Gazdasági elemzés strukturálatlan adatokon: természetesnyelv-feldolgozási esetek*

Temesvári Csanád – Horváth Beáta – Ónozó Lívia Réka

A gazdasági szövegek, mint az újságcikkek vagy a kiskereskedelmi terméknevek, alternatív, részletes és nagy gyakoriságú információforrást jelentenek, amelyek betekintést nyújthatnak a gazdasági trendekbe, és pontosabb, időszerűbb becsléseket tesznek lehetővé. Több mélytanulási modellt alakítottunk ki két különböző kutatási feladatra: 1) szentimentindex létrehozása, amely a pénzügyi és gazdasági cikkek három szentimentkategóriába sorolásából ered; és 2) a kiskereskedelmi terméknevek megfelelő vámtarifaszám-kategóriákba sorolása. Modelljeink egyrészt következetesen felülmúlták a kiskereskedelmi termékek klasszifikációjára szolgáló alapmodelleket, másrészt szentimentindexünk pontosan előre tudta jelezni a gazdasági visszaeséseket olyan esetekben, amikor nem álltak rendelkezésre magas gyakoriságú adatok.

Journal of Economic Literature (JEL) kódok: C43, C45, C60

Kulcsszavak: természetesnyelv-feldolgozás, mélytanulás, makrogazdasági nowcasting, klasszifikáció

1. Bevezetés

A természetesnyelv-feldolgozás (NLP[1]) gyors fejlődése és a nagy méretű szöveges adatok elérhetősége hatékony eszközt biztosít a közgazdászok számára az információk elemzéséhez. A közgazdaságtanban a „szöveg mint adat” paradigma az egyik legdinamikusabb módszertani terület lett, amely lehetővé teszi a kutatók számára, hogy a szentimentet és a narratívákat közvetlenül strukturálatlan forrásokból, például hírekből, vállalati jelentésekből és az online médiából nyerjék ki. A közelmúltban készült, átfogó elemzések, például *Ash – Hansen* (*2023*) és *Gentzkow és szerzőtársai* (*2019*) dokumentálták ezt az átalakulást, kiemelve a szövegalapú mérések központi

* A jelen kiadványban megjelenő írások a szerzők nézeteit tartalmazzák, ami nem feltétlenül egyezik a Magyar Nemzeti Bank hivatalos álláspontjával.

Temesvári Csanád: Magyar Nemzeti Bank, elemző. E-mail: temesvarics@mnb.hu
Horváth Beáta: Magyar Nemzeti Bank, vezető közgazdász elemző. E-mail: horvathbea@mnb.hu
Ónozó Lívia Réka: Magyar Nemzeti Bank, felügyeleti tanácsadó; Budapesti Műszaki és Gazdaságtudományi Egyetem, PhD-hallgató. E-mail: onozol@mnb.hu

Az angol nyelvű kézirat első változata 2025. március 14-én érkezett szerkesztőségünkbe.

DOI: https://doi.org/10.25201/HSZ.25.1.27

[1] Natural Language Processing

szerepét a modern empirikus közgazdaságtanban. Általánosságban elmondható, hogy a gépi tanulás, az NLP és a gazdasági előrejelzés metszéspontját ma már kulcsfontosságú kutatási területként ismerik el, amely kapcsolatban áll a nowcasting[2] és a valós idejű makrogazdasági monitorozás alapvető munkáival (*Babii et al. 2021*).

Az NLP-ben párhuzamosan végbement fejlődést a mély neurális hálózatok, legfőképpen pedig a transzformer architektúrák (*Vaswani et al. 2017*) ösztönözték. Az olyan modellek, mint a Bidirectional Encoder Representations from Transformers (BERT, *Devlin et al. 2019*) új teljesítménymércét állítottak fel a különböző szövegklasszifikációs és szentimentelemzési feladatok terén, és a Nagy Nyelvi Modell korszakában is versenyképes modelleknek számítanak (*Rostam – Kertész 2025*). Ezek az architektúrák lehetővé teszik a kontextus alapú megértést, amely eltér a hagyományos, szótár- vagy szózsák (bag-of-words) alapú módszerektől, amelyek nehezen kezelik a többjelentésű vagy tagadószavakat. A közgazdaságtanban a kutatók elkezdtek a hagyományos, kézzel készített szótárak helyett (*Tetlock 2007; Loughran – McDonald 2011*) a szavak kontextusát figyelembe vevő modellekre támaszkodni, ami a gazdasági narratívák gazdagabb, pontosabb ábrázolását eredményezte (*Nasiopoulos et al. 2025*). A transzformer alapú modellek fő alkalmazási területe a transzfer tanulás, ahol a modellt először milliárd vagy akár billió szóra vagy „tokenekre" tanítják részben felügyelt környezetben, azzal a céllal, hogy egy bemeneti tokensorozat alapján előre jelezzék a következő tokent. Ez arra ösztönzi a modellt, hogy elsajátítsa az emberi nyelv mintázatait, és ezáltal alapvető tudásbázist építsen ki. Ezt követően a modellt feladatspecifikus korpuszon, például gazdasági vagy pénzügyi szövegeken keresztül tanítják tovább annak érdekében, hogy egy adott szakterületen kiemelkedő előrejelzési teljesítményt érjen el; ezt a folyamatot „finomhangolásnak" nevezzük. Ez lehetővé teszi a végső modell számára, hogy egyszerre hasznosítsa az emberi nyelvről szerzett általános tudást, valamint hatékonyan működjön egy adott területen, például a pénzügyek (*FinBERT, Huang et al. 2022*) vagy a tudományos szövegek (S*ciBERT, Beltagy et al. 2019*) világában.

A gazdasági hírek szentimentelemzése késleltetés nélkül nyújt betekintést a közvélemény hangulatába és a piaci trendekbe (*Ónozó et al. 2024b:1*). A szöveges forrásokból származó, magas gyakoriságú mutatók generálásának képessége a gazdaságpolitikai döntéshozók és a közgazdászok számára közel valós idejű jelzéseket biztosít, megelőzve a késleltetett hivatalos statisztikákat. Uniós szinten *De Bondt – Sun* (*2025*) a ChatGPT segítségével klasszifikációs rendszert hozott létre, amely a havi globális PMI[3]-jelentésekhez héja vagy galamb szentimentet rendel. A szerzők ezeket az értékeket sikeresen alkalmazták regressziós környezetben, javítva az

[2] A nowcasting a közgazdaságtanban a nagyon közeli múlt, a jelen, valamint egy gazdasági mutató legközelebbi jövőbeli állapotának az előrejelzése.
[3] Purchasing Manager's Index

euroövezeti GDP nowcast-becsléseinek[4] pontosságát. Országos szinten *Kalamara és szerzőtársai* (*2022*) három neves brit napilap cikkei alapján hoztak létre egy szentimentindexet, amelyben mind az előfordulások számát, mind a felügyelt gépi tanulási módszereket alkalmazták. Indexük más mutatókkal kombinálva figyelemre méltó előrejelző képességgel bírt a brit közgazdászok és politikai döntéshozók által széles körben használt „proxyváltozók" tekintetében. Mind *Aguilar et al.* (*2021*), mind *Sobrino et al.* (*2020*) létrehozott egy szentimentindexet a spanyol gazdaságra vonatkozóan, kulcsszókeresés segítségével, rendre hét nagy hírportál és a Banco de España (Bank of Spain) negyedéves jelentései alapján. Mindkét index jobban teljesített a nemzeti GDP és a GDP-növekedés nowcasting előrejelzésében, mint a kérdőíves felméréseken alapuló proxyváltozók.

Mivel csak korlátozott mennyiségű, könnyen elérhető, gazdaságilag releváns szentimentadat áll rendelkezésre, a finomhangoláshoz megfelelő méretű adatbázis létrehozásához manuális annotációra van szükség. Ez az annotáció azonban jelentős manuális munkaerőt és együttműködést igényel. Ennek ellensúlyozására jelent meg az aktív tanulás (AL[5]), amely lehetővé teszi kis mennyiségű, manuálisan annotált adat és nagy mennyiségű, címkézetlen példa kombinált felhasználását. Az aktív tanulás szelektív módon azonosítja a címkézéshez leginkább informatívnak tekinthető példákat, ezáltal jelentősen csökkentve az annotációs terhet, miközben fenntartja (vagy akár növeli) a modell teljesítményét. Számos stratégia létezik a leginformatívabb adatpontok megtalálásához: például olyan példák kiválasztása, amelyek szemantikailag hasonlóak a korábban tévesen klasszifikált mondatokhoz (*Jiang et al. 2012*), vagy olyanok választása, amelyeknél a modell előrejelzése „bizonytalanabb" (*Schröder et al. 2022*), míg egyes megközelítések ezek kombinációját alkalmazzák (*Chen et al. 2011*). Bár az annotátorok még mindig többnyire emberek, egyre gyakoribb az LLM[6]-ek alkalmazása mind annotátorként, mind pedig a címkézetlen adatok közötti kiválasztást segítő módszerként – például kevésbé ígéretes adatpontok kiszűrésére vagy a címkézésre javasolt adatok rangsorolásához. Az LLM-eket új, címkézett adatbázisok létrehozására is használják (*Xia et al. 2025*). Ezeket a heurisztikákat több különböző NLP-feladat esetében is hatékonyan alkalmazták, bizonyítva hasznosságukat a gépi tanulási modellek tanításához szükséges adatok hatékony generálását illetően (*Settles 2011; Zhang et al. 2022*). Az aktív tanulás magyar nyelvű NLP-alkalmazásának egyik kiemelkedő példája *Üveges és szerzőtársai* (*2024*) írása, akik jogi dokumentumokat soroltak be jogi kategóriákba mélytanulási modellek segítségével. Az aktív tanulás alkalmazásával az alapmodell pontosságának eléréséhez szükséges adatmennyiséget akár 60 százalékkal sikerült csökkenteni.

[4] A nowcast-becslés (vagy nowcasting) egy olyan gazdasági előrejelzési módszer, amely a jelenlegi, vagy a nagyon közeli múltbeli állapotot igyekszik meghatározni, még mielőtt a hivatalos statisztikai adatok (például KSH GDP-adatok) napvilágot látnának.

[5] Active Learning Model

[6] Large Language Model, azaz a *Nagy Nyelvi Modell*, olyan mesterséges intelligencia típus, amelyet hatalmas mennyiségű szöveges adaton tanítottak be, hogy megértse és generálja az emberi beszédet.

Ez a cikk két példával illusztrálja a transzformeralapú modellek magyar nyelvű természetesszöveg-adatok klasszifikációjára történő alkalmazását. Először különböző BERT-modelleket[7] finomhangolunk, hogy szentimentpontszámokat generáljunk a hírekhez. Az adatok két jelentős magyar online hírportáltól származnak. Témamodellezést alkalmaztunk egy gazdasági és pénzügyi szempontból releváns híradatbázis létrehozásához. Különböző AL-heurisztikákat is alkalmaztunk, hogy felmérjük hatékonyságukat a modell pontosságának növelésében. Az annotációkat ChatGPT segítségével végeztük. A szentimentpontszámokat ezután magas gyakoriságú szentimentindexbe aggregáltuk. A kapott indexet prediktív képesség és időszerűség tekintetében értékeltük a legfontosabb makrogazdasági mutatókhoz viszonyítva, beleértve a bruttó hazai terméket (GDP), a beszerzésimenedzser-indexet (PMI) és a munkanélküliségi rátát. A prediktív teljesítmény értékeléséhez a Granger-oksági tesztet (*Granger 1969*) és bizonyos esetekben a Toda–Yamamoto-oksági tesztet (*Toda – Yamamoto 1995*) használtuk. Ezek olyan hipotézisvizsgálatok, amelyekkel mérhető, hogy egy idősor késleltetett értékei statisztikailag szignifikáns mértékben képesek-e előre jelezni egy másik idősor alakulását. A dinamikus idővetemítés (DTW[8], *Berndt – Clifford 1994*) módszerét használjuk a szentimentindexek és a makrogazdasági változók közötti összehangoltság mérésére. Továbbá, mivel a szövegalapú proxyváltozók és indexek hatékony módszerek a válságidőszakok előrejelzésére (*Baker et al. 2016*), küszöbértékes ARDL[9] (TADL[10], *Tong 1978*) modellt használtunk annak értékelésére, hogy a szentimentindexek válságidőszakokban (például a nagy pénzügyi válság vagy a Covid19-járvány idején) eltérően viselkednek-e. További érdekességként a betanított BERT-modellekkel elemeztük a FineWeb adatbázis egy részhalmazát, amely a CommonCrawl Repository webes keresési adathalmaza.

Második felhasználási esetünkben transzformer modelleket tanítottunk be, hogy a kiskereskedelmi üzletek nyugtáin szereplő termékek nevét vámtarifaszám-kategóriákba soroljuk. Az adatokat a Nemzeti Adó- és Vámhivataltól kaptuk kutatási célból, és azok az ország nagyobb kiskereskedelmi üzleteiből származó nyugták adatait tartalmazták. Különböző BERT-modelleket finomhangoltunk a terméknevek kategorizálására, amelynél két különböző típusú beágyazást hoztunk létre: az egyiket a szavak együttes előfordulása alapján, a másikat pedig úgy, hogy az előre betanított modell tokenizációja hozta létre a vektorreprezentációt.

A tanulmány további része a következőképpen épül fel. A *2.* és *3. fejezet* bemutatja a két általunk vizsgált felhasználási esetet; mindkettő tartalmazza a használt adatokat, a kutatás módszertanát és az eredményeket. A következtetéseket a *4*. *fejezet* tartalmazza.

[7] BERT-modellek: A BERT (Bidirectional Encoder Representations from Transformers) a Google által 2018-ban bemutatott nyelvi modell, amely forradalmasította a szövegértést. Míg a hagyományos modellek csak balról jobbra vagy jobbról balra olvasták a szöveget, a BERT *kétirányú* (bidirekcionális), így egy szó jelentését a teljes környezete (előtte és utána álló szavak) alapján értelmezi.
[8] Dynamic Time Warping
[9] Autoregressive Distributed Lag
[10] Threshold Autoregressive Distributed Lag

2. Online hírek szentimentelemzése

2.1. Adatok

Az első szöveges adatforrás két magyar hírportál gazdasági és pénzügyi híreinek gyűjteménye volt, amelyet az adott média hozzájárulásával gyűjtöttünk. A Jegybank és a hírportálok közötti megállapodásnak megfelelően a kiadók neve nem hozható nyilvánosságra, ezért azokra Médium 1 és Médium 2 néven hivatkozunk. A cikkek 1999 és 2020 közötti időszakból származnak. A gazdasági és pénzügyi szempontból releváns cikkek kiszűréséhez témamodellezést, látens Dirichlet-allokációt (LDA) alkalmaztunk. Ez a módszer a teljes korpuszon alapul, azzal a céllal, hogy a hasonló tartalmú híreket csoportosítsa. A modell azt feltételezi, hogy a cikkek látens „témák" keverékéből állnak, ahol a témák száma egy hiperparaméter. A tanítás során a modell kezdetben minden egyes cikkhez egy valószínűségi eloszlást rendel a témák között, valamint egy eloszlást rendel minden egyes témához a szavak között, majd ezeket az eloszlásokat a korpuszban megfigyelhető szavak együttes előfordulása alapján, iteratív módon frissíti. Az eljárás végén minden cikk esetében a legvalószínűbb „téma" meghatározható az eloszlás maximális elemének sorszámával. Rácskereséses módszert alkalmaztunk a hiperparaméterek, többek között a témák számának és a modell által egy iteráció során feldolgozott cikkek számának beállításához. A végső modellt a perplexitás alapján választottuk ki, ez a modell új adatokra való általánosítási képességét méri. A végső modell 16 kategóriát használt, és miután manuálisan ellenőriztük az egyes témákhoz tartozó 20 legvalószínűbb szót, közülük 13 kategória bizonyult gazdasági és pénzügyi szempontból relevánsnak. A végső tanító adatbázist úgy állítottuk össze, hogy kiszűrtük azokat a cikkeket, amelyek legvalószínűbb témája nem szerepelt a kiválasztott releváns kategóriák között.

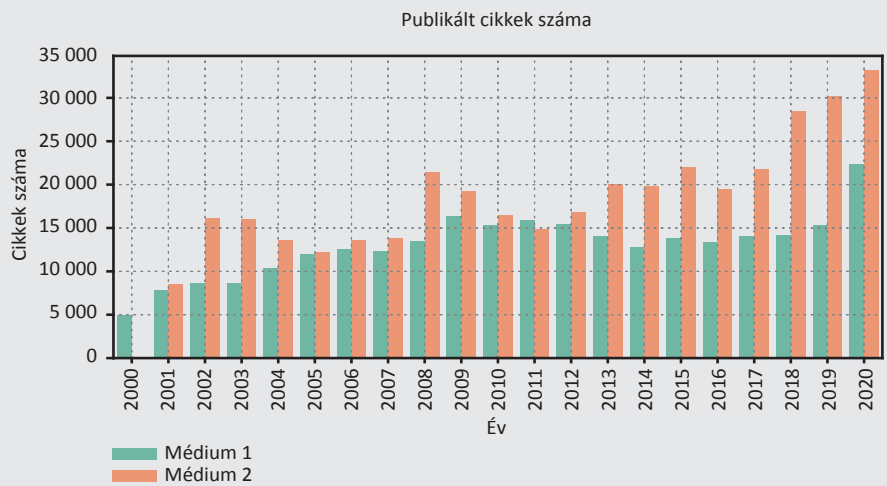
Az *1. ábra* a cikkek számának megoszlását mutatja a közzététel éve szerint, míg az *1. táblázat* a számokat médiumok szerinti bontásban mutatja be. Célunk kettős volt: Először is, egy gépi tanulási modellt szerettünk volna betanítani, amely egy adott gazdasági vagy pénzügyi témájú cikkhez pozitív, semleges vagy negatív szentimentet rendel. Másodszor, ezeket a pontszámokat havi szentimentindexbe terveztük összesíteni, hogy megvizsgáljuk, alkalmas-e a gazdasági folyamatok alakulásának előrejelzésére. Egy kicsi, de használható, címkézett adathalmaz létrehozása érdekében a 700 000 cikkben szereplő 9 millió mondatból 1 645-öt manuálisan címkéztünk többségi eljárással, amelynek során a Jegybank három gazdasági szakértője minden mondathoz hozzárendelt egy szentimentet, és a mondatok címkéjét akkor fogadtuk el, ha a három besorolás közül kettő egyezett. Ezek közül csak 87 százalék esetében volt többségi döntés, így az adatbázis 1 431-re szűkült. A tanítási-validációs tesztelési adatokhoz a fennmaradó cikkeket 70–20–10 százalékos arányban osztottuk fel, így 1 000 cikk szolgált tanító adatként, 287 a hiperparaméter-beállításhoz szükséges validációs adatként és 144 tesztelési adatként.

1. táblázat
A szövegkorpusz paraméterei

	Cikkek száma	Havi átlagos cikkszám	Első cikk	Utolsó cikk
Médium 1	293 665	1 112	2000.02.15.	2020.12.31.
Médium 2	404 894	1 533	2000.01.01.	2020.12.31.
Összesen	698 545	2 645	2000.01.01.	2020.12.31.

Forrás: Arthur et al. (2023:3)

1. ábra
A cikkek számának éves alakulása



Forrás: A szerzők számításai

2.2. Módszertan

Viszonyítási alapként szótáralapú módszert alkalmaztunk. Ez a módszer előre meghatározott szavak halmazát használja, amelyek pozitív vagy negatív gazdasági szentimentet közvetítenek. Ennek számos előnye van: egyszerűen magyarázható, számításigénye alacsony, és az eredmények könnyen értelmezhetők. A szótár összeállítása azonban nem egyszerű feladat. Azok a szavak, amelyek általános értelemben véve negatív jelentéssel bírnak, pénzügyi kontextusban elveszíthetik ezt a jelentésüket, ami miatt az általános szótár használata nem kellően hatékony (*Loughran – McDonald 2011*). Ez arra késztetett bennünket, hogy saját pénzügyi és gazdasági szótárat állítsunk össze azzal a céllal, hogy ne használjunk rögzített korpuszokat, mivel azok túlzottan korlátozták volna a szótárat, és torzították volna a konkrét adatbázisunkat. A modell úgy értékel egy mondatot, hogy megszámolja azokat a szavakat belőle, amelyek a szótárban is szerepelnek, azután összeadja a pontszámaikat (+1 a pozitív szentimentért, –1 a negatívért), majd normalizálja az értéket a teljes szótár

méretével. Ebből kapjuk egy cikk nettó szentimentpontszámát, amely az általános szentimentet fejezi ki: pozitív, ha ez a szám nagyobb, mint nulla, negatív, ha kisebb, mint nulla, és semleges egyéb esetben. Az indexet az egyes hónapok összes cikke pontszámainak átlagolásával hozzuk létre.

Az online hírek klasszifikációjára szolgáló mélytanulási megoldásunk két transzformer modellen alapul, és mindkettőjük a huBERT-re, a legnagyobb magyar webes korpuszon előre betanított BERT-modellre épül (*Nemeskey 2020*). Az egyik transzformer modell a névelem felismerésére (NER[11]) van finomhangolva (*Yang – Váradi 2023*), míg a másik a szentimentelemzésre (*Yang – Laki 2021*). Mindkét modell nyílt forráskódú, és ingyenesen elérhető a Huggingface[12] platformon. Modellünk betanítása magában foglalta az optimális hiperparaméterek optimalizálását. A számítások egyszerűsítéséhez az ingyenesen elérhető Optuna[13] szoftverkönyvtárat használtuk, és az összes különböző kombinációt a validációs halmazon mért veszteség alapján értékeltük, hogy elkerüljük a túlillesztést, beleértve a tételméretet, a modell tanításához szükséges ciklusok számát, a kiejtési (dropout) arányt és a súlycsökkenési arányt. Kísérletünk kimutatta, hogy a tételméret, azaz hogy egyszerre hány mondatot adunk a modellnek, volt a legtöbbet befolyásoló paraméter, ezért ennek a hangolása kiemelten fontos. Ez kompromisszumot teremt a jobb erőforrás-kezelés (a kis mennyiségű adat kevesebb terhelést jelent a rendszer számára) és a jobb általánosító képesség között (a kis mennyiségű adat kevesebb információs gradienst eredményez a veszteségfüggvény minimalizálása érdekében). Nem módosítottunk más hiperparamétereket, és ugyanazt a konfigurációt használtuk mind a NER, mind a szentimentértékelő modellek esetében.

Korlátozott nagyságú, annotált adatmennyiségünkkel és a címkézetlen hírcikkek nagy korpuszával az aktív tanulás alkalmazása a modell teljesítményének javítására megfelelő megoldásnak bizonyult. Az aktív tanulás egy olyan, ember által irányított módszertan, amelynek célja a gépi tanulási modell általánosítási képességeinek növelése. Ez iteratív képzési módszert jelent: miután a modellt a kezdeti képzési adatbázison tanítottuk, kiválasztjuk a címkézetlen adatok azon részhalmazát, amelyeket heurisztikus alapon a modell számára a leghasznosabbnak tartunk, hozzáadjuk ezeket a tanító adatokhoz, majd újra tanítjuk a modellt stb. Az új adatok iteratív címkézésének ezen folyamata arra irányul, hogy a legkevesebb adatot használja fel egy jól teljesítő klasszifikációs rendszer betanításához. Kísérleteink során három különböző heurisztikát használtunk. Kiindulópontnak véletlenszerű mondatokat vettünk mintául a címkézetlen mondatokból, figyelembe véve a cikkek havi eloszlását.

Első heurisztikánk a transzformer modelljeink által létrehozott vektorbeágyazást használja. Kiszámítjuk az összes címkézetlen mondat vektorbeágyazását. Ezután

[11] Named Entity Recognition
[12] https://www.huggingface.co (*NYTK/named-entity-recognition-nerkor-hubert-hungarian, NYTK/sentiment-hts5-hubert-hungarian)*
[13] https://www.optuna.org/

kiszűrjük a tesztkészletből azokat a mondatokat, amelyeknél a modell negatívnak jósolta a pozitív mondatot, és fordítva. Ezek az adatpontok jelentik a legnagyobb tanulási lehetőséget a modell számára ahhoz, hogy helyesen klasszifikálja a mondatok szentimentjét. Ezután megkerestük az összes címkézetlen mondatot, amelynek *koszinusztávolsága* ezektől kisebb, mint 0,00033, de még mindig pozitív; ezt a küszöbértéket a különböző távolságok eloszlása alapján választottuk ki. Végső adatbázisunk ezekből az új mondatokból vett mintán alapul. Az aktív tanulás ezen formája arra irányul, hogy jobb kontextust biztosítson a modell számára azáltal, hogy címkézi azokat a mondatokat, amelyek szemantikailag közel állnak azokhoz, amelyeket a modell tévesen klasszifikált.

A második heurisztika a neurális háló előrejelzését használja, és célja a modell általánosítási képességével kapcsolatos bizonytalanság értékelése. Mivel a transzformer modelleket úgy konfiguráltuk, hogy alkalmasak legyenek a klasszifikációra, a kimenet egy valószínűségi eloszlás a rendelkezésre álló szentimentkategóriák – negatív, semleges és pozitív – között. A modell bizonytalanságát a predikcióhoz tartozó *entrópia* segítségével mértük, amelyet az *(1) egyenlettel* számítottunk ki:

$$H(x) = -\sum_{i=1}^{3} p(x_i) \cdot log_2 p(x_i), \tag{1}$$

ahol $p(x_i)$ a három lehetséges kimenetel valószínűségét jelöli. Minél közelebb vannak egymáshoz ezek az értékek, annál bizonytalanabbak a modell előrejelzései. Ha egy kimenet valószínűsége sokkal magasabb a többinél, az bizalmat kelt a modell döntése iránt. Ez a megközelítés azt feltételezi, hogy azok az adatok, amelyekkel kapcsolatban a modell bizonytalan, két kategória döntési határán fekszenek, ezért ezeknek az adatpontoknak a címkézése segít a modellnek jobb pontosságot elérni. Vizsgálatunk kimutatta, hogy kb. 1,2-es entrópia értéknél a maximális valószínűség értéke észrevehetően megnőtt. Ezt a küszöbértéket használtuk azoknak a mondatoknak a mintavételéhez, amelyek entrópiája meghaladja ezt az értéket.

Végső heurisztikaként olyan mintavételi módszert alkalmaztunk, amely a mondatok beágyazását és a modell kimenetét kombinálta az úgynevezett *bizonytalanságon alapuló mintavétel* módszerével. Az összes címkézetlen mondatot a *(2)–(5) egyenletek* alapján számított *m* metrika szerint rangsoroltuk:

$$s_{LC}(x) = 1 - \max_i p(x_i) =: 1 - p(x_{\max}) \tag{2}$$

$$s_{MG}(x) = |p(x_{max}) - p(x_{max-1})| \tag{3}$$

$$D_{avg}(x) = d_{cos}(x, \overline{x_{sen}}) \tag{4}$$

$$m(x) = \left(1 - D_{avg}(x)\right) \cdot \left(0.6 \cdot s_{LC}(x) + 0.4 \cdot s_{MG}(x)\right) \tag{5}$$

ahol $s_{LC}(x)$ a *legkisebb bizalmat* jelöli, amely azt méri, hogy a modell mennyire biztos a legvalószínűbb előrejelzésében, $s_{MG}(x)$ az első és a második legvalószínűbb válasz közötti *különbséget* méri, míg $D_{avg}(x)$ a mondat beágyazásának koszinusztávolsága az átlagos $\overline{x_{sen}}$ beágyazástól.

Minden heurisztika esetében a minták 5 000 címkézetlen mondatból álltak; ezt a mennyiséget megfelelőnek tartottuk a modellek további tanításához. A címkézést a ChatGPT végezte a hivatalos OpenAI API használatával. Mivel a gazdasági hírcikkek nagy arányban tartalmaznak semleges mondatokat, az újonnan címkézett adatok összes kategóriáját a legkevesebb mondatot tartalmazó szentimenthez igazítottuk, hogy minden aktív tanulási stratégia számára kiegyensúlyozott adatbázist hozzunk létre. Az új címkézett mondatokat hozzáadtuk az eredeti tanítási adatokhoz, és a modellt a kibővített adatbázissal újratanítottuk. Az új, címkézetlen mondatok kiválasztásának és a ChatGPT segítségével történő annotálásának ezt az iterációját összesen négyszer hajtottuk végre a végleges modellek elkészítéséhez.

Az idősor-aggregációs módszertan hierarchikus struktúrát követett a mondat szintjétől a havi szintű indexekig. Először mondatszintű szentimentet nyertünk ki. A cikkszintű szentimentindexet az egyes cikkekben szereplő összes mondat szentimentpontszámának összegzésével állítottuk elő, így átfogó mérést kaptunk a szöveg általános hangulatáról. A havi szintű szentimentindexek kiszámításához az adott hónap összes cikkszintű szentimentpontszámának átlagát vettük, lehetővé téve az összehasonlítást a havi makrogazdasági adatokkal.

A szentiment értékekből összesített havi szentimentindexet három különböző makrogazdasági idősorral hasonlítottuk össze. A *bruttó hazai termék* (GDP)[14] egy adott időszak alatt egy ország határain belül előállított összes végső felhasználásra szánt áru és szolgáltatás összértékét méri. Ez egy ország gazdasági tevékenységének és általános gazdasági helyzetének egyik legfontosabb mutatója. A havi *munkanélküliségi ráta*[15] azt mutatja meg, hogy az adott hónapban a gazdaságilag aktív népességen belül mekkora arányt képviselnek a munkanélküliek. A mutatót a International Labour Organization (ILO) nemzetközi módszertana alapján számítják, és százalékos formában fejezik ki, így nemzetközileg összehasonlítható képet ad a munkaerőpiac helyzetéről. A *beszerzésimenedzser-index* (PMI)[16] egy gazdasági mutató, ami felmérések alapján méri a feldolgozóipari és szolgáltató szektor aktuális teljesítményét és üzleti hangulatát. Az index 50 pont felett bővülést, 50 pont alatt pedig zsugorodást jelez az adott ágazat gazdasági aktivitásában.

[14] Adatok forrása: MNB
[15] Adatok forrása: Eurostat
[16] Adatok forrása: Investing.com [Magyarország Feldolgozóipar Beszerzési Menedzser Index (PMI)]

A makrogazdasági változók kiválasztása szisztematikus empirikus megközelítéssel, gördülő ablakos korrelációelemzéssel történt. Az eljárás során azonosítottuk azokat a gazdasági mutatókat, amelyek a legjelentősebb együttmozgást mutatták a szentimentindexeinkkel. Ez a megközelítés lehetővé tette számunkra, hogy időben változó kapcsolatokat rögzítsünk és különböző időtartamok közötti mintákat azonosítsunk, ami a változók megbízható kiválasztását célozta. Mind a GDP, mind a munkanélküliségi ráta esetében az előző év azonos időszakához viszonyított (év/év) mutatókat használtunk, mivel ezek a mutatók alkalmasabbak a hírek dinamikáján keresztül közvetített, hosszú távú hatások rögzítésére. Az adatgenerálás folyamatából adódóan lehetnek eltérések az időbeli összehangolásban, ami befolyásolhatja az eredmények értelmezését. A havi GDP-becslés az MNB belső felhasználására készült mutatója. A munkanélküliségi adatok a Központi Statisztikai Hivataltól (KSH) származnak. A havi munkanélküliségi ráta közvetlen becsléséhez a KSH állapottérmodelleket használ a havi foglalkoztatási és munkanélküliségi adatok becsléséhez (*Horváth – Lovics 2023*). A feldolgozóipari PMI-adatokat a Magyar Logisztikai, Beszerzési és Készletezési Társaság (MLBKT) teszi közzé.

2.3. Értékelés

Modellünk hatékonyságának értékelésére több, a gépi tanulásban jól ismert metrikát alkalmaztunk. A klasszifikációhoz a súlyozott precizitást (precision score), visszahívást (recall score) és az F1 pontszámot (F1 score), valamint a kiegyensúlyozott pontossági értéket (weighted precision score) használtuk. Ezek a metrikák az összes osztályra vonatkozó értékeket kombinálják, és az egyes kategóriák mutatóinak súlyozott átlagát számolják ki, az elemek számát használva súlyként, így átfogó és kiegyensúlyozott képet adnak a modell teljesítményéről.

A generált szentimentindexhez két különböző idősorillesztési mérőszámot használtunk. Az első a Granger-féle oksági teszt (*Granger 1969*), amely egy statisztikai hipotézisvizsgálat. Ennek célja annak értékelése, hogy egy idősor előre jelezhet-e egy másikat, vagyis hogy az egyik idősor múltbeli értékei tartalmaznak-e olyan információkat, amelyek segítenek a másik idősor előrejelzésében. *X* és *Y* idősorok esetében két autoregresszív modellt építünk:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \varepsilon_t \tag{6}$$

és

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{i=1}^{p} \beta_i X_{t-i} + \varepsilon_t \tag{7}$$

A *(6)–(7) egyenletek* alapján a nullhipotézis szerint *X* nem Granger-oka *Y*-nak, ami a következővel egyenértékű:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_q = 0 \tag{8}$$

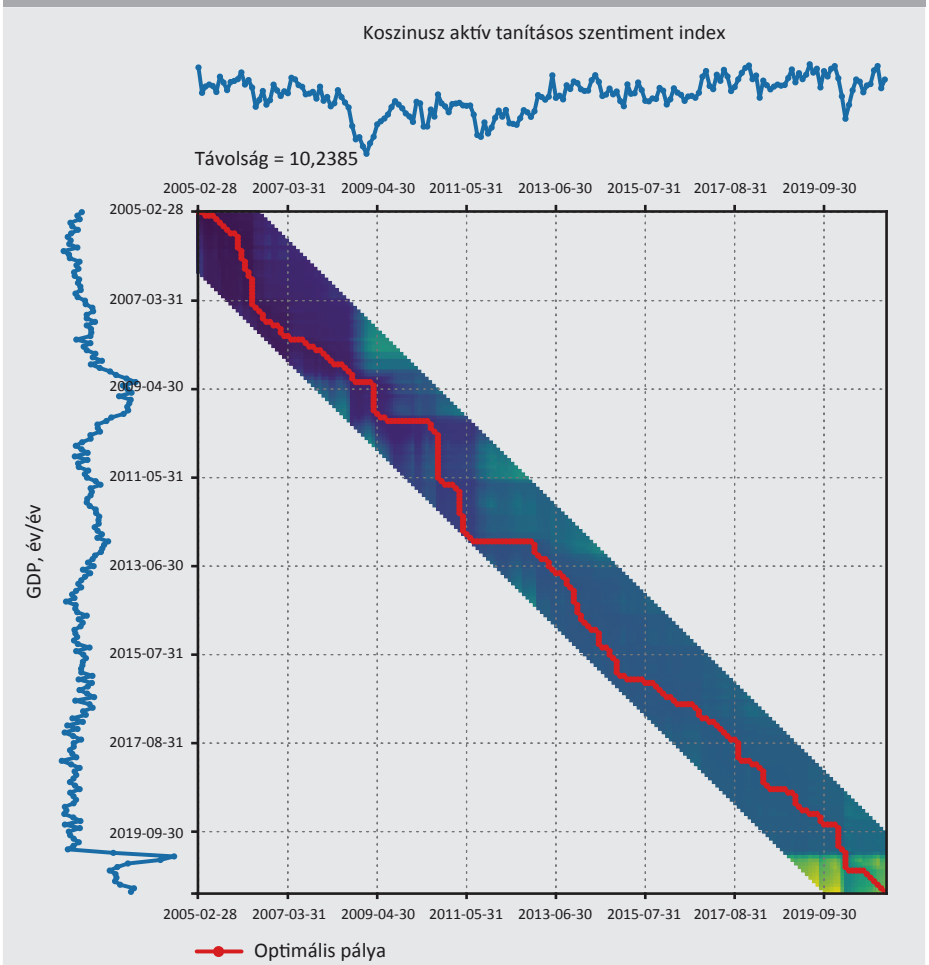
vagyis a *(8) egyenlet* alapján a β_j együtthatók mind nulla értékűek, vagyis az *X* múltbeli értékei nem számítanak az *Y* előrejelzésében. Ha bizonyos szignifikanciaszinten el tudjuk utasítani a nullhipotézist, az azt jelenti, hogy *X* késleltetett értékei statisztikailag szignifikáns előrejelző erővel rendelkeznek *Y* jövőbeli értékei számára. A Granger-oksági teszt azt feltételezi, hogy az idősorok stacionáriusak. Ezt a kiterjesztett Dickey–Fuller statisztikai teszt (ADF, *Dickey – Fuller 1979*) segítségével fogjuk ellenőrizni. Ha az ADF-teszt eredményei azt mutatták, hogy egy idősor nem stacionárius, akkor a *Toda – Yamamoto* (*1995*)-féle eljárást alkalmaztuk a Granger-oksági összefüggés tesztelésére. A módszer magában foglalja a változók közötti maximális integráció rendjének (d_{max}) meghatározását és a standard vektor autoregresszív (VAR) modell optimális késleltetés hosszának (*k*) kiválasztását. Ezt követően *VAR*($k+ d_{max}$) modellt becsülünk, majd az első *k* együtthatómátrixra Wald-tesztet alkalmazva elvégezzük az ok-okozati összefüggés vizsgálatát, azaz az utolsó d_{max} együtthatót figyelmen kívül hagyva. Ez a kiegészítés megbízható Granger-oksági tesztelést tesz lehetővé akkor is, ha a változók nem stacionáriusak vagy kointegráltak. Az autoregresszív kifejezések rendje, *p* egy olyan paraméter, amelyet magunknak kell beállítanunk. Az optimális késleltetésszám megtalálásához a Bayes-i információs kritériumot (BIC, *Schwarz 1978*) és a Hannan–Quinn információs kritériumot (HQ, *Hannan – Quinn 1979*) vizsgáltuk *p* összes értékére 8-ig, kiválasztottuk a legkisebb BIC- és HQ-értéket, majd elvégeztük a Granger-oksági tesztet ezzel a paraméterrel.

A két idősor összehangolásának értékelésére szolgáló második módszer a dinamikus idővetemítés (DTW, *Proakis – Manolakis 2007*), amely megpróbálja megtalálni a két idősor közötti optimális összehangolást. Az algoritmus az első idősor minden pontjához megpróbálja megtalálni a második idősorban azt a pontot, amelyikhez a legkisebb euklideszi távolság tartozik. Az algoritmus idősor-párokat ad eredményül, amelynek három feltételnek kell megfelelnie: 1) az egyik idősor minden pontjához találni kell a másik idősorból egy párt, 2) az első és az utolsó pontok mindig párban állnak egymással, és 3) a pároknak monoton növekvő sorrendben kell lenniük, vagyis a párok nem „keresztezhetik" egymást. A gyakorlatban egy másik, *globális korlátozásnak* is eleget kell tenni, azaz azok a pontok fogadhatók el, amelyek pozíciója a két sorozatban kellően „közel" van egymáshoz. Ha például egy pozitív *w* ablakméretet veszünk, akkor az első idősorban szereplő *j* koordinátához tartozó párosított *i* indexnek a [*j–w, j+w*] intervallumban kell lennie. Ez nemcsak a számítási terhelés csökkentését segíti, hiszen nem kell minden pontpár külön kombinációját kiszámítani – ami kvadratikus időkomplexitást eredményezne –, hanem összhangban is áll a lokális környezet elvével is: a cél, hogy időben egymáshoz közeli hasonlóságokat vizsgáljuk a proxy- és a makrogazdasági változók között.

Az algoritmus összefoglalva tehát egy pontpárosítást hoz létre a két idősor között, amelyben a végpontok párosítva vannak egymással, és a párosítás monoton növekvő mindkét idősorban. A *2. ábra* két idősor közötti optimális pálya példáját ábrázolja.

Az y-tengely a makrogazdasági változót, az x-tengely pedig a modell előrejelzéseiből generált szentimentindexet jelenti. A színes mezők az alkalmas párosításokat jelzik a globális korlát alapján, míg a színskála a párosítás költségét mutatja, amelyet az idősor két adatpontja közötti euklideszi távolság mér. A sötétebb szín alacsonyabb értéket jelöl. A piros pontok az optimális párosítások koordinátáit jelölik, a *Távolság* mező pedig a párosítások költségének összegét mutatja.

2. ábra
A koszinusz aktív tanulási index és az éves GDP-index optimális DTW-költségpályája (év/év)



Forrás: A szerzők számításai

A küszöbértékes ARDL (TADL) modellekkel azt becsültük, hogy a hírekből származó index és a makrogazdasági adatok közötti általános illeszkedés eltér-e válság, illetve normál periódusban. A TADL-modellek különösen hasznosak a rendszerbeli változások és a nemlineáris dinamikák feltérképezésében idősoros adatok esetén, mivel segítenek feltárni, hogy a mögöttes gazdasági kapcsolatok jelentősen változnak-e különböző körülmények között. A rezsimek számát és a küszöbértékeket a Bai–Perron-teszt segítségével határoztuk meg (lásd *Bai – Perron 1998*).

A TADL-modell kiterjeszti a standard küszöbértékes autoregresszív (TAR[17]) modell specifikációját azáltal, hogy lehetővé teszi egyszerre a jelenlegi és az elosztott késleltetésű tagok bevonását. Két rezsimes TADL-modell esetén a képletet a *(9) egyenlet* adja meg:

$$y_t = \begin{cases} c_1 + \sum_{i=1}^{p} \alpha_{1,i} y_{t-i} + \sum_{j=0}^{q} \beta_{1,j} x_{t-j} + \varepsilon_t, & \text{ha } y_{t-1} \leq \gamma \\ c_2 + \sum_{i=1}^{p} \alpha_{2,i} y_{t-i} + \sum_{j=0}^{q} \beta_{2,j} x_{t-j} + \varepsilon_t, & \text{ha } y_{t-1} > \gamma \end{cases}; \qquad (9)$$

ahol y_t a függő változó *t* időpontban; x_{t-j} a magyarázó változó késleltetett értékei; y_{t-j} a függő változó késleltetett értékei; $\alpha_{1,i}$ és $\alpha_{2,i}$ a függő változó autoregresszív együtthatói az 1. és 2. rezsimben; $\beta_{1,i}$ és $\beta_{2,i}$ a magyarázó változó együtthatói az egyes rezsimekben; *γ* a két rezsim közötti határt meghatározó küszöbérték; és ε_t a hibatag.

2.4. Eredmények

A *2. táblázat* összegzi a modelltanítás eredményeit. Az aktív tanulás nélkül tanított szentimentmodellt választottuk alapul, mivel az aktív tanulási modellek erre a modellre épültek, így ez volt a legkézenfekvőbb választás. Ez lehetővé tette számunkra, hogy értékeljük, milyen teljesítményt nyújtanak a különböző aktív tanulási heurisztikák. Az összességében legjobban teljesítő modell a koszinuszalapú aktív tanulásos modell volt, amely a legmagasabb pontosságot érte el. A bizonytalanságon alapuló aktív tanulásos modell a semleges mondatok kategorizálásában nyújtotta a legjobb teljesítményt, amit fontosnak tartottunk, mivel a híradások szentimenttartalma erősen a semleges érzelmek felé tolódik el. Ezt a két modellt választottuk a szentimentindexek generálásához a híradások adatbázisából. (Az optimális DTW-távolságokat a *7. táblázat* tartalmazza a *Függelékben*.) A PMI esetében mindkét aktív tanulásos modell hasonló eredményt hozott. A szótáralapú módszer érte el a legalacsonyabb távolságot az GDP esetében, míg a munkanélküliségi ráta esetében a BERT-modellek alacsonyabb távolságot mutattak, mint a szótáralapú modell.

[17] Threshold Autoregressive

2. táblázat
A szentimentklasszifikáció finomhangolásának eredményei

Modell	Precizitás (%)	Visszahívás (%)	F1-mutató (%)	Kiegyensúlyozott pontosság (%)
Alapértelmezett szentimentmodell	62,64	63,19	62,47	62,16
Bizonytalanságalapú aktív tanulásos (AL) modell	65,76	65,28	65,35	65,14
Koszinusz aktív tanulásos (AL) modell	70,77	70,14	70,29	70,19

Forrás: A szerzők számításai

Az optimális késleltetés hossz elemzése ugyanazokat az eredményeket hozta a PMI és a GDP esetében, mind a BIC-, mind a HQ-kritériumok alkalmazásával. A munkanélküliségi ráta esetében az idősor tulajdonságait figyelembe véve a HQ-eredményt használtuk. (Az eredmények a *6. táblázatban* találhatók a *Függelékben*.)

A Granger-oksági elemzés azt mutatta, hogy valamennyi hírekre vonatkozó szentimentindex olyan információkat tartalmaz, amelyek segítenek a vizsgált makrogazdasági mutatók előrejelzésében. A munkanélküliségi ráta, a koszinusz AL szentimentindex és a bizonytalanságalapú AL szentimentindex esetében a Toda–Yamamoto-módszert alkalmaztuk. (A stacionaritási tesztek eredményeit a *Függelék 5. táblázata*, míg a Granger-oksági tesztek eredményeit a *3. táblázat* tartalmazza.)

3. táblázat
A Granger-oksági teszt eredményei: OLL az 5. táblázatban feltüntetett HQ- és Granger-teszt típus alapján

	Koszinusz AL szentimentindex	Bizonytalanságalapú AL szentimentindex	Szótáralapú szentimentindex
GDP év/év	0,0678	0,063	0,0000
PMI	0,0004	0,0002	0,0000
Munkanélküliségi ráta év/év	0,0255	0,017	0,0000

Megjegyzés: év/év: éves összehasonlítás
Forrás: A szerzők számításai

A makrogazdasági változók és a szentimentindexek TADL-elemzése alátámasztja azt a megállapítást, hogy a hírindexek és a makrogazdasági adatok közötti általános összhang a válság és a válság nélküli normál időszakokban eltérő. Az alábbiakban a gazdasági mutatókat függő változóként, a koszinusz AL szentimentindexet pedig független változóként használjuk, ami a fentiek szerint a legjobban teljesítő modellt jelenti. A PMI esetében két rezsimet azonosítottunk. A küszöbérték 49,5, ami összhangban van a PMI-index értelmezésével, ahol az 50 alatti érték csökkenést jelez. Amennyiben az értékek a küszöbérték alatt vannak, mind a késleltetett, mind az egyidejű szentimentindexek statisztikailag szignifikánsak, ami előremutató és

prediktív hatást jelez. Ezzel szemben a másik rezsimben a kapcsolat szimultán együttmozgást sugall. A GDP esetében a TADL-modell három különböző rezsimet azonosít: az egyik a jelentős csökkenések időszakaihoz, a második a „normál" változások időszakaihoz, a harmadik pedig a jelentős növekedések időszakaihoz kapcsolódik. Az eredmények azt mutatják, hogy a hírek szentimentindex késleltetett változói ezekben a rezsimekben eltérő hatásokkal bírnak. A GDP csökkenésével jellemezhető rezsimben a hatások előremutató jellegűek, vagyis a késleltetett változók prediktív hatással bírnak, míg a „normál" változások időszakában a hatás statisztikailag nem szignifikáns. Ezzel szemben a növekedési időszakban a hírindexek egyidejű összefüggést jeleznek a GDP-vel. A munkanélküliségi ráta tekintetében a TADL-modell négy különböző rezsimet azonosít: egy a munkanélküliségi ráta változásának jelentős csökkenéseihez, kettő a mérsékelt változásokhoz és egy a munkanélküliségi ráta változásának hirtelen emelkedéséhez kapcsolódik. Az eredmények arra utalnak, hogy a négy rezsimben nagyon különböző folyamatok zajlanak. A hírindex késleltetett változói csak azokban az időszakokban voltak szignifikánsak, amikor a munkanélküliségi ráta magasan emelkedett, akkor viszont vezető és prediktív hatást jelzett. (A TADL-elemzés eredményei a *8. táblázatban* találhatók a *Függelékben*.)

2.5. Következtetések a Fineweb2 adatbázis alapján

Aktív tanulással támogatott transzformer modelljeink további értékeléseként egy másik magyar adatbázisból próbáltunk létrehozni indexet. A Fineweb2 adatbázis egy több billió tokenből álló, több mint 1 000 nyelven összegyűjtött szöveg adatbázis, amely a CommonCrawl nyílt webes adatbázisából tevődik össze, hogy egy tiszta, többnyelvű adatbázist hozzon létre minden NLP-feladat számára. Minden adat egy weboldal szöveges információinak pillanatképe, amelyet egy speciális előfeldolgozási folyamatnak vetettek alá, amely magában foglalja a weboldal tartalmának duplikációmentesítését, az NSFW-webhelyek[18] szűrését és a kódolási problémák kijavítását. Az előfeldolgozás részletes leírásához lásd: *Penedo et al.* (*2025*).

Az adatbázis magyar nyelvű részét választottuk ki, amely 50 millió weboldal szövegéből állt, 2013 márciusa és 2024 áprilisa közötti időbélyegekkel. A cikkek alapján tanított LDA-modellt használtuk a látens „témák" kikövetkeztetésére, és megtartottuk azokat az adatokat, amelyek ugyanabba az előre meghatározott kategóriába tartoztak. Mivel a teljes szűrt adatbázis feldolgozása észszerű időn belül nem volt kivitelezhető, úgy döntöttünk, hogy véletlenszerű mintát veszünk, amelyben minden napról 100 adatpontot választottunk ki. Ez 110 000 egyedi weboldalt eredményezett, amelyek összesen 3 665 315 mondatból álltak, amit elegendő adatnak ítéltünk az index létrehozásához. A modell-következtetés és az index létrehozása a *2.2. fejezet*-

[18] NSFW-webhelyek: Az *NSFW* egy angol mozaikszó: *„Not Safe For Work"*, ami magyarul annyit tesz: *„nem biztonságos a munkahelyen"*. Ezt a jelölést olyan internetes tartalmakra (webhelyekre, képekre, videókra vagy fórumbejegyzésekre) használják, amelyek megnyitása kínos vagy problémás lehet.

ben leírtakkal megegyező módon történt. A *Függelék 4. ábrája* a koszinusz AL modell generált indexeinek diagramját mutatja, összehasonlítva az előző évhez viszonyított GDP százalékos változásával. A szentimentindexre lineáris skálázást alkalmaztunk, így minimális és maximális értéke megegyezik a GDP-változás megfelelő minimális és maximális értékével. A szentimentindexek – úgy tűnik – követik a GDP dinamikáját, különösen a Covid19-járvány kezdetén. A Covid19-re reagálva mind a GDP, mind a szentimentindexek jelentős csökkenést mutattak. A szentimentindexek csökkenése azonban tartósabbnak tűnik. Ennek egyik lehetséges oka, hogy a hírek hosszú ideig a Covid okozta gazdasági terhekre és az ebből fakadó bizonytalanságok hosszú távú hatásaira összpontosítottak. A negatív információk folyamatos áramlása, valamint a gazdaság lassú ütemű, újbóli fellendülése hozzájárulhatott ahhoz, hogy az indexek hosszabb ideig alacsony szinten maradtak. A Granger-oksági tesztek azonban nem mutattak statisztikailag szignifikáns eredményt arra vonatkozóan, hogy bármelyik idősor Granger-oka volna a másiknak.

3. Kiskereskedelmi termékek klasszifikációja

3.1. Adatok

A kereskedelmi termékek klasszifikációs feladatához az Online Pénztárgép kivonatát használtuk, ami a Nemzeti Adó- és Vámhivatal (NAV) által biztosított, körülbelül 445 000 egyedi kereskedelmi terméknevet tartalmazó listát jelenti. Célunk egy olyan modell finomhangolása volt, amely meghatározza a termék vámtarifaszám-kategóriáját. Ezek a kategóriák Magyarországon a Kombinált Nómenklatúra (KN) szerint vannak meghatározva, amely az Európai Unió kiskereskedelmi áruk klasszifikációjának szabványa, és a Harmonizált Rendszeren (HS) alapul. A KN-kód egy nyolcjegyű hierarchikus kódrendszer, amelyben minden két további kód tovább részletezi az egyes kategóriákat. Adatbázisunkban csak 53 292 egyedi kereskedelmi termék rendelkezett érvényes négyjegyű KN-kóddal. Ezen kategóriák közül 17 négyjegyű kategóriát választottunk ki, mivel ezek a kiskereskedelmi termék kategóriák voltak, amelyek kutatásunk középpontjában álltak, és elegendő darabszámmal rendelkeztek ahhoz, hogy a finomhangolás hatékony legyen.

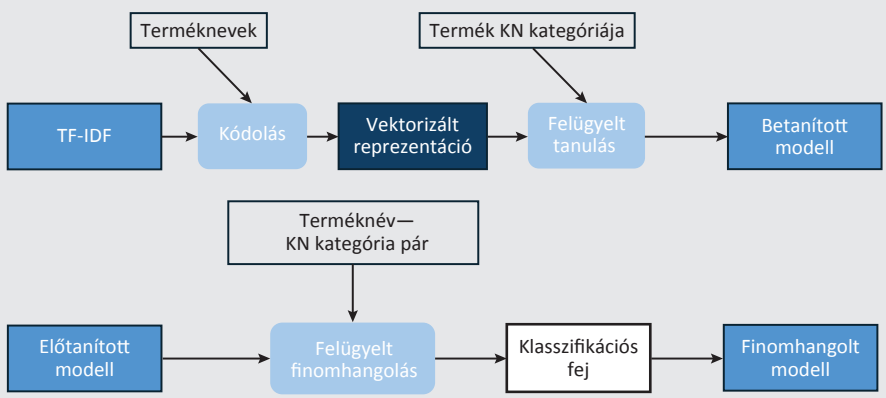
Az adatok tisztítása a következő lépésekből állt: először is eltávolításra került minden olyan jelölés, amely utalhatott arra, hogy a nyugta melyik kiskereskedőtől származik (pl. az üzlet nevét egy üzletben kizárólagosan forgalmazott termék esetében). Adatvédelmi és biztonsági okokból ezt a lépést az adóhatóság végezte el, továbbá eltávolították az üres tételeket, a termékeknek nem megfelelő elemeket (kedvezmények, kuponok) és az extra karaktereket (vesszők, kettőspontok, szóközök). Számos termék nevében szerepelt csomagolásra vonatkozó információ (göngyölegek, súly, méretek). A második lépésben két adathalmazt hoztunk létre: egyet, amelyből eltávolítottuk a csomagolásra vonatkozó extra információkat, és egyet, amelyben ezek megmaradtak, majd összehasonlítottuk a modellek teljesítményét mindkét

adathalmazra. Először az adatbázist 80–20 százalékos arányban osztottuk fel a tanító-teszt felosztáshoz, majd a tanító adatokból 10 százalékos véletlenszerű mintát vettünk a validációs megosztáshoz.

3.2. Módszertan

Két különböző lehetőséget azonosítottunk a beágyazások létrehozására. Először egy módosított TF-IDF-algoritmus[19] (*Sparck Jones 1972*) segítségével jöttek létre azok, amely a karakterek együttes előfordulását használja hasonló vektorok létrehozására hasonló terméknevekhez. Az első módszerrel betanított modelleket *vektorizált* modelleknek nevezzük. A második lehetőség szerint egy előtanított modell a tokenizációján keresztül hozza létre a termék vektorreprezentációját. Három különböző modellt hasonlítottunk össze: RoBERTa (*Liu et al. 2019*), huBERT (*Nemeskey 2020*) és PULI (*Yang et al. 2023*). Minden modell transzformer alapú, nyílt forráskódú modell, és elérhető a HuggingFace-en[20]. A finomhangolt modellek esetében – mivel a nyelvi modellek a következő token előrejelzésére vannak betanítva – módosítottuk az utolsó réteget, hogy az összhangban legyen a klasszifikációs céljainkkal. Ez a folyamat hat különböző modell létrehozásához és összehasonlításához vezetett. A módszerek végső betanítási folyamatát a *3. ábra* szemlélteti.

3. ábra
A vektorizált betanítás és finomhangolás sematikus ábrázolása



Forrás: A szerzők szerkesztése

[19] A *TF-IDF* (Term Frequency – Inverse Document Frequency) egy statisztikai mérőszám, amelyet a szövegbányászatban és az információ-visszakeresésben használnak annak meghatározására, hogy egy szó mennyire fontos egy adott dokumentumban egy dokumentumgyűjteményen (korpuszon) belül.
[20] SZTAKI-HLT/huBERT-base-cc, sentence-transformers/all-roberta-large-v1, NYTK/PULI-GPT-2

3.3. Eredmények

Az eredményeket a *4. táblázat* foglalja össze. Minden modell és betanítás esetében a teljesítmény jobb volt annál az adathalmaznál, amelyből nem távolították el a csomagolási információkat, mint annál, amelyből azokat eltávolították, ami azt jelenti, hogy ez az információ segítette a modelleket a különböző kategóriák megkülönböztetésében. Ha csak a precizitási értékeket nézzük, akkor azt gondolhatnánk, hogy a két különböző tanulási paradigma teljesítménye hasonló, de egy átfogóbb mutató, például a kiegyensúlyozott pontosság (weighted precision), a finomhangolás kiemelkedőségét mutatja a modellekből generált beágyazások használatával szemben. Ez a különbség azt jelzi, hogy a transzformereket vezérlő figyelemmechanizmus nagyobb mértékben képes megragadni a termékek leírásának különböző részei közötti szemantikai kapcsolatokat, mint egy egyszerű együttes előfordulási algoritmus. Mind a vektorizált, mind a finomhangolt modellek esetében a huBERT-alapú modell teljesített a legjobban a legtöbb mérőszámot illetően. Ez várható eredmény, mivel a PULI-modell egy generatív modell, amely kevésbé alkalmas klasszifikációs feladatokra. A PULI és huBERT modellek a RoBERTa-modellhez képest tapasztalható jobb teljesítménye annak tudható be, hogy míg a RoBERTa *többnyelvű* korpuszon volt előtanítva, addig a huBERT és a PULI kizárólag *magyar* nyelvű korpuszon előtanított modell, így elsősorban magyar nyelvű adatokkal végzett feladatokra alkalmasak.

4. táblázat
A vektorizált és a finomhangolt modell eredményei

Modell	Precizitás (%)	Visszahívás (%)	F1-mutató (%)	Kiegyensúlyozott pontosság (%)
Vektorizált RoBERTa	61,29	17,32	6,27	6,66
Vektorizált huBERT	60,45	17,35	6,34	6,67
Vektorizált PULI	61,96	17,22	6,06	6,58
Finomhangolt RoBERTa	85,84	85,73	85,69	83,42
Finomhangolt huBERT	88,02	87,97	87,93	85,82
Finomhangolt PULI	86,63	86,49	86,44	83,95

Forrás: A szerzők számításai

4. Következtetések

Tanulmányunkban a természetes nyelv feldolgozásának módszertanait vizsgáltuk mélytanulási modellekre támaszkodva, pontosabban transzformeralapú architektúrákat alkalmazva, azzal a céllal, hogy magas gyakoriságú mutatókat hozzunk létre strukturálatlan gazdasági szövegekből. Pénzügyi és gazdasági hírcikkeket, valamint kiskereskedelmi terméknéveket vizsgáltunk, hasonlóan *Kalamara és szerzőtársai* (*2022*), továbbá *Aguilar és szerzőtársai* (*2021*) munkáihoz, és ezeket működőképes klasszifikációs eszközökké alakítottuk. Ezek a források jó alapot szolgáltattak

a gazdasági mutatók becslésének támogatásához, amelyeket a használt technikákkal validáltunk.

A szentimentelemzéshez mélytanulási modelljeink hatékonyan rendelték hozzá a szentimentkategóriákat a különböző cikkekhez. Az aktív tanulásos heurisztikák alkalmazása javította a modellek általánosítási képességeit, és hatékonyabbá tette őket azáltal, hogy kevesebb finomhangoláshoz szükséges tanulási adatot használt, hasonlóan *Üveges és szerzőtársai* (*2024*) eredményeihez. A koszinuszalapú AL-modell érte el a legmagasabb általános pontosságot, míg a bizonytalanságalapú AL-modell kiemelkedő teljesítményt nyújtott a semleges mondatok klasszifikációjában, ami különösen fontos, tekintettel arra, hogy a gazdasági hírekben a szentiment erősen a semleges irányba tendál. Az így kapott szentimentindex hatékonynak bizonyult a gazdasági visszaesések előrejelzésében, különösen ott, ahol magas gyakoriságú adatok nem állnak rendelkezésre.

Az idősorelemzés komplex, rendszertől függő prediktív kapcsolatokat tárt fel a hírek szentimentindexei és a legfontosabb makrogazdasági változók között. A tanulmányunkban alkalmazott főbb módszerek a következő eredményeket mutatták:

1. Granger-oksági teszt és dinamikus idővetemítés (DTW): A Granger-okság elemzés azt mutatta, hogy az összes cikkekből generált szentimentindex statisztikailag szignifikáns információkat tartalmaz a vizsgált makrogazdasági mutatók előrejelzéséhez. Az optimális illeszkedés tekintetében a szótáralapú módszer adta a legalacsonyabb DTW-távolságot az éves GDP-változással és a PMI-vel összehasonlítva, míg az aktív tanulásos indexek jobban illeszkedtek a munkanélküliségi ráta változásához.

2. Elosztott késleltetésű küszöbértékes autoregresszív (TADL) elemzés: A rezsimváltásokat megragadó TADL-modellek megerősítették, hogy a hírek indexei és a makrogazdasági adatok közötti korreláció jelentősen eltér válság és válságmentes időszakokban.

Eredményeink azt mutatják, hogy a kiskereskedelmi termékek klasszifikációjának területén a megvalósított modellek – konkrétan a finomhangolt transzformer architektúrák (RoBERTa, huBERT, PULI) – kiemelkedő teljesítményt mutattak a referenciamodellekhez képest, és ezt a teljesítményt a kiegyensúlyozott pontossági pontszámok is alátámasztották. A magyar nyelvű korpuszon előre betanított huBERT-modell a legtöbb klasszifikációs modell közül a legjobb teljesítményt nyújtotta, ami rámutat a nyelvspecifikus előtanítás előnyeire a magyar nyelv elemzését igénylő feladatok esetében. Ezenkívül a csomagolási információk beépítése a kereskedelmi terméknevekbe javította a modell hatékonyságát, ami jelzi a modell hasznosságát a termékkategóriák megkülönböztetésében.

Összefoglalva, ez a tanulmány bemutatja a fejlett mélytanulás és NLP-technikák strukturálatlan adatokra való alkalmazásának jelentős potenciálját, értékes eszközöket kínálva a makroökonómiai szakemberek és azon piaci döntéshozók számára, akik jobb előrejelző képességet és a nemlineáris gazdasági dinamikák mélyebb megértését keresik. A jövőbeli kutatási irányok között szerepelhet egy kifinomultabb szentiment-kategóriarendszer kidolgozása speciális szentimentindexek létrehozása céljából, amelyek lehetővé tennék a gazdasági narratívák és a különböző gazdasági területekre gyakorolt hatásuk részletesebb elemzését.

Felhasznált irodalom

Aguilar, P. – Ghirelli, C. – Pacce, M. – Urtasun, A. (2021): *Can news help measure economic sentiment? An application in COVID-19 times*. Economics Letters, 199, 109730. https://doi.org/10.1016/j.econlet.2021.109730

Arthur, F.V. – Gyires-Tóth, B. – Debreczeni, M.I. – Ónozó, L.R. (2023): *Language of the Market: NLP-Driven Sentiment Analysis of the Hungarian Economy.* In: 14th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, pp. 93–98. https://doi.org/10.1109/CogInfoCom59411.2023.10397544

Ash, E. – Hansen, S. (2023): *Text Algorithms in Economics.* Annual Review of Economics, 15: 659–688. https://doi.org/10.1146/annurev-economics-082222-074352

Babii, A. – Ghysels, E. – Striaukas, J. (2021): *Machine Learning Time Series Regressions with an Application to Nowcasting*. Journal of Business & Economic Statistics, 40(3): 1094–1106. https://doi.org/10.1080/07350015.2021.1899933

Bai, J. – Perron, P. (1998): *Estimating and testing linear models with multiple structural changes*. Econometrica, 66(1): 47–78. https://doi.org/10.2307/2998540

Baker, S.R. – Bloom, N. – Davis, S.J. (2016): *Measuring economic policy uncertainty.* Quarterly Journal of Economics, 131(4): 1593–1636. https://doi.org/10.1093/qje/qjw024

Beltagy, I. – Lo, K. – Cohan, A. (2019): *SciBERT: A Pretrained Language Model for Scientific Text*. In: 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. 3615–3620, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Berndt, D.J. – Clifford, J. (1994): *Using dynamic time warping to find patterns in time series.* In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. AAAIWS'94, AAAI Press, Seattle, WA, pp. 359–370. http://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf. Letöltés ideje: 2025. május 5.

Chen, C. – Palmer, A. – Sporleder, C. (2011): *Enhancing active learning for semantic role labeling via compressed dependency trees*. In: Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, pp. 183–191. https://aclanthology.org/I11-1021.pdf. Letöltés ideje: 2025. szeptember 28.

Devlin, J. – Chang, M.W. – Lee, K. – Toutanova, K. (2019): *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1: 4171–4186. https://aclanthology.org/N19-1423.pdf. Letöltés ideje: 2024. december 10.

De Bondt, G. – Sun, Y. (2025): *Enhancing GDP nowcasts with ChatGPT: a novel application of PMI news releases*. Working Paper 3063, European Central Bank. https://doi.org/10.2866/2788332

Dickey, D.A. – Fuller, W.A. (1979): *Distribution of the estimators for autoregressive time series with a unit root.* Journal of the American Statistical Association, 74(366a): 427–431. https://doi.org/10.1080/01621459.1979.10482531

Granger, C.W.J. (1969): *Investigating causal relations by econometric models and cross-spectral methods.* Econometrica, 37(3): 424–438. https://doi.org/10.2307/1912791

Gentzkow, M. – Kelly, B. – Taddy, M. (2019): *Text as data.* Journal of Economic Literature, 57(3): 535–574. https://doi.org/10.1257/jel.20181020

Hannan, E.J. – Quinn, B.G. (1979): *The determination of the order of an autoregression*. Journal of the Royal Statistical Society: Series B (Methodological), 41(2): 190–195. https://www.jstor.org/stable/2985032

Horváth Beáta – Lovics Gábor (2023): *Havi munkaügyi adatok becslésének módszertana a KSH-ban*. Szigma, 54(3–4): 205–226. https://doi.org/10.15170/SZIGMA.54.1190

Huang, A.H. – Wang, H. – Yang, Y. (2023): *FinBERT: A Large Language Model for Extracting Information from Financial Text*. Contemporary Accounting Research, 40(2): 806–841. https://doi.org/10.1111/1911-3846.12832

Jiang, S. – Pang, G. – Wu, M. – Kuang, L. (2012): *An improved K-nearest-neighbor algorithm for text categorization*. Expert Systems with Applications, 39(1): 1503–1509. https://doi.org/10.1016/j.eswa.2011.08.040

Kalamara, E. – Turrell, A. – Redl, C. – Kapetanios, G. – Kapadia, S. (2022): *Making text count: Economic forecasting using newspaper text*. Journal of Applied Econometrics, 37(5): 896–919. https://doi.org/10.1002/jae.2907

Liu, Y. – Ott, M. – Goyal, N. – Du, J. – Joshi, M. – Chen, D. et al. (2019): *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Loughran, T. – McDonald, B. (2011): *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.* The Journal of Finance, 66(1): 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Nasiopoulos, D.K. – Roumeliotis, K.I. – Sakas, D.P. – Toudas, K. – Reklitis, P. (2025): *Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models.* International Journal of Financial Studies, 13(2), 75. https://doi.org/10.3390/ijfs13020075

Nemeskey Dávid Márk (2020): *Natural Language Processing for Language Modeling*. Ph. D. dissertation, Eötvös Loránd University, Budapest. https://doi.org/10.15476/ELTE.2020.066

Ónozó Lívia Réka – Putz Orsolya – Járási István – Gyires-Tóth Bálint (2024a): *Kiskereskedelmi terméknevek kategorizálása Kombinált Nomenklatúra szerint.* In: Berend, G. – Gosztolya, G. – Vincze, V. (eds.): XX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged, Magyarország, pp. 131–144. https://m2.mtmt.hu/gui2/?mode=browse&params=publication;34560678. Letöltés ideje: 2024. december 3.

Ónozó, L.R. – Arthur, F.V. – Gyires-Tóth, B. (2024b): *Leveraging LLMs for Financial News Analysis and Macroeconomic Indicator Nowcasting.* In: IEEE Access, Vol. 12: 160529–160547. https://www.doi.org/10.1109/ACCESS.2024.3488363

Penedo, G. – Kydlíček, H. – Sabolčec, V. – Messmer, B – Foroutan, N. – Kargaran, A.H. et al. (2025): *FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language.* Second Conference on Language Modeling. https://openreview.net/pdf?id=jnRBe6zatP. Letöltés ideje: 2025. szeptember 13.

Proakis, J.G. – Manolakis, D.G. (2007): *Digital Signal Processing: Principles, Algorithms and Applications*, 3rd Edition. Prentice-Hall International, Incorporated. https://uvceee.files.wordpress.com/2016/09/digital_signal_processing_principles_algorithms_and_applications_third_edition.pdf. Letöltés ideje: 2025. augusztus 28.

Rostam, Z.R.K. – Kertész, G. (2025): *Advances in Pre-trained Language Models for Domain-Specific Text Classification: A Systematic Review.* ACM Transactions on Intelligent Systems and Technology, 16(6), 124. https://doi.org/10.1145/3763002

Schröder, C. – Niekler, A. – Potthast, M. (2022): *Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers.* In: Muresan, S. – Nakov, P. – Villavicencio, A. (eds.): Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, Dublin, Ireland, pp. 2194–2203. https://doi.org/10.18653/v1/2022.findings-acl.172

Schwarz, G. (1978): *Estimating the Dimension of a Model.* The Annals of Statistics, 6(2): 461–464. http://www.jstor.org/stable/2958889

Settles, B. (2011): *From Theories to Queries: Active Learning in Practice.* In: Guyon, I. – Cawley, G. – Dror, G. – Lemaire, V. – Statnikov, A. (eds.): Active Learning and Experimental Design workshop in conjunction with AISTATS 2010, pp. 1–18. http://proceedings.mlr.press/v16/settles11a/settles11a.pdf. Letöltés ideje: 2025. szeptember 10.

Sobrino, N.D. – Ghirelli, C. – Hurtado, S. – Pérez, J.J. – Urtasun, A. (2020): *The narrative about the economy as a shadow forecast: an analysis using Banco de España quarterly reports.* Working Papers 2042, Banco de España. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/20/Files/dt2042e.pdf

Sparck Jones, K. (1972): *A statistical Interpretation of Term Specificity and its Applications in Retrieval*. Journal of Documentation, 28(1): 11–21. https://doi.org/10.1108/eb026526

Tetlock, P.C. (2007): *Giving content to investor sentiment: The role of media in the stock market*. Journal of Finance, 62(3): 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Toda, H.Y. – Yamamoto, T. (1995): *Statistical inference in vector autoregressions with possibly integrated processes*. Journal of Econometrics, 66(1–2): 225–250. https://doi.org/10.1016/0304-4076(94)01616-8

Tong, H. (1978): *On a threshold model*. In: Chen, C. (ed.): Pattern Recognition and Signal Processing. NATO ASI Series E: Applied Sc., (29). Sijthoff & Noordhoff, Netherlands, pp. 575–586. https://www.researchgate.net/publication/246995827_On_a_Threshold_Model_in_Pattern_Recognition_and_Signal_Processing

Üveges István – Vági Renátó – Megyeri Andrea – Fülöp Anna – Nagy Dániel – Vadász J. Pál et al. (2024): *Saving labeling cost by embracing Active Learning: a case study.* In: Berend, G. – Gosztolya, G. – Vincze, V. (eds.): XX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged, Magyarország, pp. 145–158. https://www.researchgate.net/publication/377730059_Saving_labeling_cost_by_embracing_Active_Learning_a_case_study

Vaswani, A. – Shazeer, N. – Parmar, N. – Uszkoreit, J. – Jones, L. – Gomez, A.N. et al. (2017): *Attention is All You Need.* Advances in Neural Information Processing Systems 30. Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Letöltés ideje: 2024. december 4.

Xia, Y. – Mukherjee, S. – Xie, Z. – Wu, J. – Li, X. – Aponte, R. et al. (2025): *From Selection to Generation: A Survey of LLM-based Active Learning.* In: Che, W. – Nabende, J. – Shutova, E. – Pilehvar, M.T. (eds.): Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers: 14552–14569. https://doi.org/10.18653/v1/2025.acl-long.708

Yang, Z.G. – Dodé, R. – Ferenczi, G. – Héja, E. – Jelencsik-Mátyus, K. – Kőrös, Á. et al. (2023): *Jönnek a Nagyok! BERT-Large, GPT-2, GPT-3 nyelvmodellek magyar nyelvre* (*The Big Ones are Coming! BERT-Large, GPT-2, GPT-3 language models for Hungarian*). In: 19. Magyar Számítógépes Nyelvészeti Konferencia (19th Hungarian Computational Linguistics Conference), Szegedi Tudományegyetem, Szeged, pp. 247–262. https://acta.bibl.u-szeged.hu/78417/. Letöltés ideje: 2024. december 12.

Yang, Z.G. – Laki, L.J. (2021): *Improving Performance of Sentence-level Sentiment Analysis with Data Augmentation Methods.* In: IEEE (ed.): 12th International Conference on Cognitive Infocommunications (CogInfoCom 2021): Proceedings. Institute of Electrical and Electronics Engineers (IEEE), pp. 417–422.

Yang, Z.G. – Váradi, T. (2023): *Training Experimental Language Models with Low Resources, for the Hungarian Language*. Acta Polytechnica Hungarica, 20(5): 169–188. https://doi.org/10.12700/APH.20.5.2023.5.11

Zhang, Z. – Strubell, E. – Hovy, E. (2022): *A Survey of Active Learning for Natural Language Processing.* In: Goldberg, Y. – Kozareva, Z. – Zhang, Y. (eds.): Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 6166–6190. https://doi.org/10.18653/v1/2022.emnlp-main.414

Függelék

5. táblázat
Az ADF- és KPSS-állandóságtesztek eredményei

	Az ADF p-értékei	LM-statisztikák a KPSS-hez	Integráció	Granger-teszt
GDP év/év	0,0095	0,1483	I(0)	standard
Munkanélküliség év/év	0,5325	0,4619	I(1)	Toda–Yamamoto
PMI	0,0000	0,1962	I(0)	standard
Koszinusz AL szentimentindex	0,0042	0,6134	I(1)	Toda–Yamamoto
Bizonytalanságalapú AL szentimentindex	0,0044	0,8021	I(1)	Toda–Yamamoto
Szótáralapú szentimentindex	0,0023	0,2500	I(0)	standard

Megjegyzés: Év/év: éves összehasonlítás
Forrás: A szerzők számításai

6. táblázat
A változópárok optimális késleltetési hossza a bayesi és a Hannan-Quinn-féle információs kritériumok alapján

Változópárok	BIC	HQ	Elemzésbe bevont megfigyelések
GDP év/év – Szótáralapú szentimentindex	1	1	184
GDP év/év – Koszinusz AL szentimentindex	1	1	184
GDP év/év – Bizonytalanságalapú AL szentimentindex	1	1	184
Munkanélküliség év/év – Szótáralapú szentimentindex	1	1	172
Munkanélküliség év/év – Koszinusz AL szentimentindex	1	1	172
Munkanélküliség év/év – Bizonytalanságalapú AL szentimentindex	1	1	172
PMI – Szótáralapú szentimentindex	1	3	184
PMI – Koszinusz AL szentimentindex	1	3	184
PMI – Bizonytalanságalapú AL szentimentindex	1	4	184

Megjegyzés: Év/év: éves összehasonlítás.
Forrás: A szerzők számításai

7. táblázat
DTW-értékek az összes szentimentindex és makrováltozó esetében

	Koszinusz AL szentimentindex	Bizonytalanságalapú AL szentimentindex	Szótáralapú szentimentindex
GDP év/év	10,2385	10,6563	4,4337
Munkanélküliség év/év	15,6824	16,0024	18,099
PMI	8,4366	8,8464	6,8814

Megjegyzés: Év/év: éves összehasonlítás.
Forrás: A szerzők számításai

8. táblázat
A TADL-elemzés eredményei a GDP-, a munkanélküliségi és a PMI-statisztikákra vonatkozóan a koszinusz AL index alkalmazásával

Függő változó: GDP év/év		
Küszöbérték változó: GDP év/év(–1)		
Változó	Együttható	Std. Hiba
GDP év/év(-1) < **–1,74** -- 28 megfigyelés		
C	–5,39***	0,36
KOSZINUSZ AL	–195,65***	34,06
KOSZINUSZ AL(–1)	223,39***	36,10
–1,74 <= GDP év/év(–1) < **2,44** -- 66 megfigyelés		
C	0,55**	0,25
KOSZINUSZ AL	36,10*	20,72
KOSZINUSZ AL(–1)	6,10	21,12
2,44 <= GDP év/év(–1) -- 97 megfigyelés		
C	1,75***	0,47726
KOSZINUSZ AL	62,51***	17,04485
KOSZINUSZ AL(–1)	10,22	18,2724

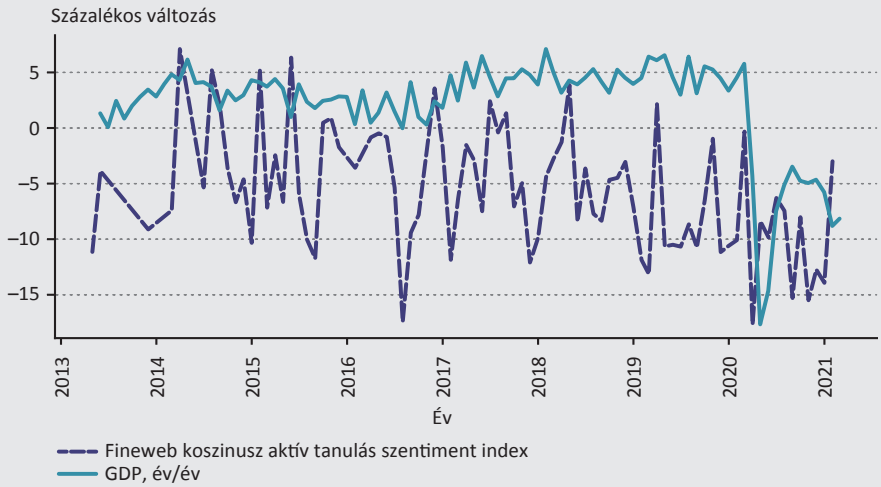
Függő változó: Munkanélküliség		
Küszöbérték változó: Munkanélküliség(–1)		
Változó	Együttható	Std. Hiba
Munkanélküliség(–1) < –10,53 -- 56 megfigyelés		
C	–26,13***	3,12
KOSZINUSZ AL	146,06*	87,13
KOSZINUSZ AL(–3)	131,54	89,80
–10,53 <= Munkanélküliség(–1) < **–2,6** -- 37 megfigyelés		
C	–3,73**	1,80
KOSZINUSZ AL	-138,57	95,55
KOSZINUSZ AL(–3)	43,02	93,84
–2,6 <= Munkanélküliség(–1) < **10** -- 58 megfigyelés		
C	1,81	1,11
KOSZINUSZ AL	–113,62**	55,37
KOSZINUSZ AL(–3)	100,81	62,56
10 <= Munkanélküliség(–1) -- 28 megfigyelés		
C	21,77***	1,52
KOSZINUSZ AL	356,83***	83,44
KOSZINUSZ AL(–3)	–307,90***	78,75

Függő változó: PMI		
Küszöbérték változó: PMI(–1)		
Változó	Együttható	Std. Hiba
PMI(–1) < **49,5** -- 29 megfigyelés		
C	46,87***	0,64
KOSZINUSZ AL	-59,70	53,46
KOSZINUSZ AL(–1)	167,05***	50,45
49,5 <= PMI(–1) -- 162 megfigyelés		
C	51,78***	0,42
KOSZINUSZ AL	70,44***	25,04
KOSZINUSZ AL(–1)	2,58	26,04

*Megjegyzés: Év/év: éves összehasonlítás, *** p < 0,01, ** p < 0,05, * p < 0,1*

Forrás: A szerzők számításai

4. ábra
A Fineweb adatbázisból származó szentimentindex és az előző évhez viszonyított GDP-változás összehasonlítása



Forrás: A szerzők számításai