Mendelova univerzita v Brně
Provozně ekonomická fakulta

Analýza sentimentu textových recenzií vybraného segmentu produktov

Bakalárska práca

Vedoucí práce:

Ing. Jan Přichystal, Ph.D.

Dorota Košťálová

Brno 2020

Obsah

1 Úvod a cieľ práce

1.1 Úvod

Dáta, ktoré prevládajú v dnešnom svete sú neštruktúrované, inými slovami, nie sú usporiadané. Každý deň sa vytvára veľké množstvo textových údajov. Je však ťažké ich analyzovať, porozumieť im a triediť ich. Najmä, tento postup je časovo veľmi náročný.

Analýza sentimentu je oblasť, ktorá analyzuje názory ľudí, názory, hodnotenia, postoje a emócie z písaného jazyka (Liu 2012).

Systémy na analýzu sentimentu sa uplatňujú takmer v každej obchodnej a sociálnej oblasti, pretože názory sú ústrednou témou takmer všetkých ľudských činností a sú kľúčovými ovplyvňovačmi nášho správania. Naše presvedčenie a vnímanie reality a rozhodnutia, ktoré robíme, sú do veľkej miery podmienené tým, ako ostatní vidia a hodnotia svet. Z tohto dôvodu, keď sa musíme rozhodnúť, často hľadáme názory druhých. To platí nielen pre jednotlivcov, ale aj pre organizácie (Liu 2012).

V dnešnej dobe už zákazníci nie sú odkázaní pri nákupe produktov na pýtanie sa názorov okolia alebo na prehľadávanie siahodlhých diskusných fór.

Úlohou analýzy sentimentu nie je vyťažovanie konkrétnych faktov, ale zisťovanie subjektívneho kontextu textov a postoja pisateľa textu k skúmanej téme. Analýza sentimentu pomáha pochopiť neštruktúrovaný text tým, že ho automaticky označí.

Z hľadiska kategorizácie sa analýza sentimentu zaraďuje do oblasti NLP (natural language processing), teda spracovania prirodzeného jazyka.

Pod pojmom sentiment rozumieme subjektívne, psychicky a sociálne podmienené, viac alebo menej pozitívne prejavy vôle, bezprostredné hnutia mysle, náklonnosť alebo odpor, ktoré nezávisia od rozumovej úvahy. Dôsledkom môže byť kladný alebo záporný postoj, ktorý môže byť aj trvalý a nepodlieha úplne vedomej kontrole. Na základe individuálnych postojov potom ľudia obhajujú svoje názory a riešia dilemy.

1.2 Cieľ práce

Cieľom práce je vytvoriť nástroj schopný určiť sentiment neštruktúrovaného textu recenzie. Nástroj musí byť schopný vyhodnotiť či sa z hľadiska sentimentu jedná skôr o negatívny, pozitívny alebo neutrálny text. Výsledky práce môžu byť prospešné pre zákazníka, ale i pre výrobcu produktov. Zákazníci vďaka výsledkom získajú prehľad veľkého množstva recenzií z rôznych stránok, ktoré by inak museli prechádzať a vyhodnocovať sami. Tento postup by mohol zabrať veľké množstvo času, v prípade, že by chceli získať čo najširší a najdôveryhodnejší prehľad. Vďaka spracovaným a vyhodnoteným dáta budú zákazníci jasne vidieť či u konkrétneho produktu prevažujú pozitívne alebo negatívne recenzie. Na základe čoho, bude uľahčený ich výber. Na strane výrobcu bude možné získaný celkový dojem zákazníkov z produktu zohľadniť pri budúcom vývoji produktov.

2 Prehľad literatúry

Od začiatku roku 2000 sa analýza sentimentu stala jednou z najaktívnejších výskumných oblastí v oblasti spracovania prirodzeného jazyka (NLP). Je tiež široko študovaná v oblasti data miningu, Web miningu, text miningu a získavania informácií. V skutočnosti sa rozšírila z počítačovej vedy na vedecké a spoločenské vedy, ako sú marketing, financie, politológia, komunikácia, medicína a dokonca aj história, a to z dôvodu jej dôležitosti pre biznis a spoločnosť ako celok (Zhang, Wang, and Liu 2018).

Existujú rôzne techniky pre rozličné úlohy sentiment analýzy, ktoré zahŕňajú metódy učenia s učiteľom, bez učiteľa a kombináciou oboch učení.

Poslednú dekádu sa hlboké učenie (deep learning) ukázalo ako mocný nástroj strojového učenia a prinieslo namodrenejšie výsledky od počítačového videnia a rozpoznávania reči po spracovanie prirodzeného jazyka (Goodfellow, Bengio, and Courville 2016). V poslednej dobe sa aplikovanie hlbokého učenia na analýzu sentimentu stalo veľmi populárne(Zhang, Wang, and Liu 2018).

Klasifikácia sentimentu na úrovni dokumentu označuje dokument s názorom (napr. recenzia produktu) ako vyjadrenie celkového pozitívneho alebo negatívneho názoru. Celý dokument považuje za základnú informačnú jednotku a predpokladá, že obsahuje stanoviska k jednej entite, napríklad ku konkrétnemu telefónu (Zhang, Wang, and Liu 2018). Následujúce práce zohľadnené v rešerši sú práve z kategórie analýzy sentimentu na úrovni dokumentu, keďže cieľom práce sú recenzie vybraných produktov.

Tang, Qin a Liu (2015a) navrhli neurónovú sieť na učenie sa reprezentácie dokumentov so zreteľom na vzťahy medzi väzbami. Najskôr sa model naučí znázornenie vety pomocou CNN (Convolutional neural network) alebo LSTM (Long short-term memory) z vnorenia slov. Potom je GRU (Gated recurrent unit) využitá na adaptívne zakódovanie sémantiky viet a ich vnútorných. Experimentálne výsledky ukazujú, že: (1) ich nervový model vykazuje vynikajúce výkony v porovnaní s najmodernejšími algoritmami; (2) GRU neurónová sieť dramaticky presahuje štandardnú rekurentnú neurónovú sieť v modelovaní dokumentov na klasifikáciu sentimentu.

Xu a kol. (2016) navrhli model LSTM s vyrovnávacou pamäťou na zachytenie celkových sémantických informácií v dlhom texte. Pamäť v modeli je rozdelená do niekoľkých skupín s rôznou mierou zabúdania. Intuíciou je umožniť skupinám pamäti, ktoré nezabúdajú, aby zachytili globálne sémantické vlastnosti, a tie, ktoré majú vysokú mieru zabudnutia, sa naučia lokálne sémantické vlastnosti. Navrhovaný CLSTM (Contextual LSTM) model prevyšuje najmodernejšie modely na troch verejne dostupných súboroch údajov o analýze sentimentu na úrovni dokumentu.

Yang a kol. (2016) navrhli HAN (Hierarchical attention network) predikciu hodnotení sentimentu na úrovni dokumentu. Model obsahuje dve úrovne mechanizmov pozornosti: jednu na úrovni slov a druhú na úrovni vety, ktoré modelu umožňujú venovať viac-menej pozornosti jednotlivým slovám alebo vetám pri zostavovaní reprezentácie a dokumentu. Experimenty uskutočnené na šiestich úlohách na klasifikáciu textu, dokázali že navrhovaná architektúra výrazne prevyšuje predchádzajúce metódy o podstatné hodnoty.

Moraes , Valiati a Neto (2013) urobili empirické porovnanie medzi SVM (Support vector machine) a ANN (Artificial neural network) pre klasifikáciu sentimentu na úrovni dokumentov, čo preukázalo, že ANN vo väčšine prípadov priniesla konkurenčné výsledky pre SVM. ANN predbehla SVM na skúšobných datasetoch filmových recenzií. ANN bola bola zriedka braná ako riešenie pre analýzu sentimentu. Kdežto SVM bol často a úspešne používaný prístup k učeniu sa v oblasti analýzy sentimentu.

Le a Mikolov (2014) navrhli paragrafový vektor (Paragraph vector), algoritmus založený na učení sa bez učiteľa, ktorý sa učí vektorové reprezentácie textov s rôznou dĺžkou, ako sú vety, odseky a dokumenty. Vektorové reprezentácie sa učia predpovedaním okolitých slov zo vzorkovaného kontextu paragrafu. Na ich experiment vykonali na viacerých úlohách na klasifikáciu textu. Použili IMBD dataset na analýzu sentimentu a Stanford Treebank dataset. Výsledky ukázali, že zvolená metóda je konkurenciou s inými najmodernejšími metódami. Zistili, že v skutočnosti má paragrafový vektor potenciál prekonať slabé stránky BoW model (bag-of-words). Napriek tomu, že práca bola zameraná na reprezentáciu textov, ich metóda môže byť aplikovaná k učeniu sa reprezentácií sekvenčných dát.

Glorot, Bordes a Bengio (2011) študovali problém adaptácie domén v klasifikácií sentimentu. Navrhli systém hlbokého učenia(deep learning) založený na DAE (Denoising Auto-Encoders) s riedkymi usmerňovacími jednotkami, ktoré môžu vykonávať extrakciu bez učiteľa textových alebo znakových reprezentacií s použitím označených aj neznačených údajov. Tieto vlastnosti sú veľmi prospešné pre doménovú adaptáciu klasifikátorov sentimentu.

Dou (2017) použil sieť hlbokých pamätí (Deep memory network) na zachytenie informácií o užívateľoch a produktoch. Navrhovaný model je možné rozdeliť na dve samostatné časti. V prvej časti sa aplikuje LSTM na naučenie sa reprezentácie dokumentu. V druhej časti sa na predpovedanie hodnotenia každého dokumentu používa sieť s hlbokou pamäťou pozostávajúca z viacerých výpočtových vrstiev (hops). V porovnaní s inými modelmi dokázali, že pri správnom nastavení dosahuje ich model vynikajúce výsledky. Malo by sa poznamenať, že z experimentu sa preukazuje, že je stále možné urobiť niekoľko zlepšení, ako napríklad lepšie zastúpenie dokumentov alebo prepracovanejšie mechanizmy pozorovania. Sú presvedčení, že ich model má obrovský potenciál a možno ho vylepšiť mnohými spôsobmi.

3 Referencie

Dou, Zi-Yi. 2017. “Capturing User and Product Information for Document Level Sentiment Analysis with Deep Memory Network.” In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 521–26.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. “Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach.”

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

Le, Quoc, and Tomas Mikolov. 2014. “Distributed Representations of Sentences and Documents.” In *International Conference on Machine Learning*, 1188–96.

Liu, Bing. 2012. “Sentiment Analysis and Opinion Mining.” *Synthesis Lectures on Human Language Technologies* 5 (1): 1–167.

Moraes, Rodrigo, JoãO Francisco Valiati, and Wilson P GaviãO Neto. 2013. “Document-Level Sentiment Classification: An Empirical Comparison between SVM and ANN.” *Expert Systems with Applications* 40 (2): 621–33.

Tang, Duyu, Bing Qin, and Ting Liu. 2015. “Document Modeling with Gated Recurrent Neural Network for Sentiment Classification.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1422–32.

Xu, Jiacheng, Danlu Chen, Xipeng Qiu, and Xuangjing Huang. 2016. “Cached Long Short-Term Memory Neural Networks for Document-Level Sentiment Classification.” *ArXiv Preprint ArXiv:1610.04989*.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. “Hierarchical Attention Networks for Document Classification.” In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–89.

Zhang, Lei, Shuai Wang, and Bing Liu. 2018. “Deep Learning for Sentiment Analysis: A Survey.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (4): e1253.