FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Ing. Ondřej Pražák
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd
Technická 8, CZ-30614 Plzeň
Tel.: (+420) 377 63 2151
E-mail: ondfa@ntis.zcu.cz

Posudek oponenta bakalářské práce

Jan Pašek
Learning of Sentence Encoding by Using Duplicate Questions from Stackoverflow

Cílem bakalářské práce Jana Paška bylo získat informace o duplicitních otázkách ze serveru Stackoverflow, vytvořit z nich datovou sadu a na této datové sadě natrénovat model hlubokého učení, který se bude učit sémantické reprezentace vět na úloze detekce duplicitních otázek. Dále se práce zabývá zpracováním úryvků kódu otázkách tak, aby se model byl schopný učit sémantické reprezentace kódu v programovacích jazycích a tyto informace zároveň použít k lepší detekci duplicit.

Práce je psaná v Anglickém jazyce velmi srozumitelnou formou a velice dobře se čte. Práce má dobrou logickou strukturu. Ač autor v teoretické části nezachází do přílišných detailů, je z textu patrné, že problematice velice dobře rozumí a detaily vynechává spíše proto, aby nebyl výsledný text příliš rozsáhlý. Je škoda, že se autor v teoretické části nevěnuje trochu více konkrétním modelům, které v praktické části používá (LSTM) a na druhou stranu popisuje několik metod, které vůbec nepoužil (například BERT). Také autor v některých případech sáhodlouze popisuje matematický vztah, který by byl jasný z jednoduchého vzorce (příklad). Práce také obsahuje několik drobných terminologických nepřesností. Kromě výše zmíněných drobných výhrad mi v práci nic nechybí, ani nepřebývá.

V práci se vyskytuje naprosto minimální množství chyb a překlepů. Její formální úroveň je vysoce nadstandardní. Grafickému zpracování dokumentu nelze vůbec nic vytknout.

Student ve své práci cituje velké množství relevantní zdrojů, jak tištěných, tak webových. Citované zroje jsou aktuální a popisují současné state-of-the-art metody.

I přes pár výše zmíněných připomínek považuji text práce za velice nadstandardní, zejména s ohledem na to, že je psán velice dobrou Angličtinou.

V praktické části student navrhl několik architektur neuronových sítí, implementoval je v nástroji *Tensorflow* a provedl poměrně velké množství experimentů pro jejich vyhodnocení. Z kódu je patrné, že autor použité technologie dobře ovládá. Metodice provedených experimentů nelze vůbec nic vytknout. Jelikož byly experimenty provedené na nově vzniklé datové sadě, není možné přímé srovnání se state-of-the-art. Práce však obsahuje srovnání s baseline modelem a dále porovnání se state-of-the-art na příbuzném SNLI datasetu. Všechny uvedené závěry jsou validní a jsou podložené relevantními daty. Dosažené výsledky vypadají velice dobře a práce je rozhodně vědeckým přínosem.

Práce splňuje zadání ve všech bodech. Přes výše zmíněné výtky práci považuji za vynikající. Svým rozsahem, obsahem i kvalitou odpovídá spíše diplomové práci.
Proto práci doporučuji k obhajobě a hodnotím klasifikačním stupněm

„výborně“.

Doplující otázky:

1. V práci uvádíte minimální počty výskytů slov (50 pro textové representace, 500 pro kód). Tyto počty jsou poměrně vysoké ve srovnání s běžnou praxí. Jak veliké jsou výsledné slovníky?
2. Kolik textu (například počet slov, znaků) obsahuje textová část vytvořené datové sady?
3. V matici záměn pro klasifikaci do tří tříd jsem si všiml, že Váš model rozpoznává negativní třídu se stoprocentní úspěšností. Proč myslíte, že je celková úspěšnost na klasifikaci do tří tříd nižší než u klasifikace do dvou tříd?

Ing. Ondřej Pražák
(oponent BP)

V Plzni 31. května 2020