Posudek oponenta diplomové práce

Autor/autorka práce: **Matěj Zeman**

Název práce: **Multi-modal emotion analysis in textual and audio data**

Obsah práce

Práce se zabývá detekcemi emocí z textových a zvukových dat. Text je rozdělen na tři hlavní části – teoretickou, popis dostupných datových sad a vlastní experimenty.

Teoretická část práce pokrývá převážně extrakci příznaků z audio signálů a textových dat.

V audio části je zmíněno několik metod pro extrakci příznaků. Největší pozornost je věnována metodě MFCC. Ostatním metodám je věnováno prostoru méně, což například u Chromagramu vede k nejasnostem (co je agregační funkce F, proč je použito právě 12 vzorků, co jsou třídy B a C apod.).

Sekce věnovaná textu je rozvržena lépe a popisuje jak starší, ale stále použitelné metody, tak novější přístupy. U slovních vektorů není ovšem vysvětleno, co reprezentuje úhel mezi nimi (resp. jejich kolmost). V sekcích 3.2.1 a 3.2.2 pro varianty Word2Vec by k lepšímu pochopení pomohla konkrétní ukázka či názornější ilustrace. Obrázek 3.3, ukazující architekturu sítě BERT, je nedostatečně popsaný a nepřehledný.

Poslední část teorie je věnována popisu některých stavebních prvků (konkrétně LSTM a CNN) neuronových sítí. Tuto část by bylo lepší zařadit před extrakci textu, vzhledem k tomu, že používá pojmy vysvětlené až v této části. LSTM architektuře je věnováno zbytečně moc prostoru, který by bylo lepší využít pro popis jiných metod (např. GRU, různé varianty konvolučních sítí). Ačkoliv by výběr mohl být širší, vzhledem k pozdějšímu použití pouze těchto popsaných metod je dostatečný.

V teoretické části se na několika místech vyskytují, pro čtenáře neznalého oboru, nevysvětlené pojmy či dříve nedefinované zkratky (cepstrum, glottal closures, morphem, Seq2Seq, self-attention atd.).

Část popisující datové sady popisuje tři sety. Kromě vlastního popisu autor i data upravuje (odstranění videa a ponechání jen zvukové složky), případně navrhuje vynechání některých tříd s nedostatečným počtem dat pro trénování sítí.

Popis vlastního řešení klasifikace emocí je zbytečné krátký a málo rozvedený. Z textu není místy patrné, co slouží jako vstup sítě. Multi-modální část bohužel zabírá pouze jednu stránku a modely tak nejsou dostatečně popsané.

Poslední část se věnuje vlastním experimentům. Autor testuje modely popsané v předchozí části práce. Je vyzkoušeno množství různých variant a kombinací předzpracování a výsledky jsou porovnány v přehledných tabulkách. Pouze u finální shrnující tabulky 7.10 došlo nejspíše ke špatnému značení sloupců (hodnoty ve sloupci „Text feature extraction method“ nedávají smysl).

Dodaný testovací program

Součástí práce je testovací program napsaný v jazyce Python. Členění kódu by mohlo být vyřešeno lépe hlavně při práci s datasety, kde jsou odlišnosti řešeny podmínkami uvnitř relativně dlouhých metod místo využití tříd, dědičnosti a popř. mechanismu dependency injection.

Práce s literaturou

Práce s literaturou je na dobré úrovni, student cituje relevantní odborné publikace, včetně publikací z posledních let.

Splnění zadání

Zadání bylo splněno.

Dotazy k práci

1) Proč byla využita v experimentech také 2D CNN? Na zvukový signál by měla stačit 1D CNN.
2) Zkoušel jste pro zpracování zvuku nějaké architektury speciálně navržené pro zvuk? Podobně jako je např. BERT a LSTM pro text.

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

V Plzni 28.5.2024 Ing. Martin Prantl, Ph.D.