FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Ing. Pavel Přibáň
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd
Technická 8, 301 00 Plzeň
Tel.: (+420) 377 63 2151
E-mail: pribanp@ntis.zcu.cz

Posudek oponenta bakalářské práce

Vojtěch Bartička
Automatické stahování smluvních podmínek z webových stránek

Obsah Práce
Práce se zabývá možnostmi automatického stahování smluvních podmínek z internetových stránek s cílem vytvořit dataset smluvních podmínek, který bude dále použit v projektu na KIV. Celá práce je výzkumného charakteru.

V první části práce je krátce popsána teorie pro stahování webových stránek (tzv. *crawling*) a následná extrakce textu. Následuje popis umělých neuronových sítí a základních pojmů, které s nimi souvisejí. V textu je správně popsána architektura konvoluční neuronové sítě, ale chybí mi zde alespoň základní matematický popis. Například v textu je uvedeno, že je aplikován tzv. *kernel* na vstupní data a výsledkem je skalár, ale není popsáno jakým výpočtem je tento skalár získán. Podobně u metod pro sémantickou reprezentaci textu (např. word2vec) bych očekával jejich matematické vyjádření.

Ve vzorečku 3.9 pro výpočet kosinové podobnosit na str. 10 chybí ve jmenovateli při výpočtu normy vektoru umocnění na druhou. Mezi metodami reguralizace neuronových sítí chybí jakákoliv zmínka o L1 nebo L2 reguralizaci.

Ve druhé části práce je popsán návrh a implementace základního postupu (programu) pro stahování a klasifikaci stránek pomocí klíčových slov do tří tříd podle jejich obsahu (*podmínky*, *ochrana dat* a *nerelevantní*). Dataset vytvořený tímto základním postupem je pak použit pro natrénování navržené konvoluční neuronové sítě. V rámci práce byly manuálně vytvořeny datasety pro ověření úspěšnosti klasifikace. Autor správně identifikuje problémy, které vznikají během trénování konvoluční neuronové sítě a také je správně eliminuje standardními postupy (např. dropout nebo redukce počtu parametrů modelu).

V textu na str. 29 je uvedeno, že byly použity předtrénované slovní vektory (fastText), ale není uvedeno odkud tyto vektory pochází (chybí reference nebo odkaz). Na str. 30 je uvedeno, že váhy modelu byly inicializovány pomocí *he_normal* inicializace, ale nikde není vysvětleno, co tato inicializace znamená.

Na závěr jsou zhodnoceny výsledky nejlepšího vytvořeného modelu, který je zkombinován s klasifikací pomocí klíčových slov. Výsledky jsou porovnány se základním modelem a je ukázáno výrazné zlepšení oproti základnímu modelu, zejména pro stránky, které byly nesprávně klasifikovány jako nerelevantní (falešně pozitivní).

Kvalita řešení a dosažených výsledků
Celkově práci považuji za zdařilou a dosažené výsledky za velmi dobré, ale nedokáži posoudit, zda kvalita klasifikace bude dostačující pro další použití v projektu zmíněném v úvodu práce. Zdrojový kód je přehledný, je vhodně rozdělen do tříd a samostatných souborů a bylo možné jej bez problému spustit.

Formální a jazyková úroveň
Formální úroveň práce je v pořádku a autor si dal na práci záležet. Dokument bakalářské práce je vysázen v LaTeX. V textu jsem objevil minimum překlepů a gramatických chyb. Autor občas používá a skloňuje anglická slova, která mají český ekvivalent (overfitting – přeučení, word embeddingy - slovní vektory apod.). Obrázky a diagramy jsou vektorové.

Práce s literaturou
Citovaná literatura je vzhledem k tématu bakalářské práce relevantní, pouze před citacemi většinou chybí mezera.

Splnění zadání
Zadání práce bylo kompletně splněno.

Dotazy k práci

1. Hledal jste nebo víte o jiných pracích, které řeší stejný nebo podobný problém ? Pokud ano, jaké postupy jejich autoři používají a jakých výsledků dosáhli?

2. Jak složité by bylo rozšířit aplikaci, aby ji bylo možné aplikovat i na jiný jazyk např. angličtinu?

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

V Plzni dne 31. května 2021

Ing. Pavel Přibáň
(oponent BP)