MAGYAR NYELVJÁRÁSOK 52 (2014): 171–182.
A DEBRECENI EGYETEM MAGYAR NYELVTUDOMÁNYI TANSZÉKÉNEK LEKTORÁLT FOLYÓIRATA

A jelentés társas aspektusának kinyerése szöveganalitikai eszközökkel

VERÉCZE VIKTÓRIA

1. Bevezetés

Azt, hogy az ember társas lény, ma már egyre több nyelveredettel, nyelvelsajátítással foglalkozó elmélet ismeri el, és veszi figyelembe. A társas közeg hatása pedig a kognitív rendszer szerveződésében is megnyilvánul. „Társadalmi környezet, társas kapcsolatok nélkül nem jöhet létre teljes értékű emberi beszéd" (RÉGER 1990: 7), a társas kapcsolatok át- meg átszövik a nyelv kialakulásának, az egyén nyelvelsajátításának, a jelentésképzésnek és mindenfajta nyelvi változásnak és változatosságnak a folyamatait. Mivel a nyelv a kognitív rendszer része, így egyszerre lesz inherensen társas és kognitív beágyazottságú, vagyis társas-kognitív. A nyelv felépítését és működését ezért jelen tanulmányomban e szerint vizsgálom azt tartva szem előtt, hogy az ember szociális lény, s mint ilyen, viselkedése, megnyilvánulásainak (szavainak és nonverbális közléseinek) jelentése is csak e kontextusba ágyazva vizsgálható.

A nyelv társas beágyazottságának számos következménye közé tartozik az is, hogy a nyelvváltozatok nem választhatók el élesen egymástól. Mivel nyelv és beszélő sem szétválasztható, ezért a beszélők közötti társas kapcsolatok hatásai egyúttal a nyelvi rendszer szerveződésében megjelenő kapcsolatok is lesznek. A beszélők különböző csoportjai a csoport szintjén olyan nyelvváltozatokat beszélnek, melyek aztán mutathatnak különbségeket például attól függően, hogy a beszélők hol élnek, melyik társadalmi rétegbe tartoznak, mi a foglalkozásuk, egynyelvűek-e, milyen neműek, milyen életkorúak stb. (vö. SÁNDOR–KAMPIS 2000: 129). Mivel azonban egy személy egyidejűleg több csoportnak is a tagja (például egy nyírségi város lakótelepének és egy fővárosi iroda munkaközösségének), ezért amikor beszélünk, akkor nyelvváltozatainkból a csoportidentitásnak megfelelően választjuk ki (gyakran nem tudatosan) az alkalmasnak tűnőt. Megnyilatkozásunk a társas jelentésen keresztül jelzi, hogy a csoporthoz tartozunk-e, illetve hogy hol foglalunk helyet a csoport hierarchiájában (lásd SÁNDOR 2002: 69).

A társas jelentést így határozza meg SÁNDOR KLÁRA: „minden egyes szóhoz és grammatikai formához társulnak olyan jelentések, amelyek megszabják, hogy az adott elemet a beszélő kikkel beszélve és milyen helyzetekben használja. Ugyanez a társas jelentés a hallgatónak azt mondja el, hogy akivel beszél, az milyen társadalmi csoportból való, melyik földrajzi területről jön, és milyennek értelmezi éppen a kettejük között fennálló viszonyt" (2001: 87). A társas jelentés tehát magában hordozza közlésünknek számos indirekt információját is, melyeket a fogadó fél akár szándékunk ellenére is sikeresen dekódolhat.

2. Automatizálás, algoritmizálás

A társas jelentés kinyerése szóbeli kommunikáció esetén nem okoz különösebb nehézséget a beszélgető felek számára, a helyzet azonban lényegesen eltérő a digitális környezetben. Napjainkban számos kutatás és fejlesztés irányul arra, hogy az emberi beszédet megértő, illetve azt reprodukáló mesterséges rendszereket hozzanak létre. Célszerű elkülöníteni egymástól a megértést és a reprodukciót, mert bár két szorosan összetartozó területről van szó, automatizálásuk eltérő stratégiákat igényel. Ahhoz, hogy az emberi beszédet tökéletesen leképező robotok jöhessenek létre, a beszédértésnek is tökéletesen kell működnie.

Előfeltevésem szerint hangzós emberi beszéd csak társas közegben jöhet létre, és mindig is magán viseli a szociális környezet folyamatosan alakító hatásait, amiből pedig egyenesen következik, hogy a mesterséges intelligenciák emberhez hasonló beszédtanulásával kapcsolatban sem hagyhatjuk figyelmen kívül a társas beágyazottságból következő hatásokat. A beszélők társas hálózata, a különböző szociális és kulturális viszonyulások, cselekvések olyan inputok lesznek így, melyeket egy gép beszélni tanulásakor, illetve a beszédfeldolgozó rendszerek fejlesztésekor is fel kell használnunk. Ezen inputok algoritmizálhatósága azonban nehézségekbe ütközik jelenlegi tudásunk szerint.

Így felvetődik a probléma másik irányból való megközelítése, vagyis hogy hogyan nyerhetjük ki a társas jelentést a kommunikációból. A szóbeli és az írásos közlések feldolgozása azonban nem teljesen azonos stratégiákat követel meg, emellett a szóbeli közlések feldolgozására is úgy kerül sor leggyakrabban, hogy első lépésként írott szöveggé alakítják őket. Koncentráljunk ezért jelen esetben az írott szövegekre, melyekből a világhálón hatalmas mennyiség áll legálisan letölthető és ezáltal vizsgálható formában rendelkezésünkre.

Ennek a hatalmas adatbázisnak a kiaknázása nem csak a nyelvészeti kutatások számára szolgálhat tanulságul. „A legkülönbözőbb internetes forrásokból származó információk, vélemények, álláspontok és meggyőződések kinyerésének igénye mára a világ számos pontján vitathatatlanul összefonódott a megalapozott stratégiai döntések meghozatalának célkitűzésével" (KOVÁCS-ÖRDÖG–PANCZA 2014: 34). Ez a fajta tudáskinyerés azonban nagy kihívást jelentő fel-

adat, mivel hatalmas mennyiségű heterogén, félig strukturált, illetve struktúra nélküli adatról van szó. Rengeteg témában találhatunk információkat, ezek típusa is nagymértékben eltérő lehet (egyszerű szöveg, táblázatok, multimédia), legtöbbjük redundáns és zajos, valamint változik. Emellett sok információhoz hiperlinkeken keresztül juthatunk el, ráadásul a web egy hatalmas közösség, tehát nemcsak adatokról beszélünk, hanem felhasználók közötti kapcsolatokról és interakciókról is.

A hasznos információk feltárásának mélysége és módszere többféle lehet attól függően, hogy pontosan milyen célra kívánjuk felhasználni a kapott adatokat. Az ún. szentimentelemzés vagy polaritásmérés egy adott szöveg érzelmi viszonyulását vizsgálja. Ez a viszonyulás általában háromkomponensű (pozitív, negatív vagy semleges), de létezik olyan elemzés is, mely öt fokozaton skálázza az értékeket a pozitív és a negatív végpontok között. Gyakran használják a vásárlók attitűdjének feltérképezésére egy adott termékkel kapcsolatban (a szentimentelemzéssel kapcsolatban alapos összefoglalást ad LIU 2012).

Az, amit emócióelemzésnek neveznek, már egy fokkal árnyaltabb képet kíván feltárni az adott szövegről. Az emócióelemzés CHARLES DARWIN azon kutatásait gondolja tovább, melyek során bizonyos emberi érzelmek eredetét az állati viselkedésből kívánta származtatni (1963). DARWIN nyomdokain haladva Paul Ekmant már leginkább az arckifejezések és egyes érzelmek összekapcsolása foglalkoztatta. Később ezekre a darwinista hagyományokra és Ekman érzelemdetektálási kísérleteire alapozva indult el az emócióelemzés nyelvtechnológiai hasznosítása, melynek során nyolc különböző érzelmi viszonyulást társítanak az adatokhoz. Ez a nyolc érzelmi kategória: a düh, a félelem, a szomorúság, az undor, a meglepettség, a várakozás, a bizakodás és az élvezet. Az eljárás során a szoftver megszámolja „az egyes kategóriákhoz tartozó szavak arányát és általában a legmagasabb értéket elérő emócióba sorolódik a vizsgált adat" (VARJÚ 2013).

Van továbbá egy olyan módszer, mely szóhasználati módok és mintázatok alapján vizsgálja a szövegeket, és ezáltal képes például a szöveg szerzőjének behatárolására. Ez pedig elvezet minket a társas jelentés egy nagyon fontos aspektusához. A társas jelentés ugyanis elmondhatja például a befogadó számára, hogy a közlő honnan származik vagy mi a foglalkozása anélkül, hogy ez expliciten, szavakkal ki lenne fejezve. Az ilyen jellegű adatok szövegből való kinyerése épp annyira fontos lehet, mint egy termékkel kapcsolatos pozitív vagy negatív hozzáállás detektálása. Egy termék értékesítése során így például optimalizálható lehet a marketingstratégia a vevők lakóhelye szerint is.

A továbbiakban egy nagyméretű adatbázison végzett automatizált elemzést kívánok bemutatni, rávilágítva a szöveganalitikának a társas jelentés kiaknázásában játszott szerepére.

3. A társas jelentés a szövegbányászatban

A szöveganalitika vagy szövegbányászat az adatbányászat egyik speciális ága. Feladata, hogy strukturálatlan adatokból strukturált adatokat hozzon létre, és azokból információt nyerjen ki (bővebben lásd még MINER és mtsai 2012). Ilyen strukturálatlan adatnak tekinthetünk például bármilyen szöveges állományt, szemben a strukturált adattal, amely gyakran számszerűsíthető és így könnyebben feldolgozható.

Jelen tanulmányban egy szöveganalitikai módszerekkel felállított döntési fát (modellt) szeretnék bemutatni, aminek segítségével kinyerhető a szöveg társas jelentésének azon aspektusa, mely a szerzővel kapcsolatban fed fel bizonyos információt a befogadó számára. Ehhez egy konkrét esetet, egy ingatlanügynökök és tulajdonosok lakáshirdetéseinek szövegeiből álló korpusz vizsgálatát veszem alapul. Az adatválasztást az indokolta, hogy a lakáshirdetések nagy számban és szabadon hozzáférhetők a magyar weben, valamint az egyes hirdetésekhez tartozó hirdetők személye nem közömbös a piac szempontjából sem. Mind a leendő lakásvásárlók, mind az ingatlanpiac eladói oldalának résztvevői számára fontos információ a hirdető kiléte. Előbbieket a bekalkulálható költségek miatt, míg az ingatlanügynökségeket az ügyfélkör bővítése szempontjából érdekelheti. A hirdető személye azonban expliciten csak kevés hirdetésben van megfogalmazva, így adott az igény egy minél pontosabb automatizált eljárásra, mely a nagy méretű adathalmazból képes kiszűrni a kellő információt.

A cél tehát a hirdetés feladójának azonosítása volt. Ez az információ pedig a szövegben impliciten megtalálható társas jelentés része. A szövegezés módja, stílusa, szó- és szófajhasználata többletinformációt kínál számunkra, melyet az emberek tudatos stratégia nélkül is kiválóan tudnak dekódolni. Példaként tekintsünk meg egy tulajdonostól és egy ügynöktől származó szöveget (a helyesírást az eredetihez híven közlöm, a személyes adatokat pedig törlöm).

„Helvécián 65m2 családi ház eladó.Gázfűtéses 2, 5 szobás beépithető tetőtér nagy kertel 1300m2 teleken.Felújításra szorul”.

„Az m0-as autó úthoz közel, de még is csendes helyen kínáljuk ezt az új építésű sorházi lakást azoknak, akik új otthont keresnek családjuk számára! Fiatalok!! a szoc.pol. csak júliusig vehető igénybe!! Ne hagyják ki ezt a lehetőséget!! Ha most lefoglalózzák, elindítják a hitelt, még megkaphatják az állami támogatást!! Ezt a lakást most még igényei, ízlése szerint alakíthatja!! [törlés] *2500 ft. értékben kiválaszthatja a burkolatok szinét, az ön ízlése szerint festik a helyiségeket! Ezen a helyen nagy kirándulásokat tehet, biztonságban tudhatja családját, gyermekeit!!, A döntésképtelenek lemaradnak erről a páratlan lehetőségről! Ön nem az!!* [törlés] *-látom ahogy a hirdetést olvassa!-* [törlés] *hívjon!! Megközelíthetőség tömeközlekedéssel: távolsági busz 5 perc. Megközelíthetőség autóval: m0-as 5 perc”*

A fenti két hirdetést elolvasva mindazok nagy biztonsággal tudják megmondani, hogy ingatlanügynök vagy tulajdonos adta-e fel őket, akik életük során olvastak már ilyen jellegű szövegeket, vagy vannak ismereteik az ingatlanpiac ezen oldaláról. A szövegek egyes szavainak klasszikus referenciális jelentéseit végigvéve ez az információ nem derül ki, hiszen egyszer sem szerepelnek sem a *tulajdonos,* sem az *ügynökség, iroda* stb. szavak, mégis van egy ilyen — társas — jelentésük is. A hirdető ki nem mondott személye csak azok számára derül ki a szövegből, akik például: beszélnek magyarul; próbáltak már Magyarországon lakást vásárolni; tudják, hogy az ügynököknek érdekében áll minél hatékonyabban felkelteni a potenciális vevő érdeklődését, hiszen hivatásszerűen foglalkoznak értékesítéssel, és mivel ez a munkájuk, feltehetően ismerik a figyelemfelkeltés stratégiáit; tudják továbbá, hogy ezzel ellentétben a tulajdonosok a saját munkájuk mellett csak ritkán foglalkoznak lakásuk értékesítésével, többnyire rutintalanul mozognak az ingatlanpiacon stb. Tehát rendelkezniük kell a megfelelő társas-kulturális információval, ami a társas jelentés dekódolásának feltétele.

Egy hazánkban élő átlagos felnőtt állampolgárról ezek általában el is mondhatók. Ahhoz azonban, hogy ezt a magától értetődő stratégiát átültessük a gépbe, és ezáltal nagy adatbázison is gyors és automatizált elemzést tudjunk lefuttatni, egy erre alkalmas hatékony modellt kellett létrehozni.

E modell létrejöttéhez a Clementine Consulting (korábbi nevén: SPSS Hungary) szövegtechnológiai és adatbányászati cég biztosította a szoftveres hátteret, valamint a személyi feltételeket. Az elemzést és a modellezést a Clementine Consulting gyakornokaként két szöveganalitikai elemző munkatársammal, Kovács-Ördög Zitával és Pancza Judittal végeztem el, segítségüket ezúton is köszönöm. A munkához a Clementine által forgalmazott IBM SPSS Modelert, egy grafikus felületű, statisztikai műveletekre, klaszterezésre és modellezésre alkalmas szoftvert, illetve a Clementine által fejlesztett Clemtext nevű magyar nyelvű szöveganalitikai eszközt használtuk fel. Az elemzés egyes részeredményeit KOVÁCS-ÖRDÖG ZITA előadásában a 2014. április 15-én Budapesten megrendezett Big Data: a Nagy Lehetőség vagy a Nagy Testvér? elnevezésű konferencián már ismertette.

4. Szóhasználati módok és mintázatok szöveganalitikai elemzése

Vizsgálatunk első lépéseként az interneten hozzáférhető adatok egy részének összegyűjtését kellett elvégezni, majd ezen adatok szűrése és bizonyos fokú strukturálása következett. Kiindulópontunk egy 17 208 rekordból és 17 mezőből álló adattábla volt, mely tartalmazta többek között a hirdetések s az ingatlanok azonosítóját, hogy tartozott-e a hirdetéshez kép, illetve magát a hirdetés szövegét is. Ez a félig strukturált adathalmaz azonban még további szűrési folyamatok lebonyolítását igényelte.

Ahhoz, hogy a modellünket betaníthassuk, illetve hogy nagy biztossággal végezhessünk elemzést az adatainkon, először is meg kellett határoznunk, hogy mely hirdetések köthetők ingatlanügynökökhöz, és melyek tulajdonosokhoz. Ez az információ elengedhetetlen a modell hatékonyságának felméréséhez, vagyis hogy valóban azt ítéli-e meg a hirdetés feladójaként, aki ténylegesen annak tekinthető. A munka több lépcsőfokból áll. Először veszünk egy tanító mintát, ahol mi magunk határozzuk meg, hogy ki ügynök és ki tulajdonos. Ezzel a mintával fogjuk betanítani a modellünket. Ezután a modellt lefuttatjuk egy tesztmintán is, ahol az megnevezi a hirdetőt. A tesztminta is olyan minta, amelyben előzetesen szintén meghatároztuk a hirdetés feladóját, ám a modell ennek az információnak nincs a birtokában a lefutáskor. Ez csak a modell hatékonyságának a visszaellenőrzéséhez szükséges, amikor is összehasonlítjuk a modell által generált hirdetőt és a korábban megadottat. Ezért elengedhetetlen az első lépésben a kézzel végzett minél pontosabb meghatározás.

A szűrés feltételeit mi állítottuk be, és a folyamat során igyekeztünk olyan jellemzőket találni, melyek markánsan jelölik, hogy a hirdetésnek ki lehetett a feladója. Ilyen fontos jellemző volt például, hogy töltöttek-e fel képet az adott hirdetéshez, hiszen tudjuk, hogy az ingatlanügynökségek elvárják munkatársaiktól a fényképek feltöltését. Másik fontos támpontunk volt a szövegek szóhasználata, azon belül is a hirdető személyére kifejezetten utaló kifejezések (például: *tulajdonostól, irodánkban, tekintse meg honlapunkon*). Az adatok ilyen előkészítése után maradt egy olyan adatbázisunk, melyben a szövegekhez már illeszkedett a hirdetés feladójának személye is. A tisztítás részeként kivettük azokat a mezőket, melyeket nem ítéltünk relevánsnak a további munkához. Így maradt 6 836 rekordunk és 5 mezőnk (nevezetesen: az ingatlan azonosítója, a hirdetés azonosítója, hogy tartalmaz-e a hirdetés képet, a hirdetés szövege, a hirdető személye). A rekordok megoszlása a tulajdonos személye szerint: 2 105 rekord ingatlanoshoz, 4 731 rekord pedig tulajdonoshoz tartozott. A szűrés során kinyert adatokból ennyiről tudtuk nagy bizonyossággal megállapítani, hogy ki adta fel a hirdetést.

Az elemzés következő fázisában a hirdetések szövegeinek részletes vizsgálata következett. A kifejezetten magyar nyelvű szövegekre kifejlesztett Clemtext[1] gyorsan és hatékonyan annotálta az adatokat, vagyis morfológiai és szófaji elemzést végzett rajtuk, és a kapott eredményeket különböző betű- és számsorokkal kódolta számunkra,[2] melyek adattáblánkban új mezőkként jelentek meg. Ezzel a szövegek strukturálása is megtörtént.

[1] A Clemtext létrejöttét a magyarlanc (ZSIBRITA–VINCZE–FARKAS 2013: 763–771) nyelvi modul integrálása tette lehetővé.

[2] A *kínálunk* szó kódja például így jelenik meg: Vmip1p---n. A „V” jelöli, hogy igéről van szó, az „m”, hogy ez egy főige, az „i”, hogy kijelentő módú, az első „p”, hogy jelen idejű, az „1”, hogy első személyű, a második „p”, hogy többes számú, az „n” pedig, hogy alanyi ragozású. A kötője-

A következő lépés részletes szó- és szófajhasználati statisztikák létrehozása volt, melyek alapján adatainkat rendezhettük. Az annotálás során egészen elemeire szegmentálódó szövegeinket egyes általunk kiemelt szófajok, illetve azok bizonyos esetei mentén kívántuk újrarendezni. Azokra a szófajokra koncentráltunk, melyek úgynevezett tartalmas szavak, szemben a funkciószavakkal, mivel a szövegek stílusa az előbbiekkel jobban jellemezhető. Így a következőket emeltük ki: főnevek; igék, ezen belül külön a felszólító módú igék, feltételes módú igék; melléknevek, ezen belül külön a fokozott melléknevek; névmások, ezen belül külön az első személyű személyes névmások, a harmadik személyű személyes névmások; határozószók, ezen belül külön a fokozott határozószók; számnevek; kötőszók. A Modeler segítségével minden szófaji csoport elemeit összeszámoltuk, illetve meghatároztuk, hogy az egyes hirdetések szövegében milyen arányban fordulnak elő.

Ezzel azonban csak általános értékeket kaptunk. Ahhoz, hogy a modellünk minél hatékonyabb és pontosabb döntéseket hozhasson, szükség volt a szófaji előfordulások eloszlásvizsgálatára. Így keletkeztek olyan mezők, melyek azt mutatják, hogy az egyes szófajok, szófajcsoportok előfordulása a számtani átlag alatti vagy feletti az adott szövegben. A további elemzésekhez már csak ezt a flag jellegű[3] információt használtuk fel, és így próbáltuk meg „betanítani" a modellünket, vagyis a modellépítéshez szükséges bemenetként már csak ezeket az információkat kapta meg a Modeler.

Fontos lépésként az adatminőség felmérése következett, melynek során megvizsgáltuk, hogy adatainkban vannak-e olyan kiugró értékek, melyek torzíthatják a statisztikánkat, és további műveleteket igényelnek, de ilyeneket nem találtunk.

Azt tapasztaltuk azonban, hogy az adatainkban a tulajdonostól származó hirdetések lényegesen nagyobb aránya mégiscsak torzíthat annyira, hogy az a döntéshozást teszi kevésbé megbízhatóvá. Így döntöttünk amellett, hogy random mintavétellel kiegyenlítjük az ingatlanosoktól és a tulajdonosoktól származó hirdetések számát, majd újra kiszámoltuk a Modelerben az egyes szófajok, szófajcsoportok átlagos előfordulását, és ismét meghatároztuk az átlag alatti és átlag feletti értékeket.

A következő lépés az adatok klaszterezése volt. A klaszterezés során tulajdonképpen információsűrítést hajtunk végre, és az adatainkat bizonyos közös tulajdonságok alapján minél homogénebb, egymást át nem fedő csoportokba

---

lek arra utalnak, hogy a magyar igéknél nem határozunk meg nemet, igenemet (passzív vagy aktív), és hogy tagadó-e. Ez a kódolás egy nemzetközi sztenderdet követ, illetve annak is a kelet-európai nyelvekre kifejlesztett változatát (részletekért lásd http://nl.ijs.si/ME/Vault/CD/docs/mte-d11f/node37.html).

[3] A flagek jelen esetben két értéket, az igazat vagy a hamisat vehették fel meghatározott kondíció függvényében. Például a főnevek esetében a következő formulát adtuk meg: „főnevek_aránya >= főnevek átlagos aránya". Azok a szövegek, ahol ez az állítás igaz, azok az „átlag feletti" értéket, míg a másik csoport az „átlag alatti"-t kapta.

rendezzük. Az egyes klaszterektől nem elvárt az azonos méret, de hatékony klaszterezéssel közel azonos méretű csoportokat kaphatunk. Bemeneteink a szófaj-, illetve szófajcsoport-gyakoriságokból generált változóink lettek. A csoportosítás célja az volt, hogy megnézzük, adataink milyen változók mentén különülnek el egymástól, s a klaszterek mutatnak-e összefüggést a hirdető személyével. Klaszterezéshez az úgynevezett TwoStep módszert választottuk, mely az elemeket hierarchikus adatszerkezetbe, fába rendezi. Az adatpontok a fa levelein találhatók, és a fa minden belső pontja megfelelhet egy klaszternek, mely a fában alatta található elemeket tartalmazza. Az eljárást azért nevezik TwoStepnek, mert két lépésben csoportosítja az adatokat. Az első lépésben egyenként megvizsgálja az eseteket, hogy a távolságfüggvény[4] alapján besorolhatók-e az előzőleg létrehozott miniklaszterekbe, vagy új miniklasztert kell, hogy képezzenek, a következő lépésben pedig már a miniklasztereket csoportosítja újra.

A TwoStep eljárás során adatainkból négy klasztert sikerült képeznünk. Bár adataink binárisan is klaszterezhetőek lennének (tulajdonosok és ingatlanügynökök mentén), mi azt szerettük volna vizsgálni, hogy a szó- és szófajhasználati mintázatok alapján milyen csoportok jönnek létre, és ezeken belül hogyan oszlik meg a tulajdonosok és az ingatlanügynökök aránya.

E négy csoportból az elsőről elmondható, hogy e csoport tagjai kevés melléknevet és határozószót, viszont sok számnevet használnak, illetve főként egyszerű mondatokban fogalmaznak, hiszen a kötőszavak szinte teljesen hiányoznak. Lényegre törő, tényszerű hirdetésekként jellemezhetjük az ide tartozókat. Amikor megvizsgáltuk a hirdetőket, azt találtuk, hogy a csoport túlnyomó részben, 86,75%-ban tulajdonosok által feladott hirdetéseket tartalmaz.

A második csoport szóhasználatát, mondhatni, az első csoport ellentéteként jellemezhetjük. Gyakoriak benne az összetett mondatok, utal erre a kötőszók nagy száma. Találunk továbbá számos megszólítást és felszólítást, valamint nagyszámú fokozott és alapfokú határozószót. Elmondhatjuk, hogy második csoportunk szövegeit választékos (már-már túlzó) stílus, részletgazdag megfogalmazás jellemzi. Ez a csoport 95,67%-ban tartalmazta ingatlanügynökök hirdetéseit.

A harmadik csoportra az igék sűrű előfordulása, illetve a fokozott határozószók átlag feletti használata jellemző. A második csoporttól a felszólító alakok, valamint a megszólítások átlag alatti előfordulása különbözteti meg. A klaszterben 76,1%-ban találtunk ingatlanosoktól származó szövegeket.

A negyedik csoportot átlagon felüli határozószó- és melléknévhasználat, míg a kötőszavak, a névmások és az igék ritkább előfordulása jellemzi. A klaszterben a tulajdonosok aránya 62,51%, míg az ügynökök által feladott hirdetéseké 37,49%.

[4] A távolság sokféle lehet. Például ha rokonokat akarunk klaszterezni, akkor a rokonsági fokok számszerűsítése is képezhet távolságokat az egyes családtagokat jelképező pontok között. Aztán pedig ennek a távolságnak valamilyen függvényesített értéke adja a csoportosítás alapját.

Így kaptuk meg azokat az eredményeket, melyek szerint az ingatlanügynökök szövegeit határozottság, választékos fogalmazás és a célcsoport gyakori megszólítása jellemzi. Ezzel szemben a tulajdonosok hirdetései lényegre törőek, és arányaikban gyakrabban említenek számadatokat.

A feladatunk ezután egy olyan döntési modell létrehozása volt, mely minél nagyobb biztossággal állapítja meg a hirdetések feladóit. Több rendelkezésre álló modell összehasonlítása után az úgynevezett C5.0 döntési fát létrehozó modellt választottunk. Ez egy olyan automatizáltan döntést hozó rendszer, amely a meghatározott bemenő adatok alapján rendszerezi az adatbázist, és mindig egy-egy adott szempont alapján dönt a továbblépésről. A modell vizualizációja leginkább egy fára emlékeztet, innen kapta a nevét is. A döntéshozás során mindig csomópontokból indul ki, ahol az adott szempont szerint kettévágja az adatokat. A modell a fa minden csomópontjában azt keresi a rendelkezésére álló változók közül, amellyel a tanulóhalmaz a célváltozóra (esetünkben ez a hirdetés feladója) nézve homogénebb csoportokat hoz létre, mint a vágás előtt. Jelen esetben a bemenő adatok a kiválasztott szófajok és szófajcsoportok átlag feletti, illetve alatti előfordulásai voltak, melyek mentén haladva a modell végül eldöntötte az egyes szövegekről, hogy ingatlanügynök avagy tulajdonos adta-e fel őket, s ezt az eredményt adta meg kimenetként.

Mivel a Modeler nem nyílt forráskódú szoftver, ezért a döntéseket irányító algoritmusok, illetve a döntési folyamat részletei nem hozzáférhetőek. Mi csak a végeredményt kapjuk meg, illetve az ahhoz vezető út bizonyos lépéseit. Azonban mivel bemenő adataink szabadon hozzáférhetőek az interneten, valamint a Modeler is bárki számára megvásárolható, munkánk megismételhetőségének megvannak a feltételei.

Modellünk döntési szempontjai közül a fokozott határozószók (pl. *fentebb*) előfordulása volt a legfontosabb. A modell minden egyes elágazásnál megmutatta azt, hogy a tanító mintájának hány százaléka esett bele az adott kritériumba, vagyis hogy pontosan milyen értékek alapján döntött az algoritmus. Például az első lépésben a modell a fokozott határozószókat vizsgálta. Ahol a fokozott határozószók aránya átlag alattinak bizonyult, ott a következőképpen alakult a hirdetők eloszlása: 86,04% tulajdonos, 13,96% ingatlanos. Ezt az algoritmus még nem találta kellően meggyőző erejűnek, úgyhogy megvizsgálta — már csak ezen a mintán — az első személyű névmások arányát. Az első személyű névmásokat átlag alatt használó szövegekben a tulajdonosok aránya 87,21%, míg az ingatlanosoké 12,79%. Ezt már elegendőnek találta a modell ahhoz, hogy azt mondja, ennek a csoportnak — vagyis ahol a fokozott határozószók és az első személyű névmások is átlag alatt fordulnak elő — tulajdonosok a hirdetői. Az első személyű névmásokat átlag felett használó szövegekben a tulajdonosok aránya 0%, míg az ingatlanosoké 100%, így ennek az ágnak — vagyis ahol a fokozott határozószók átlag alatt, míg az első személyű névmások átlag felett fordulnak elő

— a hirdetőit pedig ingatlanosokként határozta meg. A meghatározás hasonlóképpen zajlott a továbbiakban is.

A határozószókat átlag felett használó csoportot az igék előfordulása alapján osztotta tovább. Ahol az igék átlag alatti értékben szerepeltek, ott a harmadik személyű névmások szolgáltak újabb csomópontként. Ahol utóbbiak átlag felettiként reprezentálódtak, ott ingatlanos eredményt kaptunk. A másik ágon a felszólító módú igék alapján keletkeztek a következők: átlag feletti előfordulásnál ingatlanosnak bizonyult a hirdető, az átlag alatti ágat a határozószók aránya osztotta tovább. Az átlag alatti határozószók ingatlanosokhoz tartoznak, az átlag feletti aránnyal rendelkezőket pedig a fokozott melléknevekkel bontottuk tovább. Azok a hirdetők, akiknek szövegeiben átlag alatti a fokozott melléknevek előfordulása, tulajdonosnak bizonyultak, míg az átlag felettiek tovább osztódtak a főnevek gyakorisága alapján. Szintén a tulajdonosokat jellemezte a főnevek átlag alatti jelenléte, míg az átlag feletti csoport az ingatlanügynököket.

Az igék átlag feletti előfordulása nyomán keletkező ág hasonlóan összetetten alakult. A következő elágazási pontot a névmások aránya jelentette. Átlag feletti névmásaránynál ingatlanosnak bizonyult a hirdető, míg átlag alattinál a harmadik személyű névmások csoportját vette figyelembe a döntési fa. Ezek átlag feletti előfordulásakor szintén ingatlanost ítélt, míg az átlag alatti eseteket a felszólító módú igék alapján különítette el. Ahol ezek aránya átlag feletti volt, ott ingatlanost detektált a modell, míg ellenkező esetben a fokozott mellékneveket vette figyelembe. A fokozott melléknevek átlag feletti előfordulását ingatlanügynökökhöz kötötte, az átlag alattit pedig a feltételes módú igealakok mentén vizsgálta tovább. Itt ismét az ingatlanosokhoz társította az átlag feletti előfordulást. A másik ágat a főnevek aránya alapján ítélte meg. Ha a főnevek átlag alatt fordultak elő, akkor tulajdonosnak bizonyult a hirdető, míg a másik esetet a határozószók mentén csoportosítottuk tovább — az átlag feletti esetben tulajdonost, míg az alattiban ingatlanost kaptunk végeredményül.

5. Összegzés

A Modelerbe épített eszközök segítségével (Evaluation Node, Analysis Node) ellenőriztük modellünk megbízhatóságát. Előbbi grafikus értékeléssel, görbével ábrázolva, utóbbi kvantitatív értékeléssel, konkrét számadatokkal és százalékos eredményekkel mutatta meg, hogy a tanító mintán betanított modell mennyire megbízhatóan jósolta meg az eredményt a tesztmintán. Eredményként azt kaptuk, hogy modellünk az esetek mintegy 80%-ában pontosan jósolta meg a hirdető kilétét, vagyis a szófaji mintázatok alapján ugyanazt határozta meg, mint tettük mi azt más, erőforrásigényesebb és kevésbé automatizált módon a munka kezdeti fázisában. Ez a 80%-os találati arány pedig már megbízhatónak tartható.

Döntési fánk, megmutatva a döntéshozás útját, feltárta az egyes hirdetőcsoportokra jellemző szóhasználati módokat is, melyek alapján megtörténhetett az

adatok klaszterezése. Modellünk tulajdonképpen automatizáltan, megadott szófajhasználati mintázatok mentén határozta meg a hirdetők személyét, vagyis a hirdetések szövegének egy lényeges társas aspektusát.

Ahhoz, hogy a nyelvfeldolgozás, a mesterségesintelligencia-kutatás minél teljesebben tudja reprodukálni a természetes nyelveket, semmiképp nem hanyagolható el a nyelv társas beágyazottságának figyelembevétele. A társas inputok algoritmizálása azonban sok problémát vet fel. Célravezető lehet például a meglévő szövegek társas jelentésére koncentrálni, s azok kinyerésének automatizálását elvégezni. Ehhez nyújthat segítséget számunkra a szöveganalitika, ez a dinamikusan fejlődő, a tudományban és az üzleti szférában egyaránt sikerrel alkalmazott elemzési módszer.

Bár a jelen tanulmányban bemutatott kutatás elsősorban azokra a szófajokra és szófajmintázatokra alapozott, melyeket angolul „content word"-öknek, vagyis tartalmat kifejező szavaknak szokás nevezni, JAMES W. PENNEBAKER nyomán (2011) a későbbiekben érdemes lehet megvizsgálni kifejezetten a funkciószavakra (névmásokra, mondatszókra, indulatszókra, névelőkre, névutókra és kötőszókra) koncentrálva a modell hatékonyságát. Könyvében ugyanis PENNEBAKER azt veti fel, hogy ezek a méltánytalanul mellőzött szófajok még hatékonyabbak a személyiség- és stílusjegyek kirajzolásában.

Láthatjuk tehát, hogy a társas jelentés írott szövegekből való automatizált kinyerése nem lehetetlen feladat, hasznosságát pedig széles körben elismerik a tudományos köröktől az üzleti szféráig.

Irodalom

DARWIN, CHARLES ROBERT 1963. *Az ember és az állat érzelmeinek kifejezése.* Budapest, Gondolat Kiadó.

KOVÁCS-ÖRDÖG ZITA 2014. *Digitális testbeszéd.* Előadás. Elhangzott a Big Data: a Nagy Lehetőség vagy a Nagy Testvér? konferencián. Budapest, 2014. április 15.

KOVÁCS-ÖRDÖG ZITA–PANCZA JUDIT 2014. Digitális testbeszéd. *Marketingkutató 2014. tavasz:* 34–36.

LIU, BING 2012. *Sentiment Analysis and Opinion Mining.* Chicago, Morgan & Claypool Publishers.

MINER, GARY–DELEN, DURSUN–ELDER, JOHN–FAST, ANDREW–HILL, THOMAS–NISBER, ROBERT A. 2012. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Oxford, Academic Press.

PENNEBAKER, JAMES W. 2011. *The Secret Life of Pronouns: What Our Words Say About Us.* New York, Bloomsbury Publishing.

RÉGER ZITA 1990. *Utak a nyelvhez. Nyelvi szocializáció — nyelvi hátrány*. Budapest, Akadémiai Kiadó.

SÁNDOR KLÁRA 2001. Mobiltársadalom és nyelvhasználat: valami új vagy újra a régi? In: NYÍRI KRISTÓF szerk., *Mobil információs társadalom. Tanulmányok.* Budapest, MTA Filozófiai Kutatóintézete. 83–93.

SÁNDOR KLÁRA 2002. A nyelvi arisztokratizmus alkonya. In: NYÍRI KRISTÓF szerk., *Mobilközösség — mobilmegismerés*. *Tanulmányok.* Budapest, MTA Filozófiai Kutatóintézete. 67–77.

SÁNDOR KLÁRA–KAMPIS GYÖRGY 2000. Nyelv és evolúció. *Replika 40:* 125–143.

VARJÚ ZOLTÁN 2013. *Emócióelemzés, avagy Darwin és a nyelvtechnológia különös találkozása.* URL: http://kereses.blog.hu/2013/12/04/emocioelemzes_avagy_darwin_es_a_nyelvtechnologia_kulonos_talalkozasa. (2014. április 10.).

ZSIBRITA JÁNOS–VINCZE VERONIKA–FARKAS RICHÁRD 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: ANGELOVA, GALIA–BONTCHEVA, KALINA–MITKOV, RUSLAN szerk., *Proceedings of International Conference Recent Advances in Natural Language Processing*. Shouma, INCOMA Ltd. 763–771.