Internetes hírek automatikus osztályozása

Osváth Mátyás[1], Héja Enikő[1]

[1] HUN-REN Nyelvtudományi Kutatóközpont, 1068 Budapest, Benczúr u. 33.
{osvath.matyas, heja.eniko}@hun-ren.nytud.hu

Kivonat: Az elmúlt évtizedekben az online sajtó vált a hírek egyeduralkodó forrásává, amely egyre növekvő igényt teremt az automatizált tartalomkategorizálási módszerek iránt. Tudomásunk szerint ez az első olyan kutatás, amelynek célja magyar hírek téma szerinti osztályozása nagy nyelvi modellekkel. Jelen vizsgálat célja az online elérhető magyar hírekben megjelenő témák azonosítása és a hírek szövegeinek osztályozása. Az elemzés anyagát 77 magyar hírportál 5,225,195 szövege adta 2013-tól 2023-ig. A kategória-készlet elkészítésére két módszert vizsgáltunk meg: egy statikus szóbeágyazásokon alapuló módszert, valamint egy BERT alapú témamodellezést. A két módszer eredményeit manuálisan validáltuk, így jött létre a 19 elemből álló kategóriakészletet. A hírek osztályozásához három módszert alkalmaztunk: egy már létező címkeajánló rendszer mellett a nagy méretű XLM-RoBERTa és a PULI LlumiX 32K Instruct modell performanciáját vizsgáltuk, ez utóbbit zero-shot tanítással. Bár az eredmények nagy szórást mutattak, azt gondoljuk, hogy a PULI LlumiX 32K Instruct modell kis munkával továbbfejleszthető egy nagy pontosságú osztályozóvá.

1 Bevezetés

1.1 Célkitűzés

Az elmúlt évtizedekben a nyomtatott sajtó visszaszorulásával az online média vált a hírek egyeduralkodó forrásává. A jelenleg magyarul online elérhető hírek szövegeinek mérete már több milliárd tokenre rúg. Ez a tény illetve, hogy napi szinten keletkeznek újabb és újabb hírek, szükségessé teszi, hogy a híreket tartalmuk szerint kategorizálni tudjuk.

Amennyire tudjuk, ez az első olyan kutatás, amelynek célja magyar hírek téma szerinti osztályozása. Kutatásunkhoz legközelebb talán a sajtószövegek automatikus tematikus címkézését célzó címkéző algoritmus áll (Yang és mtsai, 2020), amely a hírszövegek nagy mennyiségű, manuálisan létrehozott címkéiből indult ki.

Így a jelen kutatás célja kettős: egyfelől egy nagy fedésű, ám viszonylag átlátható, jól definiált kategória rendszert kívánunk kialakítani, amely az osztályozás alapjául szolgálhat, másfelől a hagyományos gépi tanulásos osztályozó algoritmusokkal szemben (tartóvektor-gép, naiv-Bayes, véletlen erdő) azt kívánjuk vizsgálni, hogy az

újabb, nagy nyelvi modelleken alapuló eljárások milyen pontossággal képesek az osztályozás elvégzésére. Az elemzés anyagát kb. 80 magyar nyelvű hírportálon 2013 és 2023 között megjelent kb. 5 millió hír adta. A kategória-készlet elkészítése során két automatikus módszert használtunk. Az első módszer statikus szóbeágyazásokon alapult: az URL-ek releváns szegmenseinek word2vec reprezentációiból gráfokat építettünk, majd ezen gráfok részgráfjait vizsgálva automatikusan nyertük ki a kategória jelölteket. A második módszer a BERT alapú témamodellezés volt, ahol a szövegek vektortérbeli beágyazása után dimenziócsökkentést és klaszterelemzést alkalmaztunk a témák kinyerésére. A következő lépésben az automatikusan kinyert témákat manuális validálás során kategóriákba szerveztük. A végső kategóriakészlet a két módszer eredményeinek manuális egyesítésével jött létre. A kutatás második szakaszában 500 db véletlenszerűen kiválasztott hírt soroltunk be a végső 19 kategória alá. A hírek osztályozása 3 féle módon történt. Először a már említett automatikus neurális címkeajánló modell használatára került sor. A második és harmadik kísérlet során két különböző modellt használtunk: egyfelől a Meta által kiadott nagy méretű XLM-RoBERTa-t, másfelől a PULI LlumiX 32K Instruct modellt zero-shot tanítással. Eredményeink azt mutatják, hogy a PULI LlumiX 32K Instruct megközelíti a címkeajánló pontosságát, míg az XLM-RoBERTa mindössze 18% pontosságot produkál.

1.2 Szakirodalmi háttér

Tudomásunk szerint a hazai kutatások elsősorban a témamodellezésre koncentrálnak. Ezen kutatások főként Látens Dirichlet Allokációt (LDA) alkalmaztak (pl. Barna és mtsai, 2023 és Vági és mtsai 2023). De olyan kutatásokról is tudunk, amelyek BERT alapú témamodellezést használtak (pl. Nagy, 2024 és Gelei és mtsai 2023). Általánosságban elmondható, hogy a szerzők témamodellezésre használt módszereket alkalmaznak egy adott domain-specifikus korpuszra (pl. politika, vállalkozás, értékelések), melyek az így keletkezett témákkal leírhatóvá válnak. Azonban eddig nem történt olyan vizsgálat, amely a témamodellezésre szolgáló technikákat összehasonlította volna, sőt olyan magyar vizsgálatról sem tudunk, amely magyar szövegek téma szerinti osztályozásával foglalkozott volna. Kutatásunkhoz legközelebb talán a sajtószövegek automatikus tematikus címkézését célzó algoritmus áll (Yang és mtsai, 2020), amely a HVG nyomtatott, illetve online verzióinak manuálisan létrehozott címkéiből indult ki. Mivel ezek a címkék általában nem követnek központi szerkesztési elveket, hanem újságírók hozzák létre őket eseti alapon, a címkekészlet nagyon nagy, és sokszor átgondolatlan. A HVG esetében Yang és mtsai több mint 75.000 címkét használtak.

Az angol nyelvű hírek esetében már számos olyan adatbázis áll rendelkezésünkre, ahol a hírek kategóriákba vannak szervezve. Ezek közül talán a legrégebbi és legszélesebb körben használt a 20 Newsgroups Dataset, amely kb. 20,000 hírt tartalmaz, amelyek megközelítőleg egyenletesen oszlanak meg 20 különböző hírkategória között. A híradatbázis érdekessége, hogy bár egészen a legutóbbi időkig intenzíven használták

osztályozási feladatok kiértékelésére[1], amennyire tudjuk nem áll rendelkezésre részletes dokumentáció sem a híranyag gyűjtéséről, sem a kategóriák kialakításáról, sőt még a szerző sem biztos[2]. Egy további probléma a kategóriák kapcsán jelentkezik: bizonyos hírosztályok nagyon szorosan összetartoznak, míg mások sokkal távolabb vannak egymástól, és így jobban elkülöníthetőek. Egy másik ismert híradatbázistól, az AG's corpus of news articles[3]-től eltekintve, a kategorizált híradatbázisok jellemzően egy hírportál anyagait tartalmazzák, és kategóriarendszerük vagy kis számú, előre definiált elemből áll, például a BBC News (Bose, 2019) 5 előre definiált kategóriát használ, vagy az adott hírportál kategóriát használják fel, például a Guardian News (Kharosekar, 2023), de olyan is van, ahol nem derül ki egyértelműen, hogy a kategóriák hogyan kerültek kialakításra, például a multimodális N24News (Wang és mtsi, 2022), amely a New York Times híreit sorolja 24 kategóriába. A fenti megközelítések nemcsak az alkalmazott kategóriák számában térnek el, hanem abban is, hogy az osztályokba sorolt cikkek száma kategóriánként egyenletes eloszlást mutat-e vagy sem. A fentiek alapján azt mondhatjuk, hogy munkánk újszerűnek tekinthető, hiszen célja egy olyan általános kategóriarendszer részben adatvezérelt kidolgozása, amely portáltól függetlenül lehetővé teszi a hírek besorolását.

2 Munkafolyamat

2.1 Adatgyűjtés

A kutatás anyagát 5.225.195 online elérhető magyar cikk anyagából állt össze, amely a nyilvánosan elérhető Common Crawl[4] 2013 és 2023 közötti magyar nyelvű alkorpuszának a részét képezte. Az anyagot úgy állítottuk össze, hogy az URL-ekben leggyakrabban szereplő forrás domainekből kézzel kiválogattuk a híroldalakat – összesen 77-et. Az 1. Táblázat mutatja a 10 leggyakoribb forrásdomainhez tartozó URL-ek számát.

1. Táblázat: a 10 legyakoribb forrás domain és az azokhoz tartozó URL-ek száma

#	Forrás domain	URL-ek száma
1	index.hu	654.583
2	24.hu	547.370
3	hvg.hu	301.694
4	prohardver.hu	191.579
5	velvet.hu	182.266
6	444.hu	175.886
7	femina.hu	167.298

[1] Pl. https://paperswithcode.com/sota/text-classification-on-20news

[2] Általában Lang (1995)-nek tulajdonítják, de a cikk nem említi expliciten ezt az adatbázist (vö, http://qwone.com/~jason/20Newsgroups/)

[3] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

[4] https://commoncrawl.org/

8	rangado.24.hu	153.913
9	starity.hu	131.728
10	infostart.hu	122.638

2.2 A kategóriarendszer kialakítása

A kutatás első részében statikus, illetve kontextuális szóbeágyazásokat is használtunk a kategóriakészlet kialakítására. A végső kategóriakészlet ezek összevonásából született meg a manuális validálás során.

A kategória-készlet kialakítása statikus szóbeágyazásokkal

A statikus szóbeágyazásokon végzett kísérleteink azon az előfeltevésen alapultak, hogy a hírportálok esetében az URL-ek struktúrája tükrözi, hogy az adott sajtóorgánum milyen kategória alá sorolta be az adott hírt. Így az URL-ek koherensebb képet nyújtanak a kategóriákról, mint a címkék , hiszen ez utóbbiak sok esetben egyedi döntés nyomán születnek meg.

A munkaszakasz az alábbi lépéseket követte:

(1) Az összes URL-ből a megfelelő URL szakasz kivágásával létrehoztuk a kategória-jelölteket. A kategória-jelölteket a forrásdomain után közvetlenül követő legközelebbi két '/' közötti karaktersorozatként azonosítottuk.

(2) A következő lépésben a kategória-jelölteket automatikusan ékezetesítettük (Laki és Yang 2020), melynek eredményeképpen 1234 kategóriajelöltet kaptunk (pl: *önidő*, *önleány*, *öregség*, *örökhagyás*, *örülünkvincent*).

(3) Ebben a lépésben, word2vec reprezentációt rendeltünk a kategória-jelöltekhez. Ehhez a NYTK-word2vec[5] modellt használtuk. Sajnos 263 esetben nem állt rendelkezésünkre word2vec reprezentáció (pl. *whistlermamája*, *dokuexport*, *tvtorrent*, *asparhelttitkai*), így ezeket az URL töredékeket nem vettük figyelembe. Ezek jelentős része kitalált szóösszetétel volt.

(4) Héja és Ligeti-Nagy (2023) cikkben leírt módszertan alapján készítettük el a maradék 969 kategória-jelölt szomszédsági mátrixát, amelyet egy alkalmas küszöbérték (0.5) mentén binarizálva egy hurokél-mentes, irányítatlan G gráffá konvertáltunk. A G gráf 386 izolált csúcsot tartalmazott, tehát összesen 583 URL-fragmentum szolgált a kategóriák alapjául.

(5) Majd megvizsgáltuk, hogy a G gráfban vannak-e olyan részgráfok, amelyek kategória-jelölteket hírkategóriáknak megfelelő szemantikai osztályokba szervezik.

A kézi kiértékelés során összefüggő részgráfokat, teljes részgráfokat (klikkek) és k-klikk közösségeket vizsgáltunk (k=3). Az összefüggő komponensek vizsgálata során azzal a problémával szembesültünk, hogy mindig van egy óriás komponens, amely több

[5] https://nlp.nytud.hu/word2vec/cbow_3.tar.gz

kategóriát is összevon, másfelől pedig a kisebb komponensek összetartozó kategóriákat szabdalnak fel (pl. {*unicredit*, *erste*, *mkb*} és {*raiffeisen*, *cib*}).

A klikkek kiértékelése során azt találtuk, hogy a teljes részgráfok "túl érzékeny" struktúrák, vagyis ha egy teljes részgráfból valami miatt egy él hiányzik, a teljes gráf több klikkre esik szét a kategóriák indokolatlan szétdarabolását eredményezve (pl. {'2020', '2021', '2019', '2023'} és {'2021', '2022', '2023'}.

A klikkhez közel álló, ám stabilabb reprezentációt nyújtó struktúra a k-klikk közösség, amely kevésbé érzékeny az élek elhagyására. Hiszen egy k-klikk közösség azon k méretű klikkek uniója, amelyek elérhetők szomszédos (k-1 csúcsot megosztó) k-klikkeken keresztül.

Azt találtuk, hogy a k-klikk közösségek (k=3) eredményezik a legjobb szemantikai osztályokat. A 386 releváns URL szegmensből 352 került 3-klikk közösségekbe 37 csoportosítást eredményezve. Sajnos a 37 csoportból 2 szemantikailag inkoherens, feltehetőleg kisebb koherens csoportok összevonásából származik, így végül 35 szemantikailag koherens 3-klikk közösséget kaptunk, amelyek összesen 196 URL-szegmenst tartalmaztak. A 2. Táblázatban mutatunk be néhányat az automatikusan létrejött kategóriákból:

2. Táblázat: Néhány k-klikk közösségekkel létrehozott szemantikai osztály

#	Szemantikai osztályok
1	{'akciók', 'nyereményjátékok', 'promóció'}
2	{'akciók', 'programok', 'rendezvények'}
3	{'frizura', 'smink', 'szexi'}
4	{'allergia', 'cukorbetegség', 'gyermekbetegségek', 'kórkép', 'laktózérzékenység', 'savtúltengés'}
5	{'felhasználó', 'fogyasztó', 'szolgáltatás', 'termék', 'vásárló'}
6	{'baba', 'gyerek', 'kölyök', 'poronty'}
7	{'bank', 'deviza', 'devizaárfolyamok', 'hitelek', 'kötvény', 'pénzpiacok', 'részvény', 'részvényárfolyamok', 'állampapír', 'árfolyamok'}
8	{'menedzserek', 'szakértők', 'tervezők'}
9	{'blogger', 'celeb', 'klasszis', 'riporter', 'stílusikon', 'sztár', 'sztárok', 'sztárvilág', 'világsztár'}
10	{'nb1', 'nb2', 'nb3', 'rájátszás', 'tabella', 'teremfoci'}

Végül a létrejött szemantikai osztályokat kézzel validáltuk, melynek során bizonyos szemantikai osztályokat összevontunk, másokat pedig irrelevánsnak ítéltünk a feladat szempontjából, így ezeket elhagytuk.

A kategória-készlet kialakítása BERT témamodellezéssel

Másodszor, a hírekben rejlő témák és kategóriák kinyerésére a BERT alapú témamodellezést is alkalmaztunk. A témamodellezés egy nem felügyelt gépi tanulási módszer, amelyet dokumentumokban látens témák identifikálására használnak. Jelen kutatásban Grootendorst (2022) által javasolt módszert követtük az alapbeállításokkal együtt, amely az alábbi lépésekből áll. A szövegek vektortérbeli beágyazására a BERT

modell (Devlin és mtsai, 2019) variánsait használtunk. A Nemeskey (2020) által létrehozott első magyar nyelvű BERT modell, a. huBERT, és a Yang és munkatársai (2023) által fejlesztett 345 millió paraméteres PULI BERT modellt.

Következő lépésben dimenziócsökkentésre került sor az UMAP algoritmust felhasználva. Ezután egy sűrűség alapú hierarchikus klaszterezési algoritmust, a HDBSCAN-t használtuk a klasztereket létrehozására, azzal a megkötéssel, hogy egy klaszter minimum15 szót tartalmazhat, ami az a minimális elemszám, amely egy csoport klaszterként való azonosításához szükséges; az ennél kisebb elemszámú csoportokat az algoritmus zajként azonosítja. A módszer egyik fontos előnye, hogy nem szükséges előre meghatározni a klaszterek számát. A témák azonosításához c-TF-IDF (class-based TF-IDF) módszerrel végeztünk témakiválasztást. Az utolsó lépésben kézzel validáltuk az eredményeket.

2.3 Osztályozás

A kutatás második részében háromféle osztályozót alkalmaztunk a hírek kategóriákhoz való hozzárendelésére. A hírekből véletlenszerű mintavétellel választottunk ki 500 darabot, és ezeket a híreket manuálisan a kategóriákhoz rendeltük.

Először a Bevezetésben már említett automatikus neurális címkeajánló modell használatára került sor (Yang és mtsai, 2020). A modell címkéket rendel a bemeneti hírekhez a hozzájuk tartozó valószínűségi értékekkel. Mivel a modell címkekészlete és a jelenlegi kategóriarendszer között csak nagyon kicsi az átfedés (több, mint 75.000 címke vs. 19 kategória), ezért minden olyan prediktált kategóriát helyesnek tekintettünk, ahol a címke az általunk létrehozott kategóriába tartozott (pl. a pénz címke a gazdaság kategóriába tartozik).

Másodszor, egy előtanított többnyelvű RoBERTa modell továbbfejlesztett változatát használtunk osztályozásra az általunk kialakított kategóriák valószínűségeinek a meghatározásához zero-shot tanítási környezetben (Liu és mtsai, 2019). A Meta által kiadott, nagyméretű XLM-RoBERTa modellt alkalmaztuk, amely 2,5 TB méretű adaton lett előtanítva. A tanító adat 100 nyelvet tartalmazott, köztük magyart. A modell BERT architektúrán alapul és 560 millió paraméterrel rendelkezik (Conneau és mtsai, 2020). Az osztályozás finomhangolás nélkül történt, mivel még nincs megfelelő finomhangolásra alkalmas adathalmaz a magyarra.

Harmadszor egy generatív nyelvmodellt használtunk, a Nyelvtudományi Kutatóközpont által fejlesztett PULI LlumiX 32K Instruct modellt (Yang és mtsai, 2024) zero-shot tanítási környezetben. A modell azt a feladatot kapta, hogy az előre definiált kategóriák listájából a legvalószínűbb témát jósolja meg. Az alábbi prompot használtuk:

```
Milyen kategóriába tartozik az alábbi szöveg: {szöveg}.
Egy szóval válaszolj! A lehetséges kategóriák:
{kategóriák}
```

3 Eredmények

3.1 Kategóriarendszer elkészítése

A statikus szóbeágyazások esetében a kézi validálás eredménye a következő 22 téma lett: *autó*, *motor*, *sport*, *életmód*, *történet*, *bulvár*, *szórakozás*, *média*, *társadalom*, *oktatás*, *gazdaság*, *kultúra*, *politika*, *pszichológia*, *tudomány*, *világ*, *egészség*, *betegség*, *ünnepek*, *család*, *természet* és *táplálkozás*.

A huBERT és PULI BERT által feldolgozott szövegek nagyságrendileg azonos, 4.732 és 4.741 témát adtak eredményként. Ezután, a kapott témák jól, közepesen és rosszul definiált kategóriákra való szétbontására került sor manuálisan, mivel az algoritmus hajlamos nehezen értelmezhető vagy inkoherens témákat eredményezni (Grootendorst, 2022). Egy téma jól-definiált, ha kizárólag jól formált szavakat tartalmazott és szemantikailag is koherens volt. Például az egyik hazai ellátást leíró téma az alábbi kulcsszavakat adta: *ellátás*, *ellátó*, *ellátások*, *ellátáshoz*, *egészségügy*, *sürgősségi*, *családtámogatás*, *gyermekegészségügy* és *üzemorvos*. A közepesen vagy rosszul definiált témák szavai hasonló részsztringeket tartalmaznak, de szemantikailag nem koherensek. Például az egyik téma az alábbi kulcsszavakat foglalta magában: *hálózat*, *halott*, *hálószoba*, *hálólaj* és *halottasház*. A modellek által adott kimenetek kiértékelése során érdekes megfigyelés volt, hogy a két eltérő BERT modell által detektált témák nagyfokú átfedést mutattak: számos témát tehát mindegyik BERT modell felismert. Összesen 440 jól-definiált témát sikerült azonosítani (vagyis a témák kb. 10%-a volt jól-definiált). A 440 téma közül a 3. táblázatban 10 téma kerül bemutatásra. A hírek közt számos téma megjelent, köztük időjárás, közszereplők, cégek, szociális média, turizmus és asztrológia. A témák további differenciálása kérdéses.

3. Táblázat: Azonosított témák és kulcsszavak

#	Kategória	Kulcsszavak
1	Közszereplő	Orbán Viktor, Trump, Merkel, Hugh Jackman, John Kennedy
2	Szociális média	Instagram, Facebook, Netflix, Youtube
3	Cég	Apple, Coca cola, Disney, Ikea, Lego, Unicef, IBM, NVIDIA, Spotify
4	Időjárás	meteorológia, meteor, őszi, havazás, eső, riasztás, előrejelzés
5	Turizmus	turisztika, szezonális, húzóágazat, kormánybiztos, belföld, külföld, árbevétel
6	Foci	magyar, foci, futball, futballista, liga, fc, ferencváros, nemzetközi foci
7	Asztrológia	csillagjegy, horoszkóp, szex horoszkóp, pénz horoszkóp, kínai horoszkóp
8	Ezotéria	vénusz, rejtély, férfi, erogén zóna

9	Szex	szextrükk, szexfüggőség, szexvarázsló, szexpozíció, szexpóz, szerelem és szex
10	Kritika	filmkritika, cinemátrix, pókverzum, romantikus, vadidegen, avengers, színdarab, paródia

A statikus szóbeágyazásokon alapuló és a BERT témamodellezés során is azt találtuk, hogy számos téma csoportosítható egy nagyobb, felettes kategóriába, így mindkét módszer esetében a szemantikailag összetartozó témákat kategóriákba vontuk össze. A megbízhatóság növelése érdekében a kategóriák iteratív megbeszélések révén kerültek kiválasztásra. Összességében, a szóvektorok által adott témák definiáltsága megelőzte a témamodellezését: az értelmezhető kategóriák aránya meghaladta a BERT alapú témamodellezését, de a témamodellezés eredményeként új kategóriák is létrejöttek. A végleges kategóriarendszer a két módszerrel előállított témák metszetéből került előállításra. A 4. táblázatban található a 19 kategória és a hozzájuk tartozó témák.

A hírekben azonosított témák és kategóriák a magyar média által hasznosnak vélt, rendszeresen előforduló témáit mutatja be. A témamodellezés egyik hátránya, hogy feltételezi, hogy egy hír egy témáról szól, ami nem feltétlenül teljesül, és emiatt a hírekben rejlő egyedi témák felismerése problémákba ütközhet.

4. Táblázat: Témák összevonása kategóriákba

#	Kategória	Témák
1	Média	Belföld, külföld, klímaváltozás, időjárás, természeti katasztrófa, büntetés, baleset, építkezés, magazin
2	Bulvár	Sztárok, celebek, hírességek, életrajz, közszereplő, horoszkóp, ruházat
3	Ünnepek	Húsvét, karácsony, születésnap
4	Politika	Elnökválasztás, parlament, jobboldal, baloldal, demokrácia, Fidesz, választás, tiltakozás, sztrájk, miniszterelnök
5	Technológia	bio-tech, cpu, hardver, notebook, internet, robot, technológiai áttörés, telefon
6	Tudomány	Biológia, nobel-díj, űrkutatás, agrártudomány, univerzum
7	Kultúra	Zene, mozi, film, fesztivál, múzeum, fotózás, képregény, könyv, sorozat, művész
8	Gazdaság	Bank, pénz, tőzsde, segély, vállalat, részvény, kötvény, befektetés, biztosítás, hitel, kata, ingatlan, olaj, adomány, babaváró
9	Család	Szülők, anya, gyerek, baba, fiatalok, örökbefogadás, házasság, szülés
10	Egészség és életmód	Vitamin, vegan, vega, gluténmentes, táplálkozás, szépségápolás, fogyókúra, bodypozitivitás, dohányzás, testmozgás, immunrendszer, fogyókúra

11	Betegség	Covid, koronavírus, tünetek, vírus, melanoma, leukémia, cukorbetegség, allergia, influenza
12	Természet	Vulkán, óceán, természeti katasztrófák
13	Oktatás	Felsőoktatás, közoktatás, egyetem, iskola, érettségi, tudományegyetem
14	Világ	Európa, Amerika, Afrika, Törökország, Britannia, Hollywood, Izrael, Ukrajna, Oroszország
15	Sport	Foci, kézilabda, kosárlabda, jégkorong, válogatott, tenisz, világbajnokság, olimpia, stadion
16	Járművek, közlekedés	Autó, motor, bmw, mercedes, légi közlekedés
17	Társadalom	Lakás, élelmiszer, rasszizmus, transzneműség, homoszexualitás, migráció, munka, munkanélküliség, áldozat
18	Pszichológia	Párkapcsolat, depresszió, terápia, molesztálás, stressz, szakítás, álomfejtés
19	Szórakozás	Póker, szerencsejáték, gaming

3.2 Osztályozás

A hírek osztályozásához három különböző módszert alkalmaztunk egy véletlenszerűen összeállított 500 hírt tartalmazó mintán. Az automatikus neurális címkeajánló modell érte el a legnagyobb pontosságot: a hírek 79,7%-át osztályozta helyesen.

A PULI-Llumix 32k Instruct modell szintén figyelemre méltó eredményeket produkált zero-shot tanítással, hiszen a hírek 70,1%-át osztályozta helyesen, míg az XLM-RoBERTa mindössze 18,4%-os pontosságot ért el. Az alacsony pontosság hátterében áll, hogy nem találtunk magyarul elérhető adathalmazt a modell finomhangolásához.

Ugyanakkor az automatikus címkeajánló rendszer hátránya, hogy a mindössze 500 hírhez 215 különböző címkét használt fel, ami önmagában nem teszi lehetővé a hírek tartalmára vonatkozó széleskörű általánosítást. A címkék vizsgálata során azt találtuk, hogy számos címke redundáns vagy túlságosan specifikus volt ahhoz, hogy a helyes kategóriába sorolható legyen. A rendszer nagy mennyiségű, nem egységesített címkehalmazon került tanításra, ami magyarázhatja felhasznált címkék számosságát.

4 Következtetés

A magyar hírek automatikus kategorizációja a tartalmak szervezését és könnyebb hozzáférését segíti, hozzájárulva az információs ökoszisztéma jól-szervezettségéhez. Jelen tanulmány során 77 magyar nyelvű hírportál 2013 és 2023 között keletkezett híreit vizsgáltuk. A későbbi osztályozás alapjául szolgáló kategóriarendszer kialakításához egyfelől statikus szóbeágyazásokból épített gráfstruktúra részgráfjait elemeztük,

másfelől témamodellezést alkalmaztunk két különböző BERT alapú modellel. Ez a munkaszakasz a keletkezett csoportok manuális validálásával zárult, melynek során 19 kategória került kialakításra.

A második munkaszakaszban három neurális módszert vizsgáltunk abból a szempontból, hogy milyen pontossággal képesek a hírek osztályozására: egy már létező neurális címkeajánló modellt, egy BERT alapú többnyelvű modellt, valamint egy generatív utasításkövető modellt. A modellek által kimenetként adott témák kiértékeléséhez emberi munka volt szükséges. Az osztályozásra használt modellek közül a feladatspecifikus címkeajánló rendszer adta a legjobb eredményt.

Kutatásunk jelentősége, hogy egy nagy fedésű, átlátható és jól definiált kategória rendszer kialakításával hatékony módszert kínál a hírek automatikus osztályozására.

Ugyanakkor láttuk, hogy egy jelenleg is létező, elsősorban más célokat szolgáló címkéző egyelőre pontosabb eredményeket ér el a hírek osztályozásában, mint az általunk megvizsgált módszerek. Így a jelen vizsgálat egy kezdeti lépésként értékelendő, amelyet a jövőben tovább kívánunk fejleszteni. A kódbázis és osztályozásra szolgáló annotált korpuszt ennek részeként tervezzük nyilvánosságra hozni. Elsődleges célunk a jövőben annak a vizsgálata, hogy a state-of-the-art nyelvmodellek milyen eredményeket érnek el az osztályozási feladaton valamint, hogy a különböző tanítási szcenáriók és a modellek különböző beállításai milyen hatással vannak az eredményekre. További irány a klaszterelemzés által adott eredmények vizuális értelmezése, címkék eloszlásának értelmezése és a különböző modellteljesítményt mérő metrikák felhasználása. Összességében kiderült, hogy a kialakított címkekészlet és osztályozás alkalmasnak tűnik arra, hogy egy nagyobb hírkorpusz híreit automatikusan címkézzük. Remélhetőleg a kutatásunk a közeljövőben hozzájárul majd az automatikus szövegelemzési technológiák fejlődéséhez, különös tekintettel a nagy szöveges adathalmazok kezeléséhez és értelmezéséhez.

Köszönetnyilvánítás

A kutatást az MTA "Tudomány a Magyar Nyelvért Nemzeti Program" támogatta.

Bibliográfia

Barna I, Knap Á. 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. 2023;51(3):603-621. doi:10.1017/nps.2021.67

Bose, Bijoy. BBC News Classification. https://kaggle.com/competitions/learn-ai-bbc, 2019. Kaggle.

Common Crawl: https://commoncrawl.org/

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. “Unsupervised Cross-Lingual Representation Learning at Scale.” arXiv, April 8, 2020. https://doi.org/10.48550/arXiv.1911.02116.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805.

Gelei, Andrea & Fodor, Szabina & Ternai, Katalin, 2023. "Az ipar 4.0 felkészültség értékelési rendszere a témamodellezés segítségével - középpontban a kis- és középvállalatok Közgazdasági Szemle, Közgazdasági Szemle Alapítvány vol. 0(11), pages 1230-1260.

Grootendorst, Maarten. “BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure.” arXiv, March 11, 2022. https://doi.org/10.48550/arXiv.2203.05794.

Héja, Enikő; Ligeti-Nagy, Noémi. Magyar melléknevek poliszém jelentéseinek automatikus kinyerése gráfokkal. In: Berend, Gábor; Gosztolya, Gábor; Vincze, Veronika (szerk.) XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023. Szeged, Magyarország : Szegedi Tudományegyetem TTIK, Informatikai Intézet (2023) 469 p. pp. 63-76. , 14 p.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv, July 26, 2019. https://doi.org/10.48550/arXiv.1907.11692.

Kharosekar Aditya https://www.kaggle.com/datasets/adityakharosekar2/guardian-news-articles 2023, Kaggle

Nagy, Mihály. Testing an Inductive Mixed-method Computational Approach to News Frame Analysis: An analysis of Hungarian online reporting of the 2014 Russia-Ukraine conflict, IFLA International News Media Conference 2024

Nemeskey, Dávid Márk. “Natural Language Processing Methods for Language Modeling.” ELTE, 2020. https://hlt.bme.hu/en/publ/nemeskey_2020.

Vági, Renátó and Üveges, István and Megyeri, Andrea and Fülöp, Anna and Vadász, János Pál and Nagy, Dániel and Csányi, Gergely Márk (2023) Increasing access to legal information with unsupervised solutions. HUNGARIAN JOURNAL OF LEGAL STUDIES, 64 (3). pp. 456-471. ISSN 2498-5473

Yang, Zijian Győző, Attila Novák, and László János Laki. “Automatic Tag Recommendation for News Articles.” Eger, Hungary, 2020.

Yang Zijian Győző, Dodé Réka, Ferenczi Gergő, and Váradi Tamás. “Jönnek a nagyok! BERTLarge, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre.” Szeged, 2023.

Yang, Zijian Győző, Réka Dodé, Gergő Ferenczi, Péter Hatvani, Enikő Héja, Gábor Madarász, Noémi Ligeti-Nagy, et al. “The First Instruct-Following Large Language Models for Hungarian,” 247–52. Debrecen, Hungary: University of Debrecen, 2024.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022. N24News: A New Dataset for Multimodal News Classification. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 6768–6775, Marseille, France. European Language ResourcesAssociation.