A szentimentérték módosulásának vizsgálata szemantikai–pragmatikai szempontból annotált korpuszon

Szabó Martina Katalin[1,2], Nyíri Zsófi[1], Morvay Gergely[1], Lázár Bernadett[1]

[1] Precognox Informatikai Kft.
{mszabo,zsnyiri,gmorvay,blazar}@precognox.com
[2] Szegedi Tudományegyetem, Bölcsészettudományi Kar,
Szláv Intézet
szabo.martina@lit.u-szeged.hu

Kivonat: A dolgozat az emotív szemantikai tartalmú elemek egy speciális csoportját vizsgálja, kézzel annotált korpusz segítségével. Negatív emotív szemantikai tartalmú elemekként hivatkozunk azokra a kifejezésekre, amelyek képesek arra, hogy elsődleges negatív polaritásuk ellenére pozitív értékelést fejezzenek ki, vagy (pozitív és negatív) szentimentkifejezések fokozóiként szolgáljanak. A vizsgálat tárgyát képező elemek – az elméleti szempontú problematikusságuk mellett – nyelvtechnológiai szempontból is figyelemre méltóak. Automatikus kezelésük ugyanis komoly kihívást jelent mind a szentiment-, mind az emócióelemzés számára. A vizsgálati korpuszt, amelyet a kutatás céljainak megfelelően annotáltunk kézzel, magyar nyelvű twitter-bejegyzések alkotják. A korpusz létrehozásának fő célja egy olyan adatbázis megalkotása volt, amely lehetővé teszi az adott elemcsoport beható, szemantikai–pragmatikai szempontú vizsgálatát. A dolgozatban beszámolunk a korpusz létrehozásának menetéről és eszközéről, az annotálás alapelveiről, valamint a korpuszadatok vizsgálati eredményeiről is. Végül összegezzük azokat a javaslatokat, észrevételeket, amelyek figyelembe vétele véleményünk szerint hozzásegíthet a vizsgált elemek pontosabb automatikus feldolgozásához.

1 Bevezetés

A dolgozat egy az elméleti nyelvészetben is kevéssé tárgyalt, és a nyelvtechnológia számára is problematikus jelenséget, az ún. negatív emotív szemantikai tartalmú elemeket veszi górcső alá.

A vizsgált jelenség a szentimentelemzés egy kardinális részproblémájához, a polaritás, másképpen a szentimentérték módosulásához tartozik. A szentimentérték módosulásának azt a jelenséget nevezzük, amikor egy adott nyelvi elem lexikai szintű szentimentértéke nem azonos, vagy nem teljes mértékben azonos az őt magában foglaló teljes megnyilatkozás értékével [1].

Egy adott szentimentkifejezés lexikai szintű polaritása számos okból kifolyólag eltérhet a bennfoglaló megnyilatkozás polaritásától. Így például, egy pozitív kifejezés polaritása negálható (pl. *nem jó*), vagy bizonytalanná tehető (pl. *kevésbé jó*), vagy az irónia eszközével az ellentétére fordítható (pl. *ezt jól megcsináltad!*).

A szentimentérték módosulásának egy speciális típusát képezik azok az elemek, amelyek lexikai szinten negatív emotív szemantikai tartalommal rendelkeznek ugyan, azonban szenti-

mentkifejezés funkcióját betöltve, vagy pedig más szentimentkifejezések fokozóiként képesek nem negatív értékítélet kifejezésére is [2,3]. E kifejezéseknek két altípusát különböztetjük meg: az ún. *lexikai szintű értékváltás*t (1a) és az ún. *értékvesztés*t (1b) (a vizsgált elemeket kövér szedéssel emeljük ki):

1. a. **brutális** koncerten voltunk a hétvégén
   b. **brutálisan** jó volt a tegnap esti buli

Az (1a) alatti példa esetében azt látjuk, hogy a vizsgált elem aktuális polaritása ellentétes azzal, mint amelyet lexikai szinten hordoz. Ezt a jelenséget nevezzük *lexikai szintű értékváltás*nak. Az (1b) alatti példában ugyanakkor a vizsgált elem egy másik, pozitív polaritással rendelkező kifejezés fokozójának (intenzifikálójának) a funkcióját tölti be. Saját, lexikai szintű negatív értékét tehát elveszíti, és egy másik szentimentkifejezés polaritását erősíti tovább. Ezt a jelenséget nevezzük *értékvesztés*nek.

Amint arra a következő fejezetben rámutatunk (l. lentebb, 2.), a vizsgált jelenségekkel mind elméleti, mind alkalmazott nyelvészeti, különösen nyelvtechnológiai szempontból csekély számú dolgozat foglalkozik. Ugyanakkor azok tárgyalása mindkét tudományterületen fontos volna. Ami az elméleti vonatkozásokat illeti, amint azt tárgyalni fogjuk, a vizsgált elemek szemantikai viselkedéséről nincs egységes vélekedés. Ami az alkalmazott nyelvészeti, köztük a nyelvtechnológiai alkalmazásokat illeti, a polaritás módosulásának ezeket a típusait a jelenleg legelterjedtebbnek tekinthető szótáralapú szentimentelemzéssel nem lehet kezelni. A lexikai szintű értékváltásra, illetve értékvesztésre képes nyelvi elemek ugyanis rendszerint szerepelnek a szótári polaritásuknak megfelelő szentimentlexikonban. Ennek következtében automatikus kezelésük a szótáralapú elemzés során tévesen történik. Ugyancsak problematikusak a automatikus emócióelemzés szempontjából is, hiszen a szótáras elemzés ezeket a kifejezéseket azok lexikai szintű negatív tartalma alapján azonosítja (részletesebben l. [4]).

A jelen dolgozatban bemutatjuk azt a kézzel annotált korpuszt, amelyet specifikusan e probléma vizsgálatára hoztunk létre. Ismertetjük továbbá mindazokat a vizsgálati eredményeket, amelyeket a nyelvhasználati sajátságokat illetően megállapítottunk, a korpuszadatok feldolgozása és elemzése alapján. Végezetül azt is tárgyaljuk, milyen lehetőségeket látunk a vizsgálat tanulságainak nyelvtechnológiai implementációjára.

2 Szakirodalmi áttekintés

Az általunk lexikai szintű értékváltásnak, valamint értékvesztésnek nevezett jelenségekkel foglalkozó dolgozatok többsége a pszichológia, valamint az elmélet oldaláról teszi vizsgálat tárgyává a problémát [5,6,7,8,9,10]. Ami a szentimentelemzéshez kapcsolódó nyelvtechnológiai kutatásokat illeti, megállapítható, hogy amíg az ún. kontextuális polaritásváltással (pl. a negáló elemek problémája) egyre gyakrabban foglalkoznak az irodalmi tételek, addig az általunk vizsgált jelenségekre csupán csekély számú dolgozat fordít figyelmet [2,3,11].

Andor [9] alapján a jelen dolgozatban tárgyalt jelenség nem ritka a jelentésváltozások folyamataiban. Magyarázata szerint annak „leggyakoribb eseteiben a negatív jelentéstartalmú és használatú lexikális egységek pozitív irányú jelentésváltozását vagy jelentésbővülését, jelentésük kiterjesztését figyelhetjük meg” [9]. A szerző véleménye szerint a jelenség különösen „az értékítéletet, fokozást kifejező ún. intenzifikáló szavak körében jellemző”. Ugyanakkor arra is felhívja a figyelmet, hogy ezek az elemek gyakran kollokálódnak negatív polaritású melléknevekkel is amellett, hogy egyre nagyobb arányban fordulnak elő pozitív tartalmú kifejezésekben, konstrukciókban.

Andor [9] és Kugler [10] is megemlíti, hogy a jelenség az intenzifikáló elemek kapcsán az ellentétes polaritás irányába is lehetséges, azaz pozitívból negatívba (pl. tökéletesen buta, jól elhibázta). Ugyanakkor véleményük szerint ez utóbbi jóval ritkább előfordulású.

Tolcsvai Nagy [5] amellett érvel, hogy amíg az általa „hagyományosnak" nevezett jelzői csoport tagjai, így például a *kiváló*, a *nagyszerű* vagy a *csodálatos* elemek magukban hordozzák „a szóval jelölt érték valódiságát", addig az *állati* vagy a *baromi* típusú jelzők „a jelzett értékek relativitására utalnak". Ennek következtében azok jelentésében „ironizáló magatartás" mutatkozik meg. Tolcsvai Nagy [5] érvelésével ellentétben Székely [12] ugyanakkor úgy véli, hogy a vizsgált elemek szemantikailag gyakorta motiválatlanok.

A jelenség Andor [13] és Jing-Schmidt [7] vizsgálati eredményei alapján számos nyelvben megtalálható, ezért valószínűleg nyelvfüggetlen sajátság.

Kugler [10] és Jing-Schmidt [7] a jelenséget elsősorban a használat lehetséges pszichológiai oka szempontjából vizsgálja. Ennek kapcsán Kugler [10] felhívja a figyelmet a kongruencia, másképpen értékbeli egyez(tet)és tendenciájára, miszerint „a legtermészetesebb és ezért a legkönnyebben feldolgozható szerkezetekben azonos polaritású kifejezések kapcsolódnak össze. Amennyiben a kifejezés tagjai között nincs kongruencia, a szerkezet nagyobb mentális erőfeszítéssel (így szükségképpen hosszabb idő alatt) dolgozható fel. Véleményünk [11] illeszkedik Kugler [10] fentebbi megállapításához: az inkongruencia a vizsgált jelenség egyik pszichológiai motiválójának tekinthető, hiszen az így egymás mellé kerülő, ellentétes polaritású tartalmak interpretálása – az említett tendencia miatt – nagyobb figyelmet igényel a hallgatótól.

3. A vizsgálati korpusz létrehozásának és feldolgozásának a menete

A nyers korpusz, amelyből a kutatáshoz szükséges anyagot gyűjtöttük, összesen 37818 magyar nyelvű twitter-bejegyzésekből áll. A jelen kutatáshoz a korpusz azon nyelvi adataira volt szükségünk, amelyek tartalmaztak legalább egy, lexikai szintű értékváltásra vagy értékvesztésre képes elemet. A munka első fázisában ki kellett tehát nyernünk a korpuszból az ennek a szempontnak megfelelő tweeteket. A feladathoz kézzel összeállítottunk egy listát, amely lexikai szintű értékváltásra vagy értékvesztésre képes elemeket tartalmaz. A munkában egy fokozó értelmű kifejezéseket tartalmazó szótárra [14], két korpuszra [15,16], valamint internetes adatokra támaszkodtunk. Az így létrehozott szólista 109 szóalakot tartalmaz. Ezt követően automatikus módszerrel kigyűjtöttünk a korpuszból minden olyan tweetet, amely tartalmazott legalább egyet a listánkban szereplő elemek közül. Az így létrehozott korpusz összesen 610 tweetből áll.

A nyers korpusz kézi feldolgozásához a Brat nevű, online elérhető annotáló programot használtuk [17]. Az eszköz bármilyen annotálási feladatra alkalmas, a felhasználó maga konfigurálhatja a használni kívánt annotációs tageket, azok csoportjait, kapcsolatatait, és további, bevinni kívánt információkat. Teljesen személyre szabható és egyszerűen kezelhető. A config fájlt és az annotálandó szövegeket txt formátumban töltöttük fel a programba, és az annotációs fájlokat is ebben a formátumban kaptuk vissza. Az output fájlban az annotált tagek lokációi és a létrehozott kapcsolatok szerepelnek egyszerű listaként.

Az adatbázis feldolgozását a következő annotációs szempontok alapján végeztük el manuálisan: Először bejelöltük azokat az elemet, amelyek esetében a lexikai szintű szentimentérték eltért a kontextusbeli értéktől. Még ebben a lépésben döntést hoztunk arról is, hogy ez az eltérés az aktuális kontextusban lexikai szintű értékváltásnak, vagy pedig értékvesztésnek köszönhető-e, és ennek megfelelően jelöltük a vizsgált elemet, valamint a targetét vagy azt, amit módosít, tehát a frázis alaptagját. A lexikai szintű értékváltás (2a) esetében az előbbit, az értékvesztés (2b) esetében az utóbbit kerestük meg és annotáltuk. (A vizsgált elemeket ebben az esetben is kövér szedéssel, míg a további annotált kifejezéseket aláhúzással jelöljük.)

2. a. ismét **brutális** koncertet adott az énekes
   b. Valljátok be, **rohadt** jó az időérzékem!

Amint a példák mutatják, a (2a) esetében a vizsgált elem, lexikai szintű negatív polaritását elveszítve pozitív szentimentkifejezés funkcióját tölti be, és a *koncert* targetre vonatkozóan fejezi ki ezt a pozitív értékelést. Ettől eltérően, a (2b) alatti példában a vizsgált elem nem szentimentkifejezésként funkcionál, hanem a szintaktikai szerkezetben az alaptag szerepét betöltő szentimentkifejezés fokozójaként, annak szemantikai tartalmát erősítve.

A bemutatott annotálási megoldásnak az volt a célja, hogy a segítségével a jövőben vizsgálni lehessen egyrészt a fokozó szerepű, értékvesztésre képes elemeknek és az általuk módosított elemeknek, valamint a lexikai szintű értékváltásra képes elemeknek és azok targeteinek a kapcsolatait.

Az annotációban értelemszerűen csupán azokat a tweeteket annotáltuk, ahol a vizsgált elemnél lexikai szintű értékváltást vagy értékvesztést véltünk felfedezni. Azokban az esetekben tehát, ahol a vizsgált elem aktuális polaritása megegyezett annak elsődleges, azaz negatív polaritásával, nem annotáltuk, pl.

3. Egy **rohadt** kukásautó miatt áll már vagy tíz perce a troli.

A fentebb bemutatott két alaptípus annotálásán túl az értékvesztésre vonatkozóan további tageket is bevittünk, az annotált elemek további szemantikai–pragmatikai viselkedése alapján. E megoldás alkalmazása mellett korábbi vizsgálati eredményeink alapján döntöttünk: megfigyeltük, hogy amíg a negatív emotív tartalmú fokozó elemek pozitív és negatív polaritású kifejezések módosítóiként rendre deszemantizálódnak (4a-b), addig semleges melléknévi alaptagok mellett változatos szemantikai viselkedést mutatnak (4c-d) [11]. (A példákban a vizsgált elemet kövérrel, az alaptagot aláhúzással jelöltem.)

4. a. **iszonyat** jó volt ez a 4 nap, köszönöm az élményt
   b. **borzasztó** unalmas Affleck a szerepben
   c. 166 centi 46 kiló az milyen egy 14 éves lánynak? Nem vagyok **rohadt** magas? (vö. *túl*)
   d. A processzor teljesítménye elégnek tűnik, mert minden **marha** gyors

A (4c-d) alatti példák arra mutatnak, hogy semleges alaptag mellett a fokozó elem, negatív szemantikai tartalmát illetően nem feltétlenül üresedik ki, és válik puszta fokozóvá. A(4c-d) alatti tweetekben ugyanis éppen ez az elem adja hozzá a negatív értékelést a szöveghez, ellentétben a (4a-b) alatti példákkal, ahol negatív tartalommal nem számolhatunk.

Annak céljából, hogy ezeket az eseteket a korpusz alapján vizsgálni tudjuk, az intenzifikáló elemek esetében azt is annotálnunk kellett, hogy az általuk módosított elemek pozitív vagy negatív polaritásúnak, vagy pedig semlegesnek tekinthetőek-e, továbbá, hogy az intenzifikáló elemek milyen aktuális szemantikai tartalommal rendelkeznek.

Mindemellett, az összes annotált taghez létrehoztunk egy “egyéb”-kategóriát is azért, hogy a munka során esetlegesen felmerülő, sajátos jelenségeket is jelölni tudjuk egy későbbi vizsgálat céljából.

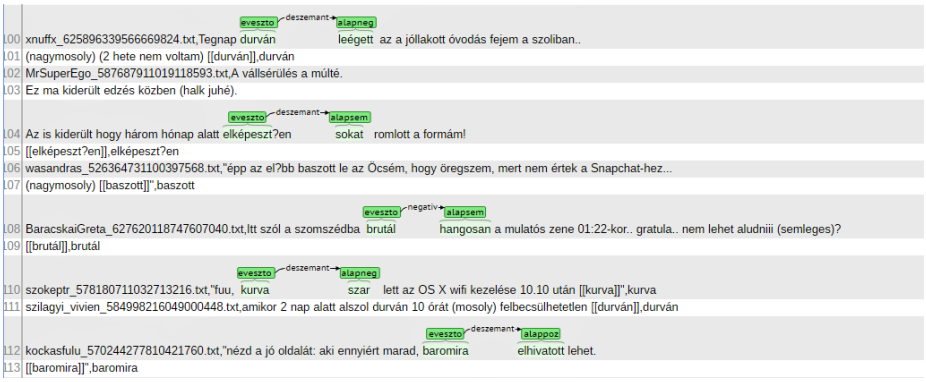
A fentebb ismertetett annotálási rendszert az alábbi táblázat foglalja röviden össze:

	Az annotált tagek és szemantikai–pragmatikai sajátságok (ez utóbbit a program technikai sajátsága okán viszonyként jelöltük):
Lexikai szintű értékváltás esetén:	lexikai szintű értékváltó elem, ami a szemantikai tartalma alapján lehetett: értékváltó vagy egyéb
	target
Értékvesztés esetén:	értékvesztő elem, ami a szemantikai tartalma alapján lehetett: deszemantizált, negatív, pozitív vagy egyéb
	alaptag, ami a szemantikai tartalma alapján lehetett: negatív, pozitív, semleges vagy egyéb

1. táblázat. Az annotálási rendszer rövid összefoglaló táblázata.

A bemutatott annotálási rendszert az annotátorok egy részletes annotálási útmutatón keresztül ismerték meg, amelyben az egyes jelenségeket példák segítségével prezentáltuk.

Az alábbi ábra egy részletet közöl a korpusz annotációjából a Brat programban:



1. ábra. Részlet a korpusz annotációjából a Brat nevű programban.

4 Eredmények

4.1 Az egyetértésmérés eredményei

Ahhoz, hogy az annotátorok közötti konszenzust mérni tudjuk, a korpusz feldolgozása előtt elvégeztünk egy egyetértésmérést a korpusz egy kisebb részletén. A méréshez összesen 100 tweetet annotáltunk, tekintettel arra, hogy már ez a mennyiség is a teljes korpusz hatod részét tette ki. Az pilot-adatokon az annotálást követően kappa-statisztikát alkalmaztunk.

A pilot-annotálás megmutatta, hogy az annotátorok a Kappa-érték szerint összesítve 0.489, azaz a Kappa-sávok alapján közepes szintű átlagos egyetértéssel dolgoztak. Az annotáció részletes vizsgálata alapján a közepes szintű eredmény alapvető oka az volt, hogy az annotálás technikai szempontból nem volt egységes: az annotátorok a munka során nem azonos kijelölési megoldásokat alkalmaztak, ami miatt számos lokáció-beli eltérés keletkezett a korpuszban. Ugyanakkor, az eredmény – kisebb részben ugyan, de – összefüggést mutatott az annotálási feladat tartalmi vonatkozásaival is, tehát azzal, hogy nem triviális, hanem szemantikai–

pragmatikai szempontból dolgoztuk fel az adatokat, ami néhol bizonytalanságot eredményezett az annotátorok körében. E fentebbi tapasztalatok alapján a korpusz feldolgozását a technikai megoldások egységesítése, valamint az annotálási alapvelvek pontosítását követően végeztük csak el.

4.2 Az annotált korpusz vizsgálati eredményei

A 610 tweetből összesen 280-at annotáltunk. Ezek voltak azok az esetek ugyanis, ahol a vizsgált elem lexikai szintű értékváltást vagy értékvesztést mutatott. A többi, nem annotált esetben a vizsgált elem megőrizte a lexikai szintű negatív polaritását az aktuális kontextusban.

Az annotációban összesen 41 esetben jelöltünk értékváltást, és 238 esetben értékvesztést, tehát intenzifikálói funkciót.

Megvizsgálva az értékváltás eseteit, a következő megállapításokat tehetjük: A két leggyakoribb elem ebben a szerepben a *durva(-n)* (13 előfordulás) és a *kemény* (8 előfordulás) volt. Számos további elemet is annotáltunk még értékváltóként, azonban ezek lexémánként átlagosan mindössze egy vagy két alkalommal szerepeltek. Szerettük volna megtudni, hogy vajon menynyire jellemző a két leggyakoribb elemre az értékváltás, azaz feltehető-e, hogy ezek az elemek szentimentkifejezés funkciójában alapvetően pozitív aktuális polaritással rendelkeznek. Megvizsgáltuk tehát ezeknek az elemeknek az összes, nem annotált előfordulását is a korpuszban. Azt tapasztaltuk azonban, hogy mindkét elem gyakran fordul elő negatív értékelés kifejezőjeként is. Ez a sajátság nyilvánvalóan megnehezíti az értékváltásra képes elemek aktuális polaritásának a helyes automatikus kezelését (a problémáról részletesebben l. lentebb, 5.).

A 238 értékvesztési esetet illetően a következő észrevételeket tehetjük: Ahogyan azt a feldolgozás menete kapcsán is ismertettük (l. fentebb, 3.), azokban az esetekben, ahol a vizsgált elem fokozó funkciót töltött be, a intenzifikálót és az általa módosított elemet további szemantikai–pragmatikai annotációval láttuk el, és a korpusz felhasználása során az így annotált sajátságokat is lekérdeztük. Az adatokat az alábbi táblázat mutatja be:

		a vizsgált elem aktuális szemantikai tartalma				
		deszem	**negatív**	**pozitív**	**egyéb**	ÖSSZESEN:
az alaptag szemantikai tartalma	**alappoz**	94	-	-	1	95
	alapneg	70	-	-	8	78
	alapseml	30	24	3	-	57
	egyéb	5	-	-	3	8
	ÖSSZESEN:	199	24	3	12	238

2. táblázat. Az annotált értékvesztési esetek részletes statisztikai adatai

Az eredmények alapján a következő megállapításokat tehetjük: A korpusz annotátorai az összesen 238 esetből 199 alkalommal vélték úgy, hogy a vizsgált elem teljesen elvesztette elsődleges

negatív szemantikai tartalmát, és az általa módosított elem mellett puszta fokozó szerepet töltött be. Ez az összes eset 83,61%-át tette ki.

A vizsgált elemek a polaritásvesztést illetően a pozitív alaptagok mellett mutatkoztak a legegységesebbnek. Megállapíthatjuk ugyanis, hogy pozitív alaptagok módosítóiként rendre elveszítik lexikai szintű tartalmukat, l. pl. (4b) fentebb. Ugyancsak kis mértékű eltérést látunk az annotációban akkor is, ha a módosított elem negatív polaritású. Összességében úgy tűnik tehát, hogy deszemantizálódnak azok a fokozó értelmű elemek, amelyek valamilyen (pozitív vagy negatív) polaritással rendelkező elemet módosítanak.

A fentebb bemutatottakkal ellentétben, semleges alaptagok módosítóiként a vizsgált elemek jelentős szemantikai változatosságot mutatnak. Az 57 eset valamivel több, mint a felében (52,63%) jelöltek az annotátorok deszemantizáltságot, azaz ítélték úgy, hogy a vizsgált elem teljesen elveszíti lexikai szintű negatív értékét. (A módosított elemeket itt is aláhúzással jelöltem.)

5. a. Jáááj, de **rettenetesen** kíváncsi lettem!
   b. Ja és Colin Farrell egy **kibaszott** nagy színészóriás

24 alkalommal (42,1%) azonban a vizsgált elem nem veszítette el negatív szemantikai tartalmát. Ezekben az esetekben tehát, annak ellenére, hogy fokozó funkciót tölt be, az aktuális kontextusban ugyanúgy negatív értékelést fejez ki, mint lexikai szinten, pl.

6. **Szörnyen** meleg van még így az éjszaka közepén is.

Végezetül, 3 olyan esetet is jelöltek a korpusz annotátorai, ahol a vizsgált elemet pozitív polaritás hordozójának értékelték, annak lexikai szintű negatív polaritása ellenére, pl.

7. mondjuk **rohadt** sok programom lesz ebben a hónapban is, de fel nem bírom fogni, hogy utána újra suli

A vizsgálat fentebbi tanulságai azért is figyelemre méltóak, mert számos, a negatív emotív szemantikai tartalmú fokozó elemekkel foglalkozó dolgozat amellett érvel, hogy azok szemantikailag rendre kiüresednek, elveszítik lexikai szintű negatív polaritásukat. Balogh [18] például úgy gondolja, hogy amennyiben „az ilyen, másodlagos fokozóelemeket egy-egy megfelelő kulcsszóhoz kapcsoljuk, elveszítik elsődleges, azaz lexikális jelentésüket és átveszik a fokozó értelmű „nagyon" adverbium jelentését." Hozzá hasonlóan érvel Jing-Schmidt [7], aki szerint a félelem érzelemmel kapcsolatos negatív emóciókifejezések esetében, fokozó szerepben a félelem szemantikai tartalom metonimikusan a magas emotív intenzitásra redukálódik. A (6) alatti tweetben ugyanakkor éppen ez az elem adja hozzá a negatív értékelést a szöveghez, ellentétben az (5) alatti példákkal, ahol nincs ilyen negatív tartalom.

Kíváncsiak voltunk, vajon milyen konkrét fokozó elemek fordulnak elő a korpuszban semleges alaptag mellett, és mutatkozik-e valamilyen eltérés abban, hogy mely elemek üresednek ki szemantikailag, és melyek nem. Megvizsgáltuk tehát mind a deszemantizált, mind a negatív polaritású csoport gyakorisági megoszlásait. A semleges alaptagok melletti deszemantizált fokozó elemek közül a leggyakoribbak a *kurva* (48), a *rohadt* (25), az *iszonyat* (19), a *baromi* (18) és a *(ki)baszott* (17) tövek, illetve ezek különböző alakváltozatai voltak. Megvizsgálva a negatív tartalmú fokozók megoszlását ugyancsak semleges alaptagok mellett azt az érdekes

sajátságot tapasztaltuk, hogy amíg közülük a négy leggyakoribb egybeesik a deszemantizáltak leggyakoribbjaival, addig a 3. leggyakoribb deszemantizált elem, az *iszonyat* negatívként egyetlen egyszer sem fordult elő a korpuszban. Ez az elem tehát minden esetben teljes értékvesztést mutat.

5 A vizsgálati eredmények felhasználhatósága az automatikus szentimentelemzésben

A vizsgált nyelvi jelenség viszonylagosan ritka előfordulása miatt az annotált korpusz mérete kicsinek mondható. Hangsúlyozzuk továbbá, hogy a pilot-korpuszon mért annotátorok közötti egyetértés csupán közepes értéket mutatott, bár – amint azt az egyetértésmérés kapcsán részleteztük (l. fentebb, 4.1) – az eredmény alapvető oka az volt, hogy az annotálás technikai szempontból nem volt egységes, és ezt a problémát orvosoltuk. Az elmondottakkal összefüggésben mégis a következő, a nyelvtechnológiai implementációra vonatkozó megállapításokat a kutatás jelenlegi szakaszában korlátozott érvényűnek tekintjük.

Az automatikus kezelés szempontjából a legkevésbé problematikusak azok a kifejezések, amelyek pozitív vagy negatív polaritású módosított elemek mellett fokozó szerepben állnak. Azt láttuk ugyanis, hogy ennek az összesen 173 esetnek a 94,79%-ában a vizsgált elem deszemantizált volt. Ez alapján azt mondhatjuk, hogy az emotív intenzifikáló elemek értékvesztése a polaritással rendelkező tagok módosítóként egyszerű reguláris szabályokkal leírható.

Problematikusabbak azonban mind a semleges tagok melletti fokozó elemek, mind pedig azok, amelyeket értékváltónak nevezünk, és pozitív szentimentkifejezés funkciójában állnak. Ezekben az esetekben ugyanis nem tudunk gépi módszerrel olyan egyértelmű kontextuális sajátságokra hagyatkozni, mint amilyenekre a fentebb tárgyalt esetben. Amint arról beszámoltunk (l. fentebb, 4.2.), megpróbáltunk gyakorisági eltéréseket felfedezni a semleges alaptagok mellett megjelenő, eltérő szemantikai sajátságokkal bíró csoportokban, azonban egyetlen kivételtől eltekintve nem találtunk érdemi különbséget: a két csoport leggyakoribb elemei megegyeznek; egyetlen kivételtől eltekintve, az *iszonyat* ugyanis csak deszemantizált változatban fordul elő ebben a pozícióban. Ez utóbbi elem szótári jelentését tehát – ezek szerint – fokozó szerepben nem kell figyelembe venni.

Ami a lexikai szintű értékváltás gyakorisági adatait illeti, a *durva* és a *kemény* a legfrekventáltabb elemek a korpuszban. Ugyanakkor megvizsgálva az összes, nem annotált előfordulásukat is, tehát azokat, ahol e kifejezések nem pozitív aktuális értéket hordoztak, azt tapasztaltuk, hogy mindkét elem ugyanolyan gyakran fordul elő negatív értékelés kifejezőjeként is. Aktuális szemantikai tartalmuk tehát lexikai szinten (így például egy szentimentszótárban) nem rögzíthető. Megfigyeltük ugyanakkor, hogy ezeknél az elemeknél, értékváltó pozícióban a tweetelők nagyon gyakran egészítik ki a szöveges megnyilatkozásaikat olyan emotikonnal, amely az aktuális értékelő jelentésre utal. Úgy gondoljuk, talán éppen azért élnek ilyen gyakran emotikonokkal ezekben az esetekben, mert a viszonylag rövid karakterhosszúságú tweetekben, a polaritásváltásra képes elemek használatakor, a közölni kívánt értékelő tartalmat egyértelművé kívánják tenni. Mindezek alapján úgy véljük, hogy a jelen kutatásban megtalált leggyakoribb értékváltó elemeket az aktuális tweet emotikonjával együtt kezelve megnő a helytálló elemzés esélye. Ezzel kapcsolatban érdemes felhívni a figyelmet [19]-re, akik ugyancsak twitter-szövegeket vizsgálva megállapítják, hogy az irónia automatikus felismerésében az emotikonok kulcsszerepet tölthetnek be.

A kutatás következő lépéseként azt tervezzük, hogy a fentebb tárgyalt sajátságokat szabályokba foglaljuk és alkalmazzuk a szótáralapú automatikus szentimentelemzéssel kombinálva, majd felmérjük, hogy azok javítanak-e, és ha igen, milyen mértékben az elemzés eredményességén. Ehhez a feladathoz rendelkezésünkre áll egy kézzel annotált szentimentkorpusz [1]], valamint egy pozitív és negatív polaritású kifejezéseket tartalmazó szentimentszótár [20]. A

korpuszt először egyszerű szóillesztéses megoldással, a szótár alapján elemezzük, és az eredményt összevetjük a kézi annotációval. Ezt követően az elemzést a szótáras módszer és a jelen munka során feltárt sajátságok kombinációjával is elvégezzük, majd ennek a megoldásnak az eredményességét is összevetjük a korpusz kézi annotációjával. Végezetül megnézzük, javult-e, és ha igen, mennyiben az automatikus elemzés eredményessége az alkalmazott szabályoknak köszönhetően.

6 Összegzés

A dolgozatban a negatív emotív szemantikai tartalmú elemek egy specifikus csoportját, az ún. lexikai szintű értékváltásra, illetve értékvesztésre képes elemeket vizsgáltuk kézzel annotált korpusz segítségével.

A vizsgálati korpuszt magyar nyelvű twitter-bejegyzésekből hoztuk létre úgy, hogy egy az erre a célra összeállított elemlista alapján kigyűjtöttük a vizsgálni kívánt elemeket tartalmazó nyelvi adatokat. Ezt követően a korpuszt egy az erre a célra felkészített eszközzel manuálisan annotáltuk. A munka során bejelöltünk minden olyan szemantikai–pragmatikai sajátságot, amellyel a korpusz későbbi, kutatásbeli felhasználását támogatni tudtuk. Az annotáció elkészülte után a korpuszt felhasználtuk a nyelvi jelenség vizsgálatára, és a feltárt sajátságokat részleteiben, példákkal együtt közöltük. Végezetül, a vizsgálat tanulságaira építve tárgyaltuk azt is, hogyan látunk lehetőséget a tapasztalatok nyelvtechnológiai implementációjára.

Ahogyan azt a dolgozatban több ízben kiemeltük, a vizsgált jelenség megfelelő kezelése bizonyos nyelvtechnológiai alkalmazások szempontjából kiemelkedően fontos volna. Így például, ezek az elemek mind a szentiment-, mind az emócióelemzésben téves következtetéseknek engedhetnek teret. Bár a kutatás alapjául szolgáló vizsgálati anyag kis méretű volt, úgy véljük, a segítségével tett megállapítások és javaslatok hozzájárulhatnak egy pontosabb, hatékonyabban működő tartalomelemző rendszer létrehozásához.[1] Ezzel összefüggésben, a munka további lépéseként tervezzük a megállapításaink és javaslataink nagyobb méretű adatbázison való vizsgálatát, igazolását.

Köszönetnyilvánítás

A jelen kutatás Az Emberi Erőforrások Minisztériuma Új Nemzeti Kiválóság Programjának támogatásával valósult meg.

[1] Természetesen egyet kell értenünk a dolgozat névtelen bírálójával abban, hogy a vizsgált jelenség viszonylagosan ritka előfordulása okán annak hatékony kezelése önmagában nem feltétlenül hoz jelentős javulást egy szentiment- vagy emócióelemző rendszer eredményességét illetően. Ugyanakkor amellett érvelünk, hogy a jelenség frekvenciája doménfüggő, és bizonyos típusú, illetve témájú szövegekben kifejezetten gyakorinak tekinthető. Így például a társalgási stílusréteghez tartozó, technológiai témájú szövegek gyakran élnek vele (pl. különböző elektronikai eszközökről, vagy azokkal kapcsolatban írt blog-, facebook- és twitterbejegyzések stb., pl. ***Brutális*** *választék,* ***durván*** *jó árak; várok egy nagyon-nagyon* ***brutális*** *iPhone 7-et; nagyon* ***durva*** *sportkocsi, kicsi, könnyű és* ***iszonyúan*** *erős* stb.), a szentiment- és emócióelemzés egyik leggyakrabban elemzett szöveganyagát pedig – gazdasági okokból – éppen ezek a nyelvi produktumok alkotják.

Irodalom

1. Szabó M. K., Vincze V.: Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai. In: Tanács A., Varga V., Vincze V., eds.: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015). Szeged, Szegedi Tudományegyetem (2015) 219−226
2. Szabó, M. K.: A polaritásváltás- és változás kezelési lehetőségei a szentimentelemzésben. Tavaszi Szél konferencia konferenciakötete. Budapest, Líceum Kiadó, Eger és Doktoranduszok Országos Szövetsége (2015a) 629−643
3. Szabó M. K.: A polaritásváltás problémája a szentimentelemzés szempontjából. In: Váradi T., ed.: IX. Alkalmazott Nyelvészeti Doktoranduszkonferencia konferenciakötete. Budapest, MTA Nyelvtudományi Intézet, (2015b) 51−61
4. Drávucz, F., Szabó, M. K., Vincze V.: Szentiment- és emóciószótárak eredményességének mérése emóció- és szentimentkorpuszokon. A jelen kötetben
5. Tolcsvai Nagy G.: A mai magyar nyelv normarendszerének egy jelentős változásáról az „ifjúsági nyelv" kapcsán. Magyar Nyelvőr 112(4) (1988) 398−406
6. Wierzbicka, A.: Australian cultural scripts − *bloody* revisited. Journal of Pragmatics, Volume 34(9) (2002) 1167–1209
7. Jing-Schmidt, Z.: Negativity bias in language: A cognitive-affective model of emotive intensifiers. Cognitive Linguistics 18(3) (2007) 417−443
8. Laczkó M.: Napjaink tizenéveseinek beszéde szóhasználati jellemzők alapján. Magyar Nyelvőr 131(2) (2007) 173−184
9. Andor J.: De durva ez a téma! – Megfigyelések a melléknévi polaritásváltásról. In Hungarológiai Évkönyv 12 (2011) 33−42
10. Kugler N.: A nyelvi polaritás kifejezésének egy mintázata, avagy milyen a félelmetesen jó? Magyar Nyelvőr 138(2) (2014) 129−139.
11. Szabó, M. K.: The usage of elements with emotive semantic content from a gender point of view. Kézirat
12. Székely G.: Egy sajátos nyelvi jelenség, a fokozás. In: Segédkönyvek a nyelvészet tanulmányozásához 66. Budapest, Tinta (2007)
13. Andor J.: Functional Studies in the Polarity and Gradation of Amplifier Adjectives and Adverbs in English. In: Andor, J., Horváth, J., Nikolov, M., eds. Studies in English Theoretical and Applied Linguistics. Pécs, Lingua Franca Csoport (2003) 43–59.
14. Tukacs, T.: Túlzásba vitt szavak. A fokozó értelmű szókapcsolatok magyar angol szótára. Budapest, Tinta (2015)
15. Váradi, T.: 2002. The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002) European Language Resources Association, Paris (2002) 385–389
16. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In Proceedings of LREC 2014 (2014)
17. Brat online annotáló eszköz (http://brat.nlplab.org/)
18. Balogh, P.: Gender-markerek a nyelvben (2009) http://webfu.univie.ac.at/wp/565
19. Carvalho, P., Sarmento, L., Silva, M. J., Oliveira, E.: Clues for Detecting Irony in User Gene-rated Contents: Oh...!! It's "so easy" ;-). University of Lisbon, Faculty of. Sciences, LASIGE. (2015)
20. Szabó M.K. 2015.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In Gecső T., Sárdi Cs. (eds.) Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához 177. Budapest, Tinta. pp. 278–285.