

## Posudek oponenta bakalářské práce

**Student:** Janko Pavel  
**Téma:** Doporučování filmů na základě uživatelských profilů ČSFD (id 22134)  
**Oponent:** Šústek Martin, Ing., UITS FIT VUT

**1. Náročnost zadání** **obtížnější zadání**  
Zadání patří dle mého názoru k obtížnějším, jelikož se zabývá doporučovacími systémy, což není náplň žádného předmětu na FIT.

**2. Splnění požadavků zadání** **zadání téměř splněno**  
Zadání v 5. bodě požaduje porovnat míru relevance doporučení s existujícími řešeními, o čemž se práce vůbec nezmiňuje. Jelikož však zadání požaduje vytvoření nové datové sady, tak mi není jasné, jakým způsobem by výsledky měly být porovnány.

**3. Rozsah technické zprávy** **je v obvyklém rozmezí**

**4. Prezentační úroveň předložené práce** **25 b. (F)**  
K logické struktuře práce mám výhrady pouze kvůli zařazení popisu klasifikace a regrese do Kapitoly 3 o dolování z dat, zařazení neuronových sítí pod doporučovací systémy a nadměrnému využívání nečíslovaných nadpisů.

Prezentační úroveň práce je však velmi špatná. Příkladem pro komisi může být Sekce 2.1 o neuronových sítích. V celé práci se objevují nepřesná nebo chybná tvrzení, což je však způsobeno tím, že student látku nepochopil, ale pouze přebírá či překládá uvedené informace z literatury, které v některých případech interpretuje svými slovy.

Jelikož student přebírá různé části z různých zdrojů, v práci se objevují různé symboly s totožným významem. Student často uvádí pojmy, které v práci přímo nevyužívá (např. rekurentní a konvoluční neuronové sítě), což však v kombinaci s tím, že tyto pojmy nejsou v práci vysvětleny, ale pouze popsány, způsobuje, že čtenář, který daným pojmům ještě nerozumí, bude zmatený. Jiným příkladem může být uvedení kapitoly o dolování z dat a i přesto, že je v této kapitole pojem dolování z dat vysvětlen, student dále tvrdí, že vydoloval data, i když došlo pouze k jejich stažení.

Logická návaznost některých částí není zřejmá, příkladem může být část K-nejbližších sousedů na straně 10, kdy student vůbec nevysvětluje, z jakého důvodu sem tento algoritmus zařadil, navíc vůbec tento algoritmus nepopisuje, ale zaměnil ho s algoritmem k-means.

Více než polovina všech matematických vztahů v sobě obsahuje chybu nebo chybí popis některých použitých symbolů. V práci se taktéž objevují některá tvrzení, které vyložene mystifikují neznalého čtenáře jako je záměna L1 a dropout regularizace nebo "v případě, že datová sada obsahuje malo dat, tak riziko přetrénování není tak vysoké".

Student často nepoužívá zažité překlady, ale vytváří nové nebo přebírá nestandardní, např. **vložky** (embeddings), **vytrénování** modelu.

Některé obrázky (např. Obrázek 2.1, 2.4 nebo 3.1) nemají dostatečné rozlišení a do závěrečné práce se tak nehodí.

V textu se vyskytují grafy, není však specifikováno, jestli se jedná o experimenty s trénovací, testovací nebo validační sadou.

Práce porovnává "subjektivní" a "objektivní" přesnost, obě metody však využívají jiné metriky (RMSE a MAE).

**5. Formální úprava technické zprávy** **75 b. (C)**  
V práci se vyskytují obvyklé jazykové nedostatky, v porovnání s ostatními pracemi se však jedná spíše o podprůměrný počet chyb.

**6. Práce s literaturou****60 b. (D)**

Student vycházel z více než 40 literárních pramenů, z nichž je většina relevantní, není mi však jasné, proč se práce odkazuje na článek "Deep neural networks for acoustic modeling in speech recognition". Informace jsou řádně odcitovány, avšak nepřesně prezentovány, což hodnotím velmi negativně.

**7. Realizační výstup****75 b. (C)**

Výsledná aplikace je funkční a doporučování je založeno na existujícím modelu. Pro určení hyper-parametrů modelu bylo využito experimentů, jejichž výsledek je zřejmý osobě pohybující se ve strojovém učení, což však nepokládám za chybu studenta. Kromě modelu bylo vytvořené také GUI a skripty pro stahování dat.

**8. Využitelnost výsledků**

Práce využívá existující model s využitím nově vytvořené datové sady, doporučovací systém lze využít v praxi.

**9. Otázky k obhajobě**

- Kolik procent z celkového počtu uživatelů a filmů ČSFD je v práci využito?
- Proč byli vybráni uživatelé, kteří hodnotili 50 až 5000 jakýchkoli filmů a nevybírali se na základě hodnocení použitých 4350 filmů?
- Jakou chybu měl použitý model a jakou chybu by měl triviální model, který by předpovídal, že každý uživatel ohodnotí každý film 4 hvězdičkami? Jak souvisí velikost chyby tohoto triviálního modelu s použitou datovou sadou?

**10. Souhrnné hodnocení****50 b. dostatečně (E)**

Realizační výstup odpovídá bakalářské práci. Text práce má však mnoho nedostatků. Přihlížím ke složitosti některých uvedených pojmů a celkově hodnotím práci 50 body (stupněm **E**). Domnívám se však, že práce je na hranici přijatelnosti a doporučuji členům komise zvážit závažnost uvedených nedostatků.

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

V Brně dne: 31. května 2019

Šustek Martin, Ing.  
oponent