DOKTORI ÉRTEKEZÉS

Kovács Balázs

Pécs, 2017

PÉCSI TUDOMÁNYEGYETEM
KÖZGAZDASÁGTUDOMÁNYI KAR

GAZDÁLKODÁSTANI DOKTORI ISKOLA

Kovács Balázs

Tőzsdei hírbányászat a magyar részvénypiacon

DOKTORI ÉRTEKEZÉS

Témavezető: dr. Kruzslicz Ferenc

Pécs, 2017

Tartalomjegyzék

Táblázatjegyzék

Ábrajegyzék

Rövidítések jegyzéke

ANN..... mesterséges neurális hálózat..... .15
ANOVA.... varianciaanalízis.....…9
AR..... abnormális hozam.....…8
AUC..... görbe alatti terület..... .90
CAPM..... tőkepiaci értékelés modellje.....…8
CAR..... kumulált abnormális hozam.....…8
CDF..... kategória-megkülönböztetési tényező..... .12
cf..... gyűjteménytámogatottság..... .64
CRISP-DM..... Cross Industry Standard Process for Data Mining..... .42
CRT..... retúrköltség..... .30
DDF..... dokumentum-megkülönböztetési tényező..... .12
DET..... felismerési hibák átváltási görbéje..... .18
df..... dokumentumgyakoriság..... .64
dt..... sablonszövegek nélkül..... 106
DGAP..... Deutsche Gesellschaft für Ad-hoc-Publizität..... .27
DJIA..... Dow Jones Industrial Average..... .14
EMH..... hatékony piacok elmélete.....…9
EN..... angol..... 125
FPR..... helytelenpozitív-arány..... .91
FTSE..... Financial Times 100 Index..... .14
FWB..... Frankfurt Stock Exchange..... .45
HP..... kifejezések tartalmazva..... 106
HSI..... Hang Seng Index..... .11
HU..... magyar..... 125
IDF..... inverz dokumentumgyakoriság..... .12
k-NN..... k legközelebbi szomszéd..... .15
KIBINFO. Kibocsátói Információs Rendszer..... .49
KKT..... Karush-Kuhn-Tucker..... .87
MSE..... átlagos négyzetes hiba..... .23
nB..... naive-Bayes osztályozó..... .77
NKY..... Nikkei225..... .14
NP..... kifejezések nélkül..... 106
nt..... sablonszövegek tartalmazva..... 106
OHLC..... nyitó-, magas-, mély- és záró-árfolyam..... .48
ROC..... vevő működési jelleggörbe..... .20
STI..... Singapore Straits Index..... .14
SVM..... támasztóvektor-gép..... .20
SVR..... támasztóvektor-regresszió..... .23
TDM..... szó-dokumentum mátrix..... .38
tf..... szógyakoriság..... .63
TPR..... helyespozitív-arány..... .91
WBAG..... Wiener Börse AG..... .54
wdf..... dokumentumon belüli gyakoriság..... .63
wdf-idf..... dokumentumon belüli gyakoriság - inverz dokumentumgyakoriság..... .39
WSJ..... Wall Street Journal..... .11

1. Bevezetés

A gazdasági modelleknek valószínűleg a legfontosabb eleme az információ, amellyel a gazdasági szereplők rendelkeznek. A közgazdaságtan embermodellje, a homo oeconomicus például minden információval rendelkezik ahhoz, hogy önérdekét legjobban szolgáló, racionális gazdasági döntését meghozza. Ezt az ideális gazdálkodót aztán olyan nagy gondolkodók kritizálták, mint Veblen, Keynes, Simon vagy Tversky, és az újabb modellek emberibb tulajdonságokat, képességeket kaptak. Az új embermodell nem feltétlenül rendelkezik minden információval, nem mindig az önérdek vezérli, esetleg nincs tisztában azzal, mi mennyire szolgálja azt, viselkedése nem racionális – jobb esetben korlátozottan racionális – és az egyes szereplők döntései eltérhetnek. Az ilyen szereplők nem feltétlenül elszigetelten cselekednek, ki vannak téve a többiek befolyásoló hatásának, vagy egyenesen a szervezeti viselkedés keretei között zajlik a döntéshozási folyamat. A piac működése az ilyen szereplők döntéseinek eredőjeként modellezhető, amiben számottevő a bizonytalanság. Ez a piacon megfigyelt mennyiségek megmagyarázását is megnehezíti, és korlátokat szab azok előrejelzésének is. A kereslet, a kínálat és az ár közötti összefüggések átlátása különösen fontos a pénzügyi piacokon, ahol a reálgazdaság számára nélkülözhetetlen források allokációja történik. A saját és/vagy mások megtakarításait e piacokon befektető professzionális szereplők számára elengedhetetlen az aktuális információk ismerete, azok helyes és pontos értelmezése, és a szükséges pénzügyi műveletek gyors lebonyolításának képessége. Ezek elsajátítása végső soron pénzben kifejezhető a rászánt idő haszonáldozati költsége, az információ várható értéke, az informatikai rendszerbe történő befektetés értéke, a kommunikációs infrastruktúra használatának díja, illetve a tranzakciós költségek alapján. A felsorolást még tovább is lehetne folytatni, amiből látszik, hogy a befektetési döntések kérdésköre mennyire interdiszciplináris terület, amely a pénzügyön túlnyúlva a kommunikáció, az informatika, a matematika, a döntéstudomány és a pszichológia különböző ágaival érintkezik.

Éppen ez a sokféle megközelítési lehetőség volt az, amely mindig is vonzott a téma iránt. A pénzügyi szakirányon folytatott tanulmányaim alatt kezdtem tudományos pályafutásom az OTDK-dolgozatommal, amelyben két nagy érdeklődési körömet, a kereskedési stratégiákat és a mesterséges intelligencia módszereket egyesítettem, és a megalkotott automatikus kereskedési rendszer teljesítményét vizsgáltam (Süle & Kovács 2013).

A mesterséges intelligencia és az adatbányászat mellett a gazdasági módszertanban régebb óta jelen lévő statisztikai és optimalizálási eszközök éppúgy érdekeltek akkoriban, és szakdolgozatomban a kereskedési döntéstámogatással foglalkoztam a technikai elemzésből nyert információk alapján. Ezek a megközelítések a befektetéshez szükséges információknak csak egy korlátozott körét vették figyelembe, nevezetesen a számszerű, főleg idősoros jellegű adatokat, mint a múltbéli árfolyamok vagy más eszközök árfolyamai, osztalékok, kamatlábak, árrések, technikai indikátorok és forgalmi adatok. Pénzügyi tanulmányaimmal párhuzamosan végeztem a gazdasági szakfordító képzést, amelyre visszatekintve a gazdasági hírek fordítása mellett a szemantika és a fordítási hibák tanulmányozása máig érezhető nyomott hagyott munkásságomban. Ekkor kezdett megfogalmazódni bennem a gondolat, hogy a gazdasági hírek tartalma bizonyos mértékben megérthető számítógép segítségével, és ezt integrálni lehet az árfolyamot magyarázó modellekbe is. A PhD-képzésre ezzel a témával jelentkeztem, és a doktori iskola megkezdésével párhuzamosan újabb módszertant sajátítottam el, a szövegbányászatot. Ahogy egyre jártasabb lettem a szövegbányászatban (Kovács & Kruzslicz 2011; Kovács 2012; Kovács et al. 2013), úgy vált egyre inkább világossá, hogyan lehet felhasználni ezt az ismeretet a kitűzött cél érdekében, és ez segített a téma szakirodalmának felderítésében is, amelyet e dolgozat *2. fejezetében* írok le részletesen, és itt csak egy gyors áttekintést adok a területről.

Az 1990-es évek végén Wüthrich alkalmazta először a hírbányászatot a tőzsdén, ő és szerzőtársai a tőzsde nyitásáig megjelent összes hír alapján jelezték előre a tőzsdeindexnek a tőzsde zárásakor várható értékét (Wüthrich, Cho, 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. 2000b). Thomas és Sycara (2000) hírek helyett tőzsdei fórumok hozzászólásaival kísérletezett, és genetikus algoritmussal behangolt szabályalapú rendszerével a következő napra készített előrejelzést. Gidófalvi (2001) az eseményvizsgálat módszertan elemeit[1] építette be a szövegbányászati modellbe, ugyanakkor ezt individuális hírek szintjén tette. Az előrejelzés időtávjának megválasztását is vizsgálta, eredményei alapján a ± 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

1 Az abnormális hozamok voltak a magyarázott változók.

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, konzisztens módon 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. A dolgozat a szöveges formában megjelent információknak a magyar tőzsdei részvényárfolyamokra gyakorolt hatását vizsgálja szövegbányászati módszertan segítségével. A disszertációt az empirikus művek közé sorolnám, melyek hipotézisei a hírekből kinyerhető információk és az árfolyamokban megnyilvánuló információk közötti kapcsolatra vonatkoznak. Így kerültek a dolgozatomba a kétnyelvű – angol–magyar – vizsgálatok Groth hatására, a saját eredmények robusztusságának vizsgálata a különböző paraméterek és szövegreprezentációk megválasztására – előbbihez hasonlóval nem találkoztam, utóbbi Schumaker hatására –, valamint az időbeliség vizsgálata – Lavrenko, Gidófalvi, részben Groth munkájához kapcsolódóan. Mielőtt a disszertációban megjelölt kérdésekkel kezdtem volna foglalkozni, több zsákutcába is belefutottam, melyekről pár sorban írnék itt a bevezetőben, hiszen ezek fontos szerepet játszottak a végső hipotéziseim és a kísérleti összeállítások kialakításában. Az egyik ilyen a részvények összetévesztésével kapcsolatos kutatásom volt.

Amikor a jelentősebb tőzsdei hírbányászattal foglalkozó cikkeket feldolgoztam, fogalmazódott meg bennem a kritika, 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. Az ilyen jellegű

kutatást az nehezíti meg általában, hogy nagyon zajos árfolyam-, volumen-, vagy árrésadatokkal kell dolgozni abban az értelemben, hogy a vizsgált időszakban a tranzakciók jelentős része a hírtől független motivációjú. A társadalomtudományos vizsgálatoknál a laboratóriumi körülmények megteremtése általában nem kivitelezhető, így az időtáv meghatározásának pontosítására úgy tűnt, az egyetlen mód, ha a mérést zavaró zajt minél jobban sikerül megmagyarázni. Egyfajta személyes paradigmaváltást köszönhetek akkori szakdolgozómnak, Szőts Leventének, akivel a Schumaker-modell devizapiaci alkalmazásán dolgoztunk (Szőts 2014), amikor érdekességként elmesélte az akkoriban nagy várakozásokkal övezett Oculus Rift nevű virtuális valóság technológiát fejlesztő cég felvásárlása utáni komikus eseményeket. 2014.03.25-én 14:30 PDT perckor publikálta a PR Newswire a Facebook és az Oculus VR között történt megállapodást az utóbbi cég 2 milliárd dollárért való felvásárlásáról. Az amerikai tőzsdék fő kereskedési időszaka utáni bejelentés hatására másnap reggel 152%-kal[2] emelkedett egy másik cég, az Oculus VisionTech Inc. részvényeinek árfolyama. A másik cégnek semmi köze nem volt az akvizícióhoz, azok partnereihez vagy az általuk képviselt technológiához, a befektetők egyszerűen összetévesztették a hasonló nevű, egyébként a tőzsdén nem jelen lévő Oculus VR-ral. Az esetben rejlő komikumot félretéve, megláthatjuk, hogy kitűnő lehetőség volt megfigyelni az információ árfolyamba való beépülésének idejét és fázisait – a korrekciót leszámítva persze –, ugyanis az árfolyamban és a volumenben lévő néhány cent mértékű zaj elenyésző volt az információ hatására jelentkező kilengésekhez képest. Feltételezve, hogy a tévedést elkövető szereplők reprezentatívak a cselekvési időt tekintve a piac összes szereplőjét illetően, hasonló esetek segítségével pontosabb mérések végezhetők el. A következő hónapokat újabb esetek keresésével töltöttem, majd nagyjából egy tucatnyi hasonló – ugyanakkor hitelesnek is tekinthető – esetet sikerült leírnom két tanulmányomban (Kovács 2014b; 2015). A statisztikai vizsgálatokhoz természetesen kevés megfigyeléssel rendelkeztem, ám az esetekből jól látszódott, hogy a hírek valóban percek, esetleg órák alatt fejtik ki látványosan hatásukat. Általában egy napon belül ki is árazódtak a kiugrások a részvények árából. Ez is hozzájárult ahhoz, hogy úgy döntöttem, hogy a korábbiaknál részletesebb lépésközöket alkalmazva meg fogom vizsgálni a hírek hatását különböző rövid időtávokon – H3-as hipotézisem.

A már említett szakdolgozat mellett egy másik is készült párhuzamosan, amely a magyar nyelvű hírek hatását vizsgálta a forint árfolyamára, ez Szabó Norbert munkája

2 Igaz, ez csak az előző napi 13,5 kanadai centes záróárról 34 centre való növekedést jelentett a délelőtt folyamán.

(Szabó 2014). A szakdolgozókkal folytatott munka során kiderült, hogy a devizaárfolyamokkal végzett kísérletek nagy hátránya, hogy nagyon sokféle hír befolyásolhatja azokat, és a híraggregáló szolgáltatások sem tudják feltétlenül az összeset összegyűjteni, illetve olyan híreket is találhatunk ezek között, amelyek árfolyamra gyakorolt hatása megkérdőjelezhető. 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. A hosszabb távú kereskedési döntéseket hozók számára ez persze nem releváns, de az adott devizapárokban napi szinten érdekelt professzionális szereplők valószínűleg ugyanazokból a sajtótájékoztatókból, közleményekből tájékozódnak inkább, mint a cikkeket készítő újságírók. Ez jelentősen megnehezíti, hogy a hírt melyik időponthoz kell rendelni. Ennek hatására fordultam az olyan hírforrások felé, amelyek deklarált módon olyan információt tartalmaznak, amely befolyásolja az árfolyamot, és korábban máshol nem volt elérhető. A Mittermayer (2004) által használt angol és a Groth és Muntermann (2008) által használt német sajtóközlemények szolgáltak mintául ehhez, így döntöttem a magyar tőzsde sajtóközleményei mellett. A BÉT tőzsdeszabályzata szerint ugyanis sajtóközlemény formájában azonnal közzé kell tenni a részvényárfolyamot befolyásoló vállalati információkat a tőzsde honlapján, amely ennél korábban máshol nem tehető közzé. E követelmények teljesülését a dolgozatomban is vizsgálom, azaz H1-es hipotézisemben kimutatom, hogy valóban befolyásolják a részvényárfolyamot, a H2-es hipotézisben pedig ellenőrzöm, hogy a publikálási folyamat megkezdése előtt nem mutatható-e ki valamilyen hatás, hiszen az azt jelentené, hogy máshol korábban megjelent az információ. Groth et al. (2014) német és angol nyelvű párhuzamos korpusszal folytatott kutatásai nyomán fogalmazódott meg bennem a H4-es hipotézisem is, amely ugyanannak a közleménynek a magyar és angol nyelvű megfelelőivel kapott eredmények összevetésére irányul.

A H5 és H6 hipotéziseim a modell robusztusságát vizsgálják a tőzsdei hírbányászati szakirodalomban tipikus kísérleteknél sokkal részletesebben. Az SVM 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 bár modellem a korábbiakhoz hasonló, azok teljes reprodukálhatóságának hiányában a modellem máshogy viselkedhet ugyanolyan paraméterek esetén, így a kísérletekhez kalibrálni kell azt.

Motivációim tisztázása után következzenek hipotéziseim még egyszer, rendszerezve:

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.

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 $-120 \le t \le 120$ perccel eltolt időpont közötti hozam nagysága alapján képzett hozamkategória – negatív, semleges, pozitív. A H1, H4, H5 és H6-os hipotézisnél rögzítetten $t = 20$ perces időeltolást alkal-

maztam, a H2-es és H3-as hipotézisnél a t rögzítését feloldottam, hiszen éppen arra vonatkozott a vizsgálat, viszont a reprezentációt és a nyelvet rögzítettem. A hírek szövegének numerikus reprezentációja alapján egy nemlineáris SVM-osztályozót[3] 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 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[4] 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.

A dolgozat *2. fejezetében* bemutatom a tőzsdei hírbányászat kialakulását és főbb képviselőit, 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*, de egy ennél is szélesebb irodalomáttekintést adok (Kovács 2014a) cikkemben, és (Pauler & Kovács 2013) monográfiánk 8.1-es alfejezetében. 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 saját modellhez használt módszerek is itt találhatók az egyes alfejezetek végén, vagy terjedelemtől függően külön alfejezetként. A modell felépítését már (Kovács 2014a) cikkemben publikáltam, viszont ott nem tértem ki részletesen az egyes részek megvalósítására. A kapott eredmények bemutatása és a hipotézisek tesztelése a 4. fejezetben olvasható, az 5. fejezetben összegzek.

Végül szeretnék köszönetet mondani mindenkinek, aki segített dolgozatom megírásában: konzulensemnek, Kruzslicz Ferencnek, kollégáimnak, kiemelten Pauler Gábornak és Hornyák Miklósnak, és végül családomnak és barátaimnak a támogatásért.

3 RBF-kernellel

4 t-próba, khi-négyzet próba

2. A szövegek alapján történő árfolyam-előrejelzés irodalma

A szöveges formában megjelent hírek árfolyamokra gyakorolt hatását vizsgáló korai tanulmányok az eseményvizsgálat – event study – módszert alkalmazták. Eseményvizsgálattal azonban nem csupán szöveges formában megjelent információk hatása vizsgálható, hanem bármilyené, amelyet eseménynek lehet tekinteni, például, ha egy részvény kereskedési volumene meghalad egy küszöbértéket, részvényfelosztás történik, vagy profit-warningot bocsát ki egy cég, stb. Ennek módszertanáról Bedő és Rappai (2004; 2006) adnak áttekintést, amely alapján röviden felvázolom a módszer lényegét. A szerzők Fama és szerzőtársaira (1969) vezetik vissza a módszertan kialakulását, akik a részvényfelosztáshoz kapcsolódó információk árfolyamba épülését vizsgálták hasonló módon. A módszer tökéletesítése a 1980–90-es évekre tehető, amely kapcsán Brown és Warner (1980, 1985), Dyckman, Philbrick, Stephans és Ricks (1984), Campbell és Wasley (1993), Barber és Lyon (1997) munkásságát emelték ki szerzők. Az eseményvizsgálat tulajdonképpen az esemény előtti és utáni helyzet statisztikai összehasonlítását jelenti. Az eseményelemzés első lépése az esemény fajtájának kiválasztása, a következő lépés az esemény időablakainak meghatározása. Az esemény előtti időablak a *normális* árfolyamot magyarázó modell paramétereinek becslésére használható – estimate window –, az eseményhez közeli időablak – event window – pedig az esemény hatásának kimutatására. Az eseményhez való időbeli viszony alapján beszélhetünk pre, illetve post event window-ról. A minta összeállítása az árfolyamokból és az eseményekből történik, de a mintából ki kell zárni azokat az eseményeket, amelyeknek nem képezhetők a fentiekben említett időablakai, például adathiány vagy nem megfelelő időbeli átlapolások miatt. A harmadik lépés annak meghatározása, hogy hogyan mérjük az árfolyamra gyakorolt hatást. Erre a célra jellemzően az abnormális hozam – AR, abnormal return – tesztet alkalmazzák, amely egy reziduális elemzés, tehát a várt és tényleges hozam közti különbségen keresztül ragadja meg a problémát. Ezt a különbséget adott időpontig kumulálva a kumulált abnormális hozamról – CAR, cumulative abnormal return – beszélünk. A várható hozam becslése általában a CAPM-modellben – capital asset pricing model, tőkepiaci értékelés modellje – (Sharpe 1964) történik, amelyben a kockázatmentes hozam szerepét a vizsgált piachoz legszorosabban kapcsolódó államkötvény, a piaci

hozamét pedig a tőzsdeindex, vagy a vizsgált részvények hozamainak átlaga játssza. A negyedik lépésben ki kell választani az alkalmazandó tesztmódszereket, amelyekkel az eseménnyel érintett és nem érintett időszakokat összehasonlíthatjuk. Itt szóba jöhetnek paraméteres statisztikai próbák, mint az egymintás t-próba, a független t-próba vagy az ANOVA – analysis of variance, varianciaanalízis –, a nem paraméteres próbák, akár a Wilcoxon előjeles rangpróba, illetve a Kruskal-Wallis teszt (Tumurkhuu & Wang 2010).

Az eseményvizsgálattal elemzett hírek általában egy binominális változóval[5] kerülnek reprezentálásra, jobb esetben ordinális skálán[6], de ritkábban előfordulhat, hogy folytonos skálán[7] jellemzik. Az esemény tálalásával, értékelésével kapcsolatos egyéb információk így elvesznek, ugyanis ezek számszerűsítése közvetlenül igencsak nehézkes. Az 1990-es évek végén jelentek meg az első olyan tanulmányok, amelyek a szöveges tartalom numerikus reprezentációját használták tőzsdei alkalmazásban (Cho et al. 1999; Cho & Wüthrich 1999; Wüthrich, Cho, et al. 1998; Wüthrich, Permunetilleke, et al. 1998). A tőzsdei hírbányászat egy folyamatosan fejlődő módszertan, amely a természetes nyelvi és adatbányászati kutatásokra épülve teszi lehetővé nagy mennyiségű szöveg feldolgozását és a belőlük származó információ hasznosítását. A hírek hatásának kutatását az árfolyam- és piacelméletek oldaláról érkező eredmények is befolyásolják, mint például a hatékony piacok elmélete (EMH, efficient-market hypothesis) (Fama 1970) és annak kritikái, illetve a magatartási pénzügy (Barberis & Thaler 2003). Ettől függően a hírek hatását vizsgálták a puszta hozamokra, volumenre, volatilitásra, árrésre és likviditásra, illetve ezek abnormális jelzőjű változataira[8], amelyek az eseményvizsgálat módszertanból kerültek a tőzsdei hírbányászatba. Például, ha egy olyan modellben vizsgáljuk a hírek hozamokra gyakorolt hatását, amely a hozamot nulla várható értékű valószínűségi változónak tekinti, akkor a nyers hozamadat megfelelő választás lehet eredményváltozónak, azonban ha az egyes eszközök hozamát külső tényezőkkel magyarázó modellbe, például a CAPM-be szeretnénk beépíteni a hírek hatását, akkor a modell alapján várható *normális* hozamhoz képest kell azt mérni, és az eredményváltozó az abnormális hozam lesz. Az árfolyamok tekintetében tehát lényeges a bennük rejlő információk kinyerésének, helyes megragadásának problémája.

5 esemény/nem esemény
6 pl. alacsony/indifferens/magas
7 pl. kvantitatív jellegű profit warning mértéke
8 pl. abnormális hozamok stb.

Ebben a fejezetben azokat a hírbányászati kutatásokat foglalom össze, amelyek mögött komolyabb, több éves munka húzódik meg, esetleg a szerző a témában írta disszertációját, illetve a későbbi tanulmányok rendszeresen hivatkoznak rájuk. Ezen kívül a szerző életművében egyszeri jellegű, vagy kevésbé jelentős munkák, vagy amelyekben a híreket osztályozó modell outputja nem közvetlenül az árfolyamra vagy a volumenre vonatkozik, megtalálhatók egy másik írásomban (Kovács 2014a). A modellek tárgyalásakor a rendezőelv a következő: előbb bemutatom a modellhez kapcsolódó publikációkat időrendben a szerzők közötti kapcsolatok kiemelésével, majd a modell különböző változatait írom le az adott változathoz kapcsolódó publikációk megjelölésével. A modellek egymással történő összevetése a 3. fejezetbe ágyazva található meg, amely ezen kívül a tőzsdei hírbányászat folyamatát és módszertanát is rendszerbe foglalja, és amelyben a saját modellemnél alkalmazott adatok és módszerek is szerepelnek.

2.1. Meghatározó modellek a tőzsdei hírbányászatban

2.1.1. A Wüthrich-féle modell

Az 1990-es évek végén jelentek meg az első tanulmányok, amelyek a tőzsdei előrejelzéshez közvetlenül az interneten keresztül publikált gazdasági hírek szövegét használták fel. A Hong Kong University of Science and Technology-n kibontakozó kutatás alapkoncepcióját 1997-ben alakította ki Beat Wüthrich két szakdolgozójával, Leung Kung Fannal és Peramunetilleke-vel (Leung Kung Fan 1997; Peramunetilleke 1997). A módszer kidolgozása során elsősorban Wüthrich tudásfeltárással és adatbányászattal kapcsolatos eredményeire támaszkodtak. 1998-ban Wüthrich és szerzőtársai – köztük az említett két szakdolgozó is – további vizsgálatokat végeztek, és eredményeiket a KDD-98[9] és az IEEE SMC[10] konferencián is bemutatták (Wüthrich, Permunetilleke, et al. 1998; Wüthrich, Cho, et al. 1998). 1999-ben a PhD értekezését író Vincent Cho és Wüthrich újabb eredményeket publikáltak a PAKDD-99[11] konferencián, valamint a JCIF[12]-ben (Cho et al. 1999; Cho & Wüthrich 1999). Az utóbbi két tanulmány jelentős részét képezi Cho doktori értekezésének is (Cho 1999). A kutatásokat összefoglaló tanulmányt Wüthrich 2002-ben jelentette meg egy könyvfejezet formájában (Wüthrich 2002). Ugyanebben az évben Peramunetilleke az ADC2002[13] konferencia tanulmánykö-

9 The 4th International Conference on Knowledge Discovery and Data Mining
10 IEEE International Conference on Systems, Man, and Cybernetics
11 Third Pacific-Asia Conference on Knowledge Discovery and Data Mining
12 Journal of Computational Intelligence in Finance
13 Thirteenth Australasian Database Conference

tetébe írt egy cikket (Peramunetilleke & Wong 2002), amely lényegében a szakdolgozatának eredményeit ismertette. A következőkben bemutatom a modellváltozatokat.

Leung Kung Fan (1997) a Hang Seng Index (HSI) záróárfolyamára következtetett a Wall Street Journal (WSJ) elektronikus változatában megjelent hírek szövege alapján. A WSJ honlapján a hírek tematikus oldalakon is elérhetők, a szerző ezek közül a címlap, a US, a Europe, az Asia és a Hong Kong oldalak tartalmát használta fel. A módszer feltételezi, hogy minden kereskedési napon a tőzsde nyitása előtt készül az előrejelzés az aznapi záróárfolyamra. Ez azt jelenti, hogy a WSJ korábban említett öt oldalán, az adott reggelen található hírek képezik a rendszer egyik inputját, míg az előző kereskedési nap záróárfolyama a másik inputját. A modell kétféle outputtal rendelkezik. A reggeli hírek alapján előrejelzést ad, hogy az aznapi záróárfolyam magasabb, alacsonyabb, vagy nagyjából változatlan lesz-e az előző záróárfolyamhoz képest. Ez a rendszer diszkrét értékű outputja, a rendszernek van folytonos outputja is. Az aznapi diszkrét árfolyam-előrejelzés és az előző záróárfolyam alapján becslést ad az aznapi záróárfolyam értékére.

Leung Kung Fan (1997) a diszkrét előrejelzéshez Probabilistic Datalog nyelven megadott szabályokat használt. A szabályokat megfogalmazhatják szakértők is, de generálhatók korábbi megfigyelések segítségével is. A szabálygeneráláshoz szükséges adatbázis korábbi kereskedési napok híreinek és záróárfolyamainak reprezentációit tartalmazza. A hírek reprezentálásához a szerző a szózsákmodell speciális változatát használta, és bizonyos kettő, három, illetve négy elemű szósorozatok, kifejezések előfordulásait azonosította a hírek szövegében. Az elemzéshez szakértők 125 szósorozatból álló listát állítottak össze. A napi hírek összessége jellemezhető egy-egy 125 elemű dokumentumvektorral, amelynek minden pozíciója egy szósorozatnak felel meg. A vektor adott pozíción lévő eleme pedig olyan súlyszám, amely a szósorozat hírekben való előfordulását jellemzi az adott napon. A kifejezések súlyozására a következő sémát alkalmazta a szerző:

$$w(i,t) = TFF(i,t) \cdot DDF(i) \cdot NF \qquad (1)$$

ahol:

i: dokumentumvektor-pozíció, amely egy adott kifejezéshez tartozik

t: a nap azonosítója, amelyhez a dokumentumvektor tartozik

$w(i,t)$: a t naphoz tartozó dokumentumvektor i kifejezéshez tartozó súlya

$TFF(i,t)$: az i kifejezés előfordulásainak száma a t naphoz tartozó hírekben

$DDF(i)$: az i kifejezés korpuszon belüli megkülönböztető erejét mérő faktor

NF : normalizációs tényező

A DDF – document discrimination factor, dokumentum-megkülönböztetési tényező – szerepét két különböző mutató is betöltheti. Az egyik az inverz dokumentumgyakoriság (IDF, inverse document frequency), a másik a kategória-megkülönböztetési tényező (CDF, category discrimination factor). A szerző kétféle normalizálást alkalmaz, az egyik esetben a napok között normalizál, és minden kifejezés esetén más tényezőt alkalmaz: $NF(i)$. A másik esetben a kifejezések között normalizálja a súlyokat, naponként eltérő normalizációs tényező szerint: $NF(t)$.

$$NF(i)=\frac{1}{\max_t(TFF(i,t)\cdot DDF(i))} \quad (2) \quad \text{illetve} \quad NF(t)=\frac{1}{\max_i(TFF(i,t)\cdot DDF(i))} \quad (3)$$

Az árfolyamok reprezentálására a folytonos skálán mért előrejelzéshez a napi hozamot választotta a szerző. A diszkrét előrejelzéshez a napi hozamokat értéküktől függően három különböző kategóriába sorolta: *fel, le, változatlan*. Ha a hozam 0,3%-nál nem kisebb, akkor a *fel* kategóriába, ha −0,3%-nál nem nagyobb, akkor a *le* kategóriába, különben a *változatlan* kategóriába kell sorolni. A küszöbszintek megállapításakor a szerző törekedett arra, hogy a három kategóriába közel azonos számú megfigyelés kerüljön.

A vizsgálat alapját képező minta az 1996. június 6. – november 14. közötti 100 kereskedési nap adatait foglalja magában. A minta tehát 100 darab, egyenként 126 elemű vektorból épül fel. Ebből 125 elem írja le az adott nap reggelén elérhető hírek tartalmát, és 1 elem az adott nap záróárfolyamának a megelőző napéhoz viszonyított nagyságát. A mintát többféleképpen osztja meg a szerző a modell paramétereinek becslésére szolgáló tanítóminta és a becsült modell tesztelésére szolgáló validációs minta között. Ez egyrészt azt jelenti, hogy kétféle megosztási arány volt: 60-40, illetve 80-20 százalék, továbbá időbeli sorrendiség alapján is kétféle megoldást alkalmaztak: az egyik megközelítés szerint az időrendben korábbi megfigyelések kerültek a tanítómintába, a másik szerint pedig véletlenszerűen.

A mintából Probabilistic Datalog szabályok nyerhetők. A Datalog egy szabálynyelv, amelyet elsősorban adatbázisokhoz használnak lekérdezőnyelvként, ennek súlyozott információk kezelésére alkalmas kiegészítését használta a szerző. A szabály tulajdonképpen egyfajta implikáció, amely két részből áll, a fejből – AKKOR – és a testből – HA. A testben különböző kifejezések kombinációi szerepelnek, a fejben pedig egy hozamkategória. Az *1. ábrán* a *le* kategóriához tartozó szabályrendszer látható.

$$
\begin{aligned}
hsi_{down}(t+1) &\leftarrow \begin{cases} stocks\,are\,mixed(t), \\ NOT[\,stocks\,were\,mixed(t)], \\ NOT[\,hang\,seng\,index\,fell(t)], \\ NOT[\,dollar\,edged\,lower(t)] \end{cases} \\
hsi_{down}(t+1) &\leftarrow dollar\,rose(t) \\
hsi_{down}(t+1) &\leftarrow political\,worries(t) \\
hsi_{down}(t+1) &\leftarrow yield\,rose(t)
\end{aligned}
$$

1. Ábra: Példa a *le* (down) kategóriához tartozó szabályrendszerre

Forrás: (Leung Kung Fan 1997, o.27)

Az algoritmus minden kategóriára egy-egy szabályrendszert állít elő. A szabálygenerálás során az egyelemű szabálytesteket addig bővíti az algoritmus újabb elemekkel, amíg a kategorizálás találati aránya javul. Ebben a lépésben a szabály fejében lévő kategóriába kerül besorolásra a megfigyelés, majd összehasonlításra kerül a tényleges kategóriával.

A tesztelés során minden megfigyelés esetén kategóriánként kiértékelésre kerülnek a szabályrendszerek, aminek eredményeként megkapjuk, hogy mekkora súllyal, mekkora hihetőséggel következnek be az egyes kimenetek, azaz kategóriák. Leung Kung Fan (1997) ezeket a súlyokat kategóriánként eltérő küszöbértékek szerint binarizálta, és ha pontosan egy kategória súlya haladta meg a küszöb értékét, akkor azt tekintette előrejelzésnek. Egyéb esetben a súlyok küszöbértékhez viszonyított relatív nagysága alapján döntött, illetve voltak még olyan típusú kísérletek is, amelyekben a három kategória súlyaiból képzett vektor legközelebbi szomszédjának tényleges kategóriáját alkalmazta. A folytonos árfolyam-előrejelzéshez lemérte mindhárom kategória átlagos hozamát különkülön, majd ezen átlagos hozamok közül az előrejelzett kategóriához tartozóval megnövelte az előző záróárfolyamot. A rendszer előrejelző képessége nem túl jó, a diszkrét előrejelzés során egyik kísérlet eredménye sem éri el az 50%-os pontosságot, a folytonos előrejelzések átlagos relatív hibája 0,62% és 0,86% között volt. Még a legkisebb relatív hiba is meghaladja a hozamok kategorizálására használt küszöbértékek távolságát, tehát 0,3% és −0,3% különbségét.

Peramunetilleke (1997) modellje sokban hasonlít Leung Kung Fanéhoz (1997), azonban nagyfrekvenciás devizaárfolyam-adatok és szalagcímek képezik az inputját. A szerző a HFDF93[14] adathalmazzal dolgozott, amely 1992. október 1. és 1993. szeptember

14 Az adathalmazt az Olsen & Associates szolgáltatta (Peramunetilleke 1997).

30. között percre pontos időbélyegzővel ellátott bid devizaárfolyamokat – USD/DEM, USD/JPY, DEM/JPY – és szalagcímeket[15] tartalmaz. A cél, hogy napon belüli előrejelzést adjon az 1, 2 illetve 3 óra alatti árváltozás előjelére[16]. A hozamok kategóriákba sorolásához a −0,023% és 0,023%-os intervallumhatárokat használta, mert így nagyjából egyenlő számosságú lett a három halmaz. Leung Kung Fanhoz (1997) hasonlóan szósorozatokkal reprezentálta a hírek szalagcímét, ám Peramunetilleke ezek szövegbeli azonosításakor szótövezést alkalmazott, illetve figyelembe vette, hogy más szavak is ékelődhetnek a sorozat elemei közé. Szakértők révén nagyjából 400 darab, legfeljebb négy tagú kifejezésből állt a lista. A kifejezések súlyozására többféle séma közül a korábban említett CDF -et és $NF(i)$ -t alkalmazó bizonyult a leghatékonyabbnak, mindazonáltal a modell a legtöbb esetben 50% alatti találati aránnyal jelzett előre. A véletlenszerű találgatáshoz, tehát kb. 33%-hoz viszonyítva szignifikánsan jobbnak ítélte az eredményeit. Különböző kísérletekben vizsgálta a periódus időtartamának megváltoztatását, illetve összehasonlította a német márka és a japán jen amerikai dollárral szembeni árfolyamának előrejelezhetőségét. Az időtartam tekintetében nem talált lényeges eltérést, a devizák közül pedig a márka bizonyult pontosabban előrejelezhetőnek, amire lehetséges magyarázat, hogy a szakértők által adott kifejezések inkább az európai folyamatokra koncentráltak.

1998-ban Wüthrich és szerzőtársai (Wüthrich, Permunetilleke, et al. 1998; Wüthrich, Cho, et al. 1998) továbbfejlesztették a Leung Kung Fan (1997) dolgozatában bemutatott modellt, illetve több, korábban nem említett részletet is bemutattak. Az előrejelzéseket 7:45-kor készíti a rendszer Hong Kong-i idő szerint, mielőtt bármely általuk vizsgált ázsiai tőzsde kinyitna. A továbbfejlesztett változatban ugyanis a HSI-n kívül a Nikkei 225 (NKY), a Singapore Straits Index (STI) értékének előrejelezhetőségét is vizsgálják, és ezen kívül a Financial Times 100 Index (FTSE) és a Dow Jones Industrial Average (DJIA vagy Dow) is a vizsgálat tárgyát képezték.

A hírek reprezentálásához használt kifejezéslista ekkor már 423 szósorozatot számlált, és ötelemű sorozatok is voltak közöttük. A kifejezések azonosítása finomodott az egy évvel korábbihoz képest, azaz a nem pontos egyezéseket is figyelembe vették, és szótövezés révén azonos alakra transzformálták a különböző formájú szavakat. A Leung Kung Fannál (1997) bemutatott súlyozási sémák közül azt alkalmazták, amelyben a

15 Money Market Headline News
16 Fel, le, változatlan.

CDF -et használták az $NF(i)$ normalizációs változattal, amit azzal indokoltak, hogy az egy évvel korábbi vizsgálatban ezzel érték el a legnagyobb találati arányt. A hozamok kategóriákba történő besorolására nagyobb abszolút értékű küszöböket használtak: a korábbi ± 0,3% helyett ± 0,5% lett.

Kitértek részletesebben a Probabilistic Datalog modell paraméterbecslésekor alkalmazott módszerre is, pontosan leírták, hogy a mintának melyik részén tanítják a modellt, és melyik részén tesztelik. Egy időbeli ablakot mozgatnak végig a mintán úgy, hogy a tanítóminta mérete 100 kereskedési nap, a tesztmintáé pedig 60[17]. A tanítóminta tehát minden tesztadat esetén más: a tesztadatot időben közvetlenül megelőző 100 megfigyelésből áll. Mivel a tanítóminta időben csúszik, ezért mind a 60 tesztadat esetén új szabályrendszert generálnak, amely felhasználásával becslést adtak a szóban forgó tesztmegfigyelés kategóriájára. A folytonos árfolyam-előrejelzés módjáról is pontosabb képet kapunk. A szerzők a szabálygenerálás után a tanítómintán elvégzik az osztályozást, majd a modell által becsült – tehát nem a ténylegesen megfigyelt – kategóriacímkék szerint elkülönítve kiszámolják az egyes hozamkategóriák átlagos hozamát, majd ezzel az átlagos hozammal korrigálják a tesztadat előtti kereskedési nap záróárfolyamát.

A szabályalapú rendszer eredményeinek összehasonlításához becsültek egy legközelebbi szomszéd (k-NN, k-nearest neighbours), kétféle mesterséges neurális hálózat (ANN, artificial neural networks) és egy regressziós módszert alkalmazó modellt is. Ezek közül a k-NN modell találati aránya meghaladta az 50%-ot a HSI esetén, egész pontosan 53% volt. A szerzők úgy gondolják, hogy spekulációs ügyletekhez döntéstámogatásra alkalmas a rendszer, és nyereséges kereskedési stratégia alakítható ki a segítségével. Az eredmények alapján várható profit számítása módszertanilag kifogásolható, mindenesetre a tesztidőszakra számított várható hozam minden esetben pozitív, és a Dow és az NKY esetében abszolút értékben is nagyobb az index hozamánál.

A modell lényegében változatlan formában maradt a PAKDD-99 konferenciáig (Cho & Wüthrich 1999), csak a kutatás fókusza változott, amely a több különböző hírforrás alapján készített előrejelzések kombinálása lett. Felmerül a hírforrások összehasonlításának kérdésköre is, amelyet két dimenzió szerint végeztek el: a minőség és a pontosság alapján. Előbbi lehet szubjektív mérték, de a szerzők az adatforrások tartalma alapján számított objektív mértéket használtak. A pontosság pedig az adatforrás felhasználásával

17 A tesztminta az 1997.12.06. és 1998.03.06. közé eső megfigyelésekből állt.

elérhető becslés pontosságát jelenti. Az előrejelzések kombinálásakor alkalmazott módszerek egy része tekintettel volt ezekre a mutatókra.

A modellben alkalmazott súlyozási sémák javítására tettek kísérletet Cho és szerzőtársai (1999). Némileg módosultak a modell egyes részei is. A hírek öt weboldal összesen 41 aloldaláról származtak, melyek mindegyike a HSI szempontjából releváns lehetett. A hírek szövegének reprezentálásához használt kifejezéslista 392 elemű volt, melyet befektetési elemzők és devizapiaci közvetítők állítottak össze speciálisan a HSI előrejelzéséhez. A tanulmány összehasonlította, hogy a kifejezések különböző súlyozása esetén az előrejelzés pontossága mekkora, illetve milyen stabil különböző hírforrásokra nézve. Az újonnan bevezetett kifejezéssúlyozási sémák a hozamkategóriák szerint különböző súlyt rendeltek a kifejezéshez. A súlyozás a kategória centroidjától való távolsággal fordítottan arányos és 0 és 1 közé transzformált. A kifinomultabb változatban a multimodális eloszlású kategóriáknál előbb klaszterekre bontották a kategóriát, és a klasztercendroidoktól vett távolság alapján határozták meg a súlyokat. Ez utóbbi változat bizonyult a legstabilabbnak a különböző hírforrásokat alapul vevő előrejelzések esetén.

2.1.2. A Lavrenko-féle modell

Victor Lavrenko és szerzőtársai 1999-ben kezdtek el dolgozni Ænalyst nevű rendszerük fejlesztésén a University of Massachusetts-en (Lavrenko et al. 1999). Egy évvel később az elkészült rendszer segítségével nyert eredményeiket a KDD 2000[18] és CIKM 2000[19] konferenciákon is bemutatták (Lavrenko et al. 2000a; 2000b). A Lavrenko-féle modellt fejlesztették tovább Fung és szerzőtársai 2002 és 2005 között. A Chinese University of Hong Kongot képviselő szerzőgárda először 2002-ben a Workshop on Data Mining and Modeling és a PAKDD 2002[20] konferenciákon publikálták eredményeiket (Fung et al. 2002b; Fung et al. 2002a). 2003-ban a CIFEr2003[21] konferencia után Fung szakdolgozata összegezte az addigi kutatást (Fung et al. 2003; Fung 2003). A szerzők 2005-ben publikálták utoljára[22] a Lavrenko-modellre épülő előrejelző rendszerüket az

18 The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Workshop on Text Mining
19 Ninth International Conference on Information and Knowledge Management
20 6th Pacific-Asia Conference
21 2003 International Conference on Computational Intelligence for Financial Engineering
22 Fung és korábbi konzulense, Yu később jelentős változtatásokkal új modellt közölt (Wu et al. 2008; 2009), illetve Yu 2010-ben a volatilitás előrejelzése irányába folytatta a kutatást (Pan et al. 2010).

IEEE Intelligent Informatics Bulletin hasábjain (Fung et al. 2005). A következőkben bemutatom az egyes modellváltozatok lényeges elemeit.

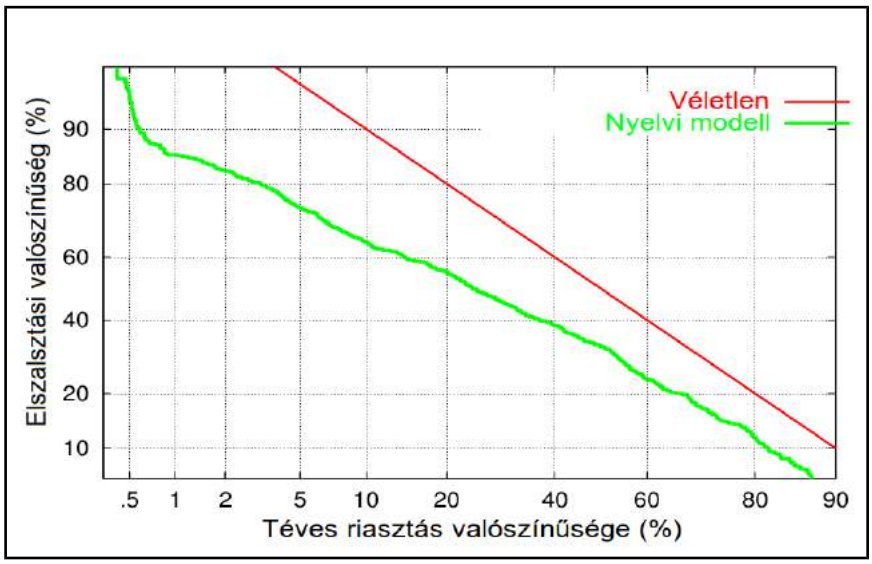
A Lavrenko-féle változatban (Lavrenko et al. 2000a; 2000b) a modell inputját képező hírek és árfolyamok a Biz Yahoo! szolgáltatásból származtak. Az adatforrás mind a 127 vizsgált részvény esetén tartalmazta a releváns híreket, összesen 38469-et, és tíz perces periódusokban az árfolyamokat 1999. október 15-től 2000. február 10-ig. A hírek szövegét a szózsákmodell szerint reprezentálták. A Wüthrich-féle modellhez képest ebben két eltérés is volt. Az egyik, hogy az azonos periódusba, pl. nap, óra stb., eső hírek szavait nem ömlesztették egybe, a hírek önálló egységet képeztek. A másik, hogy nem alkalmaztak szakértők által összeállított kifejezéslistát, az angol nyelv szavait használták. Az árfolyamok reprezentálásában jelentős különbség a Wüthrich-féle modellhez képest, hogy nem azonos hosszúságú időszakok árfolyamváltozásait – pl. napi, egyórás hozamok stb. – jelezte előre a modell, hanem változó hosszúságú lineáris trendeket. A különböző részvényárfolyam-idősorokat szakaszonkénti lineáris regresszióval osztották trendekre, mely során felülről lefelé szemléletben mindaddig két részre vágták a korábbi trendvonalat, amíg az átlagos négyzetes hiba csökkent, illetve a két új trendvonal meredeksége szignifikánsan különbözött. A trendeket meredekségük szerint különböző csoportokba sorolták – erős emelkedés, gyenge emelkedés, erős csökkenés, gyenge csökkenés, nincs javaslat – attól függően, hogy a legnagyobb meredekségű trendhez képest mekkora volt azok meredeksége[23].

A modell paramétereinek becsléséhez használt mintában a megfigyelések az egyes, különálló hírek voltak, melyeket a szózsákmodellnek megfelelően vektorokkal jellemeztek, amelyekben a dimenziók egyrészt az angol nyelv szavai voltak – a súlyozási sémára nem tértek ki –, másrészt a korábban említett trendtípusok. Ugyanaz a hír több trendhez is tartozhat a modell szerint, ám nem tértek ki, hogy az adatbázisban ezt az információt hogyan tárolták. Elképzelhetőnek tartom, hogy típusonként külön dimenziót vezettek be, vagy annyiszor ismételték meg a szóvektort, ahányféle trendtípushoz rendelhető. A hírt hozzárendelték egy trendtípushoz, ha az adott trendtípus kezdete előtt legfeljebb h órával jelent meg. Különböző kísérletekben $h=10$, $h=5$, illetve $h=1$ volt, illetve volt olyan változat is, ahol az adott trend időtartama alatt megjelent híreket rendelték a trendhez, tehát $h=0$.

23 Azok a trendvonalak, amelyek meredekségének abszolút értéke a legnagyobb meredekség 0,75-szeresénél nem volt kisebb, *erősnek* mondhatók. *Gyengék*, amelyek az iménti határérték alatt vannak, de a 0,5-szörös határnál nem kisebbek. A többi trendet a *nincs javaslat* kategóriába sorolták.

A szerzők nyelvi modelleket – language model – készítettek mind a négy trendtípushoz. A nyelvi modell egy szóhasználati mintát jelent, amely a szerzők feltételezése szerint a különböző trendtípusokhoz kapcsolódó hírek esetén különböző lehet. Feltételezik továbbá, hogy részvényenként is eltérő nyelvi modellek írhatják le az árfolyammozgásokat, ám a mintában nagyon egyenlőtlenül oszlik meg a hírek száma az egyes részvények között, így a kereskedési szimulációt kivéve nem alkalmazzák ezt a módszert. A híreket a bennük lévő szavak halmaza alapján adott valószínűséggel – illetve likelihoodok szerint – sorolják az egyes trendtípusokba.

A modell jóságát tízszeres keresztvalidációval tesztelték, mely során mind a tíz esetben a megfigyelések 90%-a véletlenszerűen került a tanítómintába. A négy nyelvi modellt minden validációs lépésben újratanították, majd külön-külön kiértékelték. A négy trendtípus előrejelezhetőségét a DET-görbék – detection error tradeoff, felismerési hibák átváltási görbéje – segítségével értékelték, melyen megfigyelhető, hogy az első- és másodfajú hibák százalékai hogyan alakulnak, ha változtatjuk azt az értékhatárt, ami fölé a likelihoodnak esni kell, hogy a megfigyelést az adott trendhez soroljuk (lásd *2. ábra*). A görbét a véletlenszerűség esetén várható és egy másik modell, a koszinusz-hasonlóság alapú osztályozó DET-görbéjéhez mérték. Az erős csökkenés esetét kivéve modelljük jobbnak bizonyult mindkét referenciához képest, az említett esetben pedig a koszinusz-hasonlóságon alapuló osztályozó teljesítménye megközelítette a nyelvi modellét.



2. Ábra: Az erős emelkedő trendhez (SURGE) tartozó 10 órás nyelvi modell DET-görbéje

Forrás: (Lavrenko et al. 2000a)

A tesztet egy kereskedési szimulációval is kiegészítették, melyben három hónapnyi tanítómintán, részvényenként és trendtípusonként egy-egy nyelvi modell paramétereit becsülték meg, majd megvizsgálták, hogy 40 nap adatain egy egyszerű kereskedési stratégiával[24] mekkora profit érhető el. A híreket ahhoz a trendhez sorolták, amelyik ideje alatt megjelentek. A megfigyelt 280000 dolláros, pozíciónként átlagosan 0,23%-os összprofitot 1000 véletlenszerű kereskedés[25] által elért profittal hasonlították össze. Az ezerből csak nyolc véletlenszerű kereskedés ért el nagyobb profitot.

Fung (2002a; 2002b) a modellt két dolog miatt kritizálta. Az egyik, hogy feltételezi, hogy a hírek hatása h óra alatt, azaz a hatékony piacok hipotézisének ellentmondva nem azonnal épül be az árfolyamba. A másik kritika szerint mivel egy adott hírt több trendhez is hozzá lehet rendelni, két ellentétes címkét is kaphat ugyanazon dokumentum. Az első kritikát illetően meg kell jegyezni, hogy Lavrenko három különböző trend-hír párosítási stratégiát is leír, amelyek közül az egyik az EMH-nak is megfelelő egyidejű hozzárendelés. A második kritikára szintén megtalálható a válasz Lavrenko munkáiban, miszerint ha a hírek hatása hosszabb ideig érvényesül – feltételezve, hogy az EMH nem érvényes –, akkor abban az időtartamban egymás után akár több különböző irányú ármozgás is megfigyelhető.

Fung 2002-es modellváltozatához a szöveges és idősoros inputadatok a Reuters-től[26] származtak. (Fung et al. 2002a; 2002b) A mintában[27] 350000 hír szerepelt, illetve napon belüli kötési árfolyamok 614 darab Hong Kong-i tőzsdén jegyzett vállalat részvényéhez kapcsolódóan. A híreket vektortérmodell alapján reprezentálták, a szavakat kisbetűssé konvertálták, szótövezésen estek át, és kiszűrték közülük a stopszavakat. Az alkalmazott súlyozási séma a következő volt:

$$w(t,d)=tf_{t,d}\cdot CDC\cdot CSC \tag{4}$$

Ahol t a kifejezés, d a dokumentum, $w_{t,d}$ a súly, $tf_{t,d}$ a szógyakoriság – term frequency –, továbbá CDC a klaszteren belüli elkülönülési együttható – inter-cluster discrimination coefficient – és CSC a klaszterek közötti hasonlósági együttható – intra-cluster similarity coefficient – , melyek képlete a következő:

24 A szimuláció során 10000 dolláros long pozíció nyílik, ha a hír alapján növekvő trend várható leginkább, short, ha negatív. A pozíció legfeljebb 1 óráig marad nyitva, de előbb záródik, ha a pozíción elérhető profit legalább 1% lesz. A kereskedési nap végén nem feltétlenül záródnak a pozíciók.

25 A véletlenszerű kereskedések ugyanabban az időpontban nyithattak pozíciót, mint a nyelvi modell alapján működő szimuláció, tehát, amikor egy hír megjelent. A long és short pozíciók közötti választás volt véletlenszerű, de olyan valószínűséggel nyithatta meg a rendszer a két pozíciót, amekkora arányban a nyelvi modellnél is előfordultak. A többi szabály megegyezett a nyelvi modellével.

26 A híreket, és az árfolyamokat a Reuters Market 3000 Extra szoftveren keresztül töltötték le.

27 2001.10.01. – 2002.04.01.

$$CDC = \left(\frac{n_{i,t}}{N_t} \right)^2 \qquad (5) \qquad\qquad CSC = \sqrt{\frac{n_{i,t}}{n_i}} \qquad (6)$$

Ahol i a klaszter, $n_{i,t}$ dokumentumgyakoriság klaszteren belül, N_t dokumentumgyakoriság a teljes korpuszban, n_i szavak száma a klaszteren belül.

Az árfolyamokat trendekre bontották, ám a Lavrenko által használt szakaszonkénti lineáris regresszió helyett saját, t-próbán alapuló felosztó-összevonó szegmentációs algoritmust alkalmaztak. A trendeket két tulajdonságuk, a meredekség és az R^2 mutató alapján három klaszterbe sorolták, majd a klaszterek átlagos meredekségének viszonya alapján hozzájuk rendelték a növekvő, csökkenő és változatlan címkéket[28]. A hírek és árfolyamok egymáshoz rendelésekor az EMH elvét követték, és a hírt ahhoz a trendhez rendelték hozzá, amelyik publikálásának idején megfigyelhető volt. Az osztályozáshoz a Hong Kong-i szerzők két bináris SVM-osztályozót (support vector machine, támasztóvektor-gép) használtak. Az egyik osztályozó a növekvő, a másik a csökkenő kategória elkülönítésére szolgált. A hírnek megfelelő dokumentumvektort kiértékeli mindkét modell, majd ha csak az egyik jelzi azt, hogy a dokumentum az adott kategóriához tartozik, akkor az lesz az előrejelzés. Ha egyik sem jelez, akkor úgy veszik, hogy nincs javaslat, ha mindkettő, akkor pedig ellentmondásos a javaslat, az adott hírt mindkét esetben elveti a rendszer. Az osztályozás minőségének megítélésére ROC-görbét (receiver operating characteristic, vevő működési jelleggörbe) alkalmaztak a szerzők, tízszeres véletlenszerű keresztvalidációval. Eredményeik szerint a modell a véletlenhez képest jobban teljesít, és az is javít az előrejelzés minőségén, ha kiszűrik azokat a dokumentumokat, amelyek az ellenkező trendkategóriába tartozó hírekhez hasonlítanak. Lavrenko-hoz hasonlóan Fung is végzett piaci szimulációt a rendszer gyakorlati alkalmazhatóságának megítéléséhez. A részvényeket a hozzájuk kapcsolódó hírek száma alapján 14 egyenlő számosságú intervallumba sorolta, és azt tapasztalta, hogy a túl kevés, vagy éppen sok hír rontja a szimuláció során elérhető profitot. A magyarázatuk, hogy a kevés hír miatt bizonytalan a modell paraméterbecslése, a sok hír pedig jelentős zajt idéz elő.

Fung később frissebb, 2003.01.20. – 2003.06.20. közötti mintán megismételte a kísérletet (Fung et al. 2003; 2005; Fung 2003). A hírszövegek reprezentálásakor statisztikai dimenziócsökkentés, χ^2-próba révén kiszűrte a nem diszkriminatív szavakat is. A tanítóminta összeállítása során újítás volt, hogy a két SVM számára külön-külön hatá-

28 A klaszterezéses módszert Lavrenko és szerzőtársai is alkalmazták, ám mivel a csoportok elkülönítésében csak a meredekség játszott szerepet, végül nem ezzel a módszerrel csoportosították a trendeket.

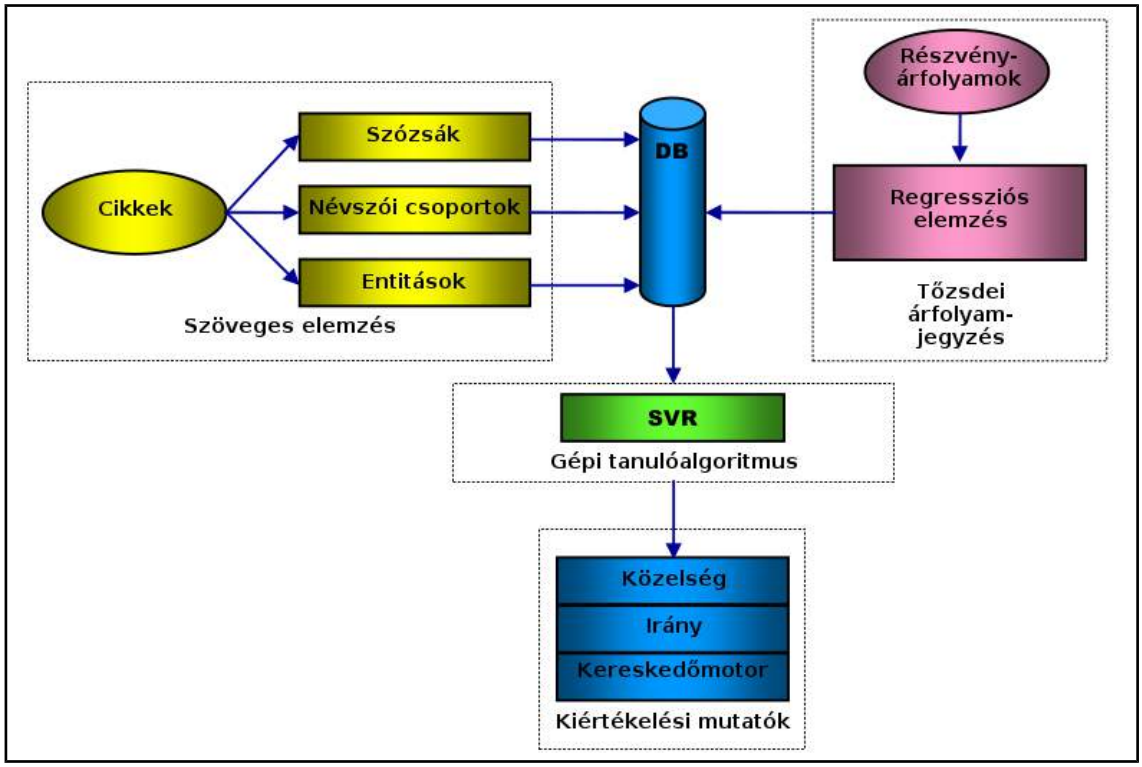
rozta meg a pozitív és a negatív tanítóadatokat. A növekvő/csökkenő kategóriához tartozó SVM pozitív tanítóadatai azok a hírek voltak, amelyek emelkedő/csökkenő trend közben jelentek meg, és tartalmaztak legalább egy olyan szót, ami diszkriminatívnak tekinthető a növekvő/csökkenő kategóriára nézve. A növekvő/csökkenő kategóriához tartozó negatív tanítóadatok azokból a hírekből állnak, amelyek nem tartalmaznak a növekvő/csökkenő kategóriára nézve diszkriminatív szavakat. A tesztelés során a kötegelt és az online megközelítés hatását is vizsgálta. A kötegelt megközelítésben a teljes minta 70%-a szolgált tanításra, 30%-a tesztelésre. Az online megközelítésben a teljes mintát tíz egyenlő részre osztotta, és minden lépésben a tanítóminta kibővült az előző lépés tesztmintájával. A tesztminta időben a tanítóminta után következő egy tizednyi megfigyelésből állt minden esetben. Az SVM- és a naiv Bayes-osztályozó mellett egy hasonlóságalapú osztályozót alkalmazott a szerző. Az osztályozók eredményének összemérését nem csak a saját mintán, hanem a Reuters-21578, és a Newsgroup-20 adatbázison is elvégezte. Az eredmények szerint a naiv Bayes-osztályozó minden mutató szerint – felidézés (recall), precizitás (precision), F1 – alul marad a másik két algoritmussal szemben. A másik két hasonló teljesítményű módszer közül a hasonlóságalapú számításigénye kisebb. A piaci szimulációs vizsgálat is új szempontokkal gyarapodott. A szimuláció során a rendszer által adott előrejelzésnek megfelelő irányú pozíciót nyitott a rendszer, amelyet m nap múlva zárt le. Az m paraméter változtatásának hatását az úgynevezett találati rátán – hit rate – mérték le. A legmagasabb találati rátát, 67,4%-ot $m=5$ esetén érte el, amely arra utal, hogy nagyjából egy hetes trendek előrejelzéséhez járulnak hozzá a hírek. Ezen kívül rögzített $m=3$ paraméter mellett Fung két benchmarkhoz mérte a hozamot, az egyik a buy-and-hold hozam, a másik pedig 1000 véletlenszerű kereskedés hozama volt. A buy-and-hold stratégiánál, azaz $-20{,}56\%$-nál magasabb hozamot ért el a hírek alapján szimulált kereskedés, 28,06%-ot, és csak 38 olyan nagyobb hozamú véletlenszerű kereskedés volt, amely ugyanabban az időpontban, ám véletlenszerűen nyitott megbízást. A szimuláció tehát a véletlennél szignifikánsan jobb hozamot ért el.

2.1.3. A Schumaker-féle modell

Robert Schumaker, Egyesült Államokbeli kutató, 2006-ban mutatta be először az AZFinText[29] nevű alkalmazását egy konferencián (Schumaker & Chen 2006), majd ez a szakirodalmi áttekintéssel egybekötött írás különböző szövegreprezentációk teljesítményre gyakorolt hatásának elemzésével kiegészülve megjelent az ACM Transactions

29 Arizona Financial Text System

on Information Systems hasábjain is (Schumaker & Chen 2009b). 2009–2010 során a szövegreprezentációk elemzését több ízben kiegészítette nyelvtani jellegű szövegjellemzők (Schumaker 2009; 2010a; 2010b), illetve szemantikai jellegű szövegjellemzők, mint az objektivitás és tájolás összehasonlításával (Schumaker et al. 2009; 2012). Másik kutatási témája a rendszer profitabilitásának megítélésére szolgáló benchmarkok körének kiválasztása, illetve a különböző stratégiáknak a modellbe történő bevonása volt (Schumaker & Chen 2008; 2010; 2011). Schumaker kísérleteihez végig ugyanazt az adathalmazt alkalmazta, így eredményei jól összehasonlíthatók egymással. A híreket a Yahoo! Finance híraggregátor szolgáltatásból töltötte le 2005. október 26. és november 28. között azon vállalatokhoz kapcsolódóan, amelyek 2005. október 3-án az S&P500 index részét képezték. E hírek közül kizárta azokat, amelyek nem kereskedési időben jelentek meg, illetve, hogy kiküszöbölje a nyitás előtti hírek befolyását a kísérleteire, kizárta azokat is, amelyek a tőzsde nyitása utáni egy órában kerültek publikálásra. Annak érdekében, hogy az AZFinText előrejelzését össze tudja hasonlítani a tényleges árfolyammal, szükséges volt, hogy a zárás előtti 20 percben megjelenő híreket is kizárja a korpuszból. Ennek eredményeképpen valamivel több mint 2800 hírből tevődött össze a korpusza. A fent említett cégek és időintervallumok esetén több mint tízmillió perces árfo-



3. Ábra: Az AZFinText rendszer felépítése

Forrás: (Schumaker & Chen 2006, o.11)

lyamadatot tartalmazott Schumaker mintája, amely egy meg nem nevezett kereskedelmi adatbázisból származott. Az AZFinText rendszer egy szövegreprezentációs komponensből, egy SVR – support vector regression, támasztóvektor-regresszió – tanulóalgoritmusból, egy előrejelző komponensből, illetve kétféle eredménymérő komponensből áll, melyből az egyik az előrejelzés pontosságát méri, a másik pedig az előrejelzésen alapuló kereskedési stratégia profitabilitását (lásd *3. ábra*).

Schumaker többféle szövegreprezentációt is összehasonlított tanulmányaiban, de a közös rendezőelvek mindegyiknél a következők voltak: a kifejezések bináris reprezentációját alkalmazta és ezzel párhuzamosan kiszűrte azokat a kifejezéseket, amelyek 3-nál kevesebbszer fordultak elő a korpuszban. Utóbbi segítségével a dimenziók száma jelentősen csökkenthető. A becsléshez az SVR algoritmust a Weka szoftver Sequential Minimal Optimization funkciójával implementálta. A magyarázó változók köre tanulmányonként eltérő volt, de minden esetben a hír publikálásakor érvényes részvényárfolyam mellett a hírek szövegéből kinyert jellemzők – kifejezések, objektivitás, tájolás stb. – töltötték be ezt a szerepet. A magyarázott változó a hír publikálása után 20 perccel megfigyelt árfolyam volt. Ez előrejelzési időtáv megválasztásában nagy szerepe volt Gidófalvi ez irányú kutatásának (Gidófalvi 2001; Gidófalvi & Elkan 2003). Közös még az említett tanulmányokban, hogy Schumaker az előrejelzés eredményének kiértékelésére az átlagos négyzetes hibát – MSE, mean squared error –, illetve a változás irányának találati pontosságát – directional accuracy – használta. Az eredményeket tízszeres keresztvalidáció alapján ellenőrizte. Az előrejelzésre alapozott kereskedési stratégiával elérhető hozam mérését Lavrenko (2000a; 2000b) nyomán alakította ki. Minden hír megjelenésekor háromféle döntés közül kellett választani a szimulátornak: ha az előrejelzett árfolyam 1%-nál kisebb mértékben tér el a jelenlegitől, akkor nem ad megbízást, különben értelemszerűen a változás előjelétől függően 1000 dolláros long vagy short pozíciót nyitott. A nyitott pozíciók 20 perc elteltével lezárásra kerültek. A tranzakciós költségeket a rendszer figyelmen kívül hagyta.

Schumaker és Chen (2006) első tanulmánya a témában tartalmaz egy alapos szakirodalmi áttekintést a korábbi pénzügyi hírbányászati rendszerekről, hogy megfelelőképpen beilleszthető legyen kutatása a korábbiak közé. Egyrészt ezt a munkát összehasonlítási alapnak szánta a későbbi tanulmányaihoz, így az előrejelzés eredménymutatói képezik a tanulmány egyik fontos részét. Másrészt a többféle lehetséges szövegreprezentációk közötti választást készíti elő ebben a munkában azzal, hogy az általános szózsákmo-

dell, a névszói csoportok – noun phrases –, illetve entitások – named entities – alkalmazása révén elérhető előrejelzési pontosságokat összehasonlította egymással és a szöveges információt mellőző regressziós modell eredményével. Az átlagos négyzetes hiba tekintetében a regressziós modell 0,072 értéket ért el, amely szignifikánsan nagyobb volt mindhárom szövegreprezentációval elérhető becslési hibánál. A legkisebb MSE az entitások esetén volt megfigyelhető, amely 0,033-as értéket jelentett. Az árfolyamváltozás irányát pusztán numerikus adatok alapján 47,6%-ban jelezte előre helyesen a rendszer, a szöveges információk alapján kapott eredmények ezt mindhárom esetben meghaladták, a névszói csoportok 50,7%-os pontosságot szolgáltattak. A kereskedési szimuláció során szöveges információk nélkül 1800 dolláros veszteség keletkezett, a többi esetben nyereséges volt a kereskedés, a névszói csoportok esetén 6353 dolláros profittal.

Schumaker és Chen (2009b) annyiban bővíti ki a korábbi munkát, hogy bevonja a vizsgált reprezentációk körébe a tulajdonneveket – proper noun – is, amely a névszói csoportok és a jól meghatározott szempontok szerint kategorizált entitások közötti átmenet. Ezen kívül a magyarázó változók köre alapján három kísérletet végzett el. Az elsőben – M1 – kizárólag szöveges információkkal jelezte előre az árfolyamot. A második modell – M2 – a hírek mellett a publikálásukkor érvényes árfolyamot is bevonta. A harmadik – M3 – a hírek mellé a numerikus információk alapján előrejelzett árfolyamot használta magyarázó változóként. Az eredmények alapján az M2-es változat bizonyult a legjobbnak minden szempontból. Az M1 modell átlagos négyzetes hibája 850 körüli, tehát a legrosszabb, míg az M2-é 0,04. A változás irányát szintén az M1 modell becsülte legrosszabbul, ez 54,6%-ot jelent, az M2 pontossága 57%-os volt. A kereskedési szimuláció az M1 modellnél negatív átlagos hozameredményt hozott, -0,05%-ot, az M2-nél 2,06% lett ez az érték. Az MSE alapján az entitások, a pontosság alapján a névszói csoportok lettek a leghatékonyabbak, a tulajdonnevek nem hoztak jobb eredményt.

Még ugyanebben az évben bekerült a vizsgált szövegreprezentációk körébe ötödikként az igékből és határozószókból álló reprezentációs séma (Schumaker 2009). Újabb eredményei szerint az MSE mutató alapján az entitások bizonyultak a legjobbnak 0,0341-es értékkel, míg az igék a másodikak 0,0356-os értékkel. Az árváltozás irányát tekintve a névszói csoportok 51,1%-át megelőzte a tulajdonnevek reprezentációs sémája, amelynek pontossága 51,4% volt. A szimuláció tekintetében az igék bizonyultak a legjobbnak 3,36%-os nyereséggel, a második helyen a tulajdonnevek álltak 2,84%-kal.[30]

30 Ezek az értékek a (Schumaker & Chen 2009b) tanulmányéval nem hasonlíthatók össze közvetlenül, mivel abban iparáganként alá volt bontva a korpusz.

A sémák elemzéséhez újabb adalék volt az a két 2010-es tanulmány (Schumaker 2010a; 2010b), amelyben a modell pontossága helyett azt vizsgálta, hogy melyek azok a szavak, amelyek a leginkább az árfolyam növekedése, illetve csökkenése irányába hatnak. Ehhez az SVR regressziós együtthatóit elemezte, és például az öt legpozitívabb hatású ige vagy határozószó a következő volt: *hereto, comparable, charge, summit, green*[31]*;* az öt legnegatívabb hatású pedig: *planted, announcing, front, smaller, crude*[32] (Schumaker 2010a). A szavak közül néhány szerepel több reprezentációs sémánál is, hiszen az általános szózsákmodell, a névszói csoportok, a tulajdonnevek és az entitások séma egymással részhalmazviszonyban állnak (Schumaker 2010b).

Schumaker azonban nem csak a nyelvtani jellegű reprezentációkat, hanem a szövegből kinyerhető szemantikai jellemzők, mint objektivitás, illetve tájolás alkalmazhatóságát is vizsgálta (Schumaker et al. 2009; 2012). Az objektivitásnak – tone – három értékét különböztetik meg a tanulmányok: objektív, szubjektív, semleges. A tájolás – polarity – szintén három értékkel rendelkezik: pozitív, negatív, semleges. Annak megállapításához, hogy egy szöveg a fenti kategóriák közül melyikbe tartozik, az OpinionFinder nevű eszközt használták. Három modellt hasonlítottak össze, melyek közül az első a szöveg kifejezéseit használta magyarázó változóként, a második ezen kívül az objektivitás mértékét is, a harmadik pedig a kifejezések mellett a tájolást vette figyelembe. Az objektivitással kapcsolatos eredményeik szerint a szubjektív hírek alapján jobban előrejelezhető az árfolyamváltozás iránya – 59% pontossággal és 3,3% hozammal –, mintha csak a szövegbeli kifejezéseket alkalmaznánk, amelynek pontossága 50,4% és hozama 2,41%. Az objektív és semleges híreknél ugyanakkor rosszabb eredményeket kapott. A tájolás kapcsán azt tapasztalta, hogy a negatív hírek hatása jobban előrejelezhető – 50,9% pontosság, 3,04% hozam –, mintha kizárólag a szavak alapján jelzünk előre. A pozitív és semleges hírek alapján ugyanakkor rosszabb az előrejelezhetőség.

A hírbányászati előrejelző rendszer profitabilitásának jobb összehasonlíthatósága érdekében két kereskedési stratégia, a momentum és az anticiklikus – angolul contrarian – stratégia várható hozamait is vizsgálták (Schumaker & Chen 2008). A szöveges előrejelzés minőségének javítása érdekében a vállalatokat iparáganként csoportosította, majd a korpuszt is ennek megfelelően kisebb részekre bontotta, ezzel kiküszöbölték azt a jelenséget, hogy bizonyos hírek egy iparág számára kedvezőek, egy másik számára viszont hátrányosak, és emiatt a hírben szereplő kifejezések árfolyamra gyakorolt hatása nem

31 Közelítőleges magyar megfelelőjük sorrendben: csatolva, hasonló, felszámít, csúcs, zöld.
32 Közelítőleges magyar megfelelőjük sorrendben: telepített, bejelent, elülső, kisebb, nyers.

állapítható meg egyértelműen. Mindkét kereskedési stratégiához két alváltozat tartozott. A momentum stratégia esetén az első alváltozatban csak az elmúlt $1 \leq t \leq 5$ hét alatt legjobb hozamokat produkáló részvények 20%-ából képeztek portfóliót. Ugyanennek a stratégiának a második alváltozata egy hibrid rendszer, amely a legjobb 20%-ba tartozó részvények híreinek szöveges elemzése alapján kereskedett. Az anticiklikus stratégia első alváltozatában csak az elmúlt t hét alatt legrosszabb hozamokat produkáló részvények 20%-ából képeztek portfóliót, míg a második, hibrid változatban e részvények hírei alapján történt a kereskedés. Volt egy harmadik stratégia is, amely tisztán a hírek szövege alapján kereskedett, hasonlóan a korábban bemutatott tanulmányokhoz. A tisztán momentum stratégia minden esetben negatív hozamot hozott, míg a tisztán anticiklikus minden esetben pozitívat, és a legnagyobb hozam 3,36% volt. Kizárólag a hírek alapján 8,5%-os hozamot ért el a szimuláció, míg a piaci hozam 5,62%-os volt, amely buy-and-hold stratégia révén volt realizálható. A hibrid rendszerek viszont ennél is jobban teljesítettek a mintán, a momentum hibrid stratégia esetén akár 20,8%-os hozam[33] is elérhető volt, az anticiklikus hibrid stratégia esetén pedig akár 13,18%[34].

A kvantitatív alapok – quantitative fund – olyan befektetési alapokat jelentenek, amelyeknél a portfólió meghatározására matematikai, statisztikai vagy mesterséges intelligencia módszereket alkalmaznak. A legjobb kvantitatív stratégiákkal szemben az AZFinText viszonylag jól teljesített, a 8,5%-os átlagos hozamát a kvantitatív alapok éves hozamok szerinti toplistáján csak négy alap előzte meg (Schumaker & Chen 2010). Az összehasonlításkor azonban figyelembe kell venni, hogy az AZFinText csak az S&P500-ba tartozó részvényekkel kereskedett. Ha így nézzük, akkor az AZFinText magasan a legnyereségesebb stratégia, ugyanis az S&P500 átlagos hozama csak 5,62% volt a vizsgált időszakban, és a kizárólag az S&P500 részvényekkel kereskedő legjobb kvantitatív alap éves hozama is csak 6,44% volt.

2.1.4. Groth-féle modell

Sven Groth, frankfurti kutató, a német nyelvű sajtóközlemények – ad hoc disclosures – negyedórás hatását a német részvénypiacon eseményvizsgálati és hírbányászati módszertannal is vizsgálta (Groth & Muntermann 2008). A következő évben nagyobb mintára is elvégezték a vizsgálatokat, és a modell kiértékelési szempontjait a kereskedési szimuláció hozamával egészítették ki (Groth & Muntermann 2009). Ezután a modellel a

33 t=1 hét esetén.
34 t=2 hét esetén.

negyed- és félórás abnormális kockázatot kezdte vizsgálni, ami együtt járt azzal, hogy nagyon egyenlőtlen megoszlású kategóriákra kellett előrejelzést készíteni (Groth & Muntermann 2010; 2011). Ezt egyrészt a kategóriákhoz tartozó tévesztési költségek változtatásával, másrészt különböző osztályozó módszerek alkalmazásával próbálták kezelni. Hasonlóan egyenlőtlen megoszlású tanítóadatokkal kellett dolgozniuk, amikor az abnormális likviditás előrejelzésével foglalkoztak, amelynek gazdasági alkalmazása a tranzakciók végrehajtásának időzítése kapcsán merülhet fel (Groth 2010; Groth et al. 2014). Utóbbi tanulmány ugyanazon korpusz angol és német változatát is összehasonlította, illetve különböző szövegreprezentációkkal is elvégezték a kísérletet. Groth 2012-ben védte meg doktori értekezését, amelynek fejezetei a fent említett munkákból álltak (Groth 2012).

Groth és Muntermann (2008) korpusza kezdetben 160 sajtóközleményből állt, amelyeket a DGAP[35] tett közzé 2008.08.01. és 2004.08.31. között, kereskedési időben. A közleményeket hat kategóriába sorolták Muntermann és Güttler (2007) nyomán: pénzügyi jelentések, várt pozitív események, várt negatív események, felvásárlások, eszközök eladása, menedzsmentet érintő változások. A vizsgálat fő célja az volt, hogy megbecsüljék, e kategóriák közül melyeket követnek abnormális árfolyam-reakciók 15 percen belül. Az eseményvizsgálat módszertan szerint a hat kategóriából csupán egy, a pénzügyi jelentések esetén figyelhető meg szignifikáns abnormális árfolyammozgás, így arra a következtetésre jutottak, hogy a hírek szövegbányászati elemzését csak e típus kapcsán érdemes folytatni. Mivel a DGAP nem sorolta be a sajtóközleményeket az említett hat kategóriába, ezért egy kétlépcsős modellt készítettek, amely az első fázisban osztályozta a híreket aszerint, hogy vajon pénzügyi jelentésről van-e szó vagy sem, majd a második lépcsőfokban a pénzügyi jelentések közé sorolt hírek árfolyamra gyakorolt hatását állapították meg. A tanítóhalmaz felcímkézése háromféle módon történt. Az első módszer volt a legegyszerűbb, amely a felső korlát[36] fölé eső 15 perces hozamokat pozitív, az alsó korlát[37] alá esőket negatív, a két korlát közöttieket semleges címkével látta el. A második módszer csak két kategóriát különböztetett meg, azaz, ha a 15 perc alatti átlagos hozam a teljes minta átlaga fölötti, akkor a címke pozitív, különben negatív. A harmadik módszer szintén csak két címkét alkalmazott, azaz ha a hír publikálást követő

35 Deutsche Gesellschaft für Ad-hoc-Publizität
36 A felső korlát 0, illetve 0,01 volt.
37 Az alsó korlát -0,01, illetve 0 volt. Utóbbi esetben a 0 hozamú megfigyeléseknek nem képeztek külön semleges kategóriát, hanem azok is a negatív kategóriába kerültek.

15 percen belül volt egy olyan jegyzés, melynek hozama kívül esik egy meghatározott intervallumon, akkor a címke pozitív, különben negatív[38]. A közlemények szövegét a szózsákmodell szerint reprezentálták, és német nyelvű szótövezés, stopszavazás, valamint gyakoriság alapú szűrés révén csökkentették a kifejezések számát, a kifejezések súlyozásakor tf-idf (term frequency–inverse document frequency, szógyakoriság–inverz dokumentumgyakoriság) sémát alkalmaztak. SVM osztályozó módszert használtak, lineáris kernellel, és a modell teljesítményét a pontosság, felidézés és precizitás mutatókkal jellemezték, és benchmarkként a default (minden megfigyelésre ugyanolyan előrejelzést adó) modellt használták, a robusztusság érdekében 10-szeres keresztvalidációt alkalmaztak. Abban a lépésben, amelyben azt kellett megállapítani, hogy a hír pénzügyi jelentés-e, 92%-os pontosságot értek el szemben a default modell 62%-ával. A második lépésben, azaz a pénzügyi jelentések árfolyamhatásának előrejelzésében, a harmadik típusú címkézéssel érték el a legmagasabb pontosságot, precizitást és felidézést is – rendre 70%, 69% és 85% –, míg ebben az esetben a default modell pontossága csak 55% körüli volt. A híreknek az árfolyam kiugrásaira, volatilitására gyakorolt hatása kapcsán tehát ígéretesebb eredményeket kaptak, mint a hozam előrejelzése kapcsán.

Groth és Muntermann (2009) nagyobb mintán is megismételték a kísérletet, amely 423 sajtóközleményt tartalmazott, melyek 2003.08.01. és 2005.07.29. között jelentek meg, és ugyancsak a DGAP-on keresztül publikálták őket. A kísérleti összeállítás lényegében megegyezett a korábbival, azzal az eltéréssel, hogy a címkézés során csak pozitív és negatív kategóriákat különböztetett meg aszerint, hogy a publikálás után 15 perccel az árfolyam növekedett vagy csökkent. A megfigyeléseknek körülbelül a 60%-a a pozitív kategóriába tartozott, így mintájuk kissé kiegyensúlyozatlan volt. Az SVM 56%-os pontossága nem érte el a default modell kb. 60%-os pontosságát. A modell kiértékelését szakterület-specifikus módszerrel, kereskedési szimulációval elérhető hozamok alapján is elvégezték. A default modellre alapozott kereskedési stratégia átlagos hozama 0,37% volt, amely az 50-50%-os valószínűséggel pozitív, illetve negatív árfolyamváltozásra számító 5000 darab véletlenszerű szimuláció átlagos 0,01%-os hozamához képest jobb volt, ám az SVM előrejelzését követő stratégia 1,05%-os átlaghozamához képest alulmaradt, annak ellenére, hogy az SVM kevésbé volt pontos, mint a default modell.

38 Egyik módszer esetén sem adták meg, hogy melyik kategóriába hány megfigyelés esett, így csak a default modell pontossága alapján lehet következtetni rá, ez alapján viszonylag egyenletes volt a pozitív és negatív kategóriák megoszlása, amikor nem volt más kategória.

Groth és Muntermann (2010; 2011) ezután az árfolyam-volatilitás előrejelzésével foglalkozott. A vizsgálatot mindkét tanulmányban a korábbi 423 elemű korpuszon végezték, a nagyfrekvenciás árfolyamadatokat a Thomson Reuters DataScope Tick History szolgáltatás révén szerezték be a Frankfurt Stock Exchange-ről. A német nyelvű szövegek előkészítése a szózsákmodell szerint történt, Porter-féle szótövezőt (Porter 1980) használtak, és a jellemzőkiválasztás a khi-négyzet statisztika alapján történt. A volatilitást a részvényárfolyam évesített hozamszórásával mérték, míg az abnormális kockázatot – amely megmutatja, hogy egy adott időpontban a várható kockázathoz képest mennyivel tért el a tényleges kockázat – az aktuális hozamszórás és az átlagos hozamszórás különbségével. Az osztályozáshoz szükséges tanítóminta összeállításakor az abnormális kockázat alapján a felső kvartilisbe eső megfigyelések a pozitív osztálycímkét kapták, a többiek a negatívat. A kísérletet elvégezték úgy is, hogy a publikálást követő 15 percben megfigyelt kockázat alapján címkézték a megfigyeléseket, és úgy is, hogy 30 perc volt a vizsgált időtartam. A fenti címkézési módszer miatt háromszor annyi negatív címke van, mint pozitív, tehát a tanítás során a pozitív kategória tévesztéséhez nagyobb költséget kell rendelni, így a megfelelő költségszint megállapítása is a vizsgálat célja volt. Az eredmények robusztusságát 10-szeres keresztvalidációval biztosították.

1. Táblázat: A volatilitás előrejelezhetősége SVM-mel 15 perces (bal) és 30 perces (jobb) időtartamra

Hibás osztályozás költsége kategóriánként		*Pontosság (%)*	*Felidézés (%)*	*Precizitás (%)*	*F_1 (%)*
negatív	*pozitív*				
0,1	0,9	77,07	8,49	100,00	15,65
0,3	0,9	78,49	31,13	64,71	42,04
0,5	0,9	77,30	47,17	55,56	51,02
0,9	0,9	69,50	78,30	43,92	56,27
0,9	0,5	61,94	88,68	38,68	53,87
0,9	0,3	52,25	96,23	34,00	50,25
0,9	0,1	42,79	100,00	30,46	46,70

Hibás osztályozás költsége kategóriánként		*Pontosság (%)*	*Felidézés (%)*	*Precizitás (%)*	*F_1 (%)*
negatív	*pozitív*				
0,1	0,9	77,54	11,32	92,31	20,17
0,3	0,9	78,49	19,81	77,78	31,58
0,5	0,9	76,83	49,06	54,17	51,49
0,9	0,9	71,87	66,98	45,81	54,41
0,9	0,5	51,30	94,34	33,33	49,26
0,9	0,3	44,92	98,11	31,04	47,17
0,9	0,1	39,72	100,00	29,36	45,40

Forrás: (Groth & Muntermann 2010, o.6)

Amennyiben a pozitív kategória tévesztési költsége jelentősen, azaz 9-szer nagyobb a negatívénál, akkor az SVM 100%-os precizitásra képes a 15 perces pozitív címkék kapcsán, illetve 92%-os precizitásra 30 perces esetben (Groth & Muntermann 2010). Ez azonban azzal jár, hogy a felidézés nagyon alacsony, 10% körüli, és a pontosság így körülbelül 77%, amit a default modell 75%-ához kell mérni. A tévesztési súlyarányok közelítésével a precizitás drasztikusan leesik, míg a felidézés javulást mutat (lásd *1. táblá-*

zat). A megfelelő költségarány kiválasztásában segít, ha az előrejelzés alapján kereskedési szimulációkat készítünk, amelyek átlagos hozama alapján lehet döntést hozni. A pozitív kategória a nagy hozamvolatilitást jelenti, amit olyan instrumentummal lehet kihasználni, amely az árfolyam nagymértékű, de tetszőleges irányú megváltozása esetén ér el nyereséget. Ez az instrumentum a long straddle opció[39]. A hipotetikus opciópiacon 15, illetve 30 perc múlva lejáró opciókat lehet vásárolni a hír publikálása idején. A szimulált stratégia csak akkor vesz fel long straddle-t, ha az SVM előrejelzése pozitív volt. A szimuláció során az opciós prémiumok becslésére a Black–Scholes-modellt (Black & Scholes 1973) használták. Benchmarkként az egyszerű long straddle stratégiát alkalmazták, amely a hír tartalmától függetlenül minden hír publikálásakor long straddle pozíciót alakít ki. 15 perces időtávon a legmagasabb hozam 7,3% volt, és ekkor a pozitív kategória tévesztési költsége 9-szerese volt a negatívénak. Ebben az esetben volt a hozamszórás is a legalacsonyabb, 5,5%. 30 perces időtávon ugyan az átlagos hozamra magasabb értéket kaptak két esetben – 8,4% 1:3-as költségaránynál, illetve 7,7% 1:9-es költségaránynál –, ám a hozamszórás is jelentősen megnőtt – rendre 12,3%, illetve 6,5%. A benchmark átlagos hozama 15 perc esetén csupán 2%, 30 perc esetén 2,8%, hozamszórása pedig rendre 5,7%, illetve 6,9% volt, tehát nem kérdéses, hogy a modell alapján menedzselt stratégia hatékonyabb. Ha az SVM-modell helyett más osztályozó módszert – naiv Bayes, k-legközelebbi szomszéd és mesterséges neurális hálózat – alkalmaztak, eredményeik hasonlók lettek, bár a naiv Bayes hozama alulmaradt a többi osztályozóval és a benchmarkkal szemben (Groth & Muntermann 2010). A legtöbb esetben nincs szignifikáns különbség a különböző osztályozók teljesítményben, de amikor igen, akkor az SVM mutatkozott a legjobbnak, a neurális hálózat pedig a másodiknak.

A hírek hatására kialakuló likviditási sokkok tanulmányozása volt a kutatás utolsó lépése, amelyhez újabb változókra volt szükség. A korpusz Groth (2010) első ilyen tanulmányában 369 sajtóközleményből tevődött össze, amelyeket 2006.01.31. és 2008.09.09. között publikáltak, míg a második tanulmányban (Groth et al. 2014) 415 közleményből, amelyeket 2006.01.31. és 2009.07.22. között németül és angolul is publikáltak. A Xetra ajánlati könyv adatokat a Reuters Tick History szolgáltatáson keresztül érte el, és ez alapján az adott pillanatbeli likviditást a CRT – cost of round trip, retúrköltség – segítségével mérte. A CRT az a költség amely egy adott pillanatban egy megadott mennyiségű részvény egyidejű megvásárlása és eladása esetén merülne fel. Az ajánlati könyv eladói

39 A stratégia azonos eszközre, azonos jellemzőkkel rendelkező put és call opciók egyidejű vételére (erre utal a long szó) épül. A straddle szó arra utal, hogy nagy ármozgást várunk. (Zeller 2001, o. 110)

és vevői oldalán elhelyezkedő ajánlatok mennyisége és árai alapján számolható. Az említett tanulmányok célja a hírek után megfigyelhető abnormális likviditás előrejelzése, amely az abnormális CRT-vel mérhető. Az abnormális CRT-t az adott időpontbeli CRT és a tíznapos CRT-mozgóátlag különbségeként definiálták. A kisebb abnormális CRT nagyobb likviditást jelent. A tanítóminta összeállításakor az alsó kvartilisbe eső megfigyelések negatív címkét kaptak – ekkor magas a likviditás –, a többi megfigyelés pozitív címkét kapott. A mintában tehát háromszor annyi pozitív címke van, mint negatív, ezért az osztályozás precizitásának növelése érdekében a negatív osztályhoz magasabb tévesztési költséget kell rendelni. A tévesztési költségek többféle összeállítását megvizsgálták SVM-osztályozóval, 10-szeres keresztvalidáció mellett.

Ha a negatív kategória tévesztési költsége jelentősen nagyobb a pozitívénál, tehát 9-szerese, akkor a negatív kategória precizitása 100%, azonban a felidézés csak 14,1%, és a 78,4%-os pontosság ekkor volt a második legmagasabb (Groth 2010). A költségarányt 1:3-ra csökkentve a precizitás 79%, a felidézés 21,2%, a pontosság pedig 78,7% lett. Az arány további csökkentésével a pontosság csökkent a precizitással együtt, míg a felidézés növekedett. Az eredmények pénzügyi szempontú kiértékelésre egy olyan szimulációt is végrehajtott, amelynek célja, hogy adott mennyiségű értékpapírt a hírt követő 15 percen belül a lehető legkisebb költséggel megvegye vagy eladja. Az SVM előrejelzésére épített lineáris stratégia lényege, hogy a részvényállományt egyenlő részekben, fokozatosan adják-veszik a hír utáni meghatározott időtartamon belül, ha az előrejelzés szerint negatív abnormális CRT várható, azaz normálisnál alacsonyabb költséggel lehet kereskedni a publikálást követően. Ha az előrejelzés pozitív, akkor a publikáláskor azonnal megtörténik az ügylet. Benchmarkként a naiv stratégia szolgált, melynek értemében a hír publikálásakor azonnal végrehajtják a teljes részvényállomány adásvételét, a hír tartalmától függetlenül. A szimuláció eredménye szerint a negatív kategóriába sorolt megfigyeléseknél a lineáris stratégiával kisebb volt a tranzakciós költség (Groth 2010).

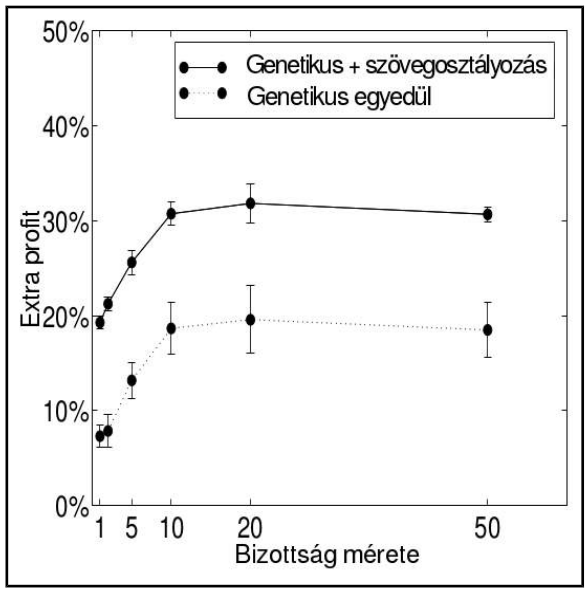
Később a fenti vizsgálatot kibővítette úgy, hogy az angol nyelvű közleményekkel is megvizsgálta ugyanazon hírek hatását, valamint különböző reprezentációkat – mint például kifejezések, szó és karakter n-grammok – alkalmazott, és a teljes szöveg helyett kizárólag a közlemények címét használta (Groth et al. 2014). A teljes szövegű hírekben a szó n-grammok, illetve a címek esetén a karakter n-grammok fontos jellemzőknek bizonyultak a khi-négyzet statisztika alapján. Angolul a teljes szövegű közlemények kevesebb kifejezéssel reprezentálhatók, mint németül, de a német nyelvű reprezentáció mind

pontosságban, mind precizitásban és felidézésben jobb értékeket produkált. Kizárólag a címeket használva inkább az angol a megfelelőbb. Az új mintán is megerősítést nyert, hogy a negatív kategóriába sorolt közlemények esetében a naiv stratégia drágább, míg pozitívak esetén olcsóbb, mint a lineáris stratégia (Groth et al. 2014). Magas 1:9-es tévesztési költségaránnyal a teljes szövegű hírek esetén kevésbé kimutatható ez a tendencia mindkét nyelven, illetve a kizárólag a közlemények címéből álló korpuszon inkább csak angol nyelven. Az eredmények alapján úgy tűnik, hogy 1:3 tévesztési költség mellett a gyakorlatban jól használható modell kapható.

2.2. További tanulmányok a tőzsdei hírbányászatról

2.2.1. Thomas

James D Thomas, a Carnegie Mellon University-ről, azt vizsgálta, hogy a Raging Bull[40] internetes tőzsdei fórumon található hozzászólások segítségével előrejelezhetők-e a részvényárfolyamok (Thomas & Sycara 2000). A fórumok közül azokat elemezte, amelyek csak egy meghatározott részvénnyel foglalkoztak. A negyven legnépszerűbb fórum közül huszonkettő fórum felelt meg ennek a kritériumnak, illetve annak a feltételnek, hogy a hozzá kapcsolódó részvény árfolyama meghaladja az egy dollárt.



4. Ábra: Az extraprofit mértéke a szöveges és szöveg nélküli információkkal generált kereskedésiszabály-bizottságok mérete alapján

Forrás: (Thomas & Sycara 2000, o.4)

40 www.ragingbull.com

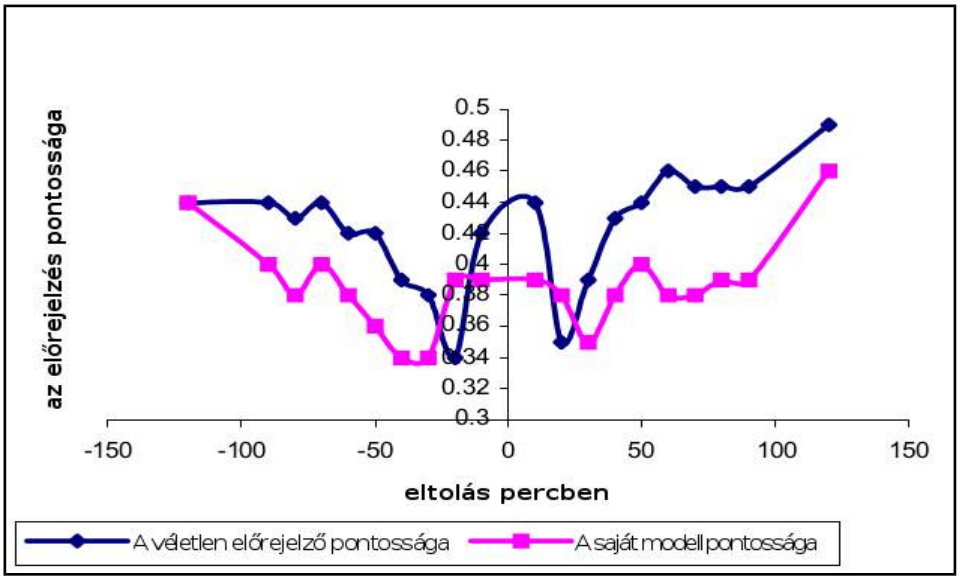
1999.01.01. és 1999.12.31. között legyűjtötte az összes hozzászólást ezekről a fórumokról, és a napi záróárfolyamokat a Yahoo! Finance weboldalról. Az árfolyamok reprezentációs egységeként a napi hozamok előjelét választotta, és két hozamkategóriába sorolta a megfigyeléseket: fel és le. A fórumokról gyűjtött hozzászólásokat naponként csoportosította úgy, hogy az előző napi zárás időpontjától a tárgynapi zárás időpontjáig elküldött hozzászólások azonos csoportba kerültek. A csoport együttes szókészletét a vektortérmodell szerint reprezentálta. A hozamkategória előrejelzésére maximum entrópia osztályozót alkalmazott, amely a t-dik napi hozzászólások alapján becsülte meg, hogy a $t+1$-dik napi hozam milyen valószínűséggel lesz pozitív. A tesztminta 200 elemű volt, a tanítóminta kezdetben 52 napnyi megfigyelést tartalmazott, majd minden periódus megbecslése után kiegészült az adott periódushoz tartozó megfigyeléssel. Az eredmények értékelésére az extra profit mértékét használta, amelyet minden periódusban az előrejelzéstől függően részvényben vagy készpénzben tartott tőkével el lehetett érni a *buy-and-hold* stratégiával szemben. A huszonkét részvényből álló mintán az átlagos extra profit −8,11%, míg a tízezer posztnál többel rendelkező részvényekre 6,91% volt. Az így nyert szöveges előrejelzés alapján kereskedési szabályokat állított elő genetikus algoritmussal. Ahogy a *4. ábrán* látható, megfelelő számú bizottság esetén a szöveges információt figyelmen kívül hagyó kereskedési szabályok átlagos extra profitja alacsonyabb a szövegek tartalmát is használóknál.

2.2.2. Gidófalvi

Gidófalvi Győző, a University of California-ról, a Lavrenko-tól kapott mintán, illetve saját adatokon is vizsgálata, hogy különböző időablakok esetén kimutatható-e szignifikánsan a hírek hatása a részvényárfolyamokra. Gidófalvi (2001) a Lavrenko-féle adatok közül csak tizenkettő NASDAQ-on jegyzett céghez tartozó tízperces felbontású árfolyamidősort és a publikálás időpontja szerinti időbélyegzővel ellátott híreket elemezte. Gidófalvi és Elkan (2003) tanulmányában saját adatokkal dolgoztak, amelyek egyperces felbontásúak voltak, továbbá a DJIA-t és az indexben szereplő 30 legnagyobb vállalat részvényárfolyamát tartalmazta a 2001.07.26. és 2002.03.16. közötti minta, és a részvényekhez tartozó hírek ugyanebből az időszakból valók. Mindkét mintából kiszűrték azokat a híreket, amelyek árfolyamra gyakorolt hatását nem lehetett megfigyelni, mivel nem állt rendelkezésre árfolyamadat. Ennek oka lehetett, hogy nem történt jegyzés az adott részvényre, vagy hétvégén, illetve ünnepnapokon jelent meg a cikk, amikor zárva volt a tőzsde, továbbá a piac nyitásától és zárásától mért idő és a vizsgált időablakok vi-

szonya nem tette lehetővé az árfolyamváltozás kiszámítását. Ez a fajta szűrés nagyjából felére-harmadára csökkentette a kezdetben rendelkezésre álló hírek számát.

A szövegek reprezentálására a vektortérmodellt alkalmazta, a szavakat szótövezte, eltávolította közülük a stopszavakat, és jellemzőkiválasztást is alkalmazott rajtuk, azaz az 1000 legnagyobb mutual information (kölcsönös információtartalom) értékkel rendelkező szót használta az elemzésben. Az árfolyamok reprezentálásához az abnormális hozamok előjele alapján megállapított kategóriákat használta. A részvény bétával korrigált normális hozamának a részvényindex hozamát tekintette. A különböző abnormális hozamok összehasonlíthatósága érdekében a részvény adott perióduson belül megfigyelt hozamát osztotta a részvény bétájával, majd ebből vonta ki a részvényindex azonos periódusban megfigyelt hozamát. Ezeket az abnormális hozamokat három kategóriába – fel, le, normális – sorolta be, aszerint, hogy a le és fel kategóriához tartozó küszöbszintek által meghatározott intervallumok közül melyikbe estek. A küszöbszintek az első tanulmányban (Gidófalvi 2001) ±0,002 értéket vettek fel, a másodikban (Gidófalvi & Elkan 2003) viszont minden kísérletben olyan értékeket választottak, hogy a három korábban említett kategória megoszlása közelítőleg rendre 25%-25%-50% legyen. A mintát kísérletenként különböző méretű tanító- és tesztmintára osztották. A hírek és az abnormális hozamok egymáshoz rendelését változó kísérleti összeállításokban vizsgálták. Minden kísérlethez definiáltak egy időablakot, amely a hírhez képest relatív kezdő- és végidő-



5. Ábra: Az időablak eltolásának hatása a véletlenszerű, illetve a hírek alapján történő előrejelzés pontosságára

Forrás: (Gidófalvi 2001)

ponttal rendelkezett. Ha a kezdőidőpont a hír megjelenése előtti, akkor a végidőpont a hír megjelenésével esett egybe, ha a végidőpont a hír utáni, akkor a kezdőidőpont esett egybe a hír megjelenésével. Az időablakok legfeljebb néhány óra hosszúak voltak minden esetben. A hírhez ezután hozzárendelte az időablak alatti abnormális hozam alapján megállapított kategóriát. Az osztályozásra naiv Bayes-módszert alkalmazott, a megvalósításhoz a Rainbow csomagot használta. Az előrejelzés minőségét a pontosság, és az egy kereskedésre jutó átlagos profit alapján értékelte. A kategorizálás pontossága egyik kísérletben sem érte az 50%-ot, az átlagos profit pedig a [−20;0] és a [0;20] időablakok esetén mutatott a véletlenhez képest szignifikánsan nagyobb értéket (*5. ábra*).

2.2.3. Koppel

Moshe Koppel, az izraeli Bar-Ilan University-ről, egy rövid tanulmányában a Multex Significant Developments korpusz összesen 12000 hírét felhasználva elemezte az S&P500 index részvényeinek árfolyamát a 2000–2002-es időszakban[41] (Koppel & Shtrimberg 2004). A hírek szövegét a vektortérmodell szerint reprezentálták, bináris súlyozást alkalmazva a szavakra, amelyek közül kiszűrték azokat, amelyek gyakorisága a korpuszban nem érte el a 60-at, és stopszavazást is végrehajtottak a szótáron. A szavak information gain (információnyereség) alapú jellemzőkiválasztáson is átestek, mely után a 100 legmagasabb értékkel rendelkező szót tartották meg az elemzéshez. Az árfolyamok reprezentálásakor a különböző kísérletekben eltérő időszakokra számított hozamokat alkalmaztak. A hozamokat előjelük szerint pozitív és negatív kategóriába sorolták. A pozitív kategóriához tartozó alsó korlát 10%, míg a negatív kategóriához tartozó felső korlát −7,8% volt, hogy a két kategóriába közel azonos számú megfigyelés essen. A korábbi tanulmányokhoz képest jelentősen nagyobb abszolút értékű küszöbök használatát azzal indokolták, hogy így kisebb a valószínűsége, hogy a hírtől független árfolyammozgások alapján kapnak címkét a hírek[42]. Kétféle kísérletet végeztek a szerzők aszerint, hogy a hírekhez hogyan rendelték hozzá az árfolyammozgásokat. Az első változatban egyidejű hozzárendelést alkalmaztak, azaz a hír publikálását megelőző nap záróárfolyama és a publikálást követő nap nyitóárfolyama közötti hozam kategóriáját rendelték a hírhez. A második változatban a publikálást követő nap nyitóárfolyama és az azt követő nap nyitóárfolyama közötti hozamkategóriát rendelték a hírhez. A mintát két-

41 2003-as adatokat is használt, de csak az osztályozó teszteléséhez.

42 Ezen kívül meg kell említeni, hogy a szerzők nem napon belüli árfolyammozgásokat vizsgáltak, hanem napi változásokat, azaz a napon belüli hozamokhoz képest jellemzően nagyobb napi hozamokra vonatkozik ez a nagyobb korlát.

féle változatban osztották tanító- és teszthalmazra. Az első változatban tízszeres keresztvalidációt alkalmaztak, a második változatban a 2000–2002-es adatok jelentették a tanítóhalmazt, a 2003-as adatok a teszthalmazt.

Az osztályozási feladathoz lineáris SVM-modellt alkalmaztak, melynek teljesítményét más módszerekkel – naiv Bayes, döntési fa – is összehasonlították. Az egyidejű hozzárendelés során tízszeres keresztvalidáció esetén az SVM átlagos pontossága 70,3% volt, míg a 2003-as teszthalmazra 65,9%. A többi módszer hasonló eredményt produkált. Megfigyelése szerint a pozitív kategória felidézése 83,3%, precizitása viszont csak 66%, ellenben a negatív hírek 77,5%-os precizitásával. A jelenséget azzal magyarázták, hogy a negatív híreknek diszkriminatív szókészlete van, amely elősegíti az elkülönítésüket, míg a pozitív híreknek nincs. Az előrejelzési típusú hozzárendelés és tízszeres keresztvalidáció során 52% pontosságot ért el a modell, ami az EMH-val konzisztens eredménynek tekinthető.

A kutatás folytatása csak 2008-ban, illetve 2011-ben következett (Généreux et al. 2008; 2011), amelyben a korábbi adatok egy részhalmazán, 6277 hírből és 464 S&P500 részvényből álló mintán végezték a kísérleteket. Ötféle szövegreprezentációt hasonlítottak össze az előrejelzés pontosságához való hozzájárulás szempontjából. A reprezentációk tartalmazták a tagadott szóalakokat is külön dimenzióként. Az első típusú reprezentáció a szótövezetlen szavakat tartalmazta, melyek legalább háromszor előfordultak a korpuszban. A második típusú reprezentáció ezek szótövezett változatait használta. A harmadik reprezentáció 420 pénzügyi kifejezés alapján jellemezte a hírek szövegét. A negyedik 123 egészségügyi metafora – pl. függőség, krónikus, felépülés –, az ötödik pedig megszemélyesítő metaforák – pl. felmászott, ugrott, száguldott, küzdött – előrejelző képességét tesztelte. A kísérletek során az unigram reprezentáció, tehát az első bizonyult a legpontosabbnak. Az unigram jellemzők kiválasztása során a dokumentumgyakoriság, az information gain és a khi-négyzet mutatók alapján kiválasztott 100 kifejezés pontosságra gyakorolt hatását is vizsgálták. A három mutató közül az information gain és a khi-négyzet hasonló eredményt mutatott, a dokumentumgyakoriságnál jobbat. A szerzők végül az information gain alapján történő dimenziócsökkentést alkalmazták. A súlyozási sémák közül a bináris és a szógyakoriság alapú súlyozással is hasonló pontosságot értek el, ezért a továbbiakban a bináris súlyozási sémát használták. Az árfolyamok reprezentálásában a (Koppel & Shtrimberg 2004) tanulmányhoz képest annyi változás történt, hogy a sztenderd küszöbértékeket $\pm 4\%$ -ban állapították meg. Azokban a kísérletek-

ben, amikor nem a különböző küszöbértékek pontosságra gyakorolt hatását vizsgálták, ezt az értéket használták. A küszöbértékek abszolút értékének $\pm 6\,\%$-ig való növelésekor a pontosság növekedett, majd 75% környékén stagnált. Az eddigi eredmények két kategóriára és egyidejű hozzárendelésre vonatkoztak, viszont, ha a kategóriák közé felvették a semleges kategóriát is, akkor a pontosság 50% körüli szintre esett vissza. A prediktív jellegű hozzárendelés esetén annál rosszabb az előrejelzés pontossága, minél távolabb van a hírhez rendelt periódus vége a hír publikálásának napjához képest. A lehető legrövidebb, két nap hosszú időszak esetén a pontosság közel 70% volt. A 2004-es eredményekkel ellentétben, nem csak a negatív kategóriához tartoztak diszkriminatív szavak, az új vizsgálatok szerint nagyjából azonos mennyiségben vannak jelen ilyen kifejezések a pozitív kategóriában is.

2.2.4. Mittermayer

Marc-André Mittermayer, az Universität Bern kutatója, 2004–2007 között vizsgálta a szövegbányászati technikák alkalmazhatóságát a tőzsdei árfolyamok előrejelzésére. Mittermayer és Knolmayer (2006b; 2007) az addig publikált tőzsdei hírbányászattal kapcsolatos modellekről review-t is készített, amelyben szerepeltek Wüthrich és tanítványainak modelljei, Lavrenko, Thomas és Gidófalvi munkái, de már a 2004-es HICSS-37[43] konferencián saját fejlesztésű, NewsCATS[44] névre keresztelt modellel állt elő (Mittermayer 2004). A NewsCATS-szel végzett kísérletekhez kezdetben használt nagyfrekvenciás, tranzakciószintű árfolyamadatai a 2002.01.01. és 2002.12.31. közötti időszakból származtak. Ugyanebből az időszakból letöltötte a PRNewswire-ön elérhető sajtóközleményeket. Ezek a közlemények olyan metaadatokkal is rendelkeznek, mint a benne foglalt vállalatok ticker szimbólumai, a tőzsde, amin jegyzik őket, és a kategória, amelybe a közlemény tartozik. Az elemzésből kizárta azokat a sajtóközleményeket, amelyeknél nem adták meg e metaadatokat, vagy több vállalatot is érintett a közlemény. Ezen kívül ötmillió dolláros alsó korlátot szabott a közleményben érintett vállalat részvényének tőzsdei forgalmára is. Az előrejelzés szempontjából nem állna rendelkezésre az árfolyam értéke, ha a NYSE-en jegyzett cégeknél 9:30 előtti vagy 15:00 utáni, valamint a NASDAQ-on 8:00 előtti és 17:00 utáni hírek is szerepelnének a mintában, ezért ezeket is kizárta.[45] A szerző azért választotta a PRNewswire-t hírforrásul, mert azt feltételezte, hogy a sajtóközlemények képezik a váratlan, de releváns tőzsdei információk elsődleges

43 Hawaii International Conference on System Sciences
44 News Categorization and Trading System
45 Minden idő Eastern Time szerint értendő.

forrását, szemben a szerkesztett hírekkel, amelyek például a Reuters-től származnak. Az imént említett szűrőfeltételek alkalmazása után 6602 közlemény maradt a mintában.

A híreket a szózsákmodellnek megfelelően reprezentálta, a szótövezett szavak közül kiszűrte a stopszavakat, majd kiválasztotta az ezer legnagyobb tf-idf értékkel rendelkezőt közülük. Az ezer jellemző alapján normalizált bináris súlyozással képezte a szó-dokumentum mátrixot (TDM, term-document matrix). A hír utáni 60 percben bekövetkezett árfolyamváltozás mértéke alapján minden hír kapott egy címkét a következők közül: jó, rossz, hatástalan. A jó hírek utáni órában az árfolyam valamikor legalább 3%-kal a hír publikálásának idején érvényes árfolyam fölé emelkedett, de átlagosan 1%-kal meghaladta azt ez idő alatt. A rossz hír utáni órában legalább 3%-os csökkenés, de átlagban 1%-os volt. A többi hír a hatástalan címkét kapta. Az így összeállt minta 347 jó hírt, 357 rossz hírt és 5898 hatástalan hírt tartalmazott. A tanítóminta minden kategóriából 200 megfigyelést tartalmazott, a maradék pedig a tesztminta részét képezte.

A NewsCATS az SVM light csomagot használta osztályozásra. 50 modellt becsült, majd ezek átlagos eredményét vette. 54–60% közötti pontosságot ért el, de a jó és rossz hírek precizitása nagyon gyenge, 4–7%-os volt. E kategóriák felidézése is sokkal gyengébbnek bizonyult a hatástalan kategóriához képest: 37–54% az 54–61%-kal szemben. A szerző szerint a gyenge eredmények oka valószínűleg a pontatlan címkézés lehet. Mittermayer a rendszer teljesítményét a piaci szimuláció révén kapott hozammal is jellemezte, amely átlagosan 0,11% volt ügyletenként, szemben a véletlenszerű kereskedés 0%-os hozamával.

A korábbi változathoz képest pontosított és javított kísérleti összeállítást Mittermayer disszertációjában (Mittermayer 2006) és az ICDM06[46] konferenciára készített tanulmányában (Mittermayer & Knolmayer 2006a) olvashatjuk. A minta ebben az összeállításban a 2002.04.01. és 2002.12.31. közötti időszak híreiből és árfolyamaiból tevődött össze. A hírforrás nem változott, azonban tisztázódott, hogy a kísérletben csak az S&P500-ban szereplő cégek adatait használta. Ezen kívül csak az árfolyamra nagy hatást gyakorló közleménykategóriákba tartozó híreket vette figyelembe. Ez összesen 989 sajtóközleményt jelentett. A hírek reprezentálására a szózsákmodell mellett egy ember által összeállított tezauruszt is alkalmazott, amely szavakat, kifejezéseket, és szósorozatokat tartalmazott. A jellemzőkiválasztás során a legnagyobb idf értékkel rendelkező

46 International Conference on Data Mining

szavak, illetve tezauruszelemek közül választotta ki az első 15%-ot. A TDM súlyozására a wdf-idf[47] sémát alkalmazta.

Az árfolyamok reprezentálásakor előbb két perc hosszúságú mozgóátlagot számolt, amelyet 15 másodpercenként léptetett úgy, hogy a hír után egy perctől három percig tartott az első átlagolási periódus, az utolsó pedig a hír után 13 perctől 15 percig. Viszonyítási alapként a hír publikálása előtt egy perc és utána egy perc közötti időszakra is kiszámolta az átlagot. A publikálás idejéhez képest jövőbeli átlagok közül a legnagyobb és a legkisebb értéket használta fel. A legnagyobb vagy legkisebb mozgóátlag-értéknek és a publikálással egy időben megfigyelt átlagértéknek a hányadosának vette a logaritmusát, és az így kapott loghozamok – melyekre a CCR_{max} , illetve CCR_{min} jelölést használták – segítségével definiálta a jó, rossz és semleges híreket, illetve a hírek címkézéséhez használt kategóriarendszer kibővült a bizonytalan kategóriával is:

$$
\begin{aligned}
&Ha \left\{ \begin{array}{c} CCR_{max} > 3\% \;\; és \;\; |CCR_{min}| < 3\% \\ vagy \\ 3\% < |CCR_{min}| \;\; és \;\; 2 \cdot |CCR_{min}| < CCR_{max} \end{array} \right\} \; akkor\ JÓ\ HÍR \\
&Ha \left\{ \begin{array}{c} CCR_{max} < 3\% \;\; és \;\; |CCR_{min}| > 3\% \\ vagy \\ 3\% < CCR_{max} \;\; és \;\; 2 \cdot CCR_{max} < |CCR_{min}| \end{array} \right\} \; akkor\ ROSSZ\ HÍR \\
&Ha \;\; CCR_{max} < 3\% \;\; és \;\; |CCRmin| < 3\% \;\; akkor\ SEMLEGES\ HÍR \\
&Különben\ BIZONYTALAN\ HÍR
\end{aligned}
\tag{7}
$$

A fenti definíciónak megfelelően a mintában 83 jó, 42 rossz, 504 semleges és 360 bizonytalan hír szerepelt. Ebből a tanítóminta összeállításakor kimaradtak a bizonytalan hírek, és mivel tízszeres keresztvalidációt alkalmazott a szerző, a többi kategóriának 90%-át használta minden esetben a tanításhoz. A tesztelésben a kimaradó 10%-ok, illetve a bizonytalan hírek szerepeltek. A modell 82%-os pontosságot ért el a kísérlet során. A piaci szimuláció ügyletenként 0,22%-os átlagos hozamot eredményezett. Az eredmények robusztusságát úgy ellenőrizték, hogy a kiinduló modellparaméterek közül egyet-egyet megváltoztatott minden újabb kísérleti összeállításban. A jellemzőkiválasztási mértékek közül a gyűjteménytámogatottság – collection term frequency – bizonyult a legjobbnak, bár az eredmények nem változtak lényegesen, 83% volt a pontosság, és 0,28% az átlagos hozam. A jellemzőkiválasztás során alkalmazott küszöbérték változása nem befolyásolta jelentősen az eredményeket, a 15% jónak bizonyult. A jellemzők szó-dokumentum mátrixon belüli súlyozására a wdf-idf bizonyult a legjobbnak. A legna-

47 wdf: within-document frequency, dokumentumon belüli gyakoriság

gyobb pontosságra, 83%-ra a nemlineáris kernelt használó SVM volt képes, amely egyben a legmagasabb átlagos hozamot is eredményezte.

2.2.5. e-Markets Group[48]

Az e-Markets Group egy ausztrál kutatócsoport neve, amelyet 2001-ben alapított Simeon Simoff, John Debenham és Ian Wilkinson. A csoport üzleti-adatbányászati fókuszú kutatásainak egyik ága a hírek árfolyamra gyakorolt hatásával foglalkozott. Bizonyos, a témához közvetve kötődő előzmények után 2005-ben mutatták be Sydney-ben a devizaárfolyamok modellezését szöveges információkkal támogató módszerüket (Zhang et al. 2005), illetve a Lavrenko-féle kísérleti összeállításhoz hasonló saját összeállításukat (Yu et al. 2005). Mindkét változatnak volt folytatása 2006-ban észak-, illetve délamerikai konferenciákon is (Zhang et al. 2006; Yu et al. 2006). 2007-ben egy könyvfejezet formájában összegezték a devizákhoz kapcsolódó kutatásaikat (Zhang et al. 2007).

A devizaárfolyam-modellezés során az euró és a dollár árfolyamát vizsgálták 2005.02.07. és 2005.07.04. között (Zhang et al. 2005). Mind a hírek, mind az árfolyamok a Bloombergtől származtak. A híreket manuálisan két részre osztották, célkorpuszra és általános korpuszra. A célkorpuszba az euró-dollárhoz kapcsolódó releváns hírek kerültek, az általános korpuszba pedig az irrelevánsak. A két korpusz közötti szóhasználati különbségeket használták fel arra, hogy a szövegreprezentációt szakterület specifikus kifejezésekkel bővítsék. A vektortéralapú szövegreprezentáció tokenekből és gyakori kifejezésekből állt. A tokenek és kifejezések szótövezésen és stopszavazáson estek át.

A hírek pozitív-negatív kategóriába sorolása három fázisban történt. Először az alapján osztályozták a híreket, hogy a dokumentum szókészlete a célkorpuszéhoz, vagy az általános korpuszéhoz van-e közelebb. Ha a célkorpuszhoz, akkor a szövegben említett országok, devizák stb. alapján eldöntötték, hogy a hírnek van-e köze az euró-dollár árfolyamhoz. Végül, ha az EURUSD szempontjából releváns hírről volt szó, akkor a hír gazdasági hatása szempontjából került besorolásra a pozitív vagy a negatív hírek közé – ehhez diszkriminancia analízist és k-közép osztályozást alkalmaztak. A hírek devizaárfolyamra gyakorolt hatását egy regressziós modellben vizsgálták, amelyben a hírek pozitivitása magyarázó változóként szerepelt. A modell eredményméréséről nem számoltak be részletesen.

A devizaárfolyamokkal foglalkozó újabb cikkekben (Zhang et al. 2006; Zhang et al. 2007) némileg módosult a koncepció, és a híreket jó, rossz, hatástalan és egyéb kategó-

48 http://research.it.uts.edu.au/emarkets/

riába sorolták be két lépésben. Először relevancia szerint osztályoztak, majd a releváns híreket sorolták be a fenti kategóriákba. A tanítóminta összeállítása sem manuálisan történt, hanem a címkéket aszerint kapták a megfigyelések – azaz a hírek –, hogy a publikálásukat megelőző és követő 12-12 órában volt-e legalább 0,1%-os árfolyamváltozás. Ha nem, akkor hatástalan, ha volt és pozitív volt, akkor jó, ha volt és negatív volt, akkor rossz címkét kapott a hír. Összesen 2589 megfigyelésből 1885 irreleváns, 200 jó, 113 rossz és 230 hatástalan, 161 egyéb hír volt. Érdemi eredményekről nem számoltak be ezekhez a kísérleti összeállításokhoz kapcsolódóan sem.

A Lavrenko-féle kísérleti összeállításhoz hasonló tanulmányaikban (Yu et al. 2005; Yu et al. 2006) több mint 2000 vállalati bejelentést, illetve az érintett AMP vállalat részvényárfolyamát gyűjtöttek össze 1998.06.15. és 2005.03.16. között. A szövegek reprezentálására a vektortérmodellt alkalmazták, amelyben a jellemzők halmaza 36 szótövezett tokenből, illetve szóbigrammból állt. A TDM-mátrix súlyozásához a tf-idf sémát választották. A részvényárfolyamokból Lavrenko-hoz hasonlóan trendidősort képeztek szakaszonkénti lineáris regresszió segítségével. A hírek fel, le vagy semleges címkét kaptak a tanítóhalmazban attól függően, hogy a kapcsolódó részvény hozama és a részvényindex hozama között mekkora és milyen eltérés mutatkozott. 464 fel, 833 le és 997 semleges címkéjű megfigyeléssel tanították SVM modelljüket, melynek megvalósításához a LibSVM csomagot használták. Az eredményeiket rendkívül szűkszavúan mutatták be, az általuk kapott 65,73%-os pontosságot is csupán a Wüthrich-féle 46%-hoz viszonyították.

3. A tőzsdei hírbányászat modellje

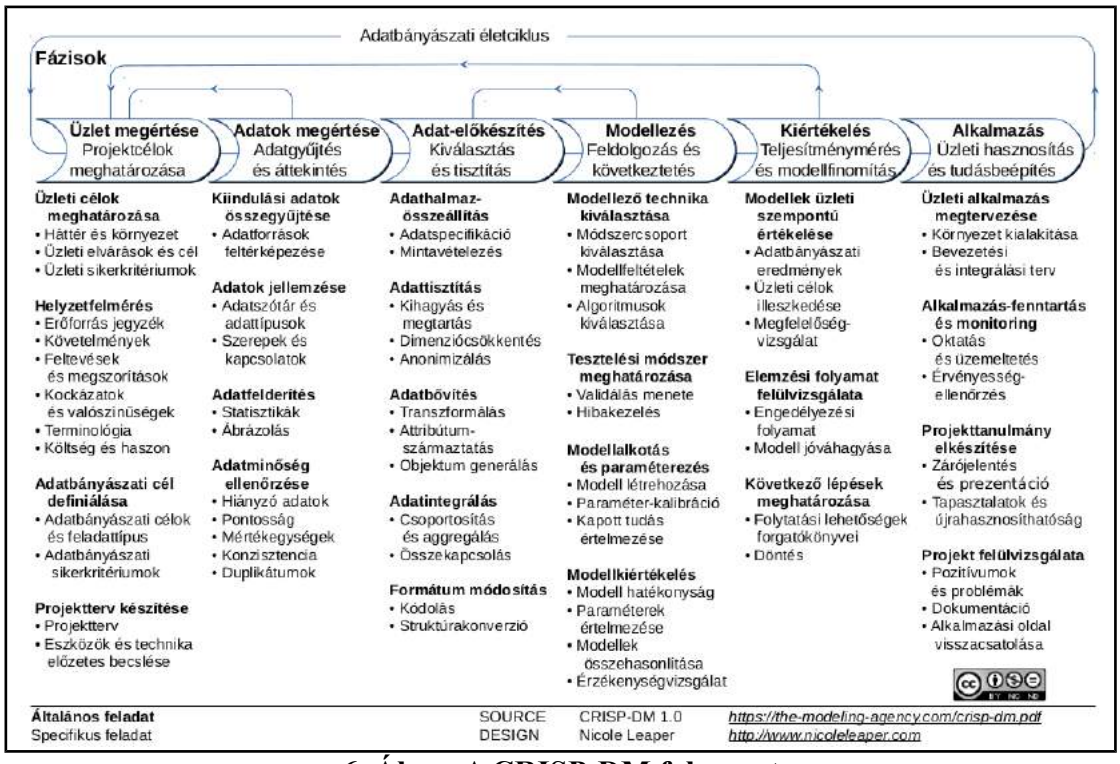
Az adatbányászat a tudásfeltárás eszközeként részben a statisztika és a gépi tanulás módszereire építkezik egy-egy adatelemzési probléma hatékony megoldásának érdekében. Az adatbányászat definícióiban hangsúlyos elem a *nagy adathalmaz* és a *hasznos információ:*

> „*Az adatbányászat az a folyamat, amellyel hasznos információ fedezhető fel automatikus módon nagy adattárakban.*" (Tan et al. 2011)

> „*Az adatbányászat egy olyan döntéstámogatást szolgáló folyamat, mely érvényes, hasznos, és előzőleg nem ismert, tömör információt tár fel nagy adathalmazból.*" (Abonyi 2006, o.10)

Az adathalmaz mérete nem értelmezhető önmagában, mindig a probléma határozza meg, hogy hol van a határ. Amikor az adathalmaz mérete akkora, hogy más adatelemzési módszerekkel az információ már nem nyerhető ki a kívánt időn belül, akkor adatbányászatra lehet szükség, de egy határon túl az is lehetséges, hogy már a hagyományos adatbányászati eszközök sem képesek erre, és a *big data* felé kell fordulni. Ami a hasznosságot illeti, az üzleti-gazdasági területen az információ hasznosságát általában kifejezhetjük pénzben is, de más területeken, mint például az orvostudomány, ez nem feltétlenül tehető meg. A hasznosság tehát szintén viszonylagos fogalom, lehet, hogy egy tudományterületen hasznos információt adó módszernek egy másik területen nincs hasznos alkalmazása. A definícióban fontos elem továbbá az adatbányászat folyamat jellege is. Az adatbányászati folyamat jól leírható fázisokból áll, amelyek a Cross Industry Standard Process for Data Mining (CRISP-DM) sztenderddel, az adatbányászati folyamat egy általános modelljével, leírhatók. A CRISP-DM kialakulása az 1990-es évek végére tehető, amelyben nagy szerepe volt az NCR Systems Engineering Copenhagen, a DaimlerChrysler AG, az SPSS Inc., és az OHRA Verzekeringen en Bank Groep B.V. által alkotott CRISP-DM konzorciumnak. 1999–2000 során megjelent a CRISP-DM 1.0-ás változata, a 2.0-ás változat fejlesztése jelenleg is folyamatban van. Az 1.0-ás szabvány részletes bemutatása megtalálható a CRISP-DM konzorcium által közzétett felhasználói kézikönyvben (Chapman et al. 2000). A CRISP-DM fő lépései a következők: üzlet megértése, adatok megértése, adatok előkészítése, modellezés, kiértékelés, alkalmazás. Az egyes fázisok logikailag a fenti sorrendben követik egymást, azonban több-

szörös visszacsatolások lehetnek szükségesek az elemzés során, ezt szemlélteti a *6. ábra*. A továbbiakban röviden kifejtem az egyes lépések lényegét, majd a fejezet további részében a tőzsdei hírbányászat módszertanát mutatom be a CRISP-DM alapján.



6. Ábra: A CRISP-DM folyamat

Forrás: (Leaper 2009 alapján fordította; Kruzslicz 2015)

Az első lépés az üzlet megértése, amely egyáltalán nem triviális, olyannyira, hogy az adatbányászati folyamat során többször is vissza kell térni ehhez a fázishoz. Meg kell ismerni az üzleti folyamatokat, a hozzájuk kapcsolódó üzleti célokat és korlátozó feltételeket, fontosabb összefüggéseket. Ezekből levezethető egy vagy több adatbányászati cél, amelyek meghatározzák a további fázisok megvalósítását. A következő lépésben meg kell érteni az adatokat, még azelőtt, hogy bármit kezdenénk velük az adatbányászati célok érdekében. Meg kell vizsgálni, hogy hol keletkeznek adatok az üzleti folyamatban, és hogyan lehet hozzáférni ezekhez, továbbá milyen minőségűek. Át kell tekinteni, hogy milyen értékeket vesznek fel a változók, mik azok főbb jellemzői. Ezek olyan ismeretek, amelyek meghatározzák, hogy milyen feladatokat kell elvégezni az adat-előkészítés során, illetve milyen modellt kell választani az adatok elemzéséhez, illetve a kiértékeléskor milyen mérőszámok lehetnek relevánsak. Az adat-előkészítést úgy kell elvégezni, hogy közben tekintettel vagyunk a modellezési fázisban alkalmazott módszer követelményeire, például bizonyos módszerek képesek a hiányzó értékek kezelésére, mások

nem. A különböző modellek eltérő adattípusú inputadatokkal dolgozhatnak, illetve eltérő típusú outputokat szolgáltathatnak, ennek megfelelően kell kiválasztani, illetve átalakítani vagy származtatni az elemzésbe bevont attribútumokat. Az adat-előkészítéskor szem előtt kell tartani, hogy a kiértékelési fázisban hogyan fogjuk mérni a modell teljesítményét, robusztusságát, és ennek megfelelően kell összeállítani a tanító- és tesztmintákat. A modellezési fázisban meg kell tervezni, hogy milyen módszerekkel, algoritmusokkal érjük el az adatbányászati célt, melyek lesznek azok a modellparaméterek, amelyeket az elemzés során változtatunk, meg kell határozni a modellek közötti választás szempontjait, és ide tartozik magának a modellnek az elkészítése. A kiértékelés során meg kell vizsgálni, hogy a modell eredménye megfelel-e a kitűzött adatbányászati és üzleti céloknak, majd ennek függvényében dönteni kell, hogy a folyamat újraindítása, a folyamat sikeres vagy sikertelen lezárása, illetve az üzleti felhasználás-e a következő lépés. Az első esetben felül kell vizsgálni az üzleti és adatbányászati célokat, illetve bármely köztes fázist. Az üzleti hasznosítás során a modell alkalmazásának feltételeit kell megteremteni, illetve törekedni kell a hatékonyságának javítására, amely újabb iterációkat tesz szükségessé. (Chapman et al. 2000)

A következő alfejezetekben a CRISP-DM módszertan alapján leírom a *2. fejezet* szakirodalom-áttekintésének szintéziseként a tőzsdei hírbányászat fázisait, majd kiválasztom a saját modellemben alkalmazott megoldásokat, és azokat módszertanilag részletesen be is mutatom.

3.1. Tőzsdei hírek üzleti célokra

A tőzsdén forgalmazott részvények adásvétele mögött rejlő motivációk közül a legismertebb a nyereségszerzés – arbitrázs, spekuláció. További motivációk lehetnek például a szavazati jog szerzése, cégfelvásárlás, a részvénykibocsátás révén történő tőkeszerzés, árfolyamkockázat fedezése stb. Devizák esetén ez utóbbi mellett fizetési kötelezettségek teljesítése, hitelfelvétel, befektetés, beruházás finanszírozása, intervenció stb. is fontos tényezők. A különböző motivációkból végrehajtott tranzakciókra különböző időtávokra vonatkozó információk lehetnek hatással. Az arbitrázs szempontjából – természeténél fogva – csak a nagyon rövid időtávú, pillanatnyi információk relevánsak, míg egy beruházási céllal történő devizaváltás szempontjából az árfolyamokat hónapokon, éveken keresztül meghatározó információkra is tekintettel kell lenni. A különböző motivációkból aggregálódó kereslet és kínálat tehát különböző időtávra vonatkozó és különböző

forrásból származó hírekre reagál, így az árfolyamot befolyásoló összes hírt szinte lehetetlen számba venni.

2. Táblázat: A tőzsdei hírbányászati kutatások üzleti motivációi

Modell	Cikk	Instrumentum	Időtáv	Motiváció
Wüthrich	(Leung Kung Fan 1997) (Wüthrich, Cho, et al. 1998) (Wüthrich, Permunetilleke, et al. 1998) (Cho & Wüthrich 1999) (Cho et al. 1999)	tőzsdeindex (HSI, NKY, STI, FTSE, Dow)	kereskedési szakasz	spekuláció
	(Peramunetilleke 1997)	deviza (USD, DEM, JPY)	1–3 óra	spekuláció
Lavrenko	(Lavrenko et al. 2000a; 2000b)	részvény	nem rögzített, de 0–10 óra körüli	spekuláció
	(Fung et al. 2002a; 2002b) (Fung 2003) (Fung et al. 2003; 2005)	részvény (HKEX)	1–7 nap	spekuláció
Schumaker	(Schumaker & Chen 2006; 2008; 2009; 2010) (Schumaker 2009; 2010a; 2010b) (Schumaker et al. 2009; 2012)	részvény (S&P500)	20 perc	spekuláció
Groth	(Groth & Muntermann 2008; 2009)	részvény (FWB[49])	15 perc	spekuláció
	(Groth & Muntermann 2010; 2011)	részvény (FWB)	15 és 30 perc	spekuláció
	(Groth 2010) (Groth et al. 2014)	részvény (DAX, MDAX, SDAX)	15 perc	tranzakciós-költség-minimalizálás
Thomas	(Thomas & Sycara 2000)	részvény (NASDAQ, NYSE)	1 nap	spekuláció
Gidófalvi	(Gidófalvi 2001)	részvény (NASDAQ)	-120–120 perc	spekuláció
	(Gidófalvi & Elkan 2003)	részvény (DJI)	-30–30 perc	spekuláció
Koppel	(Koppel & Shtrimberg 2004)	részvény (S&P500)	0–1 nap	spekuláció
	(Généreux et al. 2008; 2011)	részvény	0–28 nap	spekuláció
Mittermayer	(Mittermayer 2004)	részvény (NASDAQ, NYSE)	1 óra	spekuláció
	(Mittermayer 2006) (Mittermayer & Knolmayer 2006a)	részvény (S&P500)	15 perc	spekuláció
e-Markets Group	(Zhang et al. 2005; 2006; 2007)	deviza (EURUSD)	0–1 nap	spekuláció
	(Yu et al. 2005; 2006)	részvény (egyetlen: AMP)	nem rögzített, több nap	spekuláció

Forrás: saját szerkesztés

49 Nem tértek ki arra, hogy a Frankfurt Stock Exchange (FWB) melyik piacáról valók a részvények.

A *2. táblázatban* a témában megjelent szakcikkek mögötti üzleti motivációk, a hozzájuk tartozó időtáv és instrumentumok láthatók. Látható, hogy tipikusan a részvényárfolyamokkal foglalkozik az irodalom, a modellek elsősorban árfolyam-spekulációs indíttatásúak, de az alkalmazott időtáv nagyon változatos, a néhány percestől a több naposig terjed.

Bármilyen motivációjú is legyen a kereskedési döntés, a szükséges információhoz hozzá kell jutni valamilyen forrásból, és külön iparág épült a tőzsdei hírek és adatok szolgáltatására. A tőzsdén forgalmazott részvények kibocsátóinak törvényi kötelezettsége, hogy az őket érintő, a részvények árfolyamát potenciálisan befolyásoló eseményekről azonnal értesítsék a befektetőket. Erre a célra sajtóközleményeket – press release – bocsátanak ki, amelyeket erre szakosodott weboldalakon tesznek közzé. Magyarországon a BÉT weboldala szolgálja ezt a célt, de a cégek saját weboldalára is gyakran felkerülnek a közlemények. Más országokban erre szakosodott, úgynevezett newswire-szolgáltatást nyújtó PR cégeken keresztül szokás megtenni ezt a bejelentést, mint például a Business Wire, a PR Newswire, vagy az EQS Group AG[50]. Ezen kívül természetesen egyéb hírportálokon is találhatunk cikkeket, azonban ezek általában sajtóközleményekből és sajtótájékoztatókból szerzik be az információikat, ezért csak késve tükrözik a piacon rendelkezésre álló információk halmazát. Fontos információforrást képeznek részvények kapcsán még a rendszeres üzleti beszámolók és jelentések is. Devizák vagy tőzsdei árucikkek esetén a releváns információk különböző kormányzati, illetve iparági szervezetek és hírügynökségek által kerülnek publikálásra. A kamatdöntésektől kezdve a nyersanyagárakon és háborúkon keresztül a természeti katasztrófákig terjedhet e hírek témáinak köre. Egy szóval a potenciális dokumentumok nagyon sokfélék lehetnek. A híraggregáló szolgáltatások e problémára jelentenek megoldást azáltal, hogy a különböző forrásokból származó híreket instrumentumonként külön oldalon gyűjtik. A legismertebb híraggregáló szolgáltatások közé tartozik a Yahoo! Finance, Google Finance, vagy az Investing.com. A nagy hírügynökségek, mint a Thomson Reuters vagy a Bloomberg is nyújtanak olyan szolgáltatást, amelynek keretében híreiket instrumentumonként csoportosítják.

A tranzakció végrehajtásához pedig szükség van az aktuális vételi és eladási ajánlatokkal kapcsolatos információkra, mint az árfolyamok, mennyiségek, jutalékok és a közvetítők kondíciói. Ezek a közvetítőkön – brókerek, dealerek, bankok –, a tőzsdéken,

50 Korábban Deutsche Gesellschaft für Ad-hoc-Publizität mbH (DGAP).

illetve a velük szerződésben álló adatszolgáltatókon keresztül érhetők el. Az azonnali adatszolgáltatás általában díj fizetéséhez kötött, de a késleltetett adatok általában ingyenesek. A múltbéli adatok között a nagyobb időszakok aggregált adatai általában ingyenesek, de a napon belüli, nagyfrekvenciás adatokat meg kell vásárolni. Ezek a szempontok azért lényegesek, mert míg bizonyos hírek hosszú távon befolyásolják az árfolyamot, mások percek alatt beárazódnak, így nem mindegy, hogy milyen részletességű és minőségű idősorban keressük a hírek hatását.

Látható tehát, hogy minél frissebb az információ, és minél hatékonyabban szeretne hozzáférni egy befektető, annál nagyobb díjat kell fizetnie az adat, információ vagy hír szolgáltatójának. Ezt a többletköltséget legalább azonos nagyságú többletbevétel és kockázati prémium érdekében vállalják a befektetők. Úgy gondolom, hogy ez az oka annak, hogy a *2. táblázat* tizenhét sorából tízben a hírek napon belüli hatását modellezték, mert ilyenkor az információ költsége magasabb, és ezért fontos, hogy minél jobban kihasználják azokat. A rövid távú motivációk közül az arbitrázsban a szöveges hírek kisebb szerepet játszanak[51], így érthető, hogy a spekuláció képezi a fő üzleti motivációt a tőzsdei hírbányászattal kapcsolatban. Az egyetlen kivétel, amelyben egy tetszőleges motivációjú tranzakció időzítése, így a tranzakciós költségek minimalizálása volt a cél, tulajdonképpen szintén a spekulatív jellegű, hiszen kockázatot hordoz[52]. A fentiek értelmében a saját modellemet a spekulációs motiváció jegyében napon belüli időtávra készítettem.

3.2. Az adatok

Ennek az alfejezetnek az első részében a kiindulási adatok összegyűjtéséről írok, általánosságban. Ezt követően összesítem és jellemzem a tőzsdei hírbányászatban használatos adatforrásokat, majd bemutatom a saját modellem kiindulási adatainak összegyűjtésével kapcsolatos munkát.

3.2.1. Árfolyam- és híradatok

Ahhoz, hogy modellezhessük a különböző pénzügyi eszközökkel kapcsolatos hírek és az árfolyamuk kapcsolatát, szükség van az eszköz árfolyamának idősorára, illetve a

51 Kivételként megemlíthető a felvásárlási arbitrázs – merger arbitrage –, amelynél a bejelentett felvásárlási ár és a tényleges piaci ár közötti különbséggel érhető el kockázatmentes nyereség.

52 Groth (2010, Groth et al. 2014) a tranzakciós költségeket a CRT mutatóval mérte, amelyet adott mennyiségű részvény egyidejű eladását és vételét feltételezve számolnak, negatív CRT esetén így akár arbitrázshoz is használható lenne a rendszere.

hozzá kapcsolódó hírek korpuszára. A modellezési céloknak megfelelően meg kell határozni, hogy milyen részletességű adatokra van szükség, árfolyamok esetén például a mintavételi gyakoriságot, hírek esetén a szöveg felhasznált részeit. Ezután ki kell választani az adatok szolgáltatóját és az eszközt, amellyel hozzá lehet férni azokhoz. A metaadatok, főleg a hírek kapcsán, fontos részét képezik a modellezésnek, például a publikálás ideje, a hír típusa, milyen nyelven íródott, milyen címkékkel, annotációkkal látták el stb. Az árfolyamok esetén ismerni kell az adathiányok okát, mint például a kereskedés felfüggesztése, de azt is, hogy korrigálták-e az árfolyamot osztalékfizetéssel és részvényfelosztással, stb.

A legrészletesebb adatok az eszköz árfolyamával kapcsolatban az ajánlati könyvben – order book – találhatók. Az ajánlati könyvben az összes vételi és eladási ajánlat megtalálható árszintenként aggregált a mennyiségekkel. A könyvben szerepelnek azok az ajánlatok is, amelyeket végül nem fogadnak el, így nem kapcsolódik hozzá ügylet. A második legrészletesebben a kötéslista[53] foglalja össze az eszköz árfolyamával kapcsolatos információkat. Ez a lista csak azokat az ajánlatokat tartalmazza, amelyekből ügylet született. A kötés időpontja mellett a kötési árfolyam és mennyiség szerepel az adattáblában, ezen kívül pedig a közvetlenül az ügylet megkötése előtti legjobb vételi és eladási ajánlat, illetve mennyiség is részét képezheti. Még viszonylag részletes adatokat tartalmaznak a nagyfrekvenciás[54] OHLC[55]-táblázatok. Az OHLC-táblázatok tartalmazzák az időszak időpontját[56], amelyikben az árfolyamadatok keletkeztek, az időszakban megfigyelt nyitó- és záróárfolyamot, valamint a legmagasabbat és a legalacsonyabbat, ezen kívül az időszakon belüli kötések összmennyiségét, amely kifejezhető például részvényszámban, vagy lotban stb. Ezen árfolyamadatokat egy meghatározott időszakra viszonyítják, amely leggyakrabban egy nap. Minden OHLC-adattábla, amely hosszabb időszakra vonatkozik, a rövidebb időszakra vonatkozó OHLC-táblák aggregálásával is előállítható, tehát a napi, heti, havi-OHLC táblák rendre egyre kevésbé részletes információt hordoznak. A kevésbé likvid eszközök nagyfrekvenciás OHLC-táblái esetén gyakori, hogy sok időszak hiányzik a táblázatból, ugyanis nem történt kötés minden periódusban

53 Angolul trade log, contracts log, sales data, illetve tick-by-tick data néven is hivatkoznak rá.

54 Nagyfrekvenciás adatok alatt az egy napnál rövidebb időszakokra bontott idősorokat értem. Ezek tipikusan négyórás, egyórás, félórás, negyedórás, ötperces, egyperces felbontásúak.

55 Az OHLC az angol open, high, low és close szavak akronimája, amelyek a nyitó-, legmagasabb, legalacsonyabb és záróárfolyamra utalnak.

56 Ez megvalósítástól függően lehet az időszak kezdő- vagy záróidőpontja.

annak ellenére, hogy volt lehetőség a kereskedésre – azaz a piac nyitva tartott és nem volt felfüggesztve a kereskedés.

Az előbbiekben tárgyalt árfolyamadatok adatbázisát az adatszolgáltató állítja elő, és tőle meghatározott díjazás ellenében szerezhetők be közvetlenül a megfelelő formában. Az ajánlati könyvhöz való hozzáférésnek például több szintje van, amelyek adatszolgáltatónként eltérőek lehetnek. Ezek skálája onnan, hogy csak a legjobb ajánlatok kerülnek megjelenítésre, odáig terjedhet, hogy az összes ajánlathoz hozzáfér az ügyfél valós időben. Az árfolyamadatokhoz való hozzáférés árazása szempontjából a részletesség mellett a valósidejűség is fontos szempont. A 15 perccel késleltetett árfolyamokhoz általában ingyen is hozzá lehet férni, a valós idejűekhez viszont csak fizetés ellenében. A nem valós idejű adatokat gyakran nevezik historikus adatoknak is, amelyek a naprakész kereskedés szempontjából – értsd egy aktuális megbízás kötési árának szempontjából – kevésbé relevánsak. Ezek egy része ingyenesen hozzáférhető, például a BÉT a napi, heti vagy havi historikus OHLC-adatok letöltését ingyenesen lehetővé teszi honlapján (Budapesti Értéktőzsde Zrt. é. n. a). A historikus nagyfrekvenciás és kötési adatok viszont már nem ingyenesek, a BÉT külön díj ellenében szolgáltatja őket (Budapesti Értéktőzsde Zrt. é. n. b), illetve külső adatszolgáltatókon keresztül teszi hozzáférhetővé, melyek szintén díj ellenében közvetítik tovább az adatokat (Budapesti Értéktőzsde Zrt. é. n. c). A külső adatszolgáltató szolgáltatásán keresztül elérhető nagyfrekvenciás és kötési adatok köre is korlátozott lehet a tekintetben, hogy csak bizonyos időszakra vonatkozóan – például az utolsó három hónapra – engedélyezett a historikus adatok letöltése. Az árfolyamok után tekintsük át, hogy a hírekhez milyen módon lehet hozzáférni, milyen formában kapjuk meg az adatokat.

A gazdasági információk közzétételének leggyorsabb módja, ha a világhálón keresztül történik. Ez nem csak honlapokon, hanem elektronikus üzenetküldő rendszereken, illetve kliensszoftvereken, mobil applikációkon keresztül is történhet. A honlapokon való közzétételnek megvan az az előnye, hogy széles körben elérhető, és csak egy böngészőre van szükség hozzá. A célszoftverek ezzel szemben szolgáltatások bővebb körének nyújtására alkalmasak, ám rendszerint csak regisztrált felhasználók számára érhetők el. A híreket publikáló szervezetek esetében e két megoldás egyidejű alkalmazásával a publikációs folyamat jelentősen gyorsítható. A Budapesti Értéktőzsde például a KIBINFO rendszer segítségével minimális emberi beavatkozással végzi a tőzsdei értékpapír-kibocsátók közleményeinek publikálását. A KIBINFO rendszerhez való hozzáférés egy kli-

ensszoftveren keresztül valósítható meg, amelyet a kibocsátó szervezetek, illetve az ő kapcsolattartásra kijelölt munkatársaik használhatnak a sajtóközleményeik kezelésére. A szoftveren keresztül az összes hír egy adatbázisba kerül, majd egy órán belül automatikusan megjelenik a tőzsde honlapján. A honlapon azután bárki számára elérhetők a hírek akár visszamenőlegesen is, ám a KIBINFO rendszerhez csak a regisztrált felhasználóknak és a tőzsde munkatársainak van hozzáférése (Budapesti Értéktőzsde Zrt. é. n. d). A vállalati hírek egyik fő forrását tehát a honlapok képezik, és a hírek nagy száma miatt szükség van rá, hogy azokat különböző szempontok szerint rendszerezze a közzétevő webhely. A BÉT kibocsátói hírei például kibocsátónként és nyelvenként – magyar és angol – elkülönítve találhatók, valamint a publikálás ideje szerint tízesével külön oldalon jelennek meg, ráadásul a hírek részletes tartalmát hírenként külön lapon lehet elérni. A tőzsdei hírbányászat szempontjából tehát különösen nagy jelentősége van az olyan technológiáknak, amelyek lehetővé teszik, hogy bonyolult szerkezetű honlapokon és webszolgáltatásokon keresztül elérhető információkat kinyerjünk. A nyilvános webtartalmak összegyűjtésének és a bennük lévő információk strukturálásának technológiáit *web crawling*nak és *web scraping*nek nevezzük.

A web crawling – webes begyűjtés – technológiája révén meghatározott URL-címekről, úgynevezett *seed*ekről kiindulva a weblapokon elérhető linkek, hiperhivatkozások követése révén bonyolult struktúrájú webdokumentum-hálózat tölthető le. A crawling működését leghatékonyabban szűrőkkel, szabályokkal lehet befolyásolni, amelyek leírják, hogy melyek azok a linkek, amelyeket a begyűjtő robot, a *crawler* követhet, illetve hogy a linken keresztül mely dokumentumok tölthetők le. A szoftvermegvalósítástól függően a követési és letöltési szabályok nem feltétlenül válnak el egymástól, valamint vonatkozhatnak kizárólag a link szerkezetére, vagy a linken keresztül elérhető dokumentum tartalmára is. (Mirtaheri et al. 2013) E szabályok tipikusan szövegminták formáját öltik, hiszen a hiperhivatkozások szöveges URL-címeket tartalmaznak. Szövegminták hatékonyan leírhatók reguláris kifejezésekkel – regular expressions, röviden regex vagy regexp – vagy egy erre a célra kialakított jelölésrendszerrel. Például a *http://exam.ktk.pte.hu/redmine/.*pdf* reguláris kifejezés minden olyan URL-re illik, amely a kutatócsoportunk webhelyén található pdf-fájlokra mutat. A crawler működését lehet szabályozni más tekintetben is. Azt, hogy a crawler a linkek milyen hosszú láncolatát vizsgálja át, a keresés mélységének nevezzük. A felkeresett, vagy a letöltött linkek száma mellett korlátozható a begyűjtésre szánt idő is. A begyűjtés azonban megterhelhe-

ti a megcélzott webhely szerverét, ezért a sávszélesség és a párhuzamos letöltések számát is korlátozni célszerű. Ezen kívül az etikus begyűjtés érdekében be kell tartani a webhely robotokra vonatkozó utasításait, amely a webhely gyökerében kerül elhelyezésre, ha létezik ilyen fájl. A robots.txt nevű állomány a robotkizárási szabvány[57] alapján írja le, hogy a webhely tulajdonosa a webhely mely részeit nem szeretné, hogy a keresőmotorok felkeressék. (Koster 1994) Nem etikus begyűjtő robotok figyelmen kívül hagyhatják a robots-fájl tartalmát.

A begyűjtött HTML-állományokat a félig strukturált adatok közé sorolhatjuk, mivel bizonyos szimbólumok – *tag*ek – révén a weboldal felépítése, a rajta megjelenített információk szerepe azonosítható, ha megfelelően készítették el az oldalt. A web scraping vagy *web wrapping* alatt a weboldalak meghatározott részén lévő információk kigyűjtését és strukturálását értjük. Ennek technikai megvalósítása függ a weboldal elkészítésének technikájától, azaz, hogy a webtartalom HTML, PHP, JavaScript stb. segítségével készült-e, dinamikus vagy statikus tartalommal rendelkezik-e. Bizonyos módszerek a böngészőkhöz hasonlóan értelmezik az oldalt, esetleg képfeldolgozási eljárásokkal nyerik ki a szöveges tartalmat, majd az így kapott információval dolgoznak tovább. Egyszerű HTML-állományok esetén viszont hatékonyabb, ha szövegminták, például regex, vagy valamilyen lekérdező nyelv, például Xpath, segítségével csak az oldalstruktúra szükséges részeiből nyerjük ki az információt. (Ferrara et al. 2014)

A *3. táblázatban* összegzésre kerültek a tőzsdei hírbányászatban használatos adatforrások, külön a híreké, külön az árfolyamoké. Mindkét esetben négy szempontot vizsgáltam meg: mi az adatforrás típusa (**Típus**), a hozzáférés módja (**H**[58]), az adatelemek – egyed-előfordulások – számának nagyságrendje (**N**[59]), illetve az információ részletessége – a hírnek mely részeit tartalmazta (**R**[60]), illetve az árfolyamok milyen részletesen álltak rendelkezésre (**F**[61]).

57 Robots Exclusion Standard

58 Hozzáférés. Lehetséges értékei: p (publikus), k (korlátozott).

59 Hírek esetén: a hírek számának nagyságrendje adat-előkészítés előtt. Árfolyamok esetén: időponthoz rendelt árfolyamértékek számának nagyságrendje adat-előkészítés előtt. A hírek, illetve árfolyamok számát $a \cdot 10^n$ alakban felírva: n.

60 A szöveg elemzéshez felhasznált részei. Lehetséges értékei: C (cím), L (lead), T (teljes).

61 Az árfolyamadatok mintavételi gyakorisága, frekvenciája. Lehetséges értékei: A (ajánlati könyv), T (tick-by-tick), s (egy másodperc), $n \cdot s$ (n másodperc), m (egy perc), $n \cdot m$ (n perc), h (egy óra), $n \cdot h$ (n óra), d (egy nap). A <*h* arra utal, hogy egy óránál rövidebb időszakról van szó.

3. Táblázat: A tőzsdei hírbányászati kutatásokhoz használt adatok

Modell	Cikk	Hírforrás				Árfolyamok forrása			
		Típus	H	N	R	Típus	H	N	F
Wüthrich	(Leung Kung Fan 1997) (Wüthrich, Cho, et al. 1998) (Wüthrich, Permunetilleke, et al. 1998) (Cho & Wüthrich 1999) (Cho et al. 1999)	napilap (WSJ, FT, CNN, IHT, Bloomberg)	p	2	T	?	?	2	d
	(Peramunetilleke 1997)	kutatói adatbázis (Olsen & Associates)	k	?	C	kutatói adatbázis (Olsen & Associates)	k	?	m
Lavrenko	(Lavrenko et al. 2000a; 2000b)	híraggregáló (Biz Yahoo!)	p	4	T	pénzügyi adatszolgáltató (Biz Yahoo!)	p	?	10m
	(Fung et al. 2002a; 2002b) (Fung 2003) (Fung et al. 2003; 2005)	híraggregáló (Reuters) kutatói adatbázis (Reuters, Newsgroup-20)	k	5	T	pénzügyi adatszolgáltató (Reuters)	k	6	T
Schumaker	(Schumaker & Chen 2006; 2008; 2009; 2010) (Schumaker 2009; 2010a; 2010b) (Schumaker et al. 2009; 2012)	híraggregáló (Yahoo! Finance)	p	4	T	pénzügyi adatszolgáltató	k	7	m
Groth	(Groth & Muntermann 2008; 2009; 2010; 2011)	newswire (DGAP)	p	2	T	pénzügyi adatszolgáltató (Reuters)	k	?	<h
	(Groth 2010)				T				A
	(Groth et al. 2014)				TC				
Thomas	(Thomas & Sycara 2000)	fórum (Raging Bull)	p	2	T	pénzügyi adatszolgáltató (Yahoo! Finance)	p	2	d
Gidófalvi	(Gidófalvi 2001)	lásd (Lavrenko et al. 2000a; 2000b)		4	lásd (Lavrenko et al. 2000a; 2000b)				
	(Gidófalvi & Elkan 2003)	?	?	4	T	?	?	?	m
Koppel	(Koppel & Shtrimberg 2004)	? (Multex Significant Developments)	?	4	T	?	?	?	?
	(Généreux et al. 2008; 2011)			4	T			?	?
Mittermayer	(Mittermayer 2004)	newswire (PRNewswire)	p	5	T	?	?	9	T
	(Mittermayer 2006) (Mittermayer & Knolmayer 2006a)	newswire (PRNewswire)	p	4	T	?	?	?	15s
e-Markets Group	(Zhang et al. 2005; 2006; 2007)	híraggregáló (Bloomberg)	?	?	T	?	?	?	?
	(Yu et al. 2005; 2006)	?	?	4	T	?	?	4	d

Forrás: saját szerkesztés

A hírforrások típusai közül leggyakrabban híraggregáló weboldalakon keresztül szerezték be az elemzéshez szükséges korpuszt, ezek között volt publikus, mint a Yahoo! Finance, illetve korlátozott hozzáférésű, mint a Thomson Reuters is. A legtöbb esetben a

hírek teljes szövegét legyűjtötték, két esetben csak a címmel végeztek kísérletet. A korpuszban lévő dokumentumok száma általában ezres nagyságrendű, három esetben százas, két esetben tízezres volt. Az árfolyamadatok forrása kapcsán elmondható, hogy sokszor felületesen vagy egyáltalán nem adták meg a forrást a cikkek szerzői, de a pénzügyi adatszolgáltatók, mint a korlátozott hozzáférésű Thomson Reuters vagy a publikus Yahoo! Finance, voltak leggyakrabban megemlítve. Az adatbázis méretét illetően is gyakran homályos információkkal szolgálnak a cikkek, bár ez a modellezés szempontjából kevésbé zavaró, ugyanis az adat-előkészítés során általában a hírek számától függ a felhasznált árfolyamadatok mennyisége. A kiindulási árfolyamadatok nagyságrendje a százastól a milliárdosig terjedhet. A legrészletesebb árfolyamadat, amit használtak, az ajánlati könyv volt, a legkevésbé részletes pedig egynapos, és leggyakrabban nagyfrekvenciás adatokkal készültek a modellek.

A saját modellemhez hírforrásként a BÉT weboldalának azt a részét választottam, amelyen a részvénykibocsátók sajtóközleményeit teszik közzé, ha a fenti kategóriákba szeretném sorolni, akkor ez egy publikus hozzáférésű newswire típusú hírforrásnak tekinthető. A kiindulási korpuszom mérete 965 angol és ugyanennyi magyar nyelvű hírből állt. Az árfolyamok forrásaként a korlátozott hozzáférésű Thomson Reuters Eikon platformot használtam, amelyből a BÉT Prémium kategóriás részvényeinek egyperces részletességű OHLC adatait töltöttem le, közel hatvanezret. A továbbiakban bemutatom a kiindulási adataim összegyűjtését, jellemzését és adatminőségét.

3.2.2. A saját modell adatai

A Budapesti Értéktőzsdén a részvényeket három szekcióba lehet bevezetni: Prémium, Standard és T. A Prémium kategóriába a likvidebb részvények kerülnek, és a bevezetési feltételek is szigorúbbak. A Standard és a T kategóriában a kis és közepes vállalatok részvényei találhatók. A különbség az, hogy a T kategóriába tartozó részvények bevezetésekor nem kell nyilvános tranzakciót végrehajtani. A fenti okokból tehát a leglikvidebb, Prémium kategóriás részvények (lásd *4. táblázat*) vizsgálatát végeztem el.

4. Táblázat: Prémium kategóriás részvények a BÉT-en 2015 júliusában

Értékpapír megnevezése	Kijelzés módja (Ticker)
ANY részvény	ANY
Appeninn részvény	APPENINN
CIG Pannónia részvény	CIGPANNONIA
FHB részvény	FHB
MOL részvény	MOL
Magyar Telekom részvény	MTELEKOM
OTP Bank részvény	OTP
PannErgy részvény	PANNERGY
Rába részvény	RABA
Richter Gedeon részvény	RICHTER

Forrás: (Budapesti Értéktőzsde Zrt. é. n. h)

A BÉT árfolyamadataihoz való hozzáférés történhet végfelhasználóként és vendorként, harmadik személynek való továbbadás céljából. Utóbbi adatszolgáltató cégekkel a BÉT információszolgáltatási szerződést köt, melynek értelmében a Tőzsde meghatározott díjak ellenében különböző információs csomagokhoz biztosít hozzáférést a Wiener Börse AG (WBAG) adatszolgáltató rendszerén keresztül. Az információs csomagok között eltérés van aszerint, hogy valós időben, vagy 15 perces késleltetéssel, illetve nap végén hozzáférhetők az adatok, valamint aszerint, hogy az ajánlati könyvhöz milyen mélységig férnek hozzá. Minden csomag a BÉT összes piacát lefedi. Végfelhasználóként az adatok jellegétől függően eltérő módon lehet elérni őket. Valós idejű adatok esetén a vendorokkal előfizetői szerződés kötése szükséges. E felhasználók után a vendor havidíjat fizet a BÉT-nek. A késleltetett és napvégi adatok esetén ez nem szükséges, és akár ingyenesen közzétehetők a vendor saját felületén is. (Budapesti Értéktőzsde Zrt. é. n. e, f, g, 2013) A BÉT Prémium kategóriás részvényeinek nagyfrekvenciás árfolyamait és tranzakciós adatait a Thomson Reuters nevű vendor Eikon nevű platformján keresztül töltöttem le, melyhez a Pécsi Tudományegyetem Közgazdaságtudományi Kara biztosított hozzáférést. Az Eikon platformon keresztül legfeljebb három hónap egyperces árfolyamadatai és tranzakciós adatai érhetők el, ezért mintám 2015.04.27. és 2015.07.24. közötti három hónapos időszakot öleli fel.

5. Táblázat: A rendelkezésre álló adatok mennyisége a kiindulási mintában

BÉT-ticker	Reuters-ticker	1 perces OHLC	Tranzakciók
ANY	ANYB.BU	1653	3027
APPENINN	APPB.BU	795	1235
CIGPANNONIA	CIGP.BU	996	1565
FHB	FHBK.BU	1337	2364
MOL	MOLB.BU	13307	39990
MTELEKOM	MTEL.BU	6624	12381
OTP	OTPB.BU	19596	84780
PANNERGY	PANP.BU	1059	1614
RABA	RABA.BU	529	878
RICHTER	GDRB.BU	13273	43293

Forrás: (Thomson Reuters é. n.) alapján saját szerkesztés

Mint látható az *5. táblázatban*, a Prémium kategóriás magyar részvényeken belül is jelentős különbségek vannak az elérhető nagyfrekvenciás adatok mennyiségét illetően. Ennek oka, hogy jelentősen eltér egymástól az egyes papírok likviditása. A táblázatból egyértelműen látható, hogy az OTP-papírokkal kb. kétszer annyi tranzakció történik, mint a MOL- vagy a Richter-papírokkal, amely értékek viszont kb. három és félszeresei a Magyar Telekom tranzakciószámának. Az egyperces adatokból legfeljebb kb. 30 ezer megfigyelés állhatna rendelkezésre az adott időszakban részvényenként, tekintve, hogy 63 kereskedési napról van szó, melyek mindegyikében kis eltérésekkel 9 és 17 óra között hajtják végre a tranzakciókat[62]. A legtöbb tranzakciót lebonyolító OTP esetében például ennek kb. 65%-ával egyezik meg a megfigyelések száma. Ez azt jelenti, hogy viszonylag magas annak a valószínűsége, hogy nem tudjuk kiszámolni a hozamot két tetszőleges idősori érték között, mert az egyik hiányzik. Ilyen esetben a hiányzó értékeket a legutóbbi érvényes árfolyammal interpoláltam.

A BÉT-en kereskedett termékek kibocsátóinak vannak bizonyos információszolgáltatási kötelezettségeik, amelyeknek úgy tudnak megfelelni, hogy a Tőzsde honlapján közzéteszik sajtóközleményeiket. A BÉT bevezetési és forgalombantartási szabályzata 17-től 21-ig terjedő pontjai szabályozzák a tőzsdével való kapcsolattartás, a rendszeres információnyújtás, a rendkívüli és egyéb tájékoztatás, a nyilvánosságra hozatal és a közzététel folyamatát, ezek közül a legfontosabb szabályokat az alábbiakban ismertetem. (Budapesti Értéktőzsde Zrt. é. n. d) A kibocsátók kötelesek biztosítani, hogy a befekte-

62 Kb. 63 × 480 = 30 240 perc.

tők egyformán ugyanazokat az információkat kapják meg róluk. A Tőzsdét meg kell hívni a sajtótájékoztatókra és a sajtónak szánt anyagot egyidejűleg meg kell küldeni a BÉT-nek is. A szabályzat megkülönbözteti a rendszeres, a rendkívüli és az egyéb tájékoztatást. A rendszeres információnyújtás kiterjed többek között a pénzügyi jelentésre, a felelős társaságirányítási jelentésre, a társasági eseménynaptárra, a tulajdonosi struktúrára és a vezetők személyére. A rendkívüli tájékoztatást a kibocsátónak az információ tudomására jutását követő 30 percen belül kezdeményeznie kell a kibocsátói információs rendszeren, a KIBINFO-n keresztül. Ha kereskedési időn kívül, illetve tőzsdenapon 7:30 előtt jutott tudomására az információ, akkor 8:00-ig kell ennek eleget tennie. Az egyéb tájékoztatási kötelezettség körébe egyrészt olyan észrevételek tartoznak, amelyek a kibocsátóról szóló hírekkel kapcsolatban a kibocsátó tudomására jutottak, és amelyek a kibocsátott értékpapír értékét befolyásolhatják. Ezeket az észrevételeket legkésőbb két órán belül kell a tőzsdéhez eljuttatni. Másrészt ide tartozik még a létesítő okirat módosítása, a befektetői kapcsolattartással megbízott munkatárs megváltozása, a kibocsátott értékpapírok más szabályozott piacra történő bevezetése és a csődeljárás, amelyek kapcsán 1 tőzsdenapos határidőt ír elő a szabályzat. (Budapesti Értéktőzsde Zrt. é. n. d)

A bevezetési szabályzat 21.1.1 pontja értelmében:

> *„A rendszeres és rendkívüli tájékoztatások Közzétételi Szabályzatban meghatározottaktól eltérő módon történő nyilvánosságra hozatala nem előzheti meg az adott tájékoztatás Közzétételi Szabályzat szerint történő megjelenését.*" (Budapesti Értéktőzsde Zrt. é. n. d, o.40)

Amíg a hír nem került feltöltésre a KIBINFO rendszerébe, addig harmadik személy részére nem lehet elküldeni a tájékoztatást. Az egyéb tájékoztatások esetén is törekedni kell arra, hogy az információ a tőzsdénél is azonos időben kerüljön publikálásra. A prémium részvényekkel kapcsolatos információkat angol nyelven is közzé kell tenni, és a kibocsátó köteles a teljes anyagot azonos időben közzétenni angolul is.

A hírek közzétételével kapcsolatban a *Közzétételi Útmutató* az irányadó. (Mohai 2010) A nyilvánosságra szánt információkkal kapcsolatban a következő adatokat kell megadni: a hír alanya, a hír típusa, rövid összefoglaló, a főoldali cím, az esemény időpontja. A főoldali cím jelenik meg a BÉT honlapján címként a *Kibocsátói hírek* között. Maga a hír csatolt fájlban, doc vagy pdf formátumban kerül benyújtásra, és a hírhez kapcsolódóan beküldhetők csatolt anyagok is pdf, doc, illetve xls formátumban. A tőzsde ellenőrzi, hogy a hír típusa, címe, összefoglalója és mellékletei megfelelnek-e a hír

tartalmának, és szükség esetén kéri azok módosítását. A híreknek a tőzsde honlapjára történő kihelyezése automatikusan történik a KIBINFO rendszer segítségével. A rendkívüli és egyéb tájékoztatások és az éves jelentés az adott tőzsdenapon 8:00 és 17:15 között kerülnek publikálásra, ezt az időintervallumot a *kereskedés szempontjából fontos időszak*nak nevezik. Ezek az információk normális esetben a feltöltést követő 60. percben kerülnek publikálásra a BÉT honlapján, de ha 7:00 és 8:00 között töltik fel a hírt, akkor 8:00-kor lesz publikálva. Ha a kereskedés szempontjából fontos időszakon kívülre esik az automatikus megjelenés ideje, akkor rendszerint a következő időszak kezdetén lesz publikálva a hír, de előre is hozható, ha erre lehetőség van.

A Prémium kategóriába tartozó részvényekhez tartozó sajtóközlemények forrásául a BÉT Kibocsátói hírek című oldala szolgált. (Budapesti Értéktőzsde Zrt. é. n. i) A BÉT adott kibocsátójához tartozó kibocsátói hírek a következő általános URL-en érhetők el magyarul:

http://bet.hu/topmenu/kibocsatok/kibocsatoi_hirek/kibocsatonkenti ,

a következő URL-en pedig angolul:

https://client.bse.hu/topmenu/issuers/issuersnews/issuersnews .

Az URL-hez az *issuerid* paraméter hozzáfűzésével lehet egy meghatározott kibocsátó híreit megjeleníteni. Az issuerid a kibocsátó négyjegyű azonosítószáma, ahogy a dolgozatban vizsgált részvények kibocsátói esetében a *6. táblázatban* látható.

6. Táblázat: A BÉT Prémium részvények kibocsátójához tartozó issuerid-kódok

Ticker	issuerid
ANY	3071
APPENINN	3341
CIGPANNONIA	3356
FHB	2436
MOL	1599
MTELEKOM	1633
OTP	1604
PANNERGY	1609
RABA	1616
RICHTER	1617

Forrás: (Budapesti Értéktőzsde Zrt. é. n. i) alapján saját szerkesztés

A kibocsátói hírek oldalon egyszerre a tíz legfrissebb hír jelenik meg, a korábbi hírekhez a *pagenum* paraméter URL-hez fűzésével férhetünk hozzá. A pagenum egy ter-

mészetes szám, és mivel az összes korábbi kibocsátói hír elérhető a BÉT honlapján, az idő múlásával a pagenum paraméter felső korlátja egyre nagyobb lesz. Mivel a különböző kibocsátók különböző ideje lehetnek jelen a tőzsdén, és különböző a hírek produkálásának intenzitása is, a pagenum paraméter felső korlátja részvénykibocsátónként eltérő. A paramétereket úgy kell az URL végéhez fűzni, hogy az URL után kérdőjelet írunk, majd az egyes paramétereket ésjellel – & – választjuk el egymástól, például: *http://bet.hu/topmenu/kibocsatok/kibocsatoi_hirek/kibocsatonkenti?issuerid=3071&pagenum=1* . Minden közleménynek saját HTML-oldala van, amelyet egy kilencjegyű számmal azonosítanak, például: *http://bet.hu/topmenu/kibocsatok/kibocsatoi_hirek/119560814.html* . A *7. ábra* egy kibocsátói hír HTML-oldalának szerkezetét szemlélteti. Az oldalon megtalálható a részvénykibocsátó neve – ANY PLC –, a publikálás dátuma és időpontja – 02 Jun 2014 10:45 –, a hír kategóriája – Other Information –, a hír rövid összefoglalója – Number of voting rights… – és a hír kibocsátója, forrása is – ANY PLC –, ami nem feltétlenül egyezik meg a részvénykibocsátóval. Az oldalról érhető el a hír szövegét tartalmazó PDF állomány is – ANY140602OR01E.pdf –, illetve a hír másik nyelvi változatára mutató link – kis magyar, illetve angol zászlóval jelezve. A hírek metaadatai tehát HTML-, szöveges tartalma pedig PDF-formátumban érhetők el.

7. Ábra: Egy kibocsátói hír weboldalának képe

Forrás: (ANY PLC 2014)

Míg hírenként legfeljebb egy angol és egy magyar nyelvű HTML-oldalt kell letölteni[63], addig egy adott hír adott nyelvi változata esetén legalább egy, de akár több PDF-állomány érhető el, ugyanis több esetben mellékletek is tartoznak egy-egy közleményhez, pl. közgyűlési határozatok, adattáblázatok stb. Ugyanannak a PDF-dokumentumnak a két nyelven elérhető változatával kapcsolatban, problémát jelent, hogy nem

63 Előfordulhat, hogy csak egy nyelven bocsátották ki a hírt, és nincs párja.

ugyanaz a két fájl neve, és a PDF-eket nem látták el olyan egyedi azonosítóval, mint a HTML-oldalakat, így az azonos tartalmú, de eltérő nyelvű változatok összepárosítása nem volt triviális feladat.

A letöltéséhez a HTTrack nevű több platformon elérhető, webhelyek tükrözésére és webes begyűjtésre szolgáló szoftvert használtam. (Roche et al. é. n.) Azok a hírek kerültek a kiindulási korpuszba, amelyek 2014.07.01 és 2015.06.30 között kerültek publikálásra, hogy a szöveges adatok előkészítésekor kellő nagyságú adathalmaz álljon rendelkezésre[64]. Ebből az egyéves időszakból 1084 magyar nyelvű HTML, 1181 magyar nyelvű PDF, 966 angol nyelvű HTML és 966 angol nyelvű PDF fájl került letöltésre. Ehhez az összes érvényes[65] issuerid-pagenum kombinációval képeztem egy-egy seed URL-t, amelyet szövegfájlként adtam meg a HTTrack számára. Ezen kívül a következő szabályokat[66] adtam meg, ugyanebben a sorrendben:

-*

+*/topmenu/kibocsatok/kibocsatoi_hirek/*[0-9].html

+*/newkibdata/*[0-9]*.pdf

+*bse.hu/topmenu/issuers/issuersnews/*[0-9].html

+*bse.hu/newkibdata/*[0-9]*.pdf

Az első szabály egy tiltó szabály, azt jelenti, hogy semmilyen URL nem keresendő fel, kivéve amelyeket külön engedélyezünk. A többi szabály engedélyező szabály. A második azt jelenti, hogy ha bármely seed alatt a megadott elérési úton számokból álló HTML-oldal található, akkor azt keresse fel és töltse le. A harmadik esetében hasonló van megfogalmazva PDF-fájlokra. A negyedik és ötödik szabály felel ugyanezeknek az angol megfelelőinek a letöltéséért.

64 Később látni fogjuk, hogy a 2015.04.27. és 2015.07.24. közötti időszakra és kereskedési időben csupán 92–158 dokumentum jut a kísérleti összeállítástól függően. Ez a modellezéshez elegendő, de az adat-előkészítésnél a kevésbé gyakori kifejezések azonosítását megnehezíti.

65 Csak a 6. táblázatban látható issuerid-k és a letöltés időpontjában a 2014.07.01 és 2015.06.30 közötti híreket tartalmazó pagenum-okat értem ez alatt.

66 Nem regex, hanem a HTTrack saját nyelve. Minden szabály új sorban található.

3.3. Adat-előkészítés

Az adat-előkészítés gyakran a legidőigényesebb, legtöbb munkával járó lépés az elemzés során. A megfelelő formátumra kell alakítani az adatokat, és össze kell állítani a mintát: ki kell választani a szükséges attribútumokat, ha kell, újat kell képezni, vagy össze kell vonni meglévőket, meg kell tisztítani őket, ki kell választani a megfigyeléseket, esetleg aggregálni kell őket. A tőzsdei hírbányászati modellezés esetén gyakorlatilag az elemzésben részt vevő összes attribútum adatbővítés során keletkezik. A hírekből szövegjellemzőket – szavak, kifejezések, entitások, tájolás stb. – kell kinyerni, amellyel reprezentálható a tartalom, az árfolyamokból pedig előbb hozamokat, trendeket képeznek, vagy egyéb módon reprezentálják őket, majd ezeket általában diszkrét értékekké transzformálják. A szövegjellemzők száma általában nagyon nagy, és kívánatos lehet azok számának csökkentése, ezt jellemzőkiválasztásnak nevezik. A minta összeállításakor a hírek és a szövegek reprezentációit egymáshoz rendelik, és kizárják azokat az eseteket, amikor a párból valamelyik tag hiányos. Ezen kívül további kizárások lehetségesek, például a kevés hírrel rendelkező, vagy a kisebb forgalmú részvények esetén. Általában az egy napos, vagy annál hosszabb időtávú modelleknél a hírek reprezentációit össze is vonhatják.

A *7. táblázat* a tőzsdei hírbányászati irodalomban alkalmazott adat-előkészítési technikákat foglalja össze. A táblázat három fő csoportra osztja az előkészítési feladatokat: a szövegreprezentáció, az árfolyam-reprezentáció és minta kialakítása. A szövegreprezentáció kialakítása a szövegben rejlő információt megragadó szövegjellemzők – általában kifejezések – azonosításából, méréséből és szűréséből áll. Az árfolyam-reprezentáció az árfolyam-idősorban lévő információt megragadó jellemzők – általában hozam vagy trend – azonosításából, méréséből, és diszkretizálásából áll. A minta kialakítása kapcsán a megfigyelések szűrése a leggyakoribb feladat. Terjedelmi okokból a *7. táblázatban* csupán rövidítések szerepelnek, a rövidítések feloldása a táblázat után található.

7. Táblázat: Adat-előkészítés a tőzsdei hírbányászati irodalomban

Modell	Cikk	Szövegreprezentáció				Árfolyam-reprezentáció		Minta	
		Szj.	S. s.	Dcs.	Dsz.	I. j.	Diszkr.	Kiz.	Méret
Wüthrich	(Leung Kung Fan 1997)	szkif	tf-ddf-nf	–	125	h	±0,3% nap, 3~	–	100
	(Peramunetilleke 1997)				400		±0,023% óra, 3~		100-110
	(Wüthrich, Cho, et al. 1998) (Wüthrich, Permun-etilleke, et al. 1998)				423		±0,5% nap, 3~		160
	(Cho et al. 1999)		tf, csr, csrc, crrd		392				179
Lavrenko	(Lavrenko et al. 2000a; 2000b)	bow	?	?	?	t	0,75 min/max, 0,5 min/max, 5?	?	?
	(Fung et al. 2002a; 2002b) (Fung 2003) (Fung et al. 2003; 2005)		tf-cdc-csc, ai-wc-cc	χ^2	?		klaszterezéses felosztás, 3?	?	?
Schumaker	(Schumaker & Chen 2006; 2009) (Schumaker 2009; 2010a; 2010b)	bow, np, ne, pn	bin	cf	kb. 2600, 2800, 3700, 4300, 5300	–	–	ny1ó, z20p, átfed	2600–2900
	(Schumaker & Chen 2008; 2010)	pn			?				kb. 2800
	(Schumaker et al. 2009; 2012)	bow, obj, sen	bin		?				
Groth	(Groth & Munter-mann 2008)	bow	tf-idf	stp, stm, df	?	h, cs	0% 15 perc, 2? ±1% 15 p., 3? átlag, 2? ±2% -os csúcs 15 perc, 2?	ny, z	160, ill. kb. 60
	(Groth & Munter-mann 2009)			stp, stm	?	h	0% 15 perc 2 (60% poz.)		423
	(Groth & Munter-mann 2010; 2011)			stp, stm, χ^2	?	ak	felső kvartilis 15, ill. 30 perc, 2 (25% poz.)		
	(Groth 2010)			stp, stm	?	al	alsó kvartilis 15 perc, 2 (75% poz.)	ny15p, z15p	369
	(Groth et al. 2014)	bow, wng, cng		stm, cf, wl, χ^2	?				415

Modell	Cikk	Szövegreprezentáció				Árfolyam-reprezentáció		Minta	
		Szj.	S. s.	Dcs.	Dsz.	I. j.	Diszkr.	Kiz.	Méret
Thomas	(Thomas & Sycara 2000)	bow	tf	?	?	h	0% nap, 2?	?	?
Gidófalvi	(Gidófalvi 2001)	bow	WB	stm, stp, mi	1000	ah	±0,2% n perc, 3?	hiány, ny, z	kb. 6000
	(Gidófalvi & Elkan 2003)						többféle, 3 (25% fel, 25% le, 50% semleges)	hiány, ismét	kb 6000
Koppel	(Koppel & Shtrimberg 2004)	bow	bin	stp, cf, ig	100	h	−7,8%, +10% nap, 2~	minár, extra	kb. 850
	(Généreux et al. 2008; 2011)	bow, szkif	bin, tf	cf, stm, ig, χ^2	100 szó, 420, ill. 123 kif.	h	±4% nap, 2~ ±2% − ±10% n nap, 3~	minár, extra, seml	200–1000
Mitter-mayer	(Mittermayer 2004)	bow	nbin	stm, stp, tf-idf	1000	cs	±3% óra, 3 (347 jó, 357 rossz, 5898 hatástalan)	több, hiány, minforg, ny, z	6602
	(Mittermayer 2006) (Mittermayer & Knolmayer 2006a)	bow, szkif	wdf-idf, bin, wdf, idf	idf, cf-idf, cf, χ^2, ig, or	?	mcs	±3% / 15 perc 3+1 (83 jó, 42 rossz, 504 semleges, 360 bizonytalan)	több, ny, z, biztn	629
e-Markets Group	(Zhang et al. 2005; 2006; 2007)	bow, wng	tf	stm, stp, χ^2	?	–	–	?	kb. 2600
	(Yu et al. 2005; 2006)	bow, w2g	tf-idf	stm, ?	36	t	?, 3?	?	kb. 2300

Forrás: saját szerkesztés

A következőkben az oszlopok szerint csoportosítva láthatók a rövidítések feloldásai és rövid magyarázatai, amely egyúttal áttekintést ad a különböző gyakran alkalmazott módszerekről is. Némely rövidítés több oszlopnál is szerepel, mert például gyakran előfordul, hogy egy mutatószámot használnak a szó-dokumentum mátrix súlyozására és dimenziócsökkentésnél a szövegjellemzők rangsorolására is.

Szj.: Szövegjellemzők. A szó-dokumentum mátrix dimenziói.

- **bow**: bag of words, szózsákmodell
- **cng**: character n-gram, betű-n-gram
- **ne**: named entities, entitások
- **np**: noun phrases, névszói csoportok
- **obj**: objektivitás
- **pn**: proper nouns, tulajdonnevek
- **sen**: sentiment, tájolás
- **szkif**: emberek által összeállított szakértői kifejezéslista, általában pénzügyi kifejezések, egészségügyi vagy megszemélyesítő metaforák
- **wng**, **w2g**: word n-gram, word bi-gram, szó-n-gramm, illetve szó-bi-gramm

S. s.: Súlyozási séma. A szó-dokumentum mátrixot alkotó számok meghatározásának módszere. A kombinált módszerek kötőjellel összekapcsolva láthatók a táblázatban.

- **ai**: average-importance coefficient, átlagos fontossági együttható
- **bin**, **nbin**: bináris súlyozás, normalizált bináris súlyozás
- **cc**: cross-category coefficient, kategóriák közötti együttható
- **cdc**: inter-cluster discrimination coefficient, klaszteren belüli elkülönülési együttható
- **crrd**: cluster relevance and discrimination, diszkriminatív klaszterrelevancia
- **csc**: intra-cluster similarity coefficient, klaszterek közötti hasonlósági együttható
- **csr**: class relevance, kategóriarelevancia
- **csrd**: class relevance and discrimination, diszkriminatív kategóriarelevancia
- **ddf**: document discrimination factor, dokumentum-elkülönülési együttható
- **idf**: inverse document frequency, inverz dokumentumgyakoriság
- **nf**: normalization factor, normalizálási együttható
- **tf**: term frequency, szógyakoriság
- **WB**: Witten-Bell simítás
- **wc**: within-category coefficient, kategórián belüli együttható
- **wdf**: within document frequency, dokumentumon belüli gyakoriság

Dcs.: Dimenziócsökkentés. A szó-dokumentum mátrix dimenzióinak elő- vagy utószűrése.

- **cf**: collection frequency, gyűjteménytámogatottság, a szó gyakorisága a korpuszban
- **df**: document frequency, dokumentumgyakoriság, a szót tartalmazó dokumentumok gyakorisága a korpuszban
- **idf**: inverse document frequency, inverz dokumentumgyakoriság
- **ig**: information gain, információnyereség
- **mi**: mutual information, kölcsönös információtartalom
- **or**: odds ratio, esélyhányados
- **stm**: stemming, szótövezés
- **stp**: stop-word removal, stopszavazás
- **tf**: term frequency, szógyakoriság
- χ^2 : khi-négyzet
- **wl**: word length, a szó hossza karakterekben mérve

Dsz.: Dimenziószám. Attribútumok, dimenziók száma dimenziócsökkentés után a szó-dokumentum mátrixban.

I. j.: Idősori jellemzők. Az árfolyam-idősor alapján számított attribútumok.

- **ah**: abnormális hozam
- **ak**: abnormális kockázat, abnormális hozamszórás
- **al**: abnormális likviditás, abnormális CRT
- **cs**, **mcs**: árfolyamcsúcs, ill. az árfolyam mozgóátlagának csúcsa
- **h**: hozam
- **t**: trend

Diszkr.: Diszkretizálás. Az idősori jellemzők diszkretizálásánál alkalmazott módszer, az első adat a kategóriák osztópontjait mutatja – például $\pm 3\%$ –, mögötte látható, hogy az osztópontokat mekkora időszakra kell vonatkoztatni – például 15 perc –, ezt követi az osztópontokkal határolt kategóriák száma – jellemzően három –, és ezek mögött azok eloszlása látható. Az egyenletes eloszlás jele a hullámos vonal: ~. Ha nem ismert a kate-

góriák eloszlása, akkor kérdőjel szerepel a szám mögött. A kategóriák eloszlása abszolút számokkal van megadva, ha azokat pontosan közölték – például (347 jó, 357 rossz, 5898 hatástalan) –, különben arányokkal. Ilyenkor, ha csak két kategória van, akkor csak az egyik megoszlása került feltüntetésre – például (75% poz.).

Kiz.: Kizárás. A mintából kizárásra kerülő megfigyelések szűrőszabályai.

- **átfed**: átfedő megfigyelések kizárása
- **biztn**: bizonytalan, azaz nem pozitív, nem negatív, nem semleges hírek kizárása
- **extra**: a tőzsdeindex hozamához képest extrahozamot nem realizáló részvények kizárása
- **hiány**: hiányzó értékkel rendelkező megfigyelések kizárása
- **ismét**: ismétlődő hírhez tartozó megfigyelések kizárása
- **minár**: a minimális árfolyam alatti részvények kizárása
- **minforg**: a minimális forgalom alatti részvények kizárása
- **ny**: nyitás előtti megfigyelések kizárása
- **ny#@**: nyitás utáni # hosszúságú, @ mértékegységű időszakba tartozó, vagy korábbi megfigyelések kizárása
- **seml**: a nem pozitív és nem negatív kategóriájú megfigyelések kizárása
- **több**: egyszerre több részvényhez kapcsolódó hír kizárása
- **z**: zárás utáni megfigyelések kizárása
- **z#@**: zárás előtti # hosszúságú, @ mértékegységű időszakba tartozó, vagy későbbi megfigyelések kizárása

Méret. A megfigyelések száma a mintában, kizárások után.

Az alfejezet további részeiben az egyes adat-előkészítési módszereket tárgyalom, ami során visszautalok a *7. táblázatra*, majd végül bemutatom a saját modellnél alkalmazott technikákat.

3.3.1. A hírek szövegének előkészítése

A rendelkezésre álló hírek szövege alapján el kell készíteni a szó-dokumentum mátrixot, amely a korpusz numerikus reprezentációja. A szó-dokumentum mátrix – angolul

term-document matrix, röviden TDM – egy M×N-es mátrix[67], ahol M a szótár mérete, azaz a szövegjellemzők száma, és N a korpuszban lévő dokumentumok – esetünkben hírek – száma. A mátrix elemei az adott szövegjellemző súlyát fejezik ki az adott dokumentumban. A TDM ritka mátrix – sparse matrix –, azaz legtöbb eleme 0, ugyanis M általában nagyon nagy N-hez képest, és egy-egy dokumentumban csak M töredékének megfelelő számú szövegjellemző található meg. A TDM-mel való munka hatékonyságának növelése érdekében ezért gyakran szoktak M méretét csökkentő módszereket alkalmazni. A szövegeknek ezt a reprezentációját vektortérmodellnek, vagy szózsákmodellnek szokták hívni. (Tikk 2007) A TDM felépítéséhez először meg kell határozni a szótárt alkotó jellemzőket. Ez a lépés a szöveg kívánt részének kinyerése után annak kifejezésekre bontásával, tokenizálásával valósul meg. A legegyszerűbb tokenizálási módszer szavakra bontja a szöveget. A *7. táblázat* szövegjellemzőket összegző oszlopában is látható, hogy ez a módszer – bow rövidítéssel – a leggyakoribb, a Wüthrich-féle modellen kívül minden modell alkalmazta. Ezekből a tokenekből képezhetők kettő- vagy többelemű sorozatok, amelyeket bi-, illetve n-grammnak nevezünk[68]. Ezzel a megoldással találkozhatunk Groth et al. (2014), illetve az e-Markets Group modelljei esetében (lásd *7. táblázat*). A tokenek és n-grammok mellett szól, hogy nyelvfüggetlenek, tehát ugyanúgy alkalmazhatók angol, német és magyar nyelvre is, továbbá nem igényelnek különösebb emberi beavatkozást. A szövegben a kifejezések szakértői lista alapján is azonosíthatók, ilyenkor a többi szót figyelmen kívül is hagyjuk. Ezt a megoldást alkalmazták a Wüthrich-modellben, illetve összehasonítás jelleggel Koppel (Généreux et al. 2008; 2011), illetve Mittermayer és Knolmayer (2006a) használta még. Ez a módszer azonban már nem nyelvfüggetlen, és emberi beavatkozást igényel. Általában nem nyelvfüggetlenül, de emberi beavatkozás nélkül kinyerhetők a szövegből bizonyos nyelvtani jellemzők is, mint például a névszói csoportok, az igék, az entitások, a tulajdonnevek, vagy szemantikai jellemzők, mint például a szöveg tájolása vagy objektivitása. A nyelvtani szerepük – parts of speech – alapján képzett kifejezések a tokenekhez hasonlóan a TDM-nek egy-egy dimenziói lesznek. Az objektivitás és a tájolás általában egy folytonos vagy ordinális skálán mért mutatószám, amelyet a szöveg tokenjeinek érzelmi töltése alapján becsülnek. A Schumaker-féle modellekben láthattunk példát ezek alkalmazására.

A jellemzők körének meghatározása után ki kell számolni a TDM elemeit, ez többféle súlyozási séma szerint történhet. A legelterjedtebb módszerek a bináris súlyozás, a

67 A gyakorlatban sokszor N×M-es mátrixként használják.
68 Képezhetünk n-grammokat karakterekből is.

szógyakoriság – term frequency, tf – alapú súlyozás, és a tf-idf – term frequency and inverse document frequency – súlyozás, illetve ezek normalizált alakjai. A bináris súlyozás esetén a *j* kifejezés súlya 1 az *i* dokumentumban, ha az megtalálható benne, különben 0. A tf-súlyozásnál ugyanez az elem a *j* kifejezés *i* dokumentumban való előfordulásainak számával egyezik meg. A tf-idf súlyozás esetén a tf-súlyt korrigáljuk a szó korpuszbeli gyakoriságával:

$$\text{tf-idf}_{i,j} = \text{tf}_{i,j} \cdot \text{idf}_j = \text{tf}_{i,j} \cdot \log\left(\frac{N}{n_j}\right) \tag{8}$$

Ahol n_j azon dokumentumok száma, amelyben előfordul a *j* kifejezés. Ritkábban előforduló szavakhoz így nagyobb súly rendelhető. A normalizáláshoz a fenti mutatók értékét a dokumentum hosszával szokás leosztani, amelyre különféle távolságmértékek használhatók, a legegyszerűbb a dokumentum szavainak száma, vagy a dokumentumvektor hossza az euklideszi térben. (Tikk 2007) Ezeket az alapvető súlyozási sémákat alkalmazta a legtöbb tanulmány. Ezen kívül az osztályozási feladathoz igazítva alkalmazhatók olyan súlyozási sémák is, amelyek figyelembe veszik a kifejezések kategóriák közötti eloszlását is, ilyet látunk a Wüthrich-modellnél vagy a Lavrenko-modell Fung-féle változatánál (lásd *7. táblázat*).

A jellemzők száma a súlyozás előtt és után is csökkenthető nyelvtechnikai vagy matematikai módszerekkel. A nyelvtechnikai módszerek közül a leggyakoribb a szótövezés és a stopszavazás. Szótövezés során a tokenekből eltávolítanak minden olyan karaktert, amely ragok, képzők stb. részét képezik. Ez általában emberi beavatkozás nélkül történik szótár segítségével, nyelvfüggő módon, vagy heurisztikus algoritmusok révén, nyelvfüggő vagy nyelvfüggetlen módon. Az algoritmikus szótövezésnél nem követelmény, hogy a szótári alakra redukálódjanak a szavak, a kapott csonkok gyakran rövidebbek vagy hosszabbak. A stopszavazás során eltávolítjuk azokat a tokeneket, amelyek lényegi információt nem hordoznak. Ez tehát egy nyelvfüggő módszer, mert szükség van egy listára, amely alapján az adott nyelv stopszavai kiszűrhetők. A stopszavak általában kötőszavak, névelők stb., de ha egy szaknyelvi korpuszt elemzünk, akkor az általános nyelv más szavai is felkerülhetnek a listára. (Tikk 2007) Szótövezést alkalmaztak a Groth-modellnél és az e-Markets Group modelljeinél, ezen kívül Gidófalvi (2001; Gidófalvi & Elkan 2003), Mittermayer (2004) és Koppel bizonyos cikkeiben (Généreux et al. 2008; 2011). (Lásd *7. táblázat.*) A matematikai módszerek közül a stopszavazáshoz közel áll a szó karakterekben mért hosszán és a kifejezés gyűjteménytámogatottságán vagy

dokumentumgyakoriságán alapuló módszerek. Azért állnak közel, mert a legtöbb stopszó nagyon rövid, illetve nagyon sokszor előfordul. Ezen mutatókra alsó és/vagy felső korlátot szabva elég hatékonyan kiszűrhetők a kevésbé jelentős, vagy véletlenszerűen megjelenő, zaj jellegű szavak. Erre látunk példát a Schumaker-modell esetén, Koppel munkáiban, valamint Groth (Groth & Muntermann 2008; Groth et al. 2014) és Mittermayer és Knolmayer (2006a) bizonyos cikkeiben. (Lásd *7. táblázat.*) A szövegjellemzők számának csökkentésére az osztályozási feladathoz igazított matematikai módszerek is rendelkezésre állnak, ami alatt azt kell érteni, hogy a kifejezéseket olyan mutatókkal jellemezzük, amely azoknak a dokumentumok közötti diszkriminálóerejét méri. Ide tartozik a súlyozási sémáknál megismert tf-idf mutató is, továbbá az információnyereség, kölcsönös információtartalom, a khi-négyzet statisztika, illetve az esélyhányados. Ezek alapján a kifejezések rangsorolhatók, és a rangsorból különböző módszerekkel kiválaszthatók a végleges szövegjellemzők. A kiválasztás történhet mutató-határérték alapján, vagy úgy, hogy meghatározott számú, illetve az összesnek meghatározott százalékát jelentő kifejezést tartunk meg közülük. (Tikk 2007) A szakirodalomban ilyen fajta dimenziócsökkentésre találunk példát a Lavrenko-modell Fung-féle változataiban, Gidófalvinál, Koppelnél, Mittermayernél, Groth (Groth & Muntermann 2010; 2011; Groth et al. 2014) és az e-Markets Group (Zhang et al. 2005; 2007) bizonyos cikkeiben. (Lásd *7. táblázat.*)

A TDM dimenzióinak száma jellemzőkiválasztás után általában százas és ezres nagyságrendben mozog a tőzsdei hírbányászat esetén. A szakértői kifejezéslisták mérete mind a Wüthrich-modell esetén, mind Généreux et al. (2008; 2011) cikkeiben kb. 100–400 közötti volt. A Schumaker-modell különböző változataiban kb. 2600–5300 jellemző maradt a dimenziócsökkentés után, Mittermayer és Gidófalvi 1000-1000 kifejezéssel dolgozott, Koppel viszont jelentősen redukálta a szótár méretét, és csak 100 szót választott ki. (Lásd *7. táblázat.*) A következőkben a saját modellem híreinek előkészítését mutatom be.

3.3.1.1. A saját modell híreinek előkészítése

Ebben az alfejezetben leírom, hogyan készítettem elő a nyers korpuszt a szövegjellemzők kinyeréséhez, majd a szövegjellemzők kinyerése kapcsán röviden bemutatok egy algoritmust, de előtte kitérek az alkalmazott dimenziócsökkentési eljárásokra, és a TDM súlyozási sémájára.

A 3.2.2. alfejezetben leírtam, hogy a nyers, nyelvenként kb. 1000-1000 elemű korpuszom különálló HTML-, és PDF-fájlokban található, melyekből a hírek metaadatait és a hír szövegét kellett kinyerni mielőtt a szövegjellemzőket kiválasztottam volna. A HTML-fájlok feldolgozását RapidMinerben a *Read Documents* operátorral végeztem, amellyel TXT-formátumban olvastam be azok tartalmát, nem HTML-ként, mert így a HTML-tagek benne maradtak az állományban. A HTML-tagekre azért volt szükség, mert ezek felhasználásával volt lehetőség arra, hogy az oldalaknak egy-egy meghatározott tagek közé eső részét kinyerjem. Ezek, a *7. ábra* kapcsán említett információk az *Extract Information* operátorral lekérdezhetők Xpath nyelven, ha az *xpath queries* paraméternél a következőket állítjuk be[69]:

8. Táblázat: Az angol hírek metaadatainak XPath lekérdezései

attribute name	query expression
NewsSubject	//h:div[@class='hiralany']//h:a/text()
PublicationTime	//h:div[@class='Time']/text()
NewsType	//h:div[@class='newstype']/text()
NewsSummary	//h:div[@class='Newkib_Sum']/text()
DataDealer	//h:div[@class='datadealer']/text()[2]
HungarianNewsURL	//h:a[@class="English"]/@href[number(substring(.,string-length(.)-13,9))=substring(.,string-length(.)-13,9)]

Forrás: saját szerkesztés

Mivel az angol nyelvű változathoz tartozó lekérdezéseket tartalmazza a *8. táblázat*, az utolsó attribútum neve HungarianNewsURL lett, de a magyar nyelvű hírek esetén ezt EnglishNewsURL-re kell átírni, míg a *query expression* paraméter értékét nem kell megváltoztatni, mert a magyar nyelvű változatban is az *English* szó szerepel benne. Mivel a seed-oldalak magyar nyelvűek, és az angol híreket a magyarokon keresztül követve töltötte le a HTTrack, ezért olyan eset előfordulhat, hogy a magyar nyelvű hírnek nincs angol megfelelője, azonban fordítva nem lehetséges. A hírek azonosítóját – NewsID néven – a HTML-fájl nevéből kinyertem egy *Replace* operátorral olyan módon, hogy a fájlnévből eltávolítottam az összes olyan részt, amely nem az azonosítóhoz tartozik[70]. Ugyanezzel a módszerrel kinyerjük a HungarianNewsURL-ből a magyar, illetve EnglishNewsURL-ből az angol változat azonosítóját rendre HunNewsID és EngNewID néven.

69 A h: azonosító a HTML-névteret azonosítja RapidMinerben.

70 A *replace what* paraméternél adjuk meg a következő regexet: *.*/([0-9]*).html*, a *replace by* paraméternél pedig a következőt1: *$1*.

A letöltött PDF állományokból kizárólag a szöveges információt használtam, azaz a számszerű információk – pl. eredménykimutatás- és mérlegadatok – elhagyásra kerültek, mert egyrészt ezek elemzése nem képezi a vizsgálat tárgyát, másrészt ezek hatékonyabban kinyerhetők másféle adatforrásból. A PDF-fájlok tartalmát ugyancsak RapidMinerben nyertem ki a *Read Document* operátorral[71]. A PDF-hez tartozó hír azonosítója a fájlt tartalmazó mappa nevével egyezik meg, így a metaadatokból kiolvasható. Az azonos azonosítóhoz tartozó hírek szövegét *Aggregate* operátorral összevontam[72]. A PDF-ekből kinyert szöveg és a HTML-ekből kinyert metaadatok párosításához a *Join* operátor használható, amely kulcsként a NewsID azonosítót használja. Az angol-magyar közleménypárok a NewsID és a megfelelő HunNewsID, illetve EngNewsID kulcsként való alkalmazásával ugyancsak *Join* operátorral párosíthatók össze.

Ezt követően sor kerülhetett a szövegjellemzők körének kijelölésére. 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. Előbbi tehát a sztenderd szózsák modellnek felel meg, amelyet a *7. táblázatban* bow rövidítéssel jelöltem, utóbbi viszont – a félig automatizált jellege miatt – nem teljesen felel meg az szkif rövidítéssel jelölt szakértői kifejezéslistának. Hamarosan kitérek arra, hogy milyen algoritmust alkalmaztam ehhez, de előbb tisztázom, hogy milyen nyelvi és matematikai dimenziócsökkentést alkalmaztam, és hogyan súlyoztam a TDM-et. Minden karaktert kisbetűssé transzformáltam *Transform Cases* operátorral, és azokból a karaktersorozatokból képeztem tokeneket, amelyek csak betűket tartalmaztak[73]. Szótövezést végeztem a Snowball-szótövező angol, illetve magyar változatával[74]. Kizártam a három karakternél rövidebb tokeneket[75], továbbá azokat, amelyek 10-nél kevesebbszer fordultak elő a korpuszban[76]. A TDM súlyozására a tf-idf mértéket alkalmaztam[77], amely az egyik leggyakoribb az irodalomban.

71 Ennek *extract text only* paramétere volt beállítva, valamint a *content type* paraméter értékét *pdf*-re állítottam.

72 A *group by attributes* értékét *NewsID*-re, az *aggregation attribute* a *text* (szöveg) volt, az *aggregation functions* pedig *concatenation*.

73 A *Tokenize* operátor *mode* paraméterét *non letters* értékre állítottam.

74 A *Stem (Snowball)* operátor *language* paraméterénél mindkettő kiválasztható.

75 A *Filter Tokens (by Length)* operátor *min chars* paraméterét 3-ra, a *max chars* paraméterét pedig egy nem korlátozó, nagy számra, 9999-re állítottam.

76 Ez a *Process Documents from Data* operátor *prune method* paraméterén keresztül befolyásolható, annak értékét *absolute*-ra állítottam, illetve a hozzá tartozó *prune below absolute* paraméter értékét 10-re, a *prune above absolute* paraméter értékét pedig egy nem korlátozó, nagy számra, 9999-re állítottam.

77 A súlyozást a *Process Documents from Data* operátor *create word vector* paraméterének igazra állítása után a *vector creation* paraméter *TF-IDF* értékre való állításával értem el.

A kifejezések kinyerésére szolgáló algoritmust Mostafa (2007) algoritmusa alapján készítettem el RapidMinerben, és alkalmazás közben derült rá fény, hogy nem csupán szakkifejezések kinyerésére alkalmas, hanem az egy kaptafára íródott sajtóközlemények sablonos szövegrészeinek azonosítására is. Ez felhasználható arra, hogy egyfajta kiterjesztett stopszavazás révén a kevésbé jelentős szövegrészeket eltávolítsuk a dokumentumokból. A PDF állományokban több ilyen irreleváns szövegrész is található, mint például a kibocsátó cég által a fájlba illesztett fejléc, instrukciók, jogi felelősséget kizáró szövegek, további tájékozódási lehetőségeket bemutató bekezdések stb. E szövegrészek – továbbiakban sablonszövegek – jellemzője, hogy több dokumentumban is megtalálhatók változatlan, vagy minimálisan eltérő formában. 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. E kétfajta felosztás mind a négy lehetséges kombinációját megvizsgáltam mind angol, mind magyar nyelven. Az algoritmus leírása az *2. függelékben* olvasható. 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, ezek a számok leginkább a Schumaker-modellben alkalmazott dimenziószámoknak felelnek meg. (Lásd *7. táblázat.*)

3.3.2. Az árfolyamadatok előkészítése

A 3.2. alfejezetben bemutattam az árfolyamadatok fajtáit. Az eltérő fajtájú adatokon más-más adat-előkészítés lehet szükséges, és az elemzésnél használt modellek is különböző adat-előkészítést igényelhetnek. A következő néhány bekezdésben az árfolyammodellezésben használt leggyakoribb reprezentációkat veszem sorra. A tárgyalás során az egy periódust jellemző reprezentációktól haladok a több periódust jellemzők felé, illetve az egy adatot felhasználóktól a több adatot felhasználók felé.

A nyitó-, legmagasabb, legalacsonyabb és záróárak közvetlenül is input attribútumai lehetnek egy idősori elemzésnek, ezeket az alábbiakban a következőképpen jelölöm: p_t^x, ahol t az időperiódus sorszáma, és $x \in \{o, h, l, c\}$ az angol elnevezések kezdőbetűi[78]. Ezekkel az adatokkal közvetlenül csak egy periódusnyi időszakot tudunk jellemez-

78 Open, high, low, close.

ni, 1-1 adat felhasználásával. Ilyen reprezentációt használt a Schumaker-modell, illetve Zhang et al. (2005; 2006; 2007) bizonyos cikkekben. (Lásd *7. táblázat.*) A következő mutatószámok – az árfolyamváltozás és a hozam – egy vagy több periódusból álló időszakok jellemzésére is alkalmasak, tipikusan 2-2 adat felhasználásával. A d hosszúságú, több periódusból álló időszak alatti árfolyamváltozás képlete a következő: $D^x_{t,d} = p^x_{t+d} - p^x_t$ (9), ahol $0 < d$ egész szám. Az OHLC-adatok birtokában lehetőség van a perióduson belüli árfolyamváltozás kiszámítására is: $D_t = p^c_t - p^o_t$ (10). A hozamok számítására többféle képlet használható. A d hosszúságú időszak alatti hozam képlete: $R^x_{t,d} = \left(p^x_{t+d} - p^x_t\right) / p^x_t$ (11), ahol $p^x_t \neq 0$. Az ugyanezen időszak alatti loghozam képlete: $^{\ln}R^x_{t,d} = \ln\left(p^x_{t+d}\right) - \ln\left(p^x_t\right)$ (12), ahol $0 < p^x_{t+d}$, $0 < p^x_t$. Az egyszerű hozamképletet használta a Wüthrich-modell, Koppel és Groth és Muntermann (2008; 2009) bizonyos cikkeiben. (Lásd *7. táblázat.*) A loghozamok előnye, hogy számtani átlaggal átlagolhatók, és hosszabb időszakra aggregálhatók, ezt a tulajdonságukat használta ki Groth (2008) is.

Abban az esetben, ha rendelkezésünkre állnak referenciaértékek a vizsgált eszköz hozamait tekintve, akkor az árfolyamokat reprezentálhatjuk a tényleges és a referenciaérték eltéréseként is. A referenciahozam lehet egy másik eszköz hozama – például a kockázatmentes kamatláb, a részvényindex hozama, overnight hitel- és betéti kamatok stb. –, vagy valamilyen közgazdasági modell – például a CAPM, paritáselméletek (Moosa & Bhatti 1997), ökonometriai modellek stb. – alapján számított érték is. Ezt a referenciahozamot normális hozamnak nevezzük, az ettől való eltérést pedig abnormális hozamnak, amely tehát a következőképpen számítódik: $AR^x_{t,d} = R^x_{t,d} - \hat{R}^x_{t,d}$ (13), ahol $\hat{R}^x_{t,d}$ a normális hozam. A képletet loghozamokra is felírhatjuk, ahogy Gidófalvi (2001; Gidófalvi & Elkan 2003) tette.

A hozam mellett a volatilitást jellemző mutatószámokkal is reprezentálhatók az árfolyamok. A legegyszerűbb ilyen mutató a perióduson belüli terjedelem: $M_t = p^h_t - p^l_t$ (14). A legismertebb azonban az n darab hozamból számolt szórás, amely már kettőnél több periódus adatából kiindulva jellemzi a köztes időszakot:

$$\sigma^x_{t,d,n} = \sqrt{\frac{\sum_{i=1}^{n}\left(R^x_{t+(i-1)\cdot d,d} - \bar{R}^x_{t,d,n}\right)^2}{n}} \qquad (15)$$

Ahol $0 < n$ és egész szám, d továbbra is a periódusok hossza,
$\bar{R}^x_{t,d,n} = \left(\sum_{i=1}^{n} R^x_{t+(i-1)\cdot d,d}\right) / n$ az n darab hozam számtani átlaga.

A képletben helyesebb a loghozammal számolni azok számtani átlagolhatósága miatt. Sem a hozamszórás, sem annak négyzete, a hozamvariancia nem használatos a tőzsdei hírbányászatban, hanem inkább azoknak a referenciaértéktől való eltérésén alapuló megközelítést alkalmazzák. Ez az abnormális volatilitás, amelyet például a következőképpen mérhetünk: $AV^x_{t,d,n} = \sigma^x_{t,d,n} - \hat{\sigma}^x_{t,d,n}$ (16), ahol $\hat{\sigma}^x_{t,d,n}$ a normális szórást jelenti. A szórás helyett a varianciával is felírható a képlet. Groth & Muntermann (2010; 2011) például egy hosszabb időszak alatti periódusok átlagos hozamszórásaként definiálta a normális szórást.

Az időszakot az azt alkotó periódusok árfolyamaira illesztett regressziós egyenessel is reprezentálhatjuk, például:

$$\underset{a,b}{arg\,min} \sum_{i=1}^{n} \left(p^x_{t_i} - \hat{p}^x_{t_i} \right)^2 \tag{17}$$

A következő feltételek mellett: $\forall i \in [1,n]$ esetén $\hat{p}^x_{t_i} = a \cdot t_i + b$.

A Lavrenko-modell például a kapott regressziós egyenes a paraméterével, a meredekséggel reprezentálja az árfolyamidősort. (Lásd *7. táblázat.*)

Egy több periódusból álló időszakot jellemezhetünk az időszak alatti abnormális hozamok, illetve volatilitások aggregált (általában átlagolt vagy kumulált) értékével is. A kumulált abnormális hozam képlete például:

$$CAR^x_{t,d,n} = \sum_{i=1}^{n} AR^x_{t+(i-1)\cdot d,d} \tag{18}$$

Ez a reprezentáció elterjedt az eseményvizsgálat módszertanban – pl. (Bedő & Rappai 2004; 2006) –, azonban a tőzsdei hírbányászatban kevésbé alkalmazzák.

Ezen kívül még számos egyedi árfolyam-reprezentációval találkozhatunk, és gyakorlatilag az összes technikai indikátor is ezek közé sorolható. Achelis (2001) könyvében például több fent említett reprezentáció is szerepel mint technicista indikátor. Az egyedi megoldások közül az árfolyamcsúcs a legelterjedtebb, amely az árfolyam adott időszakon belül megfigyelt szélsőértékei alapján minősíti az időszakot. Ilyet láthatunk Mittermayer munkáiban és Groth és Muntermann (2008) egyik cikkében. (Lásd *7. táblázat.*)

A saját modellem esetén a loghozamok alapján reprezentáltam az árfolyam-idősorokat. A loghozamok diszkretizálása során három kategóriát alakítottam ki: negatív, pozitív, illetve semleges. A tőzsdei hírbányászatban leggyakrabban kettő vagy három kategóriát használnak (lásd *7. táblázat*), és mivel hipotéziseim nem zárják ki annak lehetőségét, hogy a bejelentéseknek nincs hatása a részvényárfolyamra, ezért döntöttem úgy,

hogy harmadik kategóriaként a semleges is részét képezi az árfolyam-reprezentációnak. A diszkretizálás során a lehető legegyenletesebb megoszlás kialakítására törekedtem, összhangban a Wüthrich-, illetve Koppel-modellel (lásd *7. táblázat*). Mivel a diszkretizálást különböző időablakok mellett is elvégeztem, és a mintabeli arányok, sőt a minták elemszámai is minden esetben eltérnek, a diszkretizálás eredményét csak a *3.3.3. alfejezetben* fogom bemutatni.

3.3.3. A mintába kerülő megfigyelések szűrése

A modellezés minőségének javítása érdekében a rossz minőségű megfigyeléseket érdemes kiszűrni a mintából. A megfigyeléseket minősíthetjük például abból a szempontból, hogy mennyire jól lehet őket besorolni a pozitív vagy a negatív kategóriába. Így például Mittermayer és Knolmayer (2006a) külön kategóriába sorolták a bizonytalan híreket, amelyeket kizártak, vagy például Généreux et al. (2008; 2011) a semleges – nem igazán pozitív és nem igazán negatív – híreket tekintette kizárandónak. A besorolást kissé megnehezíti, ha nem áll rendelkezésre a szükséges árfolyam-adat – bár vannak módszerek, amivel ez kezelhető. A hiányzó értékek keletkezésének több oka is lehet, de leggyakrabban azért fordul elő, mert egy periódusban nem volt kereskedés. Ilyenkor a hiányzó adatot pótolhatjuk az adott periódus előtt megfigyelt utolsó adattal, vagy abban az esetben, ha az ajánlati könyv adatai rendelkezésre állnak, használhatjuk a legjobb ajánlat értékét is. A hiányzó értéket meg is becsülhetjük valamilyen idősori, árfolyamelméleti stb. modell segítségével, ha ez nem torzítja el lényegesen a végeredményt. Ha a minta elég nagy, akkor a hiányzó értéket tartalmazó időszakot kizárhatjuk a mintából – ezt alkalmazzák leggyakrabban. Nincs is más választásunk mint a kizárás abban az esetben, ha a hírhez tartozó időablak túllóg a tőzsde nyitvatartási idején. Ezen kívül a tőzsdei hírbányászatban gyakran olyan likvid részvényekkel kereskednek, amelyek esetében a közlemények publikálásának intervalluma átfedhet a modellben alkalmazott időintervallummal, ezért ilyenkor az átfedő híreket kizárják a mintából. A kereskedési szakasz előtt kibocsátott hírek esetén ugyanez a jelenség indokolja a nyitás utáni rövid időszakban megjelent hírek kizárását. Ezen kívül, ha egy hír több vállalathoz is kapcsolódik, akkor ellentmondó hatással lehet a két árfolyamra. Ha pedig egy közleményt részben megismételnek, akkor a szöveg alapján indokolt hatása az árfolyamra elmaradhat. Ilyen, hiányzó értékre, vagy átfedő hírekre vonatkozó szűrőfeltételt alkalmaztak a Schumaker-, a

Groth-, a Gidófalvi-, és a Mittermayer-féle modelleknél, aminek eredményeként többségben ezres nagyságrendű maradt a minták mérete[79]. (Lásd *7. táblázat.*)

A rossz minőségű megfigyelések kiszűréséhez használhatók metaadatok is, mint például a szóban forgó részvény piaci kapitalizációja, átlagos forgalma, átlagos árfolyamának nagysága, vagy a tőzsdeindex hozamához mért teljesítménye, extraprofit-termelő képessége – ezek mind a kibocsátó vállalat piaci jelentőségének indikátorai. Ilyen szűrőfeltételekre látunk példát Koppel modelljénél, és Mittermayer (2004) egyik cikkében. (Lásd *7. táblázat.*)

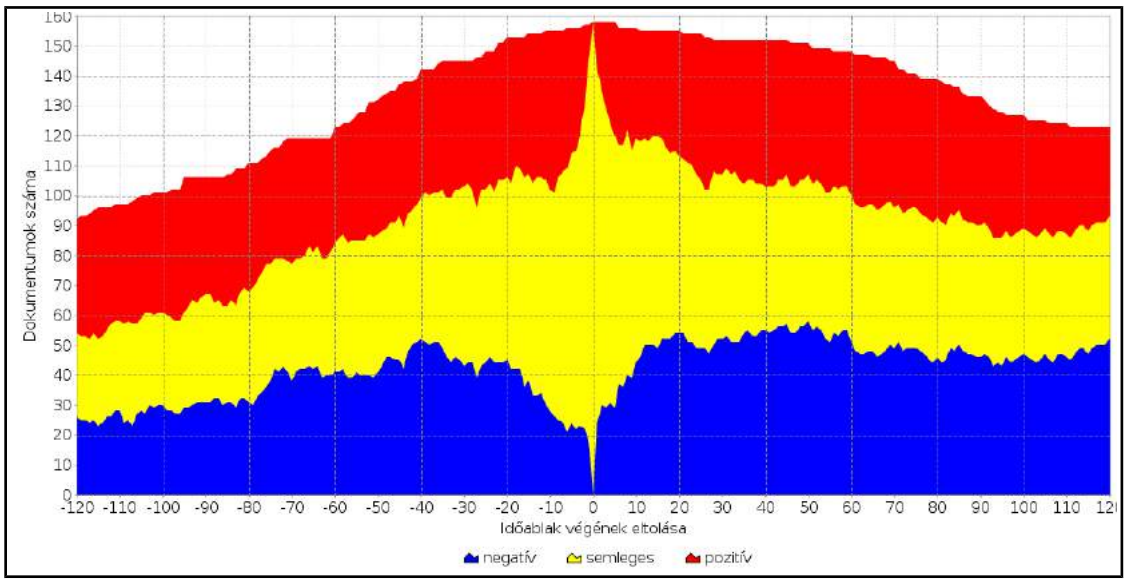
A saját modellem esetén a vizsgált időszakban[80] a kezdetben rendelkezésre álló megfigyelések száma eleve csekély volt, 330 körüli, és a minta méretének megfelelő szinten tartását tartottam szem előtt a szűrőfeltételek kialakításakor. Mivel a leglikvidebb, Prémium kategóriába tartozó magyar részvényeket elemzem, nem szűkítettem a mintát a vállalatok piaci teljesítménye alapján. 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 redukálódik. A *3.3.2. alfejezet* végén említettem, hogy a diszkretizálás függ az alkalmazott időablak hosszától, és ezért a szűrőfeltételek bemutatása szükséges a diszkretizálás eredményének megértéséhez. A következőkben ezt fogom bemutatni.

Rövidebb időtávon nagyobb az esélye annak, hogy az árfolyam változatlan marad, mint hosszú távon, ezért az időablakok hosszának változtatása miatt a hozamkategóriák definiálásához használt alsó és felső korlátok is 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[81]. Az előbbiekből következik, hogy minél rövidebb időablakot veszünk, egyrészt annál több dokumentum szerepel a korpuszban, másrészt annál nagyobb mértékben meghaladja az egyharmadot a semleges kategóriájú dokumentumok aránya.

79 Groth és Muntermann (2008) tanulmányában a mintaméret elég csekély, kb. 60 megfigyelés volt, de a többi tanulmányban 350 fölöttiek az ő mintáik is. Azért ennyivel kisebb az ő mintájuk, mert a többi tanulmányhoz képest a kisebb német piacon lévő részvényeket elemezték.

80 Emlékeztetőül: 2015.04.27. és 2015.07.24.

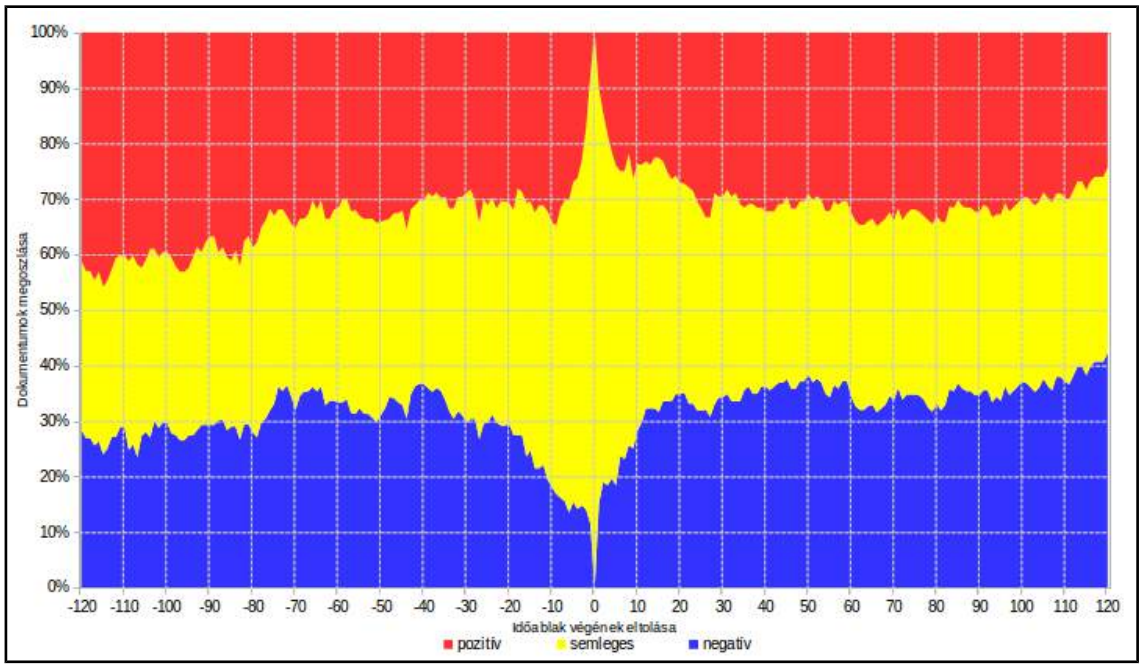
81 Ezt azonban még követi egy lépés, melynek során a mintából kizárásra kerülnek azok a közlemények, amelyek szövegét nem lehetett kinyerni a közzétett PDF-állományból, így a végső korpuszban egyharmad alatti lehet a semleges kategóriájú címkék aránya.



8. Ábra: A korpuszban lévő dokumentumok számának változása kategóriánként az időablak eltolásának függvényében

Forrás: saját szerkesztés

A *8. ábrán* megfigyelhető, hogy a korpusz mérete 92 és 158 megfigyelés között változik. A nulla perces eltolástól eltekintve – ahol 100% lenne – 81% és 30% közötti a semleges kategória aránya.



9. Ábra: A korpuszban lévő dokumentumok kategóriák szerinti megoszlásának változása az időablak eltolásának függvényében

Forrás: saját szerkesztés

A 8. ábránál jobban mutatja a dokumentumok kategóriák szerinti megoszlását a 9. ábra, amelyen látszik, hogy a hosszabb időablakoknál kevésbé egyenletes az eloszlás. A hozamkategóriák megoszlása befolyásolja a modell által elért pontosság értékelését is, hiszen a legnagyobb részarány lesz a default pontosság, másrészt a döntési határ becslésének bizonytalanságára is hat, ezért célszerű a tanítómintában a kategóriák részarányát egyenletesen tartani.

3.4. Árfolyam-modellek szöveges inputtal

Ebben az alfejezetben főleg a szöveges inputtal működő osztályozásról lesz szó. Először a tőzsdei hírbányászati irodalomban fellelhető megvalósításokat foglalom össze a *9. táblázatban*, majd külön alfejezetekben bemutatom az osztályozási modelleket, részletesebben kitérek az SVM-osztályozó működésére, körbejárom a hatékonyság mérésének kérdéskörét, a modellértékelés és -kiválasztás elveit, valamint a robusztusság, illetve az érzékenység vizsgálatának leggyakoribb módszereit. Az alfejezeteket – a korábbiakhoz hasonlóan – a saját modellnél alkalmazott módszerek, mutatók, tesztelési eljárások leírása zárja.

A *9. táblázatban* láthatók a tőzsdei hírbányászatban alkalmazott algoritmusok, validációs eljárások, hatékonyságmutatók és teszteredmények. A táblázatban használt rövidítések feloldása, illetve a módszerek felsorolása a következőkben olvasható.

FT: feladattípus. **o**: osztályozás; **r**: regresszió

A: algoritmus. A táblázatban a fő algoritmus áll az első helyen, mögötte zárójelben további algoritmusok, ha voltak.

- **ann**: artificial neural network, mesterséges neurális hálózat
- **km**: k-means, k-közép osztályozó
- **knn**: k-nearest neighbour, k-legközelebbi szomszéd
- **me**: maximum entrópia osztályozó
- **nB**: naive-Bayes osztályozó
- **regr**: lineáris regresszió
- **R**: Rocchio-osztályozó
- **svm**: support vector machine, támasztóvektor-gép osztályozó
- **svr**: support vector regression, támasztóvektor-gép regresszió
- **rb**: rule based, szabályalapú rendszer

9. Táblázat: A tőzsdei hírbányászati módszerek

Modell	Cikk	FT	A	HM	VE	TM (%)	H	B
Wüthrich	(Leung Kung Fan 1997)	o	rb	1. a 2. mre	v(v, i)	a) 40 b) 20	1b) 48,8% 2a) 35,6%	–
	(Wüthrich, Cho, et al. 1998) (Wüthrich, Permunetilleke, et al. 1998)		rb (knn, ann, regr)	a	v(i)	37,5	46,7% (53%, 43,9%, <40%)	–
	(Cho & Wüthrich 1999) (Cho et al. 1999)		rb		5kv(i)	20	51%	d34,6% e33,3%
	(Peramunetilleke 1997)		rb		v	50	51%	e33%
Lavrenko	(Lavrenko et al. 2000a; 2000b)	o	nB (knn)	det	10kv(v)	10	?	e<
	(Fung et al. 2002a; 2002b)		svm	roc	10kv(v)	10	?	e<
	(Fung 2003) (Fung et al. 2003; 2005)		svm (nB, knn)	1. f1 (p, r) 2. a	v(i)	a) 30 b) 50–10	1a) 85,7% 1b) 82,4% 2?) 67,4%	–
Schumaker	(Schumaker & Chen 2006)	r	svr	1. mse 2. a	10kv	10	1) 0,03346 2) 50,7%	1) r0,0724 2) r47,6%
	(Schumaker & Chen 2009)						1) 0,04261 2) 57,1%	1) r0,0721 2) r54,6%
	(Schumaker 2009)						1) 0,0341 2) 51,4%	–
	(Schumaker et al. 2009; 2012)						1) 0,0516 2) 59%	–
	(Schumaker & Chen 2008; 2010) (Schumaker 2010a; 2010b)			–			–	–
Groth	(Groth & Muntermann 2008)	o	svm	1. a 2. f1 (p, r)	10kv	10	1) 70%	1) d57,5%
	(Groth & Muntermann 2009)						1) 56,5%	1) d60,8%
	(Groth & Muntermann 2010)						1) 78,49% 2) 56,27%	1) e75%
	(Groth 2010)						1) 78,72%	1) e75%
	(Groth et al. 2014)						1) 81,69% 2) 59,16%	1) e75%
	(Groth & Muntermann 2011)		svm (nB, knn, ann)	1. a 2.f1 3. auc			1) 78,96% 2) 53,52% 3) 0,703	1) e75% 3) e0,5
Thomas	(Thomas & Sycara 2000)	o	me	–	v(i)	79,4	–	–
Gidófalvi	(Gidófalvi 2001)	o	nB	1. a 2. p, r	v(i)	27,7–21,8	?	?
	(Gidófalvi & Elkan 2003)					41,4–44,8	1) kb. 47%	1) e50%
Koppel	(Koppel & Shtrimberg 2004)	o	svm	1. a 2. p, r	10kv, v(i)	10 ?	1) 70,3%	1) e50%
	(Généreux et al. 2008; 2011)				10kv	10	1) 69,5%	1) e50%
Mitter-mayer	(Mittermayer 2004)	o	svm	1. r 2. p	50kv	90,9	1) 60%	1) e33%
	(Mittermayer & Knolmayer 2006a)		svm (R, knn)	1. f1 2. a	10kv	10	1) 69% 2) 83%	1) e33% 2) e33%
e-Markets Group	(Zhang et al. 2005)	o	km	–	–	–	–	–
	(Zhang et al. 2007)			a	?	?	64,1%	?
	(Zhang et al. 2006)		rb	–	–	–	–	–
	(Yu et al. 2005; 2006)		svm	a	?	?	65,73%	?

Forrás: saját szerkesztés

HM: hatékonyságmutató. Az osztályozás vagy regresszió hatékonyságát mérő mutatószám neve. Ha többet is használtak, az egyes mutatók számozva követik egymást a táblázatban.

- **a**: accuracy, pontosság
- **auc**: area under the curve, ROC-görbe alatti terület
- **det**: detection error tradeoff, DET-görbe
- **f1**: F1-mutató, a precizitás és a felidézés harmonikus átlaga
- **mre**: mean relative error, átlagos relatív hiba
- **mse**: mean square error, átlagos négyzetes hiba
- **p**: precizitás
- **r**: recall, felidézés
- **roc**: receiver operating characteristic, ROC-görbe

VE: validációs eljárás. Az osztályozó vagy regresszió paramétereinek becslésére használt megfigyelések és a hatékonyság mérésére szolgáló megfigyelések kiválasztásának módja.

- **v(x)**, illetve **#kv(x)**: egyszerű validáció, illetve keresztvalidáció. # helyére egy szám kerülhet, amely azt jelzi, hogy hányszoros keresztvalidációt végeztek a szerzők, ha x ismert, akkor helyére kerülhet:
  - **v** = véletlenszerű
  - **i** = időbeli

TM: tesztminta részaránya. Ha többféleképpen osztották meg a mintát tanító és tesztállományra, akkor vagy tartományt adtam meg, vagy ha az eredmények kapcsán külön hivatkozok az egyikre, akkor betűjellel jelöltem meg.

H: hatékonyság. Az érték előtt a hatékonyságmutató sorszáma, illetve a tesztminta méretének betűjele szerepelhet a táblázatban. Zárójelben az algoritmusoknál felsorolt egyéb módszerek hatékonysága látható, ha vannak ilyenek.

B: benchmark. A hatékonyságmutató értékeléséhez használt viszonyítási alap. A benchmark típusának rövidítése után a benchmark értéke szerepel, és ezek előtt a hozzá

tartozó hatékonyságmutató sorszáma és a tesztminta méretének betűjele szerepelhet. A benchmarkok rövidítései a következők:

- **e**: elméleti kategóriamegoszlás szerint várható találati arány
- **d**: default, mintabeli kategóriamegoszlás szerint várható találati arány
- **r**: szöveges input nélküli regressziós modell teljesítménye ugyanazon a mintán

3.4.1. Feladattípusok és algoritmusok

A módszerek áttekintéséhez, az osztályozás és a regresszió közötti különbségek tisztázásához elevenítsük fel a különböző adattípusok tulajdonságait Tan et al. (2011) alapján! A változókat[82] jellemezhetjük az általuk felvett értékekkel végezhető műveletek szempontjából, ezt skálatípusnak nevezzük. Skálatípusok tekintetében megkülönböztethetünk minőségi, ordinális, intervallum- és arányskálát. A skálatípusok ebben a sorrendben rendelkeznek az összes őket megelőző skálatípus tulajdonságaival. A minőségi, vagy másként nominális skála értékei között csak olyan összehasonlítás végezhető, amely során eldönthető, hogy a két érték azonos vagy különböző. Ordinális, vagy másként sorrendi skála esetén két értékről az is eldönthető, hogy melyik a nagyobb, azaz a két érték sorrendje is megállapítható. Az intervallum-, vagy másként különbségi skála esetén két érték között távolságot lehet számítani, azaz értelmezett a két érték különbsége. Az arány-, vagy másként hányadosskála két értékének arányát is meg lehet állapítani, azaz értelmezett az értékek hányadosa. A változók jellemzésének fontos szempontja még az értékkészlet is, azaz az a halmaz, amelyből értéket vehet fel a változó. Szorosan kapcsolódik a változók skálatípusához a változók diszkrét vagy folytonos volta. Egy attribútumot diszkrétnek nevezünk, ha értékei megszámlálhatók, ezzel szemben a folytonos változók megszámlálhatatlan halmazból vehetnek fel értéket. Jellemzően a nominális és ordinális skálával rendelkező változók diszkrétek, az intervallum- és arányskálán mért változók általában folytonosak. Az értékkészlet megadható az elemek felsorolásával, ha véges sok eleme van a halmaznak. Ez jellemzően nominális és ordinális skálatípusok esetén fordul elő. Numerikus attribútumok esetén az értékkészlet egy intervallummal is jellemezhető, azaz az értékek alsó és felső korlátjának megadásával, de természetesen nem korlátos intervallumok is megengedettek. Intervallum- és arányskálák esetén célszerű ilyen módon megadni az értékkészletet.

82 Gyakori szinonimái: attribútum, tulajdonság.

Az osztályozás és a regresszió is prediktív feladattípus, mindkettő esetén bizonyos inputváltozók[83] függvényeként írjuk fel az outputváltozót[84]. Közöttük különbséget teszünk az outputváltozó diszkrét, illetve folytonos volta szerint. A regresszió során egy folytonos attribútum értékeit kell megbecsülni az inputváltozók értéke alapján. Osztályozási feladat esetén egy diszkrét attribútum értékét becsüljük meg. (Tan et al. 2011) A szakirodalom ezt az attribútumot, illetve annak értékeit kategóriáknak, osztályoknak, illetve címkéknek nevezi. A vizsgált objektumok egyéb tulajdonságaiból következtethetünk a címke értékére, az osztályozást emiatt címkézésnek is nevezik. Vegyünk a minőség-ellenőrzés területéről egy példát, amelyben egy termék átvizsgálása után eldöntik, hogy hibátlan, javításra szorul, vagy ki kell dobni. Ebben az esetben az osztályozandó objektumok a termékek, amelyek fontos tulajdonságainak értékeit átvizsgálás során megállapítják, majd ezek alapján a három osztály valamelyikébe sorolják a terméket. E feladatot el lehet végeztetni valamilyen fizikai vagy kémiai kölcsönhatás segítségével, biológiai úton, élőlényekkel, de munkásokkal, vagy algoritmusok és mesterséges intelligencia alkalmazásával is. A probléma számítógépes megoldására különböző módszerek állnak rendelkezésre. E módszerek közé sorolható a diszkriminancia analízis (Mai 2013), a logisztikus regresszió (Bewick et al. 2005), a naiv Bayes-osztályozó (Kaur & Neelam 2014), a k-legközelebbi szomszéd (Kataria & Singh 2013), a döntési fa (Kotsiantis 2013), a szabályalapú rendszerek (Kliegr & Kuchař 2015), a mesterséges neurális hálózat (Neocleous & Schizas 2002) és az SVM (Burges 1998) módszerek. A módszerek közötti választást befolyásolja, hogy milyen skálatípusú inputváltozókkal dolgozunk, hány értékkel rendelkezik az outputváltozó, a címke értékei lineárisan szeparálhatók-e, a teljes adathalmaz helyes címkézésére törekszünk-e vagy csak a legnagyobb megbízhatóságú becslések fontosak, mekkora számítási kapacitás áll rendelkezésre a feladat elvégzéséhez és ezzel párhuzamosan mekkora adathalmazt kell felcímkézni és mekkora az algoritmus futásideje. A nem numerikus inputváltozók értékét közvetlenül képesek kezelni a döntési fa és szabályalapú módszerek, de ugyanez nem igaz például az SVM vagy a mesterséges neurális hálózat típusú osztályozókra. Nominális attribútumok numerikussá való átalakítására van lehetőség, például azáltal, hogy dummy változókat vezetünk be. Ez lehetővé teszi, hogy a kizárólag numerikus inputváltozókat megkövetelő módszerek alkalmazhatók legyenek nominális változókra is. A *9. táblázatban* jól látható, hogy a tőzsdei hírbányászat során elsősorban osztályozási feladatként mo-

83 Gyakori szinonimák: magyarázó változó, független változó.
84 Gyakori szinonimák: magyarázott változó, függőváltozó, eredményváltozó, célváltozó.

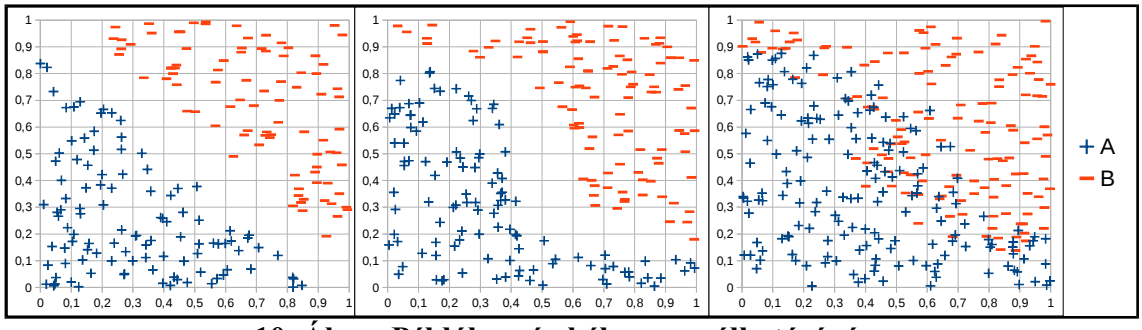
dellezik az árfolyamokat, csupán a Schumaker-modell alkalmazott regressziót. A legelterjedtebb algoritmus az SVM, annak is elsősorban a lineáris kernelű változata. Ezt használta Fung a Lavrenko-modell továbbfejlesztett változatában, Groth, Koppel, Mittermayer, illetve az eMarkets Group tagjai, Yu és szerzőtársai (lásd *9. táblázat*). A Schumaker-modell pedig ennek az algoritmusnak a regressziós változatával, az SVR-rel becsülte meg a részvények árfolyamát. A Wüthrich-féle modell szabályalapú rendszert alkalmazott, Lavrenko és Gidófalvi pedig naiv-Bayes osztályozót. (Lásd *9. táblázat.*)

Az outputváltozó lehetséges értékeinek száma alapján beszélhetünk kétcímkés – vagy másként bináris – és többcímkés osztályozásról. Bináris osztályozás esetében az objektumokat két osztály valamelyikébe soroljuk be, például: férfi-nő, igen-nem[85], növekszik-csökken[86]. Többcímkés osztályozás esetén az objektumokat három vagy több kategóriába kell besorolni, például: nő-nem változik-csökken[87], pozitív-semleges-negatív[88], doji-marubozu-hammer-shooting star[89] stb. A tőzsdei hírbányászatban mindkettőre találunk példát, Koppel modellje például elsősorban csak a pozitív és negatív hírek címkézését végzi[90], Groth-nál is találkozhatunk hasonló megközelítéssel, de az ő modelljei főleg az eseménytípusú bináris outputváltozókat[91] használták, ezen kívül Thomas és az eMarkets Groupból Zhang és szerzőtársai végeztek még bináris osztályzást. Wüthrich, Gidófalvi, Lavrenko, Mittermayer, illetve az eMarkets Groupos Yu és szerzőtársai pedig többcímkés osztályozást végeztek. (Lásd *7. táblázat*.)

Az osztályozási feladat megoldhatósága függ a címke értékeinek szeparálhatóságától. Az n darab inputváltozó terében a megfigyelések tökéletesen lineárisan szeparálhatók, ha minden címkéhez létezik legalább egy olyan $n-1$ dimenziós hipersík – a szeparálósík –, hogy e hipersík által határolt egyik féltérben a benne található összes megfigyelés címkéje azonos, de a másik féltérben található összes megfigyelés címkéjétől különböző. A gyakorlati alkalmazások esetén gyakran előfordul, hogy nem található minden címkéhez ilyen szeparálósík. Ennek az lehet az oka, hogy a különböző címkéjű megfigyelések az inputtérnek legalább egy részén *keverednek* egymással, vagy az, hogy a megfigyelések csak nemlineáris módon szeparálhatók egymástól. A *10. ábra* két folytonos inputváltozó terében szemlélteti az előbb leírtakat kétcímkés – A és B – osztályozási

85 csődelőrejelzés, sikeres ügyfélelérés, ügyfélelvándorlás stb.
86 árfolyam-, bevétel-előrejelzés stb.
87 árfolyam
88 véleménynyilvánítás
89 árfolyam-gyertya
90 Vizsgálták a semleges kategória bevezetésének hatását is.
91 Extrém magas volatilitás, illetve extrém alacsony CRT volt-e a vizsgált időszakban.

probléma esetén. Az ábra első része egy lineárisan szeparálható esetet mutat, a második rész a nemlineáris szeparációra mutat példát, a harmadik pedig olyan adathalmazt ábrázol, amely legfeljebb nagyon bonyolult nemlineáris módon szétválasztható, *összekevert* címkéjű megfigyeléseket tartalmaz. Amennyiben legalább két megfigyelésnek az input-változói rendre ugyanazt az értéket veszik fel, ugyanakkor a címkéjük különböző, akkor a megfigyelések nem szeparálhatók tökéletesen. (Elizondo et al. 2012)



10. Ábra: Példák a címkék szeparálhatóságára

Forrás: saját szerkesztés

Elsősorban kevert címkéjű osztályozás esetén van jelentősége annak, hogy az osztályozási feladatot megfogalmazhatjuk éles vagy puha szemléletben is. Éles osztályozás esetében egy megfigyelést pontosan egy osztályba sorolunk. Puha osztályozáskor minden kategória esetén megadjuk, hogy a megfigyelés mekkora megbízhatósággal tartozik az adott kategóriába. A mesterséges neurális hálózatok, a naiv Bayes- vagy a döntési fa módszerek is alapvetően puha osztályozásra képesek, és az általuk szolgáltatott megbízhatósági mértékek, valószínűségek alapján éles osztályozássá alakítható a becslésük. A kétféle szemléletnek az osztályozó hatékonyságának megítélésekor is jelentősége van. Éles esetben vagy a teljes adathalmaz helyes címkézésére törekszünk, vagy az egyik kategória összes megfigyelésének helyes címkézése a cél. Puha osztályozáskor a becslések kategóriánként sorrendezhetők megbízhatóság alapján, és csak a legnagyobb megbízhatóságú becslések alapján ítéljük meg az osztályozás hatékonyságát, tehát nem a kategória összes megfigyelését vesszük alapul. Ez például abban az esetben lehet kívánatos, ha üzleti szempontból nem éri meg az összes megfigyelt objektummal kapcsolatban erőforrásokat felhasználni, csak a legjövedelmezőbbnek ígérkezőknél. (Hernández-Orallo et al. 2012) A tőzsdei hírbányászatban az osztályozás outputját az üzleti döntés előtt mindenképpen élessé kell alakítani, így ez a jellemző megközelítés az irodalomban is, de erre még az osztályozás hatékonyságának méréséről szóló alfejezetben visszatérünk.

A saját modellem feladata, hogy a háromértékű diszkrét outputváltozót – pozitív, negatív, semleges árfolyammozgás – a sajtóközlemények reprezentálásához használt szövegjellemzők mint inputok alapján megbecsülje, tehát többcímkés osztályozási feladattípusról van szó. Ez a szakirodalomban legelterjedtebb megközelítés, amelyre a leggyakrabban alkalmazott algoritmus az SVM osztályozó volt, így a saját modellemben is ezt alkalmazom, amit még az is indokol, hogy feltételezhető, hogy a probléma lineárisan nem szeparálható tökéletesen, illetve a modell paraméterezhető, ugyanakkor becslése viszonylag gyors – ellentétben például a nem paraméterezhető naiv Bayes-osztályozóval, vagy a nagy számításigényű többrétegű perceptron – neurális hálózat – osztályozóval. A tanítást RapidMinerben a *Support Vector Machine (LibSVM)* operátor segítségével végeztem el különböző C-gamma kombinációkkal. A továbbiakban az SVM osztályozó működését mutatom be Tan et al. (2011) alapján.

3.4.1.1. Az SVM osztályozómódszer

A magyarul tartóvektor-gépnek fordítható Support Vector Machine (SVM) osztályozó módszer folytonos skálájú inputváltozók terében becsül döntési határokat. Az SVM által becsült szeparálósíkkal párhuzamosan, annak mindkét oldalán egy-egy további hipersík helyezkedik el, melyek közötti távolságot a szeparálósík margójának nevezzük. Az SVM nem kizárólag az osztályozás hibájának minimalizálására törekszik, hanem a margó maximalizálására is. Erre a problémára egy feltételes szélsőértékfeladat felírásával található megoldás, amelyben a célfüggvény a margó méretétől függ, a korlátozó feltételek pedig a megfigyelések helyes osztályozását írják elő. Nem szeparálható esetre a feltételekbe kiegészítő változók kerülnek bevezetésre, amelyek az eredeti feltételek megsértése esetén a megsértés nagyságának megfelelő értéket veszik fel. A kiegészítőváltozóknak a célfüggvényben olyan együtthatójuk van, amely rontja a célfüggvény értékét, ha a kiegészítőváltozó pozitív értéket vesz fel. Ez az együttható az SVM büntetésnek nevezett paramétere, amelyet C-vel jelölnek. Nemlineáris szeparációk esetén attribútumtranszformáció hajtható végre, azonban ez az attribútumok számának jelentős növekedését okozhatja. Ennek elkerülésére az SVM speciális transzformációs függvényeket alkalmaz, amelyeknek a legfőbb tulajdonsága az, hogy az általuk előállított transzformált térbeli vektorok belső szorzata – amire az optimum kiszámításához, illetve a megfigyelések osztályba sorolásához van szükség – meghatározható a vektorok eredeti attribútumtérbeli megfelelőinek hasonlóságából – például belső szorzatából mint hasonlósági mértékből. E meghatározáshoz használt függvényt kernelfüggvénynek nevezik.

Lássuk a fent leírtakat formalizálva (Tan et al. 2011) alapján.[92] Az inputváltozókat jelöljük rendre $x:_{1,}x_{2,}\ldots x_I$ módon, ahol I az inputváltozók száma. Az outputváltozót jelöljük y-nal, amely bináris, és értéke $+1$, ha egy megfigyelés a vizsgált osztályba tartozik, és -1, ha nem. A $+1$ és -1 címkéjű megfigyeléseket elválasztó hipersík egyenlete a következő:

$$\underline{w}\,\underline{x}+b=0 \tag{19}$$

Ahol:

$\underline{x}=[x_1, x_2, \ldots, x_I]$ az inputváltozók vektora

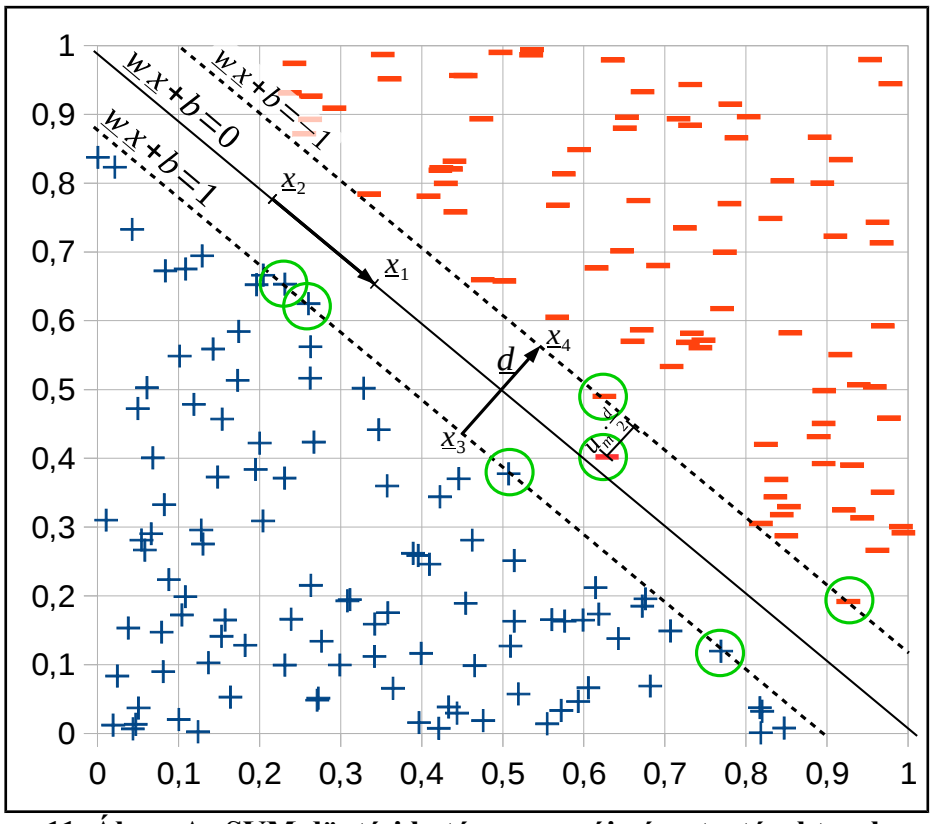
$\underline{w}=[w_1, w_2, \ldots, w_I]$ az inputváltozók együtthatóinak vektora

b konstans

Az optimális döntési határ meghatározása $\underline{w}$ és b paraméterek megfelelő megválasztásával érhető el. A döntési határ két oldalán elhelyezkedő egy-egy hipersík egyenlete a következő:

$$\underline{w}\,\underline{x}+b=1 \tag{20}$$

$$\underline{w}\,\underline{x}+b=-1 \tag{21}$$



11. Ábra: Az SVM döntési határa, margója és a tartóvektorok

Forrás: saját szerkesztés (Tan et al. 2011) alapján

92 A teljes fejezet ugyanezen forrás Tartóvektor-gép (SVM) című fejezetének kivonatolt változata, csupán a jelöléseket igazítottam a dolgozathoz. Egyéb forrás hasonló levezetéssel: (Burges 1998).

Az első hipersík a $+1$ címkéjű megfigyelések felől található, a második a -1 címkéjűek felől. Tökéletesen szeparálható esetben nincsenek közöttük megfigyelések. Tartóvektoroknak nevezik azokat a megfigyeléseket, amelyek kielégítik valamelyik margószélhez tartozó egyenletet. A két hipersík közötti távolság a margó, amelyet jelöljünk d -vel. A margó kifejezhető $\underline{w}$ hosszának függvényében. A $\underline{w}$ vektor iránya merőleges mindhárom hipersíkra. Legyen a $\underline{d}$ vektor a $\underline{w}$ -vel párhuzamos és legyen kezdőpontja a pozitív, végpontja a negatív oldalon lévő margószélhez tartozó hipersíkon, hossza éppen d . A $\underline{w}$ vektor merőlegességének belátásához vegyünk egy $\underline{x}_1$ és egy ettől különböző $\underline{x}_2$ vektort, amelyek a szeparálósíkon helyezkednek el, azaz:

$$\underline{w}\,\underline{x}_1+b=0 \tag{22}$$

$$\underline{w}\,\underline{x}_2+b=0 \tag{23}$$

Vonjuk ki egymásból a két egyenletet:

$$\underline{w}(\underline{x}_1-\underline{x}_2)=0 \tag{24}$$

Ekkor $\underline{w}$ és a szeparálósíkon fekvő $(\underline{x}_1-\underline{x}_2)$ vektor szorzatára nullát kapunk, tehát ortogonálisak. Ezzel beláttuk $\underline{w}$ vektor merőlegességét. A margó hosszának felírásához vegyük $\underline{d}$ vektor kezdőpontját, $\underline{x}_3$ -at, és végpontját, $\underline{x}_4$ -et, amelyek a megfelelő hipersíkokon találhatók:

$$\underline{w}\,\underline{x}_3+b=1 \tag{25}$$

$$\underline{w}\,\underline{x}_4+b=-1 \tag{26}$$

A két egyenletet kivonva egymásból azt kapjuk, hogy:

$$\underline{w}(\underline{x}_3-\underline{x}_4)=2 \tag{27}$$

Ebből mivel $\underline{x}_3+\underline{d}=\underline{x}_4$, azaz $\underline{x}_3-\underline{x}_4=\underline{d}$, ezért $\underline{w}\,\underline{d}=2$, amit átírhatunk a következő alakba:

$$|\underline{w}|\,d\cos 0=2 \tag{28}$$

Azaz d értéke a következőképpen függ $\underline{w}$ hosszától:

$$d=\frac{2}{|\underline{w}|} \tag{29}$$

Egy megfigyelés osztályozásához helyettesítsük be a koordinátáit a szeparálósík egyenletébe, majd ha az eredmény pozitív, akkor a pozitív kategóriába soroljuk, ha negatív, akkor a negatív kategóriába. Ezért, hogy jól elkülönüljenek egymástól az osztályok, minden megfigyelésre előírjuk, hogy a címkéjének megfelelő hipersíknak – más

szóval margószélnek – a döntési határtól távolabb eső felére essen, ám ha ez nem teljesül, egy u_m kiegészítőváltozóval mérjük a feltétel megsértését:

$$\underline{w}\,\underline{x}_m + b + u_m \geq 1 \text{ , ha } y_m = 1 \tag{30}$$

$$\underline{w}\,\underline{x}_m + b - u_m \leq -1 \text{ , ha } y_m = -1 \tag{31}$$

Ami bal oldalra rendezés és mindkét oldal y_m -mel való szorzása után leírható egyetlen egyenlőtlenséggel:

$$y_m(\underline{w}\,\underline{x}_m + b) + u_m - 1 \geq 0 \tag{32}$$

Ahol:

$m \in \{1,2,\ldots,M\}$ a megfigyelés sorszáma, M a megfigyelések száma

$\underline{x}_m$ a megfigyelés inputváltozóinak vektora

y_m a megfigyelés outputváltozója

$u_m \geq 0$ a megfigyelés osztályozásának hibája

A 11. ábra a fentieket szemlélteti egy kétdimenziós esetben, nem optimális helyzetű döntési határral – hogy egy nem nulla értékű kiegészítőváltozó is szerepeljen az ábrán. Zölddel bekarikázva a tartóvektorok láthatók[93].

A célfüggvény, amely egyszerre maximalizálja d -t és minimalizálja az osztályozás hibáját az összes megfigyelésre:

$$\min_{w,u} \; f(\underline{w},\underline{u}) = \frac{|w|^2}{2} + C\,\underline{u}\,\underline{1} \tag{33}$$

Ahol:

$\underline{u} = [u_1, u_2, \ldots, u_M]$ az eltérésváltozók vektora

$\underline{1}$ összegzővektor

$C \geq 0$ a büntetőparaméter

A fenti feltételes optimalizálási feladatra felírható a következő Lagrange-függvény:

$$\min_{\underline{w},\underline{u},\underline{\lambda},b} \; L(\underline{w},\underline{u},\underline{\lambda},b) = \frac{|w|^2}{2} + C\,\underline{u}\,\underline{1} - \sum_{m=1}^{M} \lambda_m (y_m(\underline{w}\,\underline{x}_m + b) + u_m - 1) - \underline{\mu}\,\underline{u} \tag{34}$$

Ahol $\underline{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_M]$ és $\underline{\mu} = [\mu_1, \mu_2, \ldots, \mu_M]$ a Lagrange-multiplikátorok vektorai. Továbbá érvényesek a következő Karush-Kuhn-Tucker (KKT) feltételek:

$$\lambda_m \geq 0 \text{ , } \mu_m \geq 0 \text{ , } u_m \geq 0 \tag{35}$$

$$\lambda_m (y_m(\underline{w}\,\underline{x}_m + b) + u_m - 1) = 0 \tag{36}$$

$$\mu_m u_m = 0 \tag{37}$$

93 A kiegészítőváltozós egyenlet a margó rossz oldalára eső megfigyelések esetén egyenlőség formáját ölti, így a megfigyelés támasztóvektor.

A szélsőérték létezésének elsőrendű feltételei:

$$\frac{\partial L}{\partial w_i}=0 \;\Rightarrow\; w_i=\sum_{m=1}^{M}\lambda_m y_m x_{mi} \tag{38}$$

$$\frac{\partial L}{\partial b}=0 \;\Rightarrow\; \sum_{m=1}^{M}\lambda_m y_m=0 \tag{39}$$

$$\frac{\partial L}{\partial u_m}=0 \;\Rightarrow\; \mu_m=C-\lambda_m \tag{40}$$

Ahol x_{mi} az $\underline{x}_m$ vektor i-dik koordinátája.

Ezeket behelyettesítve $L(\underline{w},\underline{u},\underline{\lambda},b)$ -be kapjuk a duális Lagrange-függvényt:

$$\max_{\underline{\lambda}} \; L_D(\underline{\lambda})=\sum_{m=1}^{M}\lambda_m-\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M}\sum_{i=1}^{I}\lambda_m\lambda_n y_m y_n x_{mi}x_{ni}=\underline{\lambda}\,\underline{1}-\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M}\lambda_m\lambda_n y_m y_n \underline{x}_m\underline{x}_n \tag{41}$$

A duális feladat numerikus megoldása során felhasználható, hogy mivel $\mu_m\geq 0$, $\lambda\geq 0$ és $C\geq 0$, ezért a harmadik elsőrendű feltétel alapján $\lambda_m\leq C$ kell, hogy legyen, ami egyszerűbbé teszi a keresést.

A fenti megoldás nemlineáris esetben azzal módosul, hogy az I dimenziójú eredeti $\underline{x}$ inputvektorok helyett magasabb J dimenziójú térbe transzformált vektorokkal kell elvégezni az optimalizálást. A transzformációról feltételezzük, hogy a segítségével kapott vektorok lineárisan szeparálhatók a J dimenziós térben. Ha a transzformációt végző függvényt a következőképpen jelöljük: $\phi(\underline{x})$, akkor a duális Lagrange-feladat a következő alakot ölti:

$$\max_{\underline{\lambda}} \; L_D(\underline{\lambda})=\underline{\lambda}\,\underline{1}-\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{M}\lambda_m\lambda_n y_m y_n \phi(\underline{x}_m)\phi(\underline{x}_n) \tag{42}$$

A λ_m együtthatók ismeretében a J dimenziós súlyvektor elemei optimális esetben:

$$w_j=\sum_{m=1}^{M}\lambda_m y_m \phi(x_m)_j \tag{43}$$

A b paraméter optimális értéke kiszámítható a transzformált térben a következő egyenletből valamely tartóvektor segítségével:

$$\lambda_m\Big(y_m\Big(\sum_{n=1}^{M}\lambda_n y_n \phi(\underline{x}_m)\phi(\underline{x}_n)+b\Big)+u_m-1\Big)=0 \tag{44}$$

Egy tetszőleges megfigyelés osztályozásához pedig a J dimenziós vektorát behelyettesítjük a szeparálósík egyenletének bal oldalába, és a kapott érték előjele alapján a pozitív vagy a negatív kategóriába soroljuk:

$$f(\underline{x}) = sgn(\underline{w}\,\phi(\underline{x}) + b) = sgn\left(\sum_{n=1}^{M} \lambda_n y_n \phi(\underline{x})\phi(\underline{x}_n) + b\right) \tag{45}$$

Ha olyan $\phi(\underline{x})$ transzformációs függvényt választunk, hogy létezik olyan $K(\underline{x}_1, \underline{x}_2)$ hasonlósági függvény, amelynek kiszámítása az eredeti, alacsonyabb I dimenziószámú attribútumtérben történik, és $K(\underline{x}_1, \underline{x}_2) = \phi(\underline{x}_1)\phi(\underline{x}_2)$, akkor $\underline{w}$, b és $f(\underline{x})$ kiszámítása jelentősen gyorsítható. Ez kerneltrükként ismeretes a szakirodalomban. Az egyik legnépszerűbb kernel a radiális bázisfüggvény, röviden rbf, amelynek képlete a következő:

$$K_{rbf}(\underline{x}_1, \underline{x}_2) = e^{-\frac{|\underline{x}_1 - \underline{x}_2|^2}{2\sigma^2}} = e^{-\gamma|\underline{x}_1 - \underline{x}_2|^2} \tag{46}$$

Bizonyos megvalósításokban a radiális bázisfüggvény sugarát a szigma paraméteren keresztül lehet befolyásolni, másokban a gamma paraméteren keresztül, melyek között a következő összefüggés áll fenn: $\gamma = \frac{1}{2\sigma^2}$. A szigma paraméter növelésével a sugár növekszik, a gamma növelésével pedig csökken. Mivel csak a tartóvektorokhoz tartozhatnak nullától különböző multiplikátorok – a KKT-feltételek miatt –, ezért a fenti kernelt csak a tartóvektorok esetén kell kiértékelni. Ha a radiális bázisfüggvények sugara jelentősen meghaladja a tartóvektorok távolságát, akkor lineáris osztályozóként viselkedik az SVM, ha pedig elenyésző a sugár, akkor legközelebbiszomszéd-osztályozóként. Tehát a gamma – vagy szigma – paraméter értékének megválasztása függ az éppen vizsgált adathalmaz tulajdonságaitól is – például az adatpontok sűrűségétől. Azt, hogy az rbf-kernellel rendelkező SVM a két szélsőséges eset közötti gamma paraméterek esetén mennyire viselkedik lineáris osztályozóként, illetve durva vagy egyre inkább finom döntési határokat képező nemlineáris osztályozóként, a C paraméter értékén keresztül lehet befolyásolni. Minél magasabb a C büntetés, a margó rossz oldalán lévő, így büntetendő tartóvektorokból a margóra eső tartóvektorok, illetve közönséges, jó oldalra eső megfigyelések válnak – persze csak egy bizonyos fokig. A nagyobb büntetés hatására a szeparálósík közelebb húzódik az egyre kevesebb tartóvektorhoz, így azok közelében a döntési határ finomabb lehet lokálisan, mivel görbületét kevesebb távol eső tartóvektor befolyásolja. Az rbf-kernelű SVM gamma és C paramétere közötti összefüggésre mutat példát a *13. ábrán* látható hőtérkép, melynek értelmezéséhez néhány fogalom bevezetése szükséges, amire a *3.4.2-es alfejezetben* kerül sor.

3.4.2. Az osztályozás hatékonyságának mérése

Az osztályozási feladat ritkán oldható meg tökéletesen, amikor is minden megfigyelés egyértelműen besorolható abba az osztályba, amelyikbe tartozik. Ennek többféle oka is lehet, mint például a kategóriák eloszlása az inputok terében, a kategóriák keveredése, a döntési határok alakja, az osztályozó módszer típusa, paraméterei, időigénye. Nemlineáris döntési határok esetén a lineáris osztályozó módszerek – például SVM lineáris kernellel, egyszerű perceptron – nem képesek tökéletesen szétválasztani a kategóriákat. Konkáv, illetve több kisebb, egymástól távol eső klaszterre tagolt térrészeket elfoglaló osztályokkal nehezen boldogul az egyébként nemlineáris osztályozó is, ha rosszul paraméterezik – például az SVM radiális bázisfüggvény-kernellel túl kicsi C és gamma paraméterekkel beállítva, vagy a többrétegű perceptron, ha túl szűk vagy túl kevés rejtett rétege van. Az inputváltozók tengelyével nem párhuzamos döntési határok a szabályalapú és döntési fa módszerek esetén megnövelik a szükséges szabályok, illetve elágazások számát és ezzel együtt a futásidőt. A különböző osztályok által elfoglalt térrészek átfedése esetén azonban nem valószínű, hogy bármilyen modell képes túltanulás nélkül tökéletesen szétválasztani a kategóriákat egymástól. Az osztályozás jóságának mértéke annál nagyobb, minél több releváns megfigyelést sorolunk a valódi kategóriába. (Tan et al. 2011)

Az osztályozás különböző teljesítménymutatói általában a releváns megfigyelés fogalmának tekintetében térnek el, tehát abban, hogy mit tekintenek viszonyítási alapnak. A leggyakoribb mértékek a pontosság (accuracy), precizitás (precision), felidézés (recall), ROC-görbe, illetve az utóbbi alapján számított AUC (area under the curve, görbe alatti terület). A pontosság a helyesen osztályozott megfigyelések aránya a teljes sokaságban. A precizitás a helyesen osztályozott, adott kategóriába tartozó megfigyelések aránya az összes, ugyanazon kategóriába helyesen és helytelenül sorolt megfigyelésen belül. A felidézés a helyesen osztályozott, adott kategóriába tartozó megfigyelések aránya az összes, ugyanabba a kategóriába valójában tartozó megfigyelésen belül. (Sokolova 2006)

$$pontosság = \frac{|H|}{|I|} \tag{47}$$

$$precizitás_c = \frac{|H_c|}{|B_c|} \tag{48}$$

$$felidézés_c = \frac{|H_c|}{|T_c|} \tag{49}$$

Ahol:

$H=(i|p_i=c_i, i\in I)$: a helyesen becsült címkéjű megfigyelések halmaza

p_i : az i jelű megfigyelés becsült címkéje

c_i : az i jelű megfigyelés tényleges címkéje

I : az ismert címkéjű megfigyelések halmaza

$H_c=(i|p_i=c_i, i\in B_c)=(i|p_i=c_i, i\in T_c)$: a helyesen becsült, c címkéjű megfigyelések halmaza

$B_c=(i|p_i=c, i\in I)$: a c címkéjűnek becsült megfigyelések halmaza

$T_c=(i|c_i=c, i\in I)$: a ténylegesen c címkéjű megfigyelések halmaza

Ahogy a *9. táblázatban* látható, a pontosságot mindegyik modellnél alkalmazták, ha nem is minden tanulmányban, vagy nem is elsődleges mutatóként. Alkalmazása mellett szól, hogy könnyen értelmezhető és tetszőleges számú kategóriára számolható, hátránya azonban, hogy a kategóriák mintabeli megoszlásától függően kell értelmezni, és két különböző eloszlású adathalmaz esetén közvetlenül nem összehasonlíthatók a pontosságok. A precizitás és a felidézés ugyancsak kedvelt mutatók, akárcsak a harmonikus átlaguk, amelyet F1 mutatónak neveznek. E mutatók közül alkalmazta valamelyiket Fung, Groth, Gidófalvi, Koppel és Mittermayer is. (Lásd *9. táblázat*.) E mutatók kategóriánként könnyen értelmezhetők, függetlenek a kategória elemszámától, azonban jelentősen különböző kategóriaeloszlás esetén nehézkes őket aggregálni, amire mikro- és makroátlagolást szokták alkalmazni.

A fenti mutatók éles osztályozás kiértékelésére alkalmasak, de ha puha osztályozóval dolgozunk, akkor az osztályba tartozás megbízhatósága, illetve valószínűsége alapján különböző küszöbszintek szerint lehet a megfigyeléseket a kategóriákba besorolni. Ha a küszöb magas, akkor várhatóan alacsony lesz a felidézés, de a küszöb csökkentésével várhatóan egyre több tévesen osztályozott megfigyelést kapunk. A kettő közötti kedvező átváltási viszony megtalálására szeretnénk törekedni a ROC-görbe vizsgálata révén. A ROC-görbe megmutatja bináris osztályozás esetén, hogy hogyan változik a felidézés, azaz a ROC-terminológiában a helyespozitív-arány – TPR, true positive rate –, és a helytelenpozitív-arány – FPR, false positive rate –, amint az osztályozás megbízhatósága szerint csökkenő sorrendbe rendezett megfigyeléseket sorban hozzávesszük a pozitív kategóriába sorolt megfigyelések halmazához. A helytelenpozitív-arány bináris osztályozás esetén mutatja a tévesen az adott – úgy mondjuk, pozitív címkéjű – kategóriába

sorolt megfigyelések arányát a ténylegesen másik – úgy mondjuk, negatív címkéjű – kategóriába tartozó megfigyeléseken belül.

$$TPR = \frac{|H_p|}{|T_p|} \qquad (50) \qquad\qquad FPR = \frac{|R_p|}{|T_n|} \qquad (51)$$

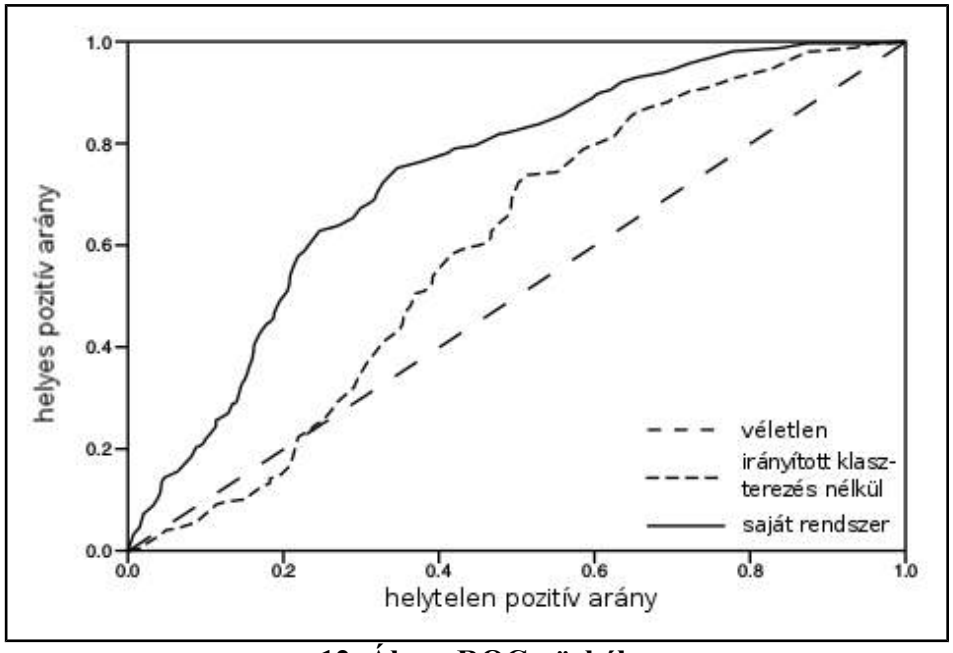
Ahol:

$C = \{p, n\}$: a címkék halmaza

p : a pozitív kategória címkéje

n : a negatív kategória címkéje

$R_c = (i | p_i \neq c_i, i \in B_c)$: a tévesen a c jelű kategóriába sorolt megfigyelések halmaza



12. Ábra: ROC-görbék

Forrás: (Fung et al. 2002b, o.490)

Az AUC a ROC-görbe alatti terület mérőszáma. Ha az osztályozó az összes megfigyelést azonos megbízhatósággal sorolná be a pozitív kategóriába, akkor a TPR és FPR mutatók azonos mértékben, egyenletesen növekednének ahogy mindig hozzáveszünk egy pótlólagos elemet a pozitív címkéjűnek becsült megfigyelések halmazához. Ekkor a TPR-FPR koordináta-rendszerben a ROC-görbe a $(0;0)$ és az $(1;1)$ pontokat összekötő egyenes lenne, és ezért az AUC mutató értéke 0,5 volna. Ha a modell az adatok egy részén nagyobb megbízhatóságot képes elérni, akkor a $(0;0)$ ponttól jobbra haladva a ROC meredeksége kezdetben 0,5-nél nagyobb, majd csökken, és az $(1;1)$ pontban 0,5-nél kisebb. Ideális esetben az origóhoz közel végtelenhez, a másik végpontnál nullához tart a ROC meredeksége, ekkor az AUC megközelíti az 1 értéket. A mutató azt

jelzi, hogy a modell mennyire alkalmas a megfigyelések rangsorolására. (Tan et al. 2011) A tőzsdei hírbányászatban is találunk néhány modellt, amely az AUC, a ROC- vagy a hozzá hasonló DET-görbe alapján jellemezte a modell hatékonyságát, ezek Lavrenko, Fung és Groth egyes tanulmányaiban olvashatók. (Lásd *9. táblázat*.)

A modell hatékonyságát célszerű olyan adathalmazon mérni, amely a tanításhoz felhasznált megfigyelésektől különböző adatokat tartalmaz, ezzel a modell túlillesztése elkerülhető. A modell hatékonyságmutatóinak mérésére szolgáló mintát tesztmintának, a paraméterek becslésére szolgálót pedig tanítómintának nevezzük. Validációnak hívjuk azt az eljárást, amikor a mintát felosztjuk tanító- és tesztmintára, majd a becsült modell teljesítményét megmérjük. A validációt elvégezhetjük úgy, hogy a felosztást nem változtatjuk, vagy úgy is, hogy többféle felosztás esetén is kiértékeljük modellt. A felosztás történhet valamely külső attribútum segítségével – például a megfigyelés időbélyegzője alapján –, vagy véletlenszerűen. A leggyakrabban alkalmazott k-szoros keresztvalidációs eljárás esetén a mintát k részre osztják, majd k-szor tanítják a modellt úgy, hogy minden alkalommal a mintának másik k-ad részét használják teszteléshez, a k darab mutatót átlagolják. (Tan et al. 2011) A tőzsdei hírbányászat legkedveltebb validációs módszere a 10-szeres keresztvalidáció, Thomas, Gidófalvi és az eMarkets Group kivételével mindegyik modellnél találunk példát az alkalmazására. A probléma időbeli jellege miatt Wüthrich, Fung, Thomas, Gidófalvi és Koppel készített olyan validációt is, amelyben a mintát úgy osztották fel, hogy a tanítómintába az időben korábbi megfigyelések kerültek, és az időben későbbieken tesztelték a hatékonyságot. (Lásd *9. táblázat*.)

Egy modell pontosságát gyakran viszonyítjuk egy másik modelléhez, amely benchmarkként szolgál. Osztályozási problémák esetén a legalapvetőbb benchmark a default modell. A default modell a tanítómintában a legnagyobb részaránnyal bíró kategória címkéjét rendeli az összes megfigyeléshez. A default modell pontossága e részaránnyal egyezik meg. A véletlen találgatással a default modell pontosságánál nem várunk magasabbat. Ennek van olyan változata is, amelyben nem a kategóriák mintabeli eloszlását használják, hanem az ismert a sokasági eloszlást. A kettő közül valamelyik benchmarkot Schumaker, Thomas és az eMarkets Group modelljeit kivéve valamennyi modellnél alkalmazták legalább egy esetben. (Lásd *9. táblázat*.) Egyéb benchmarkként lényegében bármilyen modell szolgálhat, amelyet például másik kutató készített, vagy másik osztályozóval, más paraméterekkel, más inputokra tanították.

Két modell hatékonysági mutatószámának összehasonlítása függ az alkalmazott mutatók tulajdonságaitól is. Az AUC mutatók többnyire közvetlenül is összehasonlíthatók volnának, azonban ezt csak kevés esetben adták meg a kutatók. Az F1 mutató esetén nagyságrendileg szintén alkalmazható a közvetlen összehasonlítás, azonban szigorúan véve a mikro- és makroátlagolással kapott eredmények egymással nem összehasonlíthatók – ezek átszámítására sajnos a mintaeloszlás és a részmutatók ismerete nélkül nincs lehetőség. A pontosság esetén közvetlenül csak azonos kategóriaeloszlású mintából származó értékek hasonlíthatók össze, hiszen például két 70%-os pontosságú modell nem egyenértékű, ha egyiknél a default pontosság 50% volt, másiknál pedig 33%. Két azonos keresztvalidáció során tanított modell átlagos pontosságainak összehasonlítására normális eloszlás feltételezése mellett alkalmazhatunk t-próbát. (Tan et al. 2011) Annak érdekében, hogy a különböző kutatók pontosságeredményeit össze lehessen hasonlítani, és meg tudjuk mondani, ki készített jobb minőségű osztályozást, a tanított modellnek a pontosságát a default modelléhez képest célszerű megadni. A *3. függelékben* levezetett q – pontosságminőségi – mutató ezen az elven összehasonlíthatóvá teszi a különböző modellek minőségét, bár azt nem mondja meg, hogy az eltérések szignifikánsak-e.

10. Táblázat: A tőzsdei hírbányászati modellek pontosságminőségi rangsora

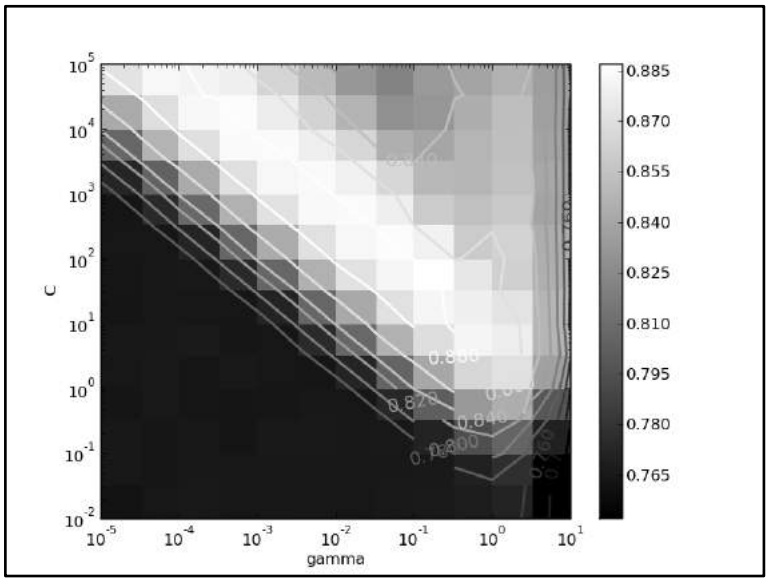
Tanulmányok	pontosság	default	teljes mintaméret	teszt-arány	q_4
(Mittermayer & Knolmayer 2006a)	83,00%	33,33%	629	10%	0,2128
(Peramunetilleke 1997)	51,00%	33,33%	100	50%	0,1031
(Cho & Wüthrich 1999) (Cho et al. 1999)	51,00%	34,60%	180	20%	0,1007
(Koppel & Shtrimberg 2004)	70,30%	50,00%	850	10%	0,0750
(Généreux et al. 2008; 2011)	69,50%	50,00%	1 000	10%	0,0702
(Groth & Muntermann 2008)	70,00%	57,50%	160	10%	0,0632
(Groth et al. 2014)	81,69%	75,00%	415	10%	0,0221
(Groth & Muntermann 2011)	78,96%	75,00%	423	10%	0,0132
(Groth 2010)	78,72%	75,00%	423	10%	0,0125
(Groth & Muntermann 2010)	78,49%	75,00%	423	10%	0,0117
(Gidófalvi & Elkan 2003)	47,00%	50,00%	6 000	45%	-0,0078
(Groth & Muntermann 2009)	56,50%	60,80%	423	10%	-0,0187

Forrás: saját szerkesztés

A *10. táblázatban* összehasonlításra kerültek azon tanulmányokban közölt legjobb pontosságmutatók, amelyek esetén a default pontosság, a teljes mintaméret és a teszt-minta aránya is ismert volt, és így ezek alapján a mutatószámok minősége kiszámolható

volt. (Lásd *7.* és *9. táblázat*.) Ez alapján a legjobb modellt Mittermayer és Knolmayer (2006a) készítette, ezután a Wüthrich-modell három tanulmánya következik (Peramunetilleke 1997; Cho & Wüthrich 1999; Cho et al. 1999), míg a legrosszabb, default alatti pontossággal rendelkező modelleket Gidófalvi és Elkan (2003), illetve Groth és Muntermann (2009) publikálták.

A modellek tömeges összehasonlítására lehet szükség, amikor a modell robusztusságát vagy érzékenységét vizsgáljuk ugyanazon a mintán. Ebben az esetben is tesztelhető az átlagos pontosságok különbözősége varianciaanalízis vagy t-próbák segítségével. Célszerű azonban emellett vizualizálni az eredményeket, hogy a modellparaméterek és a modellteljesítmény közötti összefüggések felismerhetők legyenek. Az ilyen célt szolgáló hőtérképek segítségével információt kapunk a döntési tér szerkezetét, illetve az input- és outputváltozók közötti összefüggéseket illetően. A *13. ábra* az rbf-kernelű SVM-osztályozó gamma és C paramétere, illetve a pontosság közötti összefüggést mutatja be egy ilyen hőtérképen. Az ábrán látható hőtérkép mindkét tengelye logaritmikus beosztású, és az egyes rácspontoknak megfelelő paraméterkombinációkkal tanították a modellt. A keresztvalidáció átlagos pontossága a rácspontba rajzolt téglalap világosságának felel meg. Látható, hogy ha a gamma értéke nagyon alacsony, akkor nagyon magas C érték esetén javult csak az osztályozó pontossága a mintán. A gamma növelésével egyre kisebb C paraméterek mellett is elérhető volt a javulás az osztályozó teljesítményében. Ezek arra utalnak, hogy a vizsgált problémához tartozó optimális döntési határ nem lineáris. Az



13. Ábra: A C és gamma paraméterek közötti összefüggés rbf-kernel esetén egy nemlineáris osztályozási problémára

Forrás: (Ben-Hur & Weston 2010, o.11)

ábrán az is megfigyelhető, hogy a maximális pontossághoz közeli teljesítményt nyújtó modellek a paramétertérben átlósan helyezkednek el. Ez a tulajdonság a nemlineáris optimális döntési határral rendelkező problémákra jellemző, ugyanis az optimálishoz közeli finomságú görbület csak megfelelően megválasztott gamma-C kombinációkkal érhető el. (Ben-Hur & Weston 2010) Lineárisan tökéletesen szeparálható problémák esetén ezzel szemben szinte bármely gamma-C kombinációval nagyjából azonos, 100% körüli pontosság érhető el, hiszen a szeparálósíknak nem kell görbülnie[94]. Végül megemlíteném, hogy abban az esetben, ha a kategóriák nem szeparálhatók, azaz nincs összefüggés az input és outputváltozók között, akkor valamennyi gamma-C kombináció hasonló pontosságot produkál, amely a legnagyobb elemszámú kategória részarányához – a default pontossághoz – közeli érték. Ez utóbbi két esetben a *13. ábrán* is látható átlós szerkezet nem figyelhető meg.

A saját modellem kiértékeléséhez a pontosságmutatót használom, 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. A keresztvalidációt RapidMinerben az *X-Validation* operátorral valósítottam meg, alapbeállításokkal. A pontosságot *Performance (Classification)* operátorral mértem. Az SVM gamma és C paramétereinek vizsgálata során *Optimize Parameters (Grid)* operátort használtam, a gamma esetén 0,001-től 5-ig, 20 darab logaritmikus lépésközt[95] vettem fel, a C esetén 0,1-től 5000-ig ugyancsak 20-at[96]. Az eredményeket RapidMinerben vizualizáltam hőtérképen az Advanced Charts eszköz segítségével.

3.5. A modell üzleti szempontú kiértékelése

Az előző alfejezetben láttuk, hogy a tőzsdei hírbányászati modellek hatékonysága általában felülmúlja a benchmarkot, de a legtöbb esetben jelentősen a maximum alatt maradnak. A modell elfogadásának feltétele, hogy az üzleti céloknak megfelelően lehessen felhasználni. Idézzük fel a *2. táblázatból*, hogy a legfőbb üzleti cél a spekulációs nyereségszerzés volt, illetve egy esetben a tranzakciós költségek minimalizálása. A kutatók

94 Ez természetesen függ a tanítóminta méretétől is. Másrészt számábrázolási hibák miatt nagy gamma értékek – azaz kis sugár – és kis C értékek esetén előfordulhat, hogy a szoftver által becsült pontosság 100%-tól jelentősen elmarad. Ezt a jelenséget a LibSVM csomagnál tapasztaltam.

95 A gamma tengely beosztása: 0,001; 0,0015; 0,0023; 0,0036; 0,0055; 0,0084; 0,0129; 0,0197; 0,0302; 0,0462; 0,0707; 0,1083; 0,1657; 0,2537; 0,3884; 0,5946; 0,9103; 1,3936; 2,1334; 3,266; 5.

96 A C tengely beosztása: 0,1; 0,1718; 0,2951; 0,5068; 0,8706; 1,4953; 2,5686; 4,412; 7,5786; 13,0178; 22,3607; 38,409; 65,9754; 113,3262; 194,661; 334,3702; 574,3492; 986,5623; 1694,6226; 2910,8612; 5000.

jelentős része készített valamilyen kereskedési szimulációt vagy profitkalkulációt az üzleti megfelelőség tesztelésére is. E szimulációkat foglalja össze a *11. táblázat*.

11. Táblázat: A tőzsdei hírbányászati modellek üzleti kiértékelése

Modell	Cikk	I	M	Kt	Pm	Eh	Ph	Hozam			Benchmark		
								Öh	Psz	H/P	Bt	Bé	p
Wüthrich	(Wüthrich, Permunetilleke, et al. 1998) (Wüthrich, Cho, et al. 1998)	60d	bh	–	–	1d	1d	7,5%	40	?	b&h (DJIA)	5,1%	?
Lavrenko	(Lavrenko et al. 2000a)	40d	szh spot	?	10k$	0h (1–10h)	1h	280k$	kb. 12k	0,23% (0,02–0,16%)	r (1000)	-9300$	0,8%
	(Lavrenko et al. 2000b)					0h		21k$	?	?			
	(Fung et al. 2002a; 2002b)	?	szh spot	–	–	0h	1d	?	?	?	?	?	?
	(Fung 2003)						3d	28,1%			1. b&h 2. r (1000)	1) -20,6%	2) 3,8%
	(Fung et al. 2005)						3d	18,1%					2) 7,8%
Schumaker	(Schumaker & Chen 2006)	23d	szh spot	kb. 100–300k $	1k$	20m	20m	4–6k$ (2,2–3,6%)	?	?	regr.	kb. -1800$	?
	(Schumaker & Chen 2008)			?				8,5%	1998		1. b&h (m, a) 2. k (m, a)	1) -5,5–3,4% 2) 4,5–20,8%	<5%
	(Schumaker & Chen 2010)								?		1. b&h (SP500 2. b&h (ka)	1) 5,62% 2) 4,95–24,73%	?
	(Schumaker & Chen 2009)							2,1%			regr.	-1,95%	<5%
	(Schumaker 2009)							1,6–3,4%			–	–	–
	(Schumaker et al. 2009; 2012)							0,4–3,3%					
Groth	(Groth & Muntermann 2009)	?	szh spot	–	–	15m	15m	?	236	1,05%	1. d 2. r (5000)	1) 0,37% 2) 0,01%	<5%
	(Groth & Muntermann 2010)	2y	szh opc			15m (30m)	15m (30m)	?	9–361	2,37–7,34% (3,18–8,35%)	all-long	2,03% (2,81%)	<5%
	(Groth & Muntermann 2011)								1–388	2,54–10,48% (2,89–12,42%		2,39% (3,01%)	többnyire <5%
	(Groth 2010)	kb. 2,5y	szk tr			15m	15m	?	10-52	-0,46–1,49	naiv	2,01–2,53	<10%
	(Groth et al. 2014)	kb. 3,5y							?	?		? rosszabb	
Thomas	(Thomas & Sycara 2000)	200d	szep	–	–	1d	min. 1d	-8,1–6,9%	?	?	b&h	?	?
Gidófalvi	(Gidófalvi & Elkan 2003)	4,5M	szhp	–	–	-30–30m	-30–30m	?	?	-1–1$	r (1000)	?	+/-20 perc: <10%
Mittermayer	(Mittermayer 2004)	?	szh spot	–	–	1h	1h	?	2477–2864	0,11%	r	0%	<1%
	(Mittermayer & Knolmayer 2006a)					15m	15m	72–89%	429	0,22–0,27%	r	0%	?

Forrás: saját szerkesztés

A táblázatban alkalmazott rövidítések, illetve a szimulációs technikák vázlatos áttekintése az alábbiakban olvasható:

I: Időtáv. A szimulációhoz felhasznált minta által felölelt időtartam hossza. **d**: nap, **M**: hónap, **y**: év.

M: Mutató. A megtérüléskalkuláció során kiszámított pénzügyi mennyiség.

- **bh**: becsült hozam, szimuláció nélkül
- **szep**: szimulált extraprofit buy-and-hold stratégiához képest spot ügyleteken
- **szh spot/opc**: szimulált hozam, százalékban, illetve dollárban. A vizsgált piac lehet spot, illetve opciós.
- **szhp**: szimulált hedge pozíció profitja, a piaci portfólióban és a részvényben ellentétes irányú pozíció felvétele
- **szk tr**: szimulált tranzakciós költség spot ügyleteken

Kt: Kezdőtőke. A táblázatban a **k** betű az ezer rövidítése.

Pm: Pozícióméret. A szimuláció során egy adott ügyletben kockáztatott tőke nagysága. A táblázatban a **k** betű az ezer rövidítése.

Eh: Előrejelzés hossza. Azt jelenti, hogy mely időpontra vonatkozik a szimuláció során adott megbízás alapjául szolgáló előrejelzés. **m**: perc, **h**: óra, **d**: nap.

Ph: Pozíció hossza. Ennyi ideig tartják a szimulációban a pozíciókat. **m**: perc, **h**: óra, **d**: nap.

Öh: Összes hozam. A teljes tesztidőszak alatti, dollárban összesített vagy százalékosan kifejezett hozam. A táblázatban a **k** betű az ezer rövidítése.

Psz: Pozíciók száma. A teljes tesztidőszak alatt szimulált pozíciók száma.

H/P: Hozam/pozíció. Az egy pozícióra vetített hozam.

Bt: Benchmark típusa. Ha több is volt, számozva vannak.

- **all-long**: az előrejelzéstől függetlenül long opciókba fektető stratégia
- **b&h**: buy-and-hold, a tesztidőszak elején nyitott és végén zárt long pozíció. Zárójelben egyértelműsítések, például instrumentum rövidítése, illetve **m**: momentum stratégia, **a**: anticiklikus stratégia, **ka**: kvantitatív alapok.
- **d**: default, azaz a leggyakoribb hozamváltozás-előjelnek megfelelő pozíció
- **k**: kombinált stratégia, momentum, illetve anticiklikus stratégiával kombinált szöveges előrejelzés
- **naiv**: tranzakcióidőzítést mellőző stratégia
- **r**: randomization teszt, zárójelben az ismétlések száma
- **regr.**: regresszió, szöveges input nélküli regressziós modellre épített szimuláció

Bé: Benchmark értéke. Ha több benchmark is volt, annak sorszáma után szerepel a benchmark értéke.

p: p-érték. A benchmark és a vizsgált hozam vagy költség egyezőségét tesztelő próba p-értéke.

A legelterjedtebb, spot kereskedési szimuláció során, ha a modell egy hír hatását pozitívnak jelzi, akkor vételi, ha negatívnak, akkor eladási pozíciót szimulálnak, különben pedig nem történik akció. A pozíciót általában az előrejelzett időtáv végén zárják le, de megfogalmazhatók korai zárási szabályok is, például, ha a hozam elér egy küszöbértéket. A szimulációhoz nagyfrekvenciás árfolyamokat használnak, és feltételezik, hogy a tranzakció mindig végrehajtható az adatbázisban tárolt áron. A valóságban persze ez függ a pozíció méretétől is, azonban ettől el szoktak tekinteni, akárcsak a tranzakciós költségektől. A szimulációk többsége a kezdőtőke nagyságával sem számol. Általában a pozíciók hozamát ezért százalékban adják meg, hogy a tőkétől függetlenítsék a mérést. A stratégia jellemzésére pedig általában az átlagos hozamot, és annak szórását használják, de ehhez feltételezni kell, hogy a pozíciók mérete azonos, illetve, hogy a korábbi nyereséget nem fektetjük be újra, illetve a veszteséget pótoljuk. Egyszerű hozamok és kamatos kamat esetében egyébként hibás volna az összeadás, de ha állandó mennyiségű tőkét fektetünk be, az egyszerű hozamok összeadása értelmezhető. Ha különböző rész-

vényekbe egy időben is befektethetünk a szimuláció során, akkor a pozícióméretnél kétszer, háromszor stb. nagyobb tőkével kell rendelkezni. Ez azt is jelenti, hogy az összesített pozíciónkénti hozam nem a befektetéshez felhasznált összes tőkére vonatkozik. Ezért nem találkozunk az irodalomban olyan esettel, amikor a pozíciónkénti hozamokat a buy-and-hold stratégia által elért hozammal vetik össze. Az egy pozícióra jutó átlagos hozamot általában egy vagy több másik szimuláció egy pozícióra jutó átlagos hozamával vetik össze, t-próba alkalmazásával. A tőzsdei hírbányászatban ilyen benchmark a numerikus regressziós modell, vagy valamilyen konstans típusú kereskedési döntést végrehajtó stratégia[97], illetve a randomization teszt vagy resampling – újramintavételezés. Ez utóbbi úgy működik, hogy azokban az időpontokban, amikor a vizsgált rendszer az előrejelzés alapján kereskedési döntést hoz – tehát a hírek publikálásakor –, véletlenszerűen szimulálunk döntéseket – általában az eredeti valószínűségeloszlásnak megfelelően. A véletlen kereskedés átlagos hozama t-próbával is összehasonlítható a vizsgált rendszerével, de a véletlen szimuláció többszöri ismétlésével az átlagos hozam eloszlása empirikusan is közelíthető, és így a vizsgált rendszer átlagos hozamának az empirikus eloszlásban elfoglalt helyzete alapján elvetjük – általában, ha az eloszlás alsó vagy felső 1–5 percentilisébe esik – vagy elfogadjuk azok egyezőségét.

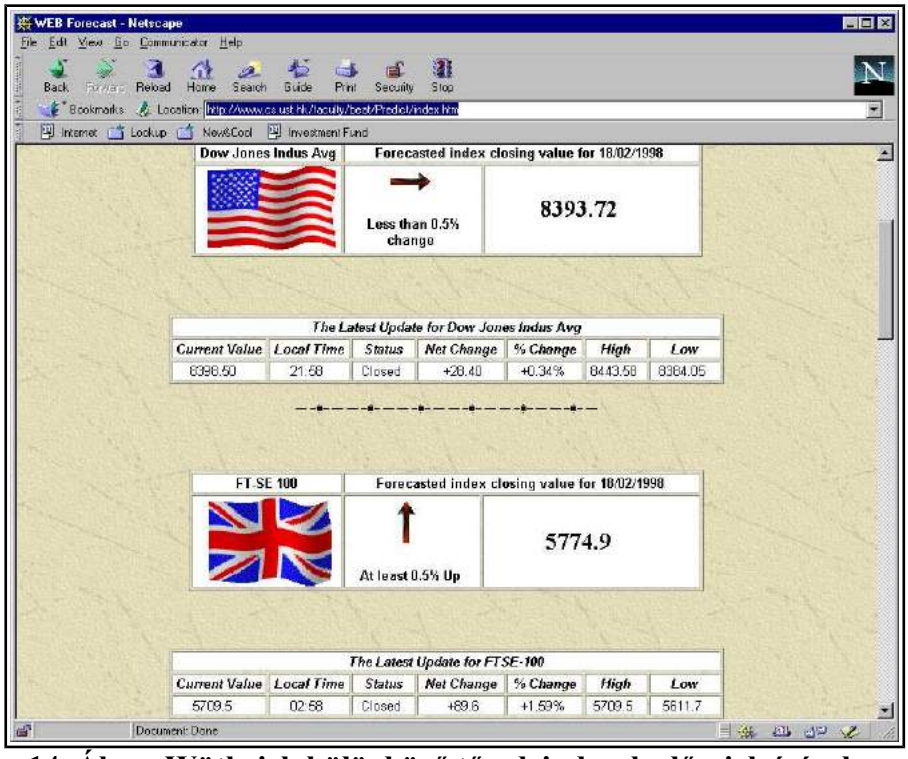
Látható a *11. táblázatban*, hogy a szimulált időtáv elég tág határok között mozog a szakirodalomban, 23 naptól 3,5 évig is terjedt – természetesen utóbbit azt tette szükségessé, hogy Groth egy kisebb részvénypiacon vizsgálódott, ahol kisebb volt a hírek publikálásának gyakorisága. Az eredmények összehasonlítása meglehetősen nehéz, ugyanis azon kívül, hogy nem azonos időszakra és instrumentumokra vonatkoznak, a tranzakciók száma és hossza is eléggé eltér az egyes esetekben, arról nem is beszélve, hogy a benchmarkok és az alkalmazott hipotézisvizsgálatok köre sem egységes. Az esetek többségében a hírbányászat révén a benchmarknál szignifikánsan jobb hozamot értek el, de elég nehéz megmondani, melyik modell a legjövedelmezőbb. Ehelyett a szimulációk közül kiemelném a legegyedibbet, amelyet Groth (2010; Groth et al. 2014) készített. A végrehajtásra váró tranzakciók időzítéséhez a tranzakciós költségek minimalizálásán keresztül közelített, és szimulációja igazolta, hogy negyed órára előrejelezhető, hogy a hírt követően lehet-e majd szignifikánsan alacsonyabb költségekkel kereskedni, és így jelentős költségmegtakarítás érhető el.

97 Például az all-long, a default és a naiv. (Lásd *11. táblázat*.)

A saját modellemben nem készítek kereskedési szimuláció típusú kiértékelést, ugyanis a különböző mintákban csak körülbelül tíz-tíz darab vagy kevesebb árfolyamváltozás éri el a BÉT-en szokásos jutalék nagyságát[98], így nem várható a modelltől sem, hogy ésszerű méretű tőkével megtérülő kereskedési stratégiát adjon. A modell jelenlegi célja a közleményekben rejlő információk hatásának kimutatása és a modell érvényességének bizonyítása, az üzleti célokra való alkalmassá tétele a jövőbeli munka részét képezi.

3.6. Üzleti hasznosítás, tudásbeépítés

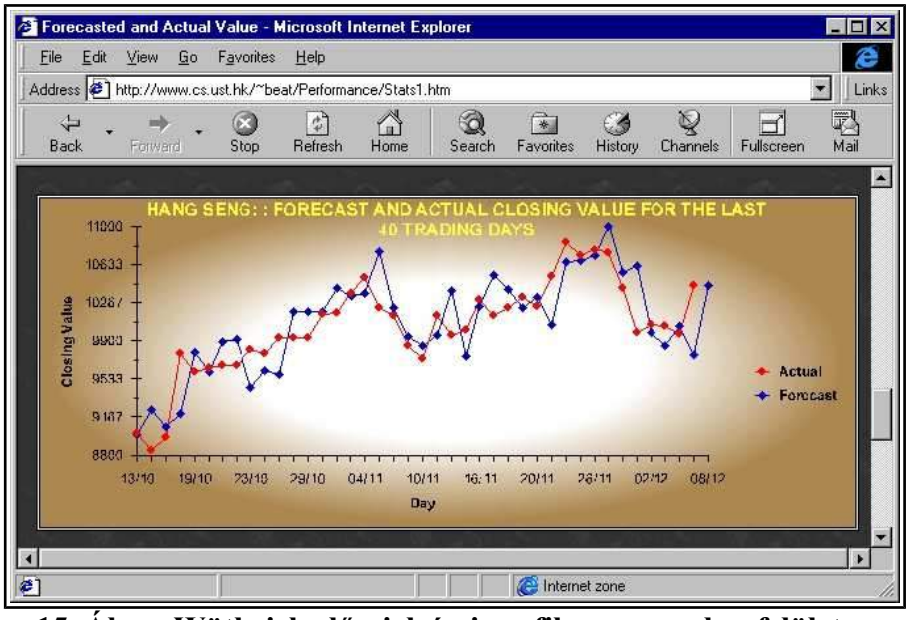
A *2. fejezetben* bemutatott modelleket elsősorban akadémiai kutatásokhoz használták, de ennek ellenére lehet találni köztük olyat, amely alapján üzleti felhasználók számára is elérhető alkalmazás készült. Például rögtön a legkorábbi, azaz a Wüthrich-modell is rendelkezett saját webes felülettel, amely a *www.cs.ust.hk/~beat/Predict* címen volt elérhető (Wüthrich, Cho, et al. 1998; Wüthrich, Permunetilleke, et al. 1998). A *14. és 15. ábrán* látható ennek a felületnek két képernyője. Az előrejelzést a tőzsdei nyitás előtt tették közzé a honlapon, és visszamenőleg is meg lehetett tekinteni.



14. Ábra: Wüthrich különböző tőzsdeindexek előrejelzésével kapcsolatos honlapja

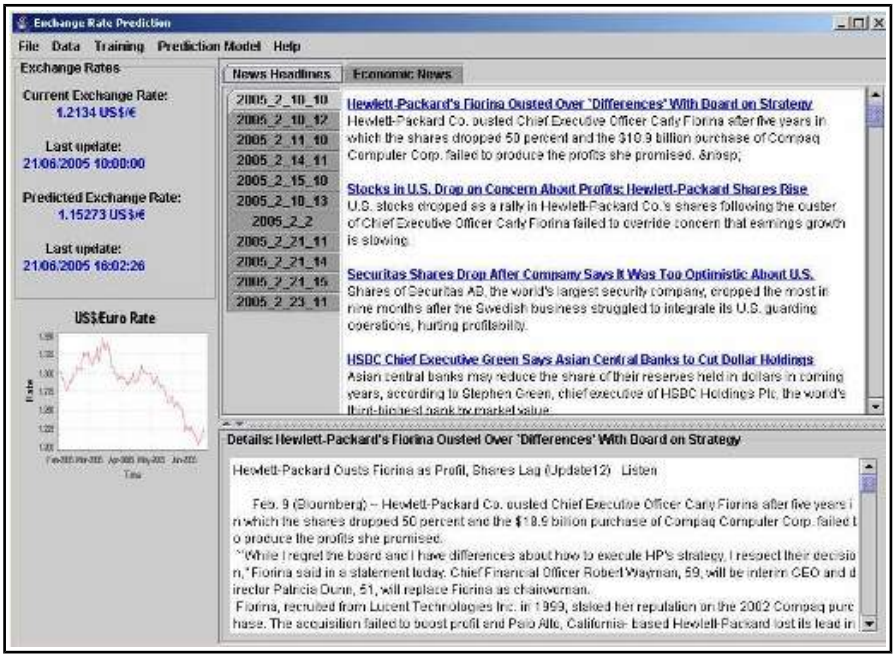
Forrás: (Wüthrich, Permunetilleke, et al. 1998)

98 Lásd például (Erste Befektetési Zrt. 2016).



15. Ábra: Wüthrich előrejelzései grafikusan a webes felületen

Forrás: (Cho et al. 1999)



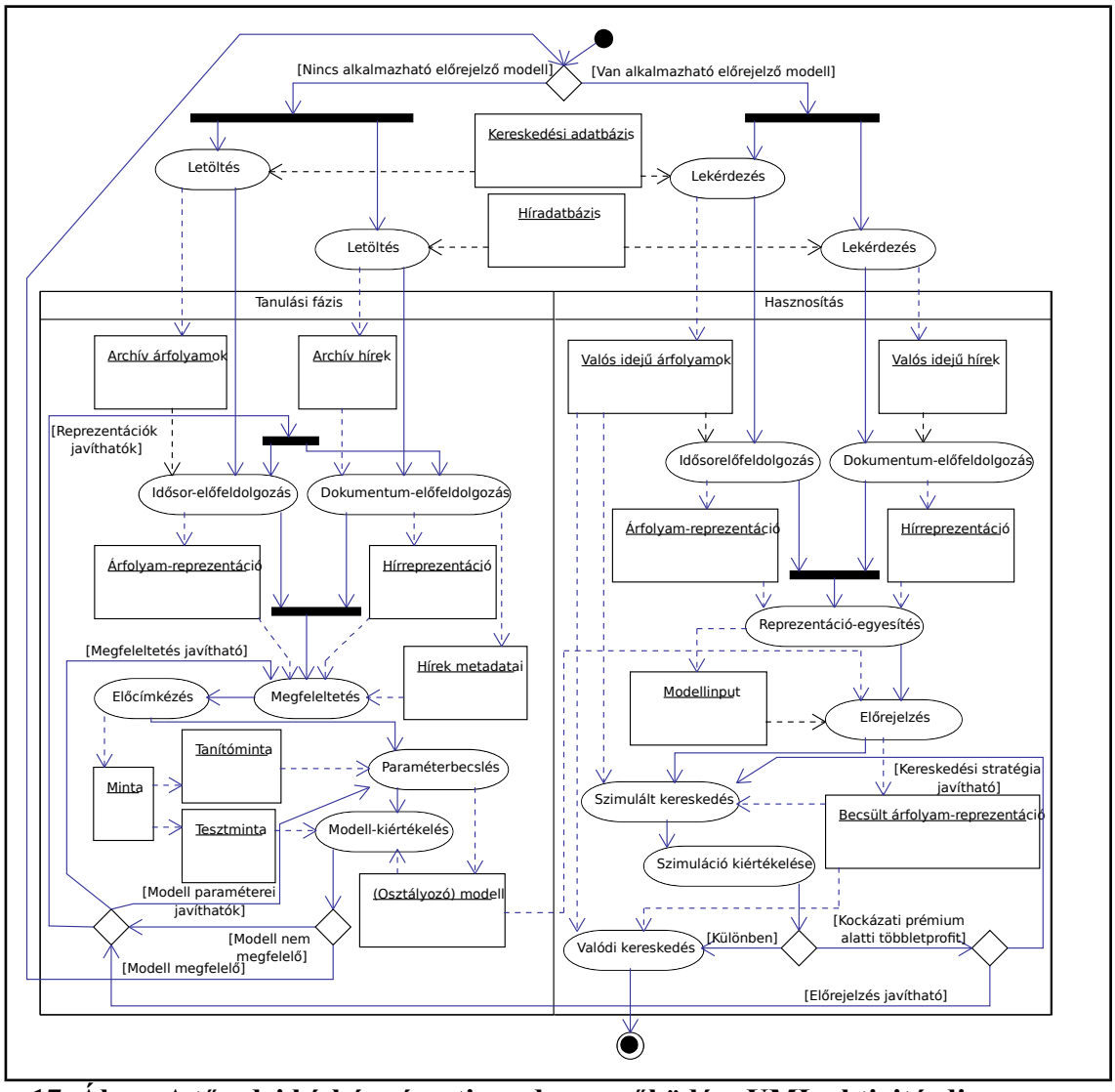
16. Ábra: Az eMarkets Group devizaárfolyam-előrejelző alkalmazása

Forrás: (Zhang et al. 2005)

Bár több modellnek saját neve is volt – pl. Ænalyst (Lavrenko et al. 2000a; 2000b), NewsCATS (Mittermayer 2004; 2006; Mittermayer & Knolmayer 2006a; 2007), AZFin-Text (Schumaker 2009; 2010b; 2010a; Schumaker & Chen 2006; 2008; 2009a; 2009b; 2010; 2011; Schumaker et al. 2009; 2012) –, de üzleti felhasználóknak készült változat-

tal kapcsolatban nem közöltek részleteket. Az eMarkets Groupnál viszont az üzleti vonal markánsabban van jelen, mint az akadémiai, bár alkalmazásuk elérhetőségével kapcsolatban keveset tudhatunk meg tanulmányaikból. A *16. ábra* a devizaárfolyam-előrejelzéshez készített szoftverük felhasználói felületét mutatja. (Zhang et al. 2005; 2007)

A saját modellem névadásával, illetve üzleti alkalmazássá fejlesztésével kapcsolatban egyelőre nem történtek erőfeszítések, de (Kovács 2014a) tanulmányomban megadtam egy teljes rendszertervet, amely a *17. ábrán* látható. Az aktivitásdiagram átfogja a teljes *3. fejezetet*, és érzékelteti, milyen komplex módon kapcsolódnak egymáshoz a modell egyes részei. A *4. fejezetben* bemutatott eredményeim az ábra bal oldalához kapcsolódnak.



17. Ábra: A tőzsdei hírbányászati rendszer működése UML aktivitásdiagramon

Forrás: saját szerkesztés

4. Tőzsdei hírbányászat a BÉT-en

Ebben a fejezetben a saját kutatási eredményeket mutatom be. Bizonyos vizsgálatok azonos típusú kísérleti összeállítást igényelnek, ezeket e fejezet bevezetőjében mutatom be, a hipotézisek vizsgálatát pedig az egyes alfejezetek tartalmazzák.

Az első típusba tartozó kísérleti összeállításokban a sajtóközlemények nyelvét, a szövegreprezentációt és az SVM paramétereit változtattam, az időablak eltolását viszont változatlanul +20 percen hagytam[99]. Az időablak méretének megállapításához Gidófalvi (2001; Gidófalvi & Elkan 2003) és Schumaker munkásságára (lásd *2.1.3 alfejezet*) támaszkodtam, de Groth munkássága (lásd *2.1.4 alfejezet*) is a hasonló hosszúságú, 15, illetve 30 perces időablakok vizsgálatára épül, és Mittermayer is dolgozott 15 perces intervallummal (Mittermayer & Knolmayer 2006a). A hozamkategóriák kialakításakor a hír publikálása utáni huszadik percben érvényes árfolyam és a publikáláskori árfolyam természetes alapú logaritmusának különbsége alapján három közel azonos méretű kategóriába soroltam a megfigyeléseket, ami megfelel a Wüthrich-féle modellben (lásd *2.1.1. alfejezet*) alkalmazott elveknek. Így 42 pozitív, 59 semleges és 54 negatív kategóriájú megfigyelés volt összesen. A negatív kategória értéktartománya $]-\infty;0[$, a pozitívé $]0;\infty[$, a semleges kategóriába a nullhozamok kerültek. A radiális bázisfüggvény kernellel rendelkező SVM-nek két paraméterét, a C-t és a gammát változtattam a *3.4.2 alfejezet* végén leírtaknak megfelelően. Összesen tehát két nyelvre, négyféle szövegreprezentációra, 21 darab C és 21 darab gamma lépésközre $2 \cdot 4 \cdot 21 \cdot 21 = 3528$ kísérleti összeállítás eredményeit hasonlítottam össze a modellek átlagos pontossága alapján. Minden kísérleti összeállítás esetén tízszeres keresztvalidációt végeztem[100]. A páros összehasonlításokhoz kétmintás t-próbát végeztem, a default pontossággal való összehasonlításokhoz egymintásat, mindkét esetben 1, 5 és 10%-os szignifikancia szintek mellett. A tömeges összehasonlítások esetén a nullhipotézisek elfogadásának és elutasításának az összetartozó kísérletek számához képesti arányából következtettem kutatási hipotéziseimre. Az arányokra különböző küszöbértékeket fogalmaztam meg: teljes egyezés, azaz 0%, illetve részleges egyezés, azaz 1, 5 és 10% ellentmondás a kutatási hipotézissel. Küszöbérték feletti ellentmondás esetén a kutatási hipotézist elvetem, különben elfogadom. Ezzel a típusú összeállítással tesztelhető a H1 – a szöveges előrejelzés pontosabb a

99 Ezért ezt a típust fix időablakosnak is nevezem.

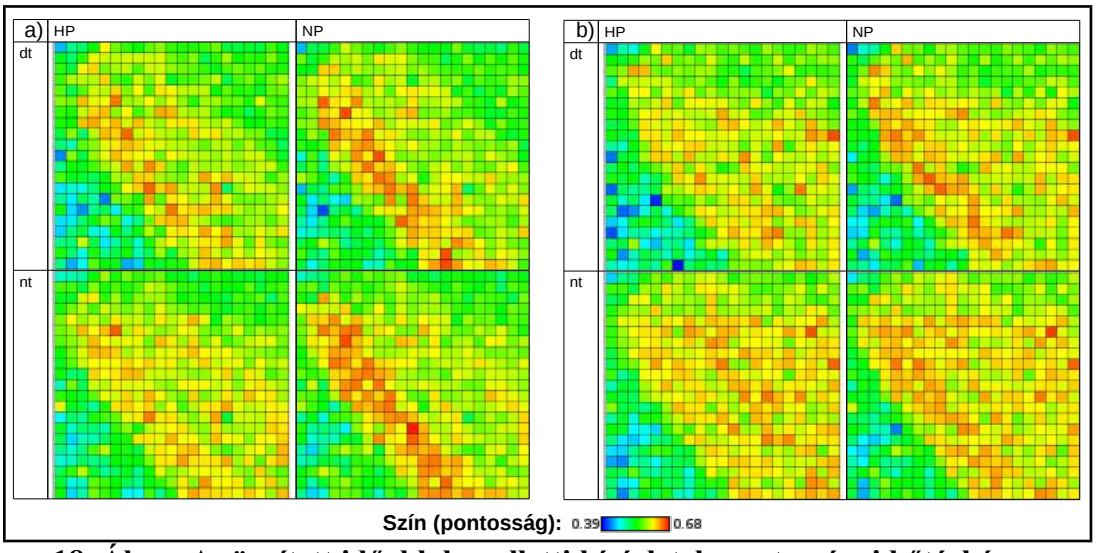
defaultnál –, a H4 – nem függ a szöveg nyelvétől a modell pontossága –, a H5 – a modell robusztus a paraméterekre – és a H6 hipotézis – a modell robusztus a szövegreprezentációra.

A második típusba tartozó kísérleti összeállításokban az SVM paramétereit és az időablak eltolását változtattam[101], míg a sajtóközlemények nyelvét és a szövegreprezentációt az első típusú kísérletek során kapott eredmények alapján választottam meg[102]: a magyar nyelvű korpuszt használtam, amelyet az (NP; nt) reprezentációnak megfelelően készítettem elő az elemzéshez. Ebben az összeállításban két olyan hipotézisemet fogom ellenőrizni, amelyek az időbeliség kérdéséhez kapcsolódnak, a H2 – a közzététel haladéktalan – és a H3 hipotézist – a modell robusztussága és minősége az időablakkal változik. A H3-as hipotézis esetén az előrejelzési időtávot 1 perctől 120 percig változtatva megvizsgálom, hogy hogyan változik a robusztus modell készítésének nehézsége, illetve a modell defaulthoz képesti pontossága. A H2-es hipotézis tesztelése hasonló ehhez, azonban ebben az esetben a bejelentést megelőző időszakra végzem el ugyanezt a vizsgálatot, tehát a bejelentés előtti 1 perctől a bejelentés előtti 120 percig változik az időtáv. A vizsgált hírek száma eltérő a különböző hosszúságú időablakok esetén, ugyanis az időablak egyik része sem eshet a kereskedési időn kívülre, amely 9:00-kor kezdődik és 17 óra után fejeződik be – az ajánlati könyv kiegyensúlyozása után, véletlenszerűen. Az egyértelműség kedvéért egységesen 17:10 percet tekintettem a kereskedési idő végének, amely a mintában található legkésőbbi árfolyamadatot – 17:08 – is magában foglalja. Míg az első típusú kísérleti összeállításban használt 20 perces időablak esetén kizárólag a nullhozamok tartoztak a semleges kategóriába, addig a hosszabb időablakok esetén a semleges kategóriához tartozó alsó és felső korlátok nullától különböző számok is lehetnek. A *3.3.3 alfejezet 8.* és *9. ábrája* mutatta be az egyes kategóriák számát és megoszlását az egyes időablakok szerint.

A következőkben bemutatom mind a fix időablakos, mind a változó időablakos típusú kísérletek eredményét, de az elemzésükre majd a külön alfejezetekben fog sor kerülni.

101 Ezért ezt a típust változó időablakosnak is nevezem.

102 Ezeket az eredményeket csak később a 4.4–4.6 alfejezetekben mutatom be, mert a hipotézisek sorrendjében a fő rendezőelv az volt, hogy előre kerüljenek az inkább piacelméleti, az információ árfolyamba való beépülésével kapcsolatosak, és hátrébb kerüljenek az inkább módszertani, a modell inputjaival és paramétereivel kapcsolatosak.



18. Ábra: A rögzített időablak melletti kísérletek pontosságai hőtérképen

Forrás: saját szerkesztés

Az első típusú kísérletek eredményeit szemlélteti és foglalja össze a *18. ábra*, melynek a) részében az angol, b) részében a magyar nyelvű korpuszon elért eredmények láthatók. A négyfajta szövegreprezentáció egy-egy $2 \cdot 2$ -es mátrixba rendezve látható az ábrán. A mátrixok oszlopfejlécében a *HP* – highlighted phrases – kód annak jelölésére szolgál, hogy a reprezentáció tartalmazza a kifejezéseket, míg az *NP* kód azt jelenti, hogy a szövegben nem kerültek kiemelésre kifejezések. A mátrixok sorfejlécében a *dt* – deleted templates – kód jelöli azt az esetet, amikor a sablonszövegek eltávolításra kerültek a reprezentációból, az *nt* kód pedig azt, amikor a sablonszövegek tartalma is részét képezi a reprezentációnak. A négy lehetséges reprezentációt rendezett párként jelöltem:

(NP; nt): nincsenek a kifejezések kiemelve és a sablonszövegek törölve, ez a normál szózsák-szövegreprezentáció;

(NP; dt): nincsenek a kifejezések kiemelve, de a sablonszövegek törölve vannak, lényegében ez is a normál szózsákmodell, csak korpuszhoz igazított stopszavazással;

(HP; nt): a kifejezések ki vannak emelve, a sablonszövegek nincsenek törölve, ez egy bővített szózsák-reprezentáció;

(HP; dt): a kifejezések ki vannak emelve, a sablonszövegek törölve vannak, ez szintén kibővített szózsákmodell, stopszavazással.

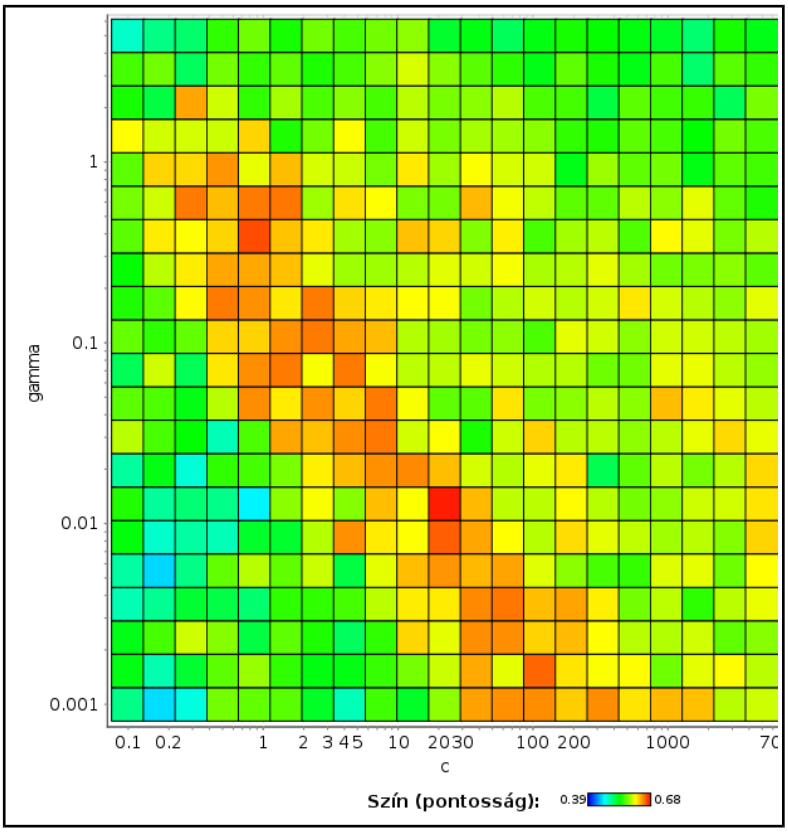
A *3.3.1.1. alfejezetben* már megadtam, hogy körülbelül hány kifejezést sikerült kinyerni, de konkrétan a húszperces vizsgálatnál a fenti reprezentációk jellemzőinek száma a *12. táblázatban* látható.

12. Táblázat: A szövegjellemzők száma nyelvenként és szövegreprezentációnként

	angol		magyar	
	HP	**NP**	**HP**	**NP**
dt	4405	4069	9132	8906
nt	4415	4079	9140	8913

Forrás: saját szerkesztés

Visszatérve a *18. ábrára*, láthatjuk, hogy nyelvenként és reprezentációnként egy-egy hőtérkép ábrázolja a modellek validáció során mért átlagos pontosságát. Helytakarékossági okokból a hőtérképek vízszintes és függőleges tengelyének beosztása nem látható, mivel az minden esetben ugyanaz[103]. Az egyértelműség kedvéért az angol nyelvű, *(NP; nt)* reprezentációjú hőtérképet külön is feltüntettem a *19. ábra* formájában.



19. Ábra: Angol nyelvű (NP; nt) reprezentáción és húszperces időzítéssel elért átlagos pontosságok különböző gamma és C paraméterek esetén

Forrás: saját szerkesztés

103 A koordináta-rendszer vízszintes tengelyén az SVM C paramétere található logaritmikus osztásközökkel. Az első osztásköz 0,1-nél, az első négyzet középpontjánál található, az utolsó 5000-nél, az utolsó négyzet középpontjánál. A függőleges tengelyen a gamma paraméter található logaritmikus osztásközökkel 0,001 és 5 között. A tengely minimális és maximális értéke viszont kívül esik e két érték által meghatározott intervallumon, hogy a négyzetek teljes terjedelmükben ábrázolhatók legyenek. Bővebben lásd a 3.4.2 alfejezetben.

Minden rácspontban egy-egy tízszeres keresztvalidáció átlagos eredményének megfelelő színezésű négyzet látható. A színskálán a kék színt az összes, 3528 darab modell legalacsonyabb pontosságát lefelé kerekítve, 0,39-hez rendeltem, a piros színt az összes modell legmagasabb pontosságát felfelé kerekítve, 0,68-hoz rendeltem. Érdemes összevetni az *19. ábrát* a *3.4.2 fejezetben* bemutatott *13. ábrával*[104]. Hőtérképeimen megfigyelhető az a *13. ábrán* láthatóhoz hasonló átlós alakzat, amely az optimálishoz közeli döntési határt produkáló modellek gamma-C kombinációit tartalmazza. Mivel a legnagyobb pontosságú modell teljesítménye sem haladta meg a 68%-ot, ebből arra lehet következtetni, hogy a probléma nem szeparálható tökéletesen az optimális döntési határral. A hőtérképeken feltűnő átlós szerkezetben a legmagasabb pontosságúhoz közeli modellek az angol nyelvű dokumentumoknál inkább egy szűkebb sávban találhatók, míg a magyar nyelvűeknél egy jóval szélesebb sávban. Ennek okát a *12. táblázattal* tudom indokolni, azaz, hogy a magyar nyelvű reprezentációban a szövegjellemzők száma több mint kétszerese az angol nyelvűben találhatóknak. Ha egy térhez további dimenziókat veszünk hozzá, a magasabb dimenziójú térben ugyanazon pontok közötti euklideszi távolság egyenlő vagy nagyobb lehet, mint az eredeti térben[105]. Az rbf-kernelű SVM szempontjából ez a margó és a döntési határ alakjának nagyobb mozgásterét jelenti adott pontosság mellett.

A *19. ábra* elemeire a következő fogalmakkal hivatkozok a dolgozatban.

Egy $p_0=(C_0,\gamma_0)$ C-gamma paraméterkombinációval rendelkező modell $p_1=(C_1,\gamma_1)$ *paramétertérbeli szomszédjának* – vagy röviden *paraméterszomszédjának* – azokat a C-gamma kombinációkat nevezem, amelyekre igaz, hogy $|r(C_0)-r(C_1)|\leq 1$, $|r(\gamma_0)-r(\gamma_1)|\leq 1$ és $C_0\neq C_1$, $\gamma_0\neq\gamma_1$, ahol $r()$ a paraméterérték sorszámát[106] megadó függvény.

Nevezzük *kvázioptimálisnak* az SVM-paramétereket kivéve minden szempontból azonos[107] modellek közötti legpontosabbtól szignifikánsan nem különböző pontosságú modelleket.

---

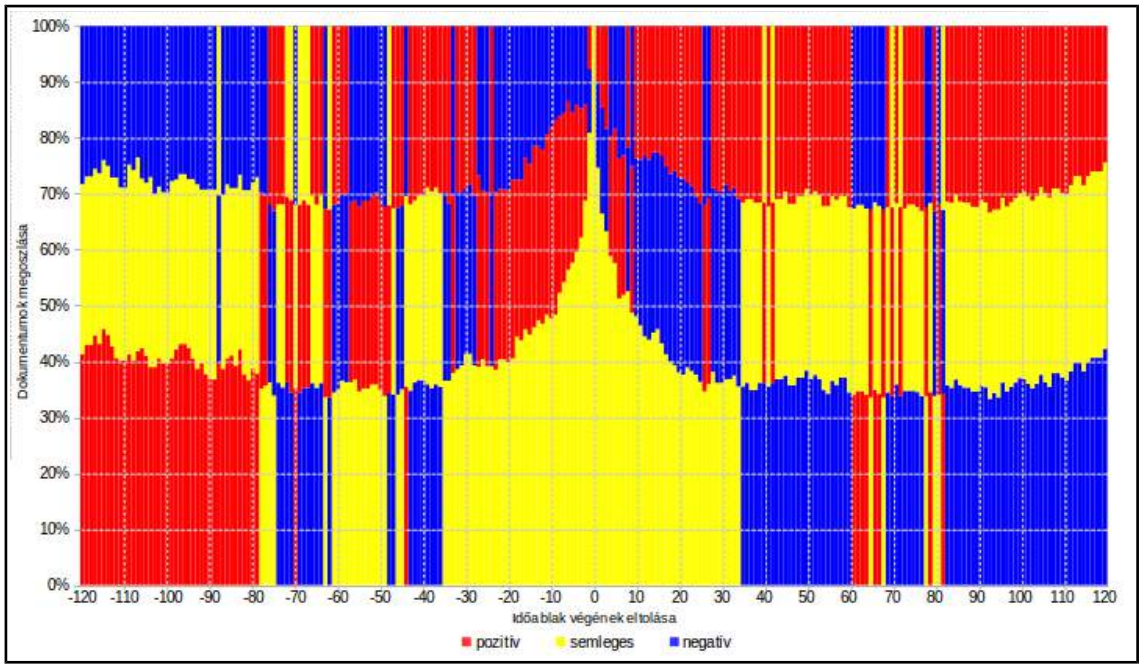
104 A két ábra közötti fő eltérés, hogy a 13. ábra vízszintes tengelyén a gamma paramétert, a függőlegesen a C paramétert mérték fel, míg az én ábráim éppen ellenkező irányú tengelyekkel készültek.

105 Legyen $\underline{a}=(a_1,a_2,\ldots,a_n)$ és $\underline{a}'=(a_1,a_2,\ldots,a_n,a'_{n+1})$, ekkor $|\underline{a}|=\sqrt{\sum_{i=1}^{n}a_i^2}$ és $|\underline{a}'|=\sqrt{\sum_{i=1}^{n}a_i^2+a'^2_{n+1}}$. Mivel mindkét szám pozitív, a négyzetre emelés nem változtatja meg a köztük lévő reláció irányát. Látható, hogy $|\underline{a}'|^2$ összegben egy nemnegatív taggal több van, következésképpen nagyobb vagy egyenlő, mint $|\underline{a}|^2$.

106 A növekvő sorrendbe rendezett paraméterlistában, lásd 95-es és 96-es lábjegyzet, 96. oldal.

107 Azonos a nyelv és azonos a szövegreprezentáció.

A változó időablakos kísérleti összeállítás esetén az eredmények kiértékelésénél tekintettel kell lenni arra, hogy a default modell pontossága az időablak eltolásától függően más-más érték, továbbá a default kategória is hol a semleges, hol a pozitív, hol a negatív. Rendezzük át a *9. ábrán* látható halmozott diagramot úgy, hogy a vízszintes tengely minden beosztása fölött az egyes kategóriáknak megfelelő sávok ne rögzített sorrendben legyenek, hanem részarány szerint felfelé csökkenő sorrendben[108]! Ekkor kapjuk a *20. ábrát*, amelyen az alsó sávban látható, hogy az egyes időtávokon mekkora volt a default kategória részaránya, illetve az alsó sáv színe mutatja, melyik kategória volt az. Láthatjuk, hogy a hír előtt és után 35–35 percen belül a semleges kategória volt a default. Ennek részarányának 33% körülinek kellene lenni, de az ilyen rövid időablakokon belül sokkal ritkábban változik az árfolyam, így nem lehet kiegyenlíteni a három kategória arányát. Ezen az intervallumon kívül viszont körülbelül a ± 90-es időablakig a három kategória részaránya viszonylag kiegyenlített, igazából bármelyik lehetne a default kategória, elméletileg 33%-os részaránnyal. Ezen az intervallumon kívül viszont a hír előtt a pozitív árfolyamváltozás, utána negatív árfolyamváltozás van többségben. Mindez úgy, hogy a semlegesek arányát 33% körül tartva szimmetrikus küszöbértékek szerint soroljuk be a hozamokat a kategóriákba.



20. Ábra: Részarány szerint rendezett halmozott diagram az árfolyam-kategóriák megoszlásáról az időablak hosszának függvényében

Forrás: saját szerkesztés

108 Volt olyan eset, amikor két kategória elemszáma azonos volt, ilyenkor az ábrán az került alsóbb sávba, amelyik a következő felsorolásban előbb szerepelt: negatív, semleges, pozitív.

A különböző időtávokra tanított modellek átlagos pontosságát és a default pontosság egyezőségét t-próbával teszteltem 10, 5, 1, és 0,1 %-os szignifikancia szinten[109]:

$$H_0: \mu_{i,t} = a_{d,t} \tag{52}$$

$$H_1: \mu_{i,t} \neq a_{d,t} \tag{53}$$

Ahol:

$\mu_{i,t}$ az i-dik modell várható pontossága, t időeltolás esetén

$a_{d,t}$ a default pontosság, t időeltolás esetén.

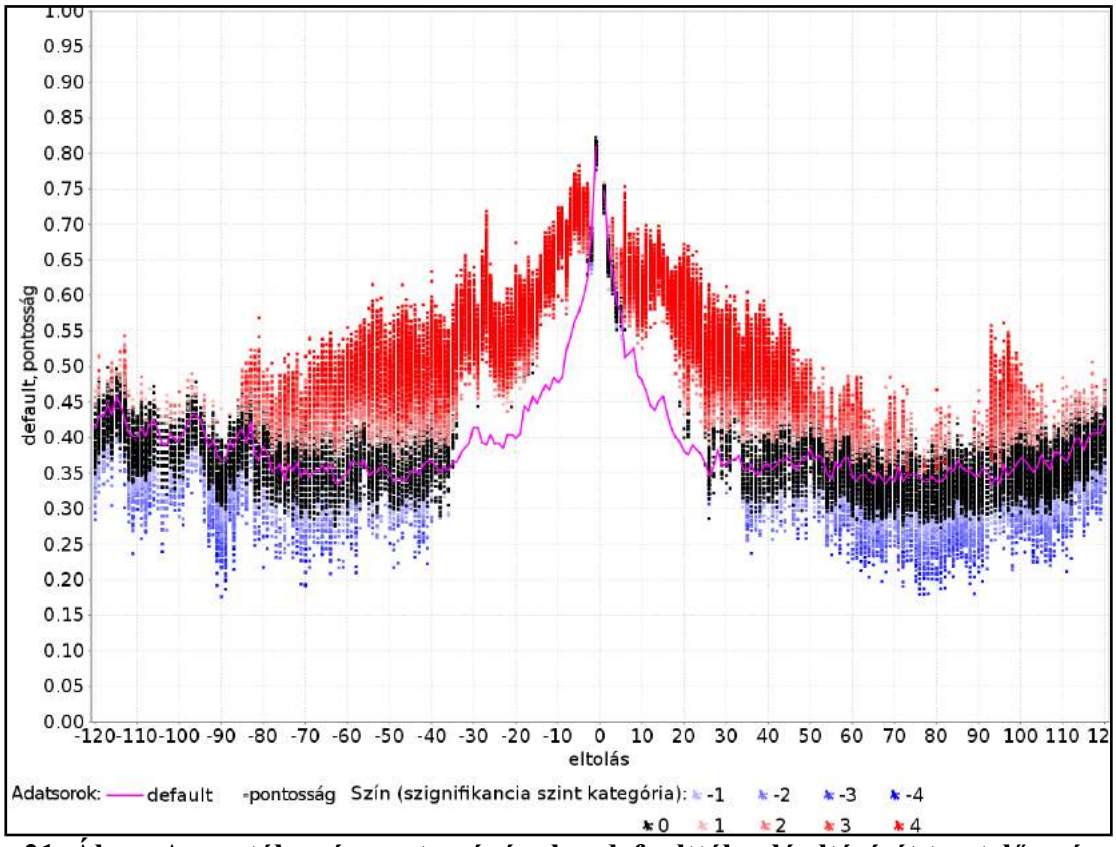
E tesztek eredménye alapján kilenc csoportba soroltam a modelleket:

- Azok a modellek, amelyek várható pontossága nem különbözött szignifikánsan a default értéktől, a 0-s jelű csoportba kerültek.
- Azon modellek, amelyek pontossága jobb volt a defaultnál 10%-os szignifikancia szinten, de 5%-on már nem, az 1-es csoportba kerültek.
- Azon modellek, amelyek 5%-on is magasabb pontosságot értek el, de 1%-on nem, azok a 2-es csoportba kerültek.
- Az 1%-on igen, de 0,1%-on nem szignifikánsak képezték a 3-as csoportot.
- Végül a 0,1%-on is szignifikánsak a 4-eset.
- Hasonlóképpen, abban az esetben, ha az SVM modell átlagos pontossága 10, 5, 1 és 0,1%-on kisebb volt a hozzá tartozó default modell pontosságánál, a modellt rendre a -1 -es, -2 -es, -3 -as, illetve -4 -es csoportba soroltam.

Minden időablak esetén, minden paraméterkombinációra tízszeres keresztvalidációval $21 \cdot 21 \cdot 10 \cdot 120 \cdot 2 = 1058400$ modell[110] tanítását végeztem el, de ennek csak tizede volt a t-próbák száma. A *21. ábrán* látható minden tesztelt modell pontossága, a default pontosságok és a tesztek eredményei. A vízszintes tengelyen az időablak hossza, a függőleges tengelyen a pontosság van felmérve. A különböző időablakokhoz tartozó default pontosságot a lila vonaldiagrammal ábrázoltam. Minden tanított modell egy színes pontnak felel meg, melynek vízszintes koordinátája az időablak, függőleges pedig a pontosság, de az ábrán nem láthatók a modell paraméterei. A pont színe a teszt eredménye alapján a fenti kategóriáknak felel meg, piros szín esetén a modell jobb a defaultnál, kék esetén rosszabb, az árnyalat a teszt erősségét mutatja. A defaulttól nem eltérő modellek színe fekete.

109 A teszteket kétoldalú próbával végeztem el, mivel első megközelítésben azt vizsgáltam, hogy van-e szignifikáns eltérés valamelyik irányban a default pontossághoz képest.

110 21 C, 21 gamma, 10 validáció, 120 hír utáni és 120 hír előtti időablak



21. Ábra: Az osztályozás pontosságának a defaulttól való eltérését tesztelő próbák eredményének változása az időablak hosszának függvényében

Forrás: saját szerkesztés

Ahogy a *20. ábra* bemutatásakor említettem, a hosszú időablakok esetén a default pontosság egyre inkább meghaladta az egyharmadot, tehát a *21. ábra* lila grafikonjának szélei *felemelkednek*, és e referenciaértékkel együtt a modellek pontossága is nőtt. Látható, hogy a modellek pontossága az időablak növelésével előbb romlik, majd javul, de már, mint a *21. ábra* mutatja, nagy részük nem haladja meg a defaultot. A *21. ábrán* az is jól látható, hogy 1–2 perces időablak esetén a legtöbb modell pontossága, hiába nagyon magas, nem különbözik a defaulttól. A különbségek akkor a legnagyobbak, amikor a hír előtti, vagy utáni időszakban kb. 5–30 perces eltolást veszünk. Ha az időablak mérete nagyjából 30–60 perc között van, még úgy tűnik, jelentős számú modell pontossága meghaladja a defaultot, de ezen túl már az alulmaradók vannak többségben.

A *21. ábráról* két tulajdonságot érdemes megállapítani, amely alapján a kutatási hipotézisek alátámaszthatók. Az egyik, hogy milyen nehéz olyan modellt készíteni, amelyik az ábra piros oldalán van az időtáv függvényében. A másik, hogy a modell pontossága és a default között mikor van a legnagyobb minőségi különbség – lényegében ez a *3. függelék* és a *3.4.2 alfejezetben* tárgyalt probléma. Előbbit az előrejelzés – azaz osztá-

lyozás – nehézségének nevezem majd, és a d-vel[111] jelölt, külön e dolgozathoz kifejlesztett mutatóval számszerűsítem. A nehézségmutató azt mutatja meg, hogy véletlenszerű paraméterbeállításokkal mi az esélye annak, hogy nem kapunk robusztus pontosságú modellt. Ha egy defaultnál pontosabb modell paraméterszomszédjai között van másik véletlennél pontosabb modell, akkor azt mondom rá, hogy robusztus pontosságú. Minden időablak esetén és minden szignifikancia szinthez képezzük a defaulttól szignifikánsan eltérő, és annál jobb pontosságú modellek halmazát, P-t, majd kivonjuk ebből azon modelleket, amelyeknek nem volt legalább egy paraméterszomszédja, amely szintén a P halmaz eleme – ez az O halmaz. Az így kapott $Q = P \setminus O$ halmaz számossága az előrejelezhetőség nehézségével inverz viszonyban van. Q számosságát a paraméterkombinációk számához viszonyítva, majd ezt egyből kivonva kapjuk a nehézségmutatót:

$$d_{t,\alpha} = 1 - \frac{|Q_{t,\alpha}|}{|G| \cdot |C|} \tag{54}$$

Ahol:

t : az időablak végének eltolása percekben mérve

α : a t-próba szignifikancia szintje

A modellek pontossága és a default közötti minőségi eltérést a *3. függelékben* levezetett, q_4-gyel jelölt, szintén e dolgozathoz kifejlesztett mutató segítségével számszerűsítem. Ez $z = 1$ paraméterrel, és az időablakokkal kapcsolatos kísérletekhez igazított jelölésrendszerrel a következő alakot ölti:

$$q_{i,t} = \log_{|I_t|+1}\left(\frac{\mu_{i,t} + \frac{1}{|I_t|}}{a_{d,t} + \frac{1}{|I_t|}}\right) \tag{55}$$

Ahol

t : az időablak végének eltolása percekben mérve

$m_i \in M$ az i-dik modell, ahol $m_i = (HU, (NP; nt), c_i, g_i)$, illetve $M = \{HU\} \times \{(NP; nt)\} \times C \times G$

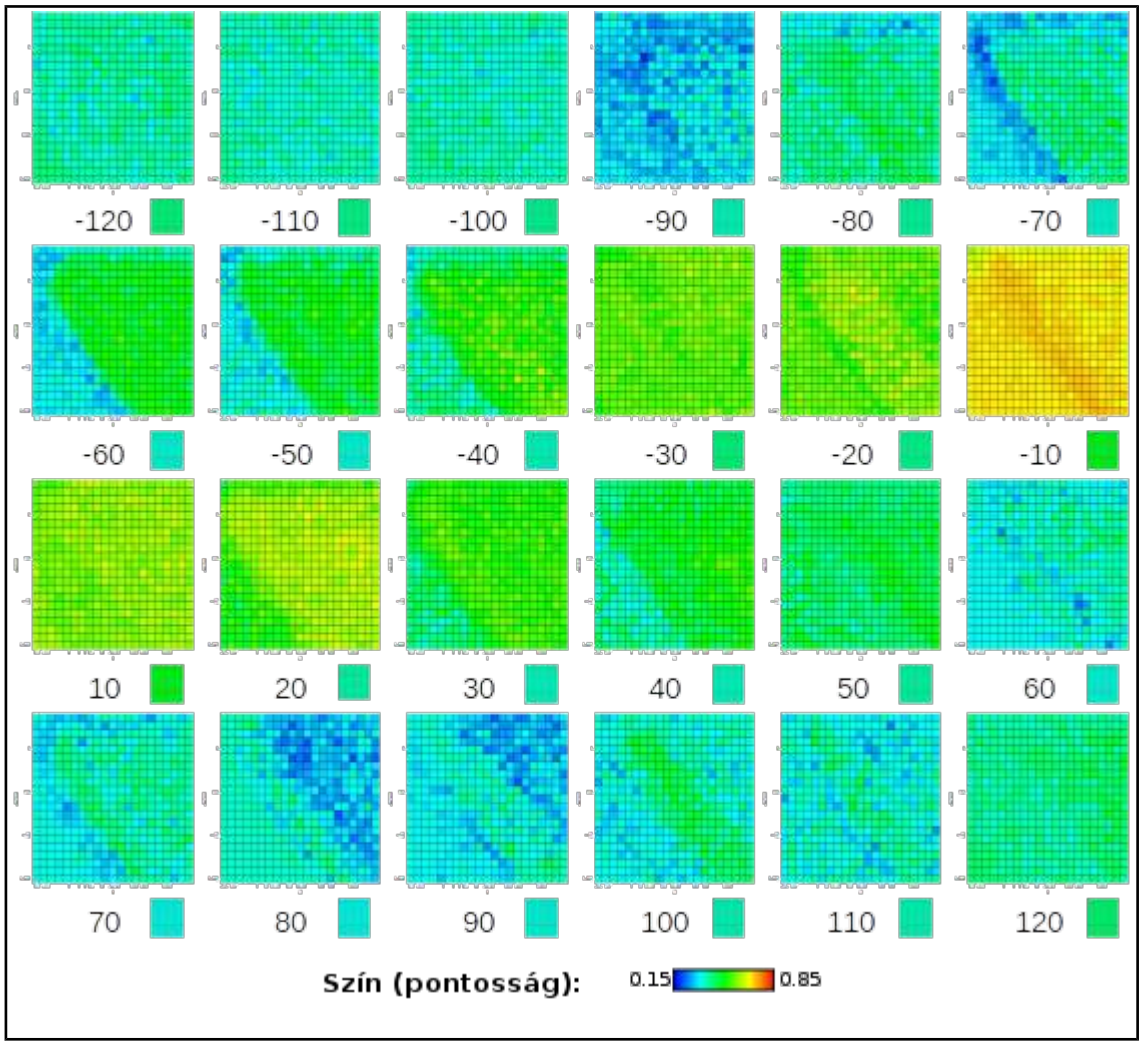
$c_i \in C$ az i-dik modell C paraméterének értéke

$g_i \in G$ az i-dik modell gamma paraméterének értéke

$0 \le \mu_{i,t} \le 1$ az i-dik modell várható pontossága t időeltolás esetén

$0 \le a_{d,t} \le 1$ a default pontosság t időeltolás esetén.

111 difficulty

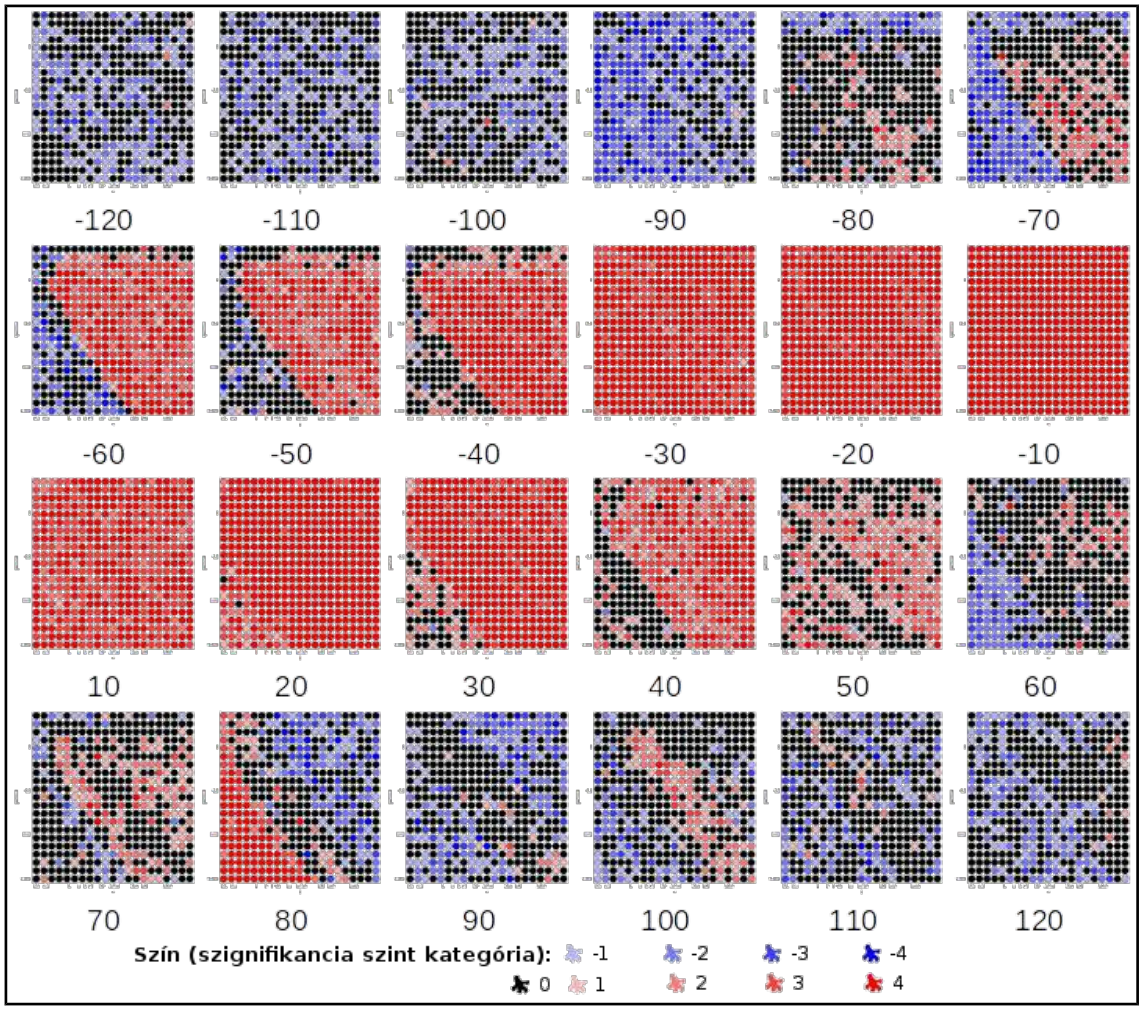


22. Ábra: Az osztályozás pontosságának változása az időablak eltolásának függvényében a C-gamma paraméterek terében

Forrás: saját szerkesztés

A *21. ábra* egyáltalán nem ad információt arról, hogy a paramétertérben hogyan alakul a modellek pontossága, ám sajnos a hőtérképes ábrázolási forma nem alkalmas a 240 különböző időablakra kapott eredmény tömör összehasonlítására, mivel indokolatlanul nagy terjedelmű volna. A *22. ábrán* minden tizedik időeltolás esetén láthatók a hőtérképek, de ezek színtartománya most egységesen 0,15–0,85.[112] Mivel a pontosságot minden eltolásnál más-más default értékhez kell viszonyítani, ezért a kis hőtérképek alatt, az eltolást percben jelző számoktól jobbra található egy-egy téglalap, benne a default pontossághoz tartozó színárnyalattal.

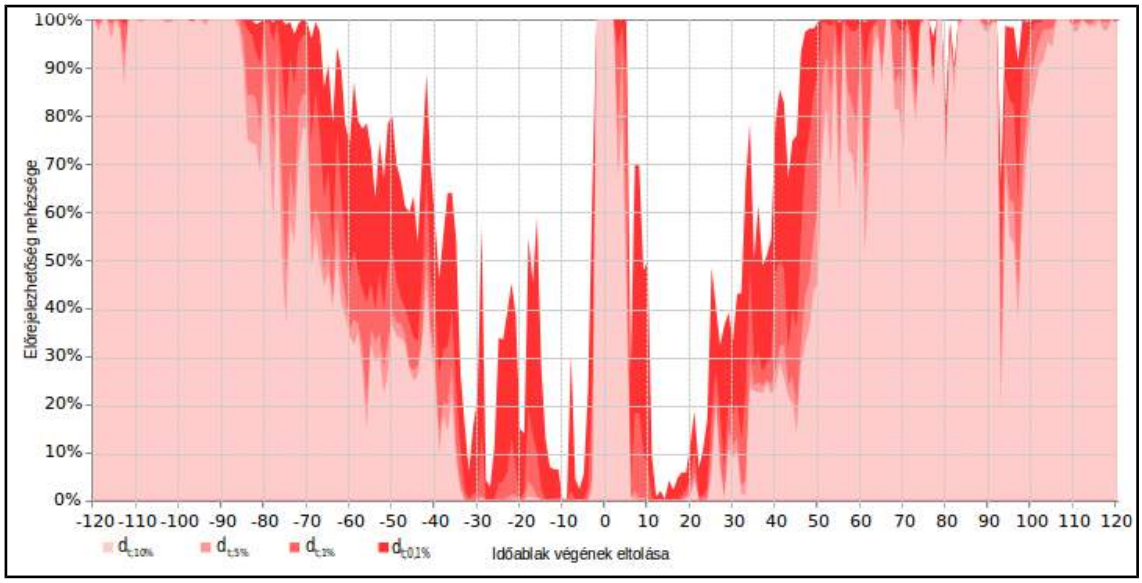
112 A 18. ábrán látható magyar nyelvű, (NP, nt) reprezentáció hőtérképe a 22. ábrán látható 20-as szám feletti térképnek felel meg, csak a színtartomány különböző.



23. Ábra: Az osztályozás pontosságának a defaulttal való egyezését tesztelő próbák eredményének változása az időablak eltolásának függvényében a C-gamma paraméterek terében

Forrás:saját szerkesztés

A *22. ábrán* látható, hogy ahogy az időtávot fokozatosan 10-ről 120 percre növeljük, úgy a paramétertérben kibontakoznak bizonyos területek, ahol jelentősen pontatlanabbak a modellek a defaultnál, ám nagyjából a 90–100. perc környékén ez a tendencia visszafordulni látszik, de a *21. ábra* alapján sejthető, hogy ezek jelentős része nem szignifikáns. A t-próbák eredményét hasonló hőtérképeken – *23. ábra* – ábrázolva jobban látszik, hogy valóban nincs szignifikáns eltérés a pozitív irányban, ez a piros árnyalatú területek eltűnésében nyilvánul meg. A d mutatók számolása e térképek segítségével szemléltethető is: ki kell számolni az összes nem egyedülálló piros árnyalatú pont által lefedett területet, majd el kell osztani a teljes rács területével. Ha minden időablakra kiszámoljuk, azt tapasztaljuk, hogy a mutató értéke egy jól látható tendenciát mutat, de több kiugrás és visszaesés tapasztalható benne, ezt vizualizálja a *24. ábra*.



24. Ábra: Az előrejelezhetőség nehézségének alakulása az időablak eltolásának függvényében

Forrás: saját szerkesztés

A *24. ábra* az előrejelezhetőség nehézségét számszerűsítő d-mutatókat ábrázolja a négy vizsgált szignifikancia szintre. Mivel a kisebb szignifikancia szint esetén kevesebb vagy ugyanannyi modellt tekinthetünk a defaulttól eltérő pontosságúnak, így a 10%-os szinthez tartozó d-mutató a legalacsonyabb, míg a 0,1%-oshoz tartozó a legmagasabb. Előbbi tehát inkább optimista, utóbbi pedig inkább pesszimista módon jellemzi a mért mennyiséget. Látható, hogy kb. ± 5 perc esetén szinte lehetetlen a pontos előrejelzés, még az optimista mérték szerint is. Hasonló igaz az egy óránál hosszabb időtávokra is.

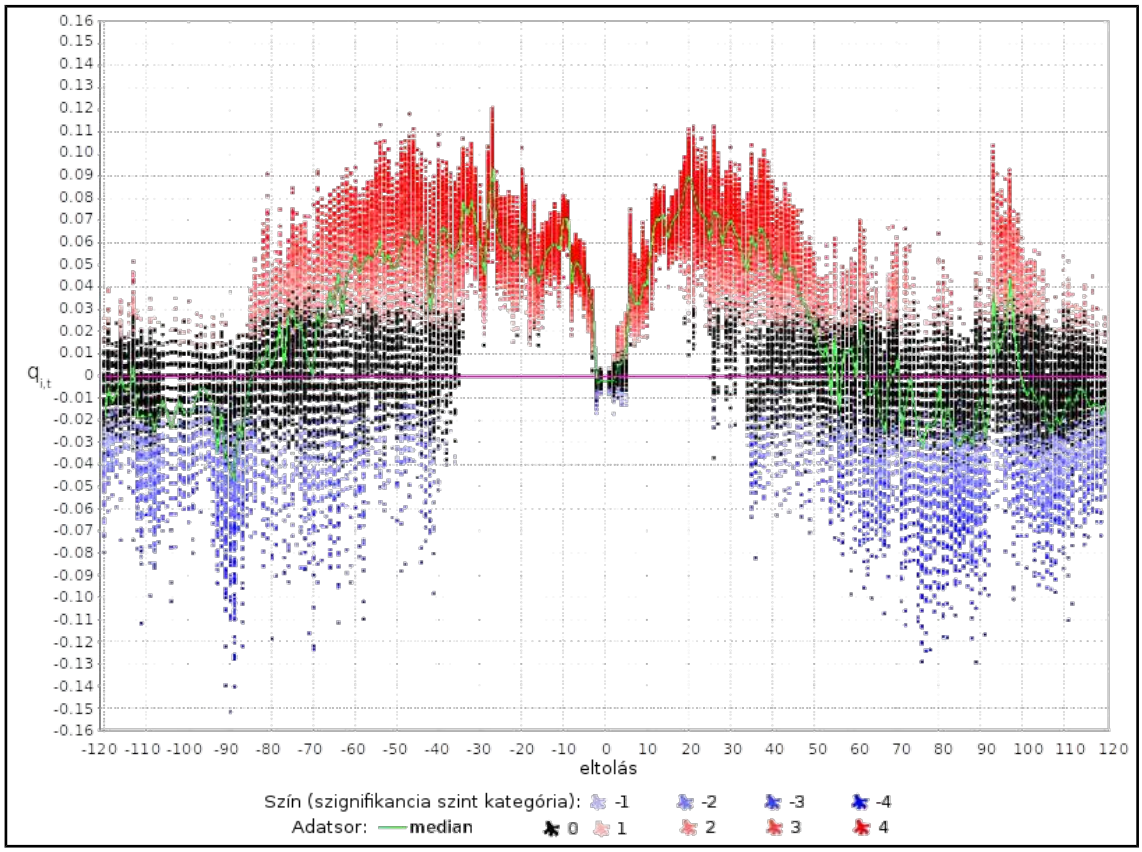
Ha minden időablakra kiszámoljuk minden modellre a minőségmutatót is – *25. ábra* –, azt tapasztaljuk, hogy a minőség $+20$, illetve -27 percig növekszik, majd a szóródás növekedése mellett csökkenő tendenciát látunk. Ennek egyik oka, hogy kb. ± 2 perc esetén a default pontosság nagyon magas volt, és a modellek átlagos pontossága gyakorlatilag megegyezett vele, majd a default pontosság gyorsan csökkent, miközben a modellek átlagos pontossága nem csökkent ilyen ütemben. Az időablak további növelésével a default csökkenése lelassult, a pontosságé viszont nem.

A legjobb minőségű modell a -27-es eltolásnál található, ez az egyetlen, amely meghaladta a 0,12-es értéket[113]. További 26 modell minősége esett a $[0{,}11;0{,}12[$ intervallumba ugyanennél az eltolásnál és egy-egy -46, -47 és -54-nél, hat pedig valamely, a következő halmazba eső eltolásnál: $t \in \{20;21;26\}$. A legjobb minőségű mo-

113 $q_{(C=113{,}32;\,\gamma=0{,}03017),\,t=-27} = 0{,}12066$

dellek kapcsán még a $[0{,}1\,;0{,}11[$ intervallumba esők megoszlását jellemzem nagy vonalakban, melyekből 311 darab van. Ebből a 128 a -27 -es eltolásnál található, 20, 21 és 26 percnél pedig rendre 45, 21 és 18 darab, végül -46, -47, -54 és -32 esetén rendre 15, 13, 13 és 12 darab. A legrosszabb modell a -90 -es eltolásnál található, $-0{,}1516$ -os mutatóval, ezután következett a -89 -es és -91 -es időablak, $-0{,}13$ -nál kisebb mutatóval. A $[-0{,}13\,;-0{,}12[$ intervallumba 15 modell esett, melyekből 8 a következő halmazba: $t\in\{-91,-89,-87,-70\}$ és 7 a következőbe $t\in\{76,77,78,89\}$.

A szakirodalom egyes modelljeire kiszámolt q-mutatókkal – lásd *3.4.2 alfejezet* – összevetve az én legjobb modelljeim alulmaradnak Mittermayer & Knolmayer (2006a) 0,2 fölötti értékével szemben, de Peramunetilleke (1997), Cho & Wüthrich (1999) és Cho et al. (1999) 0,1 körüli értékét elérik, meghaladják.



25. Ábra: Az előrejelzés minőségének alakulása az időablak eltolásának függvényében

Forrás: saját szerkesztés

A következőkben a fent bemutatott kísérletek eredményét a hipotéziseim szempontjából fogom kiértékelni.

4.1. A sajtóközlemények hatásának kimutatása

Mielőtt bonyolultabb összefüggéseket keresnénk, be kell látni, hogy a modell alkalmas a szöveges információk alapján történő árfolyam-előrejelzésre. Ahogy Gidófalvi és Elkan (2003), illetve Groth és Muntermann (2009) esetében láttuk, a modell teljesítménye alulmaradhat a véletlen találgatás várható pontosságához, a defaulthoz képest. Ha ennek oka csak annyi, hogy nem megfelelően reprezentáltuk a szövegeket, vagy az osztályozó paraméterei éppen kedvezőtlenül voltak beállítva, visszaléphetünk az adatbányászati folyamatban, és javíthatunk a modellen. Ha viszont a probléma így sem küszöbölhető ki, akkor ki kell zárni annak a lehetőségét is, hogy rossz időablakot alkalmaztunk, és ha végül nem járunk eredménnyel, arra a következtetésre kell jutnunk, hogy a sajtóközlemények szövegével nem lehet árfolyam-modellezést végezni a magyar tőzsdén. A fent említett negatív példák ellenére az irodalomban nagyrészt nem erre a következtetésre jutottak, úgyhogy első hipotézisemet ennek megfelelően fogalmaztam meg.

H1: A sajtóközlemények szövegének felhasználásával a default modellnél nagyobb pontosságú előrejelzés készíthető a BÉT prémium kategóriás részvényeire.

A hipotézist elvetem, ha az olyan modellek aránya, amelynek tízszeres keresztvalidáción mért átlagos pontossága szignifikánsan eltér a default modell várható teljesítményétől, nem haladja meg rendre a 0, 1, 5, illetve 10%-ot. E négy szint egyre szigorúbb követelményt támaszt a modellek elfogadásának. Először különböző szignifikancia szintekre egymintás t-próbával megállapítom a pontosságok és a default pontosság egyezőségét, majd szintenként kiszámolom az eltérő modellek arányát.

Az első lépésben a default modell várható pontossága a legnagyobb elemszámú kategóriába tartozó megfigyelések számának aránya a teljes mintán belül, azaz 38,06%. A null- és alternatív hipotézisek a következők minden $m_i \in M$ ($|M|=3528$) modellre:

$$H_{0,i}: \mu_i = 0{,}3806 \tag{56}$$

$$H_{1,i}: \mu_i \neq 0{,}3806 \tag{57}$$

Ahol

$m_i \in M$ az i-dik modell, ahol $m_i = (l_i, r_i, c_i, g_i)$, illetve $M = L \times R \times C \times G$

$l_i \in L$ az i-dik modell korpuszának nyelve, $L = \{HU, EN\}$ a nyelvek halmaza

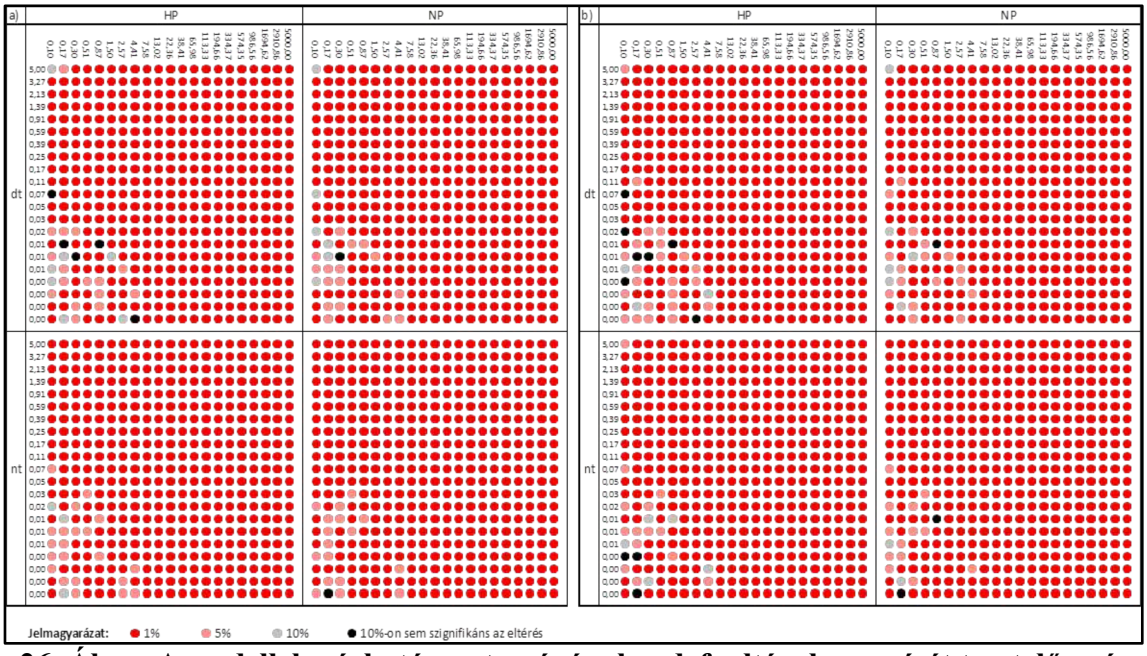
$r_i \in R$ az i-dik modell szövegreprezentációjának típusa,

$R = \{(HP; dt); (HP; nt); (NP; dt); (NP; nt)\}$ a reprezentációk halmaza

$c_i \in C$ az i-dik modell C paraméterének értéke, meghatározott értékkészletből

$g_i \in G$ az i-dik modell gamma paraméterének értéke, meghatározott értékkészletből

$0 \leq \mu_i \leq 1$ az i-dik modellre mért átlagos pontosság.



26. Ábra: A modellek várható pontosságának a defaultéval egyezését tesztelő próbák eredményei

Forrás: saját szerkesztés

A tesztek eredményét láthatjuk a *26. ábrán*, amely a *18. ábrán* látható hőtérképek pontjainak megfelelő szignifikancia szinteket mutatja. Piros színnel az 1%-on szignifikáns eltérések láthatók, 3339 eset a 3528 modellből. Rózsaszínnel és pirossal az 5%-on szignifikáns eltérés, mely 3476 modellnél található, 10%-on – piros, rózsaszín és szürke – pedig 3508 modell pontossága tekinthető különbözőnek. Bármelyik szignifikancia szintet is tekintjük, a defaulthoz képest a legtöbb modell pontosabb, hiszen a leggyengébb modell pontossága is meghaladta a 39%-ot.

13. Táblázat: A H1 hipotézis különböző feltételek melletti tesztelésének eredményei

		t-próba szignifikancia szintje		
		1%	**5%**	**10%**
		szignifikáns modellek aránya		
		94,64%	**98,53%**	**99,43%**
Elfogadási alsó küszöb az arányokra	**0%**	IGAZ	IGAZ	IGAZ
	1%	IGAZ	IGAZ	IGAZ
	5%	IGAZ	IGAZ	IGAZ
	10%	IGAZ	IGAZ	IGAZ

Forrás: saját szerkesztés

A második lépés, hogy e modellek arányát összehasonlítjuk a 0, 1, 5 és 10%-os küszöbértékekkel. 1%-os szignifikancia szinten 94,64%-os arányt kapunk, 5%-ra 98,53%-ot, 10%-ra pedig 99,43%-ot. Ezt a *13. táblázatban* láthatjuk, amelyben az IGAZ érték azt jelzi, hogy a megfigyelt arány meghaladja a küszöbértéket, és elfogadjuk a hipotézist.

A H1 hipotézist az összes szignifikancia szint és az összes alsó küszöb tekintetében el kell fogadni, azaz a sajtóközlemények szövege alapján 20 perces időablakban a véletlenhez képest pontosabb előrejelzés adható a BÉT Prémium kategóriás részvények árfolyamára. Tehát mindkét nyelven, mindegyik szövegreprezentáció és majdnem minden C-gamma paraméterkombináció esetén kinyerhető valamilyen mértékű, az árfolyam szempontjából releváns információ.

4.2. A közzétételt megelőző időablak árfolyammozgásai

A ± 120 perc hosszúságú időablakok esetén a vizsgálható dokumentumok száma a bejelentés előtti időszakban nagyjából 30-cal – tehát kb. 25%-kal – kevesebb, mint a bejelentés utáni időszakban (lásd *8. ábra*). Ha ugyanezt az összes kereskedési időn belül közzétett sajtóközlemény számához viszonyítjuk, amely 158, akkor azt mondhatjuk, hogy a sajtóközlemények több mint 40%-át a nyitás utáni két órában teszik közzé. Tehát a sajtóközleményeket gyakrabban teszik közzé a kereskedési idő elején, mint a végén. Ez azt jelenti, hogy jóval több hírt kell kizárni a mintából azért, mert nincs megfigyelhető árfolyam két órával a hír publikálása előtt, amikor a tőzsde még zárva volt. Ennek az egyenlőtlen eloszlásnak az is lehet az oka, hogy a sajtóközleményeket mégsem azonnal teszik közzé, amint az információról tudomást szereznek a kibocsátók. Ennek több oka is elképzelhető, amelyek közül például az egyik a sajtóközlemény gondos megírásához, szerkesztéséhez, vállalaton belüli jóváhagyásához szükséges információszerzési és -továbbítási munkaidő, mely eltolhatja a közlemény megjelenését. A tőzsdei nyitvatartási időn kívüli események publikálására csak közvetlenül a következő nyitás előtt van lehetőség, viszont ezeket a közleményeket kizártam a mintából, tehát ez nem lehet az oka. Mielőtt a publikálás haladéktalanságát vizsgálnák, idézzünk fel egy határidőt a tőzsdeszabályzatból: kereskedési időben az információk a feltöltést követő 60. percben kerülnek publikálásra a BÉT honlapján, a különleges esetektől eltekintve. (Lásd *3.2.2 alfejezet*.) Feltételezem, hogy ha nem töltik fel haladéktalanul a hírt a rendszerbe, akkor -60 percnél hosszabb távon is jó előrejelzést, becslést ad a modell.

H2: Nem lehet jó minőségű, illetve robusztus hírbányászati modellt készíteni olyan időablakra, amely korábbra nyúlik vissza, mintsem a sajtóközlemény átmenne a közzétételi folyamaton.

A hipotézist elfogadom, ha a 10, 5, 1 vagy 0,1%-os szignifikancia szinten számolt d mutató értéke[114] -60 percnél jelentősen hosszabb időablakok esetén magas vagy növekszik, miközben a q mutató értéke alacsony vagy romlik. A d és q mutatók kiszámítását a fejezet elején bemutattam, megvizsgálom a *24.* és *25. ábrákat*, illetve a mögöttes adatokat, és megállapítom a mutatók növekedési-csökkenési tendenciáit, végül levonom a következtetéseket.

A hír publikálása előtti 9, illetve 10 perc árfolyammozgását a legkönnyebb magyarázni a hír szövegének ex-post ismeretében $d_{-9;0,1\%}=d_{-10;0,1\%}=0{,}23\%$. A pesszimista mérték 20% alatt van még a hír előtti 5–7, 9–14, 19–20, 26–28, valamint a 31–33 perces tartományokban is. A *24. ábrán* látható, hogy a hír előtti időablakok esetén erős ugrások tapasztalhatók a d mutatókban mielőtt egyértelműen elkezdenének romlani. Mindenesetre 1%-os szinten a mutató a -4 és -34 között 20% alatti. 5%-os szinten -3 és -35 között, valamint -37 -nél, -39 -nél és -56 -nál. És végül az előrejelzés nehézségét legoptimistábban mérő, 10%-os szignifikanciához tartozó mutató értéke -3 és -39 között, valamint -56 -nál volt 20% alatt. Ennél hosszabb intervallumokra a mutató fokozatosan romlik, és -120 környékén megközelíti a 100%-ot minden szignifikancia szinten.

A *25. ábrán* látható, hogy a q mutató növekvő tendenciát mutat egészen a -27 -es időeltolásig, amelynél a legtöbb jó minőségű modell található, és a medián is itt éri el maximumát, 0,0946-ot. Ennél hosszabb időablakra a medián csökken, és a -60 -at néhány perccel meghaladó időablakokra a medián modell már nem pontosabb szignifikánsan a defaultnál.

Az eredmények alapján a H2-es hipotézist elfogadjuk, mert a -60 -as időablakon túl az előrejelzés egyre nehezebb, és a medián modell minősége is egyre romlik, a tőzsdei szabályzatban meghatározott közzétételi időn túl nem elég robusztus a hírbányászati modell pontossága.

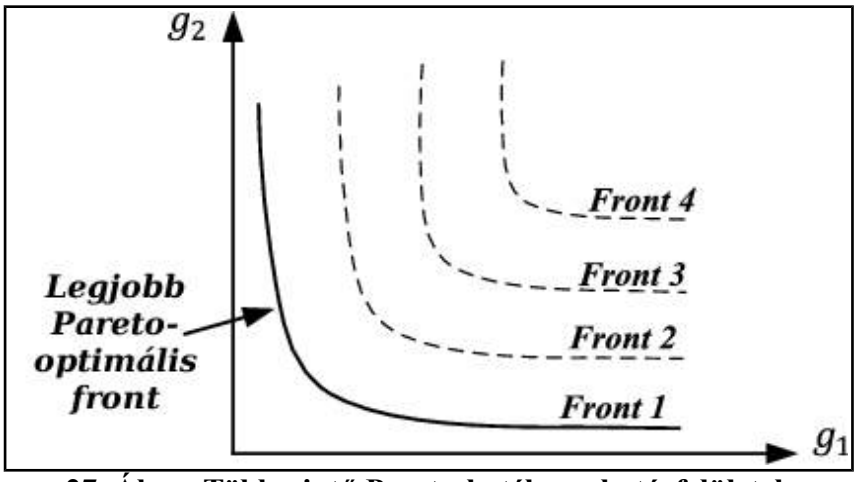
114 Azonos időablak esetén a szignifikanciaszint növelésével csökken a d mutató értéke, tehát az előrejelzés nehézségét illetően a 0,1%-os szignifikancia szinthez tartozik a legpesszimistább érték.

4.3. Az optimális időablak meghatározása

Ahogy a *3.1. alfejezetben* már felvetettem, a különböző motivációk miatt a hírek különböző időtávokon befolyásolhatják a részvényárfolyamot. A dolgozatban a hírek közvetlen közelében bekövetkező árváltozásokat vizsgálom, melynek relevanciáját az is alátámasztja, hogy a tőzsdei hírbányászati szakirodalomban a modellek jelentős része ugyancsak néhány perces esetleg órás hatásokat elemez, de konkrét esetek is megerősítik, hogy az időtáv ebben a nagyságrendben mozog (Kovács 2014b; 2015). A kérdés az, hogy ezen belül melyik időtáv az, amelyet a leginkább érdemes vizsgálni?

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

A hipotézist elfogadom, ha a 10, 5, 1 vagy 0,1%-os szignifikancia szinten számolt d mutató értéke[115] minimumhoz közeli értéket vesz fel, és a q mutató értéke egyúttal maximumhoz közeli valamely időablak esetén. A két mutató mint célfüggvény, az időablak mint inputváltozó van jelen a feladatban. A megoldáshoz szükséges d és q mutatók kiszámítását a fejezet elején bemutattam, a következőkben megvizsgálom a *24.* és *25. ábrákat*, és célfüggvényenként parciálisan jellemzem az adatokat. Második lépésben az egyváltozós vizsgálat után két célfüggvény szerint – a q mutató mediánját külön-külön mind a négy d mutatóval párosítva – megkeresem a Pareto-hatékony időeltolásokat a *27. ábrán* illusztrált módszerrel. Minden esetben három-három különböző szintű Pareto-hatékony felületet definiálok Esmaeili et al. (2016) nyomán. Először megkeressük a Pa-



27. Ábra: Többszintű Pareto-hatékony határfelületek

Forrás: (Esmaeili et al. 2016)

115 Azonos időablak esetén a szignifikanciaszint növelésével csökken a d mutató értéke, tehát az előrejelzés nehézségét illetően a 0,1%-os szignifikancia szinthez tartozik a legpesszimistább érték.

reto-hatékony pontokat a teljes halmazon, ezek képezik az első szintet, majd az első szintű Pareto-optimális megoldások által dominált pontok halmazán keressük meg a Pareto-hatékonyakat, ez lesz a második szint, ezután a második szint által dominált pontokon folytatjuk a keresést, és így tovább. A harmadik lépésben a kapott eredmények alapján elfogadom vagy elvetem a hipotézist.

Az első lépés kapcsán a negatív időablakok mutatóit már áttekintettem a *4.2 alfejezetben.* A pozitív időbeli eltolás d mutatójáról az mondható el a *24*. ábra alapján, hogy legkönnyebben a hír utáni 14 perc árfolyammozgására készíthetők modellek, 0,1%-os szignifikancia szint mellett a d mutató 0,23%, amely a minimális érték, de a 11–24 perces tartományban is 20% alatti a pesszimista mutató. Az időablak hosszának növelésével az előrejelzés egyre nehezebb[116]. Ha 1%-os szignifikancia szint mellett számoljuk ki a mutatót, akkor 12–17 perc között 0%-ot kapunk[117], azaz tetszőleges paraméterekkel jobb modellt építhetünk a defaultnál. Egyébként szinte végig 20% alatti az 1%-os nehézség a 6–32 tartományban. 5%-os és 10%-os szignifikancia szint mellett szinte ugyanazokat a tartományokat kapjuk, mint 1%-ra.

A *25. ábra* áttekintése kapcsán láttuk, hogy a -27 -es eltolásnál van a legjobb minőségű modellek zöme, továbbá a medián is itt volt a maximális. Ennél valamivel kisebb q értékeket produkáltak a 20, 21 és 26-os időablak modelljei, és ezeket követték a -46, -47, -54 és -32 eltolás modelljei. A különböző időablakokhoz tartozó medián pontosság legnagyobb értékei kissé másként alakulnak, ahogy a *14. táblázatban* látható, mert a hosszabb időszakokon nagyobb volt a szóródás.

14. Táblázat: A tíz legnagyobb medián pontosságú időeltolás

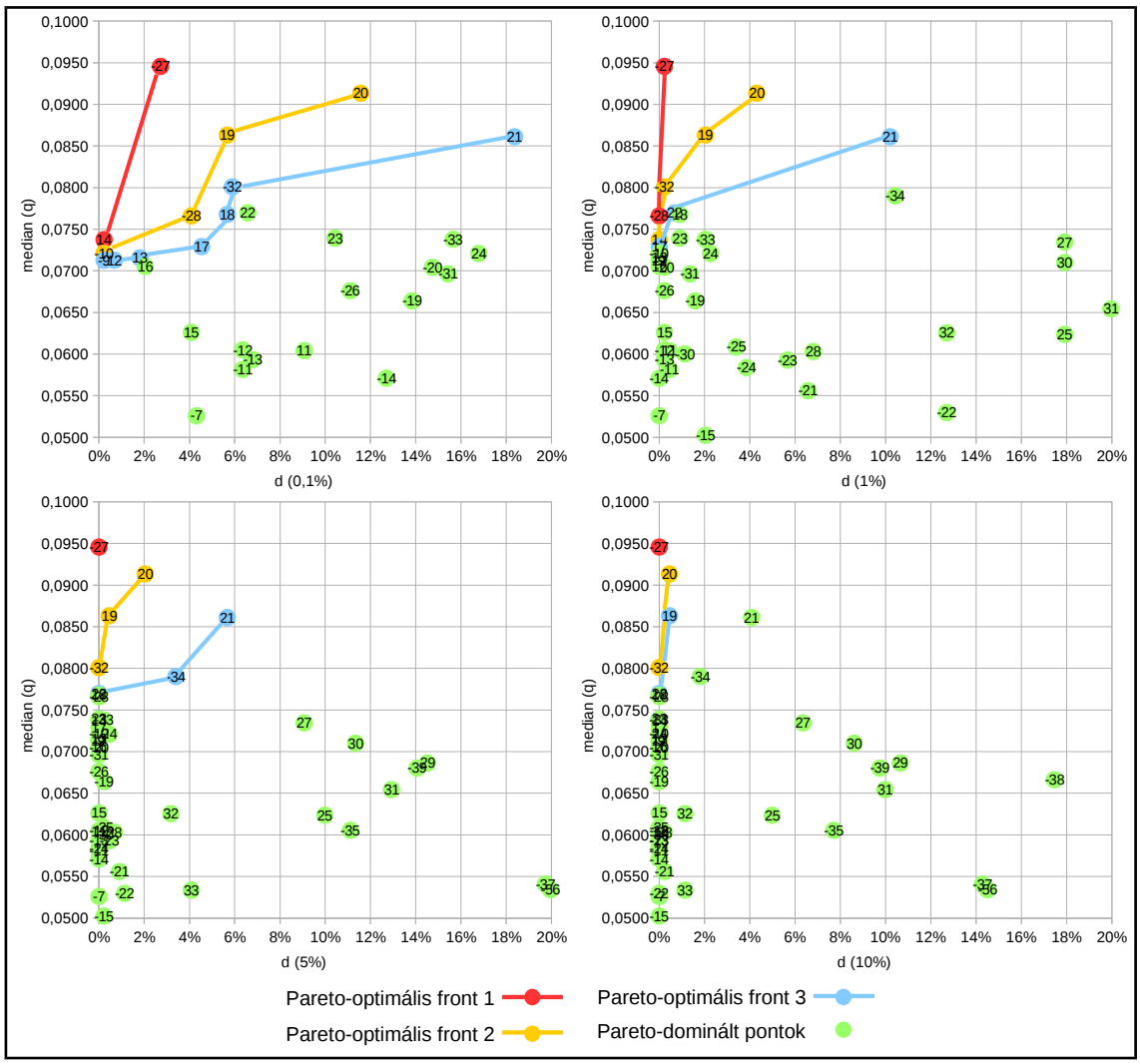
eltolás	-27	20	19	21	-32	-34	22	18	-28	26
medián (q)	0,0946	0,0913	0,0863	0,0861	0,0801	0,0790	0,0770	0,0768	0,0766	0,0754

Forrás: saját szerkesztés

Három szintet alkalmazva keresem a Pareto-optimális, -120 és 120 közötti időablakokat a maximalizálandó medián q mutató és a minimalizálandó d mutatónak a terében. Ezt a d mutató mind a négyfajta kiszámításával megvizsgáltam, az eredmények a

116 Szokatlan, hogy a 93. percben a nehézség kissé visszaesik, illetve a minőség is javul. Erre jelen vizsgálatok nem szolgáltatnak magyarázatot, hiszen a 4. függelékben bemutatott khi-négyzet próbák eredménye azt mutatja, hogy nagyon erős asszociációs kapcsolat van a néhány perccel rövidebb időablakok címkéivel, azoknál mégsem volt megfigyelhető a javulás. Valószínűleg csak koincidencia, hogy kb. a 93. percnél kezdenek a negatív címkéjű megfigyelések túlsúlyba kerülni a mintában, ami a 9. ábrán is látható.

117 Negatív irányban -5 és -10 között van ugyanilyen összefüggő tartomány.



28. Ábra: Az időablakok Pareto-dominancia viszonya a négyféle célfüggvény-kiszámítás esetén

Forrás: saját szerkesztés

15. táblázatban és a *28. ábrán* láthatók. Ez alapján a −27 perces időeltolás minden esetben az első – azaz hagyományos értelemben vett – Pareto-hatékony felületen található. A második frontnak stabil tagja a 20 perces időablak. A 19-es és a −32 -es időablak három-három esetben a második szinten, egy-egy esetben a harmadikon volt. A harmadik vonalban a 22-es és a 21-es időablakok gyakoriak.

A H3-as hipotézist az eredmények alapján el kell fogadni. Láthattuk, hogy az előrejelzés nehézsége a publikálás előtti 27 percben az egyik legalacsonyabb, miközben a minőség is itt a legmagasabb. A valamivel hosszabb, –32 perces időablakra is jobban működik a modell, mint a többire. A publikálás utáni 19–22 perces tartományban is könnyű az előrejelzés, és a magas minőségű modellek közül jelentős számú, kb. 70 da-

rab a 20, illetve 21 perces időablaknál található. Ez alapján tehát az információ a publikálás előtti fél órában kezd beépülni – általában ez fél órával követi a hír feltöltését –, majd az optimális előrejelezhetőség a publikálást követő 19–22 percig tart. Mivel rövid időtávokon a modell számára az árfolyamváltozás véletlenszerű, így közben romlik az osztályozó teljesítménye. Gidófalvi (2001; Gidófalvi & Elkan 2003) ±20 perces eredményével összehasonlítva közel másfélszeres eltérést tapasztaltam negatív irányba, míg pozitív irányba meg tudtam erősíteni ezt a számot.

15. Táblázat: Az első három Pareto-hatékonysági szintbe tartozó időablakok a négyféle célfüggvény-kiszámítás esetén

időablak	d (0,1%)	medián (q)
-27	2,72%	0,0946
14	0,23%	0,0738
-28	4,08%	0,0766
-10	0,23%	0,0721
19	5,67%	0,0863
20	11,56%	0,0913
-32	5,90%	0,0801
-9	0,23%	0,0713
12	0,68%	0,0713
13	1,81%	0,0716
17	4,54%	0,0729
18	5,67%	0,0768
21	18,37%	0,0861

időablak	d (1%)	medián (q)
-28	0,00%	0,0766
-27	0,00%	0,0946
-32	0,00%	0,0801
14	0,00%	0,0738
19	0,02%	0,0863
20	0,04%	0,0913
17	0,00%	0,0729
21	0,10%	0,0861
22	0,01%	0,0770

időablak	d (5%)	medián (q)
-27	0,00%	0,0946
-32	0,00%	0,0801
19	0,00%	0,0863
20	0,02%	0,0913
-34	0,03%	0,0790
21	0,06%	0,0861
22	0,00%	0,0770

időablak	d (10%)	medián (q)
-27	0,00%	0,0946
-32	0,00%	0,0801
20	0,45%	0,0913
19	0,45%	0,0863
22	0,00%	0,0770

Forrás: saját szerkesztés

4.4. Az információ nyelvi kódolásának jelentősége

A tőzsdei sajtóközleményeket a Prémium kategóriába tartozó vállalatoknak angolul és magyarul is kötelező közzétenni. Mint tudjuk, két nyelv között a szavak nem csak egy-az-egyhez kapcsolatban állhatnak egymással, nyelvenként eltérések vannak a szinonimák, a homonimák stb. terén. Az állandósult szókapcsolatok fordításakor sem feltétlenül ugyanazokból a szavakból állnak az egymásnak megfeleltethető kifejezések, ráadásul a szavak külön és egybeírása is eltérő lehet. Nyilván mindkét nyelv alkalmas ugyanannak az információnak a kifejezésére, de elképzelhető, hogy egy nyelven jobban lehet reprezentálni az információkat mint egy másikon.

H4: Az azonos sajtóközlemények angol és magyar nyelven közzétett változataival készített modellek pontossága között nincs szignifikáns különbség.

A hipotézist elfogadom, ha az olyan angol–magyar modellpároknak az aránya, amelyek minden más tényezőt tekintve azonosak és pontosságuk szignifikánsan nem különbözik, eléri rendre a 90, 95, 99, illetve 100%-ot. Ezen kívül, ha az arány nem éri el a küszöbértéket, akkor is elfogadom – avagy nem vetem el – a hipotézist, ha az eltérések nem szisztematikusan csak az egyik nyelvi változat javára fordulnak elő – azaz nincs preferált nyelv. A négy küszöbérték a megadott sorrendben egyre szigorúbb követelményeket támaszt az elfogadással szemben. Első lépésben kétmintás t-próbákat készítek a *18. ábra* angol és magyar nyelvű oldalán lévő 4-4 hőtérkép azonos pozíción lévő modelljeinek pontosságának egyezőségére. Második lépésben kiszámolom minden szignifikancia szintre a nem különböző pontosságú modellek arányát. Harmadik lépésben minden szignifikancia szintre megszámolom, hány esetben volt nagyobb pontossága a magyar, illetve az angol nyelvű modellnek. Negyedik lépésben összehasonlítom a kapott arányokat a küszöbértékekkel, és a magyar és angol preferenciák viszonyával.

Első lépésben meg kell állapítani, hogy a minden más tényezőt tekintve azonos angol–magyar modellpárok pontossága szignifikánsan eltér-e egymástól. A null- és alternatív hipotézisek tehát a következők minden $n \in N$ ($|N| = 1764$) modellpárosítás esetén:

$$H_{0,n}: \mu_i = \mu_j \tag{58}$$

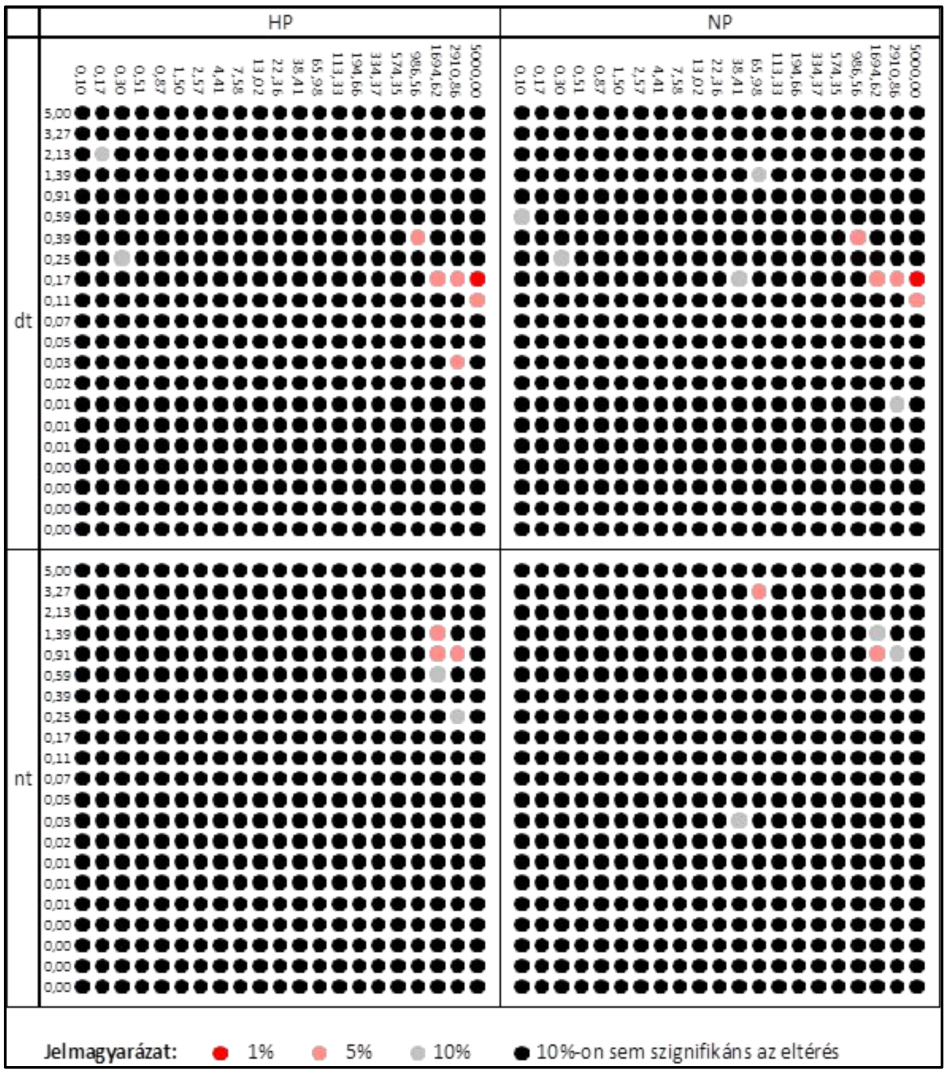
$$H_{1,n}: \mu_i \neq \mu_j \tag{59}$$

Ahol:

$N = \{n \in (M \times M) : n = (m_i, m_j), l_i \neq l_j, c_i = c_j, g_i = g_j, r_i = r_j\}$ az azonos paraméterű angol-magyar modellpárok halmaza.

A tesztek eredményét láthatjuk a *29. ábrán*, a második lépés, hogy kiszámoljuk az egyező modellek arányát. 1%-on 1762 modellpár esetén nem található szignifikáns eltérés az 1764-ből, ez 99,89%-nak felel meg[118], 5%-on[119] 1748, azaz 99,09%, 10%-on[120] 1736, azaz 98,41% a nem különböző pontosságú párok száma, illetve aránya. A harmadik lépés, hogy azokban az esetekben, ahol a küszöböt nem értük el, ellenőrizzük, hogy a szignifikánsan eltérő értékek az angol vagy a magyar változat javára figyelhetők meg. Ezt mutatja a *16. táblázat*, amely szövegreprezentációnként és összesen is tartalmazza, hogy különböző szignifikancia-szintek mellett hány olyan modellpár volt, amelyben a magyar – HU –, illetve az angol – EN – korpuszon ért el nagyobb pontosságot a modell.

118 Ezek a nem piros pontok az ábrán.
119 A nem piros vagy rózsaszín pontok.
120 A fekete pontok.



29. Ábra: Az angol és a magyar nyelvű inputra tanított modellek pontosságának egyezését tesztelő t-próbák eredményei

Forrás: saját szerkesztés

16. Táblázat: A pontosabb modellek megoszlása nyelvenként

	reprezentáció	**(HP; dt)**		**(HP; nt)**		**(NP; dt)**		**(NP; nt)**		**Összesen**	
	nyelv	**HU**	**EN**	**HU**	**EN**	**HU**	**EN**	**HU**	**EN**	**HU**	**EN**
t-próba szignifi-kancia szintje	**10%**	7	1	5	0	7	3	5	0	24	4
	5%	6	0	3	0	5	0	2	0	16	0
	1%	1	0	0	0	1	0	0	0	2	0

Forrás: saját szerkesztés

A negyedik lépés, hogy az azonos modellek arányát összehasonlítjuk a 90, 95, 99 és 100%-os küszöbértékekkel. Ezt a *17. táblázatban* láthatjuk, amelyben az IGAZ érték azt jelzi, hogy a megfigyelt arány eléri a küszöbértéket, a HAMIS pedig, hogy alatta marad.

17. Táblázat: A H4 hipotézishez kapcsolódó arányok összehasonlítása a különböző küszöbértékekkel

		t-próba szignifikancia szintje		
		1%	5%	10%
		modellek aránya		
		99,89%	99,09%	98,41%
Elfogadási alsó küszöb az arányokra	90%	IGAZ	IGAZ	IGAZ
	95%	IGAZ	IGAZ	IGAZ
	99%	IGAZ	IGAZ	HAMIS
	100%	HAMIS	HAMIS	HAMIS

Forrás: saját szerkesztés

Az összes vizsgált szignifikancia szint mellett a modelleknek legalább 95%-a magyar és angol nyelven is hasonló pontosságot ért el. A 99%-os alsó küszöböt egyedül a 10%-os szignifikancia szintnél kapott arány nem lépte túl, a 100%-ot pedig egyik sem érte el. Ezekben az esetekben a *16. táblázat* alapján meg kell vizsgálni, hogy valamelyik nyelv preferált-e. A 10%-os szignifikancia szint esetén a 99% és 100%-os küszöb mellett a H4 hipotézist nem vetem el, mert a magyar nem volt az összes esetben pontosabb[121]. A 100%-os küszöbérték esetén 1%-os és 5%-os szignifikancia szintnél kizárólag a magyar nyelvű korpuszon tanult modellek voltak pontosabbak, így ezekben az esetekben a H4 hipotézist el kell vetni. Ezt foglalja össze a *18. táblázat*, az IGAZ érték esetén fogadjuk el a hipotézist.

18. Táblázat: A H4 hipotézis elfogadása különböző feltételek mellett

		t-próba szignifikancia szintje		
		1%	5%	10%
		modellek aránya		
		99,89%	99,09%	98,41%
Elfogadási alsó küszöb az arányokra	90%	IGAZ	IGAZ	IGAZ
	95%	IGAZ	IGAZ	IGAZ
	99%	IGAZ	IGAZ	IGAZ
	100%	HAMIS	HAMIS	IGAZ

Forrás: saját szerkesztés

A H4 hipotézist az összes szignifikancia szinten el kell fogadni, ha az eltérések aránya nem éri el a jelentős mennyiségnek ítélt 1–10 százalékot. Tehát a pontosság szem-

121 Mivel hatszor annyi pontosabb modell tartozik a magyarhoz, mint az angolhoz, szubjektív megítélésem szerint a hipotézis elvetésével sem lenne indokolatlan ebben az esetben.

pontjából mindegy, hogy angol vagy magyar nyelvű korpuszt használunk az árfolyamok osztályozásához. Ha minden eltérést jelentősnek ítélünk, abban az esetben a H4 hipotézist el kell vetnünk 1%-os és 5%-os szignifikancia szintek mellett, mert a magyar nyelvű korpusz segítségével jobb eredmények érhetők el bizonyos C-gamma kombinációjú modelleknél. Ez az eredmény hasonlít Groth (2014) eredményére, aki azt állapította meg, hogy a sajtóközlemények német – tehát anyanyelvi – változatai alapján pontosabb modell készíthető, mint az angol változat alapján. Említésre méltó, hogy ezek a kombinációk, ahogy a *29. ábrán* is látszik, elsősorban a *18. ábrán* megfigyelhető átlós alakzatok kontúrjától jobbra-felfelé találhatók. Úgy tűnik tehát, hogy a magyar korpusz esetén az SVM kevésbé érzékeny a megfelelő C-gamma kombináció megválasztására. A H4-es hipotézis tesztelése során kapott eredményeim miatt használtam a H2 és H3 hipotézisekhez kapcsolódó vizsgálatoknál is a magyar nyelvet. Az eredményeket úgy interpretálhatjuk, hogy kevésbé szigorú értelemben mindegy, hogy milyen nyelvű korpuszt használunk a kísérlethez, mert hasonló eredményt kapunk; szigorúan véve viszont a magyart célszerűbb választani, mert néhány esetben nagyobb is lehet a pontosság.

4.5. Érzékenységvizsgálat az SVM paramétereire

A *3.4.1.1 alfejezetben* bemutattam, hogyan változik az rbf SVM-osztályozó szeparációjának jellege a lineáris és legközelebbiszomszéd-osztályozó között, ha a C és gamma paramétereket változtatjuk. A *13. ábra* kapcsán pedig leírtam, hogy a hőtérképes vizualizáción milyen alakzat várható lineáris, nemlineáris és nem szeparálható esetben. A *18. ábrán* látható hőtérképeimen látható alakzatok alapján arra lehet következtetni, hogy a probléma nemlineárisan szeparálható, és angol nyelv esetén inkább jobban érvényesül ez a jelleg, a magyarnál – valószínűleg a magasabb dimenziószám miatt – inkább a lineáris jelleg felé tapasztalható eltolódás. Ez alapján indokoltnak érzem, hogy a szakirodalomban általában használt lineáris kernel helyett az rbf kernelt használtam. A *18. ábrán* viszont az is látható, hogy az osztályozó nem mindig őrzi meg a maximumhoz közeli pontosságát, ha a paraméterkombinációkban egységnyi változás történik, ezért szükségesnek gondolom, hogy a tőzsdei hírbányászat irodalmát az SVM paramétereinek érzékenységvizsgálatával gazdagítsam.

H5: Az eredmények robusztusak az alkalmazott SVM osztályozási módszer beállításaira nézve.

A hipotézist elvetem, ha azon kvázioptimális modellek aránya, amelyek paramétereinek legfeljebb egységnyi megváltozásával egy másik kvázioptimális modellhez juthatunk, nem éri el a 90, 95, 99, 100%-ot. Első lépésben kétmintás t-próbával tesztelem mindegyik modellre, hogy a vele azonos nyelvű és reprezentációjú modellek között megfigyelt legnagyobb pontossággal megegyezik-e az övé – azaz, hogy kvázioptimális-e[122]. Második lépésben minden kvázioptimális modell esetén megvizsgálom, hogy a paraméterszomszédjai között van-e kvázioptimális – az ilyen tulajdonságú kvázioptimális modellekre azt mondom, hogy robusztusak. A harmadik lépésben kiszámolom a robusztus és a kvázioptimális modellek arányát. A negyedik lépésben összehasonlítom a négy, egyre szigorúbb küszöbértékkel az arányt, és ha túllépi, elfogadom a H5 hipotézist.

Először teszteljük a modelleket, hogy ugyanolyan pontosak-e, mint legnagyobb pontosságú. A null- és alternatív hipotézisek a következők minden $m_i \in M$ modell esetén:

$$H_{0,i}: \mu_i = \mu_{i^{max}} \tag{60}$$

$$H_{1,i}: \mu_i \neq \mu_{i^{max}} \tag{61}$$

Ahol:

$I_{(l,r)}$ az $l \in L$ nyelvű, $r \in R$ reprezentációjú modellek indexeinek halmaza, ahol $|I_{(l,r)}| = 441$ minden l és r esetén, és a modelleket nyelvenként és reprezentációnként eltérően számozzuk, azaz $I_{(l,r)} \cap I_{(l',r')} = \emptyset$ minden $\neg(l = l' \wedge r = r')$ esetén. Ekkor:

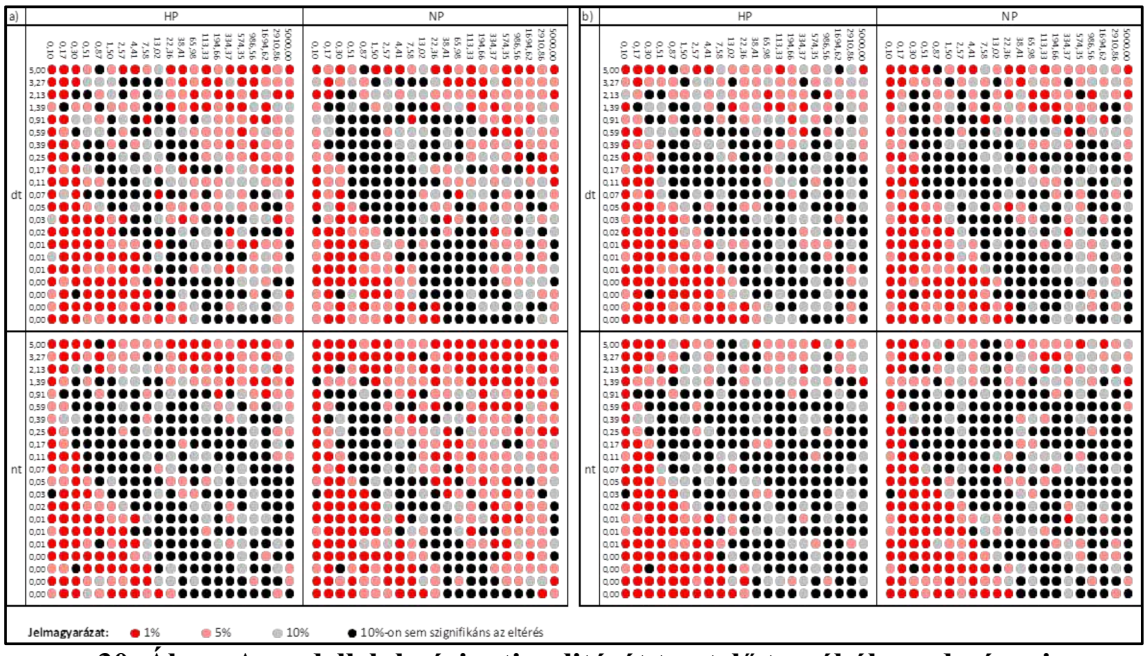
$I = \bigcup_{(l,r) \in (L \times R)} I_{(l,r)}$ az összes modell indexének halmaza, és

$|I| = \sum_{(l,r) \in (L \times R)} |I_{(l,r)}| = 3528$

$\mu_{i^{max}} = \max_j \mu_j$, ahol $(i \in I_{(l,r)}) \wedge (j \in I_{(l,r)})$

A tesztek eredményét láthatjuk a *30. ábrán*. 1% szignifikancia szinten 2686[123] modellre mondhatjuk azt, hogy nem különbözik a maximális pontosságú modellétől a pontossága[124], 5%-on 1887[125], 10%-on 1374[126] modellnek.

122 Ebbe beleértjük a maximális pontosságú modellt is.
123 3528-ra vetítve ez 76,13%.
124 Ezek a nem piros pontoknak felelnek meg.
125 Nem piros vagy rózsaszín. 53,49%.
126 Fekete pontok. 38,95%.



30. Ábra: A modellek kvázioptimalitását tesztelő t-próbák eredményei

Forrás: saját szerkesztés

A második lépés, hogy megszámoljuk, hogy ezek közül hány modellnek van kvázi-optimális paraméterszomszédja, a harmadik lépés, hogy kiszámítjuk ezek arányát. Ez a szám 1%-on 2657, amely a 2686 kvázioptimális modellnek a 98,92%-a, 5%-on 1857 – 1887-nek a 98,41%-a –, 10%-on 1346 – 1374-nek a 97,96%-a. A negyedik lépésben felírjuk a H5 hipotézis elfogadását és elutasítását összesítő *19. táblázatot*.

19. Táblázat: A H5 hipotézis elfogadása különböző feltételek mellett

		t-próba szignifikancia szintje		
		1%	**5%**	**10%**
		modellek aránya		
		98,92%	**98,41%**	**97,96%**
Elfogadási alsó küszöb az arányokra	**90%**	IGAZ	IGAZ	IGAZ
	95%	IGAZ	IGAZ	IGAZ
	99%	HAMIS	HAMIS	HAMIS
	100%	HAMIS	HAMIS	HAMIS

Forrás: saját szerkesztés

A H5 hipotézist az összes szignifikancia szinten el kell fogadni, ha az alsó küszöbértéket 90 vagy 95%-ban határozzuk meg, és el kell vetni, ha a küszöb 99% vagy 100%. Az alkalmazott SVM-osztályozó pontossága tehát viszonylag robusztus a módszer beállításaira nézve, azonban léteznek olyan paraméterkombinációk, amelyek elszigeteltek a

többi hasonló teljesítményt adótól. Emiatt az üzleti alkalmazás fázisában nem biztos, hogy csak a legpontosabb előrejelzést adó modell alkalmazására van szükség, hanem több különböző paraméterkombinációval rendelkező modell eredményeit kell összevetni. E kombinációkat célszerű a *30. ábrán* is mutatkozó, a fekete színű pontok által kirajzolt átlós területről választani.

4.6. Érzékenységvizsgálat a szövegreprezentációra

A sajtóközlemények nyelvezete eléggé eltérhet a hétköznapi nyelvétől, sok szakkifejezést használt, és vannak benne több dokumentumban megismétlődő, sablonszerű szövegek is. Az információ reprezentációja hatással lehet a modell a teljesítményére, ahogy Schumakernél (2009) láttuk. Így érdemes megvizsgálni, hogy a mintámon a normál szózsákmodellhez képest lehet-e javulást elérni, ha az eleve fölöslegesnek gondolt információt kitöröljük a reprezentációból, illetve ha a többszavas szakkifejezéseket egy egységként kezeljük, és nem az őket alkotó szavakként.

H6: Az eredmények robusztusak a szöveges inputok körének megválasztására.

A hipotézist elvetem, ha bármely két azonos nyelvű szövegreprezentáció között különbséget tapasztalok. Akkor tekintek két reprezentációt különbözőnek, ha az azonos paraméterű modelljeik pontosságát kétmintás t-próbával összehasonlítva a szignifikánsan nem különbözők aránya nem éri el a 90, 95, 99, illetve 100%-ot. Ezen kívül még az szükséges, hogy az eltérések szisztematikusan az egyik reprezentáció javára jelenjenek meg. A kutatási hipotézis elfogadásához egyre szigorúbb feltételeket jelent a négy küszöbszint.

Első lépésben t-próbával megállapítom a szignifikáns különbségeket minden modellpár esetén. Második lépésben minden szignifikancia szintre, minden reprezentációpár esetén megszámolom, hogy hány pontosabb modell tartozott az egyik, illetve a másik reprezentációhoz. Harmadik lépésben kiszámolom azonos pontosságú modellek arányát minden szignifikancia szintre és reprezentációpárra. Negyedik lépésben ezeket összevetem a küszöbértékekkel, és a küszöbérték alatti eseteknél megvizsgálom, hogy van-e preferencia a reprezentációk között, majd elfogadom vagy elvetem a H6 hipotézist.

Az első lépésben a null- és alternatív hipotézisek tehát a következők minden $k \in K$ modellpár esetén:

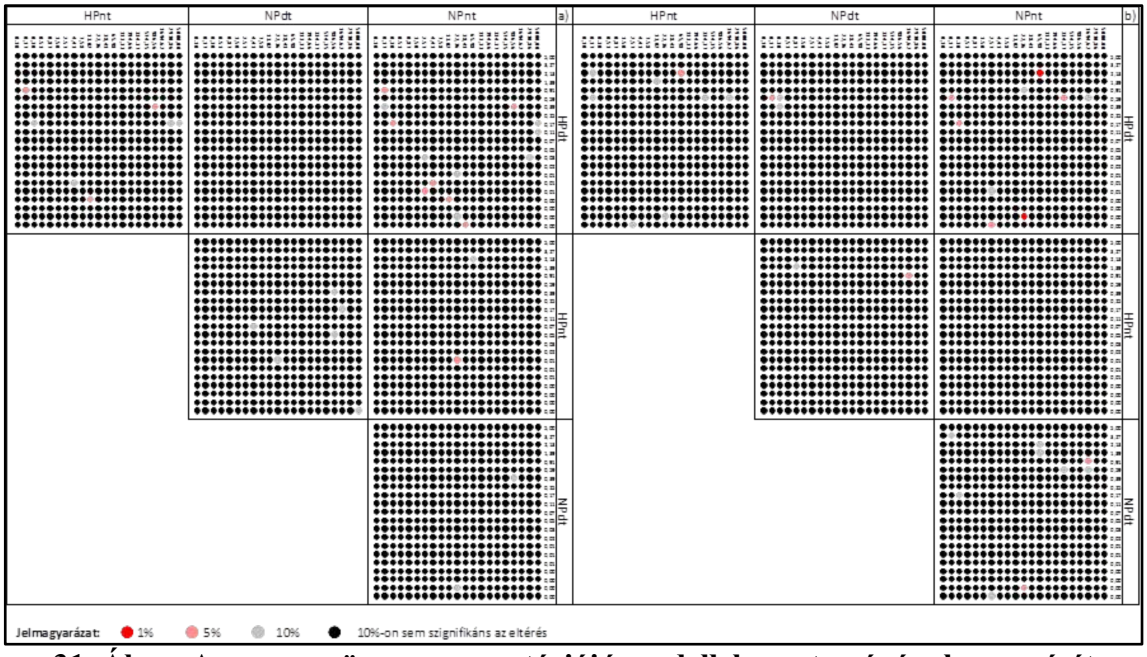
$$H_{0,k}: \mu_i = \mu_j \tag{62}$$

$$H_{1,k}: \mu_i \neq \mu_j \tag{63}$$

Ahol:

$K=\{k\in(M\times M):k=(m_i,m_j),l_i=l_j,c_i=c_j,g_i=g_j,r_i\neq r_j\}$ az azonos paraméterű, azonos nyelvű, de eltérő reprezentációjú modellpárok halmaza, melynek számossága:

$$|K|=|C|\cdot|G|\cdot\frac{|R|\cdot(|R|-1)}{2}\cdot|L|=21\cdot21\cdot\frac{4\cdot(4-1)}{2}\cdot2=5292$$



31. Ábra: Az egyes szövegreprezentációjú modellek pontosságának egyezését tesztelő t-próbák eredményei

Forrás: saját szerkesztés

A *31. ábrán* látható, hogy kizárólag a reprezentáció megváltozása a legtöbb esetben nem változtatja meg szignifikánsan a modellek pontosságát. Az ábra a) részén látható az angol nyelv esetén végzett összehasonlítások eredménye, a b) részén a magyar nyelvé.

20. Táblázat: A szövegreprezentációk közötti eltérések száma

	r_i	(HP; dt)						(HP; nt)				(NP; dt)	
	r_j	(HP; nt)		(NP; dt)		(NP; nt)		(NP; dt)		(NP; nt)		(NP; nt)	
Pontosabb:		**j**	**i**	**j**	**i**	**j**	**i**	**j**	**i**	**j**	**i**	**j**	**i**
t-próba szignifikancia szintje:	**10%**	14	1	3	0	23	0	2	6	2	0	11	0
	5%	4	0	1	0	13	0	0	1	1	0	2	0
	1%	0	0	0	0	2	0	0	0	0	0	0	0

Forrás: saját szerkesztés

A második lépésben elkészítjük a *20. táblázatot*, amelyben a háromdimenziós oszlopfejléc első sora az *i* reprezentáció, második sora a *j* reprezentáció jelét mutatja. Az

egyes *i-j* párok esetén tapasztalható különbségek iránya alapján csoportosított tesztek számát látjuk a cellákban – például 10%-os szignifikancia szint mellett a (HP; nt) reprezentáció 14 esetben volt pontosabb, mint a (HP; dt).

A *20. táblázatból* kiolvasható például, hogy az esetek többségében, ha eltérések vannak, akkor az egyik reprezentáció preferált a másikhoz képest. Ezen kívül 5%-os szignifikancia szint mellett a sablonszöveget is tartalmazó változatok – nt – pontosabbak voltak a sablontól megtisztítotthoz képest.

A harmadik lépés, hogy megszámláljuk a különböző szignifikancia szintek mellett minden reprezentáció-párosításra, az eltérést nem mutató összehasonlításokat, majd kiszámítjuk ezek arányát az összes olyan összehasonlítás számához képest, amelyek ugyanarra a reprezentáció-párra vonatkoznak. A viszonyítási alap tehát $2 \cdot 21 \cdot 21 = 882$ -vel egyenlő. Ezeket a számokat és arányokat láthatjuk a *21. táblázatban*.

21. Táblázat: Az eltérést nem mutató összehasonlítások száma és aránya a lehetséges reprezentáció-párosítások esetén

	r_i	(HP; dt)						(HP; nt)				(NP; dt)	
	r_j	(HP; nt)		(NP; dt)		(NP; nt)		(NP; dt)		(NP; nt)		(NP; nt)	
t-próba szignifikancia szintje	**10%**	867	98,30%	879	99,66%	859	97,39%	874	99,09%	880	99,77%	871	98,75%
	5%	878	99,55%	881	99,89%	869	98,53%	881	99,89%	881	99,89%	880	99,77%
	1%	882	100,00%	882	100,00%	880	99,77%	882	100,00%	882	100,00%	882	100,00%

Forrás: saját szerkesztés

A negyedik lépésben a kapott arányokat össze kell hasonlítani a különböző küszöbértékekkel, és meg kell vizsgálni, hogy ha nem éri el a küszöbértéket, akkor szisztematikus-e az eltérés az egyik reprezentációra. A *21. táblázat* arányainak a küszöbértékekkel való összehasonlításokat foglalja össze a *22. táblázat.* A táblázaton jól látható, hogy a 90 és 95%-os küszöb esetén minden szignifikancia szinten azonos teljesítményűnek kell tekintenünk a különböző reprezentációjú modelleket. A táblázat küszöb alatti – HAMIS – értékeit a *20. táblázat* alapján felül kell bírálni, majd összeállítani a *23. táblázatot*, amely a kutatási hipotézis elfogadásának és elutasításának eseteit tartalmazza.

22. Táblázat: A szignifikánsan nem különböző modellek arányának a küszöbértékekkel való összehasonlítása

	r_i	(HP; dt)								
	r_j	(HP; nt)			(NP; dt)			(NP; nt)		
t-próba szignifikancia szintje:		**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
modellek aránya:		**100,00%**	**99,55%**	**98,30%**	**100,00%**	**99,89%**	**99,66%**	**99,77%**	**98,53%**	**97,39%**
Elfogadási alsó küszöb az arányokra	**90%**	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	95%	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	99%	IGAZ	IGAZ	HAMIS	IGAZ	IGAZ	IGAZ	IGAZ	HAMIS	HAMIS
	100%	IGAZ	HAMIS	HAMIS	IGAZ	HAMIS	HAMIS	HAMIS	HAMIS	HAMIS

	r_i	(HP; nt)						(NP; dt)		
	r_j	(NP; dt)			(NP; nt)			(NP; nt)		
t-próba szignifikancia szintje:		**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
modellek aránya:		**100,00%**	**99,89%**	**99,09%**	**100,00%**	**99,89%**	**99,77%**	**100,00%**	**99,77%**	**98,75%**
Elfogadási alsó küszöb az arányokra	**90%**	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	95%	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	99%	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	HAMIS
	100%	IGAZ	HAMIS	HAMIS	IGAZ	HAMIS	HAMIS	IGAZ	HAMIS	HAMIS

Forrás: saját szerkesztés

23. Táblázat: A H6 hipotézis elfogadása különböző feltételek mellett

	r_i	(HP; dt)								
	r_j	(HP; nt)			(NP; dt)			(NP; nt)		
t-próba szignifikancia szintje:		**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
modellek aránya:		**100,00%**	**99,55%**	**98,30%**	**100,00%**	**99,89%**	**99,66%**	**99,77%**	**98,53%**	**97,39%**
Elfogadási alsó küszöb az arányokra	**90%**	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	95%	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	99%	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	HAMIS	HAMIS
	100%	IGAZ	HAMIS	IGAZ	IGAZ	HAMIS	HAMIS	HAMIS	HAMIS	HAMIS

	r_i	(HP; nt)						(NP; dt)		
	r_j	(NP; dt)			(NP; nt)			(NP; nt)		
t-próba szignifikancia szintje:		**1%**	**5%**	**10%**	**1%**	**5%**	**10%**	**1%**	**5%**	**10%**
modellek aránya:		**100,00%**	**99,89%**	**99,09%**	**100,00%**	**99,89%**	**99,77%**	**100,00%**	**99,77%**	**98,75%**
Elfogadási alsó küszöb az arányokra	**90%**	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	95%	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ
	99%	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	IGAZ	HAMIS
	100%	IGAZ	HAMIS	IGAZ	IGAZ	HAMIS	HAMIS	IGAZ	HAMIS	HAMIS

Forrás: saját szerkesztés

A H6 hipotézist az összes szignifikancia szinten el kell fogadni, ha az alsó küszöbértéket 90 vagy 95%-ban határozzuk meg. 99%-os küszöbnél is többnyire elfogadjuk a hipotézist, azonban három kivétel tapasztalható. Ezek szerint a legkevesebb feldolgozást igénylő (NP; nt) reprezentáció 5%-os és 10%-os szignifikancia szinten nem volt azonos, és jobbnak mutatkozott, mint a legtöbb feldolgozást igénylő (HP; dt) reprezentáció. Ugyancsak az (NP; nt) reprezentáció bizonyult pontosabbnak az (NP; dt) reprezentációnál, bár csak 10%-os szignifikancia szint mellett lehetett kimutatni, hogy különböznek. 100%-os küszöb esetén 1%-os szignifikancia szint mellett öt esetben elfogadjuk a H6 hipotézist, egy esetben elvetjük, ugyanis az (NP; nt) reprezentáció ebben az esetben is jobbnak mutatkozott a (HP; dt)-nél. E két reprezentáció között egyébként 5 és 10%-on is ugyanez a viszony áll fenn. Az (NP; nt) reprezentáció ezen kívül még 5 és 10%-on nem azonos, és pontosabb az (NP; dt) és a (HP; nt) reprezentáció pontosságával. Ezen kívül még különbségek tapasztalhatók 5 és 10%-on az (NP; dt) javára a (HP; dt)-hez képest. Végül 5%-on a (HP; nt) jobb a (HP; dt)-nél és az (NP; dt)-nél is.

Összefoglalva tehát ez azt jelenti, hogy ha 5 vagy 10%-nyi arányt sem tartunk jelentősnek, ami a szignifikánsan különböző pontosságú modellek aránya két reprezentáció között, akkor azt mondhatjuk, hogy nincs lényeges különbség az egyes szövegreprezentációk tekintetében. Ha már 5% vagy 1% alatti arányt is jelentősnek ítélünk, akkor azonban arra a meglepő eredményre jutunk, miszerint a legprimitívebb szövegreprezentációval legalább olyan pontos eredményt lehet elérni, mint a többivel, sőt a legkifinomultabb reprezentációhoz képest ez a viszony minden vizsgált szignifikancia szint mellett fennáll. Elmondható továbbá, hogy az egyéb páros összehasonlítások eredményei is arra utalnak, hogy ha kevésbé kifinomult reprezentációs módszert választunk, az eredmények nem romlanak szignifikánsan, és néhány esetben még javulhatnak is. Ezen eredmény az oka annak, hogy a H2 és H3 hipotéziseim tesztelésekor csak egyetlen szövegreprezentációt alkalmaztam, az (NP; nt) jelűt. Ezek szerint ugyanis legalább olyan jó, mint a többi, de kevesebb adat-előkészítést igényel.

5. Összegzés

A tőzsdei hírbányászat egy szövegbányászatra épülő módszertan, amely a gazdasági hírek szövegéből következtet egy részvény vagy más instrumentum árfolyamára vagy más kereskedési adatára. Gyökerei az 1990-es évek végére nyúlnak vissza, amikor Hong Kongban Wüthrich és szerzőtársai először készítettek napi előrejelzést a HSI index értékére. Az ezredfordulón Lavrenko fejlesztette tovább a koncepciót, és napon belüli előrejelzést készített egyes részvények árfolyamának trendjére vonatkozóan. Eközben Thomas tőzsdei fórumok hozzászólásait elemezve készített a következő napra előrejelzést részvények árfolyamára. Gidófalvi pedig úgy találta, hogy a hírek árfolyamra gyakorolt hatása a publikálás körüli ± 20 perces időablakban kimutatható. Koppel a 2000-es évek közepétől vizsgálta a hírek hangulatát és az ezt meghatározó kifejezéseket. Ugyanekkor Mittermayer sajtóközleményekből kinyert információk alapján már több mint 80%-os pontosságot elért. Az e-Markets Group nevű kutatócsoport közben devizaárfolyamok előrejelzésére szolgáló rendszert fejlesztett. 2006-tól publikált a témában Schumaker és Chen, akik több szövegreprezentációt és kereskedési stratégiát beépítettek a rendszerbe. 2008-tól elemezte Groth a német sajtóközleményeket, kezdetben az árfolyamra gyakorolt hatásukat, majd a volatilitásra és a tranzakciós költségekre. Ő végzett először kétnyelvű – angol–német – összehasonlításokat is 2014-ben.

Miután a 2. fejezetben áttekintettem a fenti szerzők munkáit, a módszertant a 3. fejezetben foglaltam egységes keretbe a CRISP-DM sztenderd alapján. 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 olyan hosszú vektorral helyettesítünk, amennyi szövegjellemzőnk van. A vektor 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.

A saját tőzsdei hírbányászati modellem hozamosztályozó típusú, 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 $-120 \le t \le 120$ perccel eltolt időpont közötti hozamkategória – negatív, semleges, pozitív. A hírek szövegének numerikus reprezentációja alapján SVM-osztályozót tanítottam rbf kernellel, melynek pontosságát 10-szeres keresztvalidációval ellenőriztem, a modell hatékonyságát az átlagos pontossággal mértem, amelyeket a saját d és q mutatókkal értékeltem. A modellt RapidMinerrel készítettem, a statisztikai teszteket Microsoft Excel 2010-ben, az ábrákat az előbbieken kívül még LibreOffice Calc-kal is. A sajtóközleményeket HTTrack-kel töltöttem le a BÉT honlapjáról. Az egyperces részvényárfolyamokat a Thomson Reuters Eikon platform segítségével szereztem be.

A H1-es hipotézisem vizsgálatakor azt vizsgáltam, hogy a szöveges előrejelzés pontosabb-e a defaultnál. 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. 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.

A H2-es hipotézisem az volt, hogy nem lehet jó minőségű, illetve robusztus modellt készíteni olyan időablakra, amely korábbra nyúlik vissza, mintsem a sajtóközlemény átmenne a közzétételi folyamaton. A hipotézis elfogadásának feltétele az volt, hogy a -60 -as időablakot megelőző időtávra nehéz előrejelezni, illetve a modellek minősége is gyenge. Eredményeim szerint a H2 hipotézist elfogadhatjuk, hiszen a -60 -as időeltoláson túl egyre nehezebb az előrejelzés, és a minőség mediánja is tovább romlik.

A H3-as hipotézis során a legrobusztusabb és legjobb minőségű modellekre vezető időablak létét feltételeztem, és ennek meghatározását többcélú optimalizálási feladatra vezettem. A H3-as 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.

A H4-es hipotézis szerint nem lényeges milyen nyelvű dokumentumokkal dolgozunk, az azonos sajtóközlemények angol és magyar nyelven közzétett változataival készített modellek pontossága között nincs szignifikáns különbség. 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. A H4 hipotézist az összes szignifikancia szinten – 1%, 5% és 10% – elfogadtam, amennyiben a különbséget nem mutató modellpárok arányának megköveteltem legalább 99%-ot. 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. Ezek alapján megállapítható, hogy az esetek jelentős többségében nem befolyásolja jelentősen a vizsgált modellek pontosságát az, hogy angol vagy magyar nyelvű korpuszt használunk az árfolyamok osztályozásához, azonban a maradék esetekben a magyar nyelvű közlemények alapján pontosabb becsléshez jutunk.

A H5 hipotézis szerint az eredmények robusztusak az alkalmazott SVM osztályozási módszer beállításaira nézve. 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 90% és 100% közötti küszöbértékeket. Eredményeim szerint kevésbé szigorúan véve[127] a kvázioptimális modellek robusztusak minden vizsgált szignifikancia szinten, de szigorúan véve[128] el kell vetnünk a hipotézist, mert kb. 1%-nyi eséllyel visszaeshet a default szintre 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.

A H6-os hipotézisem szerint az eredmények robusztusak a szöveges inputok körének megválasztására. 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. A H6 hipotézist az összes szignifikancia szinten – 1, 5 és 10% – el kell fogadni, ha a *majdnem minden* küszöbértéket 95%-ban határozzuk meg. Magasabb küszöbértékek esetén a hipotézis több párosítás esetén elvetendő, és megállapítható, hogy az egyszerűbb reprezentációt alkalmazó modellek bizonyulnak pontosabbnak. 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.

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.

127 90%-os és 95%-os küszöbértéknél.
128 99% és 100%-os küszöbnél.

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. Hasonlóan 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.

Modellem jelenlegi formájában is elég meggyőző eredményeket produkált, és úgy gondolom, hogy pontosabb előrejelzés is készíthető valamely továbbfejlesztés révén, de a teljes előrejelzést nem remélhetjük tőle. Az árfolyamot nem csak a sajtóközlemények mozgatják, hanem soha ki nem derülő, gyakorlatilag véletlenszerű egyéni motivációk, nagyszabású üzleti érdekek és persze a gazdaságban máshol történő események is. Ezért tisztában kell lenni a korlátokkal, és arra kell összpontosítani a továbbfejlesztés során, amire a modell képes, nem pedig arra, amire nem.

Felhasznált irodalom

Abonyi, János, (2006), *Adatbányászat a hatékonyság eszköze : Gyakorlati útmutató kezdőknek és haladóknak*, Budapest: ComputerBooks. Elérhető: https://www.researchgate.net/publication/264441999_Adatbanyaszat_a_hatekonysag_eszkoze.

Achelis, Steven B., (2001), *Technical Analysis from A to Z*, New York: McGraw-Hill. Elérhető: http://books.google.com/books?hl=en&lr=&id=ZproZYDvKqsC&oi=fnd&pg=PR13&dq=Technical+Analysis+from+A+to+Z&ots=AII6VIRoS2&sig=cujwHZyaCu3dAI-urVEj9ykUEm8.

ANY PLC, (2014), Number of voting rights, share capital at ANY Security Printing Company PLC. Elérhető: https://client.bse.hu/topmenu/issuers/issuersnews/117186555.html [Elérés 2015.07.25.].

Barber, Brad M. & Lyon, John D. (1997): Detecting long-run abnormal stock returns: The empirical power and speci¯cation of test statistics, Journal of Financial Economics 43, p. 341-372.

Barberis, Nicholas & Thaler, Richard (2003), A survey of behavioral finance, In *Handbook of the Economics of Finance*, Volume 1, Part B, Pages 1053-1128, ISSN 1574-0102, ISBN 9780444513632, http://dx.doi.org/10.1016/S1574-0102(03)01027-6.

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.

Ben-Hur, Asa & Weston, Jason, (2010), A User's Guide to Support Vector Machines. In O. Carugo & F. Eisenhaber (szerk.) *Data Mining Techniques for the Life Sciences*. Methods in Molecular Biology. Humana Press, o. 223–239. Elérhető: http://link.springer.com/10.1007/978-1-60327-241-4%5Cnhttp://link.springer.com/10.1007/978-1-60327-241-4₁3.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care*, 9(1), 112–118. http://doi.org/10.1186/cc3045

Black, Fischer & Scholes, Myron (1973): The Pricing of Options and Corporate Liabilities, Journal of Political Economy. 81 (3): 637–654. doi:10.1086/260062

Brown, Stephen J. & Warner, Jerold B. (1980): Measuring security price performance, Journal of Financial Economics 8, p. 205-258.

Brown, Stephen J. & Warner, Jerold B. (1985): Using daily stock returns: The case of event studies, Journal of Financial Economics 14, p. 3-32.

Budapesti Értéktőzsde Zrt., (é. n. a), Egyedi termékekre vonatkozó historikus adatok letöltése. Elérhető: http://bet.hu/magyar_egyeb/dinportl/instrdatadownload [Elérés 2016.08.08.].

Budapesti Értéktőzsde Zrt., (é. n. b), Historikus adatok díjai. Elérhető: http://bet.hu/data/cms142234/Dijtablazat_2010januar_v3.xls [Elérés 2016.08.08.].

Budapesti Értéktőzsde Zrt., (é. n. c), Vendorok listája. Elérhető: http://bet.hu/topmenu/adatszolg/vendorlista [Elérés 2016.08.08.].

Budapesti Értéktőzsde Zrt., (é. n. d), Bevezetési és Forgalomban Tartási Szabályok. Elérhető: https://bet.hu/data/cms58438/2__Konyv___Bevezetesi_es_Forgalomban-tartasi_Szabalyok.pdf [Elérés 2016.08.08.].

Budapesti Értéktőzsde Zrt., (é. n. e), Adatszolgáltatási útmutató. Elérhető: http://bet.hu/topmenu/adatszolg/adatszolg_utmutato [Elérés 2016.08.09.].

Budapesti Értéktőzsde Zrt., (é. n. f), Információs csomagok. Elérhető: http://bet.hu/topmenu/adatszolg/adatszolg_iranyelvek/informacios_csomagok/info_csomagok.html [Elérés 2016.08.09.].

Budapesti Értéktőzsde Zrt., (é. n. g), Vendor irányelvek. Elérhető: http://bet.hu/topmenu/adatszolg/adatszolg_iranyelvek/vendor_iranyelvek.html [Elérés 2016.08.09.].

Budapesti Értéktőzsde Zrt., (é. n. h), Terméklista letöltése. Elérhető: http://bet.hu/topmenu/befektetok/termekcsoportok/termeklista [Elérés 2015.07.25.].

Budapesti Értéktőzsde Zrt., (é. n. i), Kibocsátói hírek. Elérhető: http://bet.hu/topmenu/kibocsatok/kibocsatoi_hirek [Elérés 2016.08.10.].

Budapesti Értéktőzsde Zrt., (2013), Műszaki Specifikáció: BÉT Információkhoz való hozzáférés a WBAG adatszolgáltató rendszerén keresztül. Elérhető: https://bet.hu/data/cms48343/BET_Muszaki_Specifikacio.pdf [Elérés 2016.08.09.].

Burges, C.J. (1998), A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery* 2(2), pp 121–167 doi:10.1023/A:1009715923555

Campbell, C. J., & Wasley, Charles E. (1993): Measuring security price performance using daily NASDAQ returns, Journal of Financial Economics 33, p. 73-92.

Chapman, Pete, Clinton, Julian, Kerber, Randy, Khabaza, Thomas, Reinartz, Thomas, Shearer, Colin & Wirth, Rudiger, (2000), *CRISP-DM 1.0*, Elérhető: http://www.crisp-dm.org/CRISPWP-0800.pdf [Elérés 2016.08.02.].

Cho, V., Wüthrich, B. & Zhang, J., (1999), Text Processing for Classification. *Journal of Computational Intelligence in Finance*, 7(2), o.6–22. Elérhető: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.651 [Elérés 2015.01.01.].

Cho, Vincent Wing-Sing, (1999), *Knowledge discovery from distributed and textual data*. THES. Hong Kong: The Hong Kong University of Science and Technology. Elérhető: http://hdl.handle.net/1783.1/1520 [Elérés 2015.01.01.].

Cho, Vincent & Wüthrich, Beat, (1999), Combining Forecasts from Multiple Textual Data Sources. In N. Zhong & L. Zhou (szerk.) *Methodologies for Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science. Beijing, China: Springer Berlin Heidelberg, o. 174–179. Elérhető: http://dx.doi.org/10.1007/3-540-48912-6_24.

Dyckman, T., Philbrick, D., Stephans, J. & Ricks, W. E. (1984): A comparison of event study methodologies using daily stock returns: A simulation approach, Journal of Accounting Research 22, p. 1-33.

Elizondo, David A., Birkenhead, Ralph, Gamez, Matias, Garcia, Noelia, Alfaro, Esteban (2012), Linear separability and classification complexity, *Expert Systems with Applications*, 39(9), 7796-7807

Erste Befektetési Zrt., (2016), 2016. április 29-től hatályos Portfolio Online Tőzsde Befektetési Szolgáltatások üzleti DÍJJEGYZÉK. Elérhető: http://www.portfolio.hu/download/files/aaaportfolio_dijjegyzek_160428_cl.pdf [Elérés 2016.08.26.].

Esmaeili, Mobin, Sedighizadeh, Mostafa & Esmaili, Masoud, (2016), Multi-objective optimal reconfiguration and DG (Distributed Generation) power allocation in distribution networks using Big Bang-Big Crunch algorithm considering load uncertainty. *Energy*, 103, o.86–99. Elérhető: http://dx.doi.org/10.1016/j.energy.2016.02.152.

Fama, Eugene F. (1970), Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance*, 25(2), 5, 383–417. o.

Fama, Eugene, Fisher, Lawrence, Jensen, Michael & Roll, Richard, (1969), The Adjustment Of Stock Prices To New Information. *International Economic Review*, 10(1), o.1–21.

Ferrara, Emilio, De Meo, Pasquale, Fiumara, Giacomo & Baumgartner, Robert, (2014), Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, 70, o.301–323.

Fung, Gabriel Pui Cheong, Yu, J.X. & Lam, Wai, (2003), Stock prediction: Integrating text mining approach using real-time news. In *2003 International Conference on Computational Intelligence for Financial Engineering*. Hong Kong, China, o. 395–402.

Fung, Gabriel Pui Cheong, Yu, Jeffrey Xu & Lam, Wai, (2002a), Automatic stock trend prediction by real time news. In W.-K. Ching & M. K.-P. Ng (szerk.) *Advances in Data Mining and Modeling*. Hong Kong: World Scientific Publishing, o. 48–59. Elérhető: http://www.itebookshare.com/?p=242892#more-242892.

Fung, Gabriel Pui Cheong, Yu, Jeffrey Xu & Lam, Wai, (2002b), News Sensitive Stock Trend Prediction. In M.-S. Chen, P. S. Yu, & B. Liu (szerk.) *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science. Taipei, Taiwan: Springer-Verlag Berlin Heidelberg, o. 481–493. Elérhető: http://dx.doi.org/10.1007/3-540-47887-6_48.

Fung, Gabriel Pui Cheong, Yu, Jeffrey Xu & Lu, Hongjun, (2005), The Predicting Power of Textual Information on Financial Markets. *IEEE Intelligent Informatics Bulletin*, 5(1), o.1–10. Elérhető: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.3446.

Fung, Pui Cheong Gabriel, (2003), *Stock market forecasting by integrating time-series and textual information*. THES. Hong Kong: The Chinese University of Hong Kong. Elérhető: http://library.cuhk.edu.hk/record=b2418475~S15 [Elérés 2015.01.01.].

Généreux, Michel, Poibeau, Thierry & Koppel, Moshe, (2008), Sentiment analysis using automatically labelled financial news. In Khurshid Ahmad (szerk.) *LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology*. Marrakesh, Morocco, o. 38–43. Elérhető: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W6_Proceedings.pdf.

Généreux, Michel, Poibeau, Thierry & Koppel, Moshe, (2011), Sentiment Analysis Using Automatically Labelled Financial News Items. In K. Ahmad (szerk.) *Affective Computing and Sentiment Analysis*. Text, Speech and Language Technology. Springer Netherlands, o. 101–114. Elérhető: http://dx.doi.org/10.1007/978-94-007-1757-2_9.

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.

Gidófalvi, Győző & Elkan, Charles, (2003), *Using news articles to predict stock price movements*, San Diego, California. Elérhető: http://people.kth.se/~gyozo/docs/financial-prediction-TR.pdf.

Groth, Sven S., (2012), *Automation in Securities Trading: Text Mining and Algorithmic Trading*. THES. Frankfurt am Main: Johann Wolfgang Goethe-Universität.

Groth, Sven S., (2010), Enhancing Automated Trading Engines To Cope With News-Related Liquidity Shocks. In *18th European Conference on Information Systems*. Pretoria, South Africa.

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. & Muntermann, Jan, (2011), An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50(4), o.680–691. Elérhető: http://www.sciencedirect.com/science/article/pii/S0167923610001430.

Groth, Sven S. & Muntermann, Jan, (2010), Discovering Intraday Market Risk Exposures in Unstructured Data Sources: The Case of Corporate Disclosures. In *43rd Hawaii International Conference on System Sciences (HICSS)*.

Groth, Sven S. & Muntermann, Jan, (2009), Supporting investment management processes with machine learning techniques. In *Business Services: Konzepte, Technologien, Anwendungen - 9, Internationale Tagung Wirtschaftsinformatik*. Wien.

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.

Hernández-Orallo, José, Flach,Peter, Ferri, Cèsar (2012), A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss, Journal of Machine Learning Research, 13, 2813-2869

Kataria, A., Singh, M.D. (2013), A review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and Advanced Engineering*, 3(6), 354-360.

Kaur, Gurneet & Oberai, Er. Neelam (2014), A REVIEW ARTICLE ON NAIVE BAYES CLASSIFIER WITH VARIOUS SMOOTHING TECHNIQUES, *International Journal of Computer Science and Mobile Computing* – IJCSMC, 3(10), 869-878.

Kliegr T. & Kuchař J. (2015), Benchmark of Rule-Based Classifiers in the News Recommendation Task. In: Mothe J. et al. (eds) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Lecture Notes in Computer Science, vol 9283. Springer, Cham, pp 130-141

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.

Koster, Martijn, (1994), A Standard for Robot Exclusion. Elérhető: http://www.robotstxt.org/orig.html [Elérés 2016.08.09.].

Kotsiantis, S. B. (2013), Decision trees: a recent overview. *Artificial Intelligence Review* 39(4), pp 261–283, doi:10.1007/s10462-011-9272-4

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, (2012), Automatic Analysis of in-depth Interviews: Detection of Brand Names and Respondents' Gender. *SEFBIS JOURNAL*, 7(1), o.31–37. Elérhető: http://gikof.njszt.hu/sefbis/SEFBIS_Journal_7evf7szam_2012januar.pdf.

Kovács, Balázs, (2014a), 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, (2014b), 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, (2011), Tájolás és hasznosság mérése rövid szöveges üzenetekben. *ACTA AGRARIA KAPOSVÁRIENSIS*, 15(3), o.103–113.

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.

Kruzslicz, Ferenc, (2015), CRISP-DM vizuális útmutató. Elérhető: http://exam.ktk.pte.hu:81/BI/HF/CRISP-DM-visualguide-HU.pdf [Elérés 2016.08.03.].

Lavrenko, Victor, Lawrie, Dawn, Ogilvie, Paul & Schmill, Matt, (1999), *Ænalyst - Electronic Analyst of Stock Behavior*, Amherst: University of Massachusetts, Department of Computer Science. Elérhető: http://homepages.inf.ed.ac.uk/vlavrenk/doc/pitch.pdf [Elérés 2015.01.01.].

Lavrenko, Victor, Schmill, Matt, Lawrie, Dawn, Ogilvie, Paul, Jensen, David & Allan, James, (2000a), Language Models for Financial News Recommendation. In *Ninth International Conference on Information and Knowledge Management*. McLean, VA: ACM Press, o. 389–396. Elérhető: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.2004.

Lavrenko, Victor, Schmill, Matt, Lawrie, Dawn, Ogilvie, Paul, Jensen, David & Allan, James, (2000b), 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.

Leaper, Nicole, (2009), A visual guide to CRISP-DM methodology. Elérhető: https://exde.files.wordpress.com/2009/03/crisp_visualguide.pdf [Elérés 2016.08.03.].

Leung Kung Fan, Steven, (1997), *Automatic stock market predictions from World Wide Web data*. Hong Kong: The Hong Kong University of Science and Technology. Elérhető: http://hdl.handle.net/1783.1/5633 [Elérés 2015.01.02.].

Mai, Q. (2013), A review of discriminant analysis in high dimensions. WIREs Comp Stat, 5: 190–197. doi:10.1002/wics.1257

Mirtaheri, Seyed M., Dinçtürk, Mustafa Emre, Hooshmand, Salman, Bochmann, Gregor V, Jourdan, Guy-Vincent & Onut, Iosif Viorel, (2013), A Brief History of Web Crawlers. *Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*, o.40–54. Elérhető: http://dl.acm.org/citation.cfm?id=2555523.2555529.

Mittermayer, Marc-André, (2006), *Einsatz von Text Mining zur Prognose kurzfristiger Trends von Aktienkursen nach der Publikation von Unternehmensnachrichten* . THES. Universität Bern. Elérhető: http://www.dissertation.de/index.php3?active_document=buch.php3&buch=4805.

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.

Mittermayer, Marc-André & Knolmayer, Gerhard F., (2006a), NewsCATS: A News Categorization And Trading System. In C. W. Clifton et al. (szerk.) *Sixth International Conference on Data Mining (ICDM'06)*. Hong Kong: IEEE, o. 1002–1007. Elérhető: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=4053143.

Mittermayer, Marc-André & Knolmayer, Gerhard F., (2006b), *Text mining systems for market response to news: A survey*, University of Bern. Elérhető: https://www.researchgate.net/publication/228877499_Text_mining_systems_for_market_response_to_news_A_survey.

Mittermayer, Marc-André & Knolmayer, Gerhard F., (2007), Text mining systems for predicting market response to NEWS. In J. Roth, J. Gutiérrez, & A. P. Abraham (szerk.) *Proceedings of the IADIS European Conference Data Mining*. Lisbon, Portugal, o. 164–169. Elérhető: http://www.iadisportal.org/digital-library/text-mining-systems-for-predicting-market-response-to-news.

Mohai, György, (2010), A Budapesti Értéktőzsde Zártkörűen Működő Részvénytársaság Vezérigazgatójának 427/2010. sz. határozata. Elérhető: http://bet.hu/newkibdata/105599129/KUM_VH427_2010_HU.pdf [Elérés 2015.07.25.].

Moosa, I. A. & Bhatti, R. H. (1997): *International Parity Conditions: Theory, Econometric Testing, and Empirical Evidence*. U. K.: Macmillan Ltd.

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.].

Muntermann, Jan & Guettler, Andre, (2007), Intraday stock price effects of ad hoc disclosures: the German case. *Journal of International Financial Markets, Institutions and Money*, 17(1), o.1–24.

Neocleous C. & Schizas C. (2002), Artificial Neural Network Learning: A Comparative Review. In: Vlahavas I.P., Spyropoulos C.D. (eds) Methods and Applications of Artificial Intelligence. SETN 2002. Lecture Notes in Computer Science, vol 2308. Springer, Berlin, Heidelberg, pp 300-313, DOI: 10.1007/3-540-46014-4_27

Pan, Qi, Cheng, Hong, Wu, Di, Yu, Jeffrey Xu & Ke, Yiping, (2010), Stock risk mining by news. In *Conferences in Research and Practice in Information Technology Series*. o. 179–188.

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.

Peramunetilleke, Desamanya C., (1997), *A system for exchange rate forecasting using news headlines*. Hong Kong: The Hong Kong University of Science and Technology. Elérhető: http://hdl.handle.net/1783.1/5635 [Elérés 2015.01.15.].

Peramunetilleke, Desh & Wong, Raymond K., (2002), Currency exchange rate forecasting from news headlines. In X. Zhou (szerk.) *Database Technologies 2002. Proceedings of the Thirteenth Australasian Database Conference (ADC2002)*. Melbourne, Victoria, Australia: Australian Computer Society, Inc., o. 131–139. Elérhető: http://crpit.com/abstracts/CRPITV5Peramunetilleke.html.

Porter, M.F. (1980): An algorithm for suffix stripping, *Program*, 14(3), pp.130 - 137

Roche, Xavier, Philippot, Yann, Lawrie, David, Lagadec, Robert, Lera, Juan Pablo Barrio, Klueting, Rainer, Gorke, Bastian, Ferrari, Rudi, Gaza, Marcus, Ferrari, Rudi, Jokiel, Lukasz, Fernandes, Rui, Pinheiro, Pedro T. dot, Iliev, Andrei, Krakowski, Witold, Herczeg, Jozsef Tamas, Neto, Paulo, Qin, Brook, Hung, David Hing Cheong, Lin, Addy, Bramm, Jesper, Virma, Tõnu, Ström, Staffan, Köeoðlu, Mehmet Akif, Savic, Aleksandar, Nakasikiryo, Takayoshi, Sereday, Martin, Matìjèík, Antonín, Shevchuk, Andrij, Langhoff, Tobias „Spug", Rudeciur, Jadran, Miron, Alin Gheorghe & Papadakis, Michael, (é. n.), HTTrack. Elérhető: https://www.httrack.-com/.

Sharpe, W. F. (1964): Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance*, 19 (3), 425–442.

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, (2010), A Discrete Stock Price Prediction Engine Based on Financial News. *IEEE Computer*, 43, o.51–56.

Schumaker, Robert P. & Chen, Hsinchun, (2009a), A quantitative stock prediction system based on financial news. *Information Processing & Management*, 45(5), o.571–583. Elérhető: %3CGo.

Schumaker, Robert P. & Chen, Hsinchun, (2008), Evaluating a News-Aware Quantitative Trader: The Effect of Momentum and Contrarian Stock Selection Strategies . *Journal of the American Society for Information and Technology*, 59, o.255–257.

Schumaker, Robert P. & Chen, Hsinchun, (2011), Predicting Stock Price Movement from Financial News Articles. In A. Yap (szerk.) *Information Systems for Global Financial Markets: Emerging Developments and Effects*. New York: IGI Global, o. 96–128.

Schumaker, Robert P. & Chen, Hsinchun, (2009b), Textual analysis of stock market prediction using breaking financial news:The AZFinText system. *ACM Transactions on Information Systems*, 27(2), o.1–19.

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.

Schumaker, Robert P., Zhang, Yulei & Huang, Chun-Neng, (2009), Sentiment Analysis of Financial News Articles. In *20th Annual Conference of International Information Management Association*. Houston.

Schumaker, Robert P., Zhang, Yulei, Huang, Chun-Neng & Chen, Hsinchun, (2012), Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3), o.458–464. Elérhető: http://www.sciencedirect.com/science/article/pii/S0167923612000875.

Sokolova, M., Japkowicz, N., Szpakowicz, S. (2006), Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: Sattar, A., Kang, B. (eds) AI 2006: *Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science*, vol 4304. Springer, Berlin, Heidelberg, 1015-1021

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.

Szabó, Norbert, (2014), *Gazdasági hírek automatikus elemzése*. Pécs: Pécsi Tudományegyetem.

Szőts, Levente, (2014), *Az EURUSD árfolyamának előrejelzése gazdasági hírek szövegének feldolgozásával*. Pécs: Pécsi Tudományegyetem.

Tan, Pang-Ning, Steinbach, Michael & Kumar, Vipin, (2011), *Bevezetés az adatbányászatba* [on-line]., Budapest: Panem Könyvkiadó Kft. Elérhető: http://www.tankonyvtar.hu/hu/tartalom/tamop425/0046_adatbanyaszat/index.html [Elérés 2016.08.02.].

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.

Thomson Reuters, (é. n.), Thomson Reuters Eikon. Elérhető: http://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/eikon-trading-software.html.

Tikk, Domonkos (szerk.), (2007), *Szövegbányászat*, Budapest: Typotex. Elérhető: http://search.ebscohost.com/login.aspx?direct=true&db=cat03034a&AN=pec.-bibFSZ01367231&lang=hu&site=eds-live.

Tumurkhuu, Tserendash & Wang, Xiaojing, (2010), *The relationship between the profit warning and stock returns: Empirical evidence in EU markets*. Umeå School of Business. Elérhető: http://umu.diva-portal.org/smash/get/diva2:394405/FULLTEXT01.pdf.

Wu, Di, Fung, Gabriel Pui Cheong, Yu, Jeffrey Xu & Liu, Zheng, (2008), Integrating Multiple Data Sources for Stock Prediction. In J. Bailey et al. (szerk.) *Web Information Systems Engineering - WISE 2008: 9th International Conference, Auckland, New Zealand, September 1-3, 2008. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, o. 77–89. Elérhető: http://dx.doi.org/10.1007/978-3-540-85481-4_8.

Wu, Di, Fung, Gabriel Pui Cheong, Yu, Jeffrey Xu & Pan, Qi, (2009), Stock prediction: An event-driven approach based on bursty keywords. *Frontiers of Computer Science in China*, 3(2), o.145–157.

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.

Wüthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J. & Lam, W., (1998), Daily prediction of major stock indices from textual www data. In R. Agrawal & Stolorz Paul (szerk.) *Proceedings of the Fourth Knowledge Discovery and Data Mining Conference*. New York, New York: The AAAI Press. Elérhető: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.2979.

Wüthrich, Beat, (2002), Industry: predicting daily stock indices movements from financial news. In Klösgen Willi & J. M. Żytkow (szerk.) *Handbook of data mining and knowledge discovery*. New York: Oxford University Press, Inc., o. 910–920. Elérhető: http://dl.acm.org/citation.cfm?id=778339.

Yu, Ting, Jan, Tony, Debenham, John K. & Simoff, Simeon J., (2006), Classify Unexpected News Impacts to Stock Price by Incorporating Time Series Analysis into Support Vector Machine. In *The 2006 IEEE International Joint Conference on Neural Networks Proceedings*. Vancouver, BC, Canada: IEEE, o. 2993–2998. Elérhető: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1716505.

Yu, Ting, Jan, Tony, Debenham, John & Simoff, Simeon, (2005), Incorporate domain knowledge into support vector machine to classify price impacts of unexpected news . In S. J. Simoff et al. (szerk.) *Proceedings 4th Australasian Data Mining Conference* . Sydney, o. 1–11. Elérhető: http://static.googleusercontent.com/media/research.google.com/hu//pubs/archive/32722.pdf.

Zeller, Gyula (2001): *Bevezetés a származékos értékpapírpiacokba*, Pécs, Pécsi Tudományegyetem Közgazdaságtudományi Kar, ISBN: 963 6417 490

Zhang, Debbie, Simoff, Simeon & Debenham, John, (2007), Exchange Rate Modelling for E-Negotiators Using Text Mining Techniques. In J. Lu, G. Zhang, & D. Ruan (szerk.) *E-Service Intelligence*. Studies in Computational Intelligence. Springer Berlin Heidelberg, o. 191–211. Elérhető: http://dx.doi.org/10.1007/978-3-540-37017-8_8.

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/10.1007/11589990₄9.

Zhang, Debbie, Simoff, Simeon J. & Debenham, John, (2006), Identification of Important News for Exchange Rate Modeling. In M. Bramer (szerk.) *Artificial Intelligence in Theory and Practice*. IFIP International Federation for Information Processing. Santiago, Chile: Springer US, o. 475–482. Elérhető: http://dx.doi.org/10.1007/978-0-387-34747-9_49.

FÜGGELÉK

1. függelék: Név- és tárgymutató

2. függelék: Gyakori n-grammok kinyerésére szolgáló algoritmus

Az algoritmus leírása során *n-grammok* alatt tokenek tetszőlegesen hosszú, véges sorozatát értem, azaz e sorozat hossza legalább kettő, és felső korlát nincs, így akár a teljes dokumentum is tekinthető egy n-grammnak. Szemantikai szempontból az n-grammok két fő csoportba tartoznak: *gyakori szófordulatokra*, illetve *sablonszövegekre*. *Gyakori szófordulatok* alatt olyan kifejezéseket értek, amelyek egy nyelvben vagy egy nyelv meghatározott részterületén – pl. egy szakterület által használ nyelvben – egy meghatározott alsó korlátnál gyakrabban fordulnak elő és törlésük súlyos nehézségeket okozna az adott nyelv kommunikációs folyamataiban. A gyakori szófordulatok jellemzően rövidebbek, 2–5 eleműek, például ilyen a „cash-flow from investing activities" kifejezés. *Sablonszövegek* alatt olyan n-grammokat értek, amelyek egy nyelvben vagy egy szakterületen egy alsó korlátnál gyakrabban fordulnak elő és törlésük nem jelent súlyos nehézséget, mivel könnyen helyettesíthetők más összetételű n-grammokkal. Sablonszövegek jellemzően hosszabbak, akár több mondatosak is lehetnek, ilyenek pl. a szerződésminták szövegei, vagy egy vállalat sajtóközleményeiben újrahasznosított szövegrészek.

Az n-grammok kinyerését végző algoritmus főbb paraméterei a következők: *lower_limit*, *step*, *start*. A *start* paraméter az n-grammok gyakoriságának alsó korlátja az algoritmus első lépésében, a *lower_limit* paraméter pedig az algoritmus egyik megállási feltétele, az n-grammok gyakoriságának alsó korlátja az algoritmus utolsó lépésében. A *step* paraméter az egy lépésben kinyert n-grammok minimális számát jelenti. E paramétereknek eleget kell tenniük a következő feltételeknek: $1 \leq lower\_limit \leq start$, $1 \leq step$, illetve mindhárom paraméter egész szám. A *start* paraméter a korpusz elemszáma alapján heurisztikusan kerül megállapításra a folyamat elején, egész pontosan a korpuszban lévő dokumentumok számának a felének az egészrészét alkalmaztam. A másik két paraméter alapértelmezett értéke *lower_limit* = 5 és *step* = 100. Ezek a paraméterek befolyásolják a kifejezések keresésének iterációs lépéseit. Az iteráció a *start* paraméter értékétől indul és a *lower_limit* paraméter értékéig csökken. A *step* befolyásolja, hogy milyen gyorsan jut el az algoritmus a *start*-tól a *lower_limit*-ig, azonban csak közvetve, ugyanis a *step* nem az iterációs ugrás hosszával azonos. Legrosszabb esetben az iterációk száma $start - lower\_limit + 1$ lesz, különben ha egységnyi iterációs ugrással nem keletkezne *step* mennyiségű új n-gramm, akkor annyi egységnyi iterációs ugrás történik, amennyivel legalább *step* mennyiségű új n-gramm keletkezik.

Az algoritmus agglomeratív módon állítja elő a különböző hosszúságú n-grammokat. Az első lépésben azok a bi-grammok kerülnek a kifejezések halmazába, amelyek dokumentumgyakorisága eléri legalább a *start* paraméter értékét vagy, ha ilyen nincs, akkor a dokumentumgyakoriság szerint csökkenő sorba rendezett bi-grammok közül a *step* paraméternek megfelelő helyen – alapértelmezésben a századik helyen – lévő elem dokumentumgyakoriságát. Az új n-grammok kiemelésre kerülnek a szövegben.

A következő iterációban azok a bi- és tri-grammok kerülnek kiemelésre, amelyek dokumentumgyakorisága eléri legalább a dokumentumgyakoriság szerint rendezett n-gramm lista *step*-edik helyén lévő elemének dokumentumgyakoriságát. A tri-grammok esetén csak olyan tokensorozatokat veszünk figyelembe, amelyek a korábbi lépésben kiemelt valamely bi-gram és egy újabb token összekapcsolásával keletkeztek. Az új n-grammokat kiemeljük a szövegben.

A *k*-dik iterációkban azok a 2-, 3-, … (k+1)-grammok kerülnek a kifejezések halmazába, amelyek dokumentumgyakorisága eléri a *step*-edik elemét, valamint amelyek két olyan kifejezés, vagy egy tetszőleges token és egy olyan kifejezés összekapcsolásával keletkeztek, amely a *k*-dik iteráció előtt már kiemelésre került. Nevezzük az ilyen kifejezést az őt alkotó kifejezések közvetlen utódának, továbbá az őt alkotó kifejezéseket az ő közvetlen elődeinek. Az olyan kifejezéseket, amelyek közvetlen elődök, illetve utódok láncolatán keresztül állnak kapcsolatban értelemszerűen egymás elődének, illetve utódának tekinthetők. Könnyen belátható, hogy egy utód legfeljebb olyan gyakori, mint a közvetlen elődjei közül a kevésbé gyakori.

Az algoritmus előfeltétele, hogy az n-gramm jelöltek listája tartalmazzon legalább *step* számú elemet, azonban ennek nincs gyakorlati jelentősége, ugyanis a listák mérete általában több ezer elemű, míg a *step* értékét célszerű minél kisebbre választani, amennyit a hardver teljesítménye lehetővé tesz. A *step* értékének növelésével megnő annak az esélye, hogy egy ritkább tokensorozat előbb kerül kiemelésre, mint egy hasonló tokenekből álló gyakoribb tokensorozat, és ezáltal a gyakoribb tokensorozat potenciális utódainak dokumentumgyakorisága kisebb lesz a ténylegesnél. A *step* ideális értéke tehát 1, azonban az algoritmus jóval lassabb lesz, mivel többször elő kell állítani az n-grammok listáját, illetve többször át kell vizsgálni a szöveget a kinyert n-grammok kiemelése miatt. Az első iterációkban az egy lépésben kinyert kifejezések száma jobban megközelíti a *step* paraméter értékét – felülről –, mivel a leggyakoribb kifejezések száma csekély, viszont az utolsó iterációkban jóval meghaladja majd a lista mérete a *step*-et. Ez jelentő-

sen lelassítja az iterációk végrehajtását, így ebből a szempontból is célszerű meghatározni egy alsó korlátot a dokumentumgyakoriságra. Ezt szolgálja a *lower_limit*, amely ugyanakkor egy szubjektív mértéke is annak, hogy milyen gyakori n-grammokat tekintünk kifejezésnek. A *lower_limit* megállapításakor tekintettel kell lenni a korpusz méretére is, nagyobb korpusz esetén megengedhető nagyobb *lower_limit* érték is.

3. függelék: A modell pontosságalapú minőségmutatójának levezetése

A következőkben egy saját minőségmutatót konstruálok, amely alapján különböző modellek pontossága is összevethető. Ehhez egy egyszerű hétköznapi mutatóból indulok ki, majd lépésenként egy-egy hiányosságot kiküszöbölő, viszonylag triviális mutatókon keresztül haladok. A mutató kialakításánál a fő szempont, hogy alulról és felülről is korlátos legyen; értéke legyen 0, ha a modell a defaulttal azonos minőségű; pozitív legyen, ha a modell jobb, negatív, ha a default jobb; ha a modell 100%-os, a default 0%-os, akkor vegye fel a felső korlátot, fordított helyzetben pedig az alsó korlátot; és végül legyen szimmetrikus.

Az osztályozás minőségének kézenfekvő mutatószáma a pontosságmutatók abszolút különbsége:

$$q_1 = a_m - a_d \tag{64}$$

Ahol:

$a_m \in [0;1]$ a modell pontossága

$a_d \in [0;1]$ a default pontosság

A q_1 -gyel jelölt mutató a következő fontos tulajdonságokkal rendelkezik:

Ha $a_m = 100\%$ és $a_d = 0\%$, akkor $q_1 = 100\%$.

Ha $a_m = 0\%$ és $a_d = 100\%$, akkor $q_1 = -100\%$.

Ha $a_m = a_d$, akkor $q_1 = 0\%$.

Szintvonalai egyenesek, párhuzamosak és meredekségük 1, egyenletük: $a_d = a_m - q_1$ (65). Gradiense állandó és koordinátái: $\vec{g}_{q_1} = \left(\frac{\partial q_1}{\partial a_m}; \frac{\partial q_1}{\partial a_d} \right) = (1;-1)$ (66).

Ezek a százalékpontban mért különbségek csak azonos bázishoz viszonyítva összehasonlíthatók, de a kategóriák megoszlása jelentősen különböző lehet, így a_d jelentősen eltérhet az egyes esetekben. A mutató relatív változata, amely az abszolút különbséget a default pontossághoz viszonyítja:

$$q_2=\frac{a_m-a_d}{a_d}=\frac{a_m}{a_d}-1 \text{ ahol } a_d\neq 0 \quad (67)$$

A q_2 -vel jelölt mutató a következő fontos tulajdonságokkal rendelkezik:

Ha $a_m=100\%$ és $a_d\to 0\%$, akkor $q_2\to\infty$.

Ha $a_m=0\%$ és $a_d=100\%$, akkor $q_2=-100\%$.

Ha $a_m=a_d$, akkor $q_2=0\%$.

Szintvonalai az origón áthaladó egyenesek, egyenletük: $a_d=\frac{1}{q_2+1}a_m$ (68).

Gradiense: $\vec{g}_{q_2}=(\frac{1}{a_d};-\frac{a_m}{a_d^2})$ (69). Mivel $\frac{1}{a_d}>0$ és $-\frac{a_m}{a_d^2}\leq 0$, a_m növekedése, illetve a_d csökkenése q_2 értékét növeli.

E mutató hátránya, hogy $a_d\to 0$ esetén nincs felső korlátja. A mutató továbbfejleszthető a nullával való osztás elkerülésére bevezetett Laplace-korrekcióval. A korrekciót a pontosság számításakor hajtsuk végre olyan módon, hogy a mintához hozzáveszünk egy definíció szerint helyesen osztályozott megfigyelést[129].

$$a_j=\frac{|H_j|+1}{|I|+1} \quad (70)$$

Ahol

$H_j=(i|p_i=c_i, i\in I)$: a helyesen becsült címkéjű megfigyelések halmaza a $j\in\{m,d\}$ modell esetén

p_i : az i jelű megfigyelés becsült címkéje

c_i : az i jelű megfigyelés tényleges címkéje

I : az ismert címkéjű megfigyelések halmaza, $|I|>0$

Ebből a korrigált mutató a következőképpen adódik:

$$q_3=\frac{\frac{|H_m|+1}{|I|+1}-\frac{|H_d|+1}{|I|+1}}{\frac{|H_d|+1}{|I|+1}}=\frac{|H_m|-|H_d|}{|H_d|+1}=\frac{a_m-a_d}{a_d+\frac{1}{|I|}}=\frac{a_m+\frac{1}{|I|}}{a_d+\frac{1}{|I|}}-1 \quad (71)$$

A q_3 -mal jelölt mutató a következő fontos tulajdonságokkal rendelkezik:

Ha $a_m=100\%$ és $a_d=0\%$, akkor $q_3=|I|$.

Ha $a_m=0\%$ és $a_d=100\%$, akkor $q_3=-\frac{|I|}{|I|+1}$.

129 Nem lényeges, hogy melyik kategóriához, mert ez a pontosságmutató értékét nem változtatja, így az általánosságot nem sérti, ha a default kategóriához vesszük hozzá a megfigyelést

Ha $a_m = a_d$, akkor $q_3 = 0\%$.

Szintvonalai a $\left(-\frac{1}{|I|};-\frac{1}{|I|}\right)$ ponton átmenő egyenesek, egyenletük:

$$a_d = \frac{1}{1+q_3}\left(a_m + \frac{1}{|I|}\right) - \frac{1}{|I|} \quad (72).$$

Gradiense: $\vec{g}_{q_3} = \left(\frac{1}{a_d + \frac{1}{|I|}}; -\frac{a_m + \frac{1}{|I|}}{\left(a_d + \frac{1}{|I|}\right)^2}\right)$ (73)

Gradiensének első eleme pozitív, második eleme negatív az értelmezési tartományon, tehát a_m növekedésekor vagy a_d csökkenésekor q_3 nő.

E mutatónak van felső korlátja, valamint $a_d = 0$ -ra értelmezett, azonban nem normalizált a megfigyelések számára vonatkozóan. A q_3 mutatóhoz hasonló tulajdonságú szintvonalakkal rendelkező $[-z;z]$ – ahol $z>0$, de a normalizáláshoz célszerű $z=1$ -et választani – intervallumba normalizált mutatót kapunk a következő módon:

$$q_4 = \log_{\sqrt[z]{|I|+1}}\left(\frac{a_m + \frac{1}{|I|}}{a_d + \frac{1}{|I|}}\right) \quad (74)$$

A q_4 -gyel jelölt mutató a következő fontos tulajdonságokkal rendelkezik:

Ha $a_m = 100\%$ és $a_d = 0\%$, akkor $q_4 = z$.

Ha $a_m = 0\%$ és $a_d = 100\%$, akkor $q_4 = -z$.

Ha $a_m = a_d$, akkor $q_4 = 0\%$.

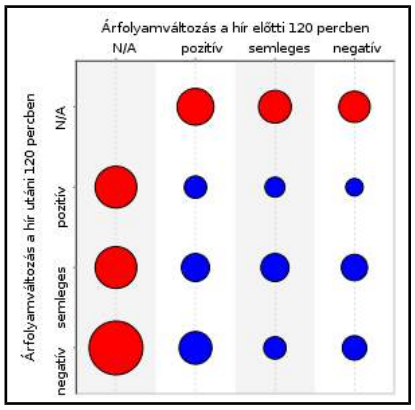
Szintvonalai a $\left(-\frac{1}{|I|};-\frac{1}{|I|}\right)$ ponton átmenő egyenesek, egyenletük:

$$a_d = \frac{1}{\sqrt[z]{|I|+1}^{q_4}}\left(a_m + \frac{1}{|I|}\right) - \frac{1}{|I|} \quad (75).$$

Gradiense: $\vec{g}_{q_4} = \left(\frac{\frac{z}{\ln(|I|+1)}}{a_m + \frac{1}{|I|}}; -\frac{\frac{z}{\ln(|I|+1)}}{a_d + \frac{1}{|I|}}\right)$ (76), melynek első eleme pozitív, második eleme negatív az értelmezési tartományon, tehát a_m növekedésekor vagy a_d csökkenésekor q_4 nő.

4. függelék: A különböző időablakok közötti asszociációs kapcsolat tesztelése

A 9. ábra −120 -as értékénél az látszik, hogy a hírek publikálása előtti két órában viszonylag nagy az árfolyam-emelkedés aránya a mintában, míg azonban a 120 -as értéknél leolvasható, hogy a publikálás utáni két órában az árfolyamcsökkenés aránya nagyobb. Ebből tehát arra lehet következtetni, hogy a közlemények kibocsátása előtti és utáni időszakok hozameloszlásai eltérő oldali asszimmetriával jellemezhetők. Ez pedig azt jelentheti, hogy a hírek egy részének hatására megfordul az árfolyam növekvő trendje. Ez utóbbi állítás ellenőrzésére készítettem el a 32. ábrán és 24. táblázatban látható kereszttáblát a hírek előtti és utáni két-két órás időszakokra, amelyből látható, hogy az egyes árfolyamtrendek hogyan változtak a hír előtti állapothoz képest a hír után. Mivel viszonylag sok azoknak a híreknek a száma, amelyeket vagy azért kellett kizárni, mert a nyitáshoz képest túl korán tették őket közzé, vagy azért, mert a zárás előtt túl későn, ezért a trendváltások csak 57 esetben vizsgálhatók meg. A 32. ábrán látható buborékdiagram mérete a 24. táblázatban található számokkal arányos, és az említett 57 megfigyelés megoszlását a kék színnel jelzett buborékok mérete jelzi.



32. Ábra: Az árfolyamtrend megváltozása a hír után

Forrás: saját szerkesztés

24. Táblázat: Az árfolyamtrend lehetséges megváltozásainak megoszlása a hír után

		Árfolyamváltozás a hír előtti 120 percben				
		N/A	pozitív	semleges	negatív	Összesen
Árfolyamváltozás a hír utáni 120 percben	N/A	0	14	11	10	35
	pozitív	18	5	4	3	30
	semleges	18	8	8	7	41
	negatív	30	11	5	6	52
	Összesen	66	38	28	26	158

Forrás: saját szerkesztés

Azt, hogy valóban van összefüggés a trendek átmenetében, khi-négyzet próbával tesztelhetjük a pozitív-semleges-negatív értékekre. Átlagosan hatnál több megfigyelés jut egy cellára, így érdemes a próbát elvégezni. A négy szabadságfokú khi-négyzet próba során kapott p-érték 85,78%, így minden ésszerű szignifikancia szinten el kell vetnünk a nullhipotézist, azaz nincs összefüggés a hír előtti kétórás ármozgás és a hír utáni kétórás ármozgás iránya között.

Ugyanilyen vizsgálatot elvégezhetünk minden lehetséges időablak-párosításra, amelyből látható, hogy két különböző időeltolás esetén mennyire különböznek az egyes dokumentumokhoz rendelt címkék – azaz mennyire különböző a modell taníthatósága a két esetben. Ez összesen $\binom{120+120}{2}=28680$ teszt elvégzését jelenti, amely gyakorlatilag csak vizuális formában áttekinthető, amire a 33. és a 34. ábra szolgál. Az ábrákon három blokk figyelhető meg. Az első blokkban azon párosítások találhatók, melynek mindkét tagja a sajtóközlemény kibocsátása előtti időszakra vonatkozik, azaz mindkét eltolás negatív. Ez a bal-alsó háromszög. A második blokk azon párosításokat mutatja, amelyek mindkét tagja a sajtóközlemény utáni időszakra vonatkozik, itt mindkét eltolás pozitív. Ez a jobb-felső háromszög. Végül a harmadik blokk azon párosításokból áll, amikor az egyik időablak a hír előtti időszakra vonatkozik, a másik a hír utánira. Ez a jobb-alsó négyzetnek felel meg. A 0 eltolást – avagy 0 hosszúságú időablakot – nem értelmezzük, ezért az ábrán a 0 abszcisszánál és ordinátánál nem látunk teszteredményeket.

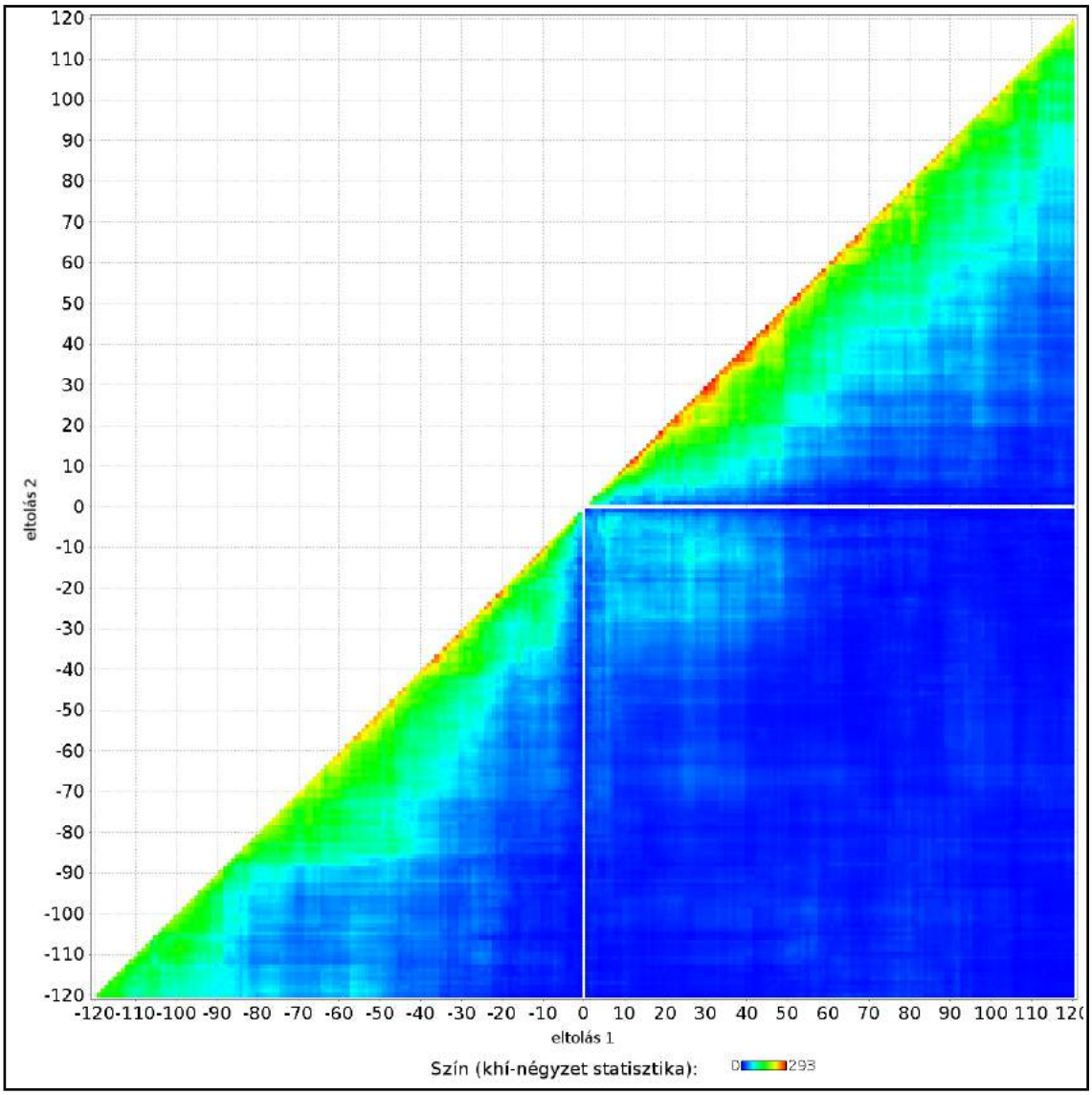
A szignifikancia vizsgálatához négy szintet különböztettem meg, mely alapján öt kategóriába soroltam a tesztek eredményét:

- a 0-s kategóriába kerültek azok, amelyek 10%-on nem voltak szignifikánsak,
- az 1-es kategóriában lévők 10%-on szignifikánsak, de 5%-on nem,
- a 2-es kategóriában lévők 5%-on szignifikáns, de 1%-on nem,
- a 3-as kategóriában lévők 1%-on szignifikánsak, de 0,1%-on nem,
- végül a 4-es kategóriában lévők 0,1%-on szignifikánsak.

E tesztek eredményét mutatja a 34. ábra. Mind a bal-alsó, mind a jobb-felső háromszög-blokk egymással részben átfedő időablakok árfolyamtrend-kategóriái közötti asszociációit jellemzi, így minél kisebb a különbség a két időablak között, annál több a közös kategória-címke, így annál hasonlóbb a megoldandó osztályozási feladat is. Ez a 33. ábrán a piros-sárga-zöld-kék sávok elhelyezkedéséből olvasható ki, illetve megerősítést nyer a a 34. ábrán, hogy szignifikánsak is. Sokkal gyengébb ez a hatás a nagyon rövid időablakok esetén, mivel ott nagy annak a valószínűsége, hogy a következő percben akkora árfolyamváltozás történik, amely elegendő ahhoz, hogy megváltozzon a trend előző egy-két percben tapasztalt iránya. Másként fogalmazva a hír után megfigyelt egy-két perc árfolyamváltozásából nem lehet következtetni a későbbi órák árfolyamváltozására. A bekezdésben leírtak illusztrálására szolgál a 35. ábra.

A 35. ábrán két-két időablak közötti kontingencia táblák alapján számolt relatív reziduumok láthatók. A buborékok mérete a relatív reziduum abszolút értékének felel meg, míg színe az előjelének. A bal oldali ábrán két, hasonló hosszúságú időablak látható, amely jól illusztrálja, hogy sokkal valószínűbb, hogy az időablak kis növelésével ugyanazokat a címkéket kapjuk. A jobb oldalin egy rövid és egy hosszú időablak címkéi közötti összefüggéseket láthatjuk, pontosabban annak hiányát.

A 33. és a 34. ábra jobb-alsó négyzetét illetően olyan időablakok közötti asszociációt vizsgálunk, amelyekben nincs átfedés, egyik ablak a publikálás előttre nyúlik, másik a publikálás utánra. A $120 \cdot 120 = 14400$ darab ilyen párosítás közül a -120 és $+120$ közöttit már megvizsgáltuk, ott nem találtunk szignifikáns összefüggést, de látható,

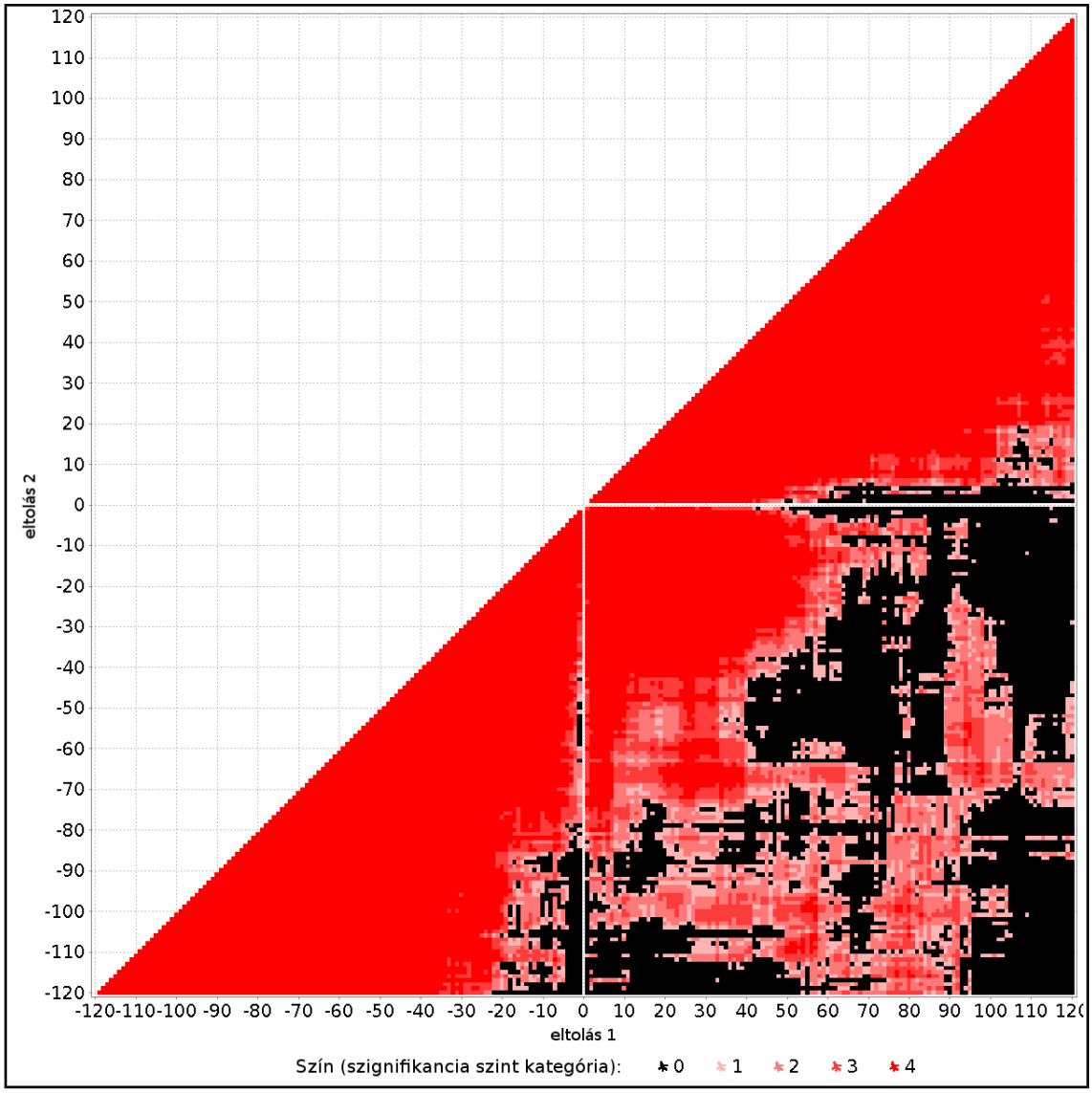


33. Ábra: A különböző eltolású időablakok árfolyamcímkéi közötti asszociációt számszerűsítő khi-négyzet mutatók értékei

Forrás: saját szerkesztés

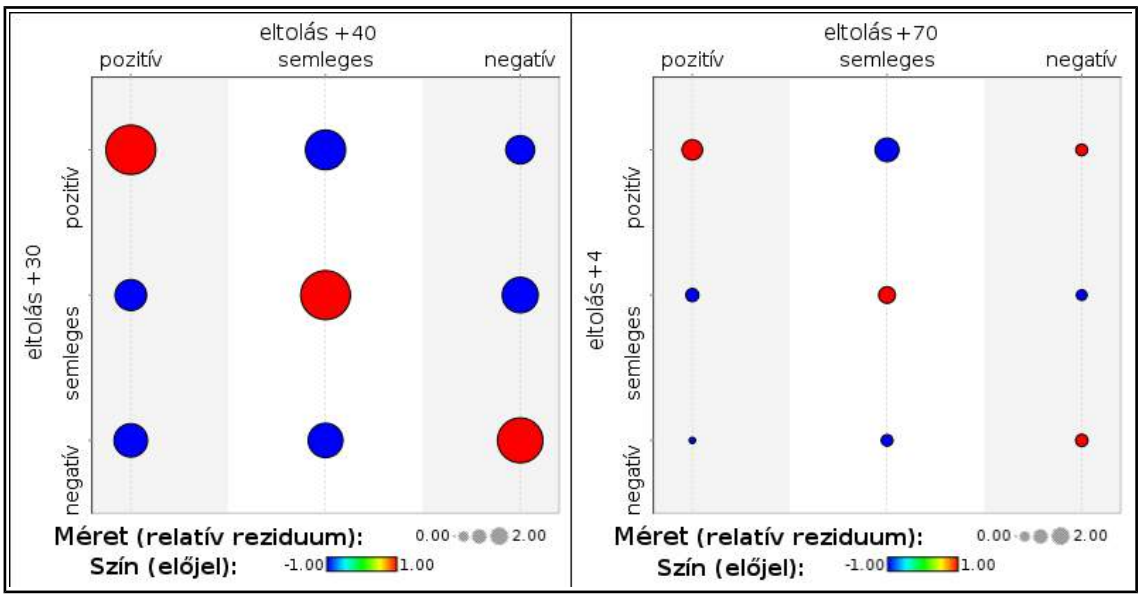
hogy különböző szignifikancia szinteken elég sok esetben van összefüggés, főleg az 50 percnél rövidebb időablakok között. Ez az összefüggés azonban nem a 35. ábrán láthatóval azonos, azaz nem azt fejezi ki, hogy a hír utáni ármozgás a hír előttivel azonos. A hír előtti és utáni árfolyamkategóriák összefüggéséről ehelyett az mondható el, hogy:

- a semleges trendek nagyobb valószínűséggel folytatódnak, minthogy a hír után pozitív vagy negatív trendbe forduljanak,
- a pozitív és negatív trendek a hír után alacsonyabb valószínűséggel váltanak át stagnálásba, ehelyett vagy folytatódnak, vagy megfordulnak inkább.



34. Ábra: A különböző eltolású időablakok árfolyamcímkéi közötti asszociációt számszerűsítő khi-négyzet próbák eredménye különböző szignifikancia szinteken

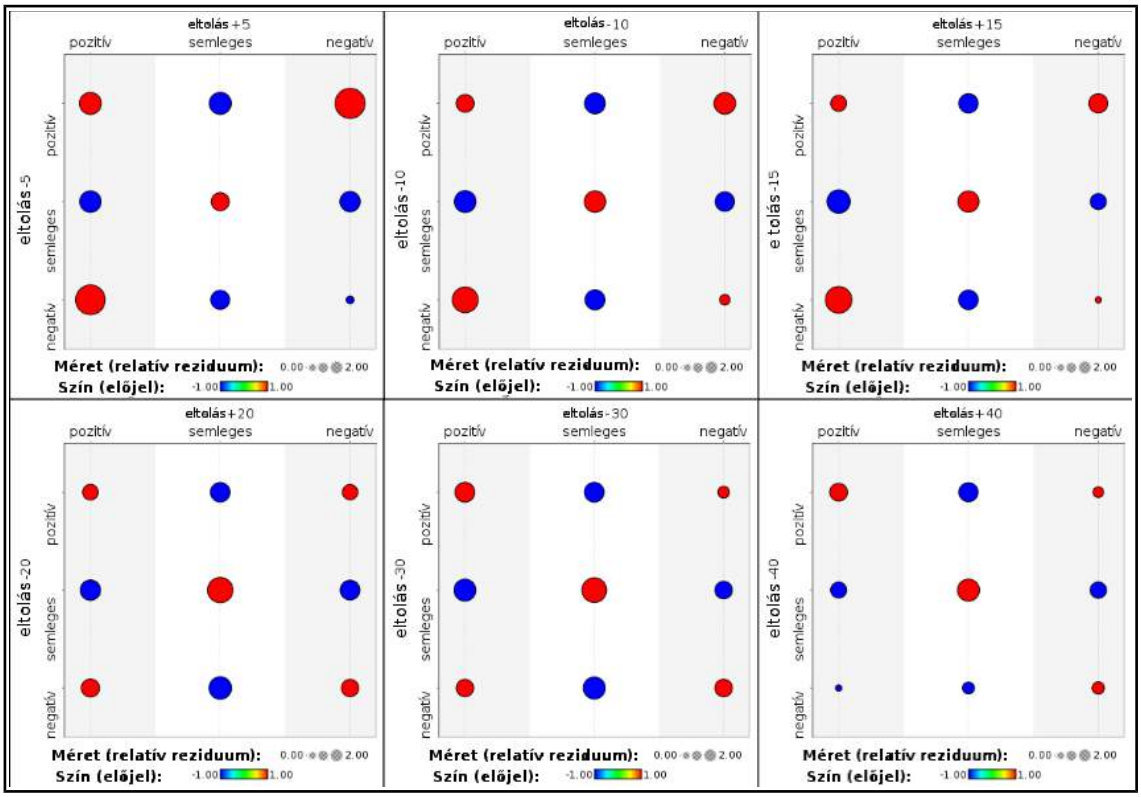
Forrás: saját szerkesztés



35. Ábra: A különböző időtávokra képzett tanítóminták címkéi közötti összefüggések (példa erős és gyenge asszociációra)

Forrás: saját szerkesztés

Erre mutat példákat a 36. ábra, amelyen a fenti szabályok a piros buborékok által kiformált X-ek alakjában jelennek meg.



36. Ábra: Példák a hír előtti és utáni időtávokra képzett tanítóminták címkéi közötti összefüggésekre

Forrás: saját szerkesztés