Statisztikai Szemle

Közzététel: 2026. június 29.

A tanulmány címe:

Polarizációs dinamikák a Magyar Országgyűlésben 1998 és 2018 között

Szerzők:
Buda Jakab Máté, az Eötvös Loránd Tudományegyetem Társadalomtudományi Kar Statisztika Tanszékének tanársegédje, az ELTE Research Center for Computational Social Science kutatócsoportjának kutatója
E-mail: buda.jakab@tatk.elte.hu
Németh Renáta, az Eötvös Loránd Tudományegyetem Társadalomtudományi Kar Statisztika Tanszékének tanszékvezető professzora, az ELTE Research Center for Computational Social Science kutatócsoportjának társvezetője
E-mail: nemeth.renata@tatk.elte.hu

DOI: https://doi.org/10.20311/stat2026.06.hu0553

Buda Jakab Máté – Németh Renáta

Polarizációs dinamikák a Magyar Országgyűlésben 1998 és 2018 között

Dynamics of polarization in the Hungarian Parliament between 1998–2018

Buda Jakab Máté, az Eötvös Loránd Tudományegyetem Társadalomtudományi Kar Statisztika Tanszékének tanársegédje, az ELTE Research Center for Computational Social Science kutatócsoportjának kutatója
E-mail: buda.jakab@tatk.elte.hu
Németh Renáta, az Eötvös Loránd Tudományegyetem Társadalomtudományi Kar Statisztika Tanszékének tanszékvezető professzora, az ELTE Research Center for Computational Social Science kutatócsoportjának társvezetője
E-mail: nemeth.renata@tatk.elte.hu

Tanulmányunkban a Magyar Országgyűlésben 1998 és 2018 között elhangzott beszédeket elemezve a szöveges adatokon végzett felügyelt és felügyelet nélküli gépi tanulás kombinációjának alkalmazási lehetőségeit mutatjuk be. A polarizáció mértékét n-gram[1] alapú XGBoost (*extreme gradient boosting*) klasszifikációs modellel mérjük, amely a két domináns párt (Fidesz és MSZP) képviselőinek beszédeit különbözteti meg. Ezt kiegészítve 16 topikból álló strukturális topikmodellt (*structural topic model*, STM) alkalmazunk a polarizáció tartalmi csomópontjainak feltárására. Az egyes topikok polarizáltságát a modell biztossága alapján határozzuk meg, és ez alapján azonosítjuk a polarizáltabb és kevésbé polarizált témákat, majd ezeket mélyebben is megvizsgáljuk. Tudomásunk szerint ilyen vagy hasonló módszerkombinációval még nem kísérleteztek korábban, de eredményeink alapján érdemes lehet más esetekben is használni, amikor csoportok nyelvhasználati különbsé-gének értelmező feltárása a kutatási cél.

Kulcsszavak: természetes nyelvfeldolgozás, polarizáció, topikmodellezés

In this study we demonstrate the potential applications of combining supervised and unsupervised machine learning methods on texts by analyzing polarization on corpus consisting of speeches delivered in the Hungarian Parliament between 1998–2018. We measure the degree of polarization using an XGBoost classification model which distinguishes between speeches given by MPs of the two dominant parties (Fidesz and MSZP) of the studied period. We also fit a structural topic model (STM) with 16 topics to find the substantive topics of polarization. We determine the polarization level of individual topics based on the XGBoost model's confidence, which allows us to identify more and less polarized topics for deeper examination. To the best of our knowledge, this is the first time that a similar methodological combinations have been used, and our results suggest it may be a good apprroach in other cases as well, when the research objective is to interpretively explore linguistic differences between groups.

Keywords: natural language processing, polarization, topic modeling

[1] Szövegben vagy adatsorban *n* egymást követő szimbólumból (jelen elemzésben szóból) álló sorozat.

Tanulmányok sora szerint egyre erősebb politikai polarizáció figyelhető meg a közéleti diskurzusokban, ami miatt a diszkurzív szituációk egyre távolabb kerülnek a habermasi kommunikatív cselekvés ideáltípusától, hiszen a megkövült nézőpontok és a rögzült reakciók megakadályozzák a nyitottságot a másik fél mondandójára. A politikai polarizációval foglalkozó tanulmányok száma a hazai *(Körösényi, 2012; Janky, 2020)* és a nemzetközi *(Fiorina–Abrams, 2008; Jensen et al., 2012; Gentzkow et al., 2019)* szakirodalomban is jelentős. (A polarizációkutatásban használt különböző NLP [*natural language processing*, természetesnyelv-feldolgozás] irányokkal kapcsolatban lásd *Németh [2023]* áttekintő cikkét.)

A polarizáció mérésében a nemzetközi szakirodalomban az utóbbi időben a gépi tanulás eszköztára is megjelent *(Goet, 2019; Baly et al., 2020; Belcastro et al., 2020)*. Ugyanakkor a hagyományos mesterséges intelligencia[2] gyakran nehezen értelmezhető, és fekete dobozként működik, amelyben az algoritmus döntéseinek okai és működése nem világos – tehát egy-egy jelenség meglétére rá tud világítani, de a megértést sokszor nem segíti elő. Jelen cikkben a gépi tanulás segítségével előállított eredményeket a politikai polarizáció nyelvben megjelenő jellemzőinek feltárására használjuk.

Korábbi elemzésünkkel összhangban *(Buda–Németh, 2024)* a polarizációs metrikát úgy definiáljuk, hogy minél jobban képes az osztályozási modell azonosítani a szerző párthovatartozását, annál nagyobb fokú a politikai polarizáció. Ezt a megközelítést használták *Bayram és szerzőtársai (2019)* is, akik az Egyesült Államok Kongresszusa Képviselőházának (*U.S. House of Representatives*) felszólalásait elemezték. Az időben előrehaladva egyre jobb hatékonyságú osztályozó algoritmusokat tudtak alkotni, amit úgy interpretáltak, hogy a politikai polarizáció egyre jobban detektálhatóvá vált a beszédek nyelvezetében. Korábbi eredményeink alapján kijelenthető, hogy ezzel a módszerrel kimutatható a vizsgált korpuszon a polarizáció növekedése. Az alábbiakban a korábbi eredmények kiegészítéseként alaposabban feltárjuk a polarizáció összetevőit. A tanulmányban arra keressük a választ, hogy a Magyar Országgyűlésben zajló hivatalos politikai diskurzusban a

[2] A mesterséges intelligencia és gépi tanulás fogalmai gyakran keverednek. Ebben a tanulmányban a mesterséges intelligencia alatt elsősorban a mély neurális háló alapú modelleket értjük, amely a gépi tanulás részét képezi. A hagyományos gépi tanulási módszerek alatt a jóval kisebb, nem neurális háló alapú modelleket értjük (mint például a tanulmányban is alkalmazott XGBoost). A modellek megmagyarázhatósága sem a hagyományos gépi tanulás, sem a mélytanulás esetén nem jellemző – ugyanakkor léteznek kifejezetten interpretálhatóságra fejlesztett gépi tanulási modellek és post-hoc interpretáló eljárások is.

vizsgált húsz év alatt melyek voltak azok a témák, amelyek tárgyalása során erősebb polarizáció figyelhető meg, azaz melyek a politikai nézetek szerint megosztóbb témák. Vizsgáljuk az eredményeink szerint kisebb polarizációt mutató témákat is, ezeket tekinthetjük azoknak, amelyekben nagyobb a konszenzus.

Módszertani szempontból új megközelítéssel kísérletezünk: a felügyelt gépi tanulással előállított polarizációs metrikát egy ugyanazon a korpuszon, de függetlenül illesztett topikmodellel előállított topikokra vetítjük, így elemezve, hogy melyek a megosztóbb és a kevésbé megosztó témák. A tartalmi megállapításokon túl az itt bemutatott módszerek kombinációja módszertani szempontból is érdekes. Klasszifikációs gépi tanulás és topikmodellezés kombinációját tudomásunk szerint korábban elsősorban szövegklasszifikációs céllal alkalmaztak. A topikmodellezéssel vagy a klasszifikációs algoritmus bemeneti változóit állítják elő, akár egyfajta dimenziócsökkentésként *(Onan et al., 2016)*, vagy bizonyos téma detektálásakor a klasszifikációs algoritmus kiegészítéseként *(Missier et al., 2016)*.

1. Módszertan

1.1. Adatok

Elemzésünk az 1998 és 2018 közötti magyar parlamenti viták hivatalos korpuszán alapul, amelyet az időszak alatt a két meghatározó párt – a Fidesz és az MSZP – képviselőinek beszédeire szűkítettünk. (Az adatok részletesebb leírását lásd: *Buda–Németh, [2024]*). A beszédeket a K-monitor (http://parlament.k-monitor.hu/) gyűjtötte össze a Magyar Országgyűlés hivatalos honlapjáról (https://www.parlament.hu/). A vizsgált húszéves időszak öt parlamenti ciklust ölel fel, változó kormányzati összetétellel: Fidesz-vezetésű koalíció (1998–2002), MSZP–SZDSZ-, illetve MSZP-kormány (2002–2010), majd Fidesz–KDNP-többség (2010–2018). A pártok ideológiai bal-jobb spektrumon elfoglalt helye ez idő alatt változott, de a kutatás szempontjából ez nem lényeges, mert a nyelvi polarizációt – a szakirodalomban gyakran használt megközelítés szerint *(Peterson–Spirling, 2018)* – a pártok nyelvhasználata közötti különbségként értelmezzük. A beszédek közül kiszűrtük az ügyrendi és az 51 szónál rövidebbeket, így a használt korpusz 89 391 beszédből – nagyjából 20 millió szóból – áll.

1.2. A szövegek előfeldolgozása

Az elemzés során használt mindkét modell szózsákmodell, azaz a szövegek szavai (azok előfordulása vagy hiánya) adják a bemeneti jellemzőket (a magyarázóváltozókat). A nyers szövegek használata nagyon nagy változóteret eredményezne, ezért a modellek teljesítményének javítása érdekében a szövegeket – az NLP[3]-gyakorlatnak megfelelően – előzetesen feldolgoztuk. Ennek célja a szavak normalizálása (toldalékok eltávolítása), a lényegtelen (például a tartalmi jelentéssel nem bíró, de gyakori, kötőszavak: a stopszavak) eltávolítása, a több szóból álló megnevezett entitások (például személyek, szervezetek) nevének összevonása, így csökkentve az adatállomány dimenzionalitását. A gyakran előforduló szókapcsolatokat (n-gramok) azonosítottuk, és egyetlen tokenné konvertáltuk. Ez a folyamat segít megragadni olyan értelmes kifejezéseket, amelyek konkrét szemantikai információt közvetíthetnek (például a szomszédos „millió" és „forint" szavakat „millió_forint"-ra cseréltük). A beszédek leiratából eltávolítottuk a formális köszöntő (például „Tisztelt Ház!") és záró mondatokat („Köszönöm a figyelmet!"), illetve a jegyző által beszúrt, a teremben elhangzó megjegyzéseket és viselkedéseket leíró részleteket (például „Taps a fideszes képviselők soraiból").

Korábbi elemzésünkhöz képest új lépés, hogy a szövegekből egy iteratív eljárás során eltávolítottuk azokat a szavakat is, amelyek a hatalomban és az ellenzékben lévő pártok képviselőinek beszédeit különböztetik meg, függetlenül attól, hogy melyik párt van éppen hatalmon, ezzel elérve, hogy a maradék szavakra illesztett klasszifikációs modell elsősorban a párthovatartozást ragadja meg. Az iteratív szóeltávolítási folyamat során olyan segédmodelleket illesztettünk, amelyek nem a párthovatartozás (Fidesz vs. MSZP), hanem a parlamenti pozíció (kormány vs. ellenzék) alapján különítik el a beszédeket mind a teljes öt ciklust felölelő korpuszon. Célunk a kormánypárti és az ellenzéki nyelvhasználat megkülönböztetése volt, függetlenül a pártok politikai nézeteitől. Minden ilyen segédmodell illesztése után a SHAP[4]-értékek *(Lundberg–Lee, 2017)* alapján azonosítottuk a legfontosabb jellemzőket, amelyeket a „hatalom nyelvének" indikátoraiként értelmeztünk. Ezeket az n-gramokat kizártuk a következő iterációk jellemzőkészletéből, lehetővé téve további fontos jellemzők feltárását. A folyamatot addig ismételtük, amíg a modell pontossága 0,6 alá nem csökkent a teszthalmazon – ez a 21. iterációnál következett be. Az eljárás során összesen 3904 n-gramot azonosítottunk a „hatalom nyelve" markereként. Bár a módszer nem veszi figyelembe a nyelvhasználat, a párthovatartozás és a kormányzati pozíció közötti potenciális, többdimenziós kapcsolatokat, hatékonyan elkülöníti a kormányzati pozícióhoz köthető nyelvi jellemzőket.

[3] *Natural language processing* – természetes nyelvfeldolgozás.
[4] *Shapley additive explanations* – Shapley additív változófontosság-mutatók.

1.3. Klasszifikációs modell

A képviselők párthovatartozásának előrejelzésére XGBoost klasszifikációs modellt illesztünk a felszólalások szövegére, ahol a modell lényegében azt becsli, hogy mekkora annak az esélye, hogy az 1-essel kódolt Fideszhez tartozik a beszédet mondó képviselő. A polarizáció mértékét ennek az előrejelzésnek a biztosságaként operacionalizáljuk. Korábbi elemzésünkben *(Buda–Németh, 2024)* egy, a modellillesztés során nem használt külön teszthalmazon validáltuk az eredményeinket, jelen elemzésben azonban egy a teljes korpuszon illesztett modellt használunk. Mivel a polarizáltságnak ezt a mértékét csak relatívan értelmezzük, tehát nem hasonlítjuk más modellek eredményéhez, ez nem okoz problémát, de lehetőséget biztosít arra, hogy a későbbiekben az itt bemutatott eredményekkel konzisztensen tudjuk az egyéni képviselők polarizáltságát is mérni, és így összehasonlíthatók legyenek az eredmények. Mivel a korpusz nem kiegyensúlyozott párthovatartozás és parlamenti ciklusok szerint, a tanítás során súlyokat alkalmaztunk, hogy a kiegyensúlyozatlanság ne befolyásolja az eredményeket. (Enélkül a modell a háttéreloszlást is megtanulná, tehát a modell eredményeinek pártok közötti összehasonlíthatósága sérülne.)

Fontos tisztázni, hogy a modell által feltárt megkülönböztető mintázatok nemcsak tartalmi különbségeket tárhatnak fel, hanem egyéb nyelvhasználatbelieket is. (Ugyanakkor az előfeldolgozás során alkalmazott iteratív tokeneltávolítási eljárás biztosítja, hogy ezek elsősorban a párthovatartozásból eredő különbségek legyenek, ne pedig a kormányzati pozícióból adódóak.) Jelen tanulmány szempontjából minden ilyen különbséget a nyelvi polarizáció részeként értelmezünk, hiszen például az is a nyelvi távolság rézét képezi, ha a két oldal képviselői szisztematikusan más szavakat használnak egy adott téma kapcsán.

A polarizáció mértékeként a modell biztosságát használtuk, amelyet úgy definiáltunk, hogy mennyire távol van a helyes irányban a modell valószínűségi becslése a 0,5-től (ami lényegében azt jelenti, hogy a modell nem tudja eldönteni, hogy melyik párthoz tartozik a beszédet mondó képviselő). A biztosság képlete:

$$\text{biztosság} = \begin{cases} 2 \cdot (p - 0{,}5), \text{ha } x = 1 \text{ (Fidesz)} \\ 2 \cdot (0{,}5 - p), \text{ha } x = 0 \text{ (MSZP)} \end{cases}, \text{ ahol } p \text{ a modell valószínűségi becslése}$$

Így a biztosság az egyes beszédek esetében elvileg –1 és +1 között alakulhat. Például amennyiben a modell valószínűségi becslése egy fideszes képviselő beszédre 0,9, akkor a biztosság 0,8, és ugyanennyi a biztosság akkor is, ha egy MSZP-s képviselő beszédére 0,1 a valószínűségi becslés. A biztosság értéke 0 közeli, amennyiben a becslés 0,5-hez van közel, és negatív, amennyiben a másik párthoz tartozónak ítéli a modell a képviselőt.

Az XGBoost a strukturált adatok kezelésében való hatékonyságáról ismert osztályozó algoritmus, különösen alkalmas szöveges osztályozási feladatokra. A modellek a szöveget nem csupán szavanként, hanem szomszédos szócsoportonként dolgozták fel, azaz n-gram-változókat alkalmaztak. A szavak sorozatainak rögzítésével az n-gramok kontextus- és szintaktikai információt szolgáltatnak a modell számára, ami javítja a szöveg jelentésének és szerkezetének megértését. Jelen tanulmányban uni-, bi- és trigramváltozókat használtunk. A beszédeket alkotó szavak gyakoriságát bevett megoldásként a TF–IDF[5]-súlyok segítségével súlyoztuk, e súlyok a szavakat a dokumentumban való gyakoriságuk alapján rangsorolják az összes dokumentumban való gyakoriságukhoz képest, ezáltal kiemelve azokat a szavakat, amelyek gyakoriak a dokumentumban, de ritkák a korpuszban.

1.4. Topikmodell

A topikmodellezés célja látens témák azonosítása dokumentumgyűjteményekben. A módszer azt feltételezi, hogy a korpusz korlátozott számú topiból áll, ahol a topikok statisztikai értelemben a szókincs feletti valószínűségi eloszlásként definiálhatóak. Tehát minden topik a szótár szavainak valamilyen gyakorisági eloszlásából áll össze, az egyes szövegek pedig az így meghatározott topikokból állnak valamilyen arányban. A strukturális topikmodellezés (STM) *(Roberts et al., 2013; Roberts et al., 2014)* lehetőséget biztosít kontextuális információk bevonására a modellbe. A bevont változók befolyásolhatják a topikok tartalmát (a szavak feletti valószínűségi eloszlást) vagy a topikok előfordulási gyakoriságát.

1.4.1. Modellspecifikáció

Az optimális topikszám meghatározásához a korpusz egy véletlen mintáján illesztettünk különböző topikmodelleket 7 és 23 közötti topikszámmal. Az illesztett modellekre megvizsgáltuk, hogy a modellezés során nem felhasznált szövegeket mennyire jól modellezik (*held-out likelihood*), és ez alapján azt kerestük, hogy van-e olyan topikszám, amely az alacsonyabb topikszámhoz képest jelentős javulást okoz, ám a topikszám további növelésével már csökken a javulás mértéke (azaz könyökpontot kerestünk). Ez alapján a 16 topikból álló modell tűnt a legmegfelelőbbnek.

[5] *Term frequency-inverse document frequency* – kifejezésgyakoriság-inverz dokumentumgyakoriság.

A modellbe bevontuk a topikok tartalmát befolyásoló kontextuális változóként a képviselők párthovatartozását, míg a topikok előfordulási gyakoriságát a párthovatartozás mellett a parlamenti ciklus is befolyásolta, az időbeli változás modellezése érdekében.

1.4.2. A topikok értelmezése

A topikok értelmezése a legrelevánsabb kifejezéseik és a legjellemzőbb beszédek alapján történt. Fontos megjegyezni, hogy a topikok nem diszkrét szövegklaszterek, az egyes szövegek különböző mértékben több topikhoz is kapcsolódhatnak. Egy topik hozzájárulása egy adott dokumentumhoz azt tükrözi, hogy az adott topik milyen mértékben játszott szerepet a dokumentum tartalmának alakításában.

1.4.3. Eredmények kombinálása: a topikok polarizáltsága

A topikok polarizáltságának meghatározásához a modell biztosságának súlyozott átlagát számítottuk ki, ahol a súlyokat az adott topik aránya adta az egyes beszédekre:

$$\text{biztosság}_t = \frac{\sum_1^N b_i \times pr_i^t}{\sum_1^N pr_i^t},$$

ahol b_i a klasszifikációs modell biztossága az *i*-edik beszédre, pr_i^t pedig a *t*-edik topik aránya az *i*-edik beszédben. Így a topikok biztosságát azok a beszédek határozták meg jobban, amelyekben nagyobb arányban fordulnak elő.

2. Eredmények

Az eredmények bemutatását a topikmodell interpretációjával, a topikok bemutatásával kezdjük, ezután térünk rá az egyes topikokra jellemző polarizáltságra és ezek alaposabb tárgyalására.

2.1. A topikok értelmezése

Mivel jelen tanulmány fő témája nem a topikmodell interpretációja, így a topikokat csak röviden mutatjuk be. Az egyes topikok értelmezése mellett itt közöljük a topikok prevalenciáját, továbbá a legjellemzőbb topikszavakat[6] és egy-egy részletet a legjellemzőbb beszédből is.

- **1. topik:** nem tartalmi parlamenti vita
  - Ez a korpusz legjellemzőbb topikja, a prevalenciája 0,214, azaz a teljes korpusz több mint 1/5-e a tartalom nélküli vitába sorolható.
  - Legjellemzőbb topikszavak: ne_haragszik (177)[7], személyeskedik (151), csuda (147), megbánt (137), butaság (132), fölszólal (125), sapka (120), végigül (120), nyikos (120), sérteget (114), előhoz (114), speciel (111), micsoda (111), beszélget (109), általánosít (108), bocsánat (104), föláll (104), tessen (103), megüt (101), fölír (101)
- **2. topik:** a törvényhozás nyelvezete
  - Prevalencia: 0,094
  - Legjellemzőbb topikszavak: vitaszakasz (398), zárószavazás (392), házszabályszerű (357), általános_vita (355), zárószavazás_előtti_módosító_javaslat (348), koherenciazavar (341), ajánlási (312), ügyrendi_javaslat (296), módosító_javaslat (293), záróvita (289), zárószavazás_előtti (276), házszabályellenes (274), törvényalkotási_bizottság (270), módosítójavaslat (267), szakaszolás (258), költségvetési_bizottság (253), alkotmányügyi_bizottság (250), módosító_indítvány (242), tárgysorozatba (231), nyelvhelyességi (230)
- **3. topik:** politikai botrányok, korrupció
  - Prevalencia: 0,082
  - Legjellemzőbb topikszavak: malév (535), cash (517), hazudik (480), üzletárs (465), hazug (363), kulcsár_attila (347), tüntető (342), brókercég (326), pofátlanság (322), kormányszóvivő (307), simicska_lajos (306), offshore (305), hazudozás (299), bukott (299), vezérigazgató (296), vip (288), hazugság (274), szánalmas (274), párttárs (270), zsíros (257)

[6] Ez alatt azokat a szavakat értjük, amelyeknek az STM-illesztés során becsült generatív modell szerint az adott topik esetén a legnagyobb az előfordulási szorzója a szó teljes korpuszbeli előfordulásához képest, a párthovatartozás hatásától függetlenül. Tehát ezek nem a topikban leggyakrabban előforduló szavak, hanem azok, amelyek gyakorisága relatíve leginkább különbözik a háttéreloszlástól, tehát a leginkább megkülönböztető szavak. A szavak mögött zárójelben ezt a becsült szorzót tüntetjük fel.

[7] A felsorolt tokenek a topikmodellben megjelenő, tehát az előtisztítás utáni normalizált tokenformák; ezek nem minden esetben felelnek meg a szövegben előforduló formá(k)nak, például a ne_haragszik token előállhat a „ne haragudjon” és a „ne haragudjanak” fordulatból is.

- **4. topik:** adók és szociális ellátások
  - Prevalencia: 0,076
  - Legjellemzőbb topikszavak: adókulcs (610), egyszerűsített_vállalkozási_adó (587), vásárlóérték (583), adóteher (560), reálkereset (549), adócsökkentés (527), adórendszer (518), nyugdíjkassza (503), áfakulcs (500), adókedvezmény (498), gyes (489), adóterhelés (486), egykulcsos (475), gyermekes_család (465), minimálbér (463), adócsomag (462), cégautó (457), adóemelés (455), áfacsökkentés (453), kisadózó (447)
- **5. topik:** közigazgatás
  - Prevalencia: 0,064
  - Legjellemzőbb topikszavak: ajánlatkérő (648), ket (576), fogyasztóvédelmi_hatóság (569), ajánlattevő (547), bejelentési (516), ügyféli (505), ügyintézési (494), fogyasztóvédelmi (469), irányelvi (460), adatkör (458), cégeljárás (455), személyazonosító (450), átültetés (448), szabadalmi (447), piacfelügyeleti (445), fióktelep (442), elektronikus (435), elektronikus_aláírás (424), lakcímnyilvántartás (423), adóigazgatási (413)
- **6. topik:** költségvetés
  - Prevalencia: 0,058
  - Legjellemzőbb topikszavak: előirányzat (729), dologi_kiadás (716), céltámogatás (691), zárszámadás (690), főösszeg (665), önhibáján_kívül_hátrányos_helyzetben_lévő (663), zárszámadási (639), adóerő (551), bérhiki (535), normatívaemelés (506), gázközművagyon (482), adósságátvállalás (475), forrásmegosztási (456), fejezeti (455), kiadási (453), állami_számvevőszék (439), önerőalap (415), forrásmegosztás (394), felhalmozási (390), főszám (390)
- **7. topik:** külpolitika, EU
  - Prevalencia: 0,057
  - Legjellemzőbb topikszavak: külpolitika (763), iraki (696), humanitárius (645), külügyi_bizottság (623), külpolitikai (553), horvátország (547), rmdsz (543), kárpátaljai (532), koszovó (531), lisszaboni (530), szerbia (526), menedékkérő (518), vajdasági (506), ukrajna (498), irak (488), ratifikációs (483), diplomácia (478), albánia (475), koszovói (475), szabadkereskedelmi (474)
- **8. topik:** törvényhozás és végrehajtás jogi felügyelete
  - Prevalencia: 0,053
  - Legjellemzőbb topikszavak: mentelmi_bizottság (973), országos_választási_bizottság (949), mentelmi_jog (925), országos_rádió_és_televízió_testület (902), választási_eljárás (883), választási_iroda (860), ítélőtábla (839), duna_televízió (831), ügyrendi_bizottság (828), országos_bírósági_hivatal

(813), országos_igazságszolgáltatási_tanács (777), vagyonnyilatkozat (775), ügyész (770), közmédium (765), alkotmányjogi (746), választási_rendszer (743), ügyészség (735), választójog (716), alkotmánybíra (710), ajánlószelvény (705)

- **9. topik:** infrastruktúrát érintő kérdések, erős hangsúly a közlekedésen
  - Prevalencia: 0,050
  - Legjellemzőbb topikszavak: szén_dioxid (816), m3_as (812), megújuló_energia (787), energiahatékonyság (744), négysávos (744), elkerülő_út (737), útszakasz (730), m5_ös (720), holtág (716), szélerőmű (712), felüljáró (710), tározó (706), megawatt (704), m0 (703), elővárosi (703), matricás (699), tranzitforgalom (698), geotermikus (696), megújulóenergia (688), körgyűrű (679)
- **10. topik:** civil, tudományos, kulturális és szociálpolitikai programok, pályázatok, finanszírozás, erős hangsúly a sporton
  - Prevalencia: 0,048
  - Legjellemzőbb topikszavak: szabadidősport (970), versenysport (965), magyar_olimpiai_bizottság (925), sportág (921), alapprogram (862), sporttörvény (831), szakszövetség (828), fogyatékosügyi (814), leader (810), regionális_fejlesztési_tanács (769), akadálymentesítés (744), polgárőr (743), sportstratégia (742), gyermek_ifjúsági (731), sport (719), olimpiai (713), sportszövetség (687), nca (667), ifjúsági (659), sportoló (656)
- **11. topik:** mezőgazdaság
  - Prevalencia: 0,047
  - Legjellemzőbb topikszavak: felvásárlási (1050), családi_gazdálkodó (1049), aranykorona (1042), kukorica (1037), kárenyhítési (1035), gabona (1035), haszonbérleti (1034), nemzeti_földalap (1032), fagykár (1031), földtulajdonos (1019), földhasználó (1019), agrárgazdasági (1016), sps (1010), növénytermesztés (1005), állattenyésztés (998), állattartó (995), földhasználat (973), állattenyésztő (965), termékpálya (964), szövetkezeti (946)
- **12. topik:** kollektív emlékezet
  - Prevalencia: 0,041
  - Legjellemzőbb topikszavak: nagy_imre (1012), megemlékezés (747), színházi (706), életmű (661), kodály (648), színház (621), mártír (598), püspök (595), kitüntetés (595), kunó (580), illyés_gyula (562), megünneplés (560), könyvtár (553), emléktábla (548), múzeum (544), arany_jános (536), emlékév (528), református (520), közgyűjtemény (517), szobor (513)

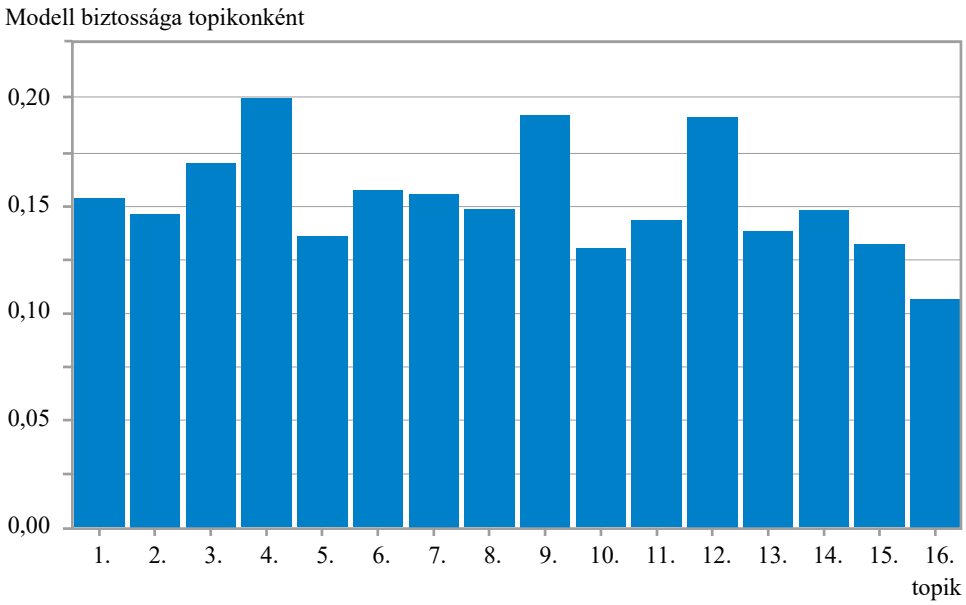
- **13. topik:** jogrend, jogállamiság, erős hangsúly a közbiztonságon
  - Prevalencia: 0,032
  - Legjellemzőbb topikszavak: büntetőjog (1365), lőfegyver (1360), szexuális (1341), családjogi (1325), bántalmazás (1294), szabálysértési_törvény (1227), család_belüli_erőszak (1225), büntető_törvénykönyv (1181), élettársi (1159), szabálysértés (1155), szabadságvesztés (1155), szabálysértési (1143), bűnmegelőzési (1119), büntetési_tétel (1097), bűncselekmény (1085), bűnözés (1077), büntetőeljárási (1068), sértett (1065), bűnmegelőzés (1061), elzárás (1041)
- **14. topik:** szakképzés és munkaerőpiac
  - Prevalencia: 0,032
  - Legjellemzőbb topikszavak: tanulószerződés (1419), tartalékos (1403), sorkatonai_szolgálat (1286), szakképző (1286), magyar_honvédség (1209), szakképzési (1200), haderő (1191), szakképzés (1179), állományú (1124), felnőttképzési (1087), nemzetőrség (1042), szakképzési_hozzájárulás (1026), szolgálatteljesítés (1017), felnőttképzés (1012), honvédelmi_törvény (1010), illetményfejlesztés (1005), közismereti (996), illetményrendszer (994), illetménykiegészítés (951), sorkatona (949)
- **15. topik:** egészségügy, nyugdíj
  - Prevalencia: 0,027
  - Legjellemzőbb topikszavak: patika (1786), járóbeteg (1767), fekvőbeteg (1757), szakorvos (1754), várólista (1740), háziorvos (1718), háziorvosi (1713), beteg (1710), szakellátás (1698), gyógyszerész (1674), betegellátás (1664), gyógyszertár (1656), mentőállomás (1641), magyar_orvosi_kamara (1629), gyógyszerellátás (1618), országos_egészségpéztár (1611), egészségbiztosítási_pénztár (1597), egészségbiztosítás (1590), kórházi (1580), gyógyszergyártó (1564)
- **16. topik:** turizmus, helyi fejlesztések
  - Prevalencia: 0,024
  - Legjellemzőbb topikszavak: balaton (1502), városliget (1387), területrendezési (1380), tó (1379), vendégéjszaka (1334), régészeti (1331), zöldterület (1330), világörökségi (1319), agglomerációs (1269), balatoni (1264), turisztikai (1237), területrendezés (1234), gyógy (1233), turizmus (1218), üdülőkörzet (1199), külterület (1197), örökségvédelmi (1191), műemlékvédelmi (1185), borvidék (1172), vízminőség (1152)

2.2. A topikok polarizáltsága

Az egyes topikok polarizáltságának megbecsléséhez kiszámoltuk a klasszifikációs modell biztosságának prevalenciával súlyozott átlagát topikonként az 1.4.3. fejezetben bemutatottak szerint. Az eredményeket az 1. ábra szemlélteti.

1. ábra

Topikok polarizáltsága
Polarization of topics



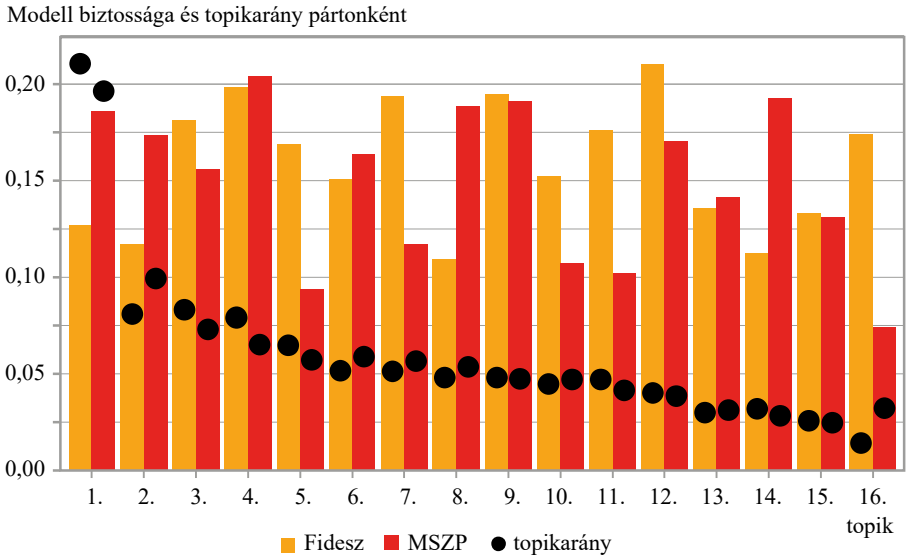
Forrás: saját szerkesztés.

Ezek a számok önmagukban nem értelmezhetők, de egymáshoz tudjuk viszonyítani őket. Az ábra alapján jól látszik, hogy az alkalmazott módszer szerint a klasszifikációs modell által felismert különbségek leginkább a 4. (adók és szociális ellátások), a 9. (infrastruktúra), a 12. (kollektív emlékezet) és a 3. (botrányok, korrupció) topikban jelennek meg, tehát értelmezésünk szerint ezek voltak a legpolarizáltabb témák a vizsgált 20 év alatt. Legkevésbé pedig a 16. (turizmus, helyi fejlesztések), a 10. (sport, civil, tudományos, kulturális és szociálpolitikai programok), a 15. (egészségügy, nyugdíj) és az 5. (közigazgatás) topikban jelentkezik a polarizáció modell által kódolt aspektusa.

Vessük össze ezeket az eredményeket a pártonként mért modell biztosságokkal (2. ábra).

2. ábra

Topikok polarizáltsága és prevalenciája pártonként
Polarization and prevalence of topics by party



Forrás: saját szerkesztés.

A 2. ábra az 1. ábrához hasonlóan a klasszifikációs modell prevalenciával súlyozott biztosságát mutatja topikonként, de most pártok szerinti bontásban. Azokra a topikokra érdemes figyelni, ahol nagy a relatív különbség a két pártra számított biztosság között, mert ez azt jelenti, hogy az egyik párthoz tartozó képviselőket viszonylag biztosan osztályozza a klasszifikáció, míg a másik párthoz tartozó képviselőket, ha erről a topikról beszélnek, kevésbé biztosan. A polarizálóbb topikok közül a 9. (infrastruktúra) és a 4. (adó és szociális ellátások) esetében alacsony a különbség a két párt biztossága között. Ezzel szemben a 12.-nél (kollektív emlékezet) és 3.-nál (botrányok, korrupció) jelentősebb különbséget tapasztalunk. Ez alapján úgy tűnik, hogy ez utóbbi két topik esetén a fideszes képviselők beszédei között vannak markánsabban azonosíthatók, polarizáltabbak, mint az MSZP-seké. Hasonló a helyzet a kevésbé megosztó 10. (sport-, civil, tudományos, kulturális és szociálpolitikai programok) és 5. (közigazgatás) topikok estén is.

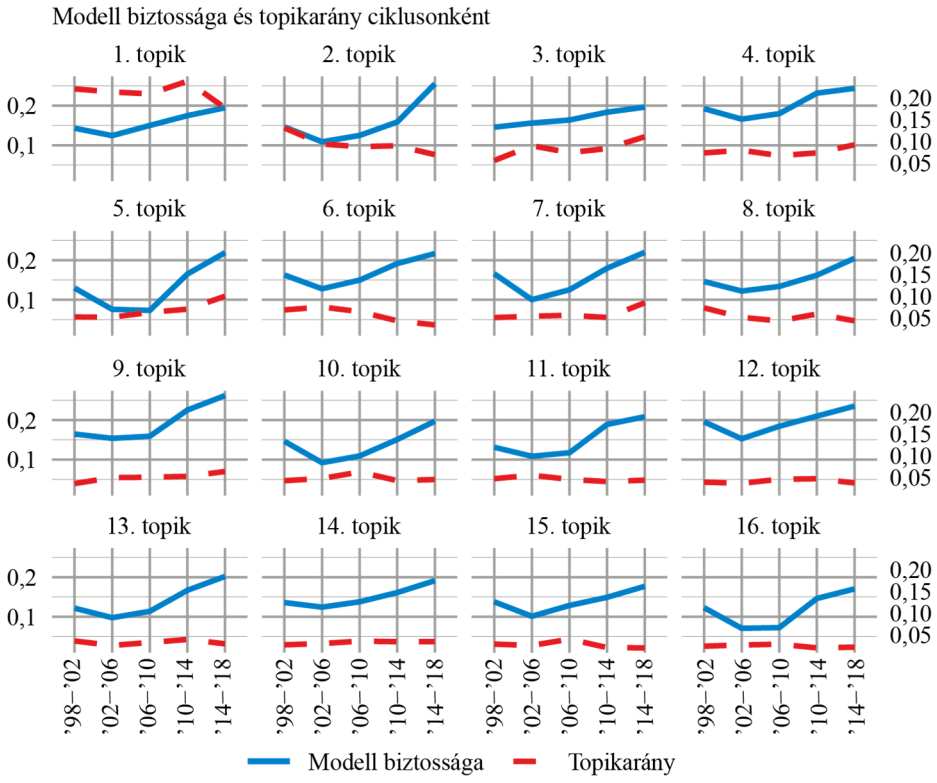
Az ábrán a prevalenciák alakulását is feltüntettük, mert amennyiben egy topik lényegesen jellemzőbb az egyik pártra, mint a másikra, akkor a biztosságokban tapasztalható hasonló különbség eredhet ebből az eloszlásbeli különbségből is. Ezzel szemben érdekes lehet, ha azt tapasztaljuk például, hogy az egyik párt esetén nagyobb egy topik prevalenciája, de a másik párt esetén biztosabb a modell. Ilyet, hogy a prevalenciák között lényegi különbség van, ám a modellbiztosságok

között fordítva áll fent lényegi különbség, csak a 16. (turizmus, helyi fejlesztések) topik esetén figyelhetünk meg. Ennek talán az lehet az oka, hogy annyira alacsony a topik általános prevalenciája (0,024), hogy a klasszifikációs modell az itt megfigyelhető szógyakoriságokat nem kódolta, de szerepet játszik az adattisztítás is: a hatalmi pozícióval összefüggő tokenek megtartásával illesztett modell esetén ennél a topiknál is az eloszlásbeli arányokat tükrözi a biztosságokban tapasztalható különbség.

Végül a 3. ábrán a topikok biztosságának és prevalenciájának alakulását láthatjuk ciklusonként. Érdemes megfigyelni, hogy a legtöbb topikban a korábbi eredményeink alapján is azonosított mintázat figyelhető meg. Az 1998–2002-es ciklus után visszaesés látható a polarizáltság itt azonosított aspektusában, majd 1-2 ciklus alatt ismét eléri ezt a szintet, és a 2010–2014-es ciklusban többnyire meghaladja, majd az utolsó vizsgált ciklusban további növekedés figyelhető meg.

3. ábra

Topikok polarizáltságának alakulása időben
Temporal evolution of topic polarization



Forrás: saját szerkesztés.

Ettől a mintázattól eltér a 9. (infrastruktúra) és a 3. (botrányok, korrupció) topik esetén tapasztalható tendencia, ahol nincs visszaesés, vagy csak minimális van az itt mért polarizáltságban az 1998–2002-es és a 2002–2006-os topik esetében. Ez a két topik az erősen polarizáltak közé tartozik, esetükben a teljes időszak alatt megfigyelt erős polarizáltság egyik összetevője épp ennek a visszaesésnek a hiánya lehet. Tehát ebben a két topikban a 20 év alatt végig meghatározó különbség figyelhető meg a két párt között, az idetartozó beszédeket alapos kvalitatív elemzésnek lenne érdemes alávetni a későbbiekben a polarizáció mélyebb megértése érdekében, de a strukturális topikmodell által azonosított megkülönböztető szavak alapján elképzelhető (Fidesz: tisza, vízpótlás, villamosítás, főút, árvízvédelmi, kerékpárút, vízkár, vasútvonal, vásárhelyi_terv, megépítés, m7, szigetköz, szennyvíz, közúthálózat, elvezetési, víztározó, szigetközi, repülőtér, környezethasználati, úthálózat; MSZP: munkatevékenység, hívódik, emlékkép, cigánytelep, illesztett, lelkület, legkiváltképp, hozzáillesztett, erdőgazdálkodó, akképpen, motivációs, közfoglalkoztatás, párosuló, nemzetiségi_önkormányzat, szétzúz, víziközmű, futamidejű, véletlenszerű, energiapolitika, közműadó), hogy a Fidesz inkább a téma beruházási, fejlesztési, míg az MSZP a környezeti és a szociális vetületére összpontosít.

3. Diszkusszió

Az eredmények azt mutatják, hogy a hatalmi helyzettől független különbségek a két párt között leginkább az adópolitikában, a szociális ellátásokban (4. topik), az infrastruktúra-fejlesztésről alkotott véleményben (9. topik), a kollektív emlékezetben (12. topik) és a botrányokról szóló diskurzusban (3. topik) mutatkoznak meg. Ez alapvetően konzisztens az egyéb ismereteinkkel, hiszen adópolitikában és kollektív emlékezetben is azt várjuk, hogy legyen markáns különbség egy magát konzervatívként és egy magát baloldaliként pozicionáló párt között. Ahogy az is konzisztens a korábbi ismeretekkel, hogy a sport sokszor kevésbé polarizált téma. Az eredmények validálásához további kvalitatív elemzésre lesz szükség. Mindenképp érdemes figyelembe vanni, hogy a klasszifikációs modellt a hatalmi pozíciót legjobban jelző szavaktól megtisztított korpuszon illesztettük, így a polarizációnak csak a hatalmi pozíciótól nagyjából független aspektusait mutatják ezek az eredmények. Feltehető, hogy amennyiben a hatalmi pozícióval összefüggő nyelvhasználati különbségeket is kódolná a modell, a 15. topik (egészségügy, nyugdíj) is

erősebb polarizáltságot mutatna, hiszen ezek olyan témák, amelyeket jellemzően erősen támad az ellenzék.

Az itt bemutatott eredmények a korábbi következtetéseinket alátámasztják és kiegészítik: a legtöbb topik esetében az általánosan felismerhető polarizációs dinamika azonosítható, de egyes topikoknál ettől jelentős eltérést tapasztaltunk.

A tanulmány a tartalmi szempontokon túl módszertani újdonsággal is szolgál: a topikmodellezés és a felügyelt gépi tanulás ilyen kombinációjával tudomásunk szerint előttünk még nem kísérleteztek, és a bemutatott eredmények szerint a módszer ígéretesnek tűnik. A fentiekben a polarizációt szubsztantív témákra bontottuk, meghaladva a hagyományos gépi tanulási megközelítést. Ez a módszerkombináció a topikmodellezésnél több részletet tár fel, és az osztályozásnál interpretálhatóbb eredményeket eredményez. A módszer használható lehet más esetekben is, ahol különböző csoportok nyelvi különbségének számszerűsítése mellett a különbségek értelmezése is a kutatási kérdés része. Például az Egyesült Államok bírósági rendszerével kapcsolatban gyakran felmerülnek a republikánusok és a demokraták által kinevezett bírók ítélkezési gyakorlatában mutatkozó különbségek. Amennyiben rendelkezésre áll megfelelő korpusz, diszkriminációkutatásra is alkalmazható lehet az eszköz: például feltárható lenne, hogy egy cég ügyfélszolgálati munkatársai különbözőképp kezelik-e a nő és a férfi ügyfeleket, és ha igen, milyen témák esetén jelentkezik markánsabban ez a különbség.

Köszönetnyilvánítás

A tanulmány a Kulturális és Innovációs Minisztérium EKÖP-24 kódszámú Egyetemi Kiválósági Ösztöndíj Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

Irodalom

Baly, R. – Martino, G. D. S. – Glass, J. – Nakov, P. (2020): We can detect your bias: predicting the political ideology of news articles. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4982–4991. https://doi.org/10.18653/v1/2020.emnlp-main.404

Bayram, U. – Pestian, J. – Santel, D. – Minai, A. A. (2019): What's in a word? Detecting partisan affiliation from word use in congressional speeches. In: *2019 International Joint Conference on Neural Networks (IJCNN).* https://doi.org/10.1109/IJCNN.2019.8851739

Belcastro, L. – Cantini, R. – Marozzo, F. – Talia, D. – Trunfio, P. (2020): Learning political polarization on social media using neural networks. *IEEE Access,* 8, 47177–47187. https://doi.org/10.1109/ACCESS.2020.2978950

Buda J. – Németh R. (2024): A felügyelt gépi tanulás alkalmazási lehetőségei szöveges adatokon. A magyar országgyűlésben 1998–2018 között elhangzott beszédek elemzése. *Statisztikai Szemle*, 102(11), 1087–1103. https://doi.org/10.20311/stat2024.11.hu1087

Fiorina, M. P. – Abrams, S. J. (2008): Political polarization in the American public. *Annual Review of Political Science,* 11, 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836

Gentzkow, M. – Shapiro, J. M. – Taddy, M. (2019): Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica,* 87(4), 1307–1340. https://doi.org/10.3982/ECTA16566

Goet, N. D. (2019): Measuring polarization with text analysis: evidence from the UK House of Commons, 1811–2015. *Political Analysis*, 27(4). https://doi.org/10.1017/pan.2019.2

Janky B. (2020): Elit diskurzus, politikai identitás és polarizáció Magyarországon. In: *Társadalmi Riport 2020* (pp. 462–477). https://doi.org/10.61501/TRIP.2020.20

Jensen, J. – Kaplan, E. – Naidu, S. – Wilse-Samson, L. (2012): Political polarization and the dynamics of political language: evidence from 130 years of partisan speech. *Brookings Papers on Economic Activity,* 2012, 1–81. https://doi.org/10.1353/eca.2012.0017

Körösényi A. (2012): A politikai polarizáció és következményei a demokratikus elszámoltathatóságra. In: Boda Zs. – Körösényi A. (szerk.): *Van irány? – Trendek a magyar politikában*, MTA TK PTI Új Mandátum, Budapest, 284–309. https://politikatudomany.tk.hun-ren.hu/uploads/files/archived/2172_IV_03_Korosenyi_Politikai_polarizacio.pdf

Lundberg S. M. – Lee S. I. (2017): A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1705.07874

Missier, P. – Romanovsky, A. – Miu, T. – Pal, A. – Daniilakis, M. – Garcia, A. – da Silva Sousa, L. (2016): Tracking dengue epidemics using twitter content classification and topic modelling. In: *Current Trends in Web Engineering*. ICWE 2016. (pp. 80–92). https://doi.org/10.1007/978-3-319-46963-8_7

Németh, R. (2023): A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, 6(1), 289–313. https://doi.org/10.1007/s42001-022-00196-2

Onan, A. – Korukoglu, S. – Bulut, H. (2016): LDA-based topic modelling in text sentiment classification: an empirical analysis. *International Journal of Linguistics and Computational Applications*, 7(1), 101–119. http://www.ijcla.org/2016-1/IJCLA-2016-1-pp-101-119-preprint.pdf

Peterson, A. – Spirling, A. (2018): Classification accuracy as a substantive quantity of interest: measuring polarization in Westminster systems. *Political Analysis,* 26(1), 120–128. https://doi.org/10.1017/pan.2017.39

Roberts, M. E. – Stewart, B. M. – Tingley, D. – Airoldi, E. M. (2013): The structural topic model and applied social science. Advances in neural information processing systems workshop on *Topic Models: Computation, Application, and Evaluation,* 4(1), https://mimno.infosci.cornell.edu/nips2013ws/slides/stm.pdf

Roberts, M. E. – Stewart, B. M. – Tingley, D. – Lucas, C. – Leder-Luis, J. – Gadarian, S. K. – Albertson, B. – Rand, D. G. (2014): Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082. https://doi.org/10.1111/ajps.12103