HunEmPoli: magyar nyelvű, részletesen annotált emóciókorpusz

Ring Orsolya[1], Vincze Veronika[2], Guba Csenge[1,3], Üveges István[1]

[1]Társadalomtudományi Kutatóközpont, Politikatudományi Intézet
Budapest, Tóth Kálmán utca 4.
[2]ELKH-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
[3] SZTE Nyelvtudományi Doktori Iskola
Szeged, Egyetem utca 2.
{ring.orsolya, uveges.istvan}@tk.hu
vinczev@inf.u-szeged.hu
csenge.guba@gmail.com

Kivonat Cikkünkben egy részletesen annotált, emócióelemzésre használható korpuszt mutatunk be, amely a projekthez kidolgozott emóciókategóriarendszer szerint, tagmondat szinten került annotálásra, alapját pedig 1008 db az Országgyűlés ülésein 2014 és 2018 között elhangzott napirend előtti felszólalás szabadon elérhető szövege jelentette, összesen 764008 token terjedelemben. Munkánkban részletesen ismertetjük az alkalmazott kategóriákat, az annotás menetét, közöljük az alapvető korpuszstatisztikai és minőségbiztosítási adatokat, valamint példát mutatunk arra, hogyan használható a léterehozott korpusz gépi érzelem- vagy szentimentazonosításra.
Kulcsszavak: Argumentumszintű szentimentelemzés, emócióanalízis, ABSA, politikai kommunikáció, parlamenti beszéd

1. Bevezetés

Bár a politikai kommunikációnak korpuszokon és NLP-módszereken alapuló kutatása párhuzamosan a szövegek online elérhetőségének növekedésével egyre nagyobb jelentőségű, a megjelent tanulmányok legtöbbször a politikusok médiában és közösségi médiában elhangzó megnyilatkozásait elemzik (Gollust és mtsai, 2020; Mariani és mtsai, 2020; Aparicio és mtsai, 2021; Wang és mtsai, 2021; Rufai és Bunce, 2020).

A parlamenti viták leiratai lényegében a beszélt nyelv ellenőrzött és szabályozott körülmények között készült átiratai, melyek szabadon elérhetőek, mivel az információszabadságról szóló törvény alapján nem vonatkoznak rájuk a szerzői jogi vagy a személyes adatok védelmére vonatkozó jogszabályok. Éppen ezért az utóbbi években több nemzetközi projekt keretében készült és készül jelenleg is korpusz parlamenti felszólalásokból[1]. Hiszen a parlamenti viták jegyzőkönyvei

[1] Ilyen például a CLARIN https://www.clarin.eu/, a Comparative Agendas https://www.comparativeagendas.net/ vagy az OPTED https://opted.eu/ Projekt

egyedi tartalmuk, szerkezetük és nyelvezetük miatt fontos forrásai a társadalomtudományi és nyelvészeti kutatásoknak.

A parlamentek a politikai kommunikáció fontos helyszínei, ahol a választott képviselők megvitatják a benyújtott törvényjavaslatokat és más országos jelentőséggel bíró ügyeket. Az itt elhangzó beszédek általában előre megtervezett beszédaktusok, amelyek által a képviselők kiemelt célja nem pusztán tájékoztatás, hanem a hallgatóság meggyőzése és támogatásuk megszerzése is. Ezen beszédek gépi érzelemelemzése nagy kihívást jelent, és bár az elmúlt években több erőforrás is elérhetővé vált, ezek nagy része angolul áll rendelkezésre és alkalmazásuk jelentős munkával járó adaptációt igényel, különösen az olyan morfológiailag gazdag nyelvek esetén, mint a magyar (Jang és Shin, 2010; Mladenović és mtsai, 2016). A magyar nyelvű szövegek hatékony elemzéséhez így felmerült az igény egy mélyen annotált emóciókorpusz létrehozására, amely a későbbiekben jól alkalmazható majd különböző gépi tanítási feladatok végrehajtásánál, nyelvi modellek tanításánál. Így egy ilyen korpusz egyszerre szolgálhatja a politikatudományi mellett a számítógépes nyelvészeti- és mesterséges intelligencia-kutatásokat is.

2. Kapcsolódó irodalom

2.1. Politikai kommunikáció

A politikai kommunikáció az érzelmi hatások kiváltásával és a különböző eszmék terjesztésével a politikai cselekvést ösztönzi. A közelmúltban lezajott technikai és társadalmi változások következtében jelentősen megszaporodott a kommunikációban résztvevők, valamint a rendelkezésre álló kommunikációs csatornák száma, mindez pedig befolyásolta a politikai kommunikáció jellegét és intenzitását is. A politikai szereplők szerepük felértékelődésére a kommunikáció professzionalizálásával reagálnak. A politikai kommunikáció egyik kiemelt jelentőségű helyszíne a parlament, ahol a választott képviselők a benyújtott törvényjavaslatok és más nemzeti jelentőségű ügyek megvitatása közben igyekeznek olyan módon artikulálni véleményüket, hogy az minél szélesebb közönséget érjen el. A parlamenti viták során különféle témák merülnek fel, érvek és ellenérvek ütköznek és ezeken keresztül alakul ki egy politikai napirend, amely aztán tematizálja a nyilvános vitákat (Bene és Nábelek, 2019). Az érzelmek politikai kommunikációban való kifejezésének kutatása az utóbbi években egyre nagyobb hangsúlyt kapott mind a nemzetközi, mind a hazai társadalomtudományi kutatásokban (Szabó, 2020; Crigler és Just, 2012; Wagner és Morisi, 2019; Settle, 2020; Richards, 2004; Haselmayer és Jenny, 2017), de a politikai és különösen a parlamenti beszédek érzelmi töltésének NLP-eszközökkel történő elemzése viszonylag új és kevéssé kiforrott (Gold és mtsai, 2018; Jafarian és mtsai, 2021), különösen a magyar nyelvre.

2.2. Emócióelemzés

A szentiment- vagy emócióelemzés célja az egyes szövegek tartalmából kinyerni azokat az információkat, amelyek értékelést fejeznek ki. Az elemzés különböző szinteken végezhető részint attól függően, hogy mi az elemzés alapegysége

(például szöveg, mondat vagy tagmondat) és meghatározzuk-e azt, hogy az érzelem mire irányul, vagy mi váltja ki azt. A szakirodalom különbséget tesz a pozitív-negatív-semleges skálán mozgó szentiment és a több kategóriával dolgozó érzelemelemzés között, melyek közül az utóbbi sokkal több információt nyújt az adott egység érzelmi töltetéről (Marcus, 2000). Bár a két fogalmat sokszor egymás szinonimájaként is használják, jelen dolgozatban mi a fenti különbségtételt alkalmazzuk a szentiment-, illetve emócióelemzés kifejezések használatánál.

A legismertebb érzelemkategória-rendszerek valószínűleg Ekman (Ekman és Wallace V. Friesen, 1982) és Plutchik (Plutchik, 1964, 1980a, 1982) nevéhez fűződnek. Ekman, aki az emberi arckifejezések kultúrákon átívelő egységes jellegét tanulmányozta, 6 (harag, undor, félelem, boldogság, szomorúság és meglepetés), Plutchik pedig 8 kategóriát határozott meg az érzelmek osztályozására (várakozás, meglepetés, öröm, bánat/szomorúság, elfogadás, undor, harag, félelem).

A pszichológiai szakirodalomban az érzelem fogalma számos kérdést vet fel, és nincsenek formális kritériumok arra vonatkozóan, hogy mit tekintünk érzelemnek és mit nem (Chapman és Nakamura, 1998; Cabanac, 2002; Griffiths, 2008). Az érzelmek empirikus elemzése ugyanis rámutatott azok összetettségére (Lakoff és Kovecses, 1987), illetve a különböző érzelmek jelentésének összefonódására a helyi kultúra sajátosságaival (Wierzbicka, 1999), valamint a speciális érzelmi kifejezések eltéréseire a különböző nyelveken (Russell, 2003). Az érzelmek komplexitása pedig nagyban nehezíti azok NLP-módszerekkel történő elemzését.

Az érzelemelemzés másik alapvető kérdése, hogy pontosan mi hordoz érzelmeket a tartalom szintjén. Ezért például több kutatásban próbálkoztak azzal (Feng et al., 2013; Loukachevitch és Levchik, 2016), hogy olyan érzelemszótárakat készítsenek, amelyek nem csak explicit, hanem konnotatív érzelemértékkel is bírnak (Feng és mtsai, 2013; Loukachevitch és Levchik, 2016). Például a kitüntetés vagy az előléptetés pozitív konnotációval bírnak, míg a munkanélküliség és a terrorizmus negatívak. Ahogy Loukachevitch és Levchik megállapította, „a konnotációs értékkel bíró, nem érzelmi töltetű szavak általában a társadalmi élet negatív vagy pozitív jelenségeiről (tényekről) közvetítenek információt" (Loukachevitch és Levchik, 2016), így ezen konnotatív szemantikai tartalmak automatikus elemzése jelentős kihívást jelent.

2.3. Magyar nyelvű szentiment- és emóciókorpuszok

A most bemutatott emóciókorpusz kialakításával a hazai szentiment- és emócióelemzési feldatokra alkalmas korpuszok, valamint a kapcsolódó kutatások két alapvető hiányosságára reflektáltunk. Az első ezek közül az eddig vizsgált doméneket illeti, tekintettel arra, hogy magyar nyelvű politikai szövegek gépi tanulással történő érzelemelemelemzésére korábban nem létezett megfelelő erőforrás, a második pedig abban keresendő, hogy az elmúlt évtizedben mindössze két olyan nyilvánosan elérhető magyar korpusz készült, amelyek szentimentelemzési feladatokra lettek annotálva, míg emócióelemzésre egy sem.

Az eddig rendelkezésre álló korpuszok közül az Opinhubank (Miháltz, 2013) megközelítőleg tízezer mondatos állománya mondatszinten, pozitív-negatív-semleges szentimentek annotációját tartalmazza, ezenfelül jelöli az adott vélemények

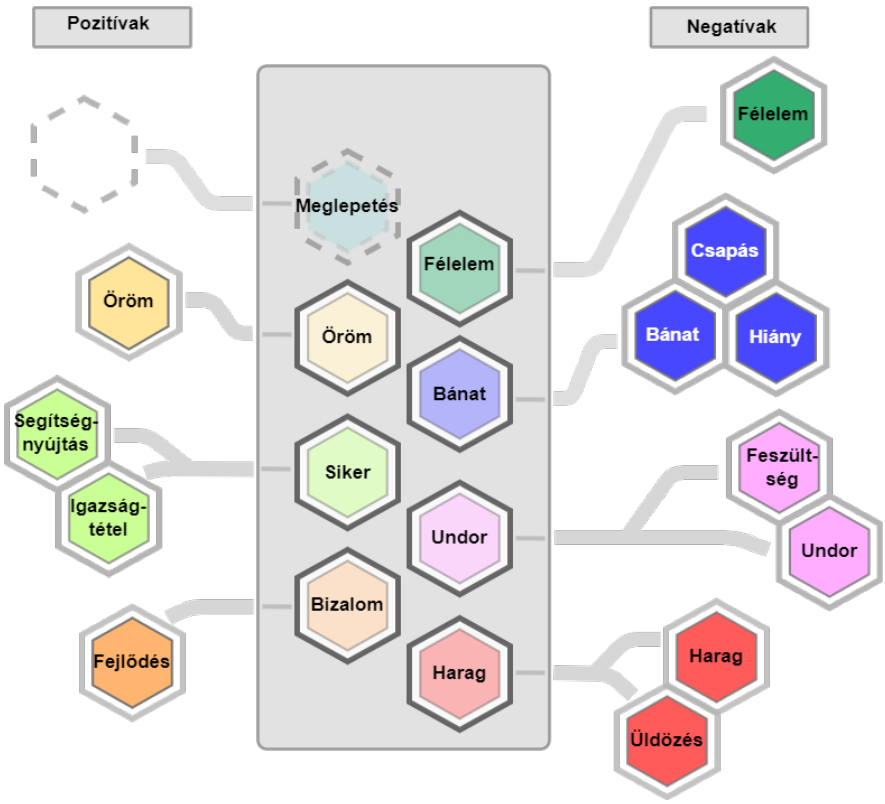
„célpontját" is a mondatokban megjelenő névelemek esetében. A forrásszövegeket ez esetben online hírforrások (híroldalak, blogok) jelentették. A HuSent korpusz (Szabó és mtsai, 2016) nagyjából 17 ezer mondatos állománya ezzel szemben online forrásból származó termékvéleményeket tartalmaz. A kapcsolódó annotáció egyebek mellett a (pozitív/negatív) szentimentértéket kifejező kifejezést és az érzelem célpontját tartalmazza.

3. A HunEmPoli korpusz létrehozása

A korpusz építéséhez a 2014-2018 közti parlamenti ciklusban elhangzó napirend előtti felszólalásokat használtuk fel. A vizsgálatunk alapjául szolgáló metaadatokkal ellátott adatbázis a Társadalomtudományi Kutatóközpont Politikatudományi Intézetének Hungarian Comparative Agendas Projektje keretében készült, kutatási célokra, regisztráció ellenében szabadon hozzáférhető. A napirend előtti felszólalások az országgyűlés plenáris ülésén hangoznak el országos jelentőségű, halaszthatatlan és rendkívüli ügyekben 2-5 perces időterjedelemben. A korpusz eltér a magyar nyelvre elérhető, beszélt nyelvi korpuszoktól, melyek spontán beszédet tartalmaznak és/vagy informális jellegűek (pl. Szabó és mtsai (2021), Vincze és mtsai (2021)), hiszen a műfajból adódóan beszélt nyelvi adatokról van szó, de ezek stílusa hivatalos, formális, a képviselők általában előre megtervezték/megírták beszédüket. A szövegek szinte minden esetben kizárólag magázó formulákat használnak, ugyanakkor a leiratok nem tartalmazzák az élőbeszédre jellemző hezitációkat, szüneteket, téves szókezdeteket. Sajátosságként ugyanakkor megemlítjük, hogy a beszédek közbeni bekiabálások is szerepelnek a leiratokban, ezeket azonban a későbbi annotáció során nem vettük figyelembe, hiszen az azokban található érzelmek nem az adott felszólalóhoz köthetők.

3.1. Az emóciókategóriák

Kutatásunk során létrehoztunk egy induktív emóció-kategóriarendszert, melyeknek kategóriái megfeleltethetőek Plutchik (1980b) emóció-kategóriarendszerének, amely nyolc osztályt különböztet meg (egy kivétellel, mivel szövegeinkben *meglepetés* nem volt azonosítható, így az az induktív kategóriarendszerből hiányzik) és átválthatóak a szentimentelemzésben használatos pozitív-negatív kategóriákra is. Erre a kibővített rendszerre azért volt szükség, mert korábbi tapasztataink szerint a politikai szövegekben található mondatok egyébként nem, vagy csak rendkívül alacsony annotátori egyetértés mellett voltak besorolhatóak az emócióelemzés során egyik leggyakrabban használt, Plutchik-féle kategóriarendszerbe, míg a bővített rendszer segítségével a korpusz magas kódolók közötti egyetértéssel volt annotálható. A végső annotációs útmutatóban összesen 12 ún. emóció-topikot (későbbiekben ET) határoztunk meg, melyek mindegyikéhez legalább három hívószót vagy kifejezést adtunk meg, hogy megkönnyítsük az annotátorok munkáját.



1. ábra: Plutchik kategóriái és a belőlük képzett emóciókategóriák a HunEmPoli korpuszban

Kapcsolódó fogalmak	Emóciótopik	Plutchik szerint	Szentiment
fenyegetettség, fenyegetés, félelem, rettegés, aggodalom	**Félelem**	Félelem	Negatív
hiány, nélkülözés, nyomor, nyomorúság, nincstelenség, gyötrelem, szenvedés, kín, sikertelenség, kudarc, negatív irányú változás	**Hiány, nélkülözés**	Bánat	
szomorúság, bánat, elkeseredettség, reménytelenség, mélabú	**Szomorúság, bánat**		
szerencsétlenség, katasztrófa, csapás	**Szerencsétlenség, csapás**		
terror, merénylet, üldözés, kegyetlenség, kegyetlenkedés, szervezett bűnözés, vandalizmus, rongálás, szándékos károkozás, bűntett, erőszak	**Terror, bűnözés**	Harag	
düh, harag, gyűlölet	**Harag**		
feszültség, zavar, érdekellentét, revans, megbüntetés	**Feszültség, zavar**	Undor	
lenézés, undor, irtózás, megvetés, gúny	**Lenézés, undor**		
javulás, közeledés, békülés, „relief”, enyhülés, fejlesztés, fejlődés, siker, pozitív irányú változás	**Javulás, fejlődés**	Siker	Pozitív
öröm, élvezet, vidámság, derű, szeretet, elfogadás, tolerancia	**Öröm**	Öröm	
segítségnyújtás, mentés, segély, gyógyítás, ellátás, kiszabadítás	**Segítségnyújtás**	Bizalom	
igazságtétel, nyomozás, törvényhozás, hivatalos szervek fellépése	**Igazságtétel, nyomozás**		

1. táblázat. Az emóciótopikok megfeleltetése a Plutchik-féle rendszer kategóriáinak és a pozitív/negatív szentimentértékeknek.

3.2. Annotálás

Az annotálást összesen hat annotátor végezte, akik részben nyelvészetet, részben politikatudományt tanuló MA- vagy PhD-hallgatók voltak. Az annotálás megkezdése előtt részletes annotálási útmutatót, valamint betanítást kaptak, majd próbaannotációkat végeztek. Az annotálást a tagtog nevű online is elérhető eszköz segítségével végeztük[2], melynek grafikus felülete megkönnyítette az annotátorok munkáját. Az annotátorok minden esetben egy-egy teljes napirend előtti

[2] https://www.tagtog.com/

felszólalást kaptak meg. Egy szöveget mindig csak egy annotátor annotált, de a korpusz 25 százalékát minőségbiztosítási célból ketten-ketten párban is feldolgozták. A minőségbiztosított fájlok esetén arra is ügyeltünk, hogy a teljes feladat során az annotátorpárok változzanak.

Az annotálást tagmondat szinten végeztük, azaz a tagmondatok szintjén volt szükséges meghatározni az ET-t, míg a tagmondatok közötti esetleges kötőszavakat nem kellett belevenni az annotációba. Több tagmondatot folyamatosan, együtt csak akkor lehetett annotálni, ha ugyanaz az ET folytatódott több tagmondaton keresztül (ebben az esetben értelemszerűen a kötőszavak is bekerültek az annotációba). Az annotálás során az annotátornak a következőket kellett megtalálnia és bejelölnie a szövegekben az ET-t tartalmazó tagmondatokon kívül : (1) az ET-t kifejező részt, azaz a kulcskifejezést, ami lehetett egy szó, de akár többszavas kifejezés is. Egy ET-n belül több kulcskifejezés is jelölhető volt. (2) az ún. argumentumot, tehát azt a dolgot, fogalmat (vagy akár annak részét vagy sajátságát), ami az azonosított ET-t kiváltja, vagy amire az ET irányul. Fontos, hogy az argumentum: (1) Nem keverendő össze az érzelem átélőjével, illetve az események elszenvedőjével. (2) Az argumentum egy vagy több szóból is állhat. Az annotáció során az argumentum jelölése csak akkor volt szükséges és lehetséges, ha az egyértelműen, szavak szintjén is megfogható volt. Rejtett névmások esetén nem jelöltük azokat. Abban az esetben, ha az argumentum egy névmás volt az adott egységben, azt ugyanúgy jelöltük, mintha tartalmas szó szerepelne a tagmondatban. Mivel az argumentum az ET-vel jelölt mondatrészen kívül is állhatott, így az annotáció során minden esetben összekötöttük az ET-t hordozó mondatrészt és a vele kapcsolatban álló argumentumot.

Az annotátorok első feladatként tehát meghatározták, hogy mely tagmondatok hordozzák a 12 ET valamelyikét, majd kiválasztották a felkínált lehetőségek közül az ennek megfelelő annotációs taget. Ezután pedig tetszőleges sorrendben bejelölték az adott ET-vel jelölt mondatrészhez tartozó kulcskifejezéseket és argumentumokat, illetve az argumentumok esetében összekötötték azt a hozzá tartozó mondatrésszel.

Az annotálás során felmerült az a probléma, hogy esetenként egy-egy (tag)mondatra több ET is illett. Pl.: *…ahol a multinacionális kereskedelmi cégek kezdeményezésével szemben az alkotmánybírósággal összefogtak a baloldali szakszervezetek*. Erre a mondatra a javulás, fejlődés ET, illetve a feszültség, zavar ET is kioszthatóvá lett volna. Ilyen esetekben az annotátorok azt az utasítást kapták, hogy azt az egy ET-t jelöljék, amelyet szerintük a kontextus jobban indokol.

Továbbá problémát okozott még bizonyos esetekben a kulcskifejezés hiánya: *ténylegesen Matolcsy György hozott döntést arról, hogy mikor mennyi pénzt kötnek le unokatestvérének bankjánál, mikor mennyi állampapírt vásárolnak unokatestvére bankjától*, vagy az olyan kulcskifejezések jelölése, ahol a kifejezés tagjai a mondaton belül távol állnak egymástól: *…hogy a háromgyermekes családoknak körülbelül csak a 8 százaléka tudja majd igénybe venni teljes mértékben a családi adókedvezményt.*. Az első esetben egyáltalán nem jelöltünk kulcskifejezést, a második példánál viszont a kulcskifejezésbe amúgy nem tartozó szavak is belekerültek a kijelölt szakaszba. Speciális esetként kezeltük az idiomatikus

kifejezéseket az argumentum jelölésének szempontjából. Az idióma alanyát és tárgyát - amennyiben azt a beszélő sem szó szerint értette - nem jelöltük külön argumentumnak, de az esetek többségében a kulcskifejezés részét képezte, pl.: *Nyilván sokan ismerik azt a közmondást, hogy más tollával nem illik ékeskedni.*

Emóciótopik	**Példa**
Félelem	*...és ez veszélyt jelent Magyarország biztonságára.*
Hiány, nélkülözés	*...miközben az embereket érintő adóterhek abban az évben is meg az azt követő évben is durván növekedtek.*
Szomorúság, bánat	*Sokukat elvitték málenkij robotra a Gulagra, nagyon sokan nem jöttek vissza közülük.*
Szerencsétlenség, csapás	*...mint a bányaszerencsétlenségben rekedt chilei bányászokat.*
Terror, bűnözés	*...leszámítva azt a szűk 10 százalékot, akik az állam vagyonát szétlopják, urizálnak, több mint 90 százalék rovására.*
Harag	*A tehetetlen dühöt érzem a kormány elhibázott bérpolitikájával és a társadalompolitikájával szemben.*
Feszültség, zavar	*Ők azok, akik fizetnek azért, hogy Trump mielőbb megbukjon.*
Lenézés, undor	*...majd fülét-farkát behúzva szépen hazaballagott, és azután az a bizonyos bankadó az ilyen iciri-piciri lett.*
Javulás, fejlődés	*...hogy megváltozik az életük; a munka tartást és méltóságot ad az embernek, javítja az önbecsülését, és nem utolsósorban pénzt hoz a házhoz.*
Öröm	*És nagyon örülünk ennek az eredménynek.*
Segítségnyújtás	*...egyrészt az üldözött, bajba jutott emberen segíteni kell...*
Igazságtétel, nyomozás	*...és ahol már az Európai Unió vizsgálóbiztosai is vizsgálódnak.*

2. táblázat. Példamondatok az egyes emóciókategóriák szerint.

3.3. Minőségbiztosítás

A korpusz minőségének biztosítása érdekében időről időre kiszámoltuk az annotátorok közti egyetértést (Cohen's Kappa). Mivel az annotátorok tagmondatok szintjén jelölték az emóciótopikokat, tokenszintű kiértékelést végeztünk, hiszen nem szerettük volna, ha az esetleges kisebb tévesztések (pl. írásjelek vagy kötőszavak jelölésében adódó eltérések) torzítják az eredményeket.

A minőségbiztosítás eredményei átlagosan 0,41 (Kappa) egyetértést mutattak, ami mérsékelt egyetértést jelent, ezért második körben egy annotátorpárral (akik az annotálás során minden összehasonlításban a legjobb eredményt produkálták) a teljes korpusz javítását elvégeztettük. Ennek eredményeként a javított korpusz átlagos egyetértése 0,7574 (Kappa) már kategóriánként erős egyetértést mutat, lásd a 3. táblázat adatai.[3] Ezen felül, az argumentumok esetében a mért

[3] A minőségbiztosításra véletlenszerűen kiválasztott fájlokban a 9-es és a 10-es ET nem fordult elő. Ennek oka az, hogy ezek a teljes korpuszban csak igen alacsony elemszámmal szerepelnek.

egyetértés 0,8947 (Kappa), míg a kulcskifejezések esetében 0,9101 (Kappa) szerint alakult, amely már a majdnem teljes egyetértést jelent.

ET	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**11**	**12**
Kappa	0,564	0,885	0,971	0,930	0,615	0,846	0,5	1	0,264	1

3. táblázat. Emóciótopikok esetében mért egyetértés - κ (1: félelem, 2: hiány, nélkülözés, 3: terror, bűnözés, 4: javulás, fejlődés, 5: feszültség, zavar, 6: lenézés, undor, 7: szomorúság, bánat 8: öröm, 11: segítségnyújtás, 12: igazságtétel, nyomozás).

3.4. Korpuszadatok

Az elkészült korpusz a korábban említetteknek megfelelően 1008 napirend előtti felszólalást tartalmaz, amelyek mindösszesen 764008 tokenből, illetve 36475 mondatból állnak. Ezek pártok szerinti megoszlását részletesen a 4. táblázat ismerteti. Jól látszik, hogy a parlamenti összetételnek megfelelően, a Fidesz képviselői szólaltak fel a legtöbbször. Ugyanakkor az is megfigyelhető, hogy a független képviselők, valamint a Jobbik felszólalásaiban szerepel átlagosan a legtöbb szó, azaz az ő felszólalásaik voltak a leghosszabbak. Ezzel ellentétben, a KDNP politikusai viszonylag röviden beszéltek.

Párt	**Token**	**Mondat**	**Token** (%)	**Mondat** (%)	**Beszéd**	**Token/ beszéd**	**Mondat/ beszéd**
LMP	156504	7430	20,48	20,37	197	794,44	37,72
KDNP	120844	6173	15,82	16,92	192	629,40	32,15
MSZP	142538	7222	18,66	19,80	193	738,54	37,42
Jobbik	160461	6900	21,00	18,92	197	814,52	35,03
Fidesz	176325	8415	23,08	23,07	221	797,85	38,08
Független	7336	335	0,96	0,92	8	917,00	41,88

4. táblázat. A korpusz alapvető statisztikai adatai.

Az 5. táblázat szemlélteti az egyes emóciótopikok pártonkénti megoszlását. Összességében a javulás számít a leggyakoribb emóciótopiknak, ez különösen a kormányoldali kommunikációban – azaz a Fidesz és a KDNP felszólalásaiban – nyilvánul meg. A második és harmadik leggyakoribb kategória, az érdekellentét és a romlás főként az ellenzéknél (LMP, MSZP, Jobbik) domináns. A 2. ábrán megfigyelhető az is, hogy e három kategória együttesen mintegy 75%-át teszik ki a korpuszban található emóciótopikoknak, azaz a többi 9 emóciótopik előfordulási gyakorisága jóval kisebb, mint e háromé (ezért is fontos, hogy az erre a három ET-re mért egyetértés magas, (0,81), azaz ezeket közel teljes egyetértéssel sikerült annotálni. E jelenség magyarázata valószínűleg a parlamenti beszédek

sajátságában rejlik: a napirend előtti felszólalások nagy hányada szól arról, hogy egy adott jelenség vagy problémakör, esetleg annak kormányzati kezelése javult-e, esetleg romlott az utóbbi időben. Ugyanakkor az érdekellentét erős jelenléte is indokolható a parlamenti vitakultúrával: a kormánypártok, illetve az ellenzéki pártok gyakran állítják szembe saját tevékenységüket a másik oldaléval, így az érdekellentétre utaló nyelvi kifejezések száma is magas a korpuszban.

	LMP	**KDNP**	**MSZP**	**Jobbik**	**Fidesz**	**Független**	**Összesen**
Félelem	133	96	87	133	161	9	619
	21,49%	15,51%	14,05%	21,49%	26,01%	1,45%	
Hiány, nélkülözés	2702	880	2370	1968	1156	56	9132
	29,59%	9,64%	25,95%	21,55%	12,66%	0,61%	
Terror, bűnözés	265	172	284	452	406	2	1581
	16,76%	10,88%	17,96%	28,59%	25,68%	0,13%	
Javulás, fejlődés	2031	2694	1923	2189	3454	164	12455
	16,31%	21,63%	15,44%	17,58%	27,73%	1,32%	
Feszültség, zavar	2202	953	1974	2271	1977	12	9389
	23,45%	10,15%	21,02%	24,19%	21,06%	0,13%	
Lenézés, undor	1044	298	988	989	575	0	3894
	26,81%	7,65%	25,37%	25,40%	14,77%	0,00%	
Bánat, szomorúság	64	84	33	43	56	1	281
	1,42%	29,89%	11,74%	15,30%	19,93%	0,36%	
Öröm	47	514	69	52	168	15	865
	5,43%	59,42%	7,98%	6,01%	19,42%	1,73%	
Harag	33	30	36	49	20	0	168
	19,64%	17,86%	21,43%	29,17%	11,90%	0,00%	
Szerencsétlenség	14	4	5	0	37	12	72
	19,44%	5,56%	6,94%	0,00%	51,39%	16,67%	
Segítségnyújtás	43	165	82	32	84	5	411
	10,46%	40,15%	19,95%	7,79%	20,44%	1,22%	
Igazságtétel	138	82	170	250	332	1	973
	14,18%	8,43%	17,47%	25,69%	34,12%	0,10%	

5. táblázat. Az emóciótopikok megoszlása pártonként.

A 6. táblázatban láthatjuk a Plutchik-féle emócióknak megfelelő kategóriákat (az eredeti annotációt leképezve ezekre), valamint a pozitív, illetve negatív érzelmi töltetű megnyilatkozások számát, szintén az eredeti annotációból leképezve. Észrevehetjük, hogy míg az ellenzéki pártoknál a negatív töltetű megnyilvánulások vannak egyértelmű fölényben, addig a kormánypártoknál más a helyzet: a Fidesznél ugyan a negatív megnyilatkozások száma magasabb, mint a pozitívaké, ám nem olyan nagy az eltérés köztük, mint az ellenzéki pártok esetében. A KDNP képviselőinek beszédeiben pedig a pozitív megnyilvánulások száma meg is haladja a negatívok számát.

	LMP	KDNP	MSZP	Jobbik	Fidesz
Félelem	133	96	87	133	161
Bánat	2780	968	2408	2011	1249
Harag	298	202	320	501	426
Undor	3246	1251	2962	3260	2552
Siker	2031	2694	1923	2189	3454
Öröm	47	514	69	52	168
Bizalom	181	247	252	282	416
Pozitív	2259	3455	2244	2523	4038
Negatív	6457	2517	5777	5905	4388

6. táblázat. Plutchik-féle emóciók és szentiment a korpuszban pártonként.

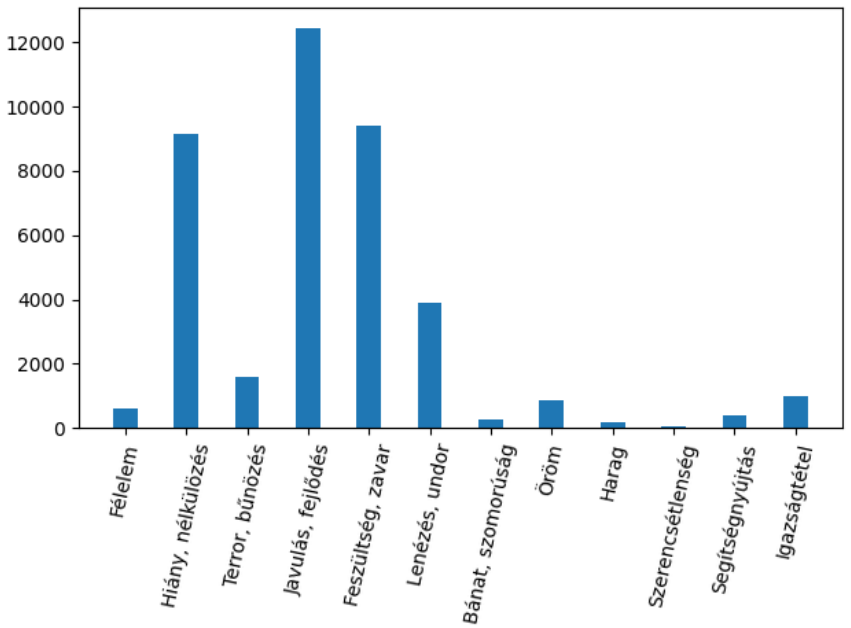
4. A korpusz felhasználhatósága

Ahogyan az 1. ábra is szemlélteti, az alkalmazott kategóriarendszer egyik nagy előnye, hogy „visszafelé kompatibilis" a szentiment- és emócióelemzéshez tradicionálisan alkalmazott, kevesebb osztállyal operáló annotálási rendszerekkel. Ez a gyakorlatban azt jelenti, hogy a választott kutatási kérdésnek megfelelően kiválasztható olyan átváltás az emóciótopikok és más rendszerek kategóriái között, amely után a korpusz tanítóadatként tud funkcionálni változatos problémák megoldásához.

Ilyen (magától értetődő) felhasználási lehetőség érzelemelemzésre alkalmas gépi tanulási modellek létrehozása, amelyre jó példa lehet Üveges és mtsai (2022), amelynek során a szerzők a korpusz egy korábbi változatának felhasználásával kíséreltek meg gépi tanítást a megadott 12 emóciótopikra. Tekintettel arra, hogy a kísérlet lefolytatásakor a korpusz minőségbiztosítása még nem volt befejezettnek tekinthető, az akkor felhasznált adatok csak mintegy a teljes szövegállomány kétharmadát tették ki (608 beszéd mondatai az 1008-ból).

A kísérlet során a huBERT modell (Nemeskey, 2020) finomhangolása történt meg a rendelkezésre álló tanítóadatok alapján aspektusszintű szentimentelemzési (ABSA) feladatra. A gépi tanult modell feladata tehát az volt, hogy a szövegegységben megjelenő érzelem-argumentum párokat kategorizálja. A kapott eredmények alapján az olyan kategóriák esetében, ahol rendelkezésre állt elégséges mennyiségű tanítóadat, a hangolt modell 0,6 - 0,8 körüli[4] F1-értékekkel teljesített. Elégséges mennyiségű adat alatt ez esetben tipikusan néhány ezer annotált mondat - aspektus pár volt értendő egy-egy emóciótopik esetében. Tekintettel az egyes kategóriák erősen kiegyensúlyozatlan eloszlására, ez mindösszesen 4 kategória esetében teljesült a korábbi, megszorítottabb részkorpuszon (Üveges és mtsai, 2022). Ezek az arányok a végleges korpuszban is hasonlóan kiegyensúlyozatlanok maradtak (vö. 5. táblázat). Itt megjegyzendő, hogy pusztán az emóciótopikok felismerése feltehetőleg ennél hatékonyabb lett volna, hiszen az emóciótopik + aspektus azonosítás alapvetően összetettebb feladatnak számít.

[4] Például a 'javulás, fejlődés' kategória esetében az első epoch alkalmával 0,86, a 'szomorúság, bánat' kategória esetén 0,7.



2. ábra: Az egyes kategóriák eloszlása a korpuszban.

Ezen felül természetesen számos további alkalmazás elképzelhető például a jelen korpuszon tanított modellek továbbhasználatával más politikai szövegek elemzése során.

5. Összegzés

Ebben a munkában bemutattuk a HunEmPolit, az első magyar, részletesen annotált, emócióelemzésre használható korpuszt. A korpusz 1008 db parlamenti napirend előtti felszólalást tartalmaz, összesen 764008 token terjedelemben, melyben összesen 39840 emóciót, 61890 kulcskifejezést és 126023 argumentumot azonosítottunk. Ismertettük az alkalmazott kategóriákat, az annotáció és a minőségbiztosítás menetét és alapelveit, emellett alapvető adatokat közöltünk az adatbázisról, valamint bemutattuk, hogyan használható érzelem- vagy szentimentazonosításra is. A korpusz kutatási célokra szabadon elérhető az alábbi linken: https://github.com/poltextlab/HunEmPoli_corpus.

A továbbiakban a korpusz adatainak társadalomtudományi szempontú elemzése, illetve újabb szövegek gépi emócióklasszifikálása mellett terveink között szerepel különféle gépi tanulási kísérletek elvégzése, valamint a huBERT-modell fimomhangolása emócióelemzési feladatokra.

Köszönetnyilvánítás

A publikációban szereplő kutatást a Társadalomtudományi Kutatóközpont és az ELKH-SZTE Mesterséges Intelligencia Kutatócsoport az Európai Unió támogatásával valósította meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. Köszönjük Szabó Martina Katalin segítségét az induktív kategóriarendszer és az annotációs elvek kidolgozásában, valamit a projektben résztvevő annotátorok munkáját is.

Hivatkozások

Aparicio, J.T., de Sequeira, J.S., Costa, C.J.: Emotion analysis of Portuguese Political Parties Communication over the covid-19 Pandemic. In: 2021 16th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–6. IEEE (2021)

Bene, M., Nábelek, F.: A politikai kommunikáció története a külföldi szakirodalomban . In: Kiss, B. (szerk.) A szavakon túl. Politikai kommunikáció Magyarországon, 1990-2015, pp. 11–29. L'Harmattan Kiadó (2019)

Cabanac, M.: What is emotion? Behavioural processes 60(2), 69–83 (2002)

Chapman, C., Nakamura, Y.: A bottom up view of emotion. In: ASSC Seminar http://server. phil. vt. edu/assc/watt/chapman1. html (1998)

Crigler, A.N., Just, M.R.: Measuring affect, emotion and mood in political communication. The Sage handbook of political communication pp. 211–224 (2012)

Ekman, P., Wallace V. Friesen, P.E.: What emotion categories or dimensions can observers judge from facial behavior In: Ekman, P. (szerk.) Emotion in the Human Face. 2nd ed., p. 39–55. Cambridge University Press, Cambridge, UK (1982)

Feng, S., Kang, J.S., Kuznetsova, P., Choi, Y.: Connotation lexicon: A dash of sentiment beneath the surface meaning. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1774–1784 (2013)

Gold, D., Bexte, M., Zesch, T.: Corpus of Aspect-based Sentiment in Political Debates. In: KONVENS (2018)

Gollust, S.E., Nagler, R.H., Fowler, E.F.: The emergence of COVID-19 in the US: a public health and political communication crisis. Journal of health politics, policy and law 45(6), 967–981 (2020)

Griffiths, P.E.: What emotions really are. In: What Emotions Really Are. University of Chicago Press (2008)

Haselmayer, M., Jenny, M.: Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Quality & quantity 51(6), 2623–2646 (2017)

Jafarian, H., Taghavi, A.H., Javaheri, A., Rawassizadeh, R.: Exploiting BERT to improve aspect-based sentiment analysis performance on Persian language. In: 2021 7th International Conference on Web Research (ICWR). pp. 5–8. IEEE (2021)

Jang, H., Shin, H.: Language-specific sentiment analysis in morphologically rich languages. In: Coling 2010: Posters. pp. 498–506 (2010)

Lakoff, G., Kovecses, Z.: The cognitive model of anger inherent. American English-In Cultural Models_in_Language and Thought-Dorothy Holland and Naomi Quinn. eds pp. 195–221 (1987)

Loukachevitch, N., Levchik, A.: Creating a general Russian sentiment lexicon. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1171–1176 (2016)

Marcus, G.E.: Emotions in politics. Annual review of political science 3(1), 221–250 (2000)

Mariani, L.A., Gagete-Miranda, J., Retti, P.: Words can hurt: How political communication can change the pace of an epidemic. Covid Economics 12, 104–137 (2020)

Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013)

Mladenović, M., Mitrović, J., Krstev, C., Vitas, D.: Hybrid sentiment analysis framework for a morphologically rich language. Journal of Intelligent Information Systems 46(3), 599–620 (2016)

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd University (2020)

Plutchik, R.: The emotions: Facts, theories and a new model. American Journal of Psychology 77, 518 (1964)

Plutchik, R.: Emotion, a psychoevolutionary synthesis. Harper and Row, New York (etc.) (1980a)

Plutchik, R.: A general psychoevolutionary theory of emotion. In: Theories of emotion, pp. 3–33. Elsevier (1980b)

Plutchik, R.: A psychoevolutionary theory of emotions. Social Science Information/sur les sciences sociales 21 (4-5), 529–553 (1982)

Richards, B.: The emotional deficit in political communication. Political Communication, 21(3), 339–352 (2004)

Rufai, S.R., Bunce, C.: World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. Journal of public health 42(3), 510–516 (2020)

Russell, J.A.: Core affect and the psychological construction of emotion. Psychological review 110(1), 145 (2003)

Settle, J.: Moving beyond sentiment analysis: Social media and emotions in political communication. The Oxford Handbook of Networked Communication. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190460518.013 20 (2020)

Szabó, G.: Emotional Communication and Participation in Politics. Intersections. East European Journal of Society and Politics 6(2) (2020)

Szabó, M.K., Vincze, V., Ring, O., Üveges, I., Vit, E., Samu, F., Gulyás, A., Galántai, J., Szvetelszky, Zs., Bodor-Eranus, E.H., Takács, K.: StaffTalk: magyar nyelvű spontán beszélgetések korpusza. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)

Szabó, M.K., Vincze, V., Simkó, K.I., Varga, V., Hangya, V.: A Hungarian Sentiment Corpus Manually Annotated at Aspect Level. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2873–2878. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1459

Üveges, I., Vincze, V., Ring, O., Guba, Cs.: Aspect-based emotion analysis of Hungarian parliamentary speeches. In: Ines, R., Gabriella, L., Christopher, K., Simone, P. (szerk.) Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022) Potsdam, Germany. University of Mannheim, University of Stuttgart (2022)

Vincze, V., Üveges, I., Szabó, M.K., Takács, K.: A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)

Wagner, M., Morisi, D.: Anxiety, fear, and political decision making. In: Oxford research encyclopedia of politics. Oxford University Press (2019)

Wang, Y., Croucher, S.M., Pearson, E.: National Leaders' Usage of Twitter in Response to COVID-19: A Sentiment Analysis. Frontiers in Communication 6, 732399 (2021)

Wierzbicka, A.: Emotions across languages and cultures: Diversity and universals. Cambridge university press (1999)