Posudek oponenta bakalářské práce

Autor práce: **Milan Vacek**

Název práce: **Named Entity Recognition in Historical Text Documents**
(Rozpoznávání pojmenovaných entit v historických textových dokumentech)

Cílem bakalářské práce Milana Vacka bylo provést sadu experimentů nad českým datasetem pro úlohu rozpoznávání pojmenovaných entit (NER). Jedná se v podstatě o klasifikační úlohu, na kterou dle zadání měly být použity neuronové sítě.

Domnívám se, že zadání svoji formou a rozsahem by bylo vhodnější na diplomovou práci. Průvodní dokument práce je psán v anglickém jazyce, což nepovažuji za šťastné vzhledem k množství gramatických chyb. Chyby jsou patrné po pouhém prolistování několika stran, nacházejí se v anglickém i českém abstraktu a pro autora to rozhodně není dobrá vizitka. Přes velmi špatnou angličtinu lze ale ve většině případů poznat, co chtěl autor říci.

Vyjadřovací schopnosti autora **nejsou na dobré úrovni.** Jednotlivé kapitoly, ale především sekce v rámci kapitol nejsou příliš dobře provázané a také mi schází informace o tom, co zrovna daná kapitola popisuje a jaký je její účel.

Za nejlepší kapitolu v práci považuji kapitolu č. 3 Relevant Existing Methods, která z mého pohledu pěkně popisuje „Related Work" zadané úlohy. **Práce s literaturou není rozhodně špatná.**

Ke zbývajícím kapitolám v práci mám ale **poměrně závažné výhrady.** Pro úlohu byl zvolen model hluboké neuronové sítě založený na kombinaci konvolučních a rekurentních vrstvách. Autor předpokládá od čtenáře již značnou pokročilou znalost problematiky neuronových sítí.

Teoretická část práce by si rozhodně zasloužila více stran než 5 stran a to i v případě, že by autor použil jiný a jednodušší klasifikátor. V celé práci jsem navíc nenašel ani jednu zmínku o algoritmu zpětné propagace chyby (tzv. Backpropagation), který je pro učení neuronové sítě zcela zásadní.

Za velmi odbytou považuji sekci 3.3, která s největší pravděpodobností popisuje analýzu výběru metody pro řešenou úlohu. Myslím, že analýza by v práci zasloužila více než jen jeden větší odstavec a navíc svou volbu metody autor opírá o výsledky dosažené jinými autory na jiných datových sadách v jiném jazyce.

Největší část práce tvoří experimenty doplněné o grafy. Popis experimentů je popsán relativně dobře, i když mi schází některé informace (např. konkrétní podoba vstupních dat) a pro plné pochopení je nutné podívat se do konkrétních zdrojových kódů. **Pořadí experimentů celkem dává smysl,** kdy je v úvodní části patrna snaha o hledání optimálních hyperparametrů a v dalších částech se postupně zkouší testovat obě popsané datové sady. **Vhodné bylo rovněž pustit experimenty několikrát** a jejich výslednou hodnotu zprůměrovat. V takovém případě bych ale do tabulek s výsledky zahrnul i chybu měření (např. směrodatnou odchylku či rozptyl vypočítaných hodnot).

Od úvodu po závěr práce obsahuje 34 stran a počet **normostran dle počtu znaků práce je 22.** Přihlédneme-li k faktu, že podstatnou část kapitoly 6 tvoří 18 grafů je rozsah práce s hlediska množství textu na hraně. Po přesunutí všech grafů do přílohy, počet stran od úvodu po závěr klesne pod 30 na 26.

V práci mi schází obrázky, diagramy, které by mnohdy přispěly k většímu porozumění a také by měly příznivý vliv na rozsah a vizuální podobu práce. V práci jsem rovněž nalezl několik nepřesností, které dle mého názoru pramení z nepochopení problematiky. Z mého pohledu by student udělal lépe, pokud by zvolil jednodušší klasifikátor, důkladně ho teoreticky popsal včetně obrázků a provedl experimenty.

Celá práce na mě působí dojmem, že by si zasloužila více péče a hlavně lepší prezentaci v doprovodném textu. Věřím, že se autor snažil, ale na výsledném dokumentu to bohužel není vidět.

Přesto všechno si myslím, že všechny body zadání byly **(s většími či menšími) výhradami splněny.**

Dotazy k práci

1. V sekci „4.1.1. *Convolutional Neural Network*" tvrdíte, že konvoluční filtr má čtvercový tvar a je složen z jedniček a nul. Váhy neuronové sítě se ale volí nejčastěji náhodně. Používáte nějaký speciální filtr a mění se jeho váhy v průběhu učení?

2. Ve všech grafech prezentujete při hledání optimálních hyperparametrů F-míru. Není ale z popisu poznat, na jakých datech je experiment prováděn. Modré křivky představují vývoj F-míry na trénovací, validační (development) nebo testovací části dat?

3. Mohl byste schématicky uvést přesnou podobu vstupů do vašeho klasifikátoru? Jsou vstupy něčím omezeny (např. délkou sekvence)?

Navrhuji hodnocení známkou **dobře** a práci doporučuji k obhajobě.

V Plzni 25.5.2020 Ing. Jiří Martínek