Posudek oponenta diplomové práce

Student: Klocok Andrej, Bc.
Téma: Analýza recenzí výrobků (id 22451)
Oponent: Doležal Jan, Ing., UPGM FIT VUT

1. Náročnost zadání — průměrně obtížné zadání

2. Splnění požadavků zadání — zadání splněno

3. Rozsah technické zprávy — je v obvyklém rozmezí

Rozsah technické zprávy je v obvyklém rozmezí.

V teoretické část je někdy uváděno zbytečně více informací (student zachází zbytečně do hloubky), které nejsou dostatečné k pochopení a ani nejsou důležité dále v technické zprávě. Tyto informace by se čtenář lépe dozvěděl z odkazovaného zdroje.

4. Prezentační úroveň předložené práce — 60 b. (D)

Rovnice uvedené v teoretické části jsou často nedostatečně popsány v textu (např. na str. 11-12), a i když jsou číslovány, jsou odkazovány nešťastně či vůbec. Příklady takových nešťastných odkazů na rovnice je možné nalézt u podnadpisu GloVe na str. 15-16. Např. odkaz na rovnici (2.13) je uveden na konci věty takto: "... dopracovať ku rovnici, ktorá stojí za GloVe 2.13." Dále např. v těle u podnadpisu TF-IDF na str. 13 je uvedena rovnice "TF_i = frekvencia term_i v dokumente", která není odkazovaná z textu.

Zkratky nejsou vždy vysvětleny při prvním výskytu (např. str. 10). Úroveň nadpisů je matoucí (např. nadpis Word2vec je na stejné úrovni jako následující napdisy CBOW, Skip-gram a GloVe, kde CBOW a Skip-gram patří logicky k pod kapitolu Word2vec).

Odkazování na kapitoly pouze pomocí čísla kapitoly (bez doprovodného textu) shledávám jako nevhodné (zvláště pro tištěnou verzi). Např. když se student na str. 28 odkazuje na kapitolu 2.2 způsobem *"V rámci teoretického rozboru bol zmienený systém Elasticsearch 2.2, ktorý sa javí ako najlepšie riešenie."*, tak čtenáře může napadnout otázka: *"Proč chce používat verzi 2.2, když aktuální verze je 7.8?"*

Student se opakuje. V návrhu uvádí to, co měl uvést a nebo již uvedl v teorii. Např. v teorii str. 19-20 a v návrhu str. 29.

Zdá se, že student nemá jasno v pojmech "přetrénování" a "předtrénování". Lze nalézt v teorii na str. 19-20 a zvláště pak str. 29, kde tyto dva pojmy uvádí různě na jedné straně.

V kapitole *3.10 Vizualizácia* mi scházel obrázek návrhu (wireframe či mockup).

Snímky výsledného produktu by bylo vhodné popsat přímo v obrázku (např. místo odkazu na popisující obrázek B.6 v příloze na str. 52, který je v angličtině).

5. Formální úprava technické zprávy — 70 b. (C)

Formální úprava technické zprávy je průměrná. Kresby vytvořené studentem jsou bitmapové a text v nich je psán malým bezpatkovým písmem (písmo se odlišuje od zbytku práce). Jazykovou stránku práce nemohu zcela posoudit, protože nemám potřebné znalosti slovenského jazyka.

6. Práce s literaturou — 80 b. (B)

Odkazy na zdroje nejsou v teoretické části vhodně umístěné. Vadí mi to např. při uvádění výpočetní složitosti nástrojů, kde by bylo vhodné vědět (prostřednictvím odkazu na zdroj), kdo a jak nástroj testoval. Některé zdroje (např. zdroj [4]) měly být spíše umístěny do poznámky pod čarou. Naopak např. poznámka pod čarou 4 na str. 7 mohla být uvedena jako zdroj dle citačních norem (s datem poslední změny a citace). Citace webových zdrojů nejsou ve formátu shodném s normou (dle normy se URL uvádí až za ISBN/ISSN).

Student občas používá nepodložená tvrzení:

- (str. 3) "Zákazníci radi zdieľajú svoje skúsenosti a názory na jednotlivé produkty a služby." - schází statistika
- (str. 31) "V mnohých publikáciách sú vytrénované modely klasifikácie sentimentu, ktoré sú doménovo špecifické." - schází alespoň příklady publikací

7. Realizační výstup

85 b. (B)

Uživatelské rozhraní klientské části výsledného produktu je v anglickém jazyce, zatímco zobrazená data jsou v českém jazyce.

Zdrojové kódy realizačního výstupu jsou dostatečně okomentovány, ale uvítal bych ještě větší množství komentářů. V části review_analysis-backend lze nastavit adresu a port serveru pouze ve zdrojovém kódu.

8. Využitelnost výsledků

Jedná se o práci kompilačního charakteru. Student vytvořil nástroj pro stahování recenzí z webového srovnavače Heureka, kterým vytvořil databázi názorů na výrobky. Dále natrénoval neuronovou síť na odfiltrování irelevantních názorů. Tato databáze je využitelná pro další výzkum a experimenty v rámci výzkumné skupiny KNOT.

Uživatelské rozhraní pro vizualizaci výsledků je s menšími obtížemi použitelné a myslím si, že není vhodné pro běžného uživatele.

9. Otázky k obhajobě

1. V technické zprávě píšete (pod napisem "Predspracovanie dát" na str. 27-28):
   """
   Síce aktuálne riešenia mapovania sekvencií do vektorového priestoru využívajú vlastné tokenizery, ako napríklad kúsky slov, je vhodné tieto dáta tokenizovať, lemantizovať, poprípade previesť do kmeňového tvaru (stem), odstrániť stop slová, pre ďalšie spracovanie.
   """,
   ale už nepíšete, proč je vhodné data tokenizovat vlastním způsobem. Mohl byste toto objasnit?
2. Na str. 29 zmiňujete pojmy "pretrénovanie" a "pred-trénovanie" v tom samém významu, což považuji za chybu (tato chyba se vyskytuje vícekrát, domnívám se tedy, že se nejedná o překlep). Mohl byste vysvětlit pojmy "přetrénování" (over-training) a "předtrénování" (pre-training) v kontextu neuronových sítí?

10. Souhrnné hodnocení

78 b. dobře (C)

Technická zpráva se hůře čte. Obsahuje množství chyb které se opakují, některé informace nejsou důležité pro pochopení realizace a některé naopak schází. Na druhou stranu je práce zajímavá a zpráva informačně bohatá. Z těchto důvodů navrhuji hodnocení stupněm C (78 bodů).

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 29. června 2020

Doležal Jan, Ing.
oponent