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，使用主动学习to optimize the human-in-the-loop 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，期间近似神经网络的贝叶斯推理即Monte Carlo dropout (MCDO)中的随机前向传递(SFP)的结果被应用于查询函数来计算不确定性。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。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。

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的性能对比。其中包括使用随机抽样算法的ExpBERT分类模型，主动学习改进的模型以及优化了不确定性抽样算法的模型，在一个Lorem ipsum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。

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，对于具有ExpBERT的神经网络文本分类器的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。

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，设计停止标准，探究交互系统是否可以提高文本分类模型性能

实现不同抽样查询算法并评估其影响，适当集成提取算法，对于不确定性抽样中的不确定性评估方式的确定

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

探索模型自身架构缺陷并完成改善

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的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4.3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 Review

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.

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.