Katona Eszter[1] – Németh Renáta[2]

Automatizált szöveganalitika a korrupció kutatásában

https://doi.org/10.18030/socio.hu.2021.1.108

Absztrakt

Tanulmányunk a természetesnyelv-feldolgozás *(Natural Language Processing, NLP)* korrupciókutatásban való felhasználását és felhasználhatóságát vizsgálja. Átfogó irodalmi áttekintésünk során a 2000 után született, automatizált szövegelemzésre épülő korrupciókutatások teljeskörű összegyűjtésére és összegzésére törekedtünk az NLP alkalmazás elterjedtségére, illetve lehetőségeire fókuszálva. Lényeges eltéréseket találtunk a felhasznált szöveges adatforrást, a korrupció mérésének módját és az elemzési megközelítést tekintve, ugyanakkor sajnálatosan kevés volt az (adatforrását, módszerét vagy mérési módját tekintve) kevert típusú tanulmány. A klasszikus, a korrupció volumenét vagy a vele kapcsolatos attitűdöt vagy percepciót leíró ill. észlelésének következményeit vizsgáló munkákon kívül találtunk a korrupció megelőzésére felhasználható eredményeket, sőt intervencióra közvetlenül alkalmasakat is. Az NLP-t csupán néhány tanulmány használta, és ezek egy része sem annyira tartalmi, mint csupán technikai feladatra. Eredményeink szerint az NLP nem nagyon elterjedt még ezen a területen, ugyanakkor az is látható, hogy gyümölcsöző lehet a használata: alternatív eszközként jól támogathatná a tradicionális kvantitatív kutatásokat. Cikkünk célja inspirációt adni az NLP társadalomtudományi felhasználására és felhívni a figyelmet annak beágyazhatóságára a meglevő tudományos diskurzusokba.

Kulcsszavak: korrupció, természetesnyelv-feldolgozás, automatizált szöveganalitika, szövegelemzés

Automated text analytics in corruption research

Abstract

Our study examines the use and possible applicability of Natural Language Processing (NLP) in corruption research. In our review, we aim to collect and summarize automated text analytics-based corruption research born after 2000. We focus on the prevalence and potential of NLP methods. We found significant differences in the textual data sources, the corruption measurement methods, and the analytical approaches used. However, there were unfortunately few mixed-type studies (in terms of data source, method, or corruption measurement method). In addition to the classic works describing of the volume of corruption or the attitude or perception related to it, we found results that can be used to prevent corruption and even be directly suitable for intervention. NLP has been used in only a few studies, and mostly only for some technical tasks. Our results show that NLP is not very widespread in this area yet. However, it can also be seen that its use can be useful and could support traditional quantitative research as an alternative tool. The aim of our article is to provide inspiration for the use of NLP in the social sciences and to draw attention to its embeddability in existing scientific discourses.

Keywords: corruption, natural language processing, automated text analytics, text analysis

1 Katona Eszter, ELTE Társadalomtudományi Kar, Research Center for Computational Social Sciences. A szerzőt munkájában az ÚNKP-19-3-II-ELTE-591 támogatta.

2 Németh Renáta, ELTE Társadalomtudományi Kar, Research Center for Computational Social Sciences. A szerzőt munkájában a K-134428 azonosítójú NKFIH pályázat támogatta.

Katona Eszter – Németh Renáta

Automatizált szöveganalitika a korrupció kutatásában

Bevezetés

A digitális forradalom az emberiség írásbeli önkifejezésének forradalma is. A közösségi média felületeken, szöveggé alakított videókon, digitalizált könyvtárakban, hatósági nyilvántartásokban felhalmozódott szövegek a társadalmi valóság sosem látott mértékű elérését teszik lehetővé. A digitális szövegeknek ez az univerzuma párhuzamosan olyan szöveganalitikai technológiákat hívott életre, melyek ma már a szociológia számára is releváns mélységű elemzésre nyújtanak lehetőséget.

Míg a szociológia klasszikus módszertana kvalitatív módon vagy egyszerű szógyakoriságok/mintázatok kvantitatív elemzésével közelítette meg a szöveges adatokat, a mintegy két évtizede megjelent természetes nyelv-feldolgozás *(Natural Language Processing, NLP)* eszközei által az egyre nagyobb tömegben képződő *(„big data")* szöveges tartalom információtartalma automatizált módon vált elérhetővé (Hirschberg–Manning 2015). Cikkünkben a korrupciókutatás területén született példákon keresztül illusztráljuk, hogy e módszerek ugyan még távol vannak a szövegek tényleges megértésétől, ám hatékonyan támogatják célzott kutatási kérdések mentén a tényleges elolvasással nem megközelíthető szövegtömegek feldolgozását. Az NLP különböző területei közül a társadalomtudományokat elsősorban az írott nyelvi források szemantikai/tartalmi elemzése érinti. A természetesnyelv-feldolgozás általunk is használt részterületére különböző kifejezésekkel utalnak: mint számítógépes nyelvészet *(computational linguistics),* automatizált szövegelemzés *(automated text analytics),* szövegbányászat *(text mining).* A szövegbányászat és az adatbányászat egymáshoz közelálló terület, de míg az utóbbi strukturált adatokkal dolgozik, addig az előbbi strukturálatlan vagy félig strukturált adatokkal.[3] A szövegbányászat új, korábban azonosítatlan információk feltárására törekszik írott forrásokban (Vijayarani et al. 2015). Mi a fenti kifejezéseket szinonimaként használjuk. Az NLP mint módszertan iránt érdeklődőknek általános áttekintésként ajánlható Aggarwal és Zhai (2012, korrekt statisztikai alapokkal) és Ignatow és Mihalcea (2016, alkalmazásorientált megközelítésben, szociológusoknak szóló) munkája, valamint Németh et al. (2020) friss hazai összefoglalója is rendelkezésre áll a módszer szociológiai lehetőségeiről.

Szakirodalmi áttekintésünk ezeknek az új technológiáknak a korrupciókutatásban való alkalmazhatóságát vizsgálja. Témaválasztásunk indoka a korrupció önmagában vett társadalmi fontosságán túl az is, hogy a korrupció látens jelenségként nehezen kvantifikálható; valamint e komplex jelenség mögött egyszerre állnak gazdasági, politikai, közigazgatási, társadalmi és kulturális tényezők, ami nem csak koncepcionálisan, de kutatásmódszertanilag is több megközelítést implikál. Mint látni fogjuk, a korrupció kutatása több oldalról kapcsolható szöveges adatokhoz. A szöveges adatok felhasználásával egy új szemléletű megközelítésre nyílik lehe-

3 A strukturált adatoknak azokat az adatokat nevezzük, amelyek hagyományosan sorokban és oszlopokban rögzítettek, könnyen kereshetők. A félig strukturált adatok olyan tulajdonságokkal bírnak, amelyek megkönnyítik az elemzést, amelyek segítségével hierarchiába rendezhetők az információk. Ilyen például egy webáruház, amely minden termékről ismétlődő struktúrában tárol adatot. Ennek segítségével az adatokat gyorsan és könnyen egy strukturált adatbázisba rendezhetjük. A strukturálatlan adatoknak egyáltalán nincsen adatbázis jellege, esetünkben lehetnek újságcikkek, blogbejegyzések, hosszabb szöveges dokumentumok.

tőség, ugyanakkor a folyamatosan termelődő szöveges adatforrások alternatívát is nyújthatnak meglevő empirikus problémákra, mint pl. a klasszikus survey kutatásokban jelentkező egyre jelentősebb válaszmegtagadásra (az új adatforrás és a survey összevetésének további szempontjairól lásd Németh 2015).

Írásunk célja a korrupciókutatók és – a korrupciókutatáson mint esettanulmányon keresztül – a szociológusok figyelmének felkeltése az NLP-technikák alkalmazásában rejlő lehetőségekre, s azok beágyazhatóságára a meglevő tudományos diskurzusokba. Azt vizsgáljuk, hogy mennyire elterjedt akár a klasszikus, akár az automatizált szöveganalitikai módszer a korrupció kutatásában. Célunk a meglévő irodalom felkutatása, a kutatásmódszertani paradigma elterjedtségének meghatározása és áttekintése. A klasszikus, nem-automatizált szövegelemzések áttekintése lehetőséget ad arra is, hogy rámutassunk az NLP-módszerekkel még nem kutatott területekre, mint kiaknázatlan lehetőségekre. Írásunk reményeink szerint hozzájárul a tradicionális társadalomkutatási módszerektől távol eső adattudomány és gépi tanulás nyelvének megértéséhez is, és további kutatásokra inspirálhat másokat.

A KORRUPCIÓ MÉRÉSE

Az automatizált szöveganalitika felhasználási lehetőségeit aszerint vehetjük számba, hogy a korrupció méréséhez hogyan járulhatnak hozzá. A korrupció mérése (lásd például Tóth–Hajdu 2018) több megközelítést követhet. A Nemzetközi Valutaalap (IMF) tanulmánya (Hlathswayo et al. 2018, magyar feldolgozása Gerő–Mikola, 2020) alapján a korrupciós indikátorok három generációját különböztethetjük meg (1. ábra) időbeli keletkezésük és módszertani megközelítésük szerint. Az alábbiakban aszerint vesszük az indikátorokat sorra, hogy milyen adatforrást használnak és milyen módszertani problémával néznek szembe – így helyezhetjük kontextusba az automatizált szöveganalitika nyújtotta mérőeszköz lehetőségeit és kihívásait.

1. ábra. A korrupció-indikátorok három generációja

Forrás: saját ábra Hlathswayo et al. 2018 és Gerő–Mikola 2020 alapján

Az első generációs indikátorok a korrupciós tapasztalatok, észlelések, illetve a korrupció jelenségéhez kapcsolódó attitűdök mérését célozzák; e kutatások általában kérdőíves adatfelvételeket használnak: szakértői értékeléseket (Korrupció Percepciós Index, CPI) és lakossági felméréseket (Global Corruption Barometer, GCB). A felmérések előnye, hogy célpopulációjukra nézve megfelelő mintavétel mellett reprezentatív eredményt adnak és garantálják az anonimitást. Ugyanakkor mérőeszközökként megvan az a hátrányuk, hogy erős nyelvi

és kulturális meghatározottságúak (lásd a korrupció, mint kifejezés országonként eltérő jelentését, Axelsson – Dahlberg 2018). Továbbá maga a survey is hatással lehet az eredményekre, például a korrupció elutasítására vonatkozó társadalmi elvárások által okozott torzítás miatt, így nem tekinthető beavatkozásmentes vizsgálatnak (a jelenség survey statisztikai hátteréről lásd Lavrakas 2008 *„social desirability bias"* szócikkét).

A második generációs mutatók empirikus megalapozottsága már kevésbé kérdőjelezhető meg, mert a személyes tapasztalatok mellett objektívebb indikátorokat is felhasználnak, ilyenek például a különböző bürokratikus intézmények teljesítményét bemutató indikátorok. Azonban az első generációs mutatókhoz hasonlóan itt is problémát jelent a kérdőíves kutatások kapcsán jól ismert, a téma érzékenységéből fakadó magas látencia (félelem a következményektől) és a fogalom definíciós nehézségéből adódó szubjektivitás. Emellett problémát jelent az, hogy a kormányzati, közigazgatási mutatók nehezen hozzáférhetők, létrejöttük módszertana nem minden esetben ismert, emellett ezek a mutatók sokkal inkább a bürokratikus kapacitásbeli különbségeket mérik, mintsem magát a korrupciót.

A harmadik generációs, *big data* alapú mutatók kevésbé kritizálhatók a szubjektivitás szempontjából. Klasszikus beavatkozásmentes mérésre alapulva a korrupció volumenét mérik és már meglévő adatokból készítenek indikátorokat, mint az ábrán jelölt, IMF által létrehozott, a korrupció médiareprezentációjára építő „news-flow index", melyről később írunk részletesen. Ezek a mérések objektív adatokon alapulnak. Jellemzően korrupciógyanús eseteket tárnak fel, ahogy Weaver (2020) tette a csúszópénzzel befolyásolt közalkalmazotti állás-betöltésekkel, vagy a korrupciós kockázatot mérik és az annak létrejöttét elősegítő körülményeket vizsgálják, mint például Fazekas és Tóth (2016). Utóbbi kutatók, feltételezve, hogy ha valaki csalni szeretne, akkor a csalás számára kedvező feltételeket hoz létre, szerződési hálózatok felderítése során azonosítottak magas közbeszerzési korrupciókockázatot mutató szervezeteket.

Utóbbi kutatások jól példázzák, hogy szöveges adatbázisokra (azaz: sajtóhírek, közbeszerzési pályázatok, közösségi média szövegeire) is épülhetnek beavatkozásmentes vizsgálatok, ahol a hagyományos kvantitatív és kvalitatív módszerek mellett – melyek szöveganalitikai megközelítése főleg szógyakoriság elemzéseket és kézi kódolást használ – eredményesen használható a *big data* elemzésre alkalmas NLP is. Valóban, a szövegbányászati megoldások társadalomtudományi elterjedésével párhuzamosan megjelentek az első olyan korrupciókutatási eredmények is, amelyek előállításában a természetesnyelv-feldolgozás (NLP) nem csupán technikai-adatgyűjtési, hanem tartalmi eredményt is előállító elemzési eszközként funkcionált. Ide tartozik például Axelsson és Dahlberg (2018) munkája, akik több tucat országot érintő nemzetközi vizsgálatban bizonyították, hogy a „korrupció" kifejezés online médiabeli előfordulása erősen korrelál más bevett észlelésalapú mutatókkal, mint például a Transparency International Korrupció Percepciós Indexe (CPI). Másik eredményük a „korrupció" kifejezéshez együtt-használatuk alapján szemantikailag társítható fogalmak körének azonosítása, ezzel a korrupció, mint erkölcsi eltérés régiónként eltérő értelmezésének megragadása.

A KUTATÁS MÓDSZERE

Elemzésünk során a *scoping review* (Arksey és O'Malley, 2005) azaz: a „hatókörbecslő" szakirodalmi áttekintés műfaját alkalmazzuk. A *scoping review* lényege egy szisztematikusabb áttekintéssel [például *systematic review* (Denyer – Tranfield 2009) vagy metaanalízis (Borenstein et al. 2009)] szemben, hogy nem pontosan definiálható témában keres irodalmat, így a kereső algoritmus nem határozható meg könnyen előre és az összegzés szempontjai sem ismertek előzetesen. A *scoping review* Arksey és O'Malley (2005) megközelítése alapján hat fázisból áll: (1) a kutatási kérdés azonosítása, (2) a releváns tanulmányok összegyűjtése, (3) a tanulmányok szelekciója, (4) az adatok bemutatása, (5) az eredmények összegzése és végül (6) az eredmények visszacsatolása szakértőkhöz, felhasználókhoz. A továbbiakban ezen a hat lépésen megyünk végig, majd az utolsó, összegző részben rámutatunk a talált munkák közül olyanokra, amelyek NLP-módszerekkel még kiaknázatlan lehetőségeket rejtenek.

2. ábra. A scoping review módszere

Forrás: saját ábra

(1) A kutatási kérdés azonosítása: célunk nagy tömegű szövegek automatizált feldolgozására épülő adatfeldolgozási és adatelemzési módszerek korrupció-kutatásbeli alkalmazásának felderítése és automatizált módszerekkel még nem vizsgált területek megismerése.

Adatgyűjtés: (2) a releváns tanulmányok összegyűjtése, és (3) a tanulmányok szelekciója. A releváns tanulmányok összegyűjtéséhez a szabadon hozzáférhető *Publish or Perish* keresőt használtuk a Google Scholar adatbázisán. A Publish or Perish kezelőfelületén definiálhatjuk, hogy a szöveg mely részében keresünk, és a találatainkat egyszerűen exportálhatjuk táblázatos formába. A keresés több adatbázis mellett definiálható (Scopus, PubMed, Microsoft Academic Search stb.), azért döntöttünk a Google Scholar mellett, mert nem csupán cikkek, könyvek, de konferencia-kötetek és más „szürke" források is elérhetőek segítségével. Ez azért volt fontos számunkra, mert az NLP területén (akárcsak általában az informatika területén) született publikációk jó részét nem publikálják folyóiratokban (ezekre a tudományos szövegekre innentől kezdve „tanulmány"-ként hivatkozunk). Hátránya azonban a program használatának, hogy a Google Scholar limitációi miatt egy lekérésre csak a legtöbbet idézett 1000 tanulmányt jeleníti meg. Ez jelen kutatásban nem jelent gondot, mert a keresésünk kevesebb mint 150 találatot eredményezett.

Az adatgyűjtést megelőzően szakértői interjút vettünk fel Hajdu Miklós korrupciókutatóval és Léderer Sándorral, a korrupció visszaszorításáért küzdő K-Monitor társalapítójával. Kiválasztásuk oka az volt, hogy szerettük volna mind a tudományos, mind pedig a civil szférában dolgozó szakértők véleményét meghallgatni arról, hogy releváns-e a szakma oldaláról a kutatási kérdés, vagyis maguk az érintett kutatók mennyire látják elterjedtnek, illetve felhasználhatónak az NLP eszközeit a korrupciókutatásban. Mindketten jó lehetőségként tekintettek a módszerre, de csak néhány, a módszert alkalmazó publikációt ismertek. A szakértői interjúkat 2019 őszén készítettük. Az interjúk körülbelül másfél óra hosszúak voltak, a beszélgetésen mindkét szerző részt vett. Interjúalanyainktól kértünk kiinduló irodalmat is, ami alapján később a keresőkifejezéseinket kiválasztottuk.

A jelen *scoping review*-ba vont tanulmányok kiválasztása során a következő kritériumokat alkalmaztunk: 2000 után született, angol nyelvű tanulmányok, melyek tartalmilag a korrupció vizsgálatát (is) célozzák, és ebben a vizsgálatban alkalmazzanak szövegelemzési módszereket. Kiinduló elképzelésünk az volt, hogy a tanulmányok címét, absztraktját és a kulcsszavakat használnánk arra, hogy megtaláljuk a keresett tanulmányokat, ám a *Publish or Perish* csak arra ad lehetőséget, hogy a tanulmány címében vagy szövegében keressünk kulcsszavakra. Így előre valószínűsíthető volt, hogy olyan tanulmányok is letöltésre kerülnek, melyek nem a vizsgált módszert használják, csak megemlítik a kifejezést. A megkeresett szakértők által ajánlott kiinduló cikkeket áttekintve arra jutottunk, hogy a címben keressük a korrupció kifejezést, a tanulmány szövegében pedig a módszerre vonatkozó keresőszavakat. A *scoping review* jellegéből adódóan szándékosan nem specifikáltuk jobban a keresést, hogy ne maradjanak ki releváns találatok, ezért (a próba-keresések tapasztalatai alapján is) számítottunk rá, hogy irreleváns találataink is lesznek, például a *corruption* angol kifejezés többjelentésű volta miatt.

Mivel elsősorban a számítógéppel támogatott automatizált szövegelemzés volt a keresés fókuszában, a 2000 után született tanulmányokra szűkítettük le a keresést, mivel 2000 után terjedt el az automatizált szövegelemzés. A keresés során a címben a korrupció kifejezést (*corruption*), a tartalomban pedig a módszer kulcsszavait kerestük. A módszer kulcsszavai a következők: *text mining* vagy *automated text analysis*, vagy *text analysis*. Az általánosabb *„text analysis"* kifejezést azért tartottuk fontosnak, hogy ne maradjon ki releváns találat: attól tartottunk, hogy nem intézményesült még az NLP módszerére állandó megnevezés[4]. Ez az eljárás persze azzal járt, hogy nem csak automatizált, hanem hagyományos szövegelemzést alkalmazó tanulmányok is bekerültek a találatok közé, valamint olyan találatok is voltak, melyek csak említik a szövegelemzést, de nem használják. Keresésünk eredménye 131 tanulmány lett.

A REVIEW EREDMÉNYE

A keresési találatok bemutatása

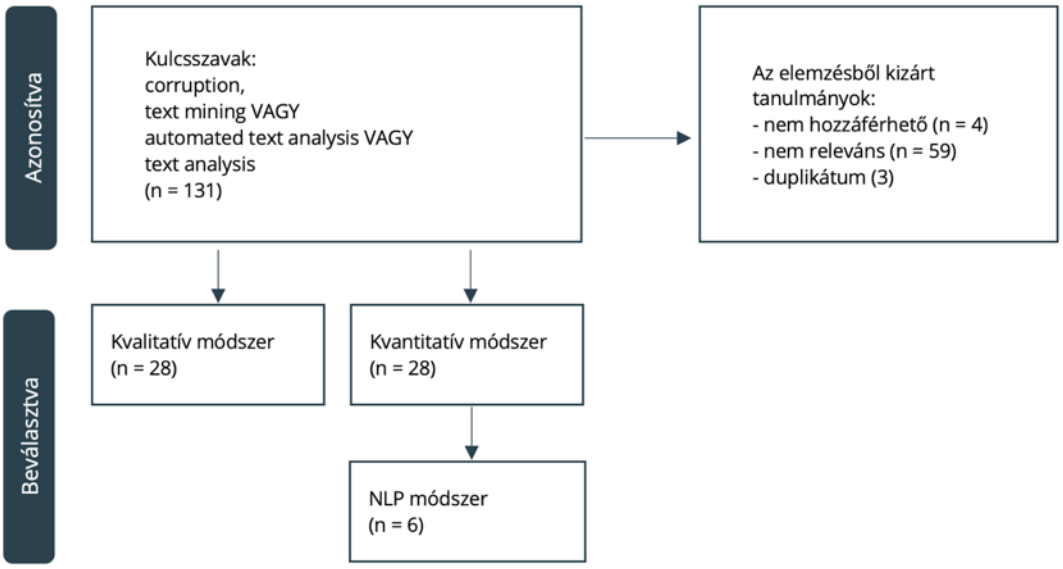
131 tanulmányt találtunk, melyek tehát a Google Scholar-on elérhetőek (3. ábra). A találatok között szakdolgozatok, doktori disszertációk is szerepelnek, ami a review minőségét nem rontja, hanem inkább javítja, hiszen az új módszerek kísérleti jelleggel gyakran először éppen ilyen munkákban jelennek meg és csak később kerülnek publikálásra. Jelen tanulmány irodalomjegyzékében terjedelemi okokból csak a reviewban idézett tanulmányok kerülnek hivatkozásra. A következő linken elérhető a találatok adatbázisa, az általunk fontosnak talált szempontok szerinti jellemzéssel együtt: https://tinyurl.com/yyahdxa5

A tanulmányok vizsgálata során kigyűjtöttük az absztraktot, ha elérhetőek voltak, a kulcsszavakat, az elemzés módszerét, az elemzés alapját nyújtó adatok típusát és a korrupció (bevezetőben bemutatott) mérési tipológiája szerinti besorolást is. Automatizált szövegelemzés esetén az adatok mennyiségét és az elemzés során használt eszközöket is feljegyeztük. Mindez elérhető a fenti linken.

Összesen 66 tanulmányt nem elemeztünk, mert vagy duplikátuma volt a talált tanulmányok egyikének (3 db), vagy nem tartozott az elemzésünk célcsoportjába (59 db). Az utóbbi, nem releváns tanulmányok jellemzően a *data corruption*, *information corruption* témában íródtak, vagy a *text analysis* szavakat tartalmazták

4 A fókusz megtartása céljából szándékosan nem vettük be viszont keresőszavaink közé azokat a klasszikus szövegelemzési módszereket, amik nem tartoznak az automatizált szövegelemzés vagy szövegbányászat területére, úgymint pl. diskurzuselemzés, tartalomelemzés, konverzációelemzés (lásd például Ignatow és Mihalcea 2016). Így találati listánk nem tartalmazza pl. a hazai korrupciókutatás egyik legfontosabb műhelyének, Tóth István Jánosnak, Szántó Zoltánnak és társaiknak kitűnő munkáit, melyek kvantitatív tartalomelemzésre épülnek, azaz előre definiált kódokat rendelve a szövegekhez vizsgálják azok struktúráit, de nem használnak szövegbányászati módszereket, lásd pl. Hajdu et al. (2016).

3. ábra. Az adatbázis bemutatása



Forrás: saját ábra

ugyan, de nem „szövegelemzés" jelentésű szókapcsolatként, más írások pedig angol absztrakttal szerepeltek, de nem angolul íródtak. A teljes gyűjteményből 4 olyan tanulmány volt, melynek teljes szövegéhez nem sikerült hozzáférnünk.

A 4. ábrán a keresési korpusszal való első ismerkedést célozva a tanulmányok absztraktjaiból készített szófelhő látható. A szófelhő elkészítése előtt az absztraktok szövegéből az NLP bevett szövegfeldolgozását követve (erről részletesen lásd: Ignatow és Mihalcea 2016) elhagytuk az angol nyelvre jellemző „túl" gyakori ún. stopszavakat, mint például a névelőket, személyes névmásokat, a megmaradó szavakat pedig egységesen kisbetűssé alakítottuk. Más szövegelőkészítési módszert (például a toldalékok levágását) nem használtunk.

4. ábra. Az absztraktokból generált szófelhő

Forrás: saját ábra

Az ábrán az 50 leggyakoribb kifejezés látható. Egy részük a kutatás egyszerű leírást meghaladó céljára utal *(„anti-corruption"),* más részük a módszertanra és adatforrásra („*case*", „*news*", „*media*") vagy az érintett szervezetekre *(„party", „companies", „government", „institutions", „organizations").* Feltűnő az „*international*" és „*countries*" kiemelt gyakorisága, tehát a téma feldolgozása sok esetben implikál országközi összevetéseket. Ami ezen kívül talán a legszembetűnőbb, az az „*Indonesia*" kifejezés megjelenése. A tanulmányokat áttekintve kiderül, hogy valóban kiugró mennyiségű elemzés foglalkozik az indonéz korrupciós helyzettel, ennek oka fel-

tételezésünk szerint az ide tartozó írások legtöbbjén affiliációként szereplő Bina Nusantara University kutatási profilja lehet; e tanulmányok keletkezését tekintve azt mondhatjuk, hogy nincsen trendje az előfordulásuknak.

A tanulmányokban használt adatforrások

A tanulmányok adatforrása elsősorban a szerkesztett és a közösségi média, kisebb részben a politikai nyilvánosság hivatalos szövegei – például törvényszövegek, parlamenti felszólalások (1. tábla). Ahadiat (2019) tanulmánya azért különösen izgalmas, mert elrugaszkodva a hagyományos adatforrásoktól, kortárs szépirodalmi novellákban vizsgálja a korrupció keretezését tartalomelemzés segítségével.

1. tábla. A tanulmányok adatforrásai

Adatforrás	Tanulmányok száma
média	27
közösségi média	5
jogi szöveg / törvény / elnöki beszéd / parlamenti felszólalás / könyvvizsgálói jelentés	11
közbeszerzés pályázat	14
egyéb (például ismeretterjesztő brossúra, interjú, szépirodalom)	8

Az adatforrások használatának alábbi bemutatása során egy-egy példát hozunk a politikai nyilvánosság különböző szintjeiről: a hivatalos politika, a média és a laikus nyilvánosság szintjéről. E szintek nyilván egymással interakcióban állnak, ami a korrupciós diskurzus tematizáltságára is hatással van. Az egyre növekvő digitalizációnak köszönhetően a nyilvános szöveges tartalmakat már nem csak az elit hozza létre, hiszen az interneten tulajdonképpen bárki megnyilvánulhat – ezáltal jobban hozzáférhetünk a laikus nyilvánosság szintjéhez is. A korrupció megértéséhez mindhárom szintet érdemes vizsgálatba vonni: erre mutatunk az alábbiakban egy-egy megközelítést. Alábbi példáinkat a kvalitatív megközelítések közül válogatjuk, mert a kvantitatív tanulmányokra, fókuszunkat követve, később jóval részletesebben kitérünk.

Az általunk gyűjtött kvalitatív szövegelemzések egyik fő eszköze Teun A. van Dijk (1985) által kezdeményezett kritikai diskurzuselemzés. E megközelítés annak megragadására alkalmas, hogy milyen nyelvi megnyilvánulásokat alkalmaznak a hatalom képviselői a saját érdekeiknek megfelelő ideológia terjesztésére (lásd például Fairclough 1992).

Hivatalos szint: elnöki beszéd

Ogunmuyiwa (2015, 2019) kritikai diskurzuselemzést is alkalmazó vizsgálatának tárgya, hogy hogyan alakult a korrupcióról szóló diskurzus az 1957–2015 közötti nigériai elnöki beszédekben és hogyan konstruálódik a beszélők korrupcióval kapcsolatos pozíciója (elkötelezettségük a korrupció ellen folytatott küzdelemben és elhatárolódásuk a korrupciós érintettségtől).

A (hagyományos) média szintje: napilapok

Touwe és Sultan (2015) célja annak kiderítése volt kritikai diskurzuselemzés alkalmazásával, hogy az indonéz Tempo hetilap hogyan számolt be egy kiválasztott korrupciós ügyről 2012 márciusa és 2013 júliusa között, és ezáltal hogyan formálta a közvélemény alakulását.

Laikus szint: közösségi média, Twitter

Mear (2016) diskurzuselemzéssel vizsgálta az *Anti Corruption International* nemzetközi korrupcióellenes civil szervezet Twitter bejegyzéseit, arra fókuszálva, hogy a szervezet jellemzően hogyan használja a „korrupció" kifejezést a közösségi médiában. Tanulmányunk végén a különböző kvalitatív módszerek automatizálhatóságával kapcsolatban erre a tanulmányra még visszatérünk.

A tanulmányokban alkalmazott korrupciómérési mód

Az általunk talált kutatások mindegyike beavatkozásmentes vizsgálat: kézenfekvő okból, hiszen nem survey-n alapulnak, hanem „*found data*" (talált adatok), azaz jellemzően valamely más célra készült dokumentumok vizsgálatán. Korrupciómérési módjuk szerint klasszifikálva őket, többen foglalkoznak korrupciós észlelések mérésével, például Hlatshwayo et al (2018) a szerkesztett média korrupció-reprezentációja alapján. Megjelenik a tanulmányok között a korrupció jelenségéhez kapcsolódó attitűdök mérése is. Bár a szokásos megközelítés itt, ahogy a Bevezetőben említettük, a survey-alapú módszer, az általunk gyűjtött, szövegelemzést használó kutatások – adatforrásukból adódóan – nem survey-alapon, hanem például közösségi média vizsgálat alapján elemzik a korrupcióval kapcsolatos társadalmi reakciókat (lásd például Niklander et al. 2016 és Li et al. 2019).

Több tanulmány közöl mérést a korrupció volumenére vonatkozóan, objektív adatokból kiindulva: a már idézett Weaver (2020) a csúszópénzzel befolyásolt közalkalmazotti állásbetöltéseket vizsgálja az álláspályázatok dokumentációjára építve. Hasonló a megközelítése azoknak a kutatásoknak, amelyek a korrupciós kockázat volumenét mérik objektív adatokból kiindulva: például Fazekas és Tóth (2016) vagy Rabuzin és Modrusan (2019) a korrupt közbeszerzések lebonyolítását elősegítő körülmények fennállását vizsgálják.

A tanulmányokban felhasznált szoftverek

Az Atlas és a WordSmith kvalitatív tartalomelemzést támogató szoftver elsősorban kulcsszókeresésre nyújt megoldást. A tanulmányok, melyek Atlast használnak, nem lépnek túl a szógyakoriságok, a szavak együttes megjelenésének elemzésén. Egy másik, több tanulmányban is használt eszköz a T-Lab, egy szövegfeldolgozást tekintve fejlettebb, nyelvi és statisztikai eszközöket egyaránt alkalmazó szoftver. Azon túl, hogy a segítségével az alapvető szöveg-előfeldolgozást is végre lehet hajtani (szövegszegmentáció, szótövezés, irreleváns szavak eltávolítása – erről bővebben lásd Németh et al. 2020), alkalmazható együttes szó-előfordulás elemzésére, kulcsszavak mintázatának feltérképezésére, a korpusz részhalmazainak összehasonlító elemzésére. Az utóbbi években a komplexebb elemzéseket (lásd például a később ismertetett topikmodellezést) lehetővé tevő, nyílt forráskódú, szabadon felhasználható, ugyanakkor programozási tudást igénylő programok, mint a Python is megjelenik a használt eszközök között.

A tanulmányokban használt elemzési módszerek

A releváns 65 tanulmányból 28 használ kizárólag kvalitatív megközelítést, a többi kevert vagy tisztán kvantitatív módszert alkalmaz. A kvalitatív megközelítésű tanulmányok nagy része diskurzuselemzést használ: sokan közülük a már említett kritikai diskurzuselemzést.

A kvalitatív megközelítésű írások többsége esetén, nyilvánvaló módon, az automatizálhatóság lehetősége fel sem merül, hiszen ezek jellemzően néhány konkrét szövegre koncentráló, mélyebb esettanulmányok. Cikkünk végén két olyan tanulmányra térünk vissza, ahol a szerző olyan elemzési módszert használ, ami támogatható lenne automatizált szövegelemzéssel.

37 tanulmány használ valamilyen formában kvantitatív megközelítést. Az adatforrások tekintetében a média és a közösségi média, a parlamenti felszólalások elemzése a kvalitatív tanulmányokhoz hasonlóan itt is megjelenik, ám a lista kiegészül a közbeszerzési pályázatok vizsgálatával. A kvantitatív módszert használó cikkeket részletesebben is tárgyaljuk. Az NLP-re támaszkodó tanulmányokra koncentrálunk, de előbb az *IMF Working Paper*-jét (Hlatshwayo et al., 2018) emelnénk ki, annak szisztematikus, átfogó jellege, és az IMF mint szereplő fontossága miatt. Az írás egy, 665 millió sajtócikket tartalmazó

nemzetközi gyűjteményből indul ki és a *big data* megközelítést követő *„cross-country news flow indices of corruption"*[5] (NIC) indexre támaszkodva jellemzi a korrupció mértékét, dinamikáját és országközi eltéréseit. A NIC mutató a korrupció média-reprezentációjából indul ki, ezzel (ahogy a Bevezetőben a harmadik generációs mérőszámokról írtuk) célja a percepcióalapú indexek hátrányainak kiküszöbölése. A korrupciós index kiszámítását 30 országra vonatkozó, specifikus keresési algoritmusok szolgálják, melyek segítségével azonosítják azokat az utóbbi évtizedekben született cikkeket, melyek adott országokban a közhivatallal való magáncélú visszaéléssel kapcsolatosak. Ennek operacionalizálásához azokat a cikkeket gyűjtik össze, amelyekben 1) az ország neve legfeljebb 8 szó távolságra helyezkedik el a korrupcióhoz kapcsolódó, előre definiált kifejezések (például *„corrupt*", „kleptoc*", „nepotism", „favoritism", „rent-seeking", „bribe*", „graft"*) valamelyikétől; 2) említik a kormányzati vonatkozású kulcsszavak (például *„government", „regime", „authorities", „public sector", „bureaucra*", „agenc*"*) valamelyikét; és 3) amelyekben nincs szó korrupcióra adott reakcióról, korrupcióellenes próbálkozásról (mert a cél a korrupciós esemény, nem pedig az antikorrupciós válasz felismerése). Az index kiszámításán túl az IMF gyűjteménye izgalmas alapja lehetne mélyebb szöveganalitikai kutatásoknak is, hiszen a korpusz széles fókuszú mind időben, mind térben – ezekre a lehetőségekre és azok korlátaira az összegzésünkben még visszatérünk.

Gyűjtésünkben kiemelt a reprezentációja (egytucat tanulmánnyal) Fazekas Mihálynak és kutatócsoportjának, akik bár az adatbázis méretét tekintve igazi *big data* megközelítést alkalmaznak, módszerük inkább a tradicionális kvantitatív eszközökhöz sorolható. Fazekas és munkatársai (2016) a hazai közbeszerzési pályázatok anyagain elemezték az eljárás korrupciós kockázatait az általuk létrehozott Magyar Közbeszerzési Adatbázist (MaKAB) használva, mely a hazai közbeszerzési eljárások adatait tartalmazza (Fazekas–Tóth 2012). Ez az adatbázis az alapja egy általuk definiált korrupciókockázat-mérő indexnek *(Corruption Risk Index*, CRI – lásd: Fazekas et al. 2016). A CRI utólagos értékelésre alkalmazható, és korrupciós kockázatot mér, nem a ténylegesen korrupciót detektálja. A CRI-t képző változók a teljes közbeszerzési folyamatot lefedik: a jelentkezési (például: egy ajánlattevős-e a szerződés), az értékelési (például: az ár helyett más kritériumok figyelembevétele) és a megvalósítási fázisra (például: történt-e szerződésmódosítás) vonatkozó információkat is figyelembe veszik. Ezek tehát a pályázatokon alapuló metaadatok, de a pályázatok szövegei nem kerülnek közvetlenül elemzésre.

A tanulmányunk fókuszában álló megközelítésre térve: a kvantitatív jellegű tanulmányok közül csupán 6 használ NLP módszereket. A következőkben e tanulmányok céljait és megoldásait részletezzük. Mivel e cikkek jellemzően felügyelt klasszifikációs modelleket és topikmodellt használnak, itt röviden összefoglaljuk e módszerek célját és megközelítését – ennél részletesebb és kifejezetten a társadalomkutatási felhasználásra vonatkozó ismertetőt Németh et al. (2020) közöl.

A gépi tanulás körébe tartozó *felügyelt klasszifikációs modellek* esetén már kiindulásnál rendelkezünk egy felcímkézett szöveghalmazzal. Például újságcikkeket kétféle címkével láttak el kódolóink (NLP-s szóhasználatban „annotátoraink"): korrupcióval foglalkozik vagy nem. A cél az, hogy az alkalmazott algoritmust (ez többféle lehet, lásd később) olyan szövegmintázatok felismerésére tanítsuk, amelyek a címkék szövegekhez rendelését automatizálhatják. Ha az automatizálás jó hatásfokú, akkor új, felcímkézetlen szövegeket is be tudunk illeszteni a kategóriarendszerbe. Elsődleges fontosságú szempont, hogy értékelni tudjuk a címkézés hatékonyságát azért, hogy információnk legyen a várható jövőbeni teljesítményről, és hogy több algoritmust használva kiválaszthassuk a legjobb modellt. Az alkalmazott algoritmus lehet a klasszikus társadalomkutatásban ismert valamely

5 Azaz a korrupció híráramlásra alapozott, országközi összevetést lehetővé tevő indexe. Az elnevezés utalhat arra a *„global news flow"* elnevezésű tudományterületre, mely bizonyos jelenségek különböző országokban tapasztalható média-reprezentációját vizsgálja.

módszer, például regresszió, vagy az NLP-ben gyakran használt, de a klasszikus módszertanban nem szereplő adattudományi algoritmusok, mint a *Random Forest, Naive Bayes, Support Vector Machine* (SVM) vagy különböző neurális hálók is (e modellekről bővebben lásd: Ignatow és Mihalcea 2016).

A gyűjtésünk során talált NLP-cikkek közül Noerlina et al. (2016) célja egy korrupciós cikkadatbázis létrehozása, így a szerzők itt technikai céllal használnak szövegbányászati módszereket. Azzal a céllal, hogy az indonéz médiában megjelenő korrupciós eseteket feltérképezzék, különböző híroldalakról gyűjtött ~400.000, vegyes témájú cikket soroltak be *Naive Bayes*-osztályozó segítségével két kategóriába aszerint, hogy melyek foglalkoznak korrupcióval, és melyek nem. A munka az adatgyűjtés kritériumait és módszerét, valamint az adattisztítás lépéseit nagyon alaposan körüljárja, ám az ennél sokkal fontosabb, a felügyelt klasszifikációhoz szükséges annotálás, tehát a korrupciós/nem korrupciós címkével való ellátás módszertanáról nem írnak. Egy másik kritikai észrevételt téve: a tanulmány szerint az elemzők 30-30 dokumentumot annotáltak, és ez alapján alkottak algoritmust a nem-annotált szövegek kategorizálására. Tekintve, hogy közel 400.000 cikk került legyűjtésre, az annotált halmaz nagyon kicsi, kézenfekvő módon a mintavételi hiba miatt várhatóan nem reprezentálja jól a sokaságot, így nem épülhet rá pontos algoritmus. Módszertanilag korrekt kutatások sokkal nagyobb, több ezres/tízezres nagyságrendű annotált adatbázist szoktak alkalmazni, lásd például Bosco et al (2015) automatikus irónia-detektálást célzó vizsgálatát, ahol 3300 politikai témájú tweet-et annotáltak aszerint, hogy tartalmaz-e iróniát.

Az általunk kiválasztott tanulmányok közül közbeszerzési pályázatok elemzésével Fazekas és társain kívül mások is foglalkoztak. Rabuzin és Modrusan (2019) célja az volt, hogy szövegbányászati technikákat alkalmazva korrupciógyanús tendereket azonosítsanak. Elemzésük Horvátországra fókuszált. A modellük nem hagyományos értelemben vett predikciós modell (ők is inkább magyarázó modellként utalnak rá), hiszen lezárult közbeszerzési pályázatok szövegezéséből indul ki, tehát ez esetben már ismert, hogy hány ajánlattevő volt. Az általuk alkalmazott modell célja annak megjóslása volt, hogy az adott pályázat „single bidder" azaz egy ajánlattevős-e (ezzel definiálva a korrupciós kockázatot). A tanulmány a leggyakoribb klasszifikációs modelleket használja: *Naiv Bayes*-t, logisztikus regressziót és *Support Vector Machines* modelleket. A kutatás során a szerzőpáros a teljes közbeszerzési pályázatnak csak egy speciális részére fókuszált: a technikai és szakmai alkalmassági követelményeket megfogalmazó részre, egészen pontosan a „technikai alkalmasság" kifejezést követő 1000 szóra. A kutatás célja tehát az volt, hogy a modell eldöntse a kiírásban megfogalmazott alkalmassági feltételek alapján, hogy feltehetően egyetlen ajánlattevő lesz-e. Összegzésük szerint a módszer használható a korrupciógyanús, „*single bidder*" pályázatok felderítésében, ám hozzá kell tenni, hogy a modellek előrejelző teljesítménye nem kiemelkedően jó. A tenderek kicsit több, mint felében sikerült helyesen besorolniuk az ismeretlen eseteket egy-, vagy többajánlattevős kategóriába. A modell teljesítménye akkor sem javult sokat, amikor a teljes adatbázis helyett a különböző (informatikai, egészségügyi, építőipari stb.) szegmensekre szétbontva alkalmazták a modellt.

A fenti tanulmányok mindegyike felügyelt klasszifikációt használt; a következő tanulmányokban *felügyelet nélküli modell* alkalmazására kerül sor. Láthattuk, hogy eddig előre definiált információkat „tanítunk meg" az algoritmusunknak, ezzel szemben a felügyelet nélküli módszerek esetén nincsen előzetes ismeretünk, nincsenek például "*singe bidder*" címkével ellátott eseteink. Klasszikus, a társadalomkutatók számára is ismert felügyelet nélküli módszer a klaszterelemzés, de a topikmodellezés is ilyen.

A *topikmodellek* (Blei és Lafferty 2009) olyan automatizált eljárások, melyek célja szöveghalmazok látens témáinak azonosítása. A szövegek néhány topik keverékeként azonosíthatók, például egy, a riói olimpia stadionépítéseiről szóló cikk 80%-ban gazdasági, míg 20%-ban sport-témát dolgozhat fel. A topikok száma és tartalma

a priori nem ismert, tehát ez a felügyelt klasszifikációval szemben nem-felügyelt módszer. Az elemzők feladata a topikok optimális számának meghatározása és a topikok interpretációja.

Pan és Chen (2018) a topikmodellek egyik altípusát, a strukturális topikmodellt használták arra, hogy megmutassák: Kínában a korrupcióval kapcsolatos panaszok nem jutnak el a felső szintű hatóságokhoz. Az elemzéshez a Kínai Kommunista Párt „J" Prefektúra Propaganda Osztályának kiszivárgott e-mailjeit elemzik, melyek online elérhetők. A gyűjteményben a 2012 és 2014 közötti időszakra vonatkozóan több mint 3000 panaszt azonosítottak. Különbséget tudtak tenni azok között a bejelentések között, amelyeket továbbítottak a tartomány vezetőihez, és amelyeket visszatartottak. A topikmodellezés legegyszerűbb típusa e panaszok látens témáit segít felderíteni anélkül, hogy előzetes ismerettel kellene rendelkezni arról, hogy a korpusz milyen tartalmak köré szerveződik. Az itt alkalmazott strukturális topikmodellbe már metaadatok is bevonhatók (például bejelentették-e a panaszt). Az elemzés során a kutatók topikmodellezés segítségével vizsgálták, hogy a panaszok milyen témában születtek, és azt, hogy a különböző témák (topikok) aránya és a témát jellemző szavak eloszlása befolyásolja-e, hogy az adott panasz bejelentésre kerül-e vagy sem. Azt találták, hogy a vezetőségi jogsértésekre vonatkozó panaszokat (mint például a sikkasztás) és egyéb korrupció-tematikájú panaszokat kisebb valószínűséggel jelentették, mint azokat, melyek az oktatás minőségére vagy a környezetszennyezésre vonatkoztak.

Muço (2019) hipotézise szerint a helyi politikusok korrupciós ügyeiről szóló információk nyilvánosságra kerülése kihat a pártjukra is. Az elemzés során brazil helyi önkormányzatok választási auditjait elemezte, melyek 2000-ben kerültek felvételre. Ezek vizsgálata során szintén nem-felügyelt eszközhöz nyúlt: főkomponens-elemzés (PCA) felhasználásával hozott létre egy olyan mutatót, amely az önkormányzatokat a legkevésbé korrupttól a legkorruptabbig sorolta. Először egy előre definiált szólista segítségével szabálytalanságra utaló kifejezéseket *(„fraud", „collusion", „procurement simulation")* keresett a szövegben, illetve megszámolta az általa súlyosként definiált szabálytalanságok előfordulását. Emellett a dokumentum hosszát és a képek számát is vizsgálta. Eredményei többek között azt mutatják, hogy a választópolgárok csak akkor veszik figyelembe az audit során nyilvánosságra kerülő problémákat, ha az elér egy bizonyos küszöböt.

A talált NLP-s cikkek között több olyan volt, ami nem annyira tartalmi, mint inkább technikai célra alkalmazta a módszert. Li et al. (2019) célja például egy olyan módszer kialakítása, ami felismeri a korrupcióhoz kapcsolódó tweeteket; ezáltal egy közösségi média alapú „*surveillance*" (felügyelő és jelentő) eszköz kialakítása. Az eszköz célzott beavatkozások (mint hatóságok számára történő jelentések) kidolgozására is felhasználható. A szerzők Twitterről gyűjtöttek adatokat olyan bejegyzéseket keresve, melyek korrupcióval kapcsolatos kulcsszavakat tartalmaztak. Ezután topikmodellt használtak (tulajdonképpen az adatok tisztításához) oly módon, hogy e modell segítségével szűrték ki az általuk zajként értelmezett, irreleváns tweeteket. Az adattisztítást követően az adatbázis egy kisebb részén manuális kódolással tárták fel a tweetek tematikáját egy tanuló adathalmazon (olyan címkéket használva, mint rendőrségi vagy egészségügyi korrupció) majd felügyelt klasszifikáció *(support vector machine, SVM)* segítségével címkézték fel a manuálisan be nem sorolt tweeteket.

Niklander et al. (2016) tanulmánya a fentiek zömével szemben nem valamilyen objektív mutató elkészítését célozza, hanem a korrupcióval kapcsolatos társadalmi attitűdökre koncentrál. A kutatás az érzelmi töltet automatizált kinyerésére alkalmas szentiment elemzéssel (erről lásd: Németh et al. 2020) vizsgálja Twitter bejegyzéseken, hogy a felhasználók hogyan reagálnak a különböző korrupciós témákra. A szentiment- és az emócióelemzés az üzleti szféra talán legnépszerűbb NLP-módszere: használják például a közösségi média hozzászólásaiban adott, a céggel vagy a cég bizonyos termékeivel, mozifilmekkel stb. kapcsolatos vélemények azonosítására. Az utóbbi időben megjelentek társadalomtudományi alkalmazások is. A szerzők két konkrét chilei

korrupciós eset kapcsán arra jutottak, hogy a reakciók nem függenek össze szorosan a hírekkel, s valódi vita helyett csak az ellenkező politikai oldal kritizálása figyelhető meg.

Kínálkozó NLP-alternatívák

Az alábbiakban néhány lehetőséget villantunk fel arra vonatkozóan, hogy milyen automatizált szövegelemzést használó alternatíva adódhat a vizsgált tanulmányok módszerére vonatkozóan. Ez nyilván nem minden kutatásra áll, sok esetben a kvalitatív módszer a megfelelő, egy esettanulmány vagy néhány eset alapos körbejárása nem lehetséges automatikus módszerekkel.

Josefsson (2014) áttekintést ad az 1994 és 2013 között született, olasz korrupcióellenes pártok választási programjai alapján azok fő témáiról és irányelveiről. A szerző manuálisan dolgozza fel a szövegeket: 9 előre definiált témát keres azokban, és súlyozza előfordulásukat aszerint, hogy milyen hangsúlyt kap az adott téma. Egy alternatív, nem-felügyelt, tehát a tematikus struktúrát induktív módon feltáró módszer lehetne ugyanerre a feladatra a topikmodell egy variánsának, az időbeli tematikus trendek vizsgálatára alkalmas dinamikus topikmodellnek a használata (lásd Katona et al. 2021). Előnye ennek a módszernek az induktív megközelítés mellett az, hogy algoritmikus módon tárja fel a dokumentumok látens témáit, így sokkal konzisztensebb módon jár el, melynek révén elkerülhetővé válna a kutatói szubjektivitás hatása. Másrészt az idő dimenziójának bevonásával lehetővé válna a topikok súlyának és tartalmának időbeli változásának (automatizált) detektálása is, mely az eredeti tanulmányban nem jelenik meg. Mivel a topikmodellezés nem címkézi automatikusan az egyes topikokat, a topikok tartalmi azonosítása a kutató feladata marad, de a módszer a topikokat jellemző kifejezések és reprezentáns szövegek megadásával mindenképpen megkönnyíti e feladatot.

Az általunk talált tanulmányok az NLP eszköztárának csak igen szűk szeletét használták, nem alkalmaztak például szóbeágyazási modellt *(word embedding model* – erről részletesen lásd: Németh et al. 2020), ami pedig a szavak jelentésének vizsgálhatóságát adná a kutató kezébe, nagyon inspiratív lehetőségekkel. A modell a vizsgált korpusz látens szemantikai struktúráját reprezentálja. Leegyszerűsítve: a korpuszunk szavait egy térben jeleníti meg, ahol a szavak elhelyezkedését jelentésük határozza meg. A szójelentést itt a szavak mondatbeli előfordulásának szűk környezete határozza meg és aszerint kerül közel vagy távol egymástól két szó a vektortérben, hogy mennyire egyezik meg ez a környezet a korpuszban.

A szóbeágyazási modell sokat hozzá tehetne például Mear (2016) elemzéséhez, aki az *Anti Corruption International* nevű, nemzetközi, korrupcióellenes NGO tweetjeit elemzi diskurzuselemzés segítségével. A szerző a „korrupció” szóval gyakran együtt előforduló kifejezéseket keresi, és ebből von le következtetést arra, hogy a szervezet jellemzően hogyan használja a kifejezést a közösségi médiában. Szóbeágyazást alkalmazva ugyanerre a kérdésre azonosíthatók lennének a “korrupció” kifejezés szinonimái is, és mivel a módszer nem-felügyelt algoritmus, így objektívebb is, hiszen nem mi választjuk ki, hogy melyek a korrupció kapcsán releváns (annak használati környezetébe tartozó) kifejezések, hanem ez automatikusan történik. A szóbeágyazási modellek segítségével többet megtudhatnánk az adott szervezet fogalom- és szóhasználatáról, megadhatnánk a „korrupció”-hoz (a szervezet nyelvhasználata alapján) jelentésben legközelebb eső kifejezéseket és vizsgálhatnánk azok klaszteresedését is. Még tartalmasabb lehetne e nyelvhasználaton alapuló elemzés, ha ezen felül összevetést adna más területen dolgozó, illetve más megközelítést követő szervezetek tweetjeinek vizsgálatával – az automatizált vizsgálat alkalmazásával e kiterjesztés nem igényelne nagy kapacitástöbbletet.

Kézenfekvő NLP-alapú kiegészítés kínálkozik Fazekas és Tóth (2016) kutatásaira is. Mint említettük, ők szöveges adatbázison (közbeszerzési kiírások korpuszán) dolgoztak, de klasszikus módszert követtek: a kiírá-

sok néhány előre definiált jegye alapján igyekeztek a korrupciós kockázatot viselő kiírásokat azonosítani. Jelen tanulmány első szerzője ez évben kezdi meg Fazekassal együttműködve a kutatás NLP-alapú kiegészítését. A közösen készítendő tanulmány egyik célja éppen annak vizsgálata lesz, hogy a korábbi modellek teljesítménye javítható-e a szöveges információ automatizált kinyerésével és annak a korrupciós kockázat előrejelzésébe történő bevonásával. Kísérletet teszünk továbbá annak a kérdésnek a megválaszolására, hogy csupán a pályázat szövegéből kiindulva lehetséges-e „megjósolni" azt, hogy mely pályázó nyeri el az adott közbeszerzést.

Az IMF „*news-flow index*"-éhez összeállított szöveges adatbázis értékes gyűjtemény, mely számos elemzési lehetőséget rejt magában. Mivel hosszú időt ölel fel (1995-2017), így a korrupció kifejezés változásának, a korrupció tematizáltságának dinamikája is vizsgálható lenne rajta. Mivel 30 ország hírforrásait tartalmazza, így nemzetközi összehasonlítás alapjául is szolgálhat: feltárhatók lennének a korrupciós szó használatának, jelentésének kulturális aspektusai és az országok közötti különbségek is. Egy másik, a big data kutatásokkal kapcsolatban általában is fontos észrevételünk, hogy bár a tanulmány mintát használ, nem egy teljes populációt, ennek ellenére a szerzők gyakran nem fogalmazzák meg expliciten, hogy a cikkek összegyűjtéséhez kiválasztott hírportálok mit kívánnak reprezentálni, nem tudjuk, hogy mire általánosíthatók az eredmények. A *survey* típusú felmérésekkel szemben, ahol a kutatás célpopulációja és reprezentativitása világosan értelmezhető, a *big data* típusú kutatásoknál ez nem egyértelmű (a *big data* kutatások reprezentativitásáról lásd bővebben: Németh 2015).

Másik kritikai észrevételünk a nagy társadalomtudományos potenciállal bíró felügyelt klasszifikációs algoritmusokat alkalmazó kutatások (lásd például: Noerlina et al. 2016) annotálására vonatkozott. Nem írnak róla részletesen, hogy kik (szakértők? laikusok?), milyen korrupció-definíciós instrukciókat követve és milyen szervezésben (egyetlen annotátor? kettős független annotálás?) annotáltak. Pedig a társadalomkutatási alkalmazásoknál éppen ezeknek a döntéseknek van kiemelt szerepe, hiszen a besorolás itt az üzleti és műszaki alkalmazásoknál sokkal komplexebb interpretációs feladatot jelent (egy bővebb reflexióként erre a problémára lásd: Németh et al. 2020).

Végül, hiányoltuk a talált tanulmányokban a módszertani rugalmasságot. Egyetlen tanulmány sem alkalmazott valódi kevert (azaz kvalitatív és kvantitatív elemeket egyaránt alkalmazó) megközelítést, ami pedig gyümölcsöző lenne. Egyrészt azért, mert a tartalmi háttértudás igazán így lenne beépíthető az automatizált elemzésekbe, azok már az induláskor, azaz az adatgyűjtés, korpuszmeghatározás, előfeldolgozás, operacionalizáció során „közelebb" kerülnének a szövegekhez. Másrészt a modellezés eredményeinek interpretációját is nagyban támogathatná a kvalitatív megközelítés: a szövegeknek a modellel támogatott, célzott módon kiválasztott részmintáinak humán – azaz nem automatizált? – feldolgozása. Hasonlóan messzi távlatokat nyithatna a különböző adatforrások vagy különböző korrupció-koncepciók kombinálása.

Összegzés

Tanulmányunkban a 2000 után született, nagy tömegű szövegek automatizált feldolgozására épülő korrupciókutatások összegyűjtésére és összegzésére törekedtünk, a természetes nyelvfeldolgozás alkalmazási elterjedtségére, illetve lehetőségeire fókuszálva. A *scoping review* módszerét alkalmaztuk, mivel diffúz, nem pontosan definiálható témában kerestünk irodalmat, így a kereső algoritmus és az összegzés szempontjai sem voltak ismertek előzetesen.

Kutatásunk limitációi közé tartozik, hogy keresési kulcsszavaink nem feltétlenül fedik le a kutatni kívánt terület egészét. Ennek oka, hogy még nincsenek terminus technicusok intézményesülve ezen a módszertani területen, a nyelvészet, statisztika, informatika, mesterséges intelligencia és kognitív tudomány határmezsgyéin (a magyar szóhasználat is cseppfolyós: szövegbányászat, automatizált/automatikus/számítógépes szöveganalitika/szövegelemzés, természetes nyelvfeldolgozás, számítógépes nyelvészet stb.). Emellett az, hogy csupán angol nyelvű találatokat elemeztünk, szintén korlátozta kutatásunkat, bár feltehető, hogy e témához tartozó fontosabb tudományos eredmények zöme megjelenik angol nyelvű nemzetközi tudományos platformon is.

Lényeges eltéréseket találtunk a felhasznált szöveges adatforrást, a korrupció-mérési módot és az elemzési megközelítést tekintve, ugyanakkor sajnálatosan kevés (adatforrását, módszerét vagy korrupció-mérési módját tekintve) kevert típusú tanulmányt találtunk. A klasszikus, a korrupció volumenét, vagy a vele kapcsolatos attitűdöt / percepciót leíró ill. észlelésének következményeit vizsgáló (Muço, 2019) munkákon kívül találtunk a korrupció megelőzésére felhasználható eredményeket (Fazekas és Tóth 2016, Pan és Chen 2018), sőt intervencióra közvetlenül alkalmasakat is (lásd: Li et al. 2019 *surveillance*-rendszerét). Az NLP-t csupán néhány tanulmány használta, és ezek egy része sem annyira tartalmi, mint csak technikai feladatra (például a korrupciós tematikájú cikkek azonosítására). Látható: az NLP nem nagyon elterjedt még ezen a területen. Ugyanakkor azt is igyekeztünk bemutatni, hogy gyümölcsöző lehet a használata, mutattunk példákat arra, amikor alternatív eszközként jól támogathatná a meglevő kutatást. Reméljük, hogy munkánkkal inspiráltuk a területen folyó kutatást, és hogy más társadalomkutatási területekre is átvihetők eredményeink.

Hivatkozások

Ahadiat, E. (2019) Corruption and Abuse of Power: A Reflection of Social Issues in Short Stories. *Social Sciences on Sustainable Development for World Challenge: The First Economics, Law, Education and Humanities International Conference.* https://doi.org/10.18502/kss.v3i14.4311

Aggarwal, Z. (2012) *Mining Text Data*. New York: Springer.

Aminudin, A. (2018) Discourse of Corruption in Case of Corruption Setya Novanto in Tempo Magazine November 2017 Edition (Discourse Analysis Teun Van Dijk). Conference paper presented at the 2nd International Symposium on Social Science, Arts and Humanities (SYSSARM 2018) Conference in Krabi, Thailand. Elérhető: https://www.researchgate.net/publication/324949870_Discourse_of_Corruption_in_Case_of_Corruption_Setya_Novanto_In_Tempo_Magazine_November_2017_Edition_Discourse_Analysis_Teun_Van_Dijk [Letöltve: 2021-04-20]

Arksey, H. – O'Malley, L. (2005) Scoping studies: towards a methodological framework. International. *Journal of Social Research Methodology: Theory and Practice,* 8(1),19–32. https://doi.org/10.1080/1364557032000119616

Axelsson, S. – Dahlberg, S. (2018) *Corruption Talk: Mapping the Word Corruption in Online Text Data Across the World.* General Conference of the European Consortium for Political Research (ECPR), Conference Paper. Elérhető: https://ecpr.eu/Filestore/paperproposal/c3de1fed-0189-4fa6-bfaf-bdde065a91de.pdf [Letöltve: 2021-04-20]

Blei, M. D. – Lafferty D. J. (2009) Topic models. Elérhető: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf [Letöltve: 2021-04-20]

Bosco, C. – Patti, V. – Bolioli, A. (2015) Developing corpora for sentiment analysis: the case of irony and senti-TUT. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15),* 4158–4162. AAAI Press. Elérhető: https://www.ijcai.org/Proceedings/15/Papers/587.pdf [Letöltve: 2021-04-20]

Borenstein, M. – Hedges, L. V. – Higgins, J. P. T. – Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester, U.K: John Wiley & Sons.

Bratu, R. – Kažoka, I. (2018) Metaphors of Corruption in the News Media Coverage of Seven European Countries. *European Journal of Communication,* 33(1), 57–72. https://doi.org/10.1177/0267323117750695

Denyer, D. – Tranfield, D. (2009). *Producing a systematic review.* In Buchanan, D. A. – Bryman, A. (Szerk.) *The Sage handbook of organizational research methods.* London: Sage, 671–689.

Dijk, T. A. (1985) *Handbook of discourse analysis*. London: Academic Press.

Fairclough N. (1992) Discourse and Text. Linguistic and Intertextual Analysis within Discourse Analysis. *Discourse & Society,* 3(2), 193–217. https://doi.org/10.1177/0957926592003002004

Fazekas M. – Tóth I. J. (2012) *A hatékony kormányzás alapvető feltétele, hogy adatokkal rendelkezzünk a kormányzásról.* Budapest: Budapesti Corvinus Egyetem. Elérhető: http://www.crcb.eu/wp-content/uploads/2014/01/kb_adatok_2010_3riport_120304.pdf [Letöltve: 2021-03-08]

Fazekas, M. – Tóth, I. J. (2016) From Corruption to State Capture: A New Analytical Framework with Empirical Applications from Hungary. *Political Research Quarterly,* 69(2), 320–334. https://doi.org/10.1177/1065912916639137

Fazekas, M. – Tóth, I. J. – King, L. P. (2016) An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research,* 22(3), 369–397.

Gerő M. – Mikola B. (2020) *A korrupcióérzékelés két arca: a hétköznapi és az állami szintű korrupció észlelésének legfontosabb meghatározói.* A Transparency International Magyarország és a Társadalomtudományi Kutatóközpont közös jelentése a korrupcióérzékelésről. Elérhető: https://transparency.hu/wp-content/uploads/2020/05/Ger%C5%91-Mikola-2020-A-korrupci%C3%B3%C3%A9rz%C3%A9kel%C3%A9s-k%C3%A9t-arca.pdf [Letöltve: 2021-03-08]

Hajdu M. – Pápay B. – Szántó Z. – Tóth I. J. (2016) *Human Assisted Content Analysis of the print press coverage of corruption in Hungary.* Projekt jelentés. ANTICORRP.

Hlatshwayo, S. – Oeking, A. – Ghazanchyan, M. – Corvino, D – Shukla, A. – Leigh, L (2018) *The Measurement and Macro-Relevance of Corruption: A Big Data Approach.* IMF Working Papers 18 (195), 1. https://doi.org/10.5089/9781484373095.001.

Hirschberg, J. – Manning, C. D. (2015) Advances in natural language processing. *Science,* 349(6245): 261–266. https://doi.org/10.1126/science.aaa8685.

Ignatow, G. – Rada, M. (2016) *Text Mining. A Guidebook for the Social Sciences*. Thousand Oaks, CA: Sage.

Josefsson, A. (2014) *Mainstream Party Strategizing on Corruption Issues – The Case of Italy.* BA thesis. Gothenburg: University of Gothenburg. Elérhető: https://gupea.ub.gu.se/handle/2077/36519 [Letöltve: 2020-07-26]

Katona E. – Kmetty Z. – Németh R. (2021) A korrupció hazai online média-reprezentációjának vizsgálata természetes nyelvfeldolgozással. *Médiakutató*, megjelenés alatt.

Lavrakas, P. J. (2008) *Encyclopedia of Survey Research Methods.* Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781412963947

Li, J. – Chen, W. – Xu, Q. – Shah, N. – Mackey, T. (2019) Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology. *IEEE Global Humanitarian Technology Conference (GHTC),* Seattle, WA, USA, 1–4. https://doi.org/10.1109/GHTC46095.2019.9033129.

Marchetti R. (2016) Methodology of the human-assisted content analysis. Elérhető: http://anticorrp.eu/publications/methodology-of-the-human-assisted-content-analysis/ [Letöltve: 2021-03-08]

Mear, C. (2016) *An International NGO Startup's Use of Social Media Technology. The Case of Anti Corruption International: A Discursive Analysis on the Organizational Use of the Term 'Corruption' on Twitter.* Master's Thesis. Trondheim: Norwegian University of Science and Technology. Elérhető: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2449754/Camilla%20Mear%20-%20MSGLOPOL%20-%202016.pdf?sequence=1&isAllowed=y [Letöltve: 2020-07-26]

Moreno, S. A. – Redondo, T. (2016) Text Analytics: the convergence of Big Data and Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence,* 3(6): 57–64. https://doi.org/10.9781/ijimai.2016.369

Muço, A. (2019) *Learn from thy Neighbor: Do Voters Associate Corruption with Political Parties?* Elérhető: https://www.dropbox.com/s/yphqpyiaue8ngoe/Learn_from_thy_neighbour.pdf?dl=0 [Letöltve: 2020- 07-26]

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

Németh, R. (2015) A számok tényleg magukért beszélnek? *Replika*, Big Data és szociológia különszám (92–93), 203–208.

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,* https://doi.org/10.1177/1609406920949338

Niklander S. – Soto R. – Crawford B. – de la Barra C.L. – Olguín E. (2016) Facilitating Analysis of Audience Reaction on Social Networks Using Content Analysis: A Case Study Based on Political Corruption. In Stephanidis C. (szerk.) *HCI International 2016 – Posters' Extended Abstracts. Communications in Computer and Information Science.* Las Vegas: Springer. https://doi.org/10.1007/978-3-319-40542-1_10

Noerlina, – Wulandhari, L. – Sasmoko, S. – Muqsith, A. – Alamsyah, M. (2017) Corruption Cases Mapping Based on Indonesia's Corruption Perception Index. *Journal of Physics: Conference Series 801.* Elérhető: https://iopscience.iop.org/article/10.1088/1742-6596/801/1/012019 [Letöltve: 2020-07-26]

Novitasari, N. – Mardikantoro, H. B. (2019) Representation of Rajawali Citra Televisi Indonesia and Indosiar Social Cognition of Journalist in the Construction of Discourse on Corruption News. *Seloka: Jurnal Pendidikan Bahasa dan Sastra Indonesia* 8(2) Elérhető: https://journal.unnes.ac.id/sju/index.php/seloka/article/view/31228 [Letöltve: 2021-03-08]

Ogunmuyiwa, H. O. (2015) A Critical Discourse Analysis of Corruption in Presidential Speeches. *International Journal for Innovation Education and Research,* 3(12), 31–50. https://doi.org/10.31686/ijier.vol3.iss12.484

Ogunmuyiwa, H. O. (2019) Analysing the discourse on corruption in presidential speeches in Nigeria, 1957–2015: Systemic functional linguistics and critical discourse analysis frameworks. Elérhető: http://etd.uwc.ac.za/xmlui/handle/11394/6674 [Letöltve: 2021-03-08.]

Pan, J. – Chen, K. (2018) Concealing Corruption: How Chinese Officials Distort Upward Reporting of Online Grievances. *American Political Science Review,* 112(3), 602–620. https://doi.org/10.1017/S0003055418000205

Rabuzin, K. – Modrusan, N. (2019) Prediction of Public Procurement Corruption Indices using Machine Learning Methods. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management 3:* KMIS, 333–340. https://doi.org/10.5220/0008353603330340

Touwe, M. – Sultan I. M. (2015) The Investigation of Tempo Weekly News Magazine in the Corruption Case of the Megaproject Hambalang Sport Facilities. *International Journal of Sciences: Basic and Applied Research,* 20(2), 233–239. https://doi.org/10.31947/kjik.v3i2.580

Vijayarani, S. – Ilamathi, M. J. – Nithya, M. (2016) Preprocessing Techniques for Text Mining- An Overview. *International Journal of Computer Science and Communication Networks,* 5(1), 7–16.

Weaver, J. (2020) Jobs for Sale: Corruption and Misallocation in Hiring. Elérhető: http://dx.doi.org/10.2139/ssrn.3590721 [Letöltve: 2021-03-08]