Médiakutató, 2026. tél
XXVII. évf. 4. szám, 57–72. o.
10.55395/MK.2026.4.5

# Az Index nyelvezetének vizsgálata a 2020-as tulajdonosváltást követően gépi tanulási módszerekkel

**Angyalffy Dávid**
Eötvös Loránd Tudományegyetem
**Németh Renáta**
Eötvös Loránd Tudományegyetem
**Buda Jakab**
Eötvös Loránd Tudományegyetem

**A szerkesztéspolitikai változások kvantitatív kimutatása a számítási módszerek fejlődésével a médiakutatás fősodrának eszköztárába emelkedett. Kutatásunk az Index tartalmainak változását vizsgálja a 2020-as tulajdonosváltást követően, gépi tanulási módszerekkel. Proxy-alapú megközelítést alkalmaztunk: három, a szöveg különböző sajátosságaira érzékeny modellt tanítottunk be a HVG és az Origo 2019-es, illetve 2021-es belföldi cikkein, majd a betanított modellekkel az Index azonos időszakból származó cikkeit soroltattuk be HVG-szerű, illetve Origo-szerű csoportokba. Az eredmények arra utalnak, hogy az Index bizonyos nyelvi-szövegszintű mintázatokban közelebb került az Origo referenciakorpuszához. Az interpretálhatósági elemzés azonosította az eltolódás stilisztikai és tematikus markereit is.**

**Kulcsszavak:** gépi tanulás, sajtónyelv, természetesnyelv-feldolgozás, tulajdonosváltás

# An analysis of the language used on Index following the 2020 change in ownership, based on machine learning methods

**The quantitative analysis of changes in editorial policy has become a mainstream tool in media research, owing to advances in computational methods. This paper examines changes in the content of Index following the 2020 change in ownership, based on machine learning methods. It employs a proxy-based approach: first, the researchers trained three models, each sensitive to different textual characteristics, on HVG's and Origo's domestic news articles from 2019 and 2021, respectively; then, they used the trained models to classify Index's articles from the same time periods into HVG-like and Origo-like groups. The findings suggest that, in terms of certain linguistic and text-level patterns, Index shares more similarities with the Origo reference corpus. The interpretability analysis also identifies the stylistic and thematic markers of this shift.**

**Keywords:** change in ownership, machine learning, natural language processing, press language

## 1. Bevezetés

A magyar médiapiac a 2010-et követő másfél évtizedben jelentős átalakuláson ment keresztül. A korábban függetlennek tekintett hírportálok tulajdonosi hátterének átrendeződése, az állami hirdetési bevételek politikailag motivált elosztása, a szerkesztőségi függetlenség fokozatos szűkülése, majd a nyugat-európaias, független, piaci és az oroszos állami médiarendszer egymás mellett élése alkotják azt a kontextust, amelyben az Index 2020-as tulajdonosváltása bekövetkezett (a szűkülő nyilvánosság evolúciójáról lásd Bajomi-Lázár 2013, 2021, Bajomi-Lázár & Horváth 2023/2025, Polyák & Urbán 2024). A tulajdonosváltást a főszerkesztő menesztése és a szerkesztőség mintegy nyolcvan tagjának felmondása követte. Az esemény a magyar sajtószabadság szempontjából emblematikus esetté vált, és széles körű nemzetközi figyelmet kapott (Bognár 2021). Kutatásunk központi kérdésfeltevése az, hogy ez a tulajdonosi átrendeződés eredményezett-e a portál nyelvi-stilisztikai jellemzőiben is kimutatható változást.

Vizsgálatunk a portál 2019-es és 2021-es cikkeire épül, de nem közvetlenül hasonlítottuk azokat össze egymással, hanem egy köztes, proxy-alapú lépést alkalmaztunk. Először klasszifikációs modelleket tanítottunk be két különböző hírportál, a HVG és az Origo cikkein, majd ezeket a modelleket használtuk az Index cikkeinek elemzésére. Ez a megközelítés politikai szövegek kvantitatív elemzése során ismert gyakorlat, amikor is ismert tulajdonságú szövegeket használnak referenciaként ismeretlen szövegek besorolásához (Laver et al. 2003, Grimmer & Stewart 2013). Előnye, hogy nemcsak azt mutatja meg, van-e különbség a két időszak között, hanem azt is, hogy a különbség milyen irányú.

Elemzésünk médiatudományi alapját Robert Entman (1993) keretezéselmélete adja. Esetünkben különösen releváns, hogy az egyes szavak és kifejezések használati gyakorisága már önmagában is hordozhat keretezési mintázatokat. Az, ha a szerkesztőség következetesen bizonyos kifejezéseket, illetve bizonyos témákat vagy eseményeket részesít előnyben, gépi tanulási módszerekkel statisztikailag mérhető és modellezhető. A napirend-kijelölési *(agenda-setting)* elmélet (McCombs & Shaw 1972) pedig magyarázatot nyújthat arra, hogy a különböző portálok eltérő szóeloszlása miért tekinthető szerkesztőségi döntések következményének, és nem pusztán nyelvi variációnak.

A keretezés és a tematizálás mechanizmusa önmagában azonban nem ad választ arra a kérdésre, hogy miért várható szerkesztőségi stílusváltás egy médium tulajdonosváltását követően. A magyar médiarendszer az elmúlt másfél évtizedben számos strukturális átalakuláson ment keresztül, amelyek növelték a médiarendszerben is megnyilvánuló politikai polarizációt (Bajomi-Lázár 2013, Bognár 2021).

Módszertanunk Justin Grimmer és Brendon M. Stewart (2013) „szöveg mint adat" megközelítéséhez köthető: a nagy szöveges korpuszok gépi elemzése lehetővé teszi olyan társadalomtudományi kérdések vizsgálatát, amelyek manuális elemzéssel nem lennének megválaszolhatók. Ugyanakkor tudni kell, hogy minden szövegmodell valamilyen, a szöveg természetével kapcsolatos egyszerűsítő feltételezésen alapul – ezért is használtunk három különböző szövegreprezentációt kutatásunkban.

## 2. Az Index és a két referenciamédium, a HVG és az Origo

A vizsgálat fókuszában levő Index és a két referenciamédium, a HVG és az Origo volt a kiinduló adatforrás.

Az Index tulajdonosi körének már az alapítástól fogva vegyes háttere volt, de akkor lett politikához igazán köthető, amikor 2006-ban Spéder Zoltán megvásárolta a portált. 2017-ben Simicska Lajos lett a portál tulajdonosa, majd a 2018-as parlamenti választásokat követően a portál gazdasági háttere ismét átalakult, amit az Index szerkesztősége a „függetlenségi barométer" létrehozásával is jelzett.[1] A fordulópont 2020 márciusában érkezett el. Vaszily Miklós, aki a kormányközeli médiarendszer kiépítésében meghatározó szerepet játszott, megszerezte az Indamedia Zrt. ötvenszázalékos tulajdonrészét (Polyák 2020), mire a szerkesztőség a „veszélyben" kategóriát jelentő szintre állította a függetlenségi barométert. 2020 júliusában menesztették a főszerkesztő

1 Index.hu Bows out: "We Really Believed in Our Newspaper and Our Mission". *Kafkadesk*, 30 July 2020, https://kafkadesk-dotorg.wordpress.com/2020/07/30/index-hu-bows-out-we-really-believed-in-our-newspaper-and-our-mission/.

Dull Szabolcsot, majd a szerkesztőség mintegy nyolcvan tagja felmondott,[2] erre a portál iránt mért olvasói bizalom is hét százalékpontot esett a Reuters Institute Digital News Report 2021-es felmérése szerint (Bognár 2021). A Telex 2023-as oknyomozó elemzése szerint az Index 2020-at követően ugyan megőrizte piacvezető szerepét, működésében azonban jól érzékelhető stratégiai változás történt.[3] A portál a látogatottság fenntartása érdekében egyre inkább a bulvárjellegű tartalmakra támaszkodott, a látogatottsági mutatók maximalizálása tulajdonosi elvárás is volt, mivel a portál így maradhatott politikailag is hasznos.

A Heti Világgazdaság (HVG) szerkesztési elveit tekintve az alapítástól kezdve a brit The Economist tartalmát vette mintául, és mind példányszám, mind olvasottság tekintetében Magyarország vezető gazdasági és politikai hetilapjává vált (Csapo-Sweet & Kaposi 1999). 2003-ban a Westdeutsche Allgemeine Zeitung médiacsoport többségi, hetvenöt százalékos részesedést szerzett a kiadóban. A német tulajdonlás 2014-ig tartott, ezután részben szerkesztőségi tulajdonosi modell alakult ki, ami megerősítette a lap autonómiáját. A Reporters Without Borders sajtószabadság-indexe és a Human Rights Watch jelentése egyaránt megerősíti, hogy a HVG a független magyar sajtó fennmaradt képviselői közé tartozott.

Az Origót 1998-ban a Magyar Telekom alapította. A fordulópont 2014-ben következett be, amikor Sáling Gergő főszerkesztőt felmentették, ezután az állami hirdetések száma is érzékelhetően nőtt a portálon.[4] A Magyar Telekom 2015 végén értékesítette az Origót a New Wave Media Kft.-nek, majd 2018-ban a portál a KESMA-konglomerátum részévé vált.[5] Az Eurotopics (2026) jellemzése szerint az Origo kormányközeli hírportál. A vizsgálatunkban szereplő 2019-es Origo-cikkek már a KESMA-korszak szerkesztési normáit tükrözik.

## 3. Médianyelvezet vizsgálata gépi tanulással

Érdemes kitérni arra, hogy a gépi tanulás révén azonosított jellemző lexikai mintázatok miként kapcsolódnak a kommunikáció és a politikai kommunikáció alapvető elméleteihez, vagyis mire lehet következtetni médiumok nyelvi eltérései alapján. Entman (1993) alapgondolata szerint a keretezés a szövegben konkrét szóválasztások, ismétlődő kifejezések és érvelési stratégiák révén valósul meg, amelyek egy hír témájának bizonyos aspektusait emelik ki. Entman a keretezést úgy határozza meg, mint bizonyos aspektusok kiválasztását, hangsúlyozását és kiemelését. Számos tanulmány hangsúlyozza, hogy a lexikai mintázatok (mint például a kulcsszavak és a kollokációk) eltérése leképezheti a médiumok időbeli változását vagy a különböző médiumok közötti eltéréseket (Gerken & van der Meer 2019, Ding et al. 2023, Mykkänen et al. 2024).

Maxwell E. McCombs és Donald Shaw (1972) napirend-meghatározási elmélete szerint a média a kiemelt témák (első szint) és az azokat keretező jellemzők (második szint) fókuszba helyezésével alakítja a közvélemény figyelmét. Ezzel kapcsolatban Markus Mykkänen és munkatársai (2024) megmutatták, hogy az automatizált szövegelemző módszerek képesek felismerni a kiemelt jelenségek – vagyis a témaeloszlások és a keretező lexikon – időbeli változásait.

A politikai rendszerek és a médiatartalom kapcsolatával Hallin és Mancini (2004) foglalkozott. Szerintük a polarizált pluralista rendszerekben (például Görögországban, Spanyolországban) a pártpolitikai hovatartozás elsősorban a kiemelt témák alakításában és a forráshasználatban mutatkozik meg: a hírműsorok témáit gyakran inkább politikai logika, mint tisztán újságírói hírérték-megítélés formálja, és az újságírók a tematizálást és az értelmezést is jelentős részben a politikai szereplőkre hagyják. Ez eltér a liberális modelltől (például Észak-Amerikától,

2 Földes, András: "Don't be Silenced" – Hungary's Biggest Independent News Site Collapses. *cz.boell.org,* 13 August 2020, https://cz.boell.org/en/2020/08/13/dont-be-silenced-hungarys-biggest-independent-news-site-collapses.

3 Fábián Tamás: Sörétes puska már volt, precíziós propagandafegyver kellett – bemutatjuk a NER-be olvasztott Index receptjét. *Telex.hu,* 2023. november 13., https://telex.hu/komplex/2023/11/13/index-ner-propaganda-media-vaszily-miklos-fidesz-fekete-szaloky-zoltan-meszaros-lorinc.

4 D. Kovács Ildikó: Megszólalt a telefon, haptákban a Telekom – az Origo széthullásának története. *Atlatszo.hu,* 2015. március 7., https://atlatszo.hu/kozpenz/2015/03/06/megszolalt-a-telefon-haptakban-a-telekom-az-origo-szethullasanak-tortenete/.

5 Nemzetstratégiai jelentőségűvé minősítette a kormány a fideszes médiabirodalmat. *Index.hu,* 2018. december 5., https://index.hu/gazdasag/2018/12/05/nemzetstrategiai_jelentoseguve_minositette_a_kormany_a_fideszes_mediabirodalmat/.

Nagy-Britanniától), ahol a pártkötődések az általuk vizsgált időszak alatt fokozatosan meggyengültek (Hallin & Mancini 2004). Hallin és Mancini modelljének a gépi tanuláson alapuló elemzésekkel való kombinálása azt jelenti, hogy ugyanazok a lexikai mintázatok a médiarendszer modelljétől függően eltérő kiemeltségi és keretezési jeleket kódolhatnak. Bár a magyar médiarendszer 2010 és 2026 közötti időszakában Hallin és Mancini egyik modelljének sem feleltethető meg közvetlenül, arra számíthatunk, hogy polarizált politikai rendszerekben erősebb összhang jelenik meg a politikai orientáció és a lexikai markerek között (Balaban et al. 2016).

A szakirodalom elismeri, hogy a „szöveg mint adat" módszerek kiforrottá váltak, de validálásra szorulnak (Gerken & van der Meer 2019, Antoniak et al. 2023). Justin Grimmer és Brendon M. Stewart (2013) is a megközelítés korlátait hangsúlyozza. Szerintük az automatikus szövegelemző eszközök használata esetén alaposan meg kell fontolni, hogy a modell valójában mit mér, beleértve a szavak sorrendjét figyelembe nem vevő „szózsák"-modellek korlátait (erről részletesebben lásd a 4.2.1. és a 4.2.2. fejezetet). Ezért használtunk vizsgálatunkban három független, eltérő szövegreprezentációra épülő modellt és tettünk erőfeszítéseket a modellek interpretálására is, vagyis arra, hogy megállapítsuk, milyen szövegjellemzők alapján soroltak HVG- vagy Origo-szerű csoportba egy-egy cikket a modellek.

A médianyelvezet kvantitatív vizsgálata a számítási módszerek fejlődésével mára a médiakutatás fősodrának eszköztárába emelkedett. Néhány konkrét példát említünk. Martin Potthast és társai (2018) például sikeresen különítették el a szélsőségesen pártos híreket stílusuk alapján a fősodortól. Ramy Baly és társai (2020) a cikkeket nyelvezetük alapján sorolták be baloldali, centrista vagy jobboldali politikai irányultság szerint, és arra figyelmeztetnek, hogy a modellek alapértelmezetten a médium stílusát és tematikus fókuszát tanulják meg, és nem a politikai irányultságát. Kutatásunkban a torzítás elkerülésére a Baly és társai által alkalmazott egyik technikát követjük mi is: az Index-cikkek nem részei a tanuló adatoknak, vagyis a tanuló modellek kizárólag a HVG és az Origo mintázatait ismerik. Fontos hangsúlyozni, hogy azt vizsgáljuk, az Index szövegei a vizsgált modellek szerint nyelvi jellemzők, azaz bizonyos stiláris és tartalmi (tematikus) jegyek mentén közelebb kerülnek-e az Origo referenciakorpuszához (az Origo-szerűbbé válás ezt jelenti), vagyis a modell önmagában nem bizonyít szerkesztéspolitikai átalakulást.

Lisa Fan és társai (2019) vezették be az „információs torzítás" fogalmát: a „lexikai torzítással" ellentétben, ahol például érzelmileg töltött jelzők jelezhetik az elfogultságot, az információs torzítás gyakran csak több forrás összevetésével válik érzékelhetővé, mivel az egyes elemek önmagukban tényszerűek. Szempontunkból ez a megkülönböztetés azért lényeges, mert az információs torzítás detektálása kontextuális modelleket igényel (ez indokolta a BiLSTM-modell bevonását kutatásunkba, a modell bemutatását lásd a 4.2 fejezetben).

A Wall Street Journal (WSJ) 2007-es tulajdonosváltását vizsgálta Allison M. Archer és Joshua Clinton (2018). Eredményük szerint ismert lapokkal összevetésben a lap véleményrovatainak politikai tartalmának aránya változatlan maradt, de a WSJ címlapján a politikai tárgyú cikkek aránya és hangsúlya megnőtt a többi laphoz képest.

A magyar nyelvű szakirodalomban Buda Jakab és Németh Renáta (2024) tanulmánya hasonlít módszertanilag kutatásunkhoz. A szerzők a magyar Országgyűlésben 1998 és 2018 között elhangzott Fidesz- és MSZP-beszédeket tanítottak gépi tanulónak, és az időben előre haladva egyre nagyobb pontosságot, azaz a politikai polarizáció erősödését detektálták. Ugyanakkor kiemelik, hogy a modell interpretációja is elengedhetetlen az értelmezés szempontjából – ezért törekszünk mi is osztályozó modelljeink interpretációjára.

## 4. Módszertan

### 4.1. Adatbázis

A három médium cikkeinek összegyűjtéséhez a SentiOne nevű, webalapú médiafigyelő és szövegelemző platformot használtuk. A SentiOne által a GDPR-előírásoknak megfelelően összegyűjtött adathalmazt a portálok belföldi híreket tartalmazó rovataira korlátoztuk (HVG, Origo: Itthon, Index: Belföld), mivel azt feltételeztük, hogy a szerkesztési eltérések a belföldi tartalmakban a legmarkánsabbak. Az adatbázis 65 088 cikkből áll, az éves és a portálonkénti megoszlást az 1. táblázat foglalja össze.

*1. táblázat*
*A letöltött cikkek megoszlása portál és év szerint*

| | 2019 | 2021 | Összesen |
|---|---|---|---|
| **HVG** | 9 969 | 7 617 | 17 586 |
| **Origo** | 10 898 | 10 765 | 21 663 |
| **Index** | 8 693 | 17 146 | 25 839 |
| **Összesen** | 29 560 | 35 528 | 65 088 |

A nyers szövegek egy többlépéses szövegtisztítási és előkészítési folyamaton estek át (a részletekről lásd cikkünknek a Médiakutató honlapján szereplő online mellékletét), ennek során kisbetűsítettünk, eltávolítottuk a jelentés nélküli és túl gyakori kifejezéseket, szótövesítettünk, és létrehoztuk a TF-IDF súlyt, amely meghatározza egy kifejezés fontosságát egy dokumentumban.[6]

A visszacsatolt neurális hálók esetében a túl gyakori kifejezések eltávolítása már nem kézenfekvő, mert ezek a szavak fontos mondatszerkezeti és grammatikai összefüggéseket hordozhatnak, amelyek a szekvenciális modell számára informatívak lehetnek (Chai 2023). Ezek alapján a BiLSTM-modell számára külön korpuszt készítettünk, amelyből kizárólag a portálnevekre expliciten utaló kifejezéseket szűrtük ki.

## 4.2. Az alkalmazott modellek

A robusztusság érdekében három modellt alkalmaztunk, ezek a szöveg más jellemzőire érzékenyek (a hangolást és a betanítást részletesen lásd az online mellékletben). A kutatási kérdés megválaszolásához a 2019-es és a 2021-es időszakra külön modelleket illesztettünk. Az elemzési dizájn kialakításakor mérlegeltük egy összevont modell lehetőségét is. Végül a két külön modell mellett döntöttünk, mivel ez lehetővé teszi a zavaró tényezők – elsősorban a két időszak közötti tematikus különbségek – jobb kontrollálását, és tisztábban megválaszolja fő kutatási kérdésünket, vagyis azt, hogy a tulajdonosi átrendeződés eredményezett-e a portál nyelvi-stilisztikai jellemzőiben is kimutatható változást. Egy összevont modell esetén a potenciális zavaró tényezők kezelése (a tematikus eltéréseken túl egyéb, nem ismert hatások is) nehezebben lenne kontrollálható. Robusztusságvizsgálatként kiegészítő modellillesztést is végeztünk: a szövegekből a HuSpacy csomag hu_core_news_lg modelljének (Orosz et al. 2023) segítségével eltávolítottuk a tulajdonneveket és egyéb névelemeket, ezeket típusukat jelölő címkékre cserélve. A hiperparaméter-keresés terét és a tanítási beállításokat (lásd az online mellékletben) nem módosítottuk, hogy a különbség kizárólag a bemenetre legyen visszavezethető. A tulajdonnév nélküli elemzés azért ad robusztusságellenőrzést, mert a tulajdonnevek (a politikusnevek, a pártnevek, az intézménynevek stb.) nagyrészt a témát, nem a stílust kódolják. Ha a modell megtanulja, hogy bizonyos nevek (például a politikusoké) gyakrabban fordulnak elő az egyik portál szövegeiben, akkor ezek alapján is tud különbséget tenni. Alább röviden ismertetjük az alkalmazott modelleket (részletesebben lásd az online mellékletet).

6 A TF-IDF *(Term Frequency – Inverse Document Frequency)* egy súlyozási módszer, amellyel azt mérhetjük meg, hogy egy szó mennyire jellemző egy adott dokumentumra egy nagyobb dokumentumgyűjteményen belül. Két részből, két logikából áll: a TF (*Term Frequency* vagyis szógyakoriság) egyszerűen megszámolja, hányszor fordul elő egy adott szó egy adott dokumentumban. Minél gyakrabban szerepel a szó abban a szövegben, annál nagyobb a TF-értéke. Ez önmagában viszont félrevezető lehet, mert például a névelők minden szövegben gyakoriak, mégsem hordoznak információt. Ezért szükséges az IDF (*Inverse Document Frequency,* vagyis inverz dokumentumgyakoriság), amely azt méri, hogy egy szó mennyire gyakori a teljes korpuszban. Ha egy szó szinte minden dokumentumban előfordul, az IDF-értéke alacsony lesz – mert nem segít megkülönböztetni a dokumentumokat egymástól. Ha viszont egy szó csak kevés dokumentumban jelenik meg, akkor az IDF-értéke magas lesz, mert az a szó valószínűleg tartalmilag megkülönböztető. Kutatásunkban tehát nem azonos súllyal szerepelt minden szó: magas TF-IDF súlyt azok a szavak kaptak, amelyek gyakoriak egy adott dokumentumban, de ritkák a teljes korpuszban – tehát valószínűleg tartalmilag releváns kifejezések az adott szövegre nézve. Ez ad alapot a szövegek vektoros reprezentációjának, amelyet aztán gépi tanulási modellek bemeneteként használunk.

#### 4.2.1. Logisztikus regresszió

A logisztikus regresszió klasszikus osztályozó, vagyis segítségével kategóriákba tudunk osztályozni megfigyeléseket. Jelen kutatásban a cikkekhez tartozó szavakra TF-IDF súlyozást alkalmaztunk (lásd a TF-IDF súly definícióját korábban). A modell kétértékű osztályozó, hiszen a HVG-hez vagy az Origóhoz való hasonlóságot kell a modellnek eldöntenie az Index cikkeinek vonatkozásában. Ez a megközelítés szózsákmodellt alkalmaz, vagyis a cikkeket szavakra bontja, tehát nem érzékeny a szósorrendre és a szintaxisra. A logisztikus regresszió eredménye az adott cikkre vonatkozó Origóhoz tartozási esély, amelyet a cikket alkotó kifejezések TF-IDF értékéből jósol meg. Ugyanakkor a modell azt feltételezi, hogy ezek a tényezők lineárisan hatnak az esély logaritmusára, ami erős megszorítás. Rugalmasabb modellek, mint a XGBoost vagy a neurális hálók gyakran jobb eredményt adnak (Fernández-Delgado et al. 2014), ezért vontuk be ezeket is a kutatásunkba.

#### 4.2.2. XGBoost

Az eXtreme Gradient Boosting (XGBoost, Chen & Guestrin 2016) modell CART-fák *(Classification and Regression Trees)* együttesére épül. Egyetlen fa a gyakorlatban rendszerint nem elég erős, ezért az XGBoost esetén a végső predikció több fa kimenetének összegeként áll elő.[7] Az XGBoost esetében is TF-IDF szóreprezentációt alkalmaztunk.

#### 4.2.3. A BiLSTM-modell

A logisztikus regresszió és az XGBoost is szózsákmodellre épül, azaz a szöveget szavak sorrend nélküli halmazaként kezeli. A neurális hálózatok megjelenésével vált lehetővé a szekvenciális minták tanulása (LeCun et al. 1989), de az igazi áttörést a visszacsatolásos (rekurrens) architektúrák hozták, amelyek képesek explicit módon kezelni az időbeli függőségeket (Elman 1990).

Ugyanakkor egy hírportál cikkének stílusa, retorikai eszközei, témafelvetései a cikk egészén átívelnek, ezért a modellnek a szöveg korábbi részeire is emlékeznie kell, amikor a későbbi szakaszokat dolgozza fel. Erre alkalmas a Long Short-Term Memory (LSTM) architektúra (Hochreiter & Schmidhuber 1997). Az egyirányú LSTM korlátja, hogy a szövegben jobbra álló szavak nem befolyásolják a bal oldali pozíciók reprezentációját. A kétirányú LSTM (BiLSTM, Schuster & Paliwal 1997) ezt a hiányosságot orvosolja.

#### 4.2.4. Értelmezés

Az LSTM fontos hátránya a nehéz értelmezhetőség, a *black-box* jelleg, vagyis az, hogy a modell a cikkeket besorolja valamelyik osztályba, de a döntésének pontos oka nehezen vagy egyáltalán nem értelmezhető az ember

7 Az XGBoost egy gépi tanulási módszer, amely sok egyszerű döntési fát épít egymás után úgy, hogy mindegyik új fa az előzőek hibáit próbálja javítani. A *gradient* arra utal, hogy a hibajavítás egy matematikai optimalizálási eljárással (gradiensmódszerrel) történik, amely hatékonyan minimalizálja a modell tévedéseit. A döntési fa lényegében olyan, mint egy kérdéssorozat, amely elágazásokban dönt. Minden elágazási pont egy egyszerű igen/nem kérdés egy adott jellemzőre, és a fa végén (a „levelein") van a végső döntés vagy becslés. Minden egyes elágazásnál úgy választja ki az algoritmus a szétválasztó feltételt, hogy a létrejövő csoportok (és végül a levelek) minél „tisztábbak" legyenek, vagyis egy levélben lehetőleg csak egyetlen osztály elemei legyenek. Ez a klasszikus regressziós elemzésnél jóval rugalmasabb megoldás, nem feltételez lineáris kapcsolatokat, nincsenek eloszlásfeltételei, és könnyen értelmezhető. A modell a CART *(Classification and Regression Trees)* algoritmusra épül, amely automatikusan megkeresi, hogy melyik kérdés (vagyis melyik változó, milyen küszöbértékkel) osztja legjobban szét az adatokat aszerint, hogy azok hasonlítsanak egymásra a célváltozó szempontjából. Ez ismétlődik lépésről lépésre, amíg a fa „kész" nem lesz. Egyetlen döntési fa gyakran túl egyszerű vagy túlilleszkedik (vagyis túl specifikusan tanulja meg a tanulóadatokat, és rosszul generalizál új adatokra). Ha viszont sok fát kombinálunk – ahol mindegyik másra fókuszál, és az előző fák hibáit korrigálja –, a modell összessége sokkal pontosabb és stabilabb lesz, mint bármelyik fa önmagában.

számára. A feketedoboz-jelleg a szózsákalapú logisztikus regresszió esetén nem jelentkezik (Ribeiro et al. 2016), ugyanakkor a figyelemmechanizmus (Bahdanau et al. 2015) lehetővé teszi, hogy megtudjuk, mely szavak járultak hozzá leginkább a döntéshez. Mi a LIME (*Local Interpretable Model-agnostic Explanations*, Ribeiro et al. 2016) módszert alkalmaztuk.[8] XGBoost esetén a feketedoboz-jelleg kevésbé jelentkezik, bár a döntési mechanizmus itt is nehezen követhető – itt a SHAP értékre (Lundberg & Lee 2017) fogunk az értelmezés során támaszkodni.

#### 4.2.5. A modellparaméterek hangolása és konfigurálása

Az adatokat tanító és teszthalmazokra osztottuk 80/20 arányban, rétegzéssel, amely biztosítja, hogy a HVG/Origo osztályok aránya a tanító és a teszthalmazban is megőrződjön. Először egy (kezdeti, alapértelmezett hiperparaméterekkel illesztett) alapmodellt vettünk, amely kiindulópontként szolgált a későbbi hiperparaméter-hangoláshoz (vagyis a kézzel beállítható paramétereket szisztematikusan variálásához, hogy megtaláljuk azt a kombinációt, amellyel a modell a legjobban teljesít). A hiperparaméter-hangolás után a legjobb paraméterekkel felépített végleges modell kiértékelése következett.

## 5. Eredmények

### 5.1. Besorolási pontosság

A modellek betanításának és teljesítményének pontos jellemzői az online mellékletben találhatók, itt csak a legfontosabb eredményeket emeljük ki. A három modelltípus, a logisztikus regresszió, az XGBoost és a BiLSTM egybehangzóan azt mutatta, hogy 2019-ben a HVG-cikkek besorolási pontossága elmarad az Origo-cikkekétől, vagyis a HVG stilisztikailag/tartalmilag heterogénebb. A 2021-es modellben ez az aszimmetria felerősödni látszik, azaz 2021-re az Origo nyelvileg egységesebb képet mutat. Ugyanakkor ez könnyen adódhat a kiegyensúlyozatlan mintaösszetételből is, hiszen 2019-ben 10 százalékkal több az Origo-cikk, 2021-ben pedig már majdnem 30 százalékkal. A HVG és az Origo szövegei mindkét vizsgált évben magas pontossággal elkülöníthetőek, miközben a 2021-es korpuszon tanított modellek minden modell esetében jobb teljesítményt értek el, mint a 2019-es korpuszon tanítottak, vagyis a távolság nőtt a két portál között.

### 5.2. A modellek által megtalált nyelvi sajátosságok

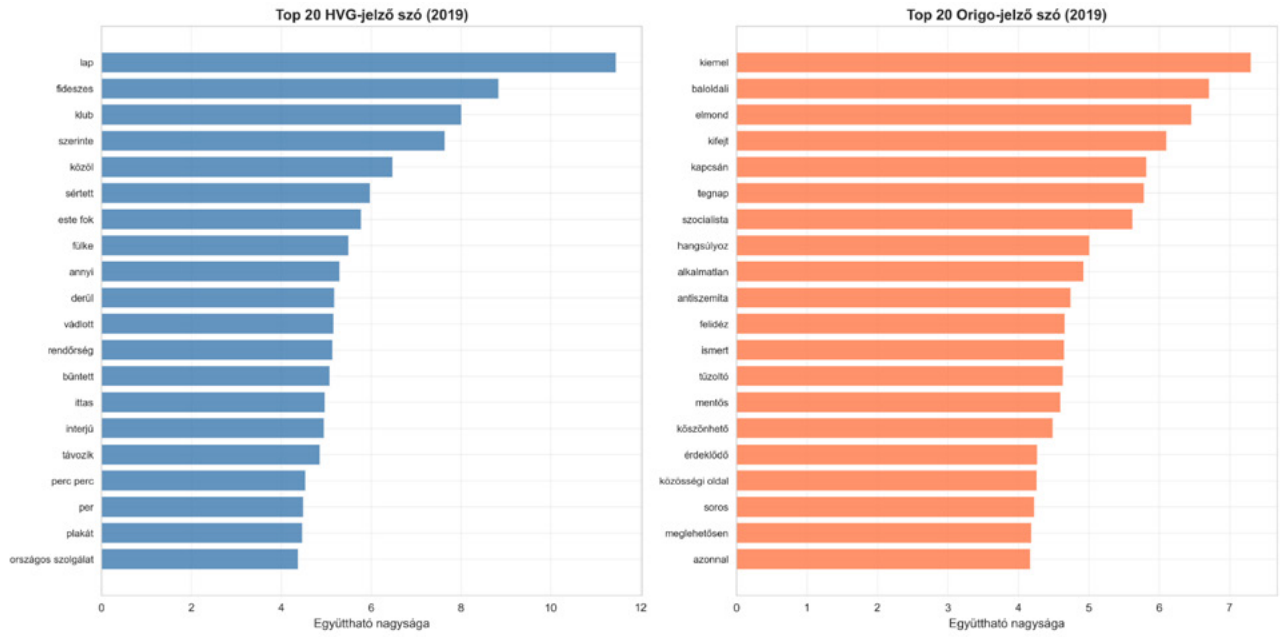
#### 5.2.1. Logisztikus regresszió

A 2019-es eredmények alapján a modell több olyan kulcsszót azonosít, amelyek jól elkülönítik a két hírforrást (lásd az 1. ábrát); itt az adott kifejezéshez tartozó érték nagysága a változó relatív fontosságukat jelzi. A HVG esetében gyakoribbak a „szerinte” és a „közöl” típusú kifejezések, míg az Origo inkább a „kiemel”, az „elmond” és

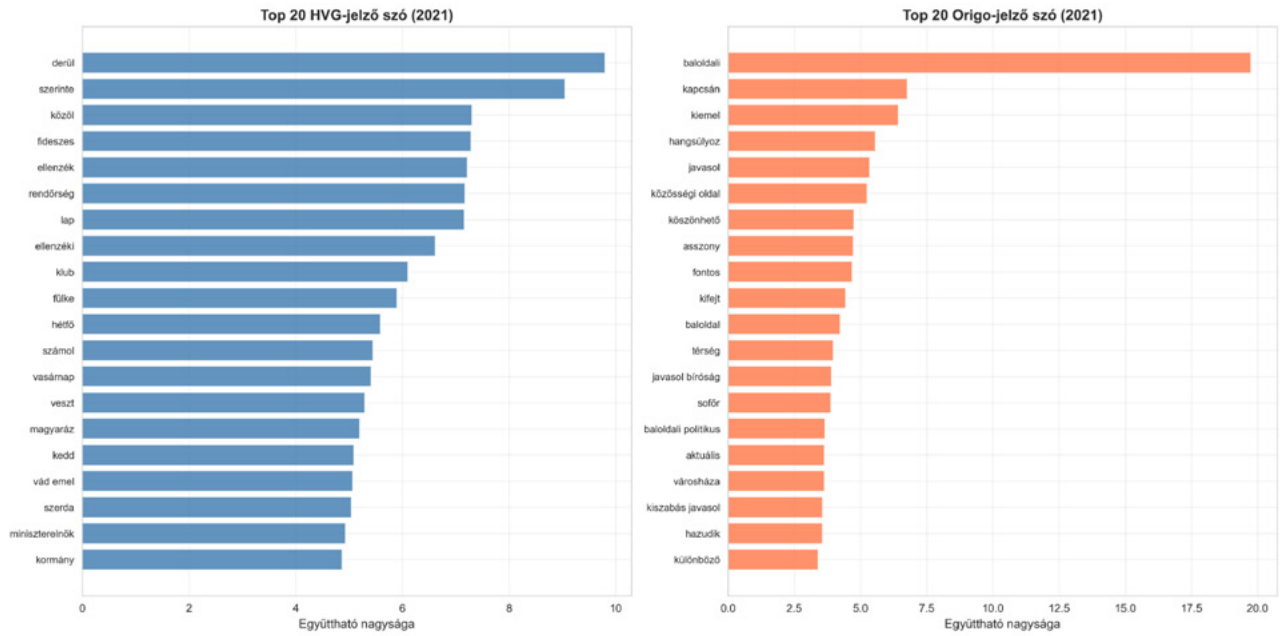
8 A LIME lokális módszer, ezért az eredmények egyedi cikkekre vonatkoznak, és nem általánosíthatók. A „lokális módszer” kifejezés arra utal, hogy az nem az egész modell globális működését próbálja megmagyarázni, hanem csak egyetlen konkrét előrejelzést. Ugyanis egy komplex modell (mint az LSTM) globálisan rendkívül bonyolult, amit egészében szinte lehetetlen egyszerűen, emberi nyelven leírni. A LIME ehelyett kiválaszt például egy adott cikket, amelyet a modell „Origo-szerűnek” minősített, majd ennek a pontnak a közvetlen közelében apró perturbációkkal új, mesterséges adatpontokat generál. Megnézi, hogy a bonyolult modell miként reagál ezekre a közeli, módosított bemenetekre. Erre a lokális adathalmazra egy egyszerű, interpretálható modellt illeszt, amely jól közelíti a komplex modell viselkedését – de csakis ebben a szűk környezetben. A LIME tehát nem állítja, hogy az általa talált szó mindig fontos az Origo-stílus felismerésében, csak azt mondja meg, hogy ennél a konkrét cikknél mely szavak befolyásolták leginkább a modell döntését.

a „kifejt” igéket használja. Emellett lexikai szinten is megjelennek eltérések (például: „ittas” vagy „részeg”). A politikai kontextusú kifejezések szintén meghatározóak mindkét portál esetében. A „fideszes” vagy az „ellenzék” a HVG-hez, a „baloldali”, a „szocialista” és a „Soros” kifejezés az Origo-hoz áll közelebb. Tematikus különbségek is megfigyelhetők: az Origo esetében a „mentős” és a „tűzoltó”, a HVG-nél a „rendőrség” jelenléte erősíti az adott kategóriába tartozást. A 2021-es eredményeket tekintve (lásd a 2. ábrát) a kulcsszavak jó része változatlan.

*1. ábra*
*A top 20 jellemző a HVG és az Origo esetében – a 2019-es modell*



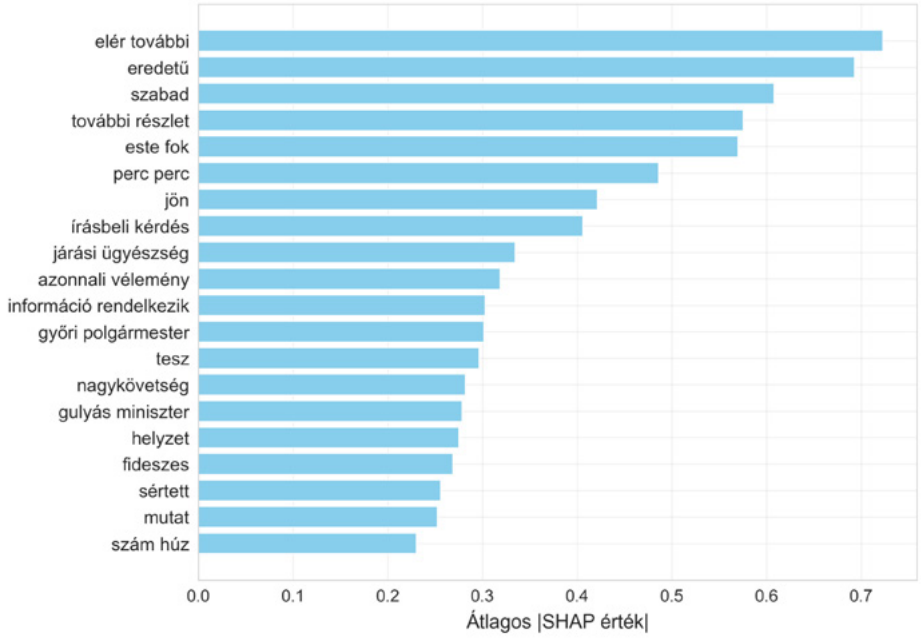

*2. ábra*
*A top 20 jellemző a HVG és az Origo esetében – a 2021-es modell*



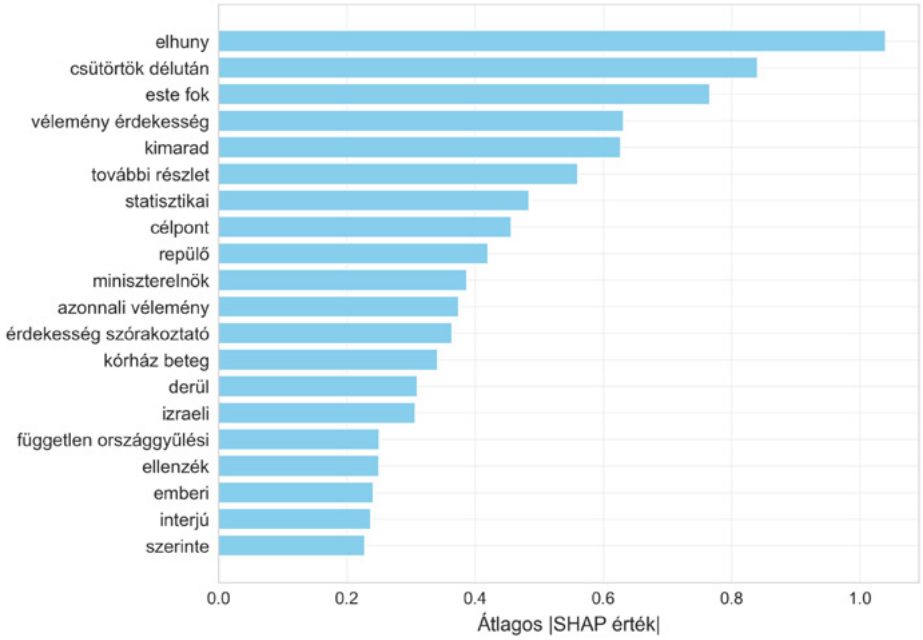
### 5.2.2. XGBoost

A HVG-osztályhoz tartozó jellemzők között (lásd a 3. és a 4. ábrát) több stabil, mindkét évben megjelenő kifejezés azonosítható. Kiemelkedő szerepet játszanak az olyan bigramok (vagyis olyan két szóból álló kifejezések), mint a „további részlet” vagy az „azonnali vélemény”. Ezek mindkét évben való jelenléte arra enged következtetni, hogy a HVG esetében a stilisztikai mintázatok időben viszonylag állandóak. Ugyanakkor bizonyos eltérések is megfigyelhetőek: a 2021-es modellben például az „elhunyt” kifejezés kiemelkedő jelentőségűvé válik, ami összefügghet a pandémiával.

*3. ábra*
*Top 20 HVG-jelző jellemző – 2019*

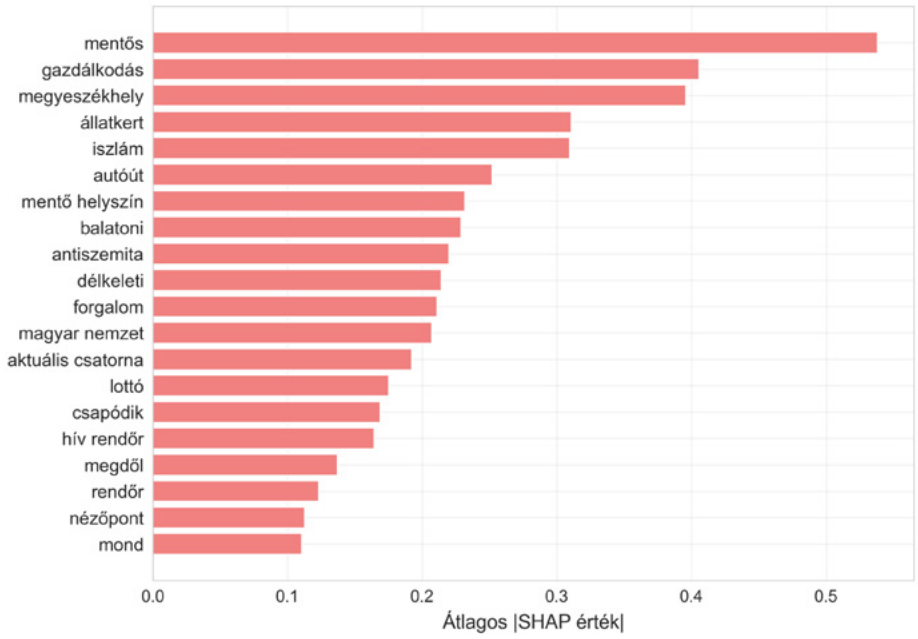



*4. ábra*
*Top 20 HVG-jelző jellemző – 2021*

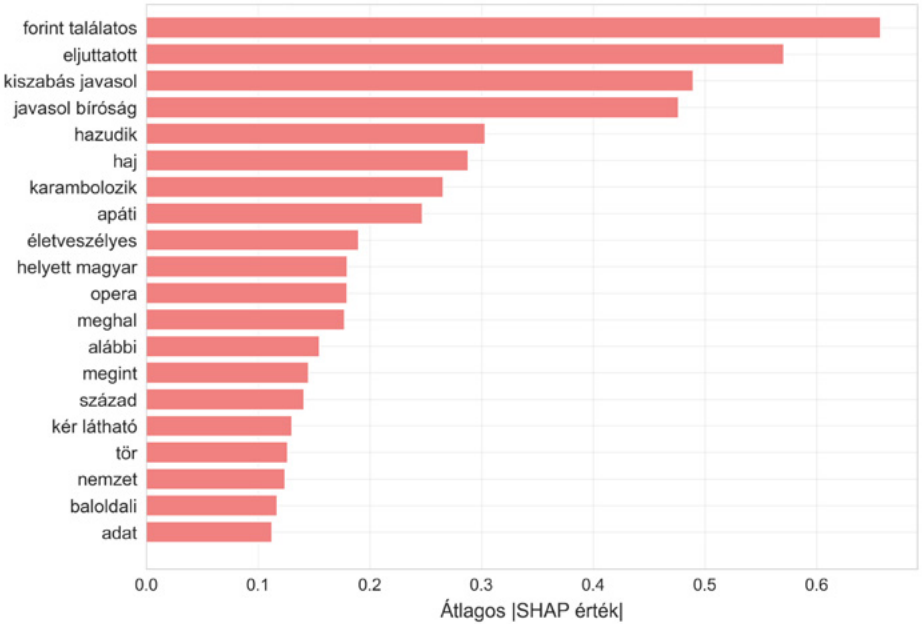


Az Origo-osztályhoz tartozó jellemzők esetében nagyobb eltérések figyelhetőek meg a két vizsgált év között (lásd az 5. és a 6. ábrát). A 2019-es modellben olyan kifejezések dominálnak, mint a „mentős", a „megyeszékhely" vagy a „gazdálkodás", amelyek inkább konkrét eseményekhez kötődő tematikus tartalmakra utalnak. Ezzel szemben a 2021-es eredményekben az olyan típusú jellemzők kerülnek előtérbe, mint a „forint találatos" vagy az „eljuttatott", ami eltérő tematikus hangsúlyokra, változékonyabb tartalmi fókuszra utalhat.

*5. ábra*
*Top 20 Origo-jelző jellemző – 2019*



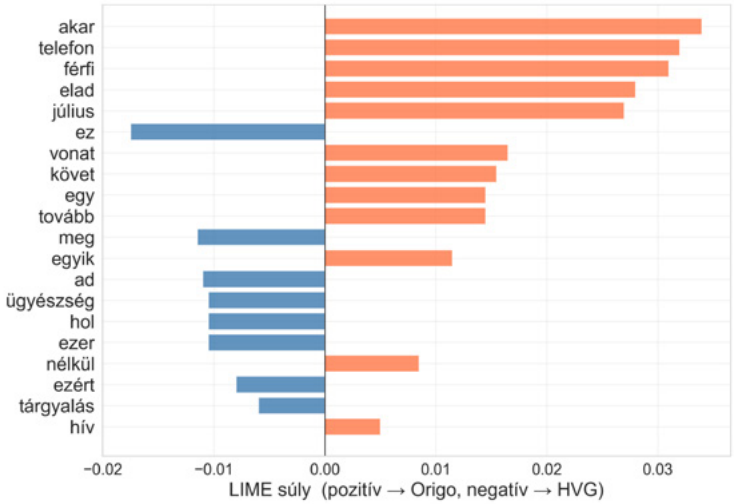

*6. ábra*
*Top 20 Origo-jelző jellemző – 2021*



### 5.2.3. BiLSTM

A BiLSTM értelmezésére a LIME-módszert alkalmaztuk. A 7. ábrán bemutatott példában a modell egy Origo-cikket nagy magabiztossággal (százszázalékos valószínűséggel) helyesen sorolt az Origo kategóriába; a döntést olyan szavak befolyásolták, mint az „akar", a „telefon", a „férfi", az „elad" és a „július". Ezek a kifejezések arra utalnak, hogy a nyelvi eltérések nemcsak stilisztikai, hanem tartalmi eltéréseket is jeleznek. A negatív irányba, tehát az Origótól elhatárolódó irányba mutató szavak (mint például az „ez", a „meg" és az „ügyészség") kisebb súllyal jelennek meg, ami tovább erősíti a döntés egyértelműségét. Az online melléklet két további példát is mutat.

*7. ábra*
*LIME-elemzés – helyes Origo-besorolás (P(Origo) = 1,000)*




## 6. Az Index-cikkek osztályozásának eredményei

Az Index cikkeinek vizsgálatára is a fent bemutatott modelleket használtuk. Részletesebb eredményeket az online melléklet közöl.

### 6.1. Logisztikus regresszió

A 2. táblázat összefoglalja a két logisztikus regressziós modell által az Index cikkeire adott besorolások megoszlását.

*2. táblázat*
*Az Index-cikkek osztályozási eredményei – logisztikus regresszió (n = 23 618)*

| Metrika | 2019-es modell, 2019-es cikkek | 2021-es modell, 2021-es cikkek |
|---|---|---|
| **Origo-szerű besorolás** | 26,29% | 40,72% |
| **HVG-szerű besorolás** | 73,71% | 59,28% |
| **P(Origo) átlag** | 0,317 | 0,419 |
| **P(Origo) medián** | 0,223 | 0,375 |

Az osztályozási eredmények alapján megfigyelhető, hogy az Origo-szerű cikkek aránya 26,29 százalékról 40,72 százalékra nőtt. A becsült P(Origo) valószínűség átlaga és mediánja szintén egyértelmű növekedést mutatott. Vagyis az Index 2021-es cikkeit a modell számottevően nagyobb arányban sorolta az Origo-szerű kategóriába.[9]

## 6.2. XGBoost

A 3. táblázat foglalja össze a két XGBoost modell által adott besorolások megoszlását. Az XGBoost esetében is megfigyelhető az Origo-szerű besorolások számának növekedése. Az Origo-szerű cikkek aránya 25,67 százalékról 34,85 százalékra nőtt, a P(Origo) valószínűségek átlaga és mediánja is növekedést mutatott, bár a logisztikus regressziónál jóval kisebb mértékűt. A hipotézisvizsgálat is eltolódást mutatott (-0,32-ről -0,15-re csökkent a d, de mindkét évre szignifikáns a két médium eltérése).

*3. táblázat*
*Az Index-cikkek osztályozási eredményei – XGBoost (n = 23 618)*

| Metrika | 2019-es modell, 2019-es cikkek | 2021-es modell, 2021-es cikkek |
|---|---|---|
| **Origo-szerű besorolás** | 25,67% | 34,85% |
| **HVG-szerű besorolás** | 74,33% | 65,15% |
| **P(Origo) átlag** | 0,326 | 0,365 |
| **P(Origo) medián** | 0,280 | 0,286 |

## 6.3. BiLSTM

A 4. táblázat ismerteti a két BiLSTM-modell besorolásainak megoszlását. Az Origo-szerűnek klasszifikált cikkek aránya 22,24 százalékról 30,40 százalékra nőtt, az átlagos és medián P(Origo) értéke is nőtt, az XGBoost-nál nagyobb mértékben. Az eltolódás iránya tehát egyértelmű. A szignifikanciateszt itt is szignifikáns mindkét évre, de csökkenő hatásnagysággal.[10]

*4. táblázat*
*Az Index-cikkek osztályozási eredményei – BiLSTM (n = 23 791)*

| Metrika | 2019-es modell, 2019-es cikkek | 2021-es modell, 2021-es cikkek |
|---|---|---|
| **Origo-szerű besorolás** | 22,24% | 30,40% |
| **HVG-szerű besorolás** | 77,76% | 69,60% |
| **P(Origo) átlag** | 0,268 | 0,343 |
| **P(Origo) medián** | 0,151 | 0,243 |

9 Statisztikai hipotézisteszttel is ellenőriztük ezt az elmozdulást, Mann-Whitney U-tesztet alkalmazva. Mivel a nagy mintaméret mellett a p-érték önmagában félrevezető lehet, ezért Cohen-féle d-hatásméretet is számoltunk (Sullivan & Feinn 2012). A teszt szignifikáns eltérést mutatott mindkét évre a két médium P(Origo) értékeinek átlaga között, de a hatásméret d = -0,36-ról -0,21-re csökkent, vagyis az Index közelebb került az Origóhoz.

10 A Cohen-féle d -0,30-ról -0,19-re csökkent.

## 7. Robusztusság-ellenőrzés

Modelljeinknek a tulajdonnevek törlésével kapott, tematikailag „semlegesített" szövegeken tapasztalt besorolási teljesítménye minimálisan tért el a fent ismertetett teljesítménymutatóktól, az Index-cikkek besorolásánál ezek a modellek erősebb különbséget mutatnak a két év között (lásd az 5. táblázatot, további részletekért pedig az online mellékletet). Vagyis eredményeink megerősítik, hogy a mért különbségek jelentős részben stilisztikai jellegűek, nem pedig a témaválasztásból fakadnak.

*5. táblázat*
*Tulajdonnevek eltávolítása után illesztett modellek eredményei*

| Modell | Év | Origo-szerű besorolás | P(Origo) átlag | P(Origo) medián | Hatásnagyság (d) |
|---|---|---|---|---|---|
| **LogReg** | 2019 | 16,9% | 0,2438 | 0,1496 | −0,36 |
| | 2021 | 33,6% | 0,3699 | 0,3034 | −0,26 |
| **XGBoost** | 2019 | 13,4% | 0,2287 | 0,1552 | −0,33 |
| | 2021 | 29,7% | 0,3279 | 0,2371 | −0,18 |
| **BiLSTM** | 2019 | 19,0% | 0,2358 | 0,1193 | −0,27 |
| | 2021 | 28,9% | 0,3463 | 0,2856 | −0,19 |

## 8. Limitációk

Elemzésünknek módszertani korlátja a téma és a retorikai stílus közötti szétválasztás nehézsége. Amennyiben az Origo és a HVG jellemzően eltérő tématartományokat fed le a belföld rovatban, a modellek részben tematikus jegyeket tanulhattak meg retorikai markerek helyett.

A mélytanulási modellek feketedoboz-jellege további korlátot jelent. Bár a logisztikus regresszió és az XGBoost globálisan is interpretálható együtthatókat adott, a BiLSTM által megtanult globális döntési logika közvetlenül nem hozzáférhető. Ebből következően a modell predikciói ugyan mutatják az eltolódás tényét és irányát, de az okokat csak közvetett módon lehet feltárni.

A szókincs időbeli változása szintén egy potenciálisan torzító tényező. A BiLSTM-modell mögött álló, előre betanított *fastText*-beágyazások egy általános magyar nyelvi korpuszból származnak, 2017 a tanító adat vége, így nem tartalmazzák azokat az újonnan keletkező szavakat vagy jelentésváltozásokat, amelyek a 2019 és 2021 közötti időszakban jelentek meg a politikai és közéleti diskurzusban (ilyen például a járványügyi szókincs). A beágyazási mátrixban nem szereplő szavak zajt vihetnek a predikcióba.

Végül a cikkhossz varianciáját is meg kell említeni. A BiLSTM esetében a hosszabb szövegek csonkolásával járó információveszteséget és az annál rövidebb szövegek feltöltésével járó zajt nehéz elkerülni. A TF-IDF-alapú modellek esetében a dokumentumhossz közvetlen hatása kisebb.

## 9. Összegzés

A három különböző komplexitású, független modell sikeresen ragadta meg az Index-cikkek stilisztikai változásának különböző szintjeit, és azonos irányba mutató, konzisztens eredményeket hozott. A 2021-es adatokon tanított modell jobban szétválasztotta a HVG-t és az Origót, vagyis 2021-re a HVG és az Origo stílusa jobban elkülöníthetővé vált, függetlenül az alkalmazott modellarchitektúrától. Mind a három modell konzisztensen azt mutatta, hogy az Index nyelvi (azaz stiláris és tartalmi) mintázatokban közelebb került az Origo referencia-

korpuszához a 2019-ről 2021-re történő átmenet során. Ugyanakkor a 2021-es cikkek többsége továbbra sem kerül dominánsan az Origo-tartományba, vagyis inkább részleges és fokozatos elmozdulás, és nem átlépés történt. A hatásméretek is kis hatást mutatnak, ennek ellenére az eredmények nem elhanyagolhatóak, hiszen csupán egyetlen év változását vizsgáltuk.

Eredményeink a magyar médianyilvánosság szélesebb átrendeződésének kontextusában is értelmezhetők. Az Index esetében megfigyelt, fokozatos stilisztikai-tartalmi közeledés az Origóhoz egy korábban tömegolvasottságú, a közvélemény szélesebb rétegeit elérő portálnak a kormányzathoz köthető médiakörnyezet nyelvi-tartalmi mintázataihoz történő fokozatos homogenizálódására utal. Ez a fajta közeledés különösen releváns azoknak a mechanizmusoknak a fényében, amelyeket a médiatulajdonlás és az állami hirdetési piac politikai felhasználásáról szóló szakirodalom ír le (Bátorfy & Urbán 2020), és amelyek a médiapiac szerkezetén keresztül közvetve is befolyásolhatják a hírfogyasztók számára elérhető témák és keretezések sokszínűségét. A választói tájékozottság szempontjából ennek különös súlyt ad, hogy a médiapluralizmus és a kormányzati elszámoltathatóság szorosan összefügg: minél inkább egy irányba tolódnak el a korábban egymástól megkülönböztethető hírforrások nyelvi-tematikus mintázatai, annál kevésbé valószínű, hogy az állampolgárok – különösen azok, akik korábban éppen az Index viszonylagos semlegességét részesítették előnyben – egymással versengő keretezéseknek legyenek kitéve.

Fontos ugyanakkor hangsúlyozni, hogy megközelítésünk nyelvi hasonlóságot mér, és nem tartalmi torzítást, politikai irányultságot vagy szerkesztőségi szándékot. Így az itt bemutatott eredmények azt jelzik, hogy a nyilvánosság szerkezetében zajló átalakulás nyoma a szövegek szintjén is kimutatható, és ez indokolttá teszi a kérdés további, közvetlenül a médiafogyasztásra és a politikai ismeretszerzésre fókuszáló kvalitatív vizsgálatát.

***Mesterséges intelligencia használata***

*A tanulmány elkészítése során a szerzők a Claude, a ChatGPT és a NotebookLM szolgáltatásait vették igénybe a programkód megírása, a kiinduló irodalmazás, a hivatkozásjegyzék formázása és a szöveg rövidítése érdekében. E szolgáltatások igénybevétele után a szerzők felülvizsgálták a szöveget és teljes felelősséget vállalnak a cikk tartalmáért.*

## Felhasznált irodalom

Antoniak, Maria, Anjalie Field, Jimin Mun, Melanie Walsh, Lauren Klein & Maarten Sap (2023): Riveter: Measuring Power and Social Dynamics Between Entities. In: Danushka Bollegala, Ruihong Huang & Alan Ritter (eds.): *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Volume 3: System Demonstrations*, pp. 377–388. Association for Computational Linguistics, https://doi.org/10.18653/v1/2023.acl-demo.36

Archer, Allison M. & Joshua Clinton (2018): Changing Owners, Changing Content: Does Who Owns the News Matter for the News? *Political Communication,* vol. 35, no. 3, pp. 353–370, https://doi.org/10.1080/10584609.2017.1375581

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio (2015): Neural Machine Translation by Jointly Learning to Align and Translate. In: Yoshua Bengio & Yann LeCun (eds.): *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, https://arxiv.org/abs/1409.0473

Bajomi-Lázár Péter (2013): The Party Colonisation of the Media: The Case of Hungary. *East European Politics and Societies,* vol. 27, no. 1, pp. 69–89, https://doi.org/10.1177/0888325412465085

Bajomi-Lázár Péter (2021): Hungary's Clientelistic Media System. In: James Morrison, Jen Birks & Mike Berry (eds.): *The Routledge Companion to Political Journalism*, pp. 103–110, Routledge, https://doi.org/10.4324/9780429284571-9

Bajomi-Lázár Péter & Horváth Kata (2023/2025): Two Journalistic Cultures in One Country. The Case of Hungary in the Light of Journalists' Discourses on Fake News. *Journalism Practice*, vol. 19, no. 5, pp. 1055–1073.

Balaban, Delia Cristina, Ioana Iancu & Meda Mucundorfeanu (2016). Framing Obama's Re-election. A Comparative Analysis of German, French and Romanian Media Coverage. *Romanian Journal of Communication and Public Relations*, vol. 16, no. 2, pp. 75–105, https://doi.org/10.21018/rjcpr.2014.2.183

Baly, Ramy, Giovanni Da San Martino, James Glass & Preslav Nakov (2020): We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In: Bonnie Webber, Trevor Cohn, Yulan He & Yang Liu (eds.): *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 4982–4991. Association for Computational Linguistics, https://doi.org/10.18653/v1/2020.emnlp-main.404

Bátorfy Attila & Urbán Ágnes (2020): State Advertising as an Instrument of Transformation of the Media Market in Hungary. *East European Politics,* vol. 36, no. 1, pp. 44–65, https://doi.org/10.1080/21599165.2019.1662398

Bognár Éva (2021): Hungary. In: Nic Newman, Richard Fletcher, Anne Schulz, Simge Andı, Craig T. Robertson & Rasmus Kleis Nielsen (eds.): *Reuters Institute Digital News Report 2021,* pp. 83–84. Reuters Institute for the Study of Journalism.

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

Chai, Christine P. (2023): Comparison of Text Preprocessing Methods. *Natural Language Engineering*, vol. 29, no. 3, pp. 509–553, https://doi.org/10.1017/S1351324922000213

Chen, Tiangi & Carlos Guestrin (2016): XGBoost: A Scalable Tree Boosting System. In: Jacques Garrigue, Gabriele Keller & Eijiro Sumii (eds.): *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 785–794. Association for Computing Machinery, https://doi.org/10.1145/2939672.2939785

Csapo-Sweet, Rita M. & Ildikó Kaposi (1999): Mass Media in Post-Communist Hungary. *International Communications Bulletin,* vol. 34, no. 1–2, https://userpages.umbc.edu/~hasegawa/aejmc-icd/spring99/mediainhungary.html

Ding, Xiaohan, Michael Horning & Eugenia H. Rho (2023): Same Words, Different Meanings: Semantic Polarization in Broadcast Media Language Forecasts Polarity in Online Public Discourse. In: Yu-Ru Lin, Meeyoung Cha & Daniele Quercia (eds.): *Proceedings of the International AAAI Conference on Web and Social Media,* vol. 17, no. 1, pp. 161–172, Association for the Advancement of Artificial Intelligence Press, https://doi.org/10.1609/icwsm.v17i1.22135

Elman, Jeffrey L. (1990): Finding Structure in Time. *Cognitive Science,* vol. 14, no. 2, pp. 179–211, https://doi.org/10.1207/s15516709cog1402_1

Entman, Robert M. (1993): Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication,* vol. 43, no. 4, pp. 51–58, https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Eurotopics (2026): *Origo. Bundeszentrale für politische Bildung*, https://www.eurotopics.net/en/197845/origo

Fan, Lisa, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, Lu Wang (2019): In Plain Sight: Media Bias through the Lens of Factual Reporting. In: Sebastian Padó & Ruihong Huang (eds.): *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* pp. 6343–6349. Association for Computational Linguistics, https://doi.org/10.18653/v1/D19-1664

Fernández-Delgado, Manuel, Cernadas Eva, Barro Senén & Amorim Dinani (2014): Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research,* vol. 15, no. 90, pp. 3133–3181.

Gerken, Fynn & Toni van der Meer (2019): Crisis Frame Dynamics: Frame Diversity in News Media and the Role of Governmental Actors. *Journal of International Crisis and Risk Communication Research*, vol. 2, no. 2, pp. 149–180, https://doi.org/10.30658/jicrcr.2.2.1

Grimmer, Justin & Brendon M. Stewart (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis,* vol. 21, no. 3, pp. 267–297, https://doi.org/10.1093/pan/mps028

Hallin, Daniel C. & Paolo Mancini (2004): *Comparing Media Systems: Three Models of Media and Politics*. Cambridge University Press, https://doi.org/10.1017/CBO9780511790867

Hochreiter, Sepp & Jürgen Schmidhuber (1997): Long Short-Term Memory. *Neural Computation,* vol. 9, no. 8, pp. 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735

Laver, Michael, Kenneth Benoit & John Garry (2003): Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review,* vol. 97, no. 2, pp. 311–331, https://doi.org/10.1017/S0003055403000698

LeCun, Yann, Bernhard E. Boser, John S. Denker, Don Henderson, Richard E. Howard, Wayne E. Hubbard & Lawrence D. Jackel (1989): Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation,* vol. 1, no. 4, pp. 541–551, https://doi.org/10.1162/neco.1989.1.4.541

Lundberg, Scott & Su-In Lee (2017): A Unified Approach to Interpreting Model Predictions. In: Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan & Roman Garnett: (eds.): *Advances in Neural Information Processing Systems 30 (NeurIPS 2017),* pp. 4765–4774. Curran Associates.

McCombs, Maxwell E. & Donald L. Shaw (1972): The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly,* vol. 36, no. 2, pp. 176–187, https://doi.org/10.1086/267990

Mykkänen, Markus, Anri Patron, Mark Badham & Chiara Valentini (2024): Leveraging NLP for Crisis Communication Management: A Case Study of News Media Analysis of the COVID-19 Pandemic in Two Nordic Countries. *Journal of Contingencies and Crisis Management*, vol. 32, no. 3, pp. 1–19, https://doi.org/10.1111/1468-5973.12611

Orosz, György, Gergő Szabó, Péter Berkecz, Zsolt Szántó & Richárd Farkas (2023): Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In: Kamil Ekštein, František Pártl & Miloslav Konopík (eds.): *Text, Speech, and Dialogue. TSD 2023. Lecture Notes in Computer Science*, vol. 14102, pp. 58–69, Springer, https://doi.org/10.1007/978-3-031-40498-6_6

Polyák Gábor (2020): The Decline and Fall of Hungary's Leading News Brand. *EJO.ch*, 3 September 2020, https://en.ejo.ch/media-politics/press-freedom/the-decline-and-fall-of-hungarys-leading-news-brand

Polyák Gábor & Urbán Ágnes (2024): Hungary: One Country, Two Media Systems. In: Aljosha Karim Schapals & Christian Pentzold (eds.): *Media Compass: A Companion to International Media Landscapes*, pp. 70–83, Wiley, https://doi.org/10.1002/9781394196272.ch8

Potthast, Martin, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff & Benno Stein (2018): A Stylometric Inquiry into Hyperpartisan and Fake News. In: Iryna Gurevych & Yusuke Miyao (eds.): *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pp. 231–240. Association for Computational Linguistics, https://doi.org/10.18653/v1/P18-1022

Ribeiro, Marco Tulio, Sameer Singh & Carlos Guestrin (2016): Why Should I Trust You? Explaining the Predictions of Any Classifier. In: Jacques Garrigue, Gabriele Keller & Eijiro Sumii (eds.): *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 1135–1144. ACM, https://doi.org/10.1145/2939672.2939778

Schuster, Mike & Kuldip K. Paliwal (1997): Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing,* vol. 45, no. 11, pp. 2673–2681, https://doi.org/10.1109/78.650093

Sullivan, Gail M. & Richard Feinn (2012): Using Effect Size – Or Why the *P* Value is Not Enough. *Journal of Graduate Medical Education,* vol. 4, no. 3, pp. 279–282, https://doi.org/10.4300/JGME-D-12-00156.1

Médiakutató, 2026. tél
XXVII. évf. 4. szám

# Online Melléklet

## Tartalomjegyzék

1. Szövegtisztítás és -előkészítés
2. Alkalmazott modellek
3. Modellek paramétereinek hangolása és konfigurálása
4. Logisztikus regressziós modell tanítása
5. Az XGBoost modell tanítása
6. BiLSTM
7. Az Index cikkek besorolása
8. Robusztusság-ellenőrzés
   8.1. A szűrési eljárás
   8.2. A modellek teljesítménye
   8.3. Az Index-cikkek osztályozásának eredményei
   8.4. Hipotézisvizsgálat
   8.5. Összevetés
Felhasznált irodalom

## 1. Szövegtisztítás és -előkészítés

### Normalizálás

A kisbetűsítés csökkenti a szókincs méretét, mivel enélkül az „Alma” és az „alma” két különböző szótártagként szerepelne a modellben, jóllehet szemantikailag ekvivalensek (Jurafsky & Martin 2026).

A webscraping során a korpuszba HTML-elemek, speciális írásjelek és karakterek, mint például hashtagek, százalékjelek, pénznemjelölők, kerülhetnek bele. Ezek a szöveg szemantikai tartalmához nem járulnak hozzá, csupán zajt képviselnek, ezért eltávolításuk általában javítja a modellek teljesítményét (Petridis 2024).

Az emotikonok különböző érzelmeket, szarkazmust, nyomatékot vagy kontextust segítenek kifejezni az írásos kommunikációban. Szentimentelemzésnél és közösségimédia-elemzésnél megtartásuk általában ajánlott, viszont online hírportálok esetében eltávolításuk inkább indokolt (Michel & Neubig 2018).

A számok kezelése nem fekete-fehér kérdés. Az egyedi, nagyon ritka numerikus értékek (telefonszámok, azonosítók, árak) erősen növelik a vektortér dimenzionalitását anélkül, hogy érdemi szemantikai tartalmat hordoznának. Ugyanakkor bizonyos kontextusban a számok informatívak lehetnek, például egy orvosi kutatásban lényeges különbség, hogy a páciensek 5%-ánál vagy 50%-ánál tapasztaltak javulást. Lehetséges köztes megoldás a számok helyettesítése egy egységes NUM-tokennel, amely megőrzi a numerikus tartalom tényét, miközben csökkenti az egyedi értékek számát (Thawani et al. 2021).

A következő normalizációs lépéseket hajtottuk végre. Először kisbetűsítettük a szövegeket. Utána eltávolítottunk minden URL-t és email címet. Ezt követően minden speciális karaktert, számot, írásjelet, emotikont szóközre cseréltünk, egyedül a magyar ábécé betűi hagytunk benne, mivel hírportálok szövegkörnyezetében ezek a karakterek megkülönböztető ereje alacsony. A cikkek címeit és tartalmát összevontuk egy mezőbe. Ezt az indokolja, hogy a cím a szerkesztői stílus egyik legsűrítettebb hordozója, amelynek kihagyása az elemzési egység fontos részét érintené. Végül minden rekordhoz egyedi azonosítót (id) rendeltünk.

## Tokenizálás

A tokenizálás az a szövegfeldolgozási lépés, amelynek során a dokumentumot tokenek sorozatára bontjuk. Ez a folyamat a szövegbányászati és természetes nyelvfeldolgozási feladatok döntő többségének alapvető eleme. Tikk (2007) meghatározása szerint a token egy karaktersorozat konkrét dokumentumbeli előfordulása, míg típusnak nevezi az azonos karaktersorozatot tartalmazó tokenek osztályát. A típusok összessége alkotja a szótárat.

Az alábbi példa szemlélteti a tokenizálás folyamatát:

> Eredeti mondat: „A kormány bejelentette az új intézkedéseket."
> Tokenek: [„a", „kormány", „bejelentette", „az", „új", „intézkedéseket", „."]

## Stopszószűrés

A természetes nyelvben nagy gyakorisággal előforduló, de tartalmi információt nem hordozó, megkülönböztető képességgel nem rendelkező szavakat stopszavaknak nevezzük. A magyar nyelvben ilyenek például „a", „és" vagy „hogy". Mivel ezek szinte minden dokumentumban megjelennek, eltávolításukkal lényegesen csökkenthető a szó-dokumentum mátrix nemzérus bejegyzéseinek száma, ami kedvező hatással van mind a tárolási igényre, mind a modell számítási hatékonyságára (Tikk 2007).

Osztályozási feladatoknál ajánlott a stopszavak eltávolítása, mivel segíti a modellt abban, hogy a megkülönböztető tokenekre összpontosítson (Orebi & Naser 2025). Ezért a bag-of-words modellek (logisztikus regresszió, XGBoost) számára készített korpuszon a spaCy csomag magyar stopszólistáját alkalmaztuk, amelyet kiegészítettünk a hírportálok szakzsargonjához tartozó szavakkal (pl. hír, tudósít, cikk), valamint azokkal a szóalakokkal, amelyek explicit módon tartalmazzák az egyes portálok nevét (pl. hvg, hvgnek, hvgben).

## Lemmatizálás

A lemmatizálást a természetes nyelvfeldolgozásban morfológiailag gazdag nyelvek esetében különösen előnyös. A stemmingtől való fő különbség abban ragadható meg, hogy amíg a stemming csupán a szóvégi toldalékokat távolítja el heurisztikusan (és nem feltétlenül eredményez értelmes szóalakot), addig a lemmatizálás minden esetben a szó szótárban rögzített alapalakját adja vissza (Jurafsky & Martin 2000).

Mindkét eljárást alkalmazzák a vektortér-modell felépítése során azzal a céllal, hogy az egyes szavak különböző morfológiai változatait közös formára hozzák, lehetővé téve egységes reprezentációjukat. A gyakorlatban ez egyedi szótárelemszám csökkenésével jár, ami mérsékli a vektortér dimenzionalitását. Angol nyelv esetén ez a csökkenés 40-70%-os is lehet, míg a magyarban egyes megfigyelések szerint akár 90%-os mértéket is elérhet (Tikk 2007). Jelen kutatásban mindkét korpuszon lemmatizálást végeztem el a spaCy csomag, hu_core_news_lg modellje segítségével, hogy a vektortér dimenzióinak redukálásával csökkentsem a modellek számítési igényét (Honnibal & Montani 2017).

A fent bemutatott előkészítési lépések végrehajtása után a cikkek az alábbi módon néznek ki a BOW korpusz esetében:

{
„id": 59,
„portal": „hvg",
„year": „2021",
„text": „per nyer jakab péter fejér megyei hírlap helyreigazítás megfizetés kötelez bíróság…"
}

### A TF-IDF-súlyozás

A TF-IDF-súlyozás két komponens szorzatából tevődik össze. A kifejezésgyakoriságból (*term frequency,* TF) és az inverz dokumentumgyakoriságból (*inverse document frequency,* IDF) (Salton & Buckley, 1988).

A kifejezésgyakoriság azt adja meg, hányszor fordul elő az adott *t* kifejezés a *d* dokumentumban. Az alapfeltevés az, hogy minél többször szerepel egy szó egy dokumentumban, annál valószínűbb, hogy a dokumentum releváns az adott szempontból. Ez a megközelítés a *bag-of-words* modellek alapja. Itt a szavak sorrendje nem számít, kizárólag előfordulásuk gyakorisága. A TF-érték önmagában azonban nem elegendő, mivel minden szót egyformán fontosnak kezel, holott bizonyos, az egész gyűjteményben elterjedt szavak (például egy autóipari korpuszban az „autó" szó) nem segítenek a dokumentumok egymástól való megkülönböztetésében (Salton & Buckley, 1988).

Ezt a problémát orvosolja az inverz dokumentumgyakoriság ($idf_t$). Ennek célja a gyűjteményszinten túl gyakori és ezért kevés megkülönböztető erővel bíró szavak súlyának csökkentése. Kiszámítása a dokumentumgyakoriságon ($df_t$) alapul, amely azt mutatja meg, hány dokumentumban szerepel az adott kifejezés:

$$idf_t = log \frac{N}{df_t}$$

ahol *N* a gyűjtemény összes dokumentumának száma. A ritka kifejezések így magas, az elterjedt szavak alacsony IDF-értéket kapnak. Yang és Pedersen (1997) empirikusan igazolták, hogy a dokumentumszintű *df*-statisztika jobb megkülönböztető erővel rendelkezik, mint a kifejezés összesített korpuszbeli előfordulási száma.

A TF-IDF-súly a két érték szorzata:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t$$

A súly akkor a legmagasabb, ha a kifejezés sokszor fordul elő az adott dokumentumban (magas *tf*), miközben a gyűjtemény egészében csak kevés dokumentumban jelenik meg (magas *idf*). A legalacsonyabb értéket azok a szavak kapják, amelyek szinte minden dokumentumban jelen vannak. Ezek szemantikailag közel állnak a korábban tárgyalt stopszavakhoz (Tikk 2007).

A TF-IDF súlyozás segítségével minden dokumentum egy vektorként ábrázolható, amelynek dimenzióit a szótár elemei, koordinátáit pedig a hozzájuk tartozó súlyok alkotják. Ez a reprezentáció alapozza meg a vektortér-modellt amelynek keretében egy *q* lekérdezéshez tartozó dokumentumpontszám a lekérdezésben szereplő szavak dokumentumbeli TF-IDF-súlyainak összegeként számítható:

$$Score(q, d) = \sum_{t \in q} tf - idf_{t,d}$$

A két Bag of Words modellnek TF-IDF reprezentáció készült unigram és bigram szinten.

## 2. Alkalmazott modellek

### Logisztikus regresszió

A klasszifikáció olyan statisztikai feladat, amelynek során megfigyeléseket sorolunk be előre meghatározott osztályokba vagy kategóriákba. A logisztikus regresszió is klasszifikációs problémákra használható, és kimenetét valószínűségként értelmezhetjük. A logisztikus regressziót az "An Introduction to Statistical Learning" könyv 4.3-as fejezete alapján mutatjuk be (James 2013).

A logisztikus regresszió bináris kimenetelű célváltozóra épül, ahol $Y \in \{0, 1\}$. A modell célja $P(Y=1|X)$ feltételes valószínűség becslése. Lineáris regresszió esetén a becsült érték:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Ez az összefüggés valószínűségi becslésre nem alkalmazható, mivel az eredmény tetszőlegesen kicsiny vagy nagy értéket felvehet. A logisztikus regresszió ezt a problémát a szigmoid (logisztikus) függvény alkalmazásával oldja meg:

$$p(X) = exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p) / [1 + exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)]$$

Ez a transzformáció garantálja, hogy $0 \leq p(X) \leq 1$ teljesül minden lehetséges prediktorértékre. A modell lineáris formában az esélyhányados (odds) természetes logaritmusát, azaz a logitet írja le:

$$log[p(X) / (1 - p(X))] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

A logit modell értelmezése szerint a prediktorok lineárisan hatnak az odds logaritmusára. Ez a linearitás a logit-térben értendő, az eredeti valószínűségi skálán a hatás szükségszerűen nemlineáris.

A logisztikus regresszió paramétereit maximális likelihood becsléssel határozható meg. Az eljárás olyan β-értékeket keres, amelyek maximalizálják a megfigyelt adatok együttes valószínűségét. A likelihood-függvény maximuma általában iteratív numerikus módszerekkel (pl. Newton-Raphson-algoritmus) érhető el.

A regressziós együtthatók a logit-skálán értelmezendők. Egy egységnyi növekedés $X_j$-ben $\beta_j$ egységgel változtatja meg az odds logaritmusát. Az együttható exponenciálisa az odds ratio (esélyhányados):

$$OR = exp(\beta_j)$$

Például $\beta_j = 0{,}7$ esetén $\exp(0{,}7) \sim 2{,}01$, ami azt jelenti, hogy az adott változó egységnyi növekedése közel megduplázza az esemény bekövetkezésének odds-át.

A becsült modell egy $\hat{p}$ valószínűséget rendel minden megfigyeléshez. Az osztályba soroláshoz egy küszöbértéket (c) kell meghatározni:

$$ha\ \hat{p} > c \rightarrow \hat{Y} = 1,$$
$$ha\ \hat{p} \leq c \rightarrow \hat{Y} = 0.$$

Az alapértelmezett küszöbérték c = 0,5, de ez a feladat természetétől és a hibák relatív költségétől függően szabadon megválasztható. Aszimmetrikus osztályeloszlás vagy különböző típusú hibák eltérő súlyozása esetén alacsonyabb (pl. 0,3) vagy magasabb (pl. 0,7) küszöb alkalmazása is indokolt lehet.

A logisztikus regresszió erősségei közé tartozik az egyszerű interpretálhatóság, a viszonylag alacsony számítási igény, valamint az, hogy kis mintákon is kielégítő teljesítményt tud nyújtani. Az együtthatók esélyhányados formájában közvetlen tartalmi értelmezést tesznek lehetővé.

Ugyanakkor a modell feltételezi, hogy a prediktorok lineárisan hatnak a logit-térben, ami komplex, nemlineáris összefüggések esetén erős megszorítás. Ilyen helyzetekben a modell teljesítménye romlik, és kiterjedt jellemzőtechnikára (feature engineering) lehet szükség. Erősen nemlineáris adatstruktúrák esetén rugalmasabb modellek, mint a XGBoost vagy a neurális hálók, általában jobb eredményt adnak (Fernández-Delgado, Cernadas, Barro & Amorim 2014).

## XGBoost

Az XGBoost egyik megkülönböztető tulajdonsága, hogy a célfüggvénye explicit regularizációs tagot tartalmaz. A tanítási célfüggvény:

$$Obj(\Theta) = \Sigma_i\, l(y_i, \hat{y}_i) + \Sigma_k\, \Omega(f_k)$$

ahol $l(y_i, \hat{y}_i)$ a veszteségfüggvény (pl. négyzetes hiba regresszióhoz, logisztikus veszteség osztályozáshoz), $\Omega(f_k)$ pedig a regularizációs tag, amely bünteti a túl komplex fákat. A regularizációs tag explicit alakja:

$$\Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2$$

ahol T a levelek száma, w a levél-pontszámok vektora, γ a minimális haszoncsökkentés egy újabb osztás elvégzéséhez (ezzel szabályozva a fa növekedését), λ pedig az L2 regularizációs együttható. Ez a formális megközelítés elősegíti a jobb általánosítást és a túlillesztés mérséklését.

Minden egyes iterációban egyetlen új fát tanít be, miközben a többi fát rögzíti, azaz az XGBoost a modellt additív módon építi fel. A t-edik iterációban a cél olyan $f_t$ megtalálása, amely minimalizálja:

$$Obj^{(t)} = \Sigma_i\, l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

A számítás hatékonyságát másodrendű Taylor-közelítéssel éri el: a veszteségfüggvényt az aktuális predikcióhoz közelíti, így a célfüggvény zárt formában optimalizálható. Ez a megközelítés gyorsabb konvergenciát biztosít az elsőrendű, csak gradienst használó, módszereknél.

A modell főbb hiperparamétereit, amikre a modellem bemutatása során hivatkozni is fogok, ismertetem. Az eta (tanulási ráta) azt szabályozza, hogy az egyes fák milyen arányban járulnak hozzá az összmodellhez. Kisebb értéke lassabb, de rendszerint jobban általánosító tanulást eredményez. A max_depth paraméter a fák maximális mélységét határozza meg: nagyobb érték összetettebb modellt jelent, de növeli a túlillesztés kockázatát. Az n_estimators a tanított fák számát jelöli, amelyet jellemzően keresztvalidációval választódik meg. A subsample és colsample_bytree paraméterek véletlen mintavételezést alkalmaznak a tanítóadatokon, illetve a jellemzőkön, ami szintén a túlillesztés ellen hat.

## BiLSTM

A valós szituációkban gyakran előfordulnak olyan adatstruktúrák, amelyekben az elemek sorrendje jelentéssel bíró információt hordoz. Ezek a szekvenciális adatok, mint például az írott vagy beszélt szövegek, az idősorok vagy a hangjelek. Egy hírszöveg esetén például egy szó valódi értelme csak a kontextusából érthető meg pontosan. Ez azzal jár, hogy ha a szórenden változtatunk az a mondanivaló változásával is járhat.

A logisztikus regresszió, az XGBoost és a többi bag-of-words alapú modell, ezzel az összefüggéssel nem számol. A szózsák-modellek egy dokumentumot a szavak előfordulásainak halmazaként kezelnek, figyelmen kívül hagyva az elemek sorrendjét és egymáshoz való viszonyát (Jurafsky & Martin 2026). Ez hatékony és jól értelmezhető megközelítés, de elveszíti a szöveg szekvenciális szerkezetéből adódó információt.

A neurális hálózatok megjelenésével lehetővé vált a szekvenciális minták automatizált tanulása. LeCun és munkatársai (1989) konvolúciós hálózatokkal, kézírásfelismerési feladaton, mutatták meg, hogy hierarchikus jegyeket lehet megtanulni nyers adatból. A természetes nyelv feldolgozásában azonban az igazi áttörést a visszacsatolásos (rekurrens) architektúrák hozták, amelyek képesek explicit módon kezelni az időbeli függőségeket (Elman, 1990).

A rekurrens neurális hálózatok (RNN) alapegysége egy visszacsatolt feldolgozó egység, amely rejtett állapottal (hidden state) rendelkezik. Ez a rejtett állapot az előző időlépések információit tárolja, és lehetővé teszi, hogy a hálózat figyelembe vegye a korábbi bemeneteket az aktuális feldolgozás során. A rejtett állapot minden időlépésben frissül az aktuális bemenet és az előző állapot alapján, így a modell képes információt továbbvinni az időben.

Ellentétben az előrecsatolt neurális hálókkal, ahol minden teljesen összekapcsolt réteg külön súlymátrixszal rendelkezik, az RNN-ek az időlépések között ugyanazt a súlymátrixot használják. Ez a paramétermegosztás hatékonyá teszi a modellt, és biztosítja, hogy bármilyen hosszúságú szekvenciára alkalmazható legyen. A rejtett állapot frissítési szabálya a következő egyenlettel írható le:

$$h_t = tanh(W_h \cdot h_{t-1} + W_x \cdot x_t + b)$$

ahol $x_t$ az aktuális bemeneti vektor, $h_{t-1}$ az előző rejtett állapot, W$h$ és W$x$ a tanítható súlymátrixok, $b$ a bias-vektor. A tanh aktivációs függvény a kimenetet [–1, 1] tartományba szorítja, ami a nullára szimmetrikus kimenet révén stabilabb gradienst biztosít a tanulás során, mint például a szigmoid függvény.

Az RNN kiterítése (unfolding) az ismétlődő struktúra időlépések mentén történő kibontása: a szekvencia minden időlépése egy önálló rétegként jelenik meg, szemléltetve az információáramlást. Ez a kibontás teszi lehetővé az időbeli visszaterjesztést (Backpropagation Through Time, BPTT), amelynek során a hibajelzés visszaáramlik az időlépéseken át, és ennek megfelelően módosulnak a háló súlyai (Werbos, 1990).

Az RNN-ek azonban kritikus problémával küzdenek: a gradiens-eltűnés (vanishing gradient) és a gradiens-robbanás (exploding gradient) jelenségével. A BPTT eljárás során a gradiens értéke az időlépések mentén szorzódik az aktivációs függvény deriváltjával és a rekurrens súlymátrix sajátértékeivel. Ha ez az érték következetesen kisebb 1-nél, a gradiens exponenciálisan csökken, ha pedig rendszeresen nagyobb 1-nél, numerikus instabilitást

okoz. A gradiens-eltűnés problémáját Hochreiter (1991) diplomamunkája formálisan azonosította és elemezte, ezentúl Bengio és munkatársai (1994) pedig további elméleti keretet adtak a jelenségnek. Pascanu és munkatársai (2013) empirikusan is demonstrálták, hogy standard SGD-vel az alap-RNN jellemzően nem képes 20 lépésnél hosszabb függőségek megtanulására.

Egy hírportál cikkének stílusa, retorikai eszközei, témafelvetései a cikk egészén átívelnek, ezért a modellnek a szöveg korábbi részeire is emlékeznie kell, amikor a késői szakaszokat dolgozza fel.

Az alap RNN erre elvileg képes lenne, de a gradiens-eltűnés miatt a gyakorlatban nem tanul meg hosszútávú függőségeket. Hochreiter és Schmidhuber (1997) erre a problémára dolgozták ki a Long Short-Term Memory (LSTM) architektúrát. Ennek az alapgondolata az, hogy vezessünk be egy külön, explicit cellaállapotot (cell state), amelyet kapumechanizmusokkal, kontrollált módon frissítünk. Így az LSTM képes szükség esetén hosszú távon is megőrizni az információt, és aktívan dönteni arról, mikor felejtsen, mikor tanuljon, és mikor adjon ki információt.

Az LSTM-cella két párhuzamos állapotot tart fenn. Az első a cellaállapot ($C_t$), amely a hosszú távú memória szerepét tölti be: ez az a vektor, amely a szekvencia egészén végigfut, és amelyre a kapuk szelektíven írnak és olvasnak. A második a rejtett állapot ($h_t$), a rövid távú, feladatorientált kimenet. Ez az ami átkerül a következő időlépésbe, és ez az osztályozó fej bemenete is (Hochreiter & Schmidhuber, 1997).

A cellaállapot frissítését és a rejtett állapot képzését három kapu (gate) szabályozza. Mindegyik kapu egy szigmoid aktivációs függvényt alkalmaz, amelynek kimenete a [0, 1] intervallumba esik. A 0 az információ elnyomását, az 1 annak teljes továbbengedését jelenti, míg a köztes értékek részleges áteresztést valósítanak meg (Hochreiter & Schmidhuber, 1997).

A felejtési kapu azt dönti el, hogy a cellaállapot melyik részét kell megőrizni és melyiket kell törölni. Bemenete az előző rejtett állapot és az aktuális bemeneti vektor, a kimenete egy [0, 1]-es súlyvektor, amellyel a cellaállapot elemeit megszorozza. Fontos megjegyezni, hogy az eredeti LSTM architektúra (Hochreiter & Schmidhuber, 1997) még nem tartalmazott felejtési kaput. Gers, Schmidhuber és Cummins (2000) vezette be, felismerve, hogy explicit mechanizmus nélkül az irrelevánssá váló információk felhalmozódhatnak a cellaállapotban.

A bemeneti kapu szabályozza, hogy milyen új információ kerüljön be a cellaállapotba. Két párhuzamos számítás zajlik benne. Az első a tényleges kapu ($i_t$) amely meghatározza, hogy melyik pozíciók frissüljenek. A második a jelölt cellaállapot ($\hat{c}_t$), ami pedig meghatározza, hogy milyen értékkel. A tanh függvény a jelölt értékeket [–1, 1] tartományba normalizálja.

A kimeneti kapu határozza meg, hogy a frissített cellaállapot melyik dimenzióit kell a rejtett állapotba, és így a következő réteg vagy lépés bemenettébe továbbítani.

A következőkben az LSTM-cella működését ismertetjük a *t*-edik időlépésben. Használt jelölések: $x_t$ az aktuális bemenet, $h_{t-1}$ az előző rejtett állapot, $C_{t-1}$ az előző cellaállapot, $\sigma$ a logisztikus szigmoid, $\odot$ az elemenkénti (Hadamard-)szorzat.

1. A felejtési kapu az előző rejtett állapot és az aktuális bemenet alapján meghatározza, mit felejtsen el a cellaállapotból:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

2. A bemeneti kapu és a jelölt cellaállapot meghatározza, milyen új értékek kerüljenek be:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\hat{c}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

3. A cellaállapot frissítése, a régi állapot kívánt részének megőrzése és az új értékek hozzáadása:

$$C_t = f_t \odot C_{t-1} + i_t \odot \hat{c}_t$$

4. A kimeneti kapu meghatározza, hogy a cellaállapot mely részei jelenjenek meg a rejtett állapotban:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \odot tanh(C_t)$$

Az alap-RNN multiplikatív frissítést alkalmaz (ami a gradiens-eltűnéshez vezet), ezzel szemben az LSTM cella-állapota összeadással frissül, így a gradiens viszonylag változatlanul képes áramlani az időlépéseken keresztül (Hochreiter & Schmidhuber, 1997).

Az LSTM legfontosabb erőssége a hosszú távú összefüggések kezelésének képessége, amelyet a hagyományos RNN-ek nem tudnak megbízhatóan ellátni. A kapumechanizmus flexibilis, vagyis a hálózat maga tanulja meg, mikor és mit felejtsen, illetve tartson meg, ami feladatspecifikus adaptálódást tesz lehetővé. Greff és munkatársai (2017) egy nagyléptékű összehasonlító tanulmányban nyolc LSTM-variánst vizsgáltak három különböző feladaton (beszédfelismerés, kézírásfelismerés, polifon zenei modellezés), és megállapították, hogy a felejtési kapu és a kimeneti aktivációs függvény a legkritikusabb elemek, ezek elhagyása jellemzően a legnagyobb teljesítménycsökkenéssel jár.

Ugyanakkor az LSTM rendelkezik gyengeségekkel. A kapumechanizmusok számos további paramétert vezetnek be a hagyományos RNN-hez képest, ami növeli a számítási és memóriaigényt, és kis adathalmazon túltanulás (overfitting) veszélyét hordozza. Emellett a belső reprezentáció nehezen értelmezhető (black-box jelleg), ami akár a bag-of-words alapú modellek egyszerű és átlátható jellegéhez képest hátrányként jelentkezik (Ribeiro et al. 2016). Az értelmezhetőség növelésére több módszer is rendelkezésre áll. A figyelemmechanizmus (Bahdanau et al. 2015) lehetővé teszi, hogy a modell a szekvencia egyes pozícióihoz súlyokat rendeljen, jelezve, mely szavak járultak hozzá leginkább a döntéshez. A LIME (Local Interpretable Model-agnostic Explanations; Ribeiro et al. 2016) módszert alkalmazzuk, amely lokálisan értelmező lineáris modelleket illeszt a fekete doboz köré, közelítő magyarázatot adva az egyes predikciókhoz.

Az egyirányú LSTM egyik további korlátja, hogy az adott időlépés rejtett állapota csak a korábbi kontextust tartalmazza, vagyis szövegben jobbra álló szavak nem befolyásolják a bal oldali pozíciók reprezentációját. Szövegek esetén ez korlátozó tényező, hiszen egy szó jelentését a tőle jobbra álló szavak is meghatározhatják.

A kétirányú LSTM (BiLSTM; Schuster & Paliwal, 1997) ezt a hiányosságot két párhuzamos LSTM-réteggel orvosolja. Az egyik a szokásos bal-jobb irányban dolgozza fel a szekvenciát, a másik jobbról balra halad. Minden időlépésben a két irány rejtett állapotát fűzik össze:

$$h_BiLSTM_t \;=\; [\rightarrow h_t \,;\, \leftarrow h_t]$$

ahol $\rightarrow h_t$ az előremutató, $\leftarrow h_t$ a visszamutató LSTM rejtett állapota a $t$-edik pozícióban. Az összefűzött vektor dimenziója így kétszerese az egyirányú változatának. Zhou és munkatársai (2016) szövegklasszifikációs kísérletekben igazolták, hogy a BiLSTM szisztematikusan felülmúlja az egyirányú változatot, különösen olyan feladatokon, ahol a mondat végi vagy dokumentum közepi kontextus visszamenőleg is értelmezi a korábbi szavakat.

## 3. Modellek paramétereinek hangolása és konfigurálása

Az adatokat tanító és teszt halmazokra osztottuk 80/20 arányban, rétegzéssel, ami biztosítja, hogy a HVG/Origo osztályok aránya a tanító és teszthalmazban is megőrződjön. Emellett a random seed-et is rögzítettük a reprodukálhatóság érdekében. A modellek kialakítása egységes, háromlépéses eljárás mentén történt. Először egy alapmodell illesztése, ezt követően hiperparaméter-hangolás, végül a legjobb paraméterekkel felépített végleges modell kiértékelése. A folyamat célja annak vizsgálata volt, hogy a hangolás milyen mértékben javítja az egyes modellek általánosító képességét.

Az elemzés teljes egésze Python nyelven készült, az Online Melléklet hivatkozásjegyzékében megtalálható a kódok GitHub linkje. A következőkben ismertetjük a főbb könyvtárakat amiket az elemzés során használtunk. A logisztikus regressziós modell, a jellemzők kinyerése, a kiértékelést segítő metrikák számítása, valamint a hiperparaméter-optimalizálás mind a scikit-learn könyvtárral valósult meg (Pedregosa et al. 2011). A gradiens-boosting alapú modellhez az XGBoost csomagot használtuk (Chen & Guestrin 2016), míg a BiLSTM architektúra készítése a TensorFlow/Keras csomaggal történt (Abadi et al. 2015). A BiLSTM modellek értelmezhetőségét a LIME (Ribeiro et al. 2016), az XGBoost modellekét pedig a SHAP (Lundberg & Lee 2017) könyvtárak segítették. Az adatokkal való dolgozáshoz a pandas és a NumPy (Harris et al. 2020), a vizualizációhoz a matplotlib (Hunter 2007) és a seaborn (Waskom 2021) könyvtárakat használtuk.

## 4. Logisztikus regressziós modell tanítása

### Tanítási konfiguráció és folyamat

Az első lépésben egy alapmodellt illesztettünk alapértelmezett paraméterekkel. A szövegeket TF-IDF reprezentációvá alakítottuk, ahol a jellemzők számát 5000-ben maximalizáltuk (max_features = 5000). L2 regularizációt alkalmaztunk (C = 1,0), az optimalizálást az lbfgs algoritmussal végezve. A konvergencia biztosítása érdekében az iterációk maximális számát 100-ról 1000-re növeltük. Az alapmodellt mind a 2019-es, mind a 2021-es korpuszra illesztettük.

A második lépésben GridSearchCV (Pedregosa et al. 2011) segítségével végeztük el a hiperparaméterek hangolását a teszt adatokon, ötszörös keresztvalidáció mellett, AUC-ROC optimalizálási célfüggvénnyel. A keresési tér a regularizáció erősségére ($C \in \{0{,}001; 0{,}01; 0{,}1; 1{,}0; 5{,}0; 10{,}0; 100{,}0\}$) és a jellemzők számára (max_features $\in$ {3000; 4000; 5000; 6000; 7000}) terjedt ki. Ez összesen 35 paraméterkombinációt jelentett, amelyeket 5-szörös cross validation használatával értékeltünk ki. Így 175 modell illesztése történt, amelyhez a GridSearchCV egy további újratanítást is hozzáadott a teljes tanítóhalmazon.

A nagyon alacsony C-értékek esetén a regularizáció túl erősnek bizonyult, aminek következtében a koefficiensek gyakorlatilag nullává váltak, és ez numerikus figyelmeztetéseket eredményezett, ugyanakkor a modell futását nem akadályozta.

A harmadik lépésben a legjobb paraméterekkel újraillesztettük a modellt (mind két esetben C = 5 és max_features = 7000), majd összehasonlítottuk az alapmodellel a teszt adatokon mért eredmények alapján. A hangolt modell mindkét év esetében javulást mutatott. 2019-ben a tesztpontosság 82,74%-ról 83,84%-ra, az AUC-ROC érték 0,9102-ről 0,9207-re nőtt. 2021-ben hasonló tendencia volt megfigyelhető, ahol a pontosság 85,80%-ról 87,26%-ra, az AUC-ROC pedig 0,9354-ről 0,9453-ra emelkedett.

A tanítási és tesztpontosság közötti különbség mindkét évben meghaladta a 7 százalékpontot, ami mérsékelt túlillesztésre utal. Ugyanakkor a teszteredmények javulása alapján a hiperparaméter-hangolás összességében pozitívan befolyásolta a modell általánosító képességét. Megfigyelhető továbbá, hogy a 2021-es adatokon tanított modellek jobb teljesítményt értek el, ami arra utalhat, hogy a két hírforrás közötti különbségek ebben az időszakban könnyebben tanulhatók. Az alacsony szórásértékek (2019: ±0,0068; 2021: ±0,0035) megerősítik, hogy a modellek stabil teljesítménnyel bírnak, és nem mutatnak erőteljes variabilitást az adatfelosztások között. Ez robusztus általánosítási képességre utal.

*1. táblázat*
*Az alapmodell és a hangolt modell teljesítménymutatói (2019 és 2021)*

| Metrika | 2019 tanítás | 2019 teszt | 2021 tanítás | 2021 teszt |
|---|---|---|---|---|
| **Pontosság** | 92,67% | 83,84% | 94,90% | 87,26% |
| **AUC-ROC** | 0,9792 | 0,9207 | 0,9885 | 0,9453 |
| **CV pontosság (átlag ± std)** | – | 83,22% ± 0,68% | – | 86,58% ± 0,35% |

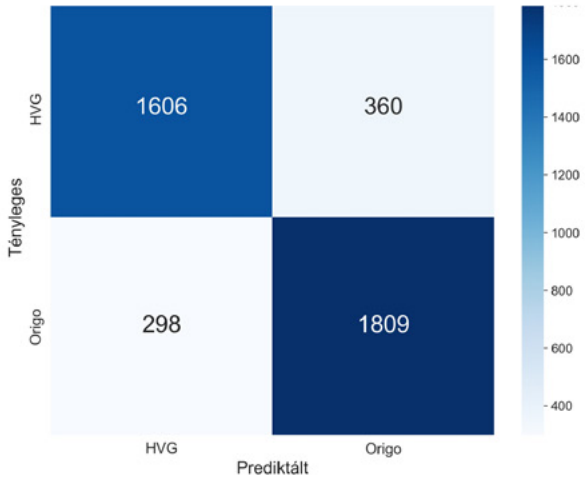
### Konfúziós mátrix

A konfúziós mátrixok alapján a 2019-es modellben a HVG-cikkek recall-értéke (81,7%) elmarad az Origo-cikkekétől (85,9%), ami arra utal, hogy a modell hajlamosabb volt a HVG-cikkeket tévesen Origo-ként besorolni. Ez utalhat arra, hogy a HVG stilisztikailag heterogénebb, ezzel megnehezíti a modell számára az egyértelmű döntési határvonal kialakítását a TF-IDF-térben.

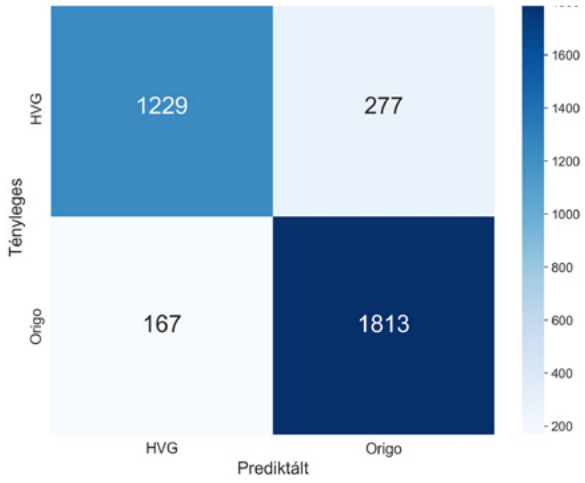
A 2021-es modellben ez az aszimmetria felerősödni látszik. Amíg a HVG recall lényegében változatlan maradt (81,6%), az Origo recall jelentősen, 91,6%-ra emelkedett. Az Origo-cikkek HVG-ként való téves besorolásának száma

közel felére csökkent (298-ról 167-re), ugyanígy a HVG-cikkek Origo-ként tévesztések is mérséklődtek (360-ról 277-re), bár kisebb arányban. Ez arra enged következtetni, hogy 2021-re elsősorban az Origo stilisztikai profilja vált markánsabbá és a modell számára könnyebben azonosíthatóvá. Emellett a két portál közötti nyelvi különbségek is nőttek.

*1a. ábra*
*Konfúziós mátrix – 2019-es hangolt modell*




*1b. ábra*
*Konfúziós mátrix – 2021-es hangolt modell*




## Jellemzőfontossági elemzés

A logisztikus regresszió egyik fő előnye az interpretálhatóság, mivel a prediktorok hatása közvetlenül értelmezhető a regressziós együtthatók alapján. Az előjel az osztályhoz való hozzájárulás irányát, az abszolút érték pedig a változó relatív fontosságát jelzi.

A 2019-es eredmények alapján (lásd a cikkben az 1. ábrát) a modell több olyan kulcsszót azonosít, amelyek jól elkülönítik a két hírforrást. Markáns különbség figyelhető meg a hivatkozási és retorikai igék használatában. A HVG esetében gyakoribbak a „szerinte" és „közöl" típusú kifejezések, míg az Origo inkább a „kiemel", „elmond" és „kifejt" igéket használja. Emellett lexikai szinten is megjelennek eltérések (pl. „ittas" vs. „részeg"), amelyek szintén hozzájárulnak a klasszifikációhoz.

A politikai kontextusú kifejezések szintén meghatározóak mind a két portál esetében. A „fideszes" vagy „ellenzék" a HVG-hez, a „baloldali", „szocialista" és „Soros" az Origo-hoz áll közelebb. Tematikus különbségek is megfigyelhetőék, amíg például az Origo esetében a „mentős" és „tűzoltó", addig a HVG-nél a „rendőrség" jelenléte erősíti az adott kategóriába tartozást.

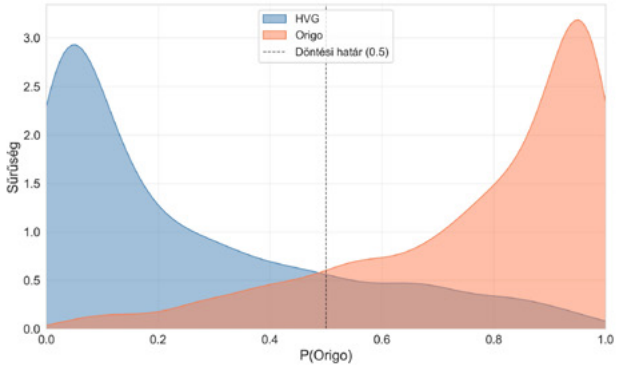
A 2021-es eredményeknél (lásd a cikkben a 2. ábrát) a retorikai és politikai különbségek továbbra is fennmaradnak, ezzel nagymértékben megerősítve a 2019-es mintázatokat.

## A becsült valószínűségek eloszlása

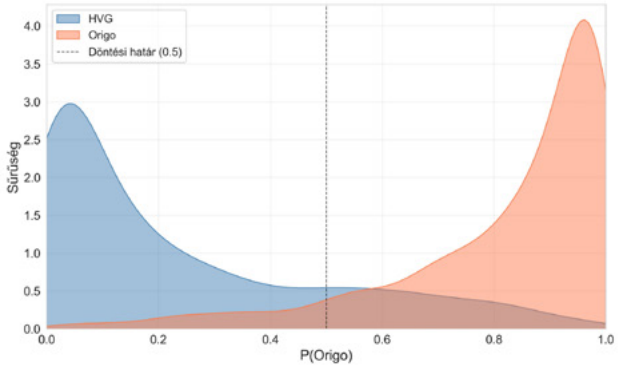
A becsült valószínűségek eloszlásának vizsgálatával megállapítható, hogy a 2019-es modell esetében a két osztály valószínűségi eloszlása jól elkülönül. Ennél a modellnél a HVG-cikkek többsége alacsony, míg az Origo-cikkek magas P(Origo) értékeket kapnak, ugyanakkor a 0,4-0,6 közötti tartományban megfigyelhető egy mérsékelt átfedés, ami bizonytalanságot jelez. A 2021-es modellnél az eloszlások szétválása még hangsúlyosabbá válik, különösen az Origo-cikkek esetében, ahol a valószínűségek erősen a 0,9–1,0 tartományba koncentrálódnak. Ezzel párhuzamosan a HVG-eloszlás továbbra is az alacsony tartományban marad, bár a középső intervallumban enyhe átfedés itt

is jelen van. A döntési határ (0,5) környezetében mindkét évben viszonylag kevés megfigyelés található, ami stabil klasszifikációs döntésekre utal. Összességében az mondható, hogy a 2021-es modell nagyobb bizonyossággal különíti el a két osztályt, összhangban magas AUC-ROC értékekkel.

*2a. ábra*
*Valószínűség-eloszlás – 2019-es modell*




*2b. ábra*
*Valószínűség-eloszlás – 2021-es modell*




## 5. Az XGBoost modell tanítása

### Tanítási konfiguráció és folyamat

Az XGBoost-modell esetében szintén TF-IDF reprezentációt alkalmaztunk, 5000 jellemzővel (max_features = 5000). A tanító és teszthalmaz arányát 80-20%-ban rögzítettük (test_size = 0,2, random_state = 67).

A hiperparaméter-hangolást GridSearchCV segítségével végeztük, ötszörös keresztvalidációval. A keresési tér a következő paramétereket tartalmazta: max_depth ∈ {4, 6, 8}, learning_rate ∈ {0,05; 0,1}, n_estimators ∈ {200; 300; 500}, colsample_bytree ∈ {0,3; 0,5}, valamint subsample ∈ {0,7; 1,0}. Ez összesen 72 paraméterkombinációt eredményezett, amelyek 5 validációs hajtáson kerültek értékelésre, így 360 modellillesztés történt.

A hangolás mindkét év esetében azonos optimális paramétereket eredményezett (max_depth = 8, learning_rate = 0,1, n_estimators = 500, colsample_bytree = 0,5, subsample = 1,0), ami arra utal, hogy a két korpusz szerkezete hasonló komplexitású döntési határokat igényelt.

Az interpretálhatóság érdekében SHAP-értékeket számítottam a teszthalmaz egy 300 elemű véletlenszerű részhalmazán. A mintavétel célja a számítási költségek csökkentése volt, miközben a magyarázatok reprezentativitása megmaradt.

### A modellek teljesítményének összehasonlítása

A 2. táblázat a két korpuszon kapott hangolt modellek teszthalmaz-szintű teljesítménymutatóit foglalja össze egymás mellett.

*2. táblázat*
*A modellek teljesítményének összehasonlítása (2019 és 2021)*

| Metrika | 2019 tanítás | 2019 teszt | 2021 tanítás | 2021 teszt |
|---|---|---|---|---|
| **Pontosság** | 98,31% | 86,50% | 99,11% | 88,76% |
| **AUC-ROC** | 0,9986 | 0,9372 | 0,9997 | 0,9555 |
| **CV pontosság (átlag ± std)** | – | 85,29% ± 0,32% | – | 88,01% ± 0,33% |

A 2019-es cikkeken tanított és tesztelt modell 86,50%-os pontosságot és 0,9372 AUC-ROC értéket ér el. A 2021-es korpuszon tanított XGBoost modell ennél magasabb teljesítményt mutat: a teszthalmaz pontossága 88,76%, az AUC-ROC értéke pedig 0,9555. A tanító és teszt eredmények közötti különbség közepes mértékű túlillesztésre utalhat, ami döntési fa alapú boosting modellek esetében gyakori jelenség. Ugyanakkor a teszthalmazon elért magas pontosság és AUC-értékek azt mutatják, hogy a modell általánosíthatósága továbbra is jó, és ezt a CV alacsony szórása is alátámasztja.

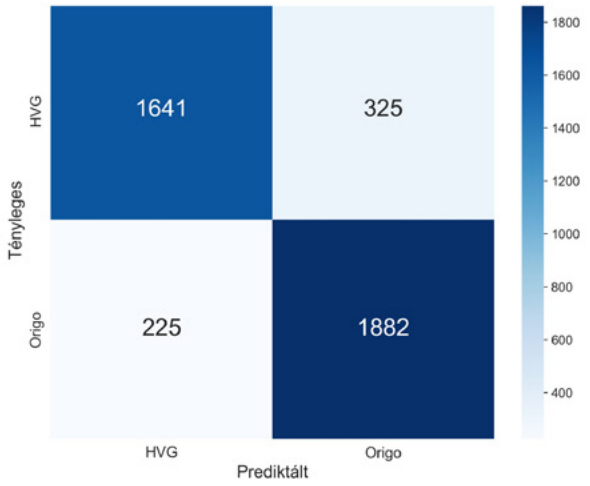
Mindkét esetben az Origo-osztály visszahívása magasabb, mint a HVG-osztályé, ami arra utal, hogy az Origo-szövegek lexikailag jellegzetesebb mintázatot mutatnak. Azt is megfigyelhetjük, hogy a visszahívás értéke nőtt, ami arra utalhat, hogy a két hírportál nyelvezete közötti különbség is nőtt.

## Konfúziós mátrixok

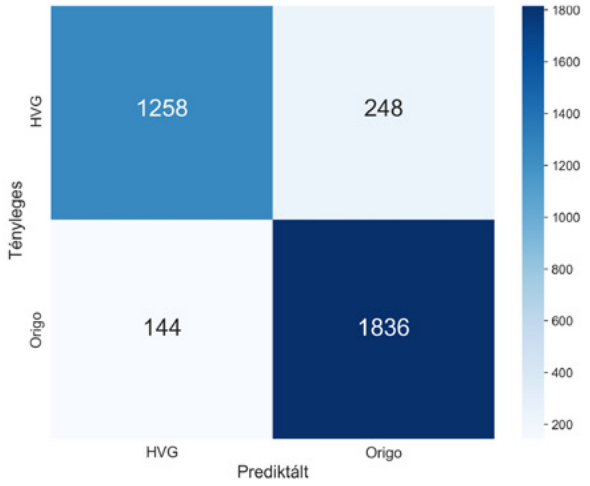
A konfúziós mátrixok alapján a 2019-es modellben a HVG-cikkek recall értéke (83,5%) elmarad az Origo cikkekétől (89,3%), ami arra utal, hogy a modell a HVG-cikkeket nagyobb arányban sorolta tévesen az Origo-kategóriába. Ez a logisztikus regressziónál megfigyelt mintázattal egybecseng, ugyanakkor az XGBoost nemlineáris feature-interakciókat is képes megragadni, ami mindkét osztályban magasabb recall értéket eredményez.

A 2021-es modellben ez az aszimmetria tovább erősödött. A HVG recall 83,5%-on stagnált, miközben az Origo recall 92,7%-ra emelkedett. Azoknak a besorolásoknak a száma, amikor az Origo-t tévesen HVG-ként sorolta be, 225-ről 144-re csökkent, ugyanígy a HVG téves besorolásainak száma 325-ről 248-ra esett vissza. A logisztikus regresszióhoz hasonlóan tehát az XGBoost eredményei is arra engednek utalni, hogy 2021-re elsősorban az Origo lexikális profilja vált elkülöníthetőbbé a TF-IDF-térben.

*3a. ábra*
*Konfúziós mátrix – 2019*



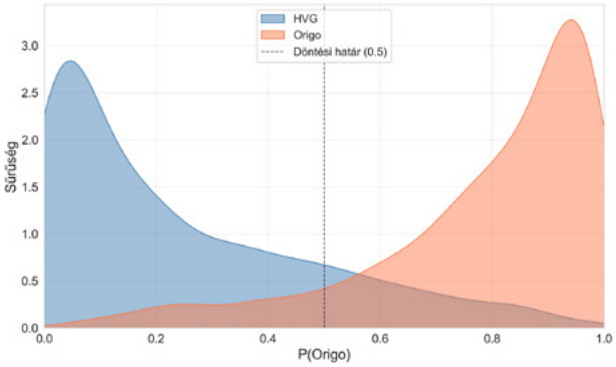

*3b. ábra*
*Konfúziós mátrix – 2021*



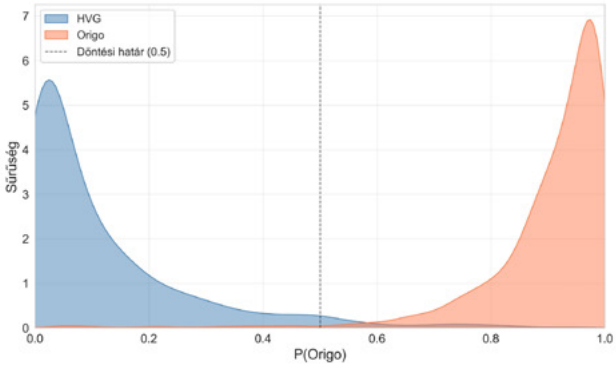

## A becsült valószínűségek eloszlása

A becsült P(Origo) valószínűségek eloszlása jól elkülöníti a két osztályt. A HVG-cikkek többsége alacsony valószínűségtartományban jelenik meg, míg az Origo-cikkek nagy része magas értékeknél koncentrálódik. A 2021-es modell esetében az eloszlások közötti különbség élesebb, ami összhangban áll a magasabb AUC-értékkel és a modell jobb diszkriminatív teljesítményével. Emellett megfigyelhető, hogy a középső, 0,4-0,6 közötti valószínűségtartományban kevesebb predikció jelenik meg, ami arra utal, hogy a 2021-es modell ritkábban ad bizonytalan besorolást.

*4a. ábra*
*Valószínűség-eloszlás – 2019*




*4b. ábra*
*Valószínűség-eloszlás – 2021*




### Jellemzők fontossága – SHAP-elemzés

A modell interpretálhatóságának érdekében mindkét korpuszon SHAP-elemzést végeztünk. A negatív SHAP-értékek a HVG-osztály irányába, míg a pozitív értékek az Origo-osztály irányába tolják a predikciót. A logisztikus regressziós modell jellemzőfontossági értékeivel összehasonlítva a SHAP-elemzés részletesebb képet ad a jellemzők hatásáról, mivel nemcsak az irányt, hanem az egyes predikciók esetében jelentkező hatás erősségét és variabilitását is megmutatja (Lundberg 2017). Ez különösen fontos nemlineáris modellek, például az XGBoost esetében.

A SHAP-elemzés ábrája alapján a HVG-osztályhoz tartozó jellemzők között több stabil, mindkét évben megjelenő kifejezés azonosítható (lásd a cikkben a 3. ábrát). Kiemelkedő szerepet játszanak olyan bigramok, mint a „további részlet”, az „azonnali vélemény” vagy az „este fok”, amelyek elsősorban szerkesztési és formátumbeli sajátosságokra utalnak. Ezek jelenléte arra enged következtetni, hogy a HVG esetében a stilisztikai és strukturális mintázatok időben viszonylag állandóak.

A 2019-es és 2021-es eredmények között ugyanakkor bizonyos eltérések is megfigyelhetőek (lásd a cikkben a 4. ábrát). A 2021-es modellben például az „elhuny” kifejezés kiemelkedő jelentőségűvé válik, ami összefügghet a pandémia időszaka alatti tematikus sajátosságával.

Az Origo-osztályhoz tartozó jellemzők esetében nagyobb eltérések figyelhetőek meg a két vizsgált év között (lásd a cikkben az 5. és 6. ábrát). A 2019-es modellben olyan kifejezések dominálnak, mint a „mentős”, „megyeszékhely” vagy „gazdálkodás”, amelyek inkább konkrét eseményekhez és helyszínekhez kötődő tematikus tartalmakra utalnak.

Ezzel szemben a 2021-es eredményekben más típusú jellemzők kerülnek előtérbe, például a „forint találatos” vagy az „eljuttatott”, amelyek eltérő tematikus hangsúlyokra utalhatnak. Ez a különbség arra enged következtetni, hogy az Origo esetében a tartalmi fókusz időben változékonyabb lehet, vagy a vizsgált korpuszok összetétele különbözik.

## 6. BiLSTM

### Tanítási konfiguráció és folyamat

A BiLSTM modell esetében fix cikk-hosszt kell alkalmazni a cikkekre a hatékony tanítás érdekében: a maximális hossz a cikkek felső 95. percentilise alapján került meghatározásra, az ennél hosszabb szövegeket csonkoltuk, az annál rövidebbeket pedig jelentés nélküli szöveggel mesterségesen töltöttük fel. A BiLSTM modell mögött álló, az egyes kifejezésekhez tartozó szóbeágyazási reprezentáció nem az újságcikkekből származik, hanem egy általános, nagy magyar szövegkorpuszon hozták létre a kifejezéseket reprezentáló fastText-vektorokat (Grave et al. 2018).

A BiLSTM-modell esetében a hiperparaméter-keresés 12 kombinációt fedett le: lstm_units ∈ {32, 64, 128}, dropout ∈ {0,25; 0,50}, learning_rate ∈ {0,001; $5\times10^{-5}$}. Az optimalizálás célfüggvénye a validációs veszteség minimalizálása volt.

A hangolás eredményei alapján mindkét évben az lr = 0,001 és a dropout = 0,25 bizonyult optimálisnak. Az LSTM egységek számában viszont eltérés mutatkozott. A 2019-es modell esetében 128, míg a 2021-esnél 64 egység adta a legjobb teljesítményt. A dropout = 0,50 minden esetben teljesítményromláshoz vezetett, ami túlzott regularizációra utal. Az alacsony tanulási ráta ($5\times10^{-5}$) mellett a modellek nem konvergáltak megfelelően a 20 epochos tanítási korlát alatt.

A tanítási görbék elemzése során a 2019-es modellnél enyhe túlillesztés volt megfigyelhető. A tanítási veszteség az epochok végéig csökkent (~0,37), míg a validációs veszteség már a 13. epoch körül tetőzött (~0,43). Ennek kezelésére két regularizációs beavatkozást is kipróbáltunk (dropout növelése 0,40-re, illetve L2 regularizáció $\lambda = 10^{-4}$ értékkel), azonban ezek alulillesztéshez vezettek, ezért a végleges modell az eredeti konfigurációt tartotta meg.

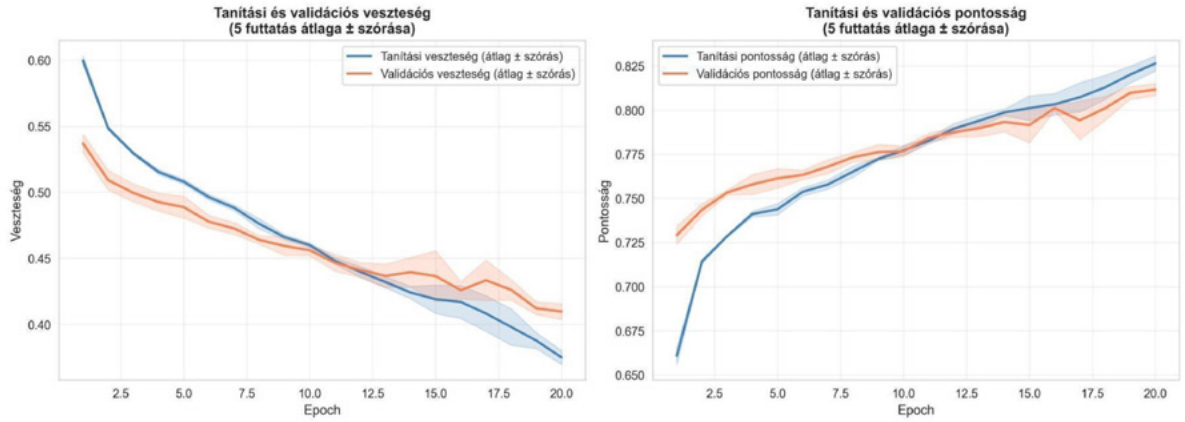
Az EarlyStopping paramétereit patience = 3 és min_delta = 0,001 értékekre állítottuk, hogy kiszűrjük a validációs veszteség kisebb oszcillációit. A végleges modellek architektúrája egy előtanított fastText beágyazási rétegből (cc.hu.300, Grave et al. 2018), egy bidirekciólis LSTM rétegből, dropout regularizációból és egy szigmoid aktivációs Dense kimeneti rétegből áll.

Az 5. és 6. ábrák a tanítási és validációs veszteség, illetve pontosság alakulását mutatják az 5 futtatás átlagának és szórásának függvényében. A 2019-es modell esetében a tanítási és validációs görbék viszonylag szűk sávban haladnak egymás közelében, ami mérsékelt túltanulásra (overfitting) utal. A 20 epochos korlát hatékonyan megakadályozta a jelentős divergenciát.

A 2021-es modell hosszabb tanítási idővel (31 epoch a legjobb futtatásnál) nagyobb különbséget mutat a tanítási és validációs görbék között, ami a modell nagyobb kapacitás-kihasználására utal, ugyanakkor a validációs veszteség továbbra is csökkenő tendenciát mutat, így a túltanulás mértéke elfogadható tartományban maradt.

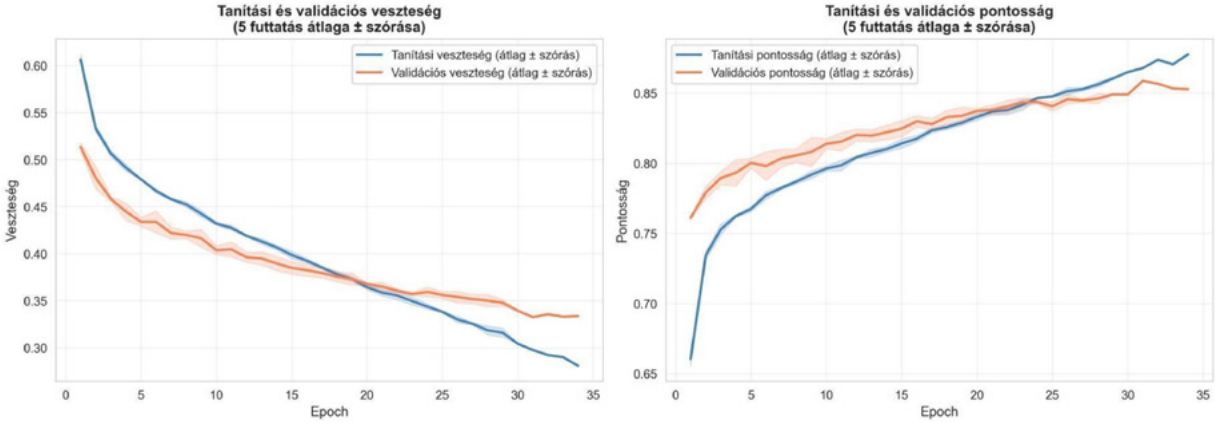
*5. ábra*

*Tanulási görbe – BiLSTM 2019 (5 futtatás átl. ± szór.)*




*6. ábra*

*Tanulási görbe – BiLSTM 2021 (5 futtatás átl. ± szór.)*



## Klasszifikációs teljesítmény

A 3. táblázat összefoglalja a BiLSTM modell teljesítménymutatóit mindkét korpuszon.

*3. táblázat*
*A BiLSTM modell teljesítménymutatóinak összehasonlítása a 2019-es és 2021-es korpuszon*

| Metrika | 2019-es korpusz | 2021-es korpusz |
|---|---|---|
| **LSTM egységek** | 128 | 64 |
| **Legjobb epoch** | 20 | 31 |
| **Legjobb validációs veszteség** | 0,3993 | 0,3325 |
| **Tanítási pontosság (best run)** | 0,8696 (86,96%) | 0,9166 (91,66%) |
| **Tanítási AUC (best run)** | 0,9487 | 0,9743 |
| **Teszt pontosság (best run)** | 0,8164 (81,64%) | 0,8587 (85,87%) |
| **Teszt AUC (best run)** | 0,9008 | 0,9308 |
| **Teszt pontosság (átlag ± szórás)** | 0,8087 ± 0,0086 | 0,8455 ± 0,0096 |
| **Teszt AUC (átlag ± szórás)** | 0,8954 ± 0,0061 | 0,9197 ± 0,0075 |

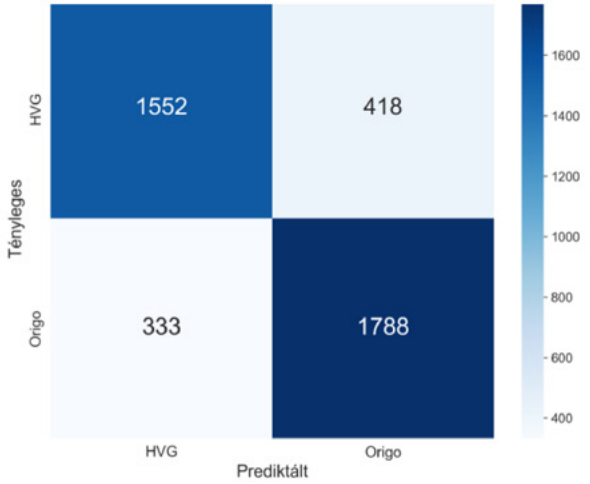
A 2021-es korpuszon tanított modell felülmúlta minden vizsgált mutatóban a 2019-es korpuszon tanított modellt. A 2019-es modell teszt pontossága 81,64%, az AUC értéke 0,9008. Ezek az értékek a 2021-es modell esetében 85,87% és 0,9308 a legjobb futtatás alapján. Az öt futtatás átlagát tekintve a pontosság 80,87% és 84,55%, az AUC 0,8954 és 0,9197. A szórás mindkét évben alacsony maradt ($\sigma_{acc} < 0{,}01$, $\sigma_{AUC} < 0{,}008$), ami az eredmények reprodukálhatóságát jelzi.

A 2021-es korpusz magasabb teljesítménye összhangban van a logisztikus regressziós és XGBoost modellek hasonló eredményeivel. Ez valószínűsíthetően a két portál közötti stilisztikai divergencia erősödésére vezethető vissza.

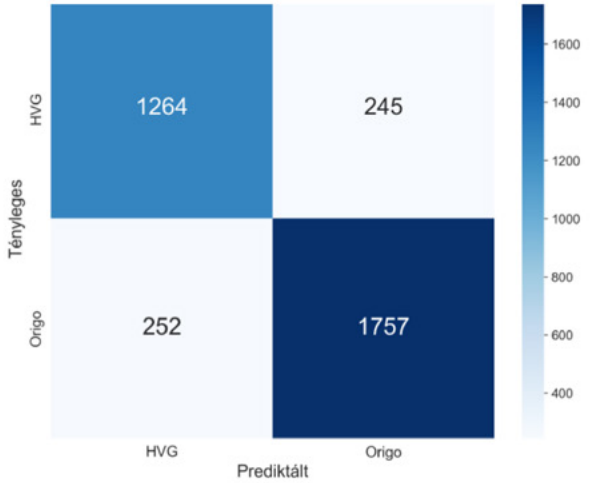
## Konfúziós mátrixok elemzése

A 7. ábrapáron a két modell konfúziós mátrixait láthatjuk. A 2019-es modell esetében a HVG-cikkek recallja alacsonyabb (78,8%), mint az Origo-cikkeké (84,3%), ami arra utal, hogy a modell hajlamosabb volt a HVG-cikkeket tévesen Origo-ként azonosítani. Ez abból következhet, hogy a HVG stilisztikai mintázatai kevesebb modell által megragadható szekvenciális mintázatot tartalmaznak.

*7a. ábra*
*Konfúziós mátrix – BiLSTM (2019)*




*7b. ábra*
*Konfúziós mátrix – BiLSTM (2021)*



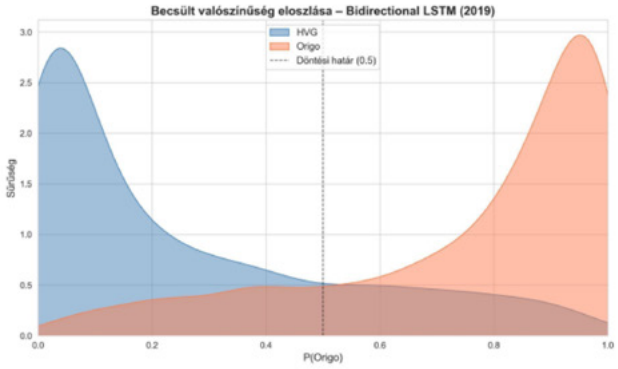
A 2021-es modellben esetében a HVG recall 83,8%-ra, az Origo recall 87,5%-ra emelkedett. Az Origo irányába mutató téves besorolások száma 418-ról 245-re csökkent, míg az ellentétes irányú tévesztések száma 333-ról 252-re mérséklődött. Ez úgyszintén azt sugallja, hogy a két portál szekvenciális nyelvi jellemzői 2021-re határozottabbá váltak.

## Becsült valószínűségi eloszlások

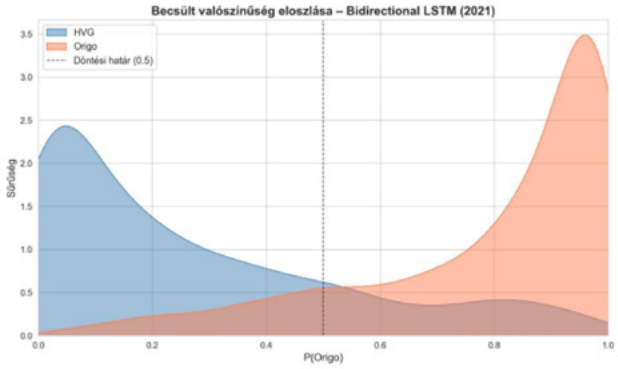
A 8. ábrapár a modellek által becsült P(Origo) valószínűségek sűrűségfüggvényeit mutatja. Mindkét ábra bimodális eloszlást mutat. A HVG-cikkek többsége alacsony P(Origo) értékekre koncentrálódik (csúcs ~0,05–0,10), míg az Origo-cikkek magas P(Origo) értékeknél sűrűsödnek (csúcs ~0,85–0,95).

A két eloszlás átfedési területe kisebb a 2021-es modell ábráján, ez a kisebb téves klasszifikációs rátával összhangban van. A 2021-es modellben a HVG sűrűségcsúcsa magasabbra nő (3,5) és a két eloszlás élesebben elválik. Ez a kalibrációs javulás összhangban van az AUC-növekedéssel, és arra utal, hogy a 2021-es modell nemcsak pontosabb döntéseket hoz, hanem konfidensebb is azokban.

*8a. ábra*
*P(Origo) eloszlás – BiLSTM (2019)*




*8b. ábra*
*P(Origo) eloszlás – BiLSTM (2021)*




## LIME-alapú interpretálhatósági elemzés

Ahogy az az elméleti szekcióban említettük, a neurális hálózatok, így az LSTM modell is, a működésükből adódóan korlátozottan értelmezhetők, ezért a predikciók mögötti döntési mechanizmus feltárásához kiegészítő módszerek alkalmazása szükséges. Ennek érdekében a modellek lokális interpretációjára a LIME módszert alkalmaztunk, amely lehetővé teszi, hogy egy-egy konkrét cikk esetében azonosítsuk azokat a szavakat, amelyek a modell döntéséhez leginkább hozzájárultak.

Fontos hangsúlyozni, hogy a LIME lokális módszer. Az eredmények egyedi megfigyelésekre vonatkoznak, és nem általánosíthatók közvetlenül a teljes adathalmazra. Ennek megfelelően az elemzés célja nem globális következtetések levonása, hanem a modell működésének szemléltetése és kvalitatív értelmezése.

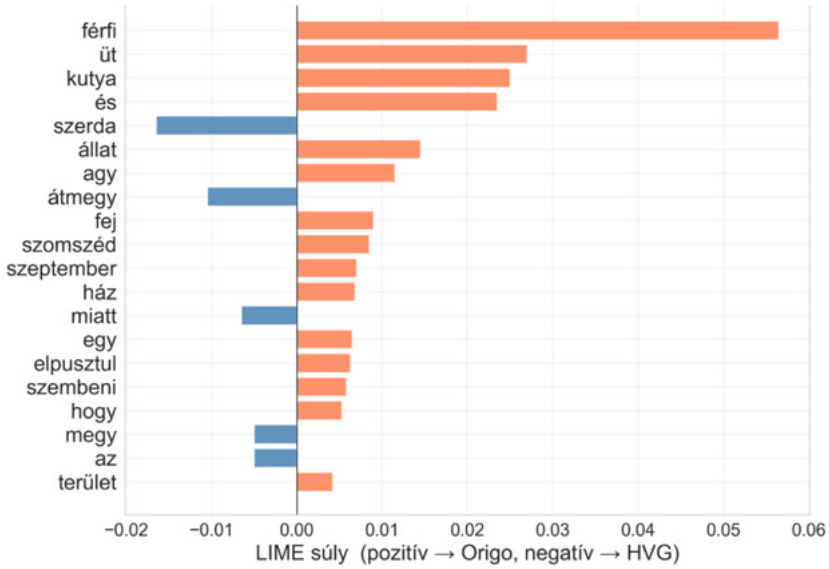
Három példát mutatunk be, amelyeket a modell predikciós bizonytalansága és a predikció helyessége alapján választottunk ki.

Az első példában (lásd a cikkben a 7. ábrát) a modell egy Origo cikket nagy magabiztossággal (P(Origo) = 1,000) helyesen sorolt az Origo kategóriába. A LIME elemzés alapján a döntést elsősorban olyan szavak befolyásolták pozitív irányban, mint az „akar”, „telefon”, „férfi”, „elad” és „július”.

Ezek a kifejezések jellemzően konkrét eseményekhez, cselekvésekhez és narratív elemekhez kapcsolódnak, ami arra utal, hogy a modell nem pusztán stilisztikai jegyeket, hanem tartalmi mintázatokat is felhasznál a klasszifikáció során. A negatív irányba ható szavak (pl. „ez”, „meg”, „ügyészség”) kisebb súllyal jelennek meg, ami tovább erősíti a döntés egyértelműségét.

Ez az eset jól szemlélteti, hogy a modell képes stabil és konzisztens mintázatok felismerésére, ugyanakkor azt is jelzi, hogy a döntés részben tematikus jegyeken alapul.

*9. ábra*
*LIME-elemzés – téves Origo besorolás (P(Origo) = 0,998)*

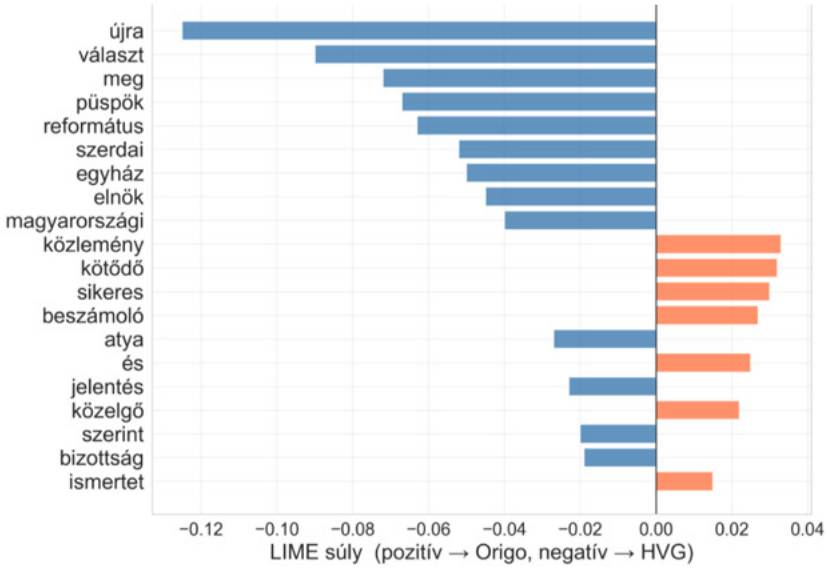



A második példában a modell egy HVG cikket tévesen, rendkívül nagy magabiztossággal (P(Origo) = 0,998) sorolt az Origo kategóriába. A LIME elemzés alapján a döntést erősen befolyásolták olyan szavak, mint a „férfi", „üt", „kutya" és „állat", amelyek gyakran bűnügyi vagy szenzációt keltő eseményekhez kapcsolódnak.

Ezzel szemben a HVG irányába ható szavak, „szerda", „az", „hogy", jóval kisebb súllyal szerepelnek, és nem képesek ellensúlyozni a domináns narratív elemek hatását.

Ez az eset arra utalhat, hogy a modell bizonyos témákhoz torzítottan az Origo kategóriát társítja. Ennek értelmében a klasszifikáció nem csak stilisztikai jegyeken alapul, hanem tematikus torzítás is megjelenik a működésében.

*10. ábra*
*LIME-elemzés – bizonytalan besorolás (P(Origo) = 0,501)*



A harmadik példában a modell egy Origo cikket közel 0,5 valószínűséggel sorolt az Origo kategóriába (P(Origo) = 0,501), ami magas fokú bizonytalanságot jelez. Az ábrán két, egymással ellentétes irányba ható szócsoport figyelhető meg.

A negatív, HVG irányú csoport olyan kifejezéseket tartalmaz, mint a „püspök", „egyház", „bizottság" és „szerint", amelyek inkább formális, intézményi és semleges hangvételre utalnak. Ezzel szemben a pozitív, Origo irányú szavak, mint a „közlemény", „sikeres", „beszámoló" , inkább narratív és értékelő jellegűek.

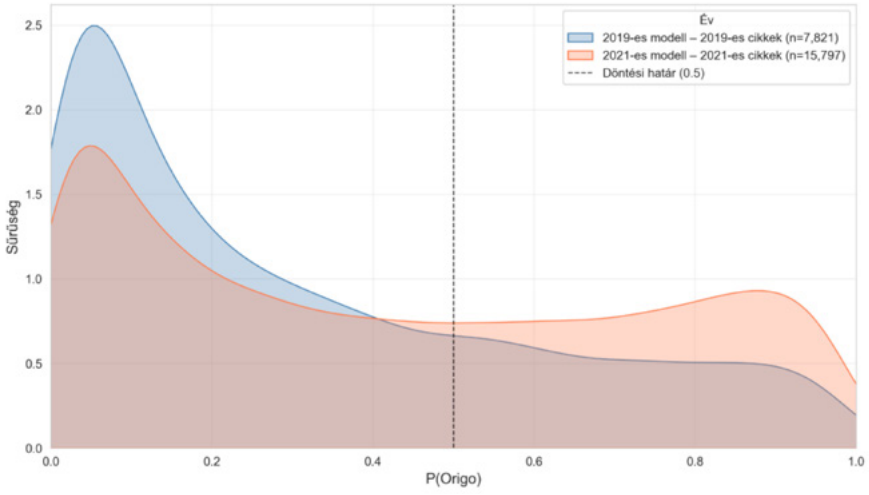
A kétféle jegy egyidejű jelenléte arra utal, hogy a cikk nyelvezete nem illeszkedik egyértelműen egyik portál stílusához sem. Ez az eset jól példázza, hogy a két hírforrás nyelvezete között bizonyos esetekben átfedések figyelhetőek meg.

## 7. Az Index cikkek besorolása

### Logisztikus regresszió – a predikciós eloszlások vizuális elemzése

A 11. ábra a becsült P(Origo) valószínűségek sűrűségeloszlását mutatja, a két modellre külön-külön.

*11. ábra*
*Index-cikkek P(Origo) eloszlása – log-reg modell, évenkénti bontásban*




Az ábrán azt figyelhetjük meg, hogy a 2019-es cikkek esetében eloszlása balra koncentráltabb és meredekebb, míg a 2021-es cikkeké laposabb és jobbra tolódik. Emellett a magas valószínűségi tartományban egy másodlagos kiemelkedés is megfigyelhető, ami arra utal, hogy a 2021-es Index-cikkek egy jól elkülöníthető része kifejezetten erős Origo-jellegű mintázatot mutat. Ezek alapján elmondható, hogy a 2021-es cikkek között nagyobb arányban találhatók olyan szövegek, amelyeket a modell közepes vagy magas valószínűséggel Origo-szerűnek minősít.

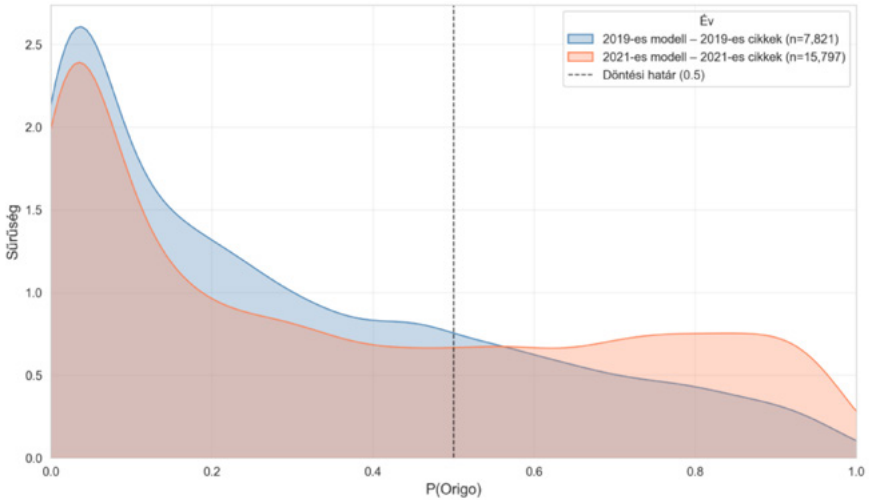
### XGBoost – a predikciós eloszlások vizuális elemzése

A 12. ábra a becsült P(Origo) valószínűségek sűrűségeloszlását mutatja évenkénti bontásban, a két modellre külön-külön.

A mind két év cikkei hasonló eredményt mutatnak a XGBoost esetében, mint a logisztikus regressziónál. A 2019-es cikkek eloszlása koncentráltabb, míg a 2021-es cikkek eloszlása jobbra tolódik, és a magasabb valószínűségi

tartományban egy enyhe kiemelkedés is megjelenik. Ebben az esetben is kijelenthetjük, hogy a 2021-es cikkek egy része a modell által azonosított Origo-jellegű mintázatokhoz közelebb áll.

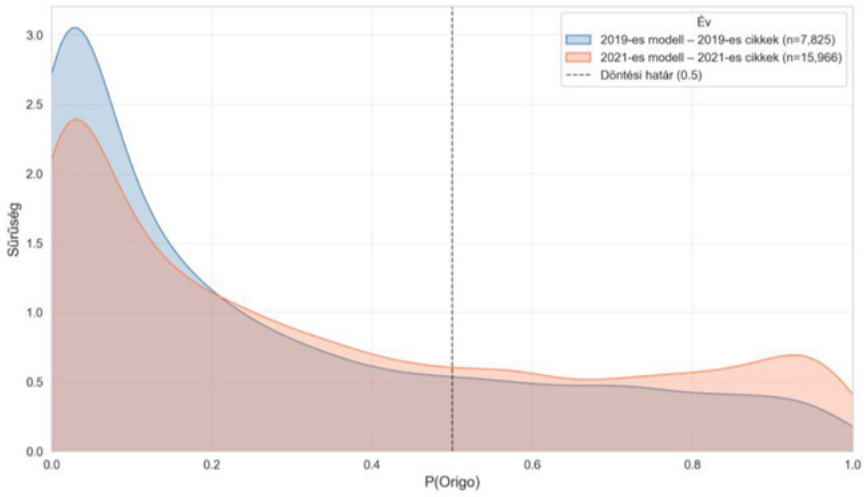
*12. ábra*
*Index-cikkek P(Origo) eloszlása – XGBoost modell, évenkénti bontásban*




## BiLSTM – a predikciós eloszlások vizuális elemzése

A 13. ábra a becsült P(Origo) valószínűségek sűrűségeloszlását mutatja évenkénti bontásban, a két BiLSTM modellre külön-külön.

*13. ábra*
*Index-cikkek P(Origo) eloszlása – BiLSTM modell, évenkénti bontásban*




A BiLSTM modellek által az Index Belföld rovat cikkeire adott P(Origo) értékek balra ferde eloszlást mutatnak. Mindkét modellváltozat esetében a módusz a 0–0,1 tartományba esik. Ez arra utal, hogy a cikkek többségét a modell inkább a HVG szerű kategóriába sorolja.

Az ábrán továbbá megfigyelhető, hogy a 2021-es cikkek görbéje a 0,2 feletti tartományban tartósan a 2019-es fölött helyezkedik el, azaz a valószínűségi tömeg eltolódása a BiLSTM modelnél is megjelenik, szintúgy a P(Origo) > 0,8 tartományban egy enyhe kiemelkedés.

# 8. Robusztusság-ellenőrzés

## 8.1. A szűrési eljárás

A névelem-felismerés a spaCy hu_core_news_lg modelljével történt, közös előfeldolgozási lépésben. A felismert tulajdonnevek típus szerinti helyettesítő tokent kaptak: „neszemely", „nehely", „neszervezet" és „neegyeb" A helyettesítés a token törlése helyett történik, így a mondathossz és a szerkezet nem torzul.

Példa egy cikk részleten:

> „Pénteken hajnali háromnegyed egykor jelentette egy polgárőr, hogy füstöl egy elhagyatott raktárépület Sopronban, a Győri úton – írja az Országos Katasztrófavédelmi Főigazgatóság (OKF) pénteki közleményében."

TF-IDF és NER-rel feldolgozott kimenet:

> péntek hajnali háromnegyed egykor jelent nehely nehely neszervezet neszervezet pénteki közlemény

A hiperparaméter-keresés terét és a tanítási beállításokat nem módosítottuk, hogy a különbség kizárólag a bemenetre legyen visszavezethető. A logisztikus regresszió és az XGBoost ugyanazt az optimumot választotta, mint a szűretlen korpuszon (logisztikus regresszió: C = 5, max_features = 7000; XGBoost: max_depth = 8, learning_rate = 0,1, n_estimators = 500, colsample_bytree = 0,5, subsample = 1,0). A BiLSTM rácskeresésében mindkét évben az lr = 0,001 és a dropout = 0,25 kombináció bizonyult optimálisnak, az LSTM egységek száma azonban mindkét évben 128 lett, szemben a kutatás 2021-es modelljének 64 egységével.

## 8.2. A modellek teljesítménye

A 4. táblázat a HVG–Origo besorolás teljesítményét foglalja össze a megszűrt korpuszon.

*4. táblázat*
*A tulajdonnév-szűrt modellek pontossága és AUC-ROC értékei*

| Metrika | LogReg 2019 | LogReg 2021 | XGBoost 2019 | XGBoost 2021 | BiLSTM 2019 | BiLSTM 2021 |
|---|---|---|---|---|---|---|
| **Teszt pontosság** | 83,94% | 86,84% | 87,83% | 89,46% | 82,05% | 82,65% |
| **Tanítási pontosság** | 93,07% | 94,92% | 98,53% | 99,14% | 86,92% | 87,20% |
| **Teszt AUC-ROC** | 0,9207 | 0,9390 | 0,9409 | 0,9581 | 0,8996 | 0,9092 |
| **Tanítási AUC-ROC** | 0,9801 | 0,9890 | 0,9984 | 0,9996 | 0,9451 | 0,9505 |

## 8.3. Az Index-cikkek osztályozásának eredményei

*5. táblázat*
*Az Index-cikkek osztályozási eredményei a tulajdonnevektől megszűrt korpuszon*

| Metrika | 2019-es LogReg 2019-es cikkeken | 2021-es LogReg 2021-es cikkeken | 2019-es XGBoost 2019-es cikkeken | 2021-es XGBoost 2021-es cikkeken | 2019-es BiLSTM 2019-es cikkeken | 2021-es BiLSTM 2021-es cikkeken |
|---|---|---|---|---|---|---|
| **Origo-szerű besorolás** | 16,9% | 33,6% | 13,4% | 29,7% | 19,0% | 28,9% |
| **P(Origo) átlag** | 0,2438 | 0,3699 | 0,2287 | 0,3279 | 0,2358 | 0,3463 |
| **P(Origo) medián** | 0,1496 | 0,3034 | 0,1552 | 0,2371 | 0,1193 | 0,2856 |

## 8.4. Hipotézisvizsgálat

*6. táblázat*
*A Mann–Whitney U-teszt és a Cohen-féle d eredményei a tulajdonnév-szűrt korpuszon*

| Modell | Tan. év | Átl. P(O) 2019 | Átl. P(O) 2021 | Δ | p | d |
|---|---|---|---|---|---|---|
| **LogReg** | 2019 | 0,2438 | 0,3356 | +0,0918 | 3,85e-173 | −0,36 |
| **LogReg** | 2021 | 0,2951 | 0,3699 | +0,0748 | 3,07e-75 | −0,26 |
| **XGBoost** | 2019 | 0,2287 | 0,3063 | +0,0776 | 1,88e-158 | −0,33 |
| **XGBoost** | 2021 | 0,2777 | 0,3279 | +0,0502 | 4,79e-39 | −0,18 |
| **BiLSTM** | 2019 | 0,2358 | 0,3091 | +0,0733 | 5,74e-114 | −0,27 |
| **BiLSTM** | 2021 | 0,2939 | 0,3463 | +0,0524 | 8,16e-51 | −0,19 |

## 8.5. Összevetés

*7. táblázat*
*A modellek teljesítménye az eredeti és tulajdonnevektől szűrt korpuszon*

| Metrika | LogReg 2019 | LogReg 2021 | XGBoost 2019 | XGBoost 2021 | BiLSTM 2019 | BiLSTM 2021 |
|---|---|---|---|---|---|---|
| **Eredeti teszt pontosság** | 83,84% | 87,26% | 86,50% | 88,76% | 81,64% | 85,87% |
| **Szűrt teszt pontosság** | 83,94% | 86,84% | 87,83% | 89,46% | 82,05% | 82,65% |
| **Eredeti AUC** | 0,9207 | 0,9453 | 0,9372 | 0,9555 | 0,9008 | 0,9308 |
| **Szűrt AUC** | 0,9207 | 0,9390 | 0,9409 | 0,9581 | 0,8996 | 0,9092 |
| **AUC-különbség** | −0,0000 | −0,0063 | +0,0037 | +0,0026 | −0,0012 | −0,0216 |

## Felhasznált irodalom

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Szoftver]. https://www.tensorflow.org/

Angyalffy, D. (2026). *Az Index.hu nyelvezetének vizsgálata a 2020-as tulajdonosváltást követően gépi tanulási módszerekkel* [Forráskód]. GitHub. https://github.com/davidangyalffy/kutatas_2026

Gers, F. A., Schmidhuber, J. & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation,* 12(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems,* 28(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Krevelen, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature,* 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen* [Diplomamunka, Technische Universität München].

Honnibal, M. & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing [Szoftver]. https://spacy.io/

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering,* 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R. Springer.* https://doi.org/10.1007/978-1-4614-7138-7

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall.

Jurafsky, D., & Martin, J. H. (2026). Speech and Language Processing (3. kiadás). Online manuscript. https://web.stanford.edu/~jurafsky/slp3/

Michel, P. & Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (543–553. o.). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1050

Orebi, T. & Naser, S. A. (2025). The impact of stop word removal on text classification: A comprehensive review. *International Journal of Academic Multidisciplinary Research,* 9(1), 128–136.

Pascanu, R., Mikolov, T. & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)* (1310–1318. o.). PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research,* 12, 2825–2830.

Petridis, C. (2024). Text classification: Neural networks vs machine learning models vs pretrained models. arXiv preprint, arXiv:2412.21022. https://doi.org/10.48550/arXiv.2412.21022.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Thawani, A., Pujara, J., Szekely, P. A. & Ilievski, F. (2021). Representing numbers in NLP: A survey and a vision. In *Findings of the Association for Computational Linguistics: NAACL 2021* (644–656. o.). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-naacl.53

Tikk, D. (Szerk.). (2007). *Szövegbányászat.* Typotex Kiadó.

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software,* 6(60), 3021. https://doi.org/10.21105/joss.03021

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE,* 78(10), 1550–1560. https://doi.org/10.1109/5.58337

Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)* (412–420. o.). Morgan Kaufmann.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H. & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (207–212. o.). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2034