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

Daniel Cífka
Experimenty s moderními modely neuronových sítí pro vícetřídní klasifikaci českého textu

Obsah práce
Práce je výzkumného charakteru a se zabývá tzv. *multi-label* klasifikací českého datasetu novinových článků. Cílem práce je provést experimenty pro daný dataset s pomocí neuronových sítí založených na architektuře Transformer.

V první části práce je popsána teorie neuronových sítí, bohužel dle mého názoru nedostatečně a pouze povrchově. Vyvolává to ve mně dojem, že student danou problematiku plně nepochopil. Některé důležité pojmy jsou pouze zmíněny, ale nevysvětleny např. *vanishing gradient*, *positional encoding*, *multi-head self-attention* a další. Za nedostatečné považuji popis architektury Transformer a její použití pro klasifikaci textu vzhledem k tomu, že je to stěžejní část celé práce. Student vůbec nepopsal teorii a matematický model klasifikační hlavy a definici klasifikační úlohy. Dále chybí jakékoliv vysvětlení fungování tzv. *subword tokenizeru*, který je pro model Transformer důležitý. Za pozitivní považuji, že se podrobně věnuje použité datové sadě a analyzuje její vlastnosti, které jsou důležité pro následné experimenty.

V práci chybí analýza dostupných modelů a existujících řešení pro klasifikaci českého textu a případné možnosti jejich využití v této práci. Autor pouze popisuje některé české monolinguální modely, které používá pro experimenty, ale není u nich jasné na základě čeho je vybral. Autor vůbec nezmiňuje multilinguální modely (např. mBERT nebo XLM-R) a ani český model Czert.

Kvalita řešení a dosažených výsledků
Kompletní experimenty jsou provedeny pouze pro jeden český model, experimenty pro ostatní modely nejsou dokončeny z údajné časové náročnosti. Myslím si, že množství provedených experimentů neodpovídá rozsahu bakalářské práce a mělo bych jich být více. Analýza chyb by měla obsahovat detailnější informace a statistiky o chybně klasifikovaných třídách. Dále mi chybí porovnání se staršími neuronovými sítěmi (např. LSTM nebo CNN). V tabulce 6.7 není zřejmé, které výsledky naměřil student.

Zdrojové kódy bylo možné spustit. Kód je komentován, ale autor zbytečně duplikuje části kódu, které mohou být společné (trénovací smyčka a načítání datové sady).

Formální úroveň
Formální úroveň práce je dostatečná. Dokument bakalářské práce je vysázen v typografickém systému LaTeX. V textu se vyskytuje množství překlepů. Tabulky a obrázky často obsahují popisky v anglickém jazyce a nejsou vysvětleny. Dále jsou v textu zaměňovány české a anglické výrazy, např. *loss function* a *ztrátová funkce*. Autor bohužel velmi často používá hovorové a neformální vyjadřování (např. *šlo ho* str. 11, *do nějaké* str. 13 a další), které je vyloženě nevhodné pro bakalářskou práci. Celkově by práce zasloužila daleko větší pozornost a pečlivost.

Práce s literaturou
Citovaná literatura je vzhledem k tématu bakalářské práce relevantní a citace v textu jsou v pořádku.

Splnění zadání
I přes výše uvedené nedostatky považují zadání práce za splněné.

Dotazy k práci

1. Co znamená pojem *vanishing gradient*?

2. Vstup použitých modelů je omezen na 512 tokenů, jak byste postupoval v případě, kdy byste chtěl, aby model zpracoval i delší texty?

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

V Plzni dne 30. května 2023

Ing. Pavel Přibáň
(oponent BP)