

# Kacsa vagy nem kacsa? Magyar nyelvű álhírek automatikus azonosítása

Vincze Veronika<sup>1,2</sup>

<sup>1</sup>Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék  
Szeged, Árpád tér 2.

<sup>2</sup>MTA-SZTE Mesterséges Intelligencia Kutatócsoport  
Szeged, Tisza Lajos körút 103.  
vinczev@inf.u-szeged.hu

**Kivonat** Az interneten olvasott hírek nem mindig származnak megbízható forrásokból, illetve nem mindig valósághűek. Cikkünkben gépi tanulási technikák segítségével különítjük el a magyar nyelvű valós híreket az álhírektől. A szakirodalom áttekintése után bemutatjuk a vizsgálati anyagot, majd ismertetjük módszereinket. Az eredmények alapján elmondhatjuk, hogy gazdag nyelvi jellemzőtérre alapuló megoldásunk 80%-os eredményességgel képes a megfelelő csoportba sorolni az ál- és valódi híreket: elsősorban a morfológiai jellemzők bizonyultak leghatásosabbnak a gépi tanulási folyamatban.

**Kulcsszavak:** hírek, álhírek, osztályozás, információkinyerés

## 1. Bevezetés

Manapság a világhálón számos dokumentum születik, melyek száma napról napra gyarapszik. E dokumentumok bárki számára hozzáférhetőek, amennyiben szerzőjük azt szabadon elérhetővé teszi. Különösen a közösségi médiában gyakori jelenség a tartalmak megosztása a többi felhasználóval. Ezzel azonban együtt jár az is, hogy sokszor nem lehetünk biztosak abban, hogy kitől származik az adott bejegyzés, hiszen az interneten sokszor csak felhasználónévvel vagy álnévvel szerepelnek az emberek, emiatt pedig az olvasott információ hitelességét sem vehetjük kétségbe, ha nem megbízható forrásból származik.

Az internetkorszak beköszöntével együtt járó változások egyike, hogy sokan például a világhálóról tájékozódnak nyomtatott sajtó olvasása vagy egy tévéhíradó megtekintése helyett (vagy mellett). Ennek megfelelően a nagyobb napilapok sok esetben saját online portált üzemeltetnek, melyre elektronikus változatban feltöltik a nyomtatott sajtóban megjelent cikkeiket, emellett pedig léteznek online hírportálok is, melyek hírei nem jelennek meg nyomtatásban, kizárólag online olvashatók. A megbízható hírportálok mellett azonban léteznek valótlan tartalmakat közlő oldalak is. Ezek egyik fajtájához tartoznak az úgynevezett hoax oldalak, melyek álhíreket, kétes valóságtartalmú állításokat, reklámokat tartalmaznak. Az oldal szerzőinek célja az olvasók megtévesztése, különféle hamis

információk terjesztése, illetve pénzszerzési tevékenység. A hoax oldalak mellett azonban léteznek olyan álhíreket közlő oldalak is, melyek elsődleges célja az olvasók szórakoztatása, és e tevékenységük nyilvánvaló az olvasó számára is. Cikkünkben ez utóbbi típusba tartozó álhírek automatikus azonosítását tűzzük ki célul: gépi tanulási technikák segítségével különítjük el a valós híreket az álhírektől. A szakirodalom áttekintése után bemutatjuk a vizsgálati anyagot, majd ismertetjük módszereinket. Az eredmények alapján elmondhatjuk, hogy gazdag nyelvi jellemzőtéren alapuló megoldásunk eredményesen képes a megfelelő csoportba sorolni az ál- és valódi híreket.

## 2. Kapcsolódó irodalom

A világban számos munka foglalkozik az álhírek automatikus felismerésével. Például megévesztő termékismertetőket automatikus azonosítását célozza [1] a szentimentelemzés szemszögéből, illetve pletykák felismerésére is született több tanulmány [2,3]. Angol nyelvű kísérletek szerint elsősorban szintaktikai [4] és nyelvi jellemzők [5] segítik az álhírek felismerését, [6] pedig részben felügyelt módszerekről számol be.

Az angol nyelvre több adatbázis is született az utóbbi időben, melyekben álhírek és valódi hírek egyaránt előfordulnak. [7] összesen 221 állítást tartalmaz annak valódisági értékével együtt, míg a Liar korpusz csaknem 13 000 rövid, politikai tartalmú állításból áll [8]. Az Emergent korpuszt 300 állítás és 2500, az állításokhoz kapcsolódó hír alkotja, valódisági értékükkel [9]. Tudomásunk szerint magyar nyelvre azonban ez az első munka, mely álhírek automatikus felismerését célozza.

Szövegek	Szövegek száma	Mondatszám	Tokenszám
Valódi hírek	40	1001	15941
Álhírek	40	719	10376

1. táblázat. A korpusz adatai.

## 3. A korpusz

Vizsgálatunkat egy 80 hírből álló korpuszon végeztük. A korpusz szövegeit az internetről töltöttük le. A valódi hírek több országos és megyei hírportálról származnak (pl. [www.index.hu](http://www.index.hu), [www.origo.hu](http://www.origo.hu), [www.delmagyar.hu](http://www.delmagyar.hu)). Az álhíreket két forrásból gyűjtöttük össze. Egyrészt valós híreket közlő napilapok április elsejei, az olvasók megtréfálására szolgáló álhíreit válogattuk ki, másrészt a [www.hircsarda.hu](http://www.hircsarda.hu) oldalról töltöttünk le cikkeket. A Hírcsárdán megjelenő hírek leginkább az aktuálpolitikai és közéleti események parodizálására épülnek, céljuk vállaltan az olvasóközönség szórakoztatása. A legtöbb hír 2017 szeptemberéből

és októberéből származik, ez alól természetesen kivételt képeznek az április elsejei hírek. A híreket véletlenszerűen válogattuk, arra azonban ügyeltünk, hogy egy adott eseményről csak egy hír szerepeljen a korpuszban.

A korpusz alapvető adatait az 1. táblázat mutatja be.

Az alábbiakban közlünk egy-egy példát valódi és álhírekre is, a szövegek eredeti helyesírását változtatlanul hagyva.

### **Moszkvában készült robbantani az Iszlám Állam<sup>1</sup>**

*Az Iszlám Állam terrorszervezet moszkvai pokolgépes merényletekre készülő alvósejtjét számolta fel az orosz Szövetségi Biztonsági Szolgálat, az FSZB – közölte hétfőn a testület társadalmi kapcsolatokkal foglalkozó osztálya.*

*A hattagú csoport forgalmas helyeken és tömegközlekedési járműveken tervezett robbantásokat. Az FSZB vasárnapi elfogási akciója során két nagy hatóerejű, házi készítésű pokolgépet, robbanószerkezetek előállításához szükséges anyagokat, két pisztolyt, két kézigránátot és kommunikációs eszközöket foglalt le a terroristáktól. Az FSZB szerint a külföldről irányított, észak-kaukázusi származású tagokból álló csoport egy világméretű kalifátus létrehozásának ideológiáját követte.*

### **Zuckerberg megvette Magyarországot<sup>2</sup>**

*„Fasza kis startupnak tűnik” – kommentálta országunkat Budapest romkocsmáit látogatva a Facebook vezére. Magyarország ezentúl Facebook appként működik tovább.*

*Lapunknak sikerült kiderítenie a váratlanul Budapesten feltűnő Mark Zuckerberg látogatásának célját. Ez pedig nem más, mint Magyarország akvizíciója, amivel a Facebook portfólióját szeretné szélesíteni a világ egyik legbefolyásosabb embere. „Ha fényképezni akarok ott az Instagram, a térképi infók végett pedig meglesz a Waze. Most pedig veszek egy országot. LOL” – indokolta a vásárlást a Facebook vezetője.*

*Elemzők szerint a vételre a kutyája beszélhette rá, aki szintén magyar származású és puli. Úgy tudjuk, hogy Magyarország Magyarország ezentúl Facebook appként működik tovább. A jelenlegi vezetést valószínűleg lecserélik és hípszter dizájnerekkel, programozókkal, egy kávéautomatával és Széles Gáborral töltik fel az igaztatótanácsot. „Yeah, Gábor is lots of fun. Különösen tetszik a perpetuum mobile alapú golfpálya koncepciója. Zseniális. Nekem jó ideje ő a háttérképem” – mondta el Zuckerberg munkatársunknak, aki kutyaeledelnek álcázva jutott a közelébe.*

*Hogy pontosan mi lehet a célja az országunkkal az egyelőre nem ismert, de valószínűleg hamarosan mobilra is optimalizálják, megjelennek rajta reklámok és bárki örökbe fogadhat rajta kidobott kismacskákat.*

<sup>1</sup> <http://www.origo.hu/nagyvilag/20171002-moszkvaban-akart-robbantani-az-islam-allam.html>

<sup>2</sup> [http://hircsarda.hu/2013/05/29/zuckerberg\\_megvette\\_magyarorszagot](http://hircsarda.hu/2013/05/29/zuckerberg_megvette_magyarorszagot)

## 4. Gépi tanulási kísérletek

Kísérleteinkben szeretnénk volna megvizsgálni, mennyire hatékonyan különíthetők el automatikusan az álhírek a valósi hírektől. Ehhez első lépésként a magyarlanc [10] elemzővel automatikusan előelemeztük a szövegeket, hozzájutva így azok szótövesített változatához, illetve morfológiai és szintaktikai elemzéséhez. A nyelvi elemzésre építve számos nyelvészeti alapú jellemzőt használtunk fel a kísérletben, figyelembe véve, hogy az álhírek és valódi hírek megkülönböztetése elsősorban szemantikai-pragmatikai kérdés, hiszen egy jól megírt álhír nem feltétlenül különbözik morfológiai és szintaktikai felépítésében a valódi hírektől. Ennek alátámasztására statisztikai szignifikanciavizsgálatokat is végzünk az egyes jellemzőkre nézve (l. 5. rész).

A felhasznált jellemzőket az alábbiakban részletezzük.

- **Statisztikai jellemzők:**
  - mondatok száma,
  - szavak száma,
  - mondatok átlagos hossza.
- **Morfológiai jellemzők:**
  - főnevek, igék, melléknevek, határozószavak, tulajdonnevek és ismeretlen szavak száma és aránya a szószámhoz képest,
  - múlt és jelen idejű igék száma és aránya az igék számához képest,
  - feltételes és felszólító módú igék száma és aránya az igék számához képest,
  - E/1. és T/1. igék száma és aránya az igék számához képest,
  - felsőfokú és középfokú melléknevek száma és aránya a melléknevek számához képest.
- **Szintaktikai jellemzők:**
  - alanyok, tárgyak, jelzők, határozók, alárendelő és mellérendelő mondatok száma és aránya.
- **Szemantikai jellemzők:**
  - pozitív és negatív jelentésű szavak száma és aránya a szószámhoz képest (a listákat [11] alapján készítettük el),
  - bizonytalanságra utaló szavak száma és aránya a szószámhoz képest ([12] alapján felállított osztályokba sorolva), érzelmekre utaló szavak száma és aránya a szószámhoz képest ([13] alapján felállított osztályokba sorolva).

A szemantikai jellemzők vizsgálatakor egyszerű listaillesztéses módszert használtunk: amennyiben az egyes szavak lemmája megegyezett bármelyik lista-elemmel, akkor találatként számoltuk.
- **Pragmatikai jellemzők:**
  - beszédaktust jelentő igék száma és aránya,
  - szó szerinti idézetek és nyilatkozatok száma és aránya.

Utóbbiakat a szövegben előforduló idézőjelek és mondatkezdő gondolatjelek számával mértük.

A feladatot bináris osztályozásként kezeltük, modellként pedig döntési fát [14] alkalmaztunk tízszeres keresztvalidációval. Alapmódszerként a többségi osztályozást választottuk, mely 50%-os pontosságot ért el.

## 5. Eredmények

Gépi tanuló rendszerünk 80%-os pontosságot ért el, azaz 80 dokumentumból 64-et sorolt be a megfelelő osztályba. A részletesebb eredmények a 2. táblázatban láthatók. Mint látható, az alapmódszernél jóval hatékonyabb eredményeket sikerült elérni. A számokból az is kiderül, hogy a rendszer jobban teljesít a valódi hírek felismerésében, míg az álhírek közül többet minősített valódi hírek, 40 esetből 9-szer tévedve.

Szövegek	Pontosság	Fedés	F-mérték
Valódi hírek	0,786	0,825	0,805
Álhírek	0,816	0,775	0,795
Összesen	0,801	0,800	0,800

2. táblázat. Eredmények.

Megvizsgáltuk az egyes jellemzőcsoportok hatékonyságát is porlasztásos módszerrel. A 3. táblázatban látottak szerint a leghatékonyabbnak a morfológiai jellemzők bizonyultak, azonban a szemantikai és szintaktikai jellemzők is pozitívan járulnak hozzá az eredményekhez.

Jellemzők	Szöveg					Különbségek			
		P	R	F	Acc	P	R	F	Acc
összes	álhírek	0,816	0,75	0,795					
	valódi hírek	0,786	0,825	0,805					
	összes	0,801	0,8	0,8	80				
morfológia nélkül	álhírek	0,488	0,5	0,494		-0,328	-0,25	-0,301	
	valódi hírek	0,487	0,475	0,481		-0,299	-0,35	-0,324	
	összes	0,487	0,488	0,487	48,75	-0,314	-0,312	-0,313	-31,25
szintaxis nélkül	álhírek	0,78	0,8	0,79		-0,036	0,05	-0,005	
	valódi hírek	0,795	0,775	0,785		0,009	-0,05	-0,02	
	összes	0,788	0,788	0,788	78,75	-0,013	-0,012	-0,012	-1,25
szemantika nélkül	álhírek	0,769	0,75	0,759		-0,047	0	-0,036	
	valódi hírek	0,756	0,775	0,765		-0,03	-0,05	-0,04	
	összes	0,763	0,763	0,763	76,25	-0,038	-0,037	-0,037	-3,75

3. táblázat. Porlasztásos eredmények. P: pontosság (precision), R: fedés (recall), F: F-mérték (F-measure), Acc: pontosság (accuracy).

Egyes jellemzők statisztikailag szignifikáns különbséget mutatnak a két csoport között, melyeket a 4. táblázatban mutatunk be részletesebben. Eszerint tehát a morfológiai jellemzők közül a határozószavak száma, illetve az ismeretlen szavak, múlt és jelen idejű igék és T/1. igék aránya, a szintaktikai jellemzők közül az alanyok és mellérendelések száma, a tárgyak aránya és a jelzők és alárendelések száma és aránya, míg a szemantikai jellemzők közül az episztemikus

és weasel típusú bizonytalanságot jelölő szavak száma (és aránya), illetve a bánatot kifejező szavak száma térnek el szignifikánsan a valódi hírek és az álhírek csoportjai között.

Jellemző	p-érték
<b>Morfológiai jellemzők</b>	
Határozószavak száma	0,0146
Ismeretlen szavak aránya	0,0001
Múlt idejű igék aránya	0,0215
Jelen idejű igék aránya	0,0184
T/1. igék aránya	0,0078
<b>Szintaktikai jellemzők</b>	
Alanyok száma	0,036
Tárgyak aránya	0,0362
Jelzők és alárendelések száma	0,0344
Jelzők és alárendelések aránya	0,0071
Mellérendelések száma	0,0459
<b>Szemantikai jellemzők</b>	
Episztemikus bizonytalanságot jelző kifejezések száma	0,0051
Episztemikus bizonytalanságot jelző kifejezések aránya	0,0183
Weasel típusú bizonytalanságot jelző kifejezések száma	0,0398
Bánatot kifejező szavak száma	0,027

4. táblázat. Szignifikáns különbségeket mutató jellemzők.

## 6. Az eredmények megvitatása

Az eredmények vizsgálata alapján elmondhatjuk, hogy gépi tanulási kísérleteink 80%-os pontossággal különítik el egymástól a valódi és álhíreket. Az is egyértelműen látszik, hogy elsődlegesen a morfológiai jellemzők bizonyultak hasznosnak, ám a szintaktikai és szemantikai jellemzők hozzáadott értéke is kiderül a 3. táblázatból. A statisztikai és pragmatikai jellemzők azonban nem jelentettek számszerűsíthető többletet a mérések során.

A fentiek alapján azt is kijelenthetjük, hogy elvárásunk – miszerint elsősorban szemantikai-pragmatikai különbségeket találunk az álhírek és valódi hírek között – nem teljesen igazolódott be. Egyrészt a morfológiai jellemzők bizonyultak leghatásosabbnak a gépi tanulási folyamatban, másrészt a statisztikailag szignifikáns jellemzők sem mutatják a szemantikai-pragmatikai jellemzők kiemelkedő szerepét. Eredményeinkből úgy tűnik, hogy a korpuszunkban található álhírek szókészlete és mondat szerkesztése is erősen eltér a valódi hírekétől: azoknál több határozószót, több jelen idejű igét és több T/1. igét használ. A mellé- és alárendelések száma azonban alacsonyabb, ami arra utal, hogy kevesebb összetett mondat található az álhírekben. Ugyanakkor a jelenre való utalások, a többes szám

első személyű igék gyakori használata is a közösségi tudatot erősíti az álhírekben, ami akár manipulatív stratégia is lehet [15], az olvasó meggyőzését célozva. Továbbá, több bizonytalanságra utaló kifejezést találhatunk a valódi hírekben, azaz az álhírekben inkább tényszerű állítások jelennek meg, ezzel is „tényként” beállítva a közölt tartalmakat. Érzelmek terén pedig különösen a bánatra és a haragra jellemző szavak használata gyakoribb a valódi hírekben, ellenben a szeretet kifejeződése az álhírekben gyakoribb, azaz a pozitív szemlélet talán jobban az álhírek sajátja.

A rendszer által tévesen osztályozott hírekre mutatunk néhány példát az alábbiakban.

Tévesnek ítélt valódi hírek:

- *Időkapszulát ástak el a Spitzbergákon*
- *A trópusi erdők több szemet bocsátanak ki, mint az amerikai gépjárműpark*
- *Észak-Korea nehézbombázókat röptet az óceán felett*

Valódinak címkézett álhírek:

- *Itt a kézműves iPhone*
- *Közmunkásokat képeznek át macskává*
- *Tarr Béla rendezői a Trónok harca 7. évadját*

A hibásan kategorizált cikkeket áttekintve megállapíthatjuk, hogy elsősorban a dokumentumok hosszúsága játszott szerepet a téves osztályozásban. A rövidebb valódi híreket gyakran álhírnek, míg a hosszabb álhíreket valódinak minősítette a rendszer. Mindemellett minél több bizonytalanságra utaló nyelvi elem volt megtalálható az álhírben, annál valószínűbb, hogy a rendszer valódi híreknek tekinti.

## 7. Összegzés

Ebben a munkában magyar nyelvű hírek és álhírek automatikus elkülönítéséről számoltunk be. Eredményeink alátámasztják, hogy gazdag – morfológiai, szintaktikai, szemantikai és pragmatikai – jellemzőtérre építve sikeresen lehetséges felismerni az álhíreket, különösen a morfológiai jellemzők játszanak fontos szerepet a folyamatban.

Jelen cikkben olyan álhírekkel dolgoztunk, melyek elsődleges célja az olvasó szórakoztatása volt, nem pedig megtévesztése. A jövőben azonban hoax típusú hírekre is szeretnénk kiterjeszteni gépi tanuló megközelítésünket, azaz az olvasót szándékosan megtévesztő, illetve manipuláló információk automatikus felismerése is célunk.

## Köszönetnyilvánítás

A tanulmány az Emberi Erőforrások Minisztériuma UNKP-17-4 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült.

## Hivatkozások

1. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, Association for Computational Linguistics (2011) 309–319
2. Ma, J., Gao, W., Wong, K.F.: Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, Association for Computational Linguistics (2017) 708–717
3. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. IJCAI'16, AAAI Press (2016) 3818–3824
4. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju Island, Korea, Association for Computational Linguistics (2012) 171–175
5. Pérez-Rosas, V., Mihalcea, R.: Experiments in open domain deception detection. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Association for Computational Linguistics (2015) 1120–1125
6. Hai, Z., Zhao, P., Cheng, P., Yang, P., Li, X.L., Li, G.: Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Association for Computational Linguistics (2016) 1817–1826
7. Vlachos, A., Riedel, S.: Fact checking: Task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, Association for Computational Linguistics (2014) 18–22
8. Wang, W.Y.: "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada, Association for Computational Linguistics (2017) 422–426
9. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, Association for Computational Linguistics (2016) 1163–1168
10. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771
11. Szabó, M.K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. In: Segédkönyvek a nyelvészet tanulmányozásához 177. Tinta, Budapest (2015) 278–285
12. Vincze, V.: Uncertainty detection in Hungarian texts. In: Proceedings of Coling 2014. (2014)
13. Szabó, M.K., Vincze, V., Morvay, G.: Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: Távtatok a mai magyar alkalmazott nyelvészetben. Tinta, Budapest (2016)
14. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA (1993)



15. Bárházi, E.: Manipuláció, valamint manipulációra alkalmas nyelvhasználati eszközök a sajtóreklámokban. *Magyar Nyelv* 104(4) (2008) 443–463