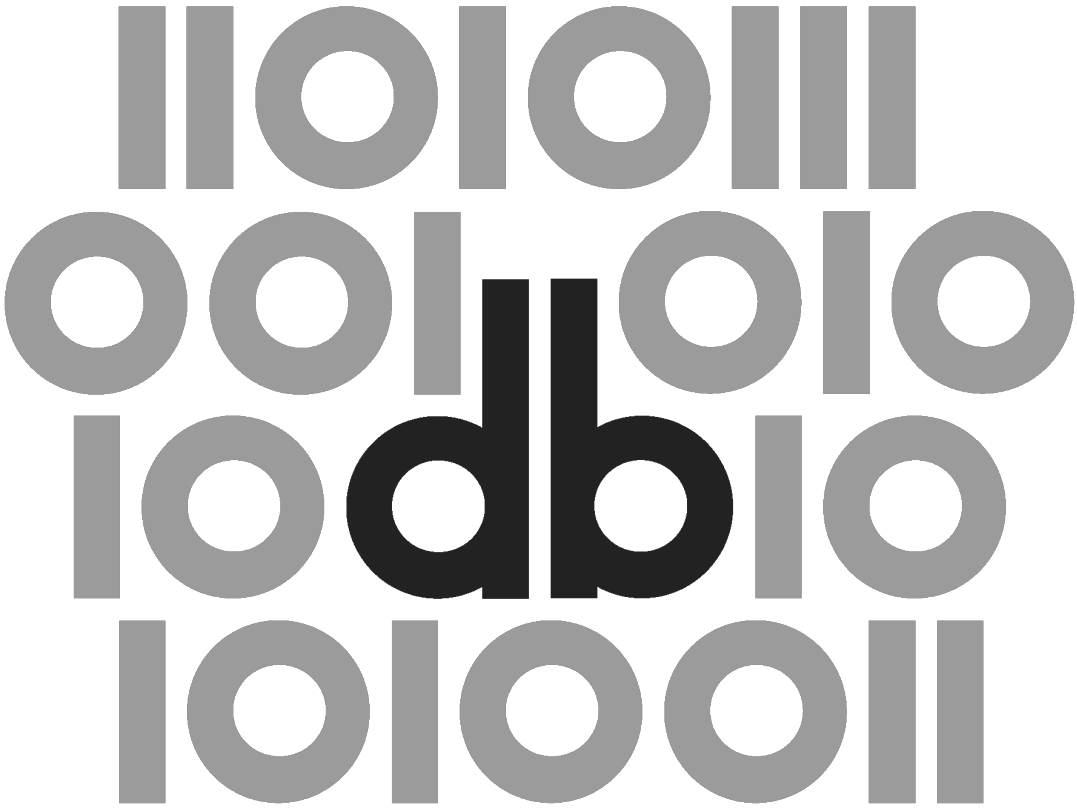
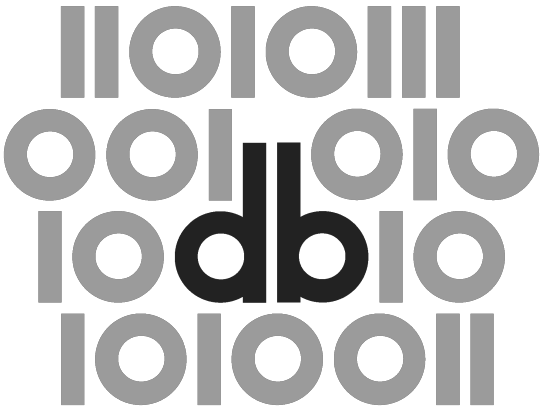


Digitális Bölcsészet
2022., hatodik szám

<DIGITÁLIS BÖLCSÉSZET>



6 (2022)

ISSN 2630-9696
DOI 10.31400/dh-hun.2022.6

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<MŰHELY>

Knap Árpád 0000-0002-4290-6025

ELTE Társadalomtudományi Kar

knap.arpad@tatk.elte.hu

Tóth Tímea Emese 0000-0002-3584-118X

ELTE Társadalomtudományi Kar

toth.emese@tatk.elte.hu

Rakovics Zsófia 0000-0002-9903-9348

ELTE Társadalomtudományi Kar

zsofia.rakovics@tatk.elte.hu

Humán annotált emóciókorpusz létrehozása aktorokhoz köthető érzelmek detektálására

Tanulmányunkban egy olyan kutatási projektet mutatunk be, amelyben egy aktorokhoz (pl. intézményekhez, személyekhez) kapcsolódó, szentimentek és konkrét érzelmek klasszifikációjára képes nyelvi modell létrehozása a célunk. A modell tanítóadatbázisát egy tízezer cikkből álló, online újságokból származó, statisztikai mintavétel segítségével összeállított, humán annotált szövegkorpusz jelenti. Az annotálás során két lépcsőben először az előforduló névelemeket, illetve aktorként funkcionáló közneveket, majd ezt követően a névelemek szövegkörnyezetében megtalálható szentiment- és érzelmi tölteteket annotáljuk. Az annotált szövegek adatbázisa jó bemeneti adatot jelenthet felügyelt klasszifikációs modellek létrehozásához. Cikkünkben ismertetjük a projekt korpuszát, a felügyelt és nem felügyelt szövegklasszifikációs eljárások sajátosságait, valamint a szentiment- és érzelemdetektálás lehetséges módszereit. Ezt követően bemutatjuk a kutatásunkban alkalmazott kétlépcsős annotálási módszertant, az ennek kialakítása során felmerült problémákat és kihívásokat, illetve azokat a kutatói döntéseket, amelyeket a létrehozni kívánt modell társadalomtudományos felhasználhatóságának érdekében hoztunk meg.

Kulcsszavak:
humán annotáció, szentimentdetektálás, érzelemdetektálás, szövegklasszifikáció, felügyelt modellek

1. Bevezetés, a projekt célkitűzései

A politikai és közéleti diskurzusok elemzése során szerzett tapasztalataink[1] azt mutatják, hogy a szövegek affektív, érzelmi töltetének automatizált meghatározása kulcsfontosságú elemzési eszközt jelentene a nagy mennyiségű szöveget értelmezni kívánó kutatók számára. A konkrét entitásokhoz, tehát személyekhez, intézményekhez vagy akár eseményekhez köthető érzelmek klasszifikálása az ilyen módszerek kiterjesztéseként fogható fel, amely segítségével az érzelmeket hordozó szavak tárgya is azonosíthatóvá válik. A jelenleg szabadon elérhető eszközkészlettel egy ilyen jellegű kutatás úgy valósítható meg, hogy névelem-felismerés segítségével azonosítjuk a kérdéses szereplőket, ezt követően a szereplők szövegkontextusát valamilyen módon definiálva leválasztjuk, majd egy szentiment- vagy érzelemszótár segítségével hozzárendeljük az affektív töltetet az egyes szövegrészekhez. Jelenlegi kutatási projektünk elsősorban a szótáras megoldások inherens gyengeségeire reflektál, illetve arra kínál megoldási javaslatot, hogy a nyelvészeti névelemfogalom kiterjesztésével a köznevekkel jelölt szereplők azonosítása is lehetővé váljon az elemzett szövegekben, amely társadalomtudományos célzatú kutatásokban kulcsfontosságú. Kutatásunkban tehát olyan, gold standard korpusz létrehozását tűztük ki célul, amely bemeneti adatként szolgálhat felügyelt nyelvi modellek tanításához. Projektünk végső célja egy klasszifikációs modell létrehozása, amelyhez transzformer-alapú nyelvi modellt tervezünk alkalmazni.

Jelen írásunk lényegében beszámoló a fenti céllal elindított kutatásunkról, amelyben bemutatjuk a projekt célkitűzéseit, módszertanát, a névelemek és az érzelmek kódolása során alkalmazott kétlépcsős annotálási folyamat logikai struktúráját, valamint reflektálunk azokra a megfontolásokra, amelyek a modell társadalomtudományos felhasználásának igényéből fakadnak. Kutatásunk jelenleg is zajlik: a 2023. januári állapot szerint a névelemeket az általunk összeállított, tízezer cikkből álló korpuszban két független kódoló annotálta, az érzelemannotálás pedig hozzávetőlegesen 20 százalékos készültségi szinten áll.

2. Korpusz és adatok

Mivel tehát a projekt céljaként meghatározott modellt elsősorban társadalomtudományos kutatásokban való felhasználásra kívánjuk optimalizálni, ezért nem csupán nyelvészeti szempontokat érvényesítettünk a korpusz kiválasztása és a módszertan összeállítása során. Az annotálandó szövegek kiválasztásakor fontos volt számunkra,

[1] Lásd például Ildikó Barna és Árpád Knap, „Antisemitism in Contemporary Hungary: Exploring Topics of Antisemitism in the Far-Right Media Using Natural Language Processing," *Theo-Web* 18, 1. sz. (2019): 75–92, https://doi.org/10.23770/TW0087; Knap Árpád, Bartha Diána, és Barna Ildikó, „Trianon és a holokauszt emlékezetpolitikai jellegzetességeinek elemzése természetesnyelv-feldolgozás használatával," *Szociológiai Szemle* 31, 4. sz. (2021): 28–62, https://doi.org/10.51624/SzocSzemle.2021.4.2; Ildikó Barna és Árpád Knap, „Analysis of the Thematic Structure and Discursive Framing in Articles about Trianon and the Holocaust in the Online Hungarian Press Using LDA Topic Modelling," *Nationalities Papers,* 2022. május 16., 1–19, https://doi.org/10.1017/nps.2021.67; Kmetty Zoltán és Knap Árpád, „Trágárság mint érzelmi válasz a COVID-19 járvány idején," in Szabó Gabriella, szerk., *Érzelmek és járványpolitizálás: Politikai érzelemmenedzserek és érzelemszabályozási ajánlataik Magyarországon a COVID-19 pandémia idején,* 173–190 (Budapest: ELTE Eötvös Kiadó, 2022).

 DOI: 10.31400/dh-hun.2022.6.4576

hogy a lehetőségekhez mérten időben és tartalmilag is változatos korpusz képezze a projekt alapját, és így a modellhez felhasználandó tanítóadatbázist. A változatosság azért kiemelt fontosságú, hogy a modell többféle stílusú, tematikájú, illetve eltérő időpontokban keletkezett szövegek klasszifikációja során is megfelelő teljesítményt nyújtson. Ezért a Digitális Örökség Nemzeti Laboratórium és az ELTE Research Center for Computational Social Science (ELTE RC2S2) kutatócsoport együttműködésében megvalósuló Webaratás projektben[2] 2021 júniusáig legyűjtött portálok tartalmaiból választottuk ki a korpuszt, rétegzett mintavétel segítségével, az alábbiaknak megfelelően.

A korpuszban négy weboldal, az *Abcúg,* a *Magyar Idők,* a *Válasz.hu* és a *VS* cikkei szerepeltek, összesen 307915 tétel. A mintavételezés során a rétegképző szempontok a megjelenés éve, a cikk rovata és annak forrása voltak, tehát itt is arra törekedtünk, hogy a megjelenés ideje és témája is – a lehetőségekhez mérten – változatos legyen. A megjelenés éve szerint három kategóriára osztottuk a dokumentumokat: 2001–2012, 2013–2016 és 2017–2020. A rovatokat hét féle kategóriára egyszerűsítettük az eredeti rovatok alapján: (1) belpolitika, közélet, vélemény; (2) színes, egészség, életmód; (3) gazdaság; (4) kultúra; (5) külpolitika, külföld vegyes; (6) sport; (7) rovat nélkül. Kizártuk a mintavételezésből azokat a rekordokat, ahol nem szerepelt a megjelenés éve (26 cikk), illetve a *Magyar Idők* és a *Válasz.hu* portálokról azokat a cikkeket, ahol nem volt rovat (9 cikk). A mintában az *Abcúg*ról 94, a *Magyar Idők*ről 5378, a *Válasz.hu*ról 2816, a *VS*-ről pedig 1714 darab cikk származik. A korpuszban szereplő cikkek végleges elemszáma a kerekítések miatt: 10002.

3. Felügyelt és nem felügyelt módszerek

A szöveganalitikában a klasszifikáció kiemelten fontos feladat. A klasszifikációs eljárások között megkülönböztetünk felügyelt *(supervised),* és nem felügyelt *(unsupervised)* algoritmusokat. A felügyelt modellek figyelembe veszik a bemenő adathalmaz metaadatait, amelyek lehetnek például emberek által megadott címkék, kategóriák – összefoglaló néven annotációk. Ebben az esetben az adathalmazt tanító illetve teszthalmazra bontjuk szét. A klasszifikációs modellt az adathalmazunk felcímkézett (tanító) részén hozzuk létre, és a teszthalmazon értékeljük ki. Gépi tanulásra építő, például neurális hálókon alapuló modellek esetében ilyenkor az algoritmus a bizonyos metaadatokkal, címkékkel rendelkező tartalmak nyelvi tulajdonságait, sajátosságait veszi figyelembe a klasszifikáció során. A létrehozott modell teljesítménye pedig elsősorban keresztvalidációval mérhető.[3]

[2] Balázs Indig, et al., „The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata," in *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020),* 33–41 (Marseille: European Language Resources Association, 2020).

[3] Németh Renáta, Katona Eszter Rita, és Kmetty Zoltán, „Az automatizált szövegelemzés perspektívája a társadalomtudományokban," *Szociológiai Szemle* 30, 1. sz. (2020): 44–62, `https://doi.org/10.51624/SzocSzemle.2020.1.`; Renáta Németh and Júlia Koltai, „The Potential of Automated Text Analytics in Social Knowledge Building," in Tamás Rudas and Gábor Péli, eds., *Pathways Between Social Science and Computational Social Science,* Computational Social Sciences, 49–70 (Cham: Springer International Publishing, 2021), `https://doi.org/10.1007/978-3-030-54936-7_3`; R. Sathya and Annamma Abraham, „Comparison of Supervised and Unsupervised Learning Algorithms

A nem felügyelt algoritmusok kizárólag az adathalmazban rejlő látens struktúrákra, mintázatokra hagyatkoznak, előzetesen hozzáadott címkéket nem vesznek figyelembe. Az ilyen nem felügyelt módszerek közé tartozik például a topikmodellezés vagy a szóbeágyazási modellek alkalmazása, amelyek képesek szövegkorpuszok látens tematikus struktúrájának, illetve látens szemantikai struktúrájának feltárására.[4] A nem felügyelt és a felügyelt algoritmusok között átmenetet képeznek a félig felügyelt *(semi-supervised)* módszerek, például a Seeded Latent Dirichlet Allocation topikmodell.[5]

Ebben a projektben, az előzetes tesztelések alapján a vektortérmodellek legújabb generációját, a transzformereken alapuló, úgynevezett kontextualizált vektortérmodelleket tervezzük alkalmazni. A korábbi, statikus vektorterekhez képest (pl. Word2vec, fastText, GloVe), ahol minden szónak egy vektortér-reprezentációja áll elő kontextustól függetlenül, a kontextualizált modellek az adott kontextushoz kötik, hogy milyen vektort kap egy szó. A transzformereken alapuló nyelvi modellek közös jellemzője, hogy nem alkalmaznak statikus beágyazást, illetve rendelkeznek bekódoló *(encoder)* valamint kikódoló *(decoder)* elemmel, amelyek különböző neurális hálókból épülnek fel. A kontextualizált modellek egyik legújabb családját a BERT (Bidirectional Encoder Representations from Transformers[6]) jelenti, amely kiugróan magas teljesítményt mutat a legtöbb természetesnyelv-feldolgozással kapcsolatos feladatban. Ennek egyik oka, hogy a modell a korábbiakhoz képest nem egy meghatározott irányban (pl. balról jobbra) olvassa be a szöveget, hanem a teljes szósorozatot egyszerre dolgozza fel, teljes környezetüket figyelembe véve tanulja meg a szavak kontextusát a, amely különösen a többértelmű kifejezések esetében hasznos. BERT-alapú modell már létezik

for Pattern Classification", *International Journal of Advanced Research in Artificial Intelligence* 2, 2 sz. (2013): 34–38, https://doi.org/10.14569/IJARAI.2013.020206.

[4] Lásd például David M. Blei, Andrew Y. Ng, and Michael I. Jordan, „Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, January (2003): 993–1022; David M. Blei and John D. Lafferty, „Topic Models," in *Text Mining,* 101–124 (Chapman and Hall/CRC, 2009); Tomas Mikolov, et al., „Efficient Estimation of Word Representations in Vector Space," arXiv, 2013. szeptember 6., http://arxiv.org/abs/1301.3781; Armand Joulin, et al., „Bag of Tricks for Efficient Text Classification," arXiv, 2016. augusztus 9., http://arxiv.org/abs/1607.01759; Piotr Bojanowski, et al., „Enriching Word Vectors with Subword Information," arXiv, 2017. június 19., http://arxiv.org/abs/1607.04606; Jeffrey Pennington, Richard Socher, and Christopher Manning, „Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1532–1543 (Doha, Qatar: Association for Computational Linguistics, 2014), https://doi.org/10.3115/v1/D14−1162; Kmetty Zoltán, „Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása," *Statisztikai Szemle* 100, 2. sz. (2022): 105–136, https://doi.org/10.20311/stat2022.2.hu0105.

[5] Bin Lu et al., „Multi-Aspect Sentiment Analysis with Topic Models," in *2011 IEEE 11th International Conference on Data Mining Workshops,* 81–88 (Vancouver, BC: IEEE, 2011), https://doi.org/10.1109/ICDMW.2011.125.

[6] Jacob Devlin, et al., „BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv, 2019. május 24., http://arxiv.org/abs/1810.04805.

 DOI: 10.31400/dh-hun.2022.6.4576

magyar nyelvre is, mint például a huBERT,[7] amelynek a szentimentek és érzelmek klasszifikációját illető használatában már van előzménye.[8]

4. Szentiment- és érzelemdetektálás

A szöveganalitikai kutatások irányainak egyik fő területét a szövegben rejlő affektív, érzelmi tartalmak automatizált felderítésével kapcsolatos munkák jelentik,[9] amelyek két alapvető iránya a szentimentelemzés és az emócióelemzés. A szentimentelemzés jellemzően negatív–semleges–pozitív tengelyen helyezi el a szöveget, míg az emócióelemzés konkrét érzelmeket különböztet meg egymástól (pl. öröm, megvetés, undor[10]). Az iránytól függetlenül, a besorolás végezhető szótáralapon, tehát a szövegben az adott szentimentkategóriával vagy konkrét érzelemmel azonosított szavak és kifejezések megkeresésével. Ez a típusú megközelítés amellett, hogy nagy arányban eredményez hamis találatokat (mivel például egyáltalán nem, vagy rossz megbízhatósággal ismeri fel a tagadást, az iróniát vagy a szarkazmust), nyelvünk agglutináló jellege miatt sem alkalmazható kielégítő teljesítménnyel magyar nyelvű szövegekre. A szótáralapú megoldás, ahogy a nevéből is következik, nagy méretű és megfelelő minőségű szentimentszótár meglétét igényli. Bár léteznek ilyen szótárak a magyar nyelvre is,[11] azonban kutatói tapasztalataink azt mutatják, hogy az általunk vizsgált korpuszok esetében ezek nem adnak megfelelő eredményeket. Az ugyancsak nehezíti a szótáralapú megközelítést, hogy bizonyos szavak jelentése szövegkörnyezettől függően nagyon eltérő lehet (a „balos" szó jelentése például egészen más politikai kontextusban, mint egy nyílászárókkal foglalkozó szakmai fórumon).

A másik megközelítést a szövegek humán (azaz emberek által végzett) annotálása, majd ezt követően felügyelt modellekkel történő klasszifikációja jelenti. Ilyenkor a klasszifikációhoz használt algoritmus, más felügyelt modellekhez hasonlóan, az egyes szentiment- vagy érzelmi kategóriákba tartozó szövegek nyelvi sajátosságait veszi figyelembe. Más nyelveken számos kutatásban sikerrel alkalmazták ezt a megközelítést. Duwairi és Qarqaz arab nyelvű tweeteken és kommenteken alkalmazott sikerrel Naive Bayes, SVM és k-legközelebbi szomszéd eljárásokat.[12] Habernal, Ptáček és Steinberger cseh nyelvű közösségimédia-tartalmakon kísérletezett különböző előfeldolgozási és

[7] Dávid Márk Nemeskey, „Introducing huBERT," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021),* 3–14 (Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2021).

[8] Zoltán Kmetty, et al., „Miniszterelnöki csata az online térben" in Böcskei Balázs and Szabó Andrea, eds, *Az állandóság változása: Parlamenti választás 2022,* 141–161 (Budapest: Gondolat Kiadó, MTA Társadalomtudományi Kutatóközpont Politikatudományi Intézet, 2022).

[9] Bing Liu, *Sentiment Analysis and Opinion Mining* (New York: Springer Cham, 2012), `https://doi.org/10.1007/978-3-031-02145-9`.

[10] Lásd például Alan S. Cowen and Dacher Keltner, „Self-Report Captures 27 Distinct Categories of Emotion Bridged by Continuous Gradients," *Proceedings of the National Academy of Sciences* 114, 38. sz. (2017): E7900–7909, `https://doi.org/10.1073/pnas.1702247114`.

[11] Lásd például Szabó Martina Katalin és Vincze Veronika, „Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai," in *XI. Magyar Számítógépes Nyelvészeti Konferencia,* 219–226 (Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2015).

[12] Rehab M. Duwairi and Islam Qarqaz, „Arabic Sentiment Analysis Using Supervised Classification," in *2014 International Conference on Future Internet of Things and Cloud* (FiCloud), 579–583 (Barcelona: IEEE, 2014), `https://doi.org/10.1109/FiCloud.2014.100`.

klasszifikációs eljárásokkal, azzal a céllal, hogy a tartalmak szentiment töltetét kategorizálják, munkájuk eredményeképpen pedig egy bárki által szabadon hozzáférhető, tízezer Facebook-bejegyzést tartalmazó annotált korpuszt is közzétettek.[13] Shi és Li kínai nyelvű szállodaértékeléseket tartalmazó korpuszon alkalmaztak sikerrel SVM modelleket a szövegek szentimentpolaritásának automatizált értékelésére.[14] Számos olyan kutatás is készült, amelyben az elérhető attribútumszelekciós és gépi tanulási algoritmusok teljesítményét vizsgálják különböző szentimentanalízishez kapcsolódó problémákon.[15]

Az utóbbi néhány évben indultak ilyen típusú, magyar nyelvre irányuló, részben társadalomtudományi vonatkozású kutatások, amelyekből nyilvánosan elérhető eredmények is születtek.[16] Projektünkkel elsődleges célunk ehhez a kutatói munkához csatlakozni, ezen túl pedig általánosságban is vizsgálni a felügyelt klasszifikációs eljárások olyan alkalmazási lehetőségeit, amelyek nem csupán a szentiment- és emócióelemzés feladata során segíthetnek, hanem más, társadalomtudományi szempontból lényeges kutatást is támogatni képesek. Mivel jelenleg még nem létezik szabadon elérhető, mindenki által használható, szentimentek és emóciók klasszifikálására alkalmas, magyar nyelvű szövegekre készített modell, kutatásunkba éppen egy ilyen algoritmus létrehozásának céljával kezdtünk bele.

5. A kétlépcsős annotálási folyamat

A projektben a korpuszban található újságcikkek szövegét kétlépcsős annotálás során látjuk el először névelemcímkékkel, majd a szöveget kisebb egységekre bontva, az annotált névelemek környezetét szentiment- illetve érzelem szempontból értékeljük.

[13] Ivan Habernal, Tomáš Ptáček, and Josef Steinberger, „Supervised Sentiment Analysis in Czech Social Media," *Information Processing & Management* 50, 5. sz. (2014): 693–707, https://doi.org/10.1016/j.ipm.2014.05.001.

[14] Han-Xiao Shi and Xiao-Jun Li, „A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning," in *2011 International Conference on Machine Learning and Cybernetics* (ICMLC), 950–954 (Guilin: IEEE, 2011), https://doi.org/10.1109/ICMLC.2011.6016866.

[15] Lásd például Yang Liu, Jian-Wu Bi, and Zhi-Ping Fan, „Multi-Class Sentiment Classification: The Experimental Comparisons of Feature Selection and Machine Learning Algorithms," *Expert Systems with Applications* 80 (2017. szeptember): 323–339, https://doi.org/10.1016/j.eswa.2017.03.042; Ajay Deshwal and Sudhir Kumar Sharma, „Twitter Sentiment Analysis Using Various Classification Algorithms," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (ICRITO), 251–257 (Noida: IEEE, 2016), https://doi.org/10.1109/ICRITO.2016.7784960; Jeremy Barnes, Lilja Øvrelid, and Erik Velldal, „Sentiment Analysis Is Not Solved! Assessing and Probing Sentiment Classification," *ArXiv:1906.05887 [Cs],* 2019. június 13., http://arxiv.org/abs/1906.05887.

[16] Ilyen a szabadon hozzáférhető OpinHuBank, amelyben bizonyos entitásokat tartalmazó mondatok, illetve az egyes mondatokra vonatkozó szentimenttöltetek annotációja szerepel (Miháltz Márton. *OpinHuBank: Szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez* [Szeged: MTA Nyelvtudományi Intézet, 2013]). Kutatásunkban az OpinHuBank adatbázisához hasonló módszertant követünk, azzal a kiegészítéssel, hogy mi konkrét érzelmeket is annotálunk a szentimentértékek mellett. Szintén megemlítendő a HunEmPoli korpusz, amelyben emóciókategóriákat annotáltak egy speciális nyelvezettel rendelkező, parlamenti beszédeket tartalmazó korpuszon (Ring Orsolya, et al., „HunEmPoli: Magyar nyelvű, részletesen annotált emóciókorpusz," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia,* 187–201 (Szeged: Szegedi Tudományegyetem, 2023).

 DOI: 10.31400/dh-hun.2022.6.4576

Az annotálás mindkét fázisában két független kódoló munkája alapján kódoljuk a szövegeket, a kérdéses esetekben pedig supervisor annotátor dönt a helyes kódolásról.[17] Az aktorokhoz kötődő szentiment- és érzelemdetektálásra alkalmazott nyelvi modell létrehozásához készülő szövegkorpusz megalkotása több ponton is kihívások elé állított minket. A kihívások egy része elméleti, másik része technikai jellegű volt. Fontosnak tartjuk, hogy az alábbiakban reflektájunk az ilyen jellegű tapasztalatainkra annak érdekében, hogy a hasonló projekteken dolgozó kutatói közösség munkájához hozzájárulhassunk.

5.1. Névelemek, aktorok annotálása

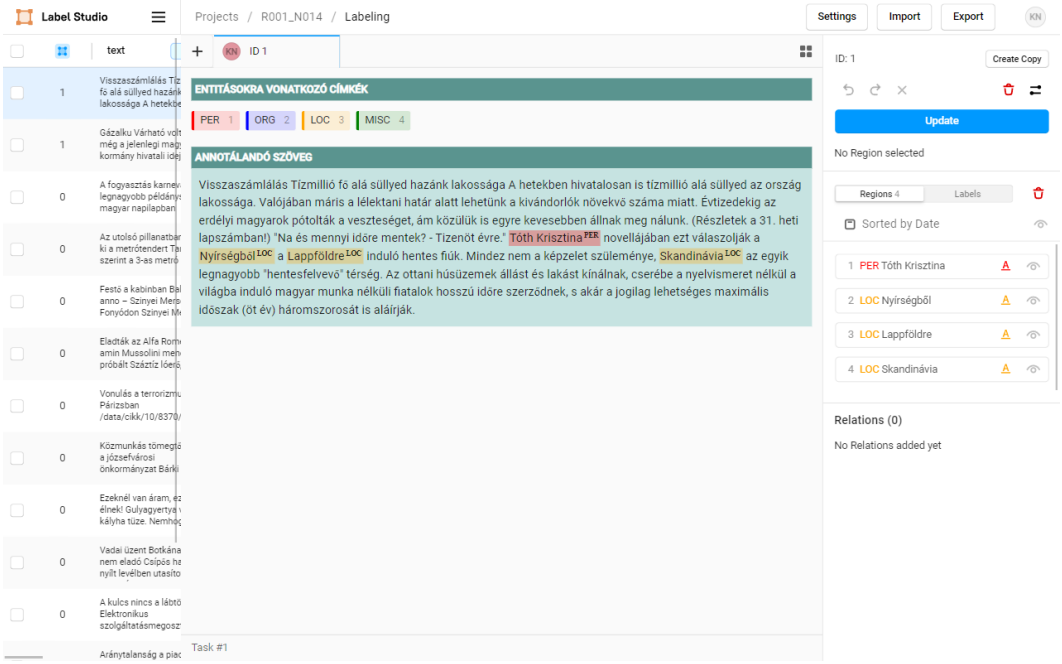
A névelemek annotálását a Simon Eszter és Vadász Noémi által összeállított NerKor annotálási útmutató[18] jelen korpuszra és feladatra kiegészített, példákkal ellátott változata alapján végezzük a *Label Studio*[19] nevű célszoftverben. A névelemek annotálása során elsősorban tulajdonneveket jelölünk a szövegekben, és az alábbi típusokat különböztetjük meg egymástól.

- PER (személynevek): valós és kitalált személyek nevei, becenevek, művésznevek, álnevek.
- ORG (szervezetnevek): intézmények, vállalatok, kormányzati hivatalok, sportcsapatok, múzeumok, egyetemek, szervezett struktúrával rendelkező szervezetek, üzletek stb. nevei.
- LOC (helynevek): országok, városok, földrészek, hegyek, folyók és tavak, tengerek, óceánok, ember alkotta építmények, például repterek, utak, gyárak nevei. A helyek egyaránt lehetnek földrajzilag vagy politikailag definiáltak.
- MISC (egyéb nevek): a fenti csoportok egyikébe sem tartozó nevek, például könyvek és festmények címei, kiállítások, konferenciák, újságok, online hírportálok és médiumok, márkák, televízió- és rádióállomások, ünnepek, programozási nyelvek, kereskedelmi útvonalak, járműmodellek nevei.

[17] Fontosnak tartjuk megjegyezni, hogy a létező automatizált eszközök alkalmazásával szemben azért döntöttünk a névelemek manuális annotálása mellett, mert úgy találtuk, hogy az automatizált megoldások nem teljesítenek megfelelően a projekt célkitűzéseihez, illetve egy *gold standard* annotált névelemkorpuszt is létre kívántunk hozni a kutatásunk során. Ezen kívül, ahogyan azt a tanulmány további részében részletesen ismertetjük, nem csupán tulajdonneveket, hanem társadalomtudományos elemzésekhez rendkívül fontos, a szövegekben aktorokként funkcionáló közneveket is annotálunk, amely a „hagyományos", nyelvészeti névelemértelmezés kiterjesztéseként értelmezhető.

[18] Eszter Simon and Noémi Vadász, „Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus," in Kamil Ekstein, Frantisek Pártl, and Miloslav Konopík, eds., *Text, Speech, and Dialogue – 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings,* Lecture Notes in Computer Science, 222–234 (New York: Springer Cham, 2021), `https://doi.org/10.1007/978-3-030-83527-9_19`.

[19] Maxim Tkachenko et al., „Label Studio: Data Labeling Software," 2020, hozzáférés 2023.03.04, `https://github.com/heartexlabs/label-studio`.



1. ábra. A *Label Studio* szoftverben kialakított, névelem-annotálásra használt felület

A névelemek annotálását minden esetben két, egymástól független kódoló végzi. Ez a gyakorlatban úgy történik, hogy a 10000 cikket tartalmazó korpuszt véletlen sorrendbe rendeztük, ezt követően 50 cikkből álló szeletekre, „pakkokra” bontottuk fel, majd ezeket a pakkokat adjuk ki annotálni a kollegáknak, akik a saját gépükön futó (tehát mások által nem elérhető), nyílt forráskódú *Label Studio* programban végzik az annotálást, majd eredményeiket egy meghatározott felhőtárhelyre töltik fel, ahol kizárólag az adott kódoló és a kutatás vezetői érik el azt. Ilyen módon tehát az annotálást végzők nem látnak rá egymás munkájára, ami a kódolások esetleges torzulását okozhatná.

A két független annotálás eredményét ezt követően a kutatás vezetői automatizált eljárásokkal hasonlítják össze. Ahol nincsen eltérés, tehát a két kódoló ugyanazon szövegrészletet annotálta, és ugyanazt a címkét használta a fentiek közül, azt az annotálást elfogadjuk. Ahol a szöveghatárokban vagy az alkalmazott címkékben eltérés mutatkozik, ott egy supervisor annotátor dönti el, hogy melyik a megfelelő a két verzió közül, illetve ahol egyik sem, ott harmadik, javított annotálást jegyez fel.

Fontos megemlíteni, hogy ezzel a folyamattal párhuzamosan egy módszertani kísérletet is végzünk, amelynek során a kódolóink egy részét arra kértük, hogy ne csupán tulajdonneveket, hanem olyan közneveket is jelöljenek a szövegekben, amelyek aktorokként funkcionálhatnak, és érzelmek kapcsolódhatnak hozzájuk. Az annotált köznevek eltérő címkéket (K_PER, K_ORG, K_LOC, K_MISC) kapnak a tulajdonnevekhez képest, hogy könnyen megkülönböztethetők legyenek a továbbiakban. Ez a gyakorlatban olyan köznevek annotálását jelenti, mint például: főpolgármester, miniszterelnök, államtitkár, szövetségi kapitány, rektor stb. Azt mondhatjuk tehát, hogy projektünkben tartalmi tekintetben általánosabban, módszertanilag specifikusabban szeretnénk az aktorokat megtalálni és hozzájuk érzelmeket társítani. Például azt szeretnénk elérni,

 DOI: 10.31400/dh-hun.2022.6.4576

hogy az általánosabb „józsefvárosi önkormányzat" kifejezés is felismerhető legyen, valamint társíthassunk hozzá szentimentértékeket és érzelmeket, annak ellenére, hogy az entitás hivatalos elnevezése „VIII. kerület Józsefvárosi Önkormányzat Polgármesteri Hivatala". Ezzel a tevékenységgel az a célunk, hogy egy kellően nagy számú, elsősorban politikai és közéleti diskurzusok elemzésekor használható köznévlistát állítsunk össze, amelynek segítségével tehát nem csupán a tulajdonnevekhez köthető, hanem az aktorokként funkcionáló köznevekhez kapcsolódó szentimenteket és érzelmeket is azonosítani tudjuk majd a modell segítségével. A szigorúan vett nyelvészeti névelemfogalommal szemben a társadalomtudományos elemzésekben ugyanis a köznevekkel jelölt, aktorként funkcionáló entitások felismerése is kulcsfontosságú, ezért szükséges egy lehetőség szerint minél bővebb köznévlistát létrehozni. A folyamat során ilyen jellegű példák nyomán elengedhetetlen volt egy olyan szabályrendszer[20] megalkotása, amely alapján pontosan eldönthető, kell-e jelölni az adott köznevet vagy sem. Ennek értelmében az a köznév annotálandó, amely egyértelműen utal valamely, az adott szövegben szereplő entitásként azonosított tulajdonnévre; illetve globálisan felismerhetőnek bizonyul és valamilyen politikai, közéleti entitást jelöl. Emellett rendkívül fontosnak tartottuk, hogy a politikai szférában előforduló, különböző politikai oldalakhoz köthető kifejezések is annotálva legyenek, így a „szocialisták", „fasiszták", illetve azok, amelyek valamilyen médiummal állnak együtt, például a „ballib sajtó". Ez utóbbi kifejezések kulcsfontosságúak olyan társadalomtudományos elemzések esetében, amelynek során például bizonyos politikai oldalak narratíváját vizsgáljuk.

5.2. Érzelem- és szentimentannotálás

A névelemek annotálását követően minden kódolt névelem esetében automatizált eljárás segítségével kiválasztjuk az adott névelem, aktor szövegkontextusát. Az erre vonatkozó előzetes vizsgálataink alapján úgy találtuk, hogy leggyakrabban az adott névelemre a +/- 1 mondatos környezetben vonatkoznak információk, érzelmek, ezért a szövegkontextust úgy definiáltuk, hogy az adott névelem előtti mondatot, azt, amelyben a névelem szerepel, illetve az azt követőt emeljük ki a teljes szövegből.[21] Ezzel a folyamattal együtt vizuálisan megjelöljük az aktuálisan annotált névelemet a szövegben, hogy az érzelemannotálás során akkor is egyértelmű legyen, melyik aktor szempontjából kell az annotálást elvégezni, ha a szövegrészletben több névelem is szerepel. Eddigi annotálásaink során úgy találtuk, hogy egy 50 cikkből álló szövegben

[20] Ahogy említettük, csak az annotálást végzők egy részét képeztük ki a köznevek annotálására. Mivel a köznevek annotálása módszertani kísérlet volt, egyúttal törekedtünk arra, hogy olyan útmutatót írjunk ezek annotálásához, amely kellőképp egyértelmű, ugyanakkor ne járjon túl nagy kapacitással egy kétséges alkalmazhatóságú útmutató értelmezése.
A köznevek annotálásának kiértékelését, illetve az annotálást végzők közti egyetértési arányok kiszámítását a köznevek annotálási folyamatának lezárultával végezzük el.

[21] A projekt tervezése során felmerült a függőségi elemzés alkalmazása az aktorok szövegkontextusának kinyerése érdekében, azonban úgy találtuk, hogy az általunk feldolgozott korpusz esetében a +/- 1 mondatos környezet megbízhatóan működik, és a függőségi elemzés implementálása az alkalmazott munkafolyamatba nagyobb költséggel járna, mint amekkora hasznot jelentene az adatok minősége szempontjából.

általában 200 és 900 között mozog az annotált névelemek száma, tehát ennek megfelelő mennyiségű, érzelmi szempontból annotálandó szövegkontextust generálunk minden egyes „pakkból".

Az annotálandó adathalmaz előállítása után a szentimenttöltetek és érzelmek bekódolása következik. Ehhez szintén a már korábban említett *Label Studio* programot használjuk, és ahogyan az előző fázisnál, ebben az esetben is két, egymástól független kódoló végzi minden egyes szövegrész annotálását. Az annotálási munkamenet az alábbiaknak megfelelően zajlik.

- A kódolók eldöntik, hogy az adott szövegrészt kódoljuk-e. Nincsen szükség a szövegek kódolására olyan esetekben, ahol például az adatok esetleges hibás legyűjtése miatt programkódot tartalmazó szövegrészletről van szó. Szintén nem szükséges a szöveg annotálása, ha például fotós nevével, egy fotó keletkezési helyével vagy helyszínével, vagy akár nevek hosszas felsorolásával (pl. futballcsapatok névsorával) találkoznak. Ezek az elemek, bár megfelelnek a névelemek annotálási kritériumainak, tartalmilag nem járulnak hozzá a szöveghez.
- Ezt követően a kódolók megállapítják a szövegről, hogy van-e benne érzelem, vagy nincs. Ha nincsen, akkor a „nincs érzelem" kódot alkalmazzák, ekkor az adott szövegrésszel nincsen további teendőjük. Fontos továbbá, hogy nem jelölnek az annotátorok érzelmeket arra a személyre vonatkozóan, aki valamilyen semleges tevékenységet folytat, például nyilatkozik. A „nincs érzelem" kód alkalmazandó olyan esetekben is, amikor például a cikk címében hiányos, tehát érzelmi szempontból nem bekategorizálható kontextusban említenek egy adott entitást.[22] Fontos kitételként szerepel az is, hogy a birtokos mondatszerkezetben szereplő birtokoshoz nem rendelhető érzelem, kizárólag a birtokhoz, amennyiben van rá utalás a szövegben.
- A szövegben jelölhető érzelem vonatkozásában negatív, semleges és pozitív szentimentkategóriák közül választhatnak a kódolók. Itt természetesen egynél több kategória is hozzárendelhető ugyanazon szöveghez, amennyiben az egyszerre tartalmaz negatív és pozitív szentimenttöltetet.
- Ezt követően a kódoló azt is bejelöli, hogy a szövegben van-e irónia. Az iróniadetektálás az aktorokként azonosítható köznevek kódolásához hasonlatos módszertani kísérlet, tehát nem központi elem a projektben. Az iróniát gyakran két ember is teljesen máshogyan értelmezi, és lényegében nincsen olyan nyelvi meghatározottsága, nyelvi jegye, amely alapján nyelvi modell segítségével azonosítható lenne. Mivel azonban az ironikus tartalmak megjelölése nem okoz jelentős többletráfordítást az annotálási folyamat során, izgalmas kísérletnek gondoltuk megpróbálkozni egy erre is alkalmas modell létrehozásával.
- A kódoló ezután annak megfelelően, hogy a szöveget a negatív, semleges vagy pozitív szentimentkategóriába sorolta, a bejelölt szentimenteknek megfelelő,

[22] Például a *Vadai üzent Botkának, a DK nem eladó* cím esetében nem egyértelműen eldönthető az író szándéka szerinti érzelemtöltet, ilyen módon pedig az olvasóból kiváltott érzelem kerülne annotálásra.

 DOI: 10.31400/dh-hun.2022.6.4576

konkrét érzelmeket tartalmazó listá(ko)n bejelöli, hogy mely érzelmek vannak jelen a szövegben. A konkrét érzelmek listáját[23] az alábbi táblázat tartalmazza.

1. táblázat. Az egyes szentimentekhez tartozó konkrét érzelmek[24]

Negatív	Semleges	Pozitív
bánat	együttérzés, szimpátia	elégedettség, öröm (elragadtatás, csodálat, rajongás, szórakozás)[25]
düh	érdeklődés, érdekesség	reménykedés, bizakodás, vágyakozás
elégedetlenség[26]	nosztalgia	
félelem, rémület, szorongás	meglepődés (szokatlanság, furcsaság)[27]	
gúnyolódás, undor, megvetés		
irigység, féltékenység		
zavartság, értetlenkedés (kellemetlenség)		

[23] Az érzelmek listájának kialakításához felhasználtuk egy korábbi projekt tapasztalatait lásd Kmetty, et al., „Miniszterelnöki csata az online térben".

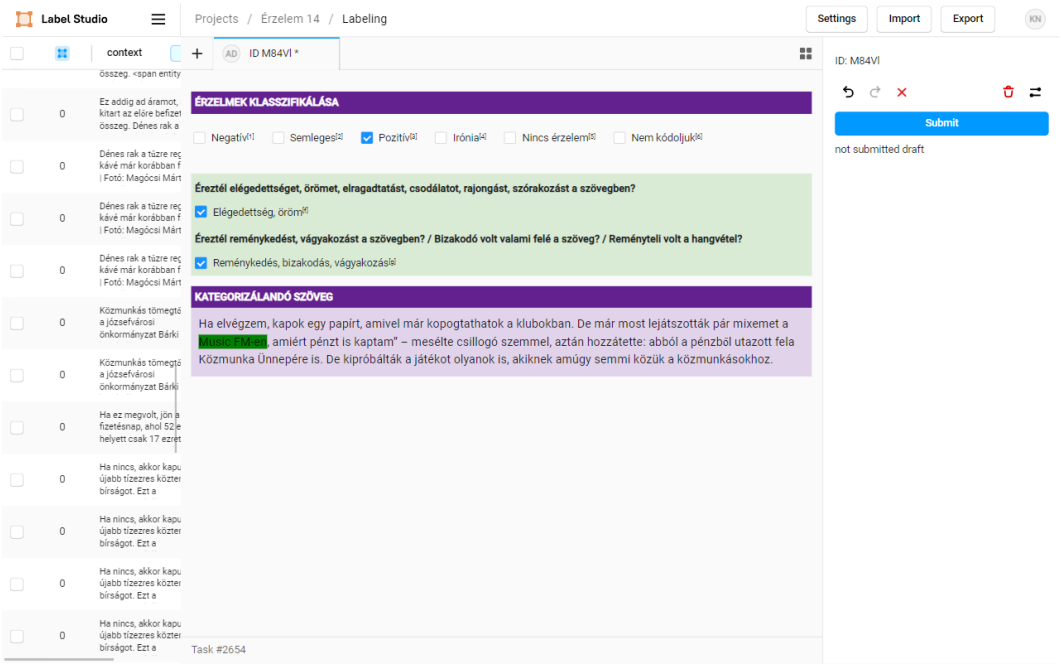
[24] Korábbi annotálási tapasztalataink alapján, amelyet magyar nyelvű internetes kommentek korpuszán végeztünk, a nemzetközileg bevett emóciókategóriák csak részben használhatók a magyar nyelvű szövegekre, a szövegek eltérő érzelmi töltete és érzelmi eloszlása miatt. Ezért ebben a projektben a korábban hivatkozott Cowen és Keltner-féle érzelmi kategóriarendszer egy módosított verzióját használjuk. Az annotálást végzők közti egyetértési arányokat az annotálási folyamat lezárultát követően tervezzük kiértékelni.
A táblázat néhány kategóriájához lábjegyzetben példamondatokat mellékeltünk.

[25] Például: „Jó erőben vagyok, mostanában minden sikerül. Ilyen egyszerű lenne? A Ferencváros már a második, megközelítette a Vasas ellen a Szusza-stadionban összeroppanó éllovas Videotont, a nemrég még sereghajtó Herczeg Andrásnak köszönhetően újra saját nevelésű játékosaira építő Debrecen*MISC* pedig sorozatban negyedik győzelmével feljött a dobogó közelébe." Ch. Gáll András, „Két hét szünet felélesztette a Fradit: Remekelt a bajnok otthonában Thomas Doll csapata, már második a tabellán," *Magyar Idők,* 2017. szept. 18., `https://www.magyaridok.hu/sport/ket--het-szunet-felelesztette-fradit-2224843/`.

[26] Például: „Tulajdonképpen a Varga Roland–Bobál Dávid-párharc döntötte el a mérkőzést, a kilenc forduló alatt tíz gólnál járó válogatott szélső kétszer is szemfülesen megelőzte a halvérű, nem a saját posztján játszó védőt, a harmadik gólt éppen a sokat szidott Botka *PER* beadásából a villámgyors Paintsil szerezte, immár a szakadó esőben, villámlás közepette." Uo.

[27] Például: „Bencsik András*PER* olyat mondott, ami két napja még elképzelhetetlen lett volna" – Vörös Szabolcs, „Bencsik András olyat mondott, ami két napja még elképzelhetetlen lett volna," *Válasz.hu,* 2018. febr. 27., `https://web.archive.org/web/20200126092543/http:/valasz.hu/itthon/bencsik-andras-olyat-mondott-ami-ket-napja-meg-elkepzelhetetlen--lett-volna-127585`.

Az annotálás során tehát hétféle negatív, négy semleges és két pozitív érzelmet különítünk el egymástól. Az érzelmi kategóriák kialakítása során figyelembe vettük korábbi tapasztalatainkat, ami a magyar nyelvű interneten elérhető szövegek – elsősorban hozzászólások – érzelmi megoszlását illeti. Ezek alapján túlnyomó a negatív érzelmek többsége a másik két kategóriához képest. Mivel a modell tanításához minden érzelemből megfelelő mennyiségű annotált szövegre van szükség, ezért bizonyos érzelmek összevonása, azokból nagyobb kategóriák létrehozása mellett döntöttünk, ahogyan az a táblázatban is látható.



2. ábra. A *Label Studio* szoftverben kialakított, szentiment- és érzelemannotálásra használt felület

Nem mehetünk el amellett, hogy az érzelmek annotálásának kihívásairól, nehézségeiről is említést tegyünk. A pszichológiában is alapvető problémakörként említik az érzelmek kategorizálhatóságát, melynek két pillére a szisztematikus kategorizáció nehézsége és az érzelmek univerzalizmusának és kultúrafüggőségének kérdésköre.[28] A pszichológia továbbá megkülönböztet alapérzelmeket és komplex érzelmeket is. Az alapérzelmek univerzálisnak tekinthetők, általában a hozzájuk társított arckifejezések okán, a komplex érzelmek észlelését és összetettségét viszont többek között a kulturális környezet és az egyén is képes befolyásolni. Lindquist (2008) szerint az érzelmek komplexitását növelheti az is, hogy milyen fogalmi kategóriákkal rendelkezik az egyén egy adott érzelem kapcsán, és hogy számára ez milyen összetevőket tartalmaz, tehát mit tud mondjuk a szorongás, vagy a félelem jelenségéről, illetve, milyen formában em-

[28] Hámori Ágnes, „Az érzelmek elemzési lehetőségei a kognitív poétikai kutatásban és korpuszfeldolgozásban," in *Nyelv, poétika, kogníció: Elmélet és módszer a poétikai kutatásban,* 139–173 (Eger: Eszterházy Károly Egyetem Líceum Kiadó, 2018).

 DOI: 10.31400/dh-hun.2022.6.4576

lékszik ezekre, hogyan használja őket.[29] Ennek okán az érzelmek felismerése egyéntől függően még az alapérzelmek vonatkozásában is nagy variabilitást mutathat, azonban annotáláskor törekednünk kell ennek standardizálására. Külön nehézséget okozhat továbbá, hogy megállapítsuk és elválasszuk az olvasás közben ránk törő érzéseket a szöveg írója által közölni kívánt érzelmektől. Ezért volt szükséges standardizálni egyrészt azt az érzelmet, amit az adott annotátornak jelölnie kell a szövegben, másrészt a csapatban dolgozók munkáját is.

Megoldásképp készítettünk egy útmutatót, amelyben rögzítettük, hogy az érzelmi töltet meghatározása a cikk írójának, szerzőjének szándéka szerint detektálandó, a legszigorúbb mondatkörnyezet alapján, vagyis az annotálás a közvetlen mondatkörnyezeti utalások figyelembevételével kell hogy történjen, nem pedig egyéni értelmezés szerint. A „cáfolata annak a balos maszatolásnak, mely szerint a Fidesz összekacsint a Magyar Gárdával" esetén a „balos maszatolás" kifejezés egyértelmű Fidesz-szimpátiát árul el, azonban arról nincs információnk, hogy a cikk szerzője a Magyar Gárdát hogyan ítéli meg, így arra a névelemre a „Nincs érzelem" címke kerül. Speciális esetnek tekinthető, amikor idéznek valakit a szövegben. Ekkor az idézett, a nyilatkozó érzelmei a mérvadóak az adott entitásra vonatkozó érzelem meghatározásához.

Végezetül fontos kiemelni, hogy a projektben elsődlegesnek tekintjük az adatok jó minőségű bekódolását, ezért számos minőségbiztosítási megfontolást is figyelembe vettünk a munkafolyamat kidolgozásakor. Ennek érdekében online oktató videókat, illetve részletes, példákkal illusztrált útmutatót készítettünk mindkét annotálási fázishoz. Szintén a minőségbiztosításhoz tartozik, hogy az annotáláson dolgozó gyakornokok bármikor elérhetik a kutatás vezetőit kérdéseikkel, a specifikus, de mindenki számára hasznos tudást közvetítő problémákat pedig megosztjuk egy közösen elérhető felületen. A folyamatok átláthatósága, követhetősége érdekében kanban rendszerű projektmenedzsment szoftvert, az adatok biztonsága érdekében felhőtárhelyet, a kommunikációhoz pedig könnyen visszakereshető munkahelyi csetplatformot használunk.

6. Összefoglalás

Ahogyan korábban már említettük, a projekthez jelenleg használt korpusz tízezer, változatos időszakokban keletkezett és eltérő témákban íródott cikkből áll. Az annotálási folyamat végén egy olyan humán annotált szövegkorpusszal fogunk rendelkezni, amely becslésünk szerint legalább 80000darab,[30] érzelmi töltetre vonatkozóan bekódolt szövegrészletet tartalmaz majd. Ezen a ponton még kérdéses, hogy ebben a 80000 szövegben pontosan milyen képet mutat majd az egyes érzelmek megoszlása. Amennyiben nagyon kedvezőtlen módon, például szélsőségesen magas lesz az érzelmeket nem tartalmazó, vagy semleges szövegek aránya, szükség lehet az annotált korpusz további

[29] Kristen A. Lindquist and Lisa Feldman Barrett, „Emotional Complexity," in Lisa Feldman Barrett, Michael Lewis, and Jeannette M. Haviland-Jones, eds., *Handbook of Emotions,* 513–530 (New York, London: Guilford Publications, 2008).

[30] Ez jelenleg természetesen csak egy becslés, amelyet a következőképpen kalkuláltunk: a jelenlegi korpusz 10000 darab cikket tartalmaz, amely 200 pakkra oszlik. Úgy találtuk, hogy egy pakkban átlagosan kb. 400 darab annotált entitás szerepel, eszerint a teljes korpuszban kb. 80000 darab entitás jelenik meg.

szövegekkel való kiegészítésére. Amennyiben erre lesz szükség, a további szövegeket olyan módon fogjuk kiválasztani, hogy egy előzetes modell segítségével azokhoz a dokumentumokhoz rendelünk majd nagyobb súlyt a mintavétel során, amelyek a modell szerint nagyobb eséllyel tartalmazzák az előzőleg nem kellő számosságban annotált érzelmi kategóriákat.

Az annotált szövegkorpusz segítségével létrehozott nyelvi modell szándékunk szerint tehát képes lesz magyar nyelvű szövegekben szentiment- és érzelmi tölteteket azonosítani. Különösen érdekesnek tartjuk olyan szentimentanalízis alkalmazását, amely névelem-felismeréssel összekövetve alkalmas a szövegben szereplő különböző aktorokhoz (pl. közéleti szereplők vagy történelmi személyiségek), eseményekhez kötődő érzelmi töltetek detektálására, tehát olyan elemzésre, amelynek során az érzelem tárgyát is lehetséges azonosítani *(entity-level sentiment analysis)*[31]. Az érzelmek detektálásának társadalmi relevanciájához, véleményünk szerint, nem fér kétség: az nemcsak politikai szövegek tartalomelemzése során, hanem a laikus diskurzusban megjelenő vélemények felderítéséhez is elengedhetetlen fontosságú eszközt jelent.

Jelen tevékenységünk célja tehát három részre osztható. Elsődlegesen cél az ismertetett nyelvi modell létrehozása, majd e modell teljesítményének lemérése, végül a projekt tapasztalatainak összegzése. A projekt jellegéből adódó célunk továbbá, hogy az említett módszereket, illetve a kutatás során gyűjtött módszertani tapasztalatainkat más kutatók számára is megismerhetőbbé, hozzáférhetőbbé tegyük, gyakorlati példákkal alátámasztva mindezeket. Nem lenne teljesíthető azonban ez a célkitűzés a létrehozott modell tartalmi fókuszú kutatásokban való, releváns elemzési eszközként való alkalmazása nélkül, amely azonban már túlmutat a jelenlegi projekt szűkebb célkitűzésein.

Köszönetnyilvánítás

Ezúton szeretnénk megköszönni a projektben dolgozó ELTE TáTK Szociológia alapszakos hallgatók munkáját, akik nélkül kutatásunk nem valósulhatna meg, illetve Dömötör Andreának az érzelemannotálási útmutató megírásában nyújtott segítségét.

Creating a Human Annotated Emotion Corpus for the Detection of Actor-related Emotions

In our study, we present an ongoing research project in which our goal is to create a language model capable of classifying sentiments and specific emotions related to actors (e.g., institutions, persons). The training database of the model is a human-annotated text corpus consisting of ten thousand articles from online newspapers, compiled using statistical sampling methods. In the project, we employ a two-phase annotation design. First, we annotate named entities and common names that function as

[31] Lásd például: Jin Ding et al., „Entity-Level Sentiment Analysis of Issue Comments," in *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering* (ICSE '18), 7–13 (Gothenburg: ACM, 2018), https://doi.org/10.1145/3194932.3194935.

 DOI: 10.31400/dh-hun.2022.6.4576

actors. Second, we annotate sentiments and specific emotions found in the context of the previously marked actors. Such a database of annotated texts can provide excellent input for creating supervised classification models. In this article, we describe the corpus of the project, the characteristics of supervised and unsupervised text classification procedures, and possible methods for sentiment and emotion detection. After that, we present the two-phase annotation methodology used in our research, the problems and challenges that arose during its development, as well as the research decisions that we made to create a model that can be used as a capable research tool in social sciences.

Keywords:
human annotation, sentiment detection, emotion detection, text classification, supervised models