PÉCSI TUDOMÁNYEGYETEM
KÖZGAZDASÁGTUDOMÁNYI KAR

GAZDÁLKODÁSTANI DOKTORI ISKOLA

Kovács Balázs

# Tőzsdei hírbányászat a magyar részvénypiacon

DOKTORI ÉRTEKEZÉS TÉZISEI

Témavezető: dr. Kruzslicz Ferenc

Pécs, 2017

# Tartalomjegyzék

**Tőzsdei hírbányászat a magyar részvénypiacon**

Kovács Balázs

Témavezető: dr. Kruzslicz Ferenc

## Absztrakt

A dolgozat a szöveges formában megjelent információknak a magyar tőzsdei részvényárfolyamokra gyakorolt hatását vizsgálja szövegbányászati módszertan segítségével. A disszertáció az empirikus művek közé sorolható, melynek hipotézisei a hírekből kinyerhető információk és az árfolyamokban megnyilvánuló információk közötti kapcsolatra vonatkoznak. A dolgozat hozzáadott értékét leginkább a kétnyelvű vizsgálatok, a saját eredmények robusztusságának vizsgálata a különböző paraméterek és szövegreprezentációk megválasztására, valamint az időbeliség vizsgálata jelentik. A hipotézisek teszteléséhez a tőzsdei hírbányászati modell hozamosztályozó változatát használtam, melynek bemeneteit a BÉT prémium kategóriás részvényeihez kapcsolódó, 2014.07.01 és 2015.06.31 közötti sajtóközlemények szövegei képezik, outputját pedig egyperces lépésközökkel a közlemény publikálásának ideje és a hozzá képest legfeljebb 120 perccel eltolt időpont közötti hozam nagysága alapján képzett hozamkategória – negatív, semleges, pozitív. A hírek szövegének numerikus reprezentációi alapján nemlineáris SVM-osztályozókat tanítottam a különböző méretű tanítómintákon, melynek pontosságát 10-szeres keresztvalidációval ellenőriztem. A különböző eredmények összehasonlításához a 10-szeres keresztvalidáció során kapott átlagos pontosságot használtam. A szöveges előrejelzés pontosabbnak bizonyult a defaultnál, ugyanis az eredményeim szerint az összes paraméterkombináció 94,64%-a esetében szignifikáns volt az eltérés 1%-on. Az optimális becslési időtáv a hírbányászati feladatra a publikálás előtt 27 perc, a publikálás után pedig 19–22 perces tartományban van, tehát némi eltérést tapasztaltam Gidófalvi ±20 perces eredményéhez képest. Ez alapján tehát az információ a publikálás előtti kb. fél órában kezd beépülni a vizsgált részvények árfolyamába, majd ez a publikálást követő kb. 20 percig tart. Mivel a közzétételi folyamat kb. egy óráig tart, ezért az ehhez kapcsolódó eredményekből az a következtetés is levonható, hogy nem lehet jó modellt készíteni a folyamat kezdete elő visszanyúló időablakra. Azt tapasztaltam továbbá, hogy az azonos sajtóközlemények angol és magyar nyelven közzétett változataival készített modellek pontossága között nincs szignifikáns különbség. Nagyon szigorúan véve a magyar nyelvű korpusz kissé pontosabb becslésre adhat lehetőséget. Az optimális eredmények elég robusztusak az alkalmazott SVM osztályozási módszer C-gamma paraméterkombinációira nézve, de kb. 1%-nyi eséllyel visszaeshet a default szintre a pontosság. Az általam vizsgált egyik szövegreprezentáció sem mutatkozott sokkal jobbnak a probléma megoldására, de szigorúbban véve megállapítható, hogy az egyszerűbb reprezentációt alkalmazó modellek pontosabbak.

**Tárgyszavak**: hír, információ, részvényárfolyam, BÉT, előrejelzés, osztályozás, szövegbányászat, támasztóvektor-gép, érzékenységvizsgálat

**JEL-kódok**: C38, C45, C51, C52, C53, C55, C58, G12, G14, G17

# 1. A témaválasztás indoklása, a kutatás céljai

Az 1990-es évek végén Wüthrich alkalmazta először a hírbányászatot a tőzsdén, ő és szerzőtársai a tőzsde nyitásáig megjelent hírek alapján jelezték előre a tőzsdeindexnek a tőzsde zárásakor várható értékét (Wüthrich et al. 1998). Egy-két évvel később Lavrenko már az individuális hírek és individuális részvények árfolyama közötti kapcsolatot modellezte, az előrejelzési időtáv pedig 0-tól 10 óráig terjedt (Lavrenko et al. 2000). Thomas és Sycara (2000) hírek helyett tőzsdei fórumok hozzászólásaival kísérletezett, és genetikus algoritmussal hangolt szabályalapú rendszerével a következő napra készített előrejelzést. Gidófalvi (2001) az eseményvizsgálat módszertan elemeit építette be a szövegbányászati modellbe, ugyanakkor ezt individuális hírek szintjén tette. Az előrejelzés időtávjának megválasztását is vizsgálta, eredményei alapján a $\pm 20$ perces eseményablak volt a legmegfelelőbb. Koppel és Shtrimberg (2004) a hírek pozitív, illetve negatív hangulatát tanulmányozta, és az ezt meghatározó kifejezésekkel is foglalkozott. A 2000-es évek közepén munkálkodott a témában Mittermayer (2004), aki sajtóközleményekből álló korpusszal végezte vizsgálatait, mivel ez olyan típusú információforrás, amely közzétételben megelőz másokat. Ezzel nagyjából egy időben az e-Markets Group nevű kutatócsoport vizsgálta a modellt a devizaárfolyamok előrejelzésére, és a fontos szakszavak kinyerése érdekében a hírkorpusz szavait egy általános korpusszal összevetve értékelte (Zhang et al. 2005). 2006-tól publikált a témában Schumaker és Chen (2006), aki a lehetséges szövegreprezentációk alkalmazásának hatását vizsgálta, konzisztensen mindig ugyanazt a korpuszt és 20 perces eseményablakot használva. Ezen kívül az előrejelzésre épített kereskedési stratégia jövedelmezőségét vizsgálta még részletesebben, összehasonlítva más stratégiák eredményével. 2008-tól Groth a német sajtóközleményekkel modellezett (Groth & Muntermann 2008), az árfolyam-előrejelzés mellett a volatilitás előrejelzésével foglalkozott, illetve kétnyelvű – angol–német – összehasonlításokat végzett 2014-ben (Groth et al. 2014).

A kutatási irányok áttekintése és rendszerezése után alakulhatott ki a dolgozat konkrét célkitűzése, hogy a szakirodalomhoz hozzáadott értékel bíró kutatási kérdések fogalmazódjanak meg. Úgy vélem, hogy a vizsgálatok nem elég alaposak az előrejelzés időtávja tekintetében. Schumaker például legtöbb esetben csak a 20 perces időtávot vizsgálta meg, amit Gidófalvi (2001) eredményeivel indokolt, aki 5 perces lépésközökkel vizsgálódott. Szükségesnek tartottam, hogy finomabb vizsgálatnak vessem alá azt az időtávot, amelyet biztosítani kell, hogy a hírek hatásukat kifejthessék.

A devizaárfolyamokkal végzett szövegbányászati kísérleteim során pedig azt tapasztaltam, hogy nagyon sokféle hír befolyásolhatja azok árfolyamát, és a híraggregáló szolgáltatások sem tudják feltétlenül az összeset összegyűjteni. Másfelől pedig ezek a fajta hírek tipikusan tele vannak régebbi információkra való hivatkozással, összefüggések keresésével, és részben ezekből fakadóan az új információhoz kapcsolódó esemény bekövetkezési idejéhez képest véletlenszerű késedelemmel kerülnek publikálásra. Ez jelentősen megnehezíti, hogy a hírt melyik időponthoz kell rendelni. A Mittermayer (2004) által használt angol és a Groth és Muntermann (2008) által használt német sajtóközlemények szolgáltak mintául ahhoz, hogy a magyar tőzsde sajtóközleményei mellett döntöttem. A BÉT tőzsdeszabályzata szerint ugyanis sajtóközleményeket azonnal közzé kell tenni a tőzsde honlapján, amely ennél korábban máshol nem jelenhet meg.

Groth et al. (2014) német és angol nyelvű párhuzamos korpusszal folytatott kutatásai nyomán vállalkoztam a közlemények magyar és angol nyelvű megfelelőivel kapott eredmények összevetésére.

Az osztályozó módszer beállítási lehetőségei jelentősen befolyásolhatják az eredményeket az adatok eloszlásától függően, ezért úgy vélem, egy-két paraméterkombináció vizsgálata alapján nem jelenthető ki egyértelműen, hogy a szöveges információk és az árfolyam közötti összefüggések a véletlennél jobban modellezhetők. A modell alkalmazhatóságát nem csak a paraméterek befolyásolják, hanem az adatok eloszlása is, amelyet pedig a magyarázó változók megválasztása határoz meg. Schumaker (2009; 2010a; 2010b) több ízben vizsgálta a különböző szövegreprezentációk hatását a modell teljesítményére, de rögzített modellparaméterek mellett tette ezt, ezért szükségesnek éreztem egy nagyobb paraméterhalmazon ellenőrizni a szövegjellemzők hatására bekövetkező teljesítményváltozásokat. Ezen kívül ezeket a vizsgálatokat az is indokolttá teszi, hogy a modellet a kísérletekhez kalibrálni kell.

## 2. Az értekezés hipotézisei

A dolgozat a szöveges formában megjelent információknak a magyar tőzsdei részvényárfolyamokra gyakorolt hatását vizsgálja szövegbányászati módszertan segítségével. A disszertációt az empirikus művek közé sorolnám, melynek hipotézisei a hírekből kinyerhető információk és az árfolyamokban megnyilvánuló információk közötti kapcsolatra vonatkoznak:

**H1**: A sajtóközlemények befolyásolják a részvényárfolyamot és a szövegük felhasználásával az a priori valószínűségnél – default modellnél – nagyobb pontosságú előrejelzés készíthető a BÉT prémium kategóriás részvényeire.

**H2**: A magyar tőzsdei sajtóközlemények haladéktalanul közzétett, új információt hordoznak. Nem lehet jó minőségű, illetve robusztus hírbányászati modellt készíteni olyan időablakra, amely korábbra nyúlik vissza, mint az az időtartam, amit a hír a KIBINFO közzétételi folyamatban eltölt.

**H3**: A modell robusztussága és minősége az időablakkal változik, és ennek optimuma meghatározható.

**H4**: Az azonos sajtóközlemények angol és magyar nyelven közzétett változatai egyformán alkalmasak a hírek árfolyamra gyakorolt hatásának vizsgálatára. Az információ nyelvi kódolásának nincs különösebb jelentősége.

**H5**: Az eredmények robusztusak az alkalmazott SVM osztályozási módszer beállításaira nézve. Egy – a legpontosabb modellhez viszonyítva – szignifikánsan jó modell paramétereinek kis megváltozásával másik szignifikánsan jó modellhez lehet jutni.

**H6**: Az eredmények robusztusak a szöveges inputok körének megválasztására. A szakterület-specifikus kifejezések azonosítása és a sablonszerű szövegrészek eltávolítása nem befolyásolják számottevően a pontosságot a szöveg szavanként való reprezentálásához képest.

# 3. Az értekezés felépítése

Az értekezés összesen öt fejezetből áll. Az 1. fejezet egy általános bevezetés, melyben tisztázom az értekezés motivációit és főbb célkitűzéseit, illetve előre vetítem a használt módszereket. A dolgozat 2. fejezetében szakirodalmi áttekintést adok, és bemutatom a tőzsdei hírbányászat kialakulását és főbb képviselőit, a 2.1 alfejezetben kiemelve a téma alapjait megteremtő Wüthrich, illetve Lavrenko munkáit, illetve Schumakert és Groth-t, akikre a saját modellem és hipotéziseim kialakításakor jelentős mértékben támaszkodtam. További jelentős kutatók munkáit tartalmazza a 2.2 alfejezet.

Az irodalomban használt módszereket és fogalmakat a 3. fejezetben rendszerezem. E módszertani fejezet felépítése a CRISP-DM sztenderdet követi, amely egy konkrét adatbányászati folyamat lépéseinek leírására szolgál. A 3. fejezet elején bemutatásra kerül a CRISP-DM folyamat, majd minden lépése egy-egy alfejezetnek felel meg, amelyek a tőzsdei hírbányászati folyamat adott lépéshez tartozó megoldásait, módszereit tekinti át. A saját modellhez használt módszerek is itt találhatók az egyes alfejezetek végén, vagy terjedelemtől függően külön alfejezetként. A modell felépítését már (Kovács 2014) cikkemben publikáltam, viszont ott nem tértem ki részletesen az egyes részek megvalósítására.

A kapott eredmények bemutatása és a hipotézisek tesztelése a 4. fejezetben olvasható. Ez a fejezet hat részfejezetre tagolódik, melyek mindegyike egy-egy hipotézissel foglalkozik. Az egyes vizsgálatok alapjául kétféle kísérleti összeállítás szolgált, melyeket a 4. fejezet elején mutatok be, majd az alfejezetekben a kísérleti eredmények eltérő szempontú elemzése révén következtetek hipotéziseimre. Az egyik kísérleti összeállítás során fix, 20 perces időablakot alkalmaztam, és a nyelv, a szövegreprezentáció, illetve az osztályozó módszer paraméterei megválasztásának hatását vizsgáltam. A másik összeállításban magyar nyelvet, és szózsákreprezentációt alkalmaztam, és az időablak hosszának, illetve az osztályozó paramétereinek megválasztásának hatását vizsgáltam. Az 5. fejezetben összegzek.

# 4. A kutatás módszertana

A hipotézisek teszteléséhez a tőzsdei hírbányászati modell hozamosztályozó változatát használtam, melynek bemeneteit a BÉT prémium kategóriás részvényeihez kapcsolódó, 2014.07.01 és 2015.06.31 közötti sajtóközlemények szövegei képezik, outputját pedig a közlemény publikálásának ideje és a hozzá képest $-120 \leq t \leq 120$ perccel eltolt időpont közötti hozam nagysága alapján képzett hozamkategória – negatív, semleges, pozitív. Hírforrásként a BÉT weboldalának azt a részét választottam, amelyen a részvénykibocsátók sajtóközleményeit teszik közzé, amely egy publikus hozzáférésű newswire típusú hírforrás. A kiindulási korpuszom 965 angol és ugyanennyi magyar nyelvű hírből állt. Az árfolyamok forrásaként a korlátozott hozzáférésű Thomson Reuters Eikon platformot használtam, amelyből a Prémium kategóriás részvényeinek egyperces részletességű OHLC adatait töltöttem le, közel hatvanezret.

A szövegjellemzők kiválasztása kapcsán fontos szempont volt, hogy mind angol, mind magyar nyelvre alkalmazható legyen a módszer, ezért választottam a tokeneket és a korpusz szóhasználata alapján heurisztikusan kinyert kifejezéseket. Szótövezést végeztem a Snowball-szótövező angol, illetve magyar változatával. Kizártam a három karakternél rövidebb tokeneket, továbbá azokat, amelyek 10-nél kevesebbszer fordultak elő a korpuszban. A TDM súlyozására a tf-idf mértéket alkalmaztam, amely az egyik leggyakoribb az irodalomban. A kifejezések kinyerésére szolgáló algoritmust Mostafa (2007) alapján készítettem el, mely nem csupán szakkifejezések kinyerésére alkalmas, hanem az egy kaptafára íródott, sablonos szövegrészek azonosítására is. Ez felhasználható arra, hogy egyfajta kiterjesztett stopszavazás révén a kevésbé jelentős szövegrészeket eltávolítsuk a dokumentumokból. A sablonszövegek előrejelezhetőségre gyakorolt hatásának vizsgálata miatt a kísérleteim kétféle változatban is elvégeztem. Az egyik változatban a sablonszövegeket benne hagytam a dokumentumokban, a másik változatban kitöröltem őket. Szintén kétféle kísérletet végeztem el, hogy megvizsgáljam, hogy a gyakori szókapcsolatok bevonásával javíthatók-e az eredmények. Az egyik változatban csak szavak, a másikban szavak és gyakori szókapcsolatok is szerepelnek. A kifejezések száma mindkét nyelven százas nagyságrendet vett fel, 200–300 körüli volt, ugyanakkor az önálló szavak száma angol nyelven 4000 körüli, magyar nyelven 8900 körüli volt.

Az árfolyam-idősorokat loghozamokkal reprezentáltam, melyek diszkretizálása során három kategóriát alakítottam ki: negatív, pozitív, illetve semleges. A diszkretizálás során a lehető legegyenletesebb megoszlás kialakítására törekedtem, összhangban a Wüthrich-, illetve Kop-

pel-modellel. A hozamkategóriák definiálásához használt alsó és felső korlátok az időablak hosszával változnak. E korlátok megállapításához a minta hozameloszlását használtam fel, és a loghozamok abszolút értékének első tercilisével és ellentettjével tettem egyenlővé őket. Ilyen módon a megfigyeléseknek legalább egyharmada a semleges kategóriába tartozik.

A közel ezer elemű kiindulási korpuszban a vizsgált időszakba[1] 330 hír tartozott. A kereskedési időn belüli hiányzó értékeket a legutóbbi érvényes árfolyammal interpoláltam. Amennyiben a hírhez tartozó időablak túllógott a tőzsde nyitvatartási idején, akkor azt a hírt kizártam az adott mintából. Ennek eredményeként a minta mérete 92–158 közöttivé redukálódott.

A hírek szövegének numerikus reprezentációja alapján egy nemlineáris SVM-osztályozót[2] tanítottam a különböző méretű tanítómintákon. A modellem kiértékeléséhez a pontosságmutatót használtam, ennek oka, hogy ez a legelterjedtebb mutató, mindig kiszámolható, tömegesen tesztelhető t-próbával, és jól vizualizálható. Az átlagos pontosságot és a szórásukat tízszeres keresztvalidáció során állapítom meg. Az SVM gamma és C paramétereinek vizsgálata során a gamma esetén 0,001-től 5-ig, 20 darab logaritmikus lépésközt vettem fel, a C esetén 0,1-től 5000-ig ugyancsak 20-at. Az eredményeket hőtérképen vizualizáltam.

A H1, H4, H5 és H6-os hipotézisnél rögzített $t=20$ perces időeltolást alkalmaztam, a H2-es és H3-as hipotézisnél a $t$ rögzítését feloldottam, hiszen éppen arra vonatkozott a vizsgálat, viszont a reprezentációt és a nyelvet rögzítettem. Ez utóbbi kísérleti összeállításban az egyes időablakokat két mutató értékével jellemeztem, melyek alapja az ahhoz az időablakhoz tartozó modellek pontossága volt. A q-val jelölt minőségmutató kifejezi, hogy a modellek mennyivel pontosabbak az a priori találati valószínűséghez – azaz a default pontossághoz – képest. A d-vel jelölt mutató az előrejelzés nehézségét számszerűsíti, és a defaultnál pontosabb modellek arányán alapul. A d-mutatóval jellemezhető, hogy milyen gyakoriak az alkalmazott SVM algoritmus két paraméterének kedvezőtlen kombinációi. Mindkét mutatóval parciálisan és egyszerre is jellemzem az időablakokat, utóbbi esetben Pareto-hatékonyság szerint.

A hírbányászati modell elkészítését teljes egészében, az elemzéseket és a vizualizációk elkészítésének nagy részét a RapidMiner 5.3.015-ös verziójával végeztem, melyhez a Text Mining Extension, Web Mining Extension és Series Extension bővítményeket telepítettem. A statisztikai teszteket[3] a RapidMiner által szolgáltatott output alapján Microsoft Excel 2010-zel készítettem. Az ábrák egy része szintén az Excel, illetve a LibreOffice Calc program 5-ös verziójának segítségével készült. A sajtóközleményeknek a BÉT honlapján elérhető változatait a HTTrack nevű webcrawler 3.48-1 verziójával gyűjtöttem le. Az egyperces felbontású részvényárfolyamokat a Thomson Reuters Eikon platform segítségével szereztem be.

# 5. A kutatás eredményei

Kutatásom egyik eredményeként tartom számon, hogy rendszereztem a szakirodalomban alkalmazott megoldásokat, és azt egységes keretbe foglaltam a CRISP-DM sztenderd alapján. Ez alapján a következő megállapításokra jutottam. A tőzsdei hírbányászati rendszerek célja általában az árfolyam rövid távú előre jelzése, melyhez hírkorpuszra és nagyfrekvenciás árfolyamokra van szükség. A hírszolgáltató szoftverén keresztül vagy webes begyűjtéssel összeállított gyűjtemény szövegét numerikusan reprezentálni kell. Ez általában a szózsákmodellel történik, de a szakértői kifejezéslisták, vagy a nyelvtani jellemzők is gyakran alkalmazott módszerek. A szózsákmodellben minden dokumentumot egy vektorral helyettesítünk, amelynek elemeit többféle súlyozási sémával meg lehet határozni, a bináristól a szógyakoriságon át a tf-idf sémáig. A vektor hosszával nő a számolásigény, így a jellemzők számának csökkentése érdekében nyelvfüggetlen és nyelvfüggő módszereket is alkalmaznak, amilyenek például a szótövezés, stopszavazás, khi-négyzet próba, szógyakorisági korlát, jellemzőrangsorolás stb. Az árfolyamokat ugyancsak reprezentálni kell, általában az árfolyamváltozás előjele alapján alakítanak ki kategóriákat, vagy egyéb jellemzőket képeznek belőle, mint például trendek, csúcsok, kilengések stb. Ha diszkrét változóval reprezentáljuk az árfolyamot, akkor osztályozás, ha folytonossal, akkor regressziós adatbányászati feladatot kell megoldani. A leggyakrabban használt algoritmus az SVM, illetve regressziós változata, az SVR, valamint a naiv Bayes-osztályozó, a szabályalapú rendszerek, a szomszédságalapú osztályozók és a neurális hálózatok. Az osztályozás jóságát általában a pontossággal mérik, de ezt a mutatót csak a default értékhez viszonyítva szabad értelmezni. Ezen kívül a precizitás, a felidézés és a belőlük számított F1 mutató is kedvelt mérőszám, de többosztályos problémákra többféle számítási módjuk létezik. A modell üzleti kiértékeléséhez kereskedési szimulációt alkalmaznak, az így kapott profitot össze lehet hasonlítani egy benchmarkportfólió vagy -stratégia hozamával. Ezen kívül újramintavételezéssel tesztelhető, hogy mekkora szignifikancia szinten tekinthető véletlenszerűnek a szimulált hozam elérése. Az utolsó fázis az üzleti implementáció, az áttekintett irodalomban nem volt jellemző.

Hipotéziseimhez kapcsolódó eredményeim az alábbiakban olvashatók.

**H1: A sajtóközlemények befolyásolják a részvényárfolyamot és a szövegük felhasználásával az a priori valószínűségnél – default modellnél – nagyobb pontosságú előrejelzés készíthető a BÉT prémium kategóriás részvényeire.**

A sajtóközlemények hatását eseményvizsgálat módszerrel is kimutatta már Bedő és Rappai (2004; 2006) , ehhez azonban az szükséges, hogy a sajtóközleményben lévő információt manuálisan előkészített változók formájában reprezentáljuk. Először tehát azt kell belátnunk, hogy a közlemények szövegéből kinyert jellemzőkkel is kimutatható a hatás, s az időtáv, a hatékonyság és a robusztusság kérdésével majd ez után lehet foglalkozni. A H1-es hipotézisem vizsgálatakor azt teszteltem, hogy a szöveges előrejelzés pontosabb-e a defaultnál, azaz a legnagyobb gyakoriságú kategóriára[4] való fogadás nyerési esélyénél. Egymintás t-próbákkal ellenőriztem, hogy a különböző paraméterek 3528 kombinációja mellett kapott átlagos pontosság szignifikánsan különbözik-e a default modell pontosságától – amely 38,06% a mintában. A hipotézist elfogadtam, mivel az eredményeim szerint az összes paraméterkombináció 94,64%-a szignifikáns volt 1%-on. A szövegbányászattal kinyert jellemzők tehát hordozzák azt az információt, amellyel az árfolyamváltozás iránya részben magyarázható a hír utáni 20 percben, és akár majdnem 30%-kal pontosabban meghatározható. Az eredmények alapján igazoltam, hogy a magyar sajtóközlemények befolyásolják a részvényárfolyamot és a szövegük felhasználásával a default modellnél nagyobb pontosságú előrejelzés készíthető a BÉT prémium kategóriás részvényeire.

**H2: A magyar tőzsdei sajtóközlemények haladéktalanul közzétett, új információt hordoznak. Nem lehet jó minőségű, illetve robusztus hírbányászati modellt készíteni olyan időablakra, amely korábbra nyúlik vissza, mint az az időtartam, amit a hír a KIBINFO közzétételi folyamatban eltölt.**

A tőzsdeszabályzat közzétételre vonatkozó előírásai szerint az információ rögzítése után legfeljebb egy óráig tart a közzétételi folyamat. Ebben az időszakban tehát már az információ létezik, és mielőtt máshol is megjelenne, bekerül a KIBINFO rendszerbe. A modell segítségével tesztelhető, hogy az információ hatása valóban nem jelentkezik ennél korábban. A H2 hipotézis elfogadásának feltétele az volt, hogy a $-60$ -as időablakot megelőző időtávra nehéz előrejelezni, illetve a modellek minősége is gyenge. Előbbit a d-mutató, utóbbit a q-mutató alapján számszerűsítettem. Eredményeim szerint a H2 hipotézist elfogadhatjuk, hiszen a

−60 -as időeltoláson túl a d-mutató jóval 20% fölötti és fokozatosan romlik, továbbá a minőség mediánja is tovább romló tendenciát mutat.

**H3: A modell robusztussága és minősége az időablakkal változik, és ennek optimuma meghatározható.**

Tételezzük fel, hogy ez az információt írásban rögzítő közlemény publikálási idejéhez viszonyítva korábbra és későbbre is tehető – az információt ugyanis kb. egy órával a publikálás előtt már írásba foglalják. Feltételezhetjük továbbá, hogy ha egy időablakot egy perccel kibővítünk vagy csökkentünk, a pontosság várhatóan csak kissé változik, viszont ha túl sok olyan percet veszünk az időablakhoz, amelyben már az információ hatása nem érvényesül, vagy túl sokat olyat hagyunk el, amelyben még érvényesülne, akkor a pontosság romlani fog. Tehát a modell akkor adja a legjobb eredményt, ha a vizsgált időablak lefedi az információ napon belüli beárazódásának időszakát. A H3-as hipotézis során a legrobusztusabb és legjobb minőségű modellekre vezető időablak meghatározását többcélú optimalizálási feladatként oldottam meg. A hipotézist az eredmények alapján el kell fogadni, az előrejelzés nehézsége a publikálás előtti 27 percben viszonylag alacsony, miközben a minőség is itt az egyik legmagasabb, a publikálás utáni 19–22 perces tartományban is könnyű az előrejelzés, és a magas minőségű modellek közül jelentős számú, is ezeknél az időablakoknál található. Ez alapján tehát az információ a publikálás előtti kb. fél órában kezd beépülni a vizsgált részvények árfolyamába, majd ez a publikálást követő kb. 20 percig tart.

**H4: Az azonos sajtóközlemények angol és magyar nyelven közzétett változatai egyformán alkalmasak a hírek árfolyamra gyakorolt hatásának vizsgálatára. Az információ nyelvi kódolásának nincs különösebb jelentősége.**

A H4-es hipotézisnél kétmintás t-próbákat alkalmaztam a minden paraméter tekintetében azonos, de eltérő nyelvű input dokumentumokat kezelő modellek pontosságának összevetésére 1764 modellpár esetén. A hipotézist elfogadtam, ha egyik nyelv sem jött ki győztesen az összehasonlításból. Az összes szignifikancia szinten – 1%, 5% és 10% – kevesebb volt a különbséget mutató összehasonlítások aránya, mint 1%, így e nem szigorú értelemben azt mondhatjuk, hogy nincs jelentős különbség a nyelvek tekintetében. Ha szigorú követelményként a fenti állítás minden esetben történő teljesülését írnánk elő, abban az esetben a H4 hipotézist el

kell vetnünk 1%-os és 5%-os szignifikancia szintek mellett, mert léteznek bizonyos C-gamma kombinációjú modellek, amelyek a magyar nyelvű korpusz segítségével jobb eredményt értek el. A magyar nyelvű közlemények alapján tehát pontosabb becsléshez juthatunk, de erre az esély elég csekély, kevesebb mint 1%. Különben egyik nyelv sem preferált a másikhoz képest.

**H5: Az eredmények robusztusak az alkalmazott SVM osztályozási módszer beállításaira nézve. Egy – a legpontosabb modellhez viszonyítva – szignifikánsan jó modell paramétereinek kis megváltozásával másik szignifikánsan jó modellhez lehet jutni.**

Az SVM-modell eredményére jelentős hatással van a paraméterek megválasztása, ha a döntési határfelület nemlineáris, ezért több paraméterkombinációra is meg kell vizsgálni az eredményeket. Ezután nem elégedhetünk meg a legjobb várható pontosságot adó modell vizsgálatával, hiszen ebben a mérésben még valamekkora bizonytalanság is van. Több olyan más paraméterkombináció lehet, amelynél nem mutatható kis szignifikánsan, hogy pontossága elmaradna a legjobb várható pontosságú mögött, ezek kvázioptimálisak. Ezek közül olyan modelleket célszerű választani, amelyek robusztusak is, tehát azok a paraméterkombinációk, amelyen bármilyen irányban egy kicsit változtatva csak nem optimális modellhez jutunk, nem preferáltak.

A hipotézis teszteléséhez először kétmintás t-próbával megállapítottam, hogy az azonos nyelvű, azonos inputváltozókkal tanított, de eltérő C-gamma paraméterkombinációjú SVM-osztályozók közül melyek kvázioptimálisak. A következő lépésben kizártam közülük a kvázioptimális paraméterszomszéddal nem rendelkezőket. A különböző szignifikancia szintekre így kapott robusztus modellek halmazának számosságát a kvázioptimális modellek számosságához viszonyítottam, és ez alapján a H5 hipotézist elfogadtam, ha az arány meghaladta a küszöbértékeket. Eredményeim szerint kevésbé szigorúan véve[5] a kvázioptimális modellek robusztusak minden vizsgált szignifikancia szinten, de szigorúan véve[6] el kell vetnünk a hipotézist, mert kb. 1%-nyi eséllyel visszaeshet a pontosság, ha egységnyit változtatunk a paramétereken. Ezért az üzleti alkalmazás fázisában a legjobb helyett több jó modell becslését érdemes összevetni, hogy a teljesítményt biztosítsuk.

**H6: Az eredmények robusztusak a szöveges inputok körének megválasztására. A szakterület-specifikus kifejezések azonosítása és a sablonszerű szövegrészek eltávolítása nem befolyásolják számottevően a pontosságot a szöveg szavanként való reprezentálásához képest.**

A kinyert szövegjellemzőkben az árfolyamváltozás magyarázása szempontjából irreleváns információk is megjelennek. Logikusnak tűnik, hogy némi manuális munkával, a szakterületre jellemző kifejezések kiemelésével, vagy a sajtóközleményekre jellemző sablonbetétek eltávolításával javítható a hasznos információk aránya. Ez segíthet a modellnek abban, hogy pontosabb legyen. A hipotézis teszteléséhez először kétmintás t-próbával megállapítottam, hogy az azonos nyelvű, azonos C-gamma paraméterkombinációjú, de más szövegreprezentációt alkalmazó modellpárok közül melyek várható pontosságai tekinthetők azonosnak. A következő lépésben minden reprezentáció-párosítás esetén az ilyen modellek számát elosztottam a hozzá tartozó összes modell számával. A hipotézist elfogadtam, ha egyik reprezentáció sem jött ki győztesen az összehasonlításból. Az összes szignifikancia szinten – 1, 5 és 10% – az tapasztalható, hogy az összehasonlításoknak kevesebb mint 3%-ánál van bármiféle eltérés. Ahol eltérés tapasztalható, ott az egyszerűbb reprezentációt alkalmazó modellek bizonyulnak pontosabbnak, tehát azok, amelyekben kevesebb a manuális munka – nincsenek kiemelve a kifejezések, vagy kiszedve a sablonszövegek. Összefoglalva tehát kevésbé szigorú értelemben nincs lényeges különbség az egyes szövegreprezentációk tekintetében, szigorú értelemben a nyers szózsák-szövegreprezentációval legalább olyan pontos eredményt lehet elérni, mint a többivel.

# 6. Jövőbeli kutatási irányok

A modell továbbfejlesztési lehetőségei között az üzleti kiértékelést segítő szimulációt a jelenlegi összeállításban nem érdemes megvalósítani, ugyanis a jutalékok meghaladják az elérhető hozamokat. Egy üzleti haszonnal járó lehetőség volna a Groth et al. (2014) által a likviditáshoz kapcsolódó tranzakciós költségek csökkentésére kifejlesztett modell megvalósítása, amelyhez ajánlati könyv szintű adatokra van szükség. Ez a modell ugyanis egyébként is végrehajtandó tranzakciók megfelelő időzítésére törekszik, és nem cél a jutalék teljes eliminálása.

Az alkalmazott szövegreprezentációk nem bizonyultak jobbnak a normál szózsákmodellhez képest, de ez nem jelenti azt, hogy nem létezik olyan reprezentáció, amely jobb volna. Érdemes lehet a szófaj, entitások és egyéb jellemzők bevonása is, ami a jelenlegi szoftverben nem elérhető beépített formában, de R vagy Python NLTK-csomagok megoldást nyújthatnak a problémára.

Szemantikai elemek is beépíthetők a modellbe, ha külső eszközöket kapcsolunk hozzá, mint például a WordNet, vagy a Wikipédia stb. Ezekkel a szövegben szó szerint nem jelenlévő fogalmak is reprezentálhatóvá válhatnak.

A vizsgálatok nagyobb magyar mintán való elvégzésére akár a jelenlegi formában is lehetőség volna, például a Standard és T kategóriás részvények bevonásával. Ennek nyilván hatása lesz majd a modell teljesítményére is, és lehetséges hogy a részvényeket valamiféleképpen csoportosítani kell majd, hogy ne okozzanak zajt az időablak meghatározása során. Ráadásul számukra nem kötelező az angol közzététel, tehát lehetséges, hogy a kétnyelvű vizsgálatról is le kell mondani ennek érdekében.

Lehetőség van az SVM-paraméterek hatásának nagyon nagy felbontás melletti vizsgálatára is. A jelenleg alkalmazott 20×20-as felbontás növelésének csak a hardverkapacitások szabnak határt elviekben. Ez azzal az előnnyel járna, hogy a robusztusság pontosabban mérhető lenne, mert a paraméterek és a pontosság „kis megváltozása” finomabban értelmezhető volna.

---

1 2015.04.27–2015.07.24.
2 RBF-kernellel
3 t-próba, khi-négyzet próba
4 Növekvő, semleges, illetve csökkenő árfolyamváltozás.
5 90%-os és 95%-os küszöbértéknél.
6 99% és 100%-os küszöbnél.

# 7. A tézisfüzetben felhasznált irodalom


Bedő, Zsolt & Rappai, Gábor, (2004), Eseménytanulmány-elemzés magyar részvényárfolyamokra - van-e értéke az árfolyamokat befolyásoló híreknek? *Szigma*, XXXV(3–4), o.107–121. Elérhető: http://www.szigma.ktk.pte.hu/index.php/letoltesek/2004-xxxv-evfolyam-3-4-szam/bedo-zsolt-rappai-gabor-esemenytanulmany-elemzes-magyar-reszveny-arfolyamokra-van-e-erteke-az-a/r%25C3%25A9szletek.

Bedő, Zsolt & Rappai, Gábor, (2006), Is there casual relationship between the value of the news and stock returns? *Hungarian Statistical Review*, (Special Number 10), o.81–99. Elérhető: http://www.ksh.hu/statszemle_archive/2006/2006_K10/2006_K10_081.pdf.

Gidófalvi, Győző, (2001), *Using News Articles to Predict Stock Price Movements*, San Diego, California: University of California. Elérhető: http://people.kth.se/~gyozo/docs/financial-prediction.pdf.

Groth, Sven S. & Muntermann, Jan, (2008), A text mining approach to support intraday financial decision-making. In *Fourteenth Americas Conference on Information Systems*. Toronto.

Groth, Sven S., Siering, Michael & Gomber, Peter, (2014), How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data. *Decision Support Systems*, 62, o.32–42.

Koppel, Moshe & Shtrimberg, Itai, (2004), Good News or Bad News? Let the Market Decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*. Palo Alto, o. 86–88. Elérhető: http://www.aaai.org/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-016.pdf.

Kovács, Balázs, (2014), Egyedi események árfolyamhatásának becslése hírszövegek elemzése alapján. *GIKOF Journal*, 10(1), o.38. Elérhető: http://gikof.njszt.hu/gikof/GIKOF_JOURNAL_2014-1.pdf.

Lavrenko, Victor, Schmill, Matt, Lawrie, Dawn, Ogilvie, Paul, Jensen, David & Allan, James, (2000), Mining of concurrent text and time series. In *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Text Mining*. Boston, MA, o. 37–44. Elérhető: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.135.5688.

Mittermayer, Marc-André, (2004), Forecasting Intraday Stock Price Trends with Text Mining Techniques. In *37th Annual Hawaii Int. Conference on System Sciences (HICSS)*. Hawaii: IEEE. Elérhető: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.6966.

Mostafa, Hatem, (2007), N-gram and Fast Pattern Extraction Algorithm. Elérhető: http://www.codeproject.com/Articles/20423/N-gram-and-Fast-Pattern-Extraction-Algorithm [Elérés 2015.07.25.].

Schumaker, Robert P., (2010a), An Analysis of Verbs in Financial News Articles and their Impact on Stock Price. In *NAACL Workshop on Social Media and Computational Linguistics*. Los Angeles.

Schumaker, Robert P., (2010b), Analyzing Parts of Speech and their Impact on Stock Price. *Communications of the International Information Management Association*, 10(3).

Schumaker, Robert P., (2009), Analyzing Representational Schemes of Financial News Articles. In *The Third China Summer Workshop on Information Systems*. Guangzhou, China.

Schumaker, Robert P. & Chen, Hsinchun, (2006), Textual Analysis of Stock Market Prediction Using Financial News Articles. In *12th Americas Conference on Information Systems*. Acapulco, Mexico.

Thomas, James & Sycara, Katia, (2000), Integrating Genetic Algorithms and Text Learning for Financial Prediction. In *Proceedings of GECCO-2000 Workshop on Data Mining with Evolutionary Algorithms*. Las Vegas. Elérhető: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.8655.

Wüthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J. & Lam, W., (1998), Daily Stock Market Forecast from Textual Web Data. In *SMC '98 Conference Proceedings*. San Diego, California: IEEE. Elérhető: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=725373&isnumber=15656.

Zhang, Debbie, Simoff, Simeon J. & Debenham, John, (2005), Exchange Rate Modelling Using News Articles and Economic Data. In S. Zhang & R. Jarvis (szerk.) *AI 2005: Advances in Artificial Intelligence*. Lecture Notes in Computer Science. Sydney, Austraila: Springer Berlin Heidelberg, o. 467–476. Elérhető: http://dx.doi.org/10.1007/11589990_49.

# 8. A szerző publikációi az értekezés témakörében

## *8.1. Megjelent publikációk*


Kruzslicz, Ferenc, Kovács, Balázs, Hornyák, Miklós, (2016), Összehasonlító klaszterjellemzés külső, szöveges források bevonásával. Statisztikai Szemle 94(11-12) o.1123-1148. Elérhető: http://dx.doi.org/10.20311/stat2016.11-12.hu1123

Kovács, Balázs, (2015), A Critique of the Assumptions Regarding Investor Confusion of Similarly Identified Stocks. The Empirical Economics Letters, 14(10), o.1027–1033. Elérhető: http://eel.my100megs.com/volume-14-number-10.htm.

Kovács, Balázs, (2014), Egyedi események árfolyamhatásának becslése hírszövegek elemzése alapján. GIKOF Journal, 10(1), o.38. Elérhető: http://gikof.njszt.hu/gikof/GIKOF_JOURNAL_2014-1.pdf.

Kovács, Balázs, (2014), Részvények a piaci hatékonyság határán: az elmúlt húsz év legnagyobb melléfogásai. In G. Rappai & Z. Schepp (szerk.) Válságtól a jóllétig: A múlt tanulságai, a jelen kihívásai és a jövő útjai. Pécs: Pécsi Tudományegyetem Közgazdaságtudományi Kar, o. 23–50.

Kovács, Balázs, Kruzslicz, Ferenc & Torjai, László, (2013), Internetes termékkritikák hasznosságának megállapítása felügyelt gépi tanulással. Szigma, 44(1–2), o.35–63. Elérhető: http://www.szigma.ktk.pte.hu/index.php/letoltesek/2013-xliv-evfolyam-1-2-szam/kovacs-balazs-kruzslicz-ferenc-torjai-laszlo-internetes-termekkritikak-hasznossaganak-megallapitasa-felugyelt-gepi-tanulassal/r%25C3%25A9szletek.

Pauler, Gábor & Kovács, Balázs, (2013), Mesterséges idegsejt hálózat alapú döntési rendszerek a devizakereskedésben, Pécs: Pauler, Gábor. Elérhető: https://www.researchgate.net/publication/259839135_Mestersges_idegsejt_hlzat_alap_dntsi_rendszerek_a_devizakereskedsben.

Süle, Attila & Kovács, Balázs, (2013), Forex devizaárfolyam előrejelző robot építése neurális hálóval. In L. Varga (szerk.) „Pollackos” TDK Füzetek. Pécs: Pécsi Tudományegyetem Pollack Mihály Műszaki és Informatikai Kar, o. 93–115.

Kovács, Balázs & Kruzslicz, Ferenc, (2011), Tájolás és hasznosság mérése rövid szöveges üzenetekben. ACTA AGRARIA KAPOSVÁRIENSIS, 15(3), o.103–113.

## 8.2. A disszertáció témaköréből tartott konferencia előadások

Kovács, Balázs, (2016), Érzékenységvizsgálattal egybekötött tőzsdei hírbányászat, 13. Országos Gazdaságinformatikai Konferencia, Dunaújváros, 2016. november 11–12.

Kovács, Balázs, (2015), Kereskedési szabályok generálása adatbányászati eszközökkel, 12. Országos Gazdaságinformatikai Konferencia, Veszprém, 2015. november 6–7.

Kovács, Balázs, (2014), Részvények a piaci hatékonyság határán - Az elmúlt húsz év legnagyobb melléfogásai, Well-being in Information Society Conference, Pécs, 2014. november 13–14.

Kovács, Balázs, (2013), Egyedi események árfolyam-hatásának becslése hírszövegek elemzése alapján, 10. Országos Gazdaságinformatikai Konferencia, Győr, 2013. november 8-9.

Pauler, Gábor, Kovács, Balázs, (2012), Transforming neuron network into money?, János Szentágothai Memorial Conference and Student Competition, Pécs, 2012. október 29–30.