Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. David Vondrák
Název práce	Rozpoznávání a klasifikace učebnic pomocí hlubokého učení
Rok odevzdání	2022
Studijní program	Informatika **Studijní obor** Softwarové systémy

Autor posudku	doc. RNDr. Pavel Pecina, Ph.D **Role** vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Obsah práce Předkládaná práce Davida Vondráka se zabývá statistickou klasifikací knih, resp. jejich specifického žánru, a to učebnic. V práci jsou řešeny celkem tři úlohy: binární klasifikace knih učebnice/neučebnice, klasifikace učebnic dle předmětu a klasifikace učebnic dle stupně vzdělání. Autor v práci navrhl řešení založené na hlubokých neuronových sítích a aplikoval je na data získaná z veřejně dostupných zdrojů, provedl zevrubnou anotaci, řadu experimentů a srovnání jednotlivých metod/architektur.

Práce je psaná česky, hlavní text je na 65 stranách, rozdělený do sedmi kapitol (včetně úvodu a závěru) a opatřený seznamem literatury, obrázků a tabulek. Po krátkém úvodu a motivaci celé práce se autor v první očíslované kapitole zabývá definicemi základních pojmů, které v práci používá (včetně např. diskuse co všechno musí splňovat kniha, abychom ji mohli považovat za učebnici). Dále pečlivě definuje všechny tři studované úlohy a specifikuje atributy dat, s kterými bude pracovat (název knihy, autor, ISBN, atp.). Všechny atributy jsou textové, převážně v češtině. ¨ Druhá kapitola je věnována metodologii, jednak metodám získávání dat, jednak metodám strojového učení a vyhodnocování experimentů. Třetí kapitola popisuje data použitá v experimentech, jejich získávání a anotaci. Data byla stažena automaticky z několika veřejně dostupných českých webových stránek, pročištěna a sloučena (s ohledem na odstranění duplicit, unifikaci autorů, nakladatelů apod.). Výsledná množina potom obsahovala přes 700 tis. záznamů a z ní byla vybrána podmnožina obsahující 15 tis. unikátních ISBN (10 tis. pro trénování a 5 tis. pro testování), která byla ručně anotována pro všechny tři úlohy.

Kapitola čtvrtá popisuje modely strojového učení použité v experimentech (Naive Bayes jako baseline a několik architektur hlubokých neuronových sítí pro porovnání), jejichž¨ výsledky jsou prezentovány v kapitole páté (včetně detailní analýzy a testů statistické signifikance). Následuje závěr se shrnutím dosažených výsledků a diskusí možností pokračování práce.

Hodnocení Předkládaná práce je výzkumně experimentální. Je vhodně strukturovaná, text přehledný, psaný čtivě (místy snad i příliš upovídaně), v podstatě bez chyb, splňuje všechny náležitosti vědecké práce. Metodologicky je práce také v pořádku, autor provedl dostatečnou rešerši, dobře zvolil způsob řešení i cíle experimentů. I přes provedené testy statistické signifikance se zdá, že v dobře natrénovaných modelech s různou architekturou/konfigurací není prakticky větších rozdílů (což ale nijak nesnižuje kvalitu práce jako takové).

Vyzdvihnout by bylo možné dva body: velmi pečlivou přípravu dat (počínaje definicí pojmu učebnice, přes podrobné anotační pokyny, analýzu mezianotátorské shody, slučování dat z různých zdrojů, detekce jazyka, atd.) a velmi podrobnou analýzu chyb (viz části 5.3.1–5.3.4).

V práci je i několik nejasností, např. ohledně použití párového t-testu pro pro testování statistické signifikance. Z textu v části 2.4.2 se zdá, že byla použita oboustranná varianta, později z textu vyplývá, že test byl jednostranný. Z části 4.1.3 je patrné, že atributy autor a nakladatel, byť jsou přirozeně textové, byly zapojeny jako kategoriální (one-hot reprezentace). Toto rozhodnutí není nijak motivováno a vysvětleno. Pro klasifikaci nových (neviděných) knih (potenciálně z jiných nakladatelství, případně od jiných autorů) by to mohlo být omezující.

Závěr Předkládaná práce Bc. Davida Vondráka zcela splňuje požadavky kladené na diplomovou práci na MFF UK.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 21. 1. 2022

Podpis: