Miről írnak a budapesti fine dining éttermek vendégei? Éttermi vendégvélemények témamodellezése neurális témamodellel

What do the customers of fine-dining restaurants write about? The themes-modelling of textual guest reviews of such restaurants with a neural topic modelling method

Szerző: Hinek Mátyás[1]

A tanulmány a budapesti fine dining éttermek szöveges vendégértékeléseinek témáit elemzi a BERTopic, egy neurális témamodellezési módszer, segítségével. A tanulmány 10.962 angol nyelvű, a Tripadvisorról származó, 2007 és 2024 márciusa között gyűjtött értékelést elemez. A hagyományos témamodellezési módszereknek korlátai vannak, különösen rövid szövegek esetében. A BERTopic a Sentence-BERT beágyazásokat kihasználva szemantikailag koherensebb témaazonosítást kínál. A vendégértékelések témamodellezése során 40 témát azonosítottunk, amelyek az éttermi szolgáltatás szinte minden aspektusát lefedik. Vizsgáltuk a számszerű vendégértékelések és az azonosított témák kapcsolatát, valamint azt, hogy az idő múlásával egyes témák aránya hogyan változott a véleményekben. A kutatás arra a következtetésre jutott, hogy bár a BERTopicnak vannak korlátai, ígéretesnek tűnik nagy mennyiségű szöveges adat elemzésében.

This paper analyses the themes of textual guest reviews of fine dining restaurants in Budapest using BERTopic, a neural topic modelling method. The study analyses 10,962 English-language reviews from Tripadvisor collected between 2007 and March 2024. Traditional topic modelling methods have limitations, especially for short texts. BERTopic offers semantically more coherent topic identification by utilising Sentence-BERT embeddings. In the topic modelling of guest reviews, 40 topics were identified covering almost all aspects of restaurant service. We examined the relationship between the number of guest reviews and the themes identified themes, and how the proportion of certain themes in the reviews changed over time. The research concluded that, although, BERTopic has limitations, it shows promise in analysing large amounts of textual data.

Kulcsszavak: neurális témamodellezés, éttermi vélemények, fine dining, BERTopic.
Keywords: neural topic modelling, restaurant reviews, fine dining, BERTopic.

1. Bevezetés

A közösségi média és a web 2.0 térnyerésével az interneten hatalmas mennyiségű szöveges ügyfél- és vendégvélemény vált hozzáférhetővé. Az ilyen típusú visszajelzések elemzése és értékelése az elmúlt másfél-két évtized során egyre nagyobb figyelmet kapott mind a gyakorlati alkalmazásokban, mind a tudományos kutatások terén.

Különösen a turisztikai szolgáltatások területén jelentős a vendégvélemények száma. Ezek a vélemények számos felületen elérhetőek, például szállásfoglalási oldalakon, keresőmotorokban, a közösségi médiában, valamint olyan komplex véleményportálokon, mint a Tripadvisor. E vélemények közvetlenül befolyásolják a potenciális utazók döntéseit, emiatt mind a szolgáltatók, mind a turizmus kutatói számára kiemelt fontosságúak.

Nagytömegű szöveges információ feldolgozására már a kétezres évek elején jelentek meg

[1] főiskolai tanár, Budapesti Gazdaságtudományi Egyetem, hinek.matyas@uni-bge.hu

gépi módszerek, amelyek részben vagy teljesen automatizált módon képesek elemezni a szöveges információkat, és az azóta eltelt időben további, egyre szofisztikáltabb eljárások születtek. Jelen tanulmányban egy új témamodellezési algoritmus, a BERTopic alkalmazásával tárjuk fel a budapesti fine dining éttermekről az elmúlt másfél évtizedben a Tripadvisor platformon megjelent több, mint tízezer vendégvélemény tipikus témáit.

2. Szakirodalmi áttekintés

2.1. HAGYOMÁNYOS TÉMAMODELLEK

A témamodellezés egyik legelterjedtebb, széles körben használt módszere a látens Dirichlet-allokáció (LDA). Az LDA egy valószínűségi algoritmus, amely azon a feltevésen nyugszik, hogy a dokumentumokat a szerzőik konkrét témákban írják meg, az adott témákhoz illeszkedő szókincset használva. Egy dokumentum több témát is tartalmazhat, és a dokumentum minden szava hozzárendelhető valamelyik témához. Az LDA ezt egy többlépéses generatív folyamatként fogalmazza meg: az adott dokumentumhoz először egy témaeloszlás kerül kiválasztásra, majd minden egyes szópozícióhoz hozzárendelnek egy témát, végül kiválasztanak egy szót az adott téma szókincsének eloszlásából. A témamodellezés során ezeket a témákat utólag kell visszafejteni. A dokumentumok témaeloszlásai, valamint a témák szóeloszlásai látens valószínűségi változókként jelennek meg, a feladat e valószínűségi eloszlások poszteriorainak becslése (BLEI 2012).

Az LDA-algoritmus az első lépésben véletlenszerűen rendel témákat a dokumentumokat alkotó szavakhoz, így létrejön egy kezdeti dokumentum

1.táblázat

Éttermi vendégvélemények probabilisztikus témamodellezésére épülő publikációk az elmúlt évekből

A tanulmány szerzői	*Vizsgált vendégvélemények (témamodellezési adatbázis)*	*Témamodellezés módszertana*	*A témamodellezés főbb eredményei*
Huang et al. 2014	*158.000 éttermi értékelés a Yelp Dataset Challenge adathalmazából*	*LDA*	*Főbb azonosított témák: kiszolgálás, ár-érték arány, dekoráció, egészségesség, elvitel, illetve időbeliség: reggeli, ebéd, vacsora.*
Park et al. 2018	*173.607 Yelp.com értékelés*	*STM (strukturális témamodell)*	*70 téma. A vizsgálat célja a „zöld" éttermek összehasonlítása a „nem zöld" éttermekkel, a két releváns topik a fenntartható ételek témában a vegán opciók, illetve a helyi/organikus összetevők.*
Westerlund et al. 2019	*3.302 vélemény az USA-ban működő indiai éttermekről*	*LDA*	*7 fő téma: szolgáltatás, ár, hangulat, kiszállítás, íz, rendelés és egyéb. Az ár és a hangulat különösen fontos tényezők az online értékelésekben.*
Kwon et al. 2020	*12.436 étterem 606.510 vendégvéleménye a Yelp.com-ról*	*STM Word2vec (vektoros, szóbeágyazásokra épülő modell)*	*65 téma, ebből 14 a vásárlói elégedettséget befolyásolja: észszerű ár, kiválóság, étel szeretete, hosszú várakozási idő, kreatív és egzotikus ételek, megközelíthetőség, középszerű íz, szolgáltatási hibák, gyenge minőség, tisztaság hiánya, udvariatlan hozzáállás, adag nagysága, a kültér hangulata.*
Zhang et al. 2022	*11.577 vendégvélemény a Yelp.com-ról (az eredményeket 14.399 mondat alapján összehasonlították emberi kódolással)*	*LDA és szentiment-elemzés, TextBlob segítségével*	*Az 5 leggyakoribb téma a Yelp.com véleményekben az ár, az étel, az idő, a szolgáltatás és a helyszín. Az emberi értékelők és az LDA által generált témák között 37%-62% egyezés volt.*
Aktas-Polat 2022	*25 étterem 22.104 Tripadvisor vendégvéleménye*	*LDA, és szentiment-elemzés osztályozása*	*5 fő témakör: szolgáltatás, élmény, meglepetés, íz és étel. A szolgáltatás témakörben volt a legmagasabb a pozitív érzelmi polaritás (96,4%), míg az ételfajtában a legalacsonyabb (87,4%).*
Kwon et al. 2022	*180 luxuskategóriába sorolt étterem 80.445 vendégértékelése a Yelp.com-ról*	*STM*	*Nagyszerű atmoszféra, ismételt látogatás, gusztustalan étel, széles menüválaszték, egyedi desszert, várakozási idő, finom étel, benyomás más vendégekről, különleges szolgáltatás évfordulóra, dekoratív bárszalon, túlárazás, negatív hírnév.*
Zhao–Liu 2023	*305.000 vendégvélemény, a Dianping.com-ról*	*Szentiment-elemzés, majd neurális hálózati modell alkalmazása*	*18 dimenzió, 5 csoportban: amelyek közül az étel volt a legfontosabb a vevői elégedettség és a visszatérési szándék szempontjából.*

Forrás: saját szerkesztés

– téma és téma – szó eloszlás. Mivel a kezdeti hozzárendelés véletlenszerű, a poszterior eloszlások valószínűségei alacsonyak lesznek, vagyis az azonos témákhoz tartozó szókincs az egyes dokumentumokban csak kevéssé fog egyezni. Ezután az eljárás Gibbs-mintavételezéssel vagy Expectation Maximization (EM, várható érték maximalizálási) algoritmussal egy iteratív folyamat során frissíti a téma-hozzárendeléseket, arra törekedve, hogy minden ismétlés során növekedjenek a poszterior valószínűségek. A folyamat addig tart, amíg a valószínűségek már nem javulnak tovább, vagy el nem érjük az előre beállított iterációs korlátot. A folyamat eredményeképp megkapjuk a dokumentumokat alkotó témák valószínűségi arányait és a hozzájuk tartozó legnagyobb valószínűségű szavakat (BLEI et al. 2003, GRIFFITHS–STEYVERS 2004, BLEI 2012).

Bár a látens Dirichlet-allokáció népszerű témamodellezési módszer, az így nyert eredmények nem mindig megbízhatóak, lévén az LDA eredményei az eljáráshoz szükséges kezdeti hiperparaméterek változtatásával módosulhatnak, illetve a Gibbs-mintavételezés miatt a megismételt futtatások is eltérő eredményekhez vezethetnek (LOVATO et al. 2015, GEORGE–DOSS 2018, RIEGER et al. 2020). A témák számát előzetesen kell megadni, amire nincs egyértelmű módszer (CHEN–DOSS 2019). Bár a különböző mérőszámok alapján (pl. szemantikai koherencia, zavarosság stb.) meg lehet becsülni valamilyen optimális témaszámot, ez a gyakorlatban nem mindig vezet az ember által is jól értelmezhető, koherens témák azonosításához. Gyakori problémát jelentenek az egyes témákba *betolakodó szavak*, amikor egy téma szavai között oda nem illők szerepelnek, de előfordulhatnak *betolakodó témák* is, amikor az algoritmus által a dokumentumokhoz társított témák nem egyeznek meg az emberi megítéléssel (CHANG et al. 2009, ZHANG et al. 2022).

A probabilisztikus témamodellezés rövid szövegek esetében gyakran rosszul teljesít, holott a közösségimédia-bejegyzések általában rövid és zajos szövegek (HA et al. 2019). A tanulmány témáját képező éttermi vendértértékelések általában rövidek, 4-5 mondat hosszúságúak, az éttermekre jellemző természetes, értékelhető szempontok gyakran specifikusak (például az olasz éttermek esetében a pizza és a tészta, míg a japán éttermek esetében a sushi). Ezek a szempontok az általános modellekkel kevésbé kezelhetők (TITOV–McDONALD 2008). E problémákra az alkalmazott modellezési eljárások fejlesztésével reagáltak a kutatók az elmúlt években, ám ezek széles körű alkalmazása nem jellemző (CHENG et al. 2014, ALBALAWI et al. 2020, QIANG et al. 2022, ZUO et al. 2023). A fenti problémák ellenére az LDA és más probabilisztikus eljárások segítségével több tanulmányban is beszámoltak az éttermi vendégvélemények témamodellezésének eredményeiről *(1. táblázat).*

2.2. NEURÁLIS NYELVI MODELLEKRE ÉPÜLŐ TÉMAMODELLEZÉS

A témamodellezési probléma megoldására az elmúlt években új eljárásokat fejlesztettek ki. Ezek közül a transzformátor-alapú, neurális hálózatokat alkalmazó modellek tűnnek a legígéretesebbnek. Az egyik legelterjedtebb transzformátor-alapú nyelvi modell a Google által fejlesztett BERT (DEVLIN et al. 2019), ami az írott szövegekből numerikus reprezentációkat, vektorokat állít elő. Az így készített szövegbeágyazásokat számos különböző nyelvi feladat megoldásához lehet alkalmazni, például szövegosztályozásra, tudományos szakcikkek témáinak azonosítására, orvosi és más tudományos publikációk elemzésére, szöveganntotációs feladatok elvégzésére, valamint a szöveggenerálás minőségének értékelésére (GLAZKOVA 2021, KOROTEEV 2021).

A BERT-alapú BERTopic egy olyan szövegosztályozási módszer, amely az írott információkat több feldolgozási lépést követően témákba, egészen pontosan klaszterekbe sorolja, amelyeket a legjellemzőbb szavaikkal reprezentál. A folyamat lépései a következők: (1) a szöveg vektortérbe történő beágyazása Sentence-BERT-tel, vagy esetleg egy hasonló nyelvi modellel; (2) a szavak és mondatok kontextusát reprezentáló vektorok dimenziócsökkentése a könnyebb és gyorsabb feldolgozhatóság érdekében; (3) a dimenziócsökkentett vektorok osztályozása klaszterező algoritmus segítségével, végül (4) a klaszterezett adatokból a c-TF-IDF (*class-based Term Frequency – Inverse Document Frequency*, klaszteralapú kifejezésgyakoriság – inverz dokumentumgyakoriság) eljárás alkalmazásával a legjellemzőbb szavak kinyerése (GROOTENDORST 2022).

Mivel viszonylag friss innovációról van szó, a BERTopic eljárás alkalmazása még nem terjedt el széles körben. Ennek ellenére néhány közelmúltban készült tanulmány már vizsgálta a BERTopic hatékonyságát. ABUZAYED és AL-KHALIFA (2021) arab nyelvű szövegeken hasonlították összse a BERTopic-ot az LDA és az NMF (nemnegatív mátrixfaktorizáció) témamodellező technikákkal, és arra a következtetésre jutottak, hogy a BERTopic általában jobb eredményeket produkál, mint a hagyományos LDA és NMF technikák.

EGGER és YU (2022) négy témamodell-algoritmus – az LDA, az NMF, a Top2Vec és a BERTopic – teljesítményét értékelték több, mint

harmincezer, a Covid19-járvánnyal és utazással kapcsolatos Twitter-bejegyzés témamodellezése során. Eredményeik azt mutatják, hogy a BERTopic és az NMF hatékonyabb a rövid szöveges adatok elemzésében, mint a Top2Vec vagy az LDA.

Egy egyetemi kutatásban több, mint hatvanezer hallgatói kurzusértékelést vizsgáltak témamodellezéssel, és megállapították, hogy a BERTopic jobb eredményeket ért el a sokféle témát és területet érintő rövid szövegek modellezésében, mint az LDA (DE GROOT et al. 2022).

OGUNLEYE és munkatársai (2023) a BERTopic algoritmust alkalmazták nigériai banki ügyfelek tweetjeinek témamodellezése során, amellyel jobb eredményeket értek el, mint a hagyományos módszerekkel.

KRISHNAN (2023) összehasonlító tanulmányban értékelte a leggyakrabban alkalmazott témamodellezési eljárásokat, beleértve a látens szemantikai elemzést (LSA), a látens Dirichlet-allokációt (LDA), a nemnegatív mátrixfaktorizációt (NMF), a Pachinko allokációs modellt (PAM), valamint a szövegbeágyazásokra épülő, szövegosztályozó Top2Vec és BERTopic eljárásokat. Eredményei alapján a vásárlói vélemények témamodellezésében a BERTopic több értelmes téma kifejtését eredményezte, és kedvezőbb eredményeket ért el, mint a többi eljárás.

ALAMSYAH és GIRAWAN (2023) a természetes nyelvfeldolgozási technikákat alkalmazták a termékminőség javítása és a hulladék csökkentése témákban a divatiparra vonatkozóan. A BERT-hez hasonló RoBERTa-modellt használták a többcímkés osztályozáshoz, a BERTopic-ot pedig a ruhatermékek fogyasztói véleményeinek témamodellezésére. A RoBERTa nagy pontosságot ért el a véleménytémák beágyazásában, míg a BERTopic-kal koherens és releváns témákat azonosítottak.

3. Módszertan

Vizsgálatunk alapvető célja azt volt, hogy fine dining éttermek szöveges vendégvéleményeiben tipikus témákat azonosítsunk a BERTopic eljárás alkalmazásával, valamint feltárjuk a vélemények egyéb jellegzetességeit. Vizsgálatunk alanyait 2023 novemberében a Michelin Guide-ban szereplő 28 budapesti étterem jelentette. Az éttermek szöveges vendégvéleményeit a Tripadvisorról válogattuk le az Apify webkaparó alkalmazás segítségével, majd az így nyert információk alapján végül 25 étteremre szűkítettük a vizsgálatot. A témamodellezés elvégzéséhez csak az angol nyelvű véleményeket tartottuk meg, melyekből összesen 10.962 darabot találtunk, 2007 és 2024 tavasza között *(2. táblázat)*.

2. táblázat

A vizsgálatba bevont éttermek angol nyelvű vendégvéleményeinek száma a Tripadvisoron 2007 és 2024 márciusa között

Étterem	*Vendég-vélemény*	*Étterem*	*Vendég-vélemény*
Arany Kaviár Restaurant	*540*	*Onyx Műhely*	*1475*
Babel Budapest	*375*	*Rutin*	*1*
Borkonyha WineKitchen	*1809*	*Salt*	*41*
Costes Downtown	*761*	*Solid*	*9*
Costes Restaurant	*749*	*Spago By Wolfgang Puck Budapest*	*86*
Essencia Restaurant - Tiago & Eva	*35*	*St. Andrea Restaurant*	*492*
Felix Kitchen & Bar	*143*	*Stand Restaurant*	*209*
Flava Kitchen & More	*18*	*Stand25 Bisztró*	*216*
Fricska 2.0	*337*	*TATI Farm To Table*	*56*
Hoppa Bistro	*1599*	*Textura*	*119*
Laurel Budapest	*163*	*Umo Restaurant & Grill*	*12*
MAK restaurant	*701*	*Zincenco Kitchen*	*232*
Nobu Restaurant	*784*	*Összesen*	*10.962*

Forrás: saját adatgyűjtés a tripadvisor.com-ról

Ugyan a BERTopic eljárás jobb eredményeket produkál a szövegek szemantikai tartalmának megragadásában, mint más eljárások, nem képes reprezentálni a dokumentumok tematikus összetettségét. Ha a szövegbeágyazás a teljes dokumentumok alapján történik, és az így kapott vektorokat klaszterezi az eljárás, akkor minden dokumentum egyetlen klaszterbe kerül, azaz minden dokumentumhoz egy téma-hozzárendelés jön létre. Ez azonban a szöveges vendégértékelések elemzésénél kifejezetten hátrány. Bár a vendégértékelések jellemzően rövidek, ám gyakran több témát is érintenek, lásd például az alábbi vendégértékelést, amely egyszerre szól a kiváló ételekről és borokról, a kiváló kiszolgálásról, az ár/érték arányról, az ételek összetevőiről, az étterem atmoszférájáról és még sok másról:

> *„If you have a chance to visit Budapest don't miss Fricska gastropub! It's always a real pleasure trip into gastronomy and also a good value for money. Excellent food with excellent wines and serving is superb. Ingredients are always fresh and best quality, dishes are sooo delicious, seasonal, healthy and exciting. Menu is changing every day so it will never get boring. When I visit Fricska for lunch or dinner*

it always makes my day. Fricska has a very nice atmosphere, staff is so friendly, polite and very professional. Sommelier Csaba can offer you great wines, their wine portfolio is huge and outstanding, mostly Hungarian wines and sparkling wines, one of the best selection I've ever seen. Daily menu at lunchtime is amazing and for very good price. I wish I could visit them every day!" (Tripadvisor vendégvélemény a Fricska 2.0 étteremről, a bejegyzés dátuma: 2017.03.03.)

Ennek kiküszöbölésére GROOTENDORST (2022) két megoldást javasolt. Az egyik lehetőség, hogy a szöveget kisebb részekre – bekezdésekre vagy mondatokra – bontjuk, és így végezzük el a témamodellezést, bár ez Grootendorst szerint nem vezet ideális témareprezentációhoz. A másik lehetőség, hogy a dokumentumok témamodellezését követően, miután kialakultak a teljes dokumentumhalmazra jellemző témák, minden dokumentumot tokenekre, lényegében szavakra bontunk. Ezután egy úgynevezett *csúszóablakot* alkalmazva, egyszerre 4-10 szót elemezve megvizsgáljuk, hogy ezek mennyiben hasonlítanak (reprezentálják) a korábban generált témákat, és így határozzuk meg egy-egy dokumentum témaeloszlását.

Jelen kutatásunk során az utóbbi eljárással az volt a probléma, hogy a vendégvélemények annyira sokféle témát tartalmaztak, hogy a próbafuttatások során a BERTopic nem volt képes klaszterezni a teljes szövegű vendégvéleményeket, így használható témák sem jöttek létre. Emiatt az első megoldást választottuk, a véleményeket mondatokra bontottuk (összesen 64.473 mondatra), és ezeket ágyaztuk be a BERT segítségével.

4. Eredmények

A BERTopic eljárás alapbeállításai közel 100 témát generáltak, ami kezelhetetlenül soknak bizonyult, rengeteg alig különböző témával. Emiatt néhány paramétert finomítottunk, például az elvárt minimális klaszternagyságot legalább 250 mondatban határoztuk meg, amivel a témaszámot 40-re csökkentettük. A 40 témából az egyik úgynevezett *outlier* téma volt, amely azokat a mondatokat tartalmazta, amelyeket a HDBSCAN klaszterezési eljárás nem osztályozott, jellemzően azért, mert túl *zajosnak* találta a szöveget. Az algoritmus a közel 65.000 mondatból majdnem 20.000 mondatot (30%) sorolt be az *outlier*-ek közé, és ez az arány nem csökkent érdemben akkor sem, amikor magasabb témaszámmal futtattuk a modellt. Az *outlier*-ek csökkentéséhez *keményebb* klaszterezési eljárásra lett volna szükség, amire szintén lehetőséget ad a BERTopic, de ebben az esetben szemantikailag egymástól távol eső mondatok is egy klaszterbe kerülhettek volna, emiatt úgy döntöttünk, hogy megtartjuk az eredeti megoldást.

3. táblázat

A BERTopic által generált első néhány téma a hozzájuk rendelt mondatok számának csökkenő sorrendjében (a -1 nem téma, a nem klaszterezett mondatokat jelenti)

Téma-szám	*Mondatok száma*	*Reprezentatív szavak*	*A ChatGPT által javasolt témamegnevezés*
-1	*19.678*	*restaurant, food, place, menu, service, great, good, experience, time, dinner*	*Excellent Budapest dining experience*
0	*5231*	*food, dishes, birthday, delicious, dinner, taste, presented, dish, meal, presentation*	*Delicious beautifully presented food*
1	*4925*	*wine, wines, pairing, wine pairing, menu, food, food wine, course, menu wine, selection*	*Wine Pairing Menus*
2	*4035*	*soup, duck, fish, beef, liver, pork, main, starter, venison, dish*	*Duck liver excellence*
3	*1941*	*table, reservation, booked, tables, advance, booking, make, reservations, make reservation, restaurant*	*Restaurant reservations during busy times*
4	*1853*	*michelin, michelin star, star, restaurant, star restaurant, michelin starred, starred, restaurants, michelin restaurant, starred restaurant*	*Michelin star restaurant reviews*
5	*1683*	*restaurant, recommend restaurant, recommend, restaurant restaurant, restaurants, highly, nice, atmosphere, highly recommend, restaurant nice*	*Excellent restaurant experiences*
6	*1647*	*service, food, food service, staff, great, service food, excellent, atmosphere, great food, friendly*	*Excellent food and service*
7	*1554*	*budapest, restaurant, best, restaurant budapest, meal budapest, meal, budapest best, visit, restaurants, budapest restaurant*	*Best Budapest restaurant experience*

Forrás: saját szerkesztés

A BERTopic a témák azonosítását azzal könnyíti meg, hogy akár többféle reprezentációs modell alapján közli az egyes klaszterek c-TF-IDF eljárással generált reprezentatív szavait, és az adott klaszterekre jellemző példamondatokat is megad. Ezek betáplálhatók nagy nyelvi modellekbe, például a ChatGPT-be, ami javaslatot tesz a témák elnevezésére *(3. táblázat)*.

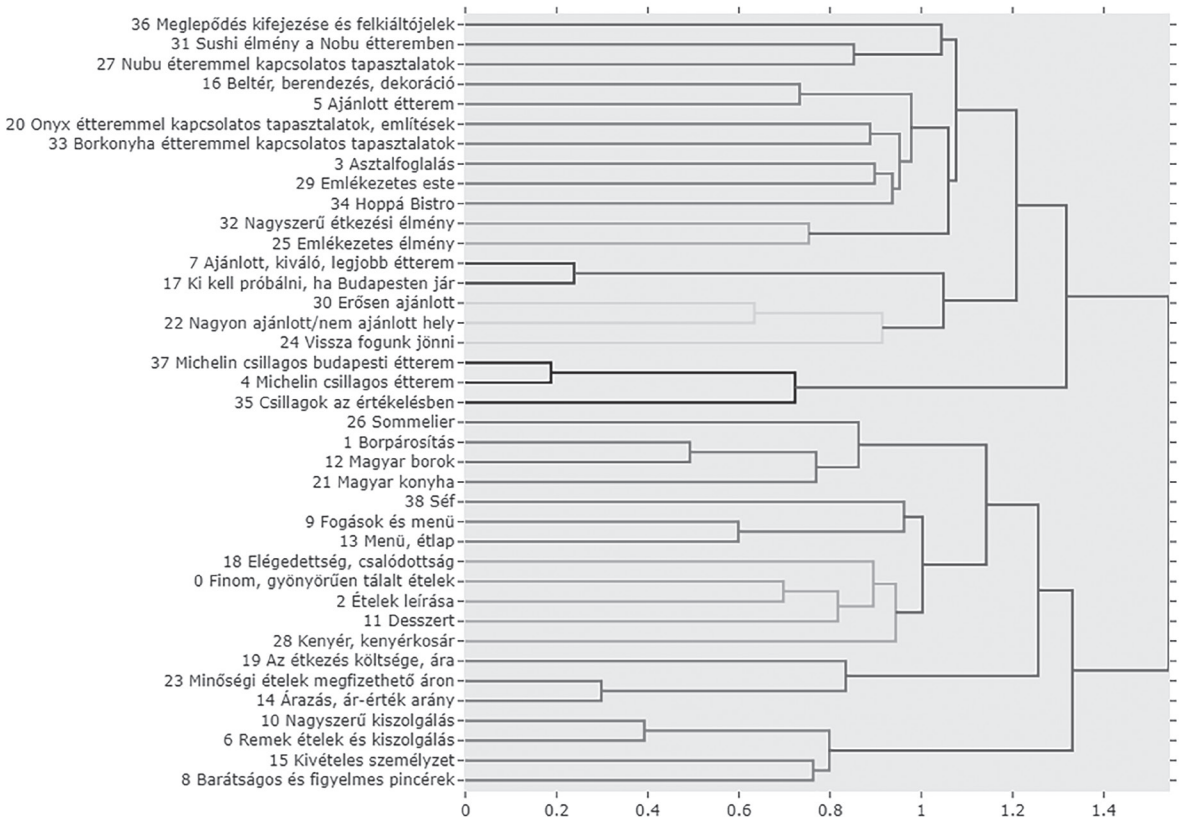
A 3. táblázat szerint a vendégértékelésekben leggyakrabban előforduló téma a 0 jelzésű, amely a reprezentatív szavak alapján a finom, gyönyörűen tálalt ételekről szól, és a példamondatok alapján a ChatGPT is erre az elnevezésre tett javaslatot. A témához az algoritmus szerint 5.231 mondat tartozik. A második leggyakoribb téma a bor és borpárosítás, amelyhez 4.925 mondat tartozik, míg a harmadik az ételek leírásával kapcsolatos téma 4.035 mondattal.

A ChatGPT által javasolt elnevezéseket nem fogadtuk el automatikusan. Az egyes témákhoz tartozó mondatokat nagyobb számban átnéztük, így a végső témaelnevezéseket a reprezentatív szavak, a témához tartozó mondatok és a ChatGPT ajánlásai alapján kutatói döntéssel határoztuk meg. Az *1. ábra* mutatja a 39 téma kutatói megnevezését és azt, hogy az algoritmus szerint mely témákat lehetne összevonni, amennyiben a témák számát tovább kívánjuk csökkenteni.

Az 1. ábrán látható, hogy főként a szemantikailag összetartozó témák összevonására kaptunk javaslatot, például a 12 Magyar borok és az 1 Borpárosítás, vagy a 32 Nagyszerű élmény és a 25 Emlékezetes élmény esetében, illetve az árazással kapcsolatos témáknál (19, 23, 24). Ám a 2 Asztalfoglalás, a 29 Emlékezetes este és a 34 Hoppá Bisztró témákat emberi értékelés alapján már nem biztos, hogy érdemes lenne összevonni. A hierarchikus klaszterezés ajánlásait figyelembe véve a hasonló témák összevonását kutatói értékelés alapján hajtottuk végre, a végső témastruktúrát a *4. táblázat* tartalmazza.

1. ábra

A BERTopic által generált 39 téma és a témák kapcsolatai (a témák hierarchikus klaszterezése) a BERTopic algoritmus szerint



Forrás: saját szerkesztés

4. táblázat

Az összevont témastruktúra és az egyes témákhoz rendelt mondatok aránya

Témák	*A témához tartozó mondatok aránya*	*Témák*	*A témához tartozó mondatok aránya*
0 Outlier és klaszterekbe nem sorolt	*31,0%*	*10 Konkrét éttermekkel kapcsolatos tapasztalatok*	*3,6%*
1 Bor, magyar borok, borpárosítás	*9,3%*	*11 Emlékezetes élmény*	*3,4%*
2 Finom, gyönyörűen tálalt ételek	*8,1%*	*12 Asztalfoglalás*	*3,0%*
3 Az étterem ajánlása	*8,0%*	*13 Desszertek leírása*	*1,7%*
4 Ételek leírása	*6,3%*	*14 Interior, berendezés, dekoráció*	*1,5%*
5 Nagyszerű kiszolgálás, remek ételek	*4,7%*	*15 Magyar konyha*	*1,0%*
6 Michelin csillagos étterem Budapesten	*3,9%*	*16 Visszatérési szándék*	*1,0%*
7 Menü, étlap, fogások	*3,8%*	*17 Sommelier*	*0,9%*
8 Személyzet, pincérek	*3,8%*	*18 Kenyér, kenyérkosár*	*0,8%*
9 Ár, árazás, ár/érték arány	*3,7%*	*19 Séf*	*0,4%*

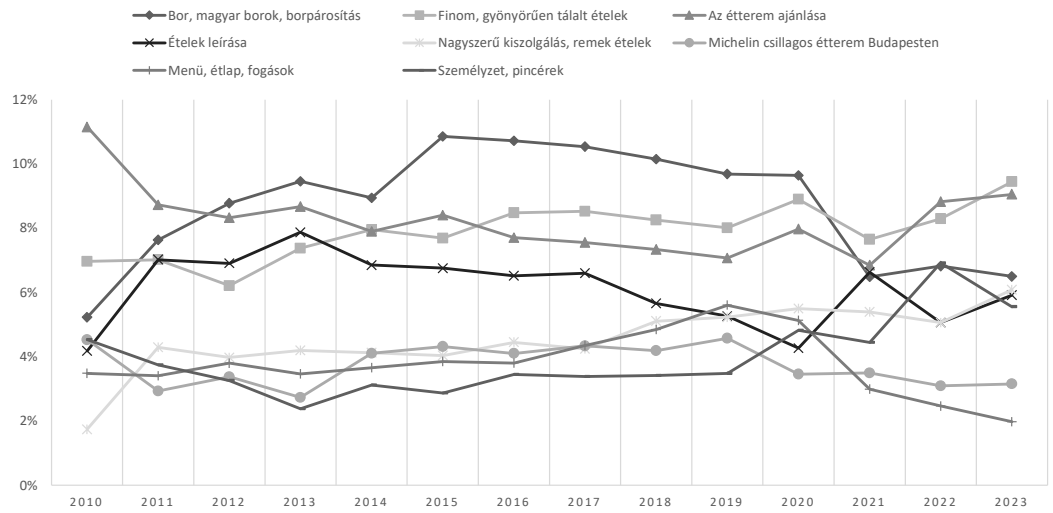
Forrás: saját szerkesztés

A témákat áttekintve látható, hogy az éttermi vendégértékeléseket témamodellező korábbi kutatásokban (1. táblázat) megjelenő témákat részben átfedő témákat kaptunk (ár/érték arány, ételek, széles menüválaszték, különleges desszertek, szolgáltatás, kiválóság, élmény, atmoszféra stb.), illetve egyes témák, különösen a kisebb arányban megjelenők egyediek, azaz kifejezetten az általunk vizsgált vendégvéleményekre jellemzőek (pl. magyar konyha, sommelier, kenyérkosár).

Míg a hagyományos témamodellek eredményeiben több esetben megjelennek negatív témák (túlárazás, nem megfelelő minőségű étel, tisztaság hiánya stb.), a BERTopic által azonosított témák között nincsenek negatív hangvételűek, noha az általunk elemzett több, mint 10.000 vendégértékelés sem mentes a negatív vendégvéleményektől. Áttekintve néhány tucat 1-es és 2-es értékelésű véleményt, a vendégpanaszok az éttermi szolgáltatások szinte minden területét érintik, ám ezek

2. ábra

Az egyes témákat tartalmazó mondatok aránya évente, a hét legnagyobb arányban megjelenő téma esetében



Forrás: saját szerkesztés

mindössze 5%-át képviselik az összes véleménynek. Mivel a BERTopic algoritmus paraméterezése során minimális klaszternagyságot kötöttünk ki, és az elemzett mondatok 30%-át az eljárás nem sorolta klaszterekbe, így nem teljesen váratlan, hogy nem jelent meg negatív tartalmú témacímke.

A számszerű vendégértékelések témánkénti átlagai között sem figyelhető meg jelentős eltérés. A 25 étteremre vonatkozó, több, mint 10.000 vendégértékelés együttes átlaga 4,6 az ötfokozatú skálán, és az egyes témák esetében is 4,3 és 4,7 közöttiek az átlagok. Csupán egy téma, az asztalfoglalás, részátlaga marad el öt tizeddel a teljes átlagtól, azaz a szolgáltatásnak ezzel a dimenziójával voltak legkevésbé elégedettek a vendégek. Azonban fontos megjegyezni, hogy a témánként képzett átlagok nem teljesen megbízhatóak, mivel az átlagolást az egyes mondatok kapcsán végeztük, miközben a számszerű vendégértékelések a teljes, több mondatot tartalmazó bejegyzésekre vonatkoztak.

Mivel a Tripadvisor értékelések esetében rendelkezésre állt az értékelés közzétételének dátuma így lehetőség nyílt arra is, hogy a témák időbeli változását vizsgáljuk *(2. ábra)*. Látható, hogy a borral kapcsolatos téma aránya az elmúlt közel másfél évtized első felében nőtt, azután csökkent. 2023-ban mintegy 6,5%-os részarányú volt. A finom, gyönyörűen tálalt ételek téma részaránya az utóbbi években 9% fölé emelkedett, és ugyancsak emelkedett az ehhez szorosan kapcsolódó téma, a nagyszerű kiszolgálás, aránya. Szintén emelkedett valamenynyit a személyzettel, pincérekkel foglalkozó mondatok aránya, míg minden más téma inkább stagnált vagy csökkent, de drámai változás egyik téma esetében sem figyelhető meg.

5. Összefoglalás

A neurális nyelvi modellekre épülő témamodellezés releváns megoldást jelent több olyan problémára, amit a probabilisztikus témamodellek rosszul kezelnek. A neurális hálózatok segítségével létrehozott vektoros beágyazások képesek megragadni a szövegek szemantikus tartalmát, ami a témamodellezés során pontosabb témakijelöléseket tesz lehetővé. Ez az éttermi vendégvélemények témamodellezésében is jól megfigyelhető, az algoritmus az egyes témákhoz koherens és értelmes szöveges információkat rendelt. A hagyományos modellekkel szemben ez a típusú témamodellezés jól teljesít a rövid szövegek esetében is. A BERTopic eljárás további előnye, hogy a témák számát nem kell előre meghatározni, az algoritmus maga csoportosítja klaszterekbe az elemzett szöveget.

Az éttermi vélemények témái jól lefedték az éttermi szolgáltatás egyes dimenzióit, az asztalfoglalástól, az ételeken, a tálaláson és a személyzet munkáján át, egészen a fine dining éttermek által nyújtott vendégélményig. Ez összecseng a korábbi kutatásokkal, sőt sok tekintetben pontosabb, és az algoritmus beállításaitól függően akár részletesebb témakatalógushoz juthatunk, mint a hagyományos módszerek esetében.

Ugyanakkor jelentős korlátok is adódnak a BERTopic alkalmazása során. Jelen vizsgálatban a legfontosabb korlátot az jelentette, hogy az egyes vendégvélemények nem kezelhetők a vizsgálat egységeként, mert az eljárás nem képes több témát azonosítani egyazon recenzióban. Ez részben orvosolható azzal, ha bekezdésekre, mondatokra bontjuk a szöveget, de a mondatonkénti témakijelölések sem mindig reprezentálják megfelelően a vélemények témaeloszlását.

A másik potenciális – bár átléphető – korlát az, hogy a BERTopic alapértelmezett klaszterezési eljárása, a HDBSCAN, viszonylag jól biztosítja, hogy szemantikailag hasonló szövegek kerüljenek egy klaszterbe, de ennek az az ára, hogy a korpusz akár több, mint 30%-a nem kerül osztályozásra.

A problémák ellenére a neurális nyelvi modellekre épülő témamodellezés ígéretes irány a nagytömegű, rövid és zajos szövegek, így a közösségi médiában megjelenő éttermi vendégvélemények feldolgozására. Gyakorlati hasznosíthatóságát elsősorban a szolgáltatók és a desztináció menedzseléséért felelős szervezetek körében vélelmezzük. A témamodellezés képes rámutatni a szolgáltatás legfontosabb, vendégek által értékelt dimenzióira, és lehetőséget ad arra, hogy a preferenciák időbeli változásait vagy az éttermek közötti különbségeket vizsgáljuk, ami hasznos lehet az éttermi és más turisztikai szolgáltatások fejlesztése során.

Felhasznált irodalom

ABUZAYED, A. – AL-KHALIFA, H. (2021): BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science.* 189. pp. 191–194. https://doi.org/10.1016/j.procs.2021.05.096

AKTAS-POLAT, S. (2022): Analysis of Fine Dining Restaurant Reviews for Perception of Customers Restaurant Service Quality. *Journal of Tourism and Gastronomy Studies*. https://doi.org/10.21325/jotags.2022.974

ALAMSYAH, A. – GIRAWAN, N. D. (2023): Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model. *Big Data and Cognitive Computing.* 7(4). 168. https://doi.org/10.3390/bdcc7040168

ALBALAWI, R. – YEAP, T. H. – BENYOUCEF, M. (2020): Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence.* 3. https://doi.org/10.3389/frai.2020.00042

BLEI, D. M. (2012): Probabilistic topic models. *Communications of the ACM.* 55(4). pp. 77–84. https://doi.org/10.1145/2133806.2133826

BLEI, D. M. – NG, A. Y. – JORDAN, M. I. (2003): Latent Dirichlet allocation. *Journal of Machine Learning Research.* 3(4-5). https://doi.org/10.1016/b978-0-12-411519-4.00006-9

CHANG, J. – GERRISH, S. – WANG, C. – BOYD-GRABER, J. – BLEI, D. (2009): Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems.* 22. https://proceedings.neurips.cc/paper_files/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

CHEN, Z. – DOSS, H. (2019): Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modeling. *Journal of Computational and Graphical Statistics.* 28. pp. 567–585. https://doi.org/10.1080/10618600.2018.1558063

CHENG, X. – YAN, X. – LAN, Y., – GUO, J. (2014): BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering.* 26(12). pp. 2928–2941. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2014.2313872

DE GROOT, M. – ALIANNEJADI, M. – HAAS, M. R. (2022): Experiments on Generalizability of BERTopic on Multi-Domain Short Text. *ArXiv.* https://arxiv.org/abs/2212.08459

DEVLIN, J. – CHANG, M.-W. – LEE, K. – TOUTANOVA, K. (2019): *BERT:* Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv abs/1810.04805.* https://doi.org/10.48550/arXiv.1810.04805

EGGER, R. – YU, J. (2022): A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology.* 7. 886498. https://doi.org/10.3389/fsoc.2022.886498

GEORGE, C. P. – DOSS, H. (2018): Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model. *Journal of Machine Learning Research.* 18(162). pp. 1–38. http://jmlr.org/papers/v18/15-595.html

GLAZKOVA, A. (2021): Identifying Topics of Scientific Articles with BERT-Based Approaches and Topic Modeling. In: Gupta, M. – Ramakrishnan, G. (szerk.): *Trends and Applications in Knowledge Discovery and Data Mining.* Springer International Publishing. pp. 98–105. https://doi.org/10.1007/978-3-030-75015-2_10

GRIFFITHS, T. L. – STEYVERS, M. (2004): Finding scientific topics. *Proceedings of the National Academy of Sciences.* 101(suppl_1). pp. 5228–5235. https://doi.org/10.1073/pnas.0307752101

GROOTENDORST, M. (2022): BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv abs/2203.05794.* https://doi.org/10.48550/arXiv.2203.05794

HA, C. – TRAN, V.-D. – NGO VAN, L. – THAN, K. (2019): Eliminating overfitting of probabilistic topic models on short and noisy text: The role of dropout. *International Journal of Approximate Reasoning.* 112. pp. 85–104. https://doi.org/10.1016/j.ijar.2019.05.010

HUANG, J. – ROGERS, S. – JOO, E. (2014): Improving Restaurants by Extracting Subtopics from Yelp Reviews. *iConference 2014 (Social Media Expo).* https://hdl.handle.net/2142/48832

KOROTEEV, M. V. (2021): BERT: A Review of Applications in Natural Language Processing and Understanding. *ArXiv.* https://arxiv.org/abs/2103.11943

KRISHNAN, A. (2023): Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis. *ArXiv.* abs/2308.11520. https://doi.org/10.48550/ARXIV.2308.11520

KWON, W. – LEE, M. – BACK, K.-J. (2020): Exploring the underlying factors of customer value in restaurants: A machine learning approach. *International Journal of Hospitality Management.* 91. 102643. https://doi.org/10.1016/j.ijhm.2020.102643

KWON, W. – LEE, M. – BOWEN J. T. (2022): Exploring Customers' Luxury Consumption in Restaurants: A Combined Method of Topic Modeling and Three-Factor Theory. *Cornell Hospitality Quarterly.* 63(1). pp. 66–7. https://doi.org/10.1177/19389655211037667

OGUNLEYE, B. – MASWERA, T. – HIRSCH, L. – GAUDOIN, J. – BRUNSDON, T. (2023): Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences.* 13(2). 797. https://doi.org/10.3390/app13020797

PARK, E. – CHAE, B., – KWON, J. (2018): The structural topic model for online review analysis: Comparison between green and non-green restaurants. *Journal of Hospitality and Tourism Technology.* 11(1). pp. 1–17. https://doi.org/10.1108/JHTT-08-2017-0075

RIEGER, J. – RAHNENFÜHRER, J. – JENTSCH C. (2020): Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype. *Natural Language Processing and*

Information Systems. 12089. pp. 118–125. https://doi.org/10.1007/978-3-030-51310-8_11

LOVATO, P. – BICEGO, M. – MURINO, V. – PERINA, A. (2015): Robust Initialization for Learning Latent Dirichlet Allocation. In: Feragen, A. – Pelillo, M. – Loog, M. (eds): *Similarity-Based Pattern Recognition.* SIMBAD 2015. Lecture Notes in Computer Science. 9370. Springer, Cham. pp. 117–132. https://doi.org/10.1007/978-3-319-24261-3_10

QIANG, J. – QIAN, Z. – LI, Y. – YUAN, Y. – WU, X. (2022): Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering.* 34(3). pp. 1427–1445. https://doi.org/10.1109/TKDE.2020.2992485

TITOV, I. – McDONALD, R. (2008): Modeling Online Reviews with Multi-grain Topic Models. *ArXiv.* https://arxiv.org/abs/0801.1063

WESTERLUND, M. – SHAIRY, Z. – LEMINEN, S. – RAJAHONKA, M. (2019): Topic modelling analysis of online reviews: Indian restaurants at Amazon.com. In: Bitran, I. – Conn, S. – Gernreich, C. – Heber, M. – Huizingh, E. – Kokshagina, O. – Torkkeli, M. – Tynnhammar , M. (eds): Proceedings of the ISPIM Connecs Ottawa Conference. ISPIM. pp. 1-14.

ZHANG, S. – LY, L. – MACH, N. – AMAYA, C. (2022): Topic Modeling and Sentiment Analysis of Yelp Restaurant Reviews. *International Journal of Information Systems in the Service Sector (IJISSS).* 14(1). pp. 1–16. https://doi.org/10.4018/IJISSS.295872

ZHAO, F. – LIU, H. (2023): Modeling customer satisfaction and revisit intention from online restaurant reviews: An attribute-level analysis. *Industrial Management – Data Systems.* 123(5). pp. 1548–1568. https://doi.org/10.1108/IMDS-09-2022-0570

ZUO, Y. – LI, C. – LIN, H. – WU, J. (2023): Topic Modeling of Short Texts: A Pseudo-Document View With Word Embedding Enhancement. *IEEE Transactions on Knowledge and Data Engineering.* 35(1). pp. 972–985. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2021.3073195

Beérkezett/Received – 13 August 2024
Elfogadva/Accepted – 19 December 2024