

Eötvös Loránd Tudományegyetem

Társadalomtudományi Kar

MESTERKÉPZÉS SZAKDOLGOZAT

Depresszió és COVID-19 – online fórumok topik modellezése

Konzulens:

Dr. Németh Renáta

Készítette:

Zaboretzky Bendegúz

S7H2WM

survey statisztika és adatanalitika

szak

2021, október

Tartalomjegyzék

Bevezetés	3
COVID-19	4
Depresszió.....	6
Topikmodellek, LDA és STM ismertetése.....	7
Látens Dirichlet-allokáció (LDA).....	7
Strukturális topikmodell (STM).....	10
Betekintés az adatbázisba	13
Elemzés	16
LDA modellek bemutatása	16
Az optimális poszt-COVID modell	19
A pre-COVID és COVID-éra összevetése	29
STM modellek bemutatása.....	36
Konklúzió.....	43
Irodalomjegyzék.....	44

Bevezetés

A depresszió minél alaposabb megértésében kulcsszerepet tölt be magával a depresszióval küzdő személy. Az ilyen helyzetben lévő személyeket, pontosabban perspektívájukat a depresszió és ahhoz szorosan kötődő problémák tekintetében hatékonyan tudjuk megközelíteni és vizsgálni online fórumokon keresztül. Ezt egy másik idei kutatás nagyszerűen meg is tette¹, melyre jelen dolgozat szorosan épített. Az újdonságot a COVID-19 megjelenése és a következtében létrejövő világjárvány, illetve annak a depressziós diskurzusra gyakorolt hatásának a vizsgálata jelenti. A dolgozat célja a korábbi kutatásra építkezni, az ott megjelent eredményeket kiegészíteni és a vezérfonalat folytatni, az említett új hatás figyelembevételével. Ebből kifolyólag jelen kutatás is topikmodellezésen alapszik, valamint NLP² (Natural Language Processing) módszereket használt az eredmények bemutatására.

¹ Németh, R., Sik, D., & Katona, E. (2021). The asymmetries of the biopsychosocial model of depression in lay discourses-Topic modelling online depression forums. *SSM-population health*, 14, 100785.

² Elsősorban LDA (Latent Dirichlet Allocation) és STM (Structural Topic Models) módszerek.

COVID-19

A COVID-19, hétköznapi nevén koronavírus, napjaink egyik legtöbb figyelmét élvező járványa. Tekintettel arra, hogy a témát részletesen taglalja számos kutató és média intézmény, szűk korlátok között, a jelen kutatás szempontjából mindenképpen szükséges aspektusokat tisztázom csupán. A WHO (World Health Organization) 2020.03.11-én jelentette világjárvánnyá, így a dolgozatomhoz felhasznált adatbázis adatainak kezdőidőpontját is ebben határoztam meg. A járvány jelentőségét egészségmegőrzési szempontokból kiindulva az alábbi két cikk részlet megfelelően szemlélteti.

<https://ourworldindata.org/coronavirus>

Quote: "Case fatality rate globally = 3.48% [based on 109,578 confirmed cases and 3809 deaths]

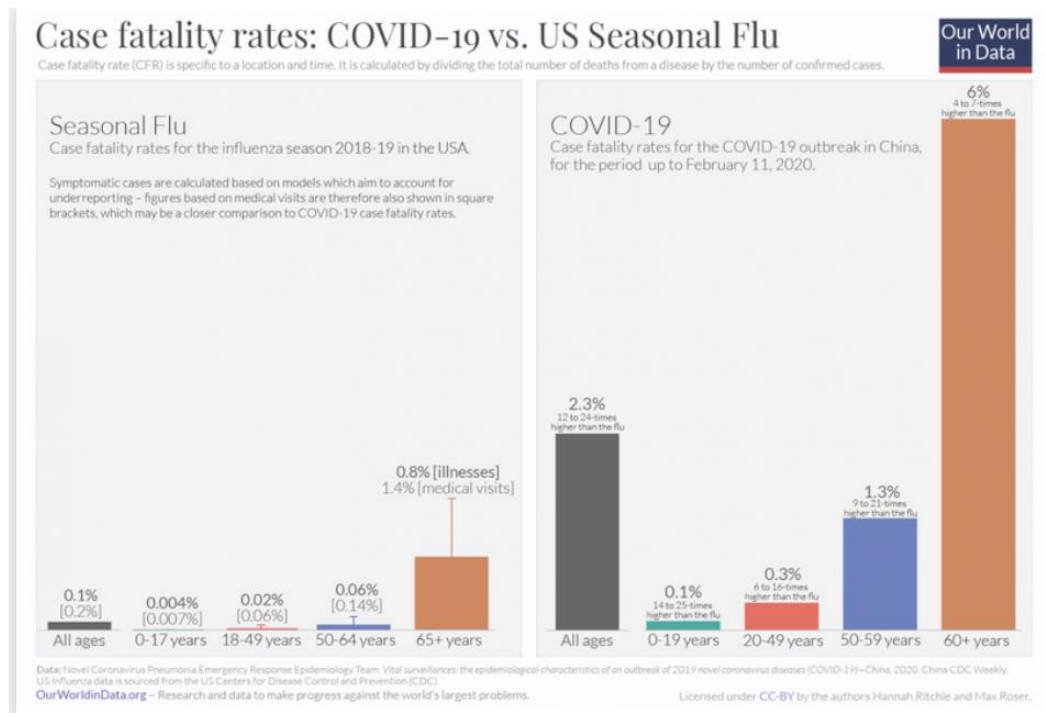
Case fatality rate in China: 3.86% [based on 80,904 confirmed cases and 3123 deaths]

Case fatality rate for the rest of the world: 2.39% [based on 28,674 confirmed and 686 deaths]"

(Last updated March 17, 2020)

The death rate for the flu is somewhere around 0.1%:

Az első részlet a mortalitás arányát hivatott bemutatni minden korcsoportra vetítve, Kína, valamint a világ többi részén, összehasonlítva a hagyományos influenza által produkált aránnyal. Globálisan közel 24-szeres értékkel rendelkezik a koronavírus.



A második részlet egy diagram, amely az előzőekben leírtakat korcsoportok szerinti bontásban mutatja be. A lábjegyzetben található linkeken³, melyek egyben a képek forrásai is, további nagyon részletes, naprakész és pontos információ található a koronavírus számos aspektusáról.

A kutatás szempontjából továbbá érdemes kiemelni a járvány visszaszorítására alkalmazott oltás előtti/melletti eljárásokat, mint például karantén vagy különböző társadalmi korlátozások (social distancing). Ezek logikus velejárója a bezártság, egyedüllét, szélső esetekben magányosság, valamint magától a vírustól való félelem és közeli hozzátartozókért való aggodás. A helyzetet gyakran inkonzisztens, megkérdőjelezhető információk sokasága övezte, mely további stressz faktorként is tekinthető.⁴ A gazdasági körülmények általános romlásának eredményeképpen, a megnövekedett munkanélküliség és létbizonytalanság szintén nem elhanyagolható szereplő a társadalom mentális egészségének szempontjából.⁵ Ezek mind befolyásoló tényezőként hatnak a depresszióval küzdőkre, tehát megalapozott az online depressziós diskurzusok változásának vizsgálata.

³ <https://ourworldindata.org/coronavirus> ; <https://sites.google.com/view/sourcescorona>

⁴L. Biester et al. (2020)

⁵L. Biester et al. (2021)

Depresszió

A depresszió, mint mentális betegség tanulmányozására, illetve a különböző kezelések kialakítására jelenleg a legátfogóbb és elfogadottabb paradigma a biopszichoszociális modell.⁶ Az említett modell, ahogy az elnevezéséből is adódik, magába foglalja a depresszió biológiai, pszichológiai és társadalmi (szociális) aspektusait, így hatékonyabban megértve, valamint kezelve a betegséget – legalábbis elméleti síkon. A gyakorlat azt mutatja, ahogyan ez kiderül Németh, R., Sik, D., és Katona, E. (2021) által publikált tanulmányból is, hogy a laikus szemlélet és a gyakorlatban előforduló megközelítések nem egészen egyeznek a szakértők által definiált modellel. Ebből kifolyólag gyakorlati oldalról, az egyes kezelések nem veszik figyelembe a teljes képet, így, ha például a biológiai szemszög van előtérben, elsősorban gyógyszeres kezelésben részesül a depresszióval küzdő egyén, míg, ha pszichológiai, akkor különböző terápiákon vesz részt. Nyilvánvaló, hogy előfordulnak kezelések, melyek ötvözik ezeket, azonban elég gyakran a társadalmi szempont ekképpen is alulmarad a másik kettővel szemben, mely komoly csökkenést jelenthet a páciens gyógyulásának hatékonysága tekintetében. Másrészt, mivel a laikus diskurzus sem rendelkezik átfogó képpel, a segítség keresését és saját betegségének, annak okainak és körülményeinek leírását pontatlanul végzi. Természetesen a segítséget az ő részükre szakértői oldalról kell biztosítani, nem elvárható, hogy „magukat kezeljék”, azonban köztudott, hogy gyógyulási hatékonyság szempontjából kiemelkedő szereppel bír a megfelelő tájékozottság és az opciók ismerete.

Amennyiben figyelembe vesszük a COVID-19 – vagy bármilyen a hétköznapi életet alapjaiban befolyásoló krízishelyzet – következményeit pszichológiai és társadalmi értelemben, úgy könnyen belátható a kapcsolat a (mentális betegségek és) depressziós diskurzus alakulásával az előző fejezet utolsó bekezdésében leírtak alapján.

⁶ Bolton, D., & Gillett, G. (2019)

Topikmodellek, LDA és STM ismertetése

A dolgozatban alkalmazott modellek mindkét esetben az NLP – természetesnyelv-feldolgozás – kategóriájába tartoznak. Természetes nyelveknek azon nyelveket nevezzük, melyek kialakulásukban és nyelvtani szerkezetükben „maguktól”, a történelem során a „véletlen” következtében jöttek létre, tudatos emberi behatás nélkül – természetesen később ezen nyelvek fejlődése is például nyelvújítás során terelhető volt. Ilyenek például a nemzetek nyelvei. Míg a mesterséges nyelvek eredete az embertől fakad, tudatos szabályok alakítják és szűkebb, konkretizált célt szolgálnak (például programnyelvek). *„Természetes nyelvek azokat a nyelveket nevezzük, amelyek spontán alakultak ki [...] az emberek közötti nyelvi kommunikáció természetes fejlődésének eredményeként [...] a mesterséges nyelvet az ember hozza létre, szabályai tudatosan tervezettek [...]”*⁷ A topikmodellek általános célja valamilyen korpusz, vagy dokumentum halmaz témáinak felismerése és kategóriákba (topikokba) való gyűjtése. Az ilyen jellegű modellek értelmezése adott topikszám esetén a topikok valószínűségének, az ezekben a topikokban előforduló szavak eloszlásának és a topikok keveredési hajlandóságának a vizsgálatából tevődik össze.

Látens Dirichlet-allokáció (LDA)

Az LDA (Latent Dirichlet Allocation – látens Dirichlet-allokáció) olyan szöveganalitikai topikmodell, mely alkalmas strukturálatlan korpusz elemzésére, valamint, ahogyan arra a neve is utal, azt feltételezi, hogy a dokumentumomkénti topikeloszlás Dirichlet-eloszlást követ. Háromszintű hierarchikus Bayesi statisztikára épülő modelltől beszélünk.⁸ Alapvetően a működése a következőképpen szemléltethető, nyilván leegyszerűsített módon: ha feltesszük, hogy a megfigyeléseink szavak, melyek dokumentumokba vannak gyűjtve, akkor minden létező dokumentumunk különböző, kis számú topik keveréke és minden jelenlévő szó hozzárendelhető a dokumentum valamelyik topikjához. Más szavakkal *„[...] létezik a témáknak egy véges halmaza, amelyben a témák statisztikailag az adott nyelv kifejezésein*

⁷ Németh R. és társai (2020)

⁸ Blei D. M. et al. (2003)

*értelmezett valószínűségeloszlásként definiálhatók.”*⁹ Itt a topikok keveredésének a minimalizálása a cél. Többek között két fontos tulajdonsága, hogy „szózsák” (bag of words) alapon, nem felügyelt tanulással prediktál. Ez tehát azt jelenti, hogy figyelmen kívül hagyja a szövegben előforduló szavak egymáshoz viszonyított sorrendjét, kizárólag szógyakoriság-eloszlást vizsgál. Valamint azt is jelenti, hogy nincs előre bekategorizált X számú bejegyzésünk, Y számú topikba, mely alapján a modell a teljes korpuszra elvégezné a predikciót. A szöveggenerálásra nézve az LDA ezeken kívül még egy erős megszorítással él, ez pedig a függetlenségi feltevés, melynek lényege, hogy korrelálatlan topikokat tételez fel.¹⁰ Az eljárás, mint általában a többi NLP-módszer, első komoly kihívása a legoptimálisabb modell kiválasztása, melyhez több mérőszám is segítséget nyújt (azonban a kutatói megfontolás és a kontextus kiemelkedő szereppel rendelkeznek). Mielőtt ezen mérőszámok bemutatására rátérnék, magának a modell működésének a matematikai elveit veszem végig. Erre kiemelt hangsúlyt kell fektetni ilyen jellegű modellek használata esetén, hiszen: *„A predikációs modellek egy része ugyanis nem is használható magyarázati séma alapjaként, mert egyfajta fekete dobozként működik: tudjuk, hogy jól prediktál, de nem tudjuk, hogy miért prediktál így vagy úgy [...]”*¹¹ Az idézett szöveg a fekete doboz jellegre hivatott felhívni a figyelmet, melyet igyekszem minél inkább semlegesíteni a dolgozatomban.

Az LDA statisztikai háttere a következőképpen épül föl:

- A szó a diszkrét adat alapvető egysége, indexelt szótár elemeként definiálva, ahol az index $\{1, \dots, V\}$. A szavakat olyan egység alapú vektorokkal ábrázoljuk, amelyeknek egyetlen komponense egyenlő eggyel és az összes többi komponens egyenlő nullával. Így az összetevők jelölésére felső indexeket használva a szótárban, a v -edik szót egy V -hosszú w vektor képviseli, így $w^v = 1$, $w^u = 0$, $u \neq v$.
- Egy dokumentum az N szóból álló sorozat, melynek jelölése $\mathbf{w} = (w_1, w_2, \dots, w_N)$ ahol w_n az n -edik szó a sorozatban.
- Egy korpusz az M dokumentumból álló gyűjtemény, melynek jelölése $D = \{w_1, w_2, \dots, w_M\}$ ¹².

⁹ Németh R. és társai (2020)

¹⁰ Blei, D., & Lafferty, J. (2006)

¹¹ Németh R. és társai (2020)

¹² Blei D. M. et al. (2003)

Ha úgy tekintünk az LDA modellre – a már korábban leírtak alapján – mint egy korpusz generatív valószínűségi modelljére, akkor a következő generatív folyamatot feltételezzük minden egyes $\mathbf{w}$ dokumentumhoz egy $\mathbf{D}$ korpuszban:

1. Válasszunk $N \sim p_{\text{poisson}}(\xi)$ szóeloszlást
2. Válasszunk $\theta \sim \text{Dir}(\alpha)$ topikeloszlást
3. Az N darab w_n szó mindegyékre:
 - a. Válasszunk egy $z_n \sim \text{Multinomial}(\theta)$ topikot
 - b. Válasszunk egy w_n szót a $p(w_n | z_n, \beta)$, a z_n topiktól függő multinomiális feltételes valószínűségből

Itt az α skalár, vagy pedig pozitív K -dimenziós vektor, hasonlóan β skalár, vagy pedig pozitív V -dimenziós vektor.¹³

A modell koherencia értékeinek kiszámítására négy alapvető mérőszám létezik. Ezek közül a legelterjedtebb és általam is alkalmazott a C_v -vel jelölt koherencia mutató. Elterjedtsége annak is köszönhető, hogy számos tesztelést követően ezt a mérőszámot találták a leginkább korrelálónak az emberi besorolással. Működésében a következőképpen néz ki. Először a szópárok együttes előfordulását számlálja meg úgynevezett 110-es szélességű „csúszó ablak” (sliding window) alkalmazásával. Ez alatt az ablak alatt azt értjük, hogy az algoritmus adott pillanatban a korpusz X számú kifejezését vizsgálja – esetünkben 110-et – majd továbbmegy a következő X számú kifejezésre. A mutató ezt követően az összes szó összes többi szóra vonatkozó NPMI (Normalized Pointwise Mutual Information)¹⁴ értéket számol, mellyel szavankénti vektorokat konstruál. Az így létrejött vektorokra és ezek összegére koszinusz hasonlóságot kalkulál, melyek átlagából kapjuk meg végül a mutató értékét.¹⁵ A koszinusz hasonlóság két vektor által bezárt szög koszinuszát jelenti.

¹³ Blei D. M. et al. (2003)

¹⁴ PMI érték kiszámítása adott szópárra: $\text{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$

¹⁵ Röder et al. (2015)

További mérőszámok:

- **C_{UCI}** : A topikot leíró szópárok együttes előfordulásának valószínűségét vizsgálja referencia-korpuszon PMI érték és 10-es szélességű csúszó ablak segítségével.¹⁶
- **C_{UMass}** : A topikot leíró szópárok együttes előfordulásának valószínűségét vizsgálja, ezen valószínűségek logaritmusának átlagát kiszámítva, valamint a szavak sorrendjét figyelembe véve.¹⁷ Ezt az értéket szemantikus koherenciának szokták hívni és a következő alfejezetben további kifejtésre kerül.
- **C_{NPMI}** : A topikot leíró szópárok együttes előfordulásának valószínűségét vizsgálja referencia-korpuszon NPMI érték és csúszó ablak segítségével, valamint a szavak környezetét is figyelembe veszi (+/- 5-ös vektorként).¹⁸

Strukturális topikmodell (STM)

Számos topikmodell a vizsgált dokumentumok között felcserélhetőséget tételez föl (exchangeability assumption), amely validitási problémát okoz, hiszen ez gyakran nem igaz. Ennek következtében számos próbálkozás történt olyan modellek létrehozására, melyek különböző metaadatok bevonásával előzetesen meghatározott dokumentum specifikus eloszlás alapján válasszák ki a témakeverési arányt, vagy a megfigyelt szavakat, ahelyett, hogy ezeket egy globális α és β alapján tennék, mint az LDA esetén. Roberts, M. E., Stewart, B. M., és Airoldi, E. M. (2016) célja az volt, hogy létrehozzanak egy modellt, mely többek között ezt kiküszöböli és általánosan használható szövegelemzők számára. Két korábbi megoldásra építettek, melyek az általános kovariáns információk témamodellekbe való beépítésével foglalkoztak: Mimno és McCallum (2008) Dirichlet-multinomiális regressziós topikmodellje, illetve Eisenstein, Ahmed és Xing (2011) eseti additív generatív modellje (sparse additive generative model). Az STM modell kiegészíti ezeket az eljárásokat azáltal, hogy engedi a topikok közötti kovarianciát és hangsúlyozza a megfigyelt dokumentum jellemzők nemlineáris funkcionális formáinak használatát.

¹⁶ Newman et al. (2010)

¹⁷ Mimno et al. (2011)

¹⁸ Aletras and Stevenson (2013)

A matematikai jelölések hasonlóak a korábban leírt LDA modelléhez:

- egyes dokumentumok: $\mathbf{d} \in \{1 \dots D\}$
- egyes szavak a dokumentumokban: $\mathbf{n} \in \{1 \dots N_d\}$
- elsődleges megfigyelések: $\mathbf{w}_{d,n}$ szavak
- szótár: $\mathbf{v} \in \{1 \dots V\}$
- topikszám: $\mathbf{k} \in \{1 \dots K\}$
- A topik prevalencia kovariánsainak mátrixát $\mathbf{X}$ jelöli, dimenziója $\mathbf{D} \times \mathbf{P}$.
- A topik tartalom kovariánsainak mátrixát $\mathbf{Y}$ jelöli, dimenziója $\mathbf{D} \times \mathbf{A}$.
- Ezeknek a mátrixoknak a sorait $\mathbf{x}_d$, illetve $\mathbf{y}_d$ jelöli.
- $\mathbf{v}$ kifejezés marginális logaritmus gyakorisága: $\mathbf{m}_v$

A generatív folyamat a következő $\mathbf{d}$ dokumentumra, $\mathbf{K}$ topik esetén, ahol a megfigyelt szavaink $\{\mathbf{w}_{d,n}\}$, a topik prevalencia mátrix $\mathbf{X}$, topik tartalom mátrix $\mathbf{Y}$, a skalár hiperparaméterek¹⁹ $\mathbf{s}, \mathbf{r}, \mathbf{p}$ és a $\mathbf{K}$ -dimenziós hiperparaméter vektor $\boldsymbol{\sigma}$:²⁰

$$\gamma_k \sim \text{Normal}_P(0, \sigma_k^2 I_P), \text{ minden } k = 1 \dots K - 1 \text{-re,}$$

$$\boldsymbol{\theta}_d \sim \text{LogisticNormal}_{K-1}(\boldsymbol{\Gamma}' \mathbf{x}'_d, \boldsymbol{\Sigma}),$$

$$\mathbf{z}_{d,n} \sim \text{Multinomial}_K(\boldsymbol{\theta}_d), \text{ minden } n = 1 \dots N_d \text{-re,}$$

$$\mathbf{w}_{d,n} \sim \text{Multinomial}_V(\mathbf{B} \mathbf{z}_{d,n}), \text{ minden } n = 1 \dots N_d \text{-re,}$$

$$\beta_{d,k,v} = \frac{\exp \left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)} \right)}{\sum_v \exp \left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)} \right)},$$

minden $v = 1 \dots V$ -re és $k = 1 \dots K$ -ra

$\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1 | \dots | \boldsymbol{\gamma}_K]$ kifejezés az egy $\mathbf{P} \times (\mathbf{K} - 1)$, az első és második egyenlet által meghatározott, topik prevalencia koefficiens mátrix. $\{\boldsymbol{\kappa}^{(t)}, \boldsymbol{\kappa}^{(c)}, \boldsymbol{\kappa}^{(i)}\}$ az a topik tartalmi koefficiens gyűjteménye, a legalsó egyenlet által meghatározva ($\mathbf{t}$ – topik; $\mathbf{c}$ – kovariáns; $\mathbf{i}$ – topik-

¹⁹ Paraméterek paraméterei

²⁰ Roberts et al. (2016)

kovariáns interakció). A harmadik és negyedik egyenlet jelöli az alapnyelvi modellt, mely engedélyezi a korrelációt a topik proporcióban logisztikus normál eloszlást alkalmazva.²¹

Tanulmányomban az R-ben használt STM csomag szemantikus koherencia (semantic coherence) érték vizsgálatot végez az optimális topikszám kereséshez. Ennek a koherencia értéknek a kiszámítása a következőképpen történik:

$$C(t; V^{(t)}) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}$$

Legyen $D(v)$ a v szótípus dokumentumgyakorisága (azaz a legalább egy v típusú tokenrel rendelkező dokumentumok száma), $D(v, v')$ pedig a v és v' szótípusok társdokumentum gyakorisága (azaz egy vagy több v típusú tokent és legalább egy v' típusú tokent tartalmazó dokumentumok száma). $V^{(t)} = (v_1^{(t)}, \dots, v_M^{(t)})$ egy olyan lista, mely M legvalószínűbb kifejezéseket tartalmazza t topikhoz. Az 1-es érték a logaritmus 0 érték felvételének elkerülése végett került hozzáadásra.²² Az egyenlet nem áll messze a korábban feltüntetett PMI érték kiszámításához használt egyenlettől, azonban a szakirodalom szerint a szakértői annotáláshoz közelebbi eredményt produkál.

²¹ Roberts, M. E., Stewart, B. M., és Airolidi, E. M. (2016)

²² Mimno et al. (2011)

Betekintés az adatbázisba

Az adatelőkészítés (preprocessing) minden kutatás kulcsfontosságú része. Meghatározza a később kapott eredményeinket, valamint ezen eredményeink létrehozása végett használni kívánt modelljeink alapjaiban körvonalazzák az előfeldolgozás folyamatát. *„Az egyes lépések pontos tartalma és sorrendje erősen függ az adott alkalmazástól is, tehát egyedi kutatói döntés függvénye. Ritkán hangsúlyozott szempont, hogy ezek a döntések (mivel az egyes lépések kölcsönösen hatnak egymásra) befolyásolják a teljes elemzés eredményét (Denny–Spirling 2018).”*²³ A szükséges alaposág és körültekintés különösen igaz ennél a szakasznál, amennyiben már meglévő kutatás módszereihez és eredményeihez igazodva, azokra építve tervezzük saját elemzésünket, mint ahogyan ez jelen esetben is történt. Mielőtt ennek részletei kifejtésre kerülnének, lépünk egyet hátrébb és indulunk az alapoktól: adatgyűjtés.

A górcső alá vett adatbázis összesen tizennégy különböző online depressziós platformról származó, adattisztítás előtt ~342.000, utána pedig ~130.000 db bejegyzést tartalmaz, melyek a SentiOne – amiről mindjárt röviden kitérek – (web scraping) nevezetű program segítségével kerültek összeállításra. A SentiOne lényegében mesterséges intelligencia alapú „természetes nyelv motorral” (natural language engine) ismeri föl az online tartalmakat, elsősorban közösségi oldalakon, blog felületeken stb. Ezzel lehetővé téve ezen tartalmak alapos vizsgálatát és adatbázisba gyűjtött letöltését. Jelen kutatásban kizárólag a letöltésen van a hangsúly, ugyanis a forráskód ismeretének a hiányában, nem alkalmazhatjuk a SentiOne által létrehozott kategorizációkat (például bejegyzést író neve). Így tehát, a nem ellenőrizhető változók kidobása után, létrejött az adattisztításra váró adatbázisom, mely tartalmazta a bejegyzés forrását, a bejegyzést író felhasználónevét, a bejegyzés dátumát, a bejegyzés tartalmát, valamint egy egyedi azonosító indexet.

A következő lépés a tényleges adattisztítás volt. Minden adatbázist érintő változás kód segítségével történt, mely a Jupyter Notebook 6.1.4-es verziójában került megírásra és futtatásra. A dolgozat során használt összes kód megtalálható saját GitHub felületemen.²⁴ Első körben a lehető legnagyobb szükségtelen adatmennyiség kiszűrése volt a cél, a már

²³ Németh R. és társai (2020)

²⁴ GitHub link: <https://github.com/WhisperPotato/Szakdolgozat>

említett dátum - 2020.03.11. - utáni adatokra való szűkítéssel. Az adattisztítás folyamatánál érdemes a nagytól a kicsiig elv mentén haladni, ezzel is csökkenthetjük az adott kódok futási idejét és ezáltal a saját munkánkat. Miután a megfelelő időszak adatait tartalmazta csak az adatbázis, eltávolításra, vagy szűrésre kerültek a duplikációk, web hivatkozások (linkek), stopszavak²⁵, valamint a húsz szónál rövidebb bejegyzések és öt karakternél rövidebb szavak (ez utóbbi csak az LDA modell futtatás alatt), így lehetővé téve az utolsó lépést: adatelőkészítés.

A kulcsfontosságú csomagok letöltése és behívása után négy alappillér következett. A tokenizáció, mely a mondatok és szavak azonosításáért felel a szövegben; A „nevezett entitás felismerés” (named entity recognition - NER Tagger), mely a névelemek felismerését szolgálta; A szótövezés, azaz stemmelés (szavak végének levágása, toldalékok eltávolítása) és lemmatizáció (eredeti szótő megadása); Végül szófaj meghatározás (part of speech - POS tagging). Az előbbihez a StanfordNERTagger, míg az utóbbiakhoz a WordNetLemmatizer és a beépített NLTK (Natural Language Toolkit) csomag került alkalmazásra, így a szövegben előforduló egyes szavaink egyrészt előre definiált csoportokba kerülhetnek gyűjtésre (például személy nevek, vagy betegségek szerint), illetve ragozott alaktól függetlenül egy szónak feleltethetőek meg és ismert a szófajuk. Ezt kihasználva felhasználónevek és személy nevek²⁶, betegségek megnevezései, továbbá minden egyéb kiemelkedő bigram (például: 5th_grade; red_meat) külön fájlba került lementésre. Szófajok közül semmi sem került törlésre, később látható módon, fontos információtartalommal rendelkeztek az igék is (szövegelemző kutatások az igéket gyakran eltávolítják).

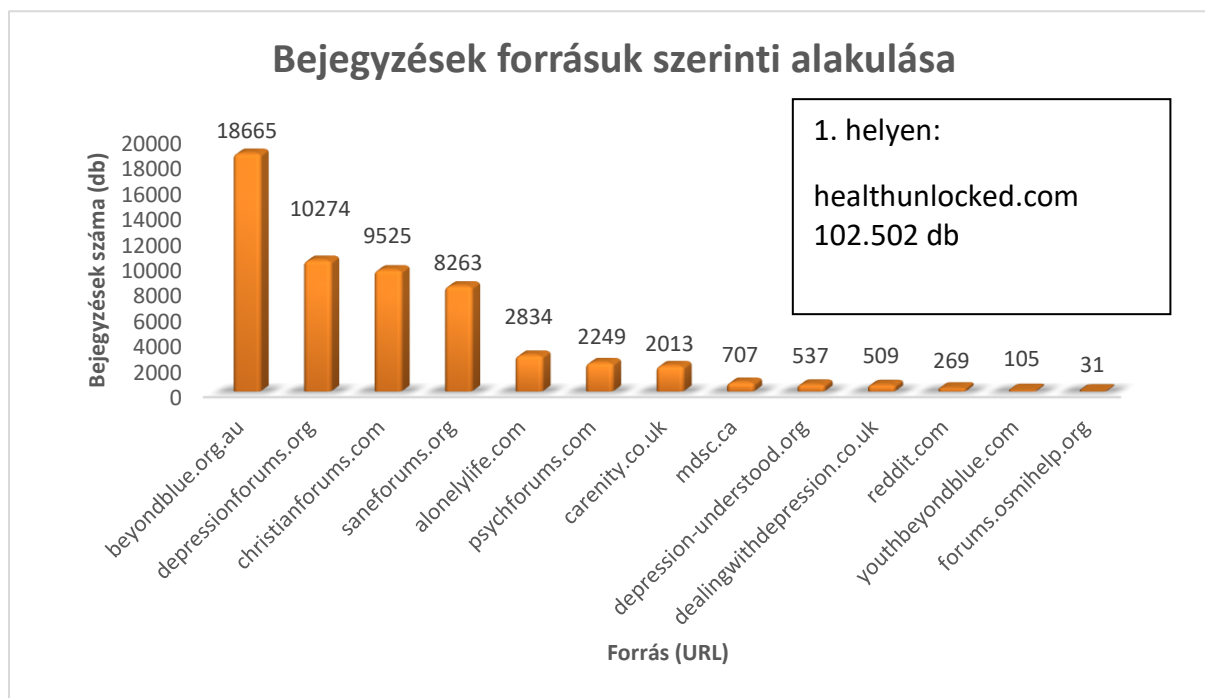
Az angol nyelv egyik sajátossága a szavak összevonásának használata (például can + not = can't, vagy you + are = you're), melyek jelentésben azonosak maradnak, csupán rövidítést szolgálnak. „Az összehúzódás a szó, szótag vagy szócsoporthoz beszélt és írott formáinak rövidített változata, amelyet a belső betűk és hangok kihagyása hoz létre.” (Wikipedia)²⁷ Erre külön figyelmet szentelve, az adatelőkészítéshez alkalmazott Python kód részeként, ennek kezelése is megtörtént, minden összevonás szétválasztásra került (például don't = do + not).

²⁵ Stopszó: tartalom nélküli szavak pl. névelők - Python saját, beépített angol stopszó listáját alkalmazva

²⁶ Külön odafigyeléssel, hogy egy tagú nevek („magányos vezetécknevek”) ne kerüljenek bele

²⁷ [https://en.wikipedia.org/wiki/Contraction_\(grammar\)](https://en.wikipedia.org/wiki/Contraction_(grammar))

A megmaradó, elemzésre alkalmas adatbázist felépítő bejegyzések forrásukat tekintve az alábbi saját készítésű diagram szerint alakultak²⁸:



Az első helyen valójában a healthunlocked.com oldal helyezkedik el, összesen 102.502 darab bejegyzéssel, azonban a szemléltethetőséget figyelembevéve kénytelen voltam ezt külön megjeleníteni.

Az STM modellhez adatelőkészítés értelmében nem történt átalakítás, pontosan ugyanazon a korpuszon futott mindkét szövegelemző modell. Ettől függetlenül szükség volt egy technikai előkészítésre, hogy RStudio-ban megtörténhessen a vizsgálat, melynek Python kódját STM.ipynb tartalmazza. STM esetén a hozzátartozó R csomag létezése miatt került sor az RStudio használatára. Ennek további részletei az 'STM modellek bemutatása' című alfejezetben lesznek olvashatóak.

Az előkészítéshez tartozó Python kódot a preprocess.ipynb tartalmazza, illetve a hozzátartozó szükséges fájlok (.tsv fájlok) szintén megtalálhatóak a GitHub felületen²⁹. Az LDA Python kódját az LDA.ipynb tartalmazza (és a szükséges fájlok szintén feltöltve) – LDA01 az első kör eredményeit, LDA02 a második kör eredményeit hivatott jelezni a fájlnévevezésekben.

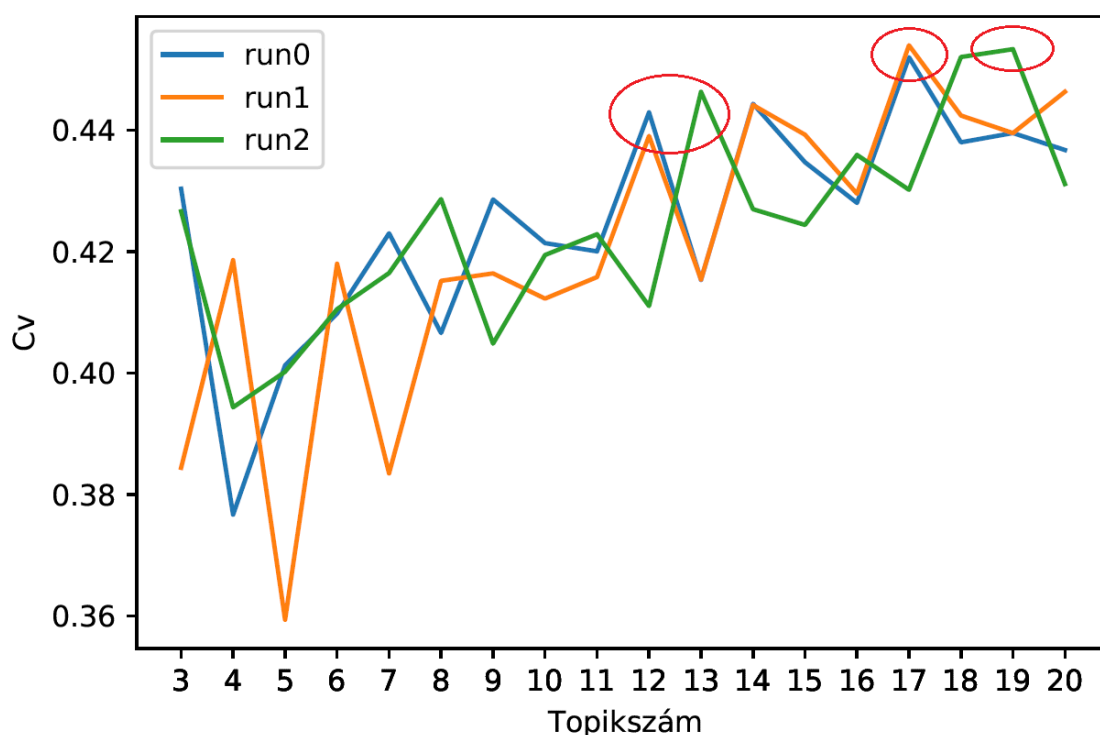
²⁸ A diagram szűrt állapotban van, az első helyen lévő healthunlocked.com túlságosan magas aránya torzítja az adatok átláthatóságát.

²⁹ <https://github.com/WhisperPotato/Szakdolgozat>

Elemzés

LDA modellek bemutatása

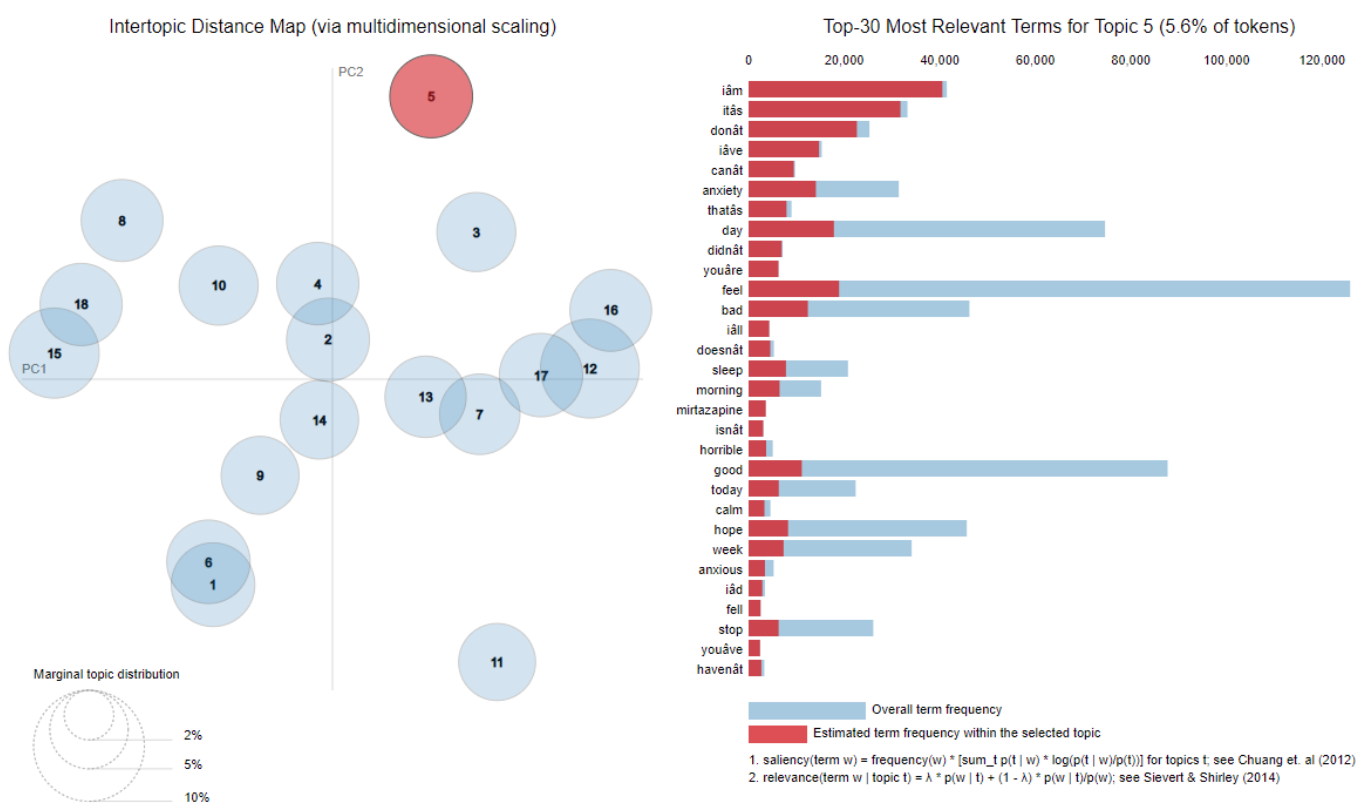
Az alábbi fejezetben először az LDA modell futtatásaim eredményeiről nyújtok teljeskörű leírást. Az LDA esetén mindenekelőtt szükség van egy optimális topikszám meghatározására. Amint azt a korábbi fejezetben kifejtettem, erre több mérőszám létezik, azonban a leggyakrabban alkalmazott a koherencia érték (coherence value – C_v) vizsgálat, ezt választottam én is. Első körös, hármas modell futtatást követően egy a már Németh Renáta és szerző társai által bemutatott ábrához nagyon hasonló vizualizációt kaptam a koherencia értékek alakulásáról a különböző topikszámok függvényében.



Az ábrán az egyes futások koherencia értékei olvashatóak le, minimum 3 és maximum 20 topikszám meghatározása mellett. Az egyes futások között, ahogyan az az LDA-ra jellemző, ugyanazon topikszám esetén eltérő koherencia értékek jöhetnek létre. Piros karikával jelöltem azokat a kiugróan magas koherencia értékeket, melyeknél többszöri futás következtében is hasonlóan magas értéket produkált a modell. Más szavakkal, a 12-13 topikszámok és a 17-18 topikszámok bizonyultak a következetesen legmagasabb koherencia értékeket előidézőknek. Mivel a 17-18 topikszámok rendelkeznek a magasabb értékkel, valamint a már sokat emlegetett Németh Renáta és szerző társai féle kutatásban is

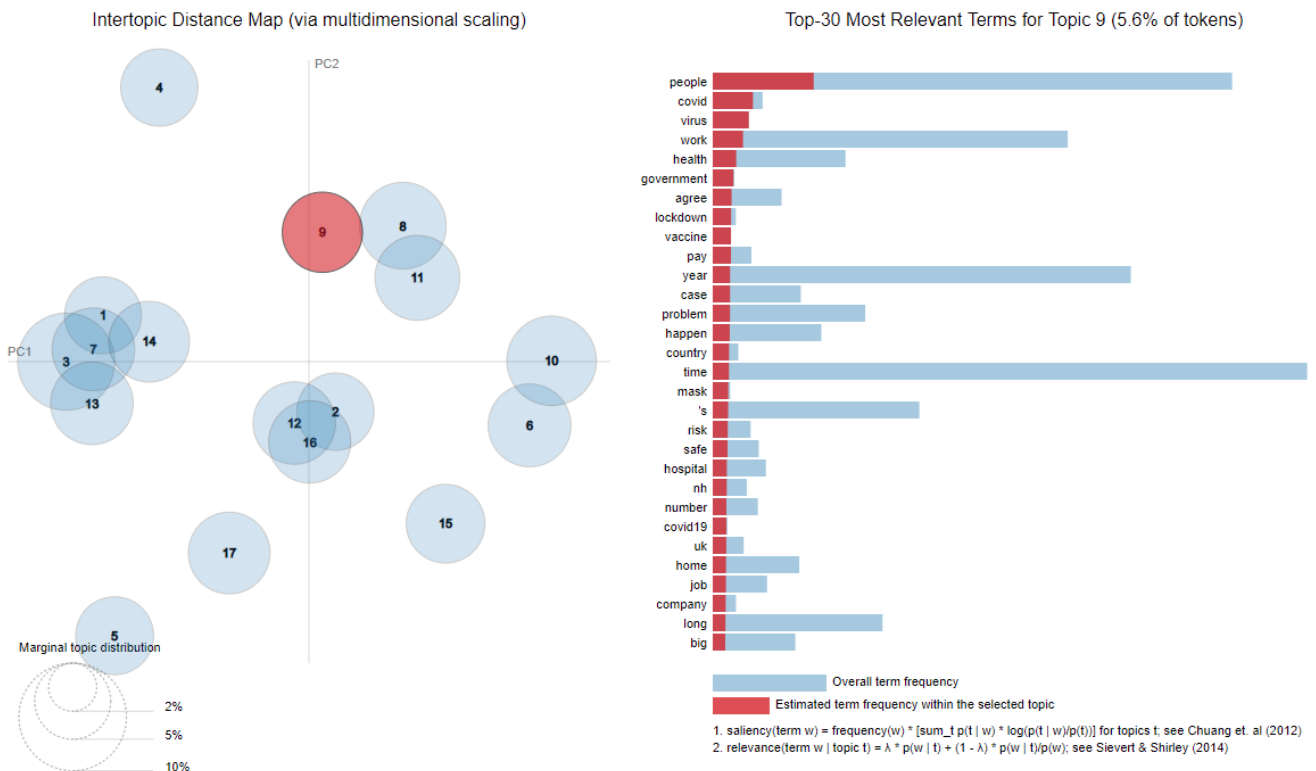
hasonlóan határoztak, jelen dolgozatban ez lett az optimális topikszám elsődleges iránymutatója. Ahhoz, hogy pontosan döntésre jussunk 17, avagy 18 lesz a megfelelő választás, volt még egy fontos szempont, amit meg kellett vizsgálni. Mint minden alapos kutatásban, önmagában a kvantitatív oldalra támaszkodni, a kvalitatív oldalt elhanyagolni nem javasolt. Tehát a 17, illetve a 18 topikszámmal rendelkező modellt mindkét esetben vizualizáltam további információszerzés és értelmezés céljából. A lábjegyzetben található linkek interaktív HTML oldalra irányítanak, ahol megnézhető az egyes topikokon belül megjelenő szavak és azok gyakorisága, vagy fordítva, adott szavak, mely topikokban jelennek meg a leginkább.

18 topikszám esetén³⁰



³⁰ <https://github.com/WhisperPotato/Szakdolgozat>

17 topikszám esetén³¹



A két ábrán két különböző topik került szándékos kijelölésre, melyek külön-külön számottevő jelentőséggel bírnak. Az első esetben az 5-ös topikot jelöltem ki. Ugyanez a topik (másik számmal) megtalálható a második modellben is. Úgy gondolom sok magyarázat nélkül is hamar észrevehető, hogy a topik nem megfelelő szavakat tartalmaz, az előkészítés során a karakter kódolás az aposztrófok esetében elcsúszott. A Python beépített contractions (és contractions.fix) modulja nem minden esetben ismeri föl az aposztróf karaktert, így nem végzi el a szükséges szétbontásokat, manuális kód megírására volt szükség. Javítást követően második körös LDA futtatással tovább optimalizálható a modell és a topikszám meghatározás. A második esetben a 9-es topikot jelöltem ki, a jövőbeli eredményekre való ígéretessége miatt. Kétséget kizáróan nevezhetjük el itt a 9-es számú topikot COVID topiknak.

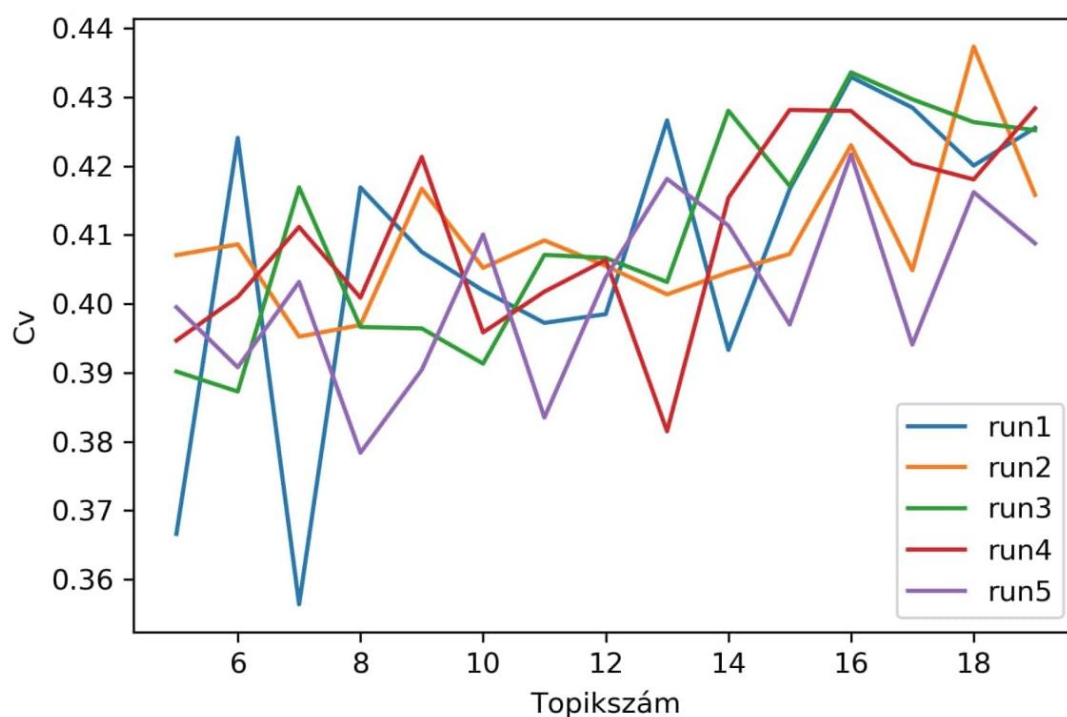
Az így előzetesen feltárt információk ismeretében, megfelelő változtatásokat bevonva (karakter kódolás és szóösszevonások – contractions – javítva; 5 az új minimum topikszám; a futási szám 3-ról 5-re emelve) került futtatásra a második körös LDA modell. A koherencia

³¹ <https://github.com/WhisperPotato/Szakdolgozat>

értékek ismételt vizsgálatával, valamint ez alapján a legoptimálisabbnak ítélt topikszámmal rendelkező modell(ek) kiválasztásával és vizualizációjával folytattam a kutatást. Tekintettel arra, hogy egyik oldalról a „legjobbnek” tűnő modell megtalálása, másik oldalról modellek összehasonlítása a cél, így ketté bontottam az elemzést és ezen alfejezetek szempontjából külön-külön választottam ki az odailő modellt. Az alábbi alfejezetekben már a végleges LDA modellek kerültek bemutatásra.

Az optimális poszt-COVID modell

A végleges modell futtatás, mint már említettem, 5 futás mellett, 5 és 20 közötti topikszám behatárolásával történt. Az alábbi ábrán láthatóak az újonnan létrejött koherencia értékek alakulása.



A korábbi eredményeket megerősítve a második futás 18-as topikszámú koherencia értéke a legmagasabb, így ebben a fejezetben ezt a modellt vizsgálom meg közelebbről. Viszont mielőbb tovább mennénk, itt emelném ki a következő fejezethez tartozó negyedik futás alatt létrejövő 19-es topikszámmal rendelkező modellt is, mely szintén magas értékkel

rendelkezik, valamint ideális megoldást nyújthat koncepcionálisan a COVID megjelenésének bemutatását illetően (eddig 18 topik³² + 1 új COVID topik = 19 topik).

A megjelenő topikok interpretációjához a következők szerint jártam el. Megnéztem minden topik legkiemelkedőbb (most salient) és legrelevánsabb kifejezéseinek (most relevant terms) alakulását akár topikon belül, akár korpuszhoz képest globálisan. Az egyes topikokhoz tartozó legmagasabb kontribúcióval rendelkező bejegyzéseket kilistáztam és (a legalább 0,85 kontribúciós értékkel rendelkezőket) egy jól működő kvalitatív módszer segítségével – olvasással – értelmeztem. A célom az volt, hogy létre tudjak hozni ezek alapján egy átlátható és összefoglaló topik táblázatot.

Az imént említett mutatók – jellegzetesség (salient) és relevancia (relevant) – matematikai definíciói az alábbiak szerint alakul. A w kifejezés k topikhoz való relevanciája, adott λ súlyparaméterrel, ahol $0 \leq \lambda \leq 1$:

$$r(w, k|\lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right)$$

A képletben szereplő ϕ_{kw} a w kifejezés ($w \in \{1, \dots, V\}$, és $k \in \{1, \dots, K\}$) k topikhoz tartozásának valószínűségét fejezi ki, ahol V a szótár kifejezéseinek számát, és p_w a w kifejezés korpuszbeli marginális valószínűségét jelenti. λ meghatározza a w kifejezés k topik alatti súlyát (logaritmus skálán mérve) a kifejezés topikon belüli előfordulásának és a korpuszbeli marginális valószínűségének arányától (lift) függően.

Továbbá idetartozik még, hogy a létrejött modellek λ paramétere interaktívan módosítható 0 és 1 értékek között. Ez a paraméter büntető jelleggel dolgozik a modellen belül, olyan értelemben, hogy figyelembe veszi az egyes kifejezések előfordulását a teljes korpuszban és nem kizárólag a topikon belül. Ha az értéke 0, akkor nagyon szigorúan büntet, ha az értéke 1, akkor nincs büntetés bevezetve, adott kifejezés relevanciája kizárólag a topikon belüli frekvencián múlik. Erre azért van szükség, mert egy a teljes korpuszban gyakran előforduló kifejezés kevesebb információ tartalommal rendelkezik kifejezetten adott topik

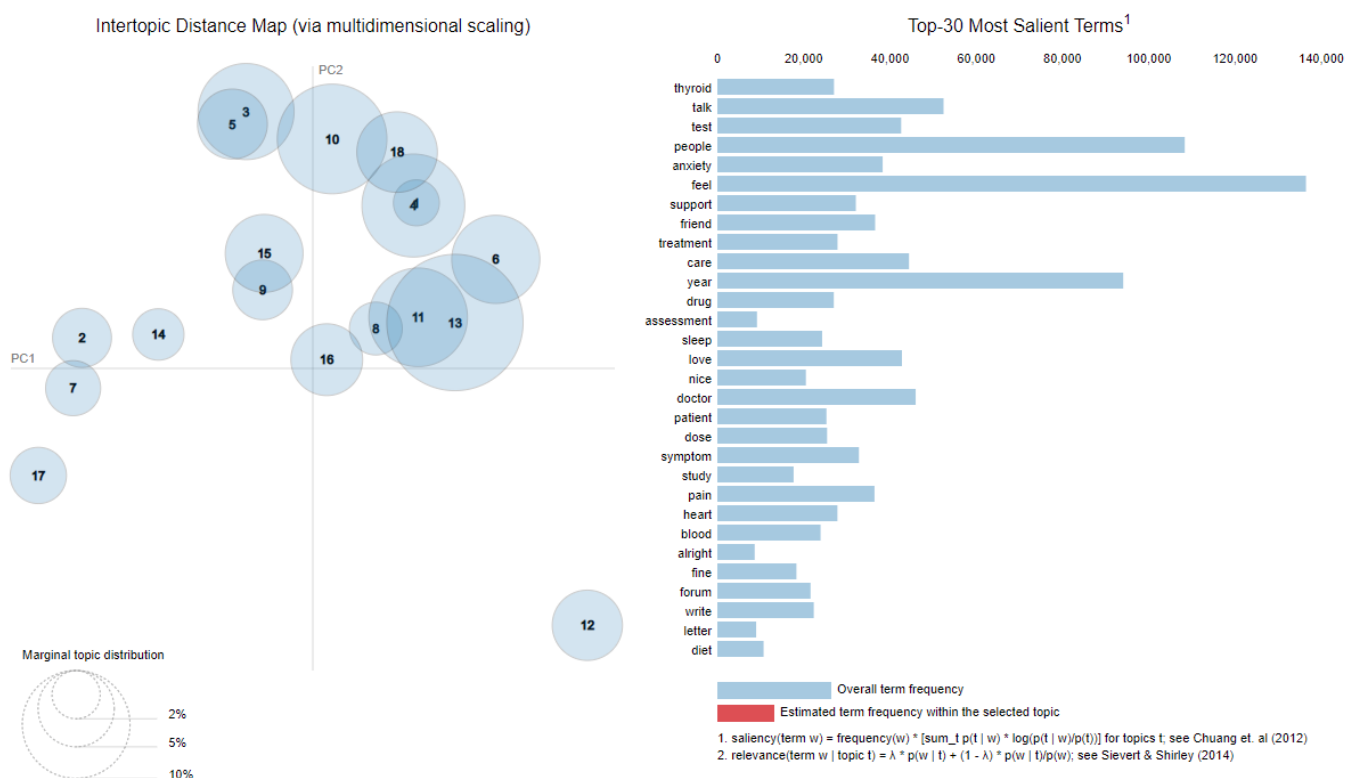
³² Ehhez persze a korábbi elemzés topikjainak összevezetése szükséges a szakdolgozatomban létrejövő topikokkal, de erre majd a következő fejezetben térek ki.

tekintetében. Minden esetben $\lambda = 0,6$ került alkalmazásra, ugyanis ez bizonyult szakirodalom szerint eredményesnek.³³

A korábbi, első körben létrejött topikok esetén feltűnt, hogy bizonyos nevek, Sharon valamint Renee, két különböző topik leginkább releváns kifejezései között megjelentek. Ebből az okból a javított modellben kiemelten figyeltem előfordulnak-e ismét, hiszen a felhasználónevek szűrve vannak, tehát ki kellett zárnom bármilyen jellegű előfeldolgozási hibát. Szerencsére nem egy elrejtett hiba, hanem teljesen más okozta a nevek megjelenését. Az új modellben már csak Sharon szerepelt, valamint megjelent egy új név - John. Kíváncsiságtól vezérelve tehát, a nevüket (Sharon, Renee, John) tartalmazó bejegyzéseket elkezdtem olvasni. Mint kiderült, Sharon (14-es topikban) egy kevésbé kedvelt, sok témához „statisztikai” megjegyzésekkel hozzászóló, feltehetően a jelen dolgozat írásának időpontjában már a platformról kizárt felhasználó, aki összesen 425-ször szerepel a teljes korpuszban. Renee (12-es topikban), összesen 975 alkalommal szerepel a szövegben. Feltehetően depresszióval foglalkozó orvos/pszichológus (sokszor Dr. Renee), saját YouTube csatornával rendelkezik, segítőkész és közkedvelt felhasználó, akihez az oldalt igénybe vevők gyakran fordulnak tanácsért, segítségért. Végül pedig elérkeztünk John-hoz (9-es topikban). Valójában ő kevésbé egy John, inkább több John-ból tevődik össze. Sok felhasználónév része a John (például XY_john_123), illetve sokan vállalják a keresztnévüket több interakciót követően és a John viszonylag gyakori az angol nyelvben. Ezenkívül előfordul még, hogy a különböző gyógynövény készítményekről folyik eszmecsere, melyek között az orbáncfű – St. John’s Wort – nagy jelentőséggel bír a hangulatszabályzó hatása miatt. Így a korpusz több, mint 1500 alkalommal említi a John nevet. Ezt tisztázva, a végleges 18-as topikszámmal rendelkező modell vizualizációja az alábbiak szerint öltött testet:

³³ Sievert & Shirley (2014); Németh, R., Sik, D., & Katona, E. (2021)

model_topics_t18_r02³⁴



Az első körös modellekhez képest a kialakult topikok általánosan jobbnak mondhatóak a releváns kifejezések alapján, valamint ezek szépen különülnek el egymástól. Kívülről befelé, általánostól a részletekig haladva, először a teljes korpusz kiugró kifejezéseire (top-30 most salient terms) vettem egy pillantást. Nem meglepő módon a depresszióhoz szorosan kapcsolódó kifejezések (ezek között például: anxiety – szorongás, treatment – kezelés, drug – gyógyszer, doctor – orvos, feel – érezni, pain – fájdalom stb.) jelennek meg elsősorban, azonban maga a depresszió, mint kifejezés nincs közöttük. Korábban nem fejtettem ki ilyen részletességgel, de az ábra jobb oldalán található sordiagram a szavak kiugrását méri a teljes korpuszban, vagy adott topikot kijelölve topikon belül. A bal oldalon található „térkép” a realizálódott topikok korpuszbeli elterjedtségét, továbbá egymáshoz viszonyított távolságát tükrözi. Jelen modellben ez a következőket jelenti. A jobb felső, második negyedben számos topik átfedésben van egymással, ezek tartalmilag közelebb állnak egymáshoz. Ebben a negyedben is megtalálható két nagyobb csoport, az egyik a 10-18-4-1 topikok csoportja – összefoglaló néven: gyógyszerek, depresszió és kezelése biomedikális oldalról; a másik a 6-11-13-8 topikok csoportja – összefoglaló néven: élet és öröm, támogatás és szocializáció. Egyelőre ezek nagyon tág és előzetesen körvonalazott témakörök, csupán a csoportosulást

³⁴ <https://github.com/WhisperPotato/Szakdolgozat>

hivatottak megérteni, később azonban minden topik megfelelően elkülönítve és megnevezve kerül vizsgálatra.

Mielőtt létrejöhetne a topik táblázat és ezáltal az egyes topikok elnevezése, részletes bemutatása és elemzése, szükség volt még egy lépésre. Ugyanazt a módszert alkalmazva, mint ami a Németh Renáta et al. féle publikációban fellelhető a kvantitatív és kvalitatív elemzés ötvözésére, az egyes topikokhoz tartozó minimum 85%-os kontribúcióval³⁵ rendelkező bejegyzéseket elolvastam – 139 bejegyzés – és az adott topikhoz releváns kifejezésekkel együtt értelmeztem. Az 1-es, 2-es, 5-ös, 8-as, 12-es, 14-es, 17-es, 18-as számú topikokhoz nem tartozik ilyen magas kontribúcióval rendelkező bejegyzés, így azok esetében az első tíz legmagasabb értékkel rendelkező poszt került olvasásra. Ezek együttese szilárd alapot és mélyreható megértést biztosított a topikok felcímkezésében és az alábbi táblázat megalkotásában.

Topik megnevezése	Topik száma	Prevalencia	Top-10 releváns kifejezés
<i>Allergia topik</i>	1	1,8%	fibro, make, mcas, good, give, year, test, thing, told, write
<i>Daganatos betegségek (rák) topik</i>	2	3%	treatment, cancer, year, doctor, cell, month, bicalutamide, study, make, life
<i>Általános medikális topik</i>	3	7,9%	test, thyroid, blood, result, dose, level, symptom, range, doctor, high
<i>Pszichológiai értelmező topik</i>	4	9,1%	feel, anxiety, time, work, thing, depression, make, good, start, thought
<i>Életmód-biomedikális topik</i>	5	4,2%	diet, food, supplement, make, iron, good, dose, people, vitamin, year
<i>Jóléti-napló</i>	6	6,7%	good, love, make, time, thing, walk, year, today, back, feel
<i>Biomedikális értelmező topik</i>	7	2,6%	study, evidence, metformin, benefit, trial, people, statin, cost, drug, good
<i>Vegyes topik</i>	8	2,4%	assessment, letter, time, report, make, assessor, drive, write, form, condition
<i>Vegyes – 2 topik</i>	9	3,1%	people, cure, year, post, make, time, symptom, write, john, disease
<i>Klasszikus biomedikális tanácsadó topik</i>	10	10,3%	pain, year, time, doctor, back, feel, month, symptom, heart, week
<i>Klasszikus pszichológiai</i>	11	8,4%	feel, support, hope, post, people, good, forum, time, talk, care

³⁵ Eredetileg nálam is 90%-os kontribúció volt tervben alsó határnak, azonban úgy nem kaptam elég bejegyzést (18 darab) alapos elemzés tekintetében.

<i>tanácsadó topik</i>			
<i>Szenvedésmonológ</i>	12	4,2%	people, talk, nice, care, feel, fine, friend, heart, alright, good
<i>Pszichológiai támogató topik</i>	13	16%	feel, time, life, thing, make, people, year, friend, work, love
<i>Biomedikális topik neurológiai fókusszal</i>	14	2,3%	year, case, time, sharon, intervention, benefit, read, side, month, tremor
<i>COVID-19</i>	15	5,2%	feel, time, life, thing, make, people, year, friend, work, love
<i>Vallási topik</i>	16	4,4%	life, people, love, jesus, make, thing, give, christian, word, lord
<i>Tudományos diskurzus laikus értelmezése</i>	17	2,7%	drug, year, people, patient, bicalutamide, story, treatment, doctor, thing, forum
<i>Medikális topik mellékhatások fókusszal</i>	18	5,6%	sleep, feel, time, anxiety, effect, night, work, start, make, doctor

Az 1-es számú topik a 4-es alatt helyezkedik el, teljes átfedésben (a statikus ábrán nehezen kivehető, érdemes a HTML dokumentumot kézbe venni), azonban sokkal csekélyebb prevalencia értékkel. A releváns kifejezések alapján egyrészt úgy tűnik, mintha egy specifikus altéma megjelenését látnánk, ami allergiára (vagy pedig a COVID-19 körüli orrspray diskurzusra) utalna. Másrészt a magas kontribúcióval rendelkező bejegyzések olvasása ezt nem támasztja alá, generikusabb, súlylal, mozgással és életmóddal kapcsolatos, nehezen behatárolható vegyes szövegek jelennek meg.

A 2-es számú topik a daganatos betegségekkel kapcsolatosan folytat diskurzust. Ezt mind a releváns szavak listája, mind pedig az olvasott bejegyzések hűen tükrözik. Támogató és tapasztalatokat megosztó céllal létrejött posztokat tartalmaz elsősorban, a medikai megközelítés itt egyáltalán nem, vagy jóval kevésbé szempont. Sok esetben gyógyíthatatlan diagnózissal rendelkező, a betegségben előrehaladott állapotban lévő felhasználók szólnak meg, valamint olyanok, akik kezelés alatt vannak, illetve korábban átestek kezeléseken. Maga a rák (cancer) kifejezés nem kizárólag ebben a topikban jelenik meg, több aspektusa is előfordul a teljes diskurzusban. Jelentős előfordulással még a 10-13-15-ös topikokban jelenik meg, ennek jelentőségét adott topikok körülírásakor fogom megemlíteni.

A 3-as számot képviselő topik, mely az általános medikális topik nevet kapta, nem szigorúan a depresszió kezeléséhez köthető, hanem bármilyen egyéb betegségek kezelése során

felmerülő aspektusokat vitatja medikai oldalról. Ilyen aspektusok lehetnek például az adott vizsgálatok eredményeinek átbeszélése (pl. egy labor lelet eredményei), vagy pedig bizonyos vitaminok, nem gyógyszernek minősülő gyógyító hatású készítmények hatásainak áttekintése. Ebből kifolyólag a fókusz kevésbé egyértelmű, viszont erős átfedésben helyezkedik el az 5-ös topikkal – erre visszatérek az 5-ös topik leírásánál.

A 4-es számú topik alapvetően pszichológiai értelmező topik, ami azt jelenti, hogy a szakértői diskurzust, valamint tanácsadást hivatott feldolgozni és megérteni. Ennek az a szerepe, hogy segítsen döntést hozni egyes kezelések között és a depresszióban szenvedő személy objektívabban, több tudással a kezében, jobban megértse és hatékonyabban mérje föl a saját helyzetét. Gyakorlatban nem elhanyagolható ezen hatásoknak a szerepe, a kezelésre szorulóknak gyakran döntenek egyéni preferencia és kontextus (életkörülmények, tapasztalatok – saját, vagy másé) függvényében egyes kezelések között (akár az egyes szakértői véleményeket háttérbe helyezve).³⁶ Ezenkívül érdekessége, hogy az 1-es számú topik felett helyezkedik el, melynek tartalmával nehezen összeköthető.

A következő, életmód-biomedikális topik, mely az 5-ös szám alatt található meg, ahogy már említettem jórészt a 3-as számú általános medikális topik alatt fekszik. Az egyes életmódbeli változtatások, mozgás, étrendek, diéták és kiegészítők hatásait veszi végig a topik, gyakran biomedikális szempontból. Nem ritka eset, hogy a tanácsot biztosító hozzászólók (olykor maguk a tanácsot kérők is) olvasottak és tájékozottak a témában, megjelent tudományos publikációkra hivatkoznak, habár ez nem egyedülálló jelenség erre a topikra (medikai topikokban közel „általános”).

A 6-os számú topik, azaz jóléti-napló mindennapi tevékenységeket követ nyomon, elsősorban pozitív energiával töltve, azzal a céllal, hogy a hozzászólók megosszák saját élményeiket, melyek örömet okoznak, az olvasók pedig ebből töltekezni tudjanak. Gyakran előfordul még napi rutin, hobbik, illetve hétköznapi tevékenységek megbeszélése.

A 7-es topik jellegét tekintve nagyban hasonlít a 4-es számú topikra, tehát értelmező topik, azonban pszichológiai fókusz helyett a biomedikális terápia kerül előtérbe. Egészen közel helyezkedik el térben a rákkal foglalkozó topik mellett, ha nem csupán az első 10 releváns kifejezést vizsgáljuk meg, akkor látható is, hogy van átfedés bizonyos mértékben ezen

³⁶ Schreiber & Hartrick (2002).

kifejezések között a két topikban. Még érdekessége a topik releváns szavainak, hogy belekeveredett néhány kevésbé kapcsolódó kifejezés, melyek a következők: esély (odds), jegy (ticket), lottó (lottery). Elképzelhető, hogy utóbbi kettő az első felől került ide, hiszen az 'esély' előfordul egészségügyi szöveggörnyezetben, azonban most magával hozhatta a szerencsejáték oldalról hozzátartozó szavakat is.

A 8-as az egyik legnehezebben értelmezhető és legkevésbé koherens topik. Releváns kifejezései kimondottan vegyesen és látszólag jórészt nem egészségügyi témából³⁷ állnak össze (például díj – award, jogosítvány – license, vezetni – drive, taxi stb.). Alacsony prevalenciával rendelkező topik, félig átfedésben a 11-es topik alatt elhelyezkedve. Úgy lehet rá tekinteni, mint a topik, amelybe azon kifejezések kerültek, melyek máshová „nem fértek be”.

A 9-es topik az előzőnek a „testvére”, abban az értelemben, hogy szintén nehezen megfogható a centrum. Vegyesen tartalmaz szabadidős, biomedikális, valamint a topik szempontjából nem értelmezhető szavakat. 15-ös, COVID-19 topik alatt helyezkedik el, szintén félig átfedésben, azonban összefüggést nem véltem felfedezni. Az elolvasott bejegyzések alapján, mind a 8-as, mind a 9-es topik esetén, semmilyen kiemelkedő észrevétel nem állapítható meg, tartalmukat tekintve hasonlítanak például a 6-os topikba tartozókhoz.

A következő klasszikus biomedikális tanácsadó névre hallgató, 10-es számot viselő topik és hozzá szorosan kapcsolódó klasszikus pszichológiai tanácsadó, 11-es számú topik a felhasználók egymás közötti segítség nyújtását, tanácsokkal való ellátását fedi le. Az előbbi nyilvánvalóan a biomedikális oldalról közelít, míg utóbbi a pszichológiai aspektusokat veszi előtérbe. Ezeknek a célja saját tapasztalatok megosztása, egymás meghallgatása és laikus támogatása, nem épít, vagy hivatkozik szigorúan szakértői tanulmányokra. Például a rák megjelenése a 10-es topikban, itt azt jelenti, hogy a betegségben szenvedők, vagy azon átesettek elmesélik saját élményeiket a különböző kezelésekről, elsősorban medikális tekintetben.

A 12-es számú topik az úgynevezett szenvedésmonológ. Az ilyen jellegű bejegyzések általában nem egy konkrét gondolati ívet hivatottak átadni, hanem inkább a felgyülemlett

³⁷ Talán a vezetéssel kapcsolatos szavak gyakrabban fordulnak elő, azonban külön topikot így sem alkotnak

negatív gondolatokat, érzéseket, stresszt és kétségbeesést adják ki magukból, ezáltal egyfajta segítséget kérve. Fontos szerepe van annak, hogy legyen egy hely, ahol lehetőség nyílik a negatív energiáktól és automatikus negatív gondolatoktól való megszabadulástól.³⁸ Habár itt meg kell jegyezni, hogy relatív kevés számú bejegyzés tartozik a topikhoz, annak ellenére, hogy a szenvedésmonológ topik kiemelt jelentőséggel bír a depressziós fórumok életében.

A soron következő 13-as topik, szintén egy pszichológiai fókusszal és a legnagyobb prevalenciával rendelkező támogatói topik. 11-es számú topikot nem meglepő módon erősen lefedi, hiszen tematikájukat tekintve is szoros az összefüggés. A releváns kifejezések alapján az mondható el, hogy a 11-es magáról a támogatói közösségről szól, míg a 13-as a családról, kapcsolatokról, esetleg munkáról beszél inkább. Pszichológiai megközelítésben itt van jelen leginkább a rák, tehát a korábban 10-es topik esetén leírtak jelen esetben ezt a perspektívát szem előtt tartva érvényesülnek.

A 14-es ismét egy biomedikális topik, kifejezetten neurológiai központtal. A fejezet elején leírtak alapján a releváns kifejezések közül sharon-t figyelmen kívül lehet hagyni. Szerkezetét és funkcióját tekintve nagyban hasonlít a 3-as számú topikra, tartalmilag viszont a neurológiai vonal következtében láthatóan szeparálódik.

A COVID-19 topik megjelenése a 15-ös szám alatt számottevő jelentőséggel bír, azonkívül jó eséllyel alátámasztja a kutatás előzetes felvetései közül azt az elképzelést, hogy a laikus depressziós diskurzusra hatással volt a világjárvány. A topik prevalencia közepes méretűnek mondható, azonban maga a COVID kifejezés számos helyen fordul elő, melyek közül a legkiemelkedőbbek gyakoriság függvényében a következő topikok: 10-es, 4-es, 11-es, 13-as. Ez azt jelenti, hogy mind biomedikális, mind pedig pszichológiai oldalról körüljárt a téma, az érintettek életkörülményeire több szempontból is hatással van. Az olvasott bejegyzések a szakirodalom által is alátámasztott szempontok szerint jelentek meg, nevezetesen egyrésztől félelem alapúak (aggódás a szerettekért, saját egészségért, jövő miatt), másrésztől elkeseredett hangvételűek (magányos, bezártság érzet, unalmas élet, fontos pillanatok elmaradása) voltak. Természetesen előfordultak heves érzelmekkel túlfűtött impulzív bejegyzések is, amelyekre jelen esetben szintén igaz, hogy a felgyülemlett stresszt,

³⁸ Michikyan (2020)

kétségbeesést és tehetetlenséget hivatottak valamelyest feloldani. A rák magasabb előfordulásának oka(i) a topikban nehezen igazolható(k). Saját benyomásra támaszkodva és célzott olvasást követően, úgy vélem, hogy két ok azonosítása lehetséges elsősorban. Az első az, hogy a daganatos betegséggel küzdők - a platformokon megszólaló - szerettei szerint az egészségügyi ellátás járvány alatti terheltsége következtében a daganatos betegek nem lettek, vagy nem megfelelően lettek gondozva (ennek kapcsán hazánkban is jelentős volt a hírforgatag). A második az, hogy a különböző még friss és gyakran bizalmat nélkülöző COVID gyógyszerek – elsősorban oltások – feltételezett mellékhatásai között a rák kialakulásának eshetőségétől tartanak leginkább a bejegyzést írók és ezen félelmüknek adnak hangot.

16-os szám alatt található vallási topik térben a pszichológiai fókusszal rendelkező támogató és értelmező topikokhoz áll a legközelebb. Ez egy alternatív módja a depresszióval való szembenézésnek, melyre korábbi kutatások³⁹ egyaránt mutattak fel eredményeket. Számos topikban fordulnak elő a hittel és vallási élettel kapcsolatos gondolatok, nem kizárólagos a jelenlétük saját topikjukra nézve – például 10-es, 13-as, 17-es, 18-as topikokban (hit - faith, imádkozni - pray, templom/vallás - church). Dolgozatom szóban forgó eredménye reményeim szerint hozzájárul a jelenleg csupán feltételezett biopszichoszociális-spirituális (biopsychosocial-spiritual) modell elterjedéséhez a depressziós laikus diskurzus vonatkozásában.

A 17-es topikba tartozó bejegyzések a tudományos diskurzus laikus értelmezését tűzték ki célul. Prevalenciáját és célját tekintve megegyezik a 4-es és 7-es topikok értelmező funkciójával, azonban térben sokkal közelebb helyezkedik el a 7-es topikhoz, tehát inkább pszichológiai vonalon megjelent tanulmányokon teszi mindezt. A diskurzusra való hatására és jelentőségére a korábban leírtak a mérvadók (4-es és 7-es topiknál).

Végül a 18-as topik az egyes (elsősorban gyógyszeres) kezelések különböző mellékhatásait veszi szemügyre medikális oldalról. A felhasználók megosztják egymással tapasztalataikat, összegyűjtött információikat bizonyos gyógyszereket illetően, ezzel segítve „normális-e”, amit a kezelés alatt állók adott időszakban éreznek, átélnek. Többször felbukkan (a biomedikális topikokban), hogy a kezelő orvos tájékoztatlanságára, vagy alaposságának

³⁹ Németh et al. (2021); Hatala (2013); Saad et al. (2017)

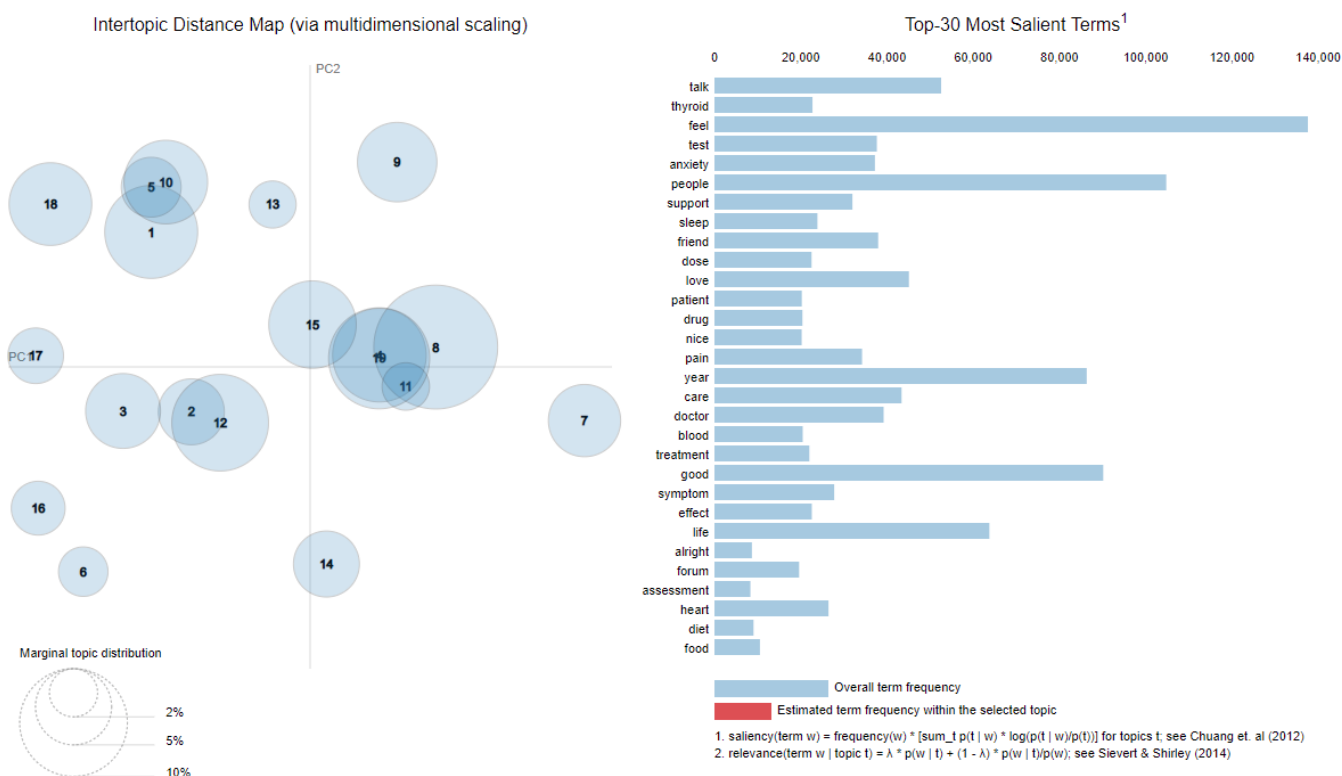
hiányára panaszkodva a közösségtől kapott tanácsok alapján próbálnak a segítségre szorulóknak a gyógyszerek közül számukra megfelelő választani.

A pre-COVID és COVID-éra összevetése

Az alábbi, 19 topikszámmal rendelkező modell vizsgálata azon egyszerű felvetésen alapszik, hogy a koronavírus megjelenésével a már létező és korábban bizonyított 18 topik mellé létrejön egy új 19. topik – COVID-19 néven. Természetesen lehet gyanítani, hogy mivel a diskurzus élő, dinamikus és szerkezetileg egymással szorosan kapcsolódó topikok halmaza, ez nem feltétlenül valósul meg ilyen egyszerűséggel az életben. Mindazonáltal az elemzés nem csupán szükséges a korábbi eredmények összehasonlítása céljából, hanem létjogosult is, hiszen a második legmagasabb koherencia értékkel rendelkező modellről beszélünk. Ami az elemzési eljárást illeti, többé-kevésbé hasonló lesz az előző fejezetben leírtakhoz, azzal a kiegészítéssel, hogy a kialakult topikok párhuzamba állítása a korábbi kutatásban⁴⁰ (melyre a dolgozat épít) létrejött eredményekkel egyaránt szempont lesz. Következésképpen olyan eredmény lesz a kezünkben az elemzés végén, mely alapján megállapítható, hogy a járvány megjelenése előtti topikok miként alakultak át, mely topikok maradtak változatlanul és melyek jelentek meg, vagy tűntek el teljesen a világjárvány alatt.

⁴⁰ Németh, R., Sik, D., & Katona, E. (2021). The asymmetries of the biopsychosocial model of depression in lay discourses-Topic modelling online depression forums. *SSM-population health*, 14, 100785.

model_topics_t19_r04⁴¹



A már jól ismert ábra most a 19 topikszámmal rendelkező modellt ábrázolja, a lábjegyzetben vizualizációt segítő interaktív linkkel ellátva. Ebben a modellben jelenlévő topikok nyilvánvalóan erős hasonlóságot mutatnak a 18 topikszámmal rendelkező, előző fejezetben tárgyalt modell topikjaival. Ennélfogva kizárólag olyan topikok esetén fogok részletes kifejtést írni, melyek az előzményekben nem szerepeltek. Erre három topiknál volt szükség, melyek a szóban forgó modellben a 11-es (kikapcsolódás) és 15-16-os (COVID), valamint a korábbi modellben a 2-es (rák) topikszámot képviselik. Először a 11-es tekintetében, a korábbi modellben a 9-es számú topikhoz képest (ami egy vegyes topik volt) itt tovább szűkítette az algoritmus a bejegyzéseket. Nagyon szorosan kapcsolódik a mindennapi tevékenységek leírásához, napi rutinhoz, kapcsolatok leírásához és a jóléti napló topikokhoz, azonban valamelyest elkülönül egy altémaként, mely kizárólag a szórakozást célozza meg (zene, költészet, YouTube). Továbbra is közel maradva helyezkedik el ez a topik csoport a COVID topikhoz, amely azzal magyarázható, hogy a járvány következtében alkalmazott megkorlátozások elsősorban a mindennapi élet ezen aspektusait érintette. A korábbi COVID topik pedig két részre osztozott, így a 15-ös és 16-os szám alatt található meg. Az első az

⁴¹ <https://github.com/WhisperPotato/Szakdolgozat>

intézkedéseket és a mindennapi életre való hatásait tárgyalja, míg a második szakirodalmi értelmező oldalról közelít.

Az összehasonlításra váró publikációban született eredmények, topik besorolás tekintetében, a soron következő szélesebb kategóriákat⁴² foglalta magába:

- Monológok (monologues):
 - ❖ Attribúciók (attributions)
 - ❖ Önfeltárások (self-disclosures)
- Interakciók (interactions):
 - ❖ Konzultációk (consultations)
 - ❖ Kvázi-terápiás jellegű részvételek (quasi-therapeutic engagements)
- Vegyes (misc):
 - ❖ Médiában valaminek az ábrázolása (media representation of...)
 - ❖ Valaminek a keveréke (mixture of...)

Table 1
Evolution of the topics while increasing their number.

		Number of topics: 7	Number of topics: 13	Number of topics: 14	Number of topics: 18	Number of topics: 19
Monologues	Attributions	Health-related attribution	Health-related attribution Partnership-related attribution Family-related attribution Work-related attribution	Health-related attribution Partnership-related attribution Family-related attribution Work-related attribution	Health-related attribution Partnership-related attribution Family-related attribution Work-related attribution	Health-related attribution Partnership-related attribution Family-related attribution Work-related attribution
		Suffering-monologues Everyday diary	Suffering-monologues Well-being monologue	Suffering-monologues Well-being monologue	Suffering-monologues Everyday diary Well-being monologue Cultural consumption reports Struggle diary Contemplative self analysis	Suffering-monologues Everyday diary Well-being monologue Cultural consumption reports Struggle diary Contemplative self analysis
	Self-disclosures					
Interactions	Consultations	Making sense of drugs Making sense of the experience	Making sense of drugs Making sense of the experience Making sense of psy-discourses	Making sense of drugs Making sense of the experience Making sense of psy-discourses Making sense of life-style intervention	Making sense of drugs Making sense of the experience Making sense of psy-discourses Making sense of psychotherapies	Making sense of drugs Making sense of the experience Making sense of psy-discourses Making sense of life-style intervention Making sense of biomedical discourse
		Spiritual support Recovery helpers counselling	Spiritual support Recovery helpers counselling Unconditional recognition	Spiritual support Recovery helpers counselling Unconditional recognition	Spiritual support Recovery helpers counselling Unconditional recognition	Spiritual support Recovery helpers counselling
	Quasi-therapeutic engagements					
Misc	Media representation of ... Mixture of ...		... mental disorders	... school and work related attributions	... mental disorders	Coaching ... mental disorders
						...suffering monologues and recovery helpers counselling ... suffering monologues and making sense of psy-discourse ... suffering monologues and biomedical discourses, plus experiences

⁴² Kép forrása: Németh et al. (2021)

Az általuk kialakított kategóriákba igyekszem saját magam által létrehozott topikjaimat megfelelően besorolni, párhuzamot felépíteni. Ezen besorolás alakulását végül értékelni fogom, mely egyben a vizsgálat tulajdonképpeni eredménye is lesz. Ehhez először ismét létre kellett hoznom egy összefoglaló topik táblázatot.

Topik megnevezése	Topik száma	Prevalencia	Top-10 releváns kifejezés
<i>Klasszikus biomedikális tanácsadó topik</i>	1	7,4%	pain, year, month, doctor, time, good, back, treatment, week, symptom
<i>Vegyes topik</i>	2	3,8%	assessment, time, report, letter, make, assessor, year, case, drive, question
<i>Biomedikális topik neurológiai fókusszal</i>	3	4,8%	people, year, patient, make, doctor, work, time, give, drug, medical
<i>Klasszikus pszichológiai tanácsadó topik</i>	4	7,5%	support, feel, hope, post, forum, people, talk, time, find, good
<i>Életmód-biomedikális topik</i>	5	3,1%	food, diet, make, good, iron, weight, sugar, supplement, thing, people
<i>Tudományos diskurzus laikus értelmezése</i>	6	2,1%	treatment, bicalutamide, year, drug, patient, people, doctor, story, thing, forum
<i>Szenvedésmonológ</i>	7	4,5%	people, talk, nice, care, feel, fine, friend, heart, alright, good
<i>Pszichológiai támogató topik</i>	8	13,2%	feel, thing, time, life, make, love, people, good, thought, friend
<i>Jóléti-napló</i>	9	5,4%	feel, anxiety, sleep, time, good, night, make, work, morning, start
<i>Medikális topik mellékhatások fókusszal</i>	10	6%	effect, work, side, medication, dose, year, time, pain, doctor, drug
<i>Kikapcsolódás topik</i>	11	1,9%	good, music, song, post, make, time, read, video, love, watch
<i>Pszichológiai értelmező topik</i>	12	8,1%	people, depression, thing, make, life, time, experience, work, problem, issue
<i>Allergia topik</i>	13	1,9%	make, fibro, good, mcas, give, year, thing, test, write, told
<i>Vallási topik</i>	14	3,8%	jesus, love, life, christian, church, lord, give, faith, people, christ
<i>COVID hatásai a mindennapokra</i>	15	6,6%	people, time, walk, work, covid, good, year, make, thing, home
<i>COVID értelmező topik</i>	16	2,5%	study, patient, people, evidence, statin, coronavirus, benefit, covid19, trial, year
<i>Biomedikális értelmező topik</i>	17	2,7%	dose, study, year, supplement, symptom, levodopa, thiamine, parkinson, effect, people
<i>Általános medikális</i>	18	5,9%	test, thyroid, blood, result, level, dose, range,

<i>topik</i>			symptom, high, levo
<i>Család és barátok – kapcsolati topik</i>	19	8,8%	year, time, life, work, family, feel, make, friend, thing, child

Amint az látható, létrejöttek egészen új topikok – COVID 1-2 topik; Kikapcsolódás topik; Kapcsolati topik – míg mások megszűntek létezni – a kettes számú Vegyes topik; Daganatos betegség topik. Az olyan korábban létező topikok melyekre az új topikok kialakulása nem volt elsődleges hatással csupán mérsékelt szerkezeti változás történt (prevalencia értékek, vagy arányaik hasonlóak maradtak). Összességében interpretálva a táblázatot, optimálisabb modellnek tekinthető a 19 topikszámú modell az előző, 18-hoz képest. Bizonyos témakörök árnyaltabban jelennek meg, valamint kevesebb a vegyes topik előfordulás. Ez is igazolja, hogy önmagában a koherencia érték nem elég erős mutató ahhoz, hogy a legmagasabb értékkel rendelkező modellen kívül más modelleket ne vizsgáljunk.

Az összevetéshez használni kívánt kategória besoroláshoz, szintén az átláthatóság miatt, újabb táblázatot hoztam létre:

Főkategória	Alkategória	Topik száma	Topik megnevezése
Monológok	<i>Attribúciók</i>	13	Allergia topik
		15	COVID hatásai a mindennapokra
		18	Általános medikális topik
		19	Kapcsolati topik
	<i>Önfeltárások</i>	7	Szenvedésmonológ
		9	Jóléti-napló
		11	Kikapcsolódás topik
Interakciók	<i>Konzultációk</i>	1	Klasszikus biomedikális tanácsadó topik
		3	Biomedikális topik neurológiai fókusszal
		4	Klasszikus pszichológiai tanácsadó topik
		5	Életmód-biomedikális topik
		6	Tudományos diskurzus

			laikus értelmezése
		10	Medikális topik mellékhatások fókusszal
		12	Pszichológiai értelmező topik
		16	COVID értelmező topik
		17	Biomedikális értelmező topik
	<i>Kvázi-terápiás jellegű részvételek</i>	14	Vallási topik
		8	Pszichológiai támogató topik
Vegyes	<i>Nyomon követés/jelentés/vezetés</i>	2	Vegyes topik

Az ábrázolt táblázat alapján számos következtetés vonható le. Az attribúciók nevezetű alkategóriában nálam külön topikként nem jelent meg a munkával kapcsolatos attribúció. A munka inkább, mint „átívelő-topik” volt jelen, azaz, számos egyéb tematikájú topik részeként jelent meg. Mindazonáltal a COVID topik szépen körvonalazódott és beilleszkedett a kategóriákba, egyrészt attribúciós, másrészt konzultációs oldalról. Még egy érdekesség, hogy az egészség mellett az allergia (ugyan kis méretű, de) külön témaként jelent meg.

Az önfeltárások alkategórián belül nem volt számottevő eltérés, habár a kulturális fogyasztási jelentések témakör nálam nem, vagy legalábbis nem ilyen formában jelentkezett⁴³. Ezzel a monológ főkategóriát le is zárhatjuk.

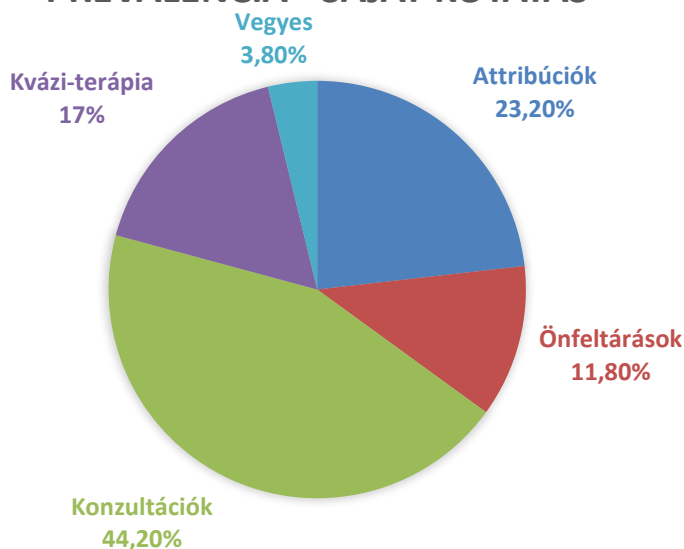
Az interakciók főkategórián belül, a konzultációk alkategóriában, mely a legtöbb topikot magába foglaló alkategória, a várt topikok mellett, itt jelenik meg ismét a COVID topik, egy egészen más szemszögből. Ami a többi topikot illeti, valójában közel ugyanazt tükrözik, – még annak ellenére is, hogy az én általam adott elnevezések olykor eltérőek – melyet Németh Renáta és szerző társai írtak le. Biomedikális oldal nálam valamivel árnyaltabban jelentkezik (míg az attribúcióknál láttuk, hogy ott viszont kevésbé árnyaltan), azonban mivel

⁴³ Talán legközelebb a kapcsolódás topikot sorolnám ide, azonban véleményem szerint ez a kettő nem egyezik meg teljesen.

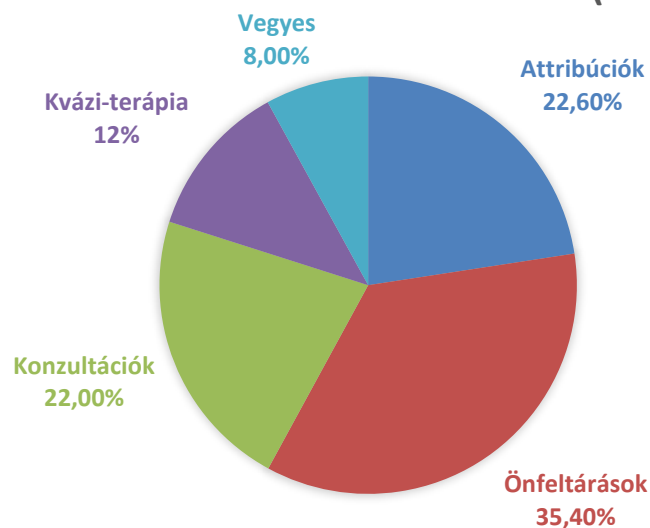
ezek a témakörök sokszor átfedésben vannak, vékony a határvonal, mikor melyik topik kap nagyobb hangsúlyt – összefoglalva azonos témacsoportot fednek le. Emellett a kvázi-terápiás jellegű részvételek alkategóriában szereplő topikok pontosan megegyeznek.

Végül elérkeztünk a vegyes főkategóriához, ahol nem lehetett és nem is állt szándékomban párhuzamot húzni az eredmények között. A korábbi kutatásban ez a szenvedésmonológ, plusz pszichológiai tanácsadó topik, plusz pszichológiai értelmező topik, plusz biomedikális és tapasztalati beszámoló topikok keveréke volt. Dolgozatomban vegyes topik néven található meg a behatárolhatatlan, – és értelemszerűen ebbe a főkategóriába sorolt – vezetéssel és valamiféle (rendszeres jelentés alapú) dokumentációval foglalkozó topik.

ALKATEGÓRIÁK SZERINTI ÖSSZESÍTETT TOPIK PREVALENCIA - SAJÁT KUTATÁS

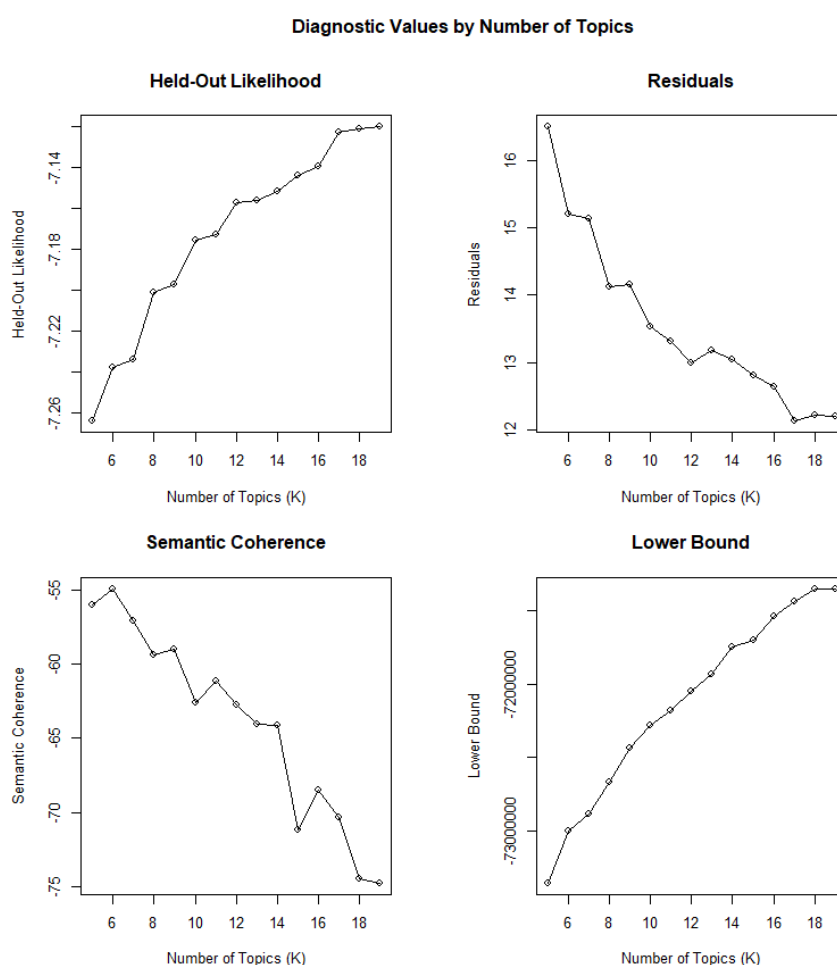


ALKATEGÓRIÁK SZERINTI ÖSSZESÍTETT TOPIK PREVALENCIA - NÉMETH ET AL. (2021)



STM modellek bemutatása

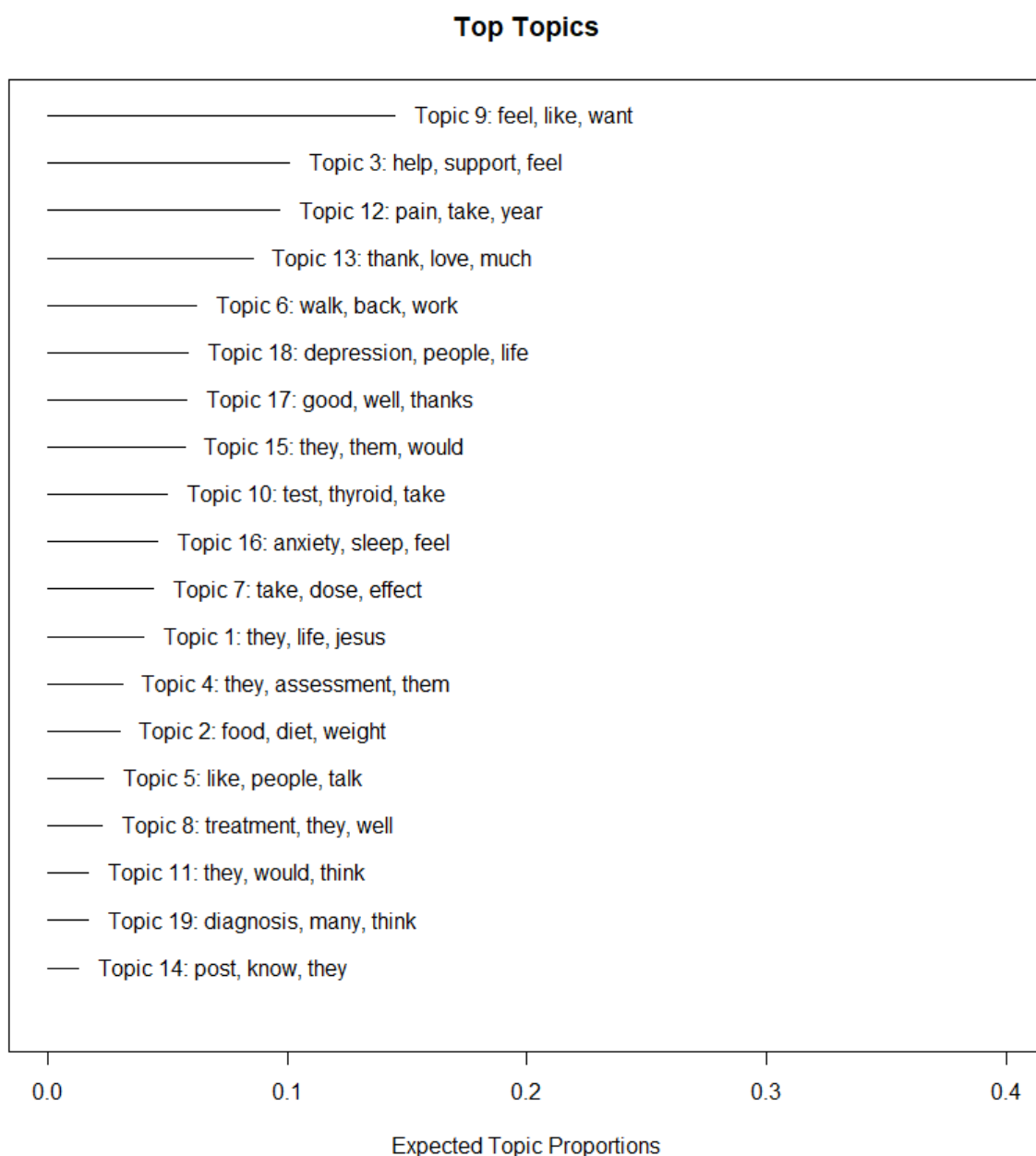
Ahogy az LDA modelleknél is látható volt, első lépésben mindig az optimális topikszám kiválasztása, azaz a megfelelő modell létrehozása történik. STM modellek esetén lehetőség van meta-változók bevonására is, ilyen jelen esetben a bejegyzés létrehozásának időpontja. Ezen változók bevonása hatással van a modellre, jó esetben javítják azt, így érdemes ezek hatását is vizsgálni. Azt a modellt fogom vizsgálni, mely tartalmazza az időt, mint folytonos meta-változót (az idő bevonása dichotóm módon külön érdekes lehet). Az alábbi ábrán diagnosztikai értékek láthatók különböző topikszámú modellek esetén.



Elsősorban a 'Strukturális topikmodell (STM)' című alfejezetben kifejtésre került szemantikus koherencia (semantic coherence) érték érdekel minket, ez az STM alapbeállítása. A grafikon egyértelmű fejlődést mutat magasabb topikszám felé haladva, nagyobb kiugrás a 10-es, 15-ös, 18-as topikszám esetén figyelhető meg, valamint a „legjobb” értékkel a 19 topikszámmal rendelkező modell rendelkezik. Egy az egyben LDA és STM koherencia értékek nem összehasonlíthatók, oly módon, hogy az egyik "jobb", mint a másik. Szakértői annotáláshoz

viszonyítva az LDA-nál használt C_v értékek általában közelebb vannak, azonban minden esetben adott kutatástól és annak szempontjaitól függően kell választani a koherencia értékek közül.

A következő két ábra egyrészt a modell topikjainak prevalencia szerinti sorrendjét, másrészt a topikokhoz tartozó releváns kifejezéseket hivatott bemutatni.⁴⁴



A táblázatból leolvasható topikok alakulása nagyon hasonló az LDA modellnél látottakhoz. Nagyrészt megfeleltethetőek az ott fellelhető topikokkal, így csak a figyelemreméltó eltérésekre térek, külön nem részletezem az egyes topikokat.

⁴⁴ A fejezetben található összes ábra és táblázat saját munka.

Topic 1: they, life, jesus, love, believe, them, would, christian, church, faith, people, come, give, lord, world, know, word, christ, thing, make
Topic 2: food, diet, weight, make, like, need, drink, they, also, them, think, sugar, would, know, good, thing, plant, water, exercise, healthy
Topic 3: help, support, feel, also, know, forum, find, health, need, please, post, talk, mental, like, people, hope, welcome, really, would, depression
Topic 4: they, assessment, them, time, would, people, letter, know, make, report, take, could, virus, even, case, write, care, drive, look, year
Topic 5: like, people, talk, want, nice, know, well, fine, care, sorry, friend, feel, heart, alright, never, really, good, need, always, they
Topic 6: walk, back, work, year, they, home, time, week, like, would, look, house, still, could, hand, come, last, take, start, month
Topic 7: take, dose, effect, drug, sleep, side, work, well, would, they, time, make, night, year, also, help, much, stop, hour, them
Topic 8: treatment, they, well, year, patient, drug, would, cancer, doctor, study, people, bicalutamide, know, cell, like, even, therapy, science, life, time
Topic 9: feel, like, want, thing, know, think, make, life, time, would, myself, really, people, even, something, they, well, work, them, thought
Topic 10: test, thyroid, take, blood, result, level, they, need, would, dose, range, high, hormone, well, symptom, doctor, patient, levo, vitamin, know
Topic 11: they, would, think, year, know, well, like, people, could, make, study, much, good, give, need, take, cost, them, symptom, cure

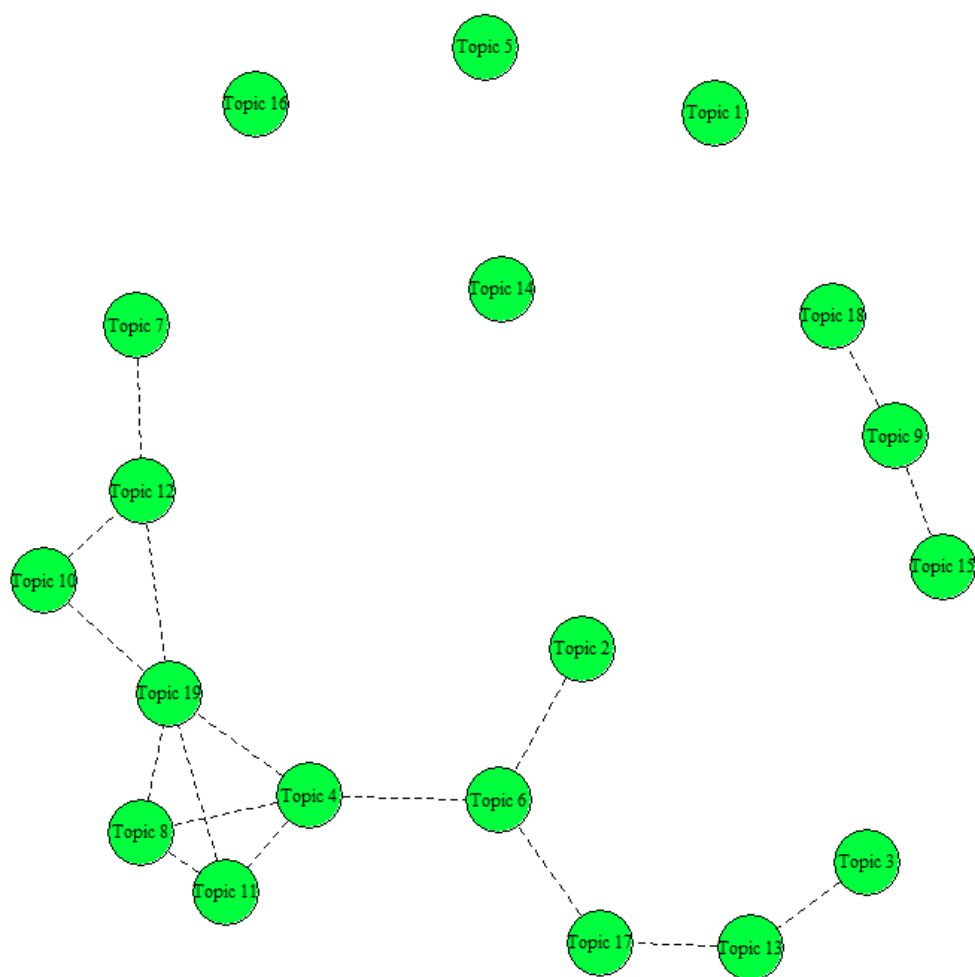
A top-10 releváns kifejezés között a COVID nem szerepel a 6-os topiknál, de ugyanez volt a helyzet a 18 topikszámú LDA modell esetén is, ettől még ő a kinevezett COVID topikunk – ezúttal egymagában. A rákkal foglalkozó (itt 8-as számot viselő) topik ismét erőteljesen körvonalazódott (releváns kifejezések között megjelenik a rák – cancer).

Topic 12: pain, take, year, they, doctor, month, week, time, feel, start, back, well, symptom, help, medication, work, could, also, give, side
Topic 13: thank, love, much, hope, well, care, time, take, know, husband, life, family, wish, give, help, best, need, keep, year, kind
Topic 14: post, know, they, would, read, time, case, think, sharon, intervention, like, last, year, make, even, comment, word, benefit, feel, take
Topic 15: they, them, would, child, family, know, parent, could, told, people, even, year, make, need, happen, want, time, thing, think, mother
Topic 16: anxiety, sleep, feel, help, think, take, well, night, like, watch, time, know, make, morning, start, good, calm, anxious, book, really
Topic 17: good, well, thanks, today, hope, luck, think, keep, look, they, really, take, great, like, morning, know, thing, time, news, them
Topic 18: depression, people, life, they, work, experience, time, many, self, thing, anxiety, change, come, question, challenge, make, learn, others, person, different
Topic 19: diagnosis, many, think, know, disease, test, lupus, rheumatologist, also, would, they, positive, well, different, treatment, diagnose, look, need, feel, take

Ezenkívül a topikok alakulásának általános szerkezetét vizsgálva elmondható, hogy ez a modell szintén nagyon közel áll (talán a legközelebb) a Németh Renáta és szerző társai által létrehozott, korábban ismertetett modellhez. Ennek fényében kifejezetten érdekes a következő, topik korrelációt szemléltető ábra, hiszen az LDA a függetlenségi feltevése okán erre nem képes. Ahhoz, hogy ezen korrelációk bemutatása érthetőbb legyen, először a már megszokott módon, egy topik összefoglaló táblázatot hoztam létre, a megfelelő elnevezésekkel ellátva őket. További észrevételek, hogy elkülönült a korábbi kapcsolati topik, családi és partnerségi attribúciós topikokra, ezenkívül néhány kisebb topik megszűnt (allergia, kikapcsolódás), míg egy – a felsoroltak mellett – napi beszámoló (everday-diary) jellegű topik is született.

Topik megnevezése	Topik száma	Prevalencia⁴⁵	Top-10 releváns kifejezés
<i>Vallási topik</i>	1	4%	jesus, love, life, christian, church, lord, give, faith, people, christ
<i>Életmód-biomedikális topik</i>	2	3%	food, diet, make, good, iron, weight, sugar, supplement, thing, people
<i>Klasszikus pszichológiai tanácsadó topik</i>	3	10,1%	support, feel, hope, post, forum, people, talk, time, find, good
<i>Vegyes topik</i>	4	3,1%	assessment, time, report, letter, make, assessor, year, case, drive, question
<i>Szenvedésmonológ</i>	5	2,3%	people, talk, nice, care, feel, fine, friend, heart, alright, good
<i>COVID hatásai a mindennapokra</i>	6	6,2%	people, time, walk, work, house, good, year, make, thing, home
<i>Medikális topik mellékhatások fókusszal</i>	7	4,4%	effect, work, side, medication, dose, year, time, pain, doctor, drug
<i>Daganatos betegségek (rák) topik</i>	8	2,3%	treatment, cancer, year, doctor, cell, month, bicalutamide, study, make, life
<i>Pszichológiai támogató topik</i>	9	14,5%	feel, thing, time, life, make, love, people, good, thought, friend
<i>Általános medikális topik</i>	10	5%	test, thyroid, blood, result, level, dose, range, symptom, high, levo
<i>Biomedikális értelmező topik</i>	11	1,7%	dose, study, year, supplement, symptom, levodopa, thiamine, parkinson, effect, people
<i>Klasszikus biomedikális tanácsadó topik</i>	12	9,7%	pain, year, month, doctor, time, good, back, treatment, week, symptom
<i>Partnerségi attribúciós topik</i>	13	8,6%	thank, love, hope, care, know, husband, life, family, give, help
<i>Biomedikális topik neurológiai fókusszal</i>	14	1,3%	year, case, time, sharon, intervention, benefit, read, side, month, tremor
<i>Családi attribúciós topik</i>	15	5,7%	year, time, life, family, feel, make, thing, child, mother, parent
<i>Jóléti-napló</i>	16	4,6%	feel, anxiety, sleep, time, good, night, make, work, morning, start
<i>Napi beszámoló topik</i>	17	5,8%	good, well, today, luck, keep, really, great, morning, time, news
<i>Pszichológiai értelmező topik</i>	18	5,9%	people, depression, thing, make, life, time, experience, work, problem, issue
<i>Tudományos diskurzus laikus értelmezése</i>	19	1,7%	treatment, diagnosis, lupus, rheumatologist, test, treatment, disease, know, think, need

⁴⁵ Az ideírt értékek megtalálhatóak a *topic prevalence_STM.pdf* fájlban, GitHub-ra feltöltve.



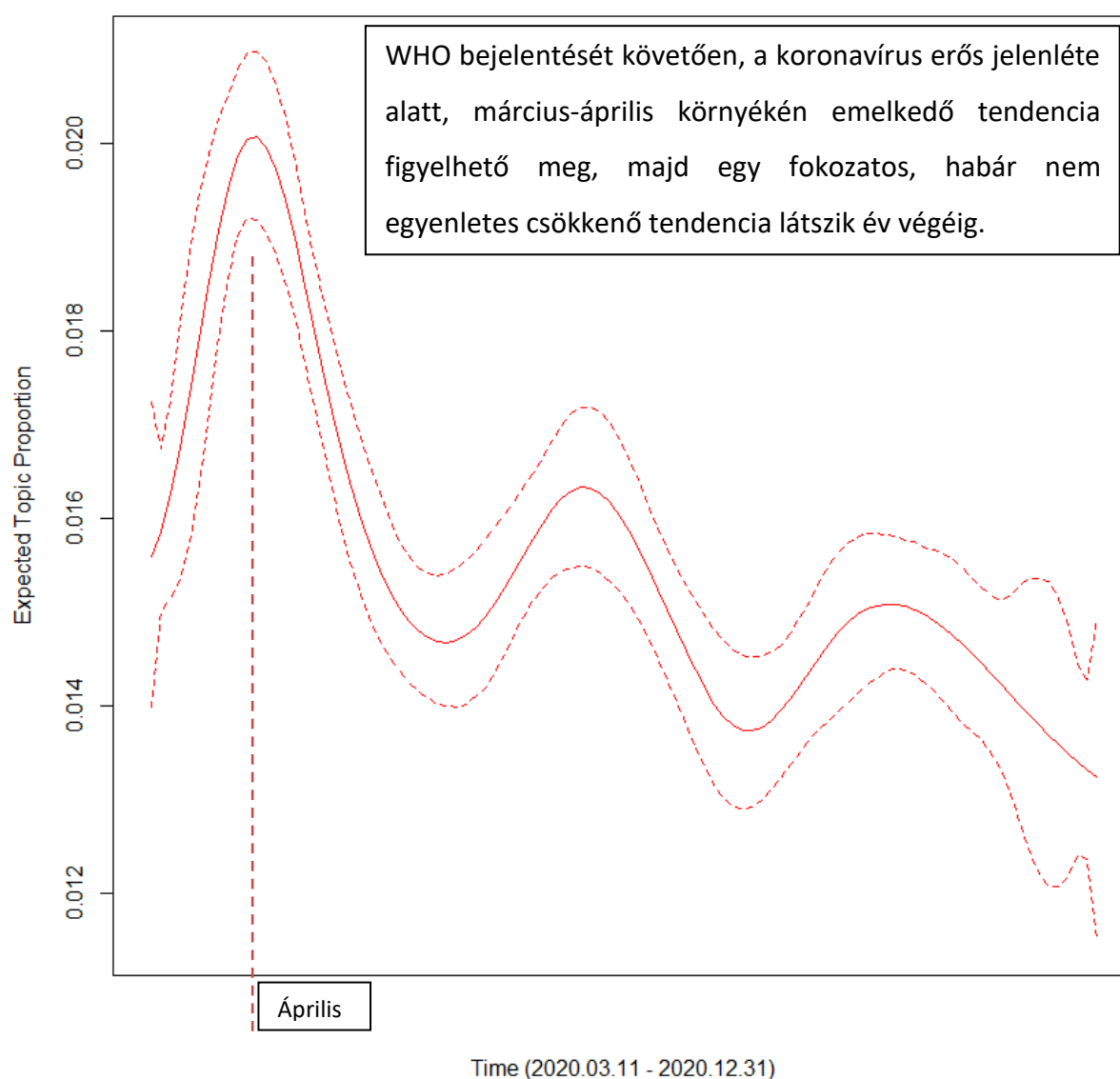
Az itt látható topik-korrelációs ábráról számos megjegyzés fogalmazható meg. Jobb oldalon a 18-as a 9-es topikkal, majd a 9-es a 15-ös topikkal korrelál. Az első kettő pszichológiai témájú topik, előbbi értelmező, míg utóbbi tanácsadó funkcióval, tehát a korreláció nem meglepő. A tanácsadó topik továbbá a családi attribúciós topikkal korrelál, mely arra utal, hogy a családi háttér, kapcsolatok fontos szerepet töltenek be az ilyen jellegű tanácsadások során. Az 1-es, 5-ös 14-es és 16-os topikok⁴⁶ nem korrelálnak a többi topikkal.

A bal oldali részt valójában három részre lehet felbontani. Jobb-alsó öt topik lesz az első rész, mely a 3-as, 13-as 17-es, valamint a 2-es és végül a fő kapcsolódási pontként szolgáló 6-os topikból áll. Az alsó hármas vonal a pszichológiai tanácsadó topiktól jut el a partnerségi topikon keresztül, a napi beszámoló topikhoz, így kapcsolódva végül a COVID topikhoz. Ez a vonal a mindennapi benyomásokat és tapasztalatokat öleli fel, ehhez „felülről” könnyen kapcsolódik az életmód-biomedikális topik.

⁴⁶ Ezek elnevezését lásd a feljebbi táblázatban.

A bal-alsóközép négyes korreláció lesz a második rész. Ide a 4-es, 11-es, 8-as és 19-es topik tartozik. Biomedikális vonal fedezhető fel, a rák topik megjelenésével. Az előző részhez a 4-es vegyes topik szolgál kapcsolódási pontként, mely sajnos túl sok információval nem szolgál.

A harmadik, bal szélső részhez szintén négy topik tartozik (19-es ide is besorolható), ezek a 7-es, 12-es, 10-es és 19-es számú topikok. Kifejezetten egy erős biomedikális topik csoportról beszélhetünk, melyek egymáshoz való korrelációja teljes mértékben logikus (sőt elvárt).



Az itt látható utolsó ábra a 19-es topik prevalencia alakulását mutatja a folytonos meta-változó (idő – day_num) függvényében a koronavírus világjárvánnyá nyilvánítását követően év végéig. Természetesen hasznos volna egy nagyobb, esetleg 5 éves időtávlatot is megvizsgálni, azonban ez egy jövőbeli kutatás feladata lesz megvalósítani.

Konklúzió

A dolgozat eredményei egybevágóan az összehasonlított tanulmány eredményeivel, véleményem szerint olykor új meglátásba helyezve azt. A koronavírus befolyásoló hatása több helyen is megjelenik, élesen nyúl bele a korábban kialakult topik szerkezetbe. Ennek pontos mikéntje még számos oldalról tovább vizsgálható volna egy komolyabb kapacitással bíró kutatás kereteiben – mind az STM által kínált további eszközök kihasználásával, mind pedig a nagyobb volumenű adatok és időintervallum felölelésével. Technikai oldalról ugyan érdekes meglátásokra tehetünk szert az STM modellnek köszönhetően, az LDA sem marad alul az eredmények pontosságát és megbízhatóságát illetően. Számomra a teljes képet a két módszer együttes alkalmazása biztosította.

Irodalomjegyzék

- Aletras, N., & Stevenson, M. (2013, March). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers* (pp. 13-22).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.
- Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, 18, 147
- Biester, L., Matton, K., Rajendran, J., Provost, E. M., & Mihalcea, R. (2020). Quantifying the effects of covid-19 on mental health support forums. arXiv preprint arXiv:2009.04008.
- Biester, L., Matton, K., Rajendran, J., Provost, E. M., & Mihalcea, R. (2021). Understanding the Impact of COVID-19 on Online Mental Health Forums. *ACM Transactions on Management Information Systems (TMIS)*, 12(4), 1-28.
- Bolton, D., & Gillett, G. (2019). *The biopsychosocial model of health and disease: New philosophical and scientific developments* (p. 149). Springer Nature.
- Eisenstein, J., Ahmed, A., & Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 1041-1048).
- Hatala, A. R. (2013). Towards a biopsychosocial-spiritual approach in health psychology: Exploring theoretical orientations and future directions. *Journal of Spirituality in Mental Health*, 15(4), 256–276.
- Hinek, M. (2021). Fesztivállátogatók véleményeinek számítógéppel támogatott tematikus modellezése—egy kísérlet eredményei: Computer-aided topic modelling based on festival-goers' opinions—results of an experiment. *Turizmus Bulletin*, 21(1), 4-12.
- Holtz, P., Kronberger, N., & Wagner, W. (2012). Analyzing internet forums. *Journal of Media Psychology*.
- Jeong, A. (2005). A guide to analyzing message–response sequences and group interaction patterns in computer-mediated communication. *Distance education*, 26(3), 367-383.

- Joglekar, S., Velupillai, S., Dutta, R., & Sastry, N. (2020). Analysing Meso and Macro conversation structures in an online suicide support forum. arXiv preprint arXiv:2007.10159.
- Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2), 254-277.
- Mandravickaitė, J., Briedienė, M., Uus, J., & Krilavičius, T. (2020). What's in the news? Identification of trending topics in alternative and mainstream Lithuanian media. In CEUR Workshop proceedings [electronic resource]: Machine learning for trend and weak signal detection in social networks and social media, Toulouse, France, February 27-28, 2020: proceedings. Aachen: CEUR-WS, 2020, Vol. 2606.
- Michikyan, M. (2020). Linking online self-presentation to identity coherence, identity confusion, and social anxiety in emerging adulthood. *British Journal of Developmental Psychology*, 38, 543–565.
- Mimno, D. M., & McCallum, A. (2008, July). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *UAI* (Vol. 24, pp. 411-418).
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).
- Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 100-108).
- Németh, R., Katona, E. R., & Kmetty, Z. (2020). Az automatizált szövegelemzés perspektívája a társadalomtudományokban. *Szociológiai Szemle*, 30(1), 44-62.
- Németh, R., Sik, D., & Katona, E. (2021). The asymmetries of the biopsychosocial model of depression in lay discourses-Topic modelling online depression forums. *SSM-population health*, 14, 100785.
- Pozsgai-Alvarez, J., & Sanz, I. P. (2021). Mapping the (anti-) corruption field: key topics and changing trends, 1968–2020. *Journal of Computational Social Science*, 1-31.

- Roberts, M. E., Stewart, B. M., & Airolidi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515), 988-1003.
- Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).
- Saad, M., De Medeiros, R., & Mosini, A. C. (2017). Are we ready for a true biopsychosocial-spiritual model? The many meanings of "spiritual. *Medicine*, 4(4), 79.
- Schreiber, R., & Hartrick, G. (2002). Keeping it together: How women use the biomedical explanatory model to manage the stigma of depression. *Issues in Mental Health Nursing*, 23(2), 91–105.
- Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. ACL workshop on interactive language learning, visualization, and interfaces. Baltimore: Association for Computational Linguistics.
- Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).
- Wesslen, R. (2018). Computer-assisted text analysis for social science: Topic models and beyond. arXiv preprint arXiv:1803.11045.