Entitásorientált véleménydetekció webes híranyagokból

Hangya Viktor, Farkas Richárd, Berend Gábor

Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail:{hangyav,rfarkas,berendg}@inf.u-szeged.hu

Kivonat Napjainkban a hírközlés jelentős hányada digitális formában történik, a híranyagokban említett entitásokra vonatkozó vélemények polaritásának automatikus meghatározása pedig komoly előnyökkel járhat. Éppen ezért munkánk során az OpinHuBank adatbázisban található entitásokra vonatkozó vélemények bekategorizálását tűztük ki feladatunkul. A javasolt megoldásunk többek között a szövegegységek dependenciaelemzésére is támaszkodva képes az entitások mondatbeli szerepének figyelembevételével pontosabb képet adni a rájuk vonatkozó véleményekről.

1. Bevezetés

A hírportálok egyre növekvő száma és kibocsátási aktivitása mind nagyobb mértékben teszi lehetővé számunkra, hogy a híranyagokban közölt entitásokkal kapcsolatos véleményeket megfigyeljük. Az ezen entitásokra (például cégekre vagy politikai szereplőkre) vonatkozó vélemények monitorozása hasznos segítséget nyújthat azok pozitív reputációjának megtartásában, felépítésében [1].

Mivel az online tartalmak száma egyre gyarapszik, így ezen feladat hatékony elvégzésére csupán automatikus rendszereket használva nyílik lehetőségünk. Munkánk során éppen ezért az OpinHuBank adatbázisát fölhasználva építettünk az entitásokkal kapcsolatos hírek polaritását detektáló rendszert. Rendszerünket úgy terveztük, hogy az a célentitásokra nézve képes legyen – az azt tartalmazó mondat mélyebb elemzésének végrehajtása által – az entitásra vonatkozó polaritás minél pontosabb meghatározására.

Az általunk kitűzött feladat nehézségei közül kiemelendő, hogy amennyiben egy mondat több entitást is tartalmaz, úgy könnyen megeshet, hogy azok egy része pozitív, míg másik részük negatív (vagy akár semleges) kontextusban kerül említésre, vagyis a mondatban található pozitív, illetve negatív konnotációjú szavak nem egyforma mértékben képesek befolyásolni egy-egy entitás véleménytöltetének végső polaritását. Példaképp tekintsük az OpinHuBank-ban található következő mondatot:

> „A befutó *Tóth Adrienn* az első sorozatban többet is hibázott, a németek utolsó versenyzője, az olimpiai bajnok *Lena Schöneborn* viszont jól célzott, s 10 mp-re csökkentette a különbséget."

Jól látható, hogy míg a mondat által közvetített vélemény Tóth Adriennre vonatkozóan negatív, addig Lena Schönebornra nézve ezzel éppen ellentétes. Az előzőekkel összhangban elmondható, hogy a *hibázott* szó jelenléte nem bír befolyásoló erővel a mondat Lena Schönebornnal kapcsolatban megfogalmazott állítás polaritására nézve. A fenti jelenség kezelésére a mondatok dependenciaelemzésének a modellünkbe történő beépítése mellett döntöttünk.

2. Korábbi munkák

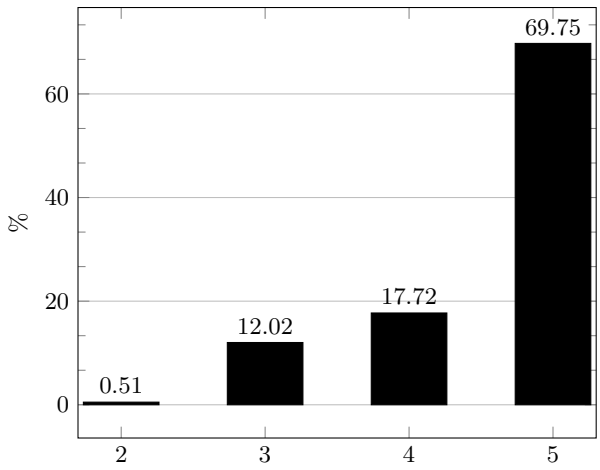
Az elmúlt években a véleménydetekció feladatát nagyfokú tudományos érdeklődés övezte, számos versenyfeladatot tűztek ki témájában [1,2]. Az angol nyelvű szövegek feldolgozására vonatkozó munkák [3,4,5] mellett a magyar nyelvű szövegekből történő véleménydetekcióra is egyre nagyobb figyelem irányul, köszönve a benne rejlő gazdasági lehetőségeknek. Cikkünkben olyan módszert mutatunk be, mely lehetőséget nyújt az egyes entitásokra irányuló vélemények beazonosítására, illetve monitorozására, azaz a velük kapcsolatos hírekben található vélemények polaritásuk szerinti kategorizálására.

Munkánk a korábbi versenykiírások közül a 2014-es SemEval aspektus-alapú véleményanalízisre vonatkozó feladatával rokonítható leginkább. A versenyfeladat kapcsán éttermekről és laptopokról szóló termékvéleményezések mondataiból az azokban említett főnévi csoportok formájában jelentkező aspektusokra (pl. kiszolgálás minősége) vonatkozó vélemények polaritását kellett meghatározni. A cél ebben az esetben tehát nem egyszerűen dokumentum- vagy mondatszintű véleményértékek meghatározása volt, hanem aspektus-mondatpárokra vonatkozóan kellett a megfelelő döntéseket meghozni. Utóbbi feladat kétségkívül nehezebb, hiszen egy mondat nyilatkozhat teljesen tárgyilagosan (vagy akár pozitívan is) a kiszolgálás minőségéről, miközben adott esetben negatív véleményeket fogalmazhat meg a termék vagy szolgáltatás árára vonatkozóan. Az ilyen és ehhez hasonló esetek kezelése a mondatok szerkezeti elemzése nélkül nem vagy csak korlátozott mértékben képzelhető el. Munkánk során mi is a verseny szervezői által megfogalmazott célhoz hasonlót tűztünk ki magunk elé, azzal a különbséggel, hogy esetünkben az OpinHuBank adatbázisban található entitások képezték a mondatok azon elemeit, melyekkel kapcsolatosan a véleménytöltet polaritását meg kívántuk határozni, nem pedig termékvéleményezések – jellemzően köznévi – aspektusai.

Módszerünk kidolgozásához és validálásához a 2013-ban létrehozott OpinHuBank [6] szöveges adatbázist használtuk fel, mely egy szabadon hozzáférhető annotált magyar nyelvű korpusz véleményelemzéshez. Az adatbázisban olyan mondatok találhatók, melyek mindegyike egy előre megadott entitással kapcsolatosak. Feladatunk tehát az volt, hogy ezen mondat-entitás párokra meghatározzuk, hogy az adott mondat tartalmaz-e az entitásra nézve pozitív vagy negatív információt. Az adatbázis közel háromnegyede hírportálok, illetve hírügynökségek oldalairól lett összegyűjtve, a fennmaradó dokumentumok forrásai blogok voltak.

3. Módszerek

Különféle modelljeinket az OpinHuBank adatbázison értékeltük ki. Az adatbázis 10006 entitás-mondatpárra tartalmazza az egyes mondatokban az entitásokra vonatkozó vélemény polaritását, melyek pozitív, negatív avagy semleges besorolásba eshetnek. Az entitás-mondatpárokat öt független annotátor sorolta be a három kategória valamelyikébe, mely feladatra vonatkozó egyezés mértékét az 1. ábra foglalja össze. Az ábrából kitűnik, hogy az entitás-mondatpárok közel 70%-ára tökéletes egyezés mutatkozott mind az öt annotátor között. A fennmaradó entitás-mondatpárok kapcsán az osztályozó modellünk tanítása és kiértékelése során osztálycímkékül az annotátorok leggyakoribb döntését vettük. Ahogy az az 1. ábrából kitűnik, az esetek több, mint 29%-ában egyértelműen meghatározható volt az annotátorok által választott osztálycímkék leggyakrabbika, és mindössze 0,5%-ban alakult ki holtverseny az annotátorok döntései kapcsán. Ezekre az esetekre – ellentmondásos jelölésükből adódóan – a továbbiakban semleges egyedekként tekintettünk. Ilyen módon az OpinHuBank adatbázisban található entitás-mondatpároknak rendre 74,9, 16,3, valamint 8,8%-ára tekintettünk semleges, negatív, valamint pozitív címkéjű példaként további vizsgálataink során. A valamilyen polaritással bíró 2511 entitás-mondatpár esetében pedig 64,9, valamint 35,1% mutatkozott negatív, illetőleg pozitív osztálybelinek. Modelljeinket egyaránt kiértékeltük azokban az esetekben, ahol feladatunkul az entitás-mondatpárok három, illetve két osztályba sorolását tűztük ki célunkul. Utóbbi esetben az adatbázisban semlegesnek mutatkozó egyedek kihagyásával tanítottuk, illetve értékeltük ki modelljeinket, melyek eredményeit a három-, illetve kétosztályos tanítás kapcsán egyaránt tízszeres keresztvalidáció eredményeként közöljük.



1. ábra. Entitás-mondatpáronként a leggyakoribb címkét választó annotátorok számának eloszlása

Munkánk során a véleményérték meghatározásának feladatára felügyelt osztályozási problémaként tekintettünk, melyben a szövegek (legfeljebb[1]) három osztályba tartozhattak annak függvényében, hogy azok egy célentitásra nézve pozitív, negatív vagy semleges információt hordoztak. Maximum entrópia alapú modelljeink paramétereinek meghatározásához a MALLET gépi tanulási programcsomagot [7] használtuk.

Az entitások kontextusának kategorizálása során nem csupán a környező szavak, szókapcsolatok jelenlétét vettük figyelembe, hanem azoknak az entitásokhoz vett relatív pozíciójának alapján történő **súlyozás**t is alkalmaztunk a modellalkotás során. Egy n hosszúságú – a célentitást az i. tokenpozíción tartalmazó – mondat j. tokenjéhez rendelt jellemző értékét a

$$\frac{1}{e^{\frac{1}{n}|i-j|}}$$

formula által határoztuk meg, így juttatva nagyobb fontossághoz a célentitás környezetében fellelhető szavakat.

Az osztályozás során az entitások kontextusának vizsgálata alkalmával a környező és kapcsolódó szavak pozitív, illetve negatív voltát is figyelembe vettük. Ennek elvégzéséhez azonban pozitív és negatív szavakat tartalmazó listákra volt szükségünk, melyekből angol nyelven több lexikon (pl. [8]) is rendelkezésre áll, ugyanakkor a magyar nyelvű szövegek esetében saját **polaritáslexikon**ok előállítására volt szükség [9]. A lexikon létrehozása során egy 2676 negatív, valamint 646 pozitív szót tartalmazó lista lefordítására került sor. A lista segítségével az egyes mondatokban található pozitív, illetve negatív szavak számából, illetve a mondatban szereplő listaelemekből külön is hoztunk létre jellemzőket ennek a jellemzőcsoportnak a kapcsán.

Mivel egyes mondatok több, adott esetben eltérő polaritású véleménnyel illetett entitást is tartalmazhatnak, ezért olyan módszerekre is támaszkodtunk, melyek képesek szeparálni a különböző entitásokra vonatkozó véleményeket, majd ezek közül kiválasztani a vizsgált entitásra nézve relevánsakat. A szövegek **dependenciaelemzés**ére támaszkodva lehetőségünk nyílt az egyes szavak pozícióin túlmenően azok nyelvtani szerepének figyelembevételére is. A dependenciaelemzés végrehajtására a `magyarlanc` [10] eszközt használtuk, egy célentitás kapcsán pedig a mondat dependenciagráfjának gyökerétől az adott entitásig elhelyezkedő szavak képezték a jellemzőcsoport által bevezetett jellemzők halmazát. Az előzőekben bemutatott, és az eredmények ismertetése során hivatkozott jellemzőcsoportokra vonatkozó rövidítéseinket a 3. táblázat foglalja össze.

4. Diszkusszió

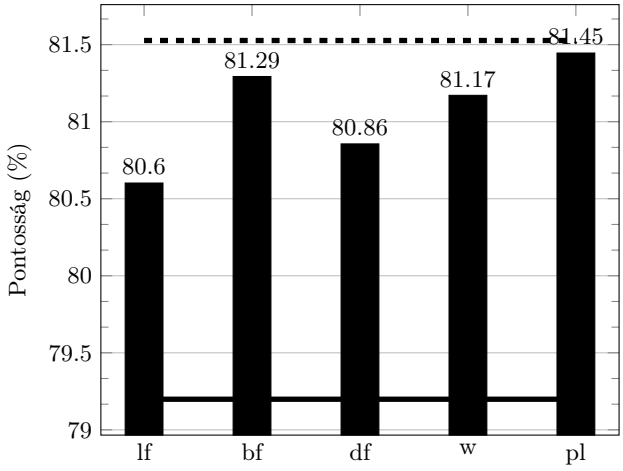
Ablációs kísérleteinket két-, illetve háromosztályos osztályozás kapcsán is kiértékeltük, melyek eredményeit a 2., illetve a 3. ábra tartalmazza. A 2. ábrán a

[1] Kétosztályos modelljeink esetében csak a pozitív, illetve negatív osztályokba történő besorolás volt a célunk.

1. táblázat. A jellemzőcsoportokra vonatkozó rövidítések

Rövidítés	Jelentés
lf	lexikai jellemzők (uni,-és bigramok)
bf	bigram jellemzők
df	dependencia alapú jellemzők
w	távolság alapú súlyozás
pl	polaritáslexikonból származó jellemzők

háromosztályos feladat megoldása során mért eredményeink láthatók, melyeket a 3. ábrán látható eredményekkel összevetve megállapíthatjuk, hogy – ahogy a várakozásoknak megfelelően – a semleges vélemények felismerését is magában foglaló feladat némileg nehezebbnek tekinthető a csupán a pozitív, illetve negatív vélemények elkülönítését megcélzó feladattól. Szaggatott vonallal azon rendszerek pontossága látható, melyek a 3. fejezetben bemutatott jellemzőtípusok mindegyikét egyidejűleg használták, folytonos vonallal pedig a – csak unigram jellemzőket használó – baseline rendszerünk eredményessége látható. Az egyes oszlopokhoz társított értékek pedig azt mutatják, hogy mennyiben változik az osztályozási feladat során elért eredményességünk, amennyiben egy-egy jellemzőcsoportot nem építünk be a jellemzőterünkbe.

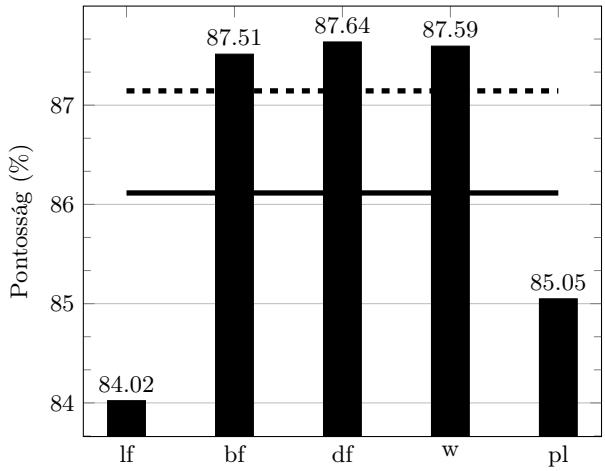


2. ábra. Ablációs kísérleteink pontosságértékei a háromosztályos feladat kapcsán

Az ábrák alapján mindkét feladat kapcsán kijelenthető, hogy a modellekből az lf jelzéssel ellátott – lexikális (uni- és bigram) – jellemzőket elhagyva figyelhető meg az osztályozási pontosság legnagyobb mértékű csökkenése. Az is elmondható, hogy ez a csökkenés javarészt az unigram jellemzők elhagyásának számlájára írható, ha ugyanis az eredményesség változásának mértékét össze-

vetjük a bf jelzésű – kizárólag a bigram jellemzőket mellőző – oszlopokéval, az eredmények hasonló fokú romlása nem volt tapasztalható.

A 2. ábra további vizsgálatából kitűnik, hogy a lexikális jellemzőket követően a dependenciaelemzésből előálló jellemzők elhagyása okozta a teljesítmény legnagyobb fokú romlását, így kijelenthető, hogy ezek alkalmazása a végső rendszer eredményességéhez nagyban hozzájárult. A dependenciaelemzésből származtatott jellemzőket követően leginkább a távolságalapú súlyozás tudott hozzájárulni a végső eredményességhez, a legkisebb hozzáadott értéke pedig a polaritásszótár használatának volt a háromosztályos feladat megoldása során az ablációs kísérleteink tanúsága szerint.

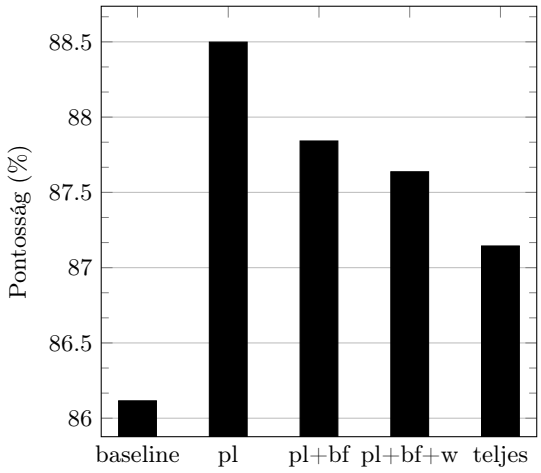


3. ábra. Ablációs kísérleteinek pontosságértékei a kétosztályos feladat kapcsán

A 2. ábrának a 3. ábrával való összevetése rámutat a két különböző – a három-, illetve kétosztályos (a neutrális osztályt tartalmazó, illetve mellőző) – tanulási feladat nagyfokú különbözőségére. Ugyan a leghasznosabbnak mindkét esetben a lexikális (főképp az unigram) jellemzők mutatkoztak, a további jellemzőcsoportok relatív hasznosságának sorrendje jelentősen eltér. Míg a háromosztályos esetben például a polaritáslexikonnak volt tulajdonítható a legkisebb szerep, addig a kétosztályos esetben a teljes rendszer szempontjából ezen jellemzők szerepe elsődlegesnek tekinthető. Ez persze nem meglepő, ha figyelembe vesszük, hogy a háromosztályos esetben a példányok közel háromnegyede a neutrális osztályba tartozó volt, a kétosztályos esetben pedig mindezen példáktól eltekintettünk, így ott kizárólag valamilyen polaritással bíró egyedek szerepeltek már csupán. Meglepő eredményként azt tapasztaltuk, hogy a bigramok figyelembevételéből, a dependenciaelemzésből, valamint a távolságalapú súlyozásból származtatott jellemzők használata – noha a háromosztályos feladat esetében mind kivétel nélkül hozzá tudtak járulni a végső rendszer eredményességéhez – nem bizonyultak

elég hatékonynak a kétosztályos feladat megoldása során, hiszen az bármelyikét elhagyva javulást tapasztaltunk ahhoz a rendszerhez képest, amelyik az összes jellemzőcsoportot egyidejűleg használta. Ezek alapján úgy tűnik, hogy alkalmazott jellemzőink egy része alkalmasnak mutatkozik annak megítélésére, hogy egy adott entitás neutrális kontextusban kerül-e említésre vagy sem, ugyanakkor kevésbé alkalmasak – a polaritás megléte esetén – annak pozitív vagy negatív voltának meghatározására.

Előző megfigyeléseinkből kiindulva a kétosztályos feladat kapcsán a jellemzőcsoportokat relatív hasznosságuk szerint sorrendbe állítva külön modelleket hoztunk létre. Ennek során megvizsgáltuk, hogy a baseline rendszer jellemzőterének – az egyre csökkenőnek mutatkozó hasznosságú jellemzőcsoportokkal történő – fokozatos bővítése milyen módon befolyásolja az eredményeket, melynek eredménye a 4. ábrán látható. A kétosztályos tanulásra vonatkozó ábrák összevetéséből magyarázatot kaphatunk arra a kontraintuitív jelenségre is, hogy a teljes jellemzőtérből csupán a polaritáslexikonra támaszkodó jellemzők elhagyásával hogyan kaphattunk a baseline rendszernél gyengébb teljesítményt (3. ábra folytonos egyenese, illetve 5. oszlopa). Ebben az esetben ugyanis az unigramokból származó jegykészleten túl csupa olyan jellemzőcsoport által került kialakításra a jellemzőkészletünk, amelyek a 4. ábra tanúsága szerint nem képesek javítani az osztályozás pontosságán.



4. ábra. A kétosztályos kiértékelés kísérleteinek pontossága

5. Konklúzió

Megközelítésünk alapvetően a szövegekben előforduló szavakon és szópárosokon alapul, azonban a pontosabb eredmények eléréséhez további – többek között a szövegek dependenciaelemzéséből, illetve polaritáslexikon alapján meghatározott – információk vizsgálatát is végrehajtottuk. A kidolgozott módszerek segítségével

a 80%-os pontosságértéket meghaladva sikerült javítani az egyszerű unigram alapú véleménykinyerésre vonatkozó osztályozás eredményein. Ezen felül olyan módszereket dolgoztunk ki, melyek hasznosak lehetnek más véleménydetekciós feladatok megoldása során is.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Amigó, E., de Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., Spina, D.: Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In: Proceedings of Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative (CLEF 2013). (2013) 333–352
2. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Association for Computational Linguistics and Dublin City University (2014) 27–35
3. Hangya, V., Berend, G., Varga, I., Farkas, R.: SZTE-NLP: aspect level opinion mining exploiting syntactic cues. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, Association for Computational Linguistics and Dublin City University (2014) 610–614
4. Hangya, V., Farkas, R.: Filtering and polarity detection for reputation management on tweets. In: Working Notes of CLEF 2013 Evaluation Labs and Workshop. (2013)
5. Hangya, V., Farkas, R.: Target-oriented opinion mining from tweets. In: Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, IEEE (2013) 251–254
6. Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013) 343–345
7. McCallum, A.K.: MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
8. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). (2010)
9. Hangya, V., Farkas, R.: Doménspecifikus polaritáslexikonok automatikus előállítása magyar nyelvre. In: MSzNy 2015 – XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2015)
10. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 368–374