A helyek arcvonásai

Szerkesztette:

Michalkó Gábor – Rátz Tamara – Donka Attila

Kodolányi János Egyetem
HUN-REN CSFK Földrajztudományi Intézet
Magyar Földrajzi Társaság
Székesfehérvár–Budapest, 2024

https://doi.org/10.15201/TA13-4

Témák és érzelmek – Szállodai vendégvélemények vizsgálata témamodellezéssel és szentimentelemzéssel

Hinek Mátyás[1]

Bevezetés

Az elmúlt években óriási tömegű strukturálatlan szöveges információ generálódott az olyan turisztikai termékek és szolgáltatások véleményezését is lehetővé tevő oldalakon, mint a Tripadvisor, a Booking.com, a Google, az Airbnb stb., nem is beszélve a közösségi média különféle platformjairól. A felhasználó által generált tartalom (User Generated Content, UGC), illetve más meghatározás szerint az online szájreklám (electronic Word of Mouth, eWoM) jelentős hatással bír a fogyasztói döntésekre (Cheung – Thadani 2012), különösen a szállodaiparban, ahol nemcsak a potenciális vendégek, utazók tájékozódnak a kiválasztott szállodával kapcsolatos tapasztalatokról, de a maguk a szállodák is követik és válaszolnak a vendégbejegyzésekre (Bore et al. 2017). A vendégvélemények különösen fontosak a szállodai szolgáltatások területén, amelyek megfoghatatlanok, igénybevételük előtt nem kipróbálhatók, minőségük nem mindig konzisztens, és maga a szolgáltatás is összetett, különböző attribútumokkal írható le (Yen – Tang 2019).

A hatalmas tömegű szöveges információ elemzésére kézi módszerekkel nincs mód – hacsak nem néhány vendégbejegyzést szeretnénk megvizsgálni – de jelenleg a generatív mesterséges intelligencia alkalmazások is csak korlátozottan képesek a vendégvéleményeket összegezni és kivonatolni.

Tanulmányunkban arra teszünk kísérletet, hogy mintegy 60 budapesti szálloda, 1-3-as minősítésű vendégvéleményeit elemezzük, algoritmizált, gépi tanulási módszerrel, feltárva, hogy melyek a közepes és gyenge minősítésű vendégértékelések legfontosabb témái. Elemzésünkben megvizsgáljuk azt is, hogy a feltárt témák előfordulása hogyan változott az időben, milyen eltérések mutatkoznak a témák tekintetében a szállodai kategóriák között, valamint szentimentelemzéssel megvizsgáljuk azt is, hogy hogyan alakul a vélemények érzelmi töltete.

1. A számítógépes témamodellezés

A vendégvélemények, újságcikkek, tudományos publikációk és más nagy tömegben rendelkezésre álló szöveges dokumentumok feldolgozása, címkézése, strukturálása még

1 főiskolai tanár, Budapesti Gazdasági Egyetem, hinek.matyas@uni-bge.hu

számítógépes eszközökkel is óriási feladat. Bár az általános internetes keresőmotorok (pl. Google, Bing), valamint a speciális keresők (pl. Google Scholar vagy a Semantic Scholar a tudományos publikációk körében) sok tekintetben nagyon jól teljesítenek, valamint az elmúlt években megjelent generatív (és egyéb) AI alkalmazások is ígéretesek, azonban számos dokumentumtípushoz, szöveges tartalomhoz, termékekről és szolgáltatásokról írt véleményező oldalak bejegyzéseihez nincs még olyan címkéző, indexáló alkalmazás, amely támogatná ezeknek a szöveges információknak különböző szempontok szerinti visszakereshetőségét, strukturálását és feldolgozhatóságát.

Részben ezeknek a problémáknak a megoldására az elmúlt két évtizedben úgynevezett témamodellező algoritmusokat fejlesztettek ki az akadémia szektorban és az informatikai iparágban. Az egyik legkorábbi alkalmazás a látens szemantikai elemzés (Deerwester et al. 1990), a nem-negatív mátrix faktorizáció (Lee – Seung 1999), a látens Dirichlet allokáció (Blei et al. 2003), illetve ennek továbbfejlesztéseként a korrelált témamodell (Blei – Lafferty 2007), valamint a strukturális témamodell (Roberts et al. 2016). Ezek az eljárások többségében statisztikai, valószínűségszámítási alapokra épülnek, és az elmúlt két évtizedben számos kutatás és vizsgálat során bizonyították alkalmazhatóságukat.

A számítógépes nyelvfeldolgozás gyors fejlődésével az elmúlt időszaban ezeknél is fejlettebb megoldások jelentek meg, amelyek a természetes nyelven megírt szövegek numerikus vektorokká alakítására épülnek. Ez az úgynevezett beágyazási technika lehetővé teszi, hogy a szöveges adatokat olyan formátumban ábrázoljuk, amelyet a gépi tanulási algoritmusok is képesek feldolgozni. A hagyományos módszerekkel ellentétben, amelyek szóhalmaz (bag-of-words) modellekre támaszkodnak és a szöveges korpuszok elemzésére nagydimenziós ritka (sparse) vektorokat alkalmaznak, a szövegbeágyazások a szavak, mondatok vagy akár egész dokumentumok szemantikai jelentését sűrű vektorokban rögzítik. Ilyen modell a Word2Vec, amely a szavak beágyazását egy sekély neurális hálózat segítségével hozza létre nagy szövegkorpuszokból, megtanulva a szókapcsolatokat (Mikolov et al. 2013). Hasonló a BERT (Bidirectional Encoder Representations from Transformers), amely jelentős előrelépést jelentett a szavak közötti kontextuális kapcsolatok gépi megértésében (Devlin et al. 2019). A még jelenleg is fejlesztés alatt álló, a BERT-et alkalmazó BERTopic témamodellező algoritmus (de Groot et al. 2022) egyes vizsgálatok alapján jobb és szemantikailag könnyebben értelmezhető témákat képes generálni, mint a hagyományos statisztikai, valószínűségszámítási alapokra épülő témamodellező algoritmusok (Egger – Yu 2022). Ugyanakkor ezeknek az eljárásoknak is vannak korlátai, és alkalmazásuk sem terjedt még el széles körben a kutatók között.

1.1. A LÁTENS DIRICHLET ALLOKÁCIÓ (LDA)

A látens Dirichlet allokáció (LDA) egy nem felügyelt gépi tanulási algoritmus, amely a dokumentumok és az azokat alkotó szavak alapján valószínűségszámítási és statisztikai módszerekkel határozza meg, hogy a szöveggyűjtemény (korpusz) egyes dokumentumait milyen témák alkotják, és az egyes témákhoz milyen valószínűséggel tartoznak a szöveggyűjtemény alapján összeállított szótár (a szókincs) szavai.

Az LDA logikája a legkönnyebben egy képzeletbeli véletlenszerű folyamattal írható le, amely bemutatja, hogy a modell hogyan feltételezi az egyes dokumentumok keletke-

zését a korpuszban. E logika szerint úgy definiáljuk a témákat, mint a korpusz szókincse feletti eloszlást. Feltételezzük, hogy a dokumentumgyűjtemény témáit még az adatok (dokumentumok és szavak) generálása előtt határozzuk meg, azaz konkrétan tudjuk, hogy milyen témákat fog tartalmazni a korpusz. Az eljárás azt is feltételezi, hogy a témákat generáljuk először, és csak azután a dokumentumokat, így minden egyes dokumentumhoz kétlépcsős folyamatban rendeljük hozzá a szavakat a következők szerint:

1) Minden dokumentumhoz véletlenszerűen választunk egy témaeloszlást.
2) A dokumentumban található minden szó esetében
    2a) véletlenszerűen kiválasztunk egy témát az 1. lépésben meghatározott témaeloszlásból;
    2b) véletlenszerűen kiválasztunk egy szót az adott téma szókincs feletti valószínűségi eloszlása alapján.

Az eljárás eredményeként az egyes létrejött dokumentumok különböző arányban fogják tartalmazni a témákat (1. lépés), és minden egyes dokumentumban lévő minden szó a dokumentumokat alkotó témákból kerül kiválasztásra (2a és 2b lépés). Az LDA vegyes tagságú modell, mert egy dokumentumhoz több téma, illetve egy szó több témához is tartozhat, más-más valószínűséggel (Blei 2012).

Az így létrejött dokumentumok és a dokumentumokat alkotó szavak lesznek az LDA eljárás megfigyelt változói, míg a témastruktúra, a dokumentumonkénti témaeloszlások és a dokumentumokat alkotó szavak témakijelölései (azaz mely témákat, mely szavak, milyen valószínűséggel alkotnak) rejtett struktúrát alkotnak. A témamodellezés célja a témák automatikus feltárása, azonosítása a dokumentumgyűjteményből. Ez úgy is felfogható, mint az előzőekben leírt dokumentum-generálási folyamat megfordítása: keressük azt a rejtett témastruktúrát, amely a dokumentumgyűjteményt generálta. Eredményül a dokumentumok témaeloszlását és a témák szóeloszlását mint valószínűségi változókat kapjuk.

A témamodellezés központi számítási problémája, hogy hogyan lehet a megfigyelt dokumentumok és szavak alapján kikövetkeztetni a rejtett témastruktúrát. A generatív folyamat egy közös valószínűségi eloszlást határoz meg a megfigyelt és a rejtett változók felett, ahol a megfigyelt változók a dokumentumok szavai, a rejtett változó pedig a témastruktúra. Az adatelemzést úgy végezzük el, hogy ezt a közös eloszlást használjuk a rejtett változók feltételes eloszlásának kiszámításához a megfigyelt változók alapján. Ezt utólagos, más néven poszterior eloszlásnak is nevezzük, a számítási probléma pedig ennek kiszámítása a dokumentumok és a szavak ismeretében.

A lehetséges témastruktúrák száma exponenciálisan nagy, így a poszterior eloszlás nem számítható ki pontosan, csak közelíthető. A témamodellező algoritmusok a poszterior közelítésére jellemzően kétféle eljárást követnek: mintavételezést és variációs algoritmust. A mintavételezésen alapuló algoritmusok, mint például a Gibbs-mintavételezés, mintákat gyűjtenek a poszteriorból és azt egy empirikus eloszlással közelítik. A variációs algoritmusok a rejtett struktúra felett egy paraméterezett eloszláscsaládot állítanak fel, majd ennek a családnak a poszterior értékhez legközelebb álló tagját keresik meg (Blei 2012).

Mindkét algoritmus jelentős számítási kapacitást igényel, a közelítő poszterior eloszlás jellemzően többszáz iterációt követően határozható meg, amely során a dokumentum – téma és a téma – szó eloszlások újra meg újra kiszámításra kerülnek, mindaddig amíg a közelítés érdemben már nem javítható tovább, vagy elérjük azt az iterációs limitet, amelyet a kutató előzetesen beállított.

1.2. A STRUKTURÁLIS TÉMAMODELL (STM)

Az eredeti LDA modell továbbfejlesztéseként Robert és munkatársai fejlesztették ki a strukturális témamodellt, amely variációs következtetési algoritmust alkalmaz a témastruktúra modellezésére (Roberts et al. 2016). Legfontosabb újításuk az, hogy a dokumentumok tematikus tartalmának mélyebb elemzésére egy hierarchikus vegyes tagsági modellt dolgoztak ki, amely lehetővé teszi a dokumentumokhoz kapcsolódó metaadatok, például a szerzők neve, a megjelenés időpontja, a vendégértékelések esetében a vendégek által adott számszerű értékelések beépítését a témamodellbe. Ezzel lehetővé válik például annak vizsgálata, hogy hogyan alakul egyes témák előfordulása az időben, milyen tematikus különbség figyelhető meg a pozitív és negatív vendégértékelések között, van-e különbség a különböző kategóriájú szállodák vendégérétkeléseinek tematikus tartalma között.

1.3. TÉMAMODELLEZÉSRE ÉPÜLŐ KORÁBBI KUTATÁSOK

Az elmúlt években számos témamodellezést alkalmazó elemzés készült vendégvélemények vizsgálatára a turizmusban. A szükséges adatokat (vendégvéleményeket) jellemzően a Tripadvisor az Airbnb, illetve az OTA-k felületeiről gyűjtötték a kutatók, és az így kapott korpuszok általában több tízezer vendégvélemény elemzését tették lehetővé. Az 1. táblázatban tekintünk át néhány olyan tanulmányt, amely LDA vagy STM algoritmussal vizsgálta a vendégvélemények jellemzőit.

1. táblázat ***LDA vagy STM témamodellezést alkalmazó, turisztikai vendégvéleményeket vizsgáló tanulmányok az elmúlt évekből***

Szerzők	A vizsgálat témája	Módszertan és fontosabb eredmények
Ding et al. 2023	Az Airbnb-vel kapcsolatos vendégvélemények elemzése, 133 ezer vélemény alapján, Kuala Lumpurban	STM, figyelembe vett kovariánsok: Airbnb lista típusa és átlagára, illetve a bérelt lakás típusa. 21 azonosított téma, amely öt dimenzióba sorolható: létesítmények, szolgáltatás, elhelyezkedés, érték és általános tapasztalat. A teljes ingatlant (lakást) igénybe vevő vendégek preferenciái több tekintetben (ár, vendéglátóval kapcsolatos kommunikáció) eltértek a megosztott ingatlan igénybe vevőkétől.
Gao et al. 2022	Airbnb és szállodai vendégvélemények összeshasonlítása, összesen 33 ezer vendégvélemény elemzésével	STM, figyelembe vett kovariánsok: szállás típusa, szállás-értékelések, árak, és az értékelések szentimentértéke. Összesen 30 témát azonosítottak, ebből 12 téma gyakoribb volt az Airbnb értékelésekben. Pozitív szentimentek olyan témákhoz kapcsolódtak, mint a házigazda ajánlása és a helyszín, míg a negatív érzelmek a tisztasághoz és a kényelemhez kapcsolódtak. Az Airbnb élmények egyedi aspektusait jelenti az interakció a házigazdákkal és az ingatlan jellemzői.

Szerzők	A vizsgálat témája	Módszertan és fontosabb eredmények
Kirilenko et al. 2021	Attrakciók (Terrakotta-hadsereg, Vörös tér és Chichen Itza) látogatói vélemények elemzése, összesen 20 ezer vélemény alapján	LDA. A negatív vélemények kihívást jelentenek az olyan nem felügyelt adatbányászati algoritmusok számára, mint az LDA. A negatív véleményekben szereplő problémák változékonysága, valamint a negatív vélemények kisebb száma megnehezíti a témák azonosítását.
Korfiatis et al. 2019	Európai légitársaságok szolgáltatásminőségének vizsgálata, 557 ezer utasvélemény alapján	STM, a modellezés során figyelembe vett kovariánsok: utasok általános elégedettsége a légitársaság által a repülés során nyújtott szolgáltatással, a kabinosztály, a repülési távolság és az értékelők (utasok) Tripadvisoron való hozzájárulásának szintje. Az eredmények feltárták a fapadosok sikerét a légitársaságok versenyében.
Hu et al. 2019	Szállodai vendégek panaszainak okai, New York-i szállodák 28 ezer értékelése alapján (ugyanannyi negatív, 1-es 2-es értékelésű, és pozitív, kizárólag 5-ös értékelésű vélemény)	STM, figyelembe vett kovariánsok: negatív értékelések és a hotel minősítése (csillag). Összesen 30 téma került kifejtésre. 10 téma aránya a negatív véleményekben jelentősen magasabb volt. Az alacsony kategóriájú szállodák esetében a létesítményekkel kapcsolatos problémák, míg a magas kategóriájú szállodák esetében a szolgáltatással és az árazással kapcsolatos problémák a legfőbb elégedetlenségi okok.
Park – Ha, 2017	Az ügyfél kiszolgálás értékelése Las Vegas egyik előkelő szállodájában, összesen 5 ezer értékelés alapján	STM, a figyelembe vett kovariáns: a vendégértékelés dátuma. Összesen 11 témát azonosítottak: személyzet által nyújtott szolgáltatások, elhelyezkedés, kommunikáció, étkezési élmény, kényelem stb. A vizsgált időszak alatt a személyzet által nyújtott szolgáltatások téma részaránya emelkedett jelentősen.
Guo et al. 2017	A szállodai szolgáltatásminőséggel kapcsolatos elégedettség, 16 ország, 26 ezer szállodájának, 267 ezer Tripadvisor vendégvéleménye alapján	LDA, összesen 30 témát (szolgáltatásdimenziót) azonosítottak, ennek kétharmada a szálloda által kontrollálható tényező (be- és kijelentkezés, kommunikáció, szobaélmény stb.). A férfiak érzékenyebb az árakra, mint a nők, az idősebb szállodai ügyfelek jobban értékelik az otthonosságot. A 2-3 csillagos szállodák vendégei több, nem ár jellegű dimenziót azonosítottak. Az 5 csillagos szállodáknak az otthonosság érzésére kellene összpontosítaniuk.

Forrás: saját szerkesztés

2. Adatgyűjtés és módszertan

Adatainkat a Tripadvisor-ról gyűjtöttük az apify.com-on elérhető Tripadvior scraper alkalmazás segítségével. Összesen közel 100 ezer, 2004. január és 2023. április közötti húsz évben született, angol nyelvű vendégvéleményt töltöttünk le olyan budapesti szállodák Tripadvisor oldlairól, amelyeken legalább ezer vendégbejegyzés volt olvasható. Ezt a korpuszt csak a negatív vélemények megtartásával tovább szűkítettük. A szakirodalom megoszlik abban a tekintetben, hogy mely vendégvélemények tekinthetők negatívnak (Kirilenko et al. 2021). Jelen kutatásban úgy döntöttünk, hogy nem csak az 1-es 2-es értékelésű véleményeket tartjuk meg, hanem a 3-as értékelésűeket is, máskü-

lönben nagyon kicsi lett volna a vizsgált korpusz, ami nem kedvező a témamodellezés szempontjából. A szűrés eredményeképpen 64 szálloda 11174 angol nyelvű vendégvéleményét tartottuk meg. A 64 szállodából 3 volt kétcsillagos, 9 háromcsillagos, 43 négycsillagos és 9 ötcsillagos.

A strukturális témamodellezéshez az R statisztikai keretrendszer (R Core Team 2021) STM programcsomagját alkalmaztuk (Roberts et al. 2019). Az adatok előkészítése során a szöveg minden szavát kisbetűssé alakítottuk át, eltávolítottuk az írásjeleket és a speciális karaktereket, kizártuk a stopszavakat (pl. névelőket és határozószavakat, amelyek érdemben nem járulnak hozzá a szöveg jelentéséhez), majd „szótöveztük" a korpusz összes szavát, azaz levágtuk a szavak összes toldalékát, a képzőket, a jelzőket és a ragokat. Végül eltávolítottuk a leggyakoribb szavakat, azokat, amelyek legalább 2000 alkalommal előfordultak a korpuszban (ilyen szó volt például a hotel). Ennek oka az, hogy a nagyon gyakran előforduló szavak jellemzően „teleszemetelik" az egyes témákat, azaz majdnem minden témában megjelennek, így megnehezítik azok azonosítását.

A dokumentumszintű kovariánsokat a dokumentumokban előforduló témák arányainak moderálására építettük be. Az ehhez alkalmazott lineáris modell a következő volt:

topikprevalencia = kategória + értékelés + értékelés * megjelenés dátuma

Az LDA és az STM témamodellező eljárásoknak az a fő jellegzetességük, hogy a témákat automatikusan derítik fel a dokumentumban, azt azonban előre meg kell adni, hogy hány témát detektáljanak. A témák optimális számát az STM csomag által ajánlott diagnosztikai mutatók alapján határoztuk meg. Az egyik ilyen mutató a szemantikus koherencia volt, amely a témák emberi értelmezhetőségét mutatja, a másik a *held-out likelihood* (visszatartott valószínűség) mutatója, amely azt mutatja meg, hogy a modell hogyan tudja megragadni az adatok mögöttes szerkezetét. Mindkét mutató esetében a magasabb diagnosztikai érték kedvezőbb, így párhuzamos értékelésük alapján az optimális témaszámot 17 témában határoztuk meg. Az ennél alacsonyabb témaszám már nagyon „összenyomta" a dokumentumok témáit, azaz az eljárás során kifejtett témák valójában több mint egy témát tartalmaztak, míg a magasabb témaszám – amit később több próbafuttatással is ellenőriztünk – a témák értelmezhetőségét jelentősen rontotta, ekkor már több olyan téma is megjelent a modellben, amelyet egyáltalán nem tudtunk értelmezni.

A poszterior eloszlás közelítése során a maximális iterációk számát 150-ben határoztuk meg, ám ennyi iterációra a futtatások során nem volt szükség, a modell hamarabb konvergált.

3. Eredmények

3.1. TÉMÁK ÉS ARÁNYAIK A VENDÉGVÉLEMÉNYEKBEN

A témamodellezés végrehajtását követően a 17 topikhoz tartozó leggyakoribb valószínűségű szavakat, valamint az úgynevezett FREX szavakat kaptuk meg, amelyek egy-egy

topikban a legnagyobb valószínűséggel fordulnak elő, ugyanakkor kizárólagosak, tehát más topikokban nem, vagy ritkán fordulnak elő (2. táblázat).

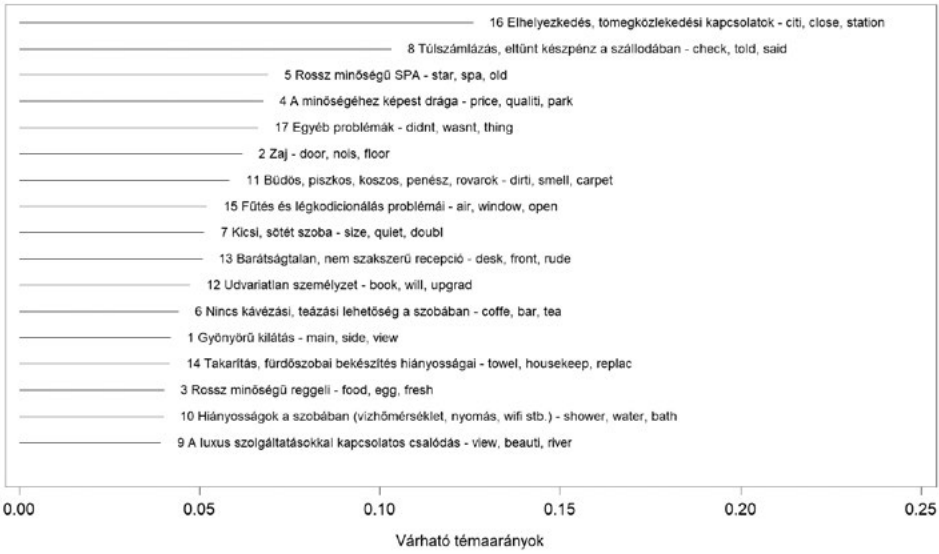
2. táblázat ***A strukturális témamodellezés során az egyes témákhoz azonosított szavak***

Téma száma	A témát alkotó legnagyobb valószínűségű szavak	FREX szavak
Topik 1	main, side, view, bridg, danub, river, build	buda, parliament, hill, pest, castl, andrassi, bridg
Topik 2	door, nois, floor, sleep, morn, next, peopl	hear, loud, alarm, woken, sound, disturb, hostel
Topik 3	food, egg, fresh, cold, fruit, order, eat	chees, bread, egg, fruit, meat, salad, sausag
Topik 4	price, qualiti, park, food, expens, better, euro	qualiti, expens, price, park, confer, car, garag
Topik 5	star, spa, old, pool, expect, rate, standard	star, spa, swim, massag, pool, sauna, outdat
Topik 6	coffe, bar, tea, tabl, facil, machin, drink	iron, tea, cup, coffe, machin, teacoffe
Topik 7	size, quiet, doubl, two, area, singl, space	size, quiet, twin, doubl, singl, larger, space
Topik 8	check, told, said, arriv, book, back, manag	card, email, credit, cash, confirm, refund, told
Topik 9	view, beauti, river, concierg, marriott, new, lobbi	york, marriott, boscolo, corinthia, intercontinent
Topik 10	shower, water, bath, use, bottl, floor, work	pressur, water, flood, shower, tub, tap, leak
Topik 11	dirti, smell, carpet, bad, chang, floor, smoke	smell, smoke, cigarett, non-smok, dust, dirti, filthi
Topik 12	book, will, upgrad, execut, custom, suit, loung	execut, kid, accor, children, famili, loung, custom
Topik 13	desk, front, rude, guest, check, call, experi	desk, front, rude, male, question, man, attitud
Topik 14	towel, housekeep, replac, tour, toilet, soap, group	shampoo, towel, soap, replac, tour, tissu, housekeep
Topik 15	air, window, open, work, hot, condit, sleep	air, condit, heat, degre, cool, temperatur, con
Topik 16	citi, close, station, metro, minut, centr, quit	metro, bus, station, centr, hop, transport, distanc
Topik 17	didnt, wasnt, thing, quit, bit, feel, dont	wasnt, feel, cant, bother, didnt, think, wouldnt

Forrás: saját szerkesztés az STM R csomag segítségével

Az eljárás során generált szavak alapján a témák azonosítása kutatói feladat. A 17 téma közül 16 azonosítása nem ütközött nehézségekbe, a 17. téma azonban nem volt egyértelműen azonosítható, így az „egyéb problémák” megnevezéssel címkéztük. A témák megnevezéseit és arányait a vendégértékelésekben az 1. ábra tartalmazza.

Az eredmények azt jelzik, hogy az 1-3-as minősítésű vendégvéleményekben nem kizárólag az elégedetlenség tényezői és vendégpanaszok jelennek meg. Érdekes módon az elhelyezkedés, tömegközlekedési kapcsolatok, amely inkább semleges téma, a leg-



1. ábra *Az azonosított témák és előfordulási arányuk a vizsgált vendégértékelésekben*
Forrás: saját szerkesztés az STM R csomag segítségével

gyakrabban fordul elő, vélhetően azért, mert a szolgáltatás egyik legfontosabb dimenziójáról van szó, az egyéb tényezőktől függetlenül. Szintén érdemes megfigyelni, hogy egy kifejezetten pozitív téma, a gyönyörű kilátás is szerepel a témák listájában, 4% körüli prevalenciával. Minden más téma (leszámítva 17. témát, ami nem volt jól azonosítható) negatív téma, de a semleges és a pozitív téma megjelenése rámutat arra, hogy a 3-as minősítésű vendégvélemények beemelése az elemzésbe, valamint a vélemények összetett jellege azt eredményezi, hogy nem kizárólag a negatív témák alkotják a korpuszt.

Ha a problémákat csoportosítjuk, akkor kiderül, hogy a szobával kapcsolatos problémák (6., 7., 10., 11., 14., 15. témák) előfordulása a teljes témaprevalencia közel negyedét képviseli. A másik jelentős arányt a túlszámlázás, az eltűnt készpénz, a nem megfelelő ár-érték arány miatti panaszok (4. és 8. téma) képviselik, együtt 17%-kos részaránnyal, míg a szállodai szolgáltatások (SPA és egyéb, 5. és 9. téma) 11%-ot, a személyzettel kapcsolatos problémák (udvariatlan, barátságtalan, nem szakszerű) 10%-ot képviselnek a korpuszban. A reggelivel kapcsolatos panaszok 4%-os részarányúak, és a szállodák számára a legtöbbször külső, környezeti adottságot jelentő zaj aránya 6,2%. Összességében azonban a szállodai szolgáltatások menedzsment és személyzet által befolyásolható elemei dominánsak a vendégpanaszokban.

3.2. A TÉMAARÁNYOK IDŐBELI VÁLTOZÁSA ÉS SZÁLLODAI KATEGÓRIÁK KÖZÖTTI KÜLÖNBSÉGEI

A vélemények megjelenésének idejét kovariánsként építettük a modellbe, így vizsgálni tudtuk, hogy hogyan változnak az egyes témák reprezentációi az idő múlásával. A legnagyobb változás az egyetlen pozitív téma, a gyönyörű kilátás esetében figyelhető meg:

amíg a vizsgált időszak elején (2004-2008 között) reprezentációja 15% körül alakult, ez az arány folyamatosan csökkent, 2023-ra 5% alá esett a vendégvéleményekben. A reprezentáció ilyen jelentős változása egyetlen téma esetében sem figyelhető meg, bár több téma részaránya is számottevően változott:

- a 16. elhelyezkedés, tömegközlekedési kapcsolatok téma reprezentációja 20%-ról, 13%-ra csökkent;
- a 3. rossz minőségű reggeli téma reprezentációja, 2%-ról 8%-ra nőtt;
- a 2. zaj téma reprezentációja 2%-ról 9%-ra nőtt;
- a 10. hiányosságok a szobában téma reprezentációja 1%-ról 5%-ra nőtt.

Más témák aránya a vendégvéleményekben nem változott ennyit, többségében csak enyhén nőttek, vagy stagnáltak, ritkább esetben kisebb mértékben csökkentek.

Ha szállodai kategóriák szerint szeretnénk megvizsgálni a témareprezentációk különbségeit, akkor az adataink elsősorban a négy- és az ötcsillagos szállodák vendégvéleményeinek összehasonlítására alkalmasak, mivel a vélemények több mint fele, 7813 vélemény négycsillagos szállodák körében született, 2085 vélemény pedig az ötcsillagosok körében. (A háromcsillagos szállodák véleményeinek száma 1035 db, a kétcsillagos szállodák értékeléseinek száma százas nagyságrendű volt.)

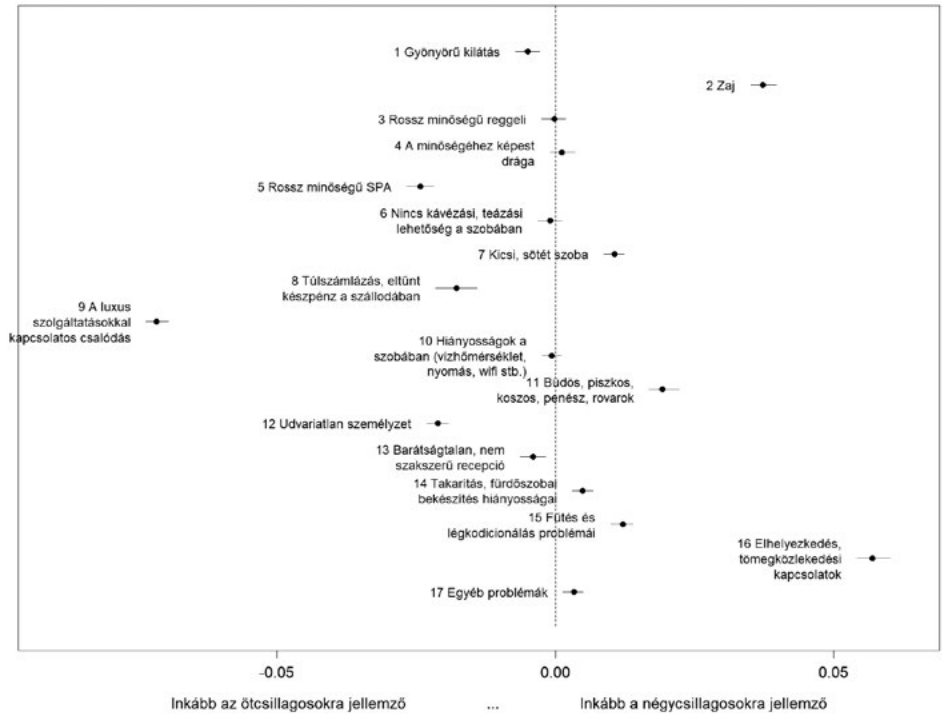
A témaprevalenciák néhány markáns esetben térnek el a négy- és ötcsillagosok között. Így

- a 9. luxus szolgáltatásokkal kapcsolatos csalódás 6%-kal magasabb arányban fordul az ötcsillagos szállodák véleményei körében, mint a korpusz egészében;
- a 10. elhelyezkedés tömegközlekedési kapcsolatok téma inkább a négycsillagos szállodák véleményeiben fordul elő, aránya közel 6%-kal magasabb, mint a korpusz egészében;
- a 2. zaj téma is inkább a négycsillagos szállodák véleményeiben fordul elő közel 4%-kal nagyobb arányban, mint a korpusz egészében.

Minden más téma jobbára kiegyensúlyozott a két szállodai kategória körében, és előfordulásuk közel van a korpusz egészére becsült előfordulási arányhoz. Ugyanakkor a spa-val kapcsolatos problémák (5. téma) inkább az ötcsillagosokra jellemzők, míg a 11. büdös, piszkos, penész, rovarok téma a négycsillagosok körében fordul elő valamivel gyakrabban (2. ábra).

3.3. A VENDÉGVÉLEMÉNYEK ÉRZELMI TÖLTETE

Mivel jórészt negatív vendégvéleményeket, panaszokat elemeztünk, így az előzetes várakozásunk az volt, hogy a vendégvélemények szentimentértékei is kedvezőtlenek lesznek. Azonban az értékelések szentimentértékei többségében inkább a semleges, enyhén pozitív tartományban találhatók (0 és 0,3 közötti pozitív értékek, ahol a 0 a semleges, a –1 az abszolút negatív, +1 az abszolút pozitív érzelmi töltetű értékelés, lásd a 3. ábrát!). Ennek egyik oka az, hogy a kritikát és a panaszt író vendégek a visszajelzésekben kevésbé használtak erős negatív érzelmi töltetű szavakat, azaz a vendégek többsége igyekezett tárgyszerűen fogalmazni az értékelések írásakor, és gyakran előfordult, hogy több témát



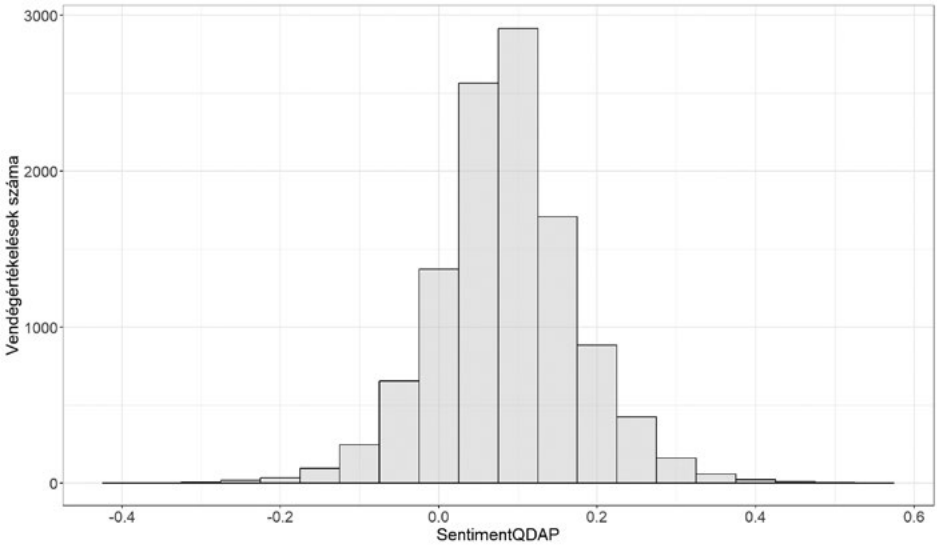
2. ábra *Témaarányok eltérései a négy- és ötcsillagos szállodák véleményei körében a korpusz átlagához viszonyítva*
Forrás: saját szerkesztés, az STM R csomag segítségével

érintve, kvázi komplex értékelés született, ami tartalmi szempontból is inkább neutrális töltetű értékeléseket eredményezett. Emellett érdemes megjegyezni, hogy az ötfokozatú skálán mért 3-as minősítésű vendégértékelések alkották a vélemények kétharmadát, ami szintén a neutrális szentimentértékek felé húzta az értékelések többségét.

Következtetések

A 64 budapesti szálloda, az elmúlt mintegy 20 évben íródott, több mint 11 ezer negatív (1-3-as minősítésű) vendértékeléseinek témamodellezése során hasonló, bár nem ugyanolyan struktúrájú és részletezettségű témákhoz jutottunk, mint amit a nemzetközi szakirodalom is feltárt, lásd egyebek mellett Hu és szerzőtársai (2019) vizsgálatát. A vizsgált értékelések nem csak negatív témákat, hanem neutrális és pozitív tényezőket is (elhelyezkedés, tömegközlekedési kapcsolatok, gyönyörű kilátás) tartalmaztak, ami jól jelzi a vendégvélemények összetettségét. Az általunk feltárt 17 téma alapján a vendégek elégedetlenségének forrásai főképp a szobával kapcsolatosak, a túlszámlázás, az eltűnt értéktárgyak és készpénz, a nem megfelelő ár-érték arány, a személyzet és az egyéb szol-

gáltatásokkal kapcsolatos tényezők mellett. Az elégedetlenségnek csak igen kis szeletét képviselték azok a (külső) tényezők, amelyekre nem tud hatni a szállodai menedzsment, ilyen például a zaj (3. ábra).



3. ábra *Az 1-3-as értékelésű szállodai vendégvélemények szentimentértékei*
Forrás: saját szerkesztés a ggplot2, valamint a QDAP szentimentszótár és R csomag segítségével

Kutatásunk tudatosan vállalt korlátokat is tartalmazott. Csak a negatív értékeléseket vizsgáltuk, ami az összes vendégvélemény mintegy tizedét alkotta, így ez az adatbázis nem feltétlenül tükrözi a teljes vendégkör véleményét, és nem tartalmazza a szállodai szolgáltatások összes, vendégek által értékelt dimenzióját. Magának az alkalmazott témamodellezési eljárásnak is vannak korlátai, a valószínűségszámítási-statisztikai módszerekre épülő, „bag-of-words" (szóhalmaz) eredményt generáló technikák sok tekintetben csak limitált témafelismeréshez vezetnek, és az így kapott eredmények gyakran nem validálhatók kézi elemzéssel, illetve nem feleltethetők meg más, például a szövegbeágyazásokra épülő témamodellezési módszerek eredményeinek. Hasonlóképpen, a szentimentelemzés is statisztikai modellen alapult, ami nem feltétlenül tükrözi a vendégek valós érzelmeit.

Összességében azonban a jelen vizsgálat megerősítette, hogy a szállodai szolgáltatásokkal kapcsolatos vendégelégedettség egy többdimenziós konstruktum, amelynek egyes elemei jól azonosíthatók a szöveges vendégvélemények automatizált elemzése során. Az elemzés rámutatott arra is, hogy mi az, amivel a vendégek elégedetlenek a szállodai tartózkodásuk során, illetve ezek a tényezők hogyan változnak az időben és a szállodai kategóriák között. Az eredmények jól hasznosíthatók a csalódásmenedzsment szállodai alkalmazása során is (Michalkó – Irimiás 2011).

Irodalom

Blei, D. M. (2012): Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.

Blei, D. M. – Lafferty, J. D. (2007): A correlated topic model of Science. *The Annals of Applied Statistics* 1(1): 17–35.

Blei, D. M. – Ng, A. Y., – Jordan, M. I. (2003): Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(4–5): 993–1022.

Bore, I. – Rutherford, C. – Glasgow, S., Taheri, B., – Antony, J. (2017): A systematic literature review on eWOM in the hotel industry: Current trends and suggestions for future research. *Hospitality – Society* 7(1): 63–85.

Cheung, C. M. K. – Thadani, D. R. (2012): The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems* 54(1): 461–470.

Deerwester, S. – Dumais, S. T. – Furnas, G. W. – Landauer, T. K. – Harshman, R. (1990): Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6): 391–407.

Ding, K. – Choo, W. C. – Ng, K. Y. – Zhang, Q. (2023): Exploring changes in guest preferences for Airbnb accommodation with different levels of sharing and prices: Using structural topic model. *Frontiers in Psychology* (14): 1120845.

Egger, R. – Yu, J. (2022): A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology* (7): 886498.

Gao, B. – Zhu, M. – Liu, S. – Jiang, M. (2022): Different voices between Airbnb and hotel customers: An integrated analysis of online reviews using structural topic model. *Journal of Hospitality and Tourism Management* (51): 119–131.

Guo, Y. – Barnes, S. J. – Jia, Q. (2017): Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management* (59): 467–483.

Hu, N. – Zhang, T. – Gao, B, – Bose, I. (2019): What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management* (72): 417–426.

Kirilenko, A. P. – Stepchenkova, S. O. – Dai, X. (2021): Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? *Tourism Management* (83): 104241.

Korfiatis, N. – Stamolampros, P. – Kourouthanassis, P. – Sagiadinos, V. (2019): Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications* (116): 472–486.

Lee, D. D. – Seung, H. S. (1999): Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755): 788–791.

Michalkó G. – Irimiás A. (2011): Csalódásmenedzsment a turizmusban: új szemlélet a turisztikai célterületek irányításában. *Marketing és Menedzsment* 45(2): 4–10.

Park, K. – Ha, S. (2017): Customer Service Evaluation based on Online Text Analytics: Sentiment Analysis and Structural Topic Modeling. *The Journal of Information Systems* 26(4): 327–353.

Roberts, M. E. – Stewart, B. M. – Airoldi, E. M. (2016): A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association* 111(515): 988–1003.

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

Yen, C.-L. A. – Tang, C.-H. H. (2019): The effects of hotel attribute performance on electronic word-of-mouth (eWOM) behaviors. *International Journal of Hospitality Management* (76): 9–18.

Online források

de Groot, M. – Aliannejadi, M. – Haas, M. R. (2022): *Experiments on Generalizability of BERTopic on Multi-Domain Short Text* (arXiv:2212.08459). arXiv. http://arxiv.org/abs/2212.08459 (letöltve: 2024. február 28.)

Devlin, J. – Chang, M. W. – Lee, K., – Toutanova, K. (2019): *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://arxiv.org/abs/1810.04805 (letöltve: 2024. február 28.)

Mikolov, T. – Chen, K. – Corrado, G. – Dean, J. (2013): *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. https://arxiv.org/pdf/1301.3781.pdf (letöltve: 2024. február 28.)

R Core Team. (2021): *R: A Language and Environment for Statistical Computing* [Software]. R Foundation for Statistical Computing. https://www.R-project.org/ (letöltve: 2024. február 28.)