Posudek diplomové práce

Diplomant: Bc. František Špaček

Předložená diplomová práce se zabývá úlohou určování sentimentu textů pomocí strojového učení a identifikací nejúspěšnějšího modelu využitelného v komerční sféře automobilového průmyslu, ve kterém diplomant pracuje. Předložená diplomová práce je tedy tvořena s ohledem na požadavky zadání konkrétní společnosti aktivní ve zmíněném průmyslu.

Uvedená práce nejprve shrnuje důvody zájmu o analýzu sentimentu textu v automobilovém průmyslu a vytyčuje tím vlastní cíl práce. Dále vysvětluje úskalí tvorby modelů strojového učení vzhledem k dostupnosti tzv. trénovacích dat se zvláštním zřetelem na licenční podmínky jejich možného komerčního užití. Hlavní náplní této práce se následně stává otázka řešení nalezení nejlepšího možného modelu pro binární klasifikaci sentimentu textů vytvořeného z dostupných dat. Jednotlivé zkoumané modely postupně zahrnují metody klasického strojového učení jako k-NN, SVM, Naive Bayes, až po dopředné/rekurentní/konvoluční neuronové sítě včetně jejich spřažení, doladění před-trénovaných transformerů, tj. nejnovějších metod, včetně zahrnutí testování různých způsobů tokenizace a vektorizace textů.

Práce má řadu pozitiv: (1) diplomant prokazuje schopnost řešit kvalitativně zadaný cíl analýzy textu strojovým učením, (2) poskytuje odpověď, který postup, metoda či který model pro řešení vytyčené úlohy v závěru pro nejefektivnější výsledky použít. Obou bodů je dosaženo systematicky, včetně testování vlivu velikosti datasetů, včetně snahy vždy alespoň minimálně vysvětlit pozorované jevy. V tomto ohledu je práce velmi pozitivní.

Primárním úskalím celé práce je ovšem její formální stránka a především jazyk, kterým je napsána. Je možné říct, že předložená práce je protkána řadou neakademických jazykových neobratností. Příklady těchto neobratností jsou častá slovní spojení zahrnující slova „nějaký/á“, např. na str. 2 "nějaká rizika", "něco podrobněji", "nějaké spektrum", na str. 33 "provede nějaké výpočty", dále nešikovná označení jako "hluší", "němí", str. 4, str. 19 "pořádné uplatnění" a řady dalších. V práci také nalezneme řadu anglických či počeštěných názvů, které by bylo vhodné dát do kurzívy nebo ideálně použít jejich české ekvivalenty. Nezřídka lze nalézt překlepy či chybějící interpunkci.

V obsahu se můžeme setkat i s určitými technickými nepřesnostmi (např. str. 19, *FastText* nutně nemusí používat pouze dvojice znaků; str. 29 SVM nehledá "osu", ale obecně tzv. nadrovinu; str. 29 implementace k-NN v knihovně *sklearn* využívá KD-strom, který umožňuje ono rychlé vyhledávání a právě jeho inicializace probíhá uvnitř volání metody *.fit(...)*, str. 36 vrstva *flatten* neprovádí transformaci, ale pouze zploštění tenzoru a další). Z hlediska metody pak vznikají otázky nad tím, zda byly pro každý model a testy využívány vždy stejné texty pro trénování a validaci, tj. zdroje variability, které mohly vést k odchylkám měřených metrik. U tabulek a obrázků s výsledky chybí detailnější popisky, především, zda se v nich uvedené výsledky týkají trénovacích/validačních/testovacích dat, případně chybí doplnění zobrazených jednotek atp.

Předložená diplomová práce je tedy kombinací velmi hezkého aplikovaného lingvistického výzkumu užívajícího strojové učení včetně neuronových sítí a umělé inteligence, zároveň je ale napsána ne příliš puntičkářskou a jazykově pramálo vyladěnou formou, která by si zasloužila velkou řadu úprav a revizí. Nelze také opomenout, že součástí práce je rovněž kompletní zdrojový kód v jazyce Python provádějící trénování jednotlivých metod a jejich evaluace. Práci proto lze, především pak skrze její technickou náročnost i přes potíže s jazykem, doporučit k obhajobě se stupněm *výborně*, tj. B.

Otázka:

- Vysvětlete cíl, princip a následný důvod navýšené kvality modelů užívající BPE tokenizátor.
- Proč byl použit model Word2Vec namísto FastText, který by řešil i slova mimo slovník?

Mgr. Vladimír Matlach, Ph.D.
V Olomouc dne 2. 1. 2024.