Doménspecifikus polaritáslexikonok automatikus előállítása magyar nyelvre

Hangya Viktor, Farkas Richárd

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

Kivonat Napjainkban a közösségi média jelentős népszerűségre tett szert, szinte bármilyen témakörben nagy mennyiségű szöveg érhető el. Ennek köszönhetően nagy figyelmet kaptak a különböző véleménydetekciós módszerek, melyek feladata szövegek osztályozása azok tartalmának polaritása alapján. A feladat megoldása során segítséget nyújtanak az ún. polaritáslexikonok, melyek az egyes szavak polaritására nézve hordoznak információkat. Munkánkban bemutatunk különböző módszereket lexikonok előállítására, valamint azok kiegészítésére és adaptálására más doménekre. Vizsgálatainkat kifejezetten számítástechnikai eszközökkel kapcsolatos véleményeken és általános hírekből származó szövegeken végeztük el, melyekből kiderül, hogy az osztályozás pontosságára nézve a megfelelő lexikon kiválasztása meghatározó.

Kulcsszavak: véleménydetekció, lexikon, természetesnyelv-feldolgozás

1. Bevezető

A közösségi média elterjedésével az embereknek lehetőségük nyílt különböző tartalmak megosztására más internetfelhasználókkal. Ennek köszönhetően nagy mennyiségben érhetőek el olyan elektronikus tartalmak, melyekben az egyes felhasználók véleményeiket fejezik ki bizonyos termékekről, ismert személyiségekről vagy cégekről. Az utóbbi években nagy figyelmet kaptak a különböző automatikus véleménydetekciós módszerek [1,2], melyeknek feladata dokumentumok polaritásának meghatározása.

A feladatra tekinthetünk úgy, mint egy osztályozási probléma, mely során egy dokumentumot pozitív vagy negatív osztályba kell sorolni. Nagy mennyiségű jelölt adat esetén egy jól bevált módszer a szövegekben előforduló szavak, esetleg szópárosok alapján történő felügyelt gépi tanulás alapú osztályozás. Ilyenkor az egyes szavak polaritását a tanító adat segítségével határozzuk meg. A módszer azonban nem kellően pontos abban az esetben, ha kevés számú tanító adat áll rendelkezésünkre, mivel az egyes szavak polaritásának megbecslése pontatlan lehet. Másik probléma, hogy sok szó nem fordul elő a tanító szövegekben, így azok polaritásáról nem tudunk mondani semmit. Az ún. polaritáslexikonok nyújtanak segítséget ilyen esetekben, melyek az egyes szavak polaritását előre adott módon tartalmazzák, így a tanítás során nem látott szavak polaritását tekintve is vannak információink. Angol nyelvre számos általános célú lexikon érhető el, magyar nyelv esetén azonban nem ilyen kedvező a helyzet.

Egyik kézenfekvő lehetőség a már létező, idegen nyelvű lexikonok lefordítása. A módszernek azonban több hátránya is van. Egyrészt a folyamat időigényes és költséges. A többjelentésű szavak különböző jelentései más-más polaritással rendelkezhetnek az egyes doménekben, ezért szükség van doménspecifikus lexikonokra, melyek a kérdéses szövegek esetén jól használhatóak. Az idegennyelvű lexikonok lefordításának másik nehézsége az, hogy az elérhető lexikonok általános célúak. Munkánkban megmutatjuk, hogy az ilyen általános célú lexikonok nem teljesítenek jól speciális doménekben. Olyan módszereket dolgoztunk ki, melyek lehetőséget nyújtanak doménspecifikus lexikonok létrehozására. A javasolt módszerek egy része kicsi, kézzel összeállított lexikon kiegészítésére alkalmas, ahol a kulcs hasonló jelentéssel bíró szavak automatikus gyűjtése. Ezen felül adott doménből származó szöveges adatok segítségével tejesen új lexikont hoztunk létre automatikus módszerrel.

Vizsgálatainkat számítástechnikai eszközökkel kapcsolatos véleményeken és általános hírekből [3] származó szövegeken is elvégeztük. Az eredményekből kiderül, hogy az osztályozás pontossága nagyban függ attól, hogy a felhasznált lexikon mely doménből származik. Az általános lexikonokkal ellentétben, a doménspecifikusak használata szignifikáns hibacsökkenést eredményez.

2. Polaritáslexikon

Nagy jelölt adathalmaz esetén lehetőség van az egyes szavak polaritásának meghatározására gépi tanulási módszerek segítségével. Ezt a tudást felhasználva korábban még nem látott szövegek esetén eldönthető azok véleménytartalma. Abban az esetben azonban, ha a jelölt adatok száma kicsi, ezek a módszerek pontatlanabbá válnak. Ilyenkor érdemes használni polaritáslexikonokat, melyek tartalmazzák a polaritással rendelkező szavakat. Ez az előtudás felhasználható az osztályozás során például arra, hogy megtudjuk, hány darab pozitív, illetve negatív szó szerepel az adott dokumentumban.

Figyelembe kell azonban venni azt a tényt, hogy az egyes szöveghalmazok különböző doménekből származhatnak. Bizonyos szavak doménenként eltérő polaritással rendelkeznek. Tekintsük a következő példákat:

– A mixer használata egyszerű és **halk** működés közben.
– Ennyi pénzért nekem túl **halk**, nagyobb hangerőre számítottam.

Az első mondat konyhai eszközökkel kapcsolatos és a *halk* szó pozitív töltetű. Ezzel szemben, a második mondat egy hangszóró értékelése, mely során a szó már negatív polaritású. Fontos tehát, hogy a véleménydetekciós feladat elvégzéséhez a megfelelő doménből származó lexikont használjuk. Mivel legtöbb esetben nem áll rendelkezésre a megfelelő lexikon és azok manuális előállítása költséges, statisztikai módszerekkel újakat kell létrehozni vagy meglévőeket kell adaptálni.

A következőkben bemutatásra kerülnek különböző megoldások lexikon létrehozására, illetve kiegészítésére, valamint ezek problémáira is rávilágítunk.

2.1. Idegen nyelvű lexikon lefordítása

Sajnos magyar nyelvre nem áll rendelkezésre szabadon elérhető (referencia) polaritáslexikon. Angol nyelvben azonban több általános célú is megtalálható [4,5]. Így egy értelemszerű megoldás egy meglevő lexikon lefordítása. Ehhez egy már angol nyelven működő véleménydetekciós rendszerben használt lexikont vettünk alapul, mely 3322 szót tartalmaz. A lexikonban az egyes szavak a $[-5, 5]$ intervallumon vett értékkel vannak jellemezve polaritásuktól függően. Az angol szavak fordítását kézzel végeztük és az adott szó összes magyar megfelelőjét felvettük a **fordított** lexikonba.

A módszernek azonban több hátránya is van. Először is a fordítás költséges és időigényes, főként abban az esetben, ha a lefordítandó lexikon mérete nagy. Többjelentésű szavak esetében sokszor nem egyértelmű, hogy melyik jelentés használandó. Például a *terrific* szó (*nagyszerű*, *szörnyű*) két ellentétes polaritású jelentéssel rendelkezik. Az eredeti lexikonban lévő érték segítségével kiválasztható a megfelelő jelentés. A *cool* szó (*hűvös*, *menő*) esetében azonban már nehezebb a döntés, mivel mindkettő jelentés lehet pozitív töltetű. Másik probléma, hogy az elérhető idegen nyelvű lexikonok általános célúak, melyek a kívánt doménben nem megfelelően használhatóak a már ismertetett okok miatt. Így fordítás során figyelembe kell venni ezt a tényt is, bizonyos esetekben szükség van az eredeti polaritási értékek megváltoztatására.

2.2. Polaritáslexikon bootstrapping módszerrel

A fent említett problémák megoldására több automatikus módszert dolgoztunk ki. Az első lexikonok létrehozására alkalmas jelölt szöveghalmaz segítségével. Egy adott w szó polaritása a következő módon számítható [6]:

$$pol(w) = PMI(w, pozitív) - PMI(w, negatív) \tag{1}$$

ahol PMI a páronkénti kölcsönös információt jelenti az adott polaritásra nézve. Számítása a következő képlettel adható meg:

$$PMI(w, p) = log_2 \frac{freq(w, p) * N}{freq(w) * freq(p)} \tag{2}$$

ahol $p \in \{pozitív, negatív\}$ a kérdéses polaritás, $freq(w, p)$ megadja a w szó p polaritású szövegekben való előfordulásainak számát, $freq(w)$ és $freq(p)$ a w szó összes előfordulásainak száma, illetve a p polaritású szövegek száma a korpuszban, N pedig a különböző szóalakok száma. Ezzel a módszerrel a korpuszban előforduló összes szóalakra kapunk egy polaritási értéket, mely a szavak adott doménen belüli megfelelő polaritását fogják tükrözni. Az értékeket úgy skáláztuk, hogy azok a $[-5, 5]$ intervallumba essenek. A továbbiakban ezzel a módszerrel előállított lexikonokra a **pmi** névvel fogunk hivatkozni.

Szükség van azonban a kérdéses doménből vagy egy ahhoz hasonlóból származó jelölt adathalmazra, melyen a statisztikai számítások elvégezhetőek. Abban az esetben, ha csak jelöletlen adathalmaz áll rendelkezésre, egy pontatlanabb

módszerrel osztályozhatjuk azt, majd ezen az automatikusan jelölt adathalmazon számíthatunk polaritáslexikont. Munkánk során végrehajtottunk egy ilyen kísérletet is, melynek eredményeit a 4. fejezetben részletezünk.

2.3. Lexikon kiterjesztése

A következőkben olyan módszereket mutatunk be, melyek kis számú szóval rendelkező lexikonokból kiindulva, bizonyos összefüggéseket felhasználva egy kiterjesztett lexikont adnak eredményül. A módszerek alapja, hogy az egyes elemekhez hasonlóan viselkedő szavakat gyűjtünk. Ily módon egyrészt a lexikonban levő szavak számát tudjuk növelni, másrészt pedig lehetőség van a lexikon domén adaptálására.

A módszerek bemenetként egy kiindulási lexikont kapnak. Mivel ilyen doménspecifikus lexikon nem létezik, ezért ennek előállítására egy félautomata módszert alkottunk meg. Első lépésként egy szóelőfordulás alapú maximum entrópia osztályozót tanítottunk a szöveges adathalmazon. A tanult modellből lehetőség van kinyerni, hogy az egyes szavak mennyire jellemzőek a pozitív, illetve negatív polaritású szövegekre. Ezt felhasználva kézzel kigyűjtöttük a legjellemzőbbeket (kb. 20 darab szó polaritásonként már elegendő). Az így gyűjtött szavakat használtuk kiindulási (**seed**) lexikonként (5, illetve -5 értékekkel), mely ily módon a kérdéses doménre jellemző szavakat tartalmazza.

WordNet. A *wordnet* egy olyan lexikai adatbázis, mely a szavakat szinonimahalmazokba sorolja, valamint ezek kapcsolatait is tartalmazza. Első módszerünk a magyar wordnetben [7] található információk alapján egészíti ki a megadott lexikont. Feltételezhetjük, hogy a kiindulási lexikonban található egyes szavak és azok szinonimáinak polaritása megegyezik. Így a kiterjesztett lexikonba felvettük ezeket a szinonimákat a megfelelő polaritási értékekkel. Előfordulhat azonban, hogy egy szó több szinonimahalmazba is tartozhat, így az többször is belekerül a kiterjesztett lexikonba, akár különböző polaritási értékekkel. Ezt úgy kezeltük, hogy az adott szó polaritási értékeinek átlagát számítottuk ki. A szinonimákon felül felhasználtuk azt az információt is, hogy mely szinonimahalmazok jelentése hasonló vagy ellentétes. Felvettük a kiterjesztett lexikonba az összes olyan szót, amely egy adott, a kiindulási lexikonban levő szó szinonimahalmazához hasonló (*similar_to*, *hyponym*[1]) vagy ellentétes (*near_antonym*) jelentéssel bíró halmaz eleme, mégpedig a hasonló jelentéssel bíró szavakat az eredeti polaritási értékkel, míg az ellentétes jelentéssel bíróakat negált értékkel. A módszer iteratívan futtatható, azaz az egyik lépésben eredménye a következő bemeneteleként használható, így tovább bővítve a lexikont. Figyelembe kell azonban venni azt a tényt, hogy az egyes kiterjesztési lépések során egyes szavak rossz polaritással is bekerülhetnek. Példa erre a számítástechnikai doménben pozitív szónak számító *csendes* alapján bekerülő *elzárkózott* szó, melyet negatív polaritással kellene felvenni. Mivel a wordnet egy általános nyelvi erőforrás, a benne előforduló szóhasonlóságok

[1] A hyponym kapcsolat nincs felvéve a wordnetben, a hypernym kapcsolat megfordításával határozható meg.

nem doménspecifikusak. Példa erre a *hangos* és az *erős* szavak, melyek nem szinonimák a wordnetben, de annak tekinthetők hangszórók véleményezése során. Emiatt a módszer egy más doménből származó kiindulási lexikont nem képes adaptálni. Abban az esetben azonban, ha a kiindulási lexikon doménspecifikus, a kiterjesztés során az adott doménre jellemző szavak és azok értékei alapján fog megtörténni.

Szavak klaszterezése. Hasonló szavak nem csak wordnet segítségével határozhatóak meg, hanem szövegek statisztikai elemzésével is. A következő módszerünkben az ún. Brown klaszterező eljárást alkalmaztuk [8], mely egy hierarchikus klaszterező eljárás. Az előző módszerhez hasonlóan, feltettük, hogy az egy klaszterbe tartozó szavak polaritása megegyező. Ennél fogva a kiterjesztett lexikonba felvettünk minden olyan szót, mely egy klaszterbe tartozik valamely a kiindulási lexikonba lévő szóval, annak megfelelő polaritási értékkel. Ennél a módszernél is bekerülhetnek szavak a lexikonba rossz polaritási értékkel, abban az esetben, ha egy klaszterbe kerülnek nem hasonlóan viselkedő szavak. Ennek egyik oka, ha túl kevés klasztert definiálunk, így sok szó kerül egy csoportba. Ellentétben a wordnettel, ennél a módszernél az egyes szavak hasonlósága a domén jellemzői alapján vannak meghatározva. Így a kiindulási lexikont (legyen szó doménspecifikusról vagy nem doménspecifikusról) kiterjesztés során az adott doménhez adaptáljuk.

3. Korpuszok

Ebben a fejezetben bemutatásra kerülnek a munkánk során felhasznált szöveges adatbázisok. Kísérleteinket általánosabbnak mondható és doménspecifikus szöveghalmazokon is elvégeztük.

3.1. Általános szövegek

Kifejezetten véleménydetekciós feladatok elvégzésére létrehozott szöveges adatbázis az *OpinHuBank* [3], mely tartalma különböző magyar nyelvű híroldalak, blogok és fórumok alapján lett összeállítva. Minden egyes szövegpéldány legalább hét token hosszú mondat, melyet öt annotátor pozitív, negatív vagy semleges kategóriába sorolt. Kísérleteinkhez csak a pozitív és negatív véleménytartalmú szövegekre volt szükségünk, ezért kiválogattuk azokat, melyek legalább három pozitív vagy negatív jelölést kaptak. Az így kapott **opinhu** adatbázis 882 pozitív és 1629 negatív mondatot tartalmazott. Az egyes mondatok esetében további információ az, hogy azok véleménytartalma kire vonatkozik, ezt azonban mi nem használtuk fel.

3.2. Doménspecifikus szövegek

Módszereinket doménspecifikus szöveghalmazon is elvégeztük. Ehhez létrehoztunk egy adatbázist, melyek számítástechnikával kapcsolatosak. Ezt az áruke-

reső[2] oldal tartalma alapján állítottuk össze. Az oldalon számos termékről találhatók értékelések, melyek közül a *számítógép* és *műszaki cikk* kategóriákba esőeket töltöttük le. Az egyes vélemények írása során a véleményezőnek meg kell adni szövegesen az adott termék előnyeit és hátrányait, melyeket rendre pozitív és negatív szövegeknek tekintettünk. Ezek a vélemények eltérő hosszúságúak, több mondatból, de akár pár szóból is állhatnak. Tesztelésünkhöz a letöltött vélemények közül kiválogattuk azokat, melyek egy mondatból állnak, azaz legalább négy token hosszúak és megfelelő mondatzáró jelet tartalmaznak. Az így összeálló **árukereső** adatbázis 3573 pozitív és 3149 negatív mondatot tartalmaz.

4. Eredmények

Cikkünk fő célja különböző lexikonok építése és azok használhatóságának összehasonlítása véleménydetekciós problémákon. Ehhez egy kétosztályos osztályozási feladatot definiáltunk, mely során a szövegeket pozitív vagy negatív címkével láttunk el. Az osztályozást maximum entrópia osztályozó segítségével végeztük el. Az egyes tanító és tesztpéldák esetében jellemzőként a bennük előforduló szavak szótöveit és a polaritási lexikonok alapján kinyert információkat használtuk. Egy lexikont annál jobbnak tekintünk egy adott szöveghalmazra nézve, minél nagyobb pontosságot érünk el segítségével a mondatok polaritásalapú osztályozása során. A következőkben polaritásos szavaknak tekintjük azokat a szavakat, melyek szerepelnek az adott lexikonban és a hozzájuk tartozó érték abszolútértéke legalább egy. Az adott polaritásos szó értékének előjelétől függően pozitív, illetve negatív a szó. A lexikonok alapján kinyerhető jellemzők a következők, melyekre egy példa látható az 1. táblázatban:

- szövegben szereplő polaritásos szavak (eredeti alakban)
- szövegben szereplő pozitív, illetve negatív szavak értékeinek összege különkülön
- szövegben szereplő polaritásos szavak értékeinek összege
- polaritásos szavak polaritásából és azokat megelőző, illetve követő szavak szótöveiből képzett párosok

1. táblázat. Példamondat és az abból kinyert jellemzők. A mondatban szereplő polaritásos szó a *jobb*, mely 5.0 értékkel szerepel a lexikonban.

Mondat:	A laptop kijelzőjének **jobb** paraméterei vannak!
Szótövek:	a, laptop, kijelző, jó, paraméter, van, !
Polaritásos szavak:	jobb
összértékek:	POZITÍV=5.0, NEGATÍV=0.0, POLARITÁSOS=5.0
szomszédság:	kijelző_POZITÍV, POZITÍV_paraméter

[2] `www.arukereso.hu`

2. táblázat. Seed lexikonok kiegészítése wordnet (wn) és klaszter alapú módszerekkel. Az első táblázat az opinhu, a második pedig az árukereső adatbázisokon mért pontosság, tízszeres keresztvalidációval.

opinhu-seed	86.2
opinhu-seed-wn-1	86.4
opinhu-seed-wn-2	85.9
opinhu-seed-wn-3	86.3
opinhu-seed-wn-4	86.0
opinhu-seed-klaszter-15	86.7
opinhu-seed-klaszter-15-t3	86.8

árukereső-seed	90.7
árukereső-seed-wn-1	90.8
árukereső-seed-wn-2	90.5
árukereső-seed-wn-3	90.8
árukereső-seed-wn-4	90.9
árukereső-seed-klaszter-18	90.8
árukereső-seed-klaszter-19-t3	90.8

A 2. táblázatokban láthatóak a lexikonkiegészítő módszerek eredményei az opinhu és árukereső adathalmazokra. Mindkettő esetében az adott adatbázis alapján kinyert seed lexikont egészítettük ki. Az egyes értékek a helyesen eltalált osztálycímkék százalékos arányát adják meg tízszeres keresztvalidációval mérve. A táblázatokban a *wn* jelölés a wordnet alapú kiterjesztésre utal, az utána szereplő szám pedig az iteráció számot jelöli. Az opinhu adatbázis esetében legnagyobb hibacsökkenést az első iteráció eredményezte, míg az árukereső esetében a negyedik. Az ötödik iterációtól kezdődően az eredmények folyamatosan romlanak, ami annak köszönhető, hogy a lexikonok egyre zajosabbá válnak. A táblázatok utolsó két sorában a klaszterezésen alapuló lexikonkiegészítés eredményei láthatók. Az egyes számok a hierarchikus klaszterezés vágási szintjét adják meg, míg a *t3* jelölés megléte esetén a klaszterekből kiszűrtünk minden olyan szót, aminek előfordulása legfeljebb három. Az opinhu adatbázis esetén a 15 mélységben történő, míg az árukereső esetében a 18 és 19 mélységben történő vágás eredményezte a legnagyobb hibacsökkenést.

3. táblázat. Különböző lexikonok segítségével elért pontosság *opinhu* és *árukereső* adatbázisokon, tízszeres keresztvalidációval.

	opinhu	árukereső
unigram	86.1	90.0
opinhu-seed-klaszter-15-t3	86.8	90.1
árukereső-seed-wn-4	86.2	90.9
fordított	88.4	90.2
opinhu-pmi	96.3	90.0
árukereső-pmi	84.3	91.9
árukereső2-pmi	-	91.0

A 3. táblázatban láthatók a kiegészítés, fordítás és bootstrapping eljárások segítségével létrehozott lexikonok eredményei. Referencia-rendszer az *unigram*, mely esetében az osztályozáshoz nem használtunk polaritáslexikont. Látható, hogy a lexikonkiegészítéssel szignifikáns hibacsökkenés érhető el, abban az esetben, ha a megfelelő doménből származó lexikon került alkalmazásra. Ellenkező

esetben is sikerült javulást elérni, viszont csak kis mértékben. Az angolról magyarra fordított általános lexikon hatása látható a táblázat *fordított* sorában. Mivel az opinhu különböző újságcikkeket tartalmaz, melyek tartalma nem csak egy adott doménnek kapcsolatosak, így az általános célú lexikon nagy (2,3%) javulást eredményezett. Ezzel szemben az árukereső adathalmaz esetében a javulás sokkal kisebb mértékű, még a 2. táblázatban látható kis elemszámú *árukereső-seed* lexikon eredményeitől is elmarad.

A 3. táblázat *pmi* végződésű soraiban láthatók az eredmények, melyek során a bootstrapping eljárással összeállított lexikonokat használtuk. Azokban az esetekben, amikor a lexikont ugyanazon szöveghalmaz segítségével állítottuk össze, mint amelyiken később az osztályozást végeztük, a kapott lexikon az adatokra nézve legpontosabbnak tekinthető, azaz egyfajta elméleti maximumot adnak meg. Fontos, hogy ilyen lexikon csak tesztkörnyezetben állítható elő. Kísérletet tettünk azonos doménből származó jelöletlen adatok felhasználására. Ehhez az árukereső oldalról letöltött olyan szövegeket, melyek egy mondatnál rövidebbek vagy hosszabbak, az egymondatos árukereső adatbázison tanult modellel automatikusan címkéztük. Az így kapott *árukereső2* alapján egy újabb lexikont hoztunk létre, mely segítségével tovább növeltük a pontosságot. Mivel az árukereső2 szöveghalmaz méretben nagyobb, mint az árukereső, ezért a lexikonban levő szavak polaritása pontosabban lettek meghatározva.

5. Összefoglalás

Munkánk során szövegek osztályozását végeztük el pozitív és negatív osztályokba azok véleménytartalma alapján. Az egyszerű szóalapú osztályozók pontosságát polaritáslexikonok felhasználásával javítottuk. Célunk olyan módszerek kidolgozása volt, melyek lehetőséget nyújtanak lexikonok automatikus létrehozására és kiegészítésére (doménadaptálására). Megmutattuk, hogy megfelelő mennyiségű, a kérdéses doménből származó jelölt vagy akár jelöletlen adat segítségével létre tudunk hozni lexikonokat automatikus módszerekkel vagy kis elemszámú kiindulási lexikont tudunk kiegészíteni szavak hasonlósága alapján. Utóbbi módszer esetén figyelembe kell venni azonban azt a tényt, hogy a kiegészítés során zajossá válhat a lexikon, ezért egy további feladat lehet az irreleváns szavak szűrése. Eredményeinkből kiderül, hogy fontos a megfelelő doménből származó lexikon használata. Általános szövegeken automatikus módszerekkel létrehozott lexikonok segítségével javítottunk az eredményeken, a legnagyobb javulást azonban a szintén általános célú fordított lexikonnal sikerült elérni. Ebből látható, hogy a legpontosabb eredményeket manuálisan összeállított lexikonok segítségével érhetjük el. Ennek oka, hogy ezek a lexikonok egyrészt kevésbé zajosak, másrészt olyan többletinformációval rendelkeznek, melyek a szövegek alapján statisztikai módszerekkel nem határozhatóak meg. A kézi összeállítás azonban hosszadalmas, költséges és a kérdéses domén ismeretét igényli. Számítástechnikával kapcsolatos domén esetében az általános célú lexikon elenyésző javulást hozott, mivel ez a lexikon nem vagy más polaritással tartalmazza a doménre jellemző véleményt kifejező szavakat. Az automatikus módszerek képesek meg-

határozni a domén jellemzőit, így ezekkel a módszerekkel létrehozott lexikonok a hibák 10%-os csökkenését eredményezték.

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. Hangya, V., Berend, G., Farkas, R.: SZTE-NLP: Sentiment Detection on Twitter Messages. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). (2013) 549–553
2. 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
3. Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013) 343–345
4. 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)
5. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics (2005) 347–354
6. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) **21**(4) (2003) 315–346
7. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and results of the Hungarian WordNet project. In: Proceedings of the Fourth Global WordNet Conference (GWC-2008). (2008)
8. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational linguistics **18**(4) (1992) 467–479