Societas et Oeconomia

Az automatizált szövegfeldolgozás szociológiai lehetőségei

Németh Renáta

Savaria University Press

AZ AUTOMATIZÁLT SZÖVEGFELDOLGOZÁS SZOCIOLÓGIAI LEHETŐSÉGEI

Societas et Oeconomia
Sorozatszerkesztő: Kovács László

AZ AUTOMATIZÁLT SZÖVEGFELDOLGOZÁS SZOCIOLÓGIAI LEHETŐSÉGEI

NÉMETH RENÁTA

Savaria University Press

Szombathely

2024

Szakmai lektor:

Tóth Gergely
Károli Gáspár Református Egyetem
Bölcsészet- és Társadalomtudományi Kar

ISBN 978-615-6489-35-7

ISSN 2631-133X

A címlapon: DeepAI generálta kép, reneszánsz festmény stílus, prompt: „Oxford college library interior with arches, a lonely robot is sitting at a desk with a huge pile of books in front of him”

Kiadja: Savaria University Press
Szombathely

Savaria University Press

„A nyelv […] irányítja az érzéseimet, kormányozza az egész szellemi lényemet; annál inkább, minél öntudatlanabbul hagyatkozom rá.”

Viktor Klemperer

KÖSZÖNETNYILVÁNÍTÁS

Ez a kötet nem jött volna létre az Eötvös Loránd Tudományegyetem Társadalomtudományi Karán működő Research Center for Computational Social Science (rc2s2.elte.hu) és tagjai nélkül. A központot 2018-ban alapítottuk Intézetünkben, Barna Ildikó és jómagam társvezetésével, azzal a céllal, hogy az akkor már a társadalomtudományokba is begyűrűző adatforradalomra reagáljunk. Kvantitatív társadalomkutatási szakemberként izgalmas kihívásként tekintettünk a számítógépes társadalomtudomány új módszereire, elsősorban a természetes nyelvfeldolgozásra, melyeket elsőként adattudománnyal foglalkozó cégeknél gyakornokoskodó tanítványaink ismertettek meg minket (ez az üzleti életből a szociológia felé történő mozgás addig szokatlan volt, de a világ más részein is hasonló módon zajlott). A kutatócsoport ennek a tudományterületnek a tartalmi szociológiai tudás felfedezésében történő kihasználását és az új módszerek szociológiai adaptálását tűzte ki célul. Ebben a vállalkozásban az intézetben dolgozó, tehát módszertani irányultságú kollégák (Buda Jakab Katona Eszter, Knap Árpád, Máté Fanni, Rakovics Márton, Rakovics Zsófia, Tóth Emese) mellett Sik Domonkos is mellettünk volt a kezdetektől, aki az elméletet mindig inspiráló módon tudta az empíriával összekötni. Csatlakoztak hozzánk az új módszereket saját területükön is hasznosnak látó vendégkutatók (Balogh Péter, Pólya Tibor, Simonovits Bori, Unger Anna) és tehetséges hallgatók is (Csomor Gábor, Zaboretzky Bendegúz). Mindannyiuknak köszönettel tartozom. Kalandos évek voltak, sok-sok tapasztalattal és sikerrel, az együtt végzett kutatómunka örömével. A kezdetben általunk tréfásan „terra incognita”-nak nevezett terület időközben ismerőssé vált – e kötet célja az évek során összegyűlt, a szociológiai kutatás kontextusában releváns módszertani tapasztalat néhány fontos eszközre koncentráló, korántsem teljes összegzése.

Szeretném megköszönni a szakmai lektor, Tóth Gergely, illetve a sorozatszerkesztő, Kovács László és a technikai szerkesztő, Szőke Viktória munkáját is.

Köszönettel tartozom végül az ELTE Társadalomtudományi Karának az alkotói szabadságért, ami lehetővé tette a könyvhöz szükséges kutatómunka elvégzését.

E könyvben több korábbi publikációm eredményét is felhasználom. Az egyszerzős cikkeknek átszerkesztett, jellemzően továbbfejlesztett verzióját közlöm, míg a többszerzős cikkek módszertani tapasztalatait több más cikkel összevetésben összegzem[1]. A felhasznált publikációk a következők:

Németh Renáta. (2015a). A számok tényleg magukért beszélnek? *Replika*, (92-93), 203-208.

Németh Renáta (2015b): Oksági következtetés az empirikus szociológiai kutatásban. *Szociológiai szemle*, (25:2), 2-30.

Németh Renáta (2021): A felügyelt gépi tanulás kihívásai a szociológiai alkalmazásokban. *Metszetek - társadalomtudományi folyóirat*, (10:3), 27-42.

Németh Renáta, Sik Domonkos, Zaboretzky Bendegúz, Katona Eszter (2023): Depression in times of a pandemic – the impact of COVID-19 on the lay discourses of e-mental health communities. *Information, communication and society*, 1-23.

Németh Renáta (2023): A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, 25 p.

Buda Jakab, Németh Renáta, Simonovits Bori, Simonovits Gábor (2022): The language of discrimination: assessing attention discrimination by Hungarian local governments. *Language Resources and Evaluation* 24 p.

Németh Renáta, Máté Fanni, Katona Eszter, Rakovics Márton, Sik Domonkos (2022): Bio, psycho, or social: supervised machine learning to classify discursive framing of depression in online health communities. *Quality and Quantity: International Journal of Methodology*, 3933-3955.

Németh Renáta, Sik Domonkos, Katona Eszter (2021): The asymmetries of the biopsychosocial model of depression in lay discourses - Topic modelling online depression forums. *SSM Population Health*, 14:100785.

[1] Ezúton nyilatkozom, hogy szerzőtársaim hozzájárultak a témában való közös kutatásaink módszertani tapasztalatainak jelen könyvben való összegzéséhez és önálló megjelentetéséhez.

Németh Renáta, Sik Domonkos, Máté Fanni (2020): Machine learning of concepts hard even for humans: the case of online depression forums. *International Journal of Qualitative Methods*, 19:1609406920949338.

Németh Renáta, Katona Eszter, Balogh Péter, Rakovics Zsófia, Unger Anna (2023): What else comes with a geographical concept beyond geography? The renaissance of the term 'Carpathian Basin' in the Hungarian Parliament. *Intersections. East European Journal of Society and Politics*, megjelenés alatt.

Buda Jakab, Németh Renáta, Rakovics Zsófia (2023): Polarization as a Measure of Text Classification Performance - Evidence from the Hungarian Parliament 1998-2020. Kézirat.

Katona Eszter, Kmetty Zoltán, Németh Renáta (2021): A korrupció hazai online média-reprezentációjának vizsgálata természetes nyelvfeldolgozással. *Médiakutató: médiaelméleti folyóirat*, (22:2), 69-88.

Sik, D., Rakovics, M., Buda, J., Németh, R. (2023). The impact of depression forums on illness narratives: a comprehensive NLP analysis of socialization in e-mental health communities. *Journal of Computational Social Science*, 1-22.

Sik Domonkos, Rakovics Márton, Németh Renáta (2023): The manifest and latent structures of medicalization and psychologization in lay depression discourses – a word embedding analysis of online forums. Kézirat, bírálat alatt.

TARTALOMJEGYZÉK

1. BEVEZETÉS

Bár a nyelv a társadalmi interakciók egy fontos eszköze, a kvantitatív társadalomkutatás – elsősorban adatgyűjtési és feldolgozási eszközök hiányában - mégsem használta igazán évtizedeken át. A helyzet az utóbbi évtizedben gyökeresen megváltozott, a szöveges adat, mint empirikus társadalomkutatási bázis használata exponenciális ütemben terjed (lásd „text as data" mozgalom). Még soha ennyi szöveg nem állt ilyen könnyen rendelkezésre a társadalomkutatás legkülönbözőbb területein, hiszen weboldalakon, közösségi média felületeken, szöveggé alakított videókon, digitalizált könyvtárakban halmozódik fel az új adatvagyon. Tulajdonképpen a társadalom minden alrendszerének létrejön specifikus szöveges leképeződése, gondoljunk csak példaként az e könyvben gyakran elemzett politikai alrendszerre: a parlamenti beszédek, viták, pártprogramok, törvények, online politikai cikkek, laikus bejegyzések mind forrásként használhatók. Míg a társadalomkutatóknak korábban kompromisszumot kellett kötniük az adatok mérete és mélysége között, a digitalizáció lehetővé tette az ilyen korlátozások elhagyását.

A szavazatokkal vagy közvélemény-kutatásokkal ellentétben e szövegek árnyaltabb vélemény kifejezését teszik lehetővé szerzőik számára. Továbbá: az online térben megjelenő szöveges adatok a közvélemény-kutatásokkal ellentétben a megfigyelt viselkedést tükrözik, nem terheltek pl. visszaemlékezési torzítással, vagy a társadalmi elvárások által okozott torzítással, ilyen szempontból érvényesebb következtetésekre adnak lehetőséget.

Az új szöveges adatvagyon társadalomkutatási szempontból fontos jellemzője, hogy a penetráció egyre szélesebb (a Hootsuit 2022-es jelentése szerint pl. a Föld lakosságának már 58%-a közösségi média felhasználó[2]), így a digitális platformok bárki számára megnyilvánulási lehetőséget biztosítanak. Ez fontos változás, hiszen korábban a nyilvánosság elé kerülő szövegek szinte kizárólag az elit tollából születtek. Jó példa utóbbi jelenségre a Google Ngram Viewer platform,[3] ahol a világ utóbbi öt évszázadban íródott többszáz-millió könyvét érjük el digitalizált formában, egyszerűbb kvantitatív elemzéseket is lehetővé tevő eszközzel, izgalmas társadalomtörténeti vizsgálatokra kapva lehetőséget, de azzal az episztemológiai korláttal, hogy mindig a korszak elit narratíváihoz férünk csak hozzá.

[2] https://www.hootsuite.com/resources/digital-trends

[3] https://books.google.com/ngrams

A szöveganalitikai társadalomkutatások egyik fontos empirikus alapja a szerkesztett és online média, ezért itt meg kell említenünk a "társadalom mediatizálódásának" kortárs folyamatát is. A mediatizáció a késő modernitásban egyszerre jelenti a média, mint saját logikával rendelkező önálló intézmény megjelenését, illetve azt a folyamatot, melynek révén a média más intézmények, mint a politika vagy a munka integrált részévé válik (Hjarvard, 2008).

A szöveges adatok mennyiségének és hozzáférhetőségének ez a forradalma jelentősen kiszélesítette az empirikusan vizsgálható társadalomkutatási kérdések körét: az egyének, csoportok és intézmények viselkedését, azok kölcsönhatását és időbeli dinamikáját naponta többmillió terrabyte-nyi digitális szöveg képezi le, s ez az adatvagyon a digitalizáció előrehaladtával egyre sokszorozódik. Az online generált adatok lehetővé teszik, hogy valós kontextusban és valós időben kövessük az emberi viselkedést, ami jelentősen túlmutat a szociológusok hagyományos kutatási módszerein.

A társadalmat leíró szöveges adatok forradalmával párhuzamosan az utóbbi tíz évben a számítási kapacitások és azzal párhuzamosan az adatok elemzésére szolgáló szöveganalitikai technológiák robbanásszerű fejlődése is bekövetkezett, s az új technológiák a szöveg feldolgozásának már releváns mélységét nyújtják. Ez a robbanás a számítástudomány és számítógépes nyelvészet után a digitális bölcsészetet és a társadalomtudományt, így a szociológiát is elérte, és egymás után jelentek meg nem csak a terület-specifikus eszközök és modellek, de a lehetőségeket bemutató írások is (Evans & Aceves; 2016, Ignatow & Mihalcea, 2017; hazánkban pl. Németh et al.; 2020, Kmetty, 2022). Az új eszközök lehetőségei inspirálóak, ugyanakkor episztemológiai és technikai korlátaikról sem feledkezhetünk meg (Németh & Koltai, 2021; Németh & Koltai, 2023).

Savage és Burrows 2007-ben az évtized egyik legtöbbet idézett szociológiai tanulmányában az empirikus szociológia közelgő válságáról írt. Azt jósolták, hogy válság következik be, ha a korábban saját módszertani szakértelméről ismert szociológia nem tud megfelelni a big data által támasztott kihívásoknak, és így elveszíti vezető szerepét. Ez nem következett be. Nyolc évvel később a British Sociological Association által kiadott „*Sociological Futures*" című könyvsorozat első tagja (Ryan – McKie, 2015) már a címében is utalt a válság végére, és fontos lehetőségeket látott a big data kutatásban, valamint a természetes nyelvfeldolgozásban (natural language processing, NLP) is. Ugyanakkor a korábban a szociológia által uralt empirikus szakértelmet tekintve egyértelmű elmozdulás figyelhető meg az akadémiai szférából az ipar felé, hiszen a terület hatalmas üzleti

lehetőségeket generál és az ipar finanszírozni is képes a szükséges fejlesztéseket. Ezért az NLP társadalomkutatási alkalmazása ebből a szempontból nyilvánvaló lépéshátrányban van.

Ez a kötet a természetes nyelvfeldolgozás szociológiai alkalmazásaiba enged bepillantást, kutatócsoportunk, az ELTE Társadalomtudományi Karán a Barna Ildikó és általam vezetett ELTE Research Center for Computational Social Science kutatócsoportban (rc2s2.elte.hu) 2018 óta folyó kutatásokon, mint esettanulmányokon keresztül. Az esettanulmányok zöme két nagy projektünkből származik. Az első „A politikai nyilvánosság rétegei Magyarországon (2001-2020)" címet viseli[4] (NKFIH K-134428), a projekt célja a hivatásos politikusi, a professzionális sajtóban megjelenő és a laikus online közbeszéd szociológiai elemzése automatizált szövegelemzés segítségével. A projekt résztvevői: Barna Ildikó, Buda Jakab, Csigó Péter, Katona Eszter, Knap Árpád, Németh Renáta (vezető kutató), Pólya Tibor, Rakovics Márton, Rakovics Zsófia, Sik Domonkos, Tóth Emese és Unger Anna. A kutatás átfogó célja a magyar politikai közbeszéd feltérképezése a 2000-es évektől napjainkig. A politikai szféra és a nyilvánosság átalakulása körvonalazza kutatásunk tartalmi keretét, a politikai diskurzus különböző rétegeit elemezzük, beleértve a hivatalos kommunikációs csatornákat (pl. parlamenti beszédeket) és az online sajtó különböző típusait is. A kutatás keretei között a diskurzusok tartalmának (a megvitatott témáknak) a vizsgálatát, valamint a nyelvhasználat/keretezés elemzését végezzük el.

Az esettanulmányokat adó másik nagy kutatásunk „A depresszió diszkurzív keretezése online fórumok közösségében" címet viseli[5], 2018 óta fut, az első három évben a Felsőoktatási Intézményi Kiválósági Program, majd egy-egy évben a Társadalmi Innovációs Nemzeti Laboratórium támogatásával. Résztvevők: Buda Jakab, Kapitány-Fövény Máté, Katona Eszter, Németh Renáta, Pólya Tibor, Rakovics Márton, Sik Domonkos (vezető kutató) és Zaboretzky Bendegúz. Kutatásunkban az NLP módszerek lehetőségeit vizsgáljuk a depresszió online betegközösségekben megjelenő egyéni szintű keretezésének megértésében. A depresszió a modernitás betegsége, kognitív keretezése társadalmi konstrukció. A keretezés határozza meg a depresszió jelentését a beteg számára, oksági magyarázatot kínálhat, sőt akár a kezelési preferenciákat is meghatározza. E téren

[4] További részletek: https://rc2s2.elte.hu/project/a-politikai-nyilvanossag-retegei-magyarorszagon-2001-2020/

[5] További részletek: https://rc2s2.elte.hu/project/a-depresszio-diszkurziv-keretezese-online-forumok-kozossegeben/

korábban elsősorban kvalitatív módon, offline szövegek (naplók, levelek, interjúk) elemzésével közelítették a keretezést. Kutatócsoportunk kiindulópontja szerint a digitális társadalom online betegközösségeinek nem-klinikai jellegű írásai jó terepet kínálnak a kérdés vizsgálatára, s hogy az automatizált szöveganalitikai módszerek jelentős kutatási potenciált jelenthetnek e téren.

Egy harmadik, kisebb volumenű kutatásból is szemlézek eredményeket. A kutatás során 2020-ban magyarországi önkormányzati hivatalok körében végeztünk kontrollált terepkísérletet annak céljából, hogy feltárjuk az online ügyintézés során esetlegesen előforduló hivatali diszkrimináció nyílt és burkolt formáit. Hétköznapi ügyintézéssel kapcsolatban fogalmaztunk meg kérdéseket, és a válasz-emaileket elemeztük – a kutatás egyediségét az adja, hogy párhuzamosan kódolókkal és gépi tanulással is feldolgoztuk őket (Csomor et al., 2021; Buda et al., 2022; Simonovits et al., 2022). A kutatás résztvevői Csomor Gábor, Hobot Péter, Németh Renáta, Simonovits Bori, Simonovits Gábor, Vig Ádám voltak.

E kötetben célom tehát, hogy a fenti kutatásokhoz kapcsolódó konkrét esettanulmányok segítségével illusztráljam az NLP, mint megközelítés lehetőségeit és korlátait, elsősorban a társadalomkutatás módszertani kontextusára, a módszertani innovációra és módszertani alternatívákra koncentrálva. Ideális olvasóm az a kvantitatív társadalomkutató, aki érdeklődik a számítógépes szövegelemzés lehetőségei iránt, és nem-technikai utat keres azok megismerésére. A kötet a megközelítés szociológiai kutatás során termelődött alkalmazási tapasztalatait rendszerezi – tehát nem kézikönyv, nem törekszik enciklopédikus teljességre, és nem tartalmazza a módszerek technikai/matematikai leírását, inkább intuitív módon közelít azokhoz, de bőséges referencia-listával ad minden fejezetben lehetőséget az érdeklődők további elmélyülésére. Ugyanezen okból (és a terület villámgyors fejlődéséből adódóan) a legkurrensebb, társadalomkutatási lehetőségeket is rejtő módszerek még nem, vagy csak érintőlegesen kerülnek tárgyalásra – itt elsősorban a nagy nyelvi modellekre (large language models, LLMs) gondolok.

Tapasztalatom szerint az NLP technikai oldalának ismertetésére kiváló források állnak rendelkezésre, de a társadalomkutatási tapasztalatokat és kihívásokat jóval kevesebb szerző tárgyalja. A társadalomkutatás alkalmazási specifikumát az adja, hogy az itt tárgyalt problémák egy évszázados kutatási paradigmába vannak ágyazva, kérdésfeltevései így lényegesen különböznek a számítástudomány vagy az ipari felhasználás kérdéseitől. Ennek a különbségnek pedig tudatában kell lennünk, amikor adaptáljuk az informatika oldaláról érkező innovációt. Ezért tűztem célul, hogy néhány példán keresztül megmutassam azt a kutatási logikát,

amit a szöveg, mint adatnak a társadalmi világ megismerése céljával történő felhasználása mögött áll.

Remélem, hogy a megközelítések, alternatív megoldások, lehetséges hibák, esetleges hiányosságok és megoldások összegzésével kiindulópontot tudok nyújtani jövőbeli kutatásokhoz. Mivel az NLP viszonylag új módszertani megközelítés, talán e kötet új kutatásokat is inspirálhat. Ezen túl az NLP-t nem ismerő társadalomtudományi kutatóknak is szeretnék bepillantást nyújtani e kutatásokba.

A kötet első fejezete *(Az NLP mint kutatási módszer társadalomtudományi kontextusa*) a szöveg, mint adat jelenséget világítja meg, a tudománytörténeti hátteret és a meglevő szövegelemzési alternatívákat is tárgyalva. A második fejezet (*Felügyelt és felügyelet nélküli tanulás*) az NLP (és általában a mesterséges intelligencia-kutatás) két fő ágát ismerteti, egymással összevetésben, majd a következő két fejezet részletesebben ki is bontja azok lehetőségeit, eszközeit. *A felügyelet nélküli tanulás* c. fejezetben a topikmodellt, annak variánsait, illetve a szóbeágyazást tárgyalom és mutatom meg több alkalmazásukat két projektünk, a depressziós fórumok ill. a hazai politikai nyilvánosság kapcsán. *A felügyelt tanulás* c. fejezet az ipari/üzleti alkalmazásokban már sokszorosan bizonyított felügyelt gépi tanulás szociológiai alkalmazásainak sajátos kérdéseit tárgyalja. A sajátosság oka, hogy ezekben az alkalmazásokban komplex fogalmak megtanulása az algoritmus feladata (pl., hogy gyűlöletbeszédet tartalmaz-e egy tweet). A felügyelt tanulás lényege, hogy előre bekódolt (gyűlöletbeszéd/nem gyűlöletbeszéd) szövegek címkézését tanulja meg az algoritmus, jellegzetes szövegmintázatokat keresve. A felmerülő kérdések: hogyan jön létre a címkézés? Hogyan lehet betanított kódolókkal elvégeztetni egy olyan hermeneutikai kihívást, mint a gyűlöletbeszéd felismerése? Segítenek-e ezen a rutinszerűen alkalmazott, részletezett annotálási irányelvek? A fejezet arra is kitér, hogyan végzik crowdsourcing platformokon a kódolást a nagy cégek, illetve ismertetem az MI-torzítást is, aminek itt az a lényege, hogy a kódolók maguk viszik be a diszkriminációt az adatokba. Az annotálás kihívásai után egy oylan projektünket ismertetem, ahol emberi kódolókat és párhuzamosan gépi tanulást is alkalmaztunk, majd két izgalmas területre térek: a tanuló modellek teljesítmény-mutatójának szociológiai értelmezési lehetőségeit tárgyalom, illetve az „*explainable AI*” kortárs problematikáját: azt, hogy hogy miért kívánatos a tanuló algoritmusok fekete dobozának kinyitása. E kérdéseket ismét kutatási tapasztalatainkkal illusztrálom.

Az *NLP a politikai polarizáció kutatásában* c. fejezet átfogó irodalmi áttekintést ad egy konkrét és izgalmas alkalmazási területen, célja, hogy szisztematikus

képet adjon módszertani oldalról a lehetőségekről és korlátokról, ugyanakkor inspirálja is az olvasót. Az NLP technikák új lehetőségeket kínálnak a politikai polarizáció nyelvi megnyilvánulásainak feltárására is: e módszerek segítségével következtethetünk egy adott szerző/beszélő politikai nézeteire, mérhetjük a polarizáció nagyságát, nyomon követhetjük annak tendenciáit, így közelebb kerülhetünk a jelenség megértéséhez. A fejezet ugyanakkor általánosabb relevanciával bír, ugyanis az ott számba vett alkalmazási kihívások (hogyan határoljuk le a vizsgálat korpuszát? hogyan definiáljuk a vizsgálat egyégét? hatékonyan működnek-e más korpuszokon is modelljeink? hogyan válasszuk meg a vizsgált szövegek jellemzőit? mi az interpretáció szerepe a gépi tanulás alkalmazásaiban? hogyan működhet együtt a kvalitatív módszer az automatizált megközelítéssel?) minden szociológiai alkalmazásban megfontolandók.

Az oksági következtetés az NLP-elemzés során c. fejezet ebben az új kontextusban tekint rá az engem régóta (Németh, 2015b, 2021) foglalkoztató problémára, az oksági következtetés lehetőségeire. Végül a záró, *Összegzés* c. fejezet a kötet tanulságait igyekszik összefoglalni.

2. Az NLP mint kutatási módszer társadalomtudományi kontextusa

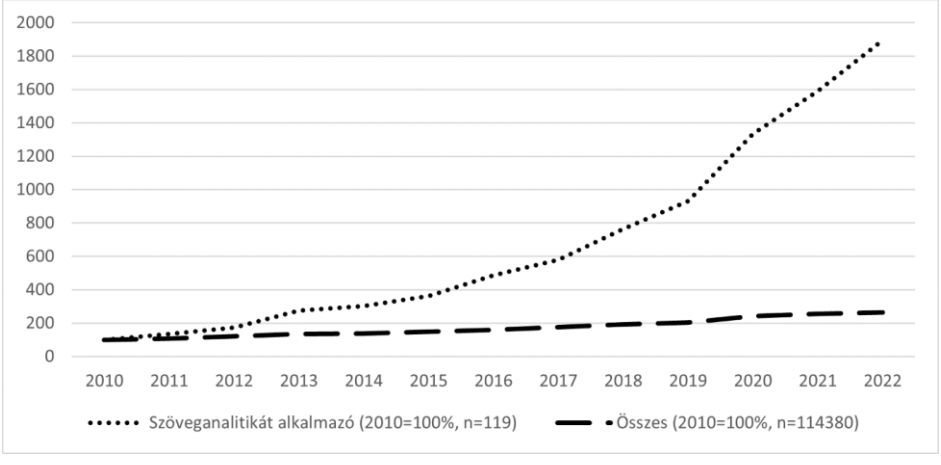
Az NLP izgalmas, és a szociológia számára perspektivikus terület az informatika, mesterséges intelligencia-kutatás és nyelvészet határán (Evans & Aceves, 2016; Ignatow & Mihalcea, 2018). Megjegyezném, hogy a módszer elnevezésével kapcsolatban sem az angol, sem a magyar terminológia nem kanonizálódott még, a szövegbányászat, számítógépes nyelvészet, automatizált szövegelemzés diffúz körvonalakkal bíró, rokon, de nem szinonim elnevezések (erről részletesebben: Németh et al., 2020). Az NLP olyan területeket is magában foglal, mint a beszédfelismerés vagy szövegek szintaktikai feldolgozása (Hirschberg & Manning, 2015), de a társadalomkutatók inkább azokat az eszközeit használják, amelyek társadalomkutatási kérdésekre adnak (új) választ – e kötetben ezekre látunk példákat. A módszert technikai oldalról is megismerni kívánóknak jó szívvel ajánlom Jurafksy és Martin (2023), Eisenstein (2019) és Silge és Robinson (2017) kötetét.

Az utóbbi évtizedben az NLP tudományos alkalmazása hatalmas növekedést produkált. Ambiciózus projektek használják az egészségügyben, üzleti alkalmazásokban, marketingben és nemzetvédelmi területen egyaránt. Az elmúlt néhány évben az NLP a társadalomtudományokban is kezdett teret nyerni, a politológiától a közgazdaságtanon át a szociológiáig (Ignatow & Mihalcea, 2017). Igaz, a társterületekhez képest a szociológiában némi késéssel indult el a változás (Edelmann et al., 2020), de a 2010-es években már világos volt a trend, például az *Annual Review of Sociology*-ban ebben az évtizedben hat olyan írás is megjelent, ami a számítógépes módszerek használatát tárgyalta a szociológiai big data kutatásban.

Ahogy egyszerű szcientrometriai elemzésem[6] (lásd 2. ábra) mutatja: míg a társadalomtudományi publikációk száma csak szolid növekedést produkált a 2010-es évtől kezdődően, addig a számítógépes szöveganalitikát alkalmazó

[6] A trendelemzés a Dimensions.ai tudománymetriai adatbázis segítségével történt. A társadalomtudományokat a „human society" kategóriával azonosítottam, ami a szociológia, politikatudomány, antropológia stb. közös kategóriája, és az ANZSRC 2020 Fields of Research (FoR) tudományosztályozáson alapul, annak 44-es kódjának felel meg. Az elemzés minden angol nyelvű publikációtípusra kiterjedt, a szöveganalitikát alkalmazó cikkeket kulcsszavas kereséssel azonosítottam, a publikáció bármely részén keresve vagylagosan összekapcsolt kulcsszavakat ("automated text" OR "natural language processing" OR "computer-assisted text" OR "computational linguistics" OR "text mining" OR "computational text").

publikációk száma exponenciálisan növekedett. Az 1. ábra a 2010-es kezdőponthoz vonatkoztatva mutatja a publikációk számának alakulását. E felfutás mögött egyszerre áll a digitális szöveges források elérhetősége és a hatalmas adatbázisok feldolgozásához szükséges technológia fejlődése.



1. ábra. A szöveganalitikát alkalmazó ill. az összes társadalomtudományos publikáció számának alakulása a 2010-es év százalékában.

Természetesen, míg az NLP egy viszonylag új interdiszciplináris terület, maga a szövegelemzés több évtizedes tradícióra tekint vissza a társadalomkutatásban. A kvantitatív szövegelemzés a két világháború között indult a tömegmédia elemzése céljából, és a háború után is folytatódott (pl. Berelson & Lazarsfeld, 1948). A kvantitatív elemzés tipikus felhasználási módja a kutató által meghatározott "kódok" szövegekben való megjelenésének számszerűsítése és a kódok kapcsolatainak azonosítása volt (lásd pl. Bales, 1950, kvantitatív tartalomelemzési iskolaként kanonizálódva Krippendorff, 1995), az 1960-as évektől kezdve számítógépek támogatása mellett (Hayes, 1960). Ezekben a korai elemzésekben a kódok mellett a nyers szövegelemeket (pl. releváns szavak) és metaadatokat (pl. szerző) is felhasználták a stilisztikai/szemantikai mintázatok feltárására. Érdemes megjegyezni, hogy a kvantitatív megközelítés pl. a tartalomelemzés esetében itt gyakran együtt jár a szöveg tényleges elolvasásával (ha a kódolás azt megkívánja), de a szöveget elsősorban adatként használja, tehát nem a szövegre magára irányul az analízis, hanem a szövegből kivont jellemzőkre.

A kvantitatív elemzések felszínesnek vélt eredményei inspirálták a kvalitatív szövegelemzés megjelenését a XX. század második felébe (Krippendorff, 1996). A

hermeneutika hagyományára építve a szövegek keletkezési kontextusának fontosságát hangsúlyozták, és expliciten távol tartották a kutatót saját *a priori* elvárásai befolyásától. Kifinomult interpretációs megközelítések alakultak ki, mint az elméletalkotás teljes folyamatának átláthatóvá tételét célzó megalapozott elmélet (Glaser & Strauss, 1967), vagy a kortárs társadalomkutatásban széles körben elterjedt kritikai diskurzuselemzés, melynek célja a tudás és a hatalom összefüggéseinek feltárása a szöveg kontextusát adó társadalmi struktúrákban (van Dijk, 1994).

A szövegre építő empirikus társadalomkutatási irányzatnak ismeretelméleti szempontból is több évtizedes előzményei vannak. Egyik ilyen előzmény a társadalomtudományok narratív fordulata (Goodson & Gill, 2011), amely elsősorban a pozitivista kutatás módszertani alternatíváját kívánta nyújtani, a szövegben látva a társadalmi jelenségek önreflexív megközelítésének lehetőségét. Több másik, ettől nem független, tudományág-specifikus előzmény a diszkurzív megközelítések társadalomkutatási megjelenése: ilyen a nyelv és politika tudománya, ahol a nyelvet a politikai cselekvés nélkülözhetetlen eszközének tekintik (Wodak & Forchtner 2017; Müller, 2008), vagy a nyelv és történelem tudománya, ahol a nemzeti történelem (Wodak, 2010) vagy a nemzeti identitás (Wodak et al., 2009) diszkurzív konstrukciójára fókuszálnak.

Ahogy láttuk, a hagyományos kvantitatív szövegelemzés is jellemzően igényelte a szöveg tényleges olvasását/megértését. Ez a megközelítés az ezredfordulót követően kezdett el megváltozni, innét kezdve a szöveg nem elolvasandó, megértendő tárgyként, hanem inkább mint az automatizált módszerek inputjaként jelent meg, anélkül, hogy ténylegesen bárki elolvasná. Az NLP társadalomkutatásban történő felhasználása tehát ehhez a közelmúltban elterjedt "szöveg, mint adat" (*text as data*, Gentzkow et al., 2019) megközelítéshez kapcsolódik, ahol a szöveg mint rendezett, jól strukturált, numerikus adatbázisba rendezett állomány adja a számítógépes algoritmusok inputját. A strukturálatlan szöveges input jól strukturált adatbázissá rendezése többféle módon történhet (egy példát mutat erre az 1. ábra, magyar parlamenti beszédek korpusza alapján). A lényeg az, hogy ez mindig egy absztrakciós lépés, olyan struktúrát célozva, amely a szöveg logikáján kívül kerül meghatározásra, az elemzési eszköz alkalmazhatósága érdekében. Például az 1. ábrán szemléltetett folyamat során a szövegekből többlépéses előfeldolgozás (*preprocessing*) során úgynevezett dokumentum-kifejezés (*document-term*) mátrix jön létre. Az előfeldolgozás itt a példában névelemfelismeréssel (*named entity recognition*, tulajdonnevek vagy kollokációk azonosításával, mint az Európai_Unió vagy a határ_túli_magyarok), kifejezésekre tördeléssel,

írásjelek és nagyon gyakori, un. stopszavak (pl. névelők) törlésével, kisbetűsítéssel és lemmatizálással (szótövekkel történő egységesítéssel) történik.

Itt a szöveget egyszerű „szózsák" (*bag of words*) modell reprezentálja, ami eltekint a toldalékolástól, a szavak sorrendjétől vagy a mondat-tagolástól, de megtartja a multiplicitást, vagyis számontartja, melyik kifejezés hányszor fordult elő a teljes szövegben.

Eredményként egy, a kvantitatív társadalomkutatásban mindennapos adatbázist kapunk (mint egy survey esetén), a sorokban kérdezettek helyett szövegekkel, az oszlopokban a hozzájuk tartozó numerikus jellemzőkkel. Ezzel a formátummal a mátrix típusú adatok elemzésére létrehozott statisztikai elemzés teljes eszköztárát tesszük elérhetővé: eloszlások összehasonlítása, dimenziócsökkentési technikák, regresszióelemzés és a többváltozós elemzés más formái, osztályozást célzó gépi tanulás stb. A klasszikus kvantitatív módszerekben jártas olvasó maga is látja ezt, ha arra gondol, hogy a mátrix oszlopai mint változók használhatók lesznek pl. egy regresszió-típusú modellben, ahol a függő változó a szöveg valamely kiemelt tulajdonsága. Például a parlamenti példában a beszéd szerzőjének párthovatartozása lehet ilyen kiemelt tulajdonság, egyfajta klasszifikációs modellként, és így a politikai ideológiákat leginkább megkülönböztet nyelvi jegyeket lesz a modell képes meghatározni. Ha a megfigyelt szöveg-mintánk egy populációt reprezentáló mintának tekinthető, akkor a felsorolt módszerek valószínűségi jellegű következtetések és becslések létrehozására is lehetőséget adnak.

A fent említett absztrakciós lépés itt tehát a szózsák-modellnek az alkalmazását jelenti. A szövegnek ez az absztrakt modellje nyilván nagy veszteséggel jár, gondoljunk csak arra, hogy a tagadás vagy az irónia így gyakorlatilag detektálhatatlan. Nyilvánvaló: ha a szöveget nem szövegként, hanem adatként használjuk, akkor jelentősen leredukáljuk az eredeti, közvetlenül értelmezhető szöveges forrást – ez az ára annak, hogy nem a szövegek elolvasása/megértése a cél, hanem mintázatok azonosítása. A statisztikában jól ismert tétel, George Box statisztikus bonmot-ja – „All models are wrong, but some are useful" – itt is érvényes, remélem, olvasóm ezzel maga is egyetért majd a könyv elolvasása után.

Ugyanakkor érdemes itt megemlíteni a politológus Benoit (2020) az előbbinél kevésbé mentegetőző megközelítését. Benoit a szöveg, mint szöveg használatát szembeállítja a szöveg, mint adat használattal. Szerinte utóbbi megközelítéssel a társadalomkutató célja vagy valamely manifeszt, a szövegben is kifejeződő jelenség vizsgálata (saját példámmal Decadri és Boussalis, 2020, a populizmus és beszédkomplexitás közötti kapcsolat vizsgálata olasz parlamenti

beszédekben), vagy olyan látens hitek és vélemények (pl. politikai álláspont) elérése, amelyek nem-verbális kifejeződése nehezen megközelíthető. A látens jellemzők kutatása esetén a kutatás célja elsősorban nem a szöveg tényleges tartalmának felderítése, hanem annak megítélése, hogy a szöveg, mint adat segítségével mit tudunk felfedni abból a látens jellemzőből, aminek a szöveg adatként egyfajta megfigyelhető következménye. Ez a megközelítés azt is megengedi, hogy a szöveg adatként történő használata olyan következtetésekre adjon lehetőséget, amit szövegként történő használatával (tehát tényleges elolvasásával) nem tudnánk elérni (!). És valóban, ilyen NLP-felhasználásra több példa is van, lásd pl. a mentális zavarok diagnózisát (Zhang et al., 2022) vagy öngyilkossági késztetések detektálását (Homan et al., 2022) a nyelvhasználatból, ahol olyan, a szöveg laikus olvasója által nem ismert jellemzőket találtak fontosnak, mint pl. az egyes szám első személyű névmás használati gyakorisága.

Forrás-dokumentumok	Névelem-felismerés és tokenizálás	Írásjelek és stopszavak eltávolítása	Kisbetűsítés és lemmatizálás
[Németh Zsolt, 1998] A nemzeti érdekérvényesítés két kiemelkedően fontos kérdéséről szeretnék beszélni, az egyik az uniós posztok ügye, a másik a határon túli magyarság kérdése.	A nemzeti érdekérvényesítés két kiemelkedően fontos kérdéséről szeretnék beszélni , az egyik az uniós posztok ügye , a másik a határon_túli_magyarság kérdése .	nemzeti érdekérvényesítés két kiemelkedően fontos kérdéséről szeretnék beszélni uniós posztok ügye határon_túli_magyarság kérdése	nemzeti érdekérvényesítés két kiemelkedő fontos kérdés szeret beszél uniós poszt ügy határon_túli_magyarság kérdés
[Mile Lajos, 2010] Az Európai Unió markánsabb, szervezettebb megjelenése a nemzetközi színtéren, a világpolitika dimenzióiban egyre érzékelhetőbbé vált, főleg a lisszaboni szerződés elfogadásaóta.	Az Európai_Unió markánsabb , szervezettebb megjelenése a nemzetközi színtéren a világpolitika dimenzióiban egyre érzékelhetőbbé vált , főleg a lisszaboni_szerződés elfogadása óta .	Európai_Unió markánsabb szervezettebb megjelenése nemzetközi színtéren világpolitika dimenzióiban érzékelhetőbbé vált lisszaboni_szerződés elfogadása	Európai_Unió markáns szervezett megjelenés nemzetközi színtér világpolitika dimenzió érzékelhető válik lisszaboni_szerződés elfogadás

A szöveg-kifejezések mátrixba rendezése

id	kifejezés				
	nemzeti	érdekérvényesítés	két	Európai_Unió	nemzet-közi
nemeth_98	1	1	1	0	0
mile_10	0	0	0	1	1

2. ábra. Szövegből mátrix.

A hagyományos kvantitatív szövegelemzés, ahogy fent láttuk, inkább csak bizonyos kifejezések vagy kódok megjelenését számszerűsítette a szövegekben. Ehhez képest nagy előrelépést jelent az NLP eszköztára, ami olyan feladatok elvégzését automatizálja, mint szövegek érzelmi töltetének azonosítása, szövegek távolságának meghatározása, szövegklaszterek létrehozása, látens tematikus struktúrák vagy látens szemantikai relációk azonosítása (Németh et al., 2020). A fent tárgyalt, jól ismert mátrix-típusú adatbázis megléte esetén ezeknek a módszereknek a jó része a megszokott kvantitatív eszközökkel elvégezhető lenne (klaszterelemzés, regresszió-elemzés stb.), de statisztikai kihívást a mátrix nagysága (potenciálisan milliós minta, többezres változószám – terminus technicus *high-dimensional data*), emiatt túlillesztés veszélye) és a mátrixon belül a jellemzően sok nulla (terminus technicus: *sparse data*, pl. mert szótárunk szavainak jó része nem szerepel adott szövegben) adja. Ezek a kihívások a survey-alapú adatok esetében, klasszikus statisztikai területen kevésbé fordulnak elő, és az adattudomány ad rájuk megoldásokat. Utóbbiakat itt nem részletezem, de pl. Eisenstein (2019) korrekt ismertetést ad róluk a szövegelemzési alkalmazások keretén belül.

A szöveg, mint adat megközelítés a társadalomkutatásban elsősorban a közvélemény-kutatásokkal szemben jelent alternatívát. A digitális szöveges adatok használatának módszertani és ismeretelméleti előnyei is vannak a hagyományos közvélemény-kutatásokkal szemben. A módszertani előny lényege, hogy a digitális forradalom a véleménynyilvánítást átcsatornázta az internetre, és a számítási módszerek hozzáférést biztosítanak ezekhez a hatalmas adatmennyiségekhez. Az ismeretelméleti előny abból adódik, hogy a digitális szöveges adatok "talált adatok" *(found data)* abban az értelemben, hogy általában valamilyen, a tudományos elemzéstől eltérő céllal készültek (Németh & Koltai, 2021). Ezért a

közvélemény-kutatási adatokkal ellentétben ezeknél nem áll fenn a válaszmegtagadás lehetősége. Továbbá, mivel a vélemény-nyilvánítást és interakciókat természetes környezetükben lehet megfigyelni, ezek az adatok a megfigyelt viselkedést tükrözik, szemben az önbevallásos válaszokkal, így a belőlük nyert álláspont feltehetően nagyobb belső érvényességgel rendelkezik, például nem terheli felidézési torzítás vagy a társadalmi elvárásból eredő torzítás (*recall bias, social desirability bias*). További előnyük, hogy valós időben elérhetők, hogy akár kövesére is lehetőséget adnak (longitudinális vizsgálat), és hogy mivel gyakorlatilag mindenkire kiterjednek, ritka jelenségek, nehezen elérhető alpopulációk vizsgálatára is lehetőséget adnak (pl. HIV-fertőzöttek társadalmi hálózata, Liu & Lu, 2018). A korábbi korszakokból származó digitalizált szövegek történeti visszatekintő vizsgálatokra is lehetőséget nyújtanak. Végül: digitalizált archívumokban elérhető, nagytömegű, klasszikusan kvalitatív módon elemzett szövegek is tárgyai lehetnek az NLP-nek (terepdokumentumok, interjúk), illetve nyitott survey-kérdések is hatékonyabban használhatók segítségével. A survey, mint módszer arra is lehetőséget ad, hogy a válaszok algoritmikus feldolgozása mellett a válaszadók szociodemográfiai hátterét, különböző attitűdjeit is elemzésbe vonjuk – ez nyilván nem lehetséges (és sokszor fájóan hiányzik is), ha pl. közösségi média használók posztjait elemezzük.

A digitális szöveges adatok felhasználása ugyanakkor kihívásokkal is jár adatminőségi szempontból. Míg a kutatóknak survey esetén ellenőrzésük van a mintavétel felett, az online gyűjtött szövegtömeg csak illúzióját kelti a teljességnek. Eközben számos torzítást tartalmazhat pl. a platform elérhetőségének korlátozásából adódóan. Figyelembe veendő tényezők például a lefedettség (ki használja a platformot?), a nehezen formalizálható mintavételi eljárások, a torzított mintaösszetétel, valamint a zaj jelenléte (Németh & Koltai, 2021). A survey hiba vizsgálatának („*total survey error*") mintájára létrehozott, a digitális nyomok hibájára vonatkozó teljes hiba keret („*total error framing*") fogalmi struktúrát biztosít e problémák azonosításához és számszerűsítéséhez (Sen et al., 2021).

A talált adat jellegnek, vagyis annak, hogy ezek az adatok nem konkrét kutatási kérdéseket szem előtt tartva jöttek létre, megvan a hátránya is: a szövegek mellől hiányoznak a survey-ek esetén rutinszerűen elérhető releváns információk, mint a szerzők társadalmi-gazdasági státusza, vagy a szöveg keletkezésének kontextusa. Az előbbi probléma egy újszerű megoldása a survey kombinálása közösségi média adatok átadásával (ennek az adatdonációs megoldásnak

hazai úttörője Kmetty Zoltán, lásd Breuer et al., 2022, a módszer alkalmazásának konkrét, kulturális fogyasztást elemző példája Kmetty & Németh, 2022).

Az NLP szociológiai alkalmazásának általánosabb korlátai is vannak, az egyik ilyen a nyelvi korlát. Az NLP-technikákat elsősorban angol nyelvre fejlesztették ki és optimalizálták, ami kihívást jelent, amikor az angolon kívüli, alulreprezentált nyelvekre alkalmazzák őket. Egy másik korlát a szociológiai kutatási kérdések megválaszolására alkalmas szövegek összetettsége. A szociológia számára releváns szövegek gyakran tartalmaznak árnyalt és kontextusfüggő jelentéseket, ami kihívást jelent az algoritmusok számára, szemben az NLP-technikák más sikeres alkalmazási területeivel, mint pl. a jellemzően strukturált és jól definiált szövegeket feldolgozó biomedikális kutatások, lásd a tudományos cikkek és klinikai jelentések automatikus feldolgozását (Doan et al., 2014).

Meg kell említeni végül, hogy etikai kérdések is felmerülnek az online gyűjtött adat körül. A szociológia számára releváns szövegek gyakran személyes és potenciálisan érzékeny információkat tartalmaznak, lásd például a közösségi médiában közzétett bejegyzéseket vagy online fórumokat. A survey-ek válaszadóival itt az egyének nem adtak explicit módon beleegyezést adataik felhasználására, és könnyen előállhat beazonosíthatóság is. Az adatvédelmi és etikai irányelvek betartása ezért kulcsfontosságúvá az NLP szociológiai kutatásokban való alkalmazásakor.

Bár a bevezetőben a kvalitatív és kvantitatív szövegelemzési megközelítést mintegy szembeállítottam egymással, meg kell jegyeznem, hogy az NLP legtöbb szociológiai alkalmazásában megkerülhetetlen a kvalitatív módszer használata az elemzés valamely pontján, leginkább a validálás és az interpretáció szakaszában. A legtöbb modell validálása ugyanis jelentős kvalitatív munkát igényel, mivel a modellek értelmezhetősége és hatékonysága sok esetben (amint majd látni fogjuk, pl. a topikmodell esetén) csak a modell által legrelevánsabbnak ítélt szövegek tényleges elolvasásával ítélhető meg. A validálás alapvető fontosságú, ha meg akarunk bízni az automatizált módszerek eredményeiben. Ha a modell outputjának értelemzése alapján nem bízunk az eredményekben, akkor dönthetünk úgy, hogy az algoritmust vagy annak paramétereit módosítjuk és megismételjük a folyamatot, amíg jobb eredményeket nem kapunk. Ezt a ciklust gyakran többször meg kell ismételni.

Hasonlóan, sokszor a modell interpretációjának támogatására is kvalitatív megközelítés használandó, hiszen a komplex NLP-modelleket nehéz értelmezni anélkül, hogy visszanyúlnánk az eredeti szövegekhez. Pl. a klasszifikáló modell

által legfontosabbnak ítélt kifejezéseket tartalmazó szövegek feldolgozásában. Az *NLP a politikai polarizáció kutatásában* c. fejezet azt is jól illusztrálja majd, miért megkerülhetetlen sok esetben a kvalitatív megközelítés, illetve, hogy menynyivel szegényebbek azok a kutatások, amelyek nem élnek ennek lehetőségével, és gyakran el is hagyják a modellinterpretációs lépést. Fontos megjegyezni, hogy a kvalitatív és kvantitatív elemzés itt nem egymás mellett, hanem együtt, egymással kombinációban valósul meg, az egyik inputként használja a másik eredményeit, vagyis nem multimódszeres, hanem kevert vagy vegyes módszeres („mixed method") kutatást látunk (Király et al., 2014).

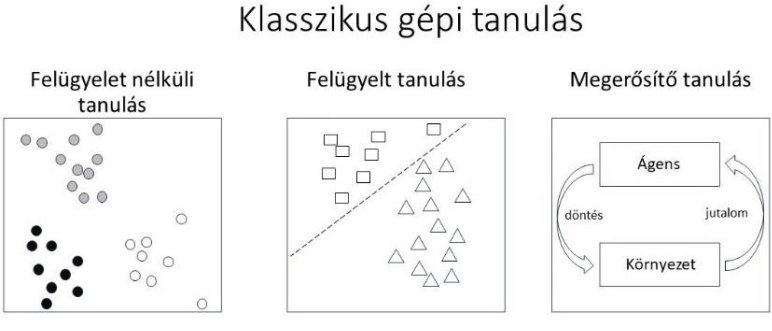
A kevert módszertani megközelítés alkalmazása expliciten megtalálható az NLP szociológia-módszertani szakirodalmában. Ignatow és Mihalcea (2017, 67–68) azon az álláspontón vannak, hogy a társadalomtudományi szövegbányászati kutatásokat általában a kvantitatív és az interpretatív elemek pragmatikus kombinációjaként végzik. Bauer és társai (2014) még tovább mennek, és azt állítják, hogy a kvalitatív/kvantitatív megkülönböztetés felszínes, és az a tévhit motiválja, hogy a jelentések vizsgálata teljesen más, mint a szavak vizsgálata.

A következő fejezetekben a kortárs társadalomkutatásban használt két nagy NLP módszercsalád, a felügyelt és felügyelet nélküli gépi tanulás néhány konkrét eszközét tárgyalom, röviden ismertetve őket és saját esettanulmányokon illusztrálva szociológiai alkalmazásukat. Nem célozván kézikönyv-szerű teljességet, nem térek ki például a network-alapú eszközök vagy a nyelvi modellek tárgyalására. Ezekről és szociológiai felhasználásukról jó összefoglalót ad Macanovic (2022).

3. FELÜGYELT ÉS FELÜGYELET NÉLKÜLI TANULÁS

Az alábbiakban röviden ismertetem a felügyelt és felügyelet nélküli gépi tanulás logikáját, majd a két megközelítés gyakorlati kivitelezését és alkalmazását mutatom meg, néhány saját esettanulmányon is. Nem térek ki a klasszikus gépi tanulás harmadik nagy típusára, a megerősítő tanulásra (*reinforcement learning*), mert bár az üzleti alkalmazásokban gyakori és jelentősége az utóbbi néhány évben egyre nő, társadalomtudományi munkákban még ritkábban találkozunk vele. A megerősítő tanulás lényege, hogy nem egyetlen döntést, hanem döntések sorozatát kell meghoznia a gépnek (pl. sakkozni tanítjuk az algoritmust). A tanuló ágens képes érzékelni és interpretálni környezetét, hibás válasz esetén negatív, jó válasz esetén pozitív jelzést („jutalmat") kap, induktív módon optimalizálva működését (lásd 3. ábra).

Szintén nem részletezem a mesterséges intelligenciának a klasszikus gépi tanuláson kívül eső új rendszert, a generatív mesterséges intelligenciát. Ennek új vonása elsősorban a céljában fogható meg: míg a fenti három klasszikus technika mintázatok felismerésére és előrejelzésre fókuszál, a generatív MI új tartalmak létrehozására (=generálására), ami ugyanakkor utánozza a betanult adatokat, lásd pl. a ChatGPT-t. A felügyelt és a megerősítő tanulás fontos szerepet játszik a generatív MI rendszerek, mint a ChatGPT létrehozásában is, hisz azok, miután hatalmas adatbázisokon, felügyelt módon előtanulnak, utána a finomhangolást megerősítő tanulással végzik, ahol a jutalmazást ember által adott visszajelzés helyettesíti, hogy az algoritmus képes legyen megragadni az emberi preferenciák finomságait.



3. ábra. A klasszikus gépi tanulás típusai.

Hasonlóan nem térek ki a gépi tanulási módszer szöveganalitikai jellegzetességeire, arra, hogy mi a „szöveg", mint input adat sajátossága a numerikus inputtal szemben, vagy hogy milyen módon keresnek a modellek mintázatokat a

szövegekben. Utóbbi kérdések iránt érdeklődő olvasóknak Németh és Koltai (2021), illetve Németh, Katona és Kmetty (2020) szociológusoknak szóló szövegbányászati bevezetőjét ajánlom.

A felügyelt és felügyelet nélküli tanulás mögött álló egzakt statisztikaelméletet Vapnik (2000) alaposan tárgyalja. Intuitíve, a felügyelt és felügyelet nélküli tanulás közötti különbség azon alapszik, hogy már létező elmélet/meglevő háttérismeret empirikus megnyilvánulásait keressük (felügyelt tanulás, ahol a „felügyelet" maga a háttérelmélet), vagy induktív módon egy még nem vizsgált téma feltárása a cél (felügyelet nélküli tanulás). A dichotómia hasonlít a klasszikus szociológiai módszertan konfirmatív/exploratív különbségtételéhez.

A felügyelet nélküli módszerek nem igényelnek előzetes feltételezéseket vagy külső ismereteket, a modell maga tanulja meg az adatok szerkezetét mintázatok keresésével. Ezzel szemben a felügyelt módszerek előzetes címkézett adathalmazt igényelnek (például egy politikai internetes fórumon található hozzászólások kategorizálásánál a címkék lehetnek "liberális" és "konzervatív", szakértői kódolás útján létrehozva). A felügyelt tanulási modellek célja, hogy megtanulják, hogyan rendeljék ezeket a címkéket a szövegekhez (a példánál maradva: hogyan automatizálható a liberális/konzervatív címkézés a korábbi szakértői kódolás alapján létrehozott csoportok szövegmintázatai alapján).

A felügyelet nélküli, szövegekre alkalmazható módszerek némelyikét a klasszikus társadalomkutatásban is gyakran alkalmazzák, ilyen pl. a klaszterelemzés. Ez legegyszerűbb esetben egy olyan vektortérben definiálható, ahol a tengelyek a szavakat jelölik, a dokumentumok vektortérbeli helyét pedig az egyes szavak dokumentumbeli gyakorisága határozza meg. De nem felügyelt módszer a nagyon népszerű topikmodell is, ami látens tematikus struktúrák létrehozására alkalmas (*topic modelling*, több változata létezik, egy szép hazai alkalmazás Barna és Knap, 2023 tollából, a Trianon-emlékévben megjelent online cikkek diskurzusainak vizsgálata), vagy a legújabb szövegbányászati megoldások alapját adó, a szavak használata alapján azok jelentésének megragadását lehetővé tevő szóbeágyazás is (*word embedding*, egy érdekes hazai alkalmazás Szabó és társai, 2020, a Pártállam c. folyóirat alapján azonosítva kulcs-kifejezések jelentésének változását). E két ismert NLP-eszközt *A felügyelet nélküli gépi tanulás* c. fejezetben, saját esettanulmányaim kapcsán bővebben is ismertetem.

Ezzel szemben a felügyelt tanulás lényege, hogy (elméletünk/háttérismeretünkre támaszkodva) előre bekódoljuk a szövegeket, majd ezeket a címkéket az algoritmus megpróbálja megtanulni. A kutató elméleti megfontolásai

befolyásolják az elemzést, hiszen a címkék, mint kategóriák meghatározása megelőzi az elemzést. Az 1. táblázat néhány elterjedt alkalmazás esetén mutatja a felügyelt tanulás logikáját: mint látható, az input nagyon változatos lehet, nem csak szöveg, de kép, vagy bármilyen más adat – ezeket ugyanúgy numerikus adattá transzformálják első lépésben, mint a szöveget. A lényeg, hogy ezeknek egy címkézett halmaza kell, hogy rendelkezésre álljon a tanítás során (spam/nem spam email-ek, pozitív/negatív termékértékelések stb.), ami alapján mintázatok azonosítása után az algoritmus képes már automatikusan címkézni új inputokat is.

Input	*output*	*alkalmazási terület*
Email	spam? i/n	spam-szűrés
termékértékelések	érzelmi töltet +/-	brand elismertség monitorozás
hozzászólás	gyűlöletbeszéd? i/n	közösségi média moderálás
radarkép	más autók pozíciója	önvezető autó
reklám, felhasználó infó	klikkel-e? i/n	online marketing
röntgenkép	rosszindulatú-e? i/n	egészségügyi diagnosztika
banki ügyfél infó	szerződés megszűnik-e? i/n	banki churn elemzés

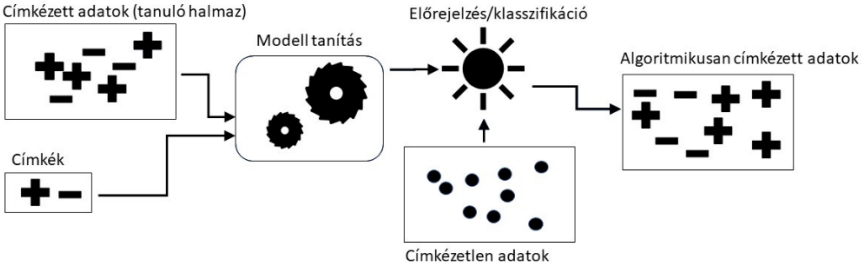
1. táblázat. A felügyelt tanulás néhány alkalmazási területe.

A táblázat is tartalmazza a gyűlöletbeszéd-detektálás problémáját. Egy konkrét, kapcsolódó társadalomkutatási példa Poletti és társai (2017) munkája, akik olasz nyelvű Twitter-üzeneteken igyekeztek automatikus gyűlöletbeszéd-felismerőt létrehozni. A gyűlöletbeszéd általuk alkalmazott definíció szerint valamely kisebbségi csoport ellen irányul és tartalmaz egyfajta illokúciós erőt, amely alkalmas a célcsoporttal szembeni erőszak terjesztésére, népszerűsítésére, alátámasztására vagy erre történő uszításra. Ezért a fogalom pontos megragadása érdekében a kódnak tartalmaznia kellett a célcsoportot (vallási, etnikai kisebbségek vagy migránsok), és a kódolók a tweetet olyan jegyekkel is felruházták, mint hogy sztereotipizál-e, tartalmaz-e agressziót, támadó-e, vagy tartalmaz-e az agressziót kendőző iróniát.

A megtanulás itt azt jelenti, hogy (viszonylag alacsony hiba mellett) az algoritmus maga is képes lesz kódokat (gyűlöletbeszéd / nem gyűlöletbeszéd) rendelni még címkézetlen szövegekhez. A tanulási folyamat pedig a kutatók által címkézett két szöveghalmaz eltérését leginkább megfogó jellegzetes szövegmintázatok keresésén alapul, ahol az előre címkézett szövegek halmazát tanuló-halmaznak (*training set*) nevezzük, lásd a folyamat logikáját a 4. ábrán. (A tanulási

folyamat valójában ennél komplexebb, a tanuló halmazon kívül szükséges egy validációs halmaz elkülönítése is a túlillesztés minimalizálása érdekében, lásd Németh, Katona, Kmetty, 2020, de ez technikai részlet, és e könyvben inkább a szociológiai alkalmazhatóságra koncentrálok).

Nagyon leegyszerűsített példán: a tanítás során használhatunk logisztikus regressziót, ahol a függő változó a bináris címke (gyűlöletbeszéd / nem gyűlöletbeszéd), numerikus magyarázó változóink pedig azt jelölik, hogy az adott nyelv szavai hányszor fordulnak elő a szövegben; a cél a legjobban illeszkedő modellhez tartozó együtthatók megtalálása. Ezután ezt a legjobbnak talált, optimális modellt alkalmazzák A gyakorlatban persze ennél jóval komplexebb modelleket alkalmaznak, komplexebb (például stiláris jegyeket reprezentáló vagy a szöveg mélyebb szemantikai viszonyait megragadó) magyarázó változókkal, a nagy mennyiségű (több ezer) magyarázó változó miatt dimenziócsökkentő megoldásokkal és a túlillesztést elkerülni hivatott megoldásokkal, de az algoritmus létrejöttének logikája ott is hasonló. A legismertebb modellek közé tartozik a regresszió-alapú modelleken kívül a Naive Bayes, a support vector machine (SVM), a random forest és a neurális hálók (bővebben: Eisenstein, 2019).



4. ábra. A felügyelt gépi tanulás logikája.

A felügyelt módszerek abban támogatják tehát a kutatókat, hogy (1) nagyobb szövegkorpuszt dolgozhassanak fel, mint amire emberi kapacitás képes lenne. További (2) cél lehet, hogy kódolásunkat kiterjesszük egy már bekódolt kisebb szöveghalmazról a teljes korpuszra. Ám a legfontosabb (inkább tudományos elvárásokhoz igazodó) felhasználási lehetőség az, hogy (3) megértsük az automatikus címkézés mögött álló szabályokat, egyszerű példán: hogy lássuk, mely szóhasználat valószínűsíti leginkább a gyűlöletbeszéd jelenlétét. Ez utóbbi cél tulajdonképpen azoknak a tartalmi-szociológiai interpretációra / magyarázatra lehetőséget adó jegyeknek a feltárását jelenti, amelyek az egyébként "fekete

dobozként" működő kódoló algoritmus mögött állnak (erről az interpretációs/magyarázati igényről bővebben, néhány saját esettanulmánnyal az interpretáció megragadására *A felügyelt gépi tanulás* c. fejezetben). A felügyelt tanulás általános tudományos célokat teljesítő felhasználását tárgyalja Molnár és társa (2024) munkája, kitérve minden fontos dimenzióra, közöttük a szakterületi tudás ('domain knowledge'), az interpretálhatóság, az okság vagy a reprodukálhatóság szerepére.

A felügyelt tanulásnak több új továbbfejlesztése létezik, ahol az eredeti logika kissé módosul. Így pl. az aktív tanulás esetén nem egy fix tanuló-halmazunk van, hanem az algoritmus maga kéri menet közben konkrét, még címkézetlen, de a tanulásban fontosnak tűnő szövegek címkézését, vagy a transzfer tanulás, amikor a címkézett adatoktól eltérő besorolási problémát kell a tanuló algoritmusnak megoldania (Eisenstein, 2019).

A felügyelet nélküli módszerek a felügyelt technikákkal együtt is használatosak. Például a felügyelet nélküli módszerek hatékony alkalmazása, ahogy fentebb említettük, gyakran dimenziócsökkentést igényel, s ezek a dimenziócsökkentési eljárások gyakran felügyelet nélküliek. Így egy felügyelt tanulási regressziós modellben magyarázó változóként használhatjuk a klaszterelemzés által adott besorolást. De ugyanígy gyakori megoldás az, amikor felügyelt modellben a szöveget alkotó szavak helyett a szóbeágyazási modell által adott szóreprezentációkat használják magyarázó változóként.

Bár saját esettanulmányt nem mutatok majd hozzá, mégis megemlítem itt a szentiment- és érzelem-elemzést, mint nagyon gyakori NLP-eszközöket, mert jól illusztrálják, hogy ugyanazon célt különböző gépi tanulási logika mentén is megvalósíthatjuk. Egy szöveg szentimentje a szerző attitűdje egy tárgyhoz (pozitív, negatív vagy semleges), míg az érzelmek a boldogságtól a haragig terjedő érzések[7]. Ez az alkalmazás mind üzleti, mind tudományos alkalmazásokban nagyon elterjedt.

A szentiment- és érzelem-elemzés végezhető felügyelet nélküli és felügyelt megközelítések alkalmazásával is. Felügyelt tanulásnál a szöveges adatok szentiment- vagy érzelem szerinti, kódolók által végzett előzetes címkézésével vizsgálhatják pl. marketingesek, hogy hogyan reagálnak a felhasználók reklámokra, szolgáltatásokra vagy termékekre, digitális bölcsészek, hogy hogyan változnak a regényben megjelenített érzelmek, vagy szociológusok, hogy hogyan

[7] Összefoglaló az alkalmazott módszerekről: Nahili, Rezeg, Kazar, 2020.

terjednek érzelmek és vélemények társadalmi hálózatokban. Felügyelt szentimentelemzésre példa hazai szociológiai kutatásban Üveges és Ring (2023) politikai szövegekre, ill. Knap és társai (2023) online média-cikkekre alapozott írása.

A felügyelet nélküli szentiment- és érzelemelemző megközelítések például a különböző állapotokhoz (pl. negatív / pozitív) tartozó szavak szótárát használhatják. A szótár-alapú módszerek automatikusan azonosítják a szótárban szereplő szavakat a szövegben, és ezért különösen hasznosak az adott fogalom nagy mennyiségű szövegben való jelenlétének azonosítására és számszerűsítésére. Az egyik legjobban kidolgozott angol nyelvű szótár az LIWC (Linguistic Word County and Inquiry, Pennebaker et al., 2015). A LIWC azonosítja a szövegben az olyan előre meghatározott kategóriákhoz tartozó szavakat, mint például az alapvető érzelmek (szomorúság, harag). Egy felügyelet nélküli, szótár-alapú szentimentelemző platform a „boldogságmérő" (lásd hedonometer.org), mellyel a Vermonti Egyetem kutatói szövegek, pl. Twitter-szövegek boldogság-szintjét mérik alkotószavaik boldogságtartalmának átlaga alapján (lásd a hedonometer.org tudományos publikációit). Az egyes szavak boldogságtartalma a kontextustól függetlenül adott, tesztalanyok korábbi pontozása alapján, pl. a *hope, hero, to win* magas pontszámmal rendelkeznek. A hedonometer kutatói szerint ezzel populációk, földrajzi régiók aktuális boldogságát objektívebben tudjuk mérni, mint a hagyományos, szubjektív elégedettségre rákérdező survey-ekkel (az eljárás elvi korlátairól lásd Németh, 2015a).

Érdemes megjegyezni, hogy szótár alapú szövegelemző módszerek nem csak az érzelem- és szentiment-detektálás területén használatosak, hanem bárhol, ahol a vizsgálandó fogalomhoz kapcsolódó szavaknak eléggé átfogó listája összeállítható. Ilyen terület például a politikatudományé: Karamshuk et al. (2016) a 2013-14-es ukrán-orosz konfliktussal kapcsolatos, a médiában és a közösségi médiában megjelenő politikai polarizáció vizsgálatához állított össze egy olyan szakértői szótárt, melynek elemei a pártos retorika indikátorainak tekinthető. Decadri és Boussalis (2020) az olasz képviselőházban elhangzott beszédek populizmusát vizsgálva állította össze szintén szakértői alapon a populista retorika szótárát.

A szótár alapú módszerekkel kapcsolatos legfontosabb korlát, hogy azon a területen működnek jól, ahol a szótárt létrehozták, vagyis domain-specifikusak. Karamshuk et al. (2016) példáján: a média, a laikus közösségi média vagy a Parlament annyira eltérő kontextusok (eltérő beszélők, eltérő célközönség, eltérő műfajok), hogy a politikai megosztottságot jelző szavak szótára is csak részben fed át. Ugyanez igaz a felügyelt tanuló módszerekre is: hatékonyságuk csak

azon a domain-en garantált, ahol fejlesztették őket; de a felügyelt algoritmusok esetén ez a tulajdonság természetszerű, ezért explicit.

Összefoglalásként: a felügyelet nélküli tanulás a szöveg látens jellemzőit fedi fel anélkül, hogy a kutató explicit módon előre meghatározná az érdeklődésre számot tartó jellemzőket. Ezek a módszerek azért értékesek, mert elméletileg hasznos, de esetleg nem eléggé kutatott vagy korábban ismeretlen jellemzőket is felfedhetnek (lásd a következő fejezetben pl. saját alkalmazási példánkat, az online depressziós fórumok topikmodell által detektált látens tematikus struktúráját). Vannak szerzők (pl. Quinn et al., 2010), akik a felügyelet nélküli és a felügyelt módszereket egymással versengő módszereknek tekintik, és a felügyelet nélküli megközelítést előnyösebbnek tartják, mert kevesebb előzetes feltevéssel él. Ez a szembeállítás véleményem szerint értelmetlen: a felügyelt és a felügyelet nélküli módszerek különböző modellek, különböző kutatási kérdések megválaszolására. Ha vannak előre meghatározott kategóriák, akkor felügyelt tanulási módszert kell használnunk, ha exploratív jellegű a kutatás, akkor felügyelet nélkülit. Ezek a módszerek nem vetélytársak, hanem egymást kiegészítő módszerek, a következő két fejezetben látni fogjuk, hogy egyaránt inspiratív alkalmazási lehetőségekkel.

4. A FELÜGYELET NÉLKÜLI GÉPI TANULÁS

Ebben a fejezetben saját esettanulmányokon keresztül két felügyelet nélküli módszert mutatok be, a topikmodellt (több variánsával) és a szóbeágyazást.

4.1. Topikmodell

4.1.1. A topikmodell általában

A topikmodell célja, hogy látens témákat (NLP-terminussal „topikokat") azonosítson korpuszunkban. A „látens témák" statisztikailag a szótár szavai feletti valószínűség-eloszlások, és egy adott topikot a benne leggyakrabban előforduló kifejezésekkel tudunk azonosítani. A modell alapgondolata abban a disztribúciós szemantikai megközelítésben gyökerezik (Firth, 1957), amely szerint az azonos kontextusban előforduló szavak általában hasonló jelentéssel bírnak. Például az olyan szavak együttes előfordulása egy online, mentális zavarokkal foglalkozó beszélgetőfórum hozzászólásaiban, mint a *blood, thyroid, doctor, diagnose*, a zavar biomedikális megközelítésére utal. Ezért a topikmodellező algoritmusok a jelentéssel kapcsolatban relációs megközelítést alkalmaznak, abban az értelemben, hogy a kifejezések együttes előfordulása fontos a jelentésük és a topikok jelentésének meghatározásában. (A relacionalitás miatt a modell képes megragadni pl. a poliszémiát is: a szó különböző felhasználási kontextusai alapján különböző topikokba fog kerülni). Ezen tulajdonsága miatt a modell éles ellentétben áll azokkal a szótár-alapú megközelítésekkel, amelyek előre definiált szótár elemeinek önmagukban vett gyakoriságát elemzik a szövegekben.

Hogy a modell statisztikai feltevései a gyakorlatban milyen témák azonosítását teszik lehetővé, a példák kapcsán látjuk majd, de már itt érdemes megjegyezni, hogy bizonyos szavak magasabb/alacsonyabb előfordulási gyakorisága nem csak a szöveg szűk értelemben vett témájától, hanem megformáltságától, műfajától, kontextusától is függ; a „látens témák" tehát ezen szempontok szerint is különbözhetnek egymástól. Ezért is (és a kifejezés magyar nyelvű szakkifejezésként történt rögzülése miatt is) használom a „topik", és nem a „téma" kifejezést.

A modell feltevései közé tartozik, hogy véges számú topik létezik, ezek „generálják" a szövegeket. A szövegek, mint topikok keverékei azonosíthatók, ahol adott szöveghez lehetőleg kevés topik, és adott topikhoz is lehetőleg kevés jellemző

szó tartozik. A modell paramétereinek becslésével meg lehet találni ezeket a topikokat (és a hozzájuk tartozó leggyakoribb, tehát a topikokat jellemző szavakat), és meg lehet becsülni azt is, hogy egy adott szöveg milyen mértékben tartozik az egyes topikokhoz. A modell által becsült topikok nem címkézettek, így a kutatónak kell ezeket a címkéket hozzárendelnie az egyes topikokhoz a bennük szereplő leggyakoribb szavak értelmezésével, illetve a topikokhoz leginkább kapcsolódó szövegek kvalitatív feldolgozásával. Ideális esetben az adott területet ismerő kutató számára kézenfekvő az értelemzés, a topikok magukért beszélnek.

Számos különböző topikmodellezési technika létezik, amelyek statisztikai feltételezéseikben különböznek egymástól; elsőként a történetileg első és talán leggyakrabban alkalmazott LDA-ra (*Latent Dirichlet Allocation*, Blei et al., 2003) mutatok esettanulmányt. Kivételes módon, ez a modell nem az üzleti életből, hanem a tudományos, mégpedig a bölcsészettudományos alkalmazásokból indulva lett népszerű. Már létrehozói, Blei és társai is a bölcsészek figyelmébe ajánlották a módszert, és valóban, nagyon korán megjelentek alkalmazásai ezen a területen, pl. történészek történeti folyóirat-elemzésre vagy napló-elemzésre, irodalmárok költői szövegek elemzésére használták. A Journal of Digital Humanities már 2012-ben különszámot szentelt a módszernek (Meeks & Weingart, 2012).

A modell illesztéséhez a dokumentum-kifejezés mátrixon kívül (a mátrixról lásd korábban *Az NLP mint kutatási módszer társadalomtudományi kontextusa* c. fejezetet) nincs szükség emberi beavatkozásra, egy kritikus kivétellel: a topikok számát előre meg kell határozni. A topikmodellek illesztése és értelmezése során ezért az egyik fő szempont a témák "helyes" vagy inkább „optimális” számának meghatározása. Ez a döntés nagyjából annyira alátámasztható objektíven, mint klaszterelemzés esetén a klaszterek számnak meghatározása. Túl alacsony topikszám esetén túl általános témákhoz jutunk, túl nagy szám esetén elaprózott, redundáns témákhoz.

Léteznek statisztikai mérőszámok is a döntés támogatására, például a perplexitás, amely azt vizsgálja, hogy milyen sikeresen prediktálja a modell az új adatokat, vagy különböző koherencia-mutatók, melyek adott topik legrelevánsabb szavai közötti szemantikai hasonlóságot próbálják különböző módon számszerűsíteni. Az optimális topikszámot ezután a különböző topikszámú modellekhez számolt mérőszámok maximuma (vagy inkább lokális maximuma) fogja kijelölni. De a statisztikai mutatóknál gyakran jobb eredményt ad a topikmodell interpretálhatóságának kvalitatív értékelése. A gyakorlatban a topikok kiválasztása így bizonyos fokú önkényességet/szubjektivitást tartalmaz.

Mielőtt saját esettanulmányomra térnék, néhány, témájában és a használt korpusz jellegében is karakteresen különböző alkalmazás bemutatásával szeretném a módszer termékenységét illusztrálni. Mützel (2015) két projektet ismertet, az első innovatív mellrák-terápiák kapcsán 25 évet átfogva elemzi tudományos publikációk szövegét, míg a Berlin-projektben 20 évre visszamenően elemzi a város éttermeinek étterem-kritikáit. Mindkét esetben az LDA topikmodell a hosszabb időtáv trendjeinek változását segített megtalálni, és arra is lehetőséget adott, hogy egy-egy fordulópontot megtalálva ott kvalitatív módon (hagyományos olvasással) forduljon a kutató a korpuszhoz. Light és Cunningham (2016) a Nobel-békedíj átvételekor elmondott beszédeket vizsgálja, melyek a korábbi (pl. keresztény) sémáktól eltolódva szerintük egyre inkább a globalizációhoz és a neoliberalizmushoz kapcsolódnak. A bölcsész Blevins (2010) a XVIII. században élt Martha Ballard naplójára alkalmazta a módszert. Médiatudományi alkalmazás Jacobi és társai (2016) munkája, akik a New York Times 1945 és 2016 közötti, nukleáris technológiáról szóló cikkeit elemezve arra jutottak, hogy az LDA nagyon hasznos eszköz hatalmas digitális szövegkorpuszok relatíve gyors tartalmi áttekintésére. A politológus Grimmer (2010) amerikai szenátorokat vizsgált abból a szempontból, hogy az általuk megjelenített topikokat mennyire igazítják a médiában párhuzamosan megjelenő témákhoz. Modellje szimultán vizsgálta a médiabeli szövegek és a szenátorok sajtóközleményeinek témáit.

4.1.2. LDA-topikmodell – A depresszió biopszichoszociális modelljének aszimmetriái online fórumokon

Szerzőtársaimmal ebben a kutatásunkban (Németh et al., 2021) a depresszió egyik legátfogóbb megközelítésének, a biopszichoszociális modellnek (Bolton & Gillett, 2019) a mintázatainak feltérképezése érdekében elemeztünk online depressziós fórumokat. Erre a biopszichoszociális modellre alapozva bírálták a társadalomtudományok a depressziónak azt a redukcionista orvosbiológiai megközelítését, amely hosszú ideig uralta a szakértői diskurzusokat. Mivel ezek a diskurzusok határozzák meg mind a mentális zavar okának keresését, mind a lehetséges megoldásokat, laikus értelmezésük központi szerepet játszik a depresszióval való megküzdésben. A „betegség- és gyógyulási narratívák” jelentősége jól ismert, és különösen a mentális zavarok esetében fontos az utóbbiak szerepe a megújult identitás megkonstruálásához. Ebben a folyamatban a társak támogatása nélkülözhetetlen, ők biztosítanak platformot az identitáskonstrukciós interakciókhoz

(Pfeiffer et al., 2011). Ezek a megfontolások adták kutatásunk tétjét: a depresszióról szóló laikus narratívák elemzésével tettünk kísérletet az online fórumok támogatási lehetőségeinek és korlátainak feltérképezésére.

A korpusz

A legnépszerűbb angol nyelvű online egészségügyi fórumokból gyűjtöttünk depresszióval kapcsolatos bejegyzéseket (N=~70 000), és kvalitatív megközelítéssel kiegészített LDA topikmodellt alkalmaztunk látens tematikus struktúrájuk feltárására. Arra kerestünk választ, hogy általában milyen tematikus klaszterek és diszkurzív minták jelennek meg a fórumokon, illetve, hogy a bio-pszichoszociális modell mely dimenziói játszanak domináns szerepet ezekben a diskurzusokban, milyen funkcióban. Az alábbiakban a kutatás módszertani tanulságaira fókuszálok, a tartalmi eredményekre csak ezek illusztrációjaként utalok.

A vizsgálat célpopulációjáról, az online depressziós fórumokról érdemes megjegyezni, hogy a felhasználók heterogén populációt alkotnak. Többségük jelenleg vagy korábban depresszióban szenvedett, más részük csak közvetve érintett (pl. családtag) vagy kíváncsi információkeresők (Nimrod, 2013). Ez egyben jelzi eredményeink általánosíthatósági korlátját is.

Az online fórumbejegyzések gyűjtéséhez a SentiOne-t, egy webalapú social listening és szövegelemző platformot használtuk. A legnépszerűbb angol nyelvű online egészségügyi fórumokat a Google kereső segítségével választottunk ki a "*depression forum*" és a "*depression online*" keresőkifejezésekkel. Ezt a keresési stratégiát az is indokolta, hogy egy, a depresszióval kapcsolatos aggodalmaira online válaszokat kereső felhasználó is hasonló keresést végezne. Mivel a Google-keresés nemcsak a legnagyobb, hanem a legelérhetőbb oldalakat is meghatározza, eredményei egyben a legszélesebb körben használtaknak is tekinthetők. A keresést a 2016-2019 között aktív, nyilvános és regisztráció nélkül elérhető fórumokra korlátoztuk. A SentiOne által a GDPR-előírásoknak megfelelően gyűjtött adathalmaz 79 889, 2016. február 15. és 2019. február 15. között közzétett bejegyzést tartalmazott, amelyek csak a nyilvánosan elérhető, a szerzők által önkéntesen megosztott posztokra terjedtek ki.

Célunk az volt, hogy csak olyan hozzászólásokat gyűjtsünk össze, amelyek kifejezetten a depressziót tárgyalják; ehhez (1) kiválasztottuk azokat a beszélgetés-folyamokat (*thread*-eket), amelyek címében vagy legalább az egyik hozzászólásában szerepelt a *depression* vagy *depressed* szó, majd (2) kiválasztottuk

azokat a hozzászólásokat, amelyek linkje, témája vagy tartalma tartalmazott egy depresszióval kapcsolatos kifejezést, például *unipolar depression* vagy *mood disorder*. A duplikált és túl rövid (20 szónál rövidebb) bejegyzések eltávolítása után a végleges korpuszunk 67 857 bejegyzést tartalmazott. A posztokat ~20 000 felhasználó írta.

A korpusz nyers formában további előfeldolgozást igényelt ahhoz, hogy elemzési célokra használható legyen (erről és az alkalmazott szózsák-modellről lásd *Az NLP mint kutatási módszer társadalomtudományi kontextusa* c. fejezetet). Ehhez a Python NLTK csomagját (Bird et al., 2009) használtuk. Lemmatizálást alkalmaztunk ugyanazon szó különböző ragozott formáinak egységesítésére, elvetettük az írásjeleket és a nagybetűket, töröltük az URL-címeket, az e-mail címeket, a korábbi hozzászólások (az azokra született válasz miatt) újraposztolt részeit és a nagyon gyakori szavakat („stopszavakat"). A szózsák-modellt korlátait tágítandó, a leggyakrabban együtt járó kétszavas kollokációkat egyetlen kifejezésként kezeltük, ilyen volt pl. a posztokban gyakran szereplő „*frontal lobe*" (homloklebeny – épp ezidőtájt, a 2019-es években merült fel ennek az egyterületnek a pszichopatológiai érintettsége depresszió esetén). Hasonlóképp egyetlen kifejezésként kezeltük a leggyakoribb mentális rendellenességek elnevezését és a tulajdonneveket.

Végül a szavak szövegeken belüli eloszlása képezte a kvantitatív elemzés numerikus bemeneti adatait (dokumentum-kifejezés mátrix alakjában), vagyis a korpuszt numerikus adatbázissá alakították át. Az elemzés bizonyos pontjain visszakanyarodtunk az érintett szövegekhez, és azokat kvalitatív módon elemeztük, kevert módszertant követve.

Módszertan

Az LDA-t a MALLET programon (McCallum, 2002) keresztül használtuk, amely Python programnyelven érhető el a gensim eszközkészlet segítségével. A topikmodelleket úgy futtattuk, hogy a topikok számát 5 és 20 között változtattuk, igyekezve nem túl széles, de mégsem túlságosan széttagolt topikokat kapni. Az algoritmus implementációja sztochasztikus elemekre támaszkodik az inicializálás során, ami némileg eltérő eredményekhez vezethet. Emiatt az instabilitás miatt minden modellt öt különböző véletlenszám-inicializálással futtattuk.

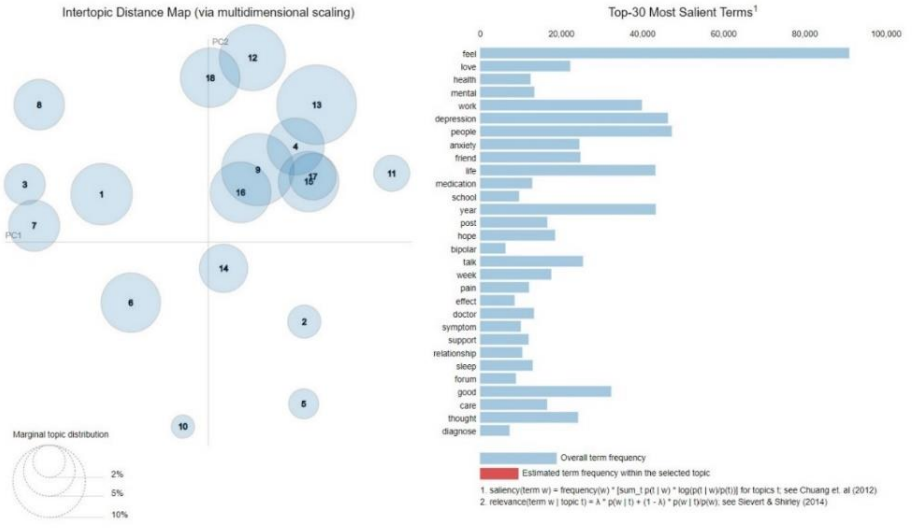
A témák optimális számának meghatározásakor először minden modellre kiszámítottunk egy koherencia-pontszámot. A több elérhető koherencia-mutató

közül a C_v mérőszámot választottuk, ami a tapasztalatok szerint teljesítményével felülmúlja az összes többi koherencia-pontszámot, ahol a teljesítményt az emberi értékelésekkel való korrelációval mérték (Röder et al., 2015). A legnagyobb koherencia-értékű (és a hozzá tartozó függvényen lokális maximumként szereplő) modelleket választottuk ki az 5-20 topikos modellek 5-féle inicializálása közül. Az így kapott (7, 13, 14, 18 és 19-es topikszámú) modelleket ezután kvalitatív módon rangsoroltuk értelmezhetőségük alapján, illetve szisztematikusan összehasonlítottuk az általuk reprezentált témastruktúrát. Az utóbbi elemzés eredménye szerint a több topikot tartalmazó modellek a kisebbekből jönnek létre topikok széthasadásával (tehát egyfajta evolúció figyelhető meg), bár némelyik modell tartalmazott értelmezhetetlen, vegyes topikokat is. Végül a 18-as topikszám bizonyult a legjobbnak az értelmezés szempontjából: nem terhelte értelmezhetetlen topik, ugyanakkor jól differenciált témákat adott. A modellek evolúciószerű egymásba-fejlődése is megerősítette a végső, 18-as modell robusztusságát.

Az értelmezés támogatása vizualizációval

Az értelmezés támogatására az LDAvis interaktív vizualizációs eszközt alkalmaztuk, itt az interaktív vizualizáció egy-egy statikus változatát mutatom be[8]. Ben Mabey pyLDAvis Python-csomagját használtuk, amely a Sievert és Shirley (2014) eredeti R implementációjának Python-beli implementációja. Az 5. ábra a topikok globális áttekintését mutatja be.

[8] Az interaktív ábra zippelt html-fileként letölthető és kipróbálható az eredeti publikációnk mellékleteként a következő linkről, érdemes kipróbálni és önálló felfedezéseket tenni: https://ars.els-cdn.com/content/image/1-s2.0-S2352827321000604-mmc1.zip



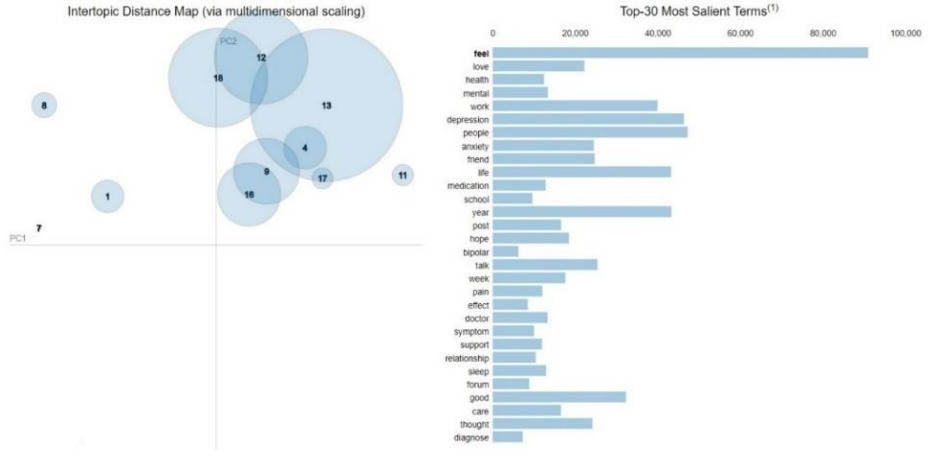
5. ábra. A topikok távolságtérképe a legrelevánsabb kifejezésekkel.

A bal oldali panel az egyes topikok korpuszbeli gyakoriságát és a topikok közötti távolságot szemlélteti. A körök területe arányos a topikok gyakoriságával (elterjedtségével) a korpuszban, ahol a gyakoriságot a korpusz adott topikhoz tartozó szavainak százalékos arányával mérjük. Az ábra egy többdimenziós skálázás kétdimenziós eredménye (a topikok páronkénti távolsága eredetileg Jensen-Shannon-divergenciával van kiszámítva, majd többdimenziós skálázás segítségével vannak kétdimenziós síkra vetítve). Egy ilyen térkép értelmezésekor a tengelyek értelmezése nem egyértelmű. Egy lehetséges megközelítés, ha nagyon távoli objektumokat veszünk, és megpróbálunk értelmezést találni a dimenziókra. Fontos megjegyezni, hogy a kétdimenziós térkép óhatatlanul leegyszerűsíti a képet, és értelmezését ki kell egészíteni kvalitatív értékeléssel is, hogy érvényesebb betekintést nyerjünk.

Az 5. ábra jobb oldali panelje egy vízszintes oszlopdiagramot mutat, amelynek oszlopai a topikmodell megértése szempontjából leginkább informatív kifejezéseket ábrázolják. Az informativitás mutatója a *saliency* (Chuang et al., 2012), ami azt méri, hogy egy kifejezés mennyi információt közvetít a témákról. Egy kifejezés *saliency* értékét a hozzá tartozó relatív gyakoriság és a megkülönböztethetőség szorzataként definiáljuk. Pl. hiába fordul elő egy szó nagyon gyakran a korpuszban, ha minden topikban megtalálható (azaz alacsony a megkülönböztethetősége): ilyenkor a szó előfordulása egy hozzászólásban nem segít annak meghatározásában, hogy a hozzászólás melyik topikhoz tartozik. Az ábra a 30 leginformatívabb kifejezést

mutatja be a jobb oldali panelen, a kék sávok a hozzájuk tartozó általános gyakoriságot jelölik. A 30-as lista jól jellemzi korpuszunk (a következőkben még részletezett) tematizáltságát: felbukkan a társas (*friend, school*), a medikális (*medication, doctor*), a pszichológiai dimenzió (*feel*, *love*), a fájdalom önkifejezése (*pain*) és a fórum interaktív, társas támogató oldala (*support, post, hope*) is.

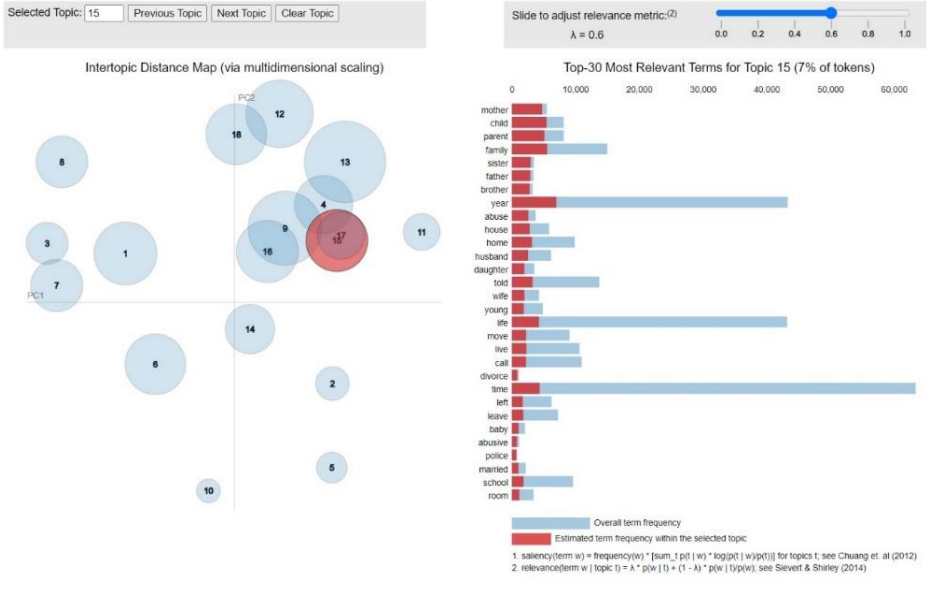
A topikok értelmezését támogatja a jobb oldalon felsorolt, leginformatívabb kifejezések topikbeli előfordulásának összevetése. Az interaktív vizualizáció erre is lehetőséget ad, a 6. ábra a *feel* kifejezés topikonkénti előfordulását szemlélteti. Látjuk, hogy a 13. topikban fordul elő a leggyakrabban, ami (egyéb információk mellett) segített abban, hogy ezt a topikot az „expresszív beszédaktusok" címkével lássuk el.



6. ábra. A „feel" kifejezés topikonkénti gyakorisága.

Egy másik fontos interaktív vizuális elem további segítséget nyújt a topikok interpretálásában. Ha bal oldalon kiválasztunk egy topikot, a jobb oldalon megjelennek a hozzá tartozó legfontosabb kifejezések. A 7. ábra a modell 15. topikjának legrelevánsabb kifejezéseit mutatja, relevanciasorrendben. Ezek a kifejezések is támogatták azt a döntésünket, hogy ezt a topikot a családdal kapcsolatos attribúcióként azonosítottuk, mivel a legrelevánsabb kifejezések a konstelláció legfontosabb szereplőit tartalmazzák (*father, mother, sister, brother*…), valamint a potenciális konfliktuspontokat (*abuse, divorce, leave*…). Egy kifejezés relevanciája (Sievert & Shirley, 2014) a topikra jellemző gyakoriságának és egy büntető tagnak az összege, ahol a büntető tag a kifejezés általános gyakoriságának

növekvő függvénye. Nyilvánvalóan azok a kifejezések, amelyek az adott topikban gyakran fordulnak elő, de a teljes korpuszban nagyon gyakoriak, kevésbé relevánsak a topik szempontjából. Az összeg egy súlyozott összeg, λ és (1-λ) súlyok használatával (λ interaktívan állítható 0 és 1 között). Minél nagyobb a λ, annál kisebb büntetés kerül bevezetésre. Ha λ egyenlő 1-gyel, akkor nem alkalmazunk büntetést, és a relevancia egyszerűen a topik-specifikus gyakoriságként van meghatározva. Mi a λ értékét 0,6-ra állítottuk be, ami jól értelmezhető eredményeket ad, és amit Sievert és Shirley (2014) is optimálisnak talált.



7. ábra. A 15. topik legrelevánsabb kifejezései
(a relevanciát λ = 0,6 értékkel számoltuk).

A 7. ábra a kifejezések topikspecifikus gyakoriságát (piros sávok) és általános gyakoriságát (kék sávok) is mutatja. Ezek az információk mélyebb megértést nyújtanak a vonatkozó kifejezések szerepéről: ha például a kék és a piros sávok hossza megközelítőleg azonos, akkor a kifejezés kizárólag a témát jellemzi. A 7. ábrán a legrelevánsabb kifejezések közé tartozik a *mother* és a *family*, azonos gyakorisággal a topikon belül, de a teljes korpuszban a *mother* kevésbé gyakori. Ez a következtetés még inkább megerősítést nyer, ha a két kifejezésre kattintunk és így feltárjuk az összes topikra vonatkozó előfordulásukat: eszerint a *mother* specifikusan erre a témára jellemző, de az *family* más topikokban is gyakori.

A topikok értelmezése: a kvantitatív és kvalitatív megközelítés kombinálása

Mivel kutatási kérdéseink összetett diszkurzív mintázatokra irányultak, a fórumbejegyzések értelmének és kommunikációs funkciójának mélyebb megértésére volt szükség. A topikmodellek felhasználói gyakran kifejezetten hivatkoznak a módszerek keverésére, lásd pl. Jacobs és Tschötschel (2019) elméleti oldalról vagy Barna és Knap (2023) egy konkrét empirikus kutatásban.

Topikonként a 30 legrelevánsabb kifejezést és a 10 legrelevánsabb hozzászólást elemeztük. Így a kvalitatív vizsgálatban összesen 180 posztot vizsgáltunk, amelyek szinte mindegyikét különböző szerzők írták. Az adott topikhoz tartozó "legrelevánsabb" posztokat úgy definiáltuk, hogy azok legalább 90%-os kontibúcióval rendelkeztek, vagyis ezek a posztok szinte kizárólag az adott topikhoz tartoztak (emlékeztetőül: a hozzászólásokat több topik együttesen generálja, adott topik szerepe adott hozzászólás létrejöttében jellemezhető a topik-kontribúcióval).

A kvalitatív elemzés során nem csak a posztok tartalmát, hanem azok kommunikációs célját is értelmeztük. Ez a perspektíva különösen termékenynek bizonyult. Valóban, a fórumbejegyzések beszédaktusoknak tekinthetők (Austin, 1975), vagyis maguk is cselekvések. Nem csak kifejeznek valamit a világról vagy önmagukról (ez a lokúciós aktus az eredeti austin-i megközelítésben, itt ilyen lehet egy objektív betegségtörténet-beszámoló), hanem tesznek is valamit a kifejezéssel (ez az illokúciós aktus, itt pl. a szenvedés-beszámoló együttérzést vált ki), és megpróbálnak hatni a másikra is azzal, hogy tesznek valamit (ez a perlokúciós aktus, itt pl. terápiás javaslatok).

Eredmények

Az eredményeket csak a módszertani megközelítés szempontjából ismertetem, arra koncentrálva, ami tanulságként más kutatásokra is átvihető.

A topikok értelemzésekor az első választóvonalat nem a szemantikai különbségek (vagyis a hozzászólás tartalmi tartalma, a beszédaktus "lokutív" dimenziója), hanem a performatív különbségek (vagyis a hozzászólások kommunikatív funkciója, a beszédaktus illokutív és perlokutív dimenziója) szerint húztuk meg. Azokat a hozzászólásokat címkéztük monológokként, amelyek a világ és az én különböző aspektusait tematizálták anélkül, hogy a többieket partnerként akarták volna bevonni az értelmezési folyamatba; azokat a hozzászólásokat pedig interakciókként, amelyek nem a világ vagy az én puszta leírására, hanem a

másokkal való diskurzusra irányultak. E kategóriákon belül két-két altípust különböztettünk meg: a monológok objektívebb attribúciókat (amelyekben a lokutív tartalom dominált) és érzelmileg terhelt önfeltárásokat (amelyekben az illokutív szándék dominált) tartalmaztak; míg az interakciók pragmatikusabb konzultációkat (amelyekben a lokutív tartalom dominált) és kvázi-terápiás elköteleződéseket (amelyekben a perlokutív kísérletek domináltak). Ez a felosztás összekapcsolható volt az 5. ábrával: pl. a monológok - önfeltárások az y-tengely felső részén, míg az interakciók az alsó részén helyezkednek el, így egy lokúciós dimenzióra (performatív funkció) következtethetünk.

A biopszichoszociális modell jelenlétére vonatkozó kutatási kérdésünkre a topikok értelemzésével választ tudtunk adni. Eredményünk szerint a biomedikális diskurzus alulreprezentált a laikus diskurzusokban, míg a pszichológiai diskurzus összességében domináns szerepet játszik. Még ha a szereplők kezdetben tisztában is vannak a depresszióhoz hozzájáruló társadalmi mechanizmusokkal (lásd a családot, párkapcsolatot, munkát reprezentáló topikokat), a gyógyulással kapcsolatos megoldások már elhanyagolják ezeket a tényezőket. Összességében tehát úgy tűnik, hogy a narratíva aszimmetrikus módon támaszkodik a biopszichoszociális modellre.

A fenti alkalmazás bár a konkrét LDA-topikmodellt használta, de a topikmodellezés más variánsai esetére is általánosítható módszertani kihívásokat és megoldásokat szemléltetett. A továbbiakban két, a szociológiai kutatások számára releváns, továbbfejlesztett variánst ismertetek kapcsolódó esettanulmányokkal: a strukturális topikmodellt és a dinamikus topikmodellt.

4.1.3. A strukturális topikmodell

A strukturális topikmodell (*structural topic model*, STM) az LDA topikmodell továbbfejlesztése, melyet társadalomkutatók fejlesztettek ki saját céljaikra (Roberts et al., 2014, 2019). A modell az LDA-val szemben lehetővé teszi, hogy kategorikus változók formájában metaváltozókat adjunk a modellhez, melyek a topikok gyakoriságát és tartalmát magyarázzák. A módszer társadalomtudományi relevanciáját jól szemlélteti, hogy politológus és szociológus létrehozói, Roberts és munkatársai (2014) az eszközt a bevándorlási kérdésekre adott nyílt végű kérdőíves válaszokra alkalmazták, hogy kimutassák pl. a pártpreferenciának a bevándorlástól való félelemmel kapcsolatos topikokra (azok elterjedtségére és keretezésére) gyakorolt hatását. Az újítás így a topikmodellt feltáró eszközből olyan

módszerré változtatta, mely lehetővé teszi alapvető szociológiai jellemzőkkel kapcsolatos összefüggések tesztelését is.

A metaváltozókon belül megadhatjuk, hogy melyek befolyásolják a topikok gyakoriságát (elterjedtségét/népszerűségét) melyek azok tartalmát (keretezését). Egy további különbség az LDA topikmodellel szemben, hogy eltörli annak a topikok korrelálatlanságára vonatkozó megszorítását. Utóbbi feltétel a legtöbb alkalmazásban valóban kontraintuitív, hiszen bizonyos topikok jellemzően felül- vagy alulreprezentáltak más topikok jelenlétében, vagyis korrelálnak egymással a szövegeken belül.

4.1.4. STM topikmodell, idő mint metaváltozó – A COVID-19 hatása az online depressziós fórumok diskurzusaira

Kutatásunk (Németh et al., 2023a) célja az volt, hogy feltárja a COVID-19 világjárvány hatását az előző fejezetben vizsgált online depresszió-fórumokon folyó laikus diskurzusokra. Mivel a pandémia komoly hatással volt a mentális zavarokkal (különösen depresszióval és szorongással) élő emberekre, ezért kutatásunk célja ezen átalakulások diszkurzív szinten megmutatkozó következményeinek feltérképezése volt. Különösen jó empirikus terepként adódott az online tér, hiszen a társadalmi távolságtartás kényszere nemcsak a mentális terhek növekedését eredményezte, hanem a szakértőkkel és más potenciális segítőkkel való személyes találkozást is akadályozta.

Az STM, mint elemzési eszköz használata kézenfekvő módon adódott, hiszen volt már ismeretünk ezen a terepen a topikmodell által megrajzolt általános tematikus struktúrákról (lásd az előző fejezetet), és mivel a pandémia előtti/pandémia alatti időszak összevetése volt a cél, az idő bináris változóként történő megjelenítése a diskurzusok topikjaira ható metaváltozóként az STM-mel lehetővé vált. A modellbe tehát egyetlen metaváltozót vontunk be, az időt, dichotomizálva pandémia előtti vs. pandémiás kategóriákba, 2020.03.11-es osztóponttal, mert az Egészségügyi Világszervezet ekkor nyilvánította a járványt globális világjárvánnyá. Mivel ésszerű feltételezni, hogy bizonyos témák mind gyakoriságukban, mind keretezésükben megváltoztak a világjárvány következtében, a modellspecifikációban megengedtük, hogy az idő mind a témák gyakoriságát, mind a tartalmát befolyásolja.

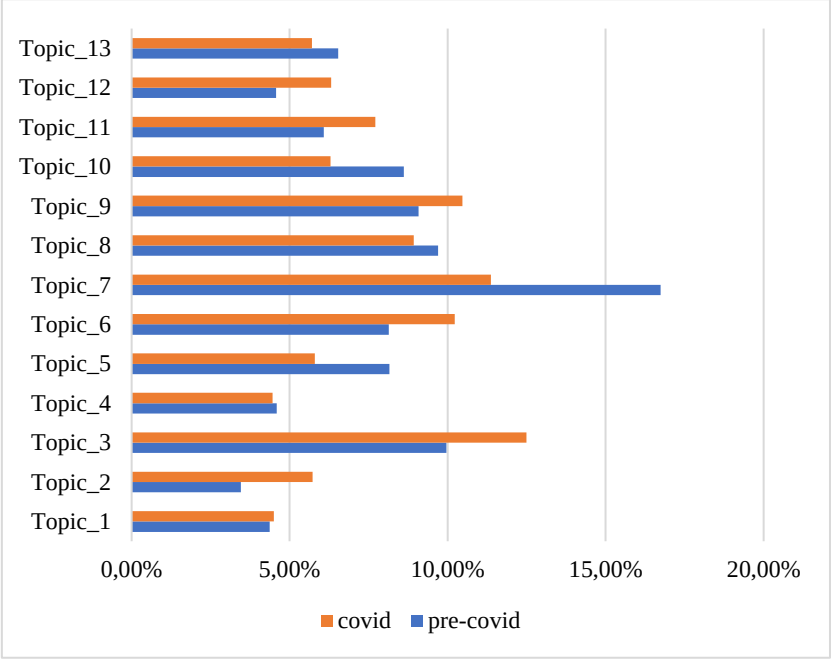
Korábbi korpuszunkat (lásd előző fejezet) egy újabb gyűjtéssel 2020. december 31.-ig, a pandémia első évének végéig terjesztettük ki. A fent leírt adattisztítás megismétlése után a kezdeti összesen 408.804 bejegyzésből 339.550

maradt a végleges adatállományban. A modelleket az STM R-csomaggal (Roberts et al., 2019) illesztettük. Az optimális topikszám meghatározását és a topikok értelmezését az előző fejezetben leírt módon végeztük, mind kvantitatív, mind kvalitatív megközelítést használva, azzal a különbséggel, hogy itt az idő hatását is külön vizsgáltuk, a következőben erre koncentrálok.

A "legrelevánsabb" szavakat a FREX mutató segítségével határoztuk meg. A FREX mutató egy adott topik esetén a szavak topikban megfigyelhető gyakoriságának és a topikra jellemző kizárólagosságának súlyozott átlaga, ahol a kizárólagosság azt számszerűsíti, hogy a szó csak az adott topikban fordul elő. A pontszám kiszámításához az STM sageLabels funkcióját használtuk. Végül egy 13 topikból álló modellt választottunk, mert diagnosztikai tulajdonságai alapján a legjobbak közé tartozott, és hasznos meglátásokat nyújtott anélkül, hogy túlságosan specifikus lett volna. A topikok között nem találtunk érdemi (0,2-nél nagyobb) korrelációt.

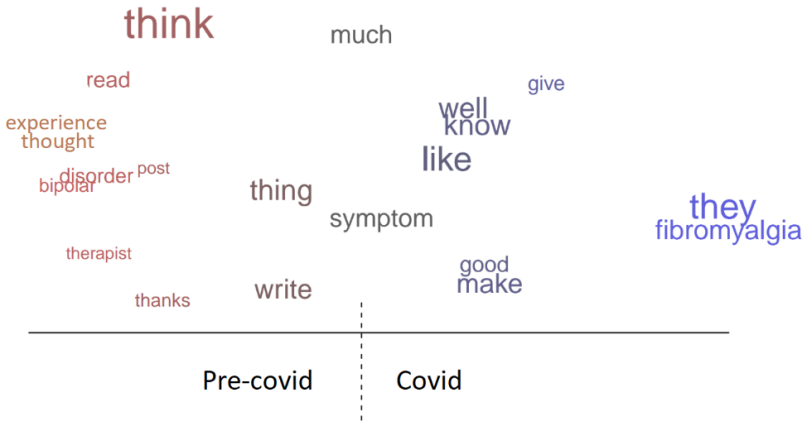
A topikok tartalma az előző fejezetben ismertetett pre-covid elemzéshez képest leginkább abban módosult, hogy egy új téma jelent meg, amely közvetlenül a világjárványhoz kapcsolódott. Ez a téma az egészségügyi intézmények kritikáját tartalmazta, melyek egyre kevésbé váltak elérhetővé az egyre kétségbeesettebb érintettek számára. Bár a kritika nem hiányzott teljesen a pandémia előtti fórumbeszélgetésekből, önálló szemantikai mintaként (azaz önálló topikként) nem azonosítottuk. Úgy tűnik, hogy a pandémiás szabályozás a diszkurzív képet is megváltoztatta: fórumokat nyitott az intézményi kritika számára.

A világjárvány hatását mérő első elemzési dimenzió, a topikok gyakoriságának változása a 8. ábrán látható. A legnagyobb csökkenés a 7. topik (melyet „szenvedésmonológok”-ként címkéztünk), míg a legnagyobb növekedés a 3. topik („mindennapi események naplója”) esetében volt megfigyelhető, bár a változások nagyságrendje nem volt nagyon jelentős.



8. ábra. Az idő hatása a topikok gyakoriságára.

A második elemzési szint az egyes topikok érdemi tartalmának pandémia által indukált megváltozására vonatkozott. Bár a pandémia előtti és a pandémiás korpusz szemantikai különbségét nem minden topik esetében találtuk azonosnak, a kisebb-nagyobb hangsúlyeltolódások mégis egyformán lehetőséget adnak az értelmezésre. Itt csak egyetlen topikot, az 1-est mutatom illusztrációként (9. ábra). Az ábra az STM "perspectives" függvényével készült, és azt mutatja, hogy a topikon belüli szavak közül melyek jelennek meg inkább az egyik vagy a másik időintervallumban. A szavak távolsága az origótól attól függ, hogy a szavak mennyire jellemzőek kizárólagosan a két időszak valamelyikére, a betűnagyság a szó gyakoriságára utal a teljes korpuszban.



9. ábra. Az 1-es topik tartalmának változása a pandémia hatására.

Az 1. topik (jellemző szavainak és posztjainak interpretálása alapján) heterogén beszélgetéseket tartalmaz arról, hogy mit jelent depresszióval élni: sem a biomedikális, sem a pszichológiai diskurzusok nem játszanak hegemón szerepet benne, de mindkét szemantikai univerzumból vesz át paneleket. Beszédaktusként szemlélve a topik főként interaktív hozzászólásokból áll, ami arra utalhat, hogy a hangsúly egy kollektív értelmezési keret megalkotásán van. A 9. ábra ennek a tartalomnak a változását mutatja a pandémia hatására. Az ábra alapján a pandémiát megelőző korszakban többnyire pszichológiai kifejezések domináltak (pl. rendellenesség, élmény, terapeuta), a pandémia-korpuszban a biomedikális kifejezések kerültek előtérbe (pl. fibromyalgia). Egy másik átalakulás az igékre vonatkozik; a pandémiát megelőző korszakban főleg mentális tevékenységeket említenek (pl. posztolni, gondolkodni, olvasni), a pandémiás korpuszban főleg instrumentális (pl. csinálni, tudni) vagy szociális (pl. adni). Ezek az átalakulások a tapasztalatok tárgyalásának hangsúlyeltolódásait jelezhetik: egyrészt úgy tűnik, hogy a biomedikális panelek központibbá válnak, másrészt a problémákat megoldó pragmatikus lehetőségeket a felhasználók gyakrabban említik (az absztrakt válaszok helyett).

Összefoglalóan: az STM gyümölcsöző megközelítésnek bizonyult ebben a projektben. A segítségével létrehozott eredmények arra utaltak, hogy a pandémia nem vette át a depressziós fórumok diszkurzív terét, mégis annak számos aspektusát átalakította: a kritika új horizontja nyílt meg, és az orvosbiológiai tekintély megerősödött.

4.1.5. STM topikmodell, idő és párthovatartozás, mint metaváltozó – A Kárpát-medencével kapcsolatos diskurzusok a magyar parlamentben

Ez az esettanulmány (Németh et al., 2023b) egy karakteresen más területen, a hazai politikai nyilvánosság kutatásában, a Parlamentben az elmúlt húsz évben elhangzott beszédek elemzésével illusztrálja az STM alkalmazhatóságát, metaváltozóként a beszélők párthovatartozását, mint a téma keretezését befolyásoló egyik legfontosabb háttérváltozót is bevonva. A "Kárpát-medence" a magyar földrajzi gondolkodás egyik legjelentősebb fogalma, de a politikai diskurzusban nemrég újra megjelenő használatát kvantitatív szövegelemzéssel még nem vizsgálták. A kifejezés kiválasztásának további motivációja az, hogy a földrajzi fogalmak sosem semlegesek, hanem különböző elképzelésekkel terheltek, ezért különböző módon simulnak politikai-ideológiai (itt: nemzetpolitikai, emlékezetpolitikai) narratívákba.

A magyar Parlamentben 1998 és 2020 között elhangzott, a kifejezést tartalmazó felszólalásokat elemeztünk. Arra keressük a választ, hogy milyen látens témák különböztethetők meg a fogalomhoz kötött diskurzusokban. Mennyiben különbözik az egyes politikai-ideológiai blokkok keretezése? Milyen változások figyelhetők meg a vizsgált időszakban? Mivel a nemzeti identitás kérdése a vizsgált időszakban igen fontos szerepet játszott a politikai vitákban, és a nemzet fogalmával kapcsolatban egyre szélesedett a politikai blokkok közötti szakadék, feltételezhető, hogy ezek a folyamatok a Kárpát-medencével kapcsolatos diskurzusokban is tükröződnek. Az STM használatának motivációja itt elsősorban az volt, hogy a beszéd előadójának politikai pozíciójának metaváltozóként való megjelenítésével a módszer képes választ adni arra a kérdésre, hogy milyen kapcsolat van a politikai pozíció és a látens témák keretezése között.

Mivel elemzésünk középpontjában a parlamenti felszólalások nyelvhasználata állt, érdemes röviden kitérni erre a diszkurzív kontextusra. A parlamentekben a beszéd hatalma a cselekvés hatalma (Ilie, 2017): amit tenni lehet, az nagymértékben függ attól, hogy mit lehet mondani. Ez fontos gondolat a Kárpát-medence kifejezés változó legitimitása/szerepe kapcsán is. Továbbá: a parlamenti beszédek egyszerre mutatnak teátrális és agonisztikus elemeket, azaz egyszerre szólnak formális és versengő célokról (Ilie, 2003). Mivel a magyar közjogi struktúrában a parlament egyszerre tölt be jogalkotói és ellenőrző funkciót, a diskurzusok ez alapján is megkülönböztethetők, ami nagyjából egybevág azzal, amit Ilie nyelvészeti szempontból megállapít: a parlamenti diskurzus műfaja több alműfajt tartalmaz, amelyek e két sajátos parlamenti célnak vannak alárendelve. A vezérszónoki

beszédek, a felszólalások és a napirend előtti felszólalások általában törvényhozási, tehát a pártok fő üzeneteit tükröző reprezentatív funkcióval rendelkeznek. Az ellenőrző funkcióval rendelkező beszédek jellemzően azonnali kérdések, azonnali válaszok, kétperces felszólalások stb., amelyek definíciójuknál fogva nem reprezentatív beszédek. A szövegelemzés szempontjából egy további fontos műfaji megkülönböztetés az, hogy a szöveg előre megírt vagy spontán. A viták emellett közönségközpontúak, mivel más képviselőkből álló valódi közönség, valamint a választókból és a médiából álló virtuális közönség előtt zajlanak.

Korpuszunk 1998. június 25. és 2020. november 23. között elhangzott beszédeket tartalmazott, a korpuszt a K-Monitor nonprofit szervezet, annak önkéntes fejlesztői és a Precognox tanácsadó cég PARLDATA projektjére támaszkodva állítottuk elő. Az elemzést a releváns beszédtípusokra szűkítettük a nem érdemi, technikai jellegű, pl. a napirenddel kapcsolatos beszédtípusok elhagyásával. Csak a pártokhoz kötődő képviselők beszédeit elemeztük, így a független képviselők, valamint a nemzeti és etnikai kisebbségek képviselői kikerültek az adathalmazból. A korpuszt kulcsszavas szűréssel azonosítottuk, csak a "Kárpát-medenc*" kifejezést tartalmazó beszédeket tartva meg (a * itt bármilyen karakterláncot helyettesít, erre a potenciális toldalékok miatt van szükség). Végleges korpuszunk 1525 beszédből állt.

Az előfeldolgozás során a korábbi fejezetekben már ismertetett adattisztítás, lemmatizálás, stopszavazás, releváns bigram- és trigram-azonosítás (pl. „határon túli magyarok”) és névelemfelismerés történt. Fontos megjegyezni, hogy ezeknek a lépéseknek egy része domain-specifikus, pl. saját Parlament-specifikus stopszó-listát fejlesztettünk.

Az STM-ben két metaváltozót használtunk, a dátumot és a politikai pozíciót. A dátum (elméleti okokból) bináris volt: a 2010-es választások előtt/után, míg a politikai pozíció egy (szintén elméletileg indokolt) három kategóriás változó volt: Fidesz / jobboldali-konzervatív blokk / baloldali-liberális blokk. A politikai pozíciót a topikok tartalmát befolyásoló metaváltozóként kezeltük, míg mind a dátum, mind a politikai pozíció befolyásolhatta a topikok gyakoriságát. A topikok optimális számának meghatározásakor a korábbi fejezetekben részletezett módon kombináltuk a kvantitatív és kvalitatív megközelítéseket, végül a 8-as topikszám mellett döntöttünk.

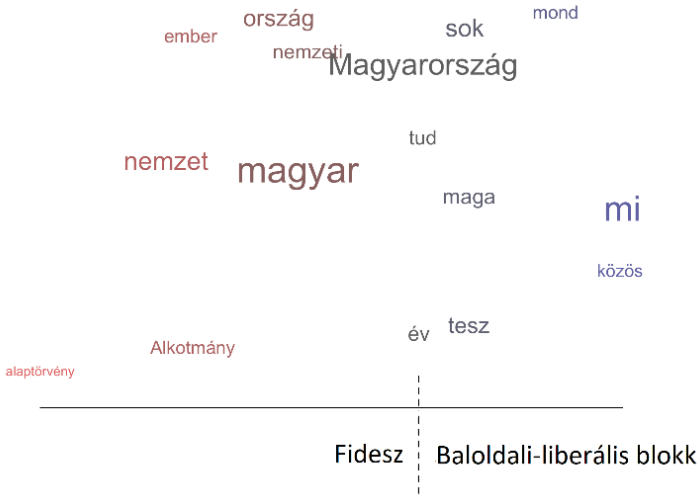
A modell robusztusságát (és általában a módszer érvényességét) mutatja, hogy a 7 és 9 topikos modellek (amelyeket a 8-topikos modelltől függetlenül

illesztettünk) értelmezése azt mutatta, hogy topikjaik megfeleltethetők egymásnak: a kisebb modell egy-egy témája vált szét a nagyobb modell új topikjait létrehozva.

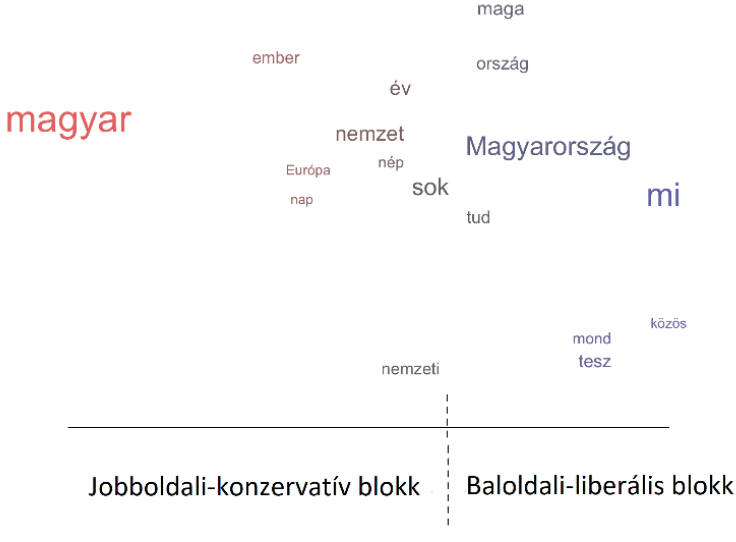
Az előző két fejezethez hasonlóan itt is fontosnak bizonyult a szövegek jelentésén túl azok pragmatikai funkciója is. Így pl. a teljes korpuszhoz képest a Kárpát-medencei alkorpuszban sokkal gyakoribbak az előre megírt vezérszónoki beszédek, felszólalások és napirend előtti felszólalások, amelyeknek jogalkotói funkciójuk van, a pártok fő üzeneteit képviselik, és előre megírtak. Ugyanakkor az alkorpuszban jóval kisebb arányban találhatók nem reprezentatív, nem ünnepi, spontán, ellenőrző funkcióval rendelkező beszédek (azonnali kérdések, azonnali válaszok, kétperces felszólalások stb.).

A topikok értelmezését a korábbi két fejezethez hasonlóan végeztük. Jellegzetes tartalmak rajzolódtak ki, mint az emlékezetpolitika, a gazdaság, a mezőgazdaság vagy a kultúra topikja. A tartalmon túl a beszédfunkció fontosságát leginkább a 3. topik, a határon túli magyarokat érintő közigazgatási és szabályozási kérdésekkel (választójog, közlekedésfejlesztés, egészségügy, lakástámogatások) foglalkozó topik szemléltette. Ez a topik meglehetősen interaktív, és ellentmondásos kérdéseket tartalmaz, amit az is jelez, hogy a releváns kifejezések között az "Ön" és a "ne haragszik" (a "ne haragudjon..." kifejezés lemmatizált formája) is a leggyakoribb kifejezések között szerepel.

A továbbiakban a politikai pozíció hatására vonatkozó néhány eredményt szemlézek, itt is inkább a módszertan potenciáljának és a lehetséges interpretációs stratégiáknak az illusztrációjaként. Egy több topikra is jellemző mintázat, hogy míg a Fidesz és a jobboldali-konzervatív képviselők saját közösségük képviseletében beszélnek (a "mi" névmás felülreprezentált), addig a baloldali-liberális oldal formálisabb, kevésbé személyes, és intézményekre ("EU", "kormány", "Magyarország") utal. Ez az ellentét tükröződik az emlékezetpolitika topikjában (lásd 10-11. ábrák), ahol a Fidesz és a jobboldali-konzervatívok etnokulturális alapon ('nemzet', 'magyar'), a baloldal pedig állampolgári alapon ('Magyarország') nevezi meg saját közösségét.

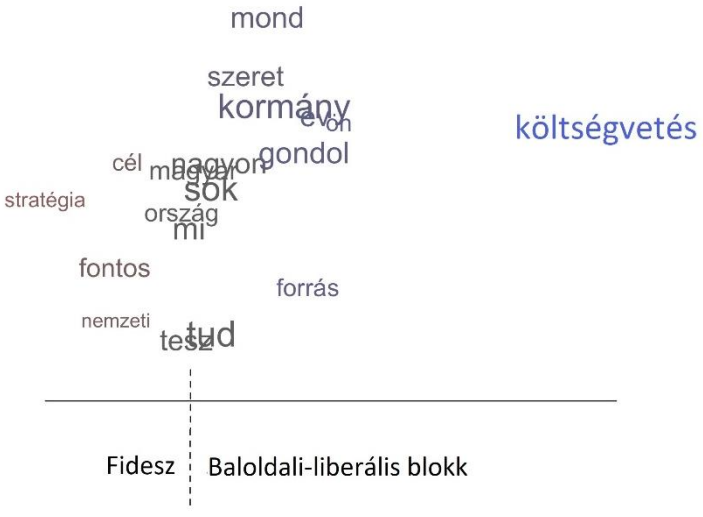


10. ábra. A politikai pozíció hatása a nemzetpolitikával kapcsolatos topik narratívájára, Fidesz vs. baloldali-liberális összevetés.

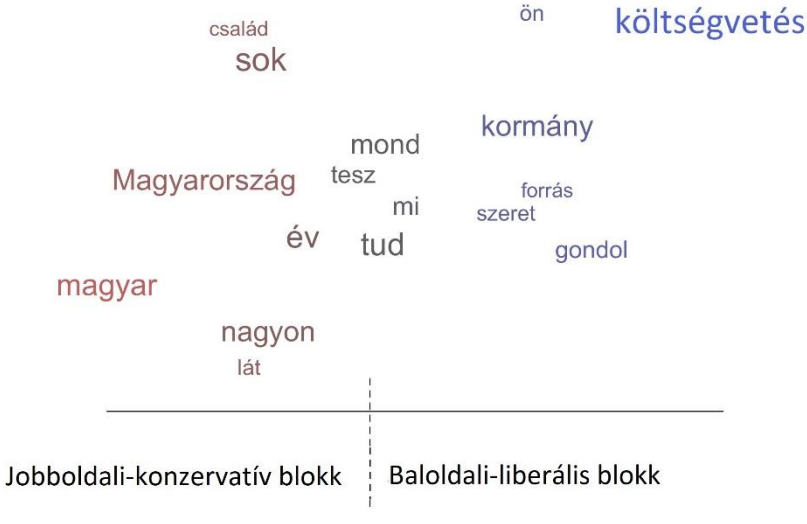


11. ábra. A politikai pozíció hatása a nemzetpolitikával kapcsolatos topik narratívájára, jobboldali-konzervatív vs. baloldali-liberális összevetés.

A jobboldali-baloldali megosztottság egy további dimenziója a gazdaságról szóló topikban jelenik meg (lásd a 12-13. ábrákat): a két jobboldali blokk érték- és érzelmi alapon, míg a baloldal racionális alapon érvel, előbbi retorikai, utóbbi bürokratikus/szakértői megközelítésben.



12. ábra. A politikai pozíció hatása a gazdasággal kapcsolatos topik narratívájára, Fidesz vs. baloldali-liberális összevetés.



13. ábra. A politikai pozíció hatása a gazdasággal kapcsolatos topik narratívájára, jobboldali-konzervatív vs. baloldali liberális összevetés.

Összefoglalóan: az STM lehetőséget adott arra, hogy a Kárpát-medence parlamenti diskurzusában azonosítsuk ugyanazon témák politikai pozíciótól függő eltérő megközelítését. Míg a baloldali-liberális pártok kevésbé személyes hangnemben beszélnek, intézményekre, érdekekre és stratégiára utalnak, addig a Fidesz és más jobboldali-konzervatív pártok beszédeiben a képviselők saját közösségük képviseletében beszélnek, értékekre, érzelmekre és kultúrára hivatkozva. Eredményeink megerősítik azt az általános megfigyelést (pl. Mouffe, 2011), hogy a baloldali-liberális blokk hisz az egyetemes racionális konszenzusban, miközben figyelmen kívül hagyja a politika érzelmi dimenzióját. Ezzel szemben a jobboldali-konzervatív blokk kollektív identitást kínál, amelyhez az emberek értékeket és érzelmeket kapcsolnak.

4.1.6. Dinamikus topikmodell – A korrupció hazai média-reprezentációjának vizsgálata

Utolsó topikmodell-esettanulmányként egy másik variáns, a dinamikus topikmodell alkalmazására mutatok példát (Katona et al., 2021), az eddigiektől karakteresen különböző domain-en (online média), témában (korrupciókutatás) és módszerrel (az exploratív elemzésen túlmutató magyarázó modellek). A korrupciókutatás rendkívül gazdag módszertani eszköz-készlettel rendelkezik, ez a megközelítés más módszerekkel szemben nem a korrupció nagyságát vagy elterjedtségét méri, hanem arról ad képet, hogy milyen tematikák mentén alakul a korrupcióról a közbeszéd egy adott országban.

Elemzésünk a korrupció hazai online média-reprezentációját vizsgálta, ahol a dinamikus topikmodell (*dynamic topic model*) használatának motivációja az volt, hogy az idő dimenziójának megjelenítésével a topikok tartalmának folyamatos változását is képesek legyünk modellezni. A dinamikus topikmodell (Blei & Lafferty, 2009) az STM-hez hasonlóan az LDA variánsa, de azzal szemben folytonos metaváltozóként kezeli az időt. Hasonlóan az STM metaváltozóihoz, itt is évről-évre változhat mind a topikok gyakorisága, mind tartalma is. A modell illesztéséhez a Python gensim csomagjának LdaSeqModel osztályát alkalmaztuk.

Szövegkorpuszunkat a K-Monitor cikkgyűjteménye adta, ami korrupciógyanús, valamint szabálytalan közpénz-felhasználással kapcsolatos ügyeket feldolgozó online sajtóbeli cikkeket tartalmaz. 26.262 cikkre épült az elemzésünk, melyek 152 portálról származtak. Ezek többsége híroldal, de bulvároldalak és

blogok is megjelentek közöttük. 2007-től, az adatbázis építésének kezdetétől 2018 augusztusáig gyűjtöttük le a cikkeket.

Vizsgálatunkban egyrészt célunk volt exploratív módon azonosítani a 2007-2018 közötti időszakra vonatkozóan a cikkek főbb témáit és a tematikus változás dinamikáját. Másrészt magyarázatokra is kísérletet tettünk annak vizsgálatával, hogy van-e kapcsolat a tematika és a médium ellenzéki/kormánypárti pozíciója között, illetve, hogy a parlamenti választások kampányidőszaka befolyásolja-e a korrupció reprezentációjának tematikáját. Mivel az elemzett időszak során megváltozott az origo.hu hírportál tulajdonjoga, természetes kísérletként adódott a tulajdonosváltás hatásának vizsgálata is. Az alábbiakban csak a korábbi fejezetekben ismertetett topikmodellezési tapasztalatokat kiegészítő módszertani tapasztalatokat szemlézem.

A szokásos előfeldolgozás után az optimális topikszám kiválasztását és a topikok interpretációját a fent ismertetett módon végeztük, kvantitatív és kvalitatív megközelítésre egyaránt támaszkodva. A választott 7 topik jól megfogható tartalmat mutatott (pl. közbeszerzési ügyek, nemzetközi kapcsolatok, önkormányzati ügyek, vállalatok) és külső információkkal (mint a 2010-es kormányváltás, konkrét politikai események) érvényességük jól alátámasztható volt. Vizsgáltuk a topikok gyakoriságának változását, klaszterződésüket, egymáshoz viszonyított távolságának változását, majd a topikok tartalmának változását. Pl. a kormányzati és nem-kormányzati szervezetek kapcsolatát érintő topik legrelevánsabb szavai között a ’minisztérium’, ’tárca’, ’közigazgatási’, ’államtitkár’, ’adat’, ’nyilvánosság’, illetve az ’alapítvány’, ’civil’, ’egyetem’, ’közérdekű’ volt (ezek a kifejezések más topikokban nem szerepeltek lényegesen). A topik tartalmi dinamikájára jellemző, hogy a ‘civil’ kifejezés 2012-ben jelenik meg és egyre fontosabbá válik a topikban. Jól interpretálható eredményeket kaptunk az egy-egy nagy portálra (index, origo, hvg) megszorított elemzésben is: itt azt vizsgáltuk, hogy az egyes portálok milyen témákat tárgyaltak, és hogyan változott ez a preferencia.

Az exploratív megközelítésen túlmutató magyarázó modellek definiálására az ad lehetőséget, hogy az egyes cikkekhez tartozó topik-kontribúciók numerikus változóként elmenthetők (értsd: mennyire vett részt adott topik a cikk létrejöttében, a 7 topikhoz tartozó kontribúciók összege cikkenként 1-gyel egyenlő). Így klasszikus adatstruktúrát kapunk: a sorokban a cikkek, az oszlopokban a hét topikhoz tartozó kontribúció, és olyan, a cikket jellemző változók, mint a cikk forrása, megjelenésének időpontja. Így a csupán néhány fontos változót tartalmazó

numerikus adatbázis birtokában a klasszikus statisztikai módszerek is alkalmazhatók voltak[9]. Konkrétan: variancia-elemzés és lineáris regresszió segítségével bizonyítható volt, hogy a parlamenti választások kampányidőszakának hatása van az egyes topikok tárgyalására (itt a kormánypárti / ellenzéki média, mint változó is a független változók között volt, hogy korrigálhassunk ennek hatására), s hogy az *Origo* tulajdonosváltásának hatása volt a korrupcióról szóló cikkek számára és tartalmára. Összességében a dinamikus topikmodell jól használhatónak bizonyult a kutatási kérdések megválaszolására, az eszköz potenciálisan más médiakutatási területeken is jól alkalmazható lehet.

4.1.7. A topikmodellezés legfontosabb módszertani tapasztalatai

Módszertani szempontból a fenti kutatási tapasztalatok alapján a topikmodell hatékonyan járulhat hozzá a szociológiai tudástermeléshez, ha érdemi elméleti kérdésekhez kapcsolódik, és ha a számítási módszereket kvalitatív értelmezéssel kombináljuk. Ez a kevert módszertanú megközelítés az eredmények validálásában és interpretálásában is hatékonynak bizonyult. Azt is láttuk, hogy az ezekhez hasonló, nagy volumenű kvantitatív kutatásokban az adatvizualizáció nemcsak illusztrálja az eredményeket, hanem a vizsgálat szerves részét képezi. Támogatja az adat-feldolgozást és az értelmezést mind az elemzés, mind a publikálás során.

Milyen „témákat" azonosítanak a topikok? A fenti példák is alátámasztják, hogy a topikok értelemzésekor egyszerre két dologra érdemes koncentrálni: a szövegek szemantikai oldalára (miről beszélnek?) és pragmatikai szintjére (milyen módon, milyen céllal, a hallgatóságra milyen hatás gyakorlásával?). Láttuk azt is, hogy érdemes itt Austin (1975) beszédaktus-elméletét alkalmazni, aki elsőként hívta fel a figyelmet arra, hogy a megnyilatkozások nem csak információt közvetítenek, hanem tényleges cselekvések, melyek sokféle funkciót (ígéret, meggyőzés stb.) töltenek be és a beszélő környezetére is hatást gyakorolnak az interperszonális kommunikáció részeként. Ezek a beszédaktusok három szinten értelmezhetők Austin szerint: mit mondunk (lokúciós aktus), mit teszünk ennek kimondásával

[9] Tehát ez egy csökkentett dimenziójú adatbázis. Az eredeti dokumentum-kifejezés adatbázis (ahol minden szó egy saját változót képez, és adott dokumenthez tartozó érték azt jelzi, hányszor fordul elő a szó a dokumentben) ehhez képest klasszikus módszerekkel nehezen elemezhető, mert rendkívül sok változót tartalmaz, amiknek viszont majdnem mindig 0 az értéke (vagyis un. 'sparse data').

(illokúciós aktus, pl. ígérünk, kérünk) illetve milyen hatást érünk el ezzel (a hallgató felé, perlokúciós aktus). Láttuk, hogy az online depressziós fórumok topikmodellezésekor (Németh et al., 2021) a topikok értelemzése két dimenzió mentén végezhető: azok tartalma ill. kommunikációs funkciója szerint, és a fő választóvonalat éppen az utóbbi jelentette, e mentén különítettünk el monológokat és interakciókat. Mindkét kategóriában további al-funkciókat azonosítottunk, pl. a monológokon belül a lokúciós aktus által dominált objektívebb, illetve az illokutív aktus által dominált érzelmileg töltöttebb önvallomásokat.

Hasonlóan fontos szempontnak bizonyult a pragmatika, vagyis a szövegek műfajának és funkciójának figyelembevétele a Kárpát-medence fogalmának parlamenti beszédekben történő használatának kutatásában is. Azt találtuk pl, hogy az összes beszédhez képest a Kárpát-medence fogalmat említő beszédek között sokkal gyakoribbak a reprezentatív, előre megírt beszédek, melyek a pártok fő üzeneteit képviselik, és kisebb arányban találhatók spontán beszédek (azonnali kérdések, azonnali válaszok, kétperces felszólalások stb.). Továbbá, a Kárpát-medence alkorpuszon belül azonosított topikok közül a határon túli magyarokat érintő közigazgatási és szabályozási kérdésekkel foglalkozó topik nem csak szemantikailag, de pragmatikailag is elkülönült a többitől, a téma sok interakciót, vitát generált. Minden alapvető fontosságú, és arra utal, hogy adott korpusz elemzésénél nem feledkezhetünk meg annak kontextusáról. A Parlament, mint kontexus esetén a beszéd elsődlegesen maga is politikai cselekvés: célja a politikai nézetek nyilvános reprezentációja és a hallgatóság befolyásolása, a meggyőzés vagy a cáfolat.

Összefoglalóan: a beszédaktus-alapú megközelítést érdemes használni topikmodellek interpretálásánál, hiszen a modell feltételezése (véges számú topik generálja a szövegeket, ahol mindegyik topik sajátos szótárral rendelkezik) alapján létrejött topikok semmiképp sem csak szemantikai, hanem pragmatikai eltéréseket is mutatnak, melyek megtalálása és értelmezése sokat adhat az eredményekhez. Erről részletesen írtunk Sik Domonkossal egy kifejezetten a topikmodellek pragmatikai fókuszú interpretációjára fókuszáló módszertani cikkünkben (Németh, Sik, 2024).

A topikmodell statisztikai outputjára térve: láttunk rá példát, hogy ha a modellt nem klasszifikációs módszerként közelítjük meg (ami a szövegeket topikokhoz sorolja), hanem a létrejött topikok szövegenkénti kontribúcióját, mint numerikus változót használjuk, olyan strukturált adatbázishoz jutunk, ahol a szövegek meta-jellemzőinek (időpontjuk, forrásuk, szerzőjük tulajdonságainak)

segítségével klasszikus statisztikai módszereket tudunk használni, azaz tartalmas szociológiai hipotéziseket tesztelni (lásd Katona et al., 2021).

A módszernek ugyanakkor megvannak természetesen a maga korlátai is. A topikmodellezés azon a feltételezésen alapul, hogy az adatok látens struktúrája véges számú topikkal reprezentálható - ez a feltételezés nem feltétlenül igaz minden szociológiai kontextusban, hiszen a társadalmi jelenségek gyakran összetettek és sokrétűek. Továbbá, a topikmodellezés megköveteli, hogy a kutató hozzon döntést a topikok számáról. Ez szubjektív és önkényes folyamat lehet, mivel nincs egyezményes módszer ennek meghatározására – bár ismertettem eljárást arra, hogyan lehet kvantitatív és kvalitatív oldalról egyaránt megtámogatni egy ilyen döntést, és hogyan lehet a döntés robusztusságát tesztelni.

Minden NLP-elemzésnél érdemes a korpusz korlátairól is beszélni. Példaként: az online depresszió-fórumokkal foglalkozó kutatásunk célcsoportját a depresszió által közvetlenül vagy közvetve érintett személyekként határoztuk meg. Ám azok, akik aktívan keresnek támogatást egy online fórumon keresztül, nem biztos, hogy reprezentatívak a teljes célcsoportra nézve. Továbbá, mivel az internethasználat összefügg a társadalmi-gazdasági státusszal, a magasabb iskolai végzettségűek felülreprezentáltak lehetnek a vizsgálatunkban. Végül, az adatvédelmi és etikai előírások miatt a keresésünk azokra a fórumokra korlátozódott, amelyek nyilvánosak és regisztráció nélkül hozzáférhetők voltak. A jelszóval védett támogató csoportok ugyanakkor biztonságosabb helyeknek tűnhetnek a felhasználók számára. Végül: csak a konkrét, depresszióval kapcsolatos kulcsszavakat tartalmazó bejegyzéseket gyűjtöttük össze, ezzel biztosítva, hogy csak olyan posztok kerülnek kiválasztásra, amelyek kifejezetten a depresszióról szólnak, azonban az eljárás a depressziót kóros hangulatzavarnak tekintő beszélgetések túlreprezentáltságához vezethet, miközben alulreprezentálja a kulturális vagy társadalmi utalásokat.

Az elemzési egység megválasztásáról. Az online depresszió-fórumok elemzésekor a hozzászólásokat önálló egységként elemeztük, miközben azok dinamikus és interaktív mintákkal rendelkező beszélgetések részei (vagyis sem az idő dimenzióját, sem a poszt-hozzászólás relációt nem vettük figyelembe, és a közös szerzővel rendelkező cikkeket sem vetettük egybe). Későbbi kutatásainkban ezért használtuk fel az időt és a szerző-azonosítót is a fórumok elemzésekor, és igyekeztünk longitudinálisan követni a felhasználókat (Sik et al., 2023a).

4.2. Szóbeágyazás

4.2.1. A szóbeágyazás általában

A topikmodell az egyes látens témák feltárását a szótár szavainak dokumentumokon belüli együttes előfordulására alapozva végezte, de ahogy hangsúlyoztuk, szózsák-modellt használt, vagyis nem vette figyelembe a szósorrendet. Egyes kutatási kérdések azonban az egyes szavak nyelvi kontextusának (azaz az őket megelőző és követő néhány szónak) a figyelembevételét igénylik. Az erre megoldásként létrehozott nyelvi reprezentációk – az úgynevezett szóbeágyazási modellek - a szavakat vektorokként ábrázolják egy nagydimenziós térben, ahol az elemzett korpuszban hasonló kontextusban előforduló szavak egymáshoz közel helyezkednek el a térben (Mikolov et al., 2013).

A módszer tehát az egyes szavakhoz rendel vektortérbeli pozíciót, amit a szó jelentésével azonosíthatunk, még pedig úgynevezett disztribúciós szemantikai megközelítésben, amikoris a jelentést kizárólag a használati környezet határozza meg. A disztribúciós szemantika fogalmát John Firth nyelvész fogalmazta meg (1957), aki szerint egy szó jelentését az a környezet azonosítja, amelyben a szó jellemzően előfordul. A szójelentést nem a szavak és a ’valóság’ elemeinek kapcsolatával, hanem azt a szavak használatával azonosító disztribúciós szemantikai megközelítés legkorábbi nyelvfilozófiai reprezentánsai Wittgenstein 1930-as évekbeli munkái.

A szóbeágyazásokat különböző algoritmusokkal lehet előállítani, melyek széles körben elérhetők nyílt forráskódú szoftverekben (magyar nyelvű összefoglalóként lásd Kmetty, 2022). Vannak a módszernek olyan általánosításai is, amelyek nagyobb szövegegységekre (mondatokra, dokumentumokra) adnak vektorreprezentációkat.

Rendkívül érdekes, hogy a szóbeágyazási modellre általában nem úgy gondolnak (létrehozói sem), mint a szövegelemzésre szolgáló önálló módszerre, hanem inkább, mint a szöveg reprezentációjának lehetséges módjára. Pl. kimutatták, hogy ez a reprezentáció – szemben az egyszerű szózsák-reprezentációval - jelentősen javíthatja a modellek teljesítményét predikciós vagy klaszterezési feladatokban (sőt, a tulajdonképpen predikciós feladatot végző ChatGPT is egyfajta szóbeágyazási modell alkalmaz). Ugyanakkor az utóbbi tíz évben sok-sok társadalomtudományi alkalmazás (pl. Garg et al., 2018; Kozlowski et al., 2019) bizonyította, hogy – mintegy mellékhatásként - a szóbeágyazási modellek alkalmasak

a szavak közötti sokrétű asszociációk ábrázolására, sőt a kulturális jelentésárnyalatok megragadására is. Úgy tűnik, hogy az emberi társadalmi kapcsolatok leképeződéseit is felfedezhetjük a szövegekben. Kozlowski és társai (2019) szociológus szemmel rendkívül inspiráló cikkében pl. egy évszázad alatt megjelent több millió könyv korpuszának feldolgozásával a bennük található szavakat egy 300 dimenziós vektortérben ábrázolták, és ezeket a vektorokat arra használták, hogy megkonstruálják a társadalmi osztályok kulturális dimenzióit, majd e nyelvi vektortérben nyomon kövessék változó pozíciójukat az idők során.

A módszer lényege heurisztikusan jól megfogható: a vektortérben két szó jelentésének közelségét a nekik megfelelő, az origóból induló vektorok által bezárt szög nagysága (pontosabban általában annak cosinusa) segítségével határozzák meg. Lásd a 14. ábrát: a ‘urologist’, ’surgeon’ és ’physician’ kis szöget zárnak be, hiszen sokszor szerepelnek hasonló környezetben. Fontos hangsúlyozni, hogy e használat-alapú definícióból következően a vektortér nem a szavak jelentésének hasonlóságán, hanem a szavak használatának kapcsolatán alapszik. A ‘man’ és a ‘woman’ például közel van egymáshoz, hiszen gyakran szerepelnek hasonló környezetben mondatainkban.

A modell által létrehozott vektortér képes tehát szemantikai kapcsolatok megragadására. A vektortér jól teljesít analógiás teszteken is, pl. a főváros relációt egy ország és fővárosa, mint vektor különbségével megadva választ kaphatunk a „*Mi Németország fővárosa?*” kérdésre:

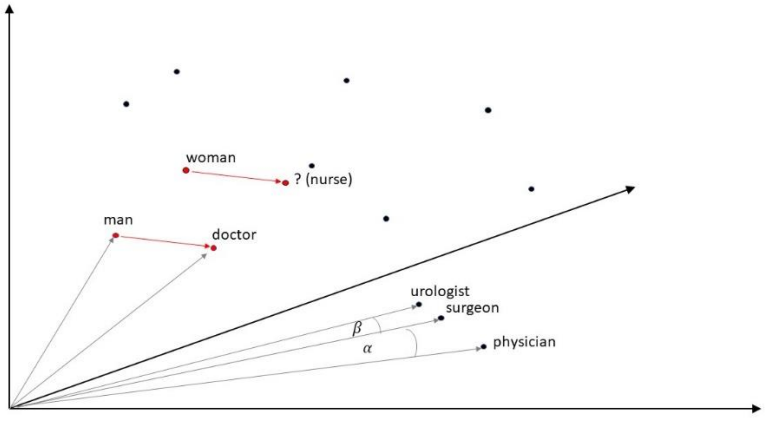
$$\textit{Franciaország} - \textit{Párizs} = \textit{Németország} - ?$$
$$? = \textit{Németország} - (\textit{Franciaország} - \textit{Párizs}) =$$
$$\textit{Németország} - \textit{Franciaország} + \textit{Párizs}$$

Ilyenkor a megfelelő vektortér-elemeket tekintve Berlin úgy adódik, hogy megkeressük az egyenlet fenti megoldásához legközelebbi vektort. Elég nagy (több százmillió szót tartalmazó korpuszokon) tanítva a modell ilyen analógiás teszteken jól működő vektorteret ad (vagyis az egyenlet jobb oldalához legközelebbi vektor valóban Berlin lesz). A vektortér teljesítményét általában is ilyen analógiás teszteken értékelik, léteznek többszáz kérdésből álló teszt-bázisok, lásd pl. Kozlowski et al., 2019.

Kozlowsky et al. (2019) társadalomtudományi kutatásokban jól hasznosítható megoldással ennél tovább ment: látens dimenziókat azonosított a vektortérben előre definiált szópárok, mint vektor-irányok segítségével (pl. man-woman,

és szinonimáik/szintaktikai variánsaik). Majd megmutatták, hogy e szópárok vektortérbeli elhelyezkedése által indukált dimenziók (a férfi – nő mellett: fekete - fehér, liberális - konzervatív) ezekben a vektortérben szorosan megfelelnek a kulturális jelentés dimenzióinak, és a szavak vetülete ezekre a dimenziókra széles körben elfogadott kulturális konnotációkat tükröz. A vektortérbeli szó-mozgások időbeli követésével a nemi, etnikai és politikai preferencia alapú asszociációk változását tudták reprezentálni az Egyesült Államokban a 20. század folyamán. Így például kirajzolták a fogyasztás és életmód (sport, ételek, zene, foglalkozás, keresztnevek) etnikai alapú differenciálódását. Értsd: a fehér-fekete tengelyre vetítve a zenei műfajokat, a hiphop a fekete oldalra, az opera a fehér oldalra esik inkább.

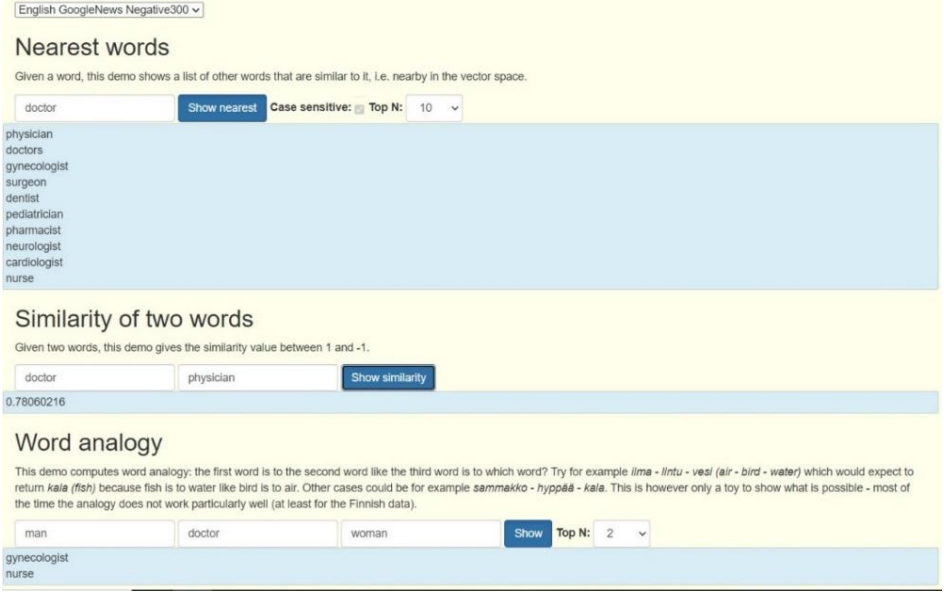
Itt tehát a társadalomtudományi kutatások szempontjából kulcsfontosságú jelenség rajzolódik ki: a szóbeágyazási vektortérben a szavak szemantikai kapcsolatai, sőt a jelentés-struktúra látens meghatározó dimenziói is megragadhatók.



14. ábra. Szavak jelentésbeli hasonlósága ill. jelentés-analógiák megjelenése a szóbeágyazási modell vektortér-reprezentációjában.

Azon felhasználók számára, akik nem tudnak helyi beágyazási modelleket illeszteni egy korpuszhoz, rendelkezésre állnak előre tanított (*pretrained*) szóvektorok, amelyeket nagy korpuszokból becsültek meg. Ezeket felhasználva az elemzők a saját korpusz elemzésekor is alkalmazhatják a nagy korpusz reprezentációját. (Ilyenkor persze komoly kérdésként merülhet fel, hogy a nagyobb szövegből átvett szójelentések hogyan alakítják a saját korpuszunk jelentését.)

Ilyen előre elkészített vektortér például a Google szakemberei által a Google News (2013-ban készült) korpuszán illesztett Word2Vec modell. Ez a vektortér szabadon elérhető, és akár programozó tudás nélkül is vizsgálható, lásd pl. a Turku NLP-group oldalán levő, nagyon egyszerűen használható demo-t: http://epsilon-it.utu.fi/wv_demo/. A 14. ábrán látható példákat magam is ebből vettem, nem csak az orvosi szakmák elnevezésének közelségét bizonyítja a demo, de azt is, hogy az analógiás teszt a korpuszban a társadalmi viszonyok lenyomataként megjelenő nemi sztereotípiákra világít rá: ami a '*man*'-nek a '*doctor*', az a '*woman*'-nek a '*nurse*'[10]. A demo ilyen analógiákra, adott szó legközelebbi környezetének listázására, és két tetszőleges szó közelségének megadására is lehetőséget ad (lásd a demot működésben közben megmutató 15. ábrát). Számos hasonló szóbeágyazási demo érhető el az interneten, más korpuszokon, más nyelveken (*word embedding* és *demo* keresőkifejezésekkel).



15. ábra. A Turku NLP group szóbeágyazási demo-ja a Google News korpuszon.

[10] Az utolsó feladatra két megoldást is közöl a demo (gynecologist és nurse), mert az analógiás szóvektor-különbség nem pontosan egy tényleges szó felé mutat a beágyazási térben, így a demo a legközelebb eső szavakat sorolja fel. Vegyük észre, hogy a gynecologist a különbség egy másik jelentését ragadja meg: a férfiak orvosa a *doctor*, a nőké a nőgyógyász (míg a *nurse* , mint megoldás a nőies/férfias foglalkozás jelentését ragadja meg).

A szóbeágyazás inspiratív szociológiai lehetőségeit jól szemléltetik hazai társadalomtudományi alkalmazásai is. Ezek között van Kmetty és társai (2021) munkája, akik a struktúrakutatások foglalkozás-hierarchiájának megfelelő mintázatot találtak angol nyelvű vektortérben a foglalkozások megnevezésének pozícióit vizsgálva. Szabó és társai (2020) a *Pártélet* c. folyóirat szocialista évtizedekben megjelent számait elemezve vizsgálták olyan kitüntetett szópárok egymáshoz való viszonyát, mint az ipar, mezőgazdaság vs. haladás, reform. Ilyés és társai (2018) pedig a magyar parlamenti beszédekben azonosították ugyanazon témák nemek szerinti eltérő keretezését, és kognitív pszichológiai aspektusból pl. a passzív/aktív kontrasztot találták párhuzamosnak a nő/férfi kontraszttal.

4.2.2. Szóbeágyazás az online depresszió-fórumok korpuszán – Probléma-megoldás vektorok kihasználása

Saját kutatásunkban (Sik et al., 2023b) az online depresszió-fórumok korpuszának alkottuk meg szóbeágyazási modelljét, azzal a céllal, hogy feltérképezzük a fórumozók szemantikai univerzumának látens szerkezetét. Konkrétabban: a fórum-felhasználók által felvetett bizonyos problémák és a hozzájuk a fórumokon kapcsolt megoldások közötti kapcsolatot szerettük volna megragadni. A releváns problémákat és megoldásokat ugyanezen korpusz topikmodellezése szolgáltatta számunkra (Németh et al., 2021).

Először a biomedikális és a pszichológiai diszkurzív struktúrák alapmodelljeit alkottuk meg probléma-megoldás szópárok kidolgozásával. Ilyen volt az orvosbiológiai diskurzushoz tartozó *symptome-test* vagy *symptome-treatment* szópár, vagy a pszichológiai diskurzushoz tartozó *trauma-therapy* és *feel-understand* szópár. A szociológiai diskurzusban a topikmodell releváns kifejezései csak társadalmi problémákat jelöltek ki, társadalmi megoldásokat nem. Ezeket a problémákat kijelölő kifejezések között volt pl. a *work, school, money, class, college, study, company, career, university, disability, office, insurance, afford, business*. Arra voltunk kíváncsiak, hogy ebben a szemantikai univerzumban milyen megoldásokat találnak/kínálnak a felhasználók a társadalmi problémák ezen formáira.

Részletesebben: három listát készítettünk a tíz legrelevánsabb szóból az orvosbiológiai, pszichológiai és társadalmi keretezében tárgyalt topikokból. Ezután a korpuszban található szavak szóbeágyazási vektorreprezentációját alkottuk meg, a skip-gram negative sampling modellt használva a gensim Python könyvtárában (Rehurek & Sojka, 2011) található word2vec (Mikolov et al., 2013)

implementáció segítségével. Mivel ez egy depresszióval kapcsolatos online fórum-korpusz, nem pedig egy általános célú korpusz, ezért vektortér-modell validálására általában használt, előre összeállított szó-analógiás tesztek (lásd fent Németország fővárosának analógiás megtalálását) teljesítménye félrevezető lehet. Ehelyett kiválasztott kulcsszavakhoz legközelebb eső szavak vizsgálata a legnagyobb cosinus-hasonlósággal rendelkező szavakra, valamint a cosinus-hasonlóságok eloszlásának általános alakjának és görbületének vizsgálatára (Schakel & Wilson, 2015) támaszkodtunk.

Ezután a pszichológiai vagy biomedikális diskurzusokban problémaként ill. megoldásként azonosított szavak minden párosítását létrehoztuk, és a probléma-megoldás-párok különbségét, mint vektort számoltuk ki. E probléma-megoldás kapcsolatok tipizálása érdekében ezután kategorizáltuk e vektorokat, mégpedig jelentésük közelsége (azaz cosinus-hasonlóságuk alapján). Ez a csoportosítás növelte a mögöttes jelentés megragadásának robusztusságát is (a fent idézett Kozlowski et al., 2019 kutatás is sok-sok szópár különbségvektorának átlagaként hozta létre pl. a nemi dimenziót: *man-woman, he-she, man-women* stb).

Ezután az így azonosított problémamegoldó vektorokat hozzáadtuk azon szavak vektoraihoz, amelyeket korábban a társadalmi okokhoz köthető problémák (pl. munka, pénz, iskola) potenciális képviselőiként azonosítottunk. Amennyiben a problémamegoldó vektorok valamilyen funkcionális kapcsolatot rögzítenek, a kiválasztott "probléma" vektorokhoz való hozzáadásuk "megoldás" vektorokat eredményezett. Más szóval: a Németország fővárosának megtalálásához vezető, fent idézett analógiás kérdésekre próbáltunk itt is válaszolni, mint: "A fájdalom [biomedikális probléma] úgy viszonyul a teszthez [biomedikális megoldás], mint a pénz [szociológiai probléma] a ...-hez?". Mivel ezek a vektorok nem pontosan egy tényleges szóvektor felé mutatnak a beágyazási térben, megnéztük a 10 legközelebbi szót minden egyes megoldási vektorhoz (mint korábban a 15. ábrán, ahol két legközelebbi szót jelölt ki a vektor).

Az eredmények részletes bemutatása helyett csak illusztrációként sorolnék fel néhány megoldást. A munka, az iskola, a karrier, az anyagi biztonság társadalmi nehézségeivel kapcsolatos biomedikális megoldások egyik visszatérő mintáját az egészséges életmód gyakorlata (testmozgás, egészséges táplálkozás stb.) alkotja. Egy másik kvázi-biomedikális kezelés a szeralapú öngyógyítási kísérletekhez, leginkább az alkohol- és marihuánafogyasztáshoz kapcsolódik. A társadalmi szenvedés pszichológiai megoldásainak egyik jellemző típusa az intim interszubjektivitáshoz kapcsolódik, olyan kifejezésekkel jelölve, mint a

gondoskodás, felelősség vagy bizalom. Ezeknek a megoldásoknak a részletes elemzése, és a hozzájuk kapcsolható fórumbejegyzések értelmezése révén hozzá tudtunk férni ahhoz a kérdéshez, hogy a társadalmi szenvedés medikalizálása vagy pszichologizálása milyen mélységben valósul meg a diskurzusokban.

4.2.3. A szóbeágyazás módszertani tapasztalatai, társadalomkutatási lehetőségek

Mint láttuk, a szóbeágyazás értékes eszköz lehet a szociológia területén. A szóbeágyazások elemzésével a kutatók betekintést nyerhetnek diskurzusok mögött kitapintható, látens szociológiai jelenségekbe/jelentésekbe. Tanulmányozhatják a nyelvben megnyilvánuló társadalmi/kulturális fogalmak jelenlétét, történeti korpusz megléte esetén időbeli fejlődését is, akár többszáz éves távlatban. Megvalósíthatjuk társadalmi előítéletek és sztereotípiák feltárását, kulturális változásokat és társadalmi átalakulásokat.

Ugyanakkor a módszernek számos korlátja is van. A szóbeágyazások nem feltétlenül ragadják meg az árnyaltabb vagy absztraktabb szociológiai fogalmakat, amelyek nem tükröződnek explicit módon a nyelvhasználati mintákban. Egy másik korlát a szójelentés értelmezése: a modell használatakor azzal a feltevéssel élünk, hogy a szó jelentése megragadható a körülötte tipikusan megtalálható lévő más szavak segítségével.

5. A FELÜGYELT GÉPI TANULÁS

Ebben a fejezetben azokat a lehetőségeket tekintem át, amelyeket a felügyelt gépi tanulás nyújthat a szociológia számára. Először az algoritmus működéséhez szükséges címkézett (szaknyelven: annotált) korpusz létrehozásának társadalomkutatás-specifikus kihívásairól írok (részben saját kutatási tapasztalatok alapján), kitérve az annotálás okozta torzítás lehetőségére is. Majd néhány olyan saját esettanulmányt mutatok be, amikor nem volt szükséges az annotálás, mert a címkék eleve adottak. Kutatási példáink alapján kitérek a predikciós modell hatékonyságának sajátos szociológiai értelmezhetőségére, végül a társadalomkutatásban kiemelt fontosságú, a modellek hatékonyságán túlmutató kérdésre: a modellek interpretálhatóságára.

5.1. Az annotálás kihívásai a szociológiai alkalmazásokban[11]

5.1.1. Motiváció

A felügyelt tanulás lényege, hogy előre bekódolt szövegek címkézését tanulja meg az algoritmus, jellegzetes szövegmintázatokat keresve. Ezeknek az ipari/üzleti alkalmazásokban már sokszorosan bizonyított algoritmusoknak a szociológiai alkalmazásai sajátos kérdéseket vetnek fel. A sajátosság oka, hogy ezekben az alkalmazásokban komplex fogalmak megtanulása az algoritmus feladata. A felmerülő kérdések: hogyan jön létre a címkézés? Hogyan lehet betanított kódolókkal elvégeztetni egy olyan hermeneutikai kihívást, mint a gyűlöletbeszéd felismerése? Segítenek-e ezen a rutinszerűen alkalmazott, részletezett annotálási irányelvek? Jobban végzi-e a kutató a besorolást, mint az egyszerű kódoló, vagyis magasabbrendű-e az egyik interpretáció, mint a másik? A fejezet arra is kitér, hogyan végzik crowdsourcing platformokon a kódolást a nagy cégek, hogy működik iparszerűen ez a humán/gép együttműködés, és milyen kérdések merülnek fel a crowsourcing interpretáció kapcsán. Végül röviden kitérek az AI-torzításra, aminek itt az a lényege, hogy a kódolók maguk viszik be a diszkriminációt az adatokba.

[11] E fejezet továbbfejlesztett változata a Metszetek c. folyóiratban megjelent, nyilvánosan elérhető cikknek: Németh, R. (2021). A felügyelt gépi tanulás kihívásai a szociológiai alkalmazásokban. *Metszetek-Társadalomtudományi folyóirat, 10*(3), 27–42.

5.1.2. A felügyelt gépi tanulás inputja: humán annotálás

Két karakteresen különböző megoldás létezik a felügyelt tanuláshoz szükséges tanuló-halmaz létrehozására. Az egyikre példa a korábban már idézett, Poletti és társai (2017) által publikált kutatás: ők kódolókat (szövegbányászati terminussal: annotátorokat) tanítottak be részletes irányelvek alkalmazásával arra, mikor minősítsenek egy szöveget agresszívnek vagy támadónak, mikor sztereotipizál és mikor irányul egy kisebbségi csoport ellen a szöveg. Itt tehát humán annotátorok vannak, akik olvasnak és interpretálnak, adott annotálási irányelveket követve.

A másik lehetőségre Jelveh, Kogut és Naidu (2014) írása példa, akik amerikai közgazdászok ideológiai pozíciójának (jobboldali/baloldali) gépi tanulását végezték el a szerzők tudományos írásai alapján. A tanuló-halmaz azokból a közgazdászoknak az írásaiból állt, akiknek pozíciója megállapítható volt külső adatokból: politikai kampány-támogatásokat ill. petíció-aláírásokat tartalmazó nyilvántartások alapján. (Vegyük észre, hogy ezek csak proxyk a politikai pozícióra, még ha ügyesen megválasztva is.) Itt tehát nincsenek humán annotátorok, nem olvasunk és nem interpretálunk, mert kész címkéink vannak.

A gyakorlatban leggyakrabban egy kutató (vagy egy kutatócsoport) kézzel végzi a kódolást, a szöveg annotálását, akárcsak a "klasszikus" kvalitatív szövegelemzés esetében. Az annotálás minőségének jelentősége kiemelkedő: a felügyelt tanuló algoritmus jó minőségben annotált tanuló-halmazból tud hatékonyan tanulni. Az annotálás persze időigényes és gyakran nem is egyszerű feladat. Az elsődleges cél a replikálhatóság, ami azt jelenti, hogy egy másik annotátor nagyon hasonló annotációkat készítene. Hovy és Lavid (2010) a következő strukturált eljárást javasolja az annotációk előállítására:

1. Határozzuk meg, hogy **milyen kategóriákba sorolva** kell annotálni. Ez általában valamilyen elmélet alapján történik (lásd: milyen jegyekkel definiálható a gyűlöletbeszéd). Itt megfelelő egyensúlyt kell találni a részletezettség/pontosság és a skálázhatóság/idővonzat között.
2. Választhatunk olyan **platform**ot, amely támogatja az annotálás adminisztrálását. Több általános célú annotációs eszköz érhető el.
3. Az annotálási feladatra vonatkozó utasítások formalizálása **annotálási irányelvek** formájában. Amennyiben az utasítások nem explicitek, a kapott annotációk szubjektív benyomásokon alapulnak majd, ami a replikálhatóságot veszélyezteti.

4. Az adatok egy kis részhalmazának kísérleti annotálása (**pilot**), több annotátorral. A pilot előzetes benyomást ad mind a megismételhetőségről, mind az annotálási irányelvek alkalmazhatóságáról. A megismételhetőséget az annotátorok döntései közötti egyezés (inter-annotator agreement) mutatóival jellemezhetjük. A konkrét annotálási eltérések vizsgálata segíthet az utasítások pontosításában, és az annotálási feladat módosításához is vezethetnek.
5. **Fő annotálás**. Érdemes legalább az adatbázis egy részét párhuzamosan annotálni, azaz két vagy több annotátorral egymástól függetlenül besoroltatni, hogy az annotátorok közötti egyezés kiszámítható legyen. Sok projektben a szövegek több címkét is kapnak, amelyek aztán összesítve egy "konszenzusos" címkévé állnak össze[12].
6. A felügyelt tanulás középpontjában az annotátorok közötti egyetértés áll: ha a kódok nem megbízhatóak, a tanuló algoritmus nem tud hatékonyan tanulni belőlük, és besorolásai sem lesznek megbízhatóak. Ezért elengedhetetlen az annotálás értékeléseként az **annotátorok közötti egyezés mutatójá**nak kiszámítása. Ha ez a mutató magas, az az annotátorok megbízhatóságát vagy magát a teljes annotációs rendszert (a besorolás értelmességét) kérdőjelezi meg. A mutató definíciójára több matematikai megoldás létezik a konkrét feladat függvényében. Az annotátorok besorolásainak egyszerű százalékos egyezése, annak ellenére, hogy nagyon széles körben használják, nem veszi figyelembe a véletlenszerűen előforduló egyezést. Ugyanis, ha két kommentátor véletlenszerűen választ két címke között, akkor a köztük lévő egyetértés várhatóan 50%-os lesz. Ezért egy jó mutató a nyers egyezési százalékot a véletlen egyezés arányához viszonyítja. Egy ilyen széles körben használt mutató a Cohen-féle kappa is.

Érdemes megjegyeznünk, hogy ha az általunk vizsgált eset nem túl specifikus, nem feltétlenül kell saját annotálást végrehajtanunk, használhatunk mások által annotált adatbázisokat is. Számtalan annotált, nyílt elérésű korpusz található az interneten, melyek elsősorban nyelvészeti feladatokra alkalmazhatók, de találhatunk Twitter bot-detektálásra szolgáló címkézett adatbázisokat is, sőt társadalomkutatási célúakat is, mint a Manifesto projekt, amely ötven ország politikai pártjainak választási programjainak annotált adatbázisát nyújta 1945-től napjainkig.

[12] Példaként lásd Danescu-Niculescu-Mizil et al. (2013).

5.1.3. Egy saját kutatási példa

Sik Domonkossal és Máté Fannival végzett kutatásunkban (Németh et al., 2020) mi is a fentihez hasonló eljárást dolgoztunk ki. Kutatásunk célja az volt, hogy különböző felügyelt tanulói algoritmusok alkalmazásával automatikusan osztályozzuk ismert nemzetközi online depresszió-fórumok bejegyzéseit aszerint, hogy abban a depresszió milyen (bio-medikális, pszichológiai vagy társadalmi) keretezését adja a felhasználó.

Az annotátorokat a téma és a módszer iránt érdeklődő társadalomtudományi szakos hallgatók közül választottuk ki. A teljes adatbázisból, ami 70 000 posztból állt, egyszerű véletlen mintavétellel 4500-at választottunk ki felcímkézendő tanuló-halmazként. Több tréninget tartottunk az annotátoroknak, és sok valós példát felvonultató, részletes annotálási irányelveket készítettünk, amelyet a pilot szakasz után is, és a fő annotálási szakaszban is folyamatosan frissítettünk. Öt címkét használtunk, mivel a három keretezés-típus mellé a "besorolhatatlan" (depresszióról van szó, de a keretezés nem azonosítható) és az "irreleváns" (nem depresszióról van szó) is hozzáadódott. Az annotálási feladat a legkevésbé sem volt triviális, ezért (1) az annotátorok szükség esetén két címkét rendelhettek a szövegekhez, egy elsődleges és egy (opcionális) másodlagos címkét. A bejegyzések 34%-a kapott második címkét legalább az egyik annotátortól. Továbbá (2) minden szöveghez két független annotátorunk volt, akik együttesen legfeljebb négy címkét adtak. A végső, konszenzusos címke „többségi szavazáson” alapult, tehát a négy címke közül a leggyakoribbat választottuk. A (nagyon kevés, 12,3%-os) kétértelmű esetek feloldására egy harmadik annotátort (egy vezető kutatót) kértünk fel. Másodlagos konszenzusos címkét is kiosztottunk, ha a „szavazásnak” volt egy egyértelmű második győztese.

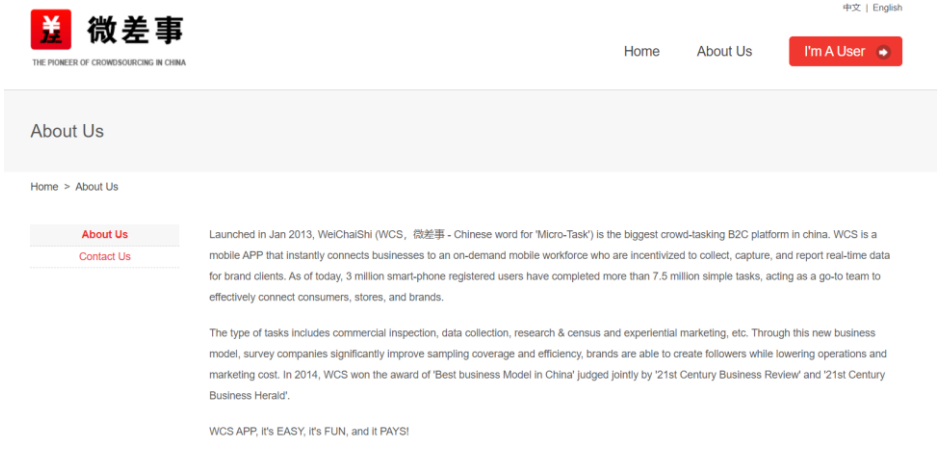
Az annotátorok közötti egyezés mutatójának meghatározására a Cohen-féle kappát használtuk, amely azt mutatja meg, hogy az annotátorok mennyivel jobban teljesítenek a véletlenszerűen besoroló annotátorokhoz képest. Tökéletesen egyező besorolásoknál az értéke 1, ezzel szemben, ha az annotátorok véletlenszerűen választják ki a címkéket, akkor a kappa egyenlő 0-val.

Annotációnk másik sajátossága az volt, hogy az annotátorok másodlagos címkéket rendelhettek a bejegyzésekhez. Ha az egyezést az elsődleges címkék egyezéseként határozzuk meg, egyszerűen elvetve a választható másodlagos címkéket, akkor egy túlságosan konzervatív mérőszámot kapunk. Ezért a kappa „liberális” változatát használtuk, az egyezést úgy definiálva, hogy az egyik elsődleges

címke megegyezik a másik annotátor által adott elsődleges vagy másodlagos címkével. A liberális kappa előnye az eredeti, konzervatív változatával szemben az, hogy figyelembe veszi a szövegek másodlagos jelentését is. Míg a liberális kappa túl optimista képet mutathat, addig konzervatív megfelelője túl szigorú értékeléshez vezet. Az "igazság" valahol a kettő között van, ezért mindkettőt bemutattuk. Konzervatív módon mérve az annotátoraink közötti egyezés 58,3%, a liberális mérőszám pedig 69,7% volt, ami elfogadható mértékű egyezést mutat.

5.1.4. Crowdsourcing annotálás

A tanulás sikerét közvetlenül befolyásolja a tanuló-halmaz mérete – hasonlóan ahhoz, ahogyan survey-ek esetén a mintanagyság a megbízhatóság megfontosabb faktora. Ezért nem ritkák a kifejezetten nagy (több tíz- vagy százezres) elemszámú annotálandó adatbázisok. Ezekben az esetekben a crowdsourcing platformokon bérelhető bedolgozó annotátorok jelentenek megoldást, például az Amazon által működtetett Mechanical Turk, a Figure Eight[13], Lighttag vagy a kínai Weichaishi (lásd 16. ábra[14]).

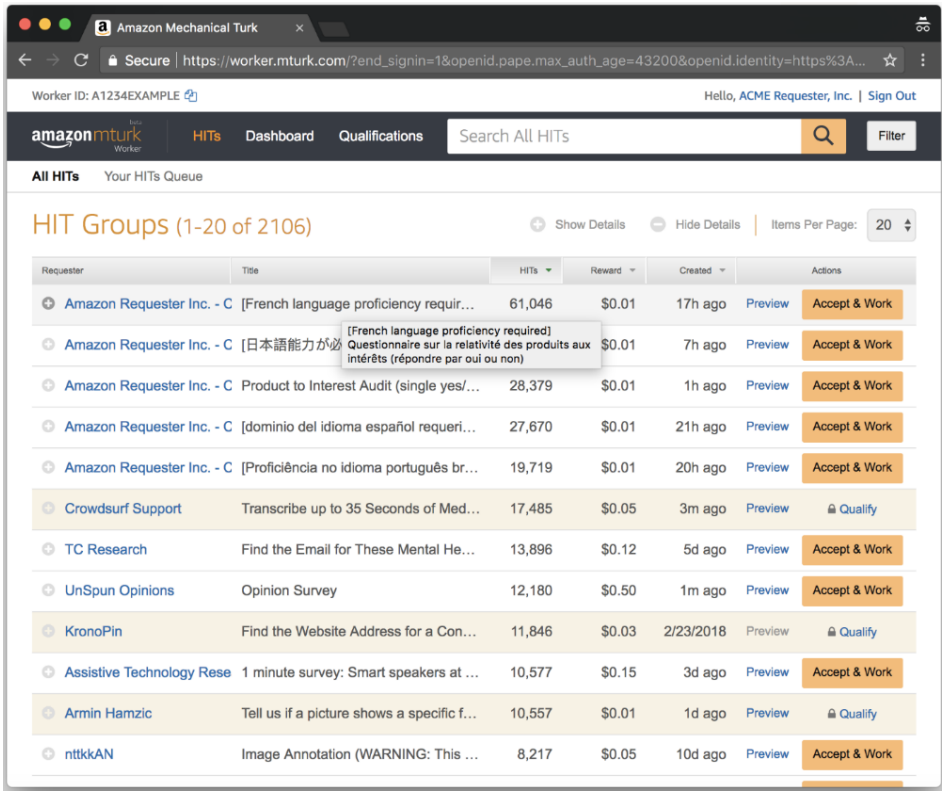


16. ábra. A WeiChaiShi kínai crowdsourcing cég beköszönő weboldala. A bedolgozók mobiltelefonján keresztüli adatgyűjtést, survey-eket és kísérleteket is kínálnak.

[13] Korábbi elnevezése CrowdFlower.

[14] Forrás: http://www.weichaishi.com/

A crowdsourcing platformon annotáltatni kívánó kutatók egyszerűen felhívást tesznek közzé a platform bedolgozói között, melyben az annotáció darabbérén kívül az annotátorok elvárt képességeit (minimális képzettségét, anyanyelvét, korábbi munkáikkal kapcsolatos elégedettségi arányt stb.) adják meg (17. ábra[15]). A viszonylag képzetlen „tömegmunkások” alkalmazása látszólag ellentétben áll a replikálhatóságra vonatkozó elvárásokkal, azonban számos, a platformot használó kutatás számol be megbízható annotációkról egyszerűbb feladatok kapcsán (pl. Snow et al., 2008).



17. ábra. Az Amazon Mechanical Turk egy oldala a felkínált munkákkal, a munkaaadók (requesters) által fix darabbérért kínált un. HIT-ekre (Human Intelligence Tasks), elemi munkaegységekre osztott feladataival.

[15] Forrás: blog.mturk.com, https://blog.mturk.com/quick-update-another-improvement-to-the-mturk-worker-experience-9cfd0b1963e7

A crowdsourcing egy tágabb terület, az ember-alapú számítás (*human-based computing*, lásd még: elosztott gondolkodás, *distributed thinking*) részét képezi. Ez olyan számítástechnikai megoldásokat foglal magába, amelyekben egy számítógép úgy látja el a funkcióját, hogy bizonyos lépéseket embereknek szervez ki, egy szimbiózisszerű interakcióban (Mühlhoff, 2019). Megfordulnak a szerepek: a gép kér fel embereket egy probléma megoldására, majd integrálja a megoldásokat. Az olyan számításigényes feladatok esetében, mint a képfelismerés, a humán annotáció fontos részét adja a mélytanuló algoritmusok tanításának.

A crowdsourcinggal kapcsolatban komoly etikai kérdések merülnek fel. Fort és Cohen (2011) már tíz éve arra figyelmeztetett, hogy az Amazon Mechanical Turk többszázezres bedolgozói tömeget működtet online, szerte a világban, jó részük él Indiában és Törökországban. A jellemző órabérek 2 dollár alatt vannak, és az általános elképzeléssel szemben a felhasználók jellemzően nem kismamák vagy diákok, aki hobbiként dolgoznak itt, hanem olyan munkavállalók, akik a megélhetéshez szükséges forrásként tekintenek munkájukra. Az alacsony béren kívül további komoly problémát jelent, hogy a munkavállalók (azaz „turkers") híján vannak az olyan alapvető munkahelyi jogoknak, mint a kollektív alku, a szakszervezet alakításának lehetősége, a munkáltatói jogsérelmek orvoslásának lehetősége, ezért rendkívül kiszolgáltatottak.

Egy friss munka (Gray & Suri, 2019) gazdaságantropológiai megközelítésben vizsgálja ezeknek a láthatatlanul dolgozó "szellemmunkásoknak" a körülményeit. A szerzők szerint az átláthatóság hiánya azt a benyomást kelti a közvéleményben, hogy a mesterséges intelligencia egyedül működteti a kortárs kényelmi szolgáltatásokat, miközben számos ponton elengedhetetlen az emberi beavatkozás: a szellemmunkások az annotáláson túl számos támogató munkát (személyazonosítás, feliratozás stb.) végeznek olyan óriások számára, mint az Amazon, a Google, az Uber és a Microsoft. A munkával járó anonimitás és rugalmasság kétségtelenül sokuknak előnyös, pl. az egyébként gyakran diszkriminált munkavállalói csoportoknak (nők, fogyatékkal élők). Ugyanakkor Fort és Cohen (2011) korábbi megfigyelései még mindig érvényesek, azzal a különbséggel, hogy ma már dolgozók millióira vonatkoznak: a szellemmunkások rendkívül alulfizetettek, kiszolgáltatottak, munkavállalói jogaik erősen sérülnek. A szerzők javaslatainak (munkavállalói juttatások bevezetése, szakszervezet alakítása, a munkavállaló és munkáltató közötti emberi kommunikációt lehetővé tevő új központok létrehozása) tétje nagy: a méltányos emberi munka jövőjét biztosítanák.

5.1.5. A humán annotálás kihívásai szociológiai alkalmazásokban

Az ipari/üzleti alkalmazásokban már sokszorosan bizonyított felügyelt gépi tanulás szociológiai alkalmazásai sajátos kérdéseket vetnek fel. A sajátosság oka, hogy ezekben az alkalmazásokban komplex fogalmak megtanulása az algoritmus feladata (lásd: gyűlöletbeszédet tartalmaz-e egy tweet), szemben az olyan könnyebben annotálható feladattal, hogy negatív vagy pozitív-e egy szolgáltatással kapcsolatos bejegyzés.

Kutatásunk (Németh et al., 2020) már idézett esete jól példázza ezt: a depresszió keretezésének eldöntése nem bizonyult egyszerű feladatnak. A pilot során, az annotálási irányelvekben felsorolt elveken alapuló értelmezést több egyéni és csoportos fordulóban gyakorolták az annotátorok. Mégis, a pilot szakasz többszöri meghosszabbítása után is elégtelen maradt az annotátorok közötti egyetértés (az elsődleges címkék százalékos egyezése 60% alatt volt). Ezen a ponton vált időszerűvé saját módszertani háttérfeltevéseink felülvizsgálata. Rá kellett jönnünk, hogy az általános kategóriák (keretezés típusok) és a konkrét hozzászólások ilyen jellegű társítása nem egyértelmű feladat. Gadamer hermeneutikai elméletét (2004) követve innét kezdve a jelentések kialakulását nyelvileg közvetített, interszubjektív, iteratív folyamatnak tekintettük, amelyben a jelentések ténylegesen egy folyamat során konstruálódnak. Módszertanilag ezt az interszubjektív folyamat az annotálási irányelvek folyamatos, iteratív frissítésében jelent meg. Ez a folyamat lényegesen különbözik azoktól a hermeneutikailag egyszerűbb üzleti alkalmazásoktól, amelyek explicit és egyértelmű kategóriák (lásd pozitív/negatív/neutrális szentiment) előre meghatározott készletét alkalmazzák. A mi eljárásunk inkább egyfajta kvalitatív kódolásként határozható meg, mivel a kategóriáinkat egy előzetes absztrakt elméletből származtattuk, és induktív módon alakítottuk ki őket a kutatás során.

Az interszubjektivitás elismerésének másik megnyilvánulása a kettős annotációra való áttérés volt. Ahelyett, hogy azt feltételeztük volna, hogy minden egyes hozzászólás egy vagy két kategóriába tartozik, amelyeket egy megfelelően képzett annotátor azonosítani tud, úgy közelítettük meg a fórumbejegyzéseket, mint amelyek többféleképpen értelmezhetők. Annak érdekében, hogy minimalizáljuk az értelmezés esetlegességét, két független annotátorral kódoltattunk minden posztot. A végső, konszenzusos címke a két annotátor kódjának egyesítésén alapult a fent leírt módon.

E fejezet későbbi szakaszában ismertetem azt a kutatásunkat, amely ennek az annotációnak az eredményére épült. Különböző felügyelt tanulási modelleket alkalmaztunk az annotáció kiterjesztésére a teljes korpuszon. Amit már itt fontos kiemelni: tapasztalataink szerint a "diszkurzív keretezés" összetett és hermeneutikailag nehéz fogalomnak bizonyul, ami nem csak az annotátorok közötti egyetértést befolyásolta, hanem a tanuló algoritmus teljesítményét is. Tapasztalatunk szerint az annotátorok közötti nézeteltérés mértéke jó becslést ad a tanuló objektív nehézségére.

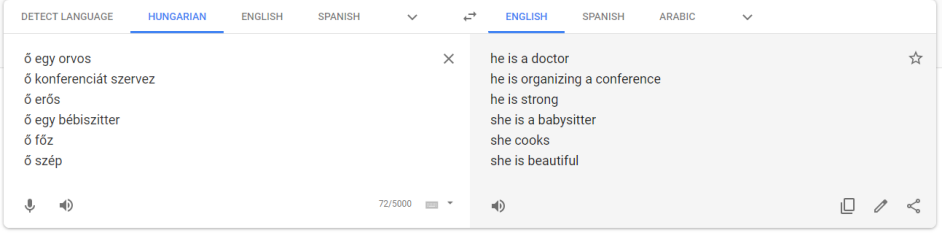
5.1.6. A Mesterséges Intelligencia torzítás és az annotálás

A Mesterséges Intelligencia torzítás (MI-torzítás, angolul AI-bias, az Artificial Intelligence rövidítéseként) lényege, hogy a nyelvtechnológia outputja maga is társadalmi torzításokat (pl. kisebbségekkel, idősekkel vagy nőkkel szembeni hátrányos megkülönböztetést) mutat, amellyel mintegy felerősíti azok társadalmi hatását (Ntoutsi et al., 2020). Az egyik legismertebb példa az Amazon kísérleti rekrutációs algoritmusának esete (Dastin, 2018). Az algoritmus létrehozásának célja az volt, hogy az 1-től 5-ig automatikusan osztályozza önéletrajzuk alapján a jelentkezőket. 2015-re azonban kiderült, az algoritmus gender-alapon torzított a szoftverfejlesztő és más technikai jellegű pozíciók esetén. Ennek oka az volt, hogy a modellt a céghez a megelőző 10 évben benyújtott jelentkezések alapján tanították, de a technológiai iparág férfi-dominált, így a jelentkezések is férfiaktól jöttek elsősorban. Emiatt az MI rosszabbul minősítette azokat a jelentkezéseket, amelyekben a „női” szó szerepelt, legyen az akár csak a „női sakkcsapat kapitány” kifejezés. A fejlesztők megkísérelték oly módon javítani, hogy az ilyen fajta nemi információtól való függést kiiktatták a tanulóból, de ez nyilván nem feltétlenül jelenti a probléma megoldását, hiszen egy ilyen rendszer más, nem vizsgált dimenziók mentén is tartalmazhat torzítást. A fejlesztők végül felhagytak ennek az algoritmusnak a használatával.

Az MI-torzítás általánosan megfogalmazva egyfajta anomália a gépi tanuló algoritmus outputjában; vagy az algoritmus fejlesztési folyamatában alkalmazott feltevésekre vagy a tanuló-halmazban rejlő inherens társadalmi torzításokra vezethető vissza. Az annotálás kapcsán a továbbiakban az utóbbira koncentrálunk.

A tanuló halmaz, ember alkotta élő szöveg lévén mindazokat a viszonyokat tükrözi vissza, amik a társadalomban is megtalálhatók. Az Amazon példáján: a tanuló-halmaz kiegyensúlyozatlan volt nemi összetételét tekintve. Hasonló

ismert példa a Google Translate esete is. Farkas Annával (Farkas & Németh, 2020) végzett kutatásunkban azt vizsgáltuk, hogy ha magyarról angolra fordítunk foglalkozásneveket tartalmazó mondatokat („Ő egy orvos”), akkor a fordító hímnemű vagy nőnemű névmást használ, s hogy ez a döntés korrelál-e a foglalkozások tényleges nemi megoszlásával illetve a magyarok survey-jel mért attitűdjével (azaz azzal, hogy inkább férfias vagy nőies foglalkozásnak vélnek valamit a magyarok). Eredményünk szerint a fordító erősen torzít a nőkkel szemben, és működése közelebb áll az attitűdökhöz, mint a tényleges foglalkozásszerkezethez (18. ábra[16]).



18. ábra. A Google Translate nemi torzítása (2020 augusztusi állapot).

A fordító torzításának oka itt is az, hogy kétnyelvű szöveghalmazokon tanul, s ha a társadalmi attitűd vagy a tényleges foglalkozásszerkezet inkább férfiasnak mutat egy foglalkozást, akkor az a szövegekben is megmutatkozik majd, így a fordító is visszatükrözi ezt az egyenlőtlenséget. (A Google mai működésében már nem detektálható ez az anomália, a visszajelzések nyomán úgy módosították az algoritmust, hogy mindkét névmást megadja a fordítás.)

Az annotáción alapuló torzítás a tanuló-halmaz szintjén vihet anomáliát a rendszerbe, kétfajta úton. Vagy az annotátorok, maguk is társadalmi normákat képviselve, tükrözik vissza a címkézésben a társadalmi viszonyokat, vagy azzal okoznak torzítást, hogy nem reprezentálják kellőképpen a társadalom egészét. Mindkét probléma egyszerre volt jelen abban a gyűlöletbeszéd-kutatásban (Sap et al., 2019), ahol a torzítást az okozta, hogy a jobbára fehér bőrű annotátorok a

[16] Forrás: Google Translate képernyőkép: https://translate.google.hu/?hl =en&tab=wT#view=home&op=translate&sl=hu&tl=en&text=%C5%91%20egy%20orvos%0A%C5%91%20konferenci%C3%A1t%20szervez%0A%C5%91%20er%C5%91s%0A%C5%91%20egy%20b%C3%A9biszitter%0A%C5%91%20f%C5%91z%0A%C5%91%20sz%C3% A9p (létrehozva: 7 August 2020)

címkézendő tweetek szerzői által használt afroamerikai dialektust hajlamosabbak voltak offenzívként megjelölni. E dialektus ugyanis több káromkodást tartalmaz, és még ha az nem is irányul más felhasználók ellen, akkor is sértőnek, offenzívnek érezték a más nyelvi normákat képviselő kódolók.

Eltávolítható-e a torzítás az algoritmusokból? Ahogy láttuk, az MI algoritmusok épp annyira lehetnek kiegyensúlyozottak bármilyen szempontból, mint amennyire a tanuló-halmaz az. „Reprezentatív" annotátor-csapat és „reprezentatív" tanuló minta kialakítására érdemes törekedni ugyan, de tökéletes reprezentativitás még elvileg is nehezen elképzelhető. Felmerül annak kérdése is, hogy valóban a nagy átlagot kell-e az algoritmusnak hoznia, nem lehet-e egy-egy kisebbségi csoport véleménye relevánsabb – gondoljunk itt akár egy tényellenőrzési feladatra, ahol a „tények" detektálása nem feltétlenül reprezentatív szavazással dönthető el.

Megkísérelhetjük továbbá nem az input, hanem az output oldalon is a javítást: a torzítás eltávolítását egy-egy dimenzió mentén (ahogy az Amazon fejlesztői tették), de számtalan látens dimenzió létezhet, ezért tökéletesen torzítatlan MI algoritmus nehezen elképzelhető – ugyanakkor törekedni kell tesztelésükre és javításukra.

5.2. Egy kísérlet: párhuzamos emberi kódolás és felügyelt tanulás – egy diszkriminációkutatási példán

Ez az alfejezet a digitális társadalomkutatás egyik fontos lehetőségéről szól, nevezetesen arról, hogy a szövegek kvalitatív emberi kódolása hogyan helyettesíthető vagy egészíthető ki gépi tanulással. A fejezet két cikkünkre hivatkozik majd, melyek ugyanannak a kutatásnak ezt a két különböző (kvalitatív és gépi tanulási) elemzését tartalmazzák (Simonovits et al., 2022; és Buda et al., 2022).

A kutatás során kontrollált terepkísérletre támaszkodva mértük fel, hogy a magyar önkormányzatok mennyire reagálnak eltérően roma és nem roma ügyfelek információkéréseire. Országos e-mailes vizsgálatot végeztünk az önkormányzatok megkeresésével, ahol az ügyfél „vélt roma" besorolását sztereotipikus roma hangzású kereszt- és családnevekkel értük el. A diszkrimináció megközelítésénél a figyelemdiszkrimináció elméleti keretét használtuk, melyet eredetileg Matejka (2013) és Bartoš et al. (2016) dolgozott ki, összekapcsolva a diszkrimináció és a korlátozott figyelem fogalmát.

Az emailekben egyszerűen megválaszolható kérdések voltak négy variációban, (környékbeli kerékpártúráról, az elérhető óvodákról, a helyi temető

nyitvatartási idejéről ill. egy esküvő lehetséges helyszíneiről). Véletlenszerűen variáltuk a feladó állítólagos etnikai hovatartozását és a kérés jellegét is, 1260 emailt küldve ki. Eredményeink szerint a románok vélt ügyfelek jelentősen (mintegy 13 százalékponttal) kisebb valószínűséggel kaptak választ.

Ha érkezett válasz, annak szövegét is elemeztük mind humán kódolók, mind gépi tanulás segítségével – és jelen megközelítésünk szempontjából a vizsgálatnak ez a nyelvi elemzési szakasza izgalmas igazán. Itt fontos szem előtt tartani, hogy (nyilván) az elsődleges diszkrimináció a nem-válaszolásnál keletkezik, a szövegekben esetleg megbújó eltérés csak erre rakódik rá.

Az e-maileket kvalitatív módon, udvariasságuk és segítőkészségük alapján kódoltuk annotátorok segítségével. Az udvariasságot úgy határoztuk meg, hogy a válasz mennyire volt tiszteletteljes, míg a segítőkészség azt jelentette, hogy a válasz milyen mértékben nyújtott információt. Mindkét tényező tekintetében statisztikailag szignifikáns, de nagyságrendjét tekintve mérsékelt eltérést találtunk: a vélt roma-nem roma csoport közötti különbségek 100 pontos skálán 5 pont nagyságrendűek voltak.

A kutatás második szakaszában azt vizsgáltuk, hogy (1) lehetséges-e a szöveges adatokban automatizált módon, emberi kódolás nélkül felismerni a megkülönböztetést és hogy (2) a gépi tanulás milyen szövegjellemzők szerint lát különbséget a két csoport között, ill. képes-e esetleg felismerni a megkülönböztetés olyan jellemzőit, amelyekre az annotátorok képzésénél alkalmazott irányelvek nem terjednek ki.

Az 1. kérdés megválaszolásakor tulajdonképpen arra kerestünk választ, hogy képes-e a tanuló algoritmus megkülönböztetni a roma és nem roma ügyfeleknek írt válasz-emaileket? Azt mondhatjuk, hogy ha modelljeink a véletlenszerű osztályozáshoz képest statisztikailag szingifikánsan jobban működnek, akkor található a szövegekben olyan jellemző, ami a vélt romákkal ill. nem-romákkal szemben eltérő bánásmódra utal. Erről a sajátos logikáról, vagyis arról, hogy a diszkrimináció meglétét és erősségét tanuló algoritmus teljesítmény-mutatójaként definiáltuk, a következő fejezetben írok bővebben más módszertani alternatívákkal összevetésben.

Többfajta modellt definiáltunk. Az első típus magyarázó változói az emailek leíró statisztikai jellemzői voltak (pl. milyen hosszú, mennyire összetett a válasz nyelvileg, milyen sok “kemény” információt (pl. számokat) tartalmaz stb.). A második modell-típus közvetlenül az e-mailek szövegét használta fel. Mindkét modellhez XGBoost algoritmust használtuk, mivel tapasztalatok szerint ez egy általánosan jól teljesítő predikciós módszer. A harmadik modell e két

modell kombinációja volt, logisztikus regressziós algoritmussal (un. ensemble modell, amely a két modell predikcióját egyesíti egyetlen közös predikcióvá).

Eredményünk szerint modelljeink a véletlenszerű osztályozáshoz képest jelentősen jobban működtek (a legjobb modellünk pontossága 61% volt), ami megerősíti a roma ügyfelekkel való eltérő bánásmód meglétét. De hogyan kell értelmezni az eltérő bánásmódot? Ha a tisztviselők másképp írnak a roma ügyfeleknek, az nem feltétlenül jelent negatív diszkriminációt: lehet, hogy például túlkompenzálnak. Ez az értelmezési bizonytalanság az oka annak, hogy célunk a második kutatási kérdés révén nemcsak a megkülönböztetés jelenlétének kitapintása volt, hanem annak interpretációja is. Vagyis annak megválaszolása, hogy mely magyarázó változók játszanak szerepet a predikciós modelljeinkben, mely szövegmintázatok korrelálnak erősen a roma vs. nem roma ügyfeleknek adott válaszokkal. Erről a módszertani problémáról, vagyis arról, hogy nem csak a modell teljesítményét vesszük figyelembe, hanem azt is, hogy leginkább mely szövegjellemzőkre támaszkodva végzi az osztályozást, alább, *A predikciós modell fekete dobozának felnyitása* c. fejezetben írok.

A legfontosabb prediktorok azt mutatták, hogy a vélt roma ügyfeleknek küldött válaszok egyrészt rövidebbek, másrészt kevesebb információt tartalmaz (pl. kisebb valószínűséggel kaptak telefonszámot), és hangnemük is formálisabb/visszafogottabb (pl. kisebb valószínűséggel szólítják őket nevükön, inkább a formális „tisztelt címzett" vagy „hölgyem" a megszólítás). Vagyis a modellek részletesebb vizsgálata arra utal, hogy a vélt roma ügyfelekkel szemben figyelemdiszkrimináció van jelen, abban az értelemben, ahogyan azt Bartos és munkatársai (2016) definiálják.

Röviden összefoglalva a két párhuzamos kutatás eredményét: mind a kvalitatív módon kódolt, mind a gépi tanulással feldolgozott adatokon ki tudtunk mutatni diszkriminációt. Vagyis a szöveges adatokon automatizált módon, emberi kódolás nélkül, tehát gyorsabban és olcsóbban is fel lehet ismerni a rejtett megkülönböztetést. A gépi tanuló részben hasonló (udvariasság, informativitás) részben más szövegjellemzők (távolságtartás) szerint látott különbséget a két csoport között, vagyis a megkülönböztetés olyan jellemzőit is felfedezhetjük segítségével, amelyekre esetleg *a priori*, az annotátorok instruálásánál nem gondolnánk.

A randomizált, kontrollált vizsgálat a legerősebb oksági bizonyítékot nyújtó design. Megmutattuk, hogy ha digitális környezetben alkalmazzák, az így keletkező hatalmas mennyiségű szöveges adatot NLP-technika segítségével hatékonyan lehet feldolgozni. Ugyanez a módszer alkalmazható lehet más kutatási design-ra, akár megfigyeléses vizsgálatokra is.

5.3. A felügyelt tanuló teljesítménye, mint szociológiailag értelmezhető fogalom operacionalizációja

Ebben a fejezetben két olyan kutatásunk tapasztalatait összegzem, amikor nem volt szükségünk annotálásra, mert a szövegek címkéi eleve adottak voltak a teljes korpuszra. A két kutatást az köti össze, hogy a felügyelt tanuló teljesítménye van a középpontban, mégpedig valamely szociológiailag értelmezhető fogalom operacionalizációjának eszközeként.

5.3.1. A predikciós teljesítmény, mint diszkrimináció-mérték

Erről a kutatásunkról (Buda et al., 2022) az előző fejezetben írtam már, itt most arra az aspektusára helyezem a fókuszt, hogy a diszkriminációt, mint predikciós problémát, a predikciós teljesítményt pedig mint diszkrimináció-mértéket kezeltük. Legjobb tudomásunk szerint a mi tanulmányunk volt az első kísérlet a diszkrimináció gépi tanulási technikákkal történő értékelésére.

A hagyományos diszkriminációkutatásban a diszkrimináció mértékét a többségi és a kisebbségi csoportok közötti, valamilyen numerikus változó szerinti átlagos különbségként szokás meghatározni. Például: milyen arányban hívták be a munkainterjú második fordulójára a munkára jelentkezőket a két csoportban, vagy milyen átlagos különbség mutatkozik a számértékű változóval mért udvariasság tekintetében két csoportnak írt válaszlevél között. Ezzel szemben a diszkrimináció jelenléte a mi megközelítésünkben egy olyan modell létezésével azonosítható, amely bizonyos hatékonysággal jósolja meg a (vélt) etnikai hovatartozást. Mit jelent a „*bizonyos hatékonyság*”? Nem kell itt nagyon magas hatékonyságra gondolni, vagyis nincs szükségünk olyan modellre, amely a válasz-emailek többségét a megfelelő kategóriába sorolja, hiszen ez azt jelentené, hogy minden tisztviselő diszkriminatív. Ehelyett elegendő, ha olyan előrejelzési pontosságot érünk el, amely statisztikailag szignifikánsan magasabb, mint a véletlen besorolás pontossága. (Az előrejelzési pontosság a jó helyre besorolt esetek arányát jelöli, 0-1 intervallumon értelmezett mutató). Nyilvánvaló, hogy ha az ügyfelek etnikai hovatartozása nem befolyásolná a válaszokat, akkor az osztályozási modell pontossága nem haladná meg a véletlenszerű osztályozását.

Ebben az összefüggésben magát a prediktív pontosságot definiálhatjuk diszkriminációs mértékként. (A standard prediktív modellezési keretrendszert követve a pontosságot természetesen nem a tanuló, hanem a teszt halmazon mértük, hogy a

túlillesztést elkerüljük). Kutatásunkban a legjobb predikciós modell a véletlenszerű osztályozáshoz képest statisztikailag szignifikánsan jobban működött, 61%-os pontossággal, erre (és a modell interpretációjára) alapozva arra jutottunk, hogy adataink statisztikailag bizonyítják a roma ügyfelekkel szembeni eltérő bánásmód jelenlétét.

A kutatás módszertani újdonsága tehát az volt, hogy a diszkriminációt predikciós problémaként kezeltük, és azt vizsgáltuk, hogy a vélt roma és nem roma ügyfeleknek írt e-maileket milyen mértékben lehet megkülönböztetni. Ez a megközelítés lehetővé tette egy diszkriminációs mérőszám meghatározását: minél nagyobb a predikciós modell képessége a (vélt) etnikai hovatartozás azonosítására, annál magasabb a diszkrimináció szintje.

5.3.2. Predikciós teljesítmény, mint a politikai polarizáció mértéke

Az *NLP a politikai polarizáció kutatásában* c. fejezetben részletezett összefoglaló tanulmányomban (Németh, 2023) a 2010 óta a témában megjelent tanulmányok (n = 154) módszertani áttekintését végeztem el annak tisztázására, hogy az NLP-kutatások hogyan konceptualizálták és hogyan mérték a politikai polarizációt. Eredményeim szerint minden harmadik tanulmány felügyelt gépi tanulást (osztályozást) alkalmazott a politikai ideológia/álláspont előrejelzésére.

Az osztályozás ebben az összefüggésben olyan módszer, amely a szerző politikai álláspontját próbálja azonosítani az általa használt szavak alapján. Ezek a gépi tanulást alkalmazó tanulmányok a politikai polarizációt vagy expliciten, vagy implicit módon, de osztályozási problémának tekintik. Itt a gépi tanuló magas predikciós hatékonysága arra utal, hogy az egyik politikai oldal által használt nyelv egyrészt egyrészt valamilyen szinten homogén, másrészt felismerhető mintázatok szerint különbözik a másik oldal által használt nyelvtől. Ahogyan a gépi tanuló segítségével az előző fejezetben diszkriminációs metrika volt meghatározható, most hasonló módon polarizációs metrika definiálható: minél jobban képes az osztályozási modell azonosítani a szerző álláspontját, annál nagyobb fokú a politikai polarizáció.

Áttekintésem (Németh, 2023) egyik ide tartozó kutatása Bayram et al. (2019), akik ezt a megközelítést követve elemezték az Egyesült Államok Kongresszusa Képviselőházának felszólalásait. Az időben előre haladva egyre jobb hatékonyságú predikciós algoritmusokat tudtak alkotni, amit úgy interpretáltak, hogy a politikai polarizáció egyre jobban detektálhatóvá vált a beszédek nyelvezetében. Gentzkow és társai (2019) hasonló logikát követő polarizációs mérőszámot javasoltak.

Saját esettanulmányként a Magyar Országgyűlésben elhangzott beszédeken végzett elemzésünk (Buda et al., 2023) tapasztalatait említeném. 1998-2018 között, öt parlamenti ciklusra illesztettünk predikciós (XGBoost) modellt, csak a Fidesz és az MSZP szónokainak megkülönböztethetőségét vizsgálva. Eredményünk szerint a ciklusonként külön illesztett modell predikciós pontossága folyamatosan javult az öt ciklus alatt (0,76-ról 0,87-re), vagyis a politikai polarizáció erősödő tendenciáját tapasztaltuk.

5.4. A predikciós modell fekete dobozának felnyitása

5.4.1. Motiváció

Ahogy korábban már említettem, egy predikciós modell interpretációjakor/magyarázatakor a cél tulajdonképpen azoknak az (itt: szövegmintázattal kapcsolatos) tényezőknek a feltárását jelenti, amelyek az egyébként "fekete dobozként" működő predikciós algoritmus mögött állnak. Lásd az előző két fejezet példáját: a modell valamilyen hatékonysággal képes a vélt roma és nem-roma ügyfeleknek írt emailek között különbséget tenni, de hogyan fogható meg ez a különbség? Vagy a polarizáció példáján: az időben előre haladva egyre könnyebb megkülönböztetni a Fidesz és az MSZP parlamenti felszólalóinak beszédét, de milyen szövegmintázatok azok, amelyek ezt a különbségtételt lehetővé teszik?

Röviden érdemes itt kitérni az előrejelzés és magyarázat kettőségére (részletesen lásd Németh, 2021), hiszen míg az üzleti és alkalmazott kutatásokban inkább az előbbi a cél (spam email detektálása, képalkotó diagnosztikai eljárás outputjának besorolása rosszindulatú/jóindulatú kategóriába stb.), a társadalomtudományok számára az utóbbi az, ami fontos, ami az elméletalkotáshoz, a tudástermeléshez hozzá tud járulni. Az adattudós üzleti felhasználó olyan modellt épít, amelynek csak az előrejelzési teljesítménye számít, és nem feltétlenül okoz neki gondot, ha a modell más szempontból fekete dobozként tekinthető.

A gépi tanulási algoritmusok készítőinek az interpretációt másodlagos célként kezelő attitűdje részben érthető is, ugyanis a ma alkalmazott gépi tanulási algoritmusok interpretálása komoly kihívás. A klasszikus, egyszerű modellek, amilyen a társadalomkutatásban is ismert regressziók, közvetlenül interpretálhatók: expliciten adott, hogy melyik magyarázó változó milyen előjellel, milyen súllyal, a többi változóval alkotott milyen interakciójával vesz részt egy predikció

létrehozásában. Ezzel szemben pl. a mélytanulási (*deep learning*) modellek működése kevésbé átlátható, mert nincsen közvetlen, jól leírható függvénykapcsolat az input és a predikció között, ugyanakkor interpretálásuk nem megoldhatatlan. Az adattudomány legújabb munkái (összefoglalóként lásd pl. Molnar 2020) éppen a prediktív modellek értelmezhetőségét, a fekete doboz kinyitását célozzák. Ezek közvetett módszerek abban az értelemben, hogy nem adnak (és nem is adhatnak) a regressziós modellekhez hasonló, közvetlenül értelmezhető kapcsolatot a bemenet és a kimenet között.

A modellek interpretálhatósága nem csak azért fontos, mert az értelmezés segít eldönteni, hogy az előrejelzések értelmesek-e, hanem azért is, mert felfedhetjük, hogy melyek azok a tényezők, amelyek a legnagyobb hatással vannak a modell döntéseire, amelyek csökkentik vagy növelik bizonyos osztályba sorolás valószínűségét, illetve, hogy milyen hatása van az a priori fontosnak tartott tényezőknek. Ezeket a modellműködéssel kapcsolatos megállapításokat az értelmezés során a szakterületi tudás kontextusába helyezhetjük, kapcsolódva a meglevő tudományos diskurzushoz, vagy ahogyan az üzleti elemzésekben fogalmaznak: *insight*-ot nyerjünk adatainkból. (Itt zárójelben megjegyezném, hogy valójában a tudományos diskurzushoz való kapcsolódás már az elemzés első szakaszában meg kell, hogy történjen, a korpusz, a modellek, a kutatási kérdés meghatározásakor).

Az interpretálhatóságnak nem csak a tudományos, de az üzleti alkalmazásokban is fontos, az utóbbi években felismert szerepe van. Ugyanis a predikciók is kevéssé megbízhatóak, ha nem tudjuk, az üzleti környezet mely paramétereire épülnek – így az üzleti/társadalmi környezet megváltozása a predikció romlását hozhatja, anélkül, hogy ennek tudatában lennénk. Ezzel hozható leginkább összefüggésbe a 2010-es évek közepének üzleti csalódásai a „big data”-ban: a korábbi hype-ot sok esetben sikertelen üzleti projektek hűtötték le. A sikertelenség oka pedig legtöbbször az algoritmikus megoldások sterilitása volt, a modellinterpretáció elmaradása, azaz a szakterületi tudás be nem építése az analitikai folyamatba. Ennek következményeként került előtérbe az *insight*, azaz az üzleti jelenség mélyebb ismerete a predikciók értelmezéséhez. E fiaskók egyik következménye volt a sok adat helyett a releváns adat megkövetelése is („*big data*” helyett „*smart data*”). Minderről bővebben korábbi összefoglaló munkámban írtam (Németh, 2021, pp. 110-122).

Ugyanez a jelenség az algoritmusokkal támogatott társadalomkutatások esetén is megfigyelhető. A hatalmas adatforrás megjelenése önmagában nem hozott átütő elemzéseket, mert a kutatások jelentős része szociológiai tudás nélkül jön létre. Számos olyan kutatás lát napvilágot, ahol a társadalmi problémákra nem

reflektáló algoritmikus megoldás légüres térben született problémára keres választ, és az eredménye nem kapcsolódik a tudományos diskurzushoz, nem ad hozzá a szociológiai tudáshoz. Például a politikai polarizáció NLP-s megközelítéseit áttekintő kutatásom (Németh 2023) szerint a polarizáció mérésére predikciós modellt alkalmazó öt tanulmányból három csak a predikciós teljesítmény optimalizálására összpontosít, és nem tér ki az interpretációra, azaz arra, hogy milyen nyelvi jellemzők játszottak szerepet a szövegek politikai ideológia szerinti szétválasztásában. Pedig egyrészt az interpretáció a modell validálásában is megkerülhetetlen: akkor tekinthető validnak a tanuló, ha plauzibilis jellemzőket használ a besorolás során. Másrészt, ha megértjük, hogy a modell miért hozott egy bizonyos döntést, és megtaláljuk azokat a kifejezéseket, amelyek a leginkább utalnak pl. a konzervatív vs. liberális álláspontokra, közelebb kerülünk a politikai polarizáció megértéséhez is.

5.4.2. Modell-interpretáció a depressziós fórumok elemzésekor

Saját, felügyelt gépi tanulásra épülő kutatásainkat a fenti tapasztalatok miatt már tudatosan a modellek interpretációjának beépítésével terveztük. Alább módszertani tapasztalatainkat összegzem, a tartalmi eredmények közül szokott módon csak egyet-egyet válogatva illusztrációként.

A depressziós online fórumok elemzésekor (Németh et al., 2022) az volt a gépi tanuló célja, hogy a posztok egy véletlen, ~4500 elemű részhalmazának annotátorok által elvégzett kódolását (miszerint bio-medikális, pszichológiai vagy társadalmi a posztban a mentális állapot keretezése, lásd e fejezetben korábban az annotálás kihívásai kapcsán) kiterjesszük a teljes ~80.000-es korpuszra. Hagyományos szózsák-modelleket, XGBoost algoritmust, szóbeágyazás-alapú modelleket és egy transzfer tanulást alkalmazó DistilBERT modellt alkalmaztunk.

A legjobb teljesítményt nyújtó DistilBERT modellünk is azok közé a modellek közé tartozik, melyeknek nem triviális az interpretációja – mi a SHAP érték (Lundberg, Lee, 2017) használata mellett döntöttünk. Egy szövegjellemző (*feature*) fontosságának értékei egyrészt hatásának nagyságával (mennyire erősen befolyásolja a jellemző az előrejelzést), másrészt irányával (pozitív vagy negatív - a jellemző növeli vagy csökkenti az előrejelzés valószínűségét) jellemezhető. A SHAP éppen ezt nyújtja: minden egyes jellemzőhöz hozzárendel egy fontossági értéket egy adott előrejelzéshez, amely alapján fontossági sorrendbe lehet állítani a jellemzőket, és a SHAP-érték előjele alapján következtetni lehet a jellemző hatásának jellegére. Összehasonlításként és a jobb megértés érdekében a logisztikus

regressziós modell értelmezését is elvégeztük. Ennek az interpretációja egyszerű és ismert (a regressziós együtthatók alapján).

Eredményeink szerint a két modell hasonló képet mutatott: a legtöbb szövegjellemző, ami a biomedikális keretezés kategóriájába sorolást valószínűsítette, az orvosi kezeléshez és az emberi testhez kapcsolódott, míg a legtöbb pszichológiai jellemző a pszichológiai terápiához, az érzelmekhez és az emberi elméhez. A szociológiai jellemzők pl. a társadalmi kapcsolatokhoz ('gyermek', 'család') vagy a társadalmi kirekesztéshez ('meleg', 'bántalmazás', 'magányos') kapcsolódtak. Akárcsak a topikmodell esetén (lásd *A topikmodellezés legfontosabb módszertani tapasztalatai* c. fejezetet), itt is előfordult inkább pragmatikai, mint szemantikai erővel bíró jellemző: a logisztikus regressziós modell szerint pl. a több főnevet tartalmazó (tehát nyelvészeti kutatások szerint formálisabb hangvételű) szövegek nagyobb valószínűséggel tartoznak a biomedikális keretezés alá – ez vélhetően a megközelítés objektívabb jellegével magyarázható.

5.4.3. Modell-interpretáció a diszkriminációs kutatásban

A hivatali diszkriminációval kapcsolatos, korábban már részletezett kutatásunkban (Buda et al., 2022) is a SHAP mellett döntöttünk. A SHAP értékek alapján az email szövegek leíró statisztikai jellemzőit, mint szövegjellemzőt használó modell esetén kiderült, hogy a válasz hossza a legfontosabb jellemző, mégpedig olyan előjellel, hogy a hosszabb e-maileket kevésbé valószínű, hogy (vélt) roma ügyfeleknek írják. A tulajdonnevek, névelők és határozók aránya hasonló irányú, de kisebb hatást fejt ki, míg pl. az írásjelek változatossága ellentétes hatást látszik kifejteni. Ezek az eredmények arra utalnak, hogy amikor egy vélt roma ügyfél választ kap, az gyakran inkább rövidebb és tömörebb.

Hogy további információkat nyerjünk a jellemzők fontosságáról, további vizsgálatot is végeztünk: egyenként kihagytunk minden egyes jellemzőt, hogy lássuk, hogyan változik a megmaradt jellemzőkből felépített modellek teljesítménye. Eredményeink szerint pl. a modell teljesítményét gyakorlatilag nem befolyásolja a névmások és az igék arányának elhagyása - ezek a jellemzők olyan információkat kódolnak, amelyek más jellemzőkben is jelen vannak. Viszont a válasz hosszát leíró jellemző nélkül a modell teljesítménye nem sokkal jobb, mint egy véletlenszerű osztályozóé.

Második modellünk a nyers szövegre, annak kifejezéseire, mint szövegjellemzőkre épült. A SHAP-értékük szempontjából legfontosabb tokenek releváns része három csoportba volt sorolható, úgymint üdvözlések és megszólítások,

hivatalos címek (pl. ’közjegyző’), és egyéb információra vagy elérhetőségre való hivatkozás (pl. ’telefonszámon’). A SHAP-értékek előjele alapján elmondható, hogy a vélt roma ügyfelek formálisabb és meglehetősen visszafogott válaszokat kaptak (pl. "tisztelt címzett" megszólítás itt gyakoribb volt a tényleges név használatával szemben, míg a "szívesen" jelző ritkább).

Végül azt megvizsgáltuk, hogy a rendelkezésünkre álló metaváltozók (azaz az emailek külső, nem a szövegükre épülő jellemzői) hogyan befolyásolják a diszkrimináció erősségét. Két ilyen metaváltozó állt rendelkezésünkre: az ügyfél neme (keresztneve alapján) és az önkormányzat településének mérete. Eredményünk szerint a figyelemdiszkrimináció magasabb szintje volt kimutatható férfi ügyféllel szemben, ill. a kisebb településeken.

5.5. A felügyelt gépi tanulás további lehetőségei

Egy kézenfekvő lehetőség, hogy a felügyelt tanulás automatizált szövegcímkézése és annak interpretálása után nem állunk meg, hiszen a felcímkézett adatbázist újabb elemzések kiindulópontjává tehetjük. Egy ilyen kutatásra példa a depressziós fórumok felhasználóinak fórum-szocializációját elemző kutatásunk (Sik et al., 2023a), ahol miután a posztokat mélytanulással automatikusan felcímkéztük (biomedikális vs. pszichológiai keretezés), megvizsgáltuk, hogy adott felhasználó egymást követő hozzászólásait tekintve milyen idősor rajzolódik ki, majd az idősor-mintázatok alapján klasztereztük a felhasználókat. Így a fórumszocializáció ideáltipikus mintázatait sikerült azonosítanunk, pl. azét a klaszterét, melynek tagjai minél több időt töltenek a depressziós fórumokon, annál inkább eltávolodnak a depresszió szakértői diskurzusaitól, alternatív narratívákat keresve. Ezután korábban létrehozott és validált topikmodellünk topikjait tettük keresztbe a szocializációs klaszterekkel, azt vizsgálva, hogy az egyes látens témák milyen tipikus mintázatokban jelennek meg az egyes klaszterek diskurzusaiban.

Végül érdemes szót ejteni a nagy nyelvi modellekről (*large language models*, LLMs, ide tartozik a ChatGPT is). Számos friss eredmény számol be e modellek felhasználhatóságáról predikciós/osztályozási feladatokban, szövegek címkézésében is, pl. Gilardi et al. (2023) akik arról számolnak be, hogy a ChatGPT teljesítménye szövegek manuális címkézésében vizsgálatuk szerint felülmúlja az emberekét (ahol az emberi kódolókat az ilyen feladatoknál megszokott módon crowdsourcing platformról, az MTurk-ről rekrutálták). Ugyanakkor alá

kell húzni, hogy ezek a modellek természetüknél fogva fekete dobozok, vagyis az általuk szolgáltatott eredmények nem magyarázhatók meg teljesen. Ez a korábban írtak fényében azt jelenti, hogy e modellek jövője és használhatósága azoknak a módszertanoknak a létrehozásán múlik, amelyek lehetővé teszik a modellek interpretációját és magyarázatát, amiből következtetni lehet arra, hogy a modellek hogyan és miért jutottak kimeneteikhez.

5.6. A felügyelt tanulás módszertani tapasztalatai

Láttuk, hogy az annotálási folyamat számára a szociológiai fogalmak egy része komoly kihívás, mert hermeneutikailag nehezebben megközelíthető esetek. Ilyen esetben, ha elfogadjuk, hogy a jelentések ténylegesen az interszubjektivitásban, egy folyamat során konstruálódnak, akkor az annotálást is inkább egy interszubjektív folyamatként tételezhetjük. Ez a szemlélet elsősorban az annotálási irányelvek folyamatos, iteratív frissítésében jelenik meg. Ilyenkor a feladat inkább egyfajta kvalitatív kódolásként határozható meg, mivel a kategóriákat egy előzetes absztrakt elméletből származtatjuk, és induktív módon alakítjuk ki őket a kutatás során. Az interszubjektivitás elismerésének másik megnyilvánulása a kettős annotációra való áttérés lehet, ahol a végső, konszenzusos címke (vagy akár címke-halmaz) a két annotátor kódjának egyesítésén alapul.

Egy ilyen kutatás lényegesen különbözik azoktól a hermeneutikailag egyszerűbb üzleti alkalmazásoktól, amelyek explicit és egyértelmű kategóriák (lásd pozitív/negatív/neutrális szentiment) előre meghatározott készletét alkalmazzák.

Tapasztalataink szépen visszatükrözik más kutatók benyomásait. Aroyo és Welty (2015) egyenesen a humán annotálás hét mítoszaként utal azokra a naiv/pozitivista elképzelésekre, melyek egyértelműen kódolható szövegeket tételeznek fel. Szerintük a humán annotálás egy elavult szemantikai eszményen alapul, ami az egyetlen helyes igazság meglétét tételezi fel. Ebből további mítoszok lettek levezetve, mint például az annotátorok közötti eltérés hátrányos volta, az a remény, hogy a megfelelően részletezett annotálási irányelvek megoldják a problémát, az a hit, hogy a szakértők besorolása mindig helyesebb, mint a laikusoké, vagy az az elvárás, hogy egyetlen kategóriába besorolható legyen a szöveg. Új szemantikai elméletet javasolnak, a „crowd truth”-ra alapozva, melynek lényege, hogy az emberi értelmezés szubjektív, és hogy az annotátorok különböző

interpretációi jó reprezentációját adják ennek a szubjektivitásnak, s az ésszerű interpretációk tartományának.

Kutatási tapasztalatunk szerint az annotátorok közötti nézeteltérés mértéke jó becslést ad a gépi tanuló objektív nehézségére, ez az eredmény általánosan is felhasználható lehet. Másik általánosítható eredményünk szerint a szöveges adatokban emberi kódolás nélkül, automatizált módon is lehetséges a diszkrimináció felismerése, és a gépi tanulás (ML) olyan megkülönböztető jegyeket is felismerhet, amelyeket az emberi kódolók esetleg nem. Legjobb tudomásunk szerint a mi tanulmányunk volt az első kísérlet a diszkrimináció ML-technikákkal történő értékelésére., de a kétfajta kódolás között lényeges különbségek mutatkoztak, vagyis mindkét megoldás felfedett olyan különbséget, amit a másik nem.

Az emberi kódolás vagy a szótáralapú megközelítések egyértelműbb értelmezést tesznek lehetővé, de hermeneutikai szempontból hátrányaik is vannak: a kódolási irányelvek vagy a szótár tükrözik a kutatók horizontját, és korlátozzák a kimutatható nyelvi különbségek körét, míg az NLP potenciálisan bármilyen különbséget megtalálhat.

Ahogy fent láttuk, a gépi tanulás nem csak a címkézés hatékonyabbá tételét teszi lehetővé, de további inspiratív lehetőségei vannak a társadalomkutatás számára. A modell predikciós teljesítménye önmagában is sok esetben szociológiai értelmezhetőséggel bír (lásd: diszkrimináció- vagy polarizáció-mérték), míg a modellek fekete dobozának felnyitása, azaz interpretálásuk új szakterületi tudást hozhat az adott probléma kapcsán.

Ezek az automatikus címkézésen túlmutató társadalomkutatási lehetőségek ritkán jelennek meg a tudományos nyilvánosságban. Így pl. a számítógépes társadalomtudomány (*computational social science*) legnagyobb seregszemléjén, az IC2S2 konferencián 2023-ban „The augmented social scientist" címmel szerveztek képzést az automatikus címkézésről[17], meg sem említve az interpretálás lehetőségét. Másik példán: Macanovic (2022) az NLP társadalomkutatási lehetőségeiről írván szintén kizárólag a kézi kódolás gépiesítéseként írja le a módszert. Hasonlóan, Kenneth Benoit, az egyik legismertebb számítógépes politológus összefoglaló írásában (2020) azt írja, hogy az osztályozásnak a társadalomtudományokban csak instrumentális értéke van, az adatok kiegészítésére szolgál (mert új címkét társít hozzájuk), de önmagában nem ad tudományosan érdekes eredményt. Példaként hozza fel, hogy egy törvényhozó párthovatartozásának osztályozása

[17] https://www.css.cnrs.fr/the-augmented-social-scientist-tutorial-at-ic2s2/

érdekes mérnöki kihívás lehet egy informatikus számára, de ez nem hoz új felismerést egy politológus számára, mivel ez az információ már ismert. Vagyis itt sem kerül szóba a modell interpretációjának lépése, aminek igen fontos hozadéka lehet a politológus számára is, lásd az egyes pártok eltérő nyelvezetére, keretezésére, tematikájára vonatkozó potenciális új információkat, amit egy ilyen interpretáció ad. De mivel maga az interpretáció/magyarázat fontossága is viszonylag újkeletű (az *explainable AI* / *interpretable AI* néhány éve meglevő törekvések), talán kevéssé ismertek még a gépi tanuló modellek értelmezésében rejlő társadalomkutatási lehetőségek is.

6. NLP A POLITIKAI POLARIZÁCIÓ KUTATÁSÁBAN[18]

6.1. Motiváció

Ebben a fejezetben átfogó irodalmi áttekintést ismertetek arról, hogyan használják az NLP-t egy kurrens társadalomkutatási területen, a politikai polarizáció nyelvi lenyomatának, a nyelvi polarizációnak a tanulmányozására, elsősorban módszertani megközelítésben. A fejezet célja kettős: egyrészt a nyelvi polarizáció kutatásának state-of-the-art megoldásait veszi számba szisztematikusan, másrészt általános, az NLP minden társadalomkutatási alkalmazásában releváns kérdéseket tárgyalok. Ezek a kérdések az NLP-t alkalmazó kutatás validitását érintő következő fő kérdései köré csoportosulnak: a szöveg-mint-adat megközelítés megfelelően közelíti-e meg a kutatás tárgyát, megfelelően reprezentálja-e a választott korpusz a vizsgált jelenséget, a szöveget adattá konvertáló módszertan teljes és érvényes képet ad-e tárgyról. E fő kérdések a következő konkrét problémák formájában jelennek itt meg: a mérés problémája (hogyan mérem a polarizációt szöveges adat esetén?), a transzferábilitás problémája (hatékonyan működnek-e más korpuszokon is modelljeink?), a kódolás kérdése (szükségesek-e emberi kódolók?), a feature kiválasztás (hogyan válasszuk meg a vizsgált szövegek jellemzőit?), a black box előrejelző modellek (tudjuk-e interpretálni a gépi tanulásos modellt?) vagy a kevert módszertan kérdése (hogyan támogathatja a validálást és interpretálást a kvalitatív módszer automatizált megközelítés esetén?). Áttekintésem több pontot jól illusztrálja a társadalomtudományi versus adattudományi alapú NLP kutatási paradigmák közötti különbséget.

[18] E fejezet továbbfejlesztett változata korábbi, nyilvánosan elérhető, angol nyelvű cikkemnek: Németh, R. (2023). A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, *6*(1), 289-313. Köszönetemet szeretném kifejezni Barna Ildikónak és Szabó Juditnak az adatgyűjtés támogatásáért és Katona Eszternek az adatvizualizációkért.

6.2. Nyelvi polarizáció – konceptualizáció és operacionalizáció

A szöveges adatforrások számának és méretének növekedésével a természetes nyelvi feldolgozás (NLP) egyre nagyobb teret nyer a politikai polarizáció kutatásában is. NLP-módszerek állnak rendelkezésünkre, amelyekkel egy adott szerző politikai nézeteire következtethetünk a közösségi médiában közzétett bejegyzéseiből, a médiában megjelent cikkeiből, politikai beszédeiből stb. A szavazatokkal vagy politikai közvélemény-kutatásokkal ellentétben a szövegek lehetővé teszik szerzőik számára, hogy árnyaltabb véleményt fejezzenek ki. Az internetes szöveges adatok a közvélemény-kutatásokkal ellentétben nem a kérdezett beszámolóját (amit számos potenciális torzítással terhel, lásd visszaemlékezési torzítás vagy társadalmi elvárások okozta torzítást), hanem a ténylegesen megfigyelt viselkedést tükrözik. Továbbá, ha az NLP-t a törvényhozók ideológiai nézeteinek vizsgálatára alkalmazzák, a politikusi beszédek kevésbé vannak kitéve pártpolitikai kontrollnak, mint a szavazás (Goet, 2017).

Ahogy Lakoff (2002) írta: a politikai világnézetek a nyelvben tükröződnek. Ezek magyarázzák a témaválasztást, a szóhasználatot, a következtetés módját, ráadásul egyes szavaknak más jelentésük van, amikor liberálisok és konzervatívok használják őket. Sőt, ami még ennél is jelentősebb: a nyelv nem egyszerűen a vélemény/ideológia lenyomata (azaz a szavakhoz vezető tettek): a fordítottja - a tettekhez vezető szavak - is működik. Kísérletek azt mutatják, hogy a pártpolitikai keretezés hatással van a közvéleményre (Chong & Druckman, 2007), és a nyelv a csoportidentitás egyik legfontosabb meghatározója (Kinzler et al., 2007). A kognitív nyelvészet (pl. Lakoff, 1997) szerint a keretezés fontosabb, mint a tartalom, és a metaforák politikai használata nagy szerepet játszik a meggyőzésben. Az értelmező politikatudomány (Bevir & Rhodes, 2016) szerint a nyelv nem semleges médium, amely a társadalmi valóságot tükrözi, hanem a valóság megkonstruálásának eszköze. Ebből a szempontból a nyelvi polarizáció kutatása kísérlet a társadalmi valóság szimbolikus kontextusának megvilágítására.

Az új technikák alkalmasak a polarizáció nagyságának mérésére, segítenek nyomon követni annak tendenciáit, közelebb visznek minket a jelenség megértéséhez. Ugyanakkor a politikai polarizáció vizsgálatára használt NLP-módszerek szétszóródva találhatók a szakirodalomban, részben azért, mert a kutatásban számos különböző tudományág vesz részt. Ezért végeztem el a (2010 óta megjelent) írások irodalmi áttekintését, részben azt remélve, hogy ez a tudományos közösség érdeklődésére általában is számot tarthat, részben az általam vezetett OTKA-

kutatás (A politikai nyilvánosság rétegei Magyarországon, 2001-2020, NKFIH-K20) empirikus megalapozása céljából. Az OTKA-kutatás eredményeit is többször idézem e könyvben.

Az Egyesült Államok az egyik legtöbbet kutatott ország a politikai polarizáció tekintetében, mégis ellentmondásos eredmények születtek arra vonatkozóan, hogy az elmúlt évtizedekben nőtt-e a polarizáció az Egyesült Államokban. Az egyik oldal (pl. Fiorina & Abrams, 2008) azt állítja, hogy az adatok kevés jelét mutatják a tömeges polarizáció növekedésének. A másik oldal szerint (pl. Abramowitz & Saunders, 2008) a bizonyítékok arra utalnak, hogy a tömeges polarizáció az 1970-es évek óta drámaian megnőtt. Ahogy Lelkes (2016) rámutat, az ellentmondásos eredményeket a polarizáció eltérő konceptualizációja és operacionalizációja okozza. A szakirodalomban ugyanis jellegzetesen eltérő megközelítésekkel találkozhatunk. Még az egyik klasszikus tanulmány (DiMaggio et al., 1996) is többdimenziós jelleggel kezeli a polarizációt. A szerzők szerint a polarizáció mérhető (1) a vélemények szóródásaként, (2) a vélemények bimodalitásaként, (3) a különböző társadalmi attitűdök közötti szoros összefüggésként, vagy (4) a társadalmi attitűdök és bizonyos egyéni jellemzők közötti korrelációként. Lelkes (2016) két másik formát különböztet meg: az észlelt polarizációt és az affektív polarizációt. A digitális szféra polarizációjáról szóló írásában Yarchi et al. (2021) a már említett definíciók mellett megkülönbözteti az interakciós polarizációt, amely arra összpontosít, hogy a homofil interakciók dominálnak-e a heterofil interakciókkal szemben (kihasználva a digitális szféra hálózati jellegét).

A nyelvi polarizációra térve: Fiorina és Adams (2008) átfogó tanulmánya a téma szakirodalmának egyik alapműve, több mint 1000 hivatkozást tartalmaz, és felsorolja a polarizáció tanulmányozására használt különböző empirikus bizonyítékokat, de nem sorolja fel a nyelvi jellemzőket. A politikai polarizáció nyelvi megnyilvánulásai (a továbbiakban röviden *nyelvi polarizáció*) csak később, főként az elmúlt évtizedben kerültek be a tudományos diskurzusba. A "politikai polarizáció" kifejezést 2012 óta használják a számítógépes nyelvészek legnagyobb konferenciáján (Annual Meetings of the Association for Computational Linguistics). Az új érdeklődés oka az, hogy a nyelvi polarizáció már régóta jelen van, de a nagyságrendje és a mögötte álló tudatos politikai-stratégiai döntések (Gentzkow et al., 2019) új jelenségek.

A nyelvi polarizáció mérésekor vagy a politikai polarizáció meglévő mérési gyakorlatát próbálják meg szöveges adatokra adaptálni, vagy új megközelítést dolgoznak ki. Az előbbi megoldás több, hagyományosan használt polarizációs

mérőszám esetében lehetséges. Például DiMaggio et al. (1996) mérései, amelyek egy numerikus változó eloszlásán alapulnak, adaptálhatók, ha a szöveges adatok valamilyen megfelelő módon numerikus adatokká alakíthatók. Más esetekben nem lehetséges a szövegekre alkalmazott megközelítés közvetlen megfeleltetése a nem szöveges adatokra alkalmazott megközelítéssel. Például Demszky et al. (2019) a nyelvi polarizáció kutatása során NLP-módszereket használ a szövegek affektív jellemzőinek a tartalomtól való elkülönítésére. A tartalom tekintetében megkülönböztetik a témaválasztást és a témaszintű keretezést. Az affektív jellemzők közvetlenül összekapcsolhatók az affektív polarizációval, a témaválasztás és a keretezés azonban minőségi jellegűek, azaz nem nyújtanak egyszerű módot a vélemények megoszlásának jellemzésére. Amint azt látni fogjuk, az interakciós polarizáció (Yarchi et al., 2021) szintén meghatározható szöveges adatokon.

Az alábbi áttekintés szerint az NLP-alapú megközelítések leggyakrabban új módszereket vezetnek be a polarizáció mérésére. A megközelítések a politikai álláspont mögöttes konceptualizációja és a rendelkezésre álló adatok szerint különböznek egymástól. Kifejezetten szöveges adatokra fejlesztették ki őket, és sok esetben olyan adattudományi logikát használnak, amely eltér a hagyományos társadalomkutatási logikától. Az NLP viszonylag új eszköz ezen a területen. Számítási módot biztosít az ideológiák vagy érzelmek automatikus azonosítására a szövegekben. Amint azt részletesen ismertetjük, a módszertani megközelítések a politikai álláspont mögöttes konceptualizációját és a rendelkezésre álló adatoktat tekintve különböznek. Így például abban térnek el az egyes megközelítések, hogy a szövegek szerzőinek rendelkezniük kell-e címkézett politikai pozícióval, vagy inkább néhány látens ismeretlen dimenziót tételezünk fel; hogy a címkézett pozíciók előre meghatározottak-e, vagy a szövegeket kézzel kell kódolni; illetve, hogy az ideológiát folytonos skálaként vagy pozíciók véges halmazaként határozzák meg. Ezek a paraméterek határozzák meg, hogyan mérjük a politikai pozíciót, és ennek alapján hogyan közelítjük meg a politikai polarizációt.

6.3. A kutatás célja

Az alábbiakban egy átfogó irodalmi áttekintést ismertetek arról, hogy hogyan használják az NLP-t a nyelvi polarizáció tanulmányozására. Meghatározom az adatforrásokat és a számítási technikákat, és áttekintem a polarizáció különböző konceptualizációit és operacionalizációját. Nem célom az összefoglalóban

érintett kutatási eredmények részletes ismertetése, és részletes szcientometriai elemzést sem nyújtok. Ehelyett a kutatási tevékenység jellegét és módszertani megközelítéseit összegzem. Az adatforrások és a számítási módszerek azonosítása mellett áttekintem is, hogy a politikai nyilvánosság mely rétegeit (hivatalos politikusi, szakértői, média, laikus) vizsgálják. Az általános empirikus megközelítést is vizsgálom: hogyan gyűjtötték a szöveges adatokat, alkalmaztak-e az automatizált módszerek mellett emberi kódolókat, és alkalmaztak-e kvalitatív megközelítést ("*close reading*") is.

Áttekintésem számos más olyan jellemzőt is tartalmaz, amely leírja a társadalomtudományi és az adattudományi alapú NLP kutatási paradigmák közötti különbséget. Ez a két megközelítés leginkább magyarázó ill. prediktív jellegűként azonosítható (Hofman et al., 2021, a statisztikai oldalról lásd Breiman, 2001). A két megközelítés közötti különbségek leginkább az alábbi jellemzők mentén azonosíthatók, a review során ezekre fogok koncentrálni: elméletbe ágyazott-e a kutatás, alkalmaz-e kvalitatív megközelítést is, kitér-e az ok-okozati összefüggésekre, interpretálja-e az eredményeket. A társadalomtudósok hagyományosan a magyarázatokat helyezik előtérbe, az elméletből levezetett oksági mechanizmusokra hivatkozva. Az adattudósok azonban inkább minél jobb teljesítményű előrejelző modellek kidolgozásával foglalkoznak, az értelmezhetőséget félretéve. A két megközelítés integrációjának lehetőségére is kitér majd az áttekintés.

Munkám célja, hogy a megközelítések, lehetséges hibák, esetleges hiányosságok és megoldások összegzésével kiindulópontot nyújtson a jövőbeli kutatásokhoz. Mivel az NLP viszonylag új módszertani megközelítés, remélem, hogy az összefoglaló új kutatásokat is inspirálhat. Ezen túl az NLP-t nem ismerő társadalomtudományi kutatóknak is szeretnék képet adni a folyamatban lévő kutatásokról.

6.4. Az irodalmi áttekintés módszertana

Az áttekintendő tanulmányok multidiszciplináris jellege, valamint a szakterületek közötti terminológiai különbségek miatt döntöttem úgy, hogy az irodalmi áttekintés módszertanának az un. átfogó összegzést (*scoping review*, Arksey & O'Malley, 2005) választom. Az átfogó összegzés a szisztematikus összegzéssel (*sytematic review*) szemben egyrészt szélesebb körű megközelítésre törekszik, másrészt nem célja, hogy konkrét kutatási kérdésekkel foglalkozzon, vagy

a tanulmányok minőségét értékelje. Ez határozza meg a review-folyamat logikáját is: pl. a keresési kifejezések előre rögzített listája helyett a keresési feltételek fokozatos bővítése érdekében a keresések többszörös iterációit kellett alkalmazni.

Az adattudomány intézményi sajátossága, hogy a kutatások jelentős része soha nem jelenik meg folyóiratokban, "csak" konferenciakötetben/preprint kiadványokban, ezeknek a platformoknak a státusza magasabb is itt, mint más tudományágakban. Ezért a kereséseket a Google Scholar segítségével végeztem, mivel ez a keresőmotor indexeli a konferencia-köteteket és preprint archívumokat is.

A 2010. január 1. és 2021. június 29. között megjelent tanulmányokat vettem figyelembe. A kezdő dátum megválasztásának oka, hogy - amint azt a Bevezetésben részletesen tárgyaltam - a nyelvi polarizáció számítástudományi módszerekkel történő megközelítése egy évtizeddel ezelőtt kezdődött.

Mivel a keresett írásokat mind témájuk, mind módszertanuk definiálja, kezdeti keresési feltételeim a "political polariz(s)ation" AND "natural language processing" voltak, majd mindkét kifejezéshez szinonimákat adtam hozzá. A szinonimákat úgy próbáltam azonosítani, hogy ellenőriztem, megtalálom-e velük az általam relevánsnak ítélt tanulmányokat. Ezért a "political polarizaion" alternatívájaként pl. a "partisanship" kifejezéseket is bevontam. Eredetileg a "partisanship" alatt az affektív vagy racionális pártidentifikációt értették (Carius-Munz, 2020), ebben az értelemben ok-okozati összefüggésben állhatott a polarizációval, bár nem tekintették azzal ekvivalensnek. Az utóbbi időben azonban számos szerző felcserélhető fogalomként használja a "partisanship" ill. "polarization" fogalmát (pl. Demszky et al., 2019).

A módszertan tekintetében: az NLP nem kanonizált fogalom, és a megközelítést kutatási területtől függően más-más néven emlegetik: a szociológiában szövegbányászat, a nyelvészetben számítógépes nyelvészet stb. Végső keresőkifejezésem a következő volt:

> ("political polariz(s)ation" OR "partisan divide" OR "language polariz(s)ation" OR "partisan rhetoric" OR "partisan polariz(s)ation" OR "partisanship" OR "partisan language" OR "polariz(s)ed language" OR "polariz(s)ed rhetoric")
> AND
> ("natural language processing" OR "text mining" OR "computational linguistics" OR "computational text analysis").

A cikkek teljes szövegén kerestem.

Arra törekedtem, hogy a lehető legátfogóbb legyek a releváns tanulmányok azonosításában. Ennek érdekében a Google adatbázisában kulcsszavas keresés útján történő keresés mellett a keresési eredménylistát kiegészítettem az általunk ismert vagy a szakirodalomban kulcsforrásként idézett fontos tanulmányokkal.

Mint általában az áttekintő vizsgálatok esetében, a keresési feltételek nehézkes, valamint a cikk teljes szövegén történő keresés miatt sok irreleváns találatot kaptam, ezért egy sor kemény és puha kizárási kritériumot határoztam meg. A kemény kritériumok szerint kizártam a nem angol nyelvű anyagokat, a szabadalmakat, a prezentációs diákat, a diplomamunkákat, a nem tudományos jellegű írásokat és a nem működő oldalakra mutató linkeket. A puha kizárási kritériumok célja az irreleváns tartalmú cikkek kizárása volt. Csak olyan cikkeket vettem be, amelyek empirikusan kutatták a nyelvi polarizációt, nem csak megemlítették azt. Ugyanakkor nem csak a politikai ideológiák körüli polarizáció elemzését vettem be a találatok közé, hanem a közpolitikai kérdések, például az éghajlatváltozás körüli polarizációt is. Ugyanakkor kizártam a következő (közvetlenül nem releváns) területeket: a nyelvi tervezéssel (language planning), a nyelvi ideológiával, a visszhangkamrákkal (echo chambers) és az ellentmondás-felismeréssel kapcsolatos tanulmányok. A módszertan tekintetében kizártam a kizárólag kvalitatív szövegelemzést és az egyszerű kvantitatív megközelítést (pl. hashtag- vagy szógyakoriságokat) alkalmazó tanulmányokat. Viszont azokat a cikkeket bevettem, amelyek elsődlegesen az ideológiai álláspontok nyelvhasználatból történő algoritmikus osztályozására összpontosítottak. A két feladatot (a polarizáció tanulmányozása és az ideológia osztályozása) nehéz szétválasztani, már csak azért is, mert - mint az áttekintésből kiderült - az osztályozási modell teljesítménye a polarizáció mérőszámaként szolgálhat.

Minden tanulmány esetén azonosítottam (ha a kérdés alkalmazható volt):

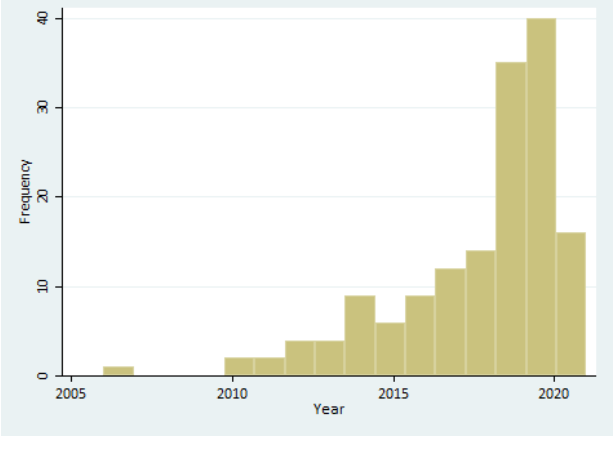
- a közzététel éve
- szerzők, hovatartozásuk
- cím
- folyóirat/kiadó
- kutatási kérdések
- ország és a vizsgált időszak
- vizsgálták-e és hogyan vizsgálták a polarizáció dinamikáját
- milyen szövegelemzési módszereket alkalmaztak (felügyelt vagy felügyelet nélküli stb.)

- ha osztályozást alkalmaztak: mi az eredmény / milyen jellemzőket használtak / hogyan jegyzetelték a képzési adatokat
- a szakterületi tudás (domain knowledge) szerepe, a szerzők megpróbálták-e értelmezni eredményeiket
- a korpusz típusa (Twitter stb.)
- a korpusz szűrése (kulcsszavas keresés stb.)
- a polarizáció konceptualizálása (pl. pozicionális, affektív)
- a polarizáció operacionalizálása
- a politikai nyilvánosság mely rétegeit vizsgálták (politikusok, média, szakértő vagy laikus közönség); ha egynél több réteget vizsgáltak: vizsgálták-e és hogyan a kapcsolatukat
- használtak-e kvalitatív szövegelemzést
- fontos módszertani újdonság.

6.5. Eredmények

6.5.1. Összefoglaló

A duplikációk törlése után a keresés 3078 egyedi találatot adott, amelyhez kézzel további 6 újságot adtunk hozzá. A kezdeti 3084 rekordból a kizárások után 154 releváns tanulmányt azonosítottunk. A 19. ábra szerint a publikációk száma az elmúlt évtizedben meredeken emelkedett.



19. ábra. A tanulmányok a közzététel éve szerint.

Az írás tárgyát képező ország		*Szerző hovatartozása*	
USA	91	USA	81
Egyesült Királyság	9	Olaszország	14
Olaszország	6	Egyesült Királyság	13
Spanyolország	6	Németország	11
Kanada	5	Katar	10
Németország	5	Kanada	8
India	3	Spanyolország	7
Skócia	3	India	6
Törökország	3	Írország	5
Ukrajna	3	Franciaország	4
Belgium	2	Izrael	4
Chile	2	Svájc	4
Franciaország	2	Bulgária	3
Izrael	2	Orosz Föderáció	3
Japán	2	Törökország	3
Norvégia	2	Hollandia	2
Oroszország	2	Japán	2
Ausztria	1	Ausztria	2
Bolívia	1	Belgium	2
Katalónia	1	Chile	2
Kolumbia	1	Norvégia	2
Ecuador	1	Marokkó	1
Indonézia	1	Svédország	1
Írország	1	Dánia	1
Észak-Írország	1	Magyarország	1
Palesztina	1	Fülöp-szigetek	1
Fülöp-szigetek	1	Szingapúr	1
Lengyelország	1	Dél-Afrika	1
Románia	1	Indonézia	1
Dél-Afrika	1	Románia	1
Nem országspecifikus	8	Lengyelország	1
		Hong Kong	1

2. táblázat. A tanulmányok tárgyát képező országok, valamint a szerzők hovatartozása szerinti országok előfordulási gyakorisága. Az országnevek a tanulmányban szereplő megnevezést követik.

A 2. táblázat a tanulmányok ország-hovatartozását mutatja, két módon: a vizsgálat tárgyát képező ország. illetve a szerzők hovatartozása szerint. (Egyes tanulmányok nem végeztek országspecifikus elemzést, míg mások egynél több országot vizsgáltak, amelyek külön-külön szerepelnek a táblázatban.)

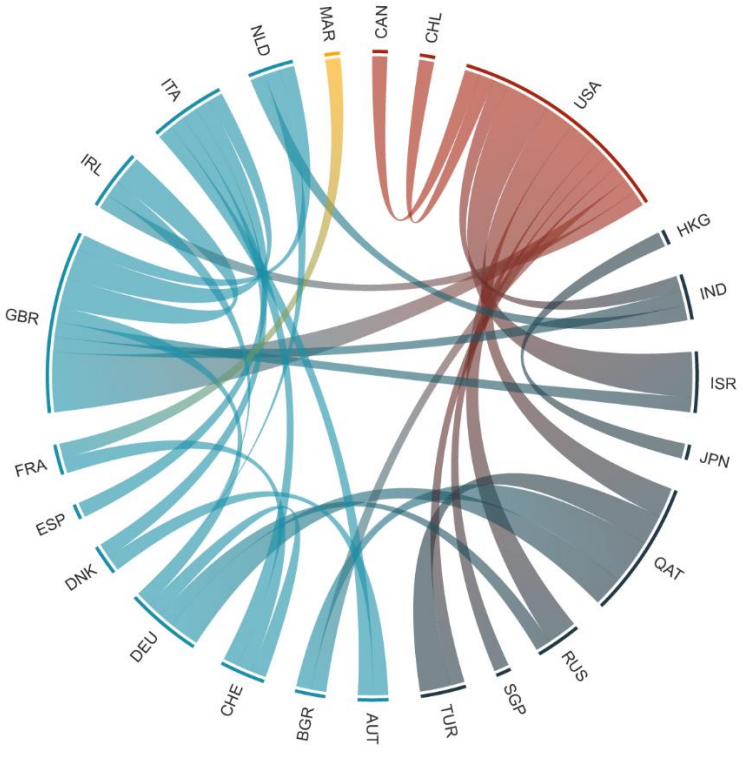
Amint a táblázat mutatja, a tanulmányok több mint felét amerikai adatok felhasználásával végezték, és az USA a tanulmányok feléhez kapcsolódó, legtöbbet publikáló ország is. Az USA-t az Egyesült Királyság és Olaszország követi a rangsorban. Érdemes megemlíteni, hogy bár Katar nem szerepel a vizsgált országok között, tíz katari szerző dolgozott (jellemzően együttműködésben) más országok tanulmányain.

A tanulmányoknak csak a tizede volt egyetlen szerző által írt. A 3. táblázat az országok közötti társszerzői együttműködést mutatja be.

A 20. ábra a 3. táblázat társszerzői kapcsolatait jeleníti meg. A gráf csúcsainak mérete az adott országhoz tartozó tanulmányok számát jelzi, az élek szélessége az együttműködésben készült publikációk számával arányos, a színek pedig a kontinenseket jelölik. Az országokat az ISO 3166 szabvány szerinti hárombetűs rövidítéssel jelöltem. Az ábra jól mutatja, hogy az abszolút számokat tekintve az USA vezet az együttműködések terén, a második helyen Nagy-Britannia és Észak-Afrika áll, de az európai és ázsiai országok is nagyon aktívak. Az arányokat tekintve azonban látható, hogy az amerikai tanulmányok 80%-ában nem volt más országból származó szerző, míg a nem amerikai szerzők szinte mind együttműködnek, jellemzően egy amerikai társszerzővel. Ezek az eredmények a kutatási téma USA-központúságát hangsúlyozzák.

Nemzetközi együttműködések			*Nemzetközi együttműködés nélküli, többszerzős publikációk*			*Egyszerzős publikációk*	
CHE	FRA	1	AUT	AUT	1	BEL	1
CHE	ITA	1	BEL	BEL	1	FRA	1
DEU	CHE	1	CAN	CAN	6	GBR	1
DEU	IRL	1	CHE	CHE	1	NOR	1
DEU	NLD	1	CHL	CHL	1	POL	1
DNK	AUT	1	DEU	DEU	5	QAT	1
GBR	DEU	1	ESP	ESP	6	SWE	1
GBR	IRL	2	FRA	FRA	1	USA	7
GBR	ITA	2	GBR	GBR	4		
GBR	NLD	1	HUN	HUN	1		
HKG	JPN	1	IDN	IDN	1		
IND	GBR	1	IND	IND	4		
IND	NLD	1	IRL	IRL	3		
ISR	GBR	1	ISR	ISR	1		
ITA	AUT	1	ITA	ITA	9		
ITA	DNK	1	JPN	JPN	1		
ITA	ESP	1	NOR	NOR	1		
MAR	FRA	1	PHL	PHL	1		
QAT	BGR	3	QAT	QAT	1		
QAT	DEU	2	ROU	ROU	1		
RUS	DEU	1	USA	USA	57		
TUR	QAT	2	ZAF	ZAF	1		
USA	BGR	1					
USA	CAN	2					
USA	CHL	1					
USA	GBR	4					
USA	IND	2					
USA	IRL	1					
USA	ISR	3					
USA	QAT	2					
USA	RUS	2					
USA	SGP	1					
USA	TUR	1					

3. táblázat. Társszerzők országok szerint.



20. ábra. Az országok társszerzői hálózata.

A 4. táblázat a tudományágak társszerzői kapcsolatait mutatja. A 21. ábra az együttműködésekre koncentrálva ezt vizualizálja hálózatként, élei és csomópontjai a 20. ábrához hasonlóan vannak definiálva, országok helyett tudományágakkal. Az ábra alján lévő függőleges élek az azonos tudományágakhoz tartozó együttműködéseknek felelnek meg.

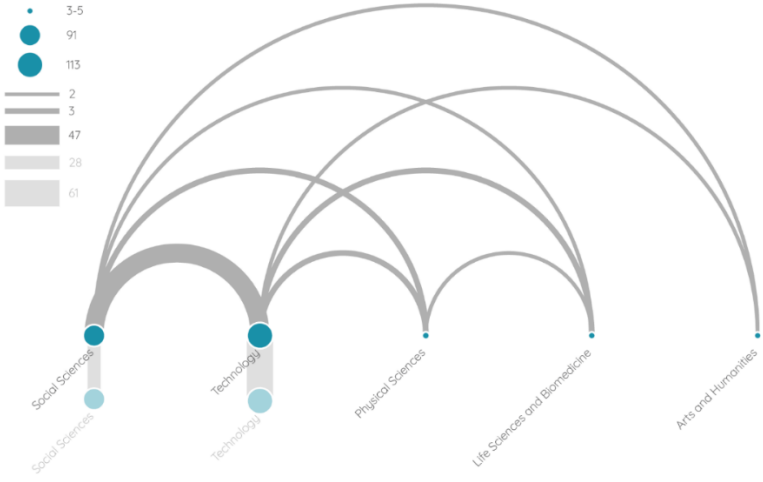
A Web of Science osztályozás[19] alapján a kutatási területeket öt nagy kategóriába soroltuk, egy kivétellel: a nyelvészetet nem a társadalomtudományokhoz, hanem a technológia kategóriájához rendeltük, mert elsősorban arra voltunk kíváncsiak, hogy a szerzők olyan interdiszciplináris csoportokban dolgoznak-e, ahol az adatelemzők (informatikusok, számítógépes nyelvészek) stb. mellett szakterületi tudósok (szociológusok, politológusok stb.) is részt vesznek.

[19] https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html

Interdiszciplináris együttműködések			*Többszerzős publikációk interdiszciplináris együttműködés nélkül*			*Egyszerzős publikációk*	
Társadalomtudományok	Technológia	47	Társadalomtudományok	Társadalomtudományok	28	Társadalomtudományok	13
Társadalomtudományok	Fizikai tudományok	3	Technológia	Technológia	61	Technológia	1
Társadalomtudományok	Élettudományok & Orvosbiológia	2					
Társadalomtudományok	Művészetek és bölcsészettudományok	2					
Technológia	Fizikai tudományok	3					
Technológia	Élettudományok & Orvosbiológia	3					
Technológia	Művészetek és bölcsészettudományok	2					
Fizikai tudományok	Élettudományok & Orvosbiológia	2					

4. táblázat. Interdiszciplináris együttműködések.

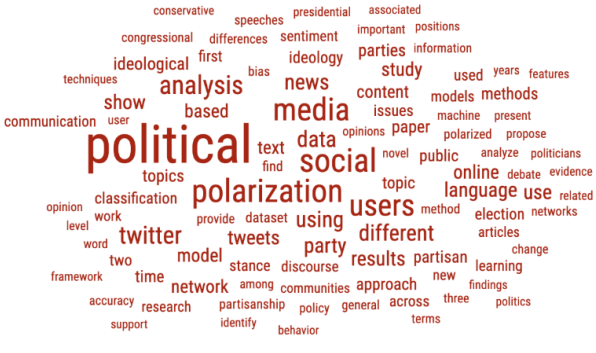
A tanulmányok kétharmada a társadalomtudományokhoz, és ugyanilyen aránya a műszaki tudományokhoz sorolódik. A másik három tudományterülethez csak néhány tanulmány tartozott. Ennek megfelelően a legtöbb együttműködés a társadalomtudományok és a technológia között jött létre. Ugyanakkor a nem interdiszciplináris munkák száma igen magas: a cikkek mintegy ötödét kizárólag társadalomtudósok, további 40%-át (!) pedig kizárólag a technológia területén kutató szerzők írták. Ha összevonjuk a fizikai tudományokat, az élettudományokat-orvostudományokat a technológiával, akkor megállapíthatjuk, hogy a cikkek 45%-át kizárólag ezekről a területeken tevékeny szerzők írták, társadalomtudományi vagy bölcsész szerzőtárs nélkül.



21. ábra. A tudományágak társszerzői hálózata.

A publikáció leggyakoribb helye az arXive pre-print szerver volt (14 tanulmány), egyébként a platformok (folyóiratok, konferencia-kiadványok stb.) igen változatosak, a tanulmányok négyötöde olyan helyen jelent meg, amely csak egyszer szerepel az adatbázisban. Ez a téma multidiszciplináris jellegét jelzi. A szerzők listája is változatos, a leggyakoribb szerzők Kareem Darwish (7 tanulmány) és Elliott Ash (4 tanulmány).

A 22. ábra a tanulmányok kivonatából készült szófelhőt mutatja be, az ábrán jól azonosíthatók a polarizációkutatás fókuszpontjai.



22. ábra. A tanulmányok absztraktjának szófelhője.

6.5.2. Adatok

Adatforrások

Az adatforrások, különböző platformok közötti különbségek gyakran fontos módszertani következményekkel járnak. Különbözhetnek aszerint, hogy mennyire hozzáférhetőek a kutatás számára, mennyire népszerűek a különböző országokban, milyen hosszúak a rajtuk megtalálható szövegek, és milyen típusú információt nyújtanak (pl. a tartalom mellett a Twitter egyéb, felhasználók közötti interakciós információt is kínál). Vagyis az adatforrás megválasztása a kutatás eredményére is hatással lehet. Ezért vizsgáltam az adatforrások típusát is. Az 5. táblázat szerint a tanulmányok több mint 40%-a Twitter-adatokat használt. Gyakori jelenség, hogy egy elemzés illusztratív, mert a tanulmány inkább módszertani újdonságot alkalmaz, így a tartalmi eredmény másodlagos, ezért mások már publikált korpuszát használja (ilyen korpuszokat használ pl. Quraishi et al., 2018: a két annotált Twitter-adatbázist eredetileg Brigadir et al., 2015, publikálta, a skót függetlenségi népszavazásról és a 2014-es amerikai félidős választásokról).

Az adatforrások az általuk lefedett időszak tekintetében is különböznek, a hosszabb időszakok lehetővé teszik a változások kimutatását (lásd a 6. táblázatot). A tanulmányok 37%-a vizsgálta az időbeli változásokat, és minden harmadik tanulmány 2 évnél hosszabb időszak adatait használta fel. A történeti kutatások jellemzően digitalizált szövegeket használnak, szemben az eleve digitálisan született szövegeket használó kortárs kutatásokkal. Lásd pl. García et al. (2020), akik digitalizált újságok alapján követik nyomon a gazdaságpolitikai bizonytalanságot Spanyolországban 1905-1945 között, a szélsőséges politikai polarizáció időszakában. Gentzkow et al. (2019) digitalizált, 1873-2016 között született amerikai kongresszusi beszédeket használt; a nagyon hosszú időszakot vizsgáló tanulmányok többsége az ő adatbázisukat használta (pl. Tucker et al., 2020).

A hosszabb időskálák a jelenben észlelt polarizáció mértékének (összevetésen alapuló) megítéléséhez is alapot nyújtanak. Jensen et al. (2012) például azt találta, hogy bár a politikai diskurzus az 1990-es évek végén polarizáltabbá vált, a polarizáció a 19. század végéhez és a 20. század nagy részéhez képest alacsony maradt.

Az időbeli változások nyomon követése úgy is lehetséges, hogy a polarizáció változásait nem egy "közös" történelmi időtávlatban, hanem az egyéni életek során mérjük. Iliev et al. (2019) például az amerikai szenátorok retorikáját

vizsgálta a hivatalban töltött idő függvényében, és azt találták, hogy a törvényhozók hasonlóbbá váltak egymáshoz a mandátumuk első néhány éve után, függetlenül a pártállástól vagy a kampányígéretektől.

Az adatforrás típusa	
Twitter	66
Kongresszusi/parlamenti felszólalások	32
Híroldalak	22
Parlamenten kívüli beszédek*	8
írásos politikai közokiratok**	7
Facebook	6
Reddit	5
Blogok	4
nem politikai szakértők által készített szövegek***	4
Újságok	3
Elnökjelöltek vitái	2
nem politikai szervezetek által készített szövegek	2
wikik (Conservapedia, RationalWiki)	2
YouTube hozzászólások	2
Online vitafórumok (4Forums.com,CreateDebate.com)	1
YouTube hozzászólások	1
könyvek/magazincikkek	1
Google Ngram	1
Google Search snippets	1
interjúk átiratai	1
politikai szakértők (politikai gondolkodók, kommentátorok)	1
Felmérés	1
Orosz közösségi hálózat "VKontakte"	1
televíziós átiratok (hírműsor)	1
Tumblr	1
WhatsApp	1

5. táblázat. Az adatforrás típusa és előfordulási gyakorisága

Ahol:

*: elnökjelöltek/elnökjelöltek beszédei, kampánybeszédek, nyilvános nyilatkozatok, elnökjelölti bejelentések,

**: pártprogramok, jelöltlisták, pártok saját platformjai, sajtóközlemények, koalíciós megállapodások,

***: bírák írásos véleményei, üzleti jelentések, tudományos cikkek.

A hosszabb időskálák is alapot nyújtanak a jelenben észlelt polarizáció mértékének értékeléséhez. Jensen és munkatársai (2012) például 127 évet vizsgáltak, és azt találták, hogy bár a politikai diskurzus az 1990-es évek végén polarizáltabbá vált, a 19. század végéhez és a 20. század nagy részéhez képest a polarizáció alacsony maradt.

A politikai nyilvánosság mely rétegeit vizsgálják		*A lefedett időszak hossza*	
csak laikus	63	keresztmetszeti adatok vagy	
csak hivatalos (politikusok)	50	2 évnél rövidebb időszak	105
csak a média	19	2-5 év	9
csak szakértő	5	5-10 év	11
média és laikus	4	10-20 év	8
hivatalos és média	4	20-50 év	10
hivatalos és laikus	3	50-100 év	2
hivatalos és szakértő	2	100+ év	9
hivatalos és média és szakértő	2		
hivatalos és média és laikus	1		

6. táblázat. A politikai nyilvánosság vizsgált rétegei és a tanulmányok által lefedett időszak hossza (gyakoriséggal).

Végül az elemzett adatok kiválasztásáról. A korpuszok tematikus fókuszálása szinte kizárólag **kulcsszavas szűréssel** történik; Lai et al. (2020) például az Egyesült Királyságban a BREXIT-népszavazásról szóló politikai vitát elemezte a Twitteren, és összegyűjtötte a #brexit hashtaget tartalmazó tweeteket. Más esetekben a téma azonosítása nem ilyen egyszerű: Green et al. (2020) a COVID-19 világjárványról szóló tweetek azonosítása érdekében teljes szótárat kellett, hogy kidolgozzon a válsággal kapcsolatos altémákra vonatkozóan. Természetesen a kulcsszavak kiválasztása nem tökéletesen objektív, de ettől még felhasználhatóak egy jól fókuszált korpusz felépítéséhez. Közösségi média platformok használata esetén a felhasználók lakóhely szerinti országa szerinti szűrést is gyakran használják.

A politikai diskurzus rétegei

A nyelvi polarizáció *a politikai nyilvánosság különböző szintjein* jelenhet meg, beleértve a politikai kommunikáció hivatalos csatornáit (pl. parlamenti beszédek), a média különböző típusait, valamint a felhasználók által generált tartalmakat (pl. közösségi média). A következőkben különbséget teszünk a

professzionális (hivatalos források és a sajtó) és a nem professzionális (laikus) kommunikáció között. Az áttekintés során egy negyedik réteg is felbukkant: a szakértői réteg (pl. bírák vagy közgazdászok által írt szövegek). A szakértői területen az ideológiai elfogultság elfogadhatatlan, ezért különösen érdekes kérdés, hogy detektálható-e itt algoritmikus eszközökkel ideológia.

A 6. táblázat szerint a legtöbb tanulmány a laikus közönséggel foglalkozott (n = 71); mindegyik közösségi média adatokat használt. Többségük tweeteket vagy posztokat vizsgált; kivételt képeztek KhudaBukhsh et al. (2020), akik a YouTube-csatornák kommentszekcióit vizsgálták, vagy Wu et al. (2019), akik Twitter bio-t (a felhasználó rövid nyilvános önjellemzését) használták a megosztottság felderítésére.

A hivatalos politika rétegével foglalkozó tanulmányok (n=60) többnyire törvényhozási beszédeken alapultak. Kivételt képeztek például Gross és munkatársai (2020), akik a németországi pártok regionális kiáltványainak adatállományát használták, ami lehetővé tette számukra a helyi szintű politika elemzését, vagy Wang és Tucker (2021), akiknek korpusza az amerikai kongresszus tagjai által kiadott sajtóközleményekből állt.

A szakértői szférát kisebb számú tanulmány (n=9) vizsgálta. Ezek között volt Jelveh et al. (2014), akik látens ideológiai elfogultságot észleltek a közgazdaságtudományi tudományos dolgozatokban, és Diaf et al. (2020), akiknek adatállománya német gazdaságkutató intézetek által kiadott konjunktúrajelentés-részletekből állt. Az elemzés egysége az előbbi esetben az egyén, az utóbbi esetben az intézmény volt. Hausladen et al. (2020) egy másik szakmai területet, a bíróságokat vizsgálták, mégpedig úgy, hogy megpróbálták osztályozni az amerikai Circuit Court döntések ideológiai irányultságát. Acree (2016) és Giglietto et al. (2018) az Ideological Book Corpus-t tanulmányozták, amely politikai gondolkodók és kommentátorok szövegeit, valamint konzervatív/liberális wikiket (Conservapedia, illetve RationalWiki) tartalmaz.

A tanulmányok nagyon ritkán vontak be több mint egy nyilvánossági réteget (mindössze 9%-uk). Serrano-Contreras és társai (2020) a politikusok által feltöltött Youtube-videókhoz fűzött megjegyzéseket vizsgálták, vagyis egyszerre a politikusi és a laikus réteget. Az ő a megközelítésük általánosan alkalmazható lenne szöveg-reakció viszony vizsgálatára. Gorrell és munkatársai (2018) hasonló felépítést követtek: a politikusok felé irányuló online abúzust vizsgálták, a politikusok tweetjeire a laikusok által adott válaszok korpuszát felhasználva.

Karamshuk et al. (2016) a 2013-14-es ukrán-orosz konfliktus idejének média és a laikus nyilvánosságbeli reprezentációját is vizsgálta, de a két réteget külön-külön kutatták. A két réteg polarizációjának hasonlóságait vagy kapcsolatukat (pl., hogy a megosztó kifejezések átterjednek-e az egyikből a másikba) nem vizsgálták. Kivételt Hofmann et al. (2020) jelentett. Két korpuszuk a médiából és a politikusok hivatalos szférájából származott, és a szerzők kifejezett célja az volt, hogy pártonként mérjék a két réteg nyelvezete közötti különbségeket. Az adott pártra vonatkozó médiakorpuszt úgy definiálták, mint a pártról szóló (és nem a párt által írt) cikkek összességét.

Acree (2016) még ennél is tovább ment, összehasonlította az általa vizsgált rétegek szerkezetét, és kimutatta, hogy a szakértői diskurzusban reprezentált ideológia gazdag és változatos, a szakmai politikai vita azonban egyetlen (bal-jobb) dimenzióba sűríti az ideológia kifejeződését. Yan és társai (2017, 2019) három különböző réteget vizsgáltak, a szakértői szférát (konzervatív és liberális wikipédiák), a média és a politikusok mellett. Fő kutatási kérdésük az volt, hogy a három réteg polarizációja mennyiben különbözik (Yan et al., 2017), és hogy az egyik rétegre fejlesztett modell mennyire vihető át másik rétegre, vagyis azok transzferábilitását vizsgálta (Yan et al., 2019). Ez utóbbival kapcsolatban kimutatták, hogy a kongresszusi jegyzőkönyveken alapuló modellek bizonyos sikerrel osztályozzák a médiából származó cikkeket is, vagyis van egy diffúziós folyamat a kongresszus és a média között. Widmer és munkatársai (2020) szintén kimutatták a média két különböző szegmense közötti diffúziót, azt vizsgálva, hogy a kábeltelevízióból érkező pártos hírközlések hogyan befolyásolják a helyi újságok által közzétett tartalmakat.

Az elemzés egysége

A legtöbb írásban a természetesen körülhatárolt szöveg az elemzés egysége, de néha ezt kisebb egységekre bontották, máskor pedig több szöveget összevontak. Ez utóbbi néhány esetben technikai okokból történik, mint például Wang és munkatársai (2017) esetében, akik ugyanazon személy minden 20 tweetjét egy szupertweetbe fűzték össze, hogy leküzdjék az egyetlen tweet korlátozott hosszát. Más esetekben a szövegek összevonásának elméleti oka van, mint Medzihorsky és társai (2014) esetében, akiknek az elemzés egysége az egyes politikai jelöltek voltak: ők az általuk kiválasztott viták során az egyes jelöltek összes megszólalását egyetlen

dokumentumként kezelték. Az így kapott szövegek várhatóan a legtöbb releváns témát lefedték, mivel minden vita egy-egy szűk témára összpontosított.

Vannak kutatók, akik több elemzési egységet használnak, mint például Lauderdale és Herzog (2016); náluk elemzési egységek voltak a beszélők általában, a vitában felszólalók, a viták, a szavak vagy maga a törvényhozás. Más tanulmányok különböző elemzési egységeket használnak, mert a magasabb elemzési egységekhez kapcsolódó polarizációt az alacsonyabb szinten mért polarizációra visszavezetve mérik, mint például Hemphill et al. (2016), akik kifejlesztettek egy mérőszámot annak jelzésére, hogy egy hashtag mennyire polarizált, és a felhasználókhoz az általuk használt hashtagek alapján rendelnek egy mérőszámot.

Külső adatok felhasználása

A szöveges adatokon alapuló gépi tanulási modellek általában csak a szöveg jellemzőit használják változóként. A modellek teljesítményének javítása érdekében külső adatok is bevonhatók az elemzésbe, vagy - ami még gyakoribb - a szövegeket a "valósággal" összekapcsoló oksági modellek létrehozására használják fel őket. Farrell (2016) például pénzügyi adatokat is felhasznál annak vizsgálatára, hogy vállalati finanszírozásban részesült. illetve nem részesült szervezetek különböznek-e az éghajlatváltozásról folytatott diskurzusuk tartalmát vagy nyelvezetét tekintve. Hasonlóképpen, Hausladen és társai (2020) az amerikai kerületi bírósági döntések ideológiai irányának előrejelzésekor azt vizsgálták, hogy a bírát jelölő elnök párthovatartozása hogyan befolyásolja a bíra döntéseit. Ash és munkatársai (2017) szintén az amerikai kerületi bírósági bírák polarizációját vizsgálták, olyan külső adatokat használva a bírákról, mint a kerület, az év, a nem és a többségi párt. E külső adatok felhasználásával olyan eredményekre jutottak, amelyek ok-okozati magyarázattal szolgálhatnak, például azt találták, hogy a jelölésre esélyes bírák kevesebb különvéleményt írtak, ha a szenátust az ellenzéki párt irányította.

Más tanulmányokban kontrollváltozóként használnak külső adatokat. Decadri és Boussalis (2020) például a populizmus és a beszédkomplexitás közötti kapcsolatot vizsgálta az olasz parlamenti beszédekben, és kontrollváltozóként a párttagságot és számos demográfiai jellemzőt használt. A kontrollváltozók bevonása segített nekik elkerülni a zavaró hatásokat. Gondoljunk arra például, hogy ha például a populista pártoknak kevesebb női képviselőjük van, és a női képviselők beszédei eltérnek a férfiakétól, akkor célszerű a nemet is bevonni a pártok közötti összehasonlító elemzésbe.

A külső adatok bevonásának másik gyakori motivációja az eredmények validálása. Lauderdale és Herzog (2016) szövegelemzéssel pozicionálta a politikai pártokat, és az eredményeket két külső viszonyítási alaphoz hasonlította: ahhoz, hogy a pártok tagjai voltak-e a kormánykoalíciónak, és hogy a szakértői megítélések alapján a bal-jobb politikai spektrumon milyen becsült pozíciót foglalnak el. Jensen et al. (2012) és Green et al. (2020) a nyelvi polarizáció mértékét egy bevett, nem nyelvi politikai polarizációs mérőszámmal (DW-NOMINATE, lásd McCarthy et al., 2006) hasonlította össze. Belcastro és munkatársai (2020) a 2018-as olaszországi általános választások és a 2016-os amerikai elnökválasztás során osztályozták a Twitter-felhasználókat, és azt találták, hogy az eredmények nagyon közel állnak a tényleges választási eredményekhez, sőt, pontosabbak voltak, mint a közvélemény-kutatások átlaga.

A közvélemény-kutatás és a szöveges adatelemzés együttes alkalmazása nagyon ritka volt: Chen et al. (2014) és Kobayashi et al. (2019) Reddit-felhasználókat toboroztak, hogy részt vegyenek egy felmérésben, és ezzel egyidejűleg összegyűjtötték a résztvevők legfrissebb Reddit-bejegyzéseit/kommentjeit is. Gauvin et al. (2016) két különböző célcsoportot ért el a kétféle adattípussal: közvélemény-kutatási adatokat (választók) és politikai pártok programjait (politikusok) használtak a választók/pártok kongruenciájának modellezésére.

A felhasznált adatokból levonható legfontosabb tanulságok

Gyakran figyelmen kívül hagyott probléma az adatok jellegének fontossága. Az adatok gyűjtésének és szűrésének módja, az adatok keletkezésének kontextusa és a szövegek műfaja mind olyan fontos tényezők, amelyek meghatározzák például a modell teljesítményét, vagy azt, hogy egy gépi tanulási modell átvihető-e egyik adatbázisról a másikba (ezt a problémát nevezik átvihetőségnek - *transferability*, vagy területközi általánosíthatóságnak – *cross-domain generalizability*).

Cohen és Ruths (2013) bebizonyította, hogy ha a korpusz definiálásához politikailag erősen megosztó hashtagek-et használunk (ami szokásos kutatási gyakorlat), akkor politikailag polarizált felhasználókkal dúsítjuk fel mesterségesen a vizsgálati populációt, amin viszont az osztályozó modell jobb teljesítménnyel működik. Eredményük arra is utal, hogy a korábban mások által közölt teljesítmény-mutatók szisztematikusan felül voltak becsülve, vagyis túl optimisták voltak. Azt is kimutatták, hogy az osztályozó modellek nem használhatók a létrehozásukkor használt politikai orientáción kívüli tartományban osztályozásra.

Más szóval, az *eltérő politikai aktivitású emberek csoportjai* nagyon különböző nyelvet is használnak. Ezért nem áll meg az a gyakori (implicit) feltételezés, mely szerint a legszélsőségesebb egyének esetén ugyanazt a jelenséget detektálhatjuk, csak könnyebben kimutatható módon. Pl. Diermeier et al. (2012), Morini et al. (2020) és Grover et al. (2019b) implicit módon ezt a feltételezést követik a kifejezettebb/extrémebb esetek kiválasztásakor. Ugyanígy Cotelo et al. (2016) is, akik csak olyan egyértelműen kódolható tweeteket vettek be elemzésükbe, amelyeket a kódolók azonos politikai címkével láttak el (vagyis ahol nem volt eltérés a kódolók besorolása között).

Yan et al. (2017) Cohen és Ruths (2013) eredményéhez hasonló következtetésre jutott három *különböző műfajú, különböző nyilvánossági rétegből származó* szövegkorpuszon. A politikai diskurzus különböző rétegeit vizsgálták, és azt találták, hogy bár a modellek jól teljesítenek a kiindulási adathalmazon belül, de az egyik adathalmazról a másikra való általánosítási képességük gyenge. Egy másik fontos eredményük azt mutatja, hogy az osztályozás sikere nemcsak terület-, hanem *időfüggő* is: az osztályozás hatékonysága csökken, ahogy a tesztadatok időben távolodnak a tanuló adatoktól. Ez kérdéseket vet fel Diermeier et al. (2012) módszertanával kapcsolatban, amelynek tanuló halmaza a 101-107. szenátusban elhangzott beszédekből áll, a teszthalmaz pedig a 108. szenátusból származik.

6.6. Módszerek

6.6.1. Szövegelemzési módszerek

Az alábbiakban röviden és intuitív módon vázolom a tanulmányokban leggyakrabban alkalmazott módszerek közül azokat, amelyek kifejezetten a politikatudományra jellemzők (Wordfish, Wordscores, Wordshoal). A többi, általános és széles körben használt NLP-módszert (topikmodell, szóbeágyazás, felügyelt gépi tanulás, szentimentelemzés) a 4-5. fejezetben már tárgyaltam.

A politológiának standard megközelítései vannak egy szöveg politikai álláspontjának meghatározására. Ezek skálázási módszerek, azaz egy numerikus dimenzió mentén helyezik el a szöveget. A **Wordscores** modellt Laver et al. (2003) dolgozta ki: a szövegeket más, eleve ismert pozícióval rendelkező szövegekhez hasonlítja, mégpedig úgy, hogy az adott szövegben található szófrekvenciákat összehasonlítja a skála szélső végpontjait meghatározó referencia-

szövegek szófrekvenciáival; ebben az értelemben felügyelt módszerről van szó. A **Wordfish-t** Slapin és Proksch (2008) fejlesztette ki, és a Wordscores-ral ellentétben nem felügyelt, mivel nem igényel a priori pozícionált szövegeket: a szófrekvenciák alapján egy mögöttes látens változót becsül, amely az ideológiát tükrözi. A **Wordshoal** modell (Lauderdale & Herzog, 2016) a Wordfish kiterjesztése: míg az utóbbi csak egy dimenziót hoz létre, az előbbi lehetővé teszi további, az ideológiát befolyásoló dimenzió(k) bevonását. A Wordshoal lehetővé teszi metaadatok bevonását is, mint például a vita tárgya, mivel a szóhasználat mintázata különböző tárgyú vitákban eltérő lehet.

6.6.2. A polarizáció operacionalizálása

A polarizáció operacionalizálásakor először a politikai pozíciók mérésének mikéntjét, majd ennek alapján a polarizáció mérésének módját kell meghatározni. A vizsgált tanulmányok jellemzően vagy ideológiai skálázást vagy osztályozást használtak egy adott szöveg politikai pozíciójának meghatározására.

A skálázást többnyire a szokásos politikatudományi megközelítésekkel végezték (lásd fent: Wordscores, Wordfish vagy Wordshoal, n=8). Gross és munkatársai (2020) a Wordfish alkalmazásával a németországi pártok helyi kiáltványainak adathalmazán azonosították a lokális pártkonfliktusok dimenzióit. Medzihorsky et al. (2014) a Wordfish segítségével ki tudta mutatni, hogy a 2012-es amerikai republikánus jelöltek távolabb kerültek a hagyományosabb republikánus ideológiától. A Wordshoalt használta például Goet et al. (2017).

A skálázás másik megközelítése az ideális pont-modellekből (*ideal point model*) származik (a DW-NOMINATE, McCarthy et al. 2006, szintén ide tartozik). A modellek a törvényhozók politikai pozícióját törvényhozási szavazataik alapján becsülik meg. E modellek általánosításának célja, hogy szövegeket is be tudjanak építeni. Nguyen et al. (2015) ezt a célt a hierarchikus ideális pont topikmodell bevezetésével kívánta elérni. Ez a modell nem csak a szavazatokat, hanem a kapcsolódó törvényjavaslat-szövegeket és maguknak a törvényhozóknak a nyelvezetét is felhasználja, és inputjai között szerepel a törvényjavaslatok témája is. Gerrish és Blei (2012) alternatív megoldása ugyanerre a problémára a témához igazított (*issue-adjusted*) ideális pont modell, amely a törvényjavaslatok tartalmát is figyelembe veszi. Az elképzelés lényege, hogy a törvényjavaslatra leadott szavazatok alapvetően a törvényhozó általános ideológiai álláspontjától függenek, amelyen esetleg (kissé) változtatnak a törvényjavaslat tartalma

szerint. Vagyis a törvényhozókhoz tartozó ideális pontok pozíciója megváltozhat attól függően, hogy milyen kérdésről szavaznak.

A skálázás alkalmazásakor a polarizáció mértéke a pozíciók eloszlásából ítélhető meg. Ebben az összefüggésben Goet (2017) egy explicit mérőszámot határoz meg, amely azt a konzisztenciát méri, amellyel a képviselők különböző politikai kérdésben a pártjuk álláspontjához közel helyezkednek el. Az "1"-es pontszám tökéletes polarizációt jelent, nulla átfedéssel a különböző pártok képviselői között.

A skálázást inkább a politikatudományi szerzők használták, míg mások *a felügyelt osztályozást* (51 tanulmány használt osztályozást, többnyire felügyelt változatot). Az osztályozás ebben az összefüggésben olyan módszer, amely a szöveg szerzője által használt szavak alapján próbálja azonosítani annak álláspontját. Ezek a tanulmányok explicit vagy implicit módon osztályozási problémának tekintik magát a polarizációt. A magas osztályozási teljesítmény arra utal, hogy az adott politikai oldal által használt nyelv homogén, és különbözik a többi oldal által használt nyelvtől. Meghatározható egy polarizációs metrika is: minél hatékonyabban azonosítja az osztályozási modell a szerző álláspontját, annál nagyobb a polarizáció. Ebben a megközelítésben a legexplicitebbek Goet (2017) és Green et al. (2020) írásai. Bayram et al. (2019) ezt a megközelítést követve elemezte a Egyesült Államok Kongresszusa Képviselőházának felszólalásait, és egyértelmű emelkedő tendenciát észlelt az osztályozási teljesítményben, ami azt jelzi, hogy a polarizáció egyre jobban detektálhatóvá vált a beszédek nyelvezetében. Gentzkow és társai (2019) sokat idézett tanulmányukban a Taddy (2013, 2015) által kidolgozott módszerekre építve egy hasonló logikát követő polarizációs mérőszámot javasolnak. Egy multinomiális beszédmodellt határoznak meg pártonként eltérő kiválasztási valószínűségekkel, a polarizációt pedig azzal mérik, hogy a modellt ismerő megfigyelő mennyire könnyen tudja kitalálni a beszélő pártját pusztán a beszélő szó-kiválasztásából. Kelly és munkatársai (2021) tovább fejlesztették ezt a módszert.

A skálázás és a felügyelt osztályozás mellett más, kevésbé elterjedt módszerek is használatban vannak. Samantray és Pin (2019) a Lelkes (2016) által kifejlesztett ideológiai divergencia mutatót használta, amely egy numerikus változó eloszlásának bimodalitása alapján jellemzi a polarizáció szintjét. Ez utóbbi változó a hagyományos politikatudományi kutatásokban közvélemény-kutatásokból származik, itt szövegek jellemzőjeként generálták. Darwish (2019) Garimella et al. (2018) nyomán Twitter-adatokon használt polarizációs mérőszámot, amely az ellentétek mértékét a csevegés-hálózat jellemzőiből méri. Budhiraja és Pal (2020) az egyes politikusokat a tweetjeik tartalma alapján szóbeágyazási vektorként

reprezentálta, majd a polarizációt a kapott pontfelhő pártok szerinti elkülönüléseként azonosította. Villa-Cox és társai (2021), általánosítva KhudaBukhsh és társait (2020), a polarizációt gépi fordítással értelmezték. Keretrendszerük úgy tekint két különböző politikai állspontot képviselő alpopulációra, mint két különböző nyelven beszélő csoportokra. Önálló kifejezéseket fordít a modell, a gépi fordítót a két populáció közösségi média-korpuszán tanítva, majd a kapott kifejezés-párokon belül a nem-azonos párok arányát azonosítja a polarizáció számszerűsíthető mértékeként. Ilyen nem-azonos, de a fordító szerint megegyező párként azonosították pl. az “all lives matter” és “black lives matter” fordulatokat. A polarizáció operacionalizálására más javaslatok is születtek, lásd Gross et al. (2013), valamint Acree et al. (2020).

Goet (2017) koherens elméleti keretet fogalmazott meg a szövegalapú polarizáció-mértékek értékeléséhez. Az ő kritériumai szerint egy érvényes szövegalapú mérőszámnak meg kell felelnie a priori elvárásainknak, például ha egy becslés kiugróan magas a többihez képest, akkor meg kell felelnie annak, amit az adott kontextusban erről a kiugró polarizációról történelmileg tudunk. Ez a munka azért fontos, mert foglalkozik a validálással is, ami az áttekintett tanulmányok zöméből teljesen hiányzik, és olyan kritériumokat ad, amelyek a gyakorlatban könnyen követhetők.

6.6.3. Az időbeli változások vizsgálatára használt módszerek

Az időbeli változásokat vizsgáló legtöbb tanulmány egyszerűen több szakaszra osztja az időintervallumot, és minden egyes szakaszon külön-külön elvégzi ugyanazt az elemzést. Például Grover et al. (2019a), Stecula és Merkley (2019), valamint Mendez et al. (2018) az érzelmek változását vizsgálták. Sapiro-Gheiler (2018), Ash et al. (2017), valamint Stecula és Merkley (2019) külön osztályozási modelleket fejlesztettek az egyes időintervallumokon, és vizsgálták teljesítményük időbeli változásait. Appelrouth (2019) a témák időbeli változását vizsgálta topikmodellek segítségével. Brigadir et al. (2015) és Rumshisky et al. (2017) szóbeágyazási modelleket alkalmaztak, és a szavak jelentésében bekövetkező változásokat, azaz a szó-távolságok és szomszédságok időbeli változását vizsgálták. Medzihorsky és munkatársai (2014) a Wordfish alkalmazásával a törvényhozási viták átiratain azonosították a politikusok ideológiai pozícióját, és vizsgálták a pozíciók időbeli változásait.

Más tanulmányok az NLP-t hagyományos, az idő modellezére alkalmas statisztikai módszerekkel kombinálták. Tsur és munkatársai (2015) például

idősoros regressziót alkalmaztak a topikmodellből származó téma-hovatartozásokra. Gross et al. (2020) lineáris mixed-effect modelleket használt, amelyekben a függő változó a Wordfish által becsült pozíció volt. Hofmann et al. (2020) idősor-modellezést alkalmazott általánosított additív modellekkel a pártok közötti lexikai különbségek összehasonlítására.

Voltak olyan tanulmányok, amelyek a strukturális topikmodellt (STM) használták, amely képes metaváltozóként közvetlenül beépíteni az időt is. Farrell és munkatársai (2016) például amerikai szervezetek által az éghajlatváltozásról írt szövegeket elemeztek egy 20 éves időszak alatt, és STM-et alkalmaztak annak vizsgálatára, hogy a szervezet más vállalatok általi támogatása hogyan befolyásolta a topikok időbeli változását.

Más szerzők, például Gross et al. (2013), Acree et al. (2020) és Iliev et al. (2019) saját statisztikai modelleket dolgoztak ki az ideológiai pozíciók időbeli változásainak azonosítására.

6.6.4. Osztályozási modellek

Felügyelet nélküli osztályozás

A felügyelet nélküli osztályozás legfőbb előnye, hogy nem igényli a vizsgált alanyok előzetes politikai-ideológiai címkézését, ezért itt nincs szükség mély szakterületi tudásra. Stefanov és munkatársai (2019) és Darwish és munkatársai (2020) a Twitteren használtak felügyelet nélküli felhasználói álláspont-detektálást (*stance detection*). Miután a felhasználókat egy alacsonyabb dimenziós térre vetítették, klaszterezést alkalmaztak, ami lehetővé tette számukra, hogy megtalálják a különböző karakteres álláspontokat reprezentáló fő-felhasználókat. A címkézett fő-felhasználókat aztán egy későbbi osztályozó modell tanítására lehetett felhasználni. Egy másik példa a felügyelet nélküli osztályozásra a klaszterelemzés, amelyet például Giglietto et al. (2018) alkalmazott.

Felügyelt osztályozás

A felügyelt osztályozást alkalmazó releváns tanulmányokat több szempont szerint dolgoztam fel: a célváltozó jellege, az annotáció módja, a strukturális információk használata és az osztályozáshoz használt változók (*features*) jellege.

A különböző modellek a politikai pozíció különböző definíciói szerint különböző **célváltozókat** használtak: mint az álláspont (pl. Fang et al., 2015), párthovatartozás (Wu et al., 2019) vagy ideológia (Ademmer et al., 2019). Specifikus célváltozót használt például Gerrish és Blei (2012), akik jogi szövegek alapján jósolták meg a szövegre leadott szavazatot. A tanulmány egyedülálló abban a tekintetben, hogy a szövegre adott reakciót jósolja meg, nem pedig a szerző ideológiai álláspontját. Cotelo és munkatársai (2016) két pártra vonatkozó álláspontot használtak célváltozóként ahelyett, hogy csak egy pártra vonatkoztatták volna (pozitív, negatív vagy semleges álláspontokkal), és az osztályozó feladata az volt, hogy a tweeteket a kilenc kombinatorikus kategória valamelyikébe sorolja be. Stecula és Merkley (2019) az éghajlatváltozással kapcsolatos vita keretezését elemezték az amerikai médiában, háromféle kerettípusra összpontosítva, és ezek segítségével határozták meg a modelljük által előrejelzett álláspontokat. Chen et al. (2014) a célváltozót egy felmérésben adott válaszok alapján határozta meg, Wang és Tucker (2021) pedig a szentiment pontszámot (*sentiment score*-t) használta célváltozóként. Gelman, és Wilson (2022) törvényhozók pártállását azok Twitter-retorikája alapján mérték, ahol a tweeteket pártos vagy egyéb (kétpárti, pártok közötti, párton kívüli) kategóriákba sorolták.

A címkék létrehozása a felügyelt osztályozó tanításához (szakterminussal: a szövegek *annotáció*ja) szintén fontos kérdés, mivel a címkézett szövegek elérése, létrehozása önmagában is kihívást jelenthet. Hausladen és munkatársai (2020) kézzel annotált korpuszt használtak az amerikai bírósági döntések ideológiai irányának előrejelzésére - az annotáció ebben az esetben nyilvánvalóan kihívást jelentett. Ha az annotációs feladat könnyen megtanítható, akkor a *crowdsourcing* is alkalmazható, mint például Lai et al. (2020) esetében, akik a CrowdFlower nevű szolgáltatást használták, vagy Wang és Tucker (2021) esetében, akik az Amazon Mechanical Turk-öt. Az annotálás és kifejezetten a crowsoursing annotálás problémájáról részletesen is írok e kötet 5.1., *Az annotálás kihívásai a szociológiai alkalmazásokban* c. fejezetében.

Sok más esetben nincs szükség manuális címkézésre, mivel a címkék külső adatokból is beszerezhetők. Kézenfekvő módon ez a helyzet (az ismert pártállású) politikusok szövegei esetében. Egy kevésbé triviális példát mutatott be Jelveh et al. (2014), akik amerikai közgazdászok által írt tudományos cikkeket vizsgáltak, és a szerzők politikai kampányadományai illetve petíció-aláírási tevékenysége alapján határozták meg azok politikai irányultságát. Zubiaga et al. (2017) Twitter-felhasználóknak a katalán függetlenségi mozgalommal kapcsolatos álláspontját

osztályozta. Címkézésük a felhasználók bio-jára támaszkodott, ahol a felhasználók azt is feltüntették, hogy melyik az a földrajzi terület, amelynek állampolgárának tekintik magukat. Ez az információ közvetlenül jelzi a felhasználók függetlenségi mozgalommal kapcsolatos álláspontját, pl. a „Països Catalans" vagy „PPCC" egyértelműen a független Katalóniára utal.

Más esetekben a címkét a nagyobb elemzési egységektől a kisebbek öröklik. Kulkarni et al. (2018) hírforrások osztályozásakor az AllSides.com (a legnagyobb médiumok politikai irányultságát értékelő platform) által biztosított címkéket használták (Baloldali, Inkább baloldali, Közép, Inkább jobboldali, Jobboldali), és a cikkekre is a hírforrásuk szerinti címkét ragasztották. Karamshuk et al. (2016) kézzel végezték a hírforrások címkézését, és a Twitter-felhasználókat az általuk megosztott hírforrások alapján címkézte. Rao et al. (2020) még tovább folytatta az öröklési láncot, és a címkéket a webdomainekről a felhasználókra, majd a felhasználókról más felhasználókra vitte át a retweet-mintázatok alapján. Kobayashi et al. (2019) azonban rámutatott, hogy a fenti öröklésen alapuló megoldások feltételezik, hogy a Twitter-felhasználók előszeretettel követnek olyan médiumokat és politikusokat, akiknek ideológiai álláspontja hasonló a sajátjukhoz, miközben ez a feltételezés nem feltétlenül áll.

A felügyelt és a felügyelet nélküli osztályozás között helyezkedik el a **félig felügyelt osztályozás**, ahol nincs szükség címkézésre, de a szakterületi ismeretekre igen. Belcastro és munkatársai (2020) neurális hálózatok segítségével derítették fel a közösségi média felhasználóinak polarizációját a választási kampányok során. Egy olyan új, félig felügyelt megközelítést hoztak létre, amely közismerten nagyon erősen polarizáló hashtagek egy kisebb halmazából létrehozott osztályozási szabályokból indul ki, majd iteratív módon új osztályozási szabályokat generál működése során.

A szövegek mellett néhány tanulmány az adatbázis *strukturális* információit is felhasználta az osztályozáshoz. Ez különösen a közösségi média-adatok esetében valósítható meg, ahol a felhasználók közötti kapcsolati háló rendelkezésre áll. Többen megállapították, hogy a strukturális információk bevonása növeli a modell osztályozási pontosságát (pl. Cotelo et al., 2016; Conover et al., 2011). Wang és munkatársai (2017) tweetek osztályozásával azt találták, hogy a legjobb modell a szövegeket és a *képeket* is integráló modell. Érdemes azonban megjegyezni, hogy ha az elsődleges cél a nyelvi polarizáció vizsgálata (a "legjobb" osztályozó meghatározása helyett), akkor magának a nyelvnek a fontosságát kell

vizsgálni. Ami viszont lényeges kérdés lehet, az az, hogy a nyelvi információ mekkora hányadát teszi ki a teljes polarizációnak.

Egy másik lényeges kérdés, hogy *milyen jellemzőket használtak* a modellek, ha szövegalapú osztályozást alkalmaztak. Leggyakrabban a szózsákos (*bag of words*, ahol minden egyes szó önállóan jelent input adatot a modell számára) megközelítést vagy n-grammokat (n hosszú karaktersorozatokat) használtak, figyelmen kívül hagyva a szintaxist (pl. Acree et al., 2020; Bayram et al., 2019). Az újabb modellek a nyelv kompozíciós aspektusát is modellezik, például neurális hálózat alkalmazásával (pl. Kulkarni et al., 2018; Belcastro et al., 2020).

Számos tanulmány a nyers szöveg mellett előre meghatározott szövegtulajdonságokat is használt az osztályozó inputjaként. Potthast et al. (2018) politikailag megosztó híreket vizsgált, és különböző stilometriai jellemzőket használt, például olvashatósági pontszámot, idézett szavak arányát, bekezdések számát stb. A hashtagekről több tanulmány is bizonyította, hogy javítják az osztályozók teljesítményét (pl. Conover et al., 2011). Mivel a hashtagek rövidek és információban gazdagok, csökkenthetik a zajt.

Néhány tanulmányban (pl. Wang et al., 2017; Zubiaga et al., 2017; Karamshuk et al., 2016; Rao et al., 2020) szóbeágyazást használtak a modell inputjának (a feature-öknek) a létrehozására. Zubiaga és társai (2017) pl. a szóbeágyazást dimenziócsökkentésre használták: a felhasználó idővonal-tartalmának szóbeágyazott reprezentációját használták feature-ként. Karamshuk et al. (2016) a szóbeágyazás egy másik fontos alkalmazását szemlélteti: kiinduló szótáruk a pártos retorika indikátorainak tekintett kifejezésekből állt, majd ezeket a szavakat a szóbeágyazási reprezentációjuk alapján a leghasonlóbbakkal egészítették ki, hogy megkapják a végső feature-készletet.

Más tanulmányok topikmodell outputját (pl. Fang et al., 2015; Rao et al., 2020), szerzői szintű szövegjellemzőket (pl. Cohen & Ruths, 2013) vagy hírforrás-szintű jellemzőket (pl. Baly et al., 2020) használtak. Baly és munkatársai (2020) azt szemléltetik, hogyan lehet különböző forrásokból származó szövegeket felhasználni: a szerzők célja cikkek politikai ideológiájának osztályozása volt, és médiaszintű jellemzőket határoztak meg (1) a médium Twitter-követőinek biojai, (2) a médiumot leíró Wikipedia-oldal tartalma, valamint (3) a médium weboldalának web-forgalmára vonatkozó információi alapján.

6.6.5. Topikmodellezés

Minden hatodik tanulmány topikmodellezést alkalmazott. Gyakran használták a módszert pl. egy téma keretezésének elemzéséhez. Így például Shen és Rosé (2019) a Reddit moderációs elveinek változására adott polarizált felhasználói válaszokat vizsgálta topikmodell alkalmazásával, és azt találták, hogy a baloldali és jobboldali beállítottságú felhasználók eltérő arányban használták mind a topikmodell által detektált témákat, mind a témákra jellemző szavakat (ez utóbbit azonosították a szerzők témán belüli keretezésként).

Tsur et al. (2015) szerint a modell által azonosított témák együttjárása is jellemzi azok keretezését. Egyik példájukban azt mutatják meg, hogy az amerikai nyilvánosságban sokat vitatott Keystone XL projektet a republikánusok a projektnek a munkaerőpiacra és a kisvállalkozásokra gyakorolt hatásával keretezik (az "Energia" témát a "Költségvetés és gazdaság" témával együtt használják).

Mások, pl. Sinno et al. (2021), a topikmodellt csak technikai okokból alkalmazták, azért, hogy tartalmilag koherens korpuszt hozzanak létre a releváns témákhoz nem kapcsolódó cikkek kihagyásával.

A topikmodell egyik verziója, a strukturális topikmodell (lásd a 4.1.3 fejezetet) azért előnyös, mert lehetővé teszi a dokumentum olyan metaadatainak beépítését, amelyek potenciálisan befolyásolhatják mind a topikok előfordulási gyakoriságát, mind tartalmát (Wesslen, 2018). Sanders et al. (2017) pl. az Egyesült Királyság 2010-2015 közötti parlamenti időszakának szakbizottsági meghallgatásait vizsgálta, és olyan dokumentumszintű metaadatokat hazsnált, mint "kamara", "párt" stb. A pénzügyi metaadatok bevonása lehetővé tette Farrell et al. (2016) számára, hogy teszteljék a vállalati finanszírozás hatását arra, hogy különböző szervezetek hogyan tárgyalják az éghajlatváltozást, azaz azt a kérdést tesztelték, hogy a finanszírozás nagyságának van-e jelentős hatása bizonyos témák előfordulási gyakoriságára.

Mások új típusú topikmodellek bevezetését javasolták, pl. Thonet at al. (2017), Trabelsi and Zaiane (2018) és Koylu et al. (2019).

6.6.6. Szentimentelemzés

Minden hatodik tanulmány (n=26) alkalmazott szentimentelemzést. Leggyakrabban *szótáralapú szentimentelemzést* (lásd e könyv 3. fejezete) alkalmaztak (pl. Coutto, 2020; Grover et al., 2019a; Grover et al., 2019b). Grover et al. (2019b) például az USA-ban a bevándorlásról szóló vitában képviselt két

ellentétes oldal közötti erkölcsi, affektív és kognitív különbségeket vizsgálta, a nyelvhasználatra koncentrálva, az LIWC szótár (lásd e könyv 3. fejezet) segítségével. Elemezték pl. a negatív érzelmekhez kapcsolódó, illetve a Gondoskodás/Károsítás és a Tisztességesség/Csalás dimenziók mentén jelentkező különbségeket. A *nem szótár-alapú módszereket* alkalmazók között van Wang és Tucker (2021), akik felügyelt gépi tanulási modelleket használtak arra, hogy sajtóközleményekhez szentiment-pontszámokat rendeljenek.

6.6.7. Szóbeágyazás

Minden nyolcadik tanulmány (n=19) alkalmazott beágyazást, főként szóbeágyazási módszereket (a módszerről lásd e könyv 4.2 fejezetét). A szavak beágyazásának gyakran alkalmazott alkalmazása a 3.3.4. pontban már említésre került, ahol egy kezdeti szótárt hasonló jelentésű kifejezésekkel bővítenek. Ezt a módszert nem felügyelt kontextusban is alkalmazták (pl. Decadri & Boussalis, 2020), hogy polarizációs szótárat kapjanak.

A modell egy másik inspiráló felhasználását példázza Brigadir et al. (2015), illetve Bonikowski et al. (2019): esetükben a szóbeágyazási vektorteret használják bizonyos jellemző kifejezések jelentésében bekövetkezett változások felismerésére. Brigadir et al. (2015) a szavak szemantikájában bekövetkező változásokat követte mind időben, mind az eltérő nézeteket valló közösségek összevetésével. Két esettanulmányt tettek közzé, amelyek a skót függetlenségi népszavazás és a 2014-es amerikai félidős választások kapcsán vizsgáltak szembenálló ideológiájú közösségeket, és a két közösséghez tartozó szóbeágyazási vektorterekben követték időbeli változását bizonyos szópárok távolságának, azaz jelentéskülönbségének. A vizsgált szavak a kampányok kulcs-témáival kapcsolatos szavak voltak. Bonikowski et al. (2019) az amerikai elnökjelöltek 2016-os kampánybeszédeinek vizsgálatát végezték el, például a "veszélyes" kifejezés szóbeágyazási szomszédságára összpontosítva, amely azt mutatja, hogy a jelölt mit tart a legsürgetőbb problémának (míg például Trump esetében a "menekültek" jelentése állt közel a "veszélyes"-hez, Clinton esetében az "előítéletek"-é).

KhudaBukhsh et al. (2020) a szóbeágyazás hatékony és jól interpretálható alkalmazását mutatta be, a szóbeágyazáson alapuló gépi fordítás alkalmazásával négy neves amerikai híroldal YouTube-csatornájának vitaszekcióján. Megmutatták, hogy a CNN és a Fox News két alközössége két különböző nyelven beszél: amit az előbbi például "demokratáknak", "biden"-nek, "kkk"-nek és

"bevándorlóknak", azt az utóbbiak "republikánusoknak", "creep"-nek, "blm"-nek és "illegálisoknak" címkézik.

Végül a szóbeágyazást is használták a kutatások felhasználók klaszterezésére, például Rashed et al. (2020) a felhasználókat szövegeik alapján egy beágyazási térben reprezentálta, majd a reprezentációkat egy alacsonyabb dimenziójú térre vetítették, amelyen klaszterelemzést alkalmaztak.

6.6.8. Több NLP-módszer kombinálása

Az alábbiakban néhány inspiráló példát mutatok olyan tanulmányokra, amelyekben több NLP-módszert együttesen alkalmaztak.

Rho és társai (2018) három, politikailag különböző médium oldaláról származó, a "#MeToo" kifejezést tartalmazó Facebook-hozzászólásokat vizsgáltak. A három oldal diskurzusainak összehasonlítása (1) LIWC-vel történt a nyelvi és affektív jegyek szempontjából, (2) szóbeágyazással a "MeToo"-val együtt előforduló közeli szavak azonosítására, és (3) egy, a szavak szövegbeli relevanciáját figyelembe vevő szósúlyozási sémával (*tfidf - term frequency inverse document frequency*) a kulcsszavak felderítésére.

Grover és munkatársai (2019a) a bevándorlásról szóló politikai vita két ellentétes oldala közötti erkölcsi, affektív és kognitív nyelvhasználati különbségeket vizsgálták, mind szentimentelemzés, mind topikmodell segítségével. Két elemzési komponenssel alkalmazták a szentimentelemzést: polaritáselemzéssel (a felhasználók véleményének pozitív, negatív és semleges kategóriákba sorolása) és érzelemelemelemzéssel (nyolc érzelem, például "harag"). A topikmodellt a tweeteken belüli legfontosabb témák azonosítására használták.

Grover et al. (2019b) LIWC és szóbeágyazás segítségével vizsgálta az amerikai bevándorláspolitikai vita oldalainak különbségét a Twitteren. Szóbeágyazást alkalmaztak az olyan kulcskifejezések szemantikai szomszédságának feltárására, mint a "bevándorló", "muszlim" vagy "menekült". Szembetűnő különbséget találtak például a "bevándorló" kifejezés szemantikai környezetét tekintve ("Alien" a bevándorlásellenes tweetek esetében vs. "Undocumented" a bevándorláspárti tweetek esetében).

Stecula és Merkley (2019) az éghajlatváltozásról szóló híradásokban megjelenő különböző keretezéseket vizsgálták. Felügyelt osztályozást használtak háromféle keretezés-típushoz tartozó célváltozókkal, és LIWC-t a keretezésekhez kapcsolódó nyelvezet vizsgálatára.

6.6.9. A szakterületi tudás szerepe az elemzésben

A frissebb tanulmányok, mind az akadémiai, mind az üzleti életben, kiemelik a szakterületi tudás (*domain knowledge*) beépítésének fontosságát az adattudományba (pl. Angelova, 2010; Liu & Zhang, 2019; Deng et al. ,2020). A szakterületi tudás az adattudományi projekt minden egyes lépésénél fontos, beleértve a kutatási kérdés megfogalmazását, az adatgyűjtést, az előfeldolgozást, a modellezést és az eredmények értelmezését. Ezért áttekintettem a review során azt is, hogy a szerzők a kutatás bármelyik szakaszában felhasználták-e a szakterületi ismeretet, például, hogy osztályozási modellek esetén a modellteljesítmény értékelése mellett megpróbálták-e értelmezni az eredményeket.

Az annotáció, ha egy társadalomtudományi kutatási kérdéshez szükséges, még a félig felügyelt osztályozás esetében is domain-ismeretet igényel: lásd pl. Belcastro et al. (2020), akik egy korlátozott számú osztályozási szabálykészletet határoztak meg olyan hashtagek alapján, amelyek bizonyos politikai oldalakra jellemzőek. Itt a releváns hashtagek kiválasztásához nyilvánvalóan szaktudásra volt szükség.

Legtöbbször az elemzés maga is érdemi kérdések mentén jön létre, ilyenkor e kérdések megfogalmazásához is tartalmi ismeretekre van szükség. Stecula és Merkley (2019) például gazdag szaktudást épített az éghajlatváltozás keretezését elemző kutatásába, amikor korábbi kutatások alapján háromféle keretezést definiált. Decadri és Boussalis (2020) a populizmus és a beszédkomplexitás közötti kapcsolatot vizsgálta Olaszországban, érdemi hipotézisek, az olasz populista retorika szótára és jól megválasztott metaadatok segítségével.

Érdemes megjegyezni, hogy az utóbbi kettő esetben a tanulmányok szerzői között nemcsak informatikusok, hanem szakterületi szakértők is voltak. Vannak példák, amik azt mutatják, hogy ha a szerzők között csak informatikusok vannak, akkor gyakran hiányzik az értelmezés, és a kutatási kérdések is inkább technikai jellegűek (pl. Tucker et al., 2020). És fordítva: ha csak társadalomtudós társszerzők vannak, akkor a módszertan meglehetősen egyszerű, bár a tanulmány tartalmilag gazdag: a kutatási kérdések kifejtésre kerülnek, a meta-változók jól megválasztottak, és az eredmények beágyazódnak egy meglévő tudományos diskurzusba (pl. Decadri & Boussalis, 2020, valamint Coutto, 2020).

Vannak olyan politikatudományi kutatási területek, ahol az intézményesített tudás rendszerezett formában áll rendelkezésre, mint például Comparative Agendas Project (CAP), melyben a hazai Társadalomtudományi Kutatóközpont

is részt vesz. A projekt olyan kódolt adatokat szolgáltat, amely lehetővé teszi a politikai ágendák érvényes összehasonlítását különböző témák és időpontok között. A CAP-ot például Praet et al. (2021) arra használta, hogy összehasonlítsa a politikai pártok Twitteren folytatott témaspecifikus kommunikációját. A CAP-hoz hasonló nemzetközi kategorizálási rendszerek nagyban támogathatják az NLP-kutatásokat.

A szakterületi tudás szerepe az interpretációs lépésben, és különösen a felügyelt osztályozás esetén a legfontosabb. Értelmezés nélkül a prediktív modellek fekete dobozok (Molnar, 2019), hiszen csak azt tudjuk, melyik megfigyelést melyik osztályba sorolt a modell, azt nem, hogy miért, mely jellemzői alapján. Ha megértjük, hogy a modell miért hozott egy bizonyos döntést, és megtaláljuk azokat a kifejezéseket, amelyek leginkább utalnak pl. a konzervatív versus liberális álláspontra, közelebb kerülünk a polarizáció megértéséhez, és eredményeinket a tudományos diskurzushoz tudjuk kapcsolni. Mindennek ellenére öt osztályozást alkalmazó tanulmányból három csak az osztályozási teljesítmény optimalizálására összpontosított, és nem tért ki arra, hogy mely nyelvi jellemzők játszottak szerepet az osztályozásban (pl. Hemphill et al., 2016; Wang et al., 2017; Baly et al., 2020).

Voltak azonban olyan cikkek is, amelyek részletesen végigvették az értelmezést. Diermeier et al. (2012) az amerikai kongresszusban vizsgálták a legjellemzőbb nyelvi jellemzőket, olyan következtetésekkel pl., hogy a konzervatív és a liberális beszédek megkülönböztetésében a kulturális referenciák fontosabbak a gazdasági referenciáknál. Egy másik példa Gentzkow et al. (2019), akik szintén nagyon alaposan végigvették a pártos mondatok értelmezését.

Nem csak az osztályozás esetében merül fel az a kérdés, hogy mely szavaknak van ténylegesen hatása. Medzihorsky és munkatársai (2014) a Wordfish segítségével követték az amerikai Republikánus Párt ideológiai eltolódását, és azonosították a leggyakrabban használt kifejezéseket, valamint azt, hogy azok milyen mértékben diszkriminálnak az eltolódás dimenziójában. Rashed et al. (2020) klaszterelemzéssel értelmezte a klaszterek közötti szemantikai különbségeket a legrelevánsabb kifejezések alapján. Rumshisky et al. (2017) részletesen értelmezte azokat a fontos, "sodródó" szavakat, amelyek (a szóbeágyazás szerint) a vizsgált időszakban a leginkább változtatták jelentésüket.

Az osztályozási modellek hatékonyságát és értelmezhetőségét Praet és munkatársai (2021) vizsgálták szisztematikusan, akik három, feature-készletük tekintetében különböző modellt határoztak meg: egy szakértői alapon definiált, CAP-alapú megközelítést, egy adatvezérelt, a bag of words módszerre épülő

megközelítést és egy másik adatvezérelt, topikmodellen alapuló megközelítést. A három módszer összehasonlítása egyértelmű kompromisszumot mutatott az értelmezhetőség és a diszkriminatív képesség között. A szakértői modell mutatta a legrosszabb előrejelzést ugyanakkor a legjobb értelmezhetőséget, és fordítva: a bag of words modell volt a leghatékonyabb, de legrosszabbul interpretálható. A három módszer kombinációja adta a legjobb megoldást.

Az osztályozási modellek értelmezhetőségével kapcsolatban érdemes megemlíteni Goet (2017) tanulmányát,melynek fő tanulsága szerint az osztályozók figyelmen kívül hagyják a dimenzionalitást (ellentétben a skálázási módszerekkel), és amikor ilyen modelleket használunk, feláldozzuk annak lehetőségét, hogy érdemi állításokat tegyünk a polarizáció mozgatórugóiról. Láttuk azonban, hogy a legjellemzőbb szavak értelmezése ad némi választ erre a kérdésre, ha nem is olyan explicit módon, mint a skálázás esetében.

Az osztályozási modellek értelmezésének egy másik, izgalmas módját például Taddy (2013), Diermeier et al. (2012), Bayram et al. 2019, Rashed et al. (2020) és Gerrish and Blei (2012) használták. Az ő megközelítésük az outlier ill. tévesen besorolt esetek felderítésére és kvalitatív vizsgálatára épült. Például Gerrish és Blei (2012), amikor az amerikai törvényjavaslatok szövegét használták a törvényhozók szavazatának előrejelzésére, és az outlier-ként azonosított képviselők esetén megállapították, hogy azért lehet rosszul modellezni őket, mert bizonyos kérdésekben (pl. külpolitika) eltérnek a bal-jobb spektrumtól. Vagyis: mivel egyes törvényhozók bizonyos kérdésekben eltérnek a pártjuktól, az ezekben a kérdésekben elfoglalt álláspontjukat az egyetlen dimenzió mentén rangsoroló modellek nem tudják megragadni.

6.6.10. Kvalitatív módszerek alkalmazása

Az áttekintett tanulmányok többsége nem alkalmazott kvalitatív módszereket. Azok közül, amelyek igen, néhányan a *validálási* szakaszban alkalmazták azokat. A topikmodellek validálása például jelentős kvalitatív munkát igényel, mivel a kutatók csak úgy tudják megítélni a modellek értelmezhetőségét és hatékonyságát, ha ténylegesen elolvassák az egyes témákat reprezentáló legrelevánsabb szövegeket (pl. Farrell et al., 2016; és Guntuku et al., 2021). Yarchi et al. (2021) új módszert mutat be, amely az NLP-t és a kvalitatív tartalomelemzést kombinálja a topikok megértéséhez. Taddy (2012) ellenpélda, itt a modell teljesen elszakad a szövegtől, az elemzés nem tér vissza a szöveghez a topikok értelmezése során.

Más tanulmányok az *értelmezés* támogatására alkalmaztak kvalitatív megközelítést. Rho et al. (2018) például diskurzuselemzést használt az összes olyan hozzászólás elemzésére, amely tartalmazza az előzetesen felderített legfontosabb kifejezéseket. Hasonlóképpen Grover et al. (2019b) a LIWC-elemzés által fontosnak talált kifejezéseket tartalmazó tweetek kvalitatív elemzését végezte el.

A szoros olvasást (*close reading*), azaz a szövegek átgondolt értelmezését például Bonikowski et al. (2019) és Sinno et al. (2021) expliciten említették, mint alkalmazott módszert. Az utóbbi munkában a szoros olvasást arra használták, hogy megértsék az annotátorok döntései mögött álló motivációkat.

Dornschneider és Todd (2020) azon kevés tanulmányok közé tartozott, amelyben a kvalitatív megközelítés dominált. Egytucat interjú segítségével vizsgálták az unionisták és nacionalisták mindennapi hangulatát egy északír városban. A gépi szentimentelemzés után kvalitatív diskurzuselemzést végeztek. Budak et al. (2016) munkájában a gépi tanulást csak technikai okokból használták a releváns dokumentumok azonosítására, a kutatás tényleges módszertanát a cikkek szoros olvasása adta.

6.7. Következtetések

Áttekintésemnek megvannak a maga korlátai. A "politikai polarizáció" sokféle definíciót tartalmaz, és a "természetes nyelvi feldolgozás" fogalma is tudományáganként eltérő. Bár a keresési kulcsszavak definiálásánál megpróbáltam ezeket a fogalmakat többféleképpen megragadni, mégis előfordulhat, hogy kimaradt néhány keresőszó, így néhány releváns írás is.

A kezdeti 3084 találatból 154 releváns írást azonosítottam. A tanulmányok száma az elmúlt évtizedben emelkedett. A legtöbb tanulmány az Egyesült Államokra összpontosított (n=91), eredményeik nemzetközi érvényességét ritkán vizsgálták. Körülbelül 40%-ban (n = 66) használtak Twitter-adatokat, és minden harmadik tanulmány felügyelt gépi tanulást alkalmazott az ideológia/állásfoglalás előrejelzésére.

Bár a tanulmányok száma az elmúlt években gyorsan nőtt, csak kisebbségük használt fel szakterület ismereteket a mélyebb megértéshez. Azok, amelyek viszont így tettek, azt mutatták, hogy az elemzés több pontján is (a legfontosabb feature-ök értelmezése, a kiugró értékek felismerése, a különböző modellek összehasonlítása stb.) szakértői értelmezésre van szükség.

A legtöbb tanulmány nem alkalmazta a szoros olvasás módszerét, és nem tárgyalta a szöveges adatokból történő oksági következtetés esetén felmerülő lehetséges problémákat (utóbbival kapcsolatban lásd bővebben a 7. fejezetet). A tanulmányok nagy aránya nem volt interdiszciplináris abban az értelemben, hogy az érintett szakterület bevonása nélkül készültek (45%), vagy éppen ellenkezőleg, csak társadalomtudósok írták őket (20%). Ezek a megfigyelések a számításintenzív kutatási módszerek társadalomtudományokon kívüli intézményesülésének következményei lehetnek.

Ugyanakkor számos inspiráló példát találtam olyan módszertani megközelítésekre, amelyek nem mutatják ezeket a hiányosságokat. Találtam példát többféle típusú adatbázis kombinált használatára: olyanokat, amelyek szövegen kívüli információkat is használnak (például közösségi média felhasználóinak kapcsolódásait), olyanokat, amelyek a szöveg metaadatait is használják, és olyanokat, amelyek a közvélemény-kutatási adatokat a közösségi média adataival kombinálják. Láttuk, hogy a kutatások a politikai nyilvánosság különböző rétegeire (politikusok, szakértők, a média vagy a laikus közönség) terjednek ki. Ugyanakkor nagyon kevés tanulmány foglalkozott a nyilvánosság egynél több rétegével, pedig a megközelítés a rétegek közötti diffúziós folyamatok tanulmányozására is alkalmas lenne.

Azt gondolom, áttekintésem alapján megállapítható, hogy az NLP-ben rejlő potenciál a politikai polarizáció kutatásában is nagy. Az áttekintés azonban általános társadalomkutatási következtetésekhez is elvezetett, elsősorban azzal, hogy érveket szolgáltat a magyarázó és prediktív modellezési paradigmák integrálása, valamint az NLP-alapú társadalomkutatás interdiszciplinárisabb megközelítése mellett is.

7. OKSÁGI KÖVETKEZTETÉS AZ NLP-ELEMZÉS SORÁN

7.1. Oksági következtetések a politikai polarizáció kutatásában: hamis korreláció, *confounder*-ek felügyelt gépi tanulás esetén

Az empirikus társadalomkutatás által alkalmazott oksági megközelítéseknek több típusa különíthető el (Németh, 2015b, 2021), itt a leggyakrabban használt megközelítést hivatkozzuk csak, a robusztus összefüggésekre alapozó, a potenciális zavaró tényezőket (*confounders*) statisztikai módszerekkel kiszűrő megközelítést (angolul még: *partialling approach*). Kitekintésként említem, hogy a kortárs tudományos megközelítések fontos alesetét, a Judea Pearl-féle oksági elemzést (Pearl, 2016) tárgyalja Molnár és társa (2024) könyvének *Causality* c. fejezete, a felügyelt tanulásra fókuszálva, de nem szöveges adatokon, hanem általános adatbázisokon.

A megközelítés kiindulópontja az az állítás, hogy a megfigyelt együttjárás nem feltétlenül jelent ok-okozati összefüggést, de az ok-okozati összefüggés valamilyen módon mindig feltételezi az együttjárást. Így a megfigyelt együttjárás bizonyos esetekben oksági összefüggést mutathat, a kérdés az, hogy az X és az azt időben követő Y változók közötti megfigyelt asszociáció mikor és milyen mértékben feleltethető meg az X-nek az Y-ra gyakorolt oksági hatásának. E probléma megoldására a megközelítés az asszociáció robusztusságának vizsgálatát javasolja: az asszociáció akkor oksági, ha új változók bevonása után (szakkifejezéssel azok "kontrollálásával") is megmarad. A kontrollálással figyelembe vett asszociáció az úgynevezett parciális asszociáció.

Fontos megjegyezni, hogy a megközelítés szociológiában megfigyelt elterjedtségének oka az, hogy a társadalomtudományok - a természettudományokkal szemben – jellemzően nem tudnak kísérletet tervezni adott probléma vizsgálatára (mert az kivitelezhetetlen vagy etikátlan), ezért megfigyeléses adatokra (*observational data*) kell, hogy támaszkodjanak. Vagyis a kezelt/kontroll szereposztás nem a kutató által végrehajtott randomizált allokáció eredménye, hanem vagy a ‘természet’ sorolja be az egyéneket e kategóriákba, vagy azok sorolják be magukat saját választásaik révén.

Ez az empirikus korlát, a megfigyeléses adatok jellemző jelenléte az alapja az alábbi, NLP-elemzésekben talált oksági következtetési problémáknak is. Lássunk erre, a kísérlet vs. megfigyelés eltérésre egy példát. A *Felügyelt gépi tanulás* c. fejezetben ismertetett kutatásunkban (Buda et al., 2022) randomizált,

kontrollált terepkísérletet végeztünk: ugyanolyan emaileket küldtünk ki önkormányzati hivataloknak, egy részüket sztereotipikus roma hangzású, másik részüket nem-roma hangzású nevekkel aláírva, e két csoportot random módon osztva szét a hivatalok között. Itt, amikor eltérést találtunk a két csoportnak írt válaszok között, biztosak lehettünk benne, hogy az eltérést az eltérő név (azaz az ügyfél vélt etnikai hovatartozása) okozza. Ugyanakkor, ha élő emailezésekhez hozzáférve, valódi hivatali emaileket és a rájuk kapott válaszokat vizsgáltunk volna (feltéve, hogy a nevek alapján ott is vélt roma és nem-roma csoportba tudjuk sorolni az ügyfeleket), nem lett volna egyértelmű az oksági következtetés. Hiszen számos potenciális zavaró tényező léphet fel, például a hazai roma lakosság iskolázottsági és egyéb státusz-mutatói átlagosan gyengébbek, mint a többségi társadalom mutatói. Mivel az iskolázottság, a társadalmi státusz egy email szövegéből elég jól megítélhető, a megfigyelt eltérés a vélt roma és nem roma ügyfelek között (részben) erre is visszavezethető lehet. Itt is a korábban már idézett Bartos et al (2016) által definiált figyelemdiszkrimináció feltevése lenne plauzibilis: a hivatalnok eltérő mértékű figyelmet szentelhet a magasabb ill. alacsonyabb státuszú ügyfélnek. Tehát a vélt etnikai hovatartozás és a diszkrimináció közötti kapcsolat ilyen megfigyeléses vizsgálatnál kevésbé támasztható alá, ill. az oksági bizonyíték erősebb lenne, ha az iskolázottságra kontrollálna az elemzés.

Bár az ok-okozati összefüggések problémáját az empirikus társadalomtudományokban széles körben tanulmányozták, az NLP-alapú kutatásokban, így speciálisan a politikai polarizáció kutatásában használt szövegosztályozásban is gyakran elhanyagolják (Németh, 2023), pedig a hibás ok-okozati következtetés nemcsak az értelmezését vagy a modellek robusztusságát érinti, de azok validitását is veszélyezteti. Ebben a fejezetben e 2023-ban publikált kutatásomban végzett irodalmi áttekintés tanulságaira támaszkodom majd.

A legkézenfekvőbb oksági megközelítés az időbeli megelőzésre való hivatkozás: Jensen és munkatársai (2012) például azt találták, hogy a politikailag polarizáló kifejezések gyakorisága a Google Books könyv-adatbázisban már azelőtt megnő, mielőtt használatuk a kongresszusi beszédekben is nőne. Bár hangsúlyozták, hogy az oksági következtetés meghaladja tanulmányuk kereteit, azt sugallják, hogy az időbeli megelőzés oksági kapcsolatra utalhat, és hogy eredményük összhangban van az elit diskurzus kongresszusi beszédre gyakorolt autonóm hatásával.

A politikai polarizáció tanulmányozása során potenciálisan szintén felmerülő oksági probléma elsősorban a zavaró tényezők kérdése, vagyis az, hogy

valóban ideológiai különbségek okozzák-e a nyelvhasználatban észlelt különbségeket, azok nem vezethetők-e vissza más különbségekre. Nagyon kevés elérhető tanulmány említi expliciten ezt a kérdést, még kevesebb próbál megoldást találni rá. Egy ellenpélda Lin és munkatársai (2006) kutatása, akik a bitterlemons.org honlapot vizsgálták, amely hetente jelentetett meg cikkeket az izraeli-palesztin konfliktussal kapcsolatos kérdésekről, mindig egy izraeli és egy palesztin szerkesztővel és vendéggel. A szerzők felügyelt osztályozást alkalmaztak a nézőpont (izraeli vagy palesztin) előrejelzésére. Mivel a nézőpont jobban előre jelezhető volt a szerkesztők, mint vendégek esetében, Lin és társai azt feltételezték, hogy az izraeli és a palesztin szerkesztők írói stílusa között lehetnek különbségek, és a modellek ezt találták meg, nem pedig a politikai nézőpontot. Annak tesztelésére, hogy a statisztikai modellek valóban a nézőpontot tanulják-e, olyan kísérleteket végeztek, amelyekben algoritmusukat szerkesztőkön tanították, majd vendégeken tesztelték, és fordítva. Az ily módon tesztelt modellek változatlanul magas predikciós pontossága arra utal, hogy a modellek ugyanazon jellemzők szerint különböztetik meg a szerkesztőket és a vendégeket, vagyis a modellek az eredeti értelemzés szerint működtek, nézőpontot tanultak, és nem szerkesztői-írói stílusokat.

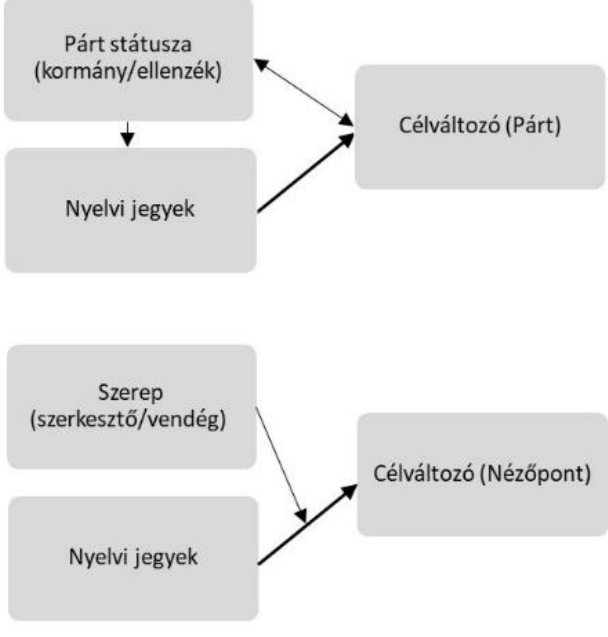
Az NLP a politikai polarizáció kutatásában c. fejezetben az kutatásokhoz felhasznált adatokból levonható tanulságok kapcsán kitértem a transzferábilitás (vagy domain-közi általánosíthatóság) problémájára, vagyis arra, amikor az egyik típusú (adott műfajú, adott időszakban született stb.) osztályozó modellek nem vihetők át, nem mutatnak jó teljesítményt másik típusú adathalmazon. Lin és társai (2006) fent éppen ehhez kapcsolódva végeztek vizsgálatot, amikor a szerkesztőkön tanított modelleket vendégeken tesztelték. Ide tartozik, és szintén a kihagyott potenciális *confounder* meglétére mutat rá Potthast et al. (2018) fontos konklúziója is, miszerint a híroldalak nyelvezete valójában nem annyira a jobb- és baloldali ideológia mentén oszlik meg, hanem inkább a mainstream/szélsőséges dichotómia, azaz tk. stilisztikai vonalak (!) mentén.

Hirst és munkatársai (2010, 2014) szintén gyenge transzferábilitást észleltek két kanadai parlamenti cikluson definiált osztályozásra nézve (az első esetben, a 36. cikluban a Liberális Párt, a másodikként vizsgált 39. ciklusban a Konzervatív Párt kormányzott). A parlamenti felszólalásokat párthovatartozás szerint osztályozták, bináris módon, csak e két pártra koncentrálva. Kimutatták, hogy amikre modelljeik valójában rátanultak, az nem az ideológia, hanem az ellenzéki ill. kormánypozícióban használt támadás és védekezés kifejezései voltak. Vagyis eredményük szerint a pártok osztályozásakor a pártstátusz (ellenzéki /

kormányzati) figyelembe veendő zavaró tényező. Az általuk használt módszerekre a következő alfejezetben térek ki.

Lin et al. (2006) és Hirst et al. (2010, 2014) egyaránt egy harmadik változóra gyanakodtak, amely mind a függő változót, mind a független változót befolyásolta, hamis összefüggést (*spurious correlation*-t) okozva, de gyanújuk csak az utóbbi esetben igazolódott be. Az előbbi esetben hatásmódosulást tártak fel: csak a modellek teljesítményszintje különbözött a szerkesztő/vendég szerepkör (a moderátor változó) függvényében, de ugyanazok a nyelvi jellemzők különböztették meg a nézőpontokat mindkét szerepkörön belül.

A harmadik változó problémáját ritkán említik az NLP-t alkalmazó tanulmányokban, ezért érdemes egy kicsit részletesebben foglalkozni vele. A zavaró tényezők és a hatásmódosítás logikáját az alábbi 23. ábra szemlélteti. Egy hasonló oksági grafikon megrajzolása és az összefüggések tesztelése (ahogyan az a hagyományos statisztikai megközelítésben szokásos) gyümölcsöző lenne szöveges vizsgálatokban is.



23. ábra. Felügyelt gépi zavaró tényező / moderátor jelenlétében. A moderátorok az ok-okozati úton helyezkednek el (vastag vonal), míg a zavaró tényezők nem.

Az oksági következtetésnek a robusztus okságra hivatkozó megközelítésen kívüli megoldásait kevés tanulmány említi a polarizációt NLP-vel megközelítő irodalmában, a kevés kivétel egyike Landeiro és Culotta (2018) is, akik részletesen tárgyalják, hogyan lehet a szövegosztályozással összefüggésben felismerni és kezelni a zavaró hatásokat, a nyelvi polarizáció mint esettanulmány kapcsán. Ők Judea Pearl (2016) statisztikai keretrendszerét használták az okság megközelítésére. Widmer és munkatársai (2020) ok-okozati bizonyítékot adtak vonatkozóan, hogy a televízió polarizált hírközlése befolyást gyakorol a helyi újságok által közzétett tartalmakra is, ők az oksági hatás megközelítésére az inkább közgazdaságtanban ismert instrumentális-változó keretrendszert alkalmazták.

7.2. A confounder-re kontrollálás módja felügyelt tanulásnál

Hogyan hajtható végre a potenciális *confounder*-re történő kontrollálás? A klasszikus statisztikában kategoriális változók kereszttáblája esetén (a Lazarsfeld-féle elaborációs technikaként ismert kutatási programban) a megoldás lényege, hogy X és Y kategoriális változók kereszttáblájában megfigyeljük az (un. nulladrendű) asszociációt, majd egy harmadik, Z változó bevonásával (a Z rögzített értéke mellett kapott feltételes kereszttáblákon) a parciális asszociációt. X és Y kapcsolatát aszerint értelmezzük, hogy a parciális kapcsolat a nulladrendűhöz képest hogyan változott (erősödött, csökkent, vagy eltűnt), illetve hogy a Z változó időben a másik kettőhöz képest hol helyezkedik el (közöttük vagy előttük). A módszer általánosított, folytonos változókra is értelmezett megfelelője a regresszió- és az útelemzés, melyek Z-re kontrollált direkt és a Z-n keresztül megvalósuló indirekt hatások megkülönböztetésére képesek.

A potenciális zavaró tényezőre történő kontrollálás felügyelt tanulás esetén legegyszerűbb megközelítésben a **rétegzett elemzés (1)** lehetne, ahol a zavaró tényező kategóriái szerinti rétegeket képez a kutató, és rétegenként külön illeszt tanuló modelleket. Lényegében ezt végezte el Jelveh et al. (2014), amikor felismerték, hogy ha a közgazdászok politikai-ideológiai álláspontját írásaik nyelvezete alapján szeretnék prediktálni, akkor az írás témája egy potenciális zavaró tényező. Valóban, az ideológiának közvetlen és közvetett hatása is van a nyelvezetre: és az előbbi az, amit a prediktáláskor szeretnénk megragadni. A közvetett hatás a témán keresztül hat, hiszen az ideológia befolyásolhatja a közgazdász kutatási területét. Pl. ha a konzervatívabb közgazdászok a makroökonómiát kutatják inkább, akkor

a makroökonómiával kapcsolatos kifejezések támogatnák a konzervatív ideológia prediktálhatóságát, így hamis korrelációt teremtve nyelvezet és ideológia között. Ezért Jelveh és társai a korpusz írásait előbb topikmodell segítségével témákhoz sorolták, majd tanuló modellek feladataként a közgazdászok ideológiájának predikcióját az egyes témákon belüli szóválasztásuk alapján definiálták.

Fent láttuk, hogy egy másik lehetséges megoldás a hamis korreláció kiszűrésére **a modell más korpuszra történő transzferálása (2)**. Hirst és társai (2010, 2014) is ezt a módszert alkalmazva fogtak gyanút azzal kapcsolatban, hogy modelljük nem a politikai ideológiát jelzi előre. Azt találták, hogy ha a korábbi parlamenti cikluson tanított modellt tesztelik a későbbi cikluson, akkor a modell korábban jó prediktív teljesítménye teljesen leromlik. Kutatótársainkkal végzett munkánkban (Buda, Németh, Rakovics, 2023) hasonló módon kíséreljük meg a magyar Parlament felszólalóinak ideológiai klasszifikációját a státusz, mint potenciális *confounder* (kormányon/ellenzékben van a párt) hatásától megtisztítani.

Egy harmadik, a fentit kiegészítő megközelítésmód is Hirst és társai tanulmányában található. A parlamenti felszólalók párthovatartozását prediktáló modelljük interpretációjára térve azt találták, hogy a korábbi, Liberális Párt kormányozta ciklusban illesztett modellben **a predikció szempontjából fontosnak talált szövegjellemzők (3)** egy része a későbbi, Konzervatív Párt által kormányozott ciklusra ‘oldalt cserélt’: ami korábban a liberálisokra volt jellemző, az a kormányváltás után konzervatív jegy lett. Ez alapján kezdtek arra gyanakodni, hogy modellük valójában nem politikai ideológiát, hanem kormányzati/ellenzéki státuszt tanul. Vegyük észre, hogy itt a modell valódi működése az értelmezésből derült ki. A modell interpretációja nélkül annak létrehozói tévesen ítélték volna meg működését. Vagyis: az algoritmus mást is tanulhat, mint amit gondolunk, hogy megtanul.

Hirst és társai egy további szellemes megoldást találtak azon gyanújuk igazolására, hogy a modell valójában érzékenyebb a státuszra, mint az ideológiára: a pártok közötti ideológiai kontrasztot azzal csökkentették, hogy az elemzésbe bevonták a baloldali pártokat is, amelyek mindkét parlamenti ciklusban ellenzékben voltak. Ha az osztályozás valóban ideológiai lenne, akkor e pártok összevonása a többi konzervatív (36. parlament) vagy liberális (39. parlament) ellenzéki párttal jelentősen rontaná a teljesítményt. Ugyanakkor, ha a pártok státusza számít, akkor ennek a lépésnek kevés hatása lesz, mivel az ellenzéki pártok többé-kevésbé megkülönböztethetetlenek. Vagyis **az oknak tételezett változó szerint csökkentették a vizsgált korpuszban a kontrasztot, míg a potenciális *confounder* szerint nem**

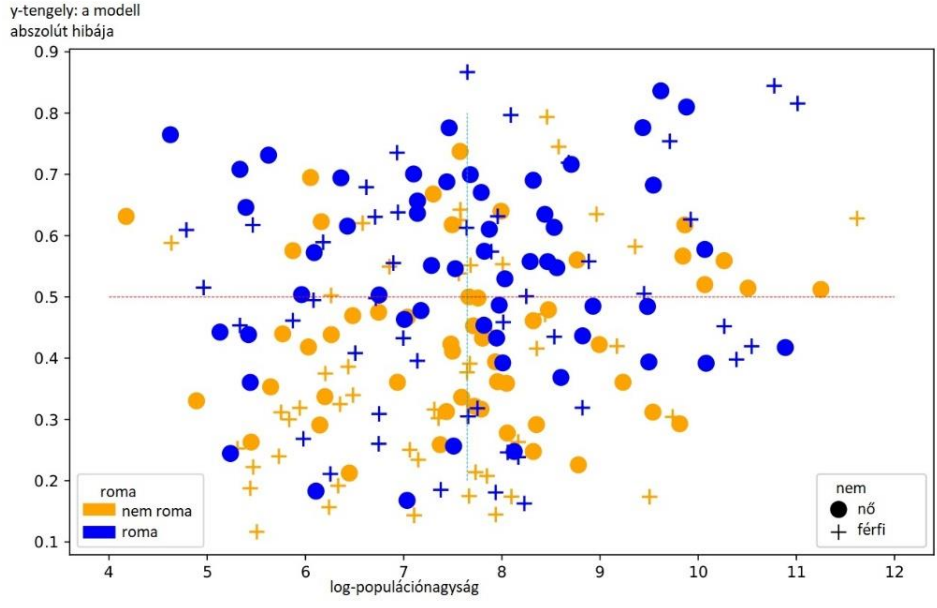
(4). Az eredmény szerint a tanuló modell teljesítménye csak kis mértékben csökkent, vagyis az utóbbi szcenáriót támasztották alá az adatok.

Végül munkatársaimmal végzett, e fejezetben is idézett kutatásunk (Buda et al., 2022) megoldását ismertetném, ahol, mivel randomizált kontrollált terepkísérletet végeztünk, a roma etnikum és a diszkrimináció közötti oksági kapcsolat esetén a hamis asszociáció fellépésének nem volt ugyan esélye, de hatásmódosító tényezőként több metaváltozó megjelenhetett. Megvizsgáltuk tehát a modellek predikciós teljesítményét ezen metaváltozók (az ügyfél neme és az önkormányzat településének nagysága) szerint. Először rétegzett elemzést végeztünk, a települések méret szerinti kettéosztásával, a medián által meghatározott vágási ponttal, illetve nemek szerint bontva a korpuszt. Mindkét modellünk jobban teljesített a férfiak körében, mint a nőknél, ugyanígy a kisebb településeken, mint a nagyobbakon. Ez azt jelenti, hogy a tanuló modellek által megállapított eltérő bánásmód erősebben érvényesül a férfiakkal szemben és a kisebb településeken.

A rétegzett elemzésen túl a modellek **individuális szintű predikciós hibáját is vizsgáltuk a hatásmódosító tényező függvényében (5)**. Lásd a 24. ábrát: itt a népességszám (az egyszerűbb ábrázolás érdekében annak logaritmusa) függvényében ábrázoltuk a deskriptív szövegjellemzőket használó modell abszolút hibáját. A modellek jobb teljesítménye a kisebb településeken és a férfiak körében itt is nyilvánvalóan megmutatkozik (lásd pl. az ábra bal alsó negyedében, tehát a kisebb hibájú predikciót mutató és kis településen élők között a férfiak számbeli fölényét).

Ez az eredményünk akkor is megerősítést nyert, amikor **a két tanuló modellt a teszt-korpuszon adott, individuális szintű, konzisztensen helytelen ill. konzisztensen helyes előrejelzéseik alapján vizsgáltuk (6)**. Azt találtuk, hogy azok az ügyfelek körében, akik esetén mindkét modell hibás előrejelzést adott (vagyis ahol a megkülönböztetés nem volt jelen, vagy legalábbis nagyon finom volt), jellemzően magasabb volt mind a nők, mind a nagyobb települések aránya is. És fordítva: ahol mindkét modell helyes előrejelzést adott (vagyis ahol a megkülönböztetett bánásmód a legkevésbé finom), ott a nők aránya viszonylag alacsony, ahogy a nagy települések aránya is.

A fentiekben hat, egymástól többé-kevésbé független módszert is felsoroltam a robusztus asszociáció mentén történő oksági következtetés támogatására. Ezek egymást kiegészítő, egymás statisztikai bizonyító erejét megerősítő módszerek arra, hogy eldöntsük, a felügyelt gépi tanulási modellek által detektált együttjárás valóban oksági kapcsolat is-e egyúttal.



24. ábra. A Buda et al. (2022) cikkünkben ismertetett, leíró szövegstatisztikákon alapuló modell abszolút hibája a település lakosságszámának logaritmusa függvényében. A pont alakja az ügyfél nemét, színe etnikai hovatartozását jelöli.

Összefoglalóan: a hamis korreláció problémája a politikai polarizáció NLP-alapú megközelítésében is előkerül néhány tanulmányban, azonban a tanulmányok zöme nem említi és nem is vizsgálja ezt a lehetséges problémát. Pedig a kevés, a problémát explikáló tanulmány kimutatta, hogy a politikai szövegek jellemzői nemcsak a szerzők politikai pozíciójától függenek, hanem más, gyakran figyelmen kívül hagyott, az előbbitől nem független tényezőktől is (például a szerzők politikai elkötelezettségétől, attól, hogy pártjuk kormányzó vagy ellenzéki pozícióban van-e, a szövegek stílusától, műfajától vagy keletkezésének idejétől). És általában, nem csak a polarizáció kutatásában is: ha a potenciális zavaró és hatásmódosító tényezőket figyelmen kívül hagyjuk, és a szövegeket kizárólag egy tényező alapján igyekszünk klasszifikálni, az súlyos hibákhoz vezethet, például túlságosan optimista predikciós teljesítményekhez vagy hamis asszociációkhoz. Mindezt nem csak az elemzés, hanem már az adatgyűjtés során érdemes szem előtt tartani, és minél heterogénebb, a lehetséges zavaró tényezők, hatásmódosítók szerint kiegyensúlyozott korpuszból kiindulni.

8. Összegzés

A hatalmas mennyiségű digitális szöveges adat elérhetősége és új elemzési potenciálja széles perspektívát jelent a szociológia számára. A számítógépes társadalomtudomány, jelen példákon a szövegbányászat várhatóan akkor lesz beépíthető a mindennapi kutatásba, ha az interdiszciplináris együttműködések széles körben elterjednek, ha a szükséges tudás és kompetencia beépül az egyetemi képzésbe. A fejlett programozási ismereteket nem igénylő, alacsony küszöbű szövegelemző platformok ugyanakkor már most megjelentek (mint például a régebb óta ismert Google Trends, Google Ngram Viewer, az európai fejlesztésű Sketch Engine, vagy az újabb Lancsbox, a Lancester University fejlesztése stb., - az utóbbi két-három évben gombamód szaporodtak a szövegelemzésben alkalmazható „no-code" eszközök), támogatva az átalakulást – ezek további fejlődése valószínűsíthető a közeljövőben.

Láttuk, hogy a szövegeken alapuló gépi tanulás is új lehetőségeket kínál a szociológia számára. A gépi tanulási megoldások egyre nagyobb része humán annotáción alapul, és az ilyen ember-gép együttműködésen alapuló számítások még erősebb felfutása várható nem csak az iparban, de a tudományban is. Az előre annotált és nyilvánosan elérhető adatbázisok már most támogatják a felügyelt tanulás saját alkalmazását.

A szociológus ugyanakkor nem csak felhasználóként, hanem a kritikai nézőpont képviselőjeként, a felügyelt gépi tanulás társadalmi hatásaira, etikai problémáira rámutató aktorként is jelen kell, hogy legyen. A crowdsourcing annotálás komoly munkaerőpiaci problémákat generál, az MI-torzítás pedig társadalmi hátrányokat erősíthet fel. Ezen túl a kötet keretein túlmutató, de zárásként mindenképp megemlítendő kérdés az internetes privacy és szabadság problémája. Kérdés, milyen hatása lesz ezekre a nagytömegű szövegek gyors feldolgozására képes technológia, nem válik-e a szövegbányászati technológia a cenzúra és megfigyelés mindenható eszközévé?

Azt remélem, esettanulmányaim bizonyították, hogy az NLP egyedülálló lehetőséget ígér a szövegekben található regularitások felfedezésére, melyek magyarázatként szolgálhatnak az elméletalkotásban is. Ugyanakkor, ahogy példáim mutatták, az NLP módszertani buktatókat is rejt. Ennek elsődleges oka az, hogy az adatok rendkívüli mennyisége és az alkalmazott módszerek összetettsége a megbízhatóság hamis érzetét keltheti. Mivel a társadalomkutatás más adatforrásaihoz, pl. a survey-hez hasonlóan az adatok itt is megfigyelésre (és nem

randomizált kísérletre) épülnek, az NLP-kutatások esetén ugyanúgy fennáll a reprezentativitás hiányának vagy az összemosódásnak (confounding) a veszélye. Az egyetlen adatbázisból vagy egyetlen internetes platformról származó adatokra épülő kutatások az általánosíthatóság kérdését vetik fel.

Azt remélem, az is kiderült írásomból, hogy az automatizált módszerek csupán kiegészítik a kutatót, és semmiképpen nem pedig helyettesítik. Továbbra is a kutatónak kell irányítania az elemzési folyamatot, számos adatgyűjtési, modellezési döntést hozva és végül interpretálva a modell eredményét. E folyamat több pontján szükség van kvalitatív jellegű, a szövegek alapos olvasását feltételező megközelítésre. Láttuk, hogy az NLP-modellek eredménye félrevezető vagy egyszerűen téves lehet, ezért megkerülhetetlen kutatói feladat az eredmények validálása is. Ez jellemzően tartalmi értékelést, így szakterületi tudást is igényel.

E megállapítások tágabb összefüggésben is vizsgálhatók. Az elmúlt években több olyan érv merült fel, mely a gépi tanulást vagy általában a mesterséges intelligenciát adaptáló területek reprodukálhatósági válságára mutattak (pl. Driggs et al., 2021). A válság okait is felmutató tanulmányok között jó néhány olyan volt (pl. Hofman et al., 2021; Hullman et al., 2022), amely azzal érvelt, hogy fel kell ismernünk: az informatika és az azt alkalmazó tudományok találkozása valójában több, mint a nagy adattárak és az azok elemzéséhez szükséges eszközök adaptálása. Ez a találkozás a különböző módszertani paradigmákkal rendelkező területek konvergenciáját is jelenti (lásd pl. a magyarázó vs. prediktív iskolákat), és a kutatás minősége attól is függ, hogy ezeket a paradigmákat hogyan integrálják produktívan.

Ugyanis az NLP-nek, és általában az adattudománynak van egy általános gyenge pontja, nevezetesen, hogy szakterületi ismeretek nélkül csak egy technikát képvisel, amely nem sok tudományos értéket nyújt a társadalomtudományok számára. Írásom számos érvet szolgáltatott a nagyobb interdiszciplinaritás felé való elmozdulás mellett. Azt gondolom, a szociológia akkor fogja kiaknázni a digitális forradalom nyújtotta lehetőségeket, ha képes megújítani módszertanát, miközben megőrzi kritikai reflexióját. Ezért volt célom e könyv megírása, mely reményeim szerint megmutatta, hogy az NLP hogyan illeszthető be szerves módon a hagyományos szociológiai módszerek eszköztárába. Ez lehet a kulcs ahhoz, hogy jobban megértsük a mai társadalmunkban végbemenő mélyreható változásokat.

9. IRODALOM

Acree, B. (2016). *Deep learning and ideological rhetoric*. The University of North Carolina at Chapel Hill University Libraries.

Acree, B. D. L., Gross, J. H., Smith, N. A., Sim, Y., & Boydstun, A. E. (2020). Etch-a-sketching: Evaluating the post-primary rhetorical moderation hypothesis. *American Politics Research*, *48*(1), 99–131.

Ademmer, E., & Stöhr, T. (2019). The making of a new cleavage? Evidence from social media debates about migration. *Kiel Working Papers*, 2140.

Angelova, G. (2010). Use of domain knowledge in the automatic extraction of structured representations from patient-related texts, In M. Croitoru, S. Ferré, D. Lukose (Eds.), *Conceptual Structures: From Information to Intelligence – Proceedings of the 18th International Conference on Conceptual Structures, ICCS 2010, Kuching, Sarawak, Malaysia, July 26-30, 2010*. (pp. 14–27). Springer.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32.

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15–24.

Ash, E., Chen, D. L., & Lu, W. (2017). Polarization of U.s. circuit court judges: A machine learning approach. *SSRN Electronic Journal*.

Austin, J. L. (1975). Lecture VIII. In *How To Do Things With Words* (pp. 94–108). Oxford University Press.

Bales, R. F. (1950). A set of categories for the analysis of small group interaction. *American Sociological Review*, *15*(2), 257 – 263.

Baly, R., Martino, G. D. S., Glass, J., & Nakov, P. (2020). *We can detect your bias: Predicting the political ideology of news articles*.

Barna, I., & Knap, Á. (2023). 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*, *51*(3), 603–621.

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, *106*(6), 1437–1475.

Bauer, M. W., Bicquelet, A. & Suerdem, A. K. (2014). Text analysis: an introductory manifesto. In Martin W., Bicquelet, Aude & Ahmet K. Suerdem (Eds.), *Textual Analysis. SAGE Benchmarks in Social Research Methods.* Sage Publications.

Bayram, U., Pestian, J., Santel, D. & Minai, A.A. (2019, July 14-19). *What's in a word? Detecting partisan affiliation from word use in congressional speeches.* [Paper presented at conference]. International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary.

Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020). Learning political polarization on social media using neural networks. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 47177–47187.

Benoit, K. (2020). *Text as data: An overview. The SAGE handbook of research methods in political science and international relations* (pp. 461-497).

Berelson, B. & Lazarsfeld, P. F. (1948). *The Analysis of Communication Content.* Universitetets Studentkontor.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In Srivastava, Sahami (Ed.), *Text mining* (pp. 101-124). Chapman and Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning research*, *3*(Jan), 993-1022.

Blevins, C. (2010). *Topic modeling Martha Ballard's diary*. https://www.cameronblevins.org/martha-ballards-diary/

Bolton, D., & Gillett, G. (2019). *The biopsychosocial model of health and disease: New philosophical and scientific developments* (p. 149). Springer Nature.

Bonikowski, B., Feinstein, Y. & Bock, S. (2019, April 12). *The Polarization of Nationalist Cleavages and the 2016 U.S. Presidential Election.* [Paper presented at conference]. UCR Political Economy Seminar.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215.

Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2022). User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, 1-20.

Brigadir, I., Greene, D. & Cunningham, P. (2015). Analyzing discourse communities with distributional semantic models. In *Proceedings of the ACM Web Science Conference, ACM, New York, NY, USA.*

Buda, J., Németh, R., Simonovits, B., & Simonovits, G. (2022). The language of discrimination: assessing attention discrimination by Hungarian local governments. *Language Resources and Evaluation*, 1-24.

Buda, J., Németh, R., & Rakovics, Zs. (2023). *Polarization as a Measure of Text Classification Performance - Evidence from the Hungarian Parliament 1998-2020.* [Unpublished manuscript]. Eötvös Loránd University.

Budak, C., Goel, S. & Rao, J.M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), 250–271.

Budhiraja, A. & Pal, J. (2020). Twitter and political culture: Short text embeddings as a window into political fragmentation. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies,* (pp. 335-336). Association for Computing Machinery.

Carius-Munz, L.M., 2020. Partisanship: conceptualizations and consequences, in: Oscarsson, H., Holmberg, S. (szerk.), *Research Handbook on Political Partisanship*. Edward Elgar Publishing, Cheltenham, pp. 47–59.

Chen, J., Hsieh, G., Mahmud, J.U. & Nichols, J. (2014). Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, New York, NY, USA.*

Cohen, R., & Ruths, D. (2013). Classifying Political Orientation on Twitter: It's Not Easy!. *Proceedings of the the 7th International AAAI Conference on Weblogs and Social Media (ICWSM-13), 7*(1), 91-99.

Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A. & Menczer, F. (2011). Predicting the political alignment of twitter users. *Proceedings of the 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing, 9-11 October, 2011, Boston, MA, USA.*

Cotelo, J.M., Cruz, F.L., Enríquez, F. & Troyano, J.A. (2016). Tweet categorization by combining content and structural knowledge. *Information Fusion, 31*, 54–64.

Coutto, T. (2020). Half-full or half-empty? Framing of UK–EU relations during the Brexit referendum campaign. *Journal of European Integration 42* (5), 695–713.

Csomor, G., Simonovits, B., & Németh, R. (2021). Hivatali diszkrimináció?: Egy online terepkísérlet eredményei= Discrimination at local governments? - Results of an online field experiment. *Szociológiai Szemle, 31*(1), 4–28.

Darwish, K. (2019). Quantifying Polarization on Twitter: The Kavanaugh Nomination. In *Social Informatics. SocInfo 2019. Lecture Notes in Computer Science*, 11864. Springer.

Darwish, K., Stefanov, P., Aupetit, M., & Nakov, P. (2020). Unsupervised User Stance Detection on Twitter. *Proceedings of the 14th International AAAI Conference on Web and Social Media, 14*(1), 141-152).

Dastin, J. (2018, October 11). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters.

Decadri, S. & Boussalis, C. (2020). Populism, party membership, and language complexity in the Italian chamber of deputies. *Journal of Elections, Public Opinion and Parties 30*(4), 484–503.

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J. & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*.

Deng, C., Ji, X., Rainey, C., Zhang, J. & Lu, W. (2020). Integrating machine learning with human knowledge. *iScience 23*(11), 101656.

Diaf, S., Döpke, J., Fritsche, U. & Rockenbach, I. (2022). Sharks and minnows in a shoal of words: Measuring latent ideological positions based on text mining techniques. *European Journal of Political Economy, 75(*102179).

Diermeier, D., Godbout, J.-F., Yu, B. & Kaufmann, S. (2012). Language and ideology in Congress. *British Journal of Political Science 42*(1), 31–55.

DiMaggio, P., Evans, J. & Bryson, B., 1996. Have American's social attitudes become more polarized? *American Journal of Sociology 102*(3), 690–755.

Doan, S., Conway, M., Phuong, T. M., & Ohno-Machado, L. (2014). Natural language processing in biomedicine: a unified system architecture overview. *Clinical Bioinformatics*, 275-294.

Dornschneider, S. & Todd, J. (2020). Everyday sentiment among unionists and nationalists in a Northern Irish town. *Irish Political Studies 36*(2), 185-213.

Driggs, D., Selby, I., Roberts, M., Gkrania-Klotsas, E., Rudd, J. H. F., Yang, G., Babar, J., Sala, E., Schönlieb, C.-B., & AIX-COVNET collaboration. (2021). Machine learning for COVID-19 diagnosis and prognostication: Lessons for amplifying the signal while reducing the noise. *Radiology. Artificial Intelligence, 3*(4), e210011.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology, 46*, 61-81.

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT Press.

Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology, 42*(1), 21–50.

Fang, A., Ounis, I., Habel, P., Macdonald, C. & Limsopatham, N. (2015). Topic-centric classification of twitter user's political orientation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 791-794). ACM Press.

Farkas, A. & Németh, R. (2020). *How to Measure Gender Bias in Machine Translation: Optimal Translators, Multiple Reference Point*. arXiv:2011.06445.

Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences of the United States of America 113*(1), 92–97.

Fiorina, M.P. & Abrams, S.J. (2008, June). Political polarization in the American public. *Annual Review of Political Science, 11*, 563–588.

Firth, J. (1957). A Synopsis of Linguistic Theory 1930-1955. In *Studies in Linguistic Analysis, Philological Society* (pp. 1–31). Oxford.

Fort, K., Adda, G. & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics 37*(2), 413–420.

Gadamer, H. G. (2004). *Truth and Method* (2nd ed.). Crossroad.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences. *Proceedings of the National Academy of Sciences of the United States of America, 115*(16), E3635–E3644.

Garimella, K., Morales, G.D.F., Gionis, A. & Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing 1*(1), 3.

Gauvin, J.P., Chhim, C., & Medeiros, M. (2016). Did they mind the gap? Voter/party ideological proximity between the BQ, the NDP and Quebec Voters, 2006–2011. *Can. J. Polit. Sci. 49*(2), 289-310.

Gelman, J. & Wilson, S. L. (2022). Measuring Congressional Partisanship and Its Consequences. *Legislative Studies Quarterly 47*(1), 225-256.

Gentzkow, M., Shapiro, J.M. & Taddy, M. (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574.

Gerrish, S.M. & Blei, D.M. (2012). How They Vote: Issue-Adjusted Models of Legislative Behavior. In F. Pereira, C.J. Burges, L. Bottou & K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25 (NIPS 2012).*

Giglietto, F., Iannelli, L., Rossi, L., Valeriani, A., Righetti, N., Carabini, F., Marino, G., Usai, S. & Zurovac, E. (2018, May 17). *Mapping Italian news media political coverage in the lead-up of 2018 general election*. SSRN Electronic Journal.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). *Chatgpt outperforms crowd-workers for text-annotation tasks.* arXiv:2303.15056.

Glaser, B., & Strauss, A. (2017). *Discovery of grounded theory: Strategies for qualitative research.* Routledge.

Goet, N.D. (2017, October 5-7). *Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015.* [Paper prepared for workshop]. Polarization, Institutional Design and the Future of Representative Democracy workshop, Harnack Haus, Berlin.

Goodson, I. F., & Gill, S. R. (2011). The narrative turn in social research. *Counterpoints, 386*, 17–33.

Gorrell, G., Greenwood, M., Roberts, I., Maynard, D. & Bontcheva, K. (2018). *Online abuse of UK MPs in 2015 and 2017: Perpetrators, targets, and topics*. arXiv:1804.01498.

Gray, M. L. & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

Green, J., Edgerton, J., Naftel, D., Shoub, K. & Cranmer, S.J. (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances, 6*(28).

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18*(1), 1–35.

Gross, J., Acree, B., Sim, Y. & Smith, N.A. (2013, September 18). Testing the Etch-a-Sketch Hypothesis: A Computational Analysis of Mitt Romney's Ideological Makeover During the 2012 Primary vs. General Elections. *APSA 2013 Annual Meeting Paper, American Political Science Association 2013 Annual Meeting*.

Gross, M. & Jankowski, M. (2020). Dimensions of political conflict and party positions in multi-level democracies: evidence from the Local Manifesto Project. *West European Politics, 43*(1), 74–101.

Grover, P., Kar, A.K., Dwivedi, Y.K. & Janssen, M. (2019a). Polarization and acculturation in US Election 2016 outcomes – Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change, 145*, 438–460.

Grover, T., Bayraktaroglu, E., Mark, G. & Rho, E.H.R. (2019b). Moral and affective differences in U.S. immigration policy debate on twitter. *Computer Supportive Cooperative Work, 28*, 317–355.

Guntuku, S.C., Purtle, J., Meisel, Z.F., Merchant, R.M. & Agarwal, A. (2021). Partisan differences in twitter language among US legislators during the COVID-19 pandemic: Cross-sectional study. *Journal of Medical Internet Research, 23*(6), e27300.

Hausladen, C.I., Schubert, M.H. & Ash, E. (2020). Text classification of ideological direction in judicial opinions. *International Review of Law and Economics 62*(105903).

Hays, D. C. (1960). *Automatic Content Analysis*. Rand Corporation.

Hemphill, L., Culotta, A. & Heston, M. (2016). #Polar Scores: Measuring partisanship using social media content. *Journal of Information Technology and Politics, 13*(4), 365–377.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261-266.

Hirst, G., Riabinin Y. & Graham, J. (2010). Party status as a confound in the automatic classification of political speech by ideology. In S. Bolasco, I. Chiari & L. Giuliano. (Eds.), *Proceedings of 10th International Conference, Journées d'Analyse statistique des Données Textuelles, 9-11 June 2010 - Sapienza University of Rome*.

Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M. & Morris, C. (2014). Text to ideology or text to party status? In B. Kaal, I. Maks, A. van Elfrinkhof. (Eds.), *From text to political positions: Text analysis across disciplines (pp. 93-116)*. John Benjamins Publishing Company.

Hjarvard, S. (2008). The mediatization of society: A theory of the Media as Agents of Social and Cultural Change. *Nordicom review, 29*(2).

Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J. & Wissik, T. (2020). Comparing Lexical Usage in Political Discourse across Diachronic Corpora. In *Proceedings of ParlaCLARIN II Workshop of the Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020* (pp. 58-65). European Language Resources Association.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

Homan, S., Gabi, M., Klee, N., Bachmann, S., Moser, A.-M., Duri', M., Michel, S., Bertram, A.-M., Maatz, A., Seiler, G., Stark, E., & Kleim, B. (2022). Linguistic features of suicidal thoughts and behaviors: A systematic review. *Clinical Psychology Review*, *95*(102161), 102161.

Hovy, E. –Lavid J. (2010). Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, *22*(1), 13–36.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). *The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning*. arXiv:2203.06498.

Ignatow, G., & Mihalcea, R. F. (2017). *An Introduction to Text Mining: Research Design, Data Collection, and Analysis* (1st edition.). SAGE Publications, Inc.

Ilie, C. (2017). Parliamentary debates. In Wodak & Forchtner (Eds.), *The Routledge handbook of language and politics* (pp. 309-325).

Iliev, I.R., Huang, X. & Gel, Y.R. (2019). Political rhetoric through the lens of non-parametric statistics: are our legislators that different? *Journal of the Royal Statistical Society: Series A (Statistics in Society), 182*(2), 583–604.

Ilyés, V., Katona, E., Morvay G., Putz, O., & Varju, Z. (2018). *Gender bias in Hungarian political discourse.* [Conference presentation]. PolText konferencia, Társadalomtudományi Kutatóközpont.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2018). Quantitative analysis of large amounts of journalistic texts using topic modelling. In Karlsson, M. & Sjøvaag, H (Eds.), *Rethinking Research Methods in an Age of Digital Journalism* (pp. 89–106). Routledge.

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology, 22*(5), 469–485.

Jelveh, Z., Kogut, B. & Naidu, S. (2014). Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1804–1809). Association for Computational Linguistics.

Jensen, J., Kaplan, E., Naidu, S. & Wilse-Samson, L. (2012). Political polarization and the dynamics of political language: Evidence from 130 years of partisan speech. *Brookings Papers on Economic Activity, 2012*(Fall), 1–81.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, (3rd ed.). Prentice Hall.

Karamshuk, D., Lokot, T., Pryymak, O. & Sastry, N. (2016). Identifying partisan slant in news articles and twitter during political crises. In *Lecture Notes in Computer Science*. Springer International Publishing pp. 257–272.

Katona. E., Kmetty, Z., & Németh, R. (2021). A korrupció hazai online médiareprezentációjának vizsgálata természetes nyelvfeldolgozással. *Médiakutató, 22*(2), 69-88.

Kelly, B., Manela, A. & Moreira, A. (2021). Text selection. *Journal of Business and Economic Statistics, 39*(4), 859-879.

KhudaBukhsh, A.R., Sarkar, R., Kamlet, M.S. & Mitchell, T.M. (2020). *We don't speak the same language: Interpreting polarization through machine translation*. arXiv:2010.02339.

Király G., Dén-Nagy I., Géring Zs., Nagy B. (2014). Kevert módszertani megközelítések: Elméleti és módszertani alapok. *Kultúra és Közösség, 5*(2), 95-104.

Kmetty, Z. (2022). Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása= How to use vector space models in social sciences. *Statisztikai Szemle*, *100*(2), 105-136.

Kmetty, Z., Koltai, J. & Rudas, R. (2021). The presence of occupational structure in online texts based on word embedding NLP models. *EPJ Data Science, 10*(1):55.

Kmetty, Z., & Németh, R. (2022). Which is your favorite music genre? A validity comparison of Facebook data and survey data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 154*(1), 82-104.

Knap, Á., Tóth, T. E., & Rakovics, Z. (2022). Humán annotált emóciókorpusz létrehozása aktorokhoz köthető érzelmek detektálására. *Digitális Bölcsészet*, (6), M-3.

Kobayashi, T., Ogawa, Y., Suzuki, T. & Yamamoto, H. (2019). News audience fragmentation in the Japanese Twittersphere. *Asian Journal of Communication, 29*(3), 274–290.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905-949.

Koylu, C., Larson, R., Dietrich, B.J. & Lee, K.-P. (2019). CarSenToGram: geovisual text analytics for exploring spatiotemporal variation in public discourse on Twitter. *Cartography and Geographic Information Science, 46* (1), 57–71.

Krippendorff, K. (1995). *A tartalomelemzés módszertanának alapjai*. Balassi Kiadó.

Kulkarni, V., Ye, J., Skiena, S. & Wang, W.Y. (2018). *Multi-view models for political ideology detection of news articles.* arXiv: 1809.03485.

Lakoff, G. (2002). *Moral politics: How liberals and conservatives think* (2nd ed.). University of Chicago Press.

Landeiro, V. & Culotta, A. (2018). Robust text classification under confounding shift. *Journal of Artificial Intelligent Research, 63*, 391-419.

Lauderdale, B.E. & Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis, 24*(3), 374–394.

Laver, M., Benoit, K. & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311–331.

Lelkes, Y. (2016). Mass Polarization: *Manifestations and Measurements. Public Opinion Quartely, 80*(S1), 392–410.

Light, R., & Cunningham, J. (2016). Oracles of peace: Topic modeling, cultural opportunity, and the Nobel peace prize, 1902–2012. *Mobilization: An International Quarterly, 21*(1), 43–64.

Lin, W.-H., Wilson, T., Wiebe, J. & Hauptmann, A. (2006). Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), June 2006* (pp. 109-116). Association for Computational Linguistics.

Liu, C., & Lu, X. (2018). Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community. *BMC medical informatics and decision making*, *18*(1), 1–10.

Liu, J. & Zhang, X. (2019). The role of domain knowledge in document selection from search results: The Role of Domain Knowledge in Document Selection From Search Results. *Journal of the Association for Information Science and Technology*, *70*(11), 1236–1247.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30.*

Macanovic, A. (2022). Text mining for social science–The state and the future of computational text analysis in sociology. *Social Science Research*, *108*(102784).

Matejka, F. (2013). Attention Discrimination: Theory and Field Experiments. *Society for Economic Dynamics 2013 Meeting Papers*, *798.*

McCallum, A. K. (2002). Mallet: A machine learning for languagetoolkit. http://mallet. cs. umass. edu.

McCarty, N.M., Poole, K.T. & Rosenthal, H. (2006). *Polarized America: The dance of ideology and unequal riches.* MIT Press.

Medzihorsky, J., Littvay, L. & Jenne, E.K. (2014). Has the Tea Party era radicalized the Republican Party? Evidence from text analysis of the 2008 and 2012 Republican primary debates. *PS Political Science & Politics, 47*(4), 806–812.

Meeks, E., & Weingart, S. B. (2012). The digital humanities contribution to topic modeling. *Journal of Digital Humanities, 2*(1), 1–6.

Mendez, G.R., Cosby, A.G. & Mohanty, S.D. (2018). Obamacare on Twitter: Online *Political Participation and its Effects on Polarisation. Teorija* in Praksa 55, 419-444.

Mikolov, T., Sutskever I., Chen K., Corrado, G. S., & Dean J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the Conference on Advances in Neural Information Processing Systems 26 (NIPS).*

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com, Morrisville, NC.

Molnar, C. & Freiesleben, T. (2024): *Supervised Machine Learning for Science.* Manuscript. https://ml-science-book.com/ , letöltve: 2024. 02. 07.

Morini, V., Pollacci, L. & Rossetti, G. (2020, June 21-24). *Capturing Political Polarization of Reddit Submissions in the Trump Era.* [paper presented at conference]. SEBD 2020, June 21-24, 2020, Villasimius, Italy.

Mouffe, C. (2011). On the political. Routledge.

Mühlhoff, R. (2019). Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning. New Media & Society, 22(10), 1868–1884.

Müller, M. (2008). Reconsidering the concept of discourse for the field of critical geopolitics: Towards discourse as language and practice. *Political Geography, 27*, 322–338.

Mützel, S. (2015). Facing big data: Making sociology relevant. *Big Data & Society, 2*(2), 2053951715599179.

Nahili, W., Rezeg, K. & Kazar, O (2020). Big Data Analytics using Supervised Learning: A Comprehensive Review of Recent Techniques. *International Journal for Research in Applied Science and Engineering Technology*, *8*(1), 305–312.

Németh, R. (2015a). A számok tényleg magukért beszélnek? Hozzászólás Dessewfy Tibor és Láng László írásához. *Replika*, (92-93), 203-208.

Németh, R. (2015b). Oksági következtetés az empirikus szociológiai kutatásban. *Szociológiai Szemle*, *25*(2), 2–30.

Németh R., Katona E. & Kmetty Z. (2020). Az automatizált szöveganalitika perspektívája a társadalomtudományokban. *Szociológiai Szemle*, *30*(1), 44–62.

Németh, R., Sik, D. & Máté, F. (2020). Machine Learning of Concepts Hard Even for Humans: The Case of Online Depression Forums. *International Journal of Qualitative Methods*, *19*(1), 1–8.

Németh, R., Sik, D., & Katona, E. (2021). The asymmetries of the biopsychosocial model of depression in lay discourses - Topic modelling online depression forums. *SSM - Population Health*, *14*(100785).

Németh, R. (2021). *Az okság alternatív fogalmi és módszertani megközelítései a szociológiában.* Savaria University Press.

Németh, R., Máté, F., Katona, E., Rakovics, M. & Sik, D. (2022). Bio, psycho, or social: supervised machine learning to classify discursive framing of depression in online health communities. *Quality and Quantity: International Journal of Methodology*, 3933–3955.

Németh, R. & Koltai, J. (2023). Natural language processing: The integration of a new methodological paradigm into sociology. *Intersections. East European Journal of Society and Politics, 9*(1), 5–22.

Németh, R., Sik, D., Zaboretzky, B., & Katona, E. (2023a, June). Depression in times of a pandemic–the impact of COVID-19 on the lay discourses of e-mental health communities. *Information, Communication & Society*, 1–23.

Németh, R., Katona, E., Balogh, P., Rakovics, Zs., & Unger, A. (2023b). What else comes with a geographical concept beyond geography? The renaissance of the term 'Carpathian Basin' in the Hungarian Parliament. *Intersections. East European Journal of Society and Politics*, megjelenés alatt

Németh, R. (2023). A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, *6*(1), 289–313.

Németh, R., & Sik, D. (2024). Beyond "latent thematic structure" - Extending the interpretation of the topic model towards pragmatics. *Intersections. East European Journal of Society and Politics*, bírálat alatt.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P., & Miler, K. (2015). Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1438–1448.

Nimrod, G. (2013). Online depression communities: members' interests and perceived benefits. *Health communication, 28*(5), 425–434.

Ntoutsi, E. et al. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, *10*(6), 1–14.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pennebaker, J.W., Booth, R.J., Boyd, R.L. & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Pfeiffer, P. N., Heisler, M., Piette, J. D., Rogers, M. A., & Valenstein, M. (2011). Efficacy of peer support interventions for depression: a meta-analysis. *General Hospital Psychiatry, 33*(1), 29–36.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V. & Bosco, C. (2017). Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In Basili, R., Nissim, M., Satta, G. (Eds.). *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome*. Accademia University Press.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 231–240). Association for Computational Linguistics.

Praet, S., Van Aelst, P., Daelemans, W., Kreutz, T., Peeters, J., Walgrave, S. & Martens, D. (2021, February 8). Comparing automated content analysis methods to distinguish issue communication by political parties on twitter. *SSRN Electronic Journal*.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science, 54*(1), 209–228.

Quraishi, M., Fafalios, P. & Herder, E. (2018). Viewpoint discovery and understanding in social networks. In *Proceedings of the 10th ACM Conference on Web Science*. ACM.

Rao, A., Morstatter, F., Hu, M., Chen, E., Burghardt, K., Ferrara, E. & Lerman, K. (2020). *Political partisanship and anti-science attitudes in online discussions about covid-19.* arXiv: 2011.08498.

Rashed, A., Kutlu, M., Darwish, K., Elsayed, T. & Bayrak, C. (2020). *Embeddings-based clustering for target specific stances: The case of a polarized Turkey*. arXiv:2005.09649.

Rehurek, R. & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Rho, E.H.R., Mark, G. & Mazmanian, M. (2018). Fostering civil discourse online: Linguistic behavior in comments of #MeToo articles across political perspectives. *Proceedings of the ACM of Human-Computer Interaction 2*, 1–28.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software, 91*, 1–40.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408).

Rumshisky, A., Gronas, M., Potash, P., Dubov, M., Romanov, A., Kulshreshtha, S. & Gribov, A. (2017). Combining network and language indicators for tracking conflict intensity. In G. Ciampaglia, A. Mashhadi, T. Yasseri. (Eds.), *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II.* (pp. 391–404) Springer International Publishing.

Ryan, L. & McKie, L. (Eds.). (2015). *An end to the crisis of empirical sociology? Trends and challenges in social research*. Routledge.

Samantray, A. & Pin, P. (2019). Credibility of climate change denial in social media. *Palgrave Commun. 5*(127).

Sanders, J., Lisi, G. & Schonhardt-Bailey, C. (2017). Themes and topics in parliamentary oversight hearings: A new direction in textual data analysis. *Stat. Politics Policy 8*(2), 153–194.

Sap, M., Card, D., Gabriel, S., Choi, Y. & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Sapiro-Gheiler, E. (2018). *"read my lips": Using automatic text analysis to classify politicians by party and ideology*. arXiv: 1809.00741

Savage, M. & Burrows, R. (2007). The Coming Crisis of Empirical Sociology. *Sociology: A Journal of the British Sociological Association, 41*(5), 885–899.

Schakel, A.M. & Wilson, B.J. (2015). *Measuring word significance using distributed representations of words*. arXiv:1508.02297.

Sen, I., Flöck, F., Weller, K., Weiß, B. & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly, 85*(S1), 399–422.

Serrano-Contreras, I.-J., García-Marín, J. & Luengo, Ó.G. (2020). Measuring online political dialogue: Does polarization trigger more deliberation? *Media and Communication 8*(4), 63–72.

Shen, Q. & Rosé, C.P. (2019). The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 58–69). Association for Computational Linguistics.

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Sik, D., Rakovics, M., Buda, J., & Németh, R. (2023a). The impact of depression forums on illness narratives: a comprehensive NLP analysis of socialization in e-mental health communities. *Journal of Computational Social Science, 6,* 781–802.

Sik, D., Rakovics, M. & Németh, R. (2023b): *The manifest and latent structures of medicalization and psychologization in lay depression discourses – a word embedding analysis of online forums.* [Unpublished manuscript]. Eötvös Loránd University.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.

Simonovits, G., Simonovits, B., Vig, A., Hobot, P., Németh, R., & Csomor, G. (2022). Back to "normal": the short-lived impact of an online NGO campaign of government discrimination in Hungary. *Political Science Research and Methods, 10*(4), 848-856.

Sinno, B., Oviedo, B., Atwell, K., Alikhani, M., & Li, J.J. (2021). *Political ideology and polarization of policy positions: A multi-dimensional approach*. arXiv:2106.14387

Slapin, J.B. & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52*(3), 705–722.

Snow, R. –O'Connor B. –Jurafsky D. –Ng A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stecula, D.A. & Merkley, E. (2019). Framing climate change: Economics, ideology, and uncertainty in American news media content from 1988 to 2014. *Frontiers in Communication, 4.*

Stefanov, P., Darwish, K., Atanasov, A. & Nakov, P. (2019). *Predicting the topical stance of media and popular Twitter users*. arXiv:1907.01260

Szabó, M. K., Ring, O., Nagy, B., Kiss, L., Koltai, J., Berend, G., ... & Kmetty, Z. (2020). Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. Historical Methods: *A Journal of Quantitative and Interdisciplinary History*, *54*(1), 1–13.

Taddy, M. (2012). On Estimation and Selection for Topic Models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR 22* (pp. 1184–1193).

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of American Statistical Association 108*(503), 755–770.

Taddy, M. (2015). Distributed multinomial regression. *Annals of Applied Statistics, 9*(3), 1394–1414.

Trabelsi, A. & Zaiane, O. (2018). Unsupervised Model for Topic Viewpoint Discovery in Online Debates Leveraging Author Interactions. In *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1).

Thonet, T., Cabanac, G., Boughanem, M. & Pinel-Sauvagnat, K. (2017). Users are known by the company they keep: Topic models for viewpoint discovery in social networks. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* (pp. 87–96). ACM.

Tsur, O., Calacci, D., & Lazer, D. (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, Beijing, China, July 26-31, 2015. pp. 1629–1638.

Tucker, E.C., Capps & C.J., Shamir, L. (2020). A data science approach to 138 years of congressional speeches. *Heliyon 6*(e04417).

Üveges, I., & Ring, O. (2023, June). HunEmBERT: a fine-tuned BERT-model for classifying sentiment and emotion in political communication. *IEEE Access, 11*, 60267–60278.

van Dijk, Teun A. (1994). Critical Discourse Analysis. *Discourse & Society 5*(4), 435–436.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory* (2nd ed.). Springer-Verlag.

Villa-Cox, R., KhudaBukhsh, A.R. & Carley, K.M. (2021). *Exploring Polarization of Users Behavior on Twitter During the 2019 South American Protests*. arXiv:2104.05611.

Wang, Y., Feng, Y., Hong, Z., Berger, R. & Luo, J. (2017*). How polarized have we become? A multimodal classification of Trump followers and Clinton followers*. arXiv:1711.00617.

Wang, R.T. & Tucker, P.D. (2021). How partisanship influences what Congress says online and how they say it. *American Political Research, 49*(1), 76–90.

Wesslen, R. (2018). *Computer-assisted text analysis for social science: Topic models and beyond*. arXiv:1803.11045.

Widmer, P., Galletta, S. & Ash, E. (2020). Media slant is contagious. Center for Law & Economics *Working Paper Series, 14/2020*.

Wodak, R., & Forchtner, B. (Eds.). (2017). *The Routledge handbook of language and politics*. Routledge.

Wodak, R. (2010). The discursive construction of history - brief considerations. Mots. *Les Langages du Politique,* (94), 57–65.

Wodak, R., de Cillia, R., Reisigl, M., & Liebhart, K. (2009). *Discursive construction of national identity* (2nd ed.). Edinburgh University Press.

Wu, P.Y., Mebane, W.R., Woods, L., Klaver, J. & Duek, P. (2019). Partisan Associations of Twitter Users Based on Their Self-descriptions and Word Embeddings. Paper prepared for presentation at the 2019 Annual Meeting of the American Political Science Association, Washington, D.C., August 29–September 1, 2019.

Yan, H., Lavoi, A. & Das, S. (2017). The Perils of Classifying Political Orientation From Text. LINKDEM@ IJCAI. Retrieved from http://ceur-ws.org/Vol-1897/paper3.pdf

Yan, H., Das, S., Lavoie, A., Li, S. & Sinclair, B. (2019). The Congressional Classification Challenge: Domain Specificity and Partisan Intensity. *In Proceedings of the 2019 ACM Conference on Economics and Computation* (pp. 71–89).

Yarchi, M., Baden, C. & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication, 38*(1-2), 98–139.

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. *NPJ Digital Medicine*, *5*(1), 46.

Zubiaga, A., Wang, B., Liakata, M. & Procter, R. (2017). *Stance Classification of Social Media Users in Independence Movements*. arXiv:1702.08388.

Bár a nyelv a társadalmi interakciók egy fontos eszköze, a kvantitatív társadalomkutatás – elsősorban adatgyűjtési és feldolgozási eszközök hiányában - mégsem használta igazán évtizedeken át. A helyzet az utóbbi évtizedben - a digitális forradalomnak köszönhetően - gyökeresen megváltozott, a „text as data" mozgalom keretében a szöveges adat, mint empirikus társadalomkutatási bázis használata exponenciális ütemben terjed.

A szöveges adatok mennyiségének és hozzáférhetőségének ez a forradalma jelentősen kiszélesítette az empirikusan vizsgálható társadalomkutatási kérdések körét: az egyének, csoportok és intézmények viselkedését, azok kölcsönhatását és időbeli dinamikáját naponta többmillió terrabyte-nyi digitális szöveg képezi le, s ez az adatvagyon a digitalizáció előrehaladtával egyre sokszorozódik.

A társadalmat leíró szöveges adatok forradalmával párhuzamosan az utóbbi tíz évben a számítási kapacitások és azzal párhuzamosan az adatok elemzésére szolgáló szöveganalitikai technológiák robbanásszerű fejlődése is bekövetkezett, s az új technológiák a szöveg feldolgozásának már releváns mélységét nyújtják. Ez a robbanás a számítástudomány és számítógépes nyelvészet után a bölcsészet- és a társadalomtudományokat, így a szociológiát is elérte.

A kötet ezeket az inspiráló lehetőségeket mutatja be, a természetes nyelvfeldolgozás (natural language processing, NLP) szociológiai alkalmazásaiba engedve bepillantást, az ELTE Társadalomtudományi Karán a Research Center for Computational Social Science kutatócsoportban 2018 óta folyó kutatásokon, mint esettanulmányokon keresztül.

Az NLP technikai oldalának ismertetésére kiváló források állnak rendelkezésre, de a társadalomkutatási tapasztalatokat és kihívásokat kevesebb szerző tárgyalja. A társadalomkutatás alkalmazási specifikumát az adja, hogy az itt tárgyalt problémák egy évszázados kutatási paradigmába vannak ágyazva, kérdésfeltevései így lényegesen különböznek a számítástudomány vagy az ipari felhasználás kérdéseitől. Ennek a különbségnek pedig tudatában kell lennünk, amikor adaptáljuk az informatika oldaláról érkező innovációt.

A könyv ideális olvasója az a társadalomkutató, aki érdeklődik a számítógépes szövegelemzés lehetőségei iránt. Ugyanakkor a társadalom empirikus megismerésének tudományos módszerei iránt érdeklődő laikusoknak is ajánlható a könyv ismeretterjesztő munkaként, hiszen a szerző konkrét példákkal szemléltet és közérthetően magyaráz.