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	Mgr. Martin Popel, PhD **Role** oponent
Pracoviště	ÚFAL MFF UK

Text posudku:

Práce zkoumá klasifikaci českých dokumentů na třech datasetech pomocí předtrénovaného modelu RobeCzech a překonává všechny dosud publikované výsledky na dvou z těchto datasetů. Třetí dataset, soubor smluv z webu Hlídač státu, byl vytvořen autorem této práce a publikován v repozitáři LINDAT, což taktéž oceňuji. Autor prokázal, že se v dané problematice dostatečně orientuje a je schopen použít moderní technologie.

Cílem práce bylo zkoumat „low-resource settings“, což je formálně splněno tím, že autor vyhodnotil všechny experimenty na různě velkých podmnožinách trénovacích dat (a navíc zkoumal i vliv množství chyb v trénovacích datech) a zjistil např., že pouze 10 % datasetu „Facebook“ stačilo k překonání výsledků modelu Czert-B (Sido et al., 2021) trénovaného na celých datech. Autor ale nezkoumal, jak se různé architektury modelu či hyperparametry vyrovnávají s menšími daty, a nezkoumal ani vliv množství textů použitých k předtrénování modelu. Proto bych low-resource neoznačil za hlavní téma práce.

Text je přehledně členěn, ale obsahuje poměrně dost gramatických chyb a stylistických neobratností, které někdy ztěžují porozumění, někdy jsou úsměvné (např. *you can also attend to yourself*). Objevují se i faktické chyby (možná překlepy) jako soft-attention místo self-attention (str. 15) a nepřesné formulace (např. *the positive posts are related only to perfume or ZOO pages*, což je se slovem *only* nepravdivé), které jsou nejspíš způsobeny snahou parafrázovat původní (citovaný) zdroj a vyhnout se přímé citaci v uvozovkách.

Oceňuji nápad na konstrukci test setu pomocí uživatelských požadavků na změnu kategorie. Tyto kategorie by tedy jednak měly být spolehlivější (ověřené lidmi) než na základě klíčových slov, jednak by to měly být právě ty případy, kdy automatická detekce dle klíčových slov (použitá pro značkování trénovacích dat) nefunguje. Takovýto test set bude zřejmě „záludnější“ než náhodný vzorek neviděných smluv (s ručně přidělenými kategoriemi). Kladně hodnotím diskuzi na stránkách 56–57, která tuto záludnost rozebírá, a manuální evaluaci na straně 59.

Z popisu datasetu smluv z Hlídače státu není zřejmé, zda má train set a test set nějaký průnik. Data jsem si stáhl a zjistil dle `idVerze`, že z 1024 testovacích smluv je 100 obsaženo i v trénovacích datech. To se obvykle bere jako metodologické selhání (či dokonce podvod, není-li to zmíněno). V tomto případě se ale zřejmě liší kategorie (což už jsem nekontroloval), tedy smlouvy z průniku spíš (uměle) snižují naměřenou úspěšnost. Každopádně by průnik měl být zmíněn v popisu datasetu na LINDATu. V textu práce to také mělo být zmíněno a diskutovány důsledky, např. porovnána úspěšnost na těch 100 vs. zbylých 924 smlouvách. Taktéž mohly být diskutovány důsledky tohoto průniku na korelace úspěšnosti na dev a test setu (Table 4.11).

V diskuzi této korelace chybí připomenutí, že dev set je vyvážený dle kategorií, kdežto test set ne, přičemž zvolená metrika Accuracy je na toto citlivá. Vzhledem k tomu, že úspěšnost klasifikace se liší napříč kategoriemi (viz Figure 4.9), může být toto dalším důvodem rozdílu mezi výsledky na dev a test setu. Pro hlubší porozumění by bylo vhodné ukázat (asi jako confusion matrix), jaké změny uživatelé navrhovali, a zda například nepřevládají změny z *IT* na *Science, research and development*, což by mohlo vysvětlit záměny tohoto typu ve Figure 4.9.

Otázka k obhajobě:

When using several windows at the same time during training, all windows are classified separately, and the result category is determined according to the average of the predictions of the individual windows. [str. 53]

Předpokládám, že stejný postup používáte krom trénování i pro predikci. Proč průměrujete až výsledky klasifikace jednotlivých oken a nikoli vstupy do poslední vrstvy (před softmax)?

Podrobnější připomínky:

- *They are known as low-resource languages and Czech is one of them.* [str. 3]
  Osobně bych češtinu neoznačil za *low-resource language*. Při hodnocení, zda je jazyk *low-resource*, je vhodné rozlišovat, jaký typ jazykových zdrojů (a pro jaký účel) máme na mysli. Pro téma práce je relevantní velikost textů pro předtrénování a velikost datasetu pro trénování (fine-tuning) dané úlohy. Neanotovaných textů jsou pro češtinu k dispozici desítky miliard slov, třeba samotný SYNv4 (citovaný v sekci 3.4) s 4GW je dokonce větší než data použitá k (před)trénování anglického modelu BERT (English Wikipedia+BookCorpus). Ani u jednoho z tří datasetů použitých v práci si nejsem jist vhodností označení *low-resource*. I v mnoha oblastech mimo zaměření práce (např. paralelní korpusy, treebanky) je pro češtinu k dispozici více dat než pro mnohé jazyky s větším počtem mluvčích.

- Popis Feed Forward Networks na str. 8 je přinejmenším zavádějící. Je opominut důležitý fakt, že se jedná o *position-wise* FF layer, tedy že x není celá sekvence, ale jen jedna pozice

(slovo). Také mi není jasné tvrzení *a connected ReLU layer with four times as many hidden units as inputs, followed by another fully connected layer without activation.*

- *To pre-train the BERT model, we do not use...* [str. 12]
  Vzhledem k tomu, že jste v rámci práce žádný model nepředtrénoval, ale jen použil existující předtrénovaný model, bylo vhodnější vyhnout se v této větě první osobě (či na to jasně čtenáře upozornit).

- Délkám příspěvků je sice věnována celá stránka 17, ale není zřejmé, jak uvedené údaje souvisejí s tématem práce, tedy zda třeba délka nějak ovlivňuje úspěšnost klasifikace. Obdobně je sice upozorněno na velmi rozdílné procento pozitivních příspěvků napříč 9 zdrojovými stránkami, ale nejsou diskutovány důsledky: Nefunguje pak sentiment analysis jen jako klasifikace zdroje stránek stránek? Jak by dopadla evaluace na test setu, kde by z každé stránky bylo stejné množství pozitivních a negativních příspěvků?

- Typ grafu použitý ve Figure 2.6 (a Fig 2.3, 2.12 a 2.15) nepovažuji za vhodně zvolený. Zdá se, že nebyla použita žádná kvantizace (bucketing). Přitom pravděpodobnost, že dva dlouhé dokumenty budou mít přesně stejný počet tokenů, je nízká. Taktéž logaritmické měřítko osy y nepovažuji v tomto případě za vhodné (zejména u Fig 2.15).

- *The corpus contains a total of 3 505 965 words, of which 50 899 are unique words, and 82 986 are unique lemmas.* [str. 19]
  Chápu, že korpus má velikost 3.5M slov, ale ty další údaje mi nejsou jasné. Čekal bych informaci o tom, kolik různých slov (slovních tvarů) a různých lemmat v korpusu je. Tomu ale neodpovídá výraz *of which* a taky, že unikátních lemmat je více[1] než unikátních slovních tvarů. Lze uvést též počet hapaxů, ale to též neodpovídá, protože lemmat vyskytujících se právě jednou by opět nemělo být více než slovních tvarů vyskytujících se právě jednou.

- *fit into one BERT window of 512 tokens, except for a negligible number of documents* [s. 20]
  Dle Figure 2.6 odhaduji, že množství dokumentů delších než 512 tokenů je nezanedbatelné. Každopádně by bylo lepší konkrétní procento delších dokumentů uvést.

- *In the Figure 2.11 we can clearly see that the contracts are evenly distributed relative to the categories.* [str. 28]
  Toto není pravda (jak dokazuje Figure 2.10). Z citovaného tvrzení i z popisku Figure 2.11 vypadla podstatná informace, že se jedná o distribuci *délek* smluv (v tokenech). Obdobně tato informace chybí i v popisku Figure 2.2. Naštěstí je to u obou obrázků uvedeno u osy y.

[1] V případě homonym to sice může nastat – např. kdyby byly v korpusu dva výskyty „stát“, jeden byl slovesem a druhý podstatným jménem, tedy každý měl jiné lemma – ale statisticky by měly převážit opačné případy, kdy různé slovní tvary mají totéž lemma.

- *However, despite these few fluctuations, the histogram tends to have a normal distribution.* [str. 32]
  Opravdu? Figure 2.15 má totiž na ose x logaritmické měřítko, bez kterého by histogram vypadal jinak (nesymetricky).

- *more than 340 000 sentences* [str. 37]
  Abstrakt Sido et al. [2021] sice taky uvádí *more than 340K of sentences*, ale to je zjevně překlep (jak dokazuje Table 2 v daném článku) – má to být *340M sentences*. Teprve pak dává smysl následující věta (*really large, as it is 50 times more compared to...*).

- *We use early stopping as a regularization technique to prevent pre-training of the model* [str. 44]
  pre-training → overfitting (či over-training)

- Ve Fig 4.4 a 4.5 u MaxEnt a Czert-B chybí informace, že se jedná o Macro-F1 (nikoli accuracy).

- Jsou metody porovnávané v Table 4.4 trénované na stejných datech a testované na stejných datech? Zřejmě ne, když jen RobeCzech používá 10-fold cv. Je tedy aspoň velikost těch dat stejná?

- *However, it is unfair to compare with Hlídač Státu on the test set, because there are mainly those contracts that they classified incorrectly, and in some sense this type of contracts was difficult for them.* [str. 55]
  Ano. Proč ale *mainly*? Vždyť test set měl být pouze z dokumentů, kde uživatelé navrhli změnu kategorie? To by v Table 4.10 u „Hlídač Státu“ mělo být accuracy na *all categories* pro test set 0 %, ale je tam 17.87 %.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 23. 8. 2021

Podpis: