Anotovanie slovníka pre analýzu sentimentu pomocou PSO

Martin Mikula, Kristína Machová

Katedra kybernetiky a umelej inteligencie, TU V Košiciach
Letná 9, 042 00 Košice

{martin.mikula, kristina.machova}@tuke.sk

Abstrakt. Internet v súčasnosti obsahuje veľké množstvo textu, ktorý obsahuje emócie a názory rôznych ľudí. Keďže je veľmi náročné analyzovať ich manuálne, v tomto príspevku sa zameriavame na automatickú analýzu sentimentu použitím slovníkového prístupu. Anotovanie slovníka je často náročný a zdĺhavý proces. Práve preto sa v tejto práci venujeme možnosti nahradenia človeka, ktorý by slovník anotoval, evolučným algoritmom, ktorý by bol použitý na nájdenie optimálnych hodnôt polarity pre slová v slovníku. Vytvorili sme dve verzie slovníka, ktorý bol použitý na analýzu sentimentu. Prvá verzia bola anotovaná človekom a pre anotovanie druhej verzie sme sa rozhodli použiť PSO. Nakoniec sme porovnali výsledky oboch verzií a slovník anotovaný pomocou PSO dosiahol porovnateľné výsledky ako slovník anotovaný človekom.

Kľúčové slová: analýza sentimentu, slovníkový prístup, particle swarm optimization.

1 Úvod

Analýza sentimentu má v dnešnej dobe veľmi významnú úlohu. Pomáha pri rozhodovaní aký produkt si kúpiť, kontrole vlastnej značky, alebo nájdením dobrého filmu. Existuje mnoho prístupov, ktoré je možné použiť na analýzu sentimentu. Niektoré prístupy sú založené na strojovom učení. Tieto prístupy používajú metódy strojového učenia, ako napr. Naivný Bayesov kvalifikátor, metódu Podporných Vektorov, metódu Maximálnej Entropie, či k-Najbližších Susedov [5], aby klasifikovali príspevky do pozitívnej alebo negatívnej triedy. V súčasnosti sú veľmi populárne aj metódy založene na hlbokom učení. V prípade hlbokého učenia, autori používajú Neurónové siete na získanie nových atribútov alebo získanie nových informácií z textov [13, 10]. My sa v tomto príspevku zameriavame na slovníkový prístup, ktorý používa slovník emocionálnych slov na určenie polarity príspevku.

Slovníky je možné rozdeliť do troch skupín, podľa toho, ako boli vytvorené. Manuálne generované slovníky sú kvalitné, keďže boli vytvorené a hodnotené ľuďmi, ale ich generovanie je veľmi časovo náročné. Na druhú stranu, slovníky generované automaticky môžu byť vytvorené veľmi rýchlo, ale ich presnosť a kvalita je často oveľa nižšia ako pri manuálne generovaných slovníkoch. Riešením toho problému sú semi-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 151-156.

automaticky generované slovníky, kedy po automatickom vygenerovaní slovníka človek skontroluje vytvorený slovník a odstráni nepresnosti.

V tomto príspevku sa zameriavame na automatické anotovanie slovníka pomocou Particle swarm optimization (PSO) v slovenskom jazyku. Tento spôsob by mohol zjednodušiť vytváranie slovníkov v nových, resp. menej používaných jazykoch. V takom prípade by stačilo iba preložiť slovník zo svetového jazyka (angličtina) a automaticky anotovať pomocou PSO.

2 Analýza sentimentu pomocou slovníkov a evolučných algoritmov

Slovníky je možné rozdeliť na dve skupiny, podľa toho aké informácie poskytujú pre analýzu sentimentu. Jednoduché slovníky iba delia slová na pozitívne a negatívne [11, 7, 8]. Ak je požadovaná komplexnejšia analýza, je potrebné použiť sofistikovanejšie slovníky. Tie obsahujú aj dodatočné informácie, ako silu polarity [14, 12] alebo skupiny slov a vzťahy medzi nimi [5]. Taktiež môžu obsahovať aj ďalšie slovná potrebné pre analýzu textu ako intenzifikátory a negátory [11, 7, 12]. Pridanie dodatočných informácií o slovách umožňuje komplexnejšiu analýzu spracovaného textu. Pridanie sily polarity umožnilo porovnávať jednotlivé slová v závislosti na zvolenej stupnici (od 0 do 1, alebo od -5 do 5). Taktiež intenzifikácia a negácia umožňujú posúvanie sily polarity pozitívnym alebo negatívnym smerom.

Existuje aj niekoľko prác venujúcich sa použitiu evolučných algoritmov v oblasti klasifikácie textov. Genetické programovanie bolo použité v práci [3] a jeho úlohou bolo nájsť efektívnu schému na váhovanie slov, ktorá by zlepšila presnosť klasifikácie. PSO bolo použité na výber atribútov [10] pre Conditional Random Field, resp. nájdenie najužitočnejších atribútov [1], ktoré najviac prispejú ku zlepšeniu klasifikácie recenzií pomocou metódy Podporných Vektorov.

3 Základná myšlienka PSO

Particle swarm optimization (PSO) je meta-heuristický algoritmus predstavený Kennedym a Eberhartom [2], ktorý napodobňuje chovanie kŕdľa vtákov. V prípade PSO sa potenciálne nazýva jedinec (particle). Skupina jedincov sa v rámci PSO vytvára tzv. populáciu. Každý jedinec sa pohybuje v problémovom priestore. Pritom nasleduje najlepšieho jedinca a uchováva si svoje najlepšie riešenie pre daný problém (*pbest*), ktoré je získané pomocou fitness funkcie. Výberom najlepšieho riešenia spomedzi skupiny jedincov získame lokálne najlepšie riešenie (*lbest*). Globálne najlepšie riešenie (*gbest*) predstavuje najlepšie riešenie z celej populácie. PSO sa skladá z dvoch základných krokov: 1. zmena rýchlosti smerom k *pbest* a g*best* a 2. zmena aktuálnej pozície. Populácia môže byť popísaná ako vektor $X_i = (x_{i1}, x_{i2}, ..., x_{iD})$. Rýchlosť jedinca môžeme popísať ako $V_i = (v_{i1}, v_{i2}, ..., v_{iD})$ a najlepšiu predchádzajúcu pozíciu jedinca ako $P_i = (p_{i1}, p_{i2}, ..., p_{iD})$. Jedinec g reprezentuje najlepšieho jedinca

v populácii a hodnota *w* predstavuje tzv. lenivostný faktor. Zmeny rýchlosti a pozície sa teda vykonávajú na základe rovníc 1, 2:

$$v_{id}^{n+1} = wv_{id}^{n} + c_1 r_1^n (p_{id}^n - x_{id}^n) + c_2 r_2^n (p_{gd}^n - x_{id}^n) \quad (1)$$

$$x_{id}^{n+1} = x_{id}^n + v_{id}^{n+1} \quad (2)$$

kde $d = 1, 2, ..., D$ je rozmer vektora, $i = 1, 2, ..., N$, kde N je počet jedincov, $n = 1, 2, ...$, je počet iterácií. Rovnice tiež obsahujú dve náhodné hodnoty r_1, r_2, ktoré zabezpečujú aby funkcia neupadla do lokálneho optima. Parametre c_1 a c_2 predstavujú faktory ovplyvňujúce pohyb voči *pbest* a g*best*. PSO skončí v momente, keď je dosiahnuté ukončovacie kritérium, alebo sa naplní počet cyklov určených na začiatku.

4 Navrhovaná metóda

V tejto práci sme vytvorili dve verzie slovníka pre analýzu sentimentu v slovenskom jazyku. Slovník bol manuálne preložený z jeho anglickej verzie [7], ktorá obsahovala 6789 slov. Tieto slová sme preložili do slovenčiny a pre každé slovo sme vyhľadali jeho synonymá a antonymá pomocou synonymického slovníka. Finálna verzia slovníka tak obsahuje 598 pozitívnych slov a 772 negatívnych slov. Stupnicu pre určenie sily polarity sme zvolili v rozmedzí od -3 (najviac negatívne slovo) do 3 (najviac pozitívne slovo).

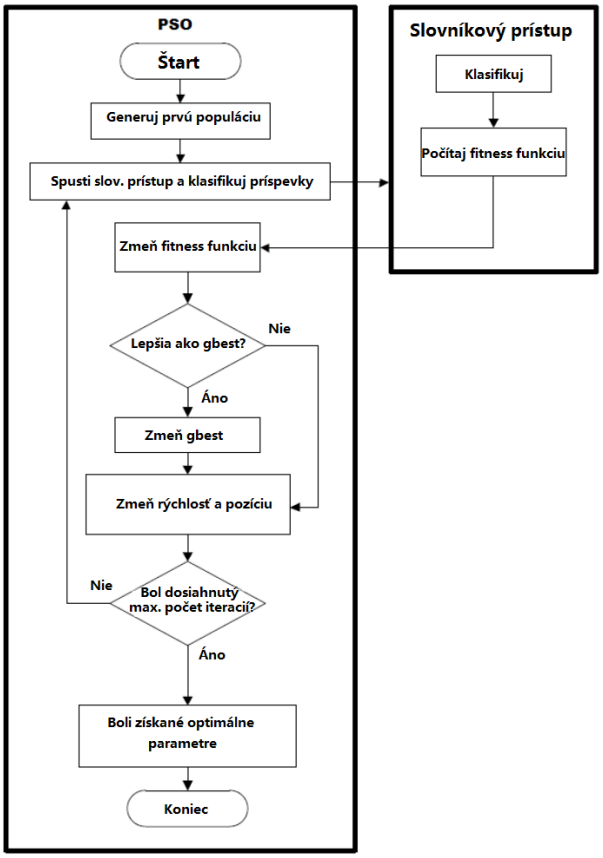
Pre automatickú anotáciu sme sa rozhodli použiť PSO. Ako ukázali viaceré výskumy, jedná sa o efektívny, robustný a jednoduchý optimalizačný algoritmus, ktorý bol aplikovaný na mnoho optimalizačných funkcií. Každý jedinec v našom prípade môže byť zapísaný vo forme vektor $X_i = (x_{i1}, x_{i2}, ..., x_{iD})$ kde $x_{id} \in \{-3,3\}$, $i = 1, 2, ..., N$, kde N je počet jedincov v populácii a $d = 1, 2, ..., D$ označuje počet slov v slovníku. Rozsah od -3 do 3 sme zvolili rovnaký, ako v prípade ľudského anotátora. Celý proces použitia PSO je možné vidieť na obrázku 1.

Prvá generácia bola generovaná náhodne. Ostatné parametre boli získané na základe experimentov a dostupnej literatúry. Tie boli nastavené nasledovne:

- počet jedincov: 15000
- počet opakovaní: 100
- $w = 0.729844$
- $c_1 = 1.49618$
- $c_2 = 1.49618$

Ako fitness funkciu sme použili makro-F1 mieru. Tá sa používa na vyhodnocovanie nevyvážených datasetov. Makro-F1 miera sa vypočíta ako priemer F1 mier pre jednotlivé triedy (pozitívnu a negatívnu). F1 miera je harmonický priemer medzi presnosťou a návratnosťou. Tie boli vyčíslené na základe výsledkov klasifikácie pomocou slovníkového prístupu. Slovníkový prístup skombinoval hodnoty polarity generované pomocou PSO so slovami zo slovníka a vytvoril dočasný slovník, ktorý bol použitý na ohodnotenie príspevkov v testovacom datasete. Na získanie polarity príspevku bol použitý algoritmus, ktorý vyhľadával slová príspevku v slovníku a na základe prira-

denej polarity upravil silu polaritu príspevku. Výsledná polarita bola porovnaná s polaritou uvedenou v datasete. Na základe toho porovnania boli vyčíslené hodnoty presnosti a návratnosti.



Obr. 1. Diagram učenia sily polarity slov pomocou PSO

5 Experimenty a výsledky

Predstavený prístup sme testovali na dvoch datasetoch. Prvý (všeobecný dataset) obsahuje 4720 príspevkov z rôznych oblastí (hodnotenie filmov, kníh a elektroniky, politika, atď.), ktoré boli získané z viacerých webových stránok. V datasete sa nachádza 2455 pozitívnych a 2265 negatívnych príspevkov. Druhý dataset (filmový dataset) bol preložený z angličtiny pomocou Google tranlator[1] a jedná sa o dataset vytvorený v práci Pang a Lee [9]. Tento dataset obsahuje 1000 pozitívnych a 1000 negatívnych

[1] https://translate.google.sk/?hl=sk&tab=wT

filmových hodnotení zo stránky rottentomatos.com. Datasety boli v procese učenia nahodne rozdelené v pomere 90:10, teda PSO hľadalo optimálne riešenie na 90% príspevkov a následne bolo toto riešenie použité na 10% neanalyzovaných dát. Výsledky experimentov je možné vidieť v tabuľke 1.

Tab. 1. Porovnanie F1 mier pre obidve verzie slovníka.

	všeobecný dataset	filmový dataset
verzia anotovaná človekom	**0,7668**	0,6294
verzia anotovaná pomocou PSO	0,7647	**0,7292**

Výsledky ukazujú, že PSO dosiahlo výsledky porovnateľné s človekom, resp. ho dokázalo trochu prekonať. Na všeobecnom datasete dosiahlo PSO výsledok veľmi blízky ľudskému anotátorovi. Na filmovom datasete dokázalo PSO prekonať človeka, čo znamená, že dokázalo nájsť vhodnejšie sily polarity pre jednotlivé slová ako človek.

6 Záver

V tomto príspevku sme predstavili prístup na analýzu sentimentu pomocou slovníkovej metódy, ktorá použila slovník anotovaný pomocou PSO. Tento slovník bol vytvorený prekladom z jeho angličtiny a v prvej verzii bol anotovaný manuálne, čo bolo časovo náročné. Druhá verzia slovníka bola anotovaná pomocou evolučného algoritmu, konkrétne PSO. Ako ukázali výsledky, verzia anotovaná pomocou PSO dosiahla v prípade všeobecného datasetu porovnateľné výsledky ako verzia anotovaná človekom a v prípade filmového datasetu lepšie výsledky ako verzia anotovaná človekom.

Do budúcna by sme chceli vylepšiť PSO použitím ďalších metód úpravy pozície a polohy a taktiež otestovať na väčšom datasete.

Literatúra

1. Basari, S. H., a kol.: Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering. 53, 453--462 (2013).
2. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, (1995), pp. 39--43.
3. Escalante, H. J., a kol.: Term-weighting Learning via Genetic Programming for Text Classification. Know.-Based Syst. 83, 176--189 (2015).
4. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06). 417--422 (2006).
5. Go, A.: Twitter Sentiment Classification using Distant Supervision. Stanford University, Available on: http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf, (2013).

6. Gupta, D. K., a kol.: PSO-ASent: Feature Selection Using Particle Swarm Optimization for Aspect Based Sentiment Analysis. NLDB. 220--233 (2015).
7. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04). ACM, New York 167--177 (2004).
8. Mohammad, S. M., Turney, P. D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence 29, 436--465 (2013).
9. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA, (2004).
10. dos Santos, C. N., Gattit, M.: Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. 69--78 (2014).
11. Stone, P., a kol.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press., (1966).
12. Taboada, M., a kol.: Lexicon-based Methods for Sentiment Analysis. Computational Linguistics. 267--307 (2011).
13. Tang, D., a kol.: Coooolll: A Deep Learning System for Twitter Sentiment Classification. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 208--212 (2014).
14. Warriner, A. B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods. 45, 1191--1207 (2013).

Poďakovanie:

Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-16-0213 a Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR, projekt č. 025TUKE-4/2015.

Annotation:

Annotation of dictionaries using PSO

The social web contains a huge amount of text with human emotions and opinions. It is very difficult to analyze them manually and in this paper, we use automatic sentiment analysis based on dictionaries. We created two versions of the dictionary. The first version was prepared manually and annotated by human, which was time-consuming. The second version used Particle Swarm optimization (PSO) for the annotation of words in the dictionaries. Both versions were tested on two datasets, a generated dataset and a movie dataset. We compared the results of the two versions and the version annotated by PSO achieved comparable results with the version annotated by human.