MI VAN A SZÁMOKON TÚL?

Magyar blue-chip vállalatok negyedéves riportjainak kvalitatív szövegalapú elemzése

Hajósi Péter
(K&H Alapkezelő Zrt.)

A tanulmány beérkezett: 2019. szeptember 15., opponálás: 2019. november 4. – november 22., véglegesítve: 2020. február 23.

ÖSSZEFOGLALÓ

A szövegek gépi feldolgozása a politikatudomány és annak határterületein is egyre nagyobb szerephez jut. Kutatásomban egyrészt áttekintem a politikatudományon belül a közpolitika területén zajló kutatásokat, másrészt megvizsgálom, hogy a szövegbányászat mennyiben alkalmazható a hazai tőzsdei vállalatok jelentéseinek elemzésénél. A négy magyar blue-chip vállalat (OTP, MOL, Richter Gedeon, Magyar Telekom) angolnyelvű negyedéves jelentéseinek szövegszerű elemzésével tesztelem azokat a hipotéziseket, miszerint *1)* a szakértők véleményét a cégek kommunikációjával kapcsolatban a riportok gépi elemzése visszaigazolja és *2)* van egyértelmű kapcsolat a riportok hangvétele és a piaci adatok – az árfolyamok elmozdulása, illetve azok volatilitása – között. Az elemzés R statisztikai csomag alkalmazásával történik, és az abban elérhető szótáralapú, a szövegek lexikai diverzitását vizsgáló és a Wordfish-módszerrel elvégzett elemzések részben visszaigazolják ezeket a hipotéziseket: szignifikáns kapcsolatot találtam a riportok hangvétele és a közzététel időpontjában mért árfolyam-elmozdulás és -volatilitás között, a négy cég közül kettő esetében (Magyar Telekom és MOL) sikerül igazolni az előzetes elemzői véleményt. Eredményeim alapján a módszertan alkalmazhatónak tűnik a hazai vállalati környezetben is. A regressziós modell magyarázó ereje ugyanakkor alacsony, ezért további elemzés szükséges ahhoz, hogy a jelentések szövegszerű vizsgálata a gyakorlatban is használható legyen. Összességében megállapítható, hogy a szövegbányászatnak van relevanciája a magyar vállalatok akár közpolitikai, akár befektetői elemzésénél, így más országok gyakorlatához hasonlóan egy központi, egységes formátumú adatbázis kiépítése előrelépés lenne a piaci szereplők hatékony, transzparens és egyenrangú tájékoztatása érdekében.

Kulcsszavak: szövegbányászat ■ transzparencia ■ tőzsde ■ Wordfish ■ hangvétel-elemzés

A szövegek gépi feldolgozása és az ilyen formában kinyerhető információk elemzése egy rohamosan fejlődő terület, amely több szakterületen, így a politikatudomány és annak határterületein is egyre nagyobb szerephez jut. Dolgozatomban áttekintem a politikatudományon belül a közpolitika területén zajló kutatásokat és kitekintésként azt vizsgálom meg, hogy a szövegbányászat mennyiben alkalmazható egy speciális területen, a hazai tőzsdei vállalatok jelentéseinek

elemzésénél. Ezzel arra mutatok rá, hogy – más országok gyakorlatához hasonlóan – milyen közösségi haszna van egy központi adatbázis működtetésének, ahol egy helyen, egységes szerkezetben teszik elérhetővé a vállalati riportokat, így könnyítve meg azok akár közpolitikai, akár befektetési elemzését.

A tőzsdei cégeket elsősorban a pénzügyi adataik alapján elemzik, miközben ma már számos kutatás irányul arra, hogy a riportokból, a menedzsment által adott interjúkból, elemzői találkozókon elhangzottakból, a sajtóban vagy a közösségi médiában megjelenő üzenetekből milyen „mélyebb" tartalmat lehet kiolvasni a cégekre vonatkozóan. Ezekből pedig későbbi vállalati eseményekre (felvásárlás, csalás, csőd, stb.), az árfolyamra (alul- vagy felülteljesítés) vagy a pénzügyi mutatókra (ROE, eladósodottság, stb.) vonatkozóan milyen előrejelzések adhatók.

A releváns irodalmat átnézve a magyar blue-chipeket (MOL, OTP, Richter Gedeon, Magyar Telekom) választom az elemzés tárgyául, mivel – tudomásom szerint – ezekre a cégekre ilyen kutatás még nem készült. Ezeket a cégeket továbbá nemzetközi összehasonlításban kevés elemző követi, a vállalatokról elérhető szöveges forrásokat így kevesen elemzik. Ezáltal nyílik nagyobb tér arra, hogy kvalitatív módszerekkel releváns információt nyerjünk ki a riportokból.

A magyar cégekre vonatkozó szöveges források ugyanakkor korlátosak. Nem érhetők el az elemzők és a menedzsment között zajló elemzői konferenciák szöveges leiratai, amelyek a szereplők „élőszavas" – így várhatóan az ilyen beszélgetések – hangvételére vonatkozó információt tartalmazhatnak. És ezeket a cégeket érdemben nem elemzik aktívan a blogokon vagy a közösségi médiában (Facebook, Amazon, Twitter stb.), így nem áll rendelkezésre megfelelő mennyiségű forrás az elemzéshez. Lényegében csak a vállalati éves és negyedéves riportok érhetők el, mint elsődleges, tehát a részvényelemzők által még fel nem dolgozott szöveges források. Korábbi nemzetközi kutatások alapján (pl. Loughran, 2016) azonban még az olyan semleges hangvételűnek tűnő szövegekből is kinyerhető felhasználható adat, mint amilyenek a vállalati jelentések.

Kutatásomban két hipotézis alapján azt tesztelem, hogy a cégek negyedéves jelentéseinek szöveges részeiből kinyerhetők-e a piaci szereplők számára értékes információk. Az egyik hipotézis, hogy ezek az információk alátámasztják a cégekről kialakult szakértői, elemzői véleményeket (szakértői összevetés). A másik, hogy kapcsolat van a jelentések szöveges hangvétele és a piac cégekkel kapcsolatos reakciója, konkrétan az árfolyam elmozdulása és annak volatilitása között (piaci összevetés). A kutatás végén az eredmények részben igazolják ezeket a feltevéseket. Módszertanilag a negyedéves riportok letöltését és előkészítését követően publikusan elérhető szótárak felhasználásával és ún. felügyelet nélküli tanulás módszerével *(WordFish)* elemzem a szövegeket. Mindvégig az R-programnyelvet és az ott elérhető programcsomagokat *(tm, quanteda)* használom.

A tanulmány a következőképpen épül fel: a szakirodalom rövid ismertetése után – az eredmények későbbi reprodukálhatósága érdekében – az adatbá-

zis bemutatása és a szöveg, mint adatforrás, elemzéshez való előkészítésének részletezése következik. Ezt követően kerül sor az elemzésre és az eredmények bemutatására a szótár- és a WordFish-módszerrel. Zárásként összefoglalom az eredményeket és felvázolom a további kutatási irányokat.*

A SZAKIRODALOM ÁLLÁSA

A szövegbányászatnak szerteágazó irodalma van: szakterületenként és az alkalmazott módszertant tekintve is rengeteg tanulmány jelent meg a témával kapcsolatban. Ráadásul a feldolgozható szövegek, mint inputok, számának rohamos növekedésével (pl. a korábban nem létező rövid Twitter- és Facebook-üzenetek) párhuzamosan ezek elérhetősége is javul, hiszen például míg korábban a szöveges források nagy része csak nyomtatott formában volt elérhető, mostanra az interneten keresztül viszonylag könnyen hozzáférhetővé váltak. Emellett az informatika és ezen belül a számítási kapacitások fejlődésével egyre hatékonyabban lehet ezeket a forrásokat feldolgozni.

A rengeteg cikk közül a szövegbányászati módszertan egyik legstrukturáltabb bemutatását a Grimmer–Stewart-szerzőpáros (2013) adja, akik – a politikatudomány területén született eredményekből szemezve – megismertetik az olvasót a rendelkezésre álló vizsgálati eszközökkel és az ezek használatával kapcsolatos általános érvényű kihívásokkal. Magyar nyelven, szintén a politikatudományra koncentrálva Sebők és társai (2016) által összeállított kötet ad részletes betekintést a kvantitatív szövegelemzés világába.

A politikatudomány területén a közpolitikai elemzéseken belül is fontos szerep jut a szövegbányászatnak, tehát a szövegek feldolgozásának. A jegybanki döntéshozók nyilatkozatainak gazdasági hatása például régóta vizsgált terület: Poole és Rasche (2003) az amerikai Fed fokozatosan egyre transzparensebbé váló kommunikációjának hatását vizsgálja, miszerint azáltal, hogy a piaci szereplők megértik a jegybankárok szándékait, a monetáris irányítás súrlódásmentesebbé és ezáltal hatékonyabbá válik. Aizenman és társai (2014) szintén a Fed esetében vizsgálják azt, hogy a monetáris bizottság (FOMC) tagjainak nyilatkozatai milyen hatással vannak a feltörekvő piacok pénz- és tőkepiacaira. Megállapításuk, hogy az elnök szavaira messze nagyobb reakció érkezik a befektetők részéről, mint az FOMC többi tagjának nyilatkozatára.

Fenti elemzések viszonylag kis mintákon (kevés szöveges forrás feldolgozásával és sok manualitással) készültek. A gépi szövegbányászat eszköztárával ugyanakkor nagyságrendileg több szövegre, mint forrásra, épülő elemzés válik elérhetővé. Így nyílt lehetőség például arra, hogy Born és társai (2014) több

* A dolgozat alapját a Rajk László Szakkollégiumban 2019 tavaszán tartott szövegbányászat-kurzuson készült elemzés adja. Ezúton köszönöm Sebők Miklós, és Kapronczay Mór PhD-kurzustartók, Bene Zsombor, CFA kollégám és két anonim opponens hozzájárulását a kutatáshoz.

mint ezer olyan pénzügyi stabilitási jelentést vagy ezzel kapcsolatos nyilatkozatot vizsgáljanak meg, amelyek 37 jegybanktól érkeztek és 15 évet fednek le. Gépi szövegelemzéssel végzett kutatásuk arra a következtetésre jut, hogy a jelentéseknek és a jegybankárok nyilatkozatainak a piaci helyzettől függően van hol kisebb, hol nagyobb hatása a részvénypiacokra.

A központi bankokon túlmenően más gazdasági közintézmények kommunikációjának, általánosságban a transzparens kommunikáció fontosságának vizsgálatához is segítséget nyújt a szövegbányászat eszköztára. Baker és társai (2016) például 12 ezer, amerikai újságból vett cikk elemzésével azt vizsgálják, hogy hogyan hat a gazdaságpolitikai lépésekkel kapcsolatos bizonytalanság a gazdasági döntésre. A cikkek kézi és gépi feldolgozásával készített mutató alapján arra a következtetésre jutnak, hogy a bizonytalanság mind a cégek, mind a makrógazdaság szintjén csökkenő foglalkoztatottsághoz és beruházási kedvhez vezet.

Az általam vizsgált, a fenti gazdasági intézményekkel szoros kapcsolatban lévő pénzügyi területen természetesen várhatóan továbbra is a számszaki adatok (makrógazdasági mutatók, árfolyamok, hangulatindikátorok stb.) maradnak az elemzés kiindulópontjai. A számokon túl ugyanakkor egyre nagyobb hangsúlyt kapnak a különböző szövegalapú források is, amelyek egy része korábban is létezett, de csak manuálisan kerültek feldolgozásra (pl. vállalati jelentések szövegei) vagy csak az elmúlt években jelentek meg (pl. pénzügyi blogok). Ezen pénzügyi témájú adatoknál használható szövegelemzési módszerek és azok eredményeiknek egy széles merítésű, „felsorolásszerű" összefoglalóját adja Kumar és társa (2016): milyen módszertannal, milyen adatbázisokon, milyen területen (pl. részvényárfolyam-előrejelzése), milyen eredményre jutottak különböző kutatások.

A kifejezetten a részvény árfolyamok előrejelezhetőségével kapcsolatos cikkek szerzői is szerteágazó eszköztárat alkalmaznak. Talán a legegyszerűbb az általam is használt lineáris regresszió, aminek nagy előnye, hogy az eredmények könnyen értelmezhetőek. Az érdekesség kedvéért érdemes ugyanakkor megemlíteni néhány összetettebb módszert is, Bing és társai (2014) például harminc, a New York-i (NYSE) és az amerikai technológiai (NASDAQ) tőzsdén jegyzett vállalat részvényárfolyamát jelzik előre sikeresen a szociális hálón (pl. a Twitteren) elérhető bejegyzések alapján, ahol a szövegek tartalmi megértésére törekvő NLP *(Natural Language Processing)* nevű gépi technikával elemzik az üzenetek hangvételét. Külön érdekesség az, ahogy a szerzők a strukturálatlan, köznyelvi Twitter-bejegyzéseket megpróbálják a gépi elemzés számára felhasználható formába hozni.

Hagenau és szerzőtársai (2013) német és brit vállalati bejelentéseknek, a pénzügyi területen leginkább elterjedt, öntanuló SVM-módszerrel *(Support Vector Machine)* való elemzésével jelzik előre a részvényárfolyamok alakulását. Li

(2010) pedig a szintén elterjedt naiv Bayes-módszer alkalmazhatóságát mutatja be amerikai vállalatok éves és negyedéves jelentéseiből vett menedzsment előrejelzések és a vállalatok későbbi eredményei közötti kapcsolat feltárásánál.

A cégekkel kapcsolatos hírek felhasználhatóságát jól mutatja, hogy néhány éve megjelentek a nagy pénzügyi hírszolgáltatók fizetős termékei, amelyek valós időben elemzik a híreket és az így feldolgozott információt elérhetővé teszik a matematikai modellekre épülő, ún. algoritmusos alapok kezelőinek. A Bloomberg például a vállalati eseményekkel kapcsolatos, a hagyományos és közösségi médiában megjelenő híreket dolgozza fel folyamatosan, és teszi elérhetővé a befektetőknek, percenkénti frissítéssel.

Az én elemzési módszertanom előre definiált szótárakra épít, tehát mások által korábban összeállított szótárakban szereplő, hangvételt kifejező szavakat keresek a vállalati riportok szövegeiben. Ennek kulcsfontosságú szempontjait Loughran és McDonald (2016) foglalja össze, ahol a pénzügyi riportokkal kapcsolatos szövegbányászati elemzések módszertani lehetőségeit, a kutatás során fellépő nehézségeket és csapdákat, illetve további kutatási irányokat összegzik a szerzők. Szintén hasznos összefoglaló Loughran és McDonald 2011-es kötete, amelyben amerikai éves beszámolókat vizsgálnak szótáralapú módszerrel. Megállapításuk, hogy a szavak eltérő jelentése és mögöttes értelme miatt az általános, pl. Harvard-érzelemszótárak nem használhatók a pénzügyi riportoknál, (semleges vs. negatív jelentés). Ennek megfelelően saját szótárakat készítenek és ezek alapján – az eredmény szerint sikeresen – vizsgálják többek között azt, hogy a szövegekből kinyert többletinformáció segíti-e az árfolyam-előrejelzést. Az elemzéshez előkészített szótárakat hozzáférhetővé is teszik.

1. táblázat. A bemutatott szótáralapú megközelítést alkalmazó cikkek rövid összefoglalója

szerző	**cím**	**módszertan**	**adatbázis**	**függő változó**
Loughran, T., McDonald, B. (2011)	When is a Liability not a Liability?	szótáralapú, saját szótár (Loughran–McDonald)	éves jelentések	abnormális hozam és eredmény
Paul C. Tetlock, Maytal Saar-Tsechansky, Sofus Macskassy (2008)	More Than Words: Quantifying Language to Measure Firms' Fundamentals	szótáralapú (Harvard IV-4)	újságcikkek	árfolyam és elemzői előrejelzési hiba
Henry, E. (2008)	Are Investors Influenced By How Earnings Press Releases Are Written?	szótáralapú, saját szótár (Henry's)	telekommunikációs és IT-cégek jelentései	abnormális hozam

Forrás: Loughran (2011), Tetlock (2008) és Henry (2008)

A szótáralapú elemzésnél gyakran idézett szerző Henry (2008), aki többek között szintén saját, számviteli területre specializált szótárral vizsgálta az amerikai telekommunikációs és IT-cégek vállalati jelentéseinek hangvételét és az árfolyamok közötti kapcsolatot. Emellett Tetlock és társai (2008) az általános, tehát nem specializált, ún. Harvard-IV-4-szótár segítségével mutat ki szignifikáns kapcsolatot az újságcikkek hangvétele és a vállalatok negyedéves eredményei, illetve az árfolyamreakció között.

AZ ADATBÁZIS BEMUTATÁSA

A világ számos országában a tőzsdei kibocsátókról rendelkezésre álló információkat egy helyen, egységes szerkezetben teszik elérhetővé, így könnyítve meg a cégek elemzését. Ilyen – jogszabály által előírt – rendszer az amerikai tőzsdefelügyelet által működtetett EDGAR (*Electronic Data Gathering, Analysis and Retrieval system*), ahol a tőzsdei vállalatok közzétételei ingyenesen letölthetők. Magyarországon 2020-tól terveznek hasonló rendszert bevezetni, így korábban sem a Magyar Nemzeti Bank, mint felügyelet, sem a Cégbíróság vagy maga a tőzsde sem működtetett ilyen szolgáltatást, így a kutatásomhoz a riportokat még manuálisan kellett letölteni és az eredeti, változatos formátumú szövegeket feldolgozni.*

Ennek megfelelően én a négy cég honlapján, angolul elérhető negyedéves riportokat töltöttem le pdf formátumban 2010 I. és 2018 IV. negyedév között, majd alakítottam át az elemezhető txt formátummá. Ezek a riportok így felölelik a 2008/09-es világgazdasági válság utáni időszakot, a magyarországi szektoradók bevezetését, majd fokozatos mérséklését és a 2015/16-ban kezdődött hazai és régiós fellendülés idejét, üzletileg tehát egy változékony időszakról beszélhetünk. A riportokból nemcsak a menedzsment értékeléseket *(MD&A section)* használom, hanem a teljes, tehát a pénzügyi eredményeket részletező részeket is magába foglaló riportokat, mivel Loughran-McDonald (Loughran, 2011: 54.) alapján a kifejezetten a menedzsment helyzetértékelése nem tartalmaz több hangvételre utaló kifejezést.

Ehhez a Magyar Telekomnál – a nemzetközi gyakorlathoz hasonlóan, de hazai cégként egyedüliként – a negyedéves riportok közzétételét követő, a cég menedzsmentje (CEO és CFO) és a befektetők és elemzők közötti telefonos konferencia leiratait is hozzátettem. Ez utóbbiak „élőszavas" információnak tekinthetők, ráadásul csak a menedzsment értékelését tartalmazzák, az elemzői kérdéseket és interpretációt nem, így továbbra is csak a Magyar Telekom

* A Magyar Nemzeti Bank által működtetett kozzetetelek.mnb.hu oldalon elérhetőek a vállalati jelentések, de ezek nincsenek egységes, könnyen feldolgozható formátumba átdolgozva.

saját magáról megosztott, elsődleges információit foglalják magukba. Ez mindösszesen 179 angol nyelvű riportot jelent.

A kutatásban – sok szerzővel ellentétben – a sokkal hosszabb, tehát több szöveget tartalmazó éves jelentések helyett a rövidebb negyedéves riportokat elemzem. Ennek oka, hogy az éves jelentések gyakorlatilag az adott évre vonatkozó, már közzétett negyedéves jelentések összefoglalói, amelyek ráadásul hónapokkal később válnak elérhetővé, így az abban rejlő információk már ismertek a befektetők számára, tehát kisebb árfolyamreakció várható azok megjelenésekor.

A nyers szövegeket a gépi elemzéshez elő kell készíteni *(preprocessing)*, ami a gépek számára feldolgozhatóvá teszi a szöveget. Ennek során a szövegekből törölni kell a szövegközöket, központozást és a számokat. Emellett a mondatkezdő nagybetűket át kell átalakítani kisbetűvé, illetve a szavak végét a nyelvtanilag még értelmezhető formába kell levágni *(stemming)* ahhoz, hogy az eltérő alakú, de azonos értelmű szavakat a gép is azonosnak értelmezze. Az elemzés felgyorsításához szükség van a nem releváns szavak (*stopword*-ök) eltávolítására (pl. névelők, kötőszavak), mivel így kisebb, de relevánsabb adatbázison történik az elemzés. *(dimenziócsökkentés).*

A pénzügyi riportokban semleges hangvételűnek tekinthető és kétértelmű szavakkal (pl. crude, well, cost) kapcsolatban kettéválasztom az adatbázist. A szótáralapú elemzésnél ezeket a szavakat – a dimenziócsökkentés és az egyértelműség végett – kiszűröm. A Wordfish-nél ugyanakkor bent hagyom őket, mivel sok közülük iparág-specifikus kifejezés, így éppen a minták szétválasztásában segít.

Az adatbázis előkészítése során fontos paraméter még, hogy a releváns szavakat előfordulások gyakorisága alapján milyen súllyal vesszük figyelembe *(szavak súlyozása)*: például a ritkábban előforduló, ezért speciálisabbnak vélt szavakat kiemelten kezeljük-e. Ezzel kapcsolatban a szótáralapú elemzésnél Loughran és McDonald (2011: 57.) alapján az ún. *term weighting scheme*-et használom (*tf-idf*), ami a túl gyakori (adott szövegben & szövegek összességében) szavak súlyát tompítja a ritkábbak javára. A Wordfish-módszernél ugyanakkor – mivel ez maga is a szavak relatív gyakoriságát vizsgálja – a torzítást elkerülendő meghagyom a szavak eredeti súlyozását.

Az elemzés során az ún. szózsák *(bag-of-words)* módszert használom, ami a szavakat vagy szókapcsolatokat önálló egységenként kezeli és a közöttük lévő jelentésbeli kapcsolatot nem értelmezi. Ezen belül a szavakat egyesével vizsgálom és nem rövidebb, 2-3-as szókapcsolatokat (ún. *bi- és trigrammok*) elemzek. Ennek oka kettős. Egyrészt a szótárak is csak különálló szavakat tartalmaznak, másrészt pedig támaszkodom Loughran és McDonald (2016: 35.) megállapítására, miszerint – az élőnyelvvel ellentétben – a pénzügyi riportokban nem jellemző a dupla tagadás („not downgraded", „not terrible earnings").

Intő jel ugyanakkor a Magyar Telekom 2014. III. negyedéves riportjában található mondat („charging the transaction fee at the time of their introduction was *not* in *breach* of any law"), ahol a tagadás („not") egyértelműen semlegesíti a későbbi – a Loughran–McDonald-szótár alapján is – negatív kifejezést („breach"). Túlmutat a dolgozat keretein, de érdemes megemlíteni, hogy a legújabb kutatások már a szövegek gépi értelmezését tűzik ki célul, tehát azok mondatszintű megértésére törekszenek. Az én kutatásomhoz ugyanakkor elég az egyszerűbb, a szavak jelentését nem vizsgáló módszertan is.

A fenti előkészítés után megmaradó leggyakoribb szavakat mutatja az alábbi ábra.

1. ábra. A 179 egyedi riportból képzett szófelhő

Megjegyzés: a kétértelmű kifejezések kiszűrése után a minimum 2000-szer előforduló szavak. *Forrás:* saját készítésű ábra

A szótáralapú elemzésnél két, publikusan elérhető szótárt használtam annak érdekében, hogy az eredmények robusztusságát ellenőrizhessem:

- Loughran és McDonald [2011]. Ez az R-ban, a quanteda package-ben beépített, az internetről letölthető verzióval összevetve megegyező összetételű szótár, ami 6 hangvételre vonatkozóan tartalmaz szavakat (negative, positive, uncertainty, litigious, modal és constraining). Az elérhető hangvételek közül csak a negatív és pozitív szavakat használtam.
- Henry's [2008]. Ez szintén elérhető R-ban. A Loughran–McDonalddal összehasonlítva, ez a szótár lényegesen kevesebb, ráadásul csak pozitív és negatív kifejezést tartalmaz. A két szótárat ugyanakkor összehason-

lítva, csak részleges az átfedés a kettő között, így a Henry's-szótár használatának is van létjogosultsága, az eredmények robusztusságának ellenőrzése végett. Sőt, látni fogjuk, hogy a Henry's szótár erősebb magyarázó erővel bír.

2. táblázat. A Loughran–McDonald és a Henry's-szótárak összevetése

pénzügyi szótárak	Loughran–McDonald	Henry's
positive	354	104
negative	2355	85
uncertainty	297	-
constraining	184	-
litigious	903	-
superfluous	56	-
Összesen	4149	189

Forrás: Loughran (2011) és Henry (2008), R-ban letöltve

ELEMZÉS ÉS EREDMÉNYEK

A szövegek előkészítése után következett a hipotézisek vizsgálata, ahol először a szakértői véleménnyel, majd a piaci adatokkal vetettem össze a szövegbányászat eredményeit.

Szakértői vélemény

Saját tapasztalat alapján az elemzők között konszenzus szokott kialakulni az egyes cégek riportjainak és összességében a befektetőkkel való kommunikációjának minőségével kapcsolatban: mennyire transzparens a kommunikáció, van-e torzítás a hangvételben, rendszerint túlságosan optimista-e vagy éppen pesszimista-e a menedzsment a jelentések során, stb.

A szakértői feltevések összegyűjtéséhez a cégen belül, részvényelemző kollégám segítségét kértem. Ketten, egymástól függetlenül soroltunk fel 2-3 véleményt a cégek kommunikációjával kapcsolatban, majd összevetettem a két listát. Az eredmények alább láthatók, ahol szürkével emeltem ki a metszeteket (pl. Magyar Telekom – transzparens, Richter – negatív/konzervatív hangvételű). Az elemzésből származó eredményekkel való összevetéshez ezeket a metszet-véleményeket használom.

3. táblázat. Szakértői vélemények az egyes cégek riportjaival kapcsolatban

Magyar Telekom	
1. vélemény	**2. vélemény**
transzparens	transzparens
	nyitott kommunikáció
egyszerű szerkezetű	egyszerűség

OTP	
1. vélemény	**2. vélemény**
komplex	részletes információk
speciális szókincsű	optimista
	régimódi

Richter	
1. vélemény	**2. vélemény**
negatív hangvételű	konzervatív hangvételű
egyszerű szerkezetű	transzparens
	befektetőbarát

MOL	
1. vélemény	**2. vélemény**
komplex szerkezetű	konzervatív
nem transzparens	kevés információ
	nem befektetőbarát

Forrás: saját készítésű ábra a szakértői vélemények alapján

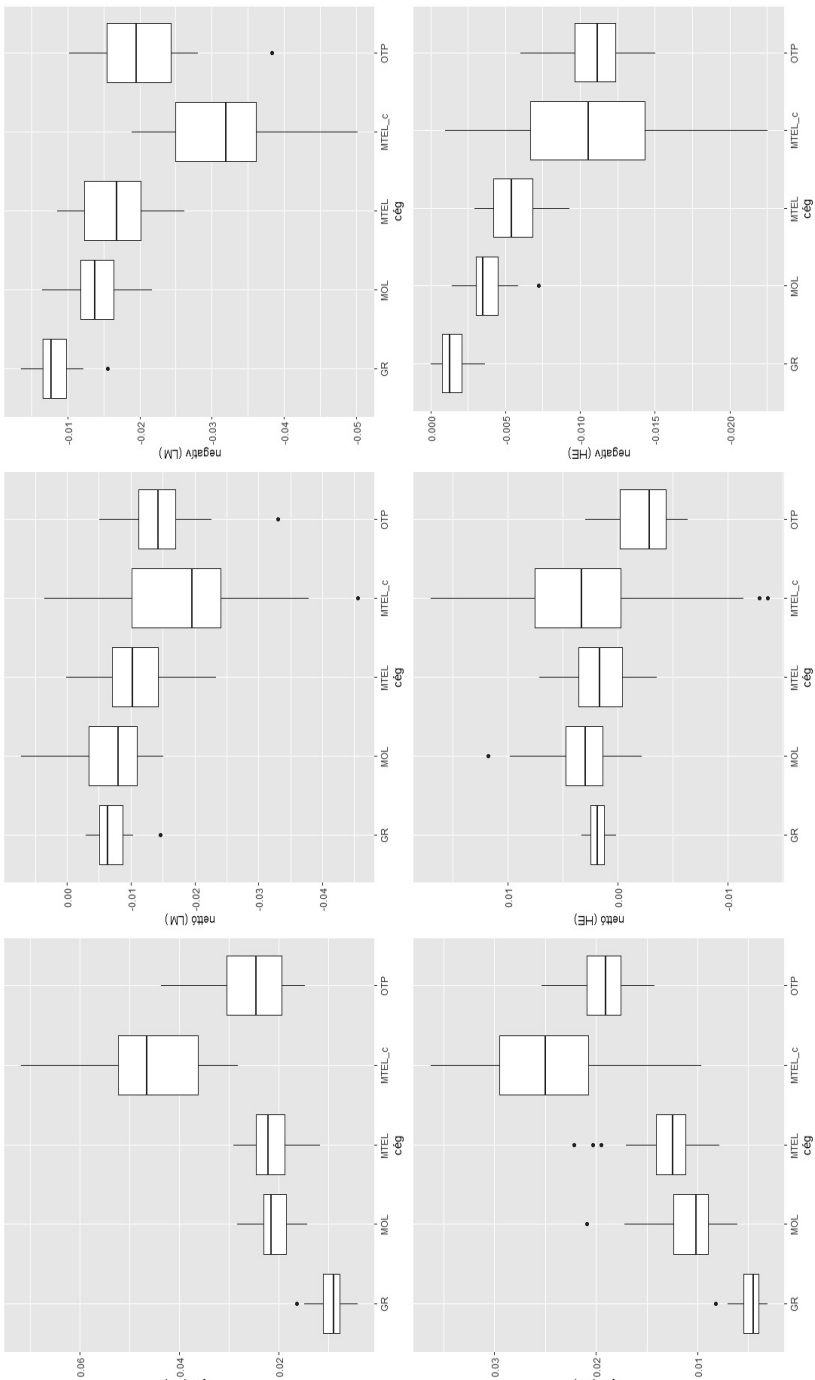
Ezek teszteléséhez első körben a szótár-módszert alkalmaztam, ahol a hangvétel-értékeket a teljes szókészletre arányosítva, százalékban jelenítem meg, ezzel küszöbölöm ki azt a torzítást, hogy a szövegek hossza, így a szavak előfordulási abszolút gyakorisága eltérő. Így az alábbi mutatókat képeztem (a hangvételre vonatkozó mutatók statisztikai leírását lásd a Mellékletben):

- teljes: pozitív és negatív összege
- nettó: pozitív és negatív különbsége
- negatív

Az eredményeket box-plottal ábrázolom *(2. ábra)*.

A grafikonból több megállapítás is leszűrhető. Jó hír, hogy a két szótár eredményeit egy-egy mutatóban (teljes/nettó/negatív), „függőlegesen" összehasonlítva, konzisztens eredményeket látunk, ami megerősíti a szótáralapú elemzés robusztusságát: a cégek egymáshoz viszonyított sorrendje a két szótárnál megegyezik. Miután azonban a Loughran–McDonald-szótárakban lényegesen több negatív szó szerepel, mint pozitív, ezért ennél a riportok hangvétele is negatívabb összehasonlítva a Henry's-szótárral, ahol a pozitív és negatív szavak aránya közel azonos.

Nem igazolódott, hogy a Richter riportjai relatíve negatívabb/konzervatívabb hangvételűek lennének. Sőt, ezek konzisztensen a legkevésbé negatív riportok (Loughran–McDonaldnál egyértelműen a nettó szentiment alapján), alacsony volatilitással. Ehhez esetleg az vezethetett, hogy a Richter a válság



2. ábra. Az egyes cégek riportjainak hangvétele (LM: Loughran–McDonald, HE: Henry's)

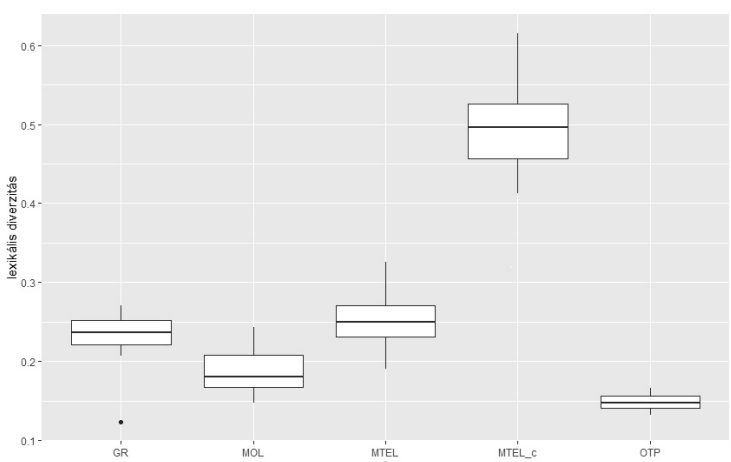
Forrás: saját készítésű ábra

időszakában is egy védettebb, a világpiaci folyamatoknak kevésbé kitett szektorban működött és a 2010-ben kivetett magyarországi szektoradókat is nagyrészt elkerülte. Igaz, itt csak a szövegek hangvételét elemzem és a konzervatív jelleg abban is megjelenhet, hogy a menedzsment szisztematikusan alulbecsüli a jövőbeli eredményeket.

A Magyar Telekom transzparens kommunikációjára több érték is utal. Egyrészt ezek a riportok tartalmazzák arányaiban a legtöbb hangvételre utaló kifejezést, ami segít a cégről valós képet kialakítani (ld. teljes hangvétel). Másrészt a telefonos konferenciák leirataiban (MTEL_c) – azon túlmenően, hogy a Magyar Telekom a négy cég közül egyedüliként teszi közzé ezeket, ezzel is segítve a transzparenciát – több hangvételre utaló kifejezés van, ami egybecseng azzal, hogy élőszövegek leiratáról beszélünk. Harmadikként, az OTP ebben a dimenzióban szintén transzparensnek tűnik, a hangvételre utaló kifejezések magas aránya miatt.

Lehetőség van a riportok lexikai diverzitásának elemzésére is, ami annak vizsgálatát jelenti, hogy hányféle szó szerepel egy-egy riportban. *(3. ábra).* Ez alapján egyrészt a Magyar Telekom és a Richter riportjaiban fordul elő a legtöbb önálló kifejezés, ezek tehát a legsokszínűbb jelentések. Kérdés, hogy ez mennyire az iparágspecifikus szakszavaknak, vagy egyéb, a cégek üzleti megítélését segítő kifejezéseknek köszönhető. Másrészt az OTP jelentései arányaiban monotonnak mondhatók. Ez valószínűleg azzal függ össze, hogy a cégnek közel tíz országban van banki leányvállalata, amelyek üzleti bemutatásához azonos szókincs is elégséges.

3. ábra. Lexikai diverzitás



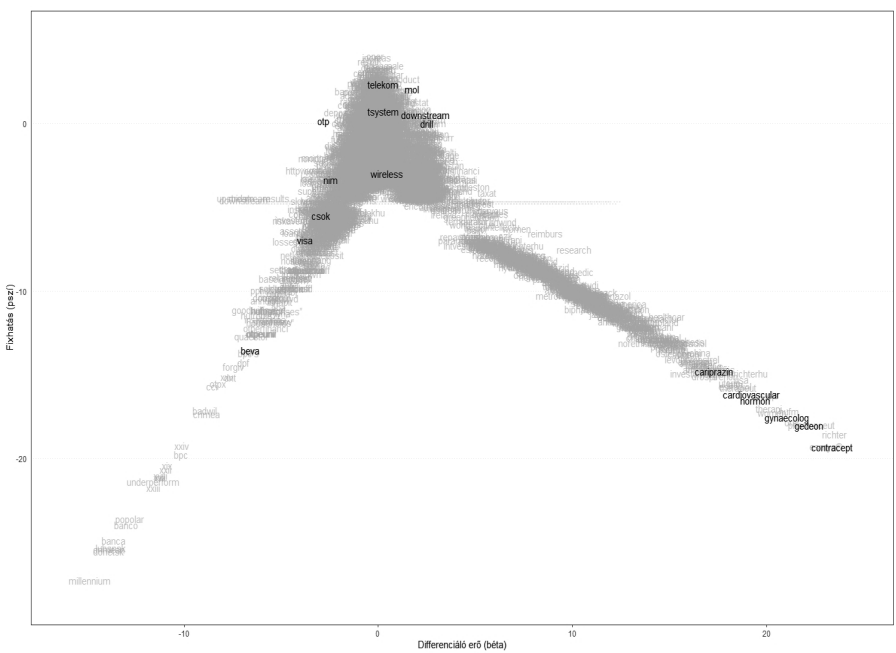
Forrás: saját készítésű ábra

Harmadrészt a Wordfish-módszerrel lehetőségünk van egy harmadik szempontból is megvizsgálni a riportokat és ezáltal tovább árnyalni a fenti megállapításokat. A Wordfish egy olyan vizsgálati módszer (bővebben lásd: Slapin, 2008), amely egy ún. felügyelet nélküli tanulást követ: nincs szükség a szövegek közötti struktúra feltárására és a csoportok előzetes definiálására, mert azt a rendszer maga alakítja ki. Ezt a vizsgálatot eredetileg politikai szövegek (pártprogramok, politikusi beszédek, stb.) jobb-bal, mint látens, skálán való osztályozásához használták, de a módszer maga alkalmazható tetszőleges szövegek közötti tartalmi távolságok, a bennük lévő szavak megkülönböztető erejének feltárására is, ahol a látens dimenziók alatt – mint látni fogjuk – jelen kutatásnál az iparágakat érthetjük.

A Wordfish eredményeinek értékeléséhez *(4. ábra)* az alábbi grafikon segít, amelynél a következő részeket érdemes megvizsgálni (mintha az Eiffel-toronyt néznénk):

- nulla közeli béta, magas pszí: ezek azok a szavak, amelyek cégtől függetlenül sok riportban előfordulnak és alacsony a megkülönböztető erejük (torony „teteje")
- szélsőséges béta, alacsony pszí: ezek a ritka, iparágspecifikus szavak (torony „lábai")

4. ábra: *A négy cég riportjainak Wordfishalapú elemzése*



Forrás: saját készítésű ábra R-ben Wordfish-módszerrel

Az ábrán a cégek néhány jellegzetes szavának kiemelésével megkíséreltem megmutatni, hogy az egyes cégek hol helyezkedhetnek el egymáshoz képest a térben. Ez alapján az OTP és a Richter sok szektorspecifikus kifejezést használ a riportjaiban, legalábbis arányaiban összehasonlítva a MOL-lal és a Magyar Telekommal, hiszen a Wordfish-módszer az OTP–Richter mentén vágja ketté a mintát. Az előző lexikai diverzitással összekapcsolva, az OTP esetében tehát az alacsony lexikai sokszínűség ráadásul magas szektorspecifikus szóhasználattal párosul. A MOL eredménye ugyanakkor érdekes: alacsony a szektorspecifikus szavak száma, miközben egy nagyon komplex, több nagyon különbözőképpen működő üzleti egységet (kitermelés, finomítás, értékesítés, vegyipar, stb.) magába foglaló vállalatról van szó.

Fontos megjegyezni ugyanakkor, hogy a Wordfish-módszer iparágakra való alkalmazása egy fontos elméleti problémát rejthet magában: a módszer ugyanis minden riportot és kifejezést egy egydimenziós rendszerbe kényszerít be (jelen esetben az OTP–Richter, tehát esetleg a bankszektor–gyógyszeripar által kijelölt skálába), miközben a négy cég négy nagyon különböző iparágat testesít meg. A Wordfish eredményei így fenntartással kezelendők és további kutatást érdemelnek.

PIACI ÖSSZEHASONLÍTÁS

A piaci adatokkal való összevetésnél azt vizsgálom, hogy a négy magyar cég negyedéves riportjaiból kinyerhető-e olyan információ, amivel a részvényárfolyamok rövid távú mozgására lehet következtetni. A negyedéves riportok választását az adja, hogy – tapasztalatok alapján – ezek tartalmaznak új, árfolyammozgató információt a piaci szereplőknek. A negyedéves riportokat jellemzően éjjel, tehát piacnyitás előtt teszik közzé. Ekkor a részvényelemzők többségében a másnap reggeli órákban, részben még piacnyitás előtt egy gyors áttekintést adnak a számokról és összevetik azokat a saját korábbi várakozásaikkal. Aznap pedig, jellemzően délután, az amerikai piacnyitásra időzítve egy 1-1,5 órás elemzői telefonos konferencia zajlik le, ahol a menedzsment (CEO, CFO) a számokat interpretálják, előrejelzéseket adnak és kérdésekre válaszolnak. Ezt követően, általában a következő napokban jellenek meg a végleges elemzői vélemények, frissített várakozásokkal és így válik teljessé az adott cég negyedéves jelentésének feldolgozása.

Az elemzés célja annak a hipotézisnek a tesztelése, hogy a szótárak segítségével lehetséges-e a jelentéseket hangvételük (tone) alapján (pozitív/optimista vs. negatív/pesszimista) gyorsan, a riportok publikálását követően elemezni. Optimális esetben az adott negyedéves riport így a számszaki adatokon túlmenően, ráadásul elfogulatlanul, tehát a cégről alkotott előzetes piaci véle-

ménytől mentesen („mindig túlzottan konzervatív/optimista a menedzsment") értékelhető.

A teszthez regressziós modellt használtam, ahol a szótárak alapján készített hangulat indikátoroknak a részvényárfolyamokkal kapcsolatos magyarázó erejét vizsgáltam. Az árfolyamok változásánál, mint függő változónál felhasználtam az elmozdulás abszolút értékét, az általános piaci trendektől megtisztított, ún. abnormális hozamot és a volatilitást.

A volatilitás, mint metaadat felhasználása külön magyarázatot érdemel. A piaci adatokkal való összehasonlításnál a legpontosabb összevetést az árfolyamreakcióval való összevetés adja, ami része az elemzésnek. Emellett ugyanakkor az árfolyamok volatilitását is érdemes vizsgálni, ami egyrészt egyszerűbben mérhető, mivel nem függ az árfolyam-elmozdulás irányától, másrészt elméletileg összefügg a piacnak a cégről alkotott ítéletével (negatív vélemény ~ nagyobb bizonytalanság a cég jövőjével kapcsolatban ~ nagyobb árfolyamvolatilitás). A volatilitásnál fontos pontosítani az adatmérés időpontját, hiszen az elemzés szempontjából nem a negyedév vége (például 2010. 03. 31.) számít, hanem a riport közzététele, ami tipikusan 30-60 nappal követi a negyedév végét (jelen példánál 2010. május). Ennek megfelelően én – némi egyszerűsítésként – a negyedévet követő 60. napon mérem meg a 30 napos volatilitást (megint, jelen példánál 2010 májusának árfolyammozgásaiból számolt volatilitást).

A szótárak alapján készített hangulat indikátorok mellett a modellbe számos, piaci és elemzői adatot, mint kontrollváltozót is beépítettem. Ezek három csoportba sorolhatók (a változók részletes leírását lásd a Mellékletben):

- Az elemzői meglepetés azt szűri ki, hogy a riportban közölt tényadatok – jelen esetben árbevételi számok – mennyiben térnek el az elemzők várakozásától és ez hogyan hat az árfolyamra és a volatilitásra: a várakozások szerint minél nagyobb a pozitív meglepetés, annál nagyobb a pozitív reakció az árfolyamban.
- Az előző napi hozam az árfolyammozgásban lévő momentumot kontrollálja: az árfolyamok sokszor autoregresszív folyamatot követnek, így az előző napi elmozdulás hat az azt követőre.
- Ugyanígy egy-egy cég vizsgálatánál érdemes a piacok általános trendjét is figyelembe venni és csak az attól eltérő, ún. abnormális, tehát az adott cégre jellemző hozamot vizsgálni.

Az eredményeket az *4.* táblázat foglalja össze.

4. táblázat. A lineáris regressziók eredményei

	abszolút hozam (t_1/t_0)	abnormális hozam (t_1/t_0)	30 napos volatilitás		
konstans	0,003	0,005	18,727***	18,121***	20,073***
	0,416	0,923	5,777	6,103	6,891
LM-negatív	0,981	0,930	–241,598	–28,825	
	–1,773	–1,893	0,894	0,143	
LM-nettó	0,636	0,527	134,403	–209,301	
	1,169	1,093	0,507	–1,005	
HE-negatív	1,558*	1,353	–138,408		–393,219
	2,000	1,962	0,363		-1,345
HE-nettó	0,853	0,708	–628,858*		–490,747*
	–1,433	–1,349	–2,156		–2,248
elemzői meglepetés (árbevétel)	0,048*	0,048*	–3,985	–1,908	–4,257
	1,995	2,248	0,338	0,158	0,366
abnormális hozam (t_0/t_{-1})		0,068			
		0,561			
abszolút hozam (t_0/t_{-1})	0,070	0,073			
	–1,039	0,636			
dummy (Richter Gedeon)	0,004	0,007	5,490	4,467	4,588
	0,709	1,406	1,897	1,610	1,656
dummy (MOL)	0,005	0,002	6,791**	6,183**	6,450**
	0,971	0,556	2,960	2,780	2,887
dummy (OTP)	0,003	0,006	6,672**	9,536***	5,745*
	0,570	1,294	2,738	4,195	2,517
korrigált R^2	3,5%	3,2%	15,4%	10,5%	15,8%

Megjegyzés: A * 5%-os, a ** 1%-os, a *** 0,1%-os szignifikancia-szintet, a t_0 pedig a riport közzétételének napját jelenti. *Forrás:* saját készítésű táblázat

Az eredményeket több szempont alapján értékelhetjük. Legfontosabb megállapítás, hogy érdekes módon a sokkal egyszerűbb, kevesebb szót tartalmazó Henry's szótár tűnik csak alkalmazhatónak, legalábbis ezen az adatbázison. Egyrészt a nettó vagy negatív HE-indikátorok szignifikáns magyarázó erővel

bírnak és a kapcsolat iránya is mindkét esetben megfelelő: az árfolyamnál a negatívabb hangvétel csökkenti a pozitív árfolyamreakciót, a volatilitásnál pedig növeli az árfolyam volatilitását. A Loughran–McDonald-szótár nem bír szignifikáns magyarázó erővel.

Emellett az árfolyamoknál, mint függő változóknál a csak negatív hangvételt mérő indikátorok használhatóbbnak tűnnek a pozitív tónust is magába foglaló nettó indikátorokhoz képest: az előbbiek magyarázó ereje vagy túllépi vagy megközelíti az 5%-os szignifikancia-értéket. A volatilitást vizsgálva ugyanakkor a nettó hangvétel tűnik alkalmazhatónak. Ez megint visszatükrözi Loughran és McDonald eredményeit (2016: 35.), akik csak a negatív szavakra koncentrálnak, mert a pozitív szavak torzítanak: a vállalati vezetők – a valós élethez hasonlóan – a negatív hírt sokszor pozitív kontextusba ágyazzák, főleg élőszóban.

A modellek magyarázó ereje ugyanakkor alacsony: az árfolyamok vizsgálatánál 3-4% körüli, ami egybecseng Loughran és McDonald (Loughran, 2011: 52.) eredményeivel. A volatilitás elemzésénél is csak 10-16%-os magyarázó erőt sikerül elérni. Az árfolyamok volatilitásánál ráadásul egy fontos endogenitási kérdés is felmerülhet: a negatív piaci hangulat és az ezzel együttjáró magasabb volatilitás ugyanis hathat a menedzsmentre, akik így negatívabb hangvételű riportot adnak ki vagy a telefonos konferencia alatt negatívabb hangvételű véleményt mondanak a cégről. Az árfolyamnál (abszolút vagy a referenciaindexhez viszonyított abnormális hozam esetében) lehetőség van az endogenitás részleges kiszűrésére, hiszen a magyarázó változók között az előző nap árfolyammozgását is figyelembe veszem.

Összességében tehát részben igazolni tudjuk a hipotézist, miszerint a riportok szótáralapú elemzésének a hazai környezetben is van létjogosultsága: pusztán a szöveges dokumentumokból is hasznos információ nyerhető ki. Az eredmények valós körülmények közötti alkalmazhatóságát részletesebben egy kereskedési modellel lehetne tesztelni, aminél például a pozitívabb hangvételű riportot közlő cégek részvényeinek megvételével és a negatívabbak eladásával lehetne az így elérhető profitot vizsgálni. Ez azonban túlmutat a jelenlegi elemzés keretein.

ÖSSZEFOGLALÁS, TOVÁBBI KUTATÁSI IRÁNYOK

Az elemzés a négy magyar blue-chip negyedéves jelentéseinek szövegszerű elemzésével, a szótáralapú és a Wordfish-módszerrel próbált választ adni a dolgozat elején feltett hipotézisekre, miszerint mennyire igazolhatók a szakértők általános véleményei a cégek kommunikációjával kapcsolatban, valamint van-e egyértelmű kapcsolat a riportok hangvétele és a piaci adatok, konkrétan az árfolyamreakció és a volatilitás között. Az elvégzett elemzések részben visz-

szaigazolták ezeket a hipotéziseket, így a módszertan alkalmazhatónak tűnik a hazai környezetben is:

- A Magyar Telekom valóban transzparens a szöveges kommunikációját tekintve és ezt a piaci adatokkal való összevetés is visszaigazolja.
- Az OTP inkább a sok azonos leánycégről kommunikál, így sok iparágspecifikus szót használ, de ezek nem a transzparenciát szolgálják. A hangvételt tekintve sok ilyen jellegű kifejezés megjelenik az OTP-riportokban, de ezek használhatóságát a piaci adatokkal való összevetés nem igazolja vissza.
- A MOL valóban „elbújik" az elemzők elől: kevés hangvétellel, alacsony diverzitással, „monotonon"kommunikál és a szektorspecifikus leírásban sem erős, miközben egy komplex iparágról beszélünk.
- A Richter inkább iparágspecifikus szavakat használ, hangulatmentesen. Szigorúan a szövegre koncentrálva nem igazolódott a cég konzervatív kommunikációjáról alkotott elemzői vélemény.

A piaci adatokkal való összevetés szintén visszaigazolja a módszer hazai alkalmazhatóságát. A külföldi eredményekhez hasonlóan a szótáralapú hangvétel, különösen a negatív tónust mérő indikátorok szignifikáns magyarázó erővel bírnak az árfolyamok elmozdulását illetően.

Fontos korlátozás ugyanakkor, hogy a modellek magyarázó ereje önmagában alacsony, így valószínűleg a minta kiterjesztésével lehet a kapcsolat erősségén javítani. Erre számos lehetőség adódik. A korábbi, válság előtti adatokat is be lehet vonni például az elemzésbe, ugyanis a 2008/09-es válság alatti riportok esetleg hangvételüket tekintve markánsabbak. Emellett egyéb, az árbevétellel kapcsolatos elemzői meglepetés szignifikanciáját nézve hasznos kontrollváltozók is beépíthetők a modellbe, amelyek jelenleg részben a hiányos adatforrás miatt estek ki (egy részvényre jutó nyereség, értékeltség, stb.).

Érdekes lehet olyan magyar cégek elemzésénél is felhasználni a kvalitatív szövegelemzési eszköztárat, amelyek egyelőre kiesnek az elemzők látóköréből vagy csak magyar nyelven adnak ki jelentéseket. Ez utóbbihoz természetesen magyar nyelvű szótárat kellene készíteni, illetve kezelni kellene a jelen dolgozatban használt angol és a magyar nyelv közötti nyelvtani különbségeket. Továbbá jelen kutatás csak napvégi, záró árfolyamokat vizsgál, miközben az elemzői információk a jelentés publikálásának napján is folyamatosan érkeznek, befolyásolva ezzel a piacok reakciót, így napon belüli *(intraday)* adatok bevonásával pontosítani lehet az egyes információforrások hatását.

A tanulmány számos továbblépései lehetőséget kínál. Felmerülhet például egy olyan endogenitási probléma, hogy ha a menedzserek tudják, hogy milyen negatív szavak számítanak, akkor próbálják elkerülni azokat. Érdekes lehet az ilyen szavak előfordulásának időbeli vizsgálata. Lehetséges az eredmények robusztusságának további vizsgálata, szélesebb validálása is, például a kettő- és magasabb fokú szókapcsolatok (*bi- és trigrammok*) elemzésével, amivel az esetlegesen előforduló tagadások hatása is figyelembe vehető. Másrészt érdekes lehet

a hangvételben érzékelhető időbeli változások vizsgálata is, ugyanis a válságot közvetlenül követő, szektoradókkal terhelt időszakot követően az utóbbi években javult a vállalatokat körülvevő üzleti környezet, amivel összevethető a riportok hangvételének esetleges javulása. Továbbá a szótárak nagyon merev megközelítésével szemben a felügyelt tanulás módszere javíthatja a módszertan alkalmazhatóságát, ahol a riportok kézi kódolása után az erre indított gépi tanulási folyamat maga alakítja ki az eredményt. Végső soron pedig egy kereskedési modellel, egy profitfüggvény segítségével tesztelhető, hogy a szótáralapú vagy más hangvétel-elemzés valós körülmények között mennyiben használható.

Összességében megállapíthatjuk, hogy a szövegbányászatnak van relevanciája a magyar vállalatok akár közpolitikai, akár befektetői elemzésénél, így más országok gyakorlatához hasonlóan egy központi, egységes formátumú adatbázis kiépítése előrelépés lenne a piaci szereplők hatékony, transzparens és egyenrangú tájékoztatása érdekében. Emellett, ha a cégek a Magyar Telekom példáját követve a telefonos, elemzői konferenciák szöveges leiratait is – szintén akár egy közös felületen – elérhetővé tennék, még árnyaltabb vizsgálatokra nyílna lehetőség, ami szintén hozzájárulna a hatékonyabb kommunikációhoz.

MELLÉKLET

A riportok hangvételét mérő, szótáralapú adatok (Loughran–McDonald és Henry's, negatív és pozitív) mellett az alábbi metaadatokat használtam az elemzéshez:

- *ID*: negyedéves riport sorszáma (1–179)
- *company*: cég neve
- *date*: melyik negyedéves időszak
- *announc_date*: a riport közlésének pontos dátuma
- a négy cég tőzsdei záróára a riport megjelenése napján, előtte és utána egy-egy nappal
- a közép-európai régiót lefedő CECE-tőzsdeindex árfolyama a céges riportok megjelenésének napján, előtte és utána egy-egy nappal, forintra átszámolva. Annak oka, hogy nem a magyar tőzsdeindexet (BUX) használtam, az, hogy abban a blue-chipek, de különösen az OTP és MOL nagyon markáns súlyt tesznek ki. A CECE-indexben ugyanakkor a négy cég közül legnagyobb súllyal is az OTP szerepel, 10,9%-kal (2019. aug.), így az index jól használható az abnormális hozam torzítatlan vizsgálatára.
- *surp_rev* és *surp_eps*: az árbevételnek és az egy részvényre jutó nyereségnek, a riportokban közölt tényadata és az elemzői várakozások közötti százalékos különbség (ún. analyst surprise), Bloombergről letöltve. Az egy részvényre jutó nyereséggel kapcsolatos értéket végül nem használtam, mert rendkívül hiányos, így nagyon leszűkítette volna az adatbá-

zist. Fontos még megjegyezni, hogy a Bloombergen sok esetben csak 1-2 elemző várakozása érhető el, így ez a kontrollváltozó nagyon zajos.
- *vol_30d*: 30-napos historikus volatilitás, negyedéves riport közlése körül
- dummy változók:
  o cég-dummyk
  o confcall-dummy: riportot követő telefonos menedzsment confcall dummy-ja (csak a Magyar Telekomnál: MTEL_c)

A fenti változók leíró statisztikáját az alábbi táblázatok mutatják:

5. táblázat. Riportok hangvétele és a közöttük lévő korreláció

	nettó hangvétel		negatív hangvétel		pozitív hangvétel	
	LM	HE	LM	HE	LM	HE
min	0,01%	–1,36%	0,35%	0,00%	0,06%	0,23%
átlag	1,04%	0,17%	1,57%	0,51%	0,54%	0,68%
medián	0,88%	0,18%	1,33%	0,39%	0,50%	0,59%
max	4,55%	1,18%	5,03%	2,25%	2,33%	1,95%

		LM			HE		
		nettó	negatív	pozitív	nettó	negatív	pozitív
LM	nettó	1,00	0,84	0,05	0,42	0,59	0,18
	negatív	0,84	1,00	0,58	0,15	0,68	0,57
	pozitív	0,05	0,58	1,00	0,36	0,37	0,78
HE	nettó	0,42	0,15	0,36	1,00	0,57	0,46
	negatív	0,59	0,68	0,37	0,57	1,00	0,47
	pozitív	0,18	0,57	0,78	0,46	0,47	1,00

Forrás: saját készítésű táblázat, Loughran (2011) és Henry (2008) alapján

6. táblázat. Piaci adatok

	árfolyamváltozás		abnormális hozam		elemzői meglepetés	
	adott/ előző nap	következő / adott nap	adott/ előző nap	következő / adott nap	árbevétel	egy részvényre eső eredmény
min	0,00%*	0,00%*	–7,75%	0,04%	0,07%	–261,21%
átlag	1,67%	1,29%	–0,12%	1,22%	4,10%	28,68%
medián	1,03%	0,97%	0,01%	1,01%	2,82%	23,89%
max	7,80%	8,97%	4,17%	4,76%	40,14%	112,64%

Megjegyzés: A *-gal jelölt adatok ellenőrizve lettek, nem történt árfolyamváltozás a két nap között.
Forrás: saját készítésű táblázat, Bloomberg alapján

IRODALOM

Aizenman, J.–Binici, M.–Hutchison, M. M. (2014): The transmission of Federal Reserve tapering news to emerging financial markets. *National Bureau of Economic Research. NBER Working Paper,* No. 19980, Issued in March 2014 (https://doi.org/10.3386/w19980).

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

Bing, L.–Chan, K. C. C.–Ou, C. (2014): Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements. *2014 IEEE 11th International Conference on e-Business Engineering,* Guangzhou, 2014, 232–239. https://doi.org/10.1109/icebe.2014.47

Born, B.–M. Ehrmann–M. Fratzscher (2014): Central bank communication on financial stability. *The Economic Journal*. Vol. 124, Issue 577, 701–734. https://doi.org/10.1111/ecoj.12039

Grimmer, J.–Stewart, B. M. (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis,* Vol. 21, Issue 3, 267–297. https://doi.org/10.1093/pan/mps028

Hagenau, M.–Liebmann, M.–Neumann, D. (2013): Automated news reading: Stock Price Prediction based on Financial News using Context-capturing Features, *Decision Support Systems,* Vol. 55, 685–697. https://doi.org/10.1016/j.dss.2013.02.006

Henry, E. (2008): Are Investors Influenced By How Earnings Press Releases Are Written? *Journal of Business Communication*, Vol. 45, Issue 4, October 2008, 363–407. https://doi.org/10.1177/0021943608319388

Kumar, B. S.–Ravi, V. (2016): A survey of the applications of text mining in financial domain. *Knowledge-Based Systems,* Vol. 114, 128–147. https://doi.org/10.1016/j.knosys.2016.10.003

Li, F. (2010): The Information Content of Forward-Looking Statements in Corporate Filings. A Naïve Bayesian Machine Learning Approach, *Journal of Accounting Research*, 2010, Vol. 48, Issue 5, 1049–1102. https://doi.org/10.1111/j.1475-679x.2010.00382.x

Loughran, T.–McDonald, B. (2011): When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, Vol. LXVI, No. 1, 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T.–McDonald, B. (2016): Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, Vol. 54, Issue 4, 1187–1230. https://doi.org/10.1111/1475-679x.12123

Loughran, T.–McDonald, B. a szótárakat hozzáférhetővé teszik. Letöltés helye: *https://sraf.nd.edu/textual-analysis/resources/* Letöltés ideje: 2019. 2. 24

Poole, W.– Rasche, R. H. (2003): The impact of changes in FOMC disclosure practices on the transparency of monetary policy: are markets and the FOMC better "synched"? *Federal Reserve Bank of St. Louis Review*, Vol. 85, No. 1. https://doi.org/10.20955/r.85.1-10

Sebők, M. (szerk.) (2016): *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban*. Budapest, L'Harmattan.

Slapin, J. B.–Proksch, S. (2008): A Scaling Model for Estimating Time-Series Party Positions from Texts, *American Journal of Political Science*, Vol. 52, No. 3, 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x

Tetlock, P. C.–Saar-Tsechansky, M.–Macskassy, S. (2008): More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, Vol. 63, Issue 3, 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x