Posudek oponenta diplomové práce

Autor/autorka práce: **Vojtěch Bartička**

Název práce: **Attribution Methods for Explaining Transformers**

Práce se zabývá netriviální úlohou z oblasti explainable AI. V práci je uvedeno, že je rozšířením autorova konferenčního článku. Tomu odpovídá i kvalita práce. Text je na dobré úrovni s minimem překlepů. Výjimkou je chybějící softmax v rovnici 2.5. Z práce jsou patrné hlubší znalosti dané problematiky. Silnou stránkou je vysvětlení jednotlivých přístupů a množství související literatury.

Důležité kroky týkající se např. přípravy datové sady jsou diskutovány a zdůvodněny. Experimenty jsou jako práce samotná systematicky rozvrženy. Z výsledků a diskuze jsou patrné možné problémy a výhody přístupů.

Po drobné úpravě v pořadí přípravy SST datasetu pracují přiložené skripty pro reprodukci výsledků dle obsáhlého readme. Skripty by však mohly být lépe strukturovány.

Zadání bylo splněno bez výhrad.

Dotazy k práci:

1. Pro vyhodnocení CTDC datasetu je zvolen výpočet dle rovnice 5.4. Jaké plynou možné problémy z porovnání top K a celého listu klíčových slov? Jak by se daly řešit?
2. Která z testovaných atribučních metod je nejméně závislá na modelu a úloze (datech)?
3. Jaká je časová náročnost u KernelSHAP atribuční metody vzhledem k počtu vzorků?

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

V Plzni 2.6.2023 Ing. Josef Baloun