Eötvös Loránd Tudományegyetem

Társadalomtudományi Kar

MESTERKÉPZÉS SZAKDOLGOZAT

Trianon emlékezetpolitikájának vizsgálata gépi tanulási és szöveganalitikai eszközökkel

Konzulens:

Dr. Németh Renáta

Készítette:

Berbekár Réka

H67RQ8

Survey Statisztika és Adatanalitika szak

2022. november

Köszönetnyilvánítás

A szakdolgozatomhoz kapott segítséget köszönöm Dr. Barna Ildikónak és Knap Árpádnak, akik rendelkezésemre bocsátották a korpuszt és segítettek kutatásuk és módszertanuk értelmezésében, Pólya Tibornak, aki biztosította a NarrCat eszköz használatát és futtatta a cikkek feldolgozását, illetve Dr. Németh Renátának, aki végigkísérte a munkámat és sok remek ötlettel támogatott.

Tartalom

1. Bevezetés

Több, mint 100 év telt el a trianoni békeszerződés aláírása óta, a Trianon közbeszédben való jelenléte azonban még mindig meglehetősen aktív. Sorra avatják az emlékműveket, számos megemlékezésnek lehetünk tanúi, a határon túli magyarok helyzete is folyamatosan téma az újságírók, politikusok körében. Trianon egy kollektív történelmi trauma, amit a magyarok Fülöp és László (2013) traumafeldolgozási modellje szerint nem dolgoztak még fel, ezt bizonyítja például a téma aktív jelenléte vagy a Trianonnal kapcsolatos szövegekben előforuduló érzelmi kifejezések száma is. A különböző politikai oldalak képviselőinek kommunikációjában megfigyelhetőek egyértelmű különbségek. Magánemberként is az a benyomásunk általában, hogy a szélsőjobboldal és a jobboldal aktívabban foglalkozik a témával, a baloldal tárgyilagosabb. Témaválasztásomat indokolja a Trianon 100. évfordulójához köthető intenzív cikkezés a témáról, illetve a politikai polarizáció növekedése is.

A szöveganalitika mint módszertani terület egyre elterjedtebb, az online felületeken rengeteg írásos anyag jelenik meg, legyen szó cikkekről, közösségi médiáról, termékleírásokról. Ezek a hosszabb-rövidebb szövegek alkalmasak arra, hogy az írójáról vagy a megjelenés helyéről következtetéseket vonjunk le. A szöveganalitikai módszerek azért fontosak, mert nagy mennyiségű írásos anyagot elemezni tudunk kvalitatív módon elemezni tudjunk

Kutatásom során arra keresem a választ, hogy milyen eltérések figyelhetők meg az online média Trianonnal kapcsolatos cikkeinek témájában és narratívájában a hírportálok politikai hovatartozásától függően. A használt adatok 2017 és 2020 közötti újságcikkek, amik tartalmazzák a „Trianon” szót. 36 hírportálról származnak az írások, amiket 4 kategóriába soroltam: szélsőjobb, kormánypárti, nem kormánypárti jobboldali és baloldali-liberális. Ami a módszertant illeti, egyrészt azt vizsgálom, hogy a NarrCat eszköz miként egészíti ki a topikmodellezés eredményeit, ehhez leíró elemzést végzek és béta regressziós modellt használok. A topikmodellezés során az algoritmus csoportokat hoz létre a szövegekből a szavak együttes előfordulása alapján, majd minden újságcikkre megad egy valószínűséget, hogy mennyire jellemző rá az adott csoport, azaz topik (Barde és Bainwad, 2017). A NarrCat egy olyan elemző szoftver, ami pszichológiai és nyelvtani formákat, kifejezéseket azonosít a szövegben, majd több dimenzióban visszaadja a kifejezések relatív gyakoriságát a szövegben. Ilyen dimenzióra példa a „Tagadás” vagy az „Én-referencia”. (László et al., 2013) Vizsgálom,

hogy mely NarrCat értékek mentén különülnek el a különböző politikai oldalakhoz köthető honlapok írásai, azaz, hogy mennyire más ezeknek az oldalaknak a hozzáállása, a narratívája Trianonnal kapcsolatban. Ugyanígy nézem azt is, hogy melyek azok a topikok, amelyek jobban jellemzők valamelyik oldalra, mely témák közösek az egyes politikai irányultságú hírportálok esetében, ki ír hasonló dolgokról és ki különbözőkről.

Szeretném továbbá bemutatni, hogy milyen újdonságot adhat hozzá a NarrCat eszköz a természetes nyelvfeldolgozás területéhez, ezen kívül vizsgálom azt is, hogy a topik és a NarrCat értékek segítségével mennyire jó klasszifikációs modellt lehet építeni. A kutatás ezen részében topikmodellezés és a NarrCat output értékei szolgálnak magyarázó változóként klasszifikációs modelljeimben, a politikai hovatartozás pedig eredményváltozóként. A kvadratikus diszkriminancia-analízis, a Support Vector Machine a Random Forest és az Extreme Gradient Boosting klasszifikációs algoritmusok optimális paramétereit igyekszem paraméterhangolással megtalálni, és ezen modellek teljesítményét hasonlítom össze a csoportba sorolás pontossága és a Cohen's Kappa mutató segítségével.

Szakdolgozatom újdonságereje főként abban rejlik, hogy a NarrCat eszköz nem volt még ilyen nagy, több, mint 11 ezer cikket tartalmazó korpuszon használva, illetve a NarrCat értékeket egyáltalán nem, a topikmodellezési értékeket is (legjobb tudomásom szerint) csak viszonylag kevés esetben használták klasszifikációs modellekhez. Olyan kutatás sem született még, ahol a topikmodellezést és a NarrCat elemzést együtt használták volna. Azt gondolom, hogy a szövegek témájának elemzését jól kiegészíti a szövegek stílusának, narratívájának nyelvtani és pszichológiai elemeket vizsgáló elemzése.

Ami dolgozatom felépítését illeti, a bevezetőt követően betekintést adok a téma és a módszertani szakirodalmi hátterébe, majd részletesen bemutatom a topikmodell és a NarrCat módszereket, a használt klasszifikációs algoritmusokat, azt követően pedig az adatokat és az adattisztítási folyamatot. Elemzésem leírását a NarrCat eredményekkel kezdem, vizsgálom, hogy azok a cikkek politikai hovatartozásától függően mennyire változnak. A topikmodellezés folyamatának és eredményének bemutatásánál arra is kitérek, hogy a topikokban milyenek az átlagos NarrCat értékek, illetve megvizsgálom a topikok politikai hovatartozás szerinti eloszlását is. Ezt követően béta regressziós modellel vizsgálom a NarrCat és a topik értékek magyarázó erejét, majd összehasonlítom a négy klasszifikációs algoritmus teljesítményét.

Szakdolgozatomat a kutatás összegzésével, korlátaival, eredményeim felhasználási lehetőségeivel zárom.

2. Szakirodalmi áttekintés

2.1. Elemzések a NarrCat eszközzel

A NarrCat narratív kategorikus tartalomelemző szoftvert a Pécsi Tudományegyetem Bölcsészettudományi karának kutatói fejlesztették, ez megmagyarázza, hogy a legtöbb NarrCatot mint módszertant használó kutatás az ő nevükhöz fűződik. Az alábbiakban bemutatok néhány elemzést, ami a NarrCat eszközzel készült. Érdemes megfigyelni, hogy ezek a kutatások viszonylag kis korpuszon folytak, a szövegeket sokszor kvalitatív módon is tanulmányozták. Olykor lehetősége nyílt a kutatóknak mondat szinten is vizsgálni a szövegeket, és a NarrCatnak csak néhány modulját használták az elemzéshez. Ezzel szemben jelen dolgozat több, mint 13 ezer újságcikkel dolgozik, hét NarrCat modult felhasználva, a NarrCat értékeket gépi tanulási eszközökkel elemezve. Ez a módszer jelentősen eltér az alábbiakban bemutatott cikkekben lévőktől, mégis ezen kutatásokkal tudom legjobban bemutatni a NarrCat felhasználását.

Ferenczhalmy, Szalai és László (2011) az ágenciát kutatták történelemkönyvek és civil elbeszélések szövegei segítségével. A vizsgált leírások fontosabb magyar történelmi eseményekről szóltak, a magyarok más nemzetekkel való összehasonlítása volt a cél ágencia tekintetében. Az ágenciát úgy definiálják, mint a fizikai és szociális környezetre való hatás képessége, de megjelenik fogalmaik között a kontroll, motiváció, cselekvések szándékossága, kitűzött cél elérésének képessége is. A NarrCat eszköz ágencia modulját használták az elemzéshez, ami a szövegekben előforduló aktivitást, passzivitást, illetve a kényszert és intenciót azonosítja. 10 történelmi eseményt vizsgáltak a kutatók, amiket pozitív, negatív és vegyes kategóriába soroltak. Pozitív volt például a honfoglalás és az államalapítás, negatív a tatárjárás és a trianoni békeszerződés, vegyes pedig a török harcok és a habsburg uralom. Két korpuszt vizsgáltak, az első általános- és középiskolai történelemkönyvek szövegeiből áll, a második egy 500 fős kutatásból származó szövegeket tartalmaz, ahol magánemberek meséltek el történelmi eseményeket saját szavaikkal. Legfőbb eredményük, hogy a magyarok mindkét korpusz esetében alacsonyabb ágenciával rendelkeztek, mint a szövegekben szereplő külső csoportok. Ez azt jelenti, hogy passzív szereplőként jelennek meg a magyarok az

elbeszélésekben, velük inkább csak megtörténnek a dolgok, kényszerítik őket a cselekvésre, nincsenek aktív hatással az eseményekre. Néhány pozitív esemény képez kivételt ezalól, például a honfoglalás és az 56-os szabadságharc pozitív mozzanatai.

Ehmann és Balázs (2015) hat magyar űrhajós naplóját elemezte, akik két hetet töltöttek el a NASA marskutató állomásán összezárva. A NarrCat pozitív és negatív érzelem, valamint a szelf- és mi referencia moduljait használták arra, hogy a hat űrhajós mentális állapotának változásait nyomon kövessék. Létrehoztak egy emocionalitás és egy csoportkohézió mutatót. Azt találták, hogy a negatív érzelmi állapot és az alacsony csoportkohézió csoporton belüli konfliktust jelent, ha a negatív érzelem és a csoportkohézió is erős, akkor pedig külső fenyegetésről van szó. Korábbi kutatásokban már igazolták, hogy elzárt kiscsoportok az együtt töltött idő harmadik negyedénél több negatív érzelmet, stresszt, konfliktust élnek meg. Ez Ehmann és Balázs kutatásánál is megjelent, pedig ott csupán két hetet töltöttek együtt a vizsgált alanyok.

Kovács Réka (2012) a párkapcsolati elégedettséget kutatta házastársak megkérdezésén keresztül, azt vizsgálva, hogy van-e különbség nem és életkor szerint közöttük. Interjút készített negyven házaspárral négy témában: megismerkedés, pozitív csúcsélmények, megoldott problémák, megoldatlan problémák. Az alanyok által elmondott összefüggő szövegek leiratai alkották a korpuszt, amin elvégezte a szerző a NarrCat elemzést. A szelf-referencia, mi-referencia, pozitív és negatív érzelmek, pozitív és negatív értékelés mérőszámokat használta a négy csoport (fiatal férfiak, fiatal nők, idős férfiak, idős nők) összehasonlítására, kevert mintás varianciaanalízist alkalmazva. A szelf-referencia mértéke nem különbözött a négy csoportban, azonban téma szerint eltérő volt. A megismerkedéshez köthető történetekben magas volt az én-referencia, ebből látszik, hogy akkor még az egyén van középpontban, később pedig már a két ember együttese. A mi-referencia a megoldatlan problémáknál alacsony, mert a kudarcérzés miatt gyengül a házastársi „mi-tudat". Ahogyan az sejthető, a pozitív kifejezések száma a megismerkedés és a pozitív élmények kategóriákban sokkal magasabb, mint a másik kettőben. A fiatal férfiaknál a megoldott és a megoldatlan problémáknál is magas a negatív érzelmi kifejezések száma, ez azt mutatja, hogy a fiatal férfiak kevésbé optimisták, alacsony önértékeléssel és nyitottsággal állnak a problémákhoz. Az értékelés is hasonló eredményt mutat, az idősebbek pozitívabban értékelték a megoldott problémákat, mint a fiatalok.

Fülöp és László (2013) a traumafeldolgozást vizsgálta Trianonról szóló újságcikkek és a NarrCat eszköz segítségével. 254 újságcikkből állt a korpusz, 1920 és 1950, valamit 1990 és 2010 közötti, ötéves intervallumokból választva. Az újságokat politikai csoportokba, bal-, közép- és jobboldali kategóriába sorolták. A NarrCat érzelem modulját használták az elemzéshez, a szövegekben megjelenő érzelmi kifejezések számát vizsgálták. Azt találták, hogy 1930 és 1940 között jelentősen csökkent, majd 1990-től kezdve enyhén emelkedő tendenciát mutat a mérőszám, de még ez a szint is alacsonynak mondható. Egy kezdeti tagadó szakaszt a kommunizmusban tabu követett, majd a rendszerváltás után elkezdett enyhén emelkedni az érzelemkifejezés, főleg a jobboldali médiában, azonban inkább ezek is inkább negatív érzelmek. Cikkükben bemutatnak egy traumafeldolgozási modellt, ami arra szolgál, hogy megvizsgáljuk valakinek az írása és viselkedése alapján, hogy mennyire dolgozott fel egy traumatikus eseményt. A gyenge traumafeldolgozásnak számos jele lehet:

- Gyakori, intenzív foglalkozás a témával
- Inkoherens, széttöredező narratíva
- Differenciálás hiánya
- Ragaszkodás kognitív és érzelmi sémákhoz
- A trauma újraélése jelenidő és indulatszavak használatával
- Szelf-fókusz, képtelenség a perspektívaváltásra
- Az események egyszerű magyarázata, komplexitás mellőzése
- Érzelmi és értékelő kifejezések magas száma, extrém szavak használata
- Kognitív kifejezések alacsony száma
- Tagadás
- Negatív szándékok projektálása külső csoportra, ellenségmegjelenítés
- Alacsony ágencia és kontroll

Összességében arra következtetnek a szerzők, hogy a veszteség tagadása és fel nem dolgozott érzelmek a jellemzők 2010-ig.

Vincze, Ilg és Pólya (2014) három különböző Trianonnal kapcsolatos korpusz vizsgáltak. Elemeztek 22, 1920 és 2000 között forgalomban lévő tankönyvet, 500 fős magánemberekből álló minta leírását az eseményről, valamint 354, 1920 és 2010 között megjelent újságcikket. A NarrCat kogníció modulját használták az elemzéshez. A szövegekben megjelenő kognitív

kifejezések, érzelmekkel kapcsolatos szavak mennyiségéből lehet következtetni arra, hogy az elbeszélő milyen mértékben dolgozott fel egy negatív eseményt. A traumát követően az érzelmi kifejezések magas száma jelenik meg általában, majd az érzelmi kifejezések helyét a kogníciós kifejezések veszik át, amikor a feldolgozás, elfogadás szakaszba lép a narrátor. Fülöp és László (2013) kutatásához hasonló eredményt kaptak mindhárom korpusz alapján, ugyanis a kogníció mérsékelt szintje magas érzelemkifejezéssel társult a szövegekben, amiből arra következtethetünk, hogy a magyarok emlékezetében Trianon traumája még közel sincs feldolgozva.

2.2. További kutatások Trianon emlékezetpolitikájával kapcsolatosan

A trianoni békeszerződés a szövetséges hatalmak és Magyarország között köttetett az I. világháború végén 1920. július 4-én. Ez az esemény traumatikus mérföldkőnek számít a nemzeti történelemben, hiszen a megállapodás mintegy kétharmaddal csökkentette Magyarország területét és lakosságát. (Vincze et al., 2014) Emiatt számos történész, közgazdász, pszichológus, szociológus foglalkozik a mai napig ennek az eseménynek a hatásával. A kutatások jelentős része írásos anyagokon keresztül vizsgálja Trianon keretezését és a trauma feldolgozását, ezek közül mutatok be most néhányat.

Zahorán (2013) Trianon közbeszédben való megjelenését vizsgálta, és a traumafeldolgozás gátjait mutatta be cikkében. Kiemeli, hogy a 2010-es évektől kezdve a kormány kommunikációjának meghatározó eleme az összefogás, szolidaritás, valamint a keserves múlt és önsajnálat helyett a közös jövőre való irányultság. A szerző úgy vélekedik, hogy a politikai elitnek visszafogottabban kellene kezelnie Trianon kérdéskörét, hogy létrejöjjön egy egységes nézőpont, és jobb viszony alakulhasson ki a szomszédos országokkal.

Kovács (2015) Trianon és holokauszt keretezését mutatja be az arról szóló tanulmányok és egy régi történelemkönyv segítségével. A Trianon mint történelmi trauma és a holokauszt-trauma kapcsolatát vizsgálta, és azt találta kutatása során, hogy ugyan a Trianon-kultusznak része volt az etnocentrizmus és antiszemitizmus, a történetírás mégis elválasztani igyekszik egymástól Trianon és a Holokauszt traumáját. Sőt, Trianont kétségkívül nagyobb traumának tartják, amit aggályokkal szemlél a cikk írója, ugyanis nem a gyógyulás, hanem a revansvágy és a nacionalizmus útját látja a történészek keretezésében.

Kutatásom közvetlen előzménye Barna és Knap (2022) kutatása a Trianon és a Holokauszt online médiában való keretezéséről. Topikmodellt használtak magyar online hírportálokon megjelent cikkek elemzéséhez. Kutatásuk célja egyrészt a Trianonnal és Holokauszttal kapcsolatos diskurzus és megjelenő tematikák azonosítása, valamint különbségek felfedezése a különböző politikai oldalak retorikájában. Láthatjuk tehát, hogy a politikai dimenziót vizsgálták ők is a cikkek témájával kapcsolatosan, egészen pontosan azt, hogy a topikokban mely honlapok, mely politikai oldalak cikkei a dominánsak. Ehhez a topikokba legmagasabb valószínűséggel tartozó cikkek politikai hovatartozását gyűjtötték ki minden topikra. Kvalitatív elemzést a topikok mélyebb megértésére használtak, elolvasták és feldolgozták a topikokra legjellemzőbb cikkek tartalmát, témáját. A szélsőjobboldal erős irredenta, revízionista nyelvezetet használ, a kormánypárti médiában megjelent cikkek enyhébbek ugyan, de közel állnak témáikban a szélsőjobbhoz. Abban is hasonlítanak, hogy Trianon felelőseit a nemzet ellenségeinek tartják, viszont a szélsőjobbra inkább jellemző, hogy kesereg a múlton, a Fidesz ezzel ellentétben egy olyan retorikát kezdett el használni, hogy a legyőzött nemzetből hős nemzet válik. A jobb és a baloldal között a legnagyobb különbség az érzelmek használata, a baloldal elhatárolódik Trianon érzelmi oldalától, és nem alakított ki tudatos emlékezetpolitikát, saját narratívát.

2.3. Szöveganalitikai és gépi tanulási módszereket alkalmazó kutatások

Dolgozatomnak fontos pontja a szövegek klasszifikációja gépi tanulási algoritmusokkal, a politikai hovatartozást igyekszem prediktálni a NarrCat és a topikhovatartozási értékekkel. Számos kutatás született az elmúlt években, ahol hasonló módszerekkel vizsgálódtak. A következőkben olyan cikkeket mutatok be, ahol a szerzők topikmodell adatokat vagy valamilyen egyéb szövegjellemzőket használtak klasszifikációra.

Huang és szerzőtársai (2021) egy közösségi finanszírozási oldalra írt kommenteket, véleményeket elemezték. Ötfokozatú skálán értékelték, hogy az adott komment mennyire pozitív (1) vagy negatív (5), és Support Vector Machine (SVM), Random Forest (RF), Gradient-Boosted Decision Trees (GBDT), Extreme Gradient Boosting (XGBoost) gépi tanulási algoritmusokat alkalmaztak, hogy a szövegeket be tudják sorolni 1-5 kategóriába. N-gram (gyakran előforduló, egymást követő szavak együttese), Word2vec (word to vector, neurális hálók segítségével szavak kapcsolatát távolságát tanulja a modell), és Term Frequency-Inverse Document Frequency (TF-IDF, szavak fontosságát méri a korpuszban) módszerekkel kinyert

szövegjellemzők voltak a független változóik. Azért, hogy javítsák a klasszifikációt, LDA topikmodellezést hajtottak végre, kialakítva 8 különböző topikot. A legjobb modell a TF-IDF + topik változókkal létrehozott modell volt, 88,1%-os pontossággal.

Liu és munkatársainak (2016.) célja Twitteren és egy másik közösségi platformon, a Weibon megjelenő „okos spamelők" (emberi profilra nagyon hasonlító profilokkal történő spamelés) kiszűrése volt. Az egyes posztok topikeloszlásából létrehoztak két mutatót, ami arra szolgált, hogy a felhasználók érdeklődését jellemezzék vele, illetve a többi felhasználóhoz képest mérték a topikeloszlásokat posztjaikban. Azt találták, hogy a tényleges felhasználók néhány téma iránt érdeklődnek, a spam fiókok pedig vagy egy témára fókuszálnak (pl. egy termék hirdetése), vagy nagyon sok, véletlenszerű témáról posztolnak. SVM, Random Forest és AdaBoost algoritmusokat használtak a szerzők, és a topik és néhány egyéb változó segítségével (posztok száma, stb.) sikerült elérniük 95,2%-os precisiont (helyes pozitív/(helyes pozitív+hibás pozitív)) és 94,6%-os recallt (helyes pozitív/(helyes pozitív+hibás negatív)).

DeBarr, Ramanathan és Wechsler (2013.) adathalász emaileket igyekezett detektálni különböző módszerekkel. Azt találták, hogy az LDA és a Random Forest algoritmusok ötvözése hasonló eredményt hoz (96,6%-os pontosság), mint a Spectral Clustering. Ez az eredmény azonban csak akkor érvényes, ha megfelelően hosszú szövegek állnak rendelkezésre, a szerzők ezért érvelnek a Spectral Clustering használata mellett.

Edaño és szerzőtársai (2022.) potilikai témájú tweetekről igyekezett SVM, XGBoost és logisztikus regresszió segítségével megállapítani, hogy stresszt tükröznek vagy sem. Ehhez végrehajtottak egy topikmodellezést, ami után meghatározták, hogy az egyes topikokban mi számít stresszornak, pl. kormányzati politika, elnök stb. Az SVM, XGBoost, Logisztikus regresszió modellezésnél a függő változó a topikhovatartozás volt, a független változók pedig a TF-IDF módszerrel kinyert szövegjellemzők. A három modell nagyon hasonló, 70% körüli pontosságot eredményezett.

3. Módszertan

3.1. NarrCat

A pszichológiai tartalomelemzés lényege, hogy az elbeszélők írásos vagy szóbeli szövegeiben akaratlanul is kódokat helyeznek el, amelyeket megfejtve következtetéseket vonhatunk le

róluk. (Osgood és Walker, 1959) Az általam használt NarrCat eszköz is a narratív pszichológiára épít, azaz írásos anyagok alapján, a szöveg narratíváját vizsgálva tudunk a szöveg írójáról következtetéseket levonni. A NarrCat egy tartalomelemző szoftver, amely egy NooJ nevű, nemzetközi, többnyelvű platformon működik, magyar nyelvű használatát a több mint 187 millió szóból álló magyar nemzeti korpusz beépítése teszi lehetővé. Az eszköz a bemenetként megadott szövegekhez különböző score-okat rendel nyelvi és pszichológiai kategóriákban úgy, hogy gépi transzformációt végez a mondatokon. Használata abban segíti a kutatót, hogy számszerűsíti bizonyos formák gyakoriságát a szövegekben, lehetővé téve így a szöveg narratívájának kvantitatív elemzését. Ezek a formák lehetnek nyelvtani szerkezetek, szavak, szókapcsolatok, szavak végződése. Egy példa a szöveg átalakítására: „A magyarok féltek a mongoloktól." -> „Saját csoport ágensének alapvető negatív érzelme a külső csoport felé a múltban." (László et al., 2013) Itt a nyelvi kategóriára példa a múlt idő azonosítása, a pszichológiai kategóriára pedig a negatív érzelem detektálása. Az eszköz kimenete azonban nem az átalakított mondatok, hanem értékek, amelyek azt mutatják meg, hogy az egyes narratív kategóriákból, formákból mennyi található meg a szövegben. Az eszköz tehát sokkal komplexebb, mint egy bag of words modell, fontos a szavak sorrendje is, ugyanis vizsgál szintaktikai viszonyokat is, illetve a végződések is jelentős szerepet kapnak a formák azonosításánál.

Kutatásom során az értékek normáltját használtam, ami az előfordulás és a szöveg szószámának hányadosa.

A NarrCat által keresett formák a következők:

- Ágencia
    - Aktivitás
    - Passzivitás
    - Kényszer
    - Intenció
- Társas referencia
    - Én referencia
    - Mi referencia
- Pszichológiai perspektíva
- Kogníció
- Érzelem
    - Pozitív
    - Negatív
- Értékelés
    - Pozitív
    - Negatív

- Tér-idő perspektíva
    - Visszatekintő forma
    - Átélő forma
    - Metanarratív forma
- Tagadás

Ágencia

Az ágenciát Szalai (2011) úgy definiálja, mint a hatékony cselekvés képessége. Az ágencia modul két almodulra bontható, aktivitásra és intencióra. Az aktivitás modulban az aktív és passzív igék arányát azonosítja az algoritmus, az aktív és passzív szerkezetre a „befejezi a kapcsolatot" és a „befejeződik a kapcsolata" mondatokat hozza példának Szalai (2011). Minél magasabb az aktivitás/passzivitás ráta, a történet szereplője annál hatékonyabb, környezetére annál nagyobb befolyással bír (Ferenczhalmy et al., 2011).

McAdams (2001) szerint az emberek intencióval rendelkező ágensek, akik a vágyaik és hiedelmeik alapján cselekednek, hogy elérjék céljaikat. Ezt a szándékosságot tetten érhetjük olyan igékkel, mint az akar, fog, kíván, de a NarrCat felismer pl. főneveket és mellékneveket is, pl. cél, terv, céltudatos. (László et al., 2013) A kényszer lehet külső és belső, és azonosítható az intenció hiányával is vagy a maga a kényszerítés, kényszerítettség jelenlétével, pl. „Nem akartam idejönni", „Muszáj volt vele beszélnem." (Narratív Pszichológia PTE, Ágencia modul, 2022)

Társas referencia

A társas referencia modul két mérőszámot ad vissza, az én és a mi referenciák arányát a szövegekben. A NarrCat azonosítja a személyes, birtokos és visszaható névmásokat (engem, miénk, magam), elemzi az igevégződéseket (csinál*juk*, csinál*nám*) és a főnevek birtokos ragozását (kalap*om*).

A mi referencia magas értékei a csoporttal való magas fokú azonosulásra és élményközösségre utalnak. Következtethetünk belőle a csoportkohézió mértékére is. Az én referenciák magas száma a személyes érintettséget és bevonódást jelzi a témával kapcsolatban. (Narratív Pszichológia PTE, Társas referencia és tagadás modul, 2022)

Pszichológiai perspektíva

A pszichológiai perspektíva akkor jelenik meg, ha az eseményeket egy karakter szemszögéből látjuk, az ő mentális állapotára történő hivatkozásokkal. Ez gyakran kognitív és érzelmi

folyamatok bemutatásával történik, pl. a „gondol” vagy „érez” igék használatával. (Vincze et al., 2013) A NarrCatban a pszichológiai perspektíva modul a kogníció, az érzelem és az ágencia-intenció modulokra épül, és a fentieknek megfelelően olyan szavakat és kifejezéseket azonosít, amik a szöveg szereplőinek belső mentális állapotát írják le. (László et al., 2013)

Kogníció

A mentális igék, pl. „általánosít”, „töpreng”, főnevek, pl. „gondolat”, „döntés”, „ötlet” és egyéb, mentális cselekvésre utaló kifejezések, pl. „felfrissíti az emlékezetét”, szövegben való előfordulásának számából áll össze a kogníciós pontszám. Azért érdemes összegezni és összehasonlítani a szövegekben való előfordulásukat, mert egy karakter mentális folyamatainak ábrázolása, (pl., mit gondol, mit hisz, miről álmodik), elősegíti a karakter iránti kognitív empátiát, könnyebben azonosulunk vele, sőt, ennek az a következménye, hogy a hibáit is könnyebben megbocsátjuk. (László et al., 2013) A mentalizáció mások mentális állapotának monitorozását jelenti, amely többek között segít az egocentrizmus, előítéletek és szociális agresszió csökkenésében, segítségnyújtás növelésében. A magas kogníciós pontszám tehát azt jelenti, hogy a szöveg írója nagy hangsúlyt fektet szereplőjének mentális állapotára, elősegítve a vele való azonosulást. (Narratív Pszichológia PTE, Kogníció modul, 2022)

Érzelem

A NarrCat három csoportba sorolja az érzelmet kifejező szavakat egy kutatók által létrehozott érzelemszótár segítségével. Pozitív szó például az „öröm”, „bizalom”, negatív a „félelem”, „szenvedés”, semleges a „meglepődés”. A pozitív és a negatív szavak teljes szószámhoz viszonyított arányát kapjuk meg az eszköz outputjaként. Ezen mérőszámok képet adhatnak arról, hogy az elbeszélő mennyire optimista, nyitott, milyen önértékeléssel rendelkezik. (László és Fülöp, 2010)

Értékelés

Az értékelés modulban szintén pozitív és negatív kategóriákban közöl pontszámokat a NarrCat, a narrátor vagy egy szereplő értékítéletét azonosítja a szótárában lévő főnevek, melléknevek, igék, határozószavak segítségével. Pozitív értékelés az elismerés, negatív a kritika, a tulajdonságok, érzelmi reakciók és az értékelő interpretációk pedig lehetnek pozitívak és negatívak is. (László et al., 2013)

Tér-idő perspektíva

Ez a modul 3 almodulra bontható, ennek megfelelően 3 pontszámot kapunk vissza kimenetként az eszköztől. A NarrCat azonosítja a mondatok formáját, ami lehet visszatekintő, átélő vagy metanarratív formátumú. Pólya és szerzőtársai (2007) egy balesettel kapcsolatos példát használnak, amint keresztül bemutatják a 3 formát:

„Az élelmiszerbolt felé vettem az utat." - retrospektív
„Látok egy autót felém tartani!": - átélő
„És nem emlékszem többre, még most sem." – metanarratív

A retrospektív forma jellemzője a múlt idő, az átélő formát folyamatos jelenidőből, térbeli határozóból és a felkiáltójelből detektálja az algoritmus, a metanarratív szerkezetet pedig a jelenidőből, az emlékszem szóból és az időhatározóból. A különböző perspektívájú szövegek különböző hatással vannak ránk még akkor is, ha ugyanaz az elbeszélés témája, egy kutatás során Pólya és szerzőtársai (2005) például azt találták, hogy a retrospektív formát használó narrátorokat alkalmazkodóképesebbnek, társadalmilag kívánatosabbnak ítélték az olvasók, mint az átélő formát használókat, akiket szorongóbbnak, de dinamikusabbnak látják. Fülöp és László (2013) pedig arról ír, hogy a jelen idő, azaz az átélő forma használata a témával való aktív elfoglaltságot, az elbeszélő aktív bevonódását mutatja.

Tagadás

A tagadás pontszám az explicit és implicit nyelvi jegyek számából tevődik össze. Explicit formák a névmások (pl. semmi, senki), határozók (pl. sehol, sosem), tagadó létigék (nincs, sincs), implicit jegyek a névutók (nélkül, kívül) és fosztóképzők (-talan, -telenül, -mentes). (Narratív Pszichológia PTE, Társas referencia és tagadás modul, 2022) Hargitai és szerzőtársai (2007) azt találták, hogy a tagadó formák gyakori használata a világ és az önmagunk leértékelését jelzi.

Sajnos arról nem elérhető információ, adat, hogy a NarrCatot hogyan validálták, és arról sem, hogy mennyire pontos az eszköz, nem tudjuk, hogy megtalál-e minden kifejezést, szót, nyelvtani formát egy adott kategóriában. Nem tudom továbbá, hogy az elütéseket hogyan kezeli az eszköz, szerencsére újságcikkeknél ez nem gyakori jelenség. Jelen szakdolgozatban nekem sem volt arra lehetőségem, kapacitásom, hogy vizsgáljam a teljesítményt, a modelleredményeket adottságnak kezeltem, pedig az eredmények értelmezéséhez és a pontosság megállapításához ez egy fontos információ lenne. Én, ha lenne lehetőségem rá, akkor úgy validálnám az eszköz teljesítményét, hogy először is kellően részletes leírásokat

készítenék humán kódolóknak arról, hogy melyik dimenzió mit jelent, és milyen szavak, formák, nyelvtani szerkezetek tartoznak egy adott kategóriához. Különböző témájú és stílusú szövegen végeztetném több annotátorral a kódolást, majd összehasonlítanám az emberi eredményeket a NarrCat eszköz eredményével. Hibák felmerülése esetén az eszközt az eredmények mentén tovább lehet javítani.

Összességében elmondhatjuk, hogy a NarrCat azáltal nyújt többet a hagyományos természetes nyelvfeldolgozási technológiáknál, hogy számos nyelvi és pszichológiai dimenzióban vizsgálja a szöveget, aminek a segítségével következtetéseket vonhatunk le a szöveg írójának, megjelentetőjének érzelmi és mentális állapotáról. Az általam használt topikmodellezés a szöveg témájának elemzéséhez kiváló, de a NarrCat segítségével ezt a témaelemzést kiegészíthetjük a szöveg érzelmi, pszichológiai, stílusbeli elemzésével is. Az olyan szövegek vizsgálatához kiemelkedően jó választás a NarrCat, ahol egy vitatott, kényes, ellentmondásos kérdésről, témáról ír a szerző, ugyanis ezekben a kérdésekben hasznos lehet, ha mögöttes információt is meg tudunk róla szerezni érzelmi, mentális állapotát illetően.

3.2. Topikmodellezés

A valószínűségi topikmodellezés célja, hogy nagyszámú dokumentum tematikáját feltárja és felcímkézze a szövegben előforduló szavak statisztikai elemzésével. (Blei, 2012) A téma az együtt előforduló szavak visszatérő mintája, egy topik tehát gyakran együtt előforduló szavakból áll. Topikmodellezés során az algoritmus összeköti az azonos kontextusból származó szavakat, és elkülöníti az egymástól külön előfordulókat. (Barde és Bainwad, 2017)

Jelen kutatás során a modellezés Látens Dirichlet Allokációval (LDA) történt, ami egy nem felügyelt statisztikai tanulási Bayesi modell. A módszer alapötlete, hogy a dokumentumok látens topikok keverékei, és a topikeloszlás a dokumentumokban Dirichlet-eloszlást követ. Ezen kívül a topikok pedig jellemezhetők a bennük előforduló szavakkal. (Blei et al., 2003)

Blei (2012) leírja, hogy az LDA egy generatív modell, ami azt jelenti, hogy az algoritmus megpróbálja meghatározni azt a háttérben meghúzódó mechanizmust, ami a szövegeket és azok témáját generálja. Az LDA generatív folyamata a változók alábbi együttes eloszlásával írható fel:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \, (\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}))$$

A képletben az első produktum a K db topik szavakon vett Dirichlet-eloszlását adja meg, a második produktum a D db dokumentum topikokon vett Dirichlet-eloszlását. A harmadik produktum pedig két valószínűségnek a szorzatából tevődik össze, mégpedig annak a valószínűségnek, hogy egy z topik megjelenik az adott dokumentumban, illetve annak a valószínűségnek, hogy egy w szó megjelenik az adott topikban. N a szavak száma.

A modellnek két hiperparamétere van, alfa a dokumentumok topikonkénti apriori eloszlásának paramétere, béta a szavak topikonkénti aprioiri eloszlásának paramétere. Ezeket a hiperparamétereket becsüljük a topikmodellezés során. Kis alfa esetén kiválaszt az algoritmus egy domináns topikot, a többinek nagyon kicsi valószínűséget allokálva. Ha alfa=1, a topikok valószínűségi eloszlása bármilyen lehet. Ahogy 1-nél nagyobbra növeljük az alfát, a dokumentumra annál nagyobb valószínűséggel feltételezzük a topikok egyenlő keverékét. Mivel a béta a szavak topikonkénti eloszlását kontrollálja, kis béta esetén kicsi lesz a szószám a topikban, nagy béta esetén pedig nagy.

A paraméteroptimalizálás során azokat az alfa és béta paramétereket keressük, amelyek maximalizálják az adatok log likelihoodját:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(w_d|\alpha, \beta)$$

Ahol egy D korpusz $\{w_1, w_2, \ldots, w_M\}$ dokumentumból áll. (Blei et al., 2003)

Az algoritmus outputként megadja minden dokumentumra, hogy melyik topikba milyen valószínűséggel tartozik, úgy is fogalmazhatnánk, hogy milyen arányban vesznek részt a topikok a dokumentumban. Azt az információt is megkapjuk, hogy melyek a topik legjellemzőbb cikkei és szavai.

A topikmodellezést megelőzi egy részletes előfeldolgozási folyamat, melynek lépései a következők:

1. URL-ek és írásjelek eltávolítása
2. Lemmatizálás
3. Part Of Speech (POS) szűrés

4. Névelemfelismerés
5. Tri- és bigramok azonosítása
6. Stopszavazás
7. Gyakorisági szűrés

(Barna és Knap, 2022)

Az URL-ek és írásjelek eltávolítását tehát a lemmatizálás, azaz szótövesítés követi, ami azt jelenti, hogy a szavakat szótári alakjukkal helyettesítjük (pl. királlyal – király). Ez azért fontos, mert különben ugyanazt a szót több alakja miatt különböző szavakként kezelné az algoritmus. Ezután a POS szűrés során eltávolíthatjuk az elemezni nem kívánt szófajokat, majd a névelemfelismerés során tulajdonneveket azonosít és hoz egységes formára az algoritmus (pl. SZTE – Szegedi Tudományegyetem). A topikmodellezés a Bag of Word megközelítésen alapszik, azaz a szavak előfordulásának sorrendje nem számít. Azonban az előfeldolgozási folyamat fontos része, hogy az összetartozó, egy kifejezést alkotó szavakat megtaláljuk, és együtt kezeljük majd a modellben (pl. kordában tart – kordában_tart). Ezeket a gyakran együtt előforduló szavakat ngramnak nevezzük, például egy két szóból álló kifejezés neve bigram. A stopszavazás olyan szavakat töröl ki a szövegekből, amiknek nincsen témaértékük, pl névelők, kötőszavak. Végezetül a gyakorisági szűrésnél eltávolíthatjuk a túl gyakori és a túl ritka szavakat a dokumentumokból, ezeknek ugyanis nem lesz jelentős hozzáadott értéke a topikok kialakításakor. (Németh és Koltai, 2021)

3.3. Elemzéshez használt algoritmusok

3.3.1. Béta regresszió

A béta regresszió alapjait Ferrari és Cribari-Neto (2004) dolgozta ki. A modellnek az a fő ismertetője, hogy a függő változó folytonos és értékkészlete 0 és 1 között van. A modell feltételezése, hogy az eredményváltozó béta eloszlású. A béta eloszlás sűrűségfüggvénye:

$$f(y;\mu,\Phi) = \frac{\Gamma(\Phi)}{\Gamma(\mu\Phi)\Gamma\big((1-\mu)\Phi\big)} y^{\mu\phi-1}(1-y)^{(1-\mu)\Phi-1}, \qquad 0<y<1,$$

ahol $0<\mu<1$; $\Phi>0$ és Γ a gamma függvény.

Ha y_1, … y_n egy véletlen minta és béta eloszlást követ μ és Φ paraméterrel, a modell a következő formában írható fel:

$$g(\mu_i) = x_i^{\mathrm{T}}\beta$$

Ahol $\beta = (\beta_1, \ldots, \beta_k)^T$ egy k×1-es vektor ismeretlen k<n paraméterekkel, $x_i = (x_{i1}, \ldots, x_{ik})^T$ a k független változót tartalmazó vektor. $g(\cdot) : (0,1) \rightarrow \mathbb{R}$ szigorúan monoton növő, kétszer deriválható linkfüggvény, ami lehet például:

- logit $g(\mu) = \log(\mu/(1-\mu))$
- probit $g(\mu) = \Phi-1\ (\mu)$, ahol $\Phi(\cdot)$ a standard normális eloszlásfüggvény
- complementary log-log $g(\mu) = \log\{-\log(1-\mu)\}$
- log-log $g(\mu) = -\log\{-\log(\mu)\}$
- Cauchy $g(\mu) = \tan\{\pi(\mu - 0.5)\}$

A paraméterbecslés maximum likelihood módszerrel történik. (Ferrari és Cribari-Neto, 2004)

3.3.2. Kvadratikus diszkriminancia-analízis

A diszkriminanacia-analízissel arra a kérdésre keressük a választ, hogy a mintatér felosztására használt algoritmusok közül melyik eredményezi a legkisebb veszteséget. A minimalizálandó veszteséget az alábbi képlettel adhatjuk meg:

$$\sum_{i=1}^{k} \pi_i L_i,$$

ahol π_i a csoportba tartozás relatív gyakoriságokból számolt apriori valószínűsége, L_i pedig a csoport vesztesége. A lineáris és a kvadratikus diszkriminancia-analízis is a Bayesi klasszifikációra épít, és feltételezi, hogy a megfigyelések normál eloszlásból származnak. Ugyanakkor a kvadratikus modell a lineárissal ellentétben nem feltételezi, hogy a kovariancia-mátrixok egyenlők a csoportokban. A kvadratikus diszkriminancia-analízis (QDA) algoritmus abba a csoportba sorolja a mintaelemeket, amelyik csoportban a következő diszkrimináns függvény a legnagyobb:

$$\delta_k(x) = \log \pi_k - \frac{1}{2}\mu_k^{\mathrm{T}} \sum\nolimits_k^{-1} \mu_k + x^{\mathrm{T}} \sum\nolimits_k^{-1} \mu_k - \frac{1}{2} x^{\mathrm{T}} \sum\nolimits_k^{-1} x - \frac{1}{2} \log \left| \sum\nolimits_k \right|$$

A μ_k-t csoportátlaggal, a σ^2-et a csoportvarianciával becsüljük. Ha ismerjük a csoportba tartozás valószínűségét, akkor azt helyettesítjük be π_k-ba, ha nem ismerjük, akkor a tanulóhalmazon tapasztalt gyakoriságokat használjuk. $\sum_k$ a k-adik csoport kovarianciamátrixa.

3.3.3. Support Vector Machine

A Support Vector Machine (SVM) egy olyan, elsősorban klasszifikációra is használható algoritmus, ahol az osztályokat úgynevezett hipersíkokkal választja el egymástól az algoritmus, amely szeparáló síkokat úgy választja meg, hogy a megfigyelések hipersíktól vett távolsága a lehető legnagyobb legyen, büntetve egy bizonyos margón belüli, valamint a rossz oldalra került elemeket. A support vektorokat azok a megfigyelés-vektorok alkotják, amelyek a margón helyezkednek el, ezek hordoznak minden releváns információt, ami a döntési határokat illeti. (Karatzoglou et al., 2006)

Az SVM előnye, hogy képes lineárisan nem szeparálható megfigyelések klasszifikációjára is kernelfüggvény segítségével. Ennek lényege, hogy az eredeti, nem lineárisan szeparálható változóteret megnöveli a kernelfüggvénnyel úgy, hogy ebben a transzformált térben már szeparálhatók a megfigyelések. (James et al., 2013) Az új megfigyeléseket a meghatározott sík segítségével osztja csoportokba a modell. A klasszifikációs függvényt le tudjuk írni a következőképp:

$$f(x) = \omega\varphi(x) + b,$$

ahol $\varphi(x)$ a magasabb dimenziós változótér, ω súlyvektor, b a küszöbérték, amit a következő regularizált rizikófüggvény minimalizálásával becsülhetünk:

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L(d_i y_i) + \frac{1}{2}||\omega||^2$$

Ahol $\frac{1}{2}||\omega||^2$ a regularizációs tag, $C\frac{1}{n}\sum_{i=1}^{n} L(d_i y_i)$ az empirikus hiba, d_i a tényleges csoporthovatartozási érték, az L pedig azt mutatja meg, hogy egy elem milyen távol van a döntési határtól hibás irányban. (Fan et al., 2018)

C (cost) a hiba nemnegatív hangolható büntető paramétere, a hibás kategorizációk száma és a túl bonyolult modell között egyensúlyt, azaz a túlillesztést kontrollálja. Ha alacsony a cost, akkor a döntési határ sima, azaz nagyobb az esélye a rossz kategorizációnak. (Karatzoglou et al., 2006)

A Lagrange-multiplikátor segítségével felírhatjuk kerneles formában a klasszifikációs függvényt:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b$$

(Fan et al., 2018)

Kerneltől függően a szigma paraméter is hangolható, ami azt kontrollálja, hogy mennyi befolyása van egy tanító mintaelemnek. Minél nagyobb a szigma, annál közelebb kell esniük más mintaelemeknek ahhoz, hogy közös csoportba kerüljenek. (Karatzoglou et al., 2006)

Szakdolgozatomban az SVM modellen belül 4 különböző kernelfüggvény használatát hasonlítom össze:

Lineáris: $K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$

Polinomiális: $K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$

Szigmoid: $K(x_i, x_{i'}) = \tanh(\gamma x_i^T x_{i'} + r)$

Radiális: $K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$

3.3.4. Random Forest

A Random Forest algoritmus döntési fákon alapszik, aminek lényege, hogy négyzetes részekre osztja az algoritmus a mintateret, és az egyes részekben található elemek átlaga vagy módusza lesz a predikció, attól függően, hogy regressziós vagy klasszifikációs fáról beszélünk.

A Random Forest először is abban különbözik a döntési fától, hogy bagging (bootstrap aggregation) módszerrel növeszti a fákat. Ez azt jelenti, hogy a megfigyeléseinkből bootstrap mintákat vesz az algoritmus, azon illeszti a fákat, majd megadja (klasszifikáció esetén) a legtöbb szavazatot kapó csoportot mint csoportpredikciót. Ha B darab bootsrap mintánk van és $\hat{f}^{*b}(x)$ a modell, akkor a becslés:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

A bootstrap minták létrehozásakor fel nem használt megfigyelések az Out Of Bag megfigyelések, ezeken a megfigyeléseken mint tesztadatokon vett hibaaránnyal (OOB hibaarány) mérhetjük a klasszifikáció pontosságát. A Random Forest fontos ismertetőjegye, hogy a döntési fákkal ellentétben nem vesz figyelembe minden változót egy fa növesztésénél.

Alapértelmezett esetben a használt változók száma $m = \sqrt{p}$, ahol p az összes változó száma. Ennek akkor van a legnagyobb jelentősége, ha van egy vagy néhány változó, ami nagyon erősen klasszifikál, ugyanis akkor hasonlóak lesznek a fák, és ha hasonlóak, akkor korrelálnak, és egy fa hozzáadása nem fog jelentős varianciacsökkenést okozni (James et al., 2013). A felhasznált változók számát az mtry paraméterrel lehet hangolni az R általam használt randomForest csomagjában.

3.3.5. Extreme Gradient Boosting

Az XGBoost algoritmus a Random Foresthez hasonlóan döntési fákon alapul. Chen és Guestrin (2016) azzal mutatják be, hogy az XGBoost mennyire hatékony és jól teljesítő megoldás klasszifikációs és előrejelzési prolémák esetében, hogy különböző gépi tanulási és adatbányászati versenyeken, kihívások során a nyertesek között leggyakrabban használt algoritmus az XGBoost. A legfontosabb jó tulajdonsága a gyorsaság, de számos előnye van ezen kívül is, például támogatja a regularizációt, keresztvalidációt, kezeli a hiányzó értékeket, illetve nagy rugalmasságot ad, hogy a felhasználó is megadhat célfüggvényeket. Az XGBoost nevében is hordozza, hogy egy boosting algoritmus, ami azt takarja, hogy ha klasszifikációs problémára használjuk (mint jelen esetben is), akkor az algoritmus döntési fákat növeszt szekvenciálisan, azaz az új fáknál mindig figyelembe veszi az előző fák hibáját, a hibás klasszifikációnak nagyobb súlyt adva. Az új fát hozzáadja a modellhez, újrakalkulálja a reziduálisokat, majd kezdődik az új fa növesztése. Ezt az algoritmust lassú tanulónak hívják, mert nem azonnal illeszti az adatokat, hanem kezdeti gyengébb modellek hibájából tanul folyamatosan. (James et al., 2013)

Az XGBoost egy regularizált célfüggvényt minimalizál, ami a következőképp írható fel:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

ahol l a veszteségfüggvény, ami a tényleges érték y_i és a (t-1)-edik iteráció becslését hasonlítja össze, $f_t(x_i)$ a predikciós függvény. Ω a regularizációs tag,

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2,$$

ahol ω a score az egyes leveleken, λ a regularizációs paraméter, γ a minimum veszteség, ami a levélcsomópont vágásához szükséges. (Chen et al., 2016)

Az XGBoost modellnek számos paramétere van, ezek általános, booster és tanuló kategóriákba oszthatók. A megfelelő modell megtalálásához a paraméterek egy részét hangolni szükséges. (Hackerearth, 2022)

Általános:

1. booster: A boosting típusát adhatjuk meg, fa, lineáris vagy dart értéket vehet fel. Jelen esetben a fa típust alkalmazom (gbtree)
2. nthread: Párhuzamos számítást lehet vele szabályozni, default értéke szerint a maximálisan elérhető magokat használja.
3. silent: Ha 0-ra van állítva, nem kapunk a futás közben üzeneteket, ha 1-re, akkor igen.

Booster:

1. nrounds: Az iterációk maximális száma.
2. eta: Tanulóráta 0 és 1 közötti értékkel. Alacsony eta esetén lassabb a számítás, lassan tanul rá mintákra az algoritmus.
3. gamma: Regularizáció mértéke, minél nagyobb a gamma, annál nagyobb a regularizáció.
4. max_depth: A növesztett fák mélységét adja meg, minél nagyobb a szám, annál komplexebb a modell, és annál nagyobb az esély a túlillesztésre.
5. min_child_weight: Megállítja a fák növesztését, ha a súlyozott levelek/csomópontok összege kisebb, mint a megadott érték.
6. subsample: A fák növesztéséhez felhasznált megfigyelések arányát adhatjuk meg itt.
7. colsample_bytree: A fák növesztéséhez felhasznált változók arányát adhatjuk meg itt.
8. lambda: Ridge regularizációt kontrolláló paraméter.
9. alpha: Lasso regularizációt kontrolláló paraméter.

Tanuló:

1. objective: itt választhatjuk ki a célfüggvényt, több csoportos klasszifikáció esetén a multi:softmax vagy a multi:softprob használható, a kettő között az a különbség, hogy az egyik csak a prediktált csoportot adja vissza, a másik minden csoportra a csoportba tartozási valószínűséget.
2. eval_metric: A modell pontosságát értékelő metrikát adhatjuk itt meg, én a merror-t, azaz a kettőnél több osztályba sorolási problémánál használatos (multiclass) hibát használtam, ami a hibásan kategorizált eseteknek és az összes esetnek az aránya.

4. Adatok, adattisztítás

Dolgozatomban online hírportálokról gyűjtött újságcikkeket vizsgálok. A korpusz a Sentione social listening platformról származik, ami egy olyan oldal, ami több más elemző és automatizációs funkciója mellett cikkeket, posztokat, kommenteket gyűjt weboldalakról. A Sentione oldalról minden újságcikk le lett töltve 2017. szeptembere és 2020. szeptembere között, ami tartalmazta a „Trianon" szót vagy annak ragozott formáját. Ezt követően kiválasztottam azokat az oldalakat, amelyek országos szintűek, és egyértelműen meghatározható a politikai hovatartozásuk. Ebben a kategorizációban segítségemre volt Szabó és Bene (2016), akik hálózatelemzéssel vizsgálták a magyar média szerkezetét. Ott a politikai kategorizációt a két szerző és egy harmadik szakértő végezte, baloldali/liberális ellenzéki, kormánypárti, radikális jobboldali, semleges/nem besorolható kategóriákat kialakítva. Munkám során támaszkodtam még az Átló (2022) kormánypárti médiát bemutató cikkére és Janky és szerzőtársainak (2019) besorolására. Elsősorban azonban Barna és Knap (2022) cikkét vettem alapul, ők balliberális, kormánypárti, szélsőjobb, nem kormánypárti jobboldali, bulvár és vegyes kategóriákat hoztak létre. Mivel csak azokat a cikkeket vizsgálom, amelyek politikailag egyértelműen besorolható oldalakhoz tartoznak, és a fenti besorolásból csak a kormánypárti, balliberális, szélsőjobb, nem kormánypárti jobboldali kategóriákat használom elemzésem során. Csak az országos szintű lapokat használtam elemzésem során, mert a megyei szintűek mind a kormányhoz köthetők, és nagy mértékben duplikált a rajtuk megjelenő tartalom. Egy jelentős kormánypárti oldalt, a Mandinert is ki kellett hagynom, mert az ott megjelent újságcikkek nem feldolgozható formában jöttek le a Sentione oldalról. Szükség volt a cikkek tisztítására, előfordult, hogy HTML tagekkel együtt szedte le a Sentione a szöveget, vagy nem csak a törzsszöveget tartalmazta a fájl, illetve szükség volt a dátum és szavak alapján meghatározott duplikátumok törlésére is. Ez azt jelentette, hogy 13 391 cikkel kezdtem meg az előfeldolgozást, amik 36 különböző honlapról származnak.

Politikai irányultság	Hírportálok	Újságcikkek száma
Kormánypárti	888.hu, borsonline.hu, demokrata.hu, echotv.hu, figyelo.hu, hirado.hu, hirtv.hu, lokal.hu, magyarhirlap.hu, magyaridok.hu, magyarnemzet.hu, origo.hu, pestisracok.hu, ripost.hu, vadhajtasok.hu, vg.hu	6225

Nem kormánypárti jobboldali	alfahir.hu, magyarhang.org, valasz.hu	461
Balliberális	168ora.hu, 24.hu, 444.hu, hvg.hu, index.hu, klubradio.hu, magyarnarancs.hu, merce.hu, nepszava.hu, nyugat.hu, qubit.hu, szombat.org	2204
Szélsőjobboldali	elemi.hu, hunhir.info, kuruc.info, nemzeti.net, vilagfigyelo.com	4501

1. táblázat. A vizsgált oldalak politikai besorolása és a cikkek száma. Forrás: Saját szerkesztés.

Érdekes megfigyelni, hogy milyen különbözők a cikkszámok az egyes politikai oldalak esetében, látszik például, hogy a szélsőjobboldali média kedvelt témája a Trianon.

5. NarrCat és topikmodell leíró eredmények

5.1. NarrCat

A NarrCat eszköz használatához szükség volt egy alap szintű adattisztításra: duplikátumok törlése, html tagek és nem a törzsszöveghez tartozó szavak, mondatok törlése. Előfeldolgozást azonban nem kellett végezni a szövegeken, ugyanis a NarrCat-nak van egy beépített előfeldolgozója. Az adattisztítást követően el is lehetett indítani az elemzés futtatását. Az eszköz online elérhető verziója ekkora korpusz elemzésére nem alkalmas, ezért a futtatást Pólya Tibor, az Eötvös Loránd Kutatási Hálózat Kognitív Idegtudományi és Pszichológiai Intézetének kutatója végezte el, akinek van teljes hozzáférése a NarrCathoz. Egy több hetes futási idő után megkaptam az eredményeket dokumentumonként excelekben. Az output fájlok tartalmát egy nagy táblába egyesítettem, aminek sorai a cikkek, oszlopai a NarrCat értékek voltak.

Az alábbi tábálatban láthatók a NarrCat dimenziók deskriptív adatai: az átlag, a medián, a szórás és a terjedelem. A minimum mindenhol 0 volt, ami azt jelenti, hogy volt olyan szöveg, ahol az adott kategóriájú kifejezésből nem volt megtalálható egy sem. A maximum két esetben, az átélő és a metanarratív formánál volt a legmagasabb, 1, ez azért fordulhatott elő, mert az átélő, metanarratív és visszatekintő mérőszámok az igék számával való osztással jönnek létre, azaz az összegük 1. Ezzel szemben a többi dimenzióban a szöveg szószámával osztunk az arányszám létrehozásakor. A legalacsonyabb maximum a pozitív érzelem esetében jelent meg. Ha a narratív formákat nem nézzük, a legmagasabb átlagos értékkel az aktív ige, a legalacsonyabbal a pozitív és a negatív érzelmek rendelkeznek. Érdekes megfigyelni, hogy a

pozitív és negatív érzelem dimenzióban az átlag egyenló, de a negatív érzelmek arányszám maximuma sokkal magasabb, mint a pozitívé. Az, hogy a mi referencia átlagos értéke kétszer akkora, mint az én referenciáé, mutatja, hogy a Trianonhoz kötődő témákban a csoport sokkal fontosabb, mint az egyén.

	Aktív ige	Passzív ige	Kényszer	Intenció	Én referencia	Mi referencia	Pszichológia	Kogníció
Átlag	0,032	0,006	0,004	0,005	0,006	0,015	0,011	0,005
Mediá n	0,031	0,005	0,003	0,004	0,002	0,010	0,010	0,005
Szórás	0,013	0,005	0,005	0,005	0,011	0,016	0,009	0,005
Min	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Max	0,125	0,083	0,077	0,074	0,143	0,241	0,133	0,100

2. táblázat. NarrCat modulok leíró adatai I. Forrás: Saját szerkesztés.

	Érzelem - pozitív	Érzelem - negatív	Értékelés - pozitív	Értékelés - negatív	Átélő forma	Metanarratí v forma	Visszatekint ő forma	Tagadás
Átlag	0,003	0,003	0,005	0,005	0,524	0,126	0,350	0,021
Mediá n	0,002	0,001	0,004	0,003	0,532	0,115	0,340	0,020
Szórás	0,004	0,004	0,005	0,006	0,132	0,079	0,134	0,014
Min	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Max	0,056	0,077	0,103	0,105	1,000	0,833	1,000	0,261

3. táblázat. NarrCat modulok leíró adatai II. Forrás: Saját szerkesztés.

5.1.1. NarrCat és politikai hovatartozás

A következőkben megvizsgáltam az átlagos NarrCat értékeket politikai oldalanként, kerestem, hogy milyen különbségek vannak a politikailag különböző hírportálokon megjelent cikkek NarrCat pontszámait tekintve. Az alábbi ábrán NarrCat modulonként és politikai oldalanként jelenik meg az átlagos NarrCat pontszám, a magas, átlag feletti értékek sárga színnel, az alacsony, átlag alatti értékek bordó színnel láthatók.

NarrCat	Szélsőjobb	Baloldali-liberális	Kormánypárti	Nem kormány-párti jobboldali	Sokasági átlag
Aktív ige	0,0305	0,0317	0,0338	0,0309	0,0322
Passzív ige	0,0054	0,0063	0,0056	0,0056	0,0057
Kényszer	0,0036	0,0039	0,0039	0,0040	0,0038
Intenció	0,0048	0,0048	0,0047	0,0049	0,0048
Én referencia	0,0066	0,0058	0,0056	0,0075	0,0060
Mi referencia	0,0168	0,0111	0,0144	0,0142	0,0146
Kogníció	0,0049	0,0054	0,0052	0,0060	0,0052
Pozitív érzelem	0,0035	0,0030	0,0031	0,0032	0,0032
Negatív érzelem	0,0028	0,0026	0,0024	0,0030	0,0026
Pozitív értékelés	0,0054	0,0048	0,0049	0,0048	0,0051
Negatív értékelés	0,0047	0,0051	0,0047	0,0054	0,0048
Átélő forma	0,5224	0,5433	0,5154	0,5497	0,5240
Metanarratív forma	0,1386	0,1185	0,1209	0,1173	0,1261
Visszatekintő forma	0,3390	0,3382	0,3636	0,3330	0,3498
Tagadás	0,0209	0,0245	0,0198	0,0268	0,0213

4. táblázat. Átlagos NarrCat értékek politikai oldalanként Forrás: Saját szerkesztés.

Az aktív igék használata a kormánypárti médiára legjellemzőbb, a passzívaké a balliberálisra. A pozitív érzelmi, és a pozitív értékelő formák használata csak a szélsőjobboldali cikkeknél átlagon felüli, ellenben a kormánypárti médiában inkább a negatív érzelmi és értékelő kifejezések dominálnak. A kogníciós pontszám csak a szélsőjobbnál átlagon aluli. Ezt egyben értelmezve a magas mi referenciák számával elmondhatjuk, hogy a szélsőjobboldali írásokban gyakran megjelenik a belső csoport, valamint kifelé kevésbé empatikusak, kevesebb a mentalizációra és a gondolatiságra utaló kifejezés. Fülöp és László (2013) már említett traumafeldolgozási modellje szerint ezekből a jelekből arra következtethetünk, hogy a szélsőjobboldal még nem dolgozta fel ezt a traumát. Erre utal a cikkek magas száma a szélsőjobboldali médiában is, nagyon sokat foglalkoznak a témával. Ezzel szemben a baloldali sajtó kommunikációjában magas kogníciós értékek és alacsony én- és mi-referencia jelenik meg, ami azt jelenti, hogy jellemző rájuk a mentalizációs képesség, valamint írásaikban nem a szelf van előtérben. Ebből az éntől és a saját csoporttól való elhatárolódásra, de kifelé empatikus magatartásra következtethetünk, tehát a traumafeldolgozás ezen csoport esetében egy sokkal előrehaladottabb szakaszban van. A kényszerre utaló szavak száma csak a szélsőjobboldal esetében átlag alatti. A kormánypárti szövegekben viszont az intenció pontszám alacsony, tehát a szélsőjobboldalra az jellemző, hogy a tudatosság jelenik meg írásaikban, a kormánypárti oldalakon pedig olyan cikkeket írnak, ahol több az elszenvedésre utaló kifejezés. A tagadó kifejezések a baloldalra és a nem kormánypárti jobboldalra jellemzők legjobban. Ami az elbeszélésmódot illeti, az átélő formát a bal és a nem kormánypárti jobboldal használja gyakran, a metanarratívot a szélsőjobb, a visszatekintőt a kormánypárti. Érdekes megfigyelni, hogy a bal- és a nem kormánypárti jobboldal NarrCat értékei gyakran

mozognak együtt, pl. magas kogníciós, negatív értékelő, átélő, tagadó pontszám, kevés aktív ige, pozitív érzelem és értékelés jellemző. A szélsőjobb narratíva-pontszámai különülnek el legjobban a másik háromtól, itt gyakoriak a szélsőségesebb értékek magas és alacsony irányban egyaránt.

5.2. Topikmodellezés

5.2.1. Előfeldolgozás

Láthattuk, hogy a NarrCatnak beépített előfeldolgozója van, ezért az adattisztítás után lehetett is futtatni a cikkek elemzését. A topikmodellezést azonban megelőzi egy komplex előkészítési folyamat, amire a topikmodellezés bemutatásánál tértem ki. Az előfeldolgozást Pythonban végeztem, a HuSpaCy könyvtár segítségével. Ennek során töröltem az URL-eket, futtattam egy szótövesítő algoritmust a cikkekre. A POS szűrés során eltávolítottam az írásjeleket és a mellékneveken, határozószavakon, számokon, főneveken és tulajdonneveken kívül minden egyéb szófajú szót, azoknak ugyanis nincs elegendő információértéke esetünkben a topikmodellezés során. A névelemfelismerés a DBpedia Spotlight segítségével történt, melynek során egységesítettem a tulajdonneveket. Ezután kimentettem a leggyakoribb trigramokat és bigramokat, azaz a leggyakrabban együtt előforduló szavakat. Ezeket a listákat validálni kell, manuálisan kell meghatározni, hogy melyik tri- és bigramoknál történjen meg a szavak összevonása, ugyanis sok esetben valójában nem összetartozó szavak követik egymást gyakran, és ezek összevonása nem szükséges. A leggyakoribb ngramok lekérése és összevonása folyamatot többször elvégeztem, hogy az ngramok minél szélesebb körét tudjam azonosítani. A stopszavazáshoz a Spacy beépített stopszólistáját használtam, a gyakorisági szűrés során pedig azokat a szavakat távolítottam el, amik háromnál kevesebbszer fordultak elő a korpuszban.

Az első modellfuttatások után, amikor az algoritmus meghatározta a topikokra jellemző leggyakoribb szavakat, vissza kellett térnem új stopszavakat, ngramokat meghatározni, mivel előfordult, hogy túl gyakori jelentéktelen szavak, vagy önmagukban jelentéktelen szavak jelentek meg. Ezt azért érdemes elvégezni, hogy javítani tudjuk a modell teljesítményét. Ekkor derült ki az is, hogy a korpusz majdnem-duplikátumokat tartalmaz. A 100%-os duplikátumok az adattisztítás során ki lettek szűrve, itt nem szószerinti volt az egyezés, de gyakran csak a cím vagy egy-egy szó különbözött a cikkekben. Valószínűsíthetően vagy az MTI-től átvett, egy az egyben leközölt hírekről, vagy hivatalos közleményről, beszédről szóló cikkekről van itt szó,

ezért lehet ekkora hasonlóság, akár a baloldali és a kormánypárti médiában megjelent két cikk között. A nagyfokú hasonlóság torzítja a topikmodellezés eredményét, ezért szükség volt a duplikátumok kiszűrésére kidolgozni egy módszert. Minden cikket összehasonlítottam minden cikkel, a hasonlóságukat Jaccard similarityvel mértem, ami közös szavak és az összes szó arányát adja meg. Megvizsgáltam magasabb és alacsonyabb egyezőségű cikkpárokat is, és végül azokat a dokumentumokat szűrtem ki, amelyeknél a mutató legalább 75%-os egyezést adott, mert úgy tapasztaltam, hogy ezek ugyanazok a hírek, ugyanabban a megfogalmazásban. Gyakran előfordult, hogy nem csak két oldalról származtak a kvázi-duplikátumok, hanem 6-7 portál is lehozta ugyanazt a hírt ugyanúgy. Miután meghatároztam a 75% hasonlóság feletti szövegeket, 1 132 olyan cikket kaptam, ami szerepelt legalább kétszer a korpuszban. Minden, ismétléseket tartalmazó cikkcsoportból véletlen sorsolással választottam egy cikket, amit megtartottam a korpuszban, a többit pedig töröltem. A kiinduló 13 391 cikkből a kvázi-duplikátumok törlése után 11 535 maradt, azaz 1 856 cikket kellett eltávolítani, ami az eredeti korpusz 13,8%-a. Ez a mennyiség jelentősnek mondható, a topikmodellezés eredményét torzítani tudta volna, ha nem teszem meg a törlést. A végleges korpusz eloszlása az alábbi táblázatban látható.

Politikai irányultság	Hírportálok	Újságcikkek száma
Kormánypárti	888.hu, borsonline.hu, demokrata.hu, echotv.hu, figyelo.hu, hirado.hu, hirtv.hu, lokal.hu, magyarhirlap.hu, magyaridok.hu, magyarnemzet.hu, origo.hu, pestisracok.hu, ripost.hu, vadhajtasok.hu, vg.hu	5246
Nem kormánypárti jobboldali	alfahir.hu, magyarhang.org, valasz.hu	419
Balliberális	168ora.hu, 24.hu, 444.hu, hvg.hu, index.hu, klubradio.hu, magyarnarancs.hu, merce.hu, nepszava.hu, nyugat.hu, qubit.hu, szombat.org	2108
Szélsőjobboldali	elemi.hu, hunhir.info, kuruc.info, nemzeti.net, vilagfigyelo.com	3762

5. táblázat. A vizsgált oldalak politikai besorolása és a cikkek száma – végleges modell. Forrás: Saját szerkesztés.

5.2.2. Modellezés

A modellezéshez egy nyílt forráskódú Python könyvtárat, a Gensimet használtam. Ahhoz, hogy meg tudjam határozni az optimális topikszámot, a topikmodellező algoritmust 5-24 topikszámra futtattam, minden topikszámra 5-ször. A modellek között a C_v mutató

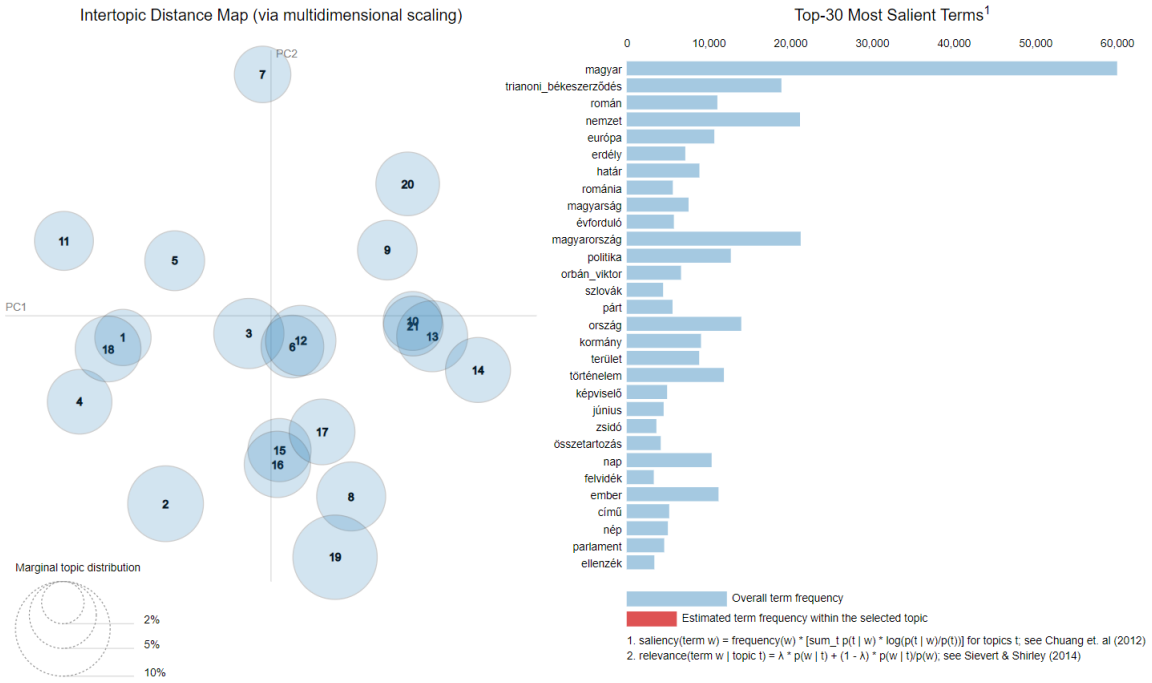
segítségével döntöttem. A C_v a topikmodell minőségét méri a szógyakorisági vektorok koszinusz távolságát felhasználva. (Röder et al., 2015) A legmagasabb C_v -vel rendelkező modellben a legnagyobb a topik-koherencia, ami a legjobb modellt jelenti. A végleges futtatáskor a 21-es topikszám rendelkezett az 5 futásra átlagosan a legmagasabb cv mutatóval (0,5025), és az összes, 100 modell közül is egy 21 topikos lett a legmagasabb koherenciájú, 0,5127. Ezen kívül megvizsgáltam a topikokat interpretálhatóság szempontjából is a rájuk legjobban jellemző szavak és cikkek segítségével. A kiválasztott 21-es topikszámú modellt jól értelmezhetőnek találtam összességében, ezért is döntöttem használata mellett.

Az alábbi ábrán megfigyelhetjük a topikokat nagyságuk szerint, két dimenzióban ábrázolva, a korpusz 30 legfontosabb szavával. Fontos kiemelni, hogy nem egyszerűen szógyakoriságot használ ez a vizualizáció, Chuang és szerzőtársai (2012) kidolgoztak egy úgynevezett saliency mutatót a szavakra a topikok jobb értelmezhetőségének érdekében. A mutató így épül fel:

$$Saliency(w) = P(w) \sum_{T} P(T|w) log \frac{P(T|w)}{P(T)}$$

Ahol w a szó, $P(T|w)$ a likelihoodja annak, hogy w szót generálta T látens topik. A szummás kifejezéssel a szerzők a szó megkülönböztethetőségét mérik, majd ezt a kifejezést megszorozzák a szó valószínűségével, azaz gyakoriságával.

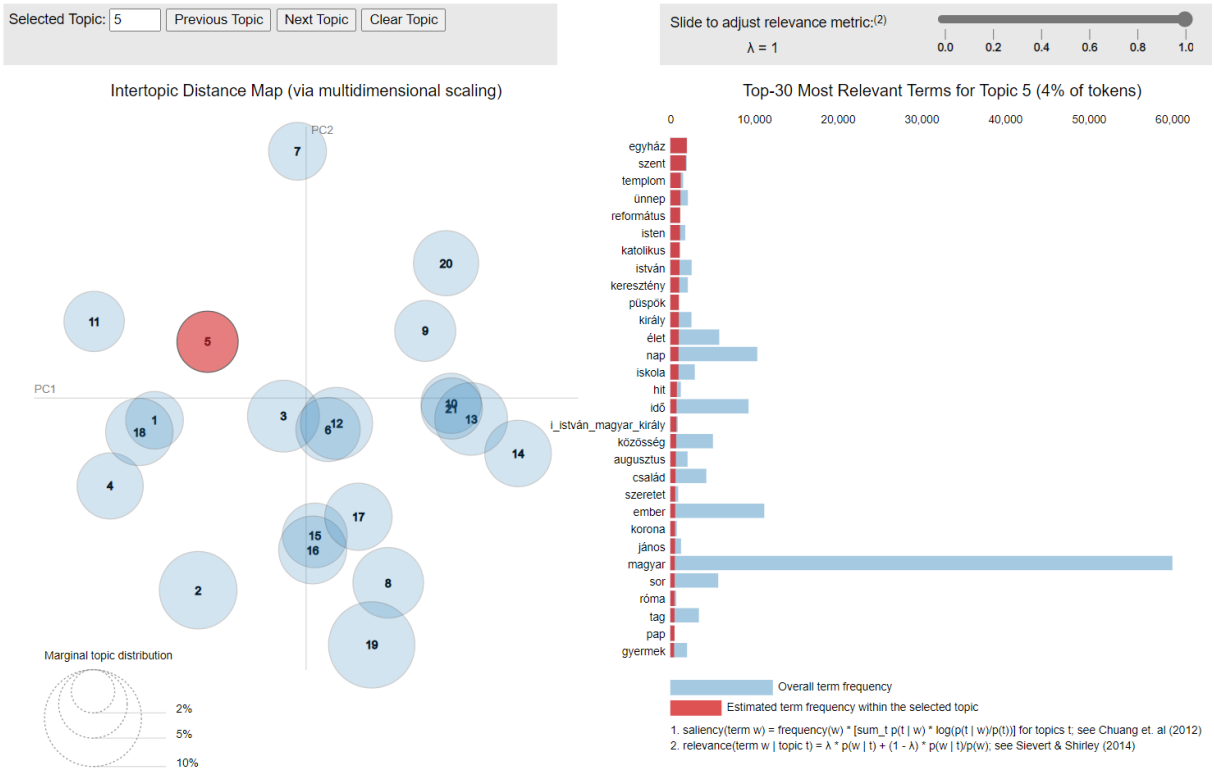
Topikok nagysága, viszonya és a korpusz legkiemelkedőbb szavai



1. ábra. Forrás: Saját szerkesztés.

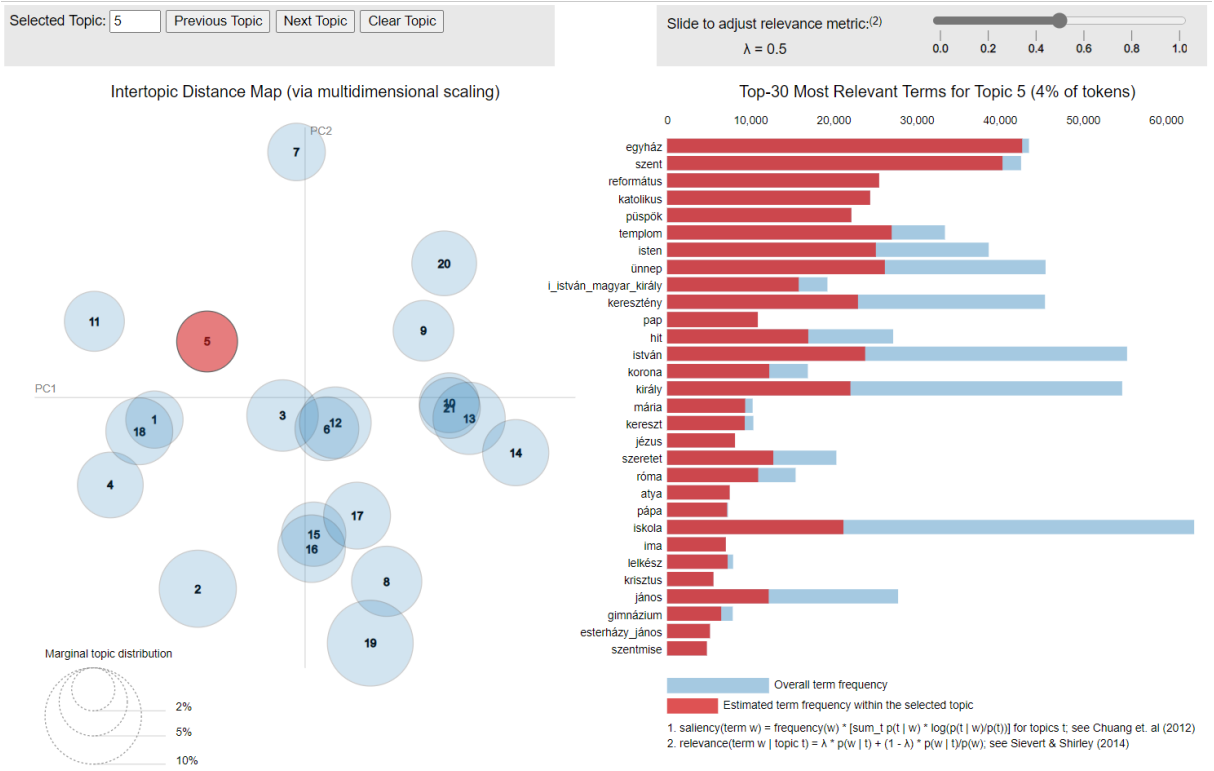
Az alábbi két ábrán láthatjuk a fenti vizualizáció változatait, amikor egy topikra nézhetjük meg a legjellemzőbb szavakat. A lambda paraméter változtatásával tudjuk állítani, hogy egy topiknál azokra a szavakra vagyunk kíváncsiak, amik ténylegesen a leggyakrabban fordulnak elő (lambda=1, első ábra), vagy a legrelevánsabbak, esetleg ennek a kettőnek a keveréke (pl. lambda=0,5, második ábra) A releváns ebben az esetben azt jelenti, hogy a szavak többi topikban kevésbé fordulnak elő, tehát kiemelten jellemzik az adott topikot. (Sievert & Shirley, 2014) Ez a vizualizáció nagyban hozzásegít a topikok értelmezéséhez.

Az 5-ös topik leggyakoribb szavai



2. ábra. Forrás: Saját szerkesztés.

Az 5-ös topik legjellemzőbb szavai



2. ábra. Forrás: Saját szerkesztés.

Amikor megkaptam a topikmodellezés outputját, a NarrCat értékeket tartalmazó táblához hozzáfűztem a politikai oldal és a topikhovatartozás információkat is, így elkészült a végleges adattáblám, melynek sorai a cikkek, oszlopai pedig a politikai csoportok, NarrCat és topikhovatartozási értékek.

5.2.3. Topikok bemutatása

A következőkben jellemzem a kapott 21 topikot egy összefoglaló címmel és 10 leggyakoribb szavukkal. Minden cikket abba a topikba sorolok, amelybe a legnagyobb valószínűséggel tartozik, és bemutatom, hogy milyen cikkek tartoznak oda, miről szólnak. Minden topikot jellemzek továbbá azzal, hogy milyen átlagos NarrCat értékeket kaptam az oda tartozó szövegekre, majd kitérek arra, hogy milyen következtetéseket vonhatunk le ezekből.

1. Topik: Szabadidő

A topik legjellemzőbb szavai: magyar, film, kiállítás, kép, történelem, történet, fotó, csapat, budapest, című

A topik legjellemzőbb cikkei futballeseményekről, filmekről, kiállításokról szólnak. Ami az átlagos NarrCat értékeket illeti a topik szövegeiben, az összes cikk átlagánál magasabb az átélő és a pozitív értékelő formák száma, alacsony az intenció és az érzelmi kifejezések száma.

2. Topik: Elbeszélések

A topik legjellemzőbb szavai: ember, kicsi, család, idő, szó, élet, dolog, gyerek, nap, szép

Azok a cikkek, amik a legnagyobb valószínűséggel ebbe a topikba tartoznak, általában nem hírek, hanem valamilyen történetet mesélnek el vagy Trianonnal kapcsolatban, vagy valamilyen hétköznapibb dologról írnak, valahogyan megemlítve Trianont. Ha megfigyeljük a NarrCat értékek átlagát ebben a topikban, akkor érződik, hogy ez a topik különbözik stílusban a többitől, például két NarrCat érték is van, ami itt a legmagasabb. Ezek az én referencia (E/1 személyű elbeszélő miatt) és a kogníció, de magas az érzelmi kifejezések, a pozitív értékelő kifejezések, az átélő és tagadó formák átlagos száma is.

3. Topik: Történelmi leírások – második világháború vége

A topik legjellemzőbb szavai: katona, magyar, horthy_miklós, budapest, szovjet, hadsereg, háború, magyarország, nap, október

A topikba tartozó cikkek, visszaemlékezések, leírások a második világháború végével és az azt követő zavaros időszakkal kapcsolatosak. A témának megfelelően a visszatekintő formák száma magas (ennél a topiknál a legmagasabb), az átélőké alacsony, az én és a mi referenciából is kevés van. Az értékelést kifejező szavak száma összességében magas, a pozitív magasabb, mint a negatív.

4. Topik: Irodalom

A topik legjellemzőbb szavai: könyv, magyar, című, író, szerző, lap, kötet, írás, irodalom, cím

Trianonnal kapcsolatos irodalmi híreket, visszaemlékező írásokat találhatunk főképp a topikban. Az intenció pontszám kétszer olyan magas, mint a kényszer, a pozitív értékelések száma szintén magas, illetve a metanarratív formák átlaga ebben a topikban a legtöbb.

5. Topik: Vallás

A topik legjellemzőbb szavai: egyház, szent, templom, ünnep, református, isten, katolikus, istván, keresztény, püspök

Vallási, egyházi élettel kapcsolatos szövegek, illetve augusztus 20-ról szóló írások alkotják a topikot. Jellemző a pozitív érzelem és értékelés, az aktív igék és a visszatekintő formák intenzívebb használata, alacsony viszont a tagadó és az intenciót kifejező formák száma.

6. Topik: Kormányzati hírek

A topik legjellemzőbb szavai: kormány, forint, támogatás, program, milliárd, miniszter, szám, millió, járvány, intézmény

A 6. topikban sok olyan cikk van, ami nem kapcsolódik szorosan Trianonhoz, kormányzati intézkedésekkel, járványhelyzettel kapcsolatos szövegek kerültek ide, valahogyan megemlítve a „Trianon" szót, pl. „A koronavírus-járvány miatt a Trianon-emlékmű felavatása augusztus 20-án lesz.". Ezek az írások eléggé tényszerűek, ennek megfelelően alacsony az én-referencia, az értékelések és a pszichológiai perspektíva pontszám (ami a kogníció és az érzelmi pontszámok összege).

7. Topik: Megemlékezések

A topik legjellemzőbb szavai: trianoni_békeszerződés, nemzet, évforduló, június, nap, összetartozás, trianon, megemlékezés, beszéd, aláírás

Ebbe a topikba tartozik a legtöbb cikk, az összes 7,08%-a, ez azért is lehet, mert a Trianon 100. évfordulója alkalmából sok helyen volt nagyszabású megemlékezés, amit azután intenzív cikkezés követett. A NarrCat score-ok közül alacsony a passzív, kényszerre és intencióra utaló kifejezések száma, kevés az én-referencia, az intenciós és a tagadó forma. Átlagon felüliek viszont a mi-referenciák, az aktív igék, a metanarratív és visszatekintő szerkezetek.

8. Topik: Gazdaság, környezet

A topik legjellemzőbb szavai: gazdaság, százalék, magyarország, ország, terület, magyar, fontos, állam, elmúlt, helyzet

Makrogazdasági, piacgazdasági, illetve környezetvédelemmel, erdőgazdálkodással kapcsolatos cikkek a legjellemzőbbek a topikra. Az érzelemkifejező szavak és én-referenciák alacsony számát figyelhetjük meg, viszonylag magas a passzív és a kényszer kifejezések mennyisége.

9. Topik: Románia

A topik legjellemzőbb szavai: román, erdély, magyar, románia, székely, autonómia, bukarest, elnök, december, kisebbség

A topik romániai híreket, erdélyi magyarsággal kapcsolatos cikkeket foglal magába. A NarrCat értékek közül nincsenek nagyon kiemelkedőek sem pozitív, sem negatív irányba, az átlagok közelében maradnak.

10. Topik: Szlovákia

A topik legjellemzőbb szavai: magyar, szlovák, felvidék, szlovákia, kisebbség, közös, nyelv, történelem, miniszterelnök, trianoni_békeszerződés

A topik hasonló a 9-eshez, csak itt nem a romániai, hanem a szlovákiai magyarság helyzetével foglalkozó írások jelennek meg. Abban is hasonlít az előző topikra, hogy nem találunk szélsőséges NarrCat értékeket.

11. Topik: Kultúra

A topik legjellemzőbb szavai: színház, című, előadás, dal, darab, program, műsor, zene, közönség, színpad

Színház, koncert, film és egyéb programajánlókat, beszámolókat tartalmaz a topik elsősorban. Magas a metanarratív formák száma, alacsony a tagadás, negatív értékelés és érzelem, a kényszer és az intenció.

12. Topik: Évfordulós beszédek, történeti leírások

A topik legjellemzőbb szavai: magyar, nép, nemzet, történelem, világ, ország, haza, erő, magyarország, idegen

Visszaemlékezések és trianoni évfordulós beszédeken elhangzottakról szóló hírek a fő alkotói a topiknak. A szövegek témája leggyakrabban a magyarság helyzete, a magyar nemzeti öntudat. A mi referenciák, a negatív érzelmi kifejezések és a tagadó formák mennyisége ebben a topikban a legmagasabb. Átlagon felüli továbbá a kényszer, intenció, értékelő formák aránya. Átélő és metanarratív elbeszélésmódot alkalmaznak főként ezek a cikkek, a visszatekintő forma nem jellemző.

13. Topik: Területi változások

A topik legjellemzőbb szavai: magyarország, magyar, terület, ország, német, világháború, határ, állam, háború, francia

Magyarország területének és lakosságának változásával foglalkoznak a topikba tartozó szövegek egyrészt Trianon után, másrészt a bécsi döntést követően. Magas a passzív igék, az intenciós és a tagadó szerkezetek előfordulása, alacsony a pozitív értékelő és érzelmi formák aránya a szövegekben.

14. Topik: Európai politika

A topik legjellemzőbb szavai: európa, nemzet, ország, magyarország, politika, európai_unió, jog, érdek, kormány, uniós

Magyarország és az Európai Unió politikai kapcsolatával, európai országokkal és egyéb EU-s ügyekkel foglalkoznak a topik cikkei, valamit megtalálhatók nagy számban a Bálványosi Szabadegyetemről szóló írások is. A kényszer és az intenció pontszámok meglehetősen magasak, alacsonyabb a pozitív érzelem és értékelés ezekben a cikkekben átlagosan.

15. Topik: Bűncselekmények

A topik legjellemzőbb szavai: zsidó, amerika, ember, ügy, mozgalom, áldozat, cigány, eset, vezető, antiszemitizmus

A cikkek nagy részében emberi jogi bűncselekményekről van szó, sok szöveg szól a cigányságról, rasszizmusról, zsidóságról, antiszemitizmusról, a George Floyd gyilkosságról és az ahhoz kapcsolódó mozgalmakról, illetve a rendőrség munkájáról. Magas az aktív igék, negatív értékelő és érzelmi kifejezések valamit a metanarratív formák aránya, alacsony a pozitív értékelés, pozitív érzelem és a passzív igék pontszám.

16. Topik: Történelmi kutatás

A topik legjellemzőbb szavai: történelem, történész, magyar, egyetem, szabadkőművesség, fontos, kérdés, trianoni_békeszerződés, téma, tudományos

Történészekről, kutatásról, történelemoktatásról szól elsősorban a topik, azon belül is sok a cikk az új nemzeti alaptantervről, Trianonhoz köthető szabadkőműves-legendákról. Ami a NarrCat értékeket illeti, magas a kogníciós és az átélő forma pontszám, alacsony az aktív ige, mi referencia, érzelmi és visszatekintő formák előfordulása.

17. Topik: Belpolitika

A topik legjellemzőbb szavai: párt, orbán_viktor, ellenzék, választás, fidesz_–_magyar_polgári_szövetség, jobbik, politikus, politika, miniszterelnök, kormány

Belpolitikáról, parlamenti pártokról, köztük lévő feszültségekről és politikusokról szóló hírek jellemzik a 17-es topikot. A NarrCat értékek egységesen az átlag közelében vannak, kiemelni talán az alacsonyabb mi referencia, pozitív értékelés és visszatekintő forma, illetve a magasabb tagadás és kogníció pontszámokat lehetne.

18. Topik: Emlékhelyek

A topik legjellemzőbb szavai: város, tér, hely, épület, település, emlékmű, név, szobor, megye, falu

Az ide tartozó cikkek új emlékművek felavatásáról, Trianon emlékének megőrzéséről szólnak. Magas a passzív igék aránya, alacsony az intencióra és a kényszerre utaló kifejezések aránya a szövegekben. Szinte az összes pontszám egyébként átlag alatti, a visszatekintő formák jelentenek még ez alól kivételt.

19. Topik: Társadalmi változások, reformok

A topik legjellemzőbb szavai: politika, társadalom, hatalom, rendszer, kérdés, valóban, ember, liberális, baloldal, helyzet

A magyar rendszerváltásról, politikai eszmék ütközéséről, jelenlegi társadalmi változásokról és külföldi országokban jelen lévő feszültségekről szólnak főleg a topik írásai. A NarrCat értékek átlag-közelinek mondhatók a topikban, a mi referenciák és az érzelmi kifejezések aránya számít viszonylag alacsonynak.

20. Topik: Nemzeti összetartozás

A topik legjellemzőbb szavai: magyar, nemzet, határ, magyarság, túli, magyarország, élő, közösség, trianoni_békeszerződés, külhoni

Kettős állampolgárságról, a határon túli magyarok és általánosságban véve a magyarság helyzetéről, jövőjéről írnak az ide tartozó cikkekben. Alacsony az én referencia érték, magas a mi referencia. Átlag alatti az összes értékelő és érzelem pontszám, a pozitív értékelések aránya kifejezetten alacsony.

21. Topik: Törvényhozás

A topik legjellemzőbb szavai: képviselő, törvény, parlament, elnök, országgyűlés, magyarország, bizottság, jog, javaslat, kormány

Elsősorban a magyar, illetve a román és egyéb magyarlakta területek parlamentjeinek működéséről, intézkedéseiről és meghozott törvényeiről írnak ezekben a cikkekben. Itt a legkisebb a passzív igék, a mi referencia és a pozitív értékelés pontszám, kevés az én referencia is, magas a kényszert kifejező formák aránya.

Az 1-es számú mellékletben látható minden topikra, hogy milyen átlagos értékek születtek azokra a cikkekre, amelyek ebbe a topikba tartoznak elsődlegesen. Oszloponként, azaz NarrCat változónként a magas értékek sárgával, az alacsonyak bordóval lettek színezve, a köztes értékek pedig a két szín közötti átmenettel jelennek meg. A fentieket összefoglalva elmondhatjuk, hogy a topikok jelentős része különböző a NarrCat értékek mentén. Ha az ágencia négy modulját vizsgáljuk (aktív, passzív igék, kényszer és intenció), a társadalmi változások, területi változások, emlékhelyek és gazdaság, környezet topikokban erős passzívitást figyelhetünk meg, azaz ezekben a témákban az cikkírók tárgyilagosan írnak, nem jellemző az aktív cselekvés narratívájukban. Ezzel szemben a kormányzati hírek, belpolitika, vallás, bűnözés topikok igéi aktívak, ezekben a témákban inkább érvényesül a cselekvő. A mi referencia aránya az évfordulós beszédek, Szlovákia, nemzeti összetartozás, vallás és megemlékezések topikokban magas, ebből arra következtethetünk, hogy ezekben a témákban kiemelten fontos a belső csoporttal való magas fokú azonosulás, élményközösség. Az alacsony kogníciós pontszám tényszerű közlésmódra, alacsony empátiára utal, meglepő, hogy a megemlékezések topikban ennyire alacsony az értéke. A negatív érzelmi kifejezések aránya az évfordulós beszédek, bűnözés, kirekesztés, társadalmi reformok, változások topikokban magas, a pozitív a vallás, irodalom, Szlovákia és a kultúra topikokban. Az évfordulós beszédek topikban a pozitív kifejezések aránya is átlag feletti, csakúgy, mint a belpolitika topikban, ez jelzi ezen topikok érzelmi túlfűtöttségét. A negatív és pozitív értékelés pontszámok is meglehetősen különbözők topikonként. Az évfordulós beszédeknél szintén nagyon magas mindkettő (pozitív magasabb), az értékelő kifejezéseket tekintve erősen pozitív még a második világháború vége, az irodalom, az elbeszélések és a vallás topikok, ugyanakkor negatív a bűnözés, kirekesztés, belpolitika és a társadalmi változások, reformok. Ezek azok a topikok tehát, ahol gyakran véleményt mond a narrátor valamiről, ami lehet tárgy, cselekedet,

esemény, alkotás, személy stb. A tagadó formák száma is az évfordulós beszédeknél, a belpolitikánál és az elbeszéléseknél magas, ami a világ leértékelését jelezheti az író részéről ezekben a témákban. A narratív forma megválasztása gyakran a témából következik, de az évfordulós beszédek, európai politika társadalmi változások, reformok és elbeszélések topikokban az átélő forma nagyszámú használata a retrospektív kárára a téma lezáratlanságára, fel nem dolgozására utal.

5.2.4. Topikok és politikai hovatartozás

A 6-os táblázatban látható, hogy melyik topikhoz melyik politikai oldalról hány cikk tartozik, a sárga szín azt jelzi, hogy az adott politikai oldalnak magas számú cikke tartozik a topikba, a bordó szín pedig, hogy kevés. A legnépesebb topik a belpolitika (17) 790 cikkel, a legkevesebb cikk, 392 pedig a 16-os, történelmi kutatás topikhoz tartozik elsődlegesen. A szélsőjobboldal cikkei a 7-es és az 5-ös topikban jelennek meg legnagyobb számban (megemlékezések, vallás), a baloldalnak a belpolitika, illetve a társadalmi változások, reformok kedvelt témája. A kormánypárti médiában a legtöbb cikk a 6-os és a 11-es, kormányzati hírek és a kultúra topikba tartozik, a jobboldal cikkei szintén a belpolitika (17) és a társadalmi változások, reformok (19) topikokban vannak legmagasabb számban jelen, hasonlóan a baloldalhoz.

Topik		Szélsőjobb	Baloldali-liberális	Kormánypárti	Nem kormány-párti jobboldali	Összesen
1	Szabadidő	100	122	245	17	484
2	Elbeszélések	177	164	251	37	629
3	Második világháború vége	181	83	185	15	464
4	Irodalom	142	100	183	28	453
5	Vallás	276	21	196	6	499
6	Kormányzati hírek	47	97	463	18	625
7	Megemlékezések	344	82	368	23	817
8	Gazdaság, környezet	93	96	208	25	422
9	Románia	256	36	244	16	552
10	Szlovákia	264	49	88	14	415
11	Kultúra	190	112	402	9	713
12	Évfordulós beszédek	261	40	235	8	544
13	Területi változások	225	87	196	16	524
14	Európai politika	130	49	234	14	427
15	Bűnözés, kirekesztés	149	90	190	9	438
16	Történelmi kutatás	108	98	170	16	392
17	Belpolitika	159	207	369	55	790
18	Emlékhelyek	171	198	270	19	658
19	Társadalmi változások, reformok	137	256	175	41	609
20	Nemzeti összetartozás	205	39	384	15	643
21	Törvényhozás	147	82	190	18	437

6. táblázat. Cikkszámok a topikokban politikai oldalanként. Forrás: Saját szerkesztés.

Az alábbi 7-es táblázat hasonló a 6-oshoz, csak százalékos megoszlást mutat, tehát azt vizsgálhatjuk meg, hogy az egyes politikai oldalak cikkei hány százaléka tartozik az adott topikhoz. A színezés itt soronként történik, a magas, átlagon felüli értékek jelennek meg

sárgával, az alacsony, átlagon aluli értékek bordóval. 7 olyan topik is van, amelyben a szélsőjobboldali hírportálok felülreprezentáltak (3, 5, 7, 9, 10, 12, 13), és van 2 olyan is, ahol mindhárom másik oldal populáción belüli megoszlása átlag felett szerepel (8, 17). Ebből arra a következtetésre juthatunk, hogy a szélsőjobboldal esetében más témák a hangsúlyosak, mint mindenki másnál. Van négy olyan téma, ami inkább a kormánypárti médiára jellemző, a kormányzat, a kultúra, nemzetközi témák és a határon túli magyarok (6, 11, 14, 20) és van egy, a törvényhozás, politika, amiben csak a kormányközeli cikkek aránya marad populációs arányon alul. Négy téma esetében a baloldal és a nem kormánypárti jobboldal cikkei populációs aránynál magasabb arányban vannak jelen, mindennapi életről szóló elbeszéléseket és az irodalomról, a történelemről és a politikai rendszerekről gyakrabban írnak (Trianon vonatkozásában) ezek az oldalak, mint a más politikai irányultsággal rendelkezők.

Topik		Szélsőjobb	Baloldali-liberális	Kormánypárti	Nem kormány-párti jobboldali	Sokasági átlag
1	Szabadidő	2,7%	5,8%	4,7%	4,1%	4,2%
2	Elbeszélések	4,7%	7,8%	4,8%	8,8%	5,5%
3	Második világháború vége	4,8%	3,9%	3,5%	3,6%	4,0%
4	Irodalom	3,8%	4,7%	3,5%	6,7%	3,9%
5	Vallás	7,3%	1,0%	3,7%	1,4%	4,3%
6	Kormányzati hírek	1,2%	4,6%	8,8%	4,3%	5,4%
7	Megemlékezések	9,1%	3,9%	7,0%	5,5%	7,1%
8	Gazdaság, környezet	2,5%	4,6%	4,0%	6,0%	3,7%
9	Románia	6,8%	1,7%	4,7%	3,8%	4,8%
10	Szlovákia	7,0%	2,3%	1,7%	3,3%	3,6%
11	Kultúra	5,1%	5,3%	7,7%	2,1%	6,2%
12	Évfordulós beszédek	6,9%	1,9%	4,5%	1,9%	4,7%
13	Területi változások	6,0%	4,1%	3,7%	3,8%	4,5%
14	Európai politika	3,5%	2,3%	4,5%	3,3%	3,7%
15	Bűnözés, kirekesztés	4,0%	4,3%	3,6%	2,1%	3,8%
16	Történelmi kutatás	2,9%	4,6%	3,2%	3,8%	3,4%
17	Belpolitika	4,2%	9,8%	7,0%	13,1%	6,8%
18	Emlékhelyek	4,5%	9,4%	5,1%	4,5%	5,7%
19	Társadalmi változások, reformok	3,6%	12,1%	3,3%	9,8%	5,3%
20	Nemzeti összetartozás	5,4%	1,9%	7,3%	3,6%	5,6%
21	Törvényhozás	3,9%	3,9%	3,6%	4,3%	3,8%

7. táblázat. Politikai oldalak cikkeinek eloszlása a topikok között. Forrás: Saját szerkesztés.

Összességében elmondhatjuk, hogy a szélsőjobboldal nagyon elkülönül témáit tekintve, de legközelebb a kormánypárti média témái állnak hozzá, pl. az évfordulós beszédek, nemzeti összetartozás, vallás, Románia egyaránt kedvelt témáik. Ugyanakkor a történelmi leírások, a társadalmi változások és az irodalom kevésbé jellemző, ezekben a témákban a baloldal és a nem kormánypárti jobboldal dominál. Természetesen van egyéb kombinációra is példa, mondjuk a szabadidő topik a baloldali és a kormánypárti sajtóra jellemző jobban, de összességében elmondhatjuk, hogy a témákat tekintve is egy polarizáltabb képet látunk.

6. Modellezés és eredmények

6.1. Béta regressziós modell

A regressziós modellel azt a hipotézist kívántam tesztelni, hogy a NarrCat értékei magyarázzák a topikhovatartozási értékeket. Ehhez 21 db béta regressziós modellt építettem, ahol a 21 topikhovatartozási mutató volt az eredményváltozó, és minden esetben a NarrCat értékek voltak a magyarázó változók. A béta regresszió használatát a függő változó (adott topikba való tartozás valószínűsége) értékkészlete, azaz a szigorúan 0 és 1 közötti értékei indokolják.

A NarrCat 19 output értéket adott vissza, ebből az összesítő értékeket töröltem (pl az „Érzelem összes" a negatív és pozitív érzelmi score-ok összege), illetve a pszichológia pontszám is a kogníció és a két érzelmi érték összegeként előáll, ezért ezt is kivettem a modellből, valamint a Visszatekintő forma változót is eltávolítottam, ez ugyanis nagyon erős negatív korrelációt mutatott az Átélő forma változóval.

Logit linkfüggvénnyel futtattam a modelleket, a mellékletben látható egy táblázat mindegyik modellre a kapott p értékekkel. A táblázat első oszlopában láthatjuk, hogy melyik topik volt abban az esetben a függő változónk. A sárgával jelölt értékek mutatják azokat a változókat, amelyek minden szokásos megbízhatósági szint mellett szignifikánsak, rózsaszínnel, amelyek 95%-os, bordóval, amelyek 90%-os megbízhatósági szinten szignifikánsak, szürkével pedig a nem szignifikáns változókat láthatjuk. A modellek teljesítményét R^2 mutatóval értékeltem, a táblázat utolsó sorában láthatjuk a mutatókat a modellekre. A mutatók mindegyik modellre alacsonyak lettek, a 2-es topik (elbeszélések) mint eredményváltozó esetében lett a legmagasabb, 0,287, illetve magas értéket kaptam még a 12-es (évfordulós beszédek) és a 19-es topikra (társadalmi változások, reformok). Ezen topikok cikkei tehát élesebben elkülönülnek a NarrCat értékek mentén. A legalacsonyabb a 9-es topiknál (Románia) volt a mutató, 0,02, illetve további, alacsony R^2-tel rendelkező modellek pedig a 10-es (Szlovákia) és a 16-os (történelmi kutatás) topikhoz tartozók. Futtattam ezután többféle linkfüggvénnyel is a modelleket, de nagyon hasonló eredményeket kaptam. Azt tehát elmondhatjuk, hogy a NarrCat értékek összességében nem magyarázzák jól a topikhovatartozást. Ebből arra következtethetünk, hogy a szöveg nyelvezete, narratívája, stílusa nem függ erősen a témától, tehát a NarrCat egy új, meglehetősen más dimenziót tud behozni a szövegek elemzésébe.

Természetesen ettől függetlenül vannak különbségek az egyes topikok között a NarrCat értékek nagyságát és relevanciáját tekintve, erről az 5.2.3.-as, „Topikok bemutatása” fejezetben írtam, és a koefficienseket megvizsgálva is hasonló következtetésre juthatunk, láthatjuk ugyanis, hogy melyik NarrCat változó hogyan járul hozzá a topikhovatartozás alakulásához.

A mellékletek között 3-as számmal láthatunk egy táblázatot, ami tartalmazza a modellekhez tartozó koefficienseket. A modell eredményváltozója az első oszlopban látható, egy modellhez tartozó koefficiens értékek egy sorban jelennek meg. A koefficiensek néhány NarrCat változóra nagyon különbözők lettek topikonként, ez azt jelenti, hogy az adott NarrCat változó egységnyi változása a topikba tartozás valószínűségét némelyik topiknál nagy mértékben megnöveli, némelyiknél pedig nagy mértékben lecsökkenti. Ilyen NarrCat értékek a passzivitás, kényszer, intenció, értékelés és érzelem. Ezzel szemben például az aktív igék, mi referencia, a tér-idő perspektíva dimenzióban megjelenő változók koefficiensei egységesebbek.

6.2. Klasszifikációs modellek

A Diszkriminancia-analízis, Support Vector Machine, Random Forest és az Extreme Gradient Boosting algoritmus teljesítményét hasonlítom össze azért, hogy teszteljem azt a hipotézist, hogy a NarrCat és topikhovatartozás változók segítségével meg tudjuk határozni, hogy politikailag milyen oldalon jelent meg egy cikk. Az összehasonlítást elsősorban a csoportba sorolás pontosságával (accuracy) mérem. A pontosság a teszthalmazon mért helyes besorolás aránya, azaz a (helyesen kategorizált megfigyelések száma) / (összes megfigyelés száma). Fontosnak tartom nézni továbbá a tanító halmazon mért pontosságot, ugyanis, ha az jóval magasabb, mint a teszthalmazon, akkor túlillesztett modellről beszélhetünk. Használom ezen kívül a Cohen’s Kappa mutatót, ami azt becsüli meg, hogy mennyivel jobban teljesít a modell a random klasszifikációhoz képest. (Peng és Choi, 2005)

Minden esetben a politikai hovatartozás a csoportképző változó, aminek négy kimenete van: kormánypárti, nem kormánypárti jobboldali, szélsőjobboldali, balliberális, tehát a fent említett módszereket többcsoportos klasszifikációra fogom használni. Az adatokat minden esetben tanító és teszthalmazra bontom, 80-20%-os arányban. A klasszifikációs modellek futtatásával célom továbbá, hogy megtudjam a független változókról, hogy melyik milyen

mértékben járult hozzá a klasszifikációs döntéshez, tehát, hogy mely változók értékei csoportosítanak hasonlóan, mint a politikai hovatartozás.

6.2.1. Diszkriminancia-analízis

Lineáris diszkriminancia-analízist (LDA) akkor lehet végrehajtani, ha a variancia-kovariancia mátrixok minden csoportra egyenlőek, azaz többváltozós homoszkedaszticitás kell, hogy fennálljon. Box-féle M tesztet használtam a homoszkedaszticitás teszteléséhez, amire kis p értéket kaptam, azaz elutasítom a nullhipotézist, miszerint a variancia-kovarianciamátrixok egyenlők. Mivel nem teljesült az LDA homoszkedaszticitás feltétele, négyzetes diszkriminancia-analízist (QDA) végeztem, ugyanis az LDA-val ellentétben a QDA azt feltételezi, hogy minden osztálynak saját kovariancia-mátrixa van.

A modellezéshez az R MASS csomagját, azon belül a qda függvényt használtam. A topikhovatartozási értékek és a NarrCat értékek voltak a magyarázó változóim, kivéve azok, amiket a regressziónál is töröltem. A 21-es topikot is törölnöm kellett modellből, mert azt jelezte az R, hogy csoportokon belül lineáris kombinációval előáll a változó.

A csoportokat a következők szerint kódoltam:

1 – Szélsőjobb
2 – Baloldali-liberális
3 – Kormánypárti
4 – Nem kormánypárti jobboldali

		Becsült csoport			
		1	**2**	**3**	**4**
Tényleges csoport	**1**	387	179	149	34
	2	74	264	78	17
	3	330	306	374	37
	4	14	45	13	6

8. táblázat. QDA modell konfúziós mátrixa. Forrás: Saját szerkesztés.

A fenti ábrán látszik a konfúziós mátrix, 44,7% lett az össz pontosság, a 95%-os konfidencia-intervallum 42,7%-46,8%. A szélsőjobboldali csoportban 51,7%, a balliberálisban 61%, a kormánypártiban 35,7%, a nem kormánypárti jobboldali kategóriában 7,7% lett a helyesen besorolt elemek aránya. Ez azt jelenti, hogy a balliberális csoportot azonosította legkönnyebben az algoritmus, a nem kormánypárti jobboldali cikkeket pedig legnehezebben.

Ez nem meglepő, hiszen ebbe a csoportba tartozik a legkevesebb szöveg, nehezen azonosítja a tulajdonságait az algoritmus. 0,21 lett a Kappa, ami gyengének számít.

6.2.2. Support Vector Machine

A support vector machine algoritmust az e1071 csomag segítségével és különböző kernelekkel futtattam: lineáris, polinomiális, szigmoid, radiális.

Lineáris:

A modell pontossága 51,4%, 60% az első (szélsőjobb) csoportban, 49% a harmadik (kormánypárti) csoportban, a másik két csoportba nem osztott az algoritmus cikket. A Kappa értéke 0,16, ami még a diszkriminancia-analízisnél is rosszabb eredmény.

Polinomiális:

A modell pontossága 51,1%, egy kicsivel rosszabb, mint lineáris esetben. 50,7%-os a pontosság az 1. csoportban, 43% a 2.-ban, 57% a 3.-ban, 12% a 4. csoportban. Ez azt jelenti, hogy polinomiális kernellel is a szélsőjobboldali és a kormánypárti cikkeket tudjuk legjobban kategorizálni, és a balliberális cikkek esetében sokkal jobb eredményt érhetünk el, mint az előző esetben. A Kappa is magasabb egy kicsivel, de még itt is csak 0,24.

Szigmoid:

Szigmoid kernellel kaptam a legrosszabb pontosságot, 42%-ot. A legjobb pontossága a kormánypárti csoportnak van, de az is csak 48%, a Kappa itt a legalacsonyabb, 0,09.

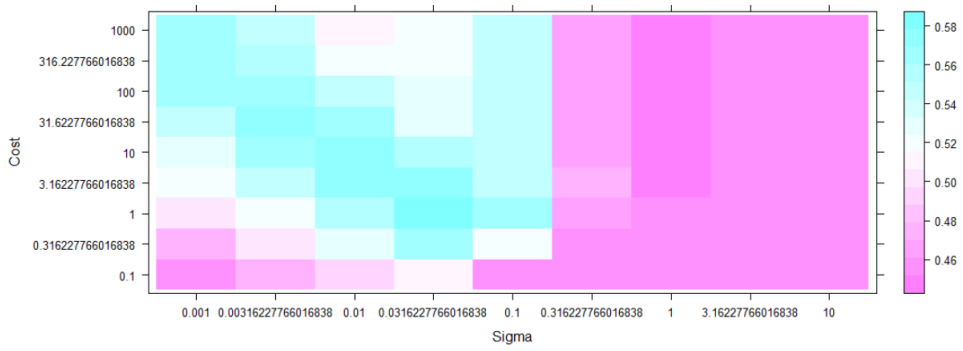
Radiális:

A radiális kernel bizonyult a legjobb választásnak, 54,6%-os pontosságot értem el vele az első futtatáskor. Az 1. csoportban 53,8% volt a helyes klasszifikáció, a 2. csoportban 48,2%, a 3.-ban 59%, a 4.-ben pedig a leggyengébb, 18,33%. Ez sajnos gyengébb teljesítmény, mintha véletlen sorsoltam volna csoportba a cikkeket, hiszen akkor 25% körüli számot kellene kapni. A Kappa 0,29.

Ezek után a radiális kerneles support vector machine modell paramétereit igyekeztem hangolni. Grid search-öt alkalmaztam a cost és a szigma paraméterekre, ami azt jelenti, hogy 10 hatványaként határoztam meg 9-9 kipróbálandó értéket, minden cost értéket minden szigma értékkel párosítva. Ez tehát 81 modellt jelentett, a modellek teljesítménye jól

megfigyelhető az alábbi ábrán, kékkel a jobban, rózsaszínnel a gyengébben teljesítő modellek láthatók szigma és cost paraméterek alapján.

Modellteljesítmény különböző szigma és cost paraméterek mellett



4. ábra. Forrás: Saját szerkesztés.

10 hajtásos keresztvalidációt alkalmaztam, azaz 10-szer futtatam ugyanazzal a paraméterpárossal a modellt, mindig a teszthalmaz másik tizedét használva mintaelemekként. A 10 futtatás átlagos pontosságát hasonlítottam össze a modellekre, az optimális paraméterpáros cost=1 és szigma=0,0316 lett. A pontosság ebben az esetben 57,8% volt. A legalacsonyabb pontosság 45,2% lett, jól látszik tehát, milyen fontos a paraméterek helyes megválasztása.

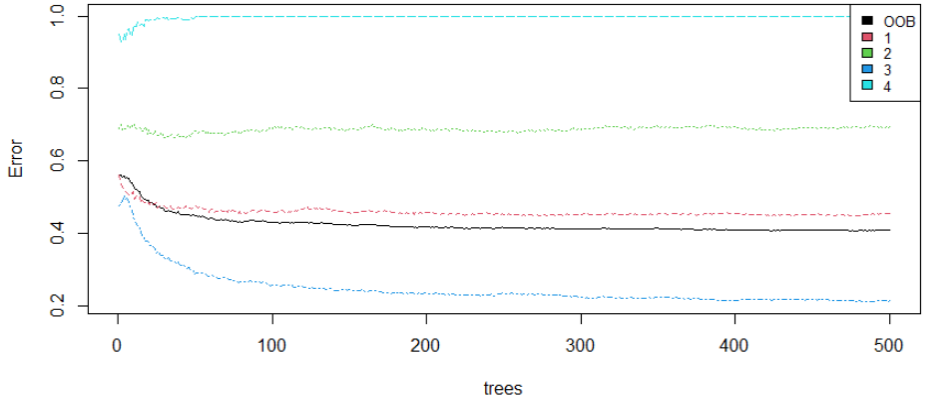
Finomítottam a grid-en és újrafuttattam az optimalizációt az előző futtatásnál optimálisként meghatározott paraméterek közelében lévő 5-5 számmal, ekkor is a cost=1, szigma=0,03162 értékekre kaptam legmagasabb pontosságot. Ha a négy csoportot nézzük, az első, szélsőjobboldali kategória tagjait 60,6%-os pontossággal sorolta be, a balliberális cikkeket 56,9%-os, a kormánypárti cikkeket 58%-os pontossággal. Nem kormánypárti jobboldali kategóriába nem került semmi, itt tehát az összes cikket rosszul kategorizálta. A Kappa 0,31-re emelkedett, azonban ez is viszonylag alacsonynak mondható.

6.2.3. Random Forest

A Random Forest modellt először alap paraméterbeállításokkal futtattam az R randomForest csomagját felhasználva. Az alapbeállítás azt jelenti, hogy a növesztett fák száma (ntrees) 500, és a változószám gyökét, azaz jelen esetben 5 változót használ minden vágásnál (mtry) az algoritmus. Ezzel azt kaptam, hogy az Out Of Bag (OOB) hibaarány 41,7%-os, az általam elkülönített tesztadatokon vett hibaarány 41,3%-os.

A 4. csoportba, azaz a nem kormánypárti jobboldali cikkek közé nem sorolt semmit, ugyanis ezen cikkek aránya nagyon alacsonynak bizonyul. A balliberális cikkeket szintén nem klasszifikálta jól, 70,1%-os hibával, a szövegek többségét kormányközelinek sorolta be. Az 1., szélsőjobboldali csoportban 45,7%-os volt a hiba, a legjobban pedig a kormányközeli média cikkeit azonosította a modell, 20,3%-os hibával. Az alábbi ábrán megfigyelhetjük, hogy a növesztett fák számának növekedésével hogyan változik a hiba. Azt látjuk, hogy 100 fánál 3 csoport hibája már nem változik, de a 3. csoport (baloldali-liberális) besorolási hibája tudott csökkenni 500 fa felhasználásáig, ezért a későbbiekben az ntrees paramétert 700-ra állítottam.

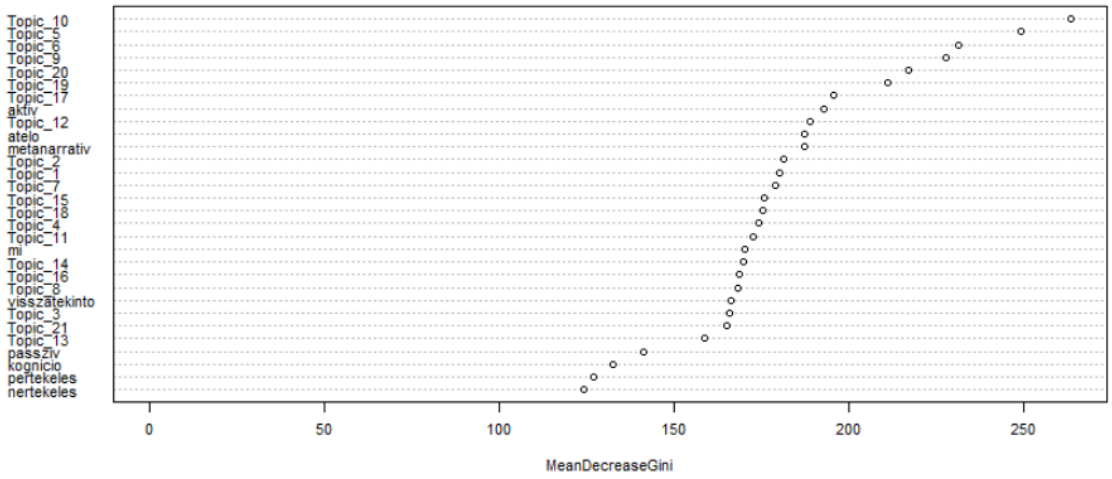
1. Random Forest modell hibája a fák számának függvényében



5. ábra. Forrás: Saját szerkesztés.

A Gini-index a csoportok közötti varianciát fejezi ki, egy vágás milyenségét mérjük vele, amikor klasszifikációs fát építünk. Annál fontosabb a változó, minél magasabb a változó használatával elért átlagos Gini-pontszám csökkenés. (James et al., 2013) Az alábbi ábrán láthatjuk a Gini-index csökkenését változónként. Jelen esetben az első 7 legfontosabb változó topikváltozó volt, és a 20 legfontosabb változó között csak 4 NarrCat érték volt, ezért elmondhatjuk, hogy a topik értékek szerint heterogénebbek a politikai csoportok, mint a NarrCat értékek szerint.

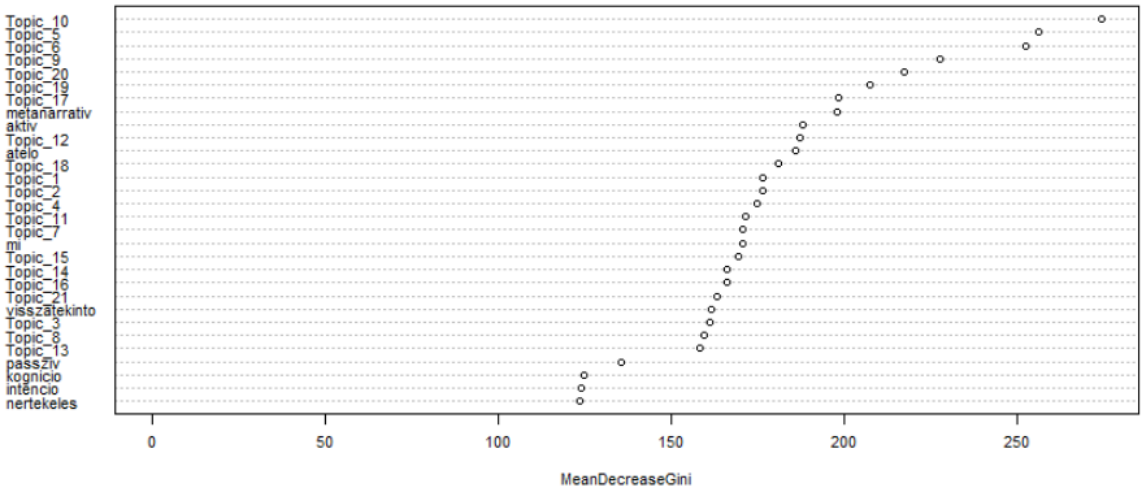
1. Random Forest modell Gini-pontszám csökkenése változónként



6. ábra. Forrás: Saját szerkesztés.

Ezt követően megpróbáltam megkeresni az mtry paraméter optimális nagyságát, ehhez futtattam több mtry értékkel a modellt, és besorolási hiba alapján értékeltem azokat. A legjobb modellt 8-as mtry-ra kaptam, 41,09% volt a besorolási hibaarány, tehát 58,91% a pontosság. Ennél a modellnél is azt tapasztaljuk, hogy az első 7, klasszifikációban Gini index szerint legnagyobb szerepet vállaló változó topikváltozó. Az ábrán ránézésre is láthatjuk, hogy a 10, 5, 9 topikok fontosak, ezek például a szélsőjobboldalra jellemzőek, a 6 és a 20 a kormánypárti médiára. Ezekkel a topikokkal tehát jól be lehet azonosítani a cikkek hovatartozását. Ami a klasszifikáció szempontjából fontos NarrCat változókat illeti, a metanarratív elbeszélési forma a szélsőjobboldalt jellemzi, az aktív igék száma a kormánypártot, az átélő a bal és a nem kormánypárti jobboldalt, a mi referenciák magas száma pedig szintén a szélsőjobboldalt, ezért ezek a változók kapnak még fontos szerepet a csoportba sorolás során. Továbbra is a harmadik és az első csoport tagjai, azaz a szélsőjobboldali és a kormánypárti cikkek kategorizálhatók legjobban.

Optimális Random Forest modell Gini-pontszám csökkenése változónként



7. ábra. Forrás: Saját szerkesztés.

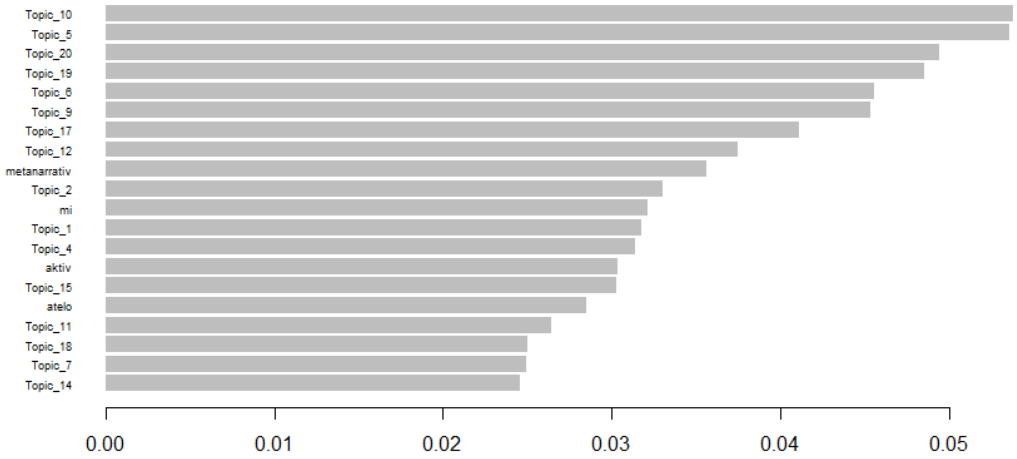
6.2.4. XGBoost

Az első modellemet a következő, alapértelmezett paraméterekkel futtattam:

- nrounds (iterációszám): 300
- eta (tanulóráta): 0,3
- gamma (regularizáció mértéke): 0
- max_depth (elágazások száma): 6
- min_child_weight (minimum elemszám egy csomóponton): 1
- subsample (használt megfigyelések aránya): 1
- colsample_bytree (használt változók aránya): 1

58,7%-os pontosságot kaptam és 0,329-es Cohen's Kappát, ami az eddigi legjobb, de alacsony érték.

Alap XGBoost modell változóinak fontossága



8. ábra. Forrás: Saját szerkesztés.

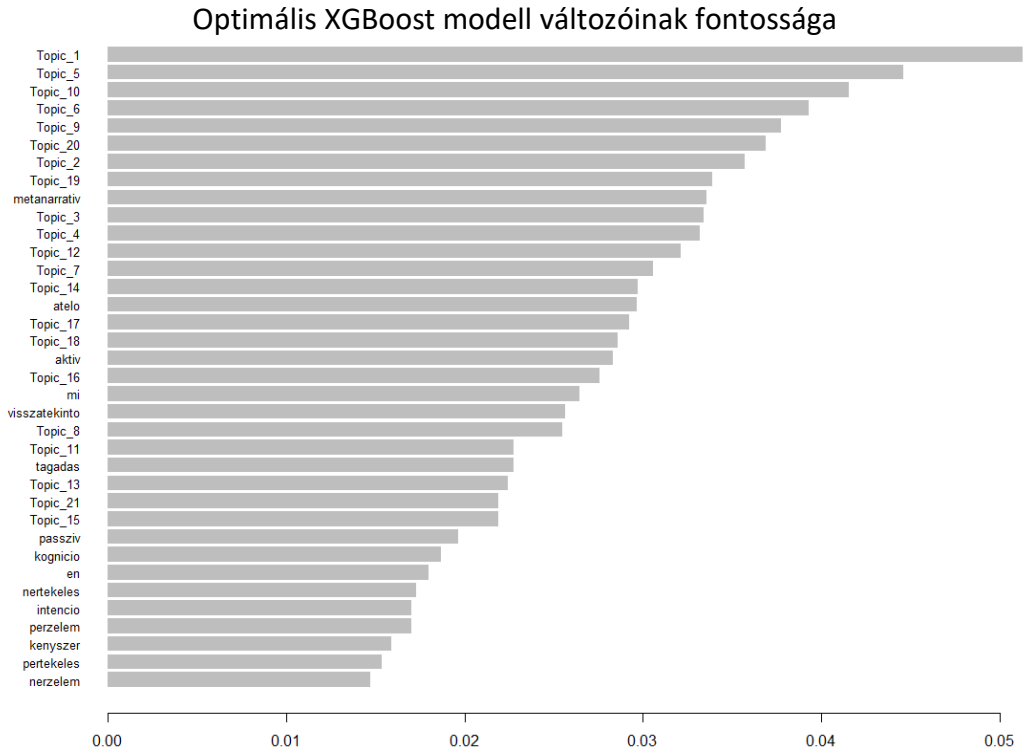
A kezdeti modell után hangoltam a paramétereket, random search módszer segítségével. 5000 db ötöst képeztem véletlenszerűen választott max_depth, eta, subsample, colsample_bytree és min_child_weight értékekkel. A max_depth paraméter 3 és 10 között lehetett, az eta 0,01 és 0,3 között, a subsample 0,7 és 1 között, a colsample_bytree 0,6 és 1 között, a min_child_weight 0 és 10 között. Ezután futtattam 5000 modellt a kapott paraméterötösökkel. A modelleket merror segítségével hasonlítottam össze, ami a többcsoportos klasszifikáció hibaaránya. A legkisebb hiba 0,41-es volt a modelleknél, azaz a klasszifikáció pontossága 59%-os volt.

Ezután változtattam a véletlenszerűen választott max_depth paraméter terjedelmét 7 és 20 közöttire, megengedve így komplexebb modelleket is, és lefuttattam újra 5000 modellt. Az optimális modell pontossága 61,16% volt, hibaaránya 38,84%. A pontosság 95%-os megbízhatósággal 59,1% és 63,2 között van, a Kappa mutató 0,37.

Az optimális modell paraméterei a következők voltak:

- max_depth: 14
- eta: 0,2685
- subsample: 0,8853
- colsample_bytree: 0,8453
- min_child_weight: 0

Ezen kívül használtam még az early_stopping_rounds paramétert, 30-ra állítottam, ami azt jelenti, hogy ha 30 iteráción keresztül nem csökken a veszteség, akkor nem várja végig az algoritmus az összes iterációt (esetemben 300-at), hanem a következő paraméteregyüttessel folytatja a modellezést. Az 5 legfontosabb változó a klasszifikációnál az 1-es, 5-ös, 10-es 6-os és 9-es topikhovatartozási változó volt, amely topikok a szabadidő, vallás, Szlovákia, kormányzati hírek, és a Románia voltak. A NarrCat változók továbbra is kisebb szerepet kapnak a klasszifikációban.



9. ábra. Forrás: Saját szerkesztés.

A tanulóhalmazon vett pontosság 1 volt, ami azt jelenti, hogy a modell teljesen illeszkedik a tanulóhalmazra. A gamma paraméterrel regularizációt vonhatunk a modellbe, ami javasolt, ha a tanulóhalmazon ennyivel nagyobb a pontosság, mint a teszthalmazon. Ezért futtattam az optimális modellparaméterekkel és különböző gamma értékekkel (0-20) az XGBoost modellt, azonban a legkisebb teszthalmazon vett hibát így is gamma = 0-ra kaptam. Ebből megállapítható, hogy regularizációt nem érdemes jelen esetben bevonni. A végleges modellel kapott konfúziós mátrixot alább láthatjuk. A modell a kormánypárti cikkeket kategorizálta legjobban, 75,4%-át jól sorolta csoportba. A szélsőjobboldali cikkek 58,4%-át klasszifikálta helyesen, a baloldal esetében pedig 40,9%-ot. A nem kormánypárti jobboldali cikkeknek pedig csupán 1,2%-át sikerült helyes csoportba sorolni, elmondhatjuk, hogy ezen cikkek klasszifikációjára nem használható az algoritmus, mivel alulreprezentáltak voltak a korpuszban, a tanulóhalmazba is nagyon kevés ilyen mintaelem került.

		Becsült csoport			
		1	**2**	**3**	**4**
Tényleges csoport	**1**	442	32	278	4
	2	64	167	177	0
	3	206	55	801	0
	4	24	17	39	1

9. táblázat. Legjobb XGBoost modell konfúziós mátrixa. Forrás: Saját szerkesztés.

7. Összegzés

Kutatásom során arra a kérdésre kerestem a választ, hogy milyen eltérések figyelhetők meg az online média Trianonnal kapcsolatos cikkeinek témájában és narratívájában a hírportálok politikai hovatartozásától függően. Ezen kívül egy módszertani kérdésem is volt, mégpedig az, hogy a NarrCat használata hogyan egészíti ki a topikmodell eredményeket és hogyan használhatók ezen eszközök output értékei klasszifikációra. Az elemzéseket több mint 11 ezer, politikailag négy kategóriába sorolható online hírportálról származó újságcikken végeztem. A négy kategória a szélsőjobboldali, a kormánypárti, a baloldali-liberális és a nem kormánypárti jobboldali.

7.1. Eredmények

Az újságcikkekre topikmodelleket és a NarrCat elemzőt futtattam. A legjobban teljesítő topikmodell 21 topikot tartalmazott, néhány példa a topikokra: Irodalom, Belpolitika, Szlovákia, Évfordulós beszédek, Megemlékezések. A 21 topikra minden szövegnek megvolt a topikba tartozási valószínűsége. A NarrCat pedig 8 dimenzióban adott vissza 16, pszichológiai vagy nyelvi kategóriában pontszámot minden szövegre, például pozitív és negatív érzelem, aktív igék, tagadás, kényszer. Megvizsgáltam, hogy a négy politikai kategória szerint milyen eltérések figyelhetők meg a topik és a NarrCat pontszámokban.

A kormánypárti médiára egy aktív, de negatív narratíva jellemző, a szélsőjobb pedig írásaiban is szélsőséges: gyakran találkozunk nagyon magas vagy nagyon alacsony értékekkel a többi oldalhoz képest. Például az alacsony kogníció, de magas mi-referencia a belső csoporttal való azonosulást, kifelé viszont az empátia hiányát jelenti. Ebből és a megjelent cikkek magas számából következtethetünk arra, hogy a szélsőjobboldali csoport esetében a Trianon mint trauma még nem feldolgozott. A baloldal értékei pont fordítottak, magas a kogníció, alacsony a mi- és én-referencia. A bal- és a nem kormánypárti jobboldal narratívája nagyon hasonló nyelvi és pszichológiai szavak, formák használatával jön létre. Ezekre a cikkekre passzív igék, kényszert kifejező formák, magas kogníciós pontszám, alacsony pozitív érzelmi és értékelő kifejezések jellemzők.

Ha megvizsgáljuk a topikokat, a szélsőjobbos média ebben is elkülönül a többi oldaltól, hét olyan topik is van, aminek legtöbb cikke szélsőjobboldalhoz köthető portálon jelent meg. A legjellemzőbbek erre az oldalra a megemlékezésekről szóló hírek, a vallás és a Szlovákia témák,

nem sokszor írnak a kormányzati hírekről, gazdaságról, szabadidőről. A kormánypárti média kedvelt témái a kormányzati hírek, a kultúra, a nemzeti összetartozás. A bal- és a nem kormánypárti jobboldal a médiában nagyon hasonló topikokról beszél, a legtöbbször a belpolitikáról és társadalmi változásokkal kapcsolatos témákról. Sok esetben egy topikra arányaiban két jellemzőbb oldal a szélsőjobb és a kormánypárti jobboldal, kevésbé jellemző a baloldal és a nem kormánypárti jobboldal, vagy éppen ez fordítva. Ez arra enged következtetni, hogy a politikai polarizáció az újságcikkek témájában is felismerhető kezd lenni.

Ahhoz, hogy megvizsgáljam, hogy a NarrCat értékek, azaz a szöveg stílusa mennyire magyarázzák a topik értékeket, vagyis a szöveg témáját, béta regressziós modelleket futtattam. Azért esett a választásom erre a modellre, mert a topikhovatartozás mint magyarázó változó 0 és 1 közötti, ami a béta regresszió sajátossága is egyben. Mivel 21 topikom lett, 21 ilyen regressziós modellt futtattam. Hiába volt a magyarázó változók köre mindig megegyező, a különböző függő változókkal jelentősen különböző modelleket kaptam. Voltak olyan NarrCat változók, amelyek koefficiensei modelltől függően kiemelkedően nagyok és kicsik is lehettek, például aktivitás, passzivitás, értékelés és érzelem változók. Azt találtam összességében, hogy nem magyarázzák jól a NarrCat értékek a topik értékeket, ez is azt támasztja alá, hogy a NarrCat eszköz egy újfajta dimenziót hoz a szövegelemzésbe, ki tudja egészíteni a topikmodell eredményeket.

Négyféle klasszifikációs algoritmust próbáltam ki, hogy csoportosítani tudjam a cikkeket a megjelentető hírportálok politikai irányultsága szerint, a Kvadratikus diszkriminancia-analízist, a Support Vector Machinet, a Random Forestet és az Extreme Gradient Boostingot. A NarrCat és topikhovatartozási értékek voltak a magyarázó változóim, a csoportképző változóm a politikai hovatartozás. Az alapján hasonlítottam elsődlegesen össze a modelleket, hogy mennyire pontosak voltak, azaz a helyesen kategorizált mintaelemek és az összes mintaelem arányát néztem a tanulóhalmazon, amely minden modellnél megegyezett. A modellek paramétereinek hangolása után megkaptam mindegyikre az optimális paramétereket és az optimális modellhez tartozó pontosságot. A legmagasabb elért pontosság a Diszkriminancia-analízis esetén 44,7%, az SVM-nél 57,8%, a Random Forestnél 58,91%, az XGBoost modellnél pedig 61,16% volt. Elmondhatjuk tehát, hogy az XGBoost modell volt a legmegfelelőbb választás. A Cohen’s Kappa is az XGBoost modellnél volt legmagasabb, 0,37. A Kappa mutató sajátossága egyébként, hogy kevésbé mutat magas értékeket kiegyenlítetlen adatok esetén,

ez magyarázza az alacsony értékét. A kormányközeli média cikkeit volt általában a legkönnyebb klasszifikálni (kivéve QDA, ott a szélsőjobboldaliakat), második legjobban pedig a szélsőjobboldalhoz köthető cikkeket. A nem kormánypárti jobboldal cikkeit egyáltalán tudták jól csoportba sorolni az algoritmusok, mert nagyon kevés számú cikk tartozott abba a csoportba, nem tudtak az algoritmusok rátanulni a csoport sajátosságaira.

A modelleknél megvizsgáltam azt is, hogy a klasszifikációban mely változók a legfontosabbak, azaz mely változók mentén heterogének a csoportok. Azt kaptam, hogy általában az első 7 legfontosabb változó topikhovatartozási változó, ami azt jelenti, hogy van néhány topik, ami mentén jól elkülönülnek a politikai csoportok. Ezek a topikok például a Szlovákia, a vallás, a kormányzati hírek vagy a társadalmi változások, reformok. Érdemes megemlíteni, hogy ez összhangban van a leíró elemzésnél megfigyeltekkel, ezek a témák azok, amik jellemzően valamelyik politikai oldal újságíróit jobban foglalkoztatják, mint másokat. A 61,16%-os pontosság ahhoz képest jónak mondható, hogy a véletlenszerű csoportosítás pontossága 25% körül kellene, hogy legyen. Láthatjuk tehát, hogy a topik és a NarrCat értékek együtt jó kiindulási alapot adnak egy szövegklasszifikációs probléma megoldásához. A NarrCat azért is egészíti ki jól a topikmodellt, mert a topikmodellel ellentétben, ahol a szavak sorrendje és ragozása nem számít, a NarrCat hasznosít a szavak sorrendjében, illetve a szóvégződésekben rejlő információt is.

7.2. A kutatás limitációi, jövőbeli kutatási irányok

Kutatásom korlátai közé tartozik, hogy ki kellett hagynom adathiba miatt egy jelentős hírportált, illetve, hogy kerültek a Trianon téma szempontjából nem releváns cikkek a korpuszba, amiket nem volt lehetőségem kiszűrni. További kutatásokban érdemes lenne határontúli magyar hírportálokat, illetve nem csak országos színtű, hanem helyi oldalakat is elemezni.

Modelleim pontosságát lehetne még javítani egyéb változók bevonásával pl. a „Term Frequency-Inverse Document Frequency”, szavak fontosságát mérő algoritmus outputjának felhasználásával, vagy a cikk írójának tulajdonságai, megjelenés idejére és helyére vonatkozó információ bevonásával. Limitációként említhetem továbbá a számítási kapacitások korlátozottságát, hiszen amennyiben több paraméter-kombinációt tudtam volna kipróbálni, lehetséges, hogy pontosabb modellt kaptam volna.

Ami az eredmények értelmezésének korlátait illeti, valójában nem tudok levonni következtetést politikai pártokról, hiszen ezt csak akkor tudnám megtenni, ha közvetlenük ők adnának utasítást arra, hogy miről hogyan írjanak a velük szimpatizáló, hozzájuk köthető oldalak. Így csak annyit tudok elemezni, hogy azok az újságírok, akik ezeknél a hírportáloknál dolgoznak, hogyan, miről írnak. Akkor lehetne pártok kommunikációját, Trianonnal kapcsolatos emlékezetpolitikáját vizsgálni, ha beszédeket, parlamenti felszólalásokat, nyilatkozatokat, pártok közösségi média oldalán megjelent posztokat elemeznénk. Ezen kívül fontos akadálynak éreztem azt is, hogy nem állt rendelkezésemre elegendő kutatás és így referenciaadat NarrCat értékekkel, ezért nem tudtam összehasonlítani, hogy milyennek számítanak az ebben a kutatásban, erre a korpuszra kapott értékek. Eredményeimet fel lehetne például használni arra, hogy a hírportálok és a politikai pártok finomítsanak kommunikációjukon, de további kutatásként érdemes lenne politikusok nyilatkozatait és közösségi médián megjelent posztjait elemezni hasonló módszertannal, hogy egy átfogóbb képet kaphassunk.

A NarrCat véleményem szerint olyan szövegek vizsgálatára megfelelő, ahol érdemes foglalkozni egy pszichológiai dimenzióval, ahol mögöttes információt ki lehet nyerni a szövegből a szöveg írójának érzelmeiről, mentális állapotáról. A topikmodell pedig akkor megfelelő választás, ha a korpuszon belül érdemes megkülönböztetni témákat. Az általam bemutatott, NarrCat + topikmodell + klasszifikáció módszertant tehát olyan esetben érdemes alkalmazni, ahol fennállnak ezek a körülmények és szeretnénk csoportba sorolni a szövegeket.

Az üzleti életben lehetne használni NarrCat és topikmodell adatokat például beérkező motivációs levelek elemzésére, versenytársak hírlevelének, posztjainak, bármilyen egyéb kommunikációjának elemzésére, illetve újságcikkeket is érdemes lehet figyelnie egy vállalatnak, hogy a gazdaság, közélet, trendek változásait azonosítani tudja. Hogy egy lehetséges klasszifikációs megoldást is említsek, a NarrCat és a topikhovatartozás értékeket lehetne használni arra, hogy egy vállalat a beérkező panaszleveleket csoportosítsa téma vagy a termék, ügyosztály érintettsége szerint. Sok internetes ajánlórendszer (pl. könyvesbolt webshopja) többek között topikmodellt is használ, érdekes lenne megvizsgálni, hogy javíthatók lennének-e az ajánlási eredmények a NarrCat használatával.

Természetesen más kutatási témában is használható lenne újságcikkek elemzésénél a dolgozatomban bemutatott módszertan, például lehetne vizsgálni más történelmi eseményt

(pl. Holokauszt) vagy valamilyen aktuális társadalmi kérdés keretezését (pl. tanárok helyzete). Érdekes lenne ezen kérdéseknél megvizsgálni, hogy politikai oldalanként a hozzájuk köthető portálok hogyan írnak ezekről a kérdésekről. A módszertant lehetne továbbá alkalmazni naplók vagy közösségi média posztok elemzésére, megtudhatnánk, hogy mennyire különböző témák érdeklik, és mennyire különbözőképpen írnak az emberek valamilyen csoporthoz tartozástól függően. Ez a csoportképző változó lehet pl. bármilyen demográfiai változó: nem, életkor, foglalkozás, jövedelmi helyzet stb.

Összességében azt gondolom, hogy a NarrCat eszköz és a topikmodellezési algoritmus együttes alkalmazásával egy igazán árnyalt képet tudunk kapni ez elemzett szövegekről és annak írójáról, ezért sok területen hasznos eredményeket lehetne elérni a használatukkal.

Mellékletek

Topik		Aktív ige	Passzív ige	Kényszer	Intenció	Én referencia	Mi referencia	Kogníció	Pozitív érzelem	Negatív érzelem	Pozitív értékelés	Negatív értékelés	Átélő forma	Metanarratív forma	Vissza-tekintő forma	Tagadás
Szabadidő	1	0,0324	0,0058	0,0035	0,0035	0,0057	0,0103	0,0049	0,0026	0,0018	0,0056	0,0042	0,4925	0,1193	0,3882	0,0178
Elbeszélések	2	0,0326	0,0065	0,0038	0,0047	0,0196	0,0181	0,0084	0,0051	0,0048	0,0073	0,0056	0,5788	0,1292	0,2921	0,0310
Második világháború	3	0,0342	0,0066	0,0027	0,0055	0,0034	0,0088	0,0042	0,0030	0,0024	0,0062	0,0050	0,4076	0,1148	0,4776	0,0203
Irodalom	4	0,0290	0,0058	0,0025	0,0042	0,0087	0,0124	0,0049	0,0038	0,0025	0,0064	0,0042	0,5354	0,1441	0,3205	0,0200
Vallás	5	0,0352	0,0052	0,0034	0,0038	0,0058	0,0177	0,0049	0,0052	0,0025	0,0073	0,0028	0,4993	0,1190	0,3816	0,0144
Kormányzati hírek	6	0,0382	0,0062	0,0047	0,0042	0,0024	0,0104	0,0039	0,0025	0,0010	0,0031	0,0030	0,4935	0,0987	0,4078	0,0156
Megemlékezések	7	0,0332	0,0038	0,0025	0,0036	0,0032	0,0194	0,0033	0,0027	0,0027	0,0052	0,0041	0,4771	0,1360	0,3869	0,0140
Gazdaság, környezet	8	0,0296	0,0064	0,0048	0,0049	0,0030	0,0144	0,0049	0,0022	0,0014	0,0053	0,0029	0,5569	0,1184	0,3247	0,0200
Románia	9	0,0331	0,0052	0,0038	0,0055	0,0045	0,0116	0,0051	0,0031	0,0027	0,0056	0,0055	0,5070	0,1243	0,3687	0,0222
Szlovákia	10	0,0317	0,0051	0,0048	0,0055	0,0074	0,0166	0,0068	0,0037	0,0028	0,0055	0,0038	0,5419	0,1216	0,3365	0,0244
Kultúra	11	0,0326	0,0052	0,0026	0,0033	0,0080	0,0125	0,0044	0,0040	0,0015	0,0052	0,0021	0,5290	0,1409	0,3302	0,0119
Évfordulós beszédek	12	0,0290	0,0059	0,0054	0,0059	0,0064	0,0352	0,0065	0,0037	0,0056	0,0070	0,0075	0,5838	0,1405	0,2757	0,0326
Területi változások	13	0,0317	0,0071	0,0038	0,0059	0,0047	0,0107	0,0047	0,0026	0,0024	0,0037	0,0047	0,4665	0,1257	0,4079	0,0240
Európai politika	14	0,0308	0,0055	0,0061	0,0068	0,0044	0,0139	0,0052	0,0026	0,0021	0,0041	0,0050	0,5817	0,1252	0,2931	0,0222
Bűnözés, kirekesztés	15	0,0354	0,0047	0,0033	0,0049	0,0067	0,0150	0,0049	0,0026	0,0040	0,0038	0,0107	0,4986	0,1329	0,3685	0,0235
Történelmi kutatás	16	0,0300	0,0057	0,0036	0,0046	0,0059	0,0103	0,0056	0,0027	0,0019	0,0046	0,0046	0,5540	0,1268	0,3192	0,0219
Belpolitika	17	0,0339	0,0056	0,0044	0,0055	0,0065	0,0114	0,0063	0,0034	0,0030	0,0044	0,0067	0,5415	0,1313	0,3272	0,0272
Emlékhelyek	18	0,0298	0,0066	0,0023	0,0040	0,0039	0,0096	0,0041	0,0027	0,0015	0,0046	0,0027	0,5060	0,1142	0,3799	0,0164
Társadalmi változások, reformok	19	0,0252	0,0061	0,0044	0,0055	0,0079	0,0161	0,0066	0,0029	0,0034	0,0051	0,0074	0,6109	0,1362	0,2529	0,0316
Nemzeti	20	0,0350	0,0064	0,0040	0,0052	0,0039	0,0197	0,0052	0,0030	0,0019	0,0038	0,0044	0,5221	0,1243	0,3536	0,0190
Törvényhozás	21	0,0313	0,0037	0,0055	0,0046	0,0033	0,0081	0,0040	0,0031	0,0023	0,0026	0,0049	0,5304	0,1151	0,3544	0,0207
Átlag		0,0322	0,0057	0,0038	0,0048	0,0060	0,0146	0,0052	0,0032	0,0026	0,0051	0,0048	0,5240	0,1261	0,3498	0,0213

1.sz. melléklet: NarrCat értékek átlaga topikonként. Forrás: Saját szerkesztés.

Topik	Aktív ige	Passzív ige	Kényszer	Intenció	Én referencia	Mi referencia	Kogníció	Pozitív érzelem	Negatív érzelem	Pozitív értékelés	Negatív értékelés	Átélő forma	Meta-narratív forma	Tagadás	R^2 mutató
1	0,8394	0,8667	0,0000	0,0000	0,0000	0,8386	0,0522	0,3972	0,4280	0,0000	0,1542	0,1097	0,9301	0,0000	0,0514
2	0,9690	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0011	0,0000	0,0000	0,0000	0,2870
3	0,0000	0,0814	0,0000	0,0000	0,0685	0,3279	0,0000	0,5555	0,0055	0,0000	0,0000	0,0000	0,0000	0,0000	0,1281
4	0,0000	0,0151	0,0000	0,0000	0,0000	0,2476	0,4113	0,0089	0,1248	0,0000	0,8485	0,0000	0,0000	0,0002	0,0680
5	0,0102	0,0631	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0577	0,0000	0,1246
6	0,0000	0,0164	0,0000	0,2430	0,0000	0,0006	0,0001	0,0062	0,0000	0,0000	0,0000	0,0116	0,0000	0,0000	0,1070
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,8331	0,0000	0,0108	0,1212	0,0000	0,0091	0,0000	0,1302
8	0,1117	0,0000	0,0000	0,7489	0,0000	0,9138	0,0020	0,0000	0,0000	0,6750	0,0000	0,0000	0,2927	0,8148	0,0557
9	0,0199	0,0005	0,2613	0,7144	0,0002	0,1275	0,0017	0,0475	0,0000	0,5431	0,0053	0,0000	0,0299	0,4446	0,0203
10	0,0728	0,0015	0,0274	0,8574	0,3867	0,0000	0,0074	0,0780	0,0002	0,8767	0,0000	0,6726	0,9595	0,0063	0,0246
11	0,0013	0,7694	0,0000	0,0000	0,0000	0,0210	0,0000	0,0000	0,0331	0,4646	0,0000	0,0000	0,0000	0,0000	0,1530
12	0,0000	0,0749	0,8864	0,1469	0,6276	0,0000	0,0160	0,1351	0,0000	0,0000	0,0000	0,0026	0,0000	0,0000	0,2448
13	0,0000	0,0008	0,0572	0,0000	0,0000	0,3321	0,2599	0,0011	0,0005	0,0000	0,4472	0,0000	0,0001	0,0000	0,0626
14	0,0094	0,0406	0,0000	0,0000	0,0000	0,0012	0,1433	0,0000	0,0002	0,0000	0,5270	0,0000	0,0000	0,0214	0,1050
15	0,0000	0,0625	0,0001	0,9046	0,0270	0,0061	0,0000	0,0449	0,0000	0,0004	0,0000	0,1291	0,0000	0,0000	0,1275
16	0,0000	0,0002	0,0297	0,0634	0,1726	0,0000	0,0000	0,0059	0,0000	0,1471	0,6660	0,0000	0,0001	0,9297	0,0355
17	0,0000	0,2598	0,0000	0,0002	0,0000	0,0000	0,0024	0,0450	0,0954	0,0000	0,0000	0,0000	0,0000	0,0000	0,0714
18	0,0349	0,0060	0,0000	0,0000	0,2011	0,6979	0,0000	0,3266	0,5557	0,0792	0,0000	0,0000	0,0000	0,0000	0,0861
19	0,0000	0,0000	0,0368	0,0001	0,0261	0,0000	0,0000	0,0000	0,4327	0,5429	0,0000	0,0000	0,0000	0,0000	0,2437
20	0,0000	0,9639	0,0112	0,3833	0,0000	0,0000	0,8759	0,8981	0,9373	0,0000	0,8541	0,0000	0,0514	0,0000	0,0652
21	0,0000	0,0000	0,0000	0,0023	0,0000	0,0000	0,0000	0,1272	0,0027	0,0000	0,0000	0,0030	0,0000	0,0863	0,0875

Minden szokásos megbízhatósági szinten szignifikáns
95%-os megbízhatósági szinten szignifikáns
90%-os megbízhatósági szinten szignifikáns
Szokásos megbízhatósági szinteken nem szignifikáns

2.sz. melléklet: Béta regressziós modellek változóinak p értékei és az R^2 mutató értéke. Forrás: Saját szerkesztés.

Topik	Intercept	Aktív ige	Passzív ige	Kényszer	Intenció	Én referencia	Mi referencia	Kogníció	Pozitív érzelem	Negatív érzelem	Pozitív értékelés	Negatív értékelés	Átélő forma	Meta-narratív forma	Tagadás
1	-2,86	-0,12	-0,25	-9,53	-11,94	3,37	0,10	-3,19	1,62	-1,50	6,66	-1,90	-0,11	-0,01	-6,00
2	-3,75	-0,02	9,68	-8,94	-11,28	19,40	2,40	11,58	12,22	12,00	13,23	3,67	0,57	0,67	7,94
3	-2,05	-2,86	2,45	-7,95	7,53	-1,37	-0,47	-8,68	-1,13	5,02	6,61	6,41	-1,67	-0,93	2,88
4	-3,16	-3,29	3,36	-18,75	-11,54	6,58	-0,52	-1,25	4,64	2,63	9,20	0,24	0,30	0,58	-2,13
5	-2,86	1,51	-2,67	-9,77	-9,63	3,09	5,45	-7,87	21,72	8,47	12,04	-5,40	-0,29	-0,19	-9,74
6	-2,83	7,04	3,62	13,19	-2,01	-5,40	-1,72	-6,89	-5,54	-12,96	-10,60	-9,16	0,17	-0,63	-6,18
7	-2,23	2,56	-22,48	-6,74	-11,51	-3,66	9,09	-14,30	0,38	15,19	-3,31	1,91	-0,57	-0,25	-11,28
8	-3,18	-0,95	9,54	11,53	0,51	-4,55	0,05	4,81	-14,14	-10,99	-0,56	-9,11	0,50	-0,10	0,13
9	-2,91	1,47	-5,38	1,87	0,63	-2,92	-0,75	-5,30	3,91	9,21	-0,87	3,72	-0,32	0,22	-0,47
10	-3,12	1,08	-4,65	3,43	0,29	-0,63	3,27	4,18	3,25	6,55	0,21	-8,53	0,03	-0,01	-1,61
11	-2,97	2,02	-0,45	-16,22	-12,01	6,52	1,14	-7,69	12,95	-4,22	1,02	-9,30	0,61	0,64	-16,59
12	-3,56	-3,45	2,35	0,20	2,13	-0,32	12,53	3,38	2,50	17,48	13,43	10,42	0,18	0,98	5,94
13	-2,41	-2,50	4,62	-3,01	8,66	-4,34	0,45	-1,78	-6,21	6,13	-6,84	-0,96	-1,00	-0,38	5,38
14	-3,65	1,48	-2,83	18,30	17,20	-7,02	1,37	2,18	-9,87	-6,51	-5,62	0,75	0,84	0,59	1,24
15	-3,31	3,31	-2,58	-6,14	0,19	1,52	-1,21	-6,15	-3,66	10,74	-4,68	27,54	-0,09	0,38	3,60
16	-3,33	-2,60	5,00	-3,30	-2,90	0,94	-2,34	6,21	-4,98	-7,27	-1,88	-0,52	0,69	0,38	0,05
17	-3,62	6,85	-1,73	8,14	6,35	-5,96	-3,45	4,93	-4,00	-3,17	-6,84	12,93	0,55	0,77	5,79
18	-2,35	-1,31	4,03	-19,68	-8,92	0,98	-0,19	-9,61	1,91	-1,14	2,40	-11,93	-0,33	-0,44	-4,85
19	-3,71	-7,59	8,66	2,97	5,83	1,44	-2,34	10,48	-10,83	1,23	0,77	12,09	1,26	0,75	10,31
20	-3,15	4,73	-0,07	3,93	1,41	-4,87	6,60	0,25	-0,24	-0,14	-8,26	-0,23	0,38	0,19	-6,56
21	-2,72	2,97	-15,42	14,14	4,85	-6,64	-4,56	-7,68	-2,90	-5,53	-17,26	5,55	-0,19	-0,46	0,99

3.sz. melléklet: Béta regressziós modellek koefficiensei. Forrás: Saját szerkesztés.

Irodalomjegyzék

Átló (2022). A kormánypárti hírmédia. https://atlo.team/igy-nez-ki-a-kormanyparti-hirmedia/ Letöltés ideje: 2022.07.13.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D. M. (2012) Probabilistic topic models. Communications of the ACM, 55(4): pp.77_84.

Barde, B. V., & Bainwad, A. M. (2017). An overview of topic modeling methods and tools. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 745-750). IEEE.

Barna, I., & Knap, Á. (2022). Analysis of the Thematic Structure and Discursive Framing in Articles about Trianon and the Holocaust in the Online Hungarian Press Using LDA Topic Modelling. Nationalities Papers, 1-19.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces (pp. 74-77).

DeBarr, D., Ramanathan, V., & Wechsler, H. (2013). Phishing detection using traffic behavior, spectral clustering, and random forests. 2013 IEEE International Conference on Intelligence and Security Informatics, 67-72.

Edaño, M. G. E., Gonzales, R. J., Laguda, R. C. B., & De Goma, J. C. Stressor Classification of Filipino Political Tweets Using LDA, SVM, XGBoost, Logistic Regression. International Conference on Industrial Engineering and Operations Management. Istanbul, Turkey

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., ... & Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy conversion and management, 164, 102-111.

Ferenczhalmy, R., Szalai, K., & László, J. (2011). Az ágencia szerepe történelmi szövegekben a nemzeti identitás szempontjából. Pszichológia, 31(1), 35-46.

Fülöp, É., Péley, B., László, J. (2011). Historical trajectory emotions in historical novels, Pszichológia, 31 (1), pp. 47-61

Hackerearth (2022). Beginners Tutorial on XGBoost and Parameter Tuning in R. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/ Letöltés ideje: 2022.10.01.

Hargitai, R., Naszódi, M., Kis, B., Nagy, L., Bóna, A., & László, J. (2007). Linguistic markers of depressive dynamics in self-narratives: Negation and self-reference. Empirical Text and Culture Research, 3, 26-38.

Huang, Y., Wang, R., Huang, B., Wei, B., Zheng, S. L., & Chen, M. (2021). Sentiment classification of crowdsourcing participants' reviews text based on LDA topic model. IEEE Access, 9, 108131-108143.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Janky, B., Kmetty, Z., & Szabó, G. (2019). Mondd kire figyelsz, megmondom mit gondolsz!. Politikatudományi Szemle 28. évf. 2. sz. p.7-33

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. Journal of statistical software, 15, 1-28.

Kovács, R. R. (2012). A házastársi kapcsolat, a párkapcsolati elégedettség narratív pszichológiai tartalomelemzéssel és pszichometriai eszközökkel történő vizsgálata, tekintettel a családi életciklusokra. Doktori (PhD) értekezés

László, J., Csertő, I., Fülöp, É., Ferenczhalmy, R., Hargitai, R., Lendvai, P., Péley, B., Pólya, T., Szalai, K., Vincze, O. & Ehmann, B. (2013). Narrative language as an expression of individual and group identity: The narrative categorical content analysis. Sage Open, 3(2), 2158244013492084.

László, J., Fülöp, É. (2010). A történelem érzelmi reprezentációja a történelemkönyvekben és naiv elbeszélésekben. Emotional Representation of History in History Textbooks and in Lay Narratives). Történelemtanítás, 1(3).

Liu, L., Lu, Y., Luo, Y., Zhang, R., Itti, L., & Lu, J. (2016). Detecting" smart" spammers on social network: A topic model approach. arXiv preprint arXiv:1604.08504.

McAdams, D. P. (2001). The Psychology of Life Stories. Review of General Psychology, 5(2), 100–122.

Narratív Pszichológia PTE, Intenció modul, 2022.
https://narrativpszichologia.pte.hu/hu/tartalom/intencio_modul
Letöltés ideje: 2022.07.13.

Narratív Pszichológia PTE, Társas referencia és tagadás modul, 2022.
https://narrativpszichologia.pte.hu/hu/tartalom/tarsas_referencia_es_tagadas_modul
Letöltés ideje: 2022.07.13.

Narratív Pszichológia PTE, Kognitív modul, 2022.
https://narrativpszichologia.pte.hu/hu/tartalom/kognitiv_modul
Letöltés ideje: 2022.07.13.

Németh, R., Koltai, J. (2021). The potential of automated text analytics in social knowledge building. In Pathways Between Social Science and Computational Social Science (pp. 49-70). Springer, Cham.

Osgood, C. E., & Walker, E. G. (1959). Motivation and language behavior: a content analysis of suicide notes. The Journal of Abnormal and Social Psychology, 59(1), 58.

Pólya, T., Kis, B., Naszódi, M., & László, J. (2007). Narrative perspective and the emotion regulation of a narrating person. Empirical Text and Culture Research, 7(3), 50-61.

Polya, T., Laszlo, J., & Forgas, J. P. (2005). Making sense of life stories: The role of narrative perspective in perceiving hidden information about social identity. European Journal of Social Psychology, 35(6), 785-796.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

Szabó, G., & Bene, M. (2016). Széttöredezett vagy összekapcsolódó? A magyar médianyilvánosság hálózatszerkezete három eset tükrében. Politikatudományi Szemle, 25(3.), 33-58.

Szalai, K. (2011). Az ágencia nyelvi jegyei: Az aktív és passzív igék szerepe a narratívumokban.

Vincze, O., Ilg, B., & Pólya, T. (2013). The role of narrative perspective in the elaboration of individual and historical traumas.