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

Author: Xinyang Song

Supervisor: Dr 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去分析文本并分配tags或者分类。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，而这些数据往往是稀缺的，并且难以获得[1]。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，目前最先进的技术还不够有效，不能被最终用户认为是可部署的。

本项目结合真实以及生成的模拟用户的更正和解释整合到分类器的纠错过程中，探索了一种解决这一限制的替代方法。利用以前的一项技术即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一起来提高分类器的性能[2]，本项目旨在研究这种技术的几个潜在的改进。By incorporating active learning principles, this research endeavors to optimize the human-in-the-loop process for error correction.

本项目的主要目标构建一个不同以往文本分类器的交互系统，通过用户反馈和解释来增强ExpBERT分类器系统的有效性。通过一个迭代过程，使用并对比基于不确定性的抽样策略以及多样抽样，选择的实例将被提交给人类注释者，他们将纠正分类器的错误，并就其纠正选择提供解释。通过这个过程产生的标记数据将被用来微调ExpBERT分类器，反复迭代最终提高其准确性和性能。

这项研究将通过研究active learning结合human-in-the-loop process for error correction的潜在好处，为推动具有解释的文本分类领域的发展以及应急系统的完善作出贡献。By harnessing the power of user explanations and actively involving human expertise, this project aims to address the challenge of limited labelled data availability and improve the effectiveness of text classifiers in real-world applications. 为了评估本方法的性能，我们将设置对比组，将此方式与未改进模型以及其他模型的性能对比。其中包括普通分类模型，带有ExpBERT的模型以及本方法，在一个Lorem ipsum的数据集上进行比较。

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

Problem statement

Research Objective

Thesis Organization

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.

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.