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]。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。

Problem statement

Research Objective

Thesis Organization

Work architecture

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.

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.