interactive machine learning and the potential of cognitive feedback. arXiv e-prints arxiv:2003.10365 [cs.HC]

Ramos G, Meek C, Simard P et  al (2020) Interactive machine teaching: a human-centered approach to building machine-learned models. Hum Comput Interact 35(5–6):413–451. <https://doi.org/10.1080/07370024.2020.1734931>

Devidze R, Mansouri F, Haug L et al (2020) Understanding the power and limitations of teaching with imperfect knowledge. In: Bessiere C (ed) Proceedings of the twenty-ninth international joint conference on artifcial intelligence, IJCAI-20. International Joint Conferences on Artifcial Intelligence Organization, 2647–2654, <https://doi.org/10.24963/ijcai.2020/367>

Active learning：

Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

Dana Angluin. 1988. Queries and Concept Learning. Machine Learning 2, 4 (1988), 319–342.

Ross D King, Kenneth E Whelan, Ffion M Jones, Philip G K Reiser, Christopher H Bryant, Stephen Muggleton, Douglas B Kell, and Stephen G Oliver. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427, 6971 (2004), 247–252.

Ido Dagan and Sean P. Engelson. 1995. Committee-Based Sampling For Training Probabilistic Classifiers. In Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995. Morgan Kaufmann, 150–157.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. (1994), 3–12.

1. Baum and K. Lang. Query learning can work poorly when a human oracle is used. In Proceedings of the IEEE International Joint Conference on Neural Networks, 1992.
2. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Networks, (2014).

Dagan and S. Engelson. Committee-based sampling for training probabilistic classifiers. In Proceedings of the International Conference on Machine Learning (ICML), pages 150–157. Morgan Kaufmann, 1995.

Ozan Sener and Silvio Savarese. “Active Learning for Convolutional Neural Networks: A Core-Set Approach”. In: 6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings. 2018.

Lewis and W. Gale. A sequential algorithm for training text classifiers. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12. ACM/Springer, 1994.

Rong Hu, Brian Mac Namee, and Sarah Jane Delany. “Active Learning for Text Classification with Reusability”. In: Expert Systems with Applications 45.C (2016), pp. 438–449.

Diversity sampling:

Peng, H., Guo, S., Zhao, D., Wu, Y., Han, J., Wang, Z., ... & Zhong, M. (2023). Query-efficient model extraction for text classification model in a hard label setting. Journal of King Saud University-Computer and Information Sciences, 35(4), 10-20.

Sener, O., Savarese, S., 2017. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489.

Bayesian active learning by disagreement

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745.

MacKay, D. (1992). Information-based objective functions for active data selection. Neural computation, 4(4):590–604.

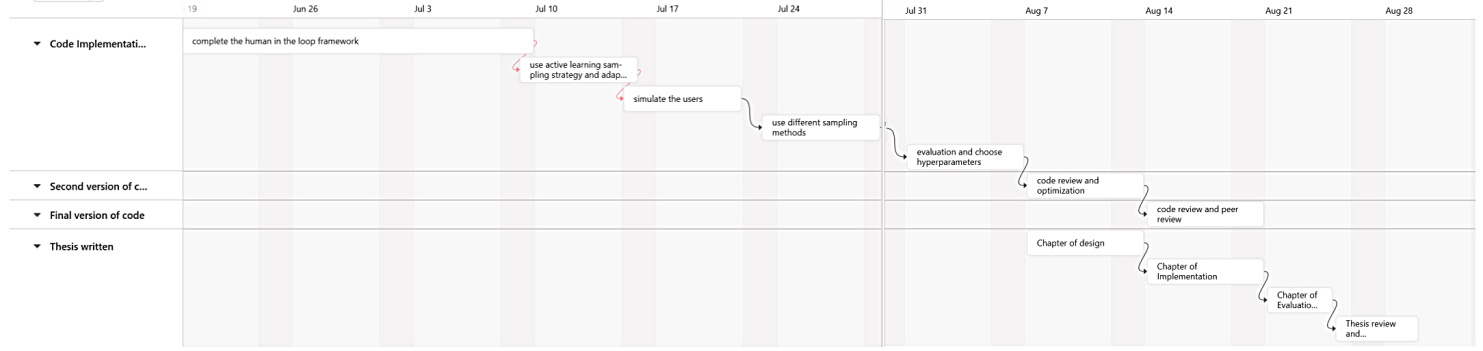
Tsymbalov, E., Panov, M., Shapeev, A.: Dropout-Based Active Learning for Regression. Analysis of Images, Social Networks and Texts pp. 247–258 (2018)

Appendix(time plan & risk assessment)

A Project Timeline

* Week 1(Jun 26 - Jul 9): I will complete the human-in-the-loop framework use active learning
* Week 2(Jul 10 - Jul 16): I will use different sampling strategies of active learning and adapt the method to the explanation generation process with a text classifier.
* Week 3(Jul 17 - Jul 23): I will simulate the users(annotators) to give explanations to unlabeled instances.
* Week 4(Jul 24 - Jul 30): I will use more advanced deep active learning methods as a sampling strategy.
* Week 5(Jul 31 - Aug 6): I will choose the best hyper-parameters and evaluation metrics for evaluation.
* Week 6(Aug 7 - Aug 13): The first version of the code will be complete, and I will carry out the code review and optimize the algorithms; At the same time, I will start writing the Chapter on design.
* Week 7(Aug 14 - Aug 20): After two iterations of the code review, I will execute the final evaluation of the code and peer review; At the same time, the Chapter of Implementation part of my thesis will be written.
* Week 8(Aug 21 - Aug 27): The Chapter on Evaluation and Conclusion will be written.
* Week 9(Aug 28 - Aug 31): Thesis review and peer review.
* Week 10(Aug 31 - Sep 5): Oral presentation will be prepared and recorded.

The visualisation format of my task timeline is as follows:



Include Gantt chart to visually show what you will do.

July 3 - july 9

10-16

17-23

24-30

1. august 6

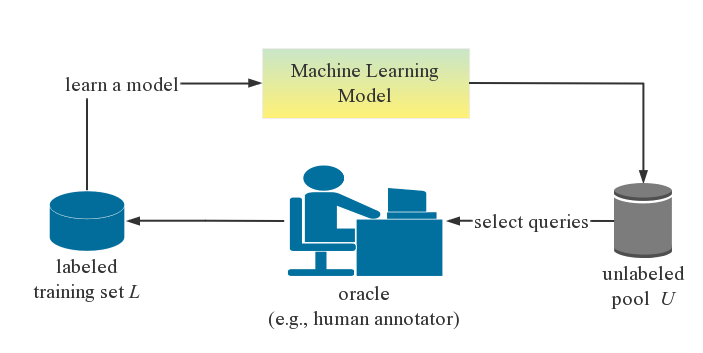
7-13 thesis part one

14-20

21-27

28-31

31-5



It should include a draft abstract: this should fit on one page, and should be a draft of what you hope to be able to say about your thesis when it is complete, several weeks into the future (that is, this is a "fairy-tale" abstract).

It should include at least two pages of coherent text (i.e., of the form you intend to write for your thesis, not rough-notes) that could be used as the opening pages of your Introduction/Overview chapter (Chapter 1 of your thesis).

It should include at least three pages of coherent text forming an initial survey/summary of relevant literature, that could be used as the basis of your Contextual Background or Literature Review chapter (typically Chapter 2 and/or 3 of your thesis).

It should include a Bibliography/References that lists all the literature sources cited in your literature survey, consistently formatted in a commonly-used style (such as APA or IEEE), and with each item in the References being complete, i.e. as you would format it it in your final submitted thesis.

It should include as an Appendix a one-page time-plan for your project, which you may choose to format as a week-by-week bullet-list, or possibly as a Gantt Chart.

It should include as an Appendix a one-page risk assessment for your project, talking about the major risks you can foresee that might plausibly occur and interfere with your plans. For each risk, state clearly what it is, what its likelihood is, what its effects/impact would be on the project, and what your intended mitigation or risk-reduction involves.