Oponentský posudek diplomové práce

Autor a název předložené práce

Bc. Jakub Maroušek: *Automatická klasifikace smluv pro portál HlidacSmluv.cz*

Téma práce

Předložená práce se zabývá automatickou klasifikací textových dokumentů. Konkrétně se jedná o veřejně dostupné smlouvy o zakázkách ve veřejné správě. Explicitně zadaným úkolem studenta bylo využít metody strojového učení a navrhnout, implementovat, otestovat a zdokumentovat softwarový modul pro existující webový portál *HlidacSmluv.cz*. Nový modul má vstupní smlouvy klasifikovat do předem daných kategorií charakterizujících doménu klasifikované smlouvy. Předpokládá se využití existujících open-source nástrojů pro automatickou klasifikaci.

Na základě zkušeností s podobnými projekty považuji toto zadání za nesnadné z několika důvodů. Zaprvé je dosti problematické definovat, jaký výstup automatické klasifikace by měl být považován za správný nebo ideální. Z toho vyplývá nejasnost optimalizačních kritérií pro trénování automatického klasifikátoru. Zadruhé, systém předem definovaných cílových kategorií je v tomto případě dosti rozsáhlý (přes 100 kategorií, navíc s ne zcela přesně vymezenou inklusivitou, viz příloha A) a je známo, že klasifikované dokumenty obecně mohou náležet do více kategorií – tento fakt klasifikační úlohu silně komplikuje. Zatřetí, při práci s “reálnými” textovými daty je třeba počítat s častým výskytem nestandardních anomálií nebo chyb. A navíc, kritickým předpokladem pro úspěšnost metod strojového učení, které zde připadají do úvahy, bývá dostatek kvalitních trénovacích dat. Se všemi těmito problémy se student ve své práci více či méně potýká.

Rešerše – data – metody – experimenty – evaluace

Struktura předložené diplomové práce je zcela standardní. První tři úvodní kapitoly (cca 20 stran) popisují zadání problému a jeho kontext a základní analýzu dostupných datových sad. Zajímavým počátečním krokem je zde analýza smluv pomocí automatického shlukování (kap. 3.2.1). Bohužel zde není vůbec popsáno, s jakými vektory použité metody pro shlukování pracovaly. Ani výsledky shlukování nejsou analyzovány podrobně a z textu tedy není jasné, co experiment se shlukováním nakonec přinesl či nepřinesl a proč.

Ideové těžiště řešení zadaného problému leží v kapitolách 4 a 5 (cca 30 stran), kde autor jednak popisuje vybrané použitelné metody strojového řešení a diskutuje jejich volbu, jednak popisuje provedené experimenty včetně testování vlivu nastavení parametrů a měření kvality výsledků. Text popisující experimenty a jejich výsledky je kupodivu velmi krátký, ačkoliv se

nepochybně jedná o klíčovou část práce. Bohužel velmi malá pozornost je věnována testovacím datům. Pokud jde o klasifikované smlouvy, což je celkový cíl celé práce, v kap. 5.1 je pouze uvedeno, že bylo náhodně vybráno 204 testovacích smluv, ale nejsou zde ani žádné detaily o rozdělení kategorií manuálně přiřazených těmto testovacím smlouvám, ani zde není žádná diskuse o dostatečnosti tohoto relativně malého souboru (počet přiřazovaných kategorií je přes 100). V práci nenacházím ani žádnou exaktní analýzu chyb. Chybně se v práci mluví o "metrikách" pro hodnocení klasifikátorů (str. 48 a dále) – použité funkce z matematického hlediska metrikami nejsou.

Krátká kapitola 6 (7 stran) popisuje praktickou implementaci a diskutuje výběr použitých softwarových technologií.

Závěrečná kapitolka "Závěr" stručně shrnuje výsledky celé práce a uvádí také prvotní zkušenosti s praktickým provozem vytvořeného modulu.

Úroveň zpracování diplomové práce, implementace a dokumentace

Formální požadavky kladené na diplomovou práci jsou splněny. Autor dostatečně cituje použitou literaturu, jejíž seznam na konci práce obsahuje 69 položek (značnou část tvoří ovšem dokumenty vystavené na webu). Nevhodně je však sestaven abstrakt, který je spíše jakýmsi extraktem ze zadání, ale vůbec neinformuje o obsahu a výsledku práce.

Práce je napsána česky a úroveň češtiny je v celé práci velmi dobrá. Lze najít pouze drobné jazykové nedostatky (např. výrazy "méně úspěšněji" na str. 57, "z řad autorů" na str. 31, nebo "od ní spustí" na str. 17, za nepříliš vhodné považuji slovo "otagovaný" (str. 16) – proč ne české "označkovaný"?, podobně místo "clustering" (str. 24) lze v českém textu psát "shlukování" apod.). Překlepy jsem nenalezl, což svědčí o značné pečlivosti autora. Rovněž po grafické stránce je zpracování standardní a výhrady mám pouze drobné. Tak např. spojovací čáry v grafech s kategoriálními hodnotami na obr. 5.1 jsou zavádějící. U obrázku 3.1 chybí popisky, lze si je však domyslet. Na str. 52 je chybná sazba ve formuli 5.8 ("F – measure").

Implementace modulu pro klasifikaci textů je zřejmě funkční, jak mi sám autor osobně předvedl. Stručná programátorská dokumentace je obsažena v příloze C. Pokud jde o použité technologie, jejich popis a zdůvodnění vyznívá solidně, ovšem k posouzení jejich vhodnosti či optimálnosti nejsem kompetentní.

Otázky k obhajobě

- V práci jste experimentálně porovnal čtyři odlišné metody strojového učení. Domníváte se, že vhodná kombinace těchto metod by mohla přinést nějaké zlepšení? V předložené práci k této otázce nenacházím žádnou analýzu, ani zkoumání, zda je úspěšnost jednotlivých metod v nějakém smyslu komplementární. Diskuse v kap. 5.6 však naznačuje, že metoda *FastText* a klasifikátor založený na klíčových slovech mají úspěšnost srovnatelnou.

- Ze zadání vyplývá, že vyvinutý modul pro klasifikaci má být nasazen pro veřejné využití. V kap. 5.6 také uvádíte, že finální řešení bylo vybráno “ve spolupráci s autorem portálu Hlídač státu”. Jaké je hodnocení Vašich výsledků ze strany provozovatele serveru *HlidacSmluv.cz*? Bylo provedeno nějaké exaktní srovnání s existujícím systémem pro klasifikaci zmíněným v kap. 2.3, který – jak uvádíte – “není dostatečný” a “nepredikuje dostatečně přesné výsledky”?

- Jak má uživatel interpretovat výstupní číselný údaj, který klasifikátor přiřazuje pěti kategoriím vyhodnoceným jako nerelevantnější? Na str. 17 uvádíte možnost kalkulovat pravděpodobnost chápanou jako “váhu” příslušné kategorie, nebo jako “míru jistoty predikce”.

- Je podle Vašeho názoru testovací sada smluv dostatečná? Svůj názor prosím zdůvodněte.

Celkové hodnocení

Předložená diplomová práce zřejmě splňuje zadání. Analýza dané problematiky, průběh práce, provedené experimenty a jejich výsledky i implementace jsou popsány srozumitelně. Práce na automatické klasifikaci veřejných textových dokumentů byla pro studenta jistě velmi cennou zkušeností. Musel se potýkat s celou řadou praktických problémů a popsaná praktická řešení je třeba ocenit. V některých částech – zejména diskuse o metodách strojového učení a jejich výběr, testování a evaluace výsledků – bych však od diplomové práce na MFF UK očekával poněkud větší preciznost.

Předloženou práci doporučuji k obhajobě.

V Praze, 4. září 2020

RNDr. Martin Holub, Ph.D.