Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Adam Szabó		
Název práce	Low-resource Text Classification		
Rok odevzdání	2021		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Milan Straka	**Role**	Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Hluboké učení dosáhlo v uplynulé dekádě znatelných úspěchů v mnoha oblastech, podstatně vylepšilo také výsledky úlohy klasifikace textu. Nejúspěšnější přístupy založené na modelu BERT navíc využívají přenos znalostí z předtrénovací fáze, což umožňuje dosáhnout velmi dobrých výsledků i s menším množství dat. Cílem diplomové práce bylo prozkoumat závislost úspěšnosti klasifikace textu na množství a kvalitě dat a zkusit navrhnout systém pro klasifikaci smluv z projektu Hlídač státu, ke kterým nejsou k dispozici ručně anotovaná trénovací data.

Práci považuji za zdařilou a aktuální, věnující se aktivně zkoumanému tématu. S použitím dodaného českého předtrénovaného modelu RobeCzech dosáhl řešitel v době odevzdání práce nejlepších známých výsledků na dvou českých datasetech a prokázal tím porozumění aktuálním metodám hlubokých neuronových sítí a zároveň technickou způsobilost k jejich využití, včetně trénování na několika GPU najednou. Řešitel dále vyhodnotil úspěšnost modelů při použití menšího množství trénovacích dat a také při použití trénovacích dat s náhodným šumem – v případě klasifikace příspěvků na sociální síti Facebook do tří tříd (pozitivní, negativní, neutrální) dosahuje model výborných výsledků i v situaci, kdy je 35% trénovacích dat náhodně uniformně zašuměno.

Kromě vyhodnocení existujících datasetů řešitel také navrhuje systém klasifikace smluv z projektu Hlídač státu. K těmto smlouvám nejsou k dispozici ručně anotovaná trénovací data, ale je k dispozici automatické přiřazení smluv do dvojúrovňové hierarchie tříd. Tato automatická anotace je sice chybová, ale výsledky se zašuměnými daty naznačují, že by i tak mohlo být možné natrénovat pomocí nich úspěšnější model. V první řadě student vytvořil a zveřejnil dataset obsahující přibližně 100 tisíc smluv s automatickou anotací. Zároveň získal z projektu Hlídač státu ruční opravy uživatelů, které tvoří testovací množinu více než tisíce smluv s ruční anotací. Dále řešitel navrhl způsob zpracování smluv pomocí technik použitých na výše zmíněných datasetech a provedl větší množství experimentů. Nejlepší vytvořený model pak na vzorku sto smluv ručně porovnal s existujícím automatickým řešením, které vylepšil z úspěšnosti 62% na 69%. Zajímavá otázka pro budoucí výzkum je, zda by nebylo možné popsaný postup iterovat a získat tak ještě úspěšnější řešení.

Práce je psána srozumitelnou angličtinou. Přestože obsahuje více jazykových chyb, než kdyby byla psána česky, tyto chyby nebrání porozumění, takže volbu angličtiny považuji za výhodu, díky které může mít práce i mezinárodní dosah.

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Datum 23. srpen 2021 **Podpis**