Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Mitchell Borchers
Název práce	Active learning in E-Commerce Merchant Classification using Website Information
Rok odevzdání	2023
Studijní program	Computer Science - Artificial Intelligence
Obor	Computer Science - Artificial Intelligence

Autor posudku	Mgr. Martin Pilát, Ph.D.	**Role**	oponent
Pracoviště	KTIML MFF UK		

Text posudku:

The main goal of the thesis was to design a machine learning model for classification of online merchants into categories and sub-categories. This goal was partially full-filled as the proposed technique is able to classify the merchant to categories only (the sub-categories are not considered in the thesis).

The whole thesis is divided into 5 chapters (plus introduction and conclusion), the first three chapters contain description of active learning, more detailed description of the xPAL method, and discussion of work related to e-commerce merchant classification. These first two chapters are very well-written, with reasonable amount of detail and can definitely be used as introduction to the area of active learning for anyone, who may be interested in the problem in the future. The chapter on related work could however contain more existing approaches – only one paper is mentioned.

The rest of the thesis contains the main contribution. Chapter 4 describes the available data and their pre-processing. Chapters 5 and 6 then contain experiments performed on the dataset and on an extended dataset. While these chapters contain a number of experiments, the writing in this part is weaker than in the introductory chapters and it is often not very clear what and how is evaluated. For example, Section 4.4.1 describes the χ-squared test and then uses it (together with tf-idf) to produce Table 4.1, however it is not very clear, how the test was used to create the table. Figure 5.3 (and others) contain a number of methods that were obviously evaluated, but their performance is never discussed in the surrounding text. Similarly, Section 6.3 describes experiments with text length filtering, however, it is never explicitly mentioned, how this filtering works (what is removed? what is retained?). Additionally, the structure of the text overall could be better. For example, it would be better to describe the LinearSVC in one of the introductory

chapters than including a 2-3 page description in the middle of the discussion about the results of the experiments.

Additionally, given that the topic of the thesis is "active learning", only small part of the experiments actually deals with active learning itself – only roughly two and half pages in Chapters 5 and 6 each. The rest of the experiments compare standard supervised learning techniques on the data.

Overall, from a machine learning point of view, the thesis does not contribute much – it is mostly applications of existing methods to data from a specific domain with the evaluation of the methods. I would expect at least some small modifications of the methods for this specific case. On the other hand, the student did a lot of other related work related to obtaining the data and annotating them. He has demonstrated that he is capable to apply the techniques for real-world problems and evaluate them. Most of my comments above were directed mostly to the structure of the text and not to the application itself. I believe the student did a lot of good work and I recommend the thesis for defense. I have only a few questions:

1. In Section 5.3 it is mentioned that the k-NN classifier works the best for $k = 8$ with cosine distance. How did you obtain these parameters? Was the tuning performed on a validation set? How was this set created?

2. Would it be possible to use some language models for the embedding instead of the tf-idf? Would you expect better results?

3. Will the results of the thesis be used in some practical application?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne June 1, 2023

Podpis: