JEL-KÉP

KOMMUNIKÁCIÓ, KÖZVÉLEMÉNY, MÉDIA

2017/1. szám

HÍRADATBÁZISOK ÉS MÉDIATARTALMAK ELEMZÉSE BIG DATA HASZNÁLATÁVAL

Szűts Zoltán

szutszoltan@gmail.com

DOI: 10.20520/JEL-KEP.2017.1.23

Absztrakt

A Big Data terminus az olyan nagy és bonyolult adathalmazokra vonatkozik, amelyek nem kezelhetők a hagyományos adatfeldolgozó és tartalomelemző eljárásokkal. Tanulmányunk azt mutatja be, hogy fejlett Big Data elemző módszerek képesek értékes információkat kinyerni a média hírek és a közösségi media üzenetek végtelen folyamából, és e módszerek használatával a társadalmi, gazdasági, kulturális folyamatok kutatói rejtett összefüggésekre találhatnak és értékes ismereteket gyűjthetnek. Tanulmányunkban először a globális konfliktusoknak a GDELT (Global Database of Events, Language, and Tone) adatbázison végzett elemzését vesszük szemügyre, majd a Google Trends and News rendszerét vizsgáljuk, végül a Twitter feltöltések Sentiment Viz rendszerrel történő elemzését mutatjuk be. A tanulmány az online Big Data elemzésekkel kapcsolatos kihivások és kérdések rövid áttekintésével zárul.

Kulcsszavak

tartalomelemzés, Big Data, közösségi média, GDELT, Google, Twitter, sentiment analysis

BIG DATA ANALYSIS OF NEWS DATABASISES AND MEDIA CONTENTS

Zoltán Szűts

Abstract

The term Big Data refers to data sets that are so large or complex that traditional data processing treatments are inadequate to deal with them. Our paper demonstrates that advanced Big Data analytic methods are able to extract valuable information from the endless stream of media news and social media messages. By using these methods, researchers of social, economic and cultural processes can reveal hidden patterns and bring valuable new insights. First we examine the Global Material Conflict Report system introduced within the framework of GDELT (Global Database of Events, Language, and Tone), then we explore the Google Trends and News system, and finally we present a Sentiment Viz analysis of Twitter feeds. The paper ends with a short overview of the most important critiques, challenges and questions of Big Data online media content analysis.

Keywords

content analysis, Big Data, social media, GDELT, Google, Twitter, sentiment analysis

JEL-KÉP

KOMMUNIKÁCIÓ, KÖZVÉLEMÉNY, MÉDIA

2017/1. szám

HÍRADATBÁZISOK ÉS MÉDIATARTALMAK ELEMZÉSE BIG DATA HASZNÁLATÁVAL[1]

Szűts Zoltán

Bevezetés

Információs társadalmunkban jelentősen megnőtt az információ értéke és ezzel együtt a tartalomelemzés fontossága, valamint az így kapott eredmények, következtetések jelentősége is. A hálózatok, a digitalizáció, a tárolási és számítási kapacitás növekedésének következményeként létrejött Big Data egy, a korábbiaknál szélesebb skálán történő médiatartalom elemzések ígéretét hordozza magában. A Big Data elnevezés azt az óriási információ mennyiséget és feldolgozását írja le, melyet a felhasználók és a hálózatba kötött digitális eszközeik generálnak, és a bárki számára elérhető számítógépek elemeznek.

A digitálisan rögzített tartalmak gépi eszközökkel könnyen kereshetők, címkézhetők, feltérképezhetők, de ezzel együtt az online médiatartalmak mennyisége és típusa folyamatosan növekszik, így mind nagyobb méretű adatmennyiséget kell feldolgoznunk.

Tanulmányunkban jó gyakorlatként online médiatartalmak – elsősorban hírek – elemzését mutatjuk a Big Data eszközeivel. Egy olyan új informatikai paradigmával állunk szemben, mely a tárhely és számítási teljesítmény növekedésének és költségcsökkenésének köszönhetően az adatok eddig nem látott volumenének valós időben történő elemzését és korábban nem is feltételezett összefüggések kimutatását biztosítja.

Ezen értelmezés szerint az online portálok, a közösségi oldalak (Facebook, Twitter, LinkedIn), az online fórumok, blogok és wikik nem csupán kommunikációs eszközként és publikációs felületként értelmezendők, de lehetővé teszik a média, a gazdaság és a kormányzati szervezetek, a kutatók, sőt az egyének számára is, hogy az ezen felületeken elérhető adatokkal Big Data elemzéseket végezzenek. Ezen elemzési folyamatnak azonban számos nehézséggel is szembe kell néznie. Az egyik ilyen kihívás (a Facebook esetében) az adatokhoz való hozzáférés korlátozása a felületet működtetők által. A másik nehézség épp a Big Data természetéből adódik, hiszen egyszerre kell valós időben különböző típusú strukturált és strukturálatlan adatot, képet, hangot, videót, adatsorokat, szöveget egységes formátumba hozni és elemezni. Éppen ezért tanulmányunk elemzéseiben az átláthatóság kedvéért kizárólag szöveges tartalmakra fókuszálunk.

Rögtön a munkánk elején fontosnak tartjuk kijelenteni, hogy az online médiatartalmakat két nagy, viszonylag élesen szétválasztható kategóriára osztjuk: egyfelől a hagyományos, professzionális kapuőri (szerkesztői) környezetben, másfelől a közösség, a mindennapi fel-

[1] A kutatást és a cikk írását a Nemzeti Média- és Hírközlési Hatóság Médiatanács Médiatudományi Intézete támogatta.

használók által létrehozott Web 2.0-ás tartalmakra. A hagyományos, kapuőr rendszerű médiacentrumok által online publikált hírek és vélemények mellett mára ugyanis a közösségi média is fontos információforrássá vált. Ezen felosztás mentén hatjuk végre a GDELT (Global Database of Events, Language, and Tone), illetve a Twitter tartalmainak Big Data elemzését is.

A tanulmány szerkezetileg az elméletek ismertetésével kezdődik, esettanulmányokkal és elemzésekkel folytatódik, végezetül pedig több fontos kérdéskörre hívja fel a figyelmet: az adatokhoz való hozzáférés és feldolgozás problémáira, illetve a módszer lehetséges torzításaira, hibáira is.

Tartalom a digitális korban

A technikai fejlődés lehetővé tette a digitálisan rögzített információk összegyűjtését, kiaknázását, rendszerezését. A széles skálájú tartalomelemzés korábban szuperszámítógépeket igényelt. A jelenben a számítási teljesítmény költségének csökkenésével a mindennapi infotechnológiai eszközökkel rendelkező felhasználók számára is biztosított ez a lehetőség (Manovich 2012). A Big Data korában a számítógépek már nem a segédeszközök szerepét töltik be, hanem önállóan látnak el feladatokat, a mesterséges intelligenciának köszönve pedig maguktól vesznek észre összefüggéseket és mintázatokat. Bár az információ mennyiség ilyen mértéke és a számítógépek mind nagyobb szerepe látszólag háttérbe szorítja az emberközpontú tartalomelemzést, az új paradigma alapja továbbra is egy ember-gép együttműködésre alapozó megközelítés. Tanulmányunkban amellett törünk lándzsát, hogy a folyamatos emberi ellenőrzés és finomhangolás melletti Big Data elemzése a leghatékonyabb módja a magy mennyiségű tartalom elemzésének az információs társadalom korában.

A közösségi médiában folyamatosan publikált adatok gyűjtése és elemzése magában hordozza az összefüggések általános és gyorsabb felismerésének ígéretét. Az adatokhoz való hozzáférés elemzés céljából alapvetően szabályozott, a platform tulajdonosának monopolát képezi, hiszen az információ, mitőbb, az abból szerzett összefüggések marketing és kereskedelmi célokra hasznosíthatók. A Twitter mikroblog platformja és a blogok általában a jelenben ez alól még kivételt képeznek, mivel az ott posztolt adatok külső szemlélők számára is hozzáférhetők, kutathatók és elemezhetők. A Facebook azonban már más üzleti modell alapján működik. És bár számos összefüggés figyelhető meg a legnagyobb közösségi oldalon posztolt információk, a felhasználók megosztási, internetes keresési, sőt vásárlási és közlekedési szokásai között is, a széles körű Big Data elemzés ezen környezetben az oldal tulajdonosának a joga. (Vagy azoké, akik ezt a jogot megvásárolják.) A Twitteren posztolt és megosztott hírek esetében a Big Data alkalmazásával például könnyen meg lehet állapítani a preferált hírforrások, megtalálni a véleményvezéreket, a megosztásokat és az interakciót elemezve.

A digitalizáció és hálózatosodás hozta változások: kereshetőség és összekapcsoltság

Az internetről szóló diskurzusban gyakran elhangzik az állítás, miszerint a globális hálózat megváltoztatta az információgyűjtés, -feldolgozás, -tárolás és -továbbítás módját. Úgy lett tervezve, hogy az adatokat digitalizált formában továbbítsa, alapvetően függetlenül a tartalom jellegétől (Szűts 2013, Benedek és Molnár 2014, Molnár 2015), ezzel összeolvasztva az eddig párhuzamosan létező médiumok tulajdonságait egy digitális csatornává (Szűts és Yoo 2013).

A kereshetőséget és összekapcsolhatóságot az elemzések során segíti a címkézés (tagging) is. Ennek során alapvetően egy szöveg, de akár egy kép, videó vagy hanganyag is meta adatokkal – címkékkel – lesz ellátva. A címkék gyakran egy alapszintű elemző algoritmus segítségével címkefelhőt alkotnak, az így kapott vizuális ábrázolásban legreprezentáltabb és így legfontosabb kifejezések a középkori ikonok univerzumából ismert szemantikus perspektíva felfogáshoz hasonlóan méretben is nagyobbak (Szűts 2012).

Számadatok

Percenként 72 órányi, naponta pedig 864 ezer órányi videó anyagot töltenek fel a felhasználók a YouTube-ra. Több mint 1,4 milliárd aktív felhasználó látogatja naponta a közösségi oldalakat, és csak a Facebookon összességében napi 4,7 milliárd percet töltenek olvasással, böngészéssel, posztolással. Napi 532 millió Facebook poszt születik, 250 millió képet töltenek fel és 30 milliárd interakció zajlik. Emellett naponta 50 millió bejegyzés jelenik meg a Twitteren (Stone 2014). Természetesen ezen adatmennyiség nagy része nem híranyag, de a töredéke is olyan nagy kvantumszámot jelent, hogy az ellenáll a hagyományos elemzési technikáknak.

A Big Data paradigmája

Az ilyen nagy mennyiségű információt már nem lehet hagyományos tartalomelemzési módszerekkel feldolgozni, hanem a Big Data paradigmarendszerében kell értelmezni. A Big Data különféle tulajdonságai, így a mennyiség (volume), a sebesség (velocity) és a változatosság (variety), illetve az adatok komplexitása számos kihívás elé állít minket, de ugyanakkor korábban nem látott összefüggések feltérképezését is ígéri.

Ahhoz, hogy az online médiatartalmak Big Data eszközeivel történő elemzését tárgyalhassuk (a tanulmány terjedelmi korlátai miatt csak rövid formában), meg kell ismerkednünk a „Nagy Adat” jelenségével. A Big Data magába foglalja a korábban soha nem látott mértékű és szinte végeláthatatlan forrásból érkező adatok rögzítését, feldolgozását, elemzését, megosztását, az összefüggések feltárását és az eredmények átlátható bemutatását. A Big Data vonzáskörébe tartozó információ mértéke többszörösen meghaladja a korábban használt adatrögzítő és feldolgozó szoftverek és számítógépek képességeit, manapság azonban már a mindennapi felhasználók számára is elérhető a technológia. A Big Data trendje mindinkább érinti a médiát is, mivel az új technológiák automatizálják és leegyszerűsítik a tartalomelemzést.

Laney (2012) szerint a Big Data-t három *V* határozza meg: a mennyiség (*V*olume), a sebesség (*V*elocity) és a változatosság (*V*ariety).

- A mennyiség a másodpercenként előállított hatalmas adatözönt jelzi. A múltban az ilyen nagymennyiségű adat még tárolási problémákat okozott. A jelenben a tárhely mérete és a tárolási és hálózati sebesség is növekedett, miközben költségük csökkent.
- A sebesség lényeges tulajdonság, hiszen az adatok nem nagy blokkokban jönnek, hanem folyamatosan, kisebb-nagyobb intenzitással, valós időben áramolnak. A hatékonyság értelmében lehetőleg valós időben kell ezen adatokat feldolgoznunk.
- A legnagyobb kihívást az adatok változatossága jelenti, hiszen az adatok eltérő formátumban érkeznek, ezeket kell strukturálni. A cél az adatfolyamok formázása az értékes információk kinyeréséhez (https://www.it-services.hu/hirek/mi-az-a-big-data). A változatosságra jellemző, hogy egyszerre érkeznek adatok a szöveges dokumentumok mellett hagyományos adatbázisokból, videó megfigyelő rendszerekből, e-mailekből, tömegközlekedési járművekből, repülőgépek motorjaiból, telefonhívásokból, és az elemző rendszereknek összefüggéseket kell felismerniük (Majkić 2014).

A Big Data egyszerre jelent szemantikai, analitikai, adattárolási és hozzáférési kihívást számunkra, hiszen napjainkban eddig nem látott nagyságrendű adatok tárolását, feldolgozását, a rejtett és váratlan összefüggések megtalálását feltételezi. Minden, ami a hálózat kontextusában születik és történik, megmarad, és ezzel visszakereshetővé, elemezhetővé válik (Csepeli 2015: 172–173).

A Big Data tartalomelemzés kihívásai

A hagyományos tartalomelemzési módszerek már nem elég hatékonyak a digitális tartalmak esetében. A gépi elemzésnél az adatmennyiségen és az időbeli tényezőkön túl egy további komoly problémával szembesülünk, ez a nyelvi kihívás: olyan egységes elemző rendszert kell felépítenünk, amely a gyakorlatban nüánsznyi különbségeket is képes felismerni.

Jelenleg már létezik több olyan magyar gépi tartalomelemző rendszer, mely a nemzetközi médiamegjelenésekben valós időben nyomon követi az általunk kívánt témákat és magyar nyelvű hírösszefoglalókat biztosít lényegében gépi fordítás alapján. Ezek a rendszerek azonban nem a Big Data eszköztárát használják, hiszen a gépi fordítás folyamata hasonlít a Google fordítójához, a címszavas keresés pedig a Google keresőjéhez. Bár a hagyományos módszerekhez képest előrelépésről van szó, fontos kiemelni, hogy a gép nem veszi észre a mintákat, nem igazán nagy adatmennyiséggel dolgozik, nem elég összetett, csak adott mennyiségű, előre beállított címszót figyel, míg egy Big Data rendszer folyamatosan tanul, folyamatosan új címszavakat vesz fel. Az említett tartalomelemző rendszerek először egy durva fordítást végeznek el, és amennyiben ennek alapján az ügyfél kéri, akkor valós, hús-vér fordító ülteti át magyarra a szöveget, ami költséges és időigényes folyamat.

Az ikonikus üzenetek és az üzenetek konnotatív szintjeinek felismerése a Big Data 3. V-jének köszönve működik, és nincs az elemzők részleges (nyelvi, képi) tudására korlátozva – nehézséget okoz az üzenet kontextusának értelmezése, de a nagyszámú minta alapján a rendszer képes felismerni. Ha a kép jelentését a gép szöveges formában rögzíti, akkor már lehetséges az összehasonlítás. A Big Data környezetében a gép a megfelelő mennyiségű tanított példa alapján felismeri a szimbolikus jelentést, illetve azt, hogy mikor kell használni.

Nehézséget okoz azonban, hogy míg az olvasók a különböző lexikális és kulturális háttértudások miatt különbözőképpen érthetik a szofisztikált üzeneteket, a gépek erre nem képesek. A legnagyobb kihívás, hogy a számítógép még képtelen felfogni az emberi nyelveket a teljes gazdagságukban, ahogy azt az emberi elme teszi. Ez a probléma önmagában is előrevetíti, hogy a leghatékonyabb Big Data elemzési rendszerek azok lesznek, melyekben szerepel az emberi értékelés és jelentést kiválasztó döntés, aminek alapján a gép képes tanulni és mind pontosabb elemzést biztosítani.

A Big Data elemzés során a következő kérdésekre kell válaszolnunk:

- mi történik, ha olyan komplex és nagymennyiségű információt kell feldolgoznunk, hogy nem alkalmazhatjuk az eddigi módszereket?
- szükség van-e valamennyi adat tárolására?
- valamennyi adatot szükséges-e elemezni?
- hogyan lehet meghatározni a legfontosabb információkat?
- milyen formában kell prezentálni az információkat?

A Big Data megjelenése előtt a teljes adatmennyiség vizsgálata lehetetlen volt, hiszen nem lehetett valamennyi cikket és tévéműsort bevenni a mintába, ezért két módszer közül választottak. Vagy véletlenszerű mintavételt kellett alkalmazni, hogy csökkentsük a vizsgált információ mennyiséget, vagy nagyszámú elemzőt kellett megbízni a munkával. Ezzel szemben a Big Data valamennyi adatot feldolgoz a gépi tanulásnak köszönhetően. Nem kell tehát többé az adatok mennyisége és mélysége közül választanunk (Manovich 2012: 466). Hasonlóképpen könnyebbé, és a kutatók számára átláthatóbbá vált az eredmények vizualizálása, mint az a Twitter médiatartalmainak elemzését tárgyaló fejezetben látni fogjuk.

A módszer hátrányai, hogy nincs lehetőség a nagy minta esetében a mintázatok, jelenségek okainak valamennyi esetben történő feltárására, ezért a gépeknek ismerniük kell a kon-

textus minél több tulajdonságát, hogy kísérletet tegyenek arra, ami az emberi elemzők számára nem okoz nagy nehézséget. Éppen ezért fontos szerepe van kategóriák alkotásának és az emberi elemzőknek is.

A sentiment analysis mint tartalomelemző módszer

A *sentiment analysis* a vélemények, érzések (érzelmek), szubjektív megnyilvánulások gépi elemzését biztosító módszer, vélemény bányászat. Tanulmányunkban a nemzetközi szakirodalomra támaszkodva a módszer elnevezésére egységesen a sentiment analysis kifejezést használjuk.

A sentiment analysis tehát számítógépes elemzése az online posztolt véleményeknek, dicséreteknek, kritikáknak, attitűdöknek és érzelmeknek, melyek egy termékkel, szolgáltatással, szervezettel vagy általában véve valamilyen jelenséggel kapcsolatosan fogalmazódnak meg. A módszer szorosan kapcsolódik a Big Datahoz, hiszen segít az összegyűjtött adatok értelmezésében.

A sentiment analysis a nyelvfelismerő rendszerek és az adatbányászat összekapcsolásának eredményeként jelent meg, és a nagymennyiségű információ elemzését segíti automatizált módon. A módszer előnye az is, hogy túllép az emberek szövegértelmezése során megjelenő előismeretein és előítéletein, és azzal a megoldással szemben, amikor például több elemző eltérően értelmez egy-egy mondatot, a gépi sentiment analysis mindig ugyanúgy sorolja be a szöveget (Zhang és Liu 2014: 1).

A sentiment analysis nem előzmények nélküli, hiszen a jó vagy rossz hírek kategóriája és a cikkek ilyen típusú besorolása már régóta ismert. A jelen hálózati környezetében azonban több tényező is nehézséget okozza. Egyrészt eddig nem ismert mennyiségű adatot kell feldolgozni. Emellett az internetes források formában, stílusban, de még helyesírásban is különböznek egymástól, ami már egyetlen nyelv, például az angol esetében is bonyolult elemzési feladatokat jelent. A legkomolyabb feladatot a szavak jelentésének kontextus függő értelmezése adja.

1. példa: *Az énekesnő azt nyilatkozta, hogy megőrül a kiskutyákért, ezért többet is örökbe fogadott.* A *megőrül* kifejezés önmagában negatív szentimentet előfeltételezne. A mondat második fele azonban a *megőrül* szónak pozitív jelentést kölcsönöz.

2. példa: *Az énekesnő azt nyilatkozta, hogy megőrül, ha folyamatosan fotózzák*. A *megőrül* kifejezés önmagában negatív szentimentet előfeltételezne, a mondat második fele azonban érthető magyarázatot is ad arra, hogy miért van negatív jelentése ebben az esetben a *megőrül* szónak.

3. példa*: Ennek a filmnek zseniálisnak kellene lennie. A története izgalmasnak tűnik, a színesek elsőrangúak, és maga a főszereplő, Jack Black is általában remekül játszik. Mégis, az egész nagyon vékony.* Ebben az esetben a gépi tanulásnak köszönve a rendszer felismeri, hogy a sok dicsérő kifejezést (*zseniális, izgalmas, elsőrangú, remekül*) az utolsó mondatban szereplő *nagyon vékony* kifejezés végül teljesen lenullázza, és a bekezdés jelentése így összességében negatív.

A sentiment analysis során hat univerzális érzelmet fedezhetünk fel: harag, undor, félelem, öröm, szomorúság és meglepődés. Egy teljes szöveg (hír) dokumentum szintű sentiment analysis-nél figyelembe kell venni, hogy az több mondatból vagy bekezdésből is állhat, melyek gyakran ellentétes érzelmeket (jelentéseket) hordoznak, és ezek összefüggése határozza meg a végső jelentést (Pang és Lee 2008).

A Big Data elemzés technikai kontextusa: gépi tanulás

2000-től mind nagyobb jelentőséggel bír a gépi tanulás a nyelvfelismerésben, a szövegértésben és az információ feldolgozásban, de használható képek feldolgozásában és felismerésében is. A gépi tanulás (*machine learning – ML*) (Samuel 1959) a Big Data alapjaihoz is hozzátartozik, hiszen lehetővé teszi, hogy az elemzéseket végző számítógépek tanuljanak. Ez a tanulás abból áll, hogy a gépek idővel mintákat vesznek észre anélkül, hogy konkrétan erre programozták volna őket. A gépi tanulásnak köszönhetően az elemzés során a számítógépek közvetlenül az adatokból jutnak ismeretekhez és oldanak meg problémákat. Az esetek többségében természetesen a számítógépeket embereknek kell tanítaniuk. Az adatokat kezdetben kutatóknak és programozóknak kell megcímkéznünk és osztályoznunk. Később azonban, ezen minta alapján, a gépek önállóan is képesek lesznek tanulni és elemezni az információkat. A gépi tanulás alapja, hogy a gépek képesek felismerni objektumokat és az őket leíró tulajdonságokat. Ez a tulajdonság alapvetően az embereket jellemzi. Egy ember általában könnyen meg tudja állapítani, ha a szöveg szarkasztikus. A gép csak akkor ismeri fel a gúnyt, ha előtte számos mintát programoztak belé. Ha azonban olyan tulajdonsággal találkozik a gép, mely nem ismerhető fel a korábbi minták alapján, tapasztalati alapján képes kikövetkeztetni ezt a tulajdonságot, akkor beszélhetünk gépi tanulásról (Shroff 2014: 127).

Példák Big Data elemzésekre

Szerkesztett online tartalom elemzése a Big Data eszközeivel

Első Big Data elemzésünkhöz nem különálló, professzionális szerkesztői környezetben előállított online médiumok tartalmait választottuk, hanem egy jelentős, azonban talán kevésbé ismert hírgyűjtő rendszer adatbázisát vizsgáltuk.

A Global Data on Event, Location and Tone (GDELT, gdelt.project.org) adatbázisban 1979-től gyűjtik a világ valamennyi jelentős hírügynöksége által megjelentett hírét, vagyis a világot érintő geokódolt események adatait. A GDELT több milliárd információs rekordját a kutatók nem tudnák feldolgozni számítógépek nélkül. A Big Data eszköztára tette lehetővé, hogy a korábbi, alapszintű számítógépes elemzések mellett a legbonyolultabb algoritmusokkal is kutathatóvá válhatnak az itt tárolt információk. Így az globalizált társadalmunkat is érintő eredmények és összefüggések valós időben jeleníthetők meg és vizualizálhatók. A bonyolultabb algoritmusok képesek egyes események előrejelzésére is. Az adatbázis maga két adasort tartalmaz. Az egyik 300 kategóriában kódolja az eseményeket, a másik tárolja az érintett személyek, helyszínek, szervezetek adatait, illetve az esemény címszavait és ezzel a kapcsolatokat is.

Az 1979. január 1-én indított adatbázis kezdetben naponta frissült, ez a gyakoriság hamarosan eléri a 15 percenkénti ciklust. A GDELT képessége, hogy a világ vezetői médiumait ilyen széles skálán szemlézi, lehetővé tette annak meghatározást, hogy egyes történésekről hogyan vélekedik a világ. A jelenben GDELT adatbázisa 250 millió hírt tartalmaz. 2013 március 31-e óta a fájlok aszerint vannak archiválva, hogy egy eseményről mikor tudósított a média, és nem az esemény története alapján. Ez a szimbolikus váltás azonban azt is jelezheti, hogy az esemény akkor számít valósnak, akkor történt meg, ha a média tudósított róla. Az események több mint 97%-ról 24 órán belül tudósít a média, de néhányat később említ meg. Egy eseményről egy hír kerül be az adatbázisba. A GDELT alapvető célja katalogizálni világszinten az eseményeket, melyek összekötik a személyeket, helyszíneket, szervezeteket, témákat, hírforrásokat egyetlen hálózatba. A rendszert működtetők meghatározása szerint a projekt célja a teljes bolygó eseményeinek kódolása egy szabadon felhasználható formátumba, hogy könnyebben észrevegyük a világunkat érintő események közti összefüggéseket, az események érzelmi telítettségét (pozitív, negatív, semleges). Így a GDELT Project

lehetővé teszi adatainak szabad használatát tudományos, üzleti vagy kormányzati célokra. A jelenben már a Google Ideas rendszere által támogatott GDELT a világ nyomtatott és elektronikus sajtóját, illetve online portáljait monitorozza több mint 100 nyelven összetett szövegfelismerő és adatbányászó alkalmazásokat használva.

A gyakorlatban az olyan mondatot, mint „Az USA kritizálta Oroszországot, hogy tegnap csapatokat küldött a Krím-félszigetre, ahol az összecsapások során 10 civil megsérült" a rendszer a következő módon tárolja: USA kritizálja Oroszországot, Oroszország csapatokat küldött (Krím), polgári áldozatok (Krím).

Az EVENT Geographic Network Visualizer lehetővé teszi a geotaggelt, földrajzi pontokhoz kötött – térképre tűzött – eredmények gyors vizualizálását (http://analysis.gdeltproject.org/module-event-geonet.html). A felhasználók által megadott kritériumok szerint a rendszer megkeresi az országok (és városok) közti összefüggéseket, és megjeleníti őket a térképen. A keresés alapján generált Google Föld kml kiterjesztésű fájl lehetővé teszi az eredmények betöltetését a Google Föld programba.

Példa: a 2013. április 1-e és 2016. április 1-e időintervallumban az Egyesült Államok és Oroszország közti tiltakozások térképen ábrázolva, ezen csak a konfliktusok jelennek meg (1. ábra)[2].

1. ábra

Az USA és Oroszország közti tiltakozások eseményeken keresztüli megjelenítése

Elemzés az európai migrációról és a témában megjelent hírek kulcsszereplőiről

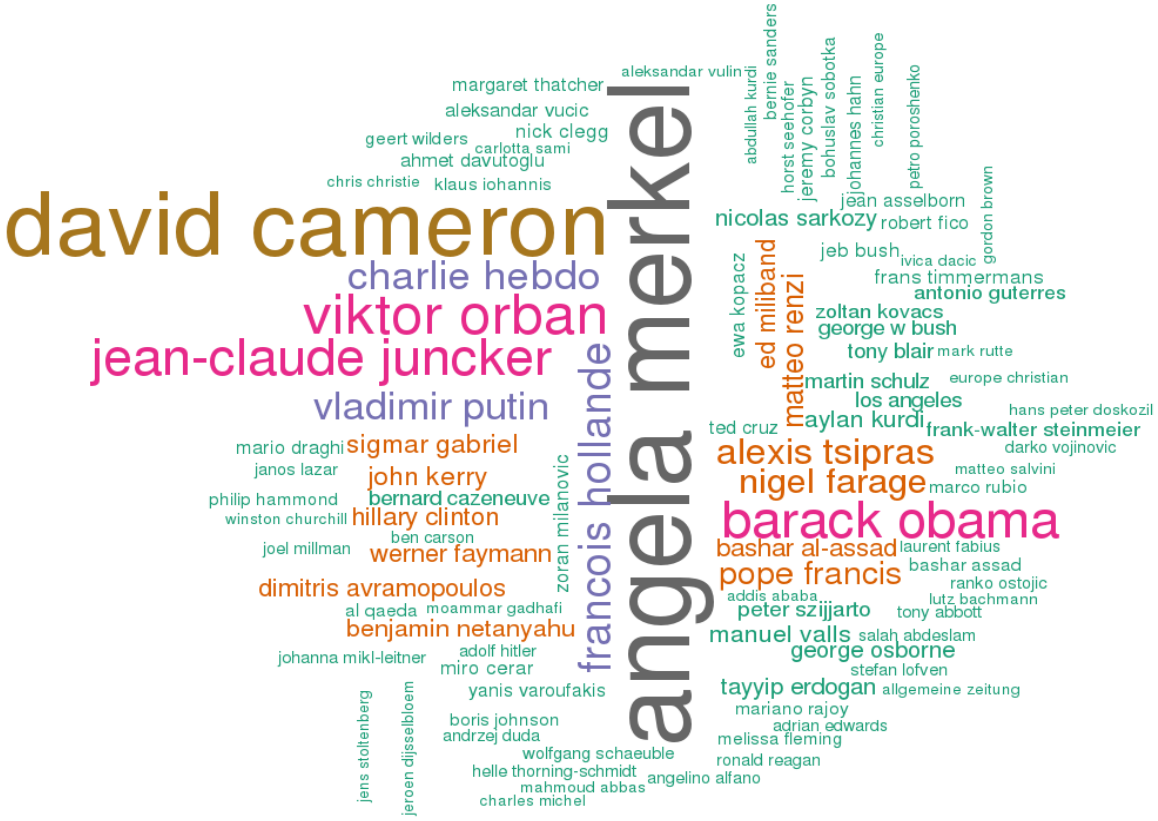
A 2015-ös év legmeghatározóbb nemzetközi eseménye a migráció volt. A témában számtalan hír és publicisztika jelent meg, a migráció témájával kezdődtek világszerte a tévéhíradók. A hagyományos tartalomelemző módszerekkel lehetetlen lenne az összes megjelenést számba venni és ebből következtetéseket levonni. A GDELT Big Data rendszere azonban alkalmas

[2] Paraméterek: Start Date = 04/01/2013, End Date = 04/01/2016, Actor1 Country: USA, Actor1 Type: Actor2 Country: RUS, Actor2 Type: Event Code: 14 (Protest), Event Quad Class:, Event Country: Weighting: NUMEVENTSx

erre. Ha például arra vagyunk kíváncsiak, hogy a 2015-ben az Európát érintő migráció kapcsán mely személyek – vezetők, politikusok – szerepeltek a legtöbbször, azaz kik határozták még a migrációról szóló híreket, akkor a 2. ábrán látható címkefelhőt[3], illetve az 1. táblázat gyakorisági listáját kapjuk.

2. ábra

Az európai migrációval kapcsolatos hírekben megjelenő nevek címkefelhője



1. táblázat

A migráció címszó kapcsán megjelenő nevek

Nevek	Gyakoriság	Százalék
1. angela merkel	9999	5,998128
2. david cameron	7934	4,759391
3. viktor orban	4543	2,725222
4. jean-claude juncker	4061	2,436084
5. barack obama	4022	2,412689
6. francois hollande	3297	1,977781
~~7. charlie hebdo~~	2864	1,718036
8. vladimir putin	2709	1,625055
9. alexis tsipras	2473	1,483486

[3] Paraméterek: Migráció, Európa, Start Date = 01/01/2015, End Date = 12/31/2015, Object Field: PERSON, Object Weight: NAMESETS, Must have ALL these keywords: europe, Must also have AT LEAST ONE of: migration, Must NOT have ANY of:

Nevek	Gyakoriság	Százalék
10. nigel farage	2322	1,392905
11. pope francis	2086	1,251335
12. matteo renzi	1757	1,053977
13. sigmar gabriel	1460	0,875814
14. john kerry	1436	0,861417
15. ed miliband	1423	0,853619
16. hillary clinton	1363	0,817627
17. bashar al-assad	1337	0,802030
18. werner faymann	1334	0,800230
19. dimitris avramopoulos	1313	0,787633
20. benjamin netanyahu	1281	0,768437
21. manuel valls	1219	0,731245
22. aylan kurdi	1182	0,709050
23. tayyip erdogan	1124	0,674257
24. nicolas sarkozy	1119	0,671258
25. george osborne	1096	0,657461
26. peter szijjarto	974	0,584276
~~27. los angeles~~	944	0,566280
28. martin schulz	916	0,549484
29. tony blair	914	0,548284
30. george w bush	901	0,540485
31. bernard cazeneuve	888	0,532687
32. zoltan kovacs	858	0,514691
33. frank-walter steinmeier	853	0,511692
34. antonio guterres	843	0,505693
35. frans timmermans	829	0,497295
36. jeb bush	800	0,479898
37. zoran milanovic	790	0,473900
38. jeremy corbyn	790	0,473900
39. nick clegg	774	0,464302
40. ewa kopacz	769	0,461302
41. miro cerar	762	0,457103
42. ahmet davutoglu	725	0,434908
43. aleksandar vucic	705	0,422910
44. robert fico	693	0,415712
45. mario draghi	689	0,413312
46. ted cruz	680	0,407914
47. jean asselborn	644	0,386318
48. yanis varoufakis	643	0,385718
49. margaret thatcher	639	0,383319
50. johannes hahn	638	0,382719
51. bohuslav sobotka	630	0,377920
52. mariano rajoy	623	0,373721
53. bashar assad	613	0,367722
54. marco rubio	591	0,354525

Nevek	Gyakoriság	Százalék
~~55. al qaeda~~	589	0,353325
56. boris johnson	576	0,345527
57. klaus iohannis	574	0,344327
58. horst seehofer	560	0,335929
59. bernie sanders	552	0,331130
60. philip hammond	549	0,329330
61. melissa fleming	543	0,325731
62. geert wilders	542	0,325131
63. johanna mikl-leitner	521	0,312534
64. tony abbott	519	0,311334
65. adolf hitler	511	0,306535
66. wolfgang schaeuble	508	0,304735
67. janos lazar	507	0,304136
68. ranko ostojic	504	0,302336
69. stefan lofven	496	0,297537
70. petro poroshenko	462	0,277141
71. helle thorning-schmidt	448	0,268743
72. salah abdeslam	446	0,267543
73. joel millman	440	0,263944
74. ben carson	436	0,261545
75. andrzej duda	434	0,260345
76. angelino alfano	432	0,259145
77. darko vojinovic	428	0,256746
78. ronald reagan	426	0,255546
79. laurent fabius	419	0,251347
80. adrian edwards	418	0,250747
~~81. europe christian*~~	400	0,239949
82. gordon brown	398	0,238749
83. lutz bachmann	388	0,232751
84. hans peter doskozil	386	0,231551
85. aleksandar vulin	385	0,230951
86. carlotta sami	384	0,230351
~~87. christian europe~~	383	0,229751
88. moammar gadhafi	378	0,226752
89. jeroen dijsselbloem	366	0,219553
90. mark rutte	361	0,216554
91. mahmoud abbas	356	0,213555
92. matteo salvini	345	0,206956
93. abdullah kurdi	344	0,206356
94. ivica dacic	338	0,202757
95. chris christie	334	0,200358
96. winston churchill	332	0,199158
97. jens stoltenberg	325	0,194959
~~98. allgemeine zeitung~~	323	0,193759
~~99. addis ababa~~	319	0,191359

Az 1. táblázat első oszlopában a 99 leggyakrabban szereplő személy neve szerepel (kisbetűkkel írva), a második oszlop a megjelenések számát, míg a harmadik oszlop a megjelenések arányát jelzi. Pontos elemzésünkhöz a táblázat némi korrekcióra szorul. Ez a korrekció is azt mutatja, hogy egyrészt ilyen nagy adatbázist már nem lehet a Big Data eszköztára nélkül elemezni, másrészt azonban emberi beavatkozásra is van szükség, ugyanis a Charlie Hebdo, Los Angeles, Al Qaeda, Europe Christian, Christian Europe, Addis Abbaba és Allgemeine Zeitung tévesen lett a személyek közé sorolva, ezeket kihúztuk. Az adatokból kiderül, hogy az európai migrációról szóló híreket kizárólag politikusok, közülük is Angela Merkel, David Cameron és Orbán Viktor dominálták. Ezen kívül az amerikai elnök az 5. helyen áll, míg az orosz a 8.-on. Egészében véve kiolvasható, hogy civil vezetők, civil véleményformálók nem játszottak domináns szerepet a migrációs hírekben, és a migrációról szóló beszédben sem. Egy ilyen típusú, összefüggésekre még kevésbé rákereső alapelemzés is olyan következtetésekhez segít minket, melyeket nem tudtunk volna észrevenni a Big Data eszköztára nélkül.

A migrációval kapcsolatban kiemelten érintett Görögország elnöke csak a 9., és a britek EU-ból való kilépése mellett élesen kampányolót Nigel Farage-t (10.) gyakorlatilag alig előzi meg a sorban. Ferenc pápa (11.) kétszer annyi, a témában megjelent hír szereplője, mint Martin Schulz (28.). Szomszédjaink közül a szerb elnök csak a 43., míg a szlovák a 44. a sorban. Érdekes módon a listában a 49. helyen megjelenik Margaret Thatcher is, aki 639 hír szereplője. Orbán Viktor mellett a magyar politkusok közül a listán szerepel még Szijártó Péter, Kovács Zoltán és Lázár János is.

Elemzés a migrációról és a témában megjelent cikkekben szereplő országokról

A migráció témájához kapcsolódóan megvizsgáltuk, hogy a hírekben mely országok szerepeltek a leghangsúlyosabban. Míg a mindennapi magyar néző úgy érzi, hogy Németország kapcsán hallani a legtöbbet a jelenségről, a hírekben világszerte az Egyesült Államok a legdominánsabb. A 3. ábrán látható címkefelhő[4] ezt azt eredményt vizualizálja, a 2. táblázatban pedig az első oszlopban az országok neve szerepel angolul, majd ezt követi az ide vonatkozó cikkek száma.

Az adatok elemzésekor elmondhatjuk, hogy Németország, mely a migráció célországa, ugyanolyan mértékben szerepel a hírekben, mint Nagy-Britannia. Ennek több oka lehet: a migráció kapcsán az elemzésekben mindig felmerül a már korábbi bevándorlók társadalmi és kulturális beilleszkedésének kérdése, ami különösen Nagy Britannia, illetve a listán 4. helyen szereplő Franciaország esetében vizsgálható. További okként felhozható, hogy Nagy-Britannia továbbra is meghatározó szerepet tölt be véleményvezérként Európában annak ellenére, hogy a migráció kérdése kapcsán egyértelműen Németország lépett elő ezen szerepbe. (Érdekes lenne 2017-ben megvizsgálni a 2016-os adatokat a GDELT adatbázisában, és összehasonlítani az egyes országok helyezésében történt változásokat.)

Az EU egyik első állomásaként funkcionáló Görögország 5. pozíciója, illetve a migrációs hullámot kibocsátó Szíria 6. helye talán nem meglepő. Hasonlóképpen Törökország vagy a szintén állomásként funkcionáló Olaszország első 10-ben való szereplése sem.

Magyarország a téma kapcsán sokat szerepelt a hírekben, az elemzésünkben, hogy 9. helyen áll. Ezzel jelentősen megelőzte a hírekben történő szereplések számában a migráció olyan kibocsátó országait, mint Afganisztán (16.), Líbia (19.), Eritrea (42!).

---

[4] Paraméterek: Migráció, Európa, Start Date = 01/01/2015, End Date = 12/31/2015, Object Field: LOC_COUNTRY, Object Weight: NAMESETS, Must have ALL these keywords: europe, Must also have AT LEAST ONE of: migration, Must NOT have ANY of:

3. ábra

A migrációval kapcsolatos hírekben megjelenő országok címkefelhője

3. ábra

Országok gyakorisága a migrációval kapcsolatos hírekben

Országok	Gyakoriság
1. United States	113451
2. Germany	53016
3. United Kingdom	52392
4. France	43324
5. Greece	42645
6. Syria	36789
7. Turkey	33098
8. Italy	33040
9. Hungary	26943
10. Russia	25922
11. Belgium	23335
12. Austria	20151
13. Iraq	18931
14. China	17098

Országok	Gyakoriság
15. Spain	14160
16. Afghanistan	13309
17. Serbia	13056
18. Sweden	12996
19. Libya	12100
20. Israel	11914
21. Poland	11887
22. Ukraine	11524
23. Australia	10092
24. Canada	10087
25. Croatia	9671
26. Ireland	9168
27. India	9057
28. Switzerland	8494
29. Iran	8420
30. Netherlands	8376
31. Czech Republic	8355
32. Macedonia	8167
33. Denmark	8051
34. Egypt	7372
35. Pakistan	7215
36. Oceans	7122
37. Romania	6880
38. Japan	6842
39. Lebanon	6668
40. Bulgaria	6284
41. Slovenia	5845
42. Eritrea	5464
43. Saudi Arabia	5309
44. Malta	5298
45. Luxembourg	5179
46. Jordan	5149
47. Nigeria	5133
48. Norway	5099
49. Republic Of [sic]	5022
50. Albania	4971
51. Mexico	4914
52. Portugal	4687
53. Somalia	4422
54. Finland	4306
55. South Africa	4230
56. Indonesia	4091
57. Malaysia	3916
58. Morocco	3729
59. Kosovo	3566

Országok	Gyakoriság
60. Sudan	3284
61. Brazil	3239
62. Latvia	3213
63. Yemen	3213
64. Philippines	3033
65. Thailand	3028
66. Bangladesh	2967
67. Tunisia	2860
68. Cyprus	2662
69. Lithuania	2562
70. Slovak Republic	2458
71. New Zealand	2419
72. Georgia	2410
73. South Korea	2402
74. Belarus	2369
75. Ethiopia	2340
76. Kenya	2338
77. United Arab Emirates	2203
78. Cuba	2172
79. Algeria	2155
80. Singapore	2066
81. Armenia	2023
82. Estonia	1993
83. Moldova	1819
84. Mali	1780
85. Venezuela	1714
86. West Bank	1687
87. Argentina	1686
88. Azerbaijan	1548
89. Qatar	1482
90. Bosnia-Herzegovina	1481
91. Niger	1424
92. Senegal	1286
93. Ghana	1259
94. Montenegro	1259
95. Hong Kong	1231
96. Uganda	1230
97. Kazakhstan	1226
98. Chile	1187
99. Colombia	1186

Ha az USA az első helyen szerepel, akkor talán jogosan merülhetne fel, hogy Mexikónak is elől kellene szerepelnie a sorban, hiszen egyrészt jelentős a fizikai migrációs kapcsolódás a két ország között, másrészt az Európába igyekvő migráció kapcsán az USA Mexikóval szemben foganasított intézkedéseit gyakran hozzák fel a cikkek példaként arra utalva, hogy az Egyesült Államok magas kerítéssel védi magát a déli szomszédjától érkezők elől. Meglepő tehát, hogy Mexikó csak az 51. helyen szerepel az országokat említő cikkek gyakorisági listáján.

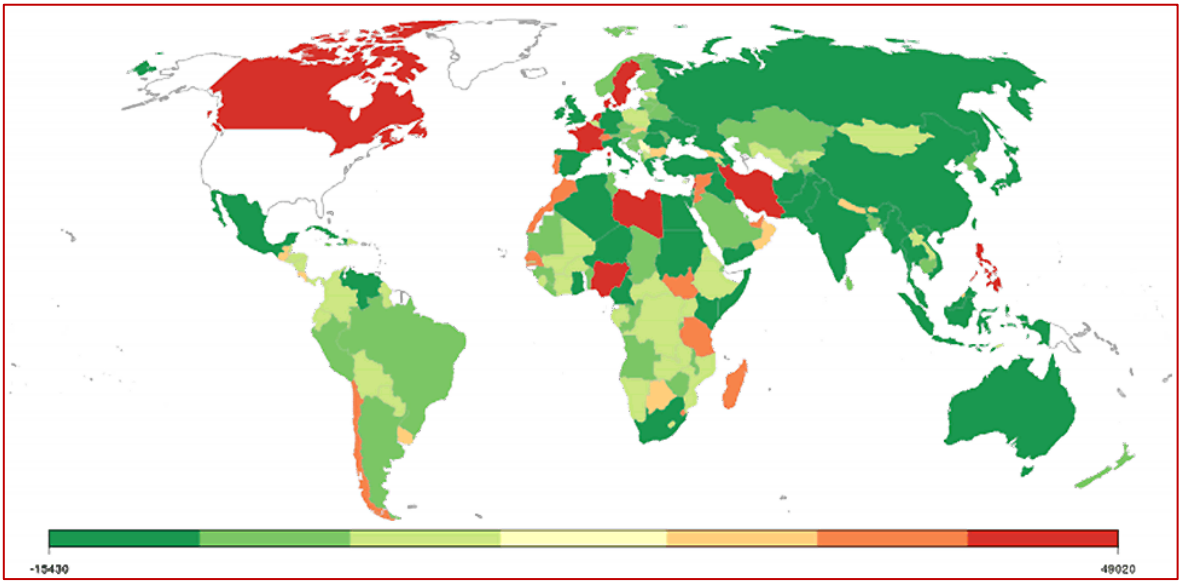
Konfliktusokra fókuszáló Big Data sajtószemle

A hosszabb időszakot átfogó elemzések mellett a GDELT Big Data rendszere az összefüggéseket a gépi tanulás segítségével feltáró napi sajtószemle készítésére is alkalmas. Az így kapott adatok elsősorban a politikai és gazdasági döntéshozók számára fontosak. Vegyük példaként a korábbiakhoz hasonlóan a 2015-ös év történéseit (GDELT Global Material Conflict 48-Hour Trend Report), és vizsgáljuk meg a két-két egymást követő napon történt konfliktusok alakulását. (2015. február 13-14. és 15-16.) Az így kapott elemzésben automatikusan megjelenik azon 10 ország neve, melyekben kiemelkedő jelentőségű konfliktusok történtek. Ahhoz, hogy pontos képet adjon a konfliktusok nagyságáról és súlyáról, a rendszer nem az események, hanem az arról történő tudósítások számát veszi alapul. Így például az Iránnal kötött atomegyezmény megkötése, bár egyetlen esemény, sajtóbeli megjelenése nagyszámú hírből épül fel, és a gyakoriság igen magas.[5]

Az így készített sajtószemle adatait a rendszer vizualizálja, térképre vetíti, ahogy ez a 4. ábrán látható.

4. ábra

Változás a konfliktusok mértékében az elmúlt 48 órában.
(Az új konfliktusok pirossal, ez enyhülők zölddel vannak jelölve)

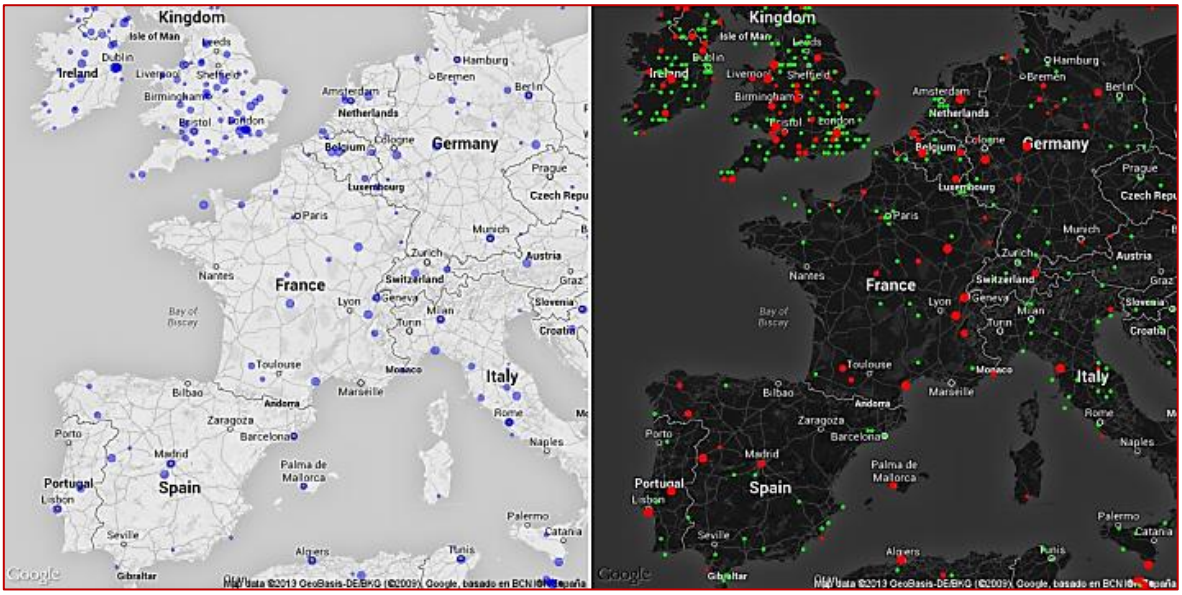


[5] Az elemzések egy Google Drive-ról érhetők el: https://docs.google.com/viewer?a=v&pid=forums&srcid=MDc3NjUxNjA1Nzg5MTQ1MTA5NTEBMDk0MDM2NjkyNjgyMjUyNjQyNTIBejl1RkRFaVJldWdKATAuMQEBdjI

A jelenséget országokra bontva újabb részletes térképeket készít a Big Data eszköztárával a GDELT adatbázisa alapján a rendszer (5. ábra).

5. ábra

Az elmúlt 48 órában született konfliktusok *A konfliktusok intenzitása, iránya*



Mindenegyes ország esetében a szürke térképen megjelenik az elmúlt 48 óra valamennyi jelentős konfliktusa a környező országok szerepeltetésével. A fekete térképen a konfliktusok intenzitása van jelölve. A konfliktusok számának jelentős növekedését a piros pontok, míg a gyengülését a zöld pontok jelzik. A pontok mérete a konfliktus méretével arányos.

A rendszernek éppen a számadatokat tároló tulajdonsága alapján lehetséges egy áldozatokat számláló használata is. Egy a médiából kapott információk alapján megbecsülhető, hogy az adott napon világszerte hány ember betegedett meg járványban, halt meg balesetben vagy háborúban, hány embernek kellett elhagynia otthonát, stb. Az adatok duplikálását megakadályozza, hogy minden eseményről csak a legteljesebb körű tudósítást rögzítik.

A GDELT konfliktusokra fókuszáló szemléjének használata során több korlátozó tényezőt is figyelembe kell vennünk. Az egyik ilyen tényező, hogy előfordulhat, a témák, helyszínek vagy személyek közti összefüggéseket nem mindig tudja megfelelően értelmezni a rendszer. A gyakori (együtt)előfordulás általában kapcsolatot feltételez, az ok-okozati következtetések azonban még a Big Data eszközeivel is nehezen fedezhetők fel. Jelenleg egy olyan cikkből, melyben együtt szerepel az Egyesült Államok, Oroszország és Szíria, illetve a légicsapások mint téma, a rendszer nem mindig tudja egyértelműen megállapítani, hogy melyik ország hajt végre légicsapásokat a másik területén, azonban további cikkeket is bevonva kikövetkeztetheti az összefüggéseket és tanulhat. Olyan esetekben, amikor egy cikk számos háttér-információt is tartalmaz, az összefüggések feltérképezése még nehezebb lehet. Így például, amikor a világsajtót bejárta egy pakisztáni lány, Malala Yousafzai hősies története, bátorságát Raoul Wallenbergével és Nelson Mandeláéval említették egy lapon, amiből természetesen nem következik, hogy ő egy eseményben vett részt az említettekkel.

Google Trendek

Az online médiatartalmak Big Data elemzése kapcsán fontos kiemelni a Google Trendek szolgáltatást. A világ legnagyobb globális digitális hálózati keresőmotorját működtető Google a Trendek szolgáltatásával egy adott téma vagy címszó múlt és jelenbeli népszerűségét mutatja be, és Big Data rendszerével következtetéseket von le a jövővel kapcsolatban. (Ezen következtetések a mindennapi felhasználók számára nem elérhetők.) A Trendek egyfajta *Zeitgeist* – korszellem – monitoraként is működik. A szolgáltatás a témákat térképen és görbén is vizualizálja, kérésre kimutatja az összefüggéseket több téma között. A hagyományos címszó kereséssel szemben a Trendek egy többdimenziós nézetét biztosítja az eredményeknek, figyelembe véve egyes témák szezonalitását, geostratégiai jelentőségét, illetve a médiában való megjelenését.

A Trendek algoritmusa egyes címszavak relatív és nem abszolút népszerűségét mutatja a keresések alapján. Az eszközt eredetileg arra fejlesztették ki, hogy megmutassa, hogyan változik egyes jelenségek népszerűsége az idők folyamán és a médiamegjelenés függvényében. A grafikonon megjelenített számok (1-től 100-ig) relatívak, az adott régió és az idő függvényei, ahol a 100-as mindig a népszerűség csúcsát jelzi. A keresésekkel egy grafikonon megjeleníthető a címszavak médiavisszhangja is.

Felkapott hírek

A Google felkapott hírek[6] szolgáltatása jó példa az online médiatartalmaknak a Big Data eszköztárával történő elemzésére. Vizsgáljuk meg a 2015-ös év 10 legfelkapottabb magyar témáit:

1. Charlie Hebdo
2. napfogyatkozás
3. Szíria
4. Oscar 2015
5. László Petra
6. migránsok
7. Simicska Lajos
8. menekültek
9. Quaestor
10. Je suis Charlie

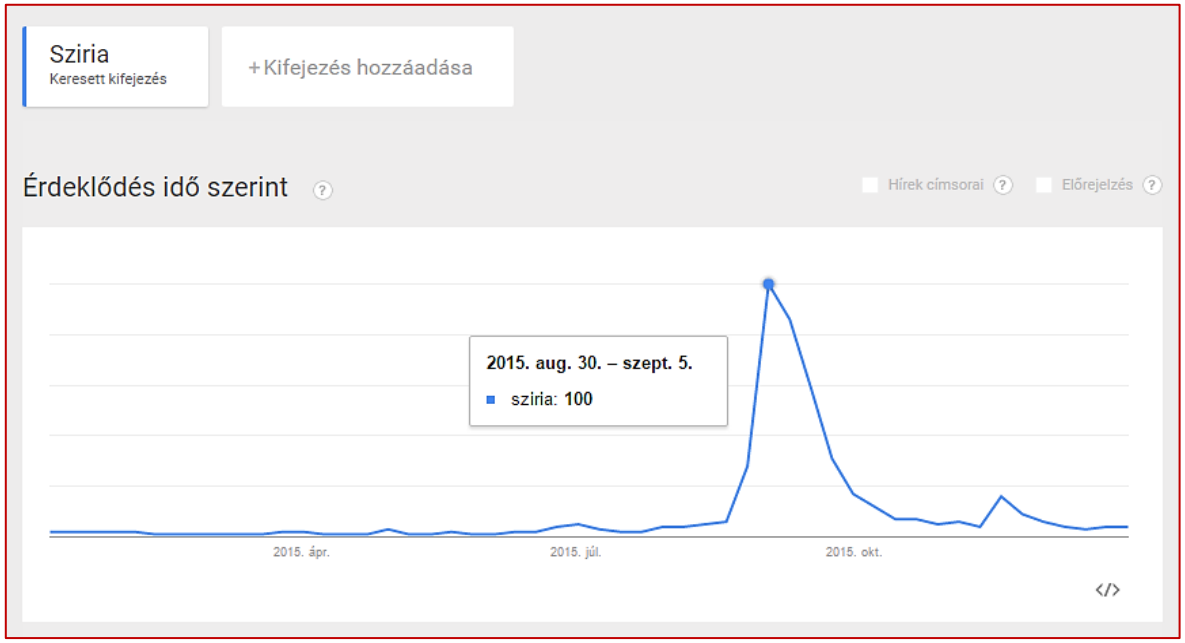
Azonnal feltűnik, hogy a migránsok és a menekültek címszavak azonos témakörbe tartoznak, azonban ellentétes jelentéstartalmat hordoznak. Ezenkívül a Charlie Hebdo és a Je suis Charlie is egyazon hírcsoportot jelöli.

Vizsgálatunkhoz először a tanulmányunkban végigvonuló vezérfonál mentén a Szíria címszót választottuk. A Google Big Data elemzése szerint a magyar médiában a téma 2015. augusztus 30-a és szeptember 5-e között csúcsosodott ki, miközben 2015 augusztusáig Szíria gyakorlatilag alig tematizálta a médiát. A hirtelen ugrás augusztusban kezdődött, és október végére gyakorlatilag lecsenget. Ez az időszak egybeesik azzal a periódussal, amikor tömegesen jelentek meg bevándorlók Magyarországon. Decemberben végül már alig szerepelt a téma a magyar hírekben (6. ábra).

[6] https://www.google.com/trends/topcharts#vm=trendingchart&cid=4a9666d1-9bfc-430a-aa4a-e56e04b4d8f1&geo=HU&date=2015&cat=

6. ábra

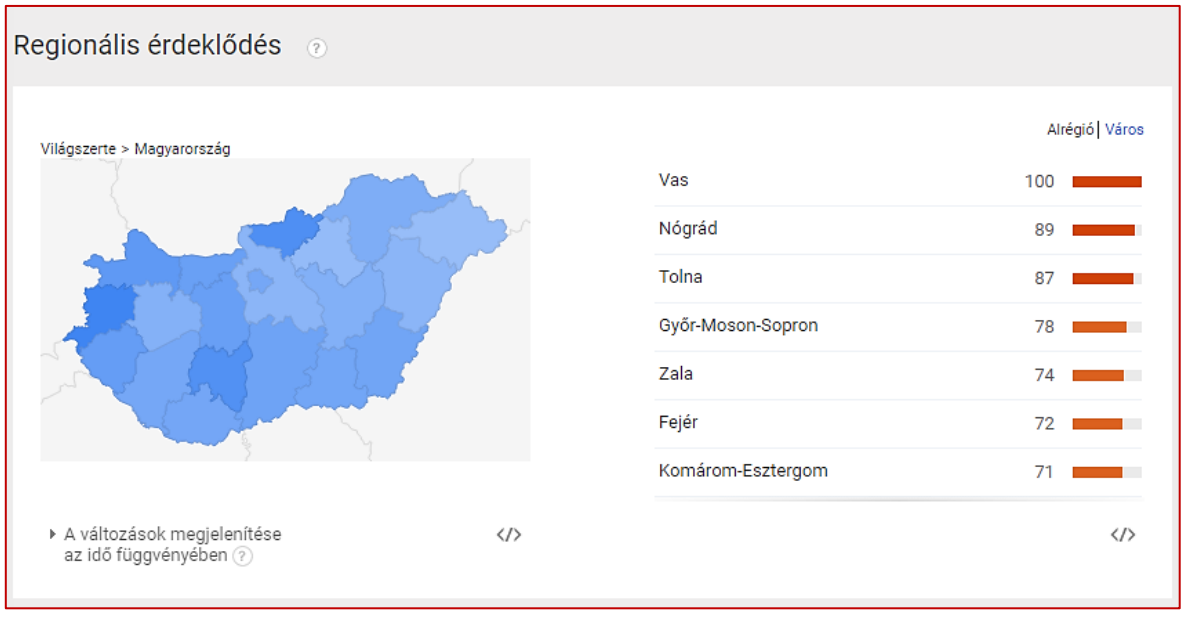
A Szíriával kapcsolatos hírek intenzitása a magyar médiában



Hasonlóképpen érdekes annak vizsgálata, hogy mely magyar régió médiumait foglalkoztatta a legjobban a téma. Meglepő eredmények születtek: Vas, Nógrád és Tolna kiemelkedik ebben a tekintetben, míg Budapest és Pest megye a lista végén szerepel (l7. ábra).

7. ábra

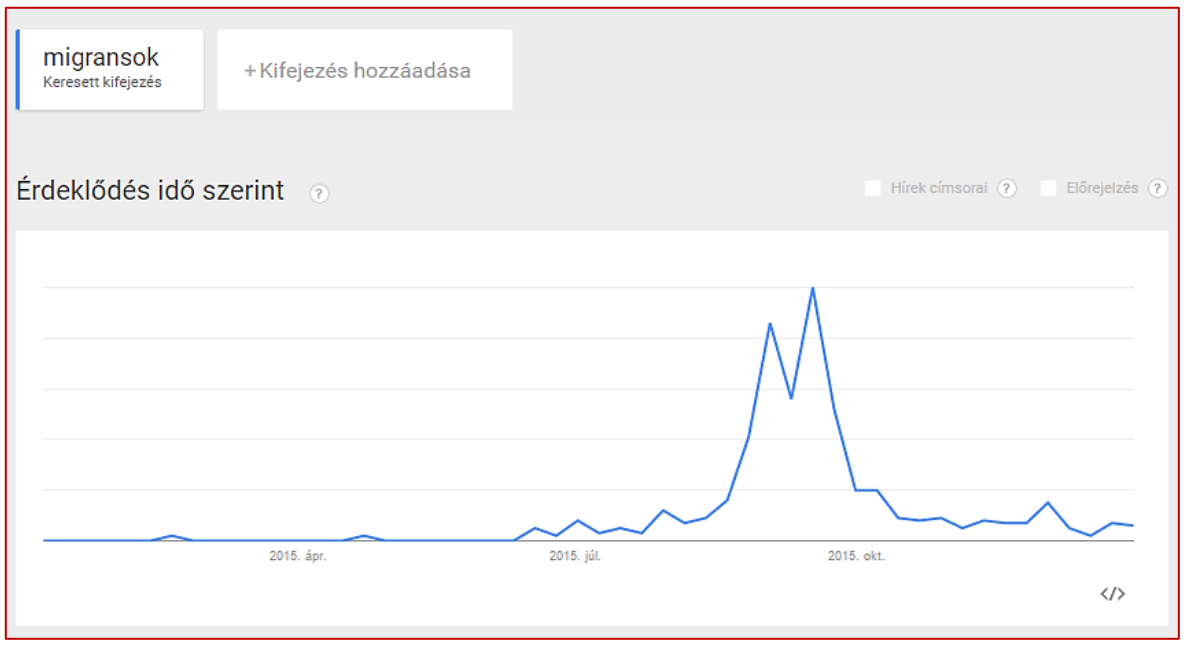
Magyar regionális médiaérintettség Szíria témájában 2015-ben



Hasonlóképpen, ha a migránsok címszó magyar médiában történő szereplését vizsgáljuk, elmondható, hogy 2015 júliusától van jelen, augusztusban kilő az érdeklődés, és ez szeptemberben egy minimális visszaesés után csak erősödik, majd októberben beállt kicsit a júliusi szint felett (lásd 8. ábra).

8. ábra

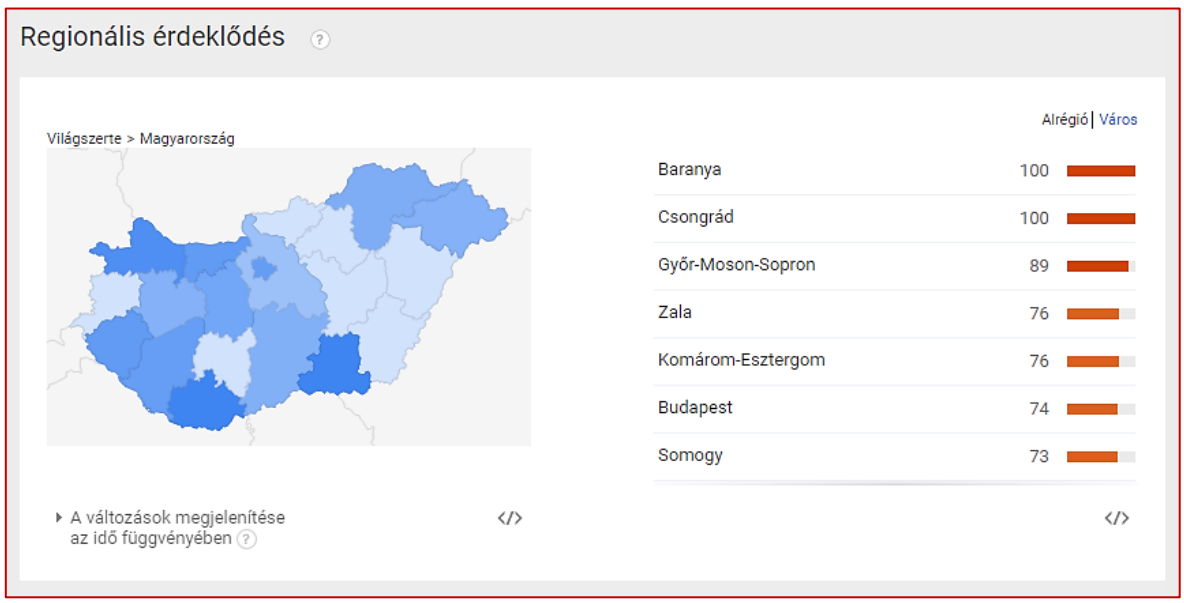
Migránsokkal kapcsolatos hírek intenzitása a magyar médiában



Az elemző joggal feltételezhette, hogy azoknak a régióknak a magyar sajtója foglalkozott a legerőteljesebben ezzel a témával, melyek földrajzilag is érintettek a témában. Ez a feltételezés igazolódott, ugyanis Baranya és Csongrád a Szerbiából és Horvátországból érkező migráns-útvonalak miatt, Győr-Moson-Sopron megye sajtója pedig az Ausztria felé távozó bevándorlók kapcsán cikkezett a legtöbbet. Érdekes módon Budapest a lista közepén szerepel (9. ábra).

9. ábra

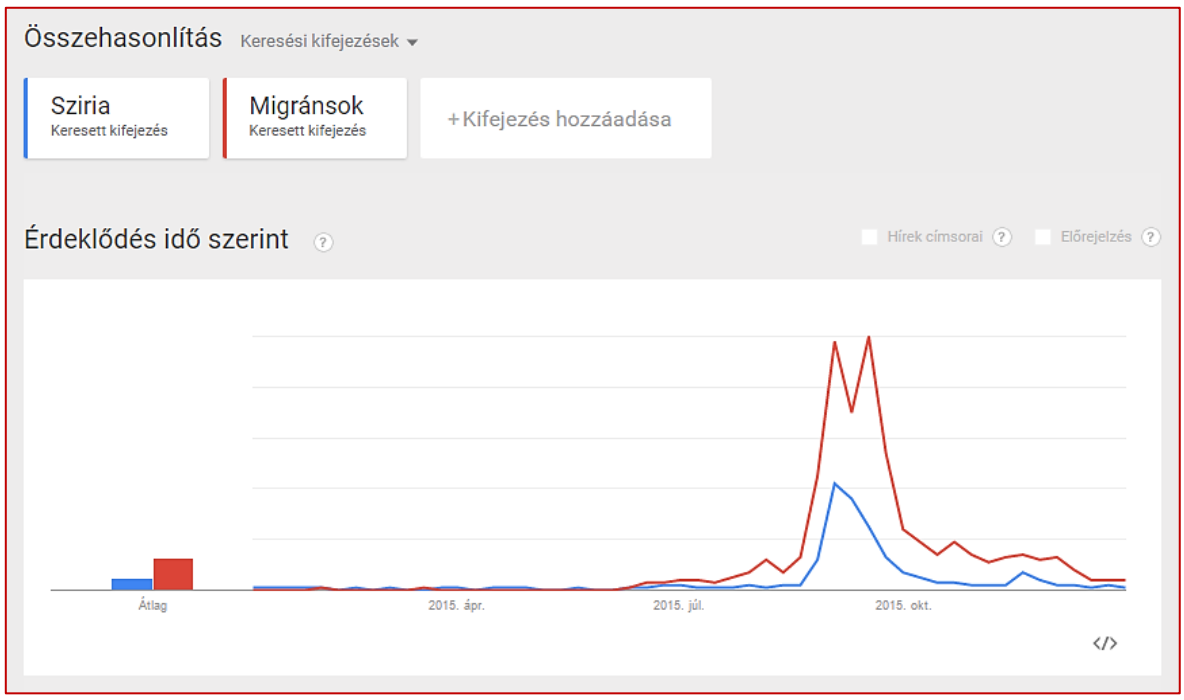
Regionális médiaérintettség a migránsok témájában 2015-ben



Összefüggéseket kerestünk a Szíria és a migránsok címszavak médiatartalmakban történő szereplésében. A két címszó gyakorlatilag együtt mozog, de a migránsok témája végül intenzívebb figyelmet kapott (10. ábra).

10. ábra

A Szíria és migránsok címszavak együttes mozgása a magyar médiában



A közösségi médiatartalmak elemzése

A közösségi média hírforrásai jelentős szerepet töltenek be a tájékoztatásban. A hírek nagy része a Facebookon jut el először a felhasználókhoz. A közösségi médiára irányuló tartalomelemzésünket a Twitter környezetében hajtottuk végre, ugyanis a Twitter rendszere lehetővé teszi a teljes tartalmának kutatását, osztályozását, míg a Facebook algoritmusa zárt, és üzleti célokra nyitják meg, így a médiakutatók sem kaphatnak teljes betekintést a legnagyobb közösségi oldal működésébe. A Twitter azért alkalmas felület a vizsgálatunkhoz, mert egyszerre ismeretségi hálózat és mikroblog-szolgáltatás, mely lehetővé teszi a felhasználóknak, hogy rövid bejegyzéseket írjanak maximum 140 karakter hosszúságban. A legújabb bejegyzések a felhasználó profilján jelennek meg, de azonnal láthatók (alapbeállításként) gyakorlatilag mindenki számára.

A Twitteren a felhasználók eltérő témákban írnak. Az életüket érintő valós eseményektől kezdve a médiatartalmak megosztásáig minden szerepel tudósításaikban. Ezek elemzésére jöttek létre az olyan platformok, mint a TUCAN, TwitterStand, Twitris, TwitInfo vagy Tweet Xplorer. Ezek képesek a különböző kutatások eredményeit vizualizálni is (Mahata és Agarwal 2014: 6).

A TUCAN (Twitter User Centric ANalyzer) képes például kimutatni az összefüggéseket egyes tweetek között, aminek köszönhetően képet kaphatunk például egy-egy felhasználót foglalkoztató témákról. Hasonlóképpen a TUCAN képes összevetni különböző felhasználók érdeklődési körét is (Grimaudo et al 2014: 65).

Elemzésünkhöz a Tweet Wiz alkalmazást választottuk, mivel ez tudja a leglátványosabban vizualizálni az eredményeket. (https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/) A Tweet Wiz a Twitter profilunkkal bejelentkezve megkeresi a keresési kifejezéssel kapcsolatban gyakorlatilag valós időben tweetelt üzeneteket, kiemeli az üzenetek címszavait, és sentiment analysis segítségével meghatározza az üzenetek olyan érzelmi töltetét is mint a szomorúság, harag, idegesség, feszültség, unalom, nyugalom, izgalom, vagy éppen a boldogság.

Egy olyan kulcsszót (személyt) választottunk első elemzésünkhöz, mely a jelenben aktívan foglalkoztatja az online közvéleményt, és nagyon gyakran szerepel az online médiatartalmakban is. Az Egyesült Államokbeli elnökválasztás kapcsán már 2016 februárjában nyilvánvalóvá vált, hogy Donald Trump valószínű republikánus elnökjelöltnek kiemelt szerepe lesz a médiatartalmakban, és ezért folyamatos téma lesz az amerikai felhasználók által politikai hírek megosztására és megvitatására használt Twitteren is. A 2016. február 25-én 18 óra 49 perckor végzett elemzés kimutatta, hogy a Donald Trumppal kapcsolatos tweetek érzelmi telítettsége alapvetően pozitív, az aktivitás tengelyén a nyugalom és a figyelem között helyezkedik el. A negatív érzelmek közül erőteljesebben csupán a feszültség és unalom manifesztálódik a tweetekben (11. ábra) Bár a felmérést nem a hagyományos fogalmak által leírható médiatartalmakban végeztük, mégis úgy éreztük, hogy az új média része, a Twitter is fontos hírforrás.

11. ábra

A Donald Trumppal kapcsolatos tweetek (2016. 02. 25. 18:49)

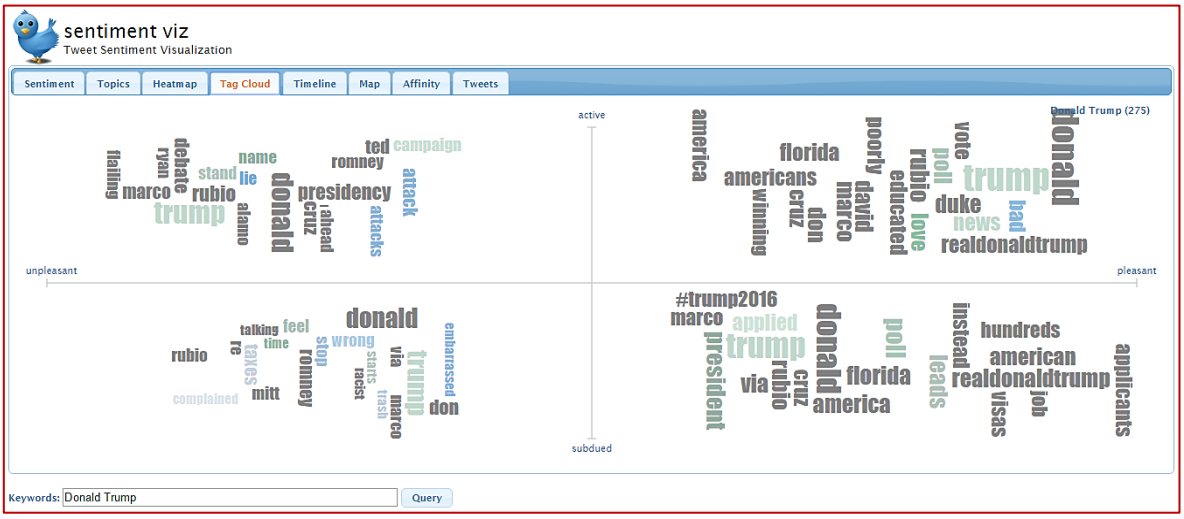
A rendszer a már ismertetett negatív-pozitív (kellemes-kellemetlen), illetve aktív-passzív tengelyen címszavak szerint is csoportosítja a tweeteket. Így az aktív-pozitív mezőben az ilyenkor általában megjelenő címszavak mellett kiemelt jelentőséggel bírnak a *Florida, hajlandó, iskolázott, szeretet* és *hírek* szavak. A pozitív és passzív mezőt meglepő módon ugyancsak a *Florida*, valamint az *elnök* és a *vízum* szavak fémjelzik. Az aktív-negatívak között szintén az *elnök* szerepel, de itt találhatók meg a *bukó*, *támadások* és *Alamo* szavak is. A passzív-negatív címszavak kisebbek, ebben a mezőben található a *stop, szégyenkező* és *adó* kifejezések. Szinte valamennyi mezőben megjelennek a republikánus vetélytársai nevei is (12. ábra).

A sentiment és címszó elemzés mellett lehetőség van a tweetek téma szerinti csoportosítására, illetve időrendbeli sorrendben történő megjelenítésére is. Hasonlóképpen hasznos közvetkeztetéseket lehet levonni – különösen a kampánystábok számára – a tweetek területi elhelyezkedéséből. Mivel a Sentiment Viz Big Data rendszer valós időben elemzi a bejegyzések hangulatát, ezért például nagyon jól lehet mérni egy-egy politikai jelölt elfogadottságát gyakorlatig minden egyes tévévita vagy kulcsfontosságú kijelentés után.

A Sentiment Viz számos témában tűnik alkalmas elemző eszköznek. Azonban ebben az esetben is nagy jelentősége van az emberi tényezőnek, ugyanis a kutatóknak kell meghatározniuk azon keresőkifejezéseket, melyek a leghatékonyabban segítik a keresést és az összefüggések feltérképezését. Képességei annyiban limitáltak, hogy a magyar felhasználók körében nem népszerű a Twitter, így a kevés üzenetből nehéz következtetéseket levonni.

12. ábra

A Donald Trump kereső kifejezéssel kapcsolatos találatok címkefelhő formájában történő vizualizálása



Kihívások és kritikák

Az elemzések után most vizsgáljuk meg az általunk választott módszer kihívásait, kritikáit és lehetséges hibáit is. Először is azt kell látnunk, hogy a közösségi média tartalmait alapvetően az Y és Z generáció tagjai állítják elő, tehát a 10-es, 20-as, 30-as éveikben járó fiatal felhasználók, akik nem reprezentálják teljes egészében a digitális írástudók összességét; ebből a szempontból az idősebb generáció rejtőzködőnek számít. Ez a minta tehát torz és hasonlít a választások idején készített minták összetételéhez, hiszen ott is azoknak a véleményét kapjuk, akik érdeklődnek a politika iránt.

Másodszor, az internetre jellemző anonimitás vagy a hamis profilok megtévesztők lehetnek, hiszen a hagyományos médiaügynökségekkel szemben, melyek azonosíthatók, nem tudhatjuk, hogy egy felhasználó hányszor szerepel a mintában (Westera 2013). Ám mivel ez a minta a Big Data környezetében nagyon nagy, nem biztos, hogy ez jelentősen befolyásolja az elemzés eredményét.

Harmadszor, a Big Data környezetében nagy a zaj, mely a kommunikáció-elméletek szerint megnehezíti a befogadást és ezzel együtt az elemzést is, továbbá elfogultságra hajlamosít (Boyd és Crawford 2012).

Negyedszer, a Big Data eszközeivel végzett elemzés azt a tévhitet keltheti a kutatókban, hogy madártávlatból mindent észre tudnak venni, ami korábban rejtett volt a szemük elől. Olyan mintázatokat és kapcsolatokat is meglátnak, melyek a valóságban nem léteznek, hanem véletlen egybeesések; a nagy mennyiségű adat ugyanis számtalan kapcsolatot létrehoz és mintázatot kirajzol (Boyd és Crawford 2012).

Összefoglalás és jövőképek

A Big Data tartalomelemzés fejlődése elvezetett ahhoz, hogy a *Narrative Science* technológiai társaság algoritmusa a már strukturált adatok (sporteredmények, tőzsdei adatok, Twitter bejegyzések, stb) segítségével képes emberi beavatkozás nélkül narratív történeteket, híreket generálni. Így például – a sport területénél maradva – egy médiaügynökség nagy mennyiségű statisztikai adat birtokában pillanatokkal a meccs vége után már hosszú elemzést publikálhat

online; korábban erre órákat kellett várni, sőt az online média megjelenése után is ez a típusú újságírás a másnap megjelenő nyomtatott sajtót jellemezte (Stone 2014: 16).

A kérdés azonban marad: kinek a birtokában lesznek az adatok? kormányok vagy vállalatok tulajdonát képezik-e majd? A Facebook algoritmusa jelenleg zárt a társadalomtudósok és hálózatkutatók előtt, a teljes képet a kapcsolatrendszerekről és felhasználók közti interakcióról ők külső szemlélőként nem láthatják. De óriási a különbség van a társadalom tagjai között is abban a tekintetben, hogy ki milyen mennyiségű adathoz férhet hozzá (Moorthy et al 2015: 75).

Irodalom

Anderson, Chris (2006) Hosszú farok – A végtelen választék átírja az üzlet szabályait. Budapest, HVG.

Benedek, András – Molnár, György (2014) Supporting the m-learning based knowledge transfer in university education and corporate sector. In: Sánchez, Arnedillo – Isaías, Pedro (eds.) *Proceedings of the 10th International Conference on Mobile Learning*. Madrid, IADIS Press, 339–343.

Boyd, Danah – Crawford, Kate (2012) Critical questions for big data. *Information, Communication and Society*, 15(–), 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878

Csepeli György (2015) A szociológia és a Big Data. *Replika*, 2015/3–4, 171–176.

Grimaudo, Luigi – Song, Han Hee – Baldi, Mario – Mellia, Marco – Munafò, Maurizio (2014) TUCAN: Twitter User Centric ANalyzer. In: Kawash, Jalal (ed.) *Online Social Media Analysis and Visualization*. New York, Springer, 63–80. https://doi.org/10.1007/978-3-319-13590-8_4

Herring, Susan C. (2010) Web content analysis: Expanding the paradigm. In: Hunsinger, J. et al (eds.) *International Handbook of Internet Research*. Springer Verlag, 2010. 233–249. https://doi.org/10.1007/978-1-4020-9789-8_14

Holsti, Ole R. (1968) Content Analysis. In: Lindzey, G. – Aronson, E. (1968 eds.) *The Handbook of Social Psychology* (Vol. II.), New Delhi: Amerind Publishing Co., 596–692.

Krippendorff, Klaus (1995) *A tartalomelemzés módszertanának alapjai.* Budapest, Balassi Kiadó.

Laney, Douglas (2012) 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (Utolsó letöltés 2016. június 19.)

Mahata, Debanjan – Agarwal, Nitin (2014) Identifying Event-Specific Sources from Social Media. In: Kawash, Jalal (2014 ed.) Online Social Media Analysis and Visualization. New York, Springer, 1–26. https://doi.org/10.1007/978-3-319-13590-8_1

Majkić, Zoran (2014) *Big Data Integration Theory*. Theory and Methods of Database Mappings, Programming Languages, and Semantics. New York, Springer. https://doi.org/10.1007/978-3-319-04156-8

Manovich, Lev (2012) Trending: The promises and the challenges of big social data. In: Gold, Matthew K. (2012 ed.) *Debates in the Digital Humanities.* Minneapolis, University of Minnesota Press. 460–475. https://doi.org/10.5749/minnesota/9780816677948.003.0047

Molnár, György (2015) Teaching and Learning in modern digital Environment. In: Szakál, Anikó (2015 szerk.) *SAMI 2015, IEEE 13th International Symposium on Applied Machine Intelligence and Informatics*. Herlany, Slovakia, 2015.01.22–2015.01.24. 213–217. https://doi.org/10.1109/sami.2015.7061878

Moorthy, Janakiraman – Lahiri, Rangin – Biswas, Neelanjan – Sanyal, Dipyaman – Ranjan, Jayanthi – Nanath, Krishnadas – Ghosh, Pulak (2015) *Big Data: Prospects and Challenges. VIKALPA The Journal for Decision Makers*, 40(1), 74–96. https://doi.org/10.1177/0256090915575450

n. a. http://www.gartner.com/it-glossary/big-data/ (Utolsó letöltés 2016. június 19.)

n. a. https://www.google.org/flutrends/about/ (Utolsó letöltés 2016. június 19.)

n. a. https://www.it-services.hu/hirek/mi-az-a-big-data/ (Utolsó letöltés 2016. június 19.)

Pang, Bo – Lee, Lillian (2008) *Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval.* 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Samuel, Arthur (1959) Some Studies in Machine Learning Using the Game of Checkers. IBM Journal, 3(3), 210–229. http://dx.doi.org/10.1147/rd.33.0210

Shroff, Gautam (2014) The Intelligent Web: Search, smart algorithms, and big data. Oxford, Oxford University Press. https://doi.org/10.5860/choice.51-6220

Stone, Martha L. (2014) Big Data for Media. (Report), Institute for the Study of Journalism, http://reutersinstitute.politics.ox.ac.uk/publication/big-data-media

Szűts Zoltán (2012) A Web 2.0 kommunikációelméleti kérdései. *Jel-Kép*, 2012/1–4. http://communicatio.hu/jelkep/2012/1_4/szuts_zoltan.htm (Utolsó letöltés 2016. június 19.) https://doi.org/10.20520/jel-kep.2012.1-4.5

Szűts Zoltán (2013) *A világháló metaforái*. Budapest, Osiris. Szűts Zoltán – Yoo Jinil (2013) A kiterjesztett valóság térhódítása. *Információs Társadalom*, 2013/1, 58–67. https://doi.org/10.22503/inftars.xvi.2016.1.1

Westera, Wim (2013) Social Media and Big Data – Cracks in the Crystal Ball? RW Connect. https://rwconnect.esomar.org/using-social-media-for-market-analysis-cracks-in-the-alleged-crystal-ball/ (Utolsó letöltés 2016. június 19.)

Zhang, Lei – Liu, Bing (2014) Aspect and Entity Extraction for Opinion Mining. In: Chu, Wesley W. (2014 ed.) Data Mining and Knowledge Discovery for Big Data – Methodologies, Challenge and Opportunities. New York, Springer. https://doi.org/10.1007/978-3-642-40837-3_1