Szentiment- és emóciószótárak eredményességének mérése emóció- és szentimentkorpuszokon

Drávucz Fanni[1], Szabó Martina Katalin[2,3], Vincze Veronika[2,4]

[1] Eötvös Loránd Tudományegyetem,
Bölcsészettudományi Kar, Nyelvtudományi Doktori Iskola
[2] Szegedi Tudományegyetem
szabo.martina@lit.u-szeged.hu, vinczev@inf.u-szeged.hu
[3] Precognox Informatikai Kft.
mszabo@precognox.com
[4] MTA-SZTE Mesterséges Intelligencia Kutatócsoport

Kivonat: A cikkben a szövegekben megbúvó szentiment-, valamint emotív szemantikai tartalmak összefüggéseit vizsgáljuk. A munka keretében egy kézzel annotált szentimentkorpuszt elemzünk két különböző kategóriaszámú emóciószótárral, valamint egy kézzel annotált emóciókorpuszt elemzünk egy szentimentszótár segítségével. Ezt követően a szótáras elemzésekkel kapott eredményeket összevetjük a korpuszok annotációjával. A vizsgálatok célja annak feltérképezése, hogy kiegészítheti-e, és ha igen, mennyiben a két tartalomelemzési megoldás egymás eredményeit, eredményességét. A bemutatott elemzések és eredmények egyedülállóak; nincs tudomásunk olyan dolgozatról, amely hasonló megoldásokat prezentálna. Ugyanakkor a dolgozatban amellett érvelünk, hogy a két elemzési módszer együttes vizsgálata hasznos és eddig ismeretlen eredményeket tárhat fel.

1 Bevezetés

A számítógépes nyelvészetben a szentimentek alatt a szerzői attitűdöt tükröző nyelvi elemeket [11], míg az emóciók alatt a szöveg szintjén tetten érhető érzelmeket értjük [10], melyeket a háttérben húzódó kognitív értékelő illetve emotív funkciók különböztetnek meg. E két fogalmi kategória ugyanakkor legfeljebb részben mutat átfedést. Ahogyan ugyanis arra Péter [7] rámutat, az értékelésnek létezik mind emocionális (1a), mind racionális (1b) típusa:

(1) a. a főnököm remek ember
b. a habbeton rossz hővezető

A szentimentelemzés vagy véleménykivonatolás (*sentiment analysis* vagy *opinion mining*) a természetesnyelv-feldolgozás részterülete, amely a szerzői attitűdöt tükröző nyelvi elemek detektálására, valamint értékének (*sentiment orientation*) és tárgyának (*target*) a megállapítására törekszik automatikus megoldások segítségével. Ezzel szemben az emócióelemzés (*emotion detection* vagy *emotion recognition*) a szövegekben megbúvó emóciótartalom kinyerését célozza. Jelen dolgozatban e két

tartalomelemzés feladatkörébe tartozó megoldás alkalmazásának eredményeinek az összefüggéseit vizsgáljuk, szótárak és kézzel annotált korpuszok segítségével.

Bár véleményünk szerint a két megoldás eredményei hatékonyan egészíthetik ki egymást, nincs tudomásunk olyan dolgozatról, amely e két módszert e szempontból, egymás összefüggésében vizsgálná. Ennek okát a következő sajátosságokban látjuk: Egyrészt, az érzelmek szövegalapú elemzésével csekély számú dolgozat foglalkozik a nyelvtechnológia tárgykörében (vö. pl. [8, 5]), ez összességében a magyar nyelvű szövegek elemzésére is igaznak tekinthető (vö. [2, 10, 12]). A nyelvtechnológusok kis vagy kisebb jelentőséget tulajdonítanak az emócióknak, mint az úgynevezett szentimenteknek, azaz a nyelvi értékelésnek, illetve az emóciókat a hazai nyelvtechnológia alapvetően a szentimentelemzés tárgykörébe utalja; a szentiment- és az emócióelemzés feladatát gyakran azonosítja egymással (vö. pl. [7: p202]). Ugyanakkor azt is érdemes megemlíteni, hogy a szövegek érzelmi szempontú tartalomelemzése komoly pszichológiai, nyelvészeti és nyelvtechnológiai kihívást támaszt a szakértők elé [2, 10, 12]. Figyelemre méltó azonban, hogy az érzelmek több más tudományos diszciplínában, így például a viselkedéstudományban vagy a pszichológiában központi szerepet töltenek be.

A jelen dolgozatban arra a kérdésre keressük a választ, hogy milyen összefüggés van a szövegekben levő emóciók és a nyelvi értékelés, másképpen a szentimentek között. Azt szeretnénk feltárni, hogy a két típusú szemantikai tartalom hogyan, illetve milyen mértékben mutat átfedést egymással, másképpen, mennyire jellemző a nyelvi értékelés és az emotív tartalmak összefonódása az általunk vizsgált szövegtípusokban, magyar nyelvű szövegekben.

Kutatási eredményeink [12] alapján amellett érvelünk, hogy a szövegekben megbúvó emóciótartalom kinyerése olyan értékes információkat hozhat a felszínre, amelyeket más tartalomelemző módszerek nem tárnak, illetve tárhatnak fel. Ezzel összefüggésben úgy véljük, hogy az emóció- és a szentimentelemzés módszere hatékonyabb, egymást kiegészítő tartalomelemző megoldáshoz vezethet.

2 A szótárak bemutatása

Az emóciókorpusz szentimentjeinek elemzéséhez egy saját készítésű szentiment-szótárat [9] használtunk. Szótárunkat részben automatikus, részben manuális módszerrel hoztuk létre, magyar nyelvű szövegek automatikus szótáralapú szentimentelemzése céljából . A szótár készítése során nem csupán mellékneveket, hanem főneveket, határozószókat és igéket is felvettük, amennyiben úgy ítéltük, hogy az adott nyelvi elemnek inherens negatív vagy pozitív szentimentértéke van. Az így elkészített szótárunk kutatási célokra szabadon hozzáférhető.[1]

A szövegbeni érzelmek elemzéséhez két emóciószótárat alkalmaztunk. Az egyik emóciószótárat két, kézzel készített, hat kategóriából álló szótárból (vö. [4, 10]) készítettük, a két szótár egyesítésével. A Mérő-féle gyűjtés eredetileg nem tartalmazott kategóriákat, a szótárak összefésülése érdekében azonban elemeit kategorizáltuk.

1 http://opendata.hu/dataset/hungarian-sentiment-lexicon

Mindkét szótár az emóciókifejezések osztályozásában Ekman és Friesen [1] érzelemkategorizálási rendszerét követi, tehát azt a hat alapérzelmet veszi alapul, amelyek arckifejezéseit a kutatások alapján kultúrafüggetlenül azonos módon produkáljuk és azonosítjuk. Az alapérzelmek a szerzők alapján a következők: az öröm, a düh, a bánat, a félelem, az undor és a meglepődés. A két szótár egyesítésével készült lexikon statisztikai adatait az **1. táblázat első sorá**ban közöljük.

A másik emóciólexikon, amellyel dolgoztunk, egy a hat kategóriás szótár nyolc kategóriásra bővített változata volt (vö. [12]). Az új kategóriarendszer létrehozását egy korábbi emóciókorpusz kézzel való annotációjának tapasztalata indokolta (vö. [10]). A korpusz létrehozásának célja, hogy az emóciók nyelvi viselkedését valós nyelvi anyagon vizsgálhassuk, valamint a szótáraink hatékonyságát tesztelhessük és fejleszthessük. A feldolgozói munka során azonban azt tapasztaltuk, hogy számos nyelvi elemet a meglevő hat kategóriával nem tudunk lefedni, ezért a munka második szakaszában nyolc kategóriával átdolgoztuk a meglevő teljes emóciólexikont. A két újonnan felvett kategória a feszültség és a vonzalom/szeretet volt.

A nyolc kategóriás lexikont a következő lépésekben hoztuk létre: Először mindkét, fentebb említett kiinduló szótárnak kézzel kialakítottuk a lexikáját a nyolc kategóriának megfelelően, egymástól függetlenül. A munka során az egyik kiinduló szótár (vö. [10]) anyagát – a már említett korpuszannotálási tapasztalatok alapján – további elemekkel is kiegészítettük. Ezt követően a két szótárat egyesítettük. Az így kialakított szótár statisztikai adatait az **1. táblázat második sora** mutatja be:

E.szótár	öröm	düh	bánat	félelem	undor	meglepődés	feszültség	vonzalom/szeretet	**Összesen:**
6 kat.	719	394	360	229	121	68	--	--	**1 891**
8 kat.	675	410	387	243	135	97	309	186	**2 442**

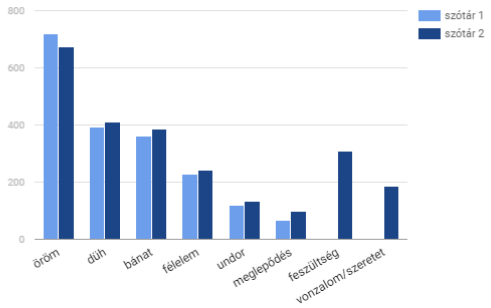
1. táblázat: Az emóciószótárak statisztikai adatai

A szótár kategóriarendszerének átszervezése az alábbi arányváltozásokat eredményezte az egyes emóciótípusokhoz tartozó elemek számában. Amint azt az 1. ábra megmutatja, egyedül az öröm kategóriába tartozó elemek száma kisebb a nyolc kategóriás szótárban, mint a hat kategóriásban. Ennek valószínűleg az az oka, hogy számos elem, amelyeket korábban az öröm-csoportban vettünk fel, megjelent az új vonzalom/szeretet kategóriában, és ezen az elemeket az öröm kategóriájából töröltük. Az összes többi emóciótípusban számbeli növekedést látni (**1. ábra**), amelynek részbeni oka az, hogy a szótár anyagát – az átkategorizáláson túl – újabb elemekkel is kiegészítettük.

3 Az elemzett korpuszok bemutatása

A kézzel annotált emóciókorpuszt kettős céllal hoztuk létre: egyrészt, hogy az emóciók nyelvi viselkedését valós nyelvi anyagon vizsgálhassuk, másrészt, hogy a

szótáraink hatékonyságát tesztelhessük és fejleszthessük [10]. Az emóciókorpusz szöveganyagát a 2014-es év folyamán keletkezett, tévés és mozis témájú blogoldalakról származó, különböző terjedelmű és szerzőségű kritikákból, hírekből, valamint kommentekből állítottuk össze, mely 15 987 mondatból és 197 707 tokenből áll. A magyar nyelvű, kézzel annotált szentimentkorpuszt termékvélemény-szövegekből hoztunk létre kutatási és fejlesztési céllal [11, 13]. Az adatbázis összesen 154 véleményszöveget, 17 059 mondatot és 251 202 tokent (központozással) tartalmaz.



1. ábra: Az emóciószótárak kategóriánkénti megoszlási arányai

4 Eredmények

Munkánk során a szentimentszótárral az emóciókorpuszt, az emóciószótárakkal pedig a szentimentkorpuszt elemeztük, azaz szótárillesztést hajtottunk végre. Amennyiben a szentimentszótárban szereplő elem illeszkedett az emóciókorpusz egyik lemmájára, akkor azt találatként értékeltük, és viszont. Az elemzések eredményeit a jelen fejezetben ismertetjük.

4.1 Az emóciókorpusz elemzése a szentimentszótárral

Az elemzéshez csupán azokat az érzelmeket vettük figyelembe, amelyek valamely konkrét emóció tagjét viselték. Amennyiben tehát egy adott elem esetében csak egy általános Emotion címke volt megadva, kihagytuk.

Az emóciókorpuszban összesen 397 annotált, emotív szemantikai tartalmú fragmentumot találtunk. Amint arról a korpusz jellemzésénél már említést tettünk (l. fentebb, 3.1), a korpuszban összesen hét emóciókategóriát annotáltunk.

Az emóciófragmentumok megoszlási arányait, valamint kategóriánként a szentimentszótárral megtalált fragmentumok számát a 2. táblázatban közöljük.

A táblázat azt mutatja meg tehát, hogy hány fragmentumot sikerült azonosítanunk a pozitív és a negatív szentimentszótárunk segítségével. A százalékos adatok azt jelzik, az összes fragmentumból hányat talált meg minimum egy elem a negatív, és hányat a

pozitív szótárból. A kapott adatok alapján a következő megállapításokat tehetjük: A szentimentszótárral az annotált emóciókifejezések megtalálási aránya jelentősnek tekinthető. A negatív szólistával az elemek 52,4%-át, a pozitív listával 41,1%-át sikerült detektálni. A legmagasabb eredményeket a negatív lexikonnal a bánat (72,3%) és a feszültség (71,2%) esetében, míg a pozitív lexikonnal az öröm (73,2%) esetében értük el.

Emóciókorpusz		**Szentimentszótárral azonosított elemek száma**	
tag	**fragmentumok**	**negatív**	**pozitív**
harag	58	35: 60,3%	22: 37,9%
feszültség	73	52: 71,2%	15: 20,5%
undor	11	8: 72,7%	5: 45,5%
félelem	25	16: 64,0%	10: 40,0%
öröm	112	25: 22,3%	82: 73,2%
bánat	83	60: 72,3%	22: 26,5%
meglepetés	34	12: 35,3%	7: 20,6%
Összesen:	**397**	**208: 52,4%**	**163: 41,1%**

2. táblázat: Az emóciókorpuszon végzett szentimentszótáras elemzés eredményei

Annak tekintetében, hogy a különböző emóciókat mely szótárral találtuk meg, és ezek között a találatok között milyen a megoszlási arány, a következőket mondhatjuk el: A találati arány a legtöbb esetben tükrözi az adott érzelem polaritását, azaz pozitív vagy negatív voltát. A meglepetés találati kiegyenlítettsége érthető, tekintettel arra, hogy ez az egyetlen alapemóció, amely pozitív és negatív egyaránt lehet. Ugyanakkor az figyelemre méltó, hogy a harag, az undor és a félelem emóciófragmentumokat is nagy arányban detektálja a velük ellentétes polaritású szentimentlexikon.

A **3. táblázat** megmutatja az eredményeket úgy, ha az emóciókategóriákat polaritásuk alapján összevonjuk. A meglepetés kategóriát, a fentebb említett ok miatt külön sorként tüntetjük fel.

Emóciókorpusz		**Sz.szótárral azonosított elemek száma**	
tag	**fragmentumok**	**negatív**	**pozitív**
negatív	250	171: 68,4%	74: 29,6%
pozitív	112	25: 22,3%	82: 73,2%
meglepetés	34	12: 35,3%	7: 20,6%

3. táblázat: Az emóciókorpuszon elemzése összevont szentimentszótárral

Úgy véljük, hogy a jelenség oka – legalább részben – a negáló elemek használatában keresendő: valószínűleg gyakran fejezzük ki negatív érzelmeinket pozitív polaritású elemek tagadásával, és ezekben az esetekben a fragmentum polaritása nem egyezik meg a benne szereplő emóciókifejezés polaritásával

(pl. *egyáltalán nem örülök, nem volt elragadtatva*). A tapasztalatokat alaposabb vizsgálat tárgyává kívánjuk tenni a jövőben.

4.2 A szentimentkorpusz elemzése az emóciószótárral

4.2.1 Illesztés

A szótáralapú elemzés során a hat, majd a nyolc kategóriás emóciószótárat is a szentimentkorpuszra illesztettük. A szentimentkorpusz 15 675 fragmentumából a hat és a nyolc emóciót tartalmazó szótár alapján összesen 4 310 és 4 380 fragmentumot találtunk meg. Egy fragmentumon belül bizonyos esetekben több illeszkedése is volt a szótárnak. A hat kategóriás szótár esetében összesen 4586 illeszkedése volt a 4 310 fragmentumnak, illetve a nyolc kategóriás szótárnál 4 682 illeszkedés 4 380 fragmentumban. Részletesen lásd a **4-5. táblázatok**at.

Szentimentkorpusz		**Emóciószótárral azonosított fragmentumok**						
tag	**fragmen-tumok**	**bánat**	**düh**	**félelem**	**undor**	**megle-petés**	**öröm**	**Össz.:**
negatív	8 465	386; 4.6%	54; 0.6%	225; 2.7%	556; 6.6%	40; 0.5%	632; 7.5%	1 700; 20,1%
pozitív	7 210	86; 1.2%	18; 0.2%	37; 0.5%	392; 5.4%	97; 1.3%	2 063; 28.6%	2 610; 36,2%
Össz.:	**15 675**	**472; 3.0%**	**72; 0.5%**	**262; 1.7%**	**948; 6.0%**	**137; 0.9%**	**2 695; 17.2%**	**4 310; 27.5%**

4. táblázat: A hat kategóriás emóciószótár illesztése a szentimentkorpuszon

Sz.korpusz		**Emóciószótárral azonosított fragmentumok**								
tag	**frag-men-tumok**	**bánat**	**düh**	**félelem**	**feszült-ség**	**undor**	**meg-lepetés**	**öröm**	**sze-retet**	**Össz:**
negatív	8 465	386; 4,6%	41; 0,5%	225; 2,7%	52; 0,6%	556; 6,6%	40; 0,5%	545; 6,4%	101; 1,2%	1 739; 20,5%
pozitív	7 210	86; 1,2%	13; 0,2%	36; 0,5%	34; 0,5%	392; 5,4%	97; 1,3%	1 786; 24,8%	292; 4,0%	2 641; 36,6%
Össz:	**15 675**	**472; 3,0%**	**54; 0,3%**	**261; 1,7%**	**86; 0,5%**	**948; 6,0%**	**137; 0,9%**	**2 331; 14,9%**	**393; 2,5%**	**4 380; 27,9%**

5. táblázat: A nyolc kategóriás emóciószótár illesztése a szentimentkorpuszon

A nyolc kategóriás emóciószótár, bár 29,1%-kal nagyobb a hat kategóriásnál, csupán 70-nel (0,4%) több fragmentum annotációját eredményezte.

A szentimentkorpuszbeli fragmentumoknak körülbelül a negyedében (27,9% illetve 27,5%) azonosítottunk a két szótár segítségével emóciókifejezést. A korpusz szövegtípusaira tekintettel, amely termékvéleményeket és twitter bejegyzéseket, tweeteket tartalmazott megállapítható, hogy az emóció nem tipikus, illetve domináns formája a

szentiment kifejezésének, tehát a szentimentkorpusz a fenti eredmények alapján többnyire tárgyilagosnak tekinthető. A kapott adatokra támaszkodva a jövőben az objektivitás kérdését tovább kívánjuk vizsgálni más tartalomelemzési megoldásokkal (pl. funkciószó-megoszlás) is hasonló domainen.

A pozitív szentimentet tartalmazó fragmentumokban 16,1%-kal volt magasabb az emóciókifejezések aránya, ezen belül nem meglepő módon a legtöbb illeszkedést a pozitív emóciók (öröm, szeretet) eredményezték. Az aránytalanság lehetséges magyarázata, hogy a pozitív érzelmek kifejezése kisebb lexikai változatosságot mutat, ezért a szótárral való illesztés a szűkebb nyelvi eszközkészletet nagyobb arányban képes azonosítani. Ezt alátámasztja az emóciószótárban felülreprezentált negatív emóciók száma is. Lehetséges magyarázat még a politikai korrektség jelensége, azaz hogy az adatközlők a negatív véleményt a nyilvánosság miatt árnyaltabban, indirekt módon fejezik ki – vagy akár vissza is tartják. Ezenkívül feltételezhető, hogy mivel a korpusz alapjául szolgáló szövegek szerzői a felületet online közösségi felületként használják, ezért a közösségből való kizárás elkerülése végett tartózkodnak a potenciálisan megosztó, azaz szélsőséges, direkt érzelemkifejezéstől. E feltételezés alátámasztására más jellegű (nem online) közlésekből összeállított korpusz összehasonlító elemzése szükséges, amely feltárhatja, hogy a jelenség milyen mértékben jellemző a korpuszra, vagy általában a magyar nyelvű, írott formában megjelenő érzelemkifejezésre.

Minden egyes emóciónak mind a két szentiment fragmentumaiban volt illeszkedése, a legtöbb illeszkedése az öröm emóció kifejezésének, míg legkevesebb a düh emóció kifejezésénél figyelhető meg.

Tetten érhető a korreláció az illeszkedések emóciójának polaritása és az illeszkedő szentiment fragmentum polaritása között, azaz például a bánat, illetve az öröm ~4-szer nagyobb arányban illeszkedik a negatív, mint pozitív szentimentű fragmentumokra. A meglepődés az elemzett korpuszban többségében pozitív polaritású, tehát nem várt jó dologra vagy tulajdonságra utalt, mivel 2,5-szer nagyobb arányban illeszkedett a pozitív szentimentekre. Az undorhoz tartozó illeszkedéseknek csak 58,6%-a volt negatív, amelynek hátterében a negáció, vagy ellentételezés alkalmazása is lehet, ennek elemzését lásd később.

A következő szakaszban megvizsgáljuk, hogy ezen korrelációkat kihasználva a szentimentek polaritását az emóciószótárral mennyire pontosan lehet előrejelezni.

4.2.2 Előrejelzés illesztés alapján

Az előző szakaszban az illeszkedő emóciókifejezések és a szentiment fragmentum tag polaritása közötti korreláció alapján arra adódik lehetőség, hogy előrejelezzük a szentimentet az illeszkedő emóció fragmentumok segítségével.

Emóciónként az egyes mondatokra akkor jeleztünk előre az adott emócióval azonos polaritású szentiment taget, ha volt illeszkedés az adott emócióhoz tartozó fragmentumokkal. A meglepetés polaritása – ahogy arra már korábban is utaltunk (l. fentebb, 4.2.1) – nem egyértelmű, mégis a fentebb bemutatott vizsgálat során azt tapasztaltuk, hogy a jelen korpuszban pozitív polaritású öröm emócióhoz hasonlóan felülreprezentált a pozitív szentimentű mondatok között, így a meglepetés emóciókifejezés illeszkedése esetén pozitív szentimentet jeleztünk előre.

Az emóciónkénti előrejelzés mellett kombinált előrejelzést is végeztünk, ahol bármely negatív emóciókifejezés illeszkedése esetén a negatív szentiment taget jeleztünk

előre. Az emóciónkénti, valamint a kombinált módszerrel kapott előrejelzési eredményeket lásd a **6-7. táblázatok**ban.

A várakozásoknak megfelelően a pontosság lényegesen magasabb a fedésnél, hiszen az emócióilleszkedés a szentiment-fragmentumok kevesebb mint 30%-ban fordult elő (l. fentebb, 4.2.1). A nyolc kategóriás emóciószótár kombinált pontossága alacsonyabb a hat kategóriás szótárénál, míg fedése a kombinált esetében magasabb, aminek hátterében a nagyobb szótárméret valószínűsíthető.

Emóciószótár	Szentiment	Pontosság (P)	Fedés (F)	F1
bánat	negatív	81,8%	4,6%	8,6%
düh	negatív	75,0%	0,6%	1,3%
félelem	negatív	85,9%	2,7%	5,2%
undor	negatív	58,6%	6,6%	11,8%
meglepetés	pozitív*	70,8%	1,3%	2,6%
öröm	pozitív	76,5%	28,6%	41,7%
Kombinált	**negatív**	**67,3%**	**12,8%**	**21,5%**
	pozitív	**76.0%**	**29.4%**	**42.4%**

6. táblázat: A hat kategóriás emóciószótár osztályozási eredményei

A nyolc kategóriás szótár pontossága az egyes emóciók esetén mindössze 2 százalékponton belüli eltérést mutatott a hat kategóriás szótárral való illesztéshez képest. A legnagyobb pontosságot mindkét szótár esetén a félelem emóció mutatta (85,6% illetve 86,2%), míg a legalacsonyabb pontosságot az undor (58,6%).

A hibaelemzéssel megállapítottuk, hogy az undor emóció 391 pozitív szentimentű fragmentumra illeszkedett (fals pozitív hiba), amelyből 358 (92%) fragmentum tartalmazott negációt (például: „Az íze nem rossz”, „Ezzel a szaloncukorral biztosan nem fog rosszul járni”). Ez alapján valószínűsítettük, hogy a negáció figyelembevételével tovább javítható az előrejelzés pontossága.

A hat és nyolc kategóriás szótár esetében is megfigyelhető, hogy míg a pozitív szentimentű fragmentumok több mint negyedében illeszkedett vele azonos polaritású emóciókifejezés, addig a negatív szentimentet tartalmazó fragmentumok esetében csak a nyolcada mutatott azonos polaritású illeszkedést. Ez arra enged következtetni, hogy a korpuszban diverzebbek, konfúzabbak a negatív érzelmeket kifejező nyelvi szerkezetek.

A hat kategóriás emóciószótárban az undor érzelmet kifejező emóciófragmentumok negatív szentimentek előrejelzésének a pontossága önmagában a legalacsonyabb (58,6%). Ha figyelembe vesszük a tagadószavakat, ez az érték lényegesen javul (93,9%). További kutatási célként megfogalmazható, hogy érdemes megvizsgálni az undort kifejező emóciófragmentum azon eseteit, amikor negáció kapcsolódik hozzá.

A meglepetés emóciószótár a szentiment korpuszon jobban jelezte előre a pozitív szentimentet, mint a negatívat annak ellenére, hogy az emóciót semlegesnek gondoljuk. A meglepetés emóciószótár alapvetően semlegesnek tekinthető (pl. *nem számít rá, hihetetlen*), de van néhány polarizált kifejezés (pl. *csodál, megilletődés, szörnyülködés, lefagy*) is.

Emóciószótár	Szentiment	Pontosság (P)	Fedés (F)	F1
bánat	negatív	81,8%	4,6%	8,6%
düh	negatív	75,9%	0,5%	1,0%
félelem	negatív	86,2%	2,7%	5,2%
feszültség	negatív	60,5%	0,6%	1,2%
undor	negatív	58,6%	6,6%	11,8%
meglepetés	pozitív*	70,8%	1,3%	2,6%
öröm	pozitív	76,6%	24,8%	37,4%
szeretet/vonzalom	pozitív	74,3%	4,0%	7,7%
Kombinált	**negatív**	**66,9%**	**13,2%**	**22,1%**
	pozitív	**75,7%**	**29,4%**	**42,4%**

7. táblázat: A nyolc kategóriás emóciószótár osztályozási eredményei

A hat és nyolc kategóriás emóciószótárak szentiment előrejelzése hasonló eredményeket mutatott. Mind a nyolc, mind a hat kategóriás szótárban voltak olyan fragmentumok, amelyekben az emóció több szótárilleszkedést is mutatott. A nyolc kategóriás szótárnál 98 olyan fragmentum volt, amely több emóció esetén is mutatott szótárbeli illeszkedést, (pl: „*Borzalmas* gagyi csoki darabok vannak benne", „*Csodá-san* kipárnázott nagyon jó futócipő") de ezek túlnyomórészt azonos szentiment-polaritású mondatban fordultak elő, azaz jól osztályoztak. Ezek az elemek (a példákban eltérő szedéssel emeltük ki) a különböző szótáralapú tartalomelemzési feladatokban azért problémásak, mert a lexikai szintű emotív tartalmuknak, illetve polaritásuknak megfelelően szerepelnek az emóció- vagy a szentimentszótárban, és szótáras elemzésük is ennek megfelelően történik. A problémáról l. [14].

Kivételként hozható fel például a negatív szentiment fragmentumok és az öröm emóció illesztése („mégsem nyerte el a *tetszésünket*", „Az amúgy nagyon jó reklámai miatt már-már jó sörnek tűnő [márkanév] a vakteszten jó nagyot bukott") vagy a pozitív szentiment-fragmentum és undor emóció illesztése („*rossznak* sem *rossz*").

A magas pontosság azt sejteti, hogy az emóció és a szentiment egymástól nem független, viszonyuk úgy írható le, hogy az emócióból következik a szentiment, azaz az emóciók az egyes szentimentek alfajai.

A tapasztalatok alapján az előrejelzéseket megismételtük úgy, hogy az emóciószótár-beli illeszkedés és negáció együttes jelenléte esetén is az adott emócióval ellentétes polaritású szentimentet jeleztük előre, azaz negáció jelenléte esetén figyelmen kívül hagytuk az illeszkedést. Például az „íze nem *rossz*" fragmentum esetén bár illeszkedik a „rossz" kifejezés ami az undor emóciószótárban megtalálható, a negáció miatt mégis az undorral ellentétes, tehát pozitív szentimentet jelzünk előre. Az így kapott eredményeket lásd a **8-9. táblázatok**ban.

A várakozásoknak megfelelően a negáció figyelembevétele még az együttes előfordulás naiv megközelítésével is átlagosan 14,9, illetve 13,2 százalékponttal növelte az emóciónkénti pontosságot, a fedés törvényszerű csökkentése mellett.

A tagadószavak figyelembevétele nem javította ugyanakkor a meglepetés emóció illeszkedésének pontosságát, ahol enyhe csökkenés volt (kevesebb, mint 1%). Ezt

részben magyarázhatja, hogy az emóciószótárban vannak olyan kifejezések, amelyek negációt tartalmaznak. Ezt a jelen (naiv) módszer nem kezelte megfelelően.

Emóciószótár	Szentiment	Pontosság (P)	Fedés (F)	F1
bánat	negatív	89,8%	3,5%	6,8%
düh	negatív	93,9%	0,5%	1,1%
félelem	negatív	90,0%	2,4%	4,7%
undor	negatív	93,9%	6,0%	11,3%
meglepetés	pozitív*	70,1%	1,2%	2,4%
öröm	pozitív	88,7%	27,1%	41,5%
Kombinált	**negatív**	**91,1%**	**11,0%**	**19,7%**
	pozitív	**87,6%**	**27,8%**	**42,2%**

8. táblázat: A hat kategóriás emóciószótár eredményei negáció kezelésével

Emóciószótár	Szentiment	Pontosság (P)	Fedés (F)	F1
bánat	negatív	89,8%	3,5%	6,8%
düh	negatív	92,5%	0,4%	0,9%
félelem	negatív	90,0%	2,4%	4,7%
feszültség	negatív	90,5%	0,4%	0,9%
undor	negatív	93,9%	6,0%	11,3%
meglepetés	pozitív*	70,1%	1,2%	2,4%
öröm	pozitív	88,7%	23,5%	37,2%
szeretet	pozitív	88,7%	3,8%	7,3%
Kombinált	**negatív**	**91,0%**	**11,3%**	**20,1%**
	pozitív	**87,6%**	**27,8%**	**42,2%**

9. táblázat: A nyolc kategóriás emóciószótár eredményei negáció kezelésével

A negációra történő szűrés után körülbelül egyforma volt a pontossága a hat kategóriás és nyolc kategóriás szótárnak, természetesen az előfordulási gyakoriság minden emóció esetében másképp alakul. A meglepetés ennél az esetnél is kivételt képez, az *elképesztő(en), hihetetlen(ül)* intenzifikálói funkcióban való használata miatt, amikor is a vizsgált elemek nem a lexikai szintű polaritásukat hordozzák, hanem puszta fokozó szerepet töltenek be (vö. [14]).

4.2.3 Részletes hibaanalízis

A szentimentkorpusz emóciószótárakkal történő vizsgálata során az előrejelzési hibákat és kapcsolódó észrevételeket a következő kategóriákba soroltuk. *Szótári hibák*nak neveztük azokat az eltéréseket, amikor az illesztés a szótár tartalmára visszavezethetően elmaradt, vagy nem a megfelelő emócióra vonatkozott:

- Hiányos szótár vagy téves illesztés állandósult szókapcsolatok, szólások esetén: „A többi olyan , mint halottnak a *csók*”, „Te *jó* ég”, „wow”,)
- Azonos alakú („Folyékony *zsír*”, „*csípi* a torkom”)

- Többtagú kifejezések („kellemes *csalódás*”, „nem *rossz*_”)
- Félrekategorizálás („Jaj” - a bánatban van, de meglepődésnél nincsen)
- Eltérő szófaj („feldobottság” vs. „feldob”; „szeretet” vs. „szeret”)
- Hangulatfestő („fúúúúj”, „ööö”)
- *Módszertani hibák*nak neveztük azokat a jelenségeket, amikor a szótár megfelelő volt, de az illesztésnél eltérés volt tapasztalható („*Tökéletesen* közepes”, „csak *fél* ponttal csúszott le a harmadik helyről”; nincs, de kellene „jól esett”, „meg vagyok elégedve”, „bejön”, „fúj”).

A következő csoportba tartoznak azok a hibák, amelyek különböző nyelvi jelenségek nem megfelelő kezeléséből adódtak:

- Óhajtó értelem (pl. „lehetne sokkal *jobb* is”)
- Szarkazmus (pl. „ha használhatatlan zsebkendőre vágyik , ez lesz az ön *ideális* ár / érték arányú terméke”)
- Kulturálisan kódolt jelenségek („műanyag íz”, „*kipirosítva* az orrot”)
- Kontextus ismeretére lenne szükség a pontos meghatározáshoz („Ebben *tuti*, hogy 20%-os ecet van”)
- Ellentétes polaritású határozó (pl. „*iszonyú* trendi”, „*borzasztó* finom amellett”, „*Elképesztően* kellemetlen íze van”, ”fázott is *rendesen* a lábam„)
- Burkolt negáció, viszonyítás („Voltak *jobbak*”, „Több rosszat kapott, mint *jót*”, „feltételeztem, a hús is *jó* benne”, „pont annyira *keserű*, amennyire kell”)
- „Jó”, „szépen” mint negatív-nyomatékosító (pl. „Sűrű a *jó* sok zselatintól”, „az a mennyiség most *szépen* ki is engedett”)
- Kettős tagadás („túl bizarr , hogy *utálni* lehessen”)

5 Összegzés

Munkánkban a szövegekben megbúvó szentiment-, valamint emotív szemantikai tartalmak összefüggéseit vizsgáltuk, melynek során kézzel annotált szentiment- illetve emóciókorpuszt elemzünk különböző emóció- illetve szentimentszótárakkal. A szótáras elemzésekkel kapott eredményeket összevetése a korpuszok annotációjával rejtett összefüggésekre mutatott rá. Például az egyszavas emóciószótár-illeszkedések a szentimentkorpuszon a fragmentumok alig negyedében fordultak elő, de már pusztán a negáció együttes előfordulásának figyelembevételével is 90% körüli pontosságot mutattak a szentiment polaritásának azonosításában. A manuális hibaanalízis több módszertani, nyelvi, pragmatikai és kognitív okot tárt fel, melyek magyarázzák a pontosságot csökkentő fals illeszkedéseket és lehetővé teszik a bemutatott és hasonló módszerek továbbfejlesztését. Ugyanakkor az elemzés során nem került azonosításra olyan fragmentum, amely a kontextus ismerete nélkül is egyértelműen a jelölt szentimenttel ellentétes polaritású érzelmet jelenített volna meg, így jelen korpusz esetében az emóciók a szentimentek aleseteinek tekinthetők a fragmentumok szintjén.

A bemutatott elemzések és eredmények egyedülállóak; nincs tudomásunk olyan hazai munkáról, amely hasonló megközelítést vizsgálna. A vizsgálat során feltérképeztünk olyan kutatási, alkalmazási és továbbfejlesztési lehetőségeket, amelyekben kiegészítheti egymást a két tartalomelemzési megoldás.

Köszönetnyilvánítás

Jelen kutatás az Emberi Erőforrások Minisztériuma (EMMI) Új Nemzeti Kiválóság Program (ÚNKP) támogatásával valósult meg.

Bibliográfia

1. Ekman, P., Friesen, W.V. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* 1: 49–98.
2. László J., Ehmann B. 2004. A narratív pszichológiai tartalomelemzés új eljárása: A LAS-Vertikum. In: Erős F. (szerk.): *Magyar Pszichológiai Szemle Könyvtár: Az elbeszélés az élmények kulturális és klinikai elemzésében.* Akadémiai Kiadó, Budapest. 75−87.
3. Liu, B. 2012: *Sentiment Analysis and Opinion Mining.* Draft. Elérhető: http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
4. Mérő L. 2010. *Az érzelmek logikája*. Tericum, Budapest
5. Mulcrone, K. 2012. Detecting Emotion in Text. Elhangzott: UMM CSci Senior Seminar Conference. Amerikai Egyesült Államok, University of Minnesota: Morris. 2012. ápr. 28. https://wiki.umn.edu/pub/UmmCSciSeniorSeminar/Spring2012Talks/KaitlynM ulcrone.pdf
6. Péter M. 1991. *A nyelvi érzelemkifejezés eszközei és módjai.* Tankönyvkiadó, Budapest.
7. Pólya T., Csertő I., Fülöp É., Kővágó P., Miháltz M., Váradi T. 2015. A véleményváltozás azonosítása politikai témájú közösségi médiában megjelenő szövegekben. In: *XI. Magyar Számítógépes Nyelvészeti Konferencia*. 198–209.
8. Strapparava, C., Mihalcea, R. 2008. *Learning to identify emotions in text.* SAC 2008. http://web.eecs.umich.edu/~mihalcea/papers/strapparava.acm08.pdf
9. Szabó M.K. 2015. Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In: *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához* 177. Tinta, Budapest. 278–285.
10. Szabó M.K., Morvay G. 2015. Emócióelemzés magyar nyelvű szövegeken. In: *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához* 177. Budapest, Tinta. pp. 286–292.
11. Szabó M. K., Vincze V. 2015. Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai, In: *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*. Szegedi Tudományegyetem, Szeged. 219−226.
12. Szabó M.K., Vincze V., Morvay G. 2016a. Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: *Távlatok a mai magyar alkalmazott nyelvészetben.* Budapest: Tinta
13. Szabó M.K., Vincze V., Simkó K., Varga V., Hangya V. 2016b. A Hungarian Sentiment Corpus Manually Annotated at Aspect Level. In: *Proceedings of LREC 2016.* Portoroz, Szlovénia Portoroz: European Language Resources Association (ELRA). 2873-2878.
14. Szabó M.K. *The usage of elements with emotive semantic content from a gender point of view.* Kézirat.