POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Vysvětlování výstupu modelů zpracování přirozeného jazyka pro úlohu ověřování faktů
Jméno autora:	**Eliška Kopecká**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra pocitacu
Oponent práce:	Gustav Šír
Pracoviště oponenta práce:	Katedra pocitacu

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Zadani prace, tak jak bylo sepsano, dava dostatek prostoru ke splneni, coz je vhodne, s prihlednutim k problematicke specifikaci ulohy interpretability ve strojovem uceni.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Bez vyhrad, studentka vhodne voli jednu ze 2 hlavnich komponent existujici pipeline a testuje jeji interpretabilitu.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Postup mi prijde zcela odpovidajici. Studentka provadi sirokou resersi pojmu interpretability, a voli 2 nejpouzivanejsi metody pro lokalni vysvetlovani vystupu modelu, adaptuje je pro dany problem a dukladne testuje, s prihlednutim k omezenim plynoucim z daneho prostredi (narocnost testovani).

Odborná úroveň **A - výborně**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Prace nepredstavuje z hlediska odbornosti zadne prevratne koncepty, ale to je dane i problematikou interpretability vytycenou v zadani. Navic studentka k problematice pristupuje velmi peclive, spravne upozornuje na zaludnosti teto domeny, a nabizi nektera i docela kreativni reseni. Sice pouziva vesmes standardni metody/knihovny, avsak predstavuje i nektere vlastni modifikace, a resi i testovani v „produkci“. Vse je navic zpracovano velmi peclive.

Formální a jazyková úroveň, rozsah práce **A - výborně**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Zde je prace zcela vyborna, text je psany velmi peclive se zjevnym durazem na porozumeni ctenare. Obcas pusobi mozna az trochu moc pribehove na odborny text, ale osobne to hodnotim kladne. Anglictina je skvela, typografie velmi pekna, vcetne grafiky a vyberu barev.

Výběr zdrojů, korektnost citací **A - výborně**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Bez vyhrad.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

Další komentáře a hodnocení
Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.
Viz celkove hodnoceni.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Prace resi problem interpretace modelu pro overovani faktu v ceskem jazyce, k cemuz adaptuje 2 klasicke, lokalni, agnosticke metody. Pozitivne hodnotim i vlastni modifikaci s embeddings (text- augmented LIME). Tyto metody pak testuje s prihlednutim k omezenym moznostem evaluace (za me idealni postup k testovani problematickeho konceptu interpretability v dane uloze je pak hezky nastinen ve future work – s tim se plne ztotoznuji).

Celkove se studentka s touto zaludnou problematikou poprala velmi obstojne a vysledek pusobi profesionalne.

Poznamky:

Osobne nemyslim, ze je hned nutne uchylovat se pouze k lokalnim, neprimym metodam jako LIME a resit (diskutovat) pak problemy s fidelitou (o cemz svedci i lepsi vysledek SHAP), ale lepe primo trackovat gradient/feature saliency v danem transformeru, k cemuz uz lze kdyztak take nalezt hotove knihovny [1].

Otazky:

- Naznacujete ze fasttext embeddings nejsou prilis vhodne, protoze nezohlednuji dany problem. Nebylo by tedy lepsi pro vysvetlovani pouzit primo embeddings z daneho transformeru?
- Neslo by pouzit bag of embeddings namisto bag of words a delat perturbaci primo v prostoru embeddings, namisto hledani nejblizsiho slova?
- Byl vyvoj zobrazene aplikace soucasti Vasi prace?
- Chapu spravne ze pocet testeru kvality interpretaci byl 2?

Par typos:

trash-hold :)

only additive possible additive

(e.g..

paremtrizations

[1] https://github.com/jalammar/ecco

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně.**

Datum: 6.6.2022 Podpis: