FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Ing. Jiří Martínek, Ph.D.
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd
Technická 8, 301 00 Plzeň
Tel.: (+420) 377 632 446
E-mail: jimar@kiv.zcu.cz

Posudek oponenta bakalářské práce

Václav Tran
Automatické vytváření souhrnů historických dokumentů

Obsah práce
Obsahem bakalářské práce Václava Trana bylo prozkoumat metody automatické sumarizace textu, dostupné související datové sady a provést experimenty s vybranými modely. Součástí práce bylo i vytvoření datové sady českých historických dokumentů.

V úvodu práce je vysvětlena motivace, v následné kapitole pak stručně neuronové sítě. Následuje pak rešerše dostupných datových sad s jejich porovnáním (*Tabulka 3.1*). V kapitole 4 autor popisuje dostupné metody a modely opřené o reference na vědecké články. Zde mi schází aspoň zmínka o metodě *Latent Semantic Analysis* – LSA, která se svého času používala rovněž pro sumarizaci. Na druhou stranu oceňuji, že se student společně s vedoucím zaměřili na moderní metody a velké jazykové modely, které jsou dnes state of the art v oblasti zpracování přirozeného jazyka. Následují návrh a popis experimentů s prezentací výsledků. K těmto kapitolám **nemám žádné zásadní výhrady**.

Kvalita řešení a dosažených výsledků
Práce je čtivá, poměrně dobře napsaná a myslím si, že by obstála i jako diplomová práce. Práce obsahuje rešerši metod, analýzu souvisejících prací (*related work*) a dobře navržené a provedené experimenty. V sekci *5.3 Creating Dataset for Model Evaluation* je zmínka o „Domažlických listech“ jakožto součást datové sady. Nicméně hned v následující větě: (*While constructing the dataset, we only used the texts from the journal „Posel od Čerchova“*) je řečeno, že se tento zdroj dat nepoužije, ale chybí důvod, proč tomu tak je.

Dodané zdrojové kódy jsem po nainstalování potřebných knihoven dokázal spustit. Implementace prostřednictvím Jupyter notebooků je pro tento typ práce dle mého názoru vhodná a uživatel může postupně spouštět jednotlivé buňky s kódy. Ocenil bych, že student (ve spolupráci s vedoucím) dokázal proniknout do trénování velkých jazykových modelů, které mají několik miliard parametrů a pro jejichž trénování je zapotřebí výkonných výpočetních clusterů (MetaCentrum).

Formální úroveň
Práce je vysázena v typografickém systému LaTeXa je psaná dobrou angličtinou, jediný překlep jsem nalezl v české verzi abstraktu. Zrušil bych číslování stránky v obsahu. Z hlediska formální úrovně mám **výhradu k rovnicím a vzorcům**, které nejsou číslovány. V sekci *4.4 Evaluation Metric* je nevhodně použit znak $\times$ (vektorový součin) jakožto obyčejné násobení. V té samé sekci bych očekával lepší vysvětlení metrik, např. u popisu *ROUGE-N* se bez předchozího vysvětlení „objeví“ precision, recall a F1-score a není popsáno, jak to souvisí s popisovanou metrikou.

Reference a práce s literaturou
Práce s referencemi a literaturou bez problémů. Jedinou výtku mám, že se v nemalé míře mezi referencemi objevují záznamy s arXivu (nerecenzované zdroje). Upřednostnil bych oficiální citace z recenzovaných sborníků a časopisů (obzvlášť pokud jsou již několik let staré a zcela určitě je možné je dohledat).

Závěr a hodnocení

Zadání z mého pohledu je splněno ve všech jeho bodech. K práci nemám žádné zásadní výhrady. Navrhuji známku „**výborně**“ a práci **doporučuji k obhajobě**.

Dotazy k práci

1. Z jakého důvodu byly z experimentů a z datové sady vyřazeny „Domažlické listy“?
2. Plánujete s vedoucím využít dosažené výsledky pro publikaci článku na vědecké konferenci?

V Plzni dne 22. května 2024

Ing. Jiří Martínek, Ph.D.
(oponent BP)