Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Hynek Kydlíček
Název práce	Implicit Information Extraction from News Stories
Rok odevzdání	2023
Studijní program	Informatika
Specializace	Umělá inteligence
Autor posudku	Jindřich Libovický Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání	X			
Splnění zadání	X			
Rozsah práce *... textová i implementační část, zohlednění náročnosti*	X			

Bakalářská práce skládá ze dvou logických celků: Prvním je příprava rozsáhlého datasetu žurnalistických textů z českých zpravodajských serverů mezi lety 2000 a 2022. V druhé části práce se student věnoval vývoji modelů strojového učení založených na předtrénovaných jazykových modelech pro češtinu (RobeCzech a FERNET-NEWS) a jejich porovnání s komerčně dostupným systémem GPT-3. Porovnání s lidskou anotací ukazuje, že jazykové modely jsou schopny klasifikovat většinu vlastností novinových článků lépe než lidští hodnotitelé.

Student splnil zadání v plném rozsahu a v mnoha ohledech předčil původní plány. Dataset a ukázky úloh, které se s pomocí datasetu dají řešit student a vedoucí práce zpracovali do článku, který je přijatý k publikaci na konferenci Text, Speech and Dialog.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*	X			
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*	X			
Analýza	X			
Vývojová dokumentace	X			
Uživatelská dokumentace	X			

Práce není softwarovým dílem v klasickém slova smyslu, jak předpokládá tento hodnotící formulář. Příprava datasetu pro klasifikaci zpravodajských textů je spíše data science. Spíše než o detailní analýzu a následnou implementaci se jedná o iterativní proces, kdy se předpoklady o tom, jaká by data měla být střetávají s realitou. To je velmi dobře popsáno v kapitole 2, kterou bychom podle tohoto formuláře mohli považovat za vývojovou dokumentaci. Stejně kvalitně se popsána i následná experimentální práce (kapitoly 3–4), kde bylo cílem vyvinout klasifikátor pomocí předtrénovaných jazykových modelů pro češtinu a porovnat tyto klasifikátory s jazykovým modelem GPT-3 a lidským hodnocením.

Drobné nedostatky a faktické nepřesnosti lze najít v kapitole 1. V popisu architektury Transformer chybí dopředné vrstvy (s. 14), jazykový model ELMo používá rekurentní neuronové sítě, nikoli Transformer (s. 15).

Implementační část práce

	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*	X			
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*	X			
Stabilita implementace	X			

Příprava datasetu je výpočetně náročná, proto je řešená pomocí distribuované architektury, která paralelně stahuje a zpracovává stránky z projektu CommonCrawl. Úloha je složitější tím, že stejné články mohou mít odlišné URL nebo se z různých jiných důvodů vyskytují v CommonCrawlu vícekrát. Proto je potřeba zpracování dat mezi sebou synchronizovat. Kód je psaný na míru výpočetním clusteru AIC provozovaným ÚFAL, ale měl by fungovat na všech linuxových clusterech pracujícím s rozvrhovacím systémem Slurm.

Práce se s jazykovými modely a dalšími nástroji pro strojové učení pomocí nástrojů, které jsou v komunitě běžné (Huggingface Transformers, PyTorch Lightning). Při vývoji modelů student prokázal, že velmi dobře orientuje v současné literatuře o jazykových modelech a zvládá implementovat a experimentálně ověřovat nejnovější metody.

Celkové hodnocení Výborně

Práci navrhuji na zvláštní ocenění Ne

Datum 19.6. 2023

Podpis