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。

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。

本论文的主要结论如下：

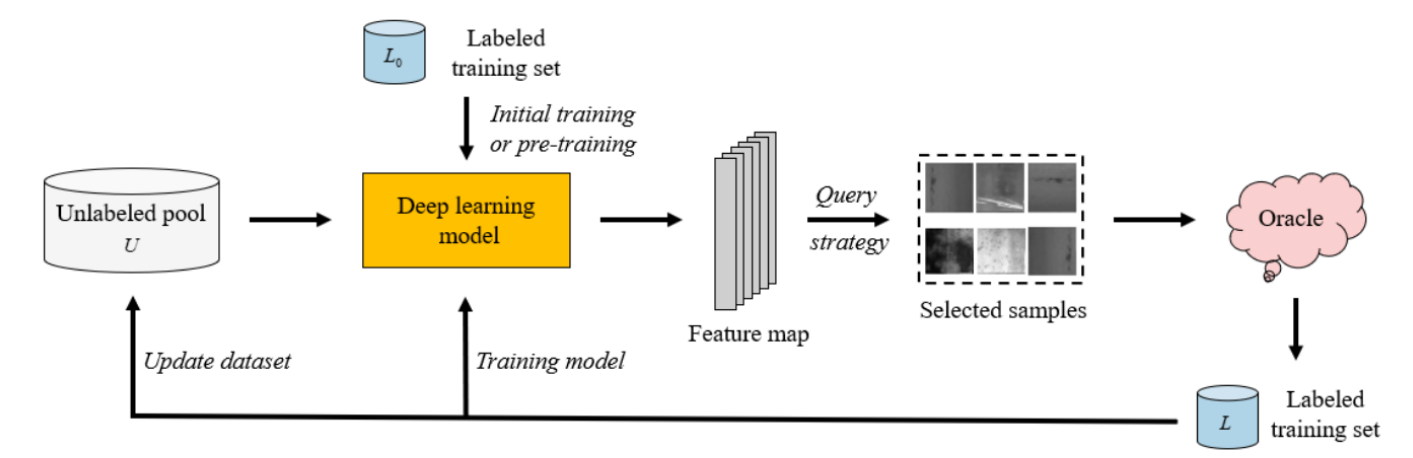
Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum Blah blah blah. Lorem ipsum Lorem ipsum 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]，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，较新的方法使用贝叶斯扩展，通过drop-out实现不确定性估计，或者使用概率NN来测量不确定性，从而完成不确定性抽样或者模型的训练。同时，除了抽样策略方面可以应用贝叶斯原理[10]，模型本身的难以解释性也可以采用贝叶斯深度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。贝叶斯方法能够提供一种统计学的框架，对不确定性进行评估，并利用先验知识和后验推理来增强模型的可解释性。将贝叶斯方法与主动学习相结合，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，贝叶斯主动学习结合了贝叶斯方法、深度学习和主动学习的优势，具有代表性。它能够在大规模未标记数据中提取丰富的特征表示，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，在深度学习中特征学习以及分类器优化同步完成。但是仅仅在主动学习中对深度学习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：

建立基于池的human-in-the-loop主动学习框架，设计中止标准，探究交互系统是否可以提高文本分类模型性能

实现根据不确定性，多样性等不同抽样查询算法并评估其影响，适当集成提取算法

探索优化架构缺陷使用贝叶斯深度学习结合inference方法完成主动学习框架的改善，并探究其是否可以有效改进原有优化架构的性能

建立注释者评估模拟算法，以及真人注释者模拟计划

探究在噪音环境下先进ExpBERT文本分类架构的强健性

挑战：

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。

第一个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2.5节对不确定性算法评估方式以及其优化方法的介绍，最后一节是文本分类框架利用贝叶斯深度学习框架改进的探索性方法的介绍。

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BNN等其他可替代性模型的设计理念。

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

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]。

Active learning

Bayesian active learning

Small-Text active learning

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]

Appendix(time plan & risk assessment)

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.

Much of the focus on emergency time response systems for social platforms has been on creating better text classification algorithms to learn from the data. However, obtaining useful annotated datasets can prove difficult [4]. The throughput of general social platform data is not able to accomplish bulk annotation, so many weakly supervised forms that enable the extraction and annotation of data with high information value at a controlled cost have been widely used in classification projects [5,6]. To reduce the cost of annotation, active learning is used in the backend of social networking sites to accomplish annotation. Active learning is an effective method to solve these problems by selecting unannotated samples with high information content to be annotated by experts [7, 8]. Querying the most informative instances using evaluation algorithms with different sampling is probably the most popular method in active learning. As a result, query strategies have naturally become a hot topic of research in active learning algorithms, with a variety of optimisation algorithms emerging. Putting annotators into an optimisation loop ultimately achieves high accuracy.

At the same time, neural text classifiers are often not comfortable with early uncertainty sampling [9] and are often considered overconfident in their output of results, and ineffective responses to the complexity of NNs continue to be an important area of research. This is because it has no inherent measure of vulnerability. In contrast, newer approaches use Bayesian extensions to achieve uncertainty estimates via drop-out or use probabilistic NNs to measure uncertainty.

Firstly, based on the above background we can conclude that interpretation-based neural text classification models need to be trained using a small but informative number of instances and can exploit the maximum potential of the model parameters during training. Secondly, how to select instances with large information content to supply to human annotators and detect the posterior performance of the instances is also the focus of this experiment. Secondly, due to the labour cost, how to simulate the processing process of human annotators is also an effective way to reduce labour. Finally, the exploration of structural improvements to the neural network text classifier with ExpBERT will also have a benign impact on the retrieval of contingency times on social platforms.

Therefore, the following research objectives are combined with the above issues:

Establish a pool-based human-in-the-loop active learning framework, design stopping criteria and explore whether interactive systems can improve text classification model performance

Implement different sampling query algorithms and evaluate their impact, appropriately integrate extraction algorithms, and for uncertainty in uncertainty sampling determine how uncertainty is evaluated

Development of an annotator evaluation simulation algorithm, and a live annotator simulation scheme

Exploring the robustness of the advanced ExpBERT text classification architecture in a noisy environment

Exploring architectural flaws in the model itself and completing improvements

Thesis Organisation

The overall structure of the thesis based on the research objectives consists of six modules: Introduction, Background, Design, Implementation, Evaluation and Conclusion.

The introduction phase in Chapter 1 provides an overview of the background to the project, presenting the motivation for the experiments in light of the background, followed by a brief description of the approach to the implementation of the overall interactive system, listing the aims of the research and the challenges faced.

In Chapter 2 a detailed introduction to the technology of the overall interactive system and its application will be provided, namely the application of the interactive system to a text classification system (Section 2.1), and Section 2.2 provides an introduction to the framework of active learning. Due to the multitude of example query methods in active learning, section 2.4 presents an introduction to the sampling query algorithm in combination with pseudo-code and formulas as well as an introduction to the way uncertainty algorithms are evaluated and their optimisation methods in section 2.5. The final section presents an improved exploratory approach to the text classification framework.

In the first section of the design chapter the overall framework and the design details of the method are described. Due to the importance of the query strategy, section 3.2 will focus on the query strategy adopted in this report has been designed for the evaluation of the representativeness of the examples, and finally on the design philosophy of other alternative models.

The chapter on implementation will first give an overview of the experimental environment, introduce the elements in the database and their properties etc. in section 4.2, and then describe the process of implementing the simulated user operations. Since the model parameters as well as the sampling strategy are to be evaluated in the experiment, evaluation criteria need to be described in section 4.3 and finally an overview of the development resources is given.

The chapter on evaluation begins with a description of the selection process of the hyperparameters, the presentation and analysis of the evaluation results for the validation and test sets, and finally completes with an evaluation of the effect of noise.

In the final chapter a summary of the previous experiments is presented, followed by an analysis of the limitations of the overall model, and finally a discussion of further enhancements and future work.