Szociológiai Szemle 29(4): 115–121.

Digitális társadalomkutatások az ELTE-n

Beszámoló A társadalom kutatásának új útjai című workshop digitális társadalomkutatással foglalkozó szekcióiról

Németh Renáta – Barna Ildikó

https://doi.org/10.51624/SzocSzemle.2019.4.5

A társadalom kutatásának új útjai címmel zajlott workshop az ELTE Társadalomtudományi Karán 2019. október 11-én (absztraktfüzet és az előadás diái az rc2s2.eu oldalon találhatók). Az eseményen a Felsőoktatási Intézményi Kiválósági Program (FIKP) által finanszírozott egy éve folyó munka eredményeit mutatták be a kutatók. Beszámolónkban a workshop digitális társadalomkutatásokat bemutató két szekcióját ismertetjük, melyek egy új kutatási projektünk köré csoportosultak. Az első szekció a karon egy éve alakult Research Center for Computational Social Science (RC^2S^2) kutatócsoportunk automatizált szöveganalitikai kutatásai köré épült, a második szekcióban pedig kutatócsoportunk együttműködő partnere, az ELTE Bölcsészettudományi Kara Digital Humanities (DH) kutatóközpontja képviselői mutatkoztak be. Elsőként Németh Renáta, az RC^2S^2 társvezetője (Koltai Júlia társszerzőségével) adott elő *Automatizált szöveganalitika a szociológiai kutatásban – kihívások és lehetőségek* címmel. Az előadás két párhuzamos feladatot célzott egyszerre: kontextusba helyezte az automatizált szöveganalitikát mint módszert a szociológián belül, másrészt – ettől nem függetlenül – ismertette az első szekcióban bemutatkozó kutatócsoport céljait. Az automatizált szöveganalitika módszerének motivációjaként elhangzott, hogy a digitális forradalom a digitális önkifejezés forradalma is, ami nem csak bőséges szöveges forrást bocsát a kutató rendelkezésére, de megnyilvánulási platformot is biztosít bárkinek. Ez a jelenség élesen szemben áll a korábbi társadalmakkal, amikor a nyilvánosságba került szövegek szinte kizárólag az elit narratívái voltak. Az átalakulás új adatforrást generál nemcsak (a lehetőséggel az első pillanattól élő) üzleti szférának, de a társadalomtudománynak is. E szöveges adatok előnye a szociológia számára, hogy a közvetlenül megfigyelhető társas viselkedéshez tartoznak (szemben az önbevallásra épülő survey-jel), ami növeli érvényességüket. További előnyei nyilvánvalóak (real-time, longitudinális követésre is alkalmazható, kis alpopulációk vizsgálatára is elegendő adatot szolgáltat, nem csak 'born-digital', de digitalizált szövegtárak is kialakíthatók). Ez az adatforrás azonban nem mentes az adatminőségi (például külső és belső érvényességre, reprezentativitásra

vonatkozó) korlátoktól sem. E kihívások nem érvénytelenítik a digitális elemzéseket, de szükségessé teszik az adatok létrejöttének megértését és a klasszikus empirikus adatokkal való (egymást kiegészítő és nem érvénytelenítő) relációjuk felismerését.

Az előadó elmondta, hogy a digitális szövegek automatizált elemzését gyakran szkepszis övezi a szociológia oldaláról, miközben az üzleti/technológiai kutatások nagy számban hoznak létre, mintegy melléktermékként, társadalomtudományi szempontból is interpretálható eredményeket. A szkepszis egyik oka ugyanabban a jelenségben azonosítható, mint ami a digitális adatok üzleti felhasználásával kapcsolatos csalódásokat (a 'big data hype' görbe lezuhanását) is létrehozta: az adat- vagy szövegbányászatra épülő üzleti kutatások egy része az adott vállalat rendszerétől izoláltan kezelte az algoritmussal optimalizálandó üzleti problémát. Nem jelenítette meg a tényleges szakterületi tudást (*domain knowledge*) sem a megoldandó probléma kijelölésekor, sem a felhasznált adatok kiválasztásánál, sem a kidolgozott megoldás üzleti adaptálásánál. A domain knowledge jelentőségének felismerése az automatizált szöveganalitikát használó társadalomkutatások sikerének, tudományos diskurzushoz való hozzákapcsolásának is kulcsa. Jelenleg ugyanis számos digitális társadalmi adatot használó kutatásban alacsony a szociológia kontribúciója. Ez intézményesülési problémákra is visszavezethető: a szükséges empirikus szakértelem eltolódott az akadémiai szférától a technológia/ipar felé, és viszonylag magas a belépési küszöb erre a területre (bár léteznek az átlépésére intézményesített próbálkozások könyvkiadók, egyetemek oldaláról).

Az előadó hangsúlyozta: az RC^2S^2 kutatócsoport létrehozásával nem csupán a digitális adatok kutatásának új útjait szándékozzák megtalálni, mintegy módszertani kísérletként, hanem új szociológiai tudás létrehozását is célozzák. Meggyőződésük, hogy a szociológia következő évtizedeiben a computational irányzat megkerülhetetlen lesz. A kutatócsoport tudatos célja a domain knowledge inkorporálása a kutatásba és a szükséges empirikus tudás megszerzése társintézményekkel és üzleti cégekkel való kapcsolatépítéssel, új kutatói kapacitások bevonásával. A hazai és nemzetközi intézményesülés elősegítését oktatási programok fejlesztésével (lásd pl. a Survey Statisztika és Adatanalitika MSc-t), nemzetközi és hazai konferencia-részvételekkel és publikációs platformok (folyóirat-különszámok, konferenciaszekciók) kezdeményezésével igyekeznek tudatosan támogatni.

Másodikként Barna Ildikó (az RC^2S^2 társvezetője) és Knap Árpád *A magyarországi antiszemitizmus vizsgálata NLP topikmodell segítségével* című előadása hangzott el. Az előadók kiemelték, hogy Magyarországon az antiszemitizmus szintje mindig is a legmagasabbak közé tartozott Európában. Reprezentatív felmérések szerint a népesség körülbelül 33–40 százaléka tekinthető antiszemitának. A zsidók körében felvett kutatások adatai azt mutatják, hogy annak ellenére, hogy ma a korábbiaknál sokkal kevesebben tapasztalnak antiszemitizmust, az antiszemitizmus percepciója jelentősen nőtt. A különbség a tapasztalat és a percepció között több okra vezethető vissza, amelyek közül az egyik az online antiszemitizmus terjedése. Ez vezette

a kutatókat az online antiszemitizmus vizsgálatával foglalkozó kutatási projektjük elindításához, az előadásban ennek első eredményeit mutatták be.

Nagy mennyiségű szöveges adatot vizsgáltak a természetesnyelv-feldolgozás (Natural Language Processing, NLP) módszerével. Adatforrásukat a szélsőjobboldali Kuruc.info hírportálon található, feltételezhetően antiszemita tartalmú (= a 'zsidó' kifejezést tartalmazó) 2 289 cikk és 51 060 hozzászólás alkotta. A Kuruc.info oldalra két ok miatt esett a választásuk. Egyrészt ennek az internetes hírportálnak komoly szerepe volt a szélsőjobboldali eszmék elterjesztésében. Másrészt, bár jelentősége az utóbbi időben csökkent, 2018-ban még mindig a teljes népesség 10 százaléka és a Jobbik szavazóinak 23 százaléka látogatta legalább időnként a portált.

Az elemzésben Latent Dirichlet Allocation (LDA) topikmodellezést használtak. E modell adott szöveghalmaz látens tematikus struktúrájának feltárására alkalmas, outputja a látens témák ('topikok') előfordulási aránya és jellemző kifejezéseinek listája. Az előadók jellemezték az így azonosított topikokat, bemutatták, hogy ezek a topikok hogyan kapcsolódnak az antiszemitizmus különböző fajtáihoz, úgymint a rasszista, az összeesküvést feltételező, a másodlagos és az új antiszemitizmushoz. Végül összehasonlították a cikkek és a hozzászólások korpuszainak látens tematikus szerkezetét. Az elemzés során sikerült jól interpretálható topikokat azonosítaniuk, és kiderült, hogy a cikkek és a kommentek eltérő karakterrel rendelkeznek. A legfontosabb konklúziók között elhangzott, hogy ez a megközelítés (hiszen nem önbevallásra, hanem kvázi-résztvevő megfigyelésre épít) a survey módszerrel nem mérhető zsidóellenes narratívákat (mint a rasszista) is képes azonosítani. Az előadók elmondták, hogy kutatásukat a jövőben tematikus és módszertani szempontból is bővíteni fogják. Egyrészt az antiszemitizmus rejtettebb formáival és az online gyűlöletbeszéd más típusaival fognak foglalkozni, másrészt ezek az új témák újfajta módszerek alkalmazását is szükségessé teszik majd.

Kmetty Zoltán és Koltai Júlia előadása *A kulturális jelenségek megértése a természetesnyelv-feldolgozás lehetőségeivel* címet viselte. Előadásukban az elmúlt években széles körben elterjedt neurális hálókon (mesterséges intelligencián) alapuló szóbeágyazási modellek társadalomtudományi alkalmazására fókuszáltak. Szóbeágyazási modelleket széles körben alkalmaznak más területeken, úgymint szótárkészítésnél, videóajánlási rendszerek kialakításakor vagy termékértékelések elemzésére; az emberi viselkedés és kultúra megértésében azonban eddigi alkalmazásuk korlátozott. Mivel azonban a rendelkezésre álló digitális adatok (szövegek) hatalmas mennyisége sok információt szolgáltat preferenciáinkról, választásainkról és gondolkodásunkról, nagy potenciál rejlik társadalomtudományi hasznosíthatóságukban. Az előadásban számos példát mutattak be a szóbeágyazási modellek ezen területen történő felhasználására, de kitértek a modellezés módszertanára, a megoldandó problémákra és a további fejlesztési irányokra is. A nemzetközi példák mellett három saját elemzést is bemutattak. Az első elemzésükben előretrénelt szövegvektorokat felhasználva bemutatták, hogy online szövegek alapján hogyan lehet rekonst-

ruálni az egyes országokra jellemző korrupciós típusokat. A csak szövegbányászati módszerrel előállított országszintű korrupciós indikátoruk közel 0,8-as korrelációt mutatott a Transparency International korrupciós indexével. Második példájukban a travelgram hashtag alatt megjelent Instagram-posztok elemzését mutatták be. A szóbeágyazási modellt ebben az esetben arra használták, hogy azonosítsák az egyes városokra jellemző utazási mintákat. Utolsó példájukban pedig szintén Instagram-adatok elemzését mutatták be, arra fókuszálva, mennyiben ad hasonló/eltérő eredményt, ha ugyanazt a korpuszt topikmodellezéssel vagy szóbeágyazással dolgozzák fel. Mind a három saját példa jól mutatta, hogy milyen hatalmas potenciál rejlik ennek a módszernek a társadalomtudományi adaptálásában.

Sik Domonkos és Máté Fanni (Németh Renáta és Katona Eszter társszerzőségével készült) előadása *A depresszió keretezése online betegfórumokon – a társadalmi szenvedés kutatása automatizált szöveganalitikai perspektívában* címet viselte. Sik Domonkos kiemelte: a téma szociológiai tétje nagy, mivel a depresszió a modernitás (nép)betegsége, a társadalmi szenvedés egy formája. A kutatás célja az automatizált szöveganalitika lehetőségeinek kiaknázása az online depressziófórumok (a depresszió laikus diskurzusainak) elemzésében egy ez ideig jellemzően kvalitatív szövegelemzéssel megközelített témában.

Az előadás a kutatás első szakaszát ismertette, melynek középpontjában a depresszió egyéni értelmezése (keretezése) áll. E keretezés maga is társadalmi konstrukció, s nem csupán a depresszió jelentését, interpretációját és oktulajdonítását határozza meg, hanem a lehetséges megelőzést és a preferált kezelési módokat is. A depresszió tudományos diskurzusában három keretezés különül el: a biológiai-medikális, a pszichológiai és a szociológiai aspektusra épülő. A kortárs szociológiai megközelítések egyik kérdése például, hogy hogyan transzformálódik a pszichoterápia során az eredetileg társadalmi keretezés a terapeuta által jobban uralt, pszichéhez kapcsolt keretezéssé. Ezen értelmezési módok megjelenését vizsgálták a kutatók a legnépszerűbb angol nyelvű online, depresszióval kapcsolatos fórumokon született körülbelül 80 000 poszt elemzésével. Gépi tanulást alkalmaztak: algoritmusuknak a kódolók által a három keretezési kategóriába besorolt 4500 poszt klasszifikációs szabályait tanították meg. E kísérlet tétje nagy: a (humán kódolók inputjára épülő, vagyis felügyelt) gépi tanulás a mesterségesintelligencia-kutatás középpontjában áll, betanított munkások ezreit alkalmazzák online platformokon egyszerű osztályozási feladatokra az ipari/üzleti alkalmazások során. Nyitott kérdés, hogy vajon a szociológiai kutatások számára releváns, hermeneutikailag nagyobb kihívást jelentő osztályozási feladatok milyen sikerrel algoritmizálhatók – ez az automatizált szöveganalitika szociológiai alkalmazásának lehet kulcsa. Az előadás az emberi besorolás tapasztalatait tárgyalta, legfontosabb következtetései között volt annak felismerése, hogy a kódolók aktív értelmezését megkövetelő téma esetén a a nem egyértelmű besorolás, a kódolók bizonyos fokú eltérése, ezen eltérés megértése és a feldolgozási módszerbe építése, és magának az értelmezésnek/kódolásnak a megér-

tése megkerülhetetlen és sok elméleti-technikai kihívást jelentő feladat. Ez vélhetően nem csak a depressziós keretezés, hanem minden (hermeneutikailag nehezebb fogalmakat vizsgáló) szociológiai gépi tanulás sajátja.

Katona Eszter Máté Fannival (Németh Renáta és Sik Domonkos társszerzőségével) a szekció utolsó előadásában *(A depresszió keretezése online betegfórumokon – felügyelt tanulás szubjektív témában)* ugyanezen depressziókutatás kapcsán az emberi besorolásra épülő automatizált osztályozó modelleket tárgyalta: felügyelt tanulóalgoritmusuk segítségével alkottak a kódolók által korábban besorolt posztok mintázatai alapján osztályozó modelleket. Az általuk alkalmazott klasszifikációs modellek (Support Vector Machine, Naive Bayes, döntési fa, logisztikus regresszió) a tapasztalatok szerint jól teljesítenek olyan üzleti feladatokon, mint a fogyasztók adott áruval kapcsolatos posztjának elégedettség szerinti besorolása, de itt is kérdés, hogy a szociológia számára releváns, komplexebb fogalmak esetében is használhatók-e. Előadásukban bemutatták a (szöveganalitikai kutatások komoly részét adó) szöveg-előfeldolgozás során hozott döntéseket. Ismertették, hogy a klasszifikációs modell illesztése során milyen szempontokat vettek figyelembe, milyen úton jutottak el a végső modellekig, hogyan mérték a modell predikciós pontosságát (a kódolók pontosságával összevetésben), s melyek azok a nyelvi kifejezések, amik meghatározzák, hogy egy-egy bejegyzés nagy valószínűséggel milyen keretezéshez tartozhat. Az előadás egyik legfontosabb következtetése az volt, hogy a szociológiai keretezés kevésbé jól prediktálható, mint a biológiai-medikális, illetve pszichológiai keretezés. Ez arra mutathat, hogy a depresszió szociológiai diskurzusa kevésbé kidolgozott, míg a pszichológiai vagy a biológiai-medikális keretezés a nyilvánosságban is széles körben használt intézményesített szókészlettel bír. Ez (a jövőben még erősebb alátámasztást igénylő) különbség a depresszió diszkurzív mezejének hierarchikus szegmentációjaként értelmezhető.

A második szekció a társadalomkutatásban használható digitális adattárak építéséről szólt. Az első előadó, az ELTE BTK Digitális Bölcsészet Központ társvezetője, Palkó Gábor előadásában *(Szemantikus adathálózatok építése életrajzi adatokból: a WikiDATA modell tudományos célú alkalmazása)* egy általuk épített szemantikus prozopográfiai (kollektív életrajzi) adathálózatot ismertetett, amely a WikiDATA adatelemeivel közösen is vizsgálható. A HECE projekt alapja egy jelenleg is folyó kutatás (előzménye a 2014–2019 között zajlott MTA–ELTE Lendület pályázat Humanizmus Kelet-Közép-Európában, Történeti Értelmiségi Hálózatok címmel, Kiss Farkas Gábor vezetésével). A kutatás célja az 1420 és 1620 között a Magyar Királyság területén született irodalmi művek és szerzőik értelmiségi karriermintázatainak vizsgálata. Bár a prozopográfia egyik fő műfaja a XXI. században még mindig a könyvformátumú enciklopédia és a lexikon, ezeknél a nagy érdeklődésre igényt tartó kapcsolathálózatok reprezentációja nagyon korlátozott. A megoldás egy prozopográfiaként szolgáló adatbázis létrehozása, amely más, külső adatbázisokból származó adatokkal is összekapcsolható. Az előadásból kiderült az is, hogy a

külső adatbázis céljára a kutatók a WikiDATA-t tartják a legalkalmasabbnak, még akkor is, ha az adatminőséggel a tudományos kutatás horizontjából fel is merülnek problémák. A WikiDATA a világ legnagyobb enciklopédiája és legösszetettebb prozopográfiai adathálózata. A HECEdata projekt keretében egy a WikiDATA mintáját követő önálló adathálózat került kialakításra. Az ELTE DH vezetésével a kutatók egy mintegy hatvan adatmezőt tartalmazó adatbázis-szerkezetet hoztak létre, amelyet doktori hallgatók közreműködésével töltenek fel életrajzi adatokkal. A későbbiekben a kutatócsoport szerkesztésében megjelenő lexikon szövegéből félautomatikus módszerrel nyerik majd ki a könyvészeti adatokat. Egy ilyen adatbázis lehetőséget biztosít az adatbázison belüli és – a létrehozott adatkapcsolatokon keresztül – külső adatbázisokban mintázatok gépi intelligenciával való feltárására.

A szekció második előadásában Indig Balázs, az ELTE DH nyelvtechnológusa (*Ki szűken arat, bőven arat – Egy újelvű, strukturált webaratási módszer céljai és első tapasztalatai* címmel) azt a projektet mutatta be, mely kutatócsoportunk számára is előállít társadalomkutatásban használható webes korpuszokat. A webaratás tudományos szükségessége mellett érvelt, majd lépésenként mutatta be magát a folyamatot, kifejezetten bölcsész- és társadalomtudományi célú felhasználásokra felkészítve azt. Elmondta, hogy a gépek kapacitásának növekedésével és a Web 2.0 által robbanásszerűen megnőtt adatmennyiséggel nagyot nőtt a webaratás jelentősége. A korábbi várakozásokkal ellentétben napjainkban reneszánszukat élik a statikus, Web 1.0-át idéző felépítéssel rendelkező oldalak, melyek tovább fűtik ezt a növekedést. Egyre digitalizálódó világunkban viszont felmerül a kérdés, hogy a tudósok számára hogyan őrződnek meg kutatható formában ezek az egyre volatilisebb tartalmak. Több független, mindenki számára elérhető kezdeményezés közül kiemelendő az Internet Archive, amely számos kiválasztott weboldalt heurisztikus alapon ment le úgynevezett webarchívumokba, különböző időintervallumokban vissza-visszatérve a már meglátogatott címekre. A lementett, már nem elérhető tartalmak – habár korrajznak tekinthető általános képet adnak az oldalakról, de – rendszerszintű kutatásra a webaratás durva felbontása miatt és pontos célja hiányában alkalmatlanok. Az előadó és társai célja pontosan célzott, részletgazdag és jól kutatható achívumok létrehozása a gondosan kiválasztott weboldalakból, hogy a lementett anyagból kinyerhessük a szöveges tartalmat és metaadatokat, melyek így egy egységes, strukturált adattömbbé állhatnak össze. Munkájukhoz felhasználják az Internet Archive által fejlesztett és a nemzetközi közösségben nagy népszerűségnek örvendő, ISO-szabvány szerinti WARC formátumot, amelynek minden további feldolgozása újrafuttatható, javítható. A módszer lényege, hogy statisztikai információkat felhasználó heurisztikák helyett a weboldalak visszatérő, emberek számára készült mintázatait használjuk fel a fontos, mentendő címek rendszerszerű kiválogatására. Az előadó a módszer skálázhatóságának vizsgálata céljából nyolc különböző hírportálon indított kísérletük tapasztalatait mutatta be, amit az ismertetett módszer továbbfejlesztése és további weboldalak bevonása fog követni. Ismertette továbbá jövőbeli

terveiket és a módszer várható alkalmazási területeit, melyeket a jelenlegi eredmények tükrében állapítottak meg.

A workshop mintegy száz résztvevővel (köztük egyetemünk vezetői, más karok oktatói, a társegyetemek és az MTA képviselői, továbbá a szöveganalitikai kutatásainkban érintett partnercégek képviselői jelenlétében) zajlott. Az előadások számos kérdést és hozzászólást generáltak, ugyanabba az irányba mutatva: a digitális adatelemzés, az automatizált szöveganalitika forradalmi változásokat hoz az ipari, üzleti és tudományos alkalmazásokban egyaránt. A szociológia akkor tudja ennek a fejlődésnek a lehetőségeit kihasználni, ha képes lesz megújítani kutatási módszertanát kritikai reflexiójának megőrzése mellett. A workshop bemutatott előadásai ennek támogatását célozták.