

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?



Granát Marcell Péter

Neptun-kód: AYCOPF

Közgazdasági elemző MA

I.évfolyam

Mazzag Bálint

Neptun-kód: G1UTIO

Alkalmazott Közgazdaságtan BA

IV. évfolyam

Témavezető: Vékás Péter

2021. április 12.

Absztrakt

Kutatásunkban szótáralapú szövegelemzési módszerekkel megvizsgáljuk, hogy a különböző európai országok sajtótermékeiben megjelenő, koronavírusról szóló hírek milyen szentimenttel rendelkeztek az elmúlt egy évben, illetve mi befolyásolta ennek alakulását. Ennek tárgyalásához ökonometriai módszerekkel elemezzük mind a fertőzöttek és halálozások számát, mind pedig a gazdasági mutatók havi szentiment idősorokra gyakorolt hatását.

Eredményeink alapján elmondható, hogy mind az emberek gazdaságról formált véleménye, mind a járványhelyzet jelentősen befolyásolta a média hangulatát az elmúlt egy évben. A kezdeti negatív hangulatot, melyet a korai cikkek keltettek, 2020 nyarára leváltották a neutrálisabb témák, és a halottak számára is egyre érzéketlenebbé vált az európai média.

Kutatásunk szakirodalomhoz hozzáadott értéke, hogy az európai országok koronavírusra adott reakcióit szövegelemzéssel kutató irodalmat a kézirat lezárásának dátumakor nem találtunk, továbbá a témakör szövegelemzéséhez alkalmazható specifikus szentiment szótárat alkottunk és tettünk közzé, az elemzéshez alkalmazott szentiment idősorokon túl.

Tartalomjegyzék

1	Bevezetés	4
2	Elméleti megfontolás	5
3	Adatok bemutatása.....	9
3.1	Korpusz leírása.....	9
3.1.1	Google Translate használata	11
3.2	Járvánnyal kapcsolatos és gazdasági adatok bemutatása.....	12
4	Szövegelemzés.....	13
4.1	Szentimentelemzés.....	13
4.2	TF-IDF elemzés	18
4.3	Témamodellezés	20
5	Ökonometriai modellek	23
5.1	Leíró statisztikák.....	23
5.1.1	Bizonytalanságkerülés	26
5.2	Longitudinális elemzés	27
6	Konklúzió.....	30
7	Irodalomjegyzék	31
8	Függelék.....	34

Ábrák jegyzéke

1. ábra.	A tanulmány felépítése.	8
2. ábra.	Leggyakrabban előforduló szavak a magyar nyelvű cikkekben a fordítást megelőzően és azt követően.....	11
3. ábra.	Az új esetek száma és a szentimentet tartalmazó szavak száma naponta	13
4. ábra.	Szentimenttel bíró szavakkal korreláló szavak hálójá.	15
5. ábra.	A szentiment alakulása országoként.....	17
6. ábra.	A cikkekben havonta leginkább jellemző szavak, TF-IDF értékek alapján. ...	19
7. ábra.	Leggyakoribb szavak topikonként	21
8. ábra.	Topikok jellemzése szentiment, dátum és gyakoriság szerint.....	22
9. ábra.	Regressziós fa	24
10. ábra.	A hírekben megjelenő szentiment és gazdaság érzékelési index közötti kapcsolat	25
11. ábra.	A teljes korpusz leggyakoribb szavai, havonta	34

12. ábra. Leggyakrabban előforduló pozitív és negatív szentimenttel rendelkező szavak szófelhője	35
13. ábra. Szentimentet tartalmazó szavak időbeni alakulása, 7 napos mozgóátlaggal	36
14. ábra. Témák optimális számának megválasztása	36
15. ábra. Topikok relatív megoszlásának időbeni dinamikája	37

Táblázatok jegyzéke

1. Táblázat. Letöltött híroldalak bemutatása	10
2. Táblázat. Regressziós eredmények időszakra való kontrollálás nélkül	28
3. Táblázat. Becsült regresszió paraméterek kontrollálva az időszakokra.....	29
4. Táblázat. A leggyakoribb olyan szavak a korpuszban, melyek pontszámát megváltoztattuk a koronavírus-specifikus szótárban	35

1 Bevezetés

A koronavírus-járvány hatása az egész világon érzékelhető, az elmúlt évek egyik legjelentősebb gazdasági és társadalmi válságát okozta. A vírus a társadalom jövőjére nézve is bizonytalanságot és további kérdéseket ébresztett és egyre nagyobb kockázati tényezőt jelent. A WHO (Egészségügyi Világszervezet) 2020. január 30-án nyilvánította az új koronavírus-járványt (COVID-19) világjárvánnyá, a kézirat lezárásának időpontjában pedig globálisan több, mint 120 millióan fertőződtek meg és több, mint 2 és fél millió emberáldozatot követelt a betegség. (WHO, 2020) A járvány megelőzésére és visszaszorítására a szervezet legfőbb javaslatai között találhatjuk a maszkviselést, otthoni karantént, valamint távolságtartást, amelyeket a világ legtöbb országában a járvány megjelenése után már jogszabályként fogalmaztak meg. Számos esetben teljes országokat zártak le a járvány terjedésének megfékezésére, amelyek több hónapig is eltarthattak.

Elemzésünkben arra keressük a választ, hogy azon legnagyobb médiumok, amelyekből az európai országok lakosainak legnagyobb többsége tájékozódik, milyen módon mutatták be a járványt, a cikkek szövegeiben milyen kontextusban szerepelt a pandémia. Ennek bemutatását azért találtuk fontosnak, mivel általánosságban elmondhatjuk, hogy a sajtóorgánumok hajlamosak túlreagálni egy-egy hírt, eseményt vagy válságot. Az pedig, hogy a koronavírussal kapcsolatban milyen érzelmek társultak a hírekhez befolyásolhatja az olvasókat, akik ez alapján hozhattak döntéseket, valamint a mentális egészségükre is hatással lehetett. Kutatjuk azt is, hogy az adott országok fertőzési, vagy halálozási mutatószámait, vagy éppen az országok gazdasági teljesítményét mennyire reflektálta az adott hónapban a koronavírussal kapcsolatos cikkek érzelmi töltete. Ez abból a szempontból lehet különösen fontos, hogy megvizsgáljuk, hogy valóban érvényes-e az a hipotézisünk, miszerint a vírus lefolyásának objektív és valós bemutatása az újságírók feladata volt.

Hipotézisünk szerint azonban a koronavírusról való tudósítás a járvány hosszú, már több, mint egy éves lefolyása során a médiában gyakori jelenséggé aposztrofált „fáradtság” állapotába került. Ennek következtében egy idő után az újdonsággal nem szolgáló hírek kevésbé lesznek polarizáltak, nem jelentenek akkora szenzációt. Ezt a feltételezésünket vizsgáljuk meg panelelemzéssel az általunk választott országokra, amelynek keretében figyelembe kell vennünk nem csak a járvány terjedésével kapcsolatos fertőzöttségi adatokat, hanem különböző gazdasági mutatószámokat is.

2 Elméleti megfontolás

A szövegelemzéshez kapcsolódó kutatások rendkívül széles skálán mozognak, mivel a korpuszok általában bármilyen szöveges anyaghoz, például honlapok tartalmához, publikációs szövegekhez, de akár a nyomtatott sajtóhoz, vagy a publikusan elérhető közösségi média tevékenységekhez is kapcsolódhatnak. A koronavírushoz kapcsolódó kutatások legnagyobb része a közösségi médiához kötődik, így a Twitteren közzétett tweetekre, illetve Facebook posztokra, vagy kommentekre épül, mivel ezzel vizsgálhatják meg a szerzők az emberek közvetlen és azonnali reakcióját a járványra, vagy éppen az országukban alkalmazott korlátozásokra, szabályokra. Ilyen például Mansoor et al. (2020) írása, akik a távmunkáról és a távoktatásról szóló indiai tweeteket, valamint Boon-Itt et al. (2020) kutatása, amelyben az amerikai tweetek időbeli alakulását, szentimentjét és főbb témáit elemezték.

Ezzel szemben ugyanakkor a mi írásunkban leginkább a hagyományos médiát vesszük figyelembe, így a kiválasztott európai országok állami online sajtótermékeiből töltöttünk le cikkeket. Ezáltal mi nem közvetlenül figyeljük meg és elemezzük az emberek reakcióit, hanem a sajtótermékek által átadott üzenetek alakulását vizsgáljuk, amelyek szerintünk szintén indikátorai és mérőszámai lehetnek egy ország általános érzelmi hangulatának. A vizsgált országokban egyre nagyobb arányban, átlagosan 77%-ban az online médiából értesülnek az állampolgárok a hírekről, amelyekbe beletartoztak az elmúlt egy évben a vírussal kapcsolatos legújabb kutatások, adatok és intézkedések is. A legfőbb hírforrások általában egy-egy nagyobb médiakonglomerátum tagjai voltak, például Magyarországon az index.hu honlapja, ugyanakkor az állami médiumok számos országban nagyobb szerepet kaptak, így például a BBC az Egyesült Királyságban. (Reuters Institute, 2021) Ezen kívül természetesen a közösségi média is fontos csatornája volt a híradásoknak, többek között, Magyarország miniszterelnöke is sok esetben a Facebookon jelentette be a korlátozásokat, ezzel is mutatva az online közösségi tér egyre növekvő erejét.

Korpuszunk olyan szempontból kiemelendő, hogy az európai országok koronavírusra adott reakcióit szövegelemzéssel kutató irodalmat a kézirat lezárásának dátumakor nem találtunk.

A harmincegy országból kapott adatokat legtöbbször az országok saját nyelvén tudtuk letölteni, így a korpuszok elemzésének komplexitása jelentősen megnőtt. Többnyelvű korpuszok elemzésére a szakirodalomban két főbb típusú megközelítést láttunk: vagy mindegyik nyelvet külön elemezzük, mint például Dashtipour et al. (2016) vagy Volkova et al. (2013) írásaiban, vagy pedig egy közös nyelvre alakítjuk át az összes szöveget, mint Can et al.

(2018) vagy Denecke (2008) kutatásaiban. Mivel a későbbiekben bemutatásra kerülő szentimentelemzéshez egy szentimentszótár használata is szükséges, amely jelentős eltérést mutathat a nyelvek között, ezzel összehasonlíthatatlanná téve az adatainkat, így a második opció mellett döntöttünk. Ezen módszerhez a gépi fordításra képes Google Translate eszközt használtuk, amelyet korábban számos kutatásban is használtak már (Balahur és Turchi, 2014), valamint a módszer hatékonyságát is bemutatták több esetben. De Vries és társai például 2018-as cikkükben az Európai Parlament szövegeinek párhuzamos korpuszait vizsgálva jutottak arra a megfigyelésre, hogy a szózsák-moddellrel készült dokumentum-kifejezés mátrix, valamint a témamodellezési algoritmus szinte azonos eredményeket mutat a Google Translate-tel lefordított korpuszaikon, mint a professzionális fordítók által leírt dokumentumokon.

Mivel kutatásunkban szintén a szózsák-moddellre alapozva alkotjuk meg az országokhoz tartozó szentiment idősorokat, valamint egyéb szövegelemzési lépéseinkben is erre fogunk támaszkodni, így végül minden ország cikkeit a Google Translate program segítségével, angolra fordítottuk le.

A szózsák-moddell (bag-of-words) a szöveg egyéb tulajdonságait (például a szemantikus jelentést, vagy a nyelvtant) kizárva, mindössze a szavak összességére, valamint a különböző szavak eltérő gyakoriságaira és előfordulásaira összpontosít. A módszert először 1954-ben Zellig Harris mutatta be (Harris, 1954), amely utána számos elemzésnek, így az általunk felhasznált szentimentelemzésnek is alapjául szolgál.

A szentimentelemzés a szövegelemzés egy klasszifikációs típusa, amelyben a szövegből annak érzelmi töltetét kívánjuk kinyerni, így például meg tudjuk állapítani, hogy az adott szó, mondat vagy bármilyen szövegrészlet mennyire pozitív vagy negatív töltetű, vagy éppen milyen érzelmeket tartalmaz. A szentimentelemzés legegyszerűbb formája szótáralapú, amelyben előre meghatározott szavak előfordulási gyakoriságai alapján rendelhetünk a szöveghez pontszámot, amelyeket utána átlagolunk. Ilyen típusú szótárakat előre kalibráltan találhatunk az irodalomban: a gépi tanulási módszerekkel kalibrált, valamint a szavak szemantikai tulajdonságait is figyelembe vevő SentiWordNet szótárat (Baccianella et al, 2010); a pozitív, negatív szentimentet, valamint hat alapérzelmet is megkülönböztető NRC lexikont (Mohammad et al. 2013); vagy a szavak szentimentjét 5-től -5-ig terjedő skálán pontozó AFINN lexikont is (Nielsen, 2011). Az írásunkban végül a Bing lexikont (Hu és Liu, 2004) használjuk, amely az egyik elsőként kalibrált szótár volt az irodalomban és amely mindösszesen pozitív és negatív kifejezéseket tartalmaz, amelyeknek +1-es és -1-es pontszámokat adtunk.

Ugyanakkor fontos kiemelni, hogy azon elemzésekben, amelyekben a korpusz egy bizonyos témakört vesz a középpontjába, így például gazdasági vagy politikai töltetű szövegek vizsgálatánál, a szerzők saját maguk által, általában géppel vagy kézzel készített szótárakkal dolgoznak. (Drávucz et al., 2017) Így mi is elkészítünk egy, a koronavírussal kapcsolatos szövegekre használható szótárt a Bing szótárból kiindulva, majd pedig az ebből kapott szentimentek napi és havi idősoraival dolgozunk tovább.

A szövegelemzésünk további lépéseiben szintén a szózsák-modellre támaszkodunk és TF-IDF elemzést, valamint topik modellezést végzünk el. A TF-IDF (term frequency - inverse document frequency) elemzés a kifejezések gyakoriságait és az azokat tartalmazó dokumentumok gyakoriságainak reciprokait felhasználó szövegelemzési módszer. Ezzel leginkább azon szavakat, vagy egyéb n-gramokat¹ szűrhetjük ki, amelyek egy adott dokumentumrészletben fontosak, ezzel pedig a részkorpuszok jellemzőit azonosíthatjuk. A módszer lényege, hogy nem csak az n-gramok előfordulási gyakoriságát (TF) vizsgáljuk meg a fontosság méréséhez, hanem azt is, hogy hány alkorpuszban található meg és ennek inverzét mérhetjük az IDF értékkel, amelynek magas értéke azt mutatja meg, hogy mennyire specifikus az adott n-gram a dokumentumok között. Amennyiben tehát a két mérőszámot összeszorozzuk, megkaphatjuk, hogy melyek azok a szavak, amelyek kevés dokumentumrészletben fordulnak elő mégis nagyobb gyakorisággal rendelkeznek a teljes korpuszra vetítve. Ez a módszer ugyanakkor nem csak a korpuszelemzésben használatos, hanem számos egyéb információelméleti kutatásnak is része (Aizawa, 2003), valamint akár a keresőmotorok optimalizálásában is szerepet játszhat. (Ntoulas et al., 2004)

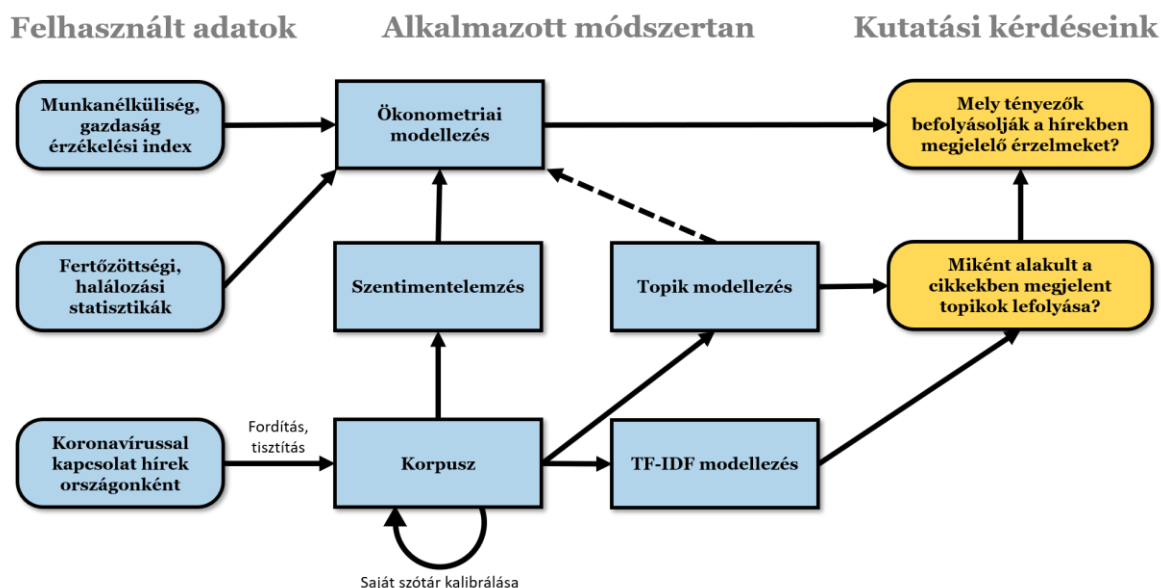
A topik modellezés alapvető célja a korpuszban található legjobban elkülöníthető témák megtalálása, és az ahhoz tartozó főbb n-gramok azonosítása. Ezen módszer leginkább a cikkek mögött található szemantikus tartalmat segíti azonosítani, így pedig könnyebben megfigyelhetjük, hogy milyen témájú médiatartalmak kerültek publikálásra a vizsgált időszakban az európai országokban. A látens, tehát a háttérben meghúzódó, szemantikai elemzés az előző fejezetben tárgyalt TF-IDF metódus továbbfejlesztése volt Deerwester et al. (1990) által, mivel addig mindösszesen a megfigyelt gyakoriságok alapszintű elemzésével foglalkozott a szakirodalom, míg a kifejezés-dokumentum mátrix magasabb szintű elemzésével a háttérben meghúzódó szemantikus összefüggéseket is ki lehet mutatni. A topik

¹ Egymás közvetlen szomszédságában előforduló szópárok.

modellezést pedig már a koronavírus témájában is gyakran használták az irodalomban, így például Älgå et al. (2020) a járvánnyal kapcsolatos kutatói publikációk témáit vizsgálta meg, valamint Tao et al. (2020) tweeteket vizsgáltak globálisan, amelyekből az emberek főbb problémáit próbálták leszűrni a járvánnyal kapcsolatban. Jelen kutatásunkban mi is hasonló típusú elemzésre használjuk a topik modellt. A cikkekben fellelhető témakörök lehatárolása, valamint ezek időbeni megjelenítése képezik kutatásunk részét.

A korpusz elemzése után a koronavírussal kapcsolatos fertőzöttségi és halálozási adatokat egy panelelemzéssel vizsgáljuk. A gazdasági szakirodalomban fellelhető cikkek többsége a napi új fertőzés- és halálozet-számokkal magyarázta a gazdaság tevékenységét. Például Al-Awadhi et al. (2020) a napi tőzsdeindexek hozamait magyarázta a koronavírus változóival, valamint cégek mérleg- és piaci adataival. Topcu és Gulal (2020) cikkében az eredményváltozók a fejlődő országok tőzsdeindexei voltak, a magyarázóváltozók között pedig szintén megjelentek egyéb piaci mutatók is, mint például az olajárindex, vagy a helyi devizák dollárral szembeni árfolyamai. Ezen kívül ugyanakkor számos kutatás szólt a korlátozások sikerességének felméréséről, amelyekben a panelelemzések mellett térstatisztikai módszerek is fellelhetőek voltak. (Oehmke et al., 2021 és Sarkodie és Owusu, 2020)

A saját kutatásunkban a felsoroltakhoz hasonlóan az új fertőzöttek és halálozások napi és havi mérőszámaival dolgozunk, valamint egyéb gazdasági adatokat is bevonunk a panelmodellbe, így például a munkanélküliséget, valamint az Eurostat által létrehozott gazdasági érzékelési indikátort (ESI) használtuk. (Gelper és Croux, 2010) Az említett adatokat, módszertani lépéseket és kutatási kérdéseket az 1. ábra jeleníti meg vizuálisan.



1. ábra. A tanulmány felépítése.

3 Adatok bemutatása

3.1 Korpusz leírása

Korpuszunkat az Európai Unió jelenlegi 27 tagországa mellett még az Egyesült Királyság, Izland, Norvégia és Svájc különböző hírforrásaiból gyűjtöttük. Minden országból egy médiumot választottunk: amennyiben létezett állami hírforrás, valamint ennek online formája webes adatgyűjtésre (webscrapingre) alkalmas volt, úgy ezt választottuk ki (az 1. táblázatban jelöltek szerint ez a vizsgált 31 országból 19-ben volt megtalálható²). Amennyiben nem, úgy a legtöbbet olvasott online hírportálok közül a legnagyobb, az általunk ismert módszerekkel letölthető oldalt jelöltük ki. Bármilyen portált is választottunk, amennyiben megtalálható volt, a koronavírussal kapcsolatos híreket összegyűjtő aloldal (ilyen volt például Magyarországon a www.hirado.hu/rovatok/koronavirus honlap), úgy mindösszesen az ott található cikkeket töltöttük le, míg egyéb esetekben a *covid* és - az adott oldal nyelvére lefordított - *koronavírus* kifejezésekre rákeresve az összes kiadott találatot.

² Belgium esetében az ország hivatalos nyelveinek beszélőinek számából adódóan mind holland, mind pedig francia nyelvű állami hírportált is letöltöttünk, ezen cikkeket pedig ömlesztve kezeltük.

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?

Ország	Letöltött hírforrás	Állami?	Kezdő dátum	Végdátum	Cikkek száma
Ausztria	Die Presse	Nem	2020-01-09	2021-01-31	6580
Belgium (Holland)	VRT	Igen	2020-02-04	2021-01-31	12239
Belgium (Francia)	RTBF	Igen	2020-02-04	2021-01-31	12239
Bulgária	Bulgarian National Television	Igen	2020-03-07	2021-01-31	3188
Svájc	SWI swissinfo.ch	Nem	2020-01-20	2021-01-31	1009
Ciprus	Cyprus Mail	Nem	2020-01-20	2021-01-31	4622
Csehország	Česká televize	Igen	2020-03-13	2021-01-31	2997
Németország	DER SPIEGEL	Nem	2020-02-02	2021-01-31	8225
Dánia	Politiken	Nem	2020-02-11	2021-01-31	2599
Észtország	ERR	Igen	2020-01-14	2021-01-31	3926
Görögország	ERT	Igen	2020-01-20	2021-01-31	1283
Spanyolország	RTVE	Igen	2020-01-16	2021-01-31	7188
Finnország	Yle News	Igen	2020-03-19	2021-01-31	9508
Franciaország	France 24	Igen	2020-01-06	2021-01-31	3001
Horvátország	Večernji list	Nem	2020-01-28	2021-01-31	8263
Magyarország	hirado.hu	Igen	2020-01-09	2021-01-31	9063
Izland	RÚV	Igen	2020-02-01	2021-01-29	6369
Olaszország	la Repubblica	Nem	2020-01-11	2021-01-31	47643
Litvánia	LRT televizija	Igen	2020-01-20	2021-01-30	10906
Luxemburg	L'essentiel	Nem	2020-01-09	2021-01-31	1489
Lettország	LSM	Igen	2020-02-27	2021-01-30	2671
Málta	TVM	Igen	2020-03-13	2021-01-31	2557
Hollandia	NOS	Nem	2020-01-09	2021-01-31	4858
Norvégia	NRK	Igen	2020-01-06	2021-01-31	3223
Lengyelország	TVP	Igen	2020-01-14	2021-01-31	8660
Portugália	RTP	Igen	2020-03-10	2021-01-31	24535
Románia	TVR	Igen	2020-02-22	2021-01-31	7646
Svédország	SVT	Igen	2020-01-22	2021-01-31	4545
Szlovénia	RTV Slovenija	Igen	2020-01-09	2021-01-31	2494
Szlovákia	Nový Čas	Nem	2020-02-06	2021-01-31	14801
Egyesült Királyság	The Guardian	Nem	2020-01-18	2021-01-27	3716

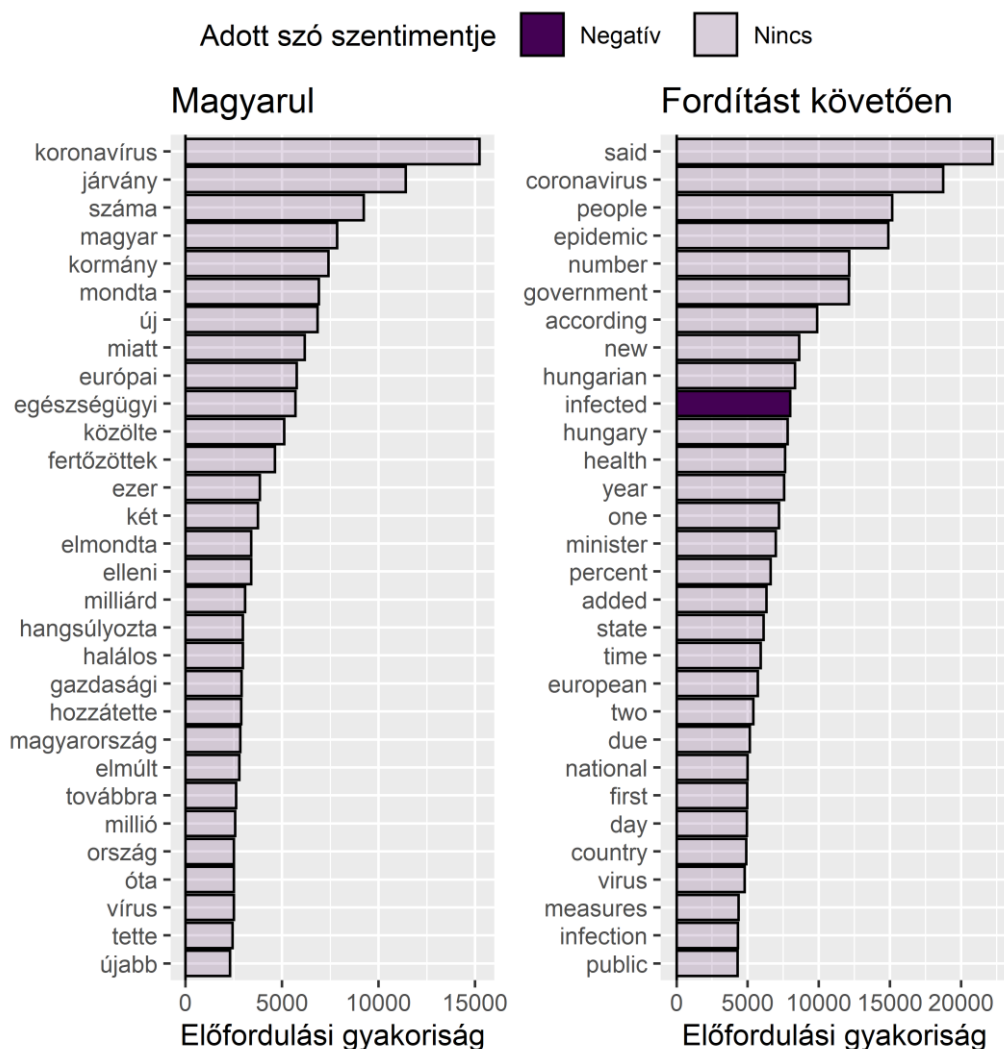
1. Táblázat. Letöltött híroldalak bemutatása.

Az 1. táblázatban megtalálhatóak az adott országokhoz tartozó hírportálok nevei, továbbá a letöltött cikkek száma és nyelve, valamint az első, illetve utolsó cikk dátuma. Az adatok vizsgálatát a 2020. január 1-től 2021. január 31-ig tartó intervallumon végezzük, ugyanakkor láthatjuk, hogy mindegyik országban máskor kezdődött a koronavírussal kapcsolatos hírek átfogó publikációja. A leíró statisztikák alapján elmondhatjuk, hogy a kijelölt időszak alatt átlagosan 7 813 cikk került megírásra a vizsgált országokban, és a teljes

körpuszban 479 190 különböző token (szövegből kinyert szó) található meg, összes számuk pedig 101 612 671, így egy cikkben átlagosan körülbelül 431 szó fordult elő. Az adatok letöltéséhez használt R kódok az alábbi linken érhetőek el: <https://github.com/MarcellGranat/CoronaSentiment>.

3.1.1 Google Translate használata

Az adatok letöltése után a szakirodalomban említettek szerint gépi fordítási eszközzel, a Google Translate programmal alakítottuk egynyelvűre a körpuszunkat. Ezt a fordítást a cikkek címeire, valamint teljes szövegére alkalmaztuk, amelyhez a Google Translate dokumentum fordító szolgáltatását használtuk fel.



2. ábra. Leggyakrabban előforduló szavak a magyar nyelvű cikkekben a fordítást megelőzően és azt követően.

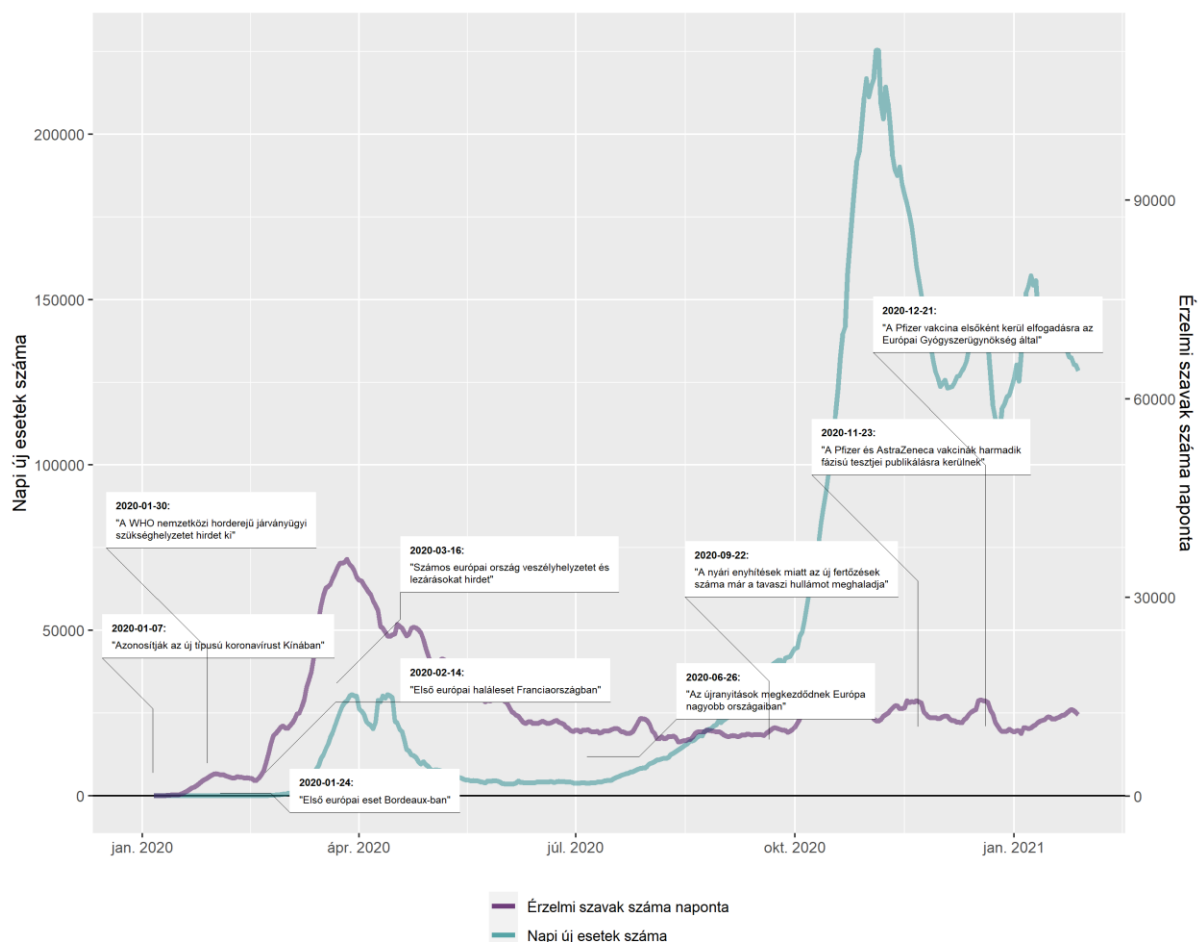
Ahhoz, hogy bemutassuk, hogy a Google Translate-el való fordítás valóban nem alakította át a cikkek mondatait és szavait, a 2. ábrán látható a magyar nyelvű korpusz és annak lefordított verziójának leggyakoribb szavai, külön kiemelve a szentimenttel rendelkezőket. Fontos megjegyezni, hogy ez az ábra mindkét korpusz töltelékszavait (stopword), tehát a gyakran használt, tartalommal nem rendelkező szavait (például: névelők, névmások, számnevek) kiszűrte. Elmondhatjuk, hogy a leggyakoribb magyar szavak közül mindösszesen kilenc darab nem található meg az angol szavak között, amelyek mindegyike vagy névmás, vagy pedig számnév írott formája, amelyek az angol szövegben kiszűrésre kerültek. A leggyakoribb angol nyelvű szavak közül pedig mindössze egy rendelkezik szentimenttel a saját szótárunk szerint.

Havi bontásban a leggyakrabban előforduló szavak a teljes korpuszban a természetesen a *koronavírus* és a *covid* kifejezések voltak, azonban az előbbi inkább a korábbi periódusokban volt gyakoribb, és az egyszerűbb *covid* megfelelője pedig fokozatosan jelent meg, ahogyan a szóhasználatunkba egyre gyakrabban került be ez az elnevezés. Ezt a folyamatot mutatja meg a függelék 11. ábrája.

3.2 Járvánnyal kapcsolatos és gazdasági adatok bemutatása

A járvány adatai közül a koronavírus hullámainak azonosításához elsődlegesen használt új fertőzések és új halálozások napi és havi adataival dolgoztunk, amelyeket a Johns Hopkins University oldaláról (Dong et al., 2020) töltöttünk le. Ahhoz, hogy az adatokat egy panelelemzésbe be lehessen illeszteni, normálni is kellett az országok lakosainak számával, így minden esetben ezer főre vetítettük ezt az adatot.

A 3. ábrán a fertőzési számok alapján azonosítottuk a járvány európai szintű hullámain, így 2020 június és július hónapjai számítottak az első hullám végének, míg ez év augusztusában kezdődött a második hullám. A WHO kezdeti bejelentéseit, illetve a főbb eseményeket jelöltük a grafikonon.



3. ábra. Az új esetek száma és a szentimentet tartalmazó szavak száma naponta

A felhasznált gazdasági adatokat az Eurostat honlapjáról töltöttük le. A célunk az volt, hogy minél gyakoribb rendszerességű általános, gazdasági situációt leíró adatot gyűjtsünk, így a munkanélküliségre, valamint az ESI-re esett a választásunk. Míg az előbbi az országok hivatalos statisztikai közleményeiből kerül publikálásra, addig az ESI az Európai Bizottság saját, primer felméréseken alapuló mérőszáma³.

4 Szövegelemzés

4.1 Szentimentelemzés

A szentimentelemzés szótáralapú módszeréhez a szakirodalom alapján saját szótárt kalibráltunk, mivel a koronavírus-specifikus szavak közül számos olyan volt, amely nem a

³ Svájc, Izland és Norvégia esetében az adatok az Eurostat adatbázisa miatt hiányoznak, míg az Egyesült Királyság a Brexit véglegesítése miatt szüntette be az adatszolgáltatást 2020 novemberében.

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?

megszokott jelentésével használunk ilyen kontextusban⁴. Erre a legjobb példa a *pozitív* és *negatív* szavak jelentései, amelyek az eredeti szótárban 1-es és -1-es pontszámmal szerepeltek, ebben a sorrendben. Ugyanakkor a koronavírus specifikus korpuszban ez a két pontszám megfordul, a pozitív tesztek egyértelműen negatív szentimentet hordoznak, a negatív tesztek és esetek pedig pozitívat. Hasonlóan kicseréltük a növekedést és csökkenést jellemző szavak szentimentjét, amely eredeti értelmükben pozitív és negatív tartalommal rendelkeznek. Ez azonban a jelen kontextusban megfordul, mivel a két szót leggyakrabban a fertőzöttség és a halálozási számokkal kapcsolatban használják a cikkekben. A lexikon átírásához az eredeti szótárban található 5702 szóból a korpuszban előforduló leggyakoribb 2000-et ellenőriztük le. A szövegekben előforduló leggyakoribb szavak, amelyeknek eredeti és az új szótár szerinti pontszámai eltérnek, a függelék 4. táblázatában találhatóak. A teljes szótár az alábbi linken érhető el:

https://github.com/MarcellGranat/CoronaSentiment/blob/main/bing_dictionary_for_coronavirus.xlsx.

A függelék 12. ábráján bemutatjuk a pozitív és negatív szavak közül a leggyakoribbakat, valamint a korábbi, 3. ábrán látható koronavírus hullámokkal összefüggésben érdemes megfigyelni a szentimenttel rendelkező szavak alakulását. 2020 áprilisában volt a legmagasabb ennek a száma, amely a cikkek számának tetőzésével esik egybe, tehát az első hullám alatt több cikk született a járványról, mint bármikor máskor a későbbi időszakban. Ezzel párhuzamosan ugyanakkor a szentimenttel rendelkező szavak aránya folyamatosan csökken a cikkekben, így tehát elmondhatjuk, hogy egyre kevésbé polarizáltak az írók cikkeit.

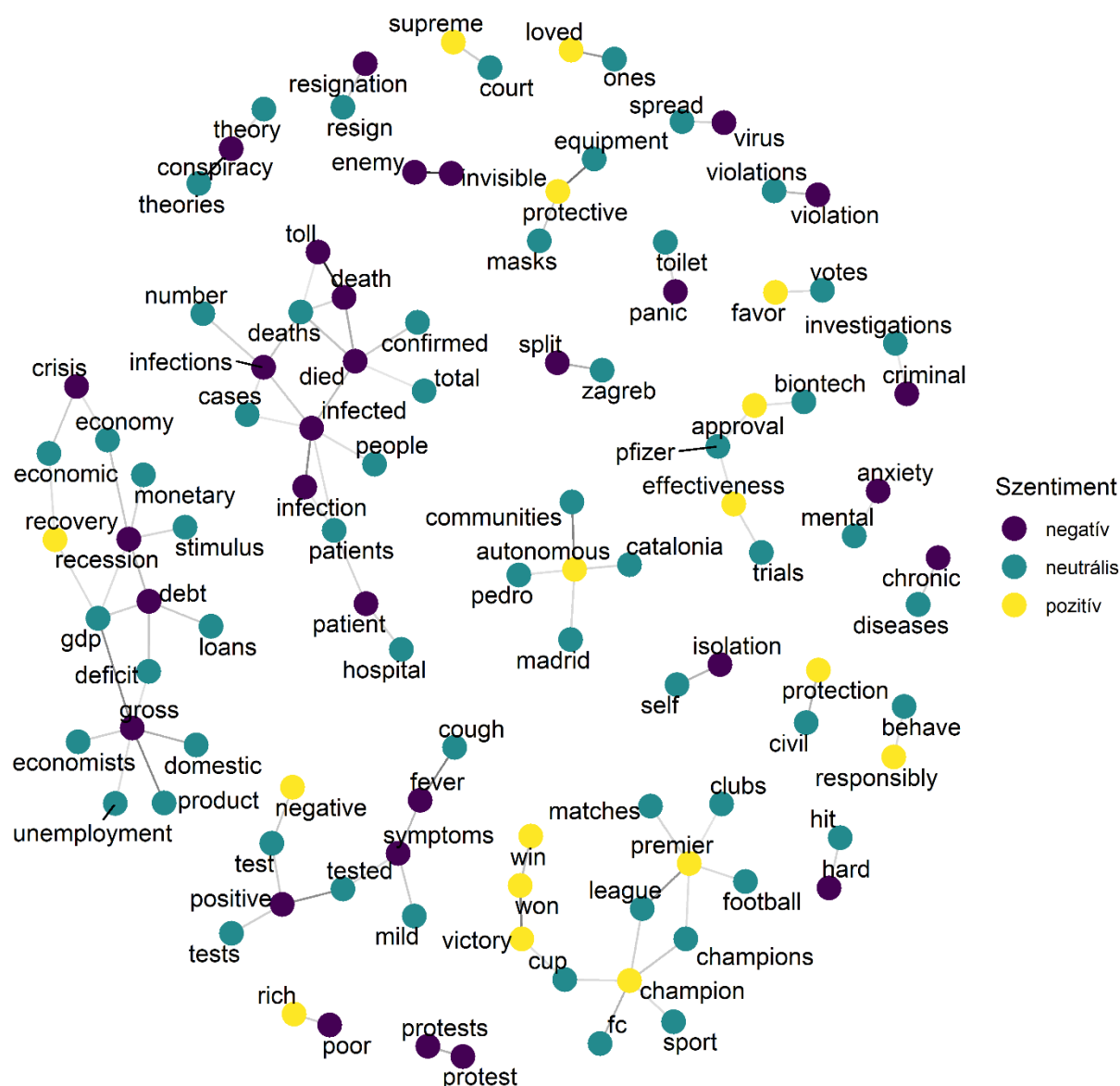
Ezek után megvizsgáltuk a különböző szavak korrelációját a korpuszon belül. Két szó (X és Y) közötti korrelációt Silge és Robinson (2017) könyve alapján az alábbi módszer szerint számolhatjuk:

	Tartalmazza Y szót	Nem tartalmazza Y szót	Összesen
Tartalmazza X szót	n_{11}	n_{10}	$n_{1.}$
Nem tartalmazza X szót	n_{01}	n_{00}	$n_{0.}$
Összesen	$n_{.1}$	$n_{.0}$	

$$\phi = \frac{n_{11}n_{00} - n_{01}n_{10}}{\sqrt{n_{1.}n_{0.}n_{.1}n_{.0}}},$$

ahol n az összes cikk mennyiségét jelenti, országonként ömlesztve.

⁴ Ugyanakkor fontos kiemelni, hogy új szavakat nem adtunk hozzá a szótárhoz, valamint mindösszesen két szót vettünk ki a szótárból, az aktuálpolitikai vonatkozása miatt. Ezek a *trump* szó (eredeti értelemben: *legyőzni*, a szótárban pozitív jelentésű); valamint a *vice* szó (eredeti értelemben: *bűn*, a szótárban negatív jelentésű).



4. ábra. Szentimenttel bíró szavakkal korreláló szavak hálójája.

A korpusz korábban leírt méretéből ugyanakkor adódik, hogy ha minden tokent belevennénk a korrelációs mátrixba, az jelentős mennyiségű adatot jelentene, így 100 véletlenszerűen kiválasztott cikket vizsgáltunk csak meg országonként. A 4. ábrán a vizsgált szavak korrelációs hálójája látható. Ez megmutatja, hogy a szentimenttel rendelkező szavak és a neutrális szavak közül melyek azok, amelyeknek a legnagyobb volt egymással a korrelációja. Láthatjuk, például, hogy az egyik csoport leginkább a vírus terjedésével és adataival kapcsolatban alakult ki. Eerre utaló szavak például az *infected*, *died*, *death*, amelyekhez a *patients* és *people* szavak kapcsolódnak. Egy másik csoportot a gazdasági adatokkal

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?

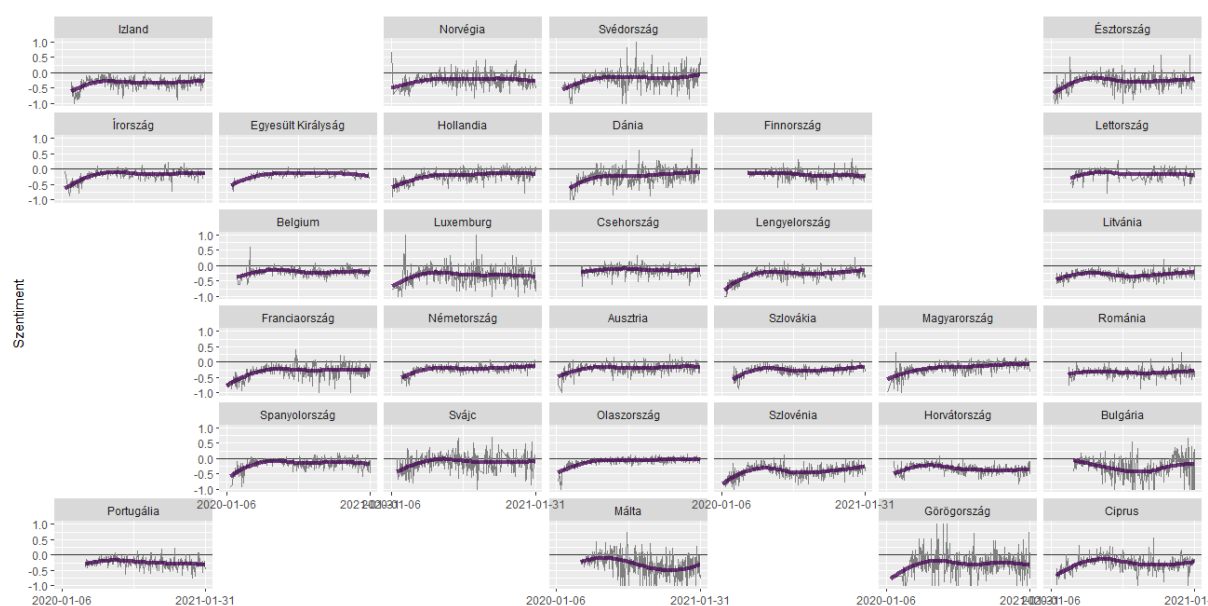
kapcsolatban fedezhetünk fel, leginkább negatív szavakból. Ilyenek például a *recession* és *debt* szavak, amelyek az *economy*, *monetary* szavakhoz kapcsolódnak. Ugyanakkor érdemes megjegyezni, hogy a pozitív töltetű *recovery* is a csoport tagját képezi.

Ezen kívül a csoportokból még kitűnnek a sporttal kapcsolatos szókapcsolatok. Ezek ugyan nem a koronavírushoz szorosan kapcsolódó bigramok, ugyanakkor az érzelmi töltetük az általunk használt Bing szótárban pozitív. Ilyen például a *win*, *victory*, *champion*. Ezeken kívül több olyan kapcsolatot is találhatunk, amely az azonos szótövek miatt jöhetett létre, ugyanakkor a korreláló párok közül fontos kiemelni a mentális egészséggel foglalkozókat: *mental*, *anxiety* és *loved*, *ones* valamint a védekezést bemutatókat: *behave*, *responsibly* és *invisible*, *enemy*. Érdekes párosokat alkot még a *toilet* és *panic*, amely a vírus eleji pánikvásárlásra utalhat, valamint a *rich*, *poor* ellentétpár összefüggése.

A már teljesen angol nyelvű korpusz szentimentelemzésekor a cikkeket pontszámokkal láttuk el. Ezeket a pontszámokat az alábbi módszerrel kapjuk meg:

1. A cikkeket tokenekre (a mi esetünkben szavakra) bontjuk.
2. Minden szó mellé hozzárendeljük a szótárban található pontszámát. Amelyik szóhoz nem tartozik pontszám, amellé nem rendelünk adatot.
3. Azon szavak pontszámait, amelyek szentimenttel rendelkeznek, kiátlagoljuk és így kapjuk meg a cikk szentiment pontszámát.

Annak érdekében, hogy az országokra vetített szentiment idősorokat vizsgálni tudjuk, ezeket az így megkapott pontszámokat országonként csoportosítva, naponta és havonta is a cikkekben található szentimentes szavak számával súlyozva átlagoljuk.



5. ábra. A szentiment alakulása országokként

Az 5. ábra földrajzi elhelyezkedésük szerinti elrendezésben ábrázolja az európai országok fentebb leírt módszertan szerint kiszámított, a panelelemzéshez is használandó szentimentidősorait. Ugyanitt látható, hogy a legtöbb országban a januári és februári hónapokban a cikkekben található érzelem negatív volt, majd ez több országban az idő elteltével növekedni kezdett. A cikkekben található érzelem növekedése jelenthetné a koronavírus helyzet javulását, azonban a korábbi grafikonokon láthattuk, hogy Európában szeptemberben indult el a második hullám, amely így ellentmondásra vezet. Az első hipotézisünk ezzel kapcsolatban az, hogy az első hullám reakciója után a második hullámra már a sajtó nem reagált ugyanolyan mértékben, így a cikkekben közvetített üzenetek sem tükrözték megfelelően az aktuális járványhelyzetet.

Ezt az eredményt két irányból is magyarázhatjuk: egyrészt az első hullámot túlreagálta a sajtó és a második hullámot már higgadtabban mutatta be, így az állampolgárok is neutrális szövegkörnyezetből értesülhettek a fontosabb hírekről. Másrészt viszont előfordulhat, hogy a nagyobb negatív szentiment visszatartó erővel rendelkezhetett volna, így az első hullámban megfelelő mennyiségű negatív szentimentet közvetítettek a médiumok és augusztus végétől kezdve az emberek már hozzászótkak a vírus jelenségéhez, így nem írtak olyan sok cikket a témáról és amelyek íródtak, azok is kevésbé negatív töltettel.

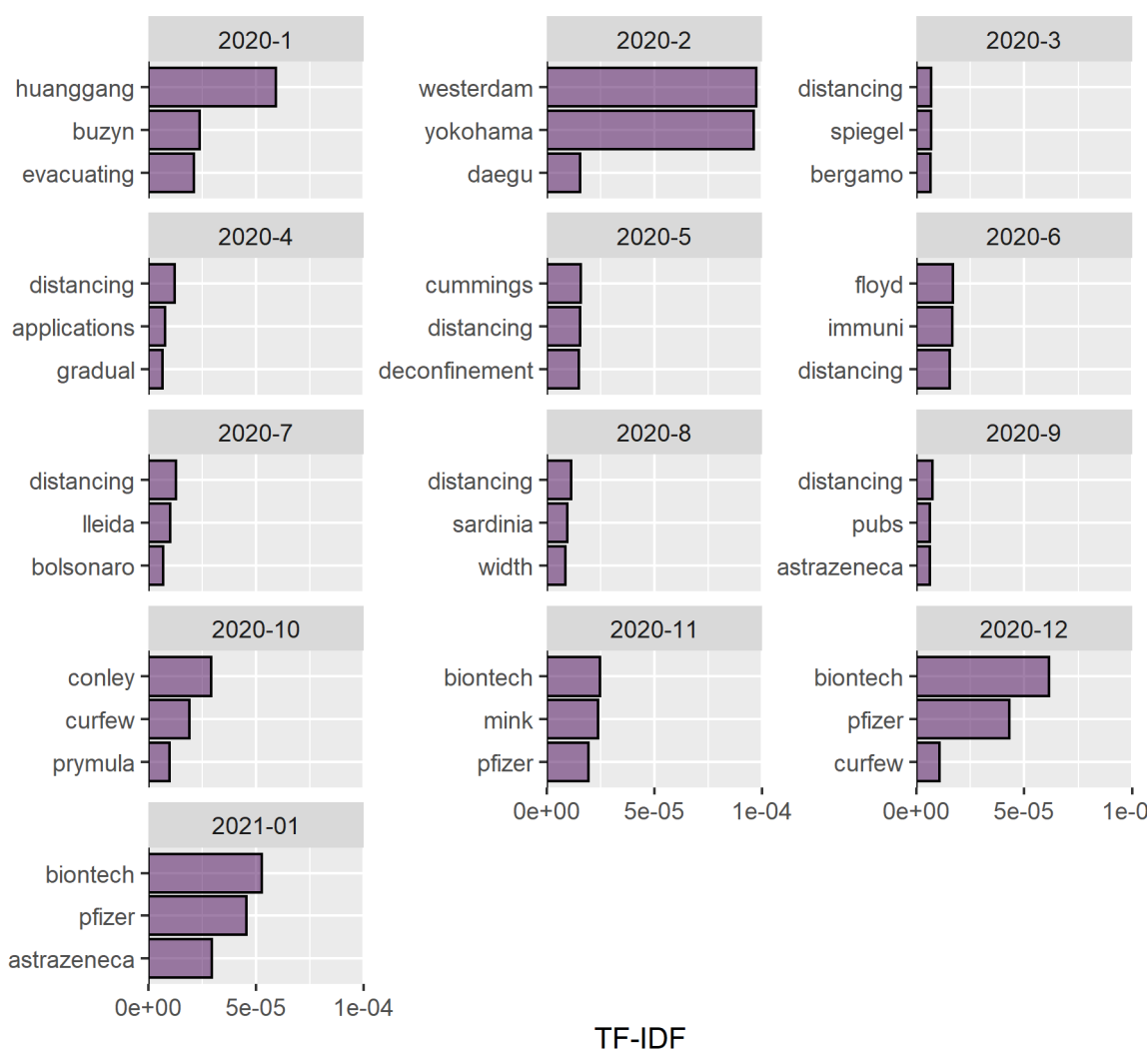
4.2 TF-IDF elemzés

A TF-IDF mutató a szakirodalmi ismertetőben leírtak szerint, azt írja le, hogy egy adott kifejezés mennyire specifikus egy vagy több adott dokumentumra nézve. Pontos számítási módja tehát az alábbi: egy adott szó TF mutatóját annak egy adott dokumentumrészletben vett előfordulási gyakorisága adja. Az IDF értékét pedig úgy kapjuk, hogyha az összes dokumentumrészlet számát elosztjuk azoknak a dokumentumrészletek számával, melyekben szerepel az adott kifejezés, majd vesszük ennek a logaritmusát. (Holecz és Szűcs, n.a.) Hogyha a két mérőszámot összeszorozzuk, megkapjuk, hogy egy kifejezés egy adott dokumentumrészletben mennyire specifikus, azaz mennyire jellemzi azt az adott részletet.

Különböző megfontolások alapján oszthatjuk részekre a teljes korpuszunkat, amelyből a két fő dimenzió az országok menti, valamint a havi feldarabolás lenne. Míg az előbbi felosztás szerint az országok cikkeire jellemző szavakat kaphatjuk meg, úgy az utóbbi megoldás a havonta aktuális tokenekre, valamint felkapott témákra hívná fel a figyelmet. Mindkettőt kipróbálva az előbbi megoldás a Google Translate fordításai miatt vagy nem angol szavakat, vagy pedig az országok nagyobb városainak neveit és politikusait mutatta ki, így inkább a havi felosztást mutatjuk be.

A rendelkezésre álló adathalmaz mérete miatt véletlen mintavételt alkalmaztunk a korpuszon, amelynek keretében minden országból 1000 cikket választottunk ki a letöltött periódusból. Ezzel nem csak az országok közötti egyenlőtlenségeket küszöbölhettük ki a cikkek és ennek megfelelően a tokenek számában, hanem a memória használata sem haladta meg a számítógép elérhető kapacitását. Az elemzésben található idegen szavak miatt azon szavakat leszűrtük, amelyek kevesebb, mint három ország cikkeiben fordulnak csak elő. Ezen kívül egy adott hiperparaméter, jelen esetben 0,00005 feletti előfordulási gyakoriságú szavakat vettük csak figyelembe, amelyek az adott hónapban megjelenő összes szónak minimum 0,0005%-át teszik ki a korpuszban.

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?



6. ábra. A cikkekben havonta leginkább jellemző szavak, TF-IDF értékek alapján.

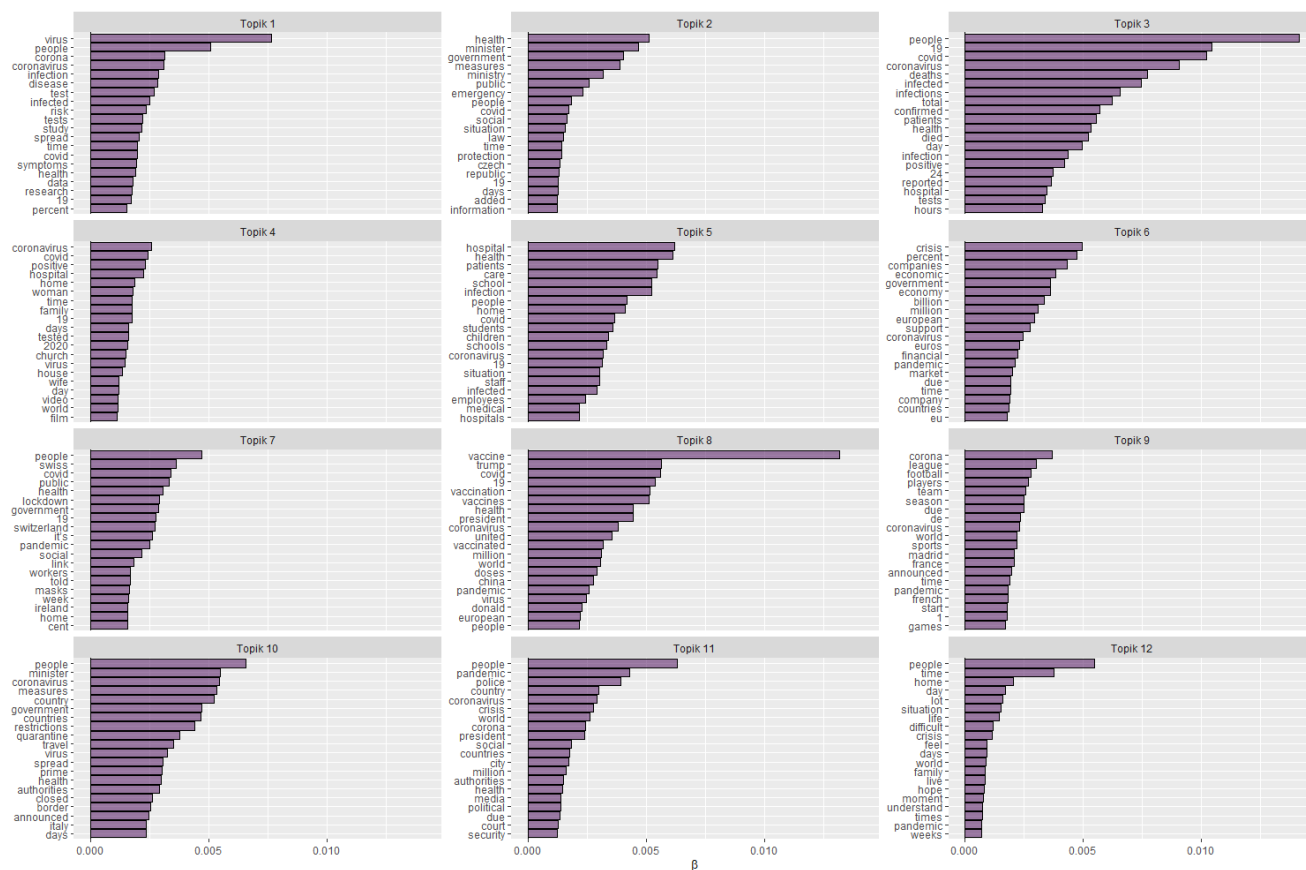
A 6. ábrán láthatóak azon kifejezések, amelyek az egyes hónapok legfontosabb szavai voltak, tehát a TF-IDF mutatóik havonta a legmagasabbak. Ebből az ábrából kivehetőek a havi “slágertémák”, vagy botrányok, amelyek körül a teljes európai sajtó forgott. Ilyen például 2020 januárjában és februárjában a Japán-tengeren ragadt óceánjáró hajók esete, 2020 tavaszán a *social distancing*, tehát a közösségi távolságtartás kifejezés elterjedése, valamint a novemberi tömeges nyérc levágás Dániában. Ugyanakkor fontos kiemelni azt is, hogy az utolsó hónapokat dominálta az oltások megjelenése, így csak a gyártók neveit találhatjuk a szavak között. Konklúzióként tehát elmondhatjuk, hogy a havi botrányok mellett, a járvány elleni védekezés különböző módszerei (távolságtartás, vakcinák) folyamatosan helyet kaptak a legfontosabb szavak között.

4.3 Témamodellezés

A topik modelling elemzés célja, hogy a korpuszban található cikkeket kategorizálja, valamint megállapítsa, hogy melyik cikk milyen főbb téma köré csoportosul. Ennek a legfőbb matematikai alapja a látens Dirichlet allokáció (Blei et al., 2003), amely a korábbi elemzésekhez hasonlóan a szózsák-modellre épül. Ezen számítás kiinduló gondolata, hogy a teljes szöveg különböző háttérben meghúzódó szemantikai egységekből, azaz témákból épül fel, amelyek különböző arányban jelennek meg akár egy-egy cikkben is. Továbbá, minden topik szavak összességéből áll, amelyek különböző mértékben definiálják az adott témát. Fontos kiemelni, hogy egy szó akár több témában is megjelenhet. Ezen kiinduló axiómák szerint építhetjük fel a valószínűségi eloszlásra épülő modellt, amely legfőképpen a korpusz kifejezés-dokumentum mátrixát használja a becslésre. Ezen kívül azonban több hiperparamétert is tartalmaz, melyek közül a legfontosabb a topikok száma.

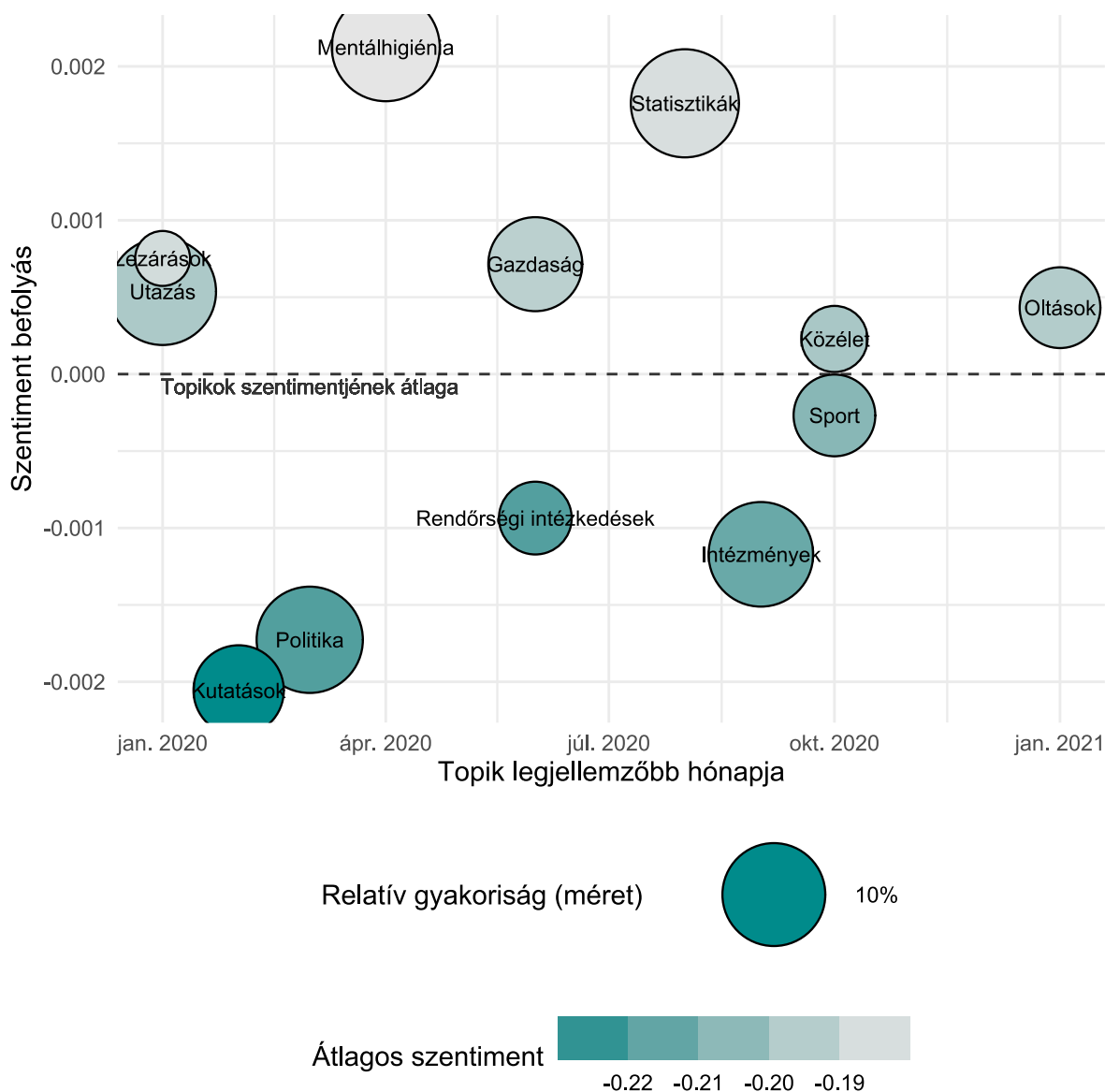
Ennek meghatározása az egyik legfontosabb lépés a becslésben, ehhez pedig a korábban bemutatott szakirodalom általában 10 és 16 közötti értéket ad meg. Ahhoz, hogy matematikai módszerrel határozzuk meg az optimális topik számot, egy rendkívül memóriaintenzív számítást kell elvégeznünk, amely a topikok koherenciáját méri log-likelihood módszerrel. Ezt a módszert 16 topik számig tudtuk lefuttatni a már korábban is használt, országonkénti 1000 cikkre véletlen mintánkon, azonban a függelék 14. ábráján látható módon, a mérőszám nem konvergált optimumhoz. Amennyiben viszont 100 cikket vettünk minden országból, az optimum 70 topiknál kalibrálódott. Ezen tesztek után úgy döntöttünk, hogy olyan topik számot veszünk alapul, amely még jól ábrázolható, illetve a korábbi, koronavírussal kapcsolatos szakirodalommal összhangban van, így végül 12 topikot modelleztünk.

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?



7. ábra. Leggyakoribb szavak topikonként

A 7. ábrán, a témák azonosításakor láhattuk, hogy a tizenkét témakörből néhány egyértelműen felismerhető: a 2-es téma az országok vezetőit, valamint azok intézkedéseit emeli ki, a 3-as téma a statisztikai leírásokra fókuszál, az 5-ös a szociális intézmények (kórházak, iskolák) helyzetét mutatja be, a 6-os topik a gazdasági híreket tömöríti, míg a 8-as témakör az oltásokról, továbbá ezek külpolitikai jelentőségéről szól. Végül a 9-es topik a sportvilágról számol be, a 10-es pedig az utazási korlátozásokról. A személyesebb, életmóddal kapcsolatos híreknek számító, otthonlétről és családi életéről szóló híreket a 4-es, illetve a 12-es témákban találhatjuk.



8. ábra. Topikok jellemzése szentiment, dátum és gyakoriság szerint

Az előző állításunkat kiegészíti a következő, 8. ábrán a topikok átlagos szentimentjének és a topikok szentimentre való befolyásának⁵ alakulása. Ezen grafikon alapján egyrészt láthatjuk, hogy a topikok közül az általános kutatásokat tartalmazó, valamint a rendőrségi intézkedésekkel foglalkozó témáknak volt a legalacsonyabb átlagos szentimentje, azonban hogyha a témák szentimentre való befolyását nézzük meg, az y tengelyen láthatjuk, hogy súlyozottan melyeknek volt a legnagyobb negatív hozzájárulása összességében. Ebből pedig szintén a kutatásokkal, valamint a kórházakkal és iskolákkal, továbbá a politikával foglalkozó

⁵ A szentimentre való befolyást az alábbi szerint számoltuk: a topik érzelmi pontszámának a korpusz átlagos szentimentjétől való eltérését súlyoztuk a téma relatív gyakoriságával

hírek voltak a negatív szempontból legfontosabbak. Olyan témák, amelyeknek az átlagos érzelmi töltete kevésbé volt negatív, de mégis a gyakoriságuk miatt nagyban hozzájárultak a teljes szentimenthez, a rendőrségi intézkedésekkel és a sporttal foglalkozó témát emelhetjük ki. Az x tengelyen található hónapok azt jelentik, hogy az adott topik melyik hónapban volt a legjellemzőbb. A mentálhigiénéhez kapcsolódó cikkek jelentős pozitív töltetet jelenítettek meg, amelyek a lezárások első időszakában, áprilisban voltak a leggyakoribbak. A sport szeptemberben, a legnagyobb sportesemények újraindításakor, míg az utazással foglalkozó már 2020 januárjában felkapott téma volt. A gazdasági hírek júniusban voltak a leggyakoribbak, míg a vakcinával kapcsolatosak 2021 januárjában tetőztek. Az egyes témák gyakoriságának pontos időbeli eloszlása a függelék 15. ábráján tekinthető meg.

5 Ökonometriai modellek

5.1 Leíró statisztikák

A következőkben a hírekben megjelenő szentimentet befolyásoló tényezők modellezését tárgyaljuk. Ehhez a már korábban bemutatott havi szentimenteket modellezzük ökonometriai eszközökkel.

Ahogy az az 5. ábrán jól kivehető, a hírekben megjelenő szentiment nagyon alacsony volt 2020 januárjában, majd fokozatosan emelkedett. Ennek és további könnyen interpretálható mintázatoknak azonosítására regressziós fa modellt készítünk, amelyben kimeneti változó a havi szentiment, prediktorként pedig az új esetek és halálozások számát, mind abszolút mind ezer főre eső értékben szerepeltettük kiegészítve a munkanélküliséggel, gazdasági érzékelési indikátorral és az idő változójával. A regressziós fa lényege, hogy a prediktorokat felhasználva úgy kategorizálja a megfigyeléseket, hogy a kimeneti változó tekintetében minél homogénebb csoportokat hozzon létre. Bár a modell figyelmen kívül hagyja az adattáblánk longitudinális jellegét, így kutatási kérdésünk megválaszolására nem elégséges ⁶, számos könnyen értelmezhető leíró statisztikával szolgál a modell. Az eredményt a 9. ábra ismerteti.

A regressziós fa alapján a legelső csoportosítási lehetőség, hogy januári megfigyelésről van-e szó. ⁷ Ez a konklúzió olvasható le a 3. ábráról is, azaz januárban és februárban

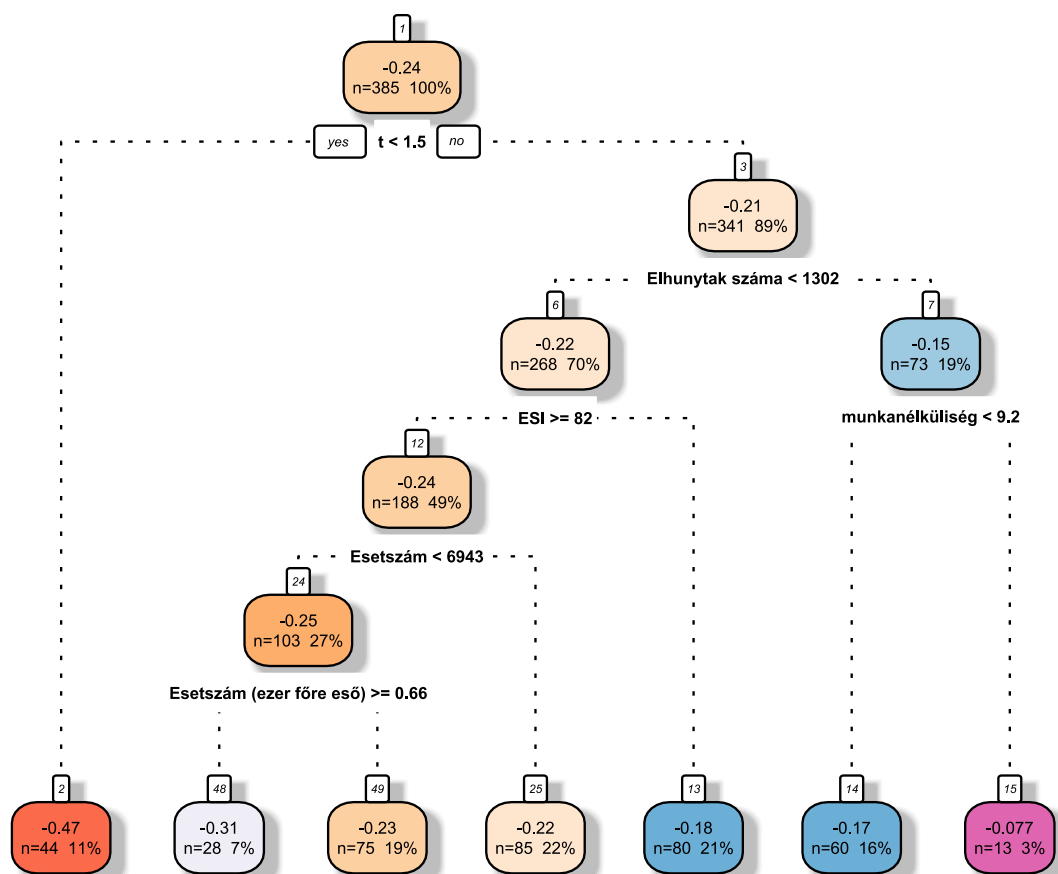
⁶ Ebből kifolyólag a modell komplexitási paraméterének optimalizálásától jelenleg eltekintünk és azt azon megfontolásból állítjuk 0,012-re, hogy egy jól ábrázolható fát kapjunk.

⁷ A modellben a $t = 0$, amennyiben 2020 januári a megfigyelés, $t = 1$, ha 2020 februári, stb.

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?

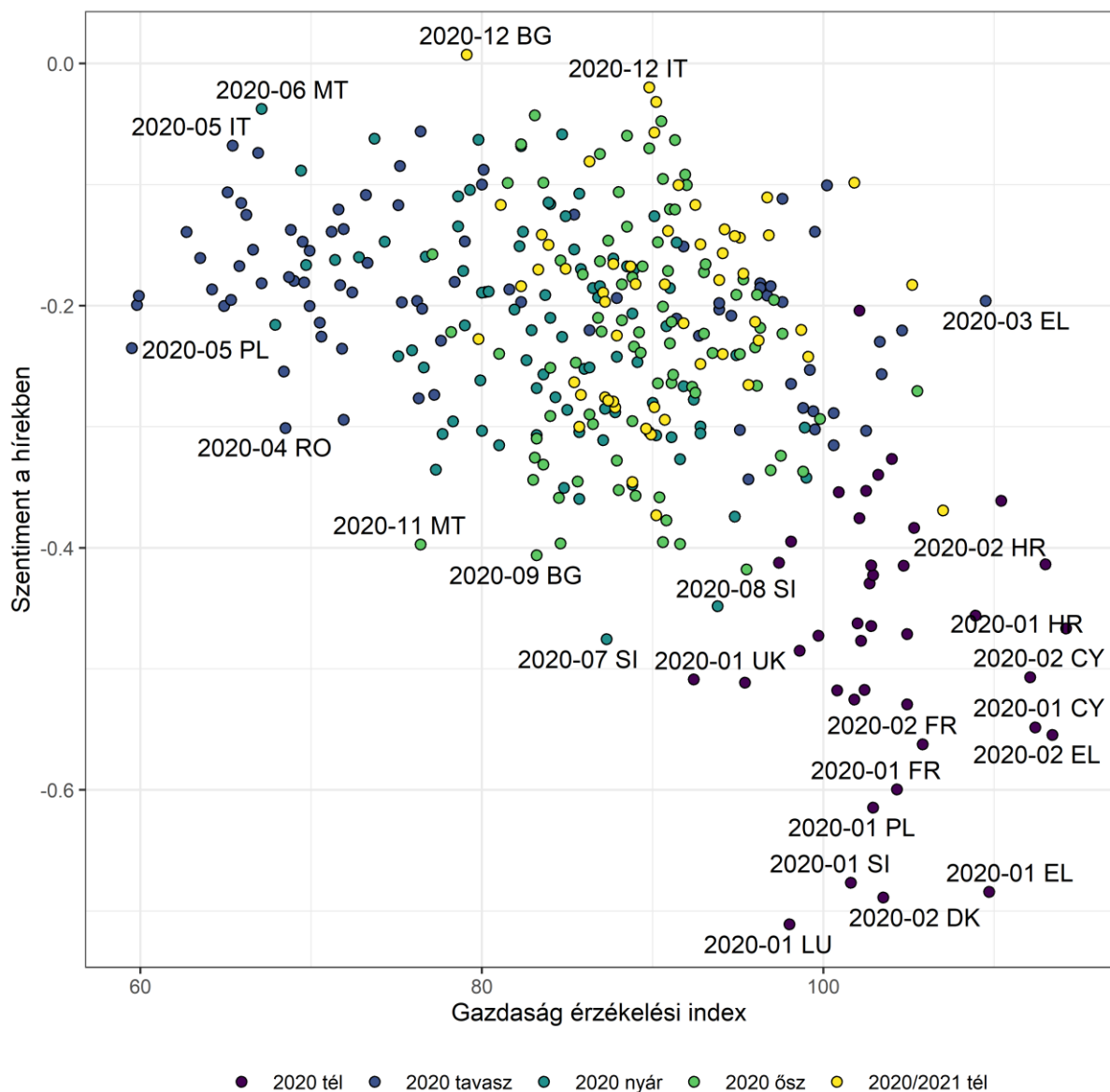
számottevően negatívabb volt a cikkekben megjelenő érzelmi töltet, míg ehhez nem párosult akkor még a koronavírus esetek növekedése.

A fa további csoportosításai között megjelenik, hogy a havi szentiment átlagosan magasabb azon megfigyelések között, amikor az elhunytak száma magasabb, továbbá a munkanélküliségi ráta növekedése és gazdaságérzeléslési index csökkenése is az érzelmi töltet emelkedésével jár együtt. Ezekből a statisztikákból az látszik, hogy az európai média kommunikációja főként a krízishelyzetek idején közöl le pozitív töltettel rendelkező híreket. Ezt alátámasztja a korábbi fejezetben leírt témamodelllezési eljárással nyert eredményünk: az első hónapokban már megjelentek a koronavírusról szóló kutatások, amelyek erős negatív töltettel rendelkeztek, ezzel szemben a 2021 telén lényegesen magasabb a halálozások száma, miközben a hírfolyamokban már növekszik az vakcináról szóló cikkek aránya, ami felfelé mozdtítja el a szentimentet.



9. ábra. Regressziós fa

A regressziós fa modellje azonban figyelmen kívül hagyja az adatok longitudinális jellegét. Érdekes felfigyelni az egyes indikátorok közös időbeli trendjére az országok között. A 10. ábra a gazdaság érzékelési index és a médiában megjelenő szentiment közötti kapcsolatot mutatja be évszakok szerinti bontásban.



10. ábra. A hírekben megjelenő szentiment és gazdaság érzékelési index közötti kapcsolat

A 10. ábrából kirajzolódik, hogy a hírekben megjelenő szentimentek és a gazdaság érzékelési index között is közös trend azonosítható, és a két változó erős szignifikáns

korrelációban áll egymással ($r^8 = -0,51$). Az ellentétes irányú együttmozgás azonban időszakok között eltérő erősséggel jelenik meg: legjelentősebben 2020 nyarán ($r = -0,48$), azonban ugyanez a jelenség 2020 telén még egyáltalán nem volt erős ($r = -0,063$).

Hasonló módon a járványhelyzet és a médiában megjelenő szentiment közötti összefüggés is eltérő módon alakult az egyes évszakokban. Nyáron negatív irányú együttmozgás jellemzi az esetszámot és a szentiment közötti kapcsolatot ($r = -0,15$), ezzel szemben 2020 telén még éppen azokra a megfigyelésekre volt jellemző a pozitívabb hangulat, amelyekben az esetszám magasabb volt ($r = 0,38$).

Ennek kezelésére érdemes olyan ökonometria modelt készíteni, amely kontrollál az egyes évszakokra, és időszakon belül mutatja ki a járványügyi és gazdasági folyamatok hírekre gyakorolt hatását.

5.1.1 Bizonytalanságkerülés

Következtetésként levonhatjuk, hogy az európai médiában megjelenő cikkekből észrevehető tendencia, miszerint 2020 telén még nagyon alacsony volt az általános szentiment, rámutat a koronavírus válságra adott reakció egyik fontos pontjára. Mivel a válság kezdetekor még keveset tudtunk annak pontos végkimeneteléről, így az ezzel kapcsolatos cikkek is annyira negatív hangulatban íródtak, hogy az szignifikánsan kimutatható lett. Ezzel szemben 2020 tavaszától a gazdasági érzékelésben és a médiatartalmak érzelmi töltetében egy olyan közös trend jelent meg, amely az írásos reakciók polarizáltságának mérséklődésére utal.

Ennek magyarázatára szolgálhat a viselkedési közgazdaságtanban is ismert bizonytalanságkerülés. Ellsberg (1961) szerint az emberek bizonytalanságkerülése azt jelenti, hogy nem szeretjük azokat a szituációkat, melyek során nem tudjuk megbecsülni a kimenetelek valószínűségét.

Jelen esetben ez a szituáció 2020 januárjában volt elmondható az európai sajtóra és állampolgárookra is, mivel senki nem tudta megmondani biztosan, hogy a világjárványnak mekkora hatása lesz és pontosan milyen kockázatot visel majd a társadalom. Volt olyan sajtóhír is, amely az 1918-as spanyol náthánál is nagyobb veszteségeket prognosztizált⁹, de ennél óvatosabb becslések is megjelentek. Ugyanakkor, amikor 2020 márciusában a veszteségek

⁸ Lineáris korrelációs együttható.

⁹ A spanyol nátha nagyjából 100 millió halálos áldozatot követelt világszerte 1918 és 1920 között. (Devlin, 2020) A kézirat lezárásakor az új típusú koronavírusban nagyjából 3 millióan hunytak el globálisan.

elkezdődtek már sejteni lehetett a potenciális amplitúdóját a járványnak, még akkor is, hogyha az első hullám elején tartottunk. Ezzel pedig az emberek a járvány kockázatáról képet kaphattak, így a bizonytalanság már kevésbé jellemezte a vírus maximális mértékéről alkotott véleményeket.

Hipotézisünk szerint a 2020 telén íródott cikkekben ez, a bizonytalanságtól való félelem magas szinten jelen volt, így ez is okozhatta a jelentősen alacsonyabb szentimentet. Ugyanakkor, amint a bizonytalanság már kevésbé jellemezte a járványról alkotott becsléseket, akkor az érzelmi töltet is moderálódott.

5.2 Longitudinális elemzés

A következőkben longitudinális (panel) modellezést végzünk, amely során a hírekben megjelenő szentimentet befolyásoló tényezőket vizsgáljuk, figyelembe véve az előzőekben feltártak alapján, hogy a változók jelentős közös időbeni trenddel rendelkeznek. Az évszakokra való kontrollálás segítségével arra a kérdésre keressük a választ, hogy adott időszakon belül a magasabb esetszám vagy eltérő gazdasági helyzet statisztikailag szignifikáns hatással bír-e a hírekben megjelenő szentimentre. Az időszakokra való kontrollálás azonban szükségtelenül csökkentheti a paraméter becslések pontosságát, amennyiben redundánsnak bizonyulnak. Ennek megfelelően leközzöljük azokat a modelleket is, amelyek nem kontrollálnak az időszakra.

A felhasznált prediktorok köre azon megfontolásból esett a tanulmányunkban leközzöltre, mert az esetszám és a halálozás, illetve a gazdaság érzékelési index között erős korreláció van. A változók közötti multikollinearitás miatt érdemes az új információt már nem tartalmazókat kihagyni. Előzetes kétváltozós elemzésünk során kiderült, hogy a szentiment lényegesen jobban együtt mozog az ezer főre eső halálozással ($r = 0,23$), mint az ezer főre vetített esetszámmal ($r = 0,14$), illetve a gazdaságérzékelési indexszel ($-0,51$), mint a munkanélküliséggel ($0,05$).

A panelökonometriai elemzés első lépése, hogy meghatározzuk az optimális modell típust. A lehetséges panelmodellek közül az alábbiakat vetjük össze: ömlesztett modell, fixhatás-modell, véletlenhatás-modell. A három modell közötti választás statisztikai tesztek alapján történik. Elsőként a Chow-tesztet végezzük el, mely a regressziós modellünkben az országok egyedi konstansain vizsgálja F-próbával azt a nullhipotézist, hogy az összes ország állandó paraméterének modellből való kihagyásával nem veszítünk információt. Amennyiben a hipotézist nem tudjuk a bevett szignifikancia szinteken elutasítani úgy az ömlesztett modellt alkalmazzuk. Ellenkező esetben a Hausman-tesztet, melynek nullhipotézise, hogy a

véletlenhatás-modell felhasználása preferált a fix-hatás modellhez képest. Mindkét teszt nullhipotézisének elutasítása esetén a fixhatás-modell becslésére kerül sor. (Wooldridge, 2016)

Első regressziókban nem kontrollálunk az évszakokra és a halálozást, illetve a gazdasági szentimentet szerepeltetjük prediktorként. Az elvégzett Chow és Hausman tesztek alapján a fixhatás modell használata megfelelő és a korrigált determinációs együtthatónk (R^2) 28%. A becsült paramétereket a 2. táblázat ismerteti.

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
Halálozás	0,1992	0,0422	4,7184	0,00%
ESI	-0,0056	0,0005	-11,2914	0,00%

2. Táblázat. Regressziós eredmények időszakra való kontrollálás nélkül

A 2. táblázatban látható eredmények alapján kiolvasható, hogy mind a halálozás, mind a gazdasági érzékelési index szignifikánsan együtt mozog a szentimenttel. A statisztikai szignifikancia mellett az ezer főre eső halálozás becsült hatásának mértéke is jelentősnek tűnik, ugyanis annak 1 fővel való növekedése az ESI-re kontrollálva átlagosan 0,2 egység szentiment csökkenéssel jár együtt. Az ESI hatása is hasonló mértékben jelentős.¹⁰

A leírt hatások azonban eltűnnek amennyiben az évszakokra is kontrollálunk a modellben dummy változókkal. Az évszakokhoz tartozó paraméterek úgy értelmezhetőek, hogy a szentimentértékek várhatóan (minden más változatlansága mellett) annival magasabbak, mint 2020 telén. Továbbá azon elméleti megfontolásból, hogy az idő elteltével egyre kevésbé reagál a hírekben megjelenő szentiment a valós halálozási esetszámra a modellbe bevontuk a halálozási arányszám idővel (azaz a 2019 decemberétől eltelt hónapok számával) vett interakcióját is. Ezzel a lépéssel a magyarázott variancia jelentősen megnő, a modellnek így már 64% az R^2 -e. A becsült paramétereket a 3. táblázat tartalmazza.

¹⁰ Bár a becsült együtthatók eltérőek, azoknak egységnyi változása egészen más skálán van mérve. Míg az ezer főre vetített esetszám esetén a néhány főnyi emelkedés is nagynak számít, addig az ESI esetében a 40 egységnyi csökkenés is megfigyelhető.

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
Halálozás	-0,2136	0,1132	-1,8872	6,01%
ESI	-0,0018	0,0005	-3,6821	0,03%
tavasz	0,2569	0,0161	15,9565	0,00%
nyár	0,2228	0,0152	14,6746	0,00%
ősz	0,2119	0,0140	15,1432	0,00%
tél	0,2361	0,0181	13,0721	0,00%
Halálozás × eltelt hónap	0,0297	0,0117	2,5328	1,18%

3. Táblázat. Becsült regresszió paraméterek kontrollálva az időszakokra

A modell becsült paramétereiből látható, hogy mind az évszakok kétértékű változói, mind a gazdasági szentiment szignifikáns magyarázóváltozónak bizonyult. A halálozás hatása is szignifikáns 10%-os szignifikancia szinten (az interakció okozta multikollinearitás miatt szignifikánsnak tartjuk), a becsült hatás mértékét tekintve pedig szintén jelentősnek mondható.

Az évszakok egyes becsült átlagos hatásai között nincsen különbség, amelyből azt a következtetést vonhatjuk le, hogy csupán 2020 telén látunk szignifikánsan alacsony szentiment értékeket. Ez a jelentős különbség nem indokolható sem az esetszámmal, sem a gazdaság helyzetéről vélekedéssel, így ezt a média túlreagálásának tekinthetjük.

A gazdaság érzékelési indexről továbbra is az mondható el, hogy negatívan mozog együtt a hírekben megjelenő szentimenttel. A gazdaság érzékelési index és a hírekben megjelenő szentiment közötti kapcsolat évszakonkénti elemzéséből az derül ki, hogy a legerősebb negatív korreláció 2020 tavaszán ($r = -0,38$) és nyarán ($r = -0,48$) mérhető, míg őszzel ez az ellentétes irányú együttlmozgás egyáltalán nem jellemző ($r = -0,09$). Tavasszal és nyáron a gazdaság érzékelési index a megelőző időszakok értékeinél kiugróan alacsonyabb volt, illetve éppen megnőtt a gazdaságról szóló hírek aránya a médiában, amelyek az átlagnál sokkal magasabb szentimenttel rendelkeztek. Mindez hozzájárul ahhoz, hogy 2020 tavaszán magas szentimentet és alacsony gazdaság érzékelési indexet találunk. Összeségében levonható konklúzióként, hogy a média pozitívabb hírekkel reagált az emberek gazdaságról alkotott negatív képére.

Időszakokra kontrollálva az esetszám és az interakció becsült hatásai alapján azt mondhatjuk, hogy bár a hírek hangulata negatívabbra fordul az ezer főre eső halálozás növekedésével, ez a hatás egyre csökken az idő előrehaladtával. Pontbecslés alapján 2020 ősztől eltűnik a halálozás hírekre gyakorolt hatása. Így tehát azt mondhatjuk, hogy a média korai túlreagálása után az európai második hullám elejétől a sajtóban megszokottá vált a koronavírus, és a cikkekben hordozott szentiment sem reagált már a növekvő halálozási számra.

Erről a viselkedésről eszünkbe juthat Aiszóposz fabulája a farkast kiáltó fiúról, aki akkor figyelmeztetett veszélyre és kért segítséget, amikor még nem érte valós fenyegetés, így harmadjára, amikor tényleg megtámadta az állat már nem segített neki senki. (Aiszóposz, 1957) A média 2020 januárjában olyan negatívnak állította be a koronavírus-járványt, hogy az őszi második hullám súlyosságára az emberek már gyengébben reagáltak. Nem csak a halálozási számokra voltak kevésbé érzékenyebbek, de a szabályozásokat sem annyira vették figyelembe. (Radó, 2021)

6 Konklúzió

Összességében megállapíthatjuk, hogy az európai sajtó koronavírus-járvánnyal kapcsolatos híreiben mind azok témáiban, mind pedig szóhasználatában és az átadott érzelmében jelentős időbeni változás volt tapasztalható. Az első hónapokban, tehát 2020 januárjában a későbbi hónapokhoz képest kirívóan negatív szentiment társult, amit nem indokolt a járvány akkori állapota vagy a gazdaság helyzete. A különböző témakörök előfordulásának és szentimentjének vizsgálatával bemutattuk, hogy a korai kutatásokat ismertető cikkeket, amelyek jelentősen csökkentették a médiában megjelenő hangulatot, tavaszra neutrálisabb a mentálhigiénéről szóló és az emberek fejében lévő gazdasági hanyatlásra válaszul születő gazdasági témájú hírek váltották fel.

A hírek szentimentjét a gazdaság mellett a járványhelyzet is nagymértékben befolyásolta. Kutatásunkban empirikusan támasztottuk alá, hogy az ezer főre eső halálozás növekedésére a hírek negatívabb hangulattal telnek meg, azonban ez a mechanizmus az idő előrehaladtával eltűnni látszik, melynek részben magyarázata, hogy Európa-szerte növekszik a vakcinák témakörében íródó cikkek aránya. Másrészt viszont a cikkekben felfedezhető az is, hogy a járvány kezd megszokottá válni a közbeszédben, így kevésbé jelent újdonságot és a médiában kevésbé polarizált szóhasználat övezi.

7 Irodalomjegyzék

- Aiszóposz. (1957). A tréfás pásztor. In: Hausrath, A. (szerk.) *Corpus Fabularum Aesopicarum*.
- Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1), 45-65.
- Al-Awadhi, A. M., Alsaifi, K., Al-Awadhi, A., & Alhammadi, S. (2020). Death and contagious infectious diseases: Impact of the COVID-19 virus on stock market returns. *Journal of behavioral and experimental finance*, 27, 100326.
- Älgå, A., Eriksson, O., & Nordberg, M. (2020). Analysis of Scientific Publications During the Early Phase of the COVID-19 Pandemic: Topic Modeling Study. *Journal of medical Internet research*, 22(11), e21559.
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, No. 2010, pp. 2200-2204).
- Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1), 56-75.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.
- Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4), e21978.
- Can, E. F., Ezen-Can, A., & Can, F. (2018). Multilingual sentiment analysis: An RNN-based framework for limited data. *arXiv preprint arXiv:1806.04511*.
- Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4), 757-771.
- De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, 26(4), 417-430.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391-407.
- Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th international conference on data engineering workshop* (pp. 507-512). IEEE.

Devlin, H. (2020). Four lessons the Spanish flu can teach us about coronavirus. <https://www.theguardian.com/world/2020/mar/03/four-lessons-the-spanish-flu-can-teach-us-about-coronavirus> (Letöltés dátuma: 2021. 04. 12.)

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5), 533-534.

Drávucz, F., Szabó, M. K., & Vincze, V. (2017). Szentiment-és emóciósztárak eredményességének mérése emóció-és szentimentkorpuszokon. Vincze, V.(szerk.) *XIII. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY*, 228-239.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics*, 643-669.

Gelper, S., & Croux, C. (2010). On the construction of the European economic sentiment indicator. *Oxford Bulletin of Economics and Statistics*, 72(1), 47-62.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162.

Holecz, M., Szűcs, Á. (n.a.). Politikai kötődés automata meghatározása az online sajtóban. *Távlatok a mai magyar alkalmazott nyelvészetben*, 233.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

Mansoor, M., Gurumurthy, K., & Prasad, V. R. (2020). Global Sentiment Analysis Of COVID-19 Tweets Over Time. *arXiv preprint arXiv:2010.14234*.

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Ntoulas, A., Cho, J., & Olston, C. (2004, May). What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web* (pp. 1-12).

Oehmke, T. B., Post, L. A., Moss, C. B., Issa, T. Z., Boctor, M. J., Welch, S. B., & Oehmke, J. F. (2021). Dynamic panel data modeling and surveillance of COVID-19 in metropolitan areas in the United States: Longitudinal trend analysis. *Journal of medical Internet research*, 23(2), e26081.

Radó, A. (2021). A halálozási számok az egekben, de a magyarok még az enyhe korlátozásokat sem veszik elég komolyan. <https://qubit.hu/2021/04/12/a-halalozasi-szamok-az-egekben-de-a->

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?

magyarok-meg-az-enyhe-korlatozasokat-sem-veszik-eleg-komolyan (Letöltés dátuma: 2021.04.12.)

Reuters Institute. (2021): Reuters Institute Digital News Report 2020. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf (Letöltés dátuma: 2021.03.23.)

Sarkodie, S. A., & Owusu, P. A. (2020). Investigating the cases of novel coronavirus disease (COVID-19) in China using dynamic statistical techniques. *Heliyon*, 6(4), e03747.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. " O'Reilly Media, Inc."

Tao, G., Miao, Y., & Ng, S. COVID-19 Topic Modeling and Visualization.

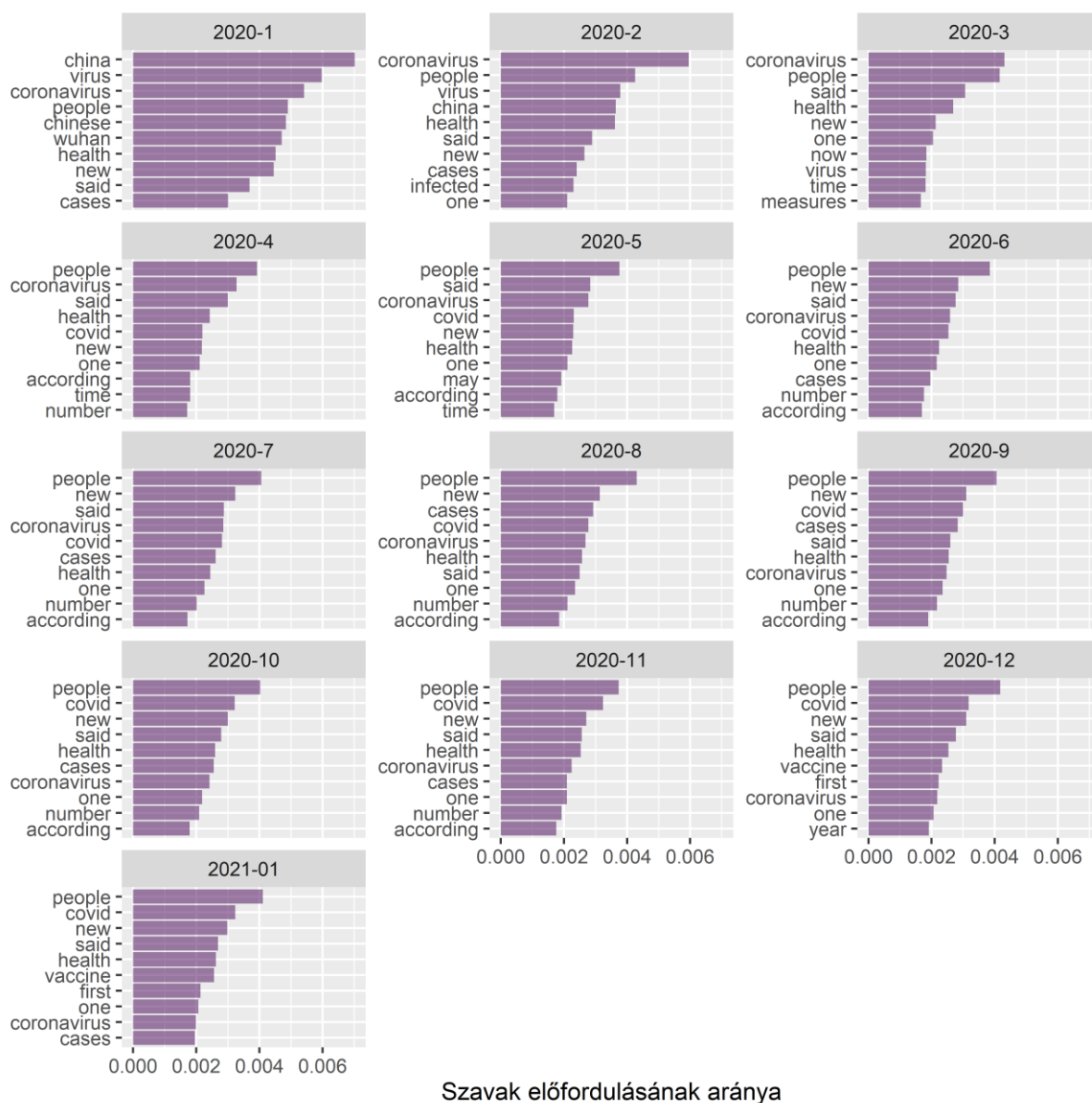
Topcu, M., & Gulal, O. S. (2020). The impact of COVID-19 on emerging stock markets. *Finance Research Letters*, 36, 101691.

Volkova, S., Wilson, T., & Yarowsky, D. (2013, October). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1815-1827).

WHO. (2020). Coronavirus disease (COVID-19). <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> (Letöltés dátuma: 2021.04.09.)

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education.

8 Függelék

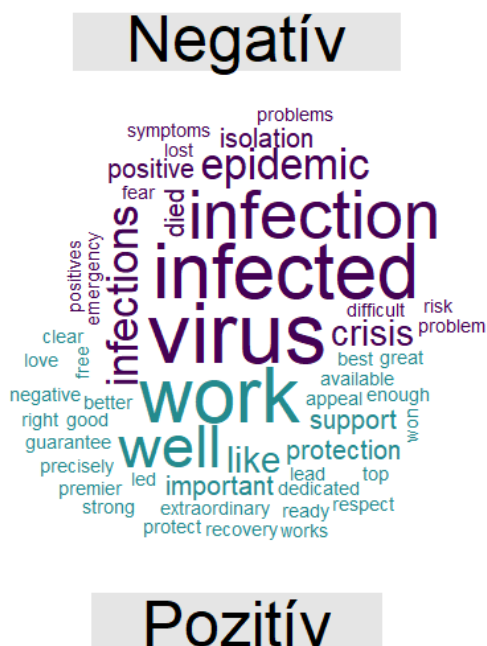


11. ábra. A teljes korpusz leggyakoribb szavai, havonta

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?

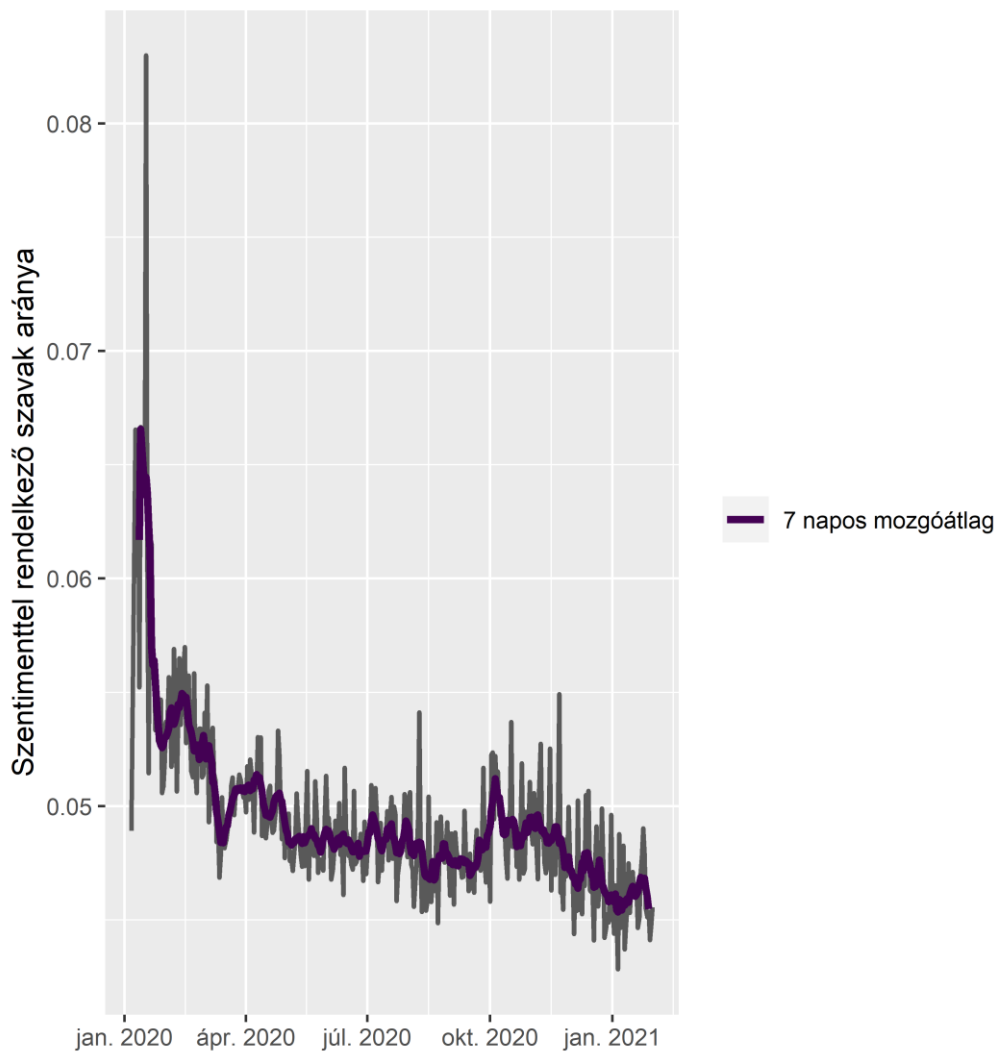
Szó	Gyakoriság	Régi szentiment	Új szentiment
positive	80778	1	-1
negative	25899	-1	1
patient	22501	1	-1
significant	15156	1	-1
fall	12073	-1	1
rapid	11753	1	-1
fell	10968	-1	1
decline	10752	-1	1
positives	10751	1	-1
fans	7820	1	-1
exceeded	7073	1	-1
fallen	5221	-1	1
sharp	3830	1	-1
falling	3774	-1	1
tough	3382	1	-1
exceed	3343	1	-1
defeat	3267	1	-1
falls	2984	-1	1
declining	2855	-1	1
exceeds	2300	1	-1
tougher	1661	1	-1

4. Táblázat. A leggyakoribb olyan szavak a korpuszban, melyek pontszámát megváltoztattuk a koronavírus-specifikus szótárban

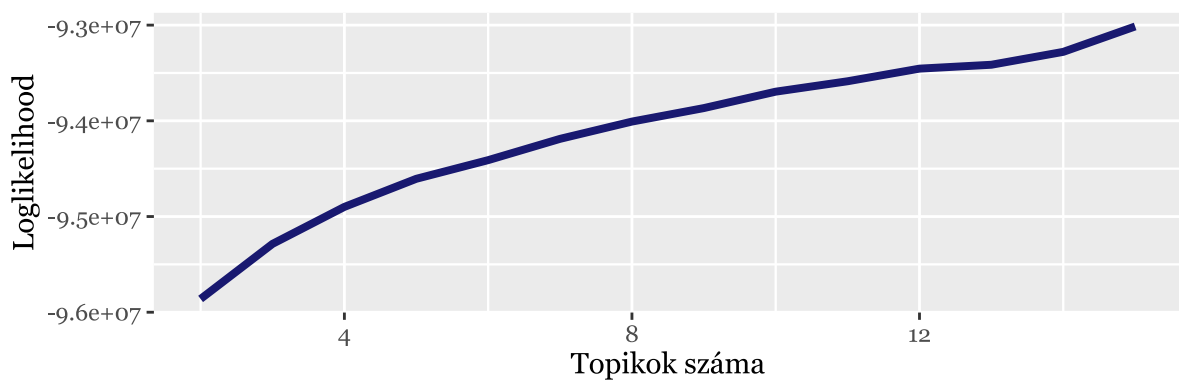


12. ábra. Leggyakrabban előforduló pozitív és negatív szentimenttel rendelkező szavak szófelhője

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?

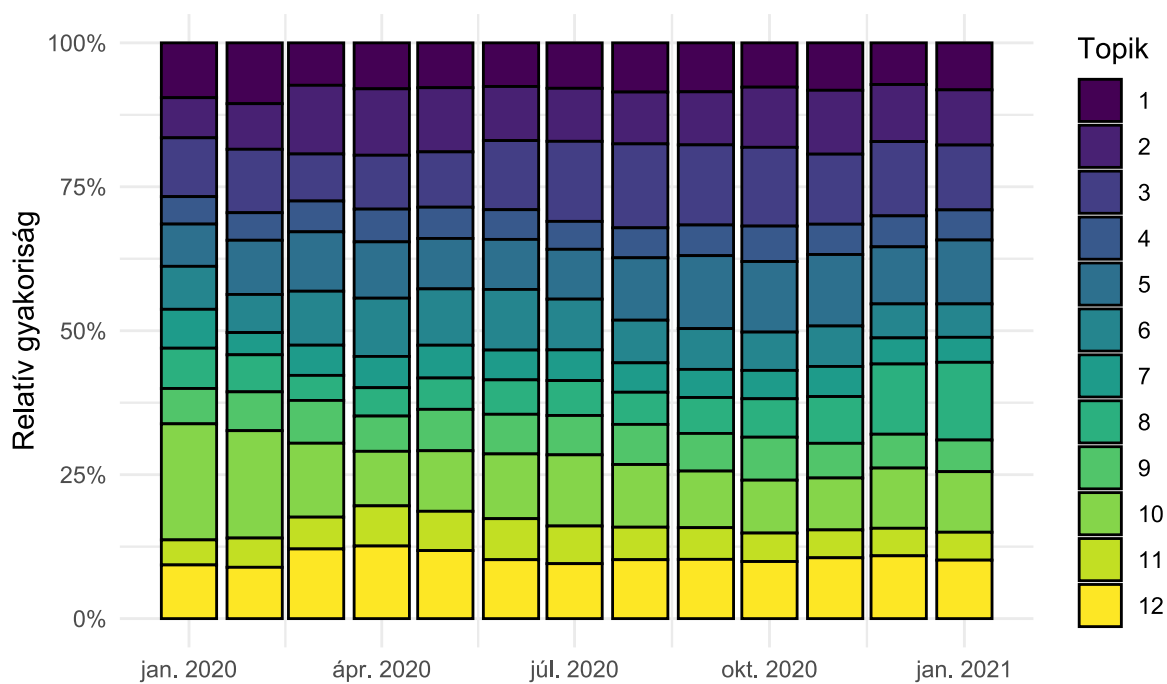


13. ábra. Szentimentet tartalmazó szavak időbeni alakulása, 7 napos mozgóátlaggal



14. ábra. Témák optimális számának megválasztása

Reális képet festenek-e a koronavírus-járvány súlyosságáról az európai sajtó hírei?



15. ábra. Topikok relatív megoszlásának időbeni dinamikája