# INTERNETES TERMÉKKRITIKÁK HASZNOSSÁGÁNAK MEGÁLLAPÍTÁSA FELÜGYELT GÉPI TANULÁSSAL[1]

KOVÁCS BALÁZS – KRUZSLICZ FERENC – TORJAI LÁSZLÓ
*PTE KTK – BDE Research Kft.*

Az elmúlt évek során az Internet a vállalati marketing funkció számára is az egyik legfontosabb információforrássá nőtte ki magát. Ennek megfelelően egyre több kutatás foglalkozik az internetes felhasználók által generált dokumentumok hasznosítási lehetőségével is. A termékkritikákban (reviews, comments) rejlő információ kinyerését célzó kutatási irányok egyike, az ún. koncepció kinyerés (concept extraction), ami többek közt a termékekre vonatkozó fogyasztói ítéleteket tárja fel és elemzi. A vizsgálat fókusza lehet a felhasználói hozzászólások tartalma, de azok minősége, hasznossága is. Cikkünkben áttekintjük a termékkritika-hasznosság fogalmához kapcsolódó különböző értelmezési megközelítéseket. Célunk, hogy a termékkritikák hasznosságának automatikus megállapításához dolgozzunk ki egy mesterséges neurális hálózatot és egy Support Vector Machine-t alkalmazó felügyelt tanulási eljárást, melyben különböző szövegjellemző halmazokat használunk a tanítás során.

## 1 Bevezetés

Az internetes technológiákra épülő üzleti megoldások terjedésével rohamosan nő az olyan típusú weblapok száma, ahol egy termékről leírhatjuk tapasztalatainkat, véleményünket vagy valamilyen skálán minősíthetjük azt. Lehet ez egy webáruház, egy fórum vagy akár egy elektronikus szaklap is. Az ilyen típusú, online felhasználók által írt véleményeket digitális szájreklámnak/szóbeszédnek (word of mouth) tekinthetjük (Dellarocas, 2003), amik jelentős befolyással bírnak a vásárlási döntésekre, és így az értékesítési eredményekre is (Duan és Whinston, 2008). Zhu és Zhang (2010) azt találták, hogy az internethasználók 24%-a tájékozódik online termékkommentekből, mielőtt offline vásárol. Ugyanakkor nehezíti az információgyűjtést, hogy az említett termékvélemények szétszórtan és rosszul strukturáltan jelennek meg a világhálón, ráadásul egy átlagos felhasználó nehézségekbe ütközhet, ha a dokumentumokban található információk hitelességét akarja megítélni.

A termékvélemény-keresők a fenti problémára kínálnak megoldást: a potenciális vásárlók információgyűjtési tevékenységét hivatottak leegyszerűsíteni és hatékonyabbá tenni. Mindezt azzal érik el, hogy

[1]Jelen tanulmány a „TÁMOP-4.2.2.C-11/1/KONV-2012-0005, Jól-lét az információs társadalomban" pályázati projekt támogatásával készült. Kovács Balázs kutatómunkáját részben a Felkai András ösztöndíj tette lehetővé, melyet a Citibank kezdeményezésére az Alapítvány a Pénzügyi Kultúra Fejlesztéséért hozott létre. Beérkezett: 2012. november 23. E-mail: kovacsb@ktk.pte.hu, kruzslic@ktk.pte.hu, torjai.laszlo@bde.hu.

- a termékvéleményekkel kapcsolatos információkat kinyerik a weblapokról,
- feldolgozzák, rendszerezik, értékelik azokat, valamint
- megfelelő felületet biztosítanak az eredmények lekérdezéséhez és megjelenítéséhez.

Ezen megoldások alkalmazhatósága azonban túlmutat a potenciális vásárlók információéhségének kielégítésén. Üzleti szervezetek számára az egyik legkézenfekvőbb felhasználási terület a vevőelégedettség-vizsgálat, amit az érzelmi tájolás elemzés (sentiment orientation analysis) segítségével lehet elvégezni (Li és Wu, 2010). Ehhez az egyes kritikákban megfogalmazott véleményeket valamilyen minőségi skálán (például jó-semleges-rossz vélemény a termékről) kell ábrázolni. Az adott termékre vonatkozó érzelmi tájolások megfelelő összegzésével az individuális fogyasztók véleményét aggregált vásárlói preferenciává alakíthatjuk át (Decker és Trusov, 2010). Amennyiben a felhasználók regisztrálva közölnek véleményeket, úgy az összegyűjtött ügyféladatbázist felhasználva személyre szabott reklámok készíthetők és juttathatók el a címzettekhez (Cheung és sztsai, 2003). Hasonló megközelítés alkalmazható az ügyfélszolgálatokra beérkező üzenetek (ügyfélpanaszok, elismerések stb.) elemzése kapcsán is (Burk, 2007; Coussement és Van den Poel, 2008).

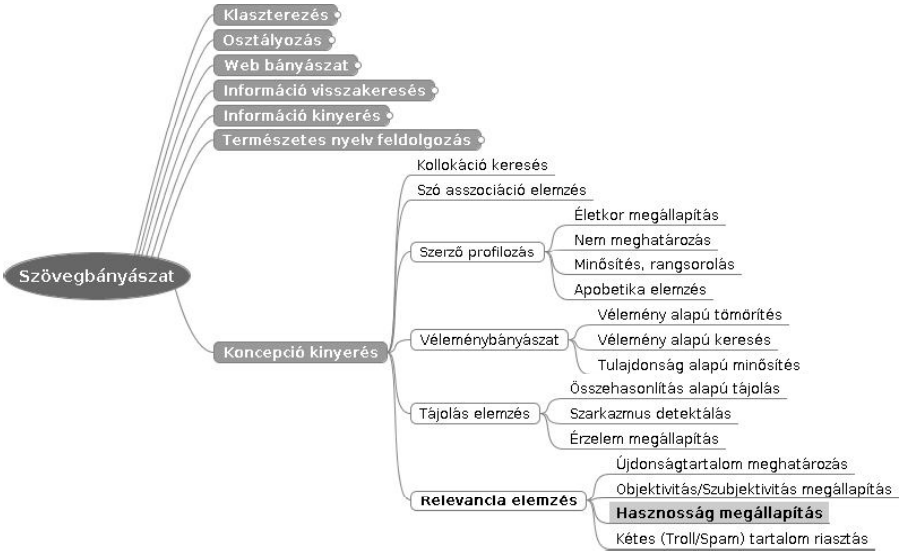
A kutatások egy másik iránya nem a hozzászólásokban található vélemények érzelmi tájolását elemzi, hanem a bejegyzés minőségét, információtartalmát, hitelességét. A továbbiakban összefoglalóan a hozzászólások hasznosságaként hivatkozunk e kategóriákra. Ezen elemzések elvégzését több dolog is motiválhatja:

- Gyakorlatilag lehetetlen feladat egy potenciális vásárló számára, hogy az összes releváns hozzászólást elolvassa, főként olyan népszerű termékek esetén, melyekről felhasználók ezrei mondják el saját véleményüket. Csak a leghasznosabb hozzászólásokat kell megjeleníteni számára.
- A termékkommentek szerzői vagy a megjelenítő weblapok különböző ,,hitelességi szinten" helyezkedhetnek el, születnek hozzászólások félrevezetési céllal és találkozhatunk spamekkel is (Duan és Zirn, 2012; Xie és sztsai, 2012).
- Az aggregált vásárlói preferencia felmérésénél a különböző információtartalmú, minőségű vagy korú bejegyzéseket más-más súllyal érdemes figyelembe venni.
- Stb.

A termékkritikákat közlő weboldalak egy része lehetőséget ad arra, hogy a felhasználók értékeljék az általuk olvasott hozzászólások hasznosságát. Az Amazon.com-on például arra az eldöntendő kérdésre kell választ adniuk az olvasóknak, hogy hasznosnak találták-e a kritikát (Was this review helpful to you?). Az így kapott eredményeket fel lehet használni a fenti célokra,

de számos esetben félrevezető lehet ezek alkalmazása: a bejegyzések jelentős százalékára például nem, vagy alig érkezik visszajelzés, ami – függetlenül a hozzászólás valódi hasznosságától – csökkenti annak esélyét, hogy újabb felhasználók olvassák és értékeljék azt (Kim és sztsai, 2006; O'Mahony és Smyth, 2010) (ezen problémáról bővebb leírást adunk a következő szakaszban). Nehézséget okozhat az is, hogy különböző forrásból származó hozzászólások más módszerrel és olvasói bázis által kerülnek értékelésre, így a nyert mutatók nem összehasonlíthatók. Ezen korlátok motiválták a hozzászólások hasznosságának automatikus (gépi) meghatározására irányuló törekvéseket.

Az *1. ábra* a szövegbányászati, azon belül pedig a koncepció kinyerési kutatások egy lehetséges kategorizálását mutatja.




*1. ábra.* A „hasznosság megállapítás" helye a szövegbányászati kutatásokban

Tanulmányunk célja, hogy egy olyan gépi tanuláson alapuló eljárást dolgozzunk ki, melynek segítségével automatizálni lehet a termékekre vonatkozó internetes hozzászólások hasznosságának megállapítását. Az eljárás kidolgozásához magyar nyelvű, mobiltelefonokra adott kommentek 1000 elemű korpuszát használtuk fel.

Cikkünk második szakaszában röviden összefoglaljuk, hogy a nemzetközi irodalomban milyen megközelítések és megoldások születtek a kommentek hasznosságának értékelésére. A harmadik szakaszban sorra vesszük azon szövegjellemzőket, melyek felhasználhatók a hozzászólások hasznosságának becsléséhez. A negyedik szakaszban szűkítjük ezen jellemzők körét, hogy használható méretű adathalmazzal hajthassuk végre a felügyelt tanítást. Az ötödik szakaszban bemutatjuk a felügyelt tanulási eljárást, a hatodik szakaszban pedig ismertetjük a tanulás eredményeit. A tanulmány végén összefoglaljuk elemzéseinket, és további kutatási lehetőségeket vázolunk fel.

## 2 Irodalmi áttekintés

Jelen szakaszban áttekintést adunk az irodalomban fellelhető azon eredményekről, melyek a termékkritikák hasznosságát értékelik.

Pan és Zhang (2011) olyan jellemzőket kerestek kutatásuk során, melyek érdemben befolyásolják a termékkritikák olvasók által észlelt hasznosságát (perceived review helpfulness/usefulness). Az Amazon.com oldalon fellelhető hozzászólásokat elemezték használati és élvezeti terméktípusok esetén, ahol az észlelt hasznosság értékét a felhasználói visszajelzésekből származtatták. Kvantitatív elemzésük során kimutatták, hogy a pozitív tartalmú és a hosszabb termékvéleményeket hasznosabbnak tartják a felhasználók, de a kapcsolat erőssége eltér terméktípusonként. Felhívják a figyelmet arra, hogy – a racionális magyarázatokon túl – ennek oka lehet egyrészt, hogy egy olvasó általában már rendelkezik vásárlási szándékkal a kritikák olvasása közben, így egy pozitív tartalmú véleményt megerősítésként él meg, másrészt – a kognitív disszonancia csökkentése érdekében – a hosszabb kritika elolvasása miatti nagyobb erőfeszítést magasabb észlelt hasznossági szinttel ellensúlyoz. A szerzők kvalitatív elemzésük során a termékvélemény írójának innováció-elfogadási szintjét is számszerűsítették, majd kimutatták, hogy a magas szinten állók (véleményvezérek) hozzászólásait hasznosabbnak vélik az olvasók, de a túlságosan innovatívak már olyan extrém véleménnyel bírnak, ami csökkenti kritikájuk elfogadottságát.

Az észlelt hasznosság (helpfulness) környezetfüggő jellegét emeli ki Danescu-Niculescu-Mizil és sztsai (2009) eredménye, akik azt találták, hogy az egyes hozzászólások észlelt hasznosságát a termékről elérhető többi kritika száma és minősége is befolyásolja: kevés elérhető komment esetén felértékelődik azok információtartalma és így hasznossága is.

Kim és sztsai (2006) az Amazon.com weboldalról gyűjtött termékkritikák hasznosságát (helpfulness) azok szövegjellemzői segítségével modellezték. Kutatásuk célja – a változók közötti összefüggések feltárásán túl – az volt, hogy képessé váljanak új hozzászólások hasznosságának automatikus becslésére. A szerzők a termékkritikákra adott olvasói szavazatokból (hasznos vs. nem hasznos) számolt aránnyal jellemezték a tanítóhalmaz hozzászólásait. Elemzésükben MP3 lejátszókra és digitális kamerákra koncentráltak, a hasznosságot pedig a következő szövegjellemzőkkel próbálták leírni:

- strukturális jellemzők (a hozzászólásban szereplő tokenek, mondatok száma, mondatok átlagos hossza, kérdő és felkiáltó mondatok aránya, kiemelések és sortörések száma stb.),
- lexikális jellemzők (tokenek és bigramok tf-idf értéke stb.),
- szintaktikai jellemzők (főnevek, igék és melléknevek aránya stb.),
- szemantikai jellemzők (termékjellemzőkre vonatkozó objektív adatok, pozitív és negatív értelmű szavak száma stb.),
- metaadat jellemzők (termékminősítő csillagok száma, az átlagértéktől való eltérés nagysága stb.).

Az SVM regresszió keretében lineáris, polinomiális, valamint radiális bázisfüggvényt is kipróbáltak, melyek közül az utóbbi teljesített legjobban. Az elemzés eredménye azt mutatta, hogy leginkább a strukturális (hozzászólás hossza) és a metaadat jellemzők határozták meg a hasznosságot.

Az a megoldás, hogy a hozzászólások hasznosságára érkező olvasói szavazatok kerüljenek felhasználásra a gépi tanulás során több kritikát is kapott. Cao és sztsai (2011) például azt vizsgálták, hogy milyen tényezők hatására kapnak egyes termékkritikák sok, míg mások kevés szavazatot (helpfulness és unhelpfulness szavazatok összege) akár olyan esetekben is, amikor ezen kritikák „objektív" hasznossága megegyezik. Adatforrásként a CNET Download.com oldalát használták fel, ahol szoftvertermékekhez kapcsolódó bejegyzések találhatók.

Elemzésük során három jellemzőcsoportot vizsgáltak:

- alapjellemzők: kritika megjelenésének időpontja, a termék értékelése stb.,
- stilisztikai jellemzők: mondatok hossza, használt szavak halmaza stb.,
- szemantikai jellemzők: több szavazatot kaphat-e például a „bölcs befektetés ez a szoftver" kifejezés, mint az „ez egy jó szoftver".

Elemzésük azt mutatja, hogy a kapott szavazatok száma jelentős szórással rendelkezik, és a szemantikai jellemzők bírnak a legnagyobb befolyással erre, ami eltér a korábbi tanulmányok által ezen jellemzőknek tulajdonított szereptől. Emellett azt találták, hogy az extrém vélemények több szavazatot vonzanak, mint a vegyes vagy semleges hozzászólások.

Az Amazon.com oldalán található szavazatok alkalmazásának első jelentős kritikáját Liu és sztsai (2007) fogalmazták meg. Kutatásuk célja az alacsony minőségű (low-quality) termékkritikák megtalálása volt annak érdekében, hogy az adott termékre vonatkozó véleményösszegzés során figyelmen kívül hagyhassák azokat. Elsőként azt állapították meg, hogy az Amazon.com oldalon alkalmazott – olvasói szavazatokra építő – értékelő módszer három tényező miatt is torzított eredményt ad:

- Az olvasók jobban szeretnek pozitív szavazatot (helpful) adni, így bizonyos hozzászólások 100%-ban hasznos címkét kapnak, miközben csak egy rövid értékelést adnak a termékről. („imbalance vote" torzítás)
- Nagyszámú korábbi szavazat túlzott mértékben kelti a felhasználókban azt a képzetet, hogy a termékkritika „hiteles", függetlenül annak tényleges minőségétől. Ez tovább növeli a szavazatok számát, és csökkenti a szavazók objektivitását. („winner circle" torzítás)
- A termék piacra dobásának időpontjához közel megjelenő hozzászólások több szavazatot kapnak, mint a magasabb minőségű, de később megjelenő kritikák. („early bird" torzítás)

A fenti torzítások miatt a termékkritikák hasznosságának – a bináris SVM tanításához szükséges – címkézését manuálisan végezték el. Az automatikus értékeléshez specifikálták a hozzászólás-minőség (quality) mérésének sztenderd módját, melynek során három fő faktort azonosítottak. Az informativitást (informativeness) a mondatokra, szavakra és termékjellemzőkre vonatkozó mutatókkal mérték, mint például azok hossza, száma, előfordulásuk gyakorisága a szövegben és a címben. Az olvashatóságot (readability) a bekezdések számával, a bekezdések átlagos hosszával, valamint a szövegelválasztó jelek számával jellemezték. A szubjektivitást (subjectiveness) a pozitív, illetve negatív mondatok arányával, valamint a szubjektív (felhasználói véleményt tartalmazó) mondatok számával jellemezték. Eredményként azt kapták, hogy leginkább az informativitás jellemzők alapján lehet következtetni a kritikák minőségére, és a szubjektivitás faktor csak minimális mértékben járul hozzá a becslés pontosságának javulásához.

Chen és Tseng (2011) egy olyan módszert dolgoztak ki, mellyel a termékkritikák információminőségét (quality of information) értékelik. A hozzászólások jellemzőinek strukturálására egy – más területeken már sikeresen alkalmazott – információminőség (information quality – IQ) keretrendszert használtak fel, majd két különböző, többosztályos SVM-mel (lineáris kernel mellett) értékelték azokat. A több weboldalról származó hozzászólásokat manuálisan sorolták a következő csoportokba: magas minőségű (high-quality), közepes minőségű (medium-quality), alacsony minőségű (low-quality), másolat (duplicate) és spam. Az IQ keretrendszer hierarchikus felépítésű, kilenc dimenzió mentén 51 mutatót alkalmaz a termékkritikák jellemzésére. A dimenziók az alábbiak:

- hihetőség (believability),
- véleménymentesség (objectivity – szubjektivitás ellentéte),
- elismertség (reputation – szerző elismertsége),
- relevancia (relevancy),
- időszerűség (timeliness),
- teljesség (completeness – termék teljes körű bemutatása),
- információmennyiség (appropriate amount of information),
- érthetőség (ease of understanding),
- tömörség (concise representation).

A 10-szeres keresztvalidáció során azt tapasztalták, hogy a véleménymentesség és az információmennyiség jellemzők bírnak a legnagyobb magyarázó erővel a hozzászólások osztályozása során.

Wu és sztsai (2010) olyan hozzászólás-jellemzőket kerestek, melyek segítségével kiválaszthatók a kétes termékkritikák (suspicious reviews). Ezek olyan bejegyzések, melyeket nem valós felhasználók, hanem például forgalmazók vagy azok versenytársai írtak azzal a céllal, hogy a valóságosnál jobb vagy éppen rosszabb fényben tüntessék fel a termékeket. A kutatás során a

szerzők a TripAdvisor oldalán található, szálláshelyekre vonatkozó kritikákat elemezték, és azt találták, hogy egy bejegyzés kétes jellegét az jelzi leginkább, ha az pozitív és rögtön egy negatív után következik.

A fent bemutatott megoldások alapvetően a termékkritikák szövegjellemzőit használták fel arra, hogy felügyelt tanulás segítségével modellezzék azok hasznosságát. Tsur és Rappoport (2009) egy merőben új technikával próbálták meg értékelni az online értékesített könyvekre vonatkozó hozzászólások hasznosságát (helpfulness), ahol akkor tekintettek egy bejegyzést hasznosnak, ha az támogatta az olvasó vásárlással kapcsolatos döntését. A probléma kezelésére fejlesztették ki a RevRank algoritmust. Elsőként a vizsgált termékről írt kritikákban azonosították a legfontosabb kifejezéseket, vagyis a nem túl gyakori, de az adott termékre vonatkozóan magas információtartalommal bíró szavakat, szóösszetételeket, majd ezek összegzéséből létrehoztak egy virtuális mag hozzászólást (virtual core review). A bejegyzések hasznosságának értékelése az ,,optimálisnak" tekinthető mag hozzászóláshoz mért hasonlóság alapján történik. Az eljárás nagy előnye az irodalomban található többi megoldáshoz képest, hogy teljes mértékben felügyelet, azaz manuális címkézés nélkül tud működni.

Lu és sztsai (2010) egy számottevően új elemet vontak be a hasznosság értékelésének módszertanába. A szövegjellemzők mellett a termékkritikák szerzőinek társas(ági) hálózatát (social network), mint környezeti információt is figyelembe vették a hozzászólások minőségének (quality) becslésekor. A következő két feltételezéssel éltek:

- A termékkritika minősége függ a szerző minőségétől.
- Egy szerző minősége függ a hálózatban vele kapcsolatban álló szerzők minőségétől, ugyanis a hálózati kapcsolatok egyfajta bizalmat, barátságot fejeznek ki.

Adatforrásként a Ciao UK1 oldal bejegyzéseit használták fel, ahol az olvasóknak lehetősége van a kritikák értékelésére, és arra is, hogy a számukra tetsző szerzőket hozzáadják saját bizalmi körükhöz (circle of trust). A csak szövegjellemzőket tartalmazó regressziós függvényt mint alapmodellt kiegészítették a társasági hálózatot figyelembe vevő változókkal, melynek eredményeként szignifikáns módon javult a modell becslési pontossága.

Az irodalmi összefoglaló végén két olyan cikket említünk meg, melyek nem termékkritikák, hanem más online dokumentumok kapcsán végeznek a fentiekhez hasonló elemzéseket.

Siersdorfer és sztsai (2010) dolgoztak ki elsőként automatikus osztályozó eljárást YouTube hozzászólások közösség általi elfogadására vonatkozóan (accepted or not accepted by the community). 67 000 YouTube videóra adott több mint hatmillió komment kapcsán vizsgálták, hogy van-e összefüggés azok elfogadottsága, a megjelenő tokenek gyakorisága, az olvasói szavazatok száma és a videó tartalmának kategóriája (zene, politika stb.) között.

Pon és sztsai (2011) célja egy olyan rendszer (iScore) felépítése volt, ami képes kiszűrni az olvasó számára érdektelen (uninteresting) híreket az interneten. Bár a termékekre vonatkozó kommentek és a hírek nem kezelhetők

azonos módszerrel, de az érdekesség-érdektelenség (interesting-uninteresting) koncepciójának bevezetése felhasználható a termékkritikák értékelése esetén is. A szerzők három relevancia kategóriát azonosítottak, melyek alapján érdekesnek (interesting) tekinthető egy cikk egy adott felhasználó számára. A kognitív relevancia (cognitive relevance) akkor teljesül, ha a hír informatív, újszerű és magas minőségű. A szituációs relevancia (situational relevance) feltétele, hogy támogassa a döntéshozatalt és csökkentse a bizonytalanságot, míg a motivációs relevanciához (motivational relevance) az szükséges, hogy illeszkedjen a felhasználó céljaihoz, szándékához. A fentiekből látszik, hogy az érdekesség fogalma sokkal komplexebb, mint egyszerűen a felhasználó érdeklődési területének való megfelelés.

Ahogy az irodalmi példákból is látszik, a „termékkritikák hasznossága" erősen szubjektív kategória. Függ a felhasználó céljától (például vásárlási döntés támogatása, általános információgyűjtés), és egyéb preferenciáitól, amik ráadásul változhatnak az idő múlásával (például változik a felhasználó tudásszintje a kritikák olvasása során). Elemzésünk során ezért egyetlen felhasználót kértünk meg arra, hogy értékelje a hozzászólásokat ötfokozatú Likert-skálán („Mennyire találtad hasznosnak a hozzászólást?"). Nem törekedtünk tehát a válaszadó céljainak, motivációinak előzetes feltárására, a hasznossági értékek az aktuális, egyéni preferenciákat tükrözik. Megjegyezzük, hogy a válaszadó (és általában a hasonló kérdésekre választ adó személy) nem feltétlenül tudja szétválasztani az egyéni hasznosságra vonatkozó és a hozzászólás információtartalmára (hasznosságpotenciál) vonatkozó értékelését.

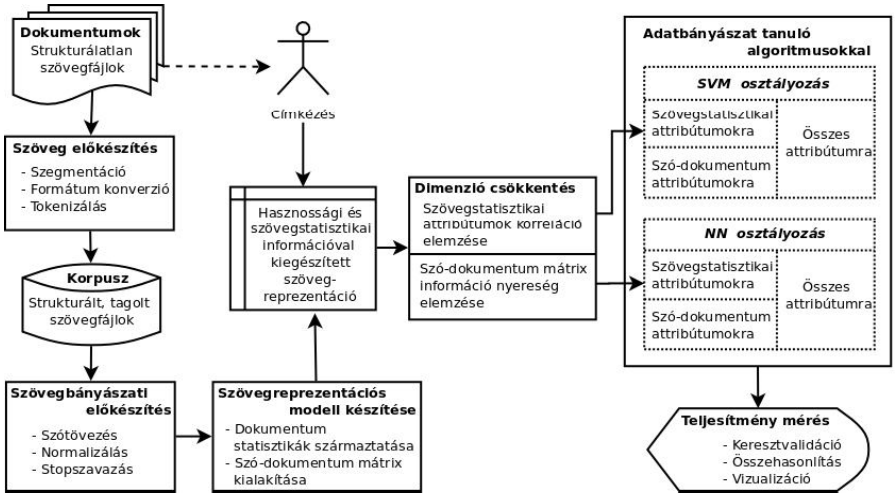
# 3 A kommentek előfeldolgozása és reprezentációs modelljei

Vizsgálataink során 1000 darab, mobiltelefonokkal kapcsolatos magyar nyelvű termékkritikát gyűjtöttünk össze különböző weboldalakról. A hozzászólásokat egy kísérleti alany manuális úton címkézte fel oly módon, hogy a következő egyszerű kérdésre kellett választ adnia: „Mennyire találtad hasznosnak a hozzászólást?". Az ötfokozatú Likert-skálán kapott válaszokat ezután bináris mutatóvá konvertáltuk: a 4-es és 5-ös értékkel bíró hozzászólásokat „hasznos"-nak, az 1 és 3 közötti értékkel rendelkezőket pedig „nem hasznos"-nak tekintettük. Az alábbiakban – illusztrációként – bemutatunk egy hasznosnak és egy nem hasznos tekintett kommentet:

Egy hasznos hozzászólás: *„Igazi üzleti telefon. Hibátlan. Nagy adatmennyiségnél (500 sms-nél picit belassul) 3 évet használtam gond nélkül, most fater használja, 3 elő-és hátlap csere volt, 1 akksi eddig, és bírja. Jó akksi, memória kapacitás. Az árát még mindig tartja. (Nem az újkori árhoz képest, hanem amit kb. 3-4 éve ért el) egyszer pancsizott, az öreg beesett a stégről a vízbe. Szétkaptuk, kiszáradt, azóta eltelt 2 év. Működik. Előnyök: Ergonómia, tudás, memória. Hátrányok: Üzleti telefonként ugyan nem volt, de talán a kamera picit gyenge, amit soha nem használtam."*

Egy nem hasznos hozzászólás: *„Én most fogok kapni ilyet narancsban. Nagyon várom, szerintem nagyon jó. Semmi bajom vele, csak egy picit tucatteló”*

A tanulási folyamat célja, hogy olyan – minden komment esetén számszerűsíthető – szövegjellemzőket találjunk, melyek kombinációja magyarázza a manuális címkézéssel nyert hasznossági értékeket. Ezen szövegjellemzők aztán felhasználhatók az előre nem címkézett kommentek hasznosságának automatikus megállapításához. A fenti cél elérésére kidolgozott eljárás lépéseit mutatja be a *2. ábra.*




*2. ábra.* Kommentek hasznosságának automatikus megállapítására kidolgozott eljárás

A folyamat első lépéseként a rendelkezésre álló 1000 hozzászólás szövegének előfeldolgozását kellett elvégezni, hogy számszerűsítésre kerülhessenek a potenciális magyarázó mutatók. Mivel a termékkritikák gyűjtése során a különböző metaadatok (például szerző, dátum) csak részlegesen álltak rendelkezésre, valamint nem minden weblap ad lehetőség a kommentek formázására – és ha van is rá mód, akkor a gyorsaság miatt a felhasználók ritkán élnek ezzel a lehetőséggel –, ezért mind a metaadatokat, mind a formázásban rejlő információkat figyelmen kívül hagytuk. A hasznosság manuális megállapításához és gépi becsléséhez mindössze a hozzászólások formázatlan szövegét használtuk fel. Ezen forrásból az alábbi szövegváltozatokat készítettük el, ahol a feldolgozottsági állapotok egymás után következnek és egymásra épülnek:

1. Nyers szöveg (az eredeti tartalom, változatlan HTML formátumban)
2. Folyó szöveg (a HTML tartalom egyszerű szöveges TXT formátumra alakítása)
3. Unikód szöveg (a folyó szöveg dokumentumainak egységes karakterkódolása)

4. Tokenizált szöveg (az unikód szöveg szó és írásjel blokkokra tagolása)

5. Szótövezett szöveg (a tokenizált szöveg szótövezett változata)

6. Normalizált szöveg (a szótövezett szöveg kisbetűsre alakítása és egyéb egységesítése)

7. Stopszavazott szöveg (a normalizált szöveg gyakori töltelék szavainak törlésével előállított szöveg)

Ezen transzformációk közül a karakterkódolás, a szótövezés és a stopszavazás csak nyelvfüggő módon és eszközökkel valósítható meg. Az utóbbi kettő végeredményét ráadásul erősen befolyásolja a tömörítési fok megválasztása, azaz az alkalmazott szótőkeresés mélysége és a stopszó lista nagysága. Ezek a módszerek kizárólag a nyelvtani szabályok ismeretére építve nem is mindig egyértelműek, például az ,,almát" kifejezés szótöve a szövegkörnyezettől függően ,,alma" de akár ,,alom" is lehet. Az előfeldolgozás eredményeként előfordult, hogy bizonyos kommentek túlzottan lerövidültek, kiürültek. Az így keletkező ,,üres" hozzászólások (az adathalmaz 10,29%-a) általában semmiféle hasznossággal nem rendelkeznek, így azokat kizártuk a további vizsgálatokból, hiszen csak ,,látszólag" javítanák az osztályozó módszerek pontosságát.

A hozzászólások jellemzéséhez alkalmazott mutatók értéke függ attól, hogy a szövegek mely feldolgozottsági állapotát használva számszerűsítjük azokat. Általános szabályként a mutatók értékét mindig abból a legmagasabb szinten feldolgozott szövegváltozatból kalkuláltuk, ahol az még éppen értelmezhető volt. Például a kis és nagy betűk számát a normalizált változat előtti szótövezett változat alapján, míg az írásjelek számát az Unikód szövegváltozatból állapítottuk meg.

Az irodalomkutatás alapján összegyűjtöttük, jelöléstechnikailag egységesítettük és csoportosítottuk azon szövegjellemzőket, melyeket idegen nyelvű szövegek esetében már felhasználtak a hasznosság gépi megállapításához. A kvantitatív mérőszámok jellegzetessége illetve forrása szerint az alábbi kategóriákat különítettük el.

1. Strukturális jellemzők: egy adott dokumentumhoz tartozó, annak értelmezése nélkül származtatható statisztikai mutatók (például NWRD – a szavak száma, DWRD – a különböző szavak száma).

2. Lexikális jellemzők: a dokumentumokat egy egységes gyűjtemény (korpusz) részének tekintve, a szövegelemek dokumentumok közötti megoszlásának mérőszámai (például UTDF – a szógyakoriságokat tartalmazó szó-dokumentum mátrix).

3. Szintaktikai jellemzők: a szövegek helyességének, a különféle nyelvtani szabályoknak való megfelelés és nyelvtani osztályokba való besorolások mutatói (például NSMD – mosolykódok (smiley-k) száma a dokumentumban).

4. Szemantikai jellemzők: a szavak és mondatok értelmezését is felhasználó mutatók. Jellegüknél fogva ezek a mutatók erősen függnek az olvasó szubjektumától is (például: LOPD – a dokumentum pozitív, negatív vagy semleges tájolása).

| | | Hasznos komment | Nem hasznos komment |
|---|---|---|---|
| *Strukturális jellemzők* | | | |
| Mondat jellemzők | | | |
| Mondatok száma a dokumentumban | nsnd | 3 | 2 |
| Átlagos mondathossz szavakban mérve | aslw | 27,33 | 8 |
| Átlagos mondathossz betűkben mérve | aslc | 168 | 50 |
| Szó jellemzők | | | |
| Szavak száma a dokumentumban (írásjel tokenek nélkül) | nwrd | 82 | 16 |
| Szavak számának negyedik gyöke (n4wrd=nwrd$^{1/4}$) | n4wrd | 3,01 | 2 |
| Különböző szavak száma a dokumentumban | dwrd | 63 | 16 |
| A szöveg lexikális sűrűsége (dwrd/nwrd) | lxdn | 0,77 | 1 |
| Nagybetűvel kezdődő szavak száma a dokumentumban | nfcwd | 9 | 1 |
| A csupa nagybetűs szavak száma a dokumentumban | nacwd | 5 | 0 |
| Komplex (3 vagy több szótagú) szavak száma | ncwd | 35 | 5 |
| Komplex szavak részaránya (ncwd/nwrd) | rcwd | 0,43 | 0,31 |
| Átlagos szóhossz karakterekben mérve | awlc | 5,95 | 5,94 |
| Átlagos szóhossz szótagokban mérve (nsyd/nwrd) | awly | 2,32 | 2,19 |
| Szótag jellemzők | | | |
| Szótagok száma a dokumentumban | nsyd | 190 | 35 |
| Betű jellemzők | | | |
| Karakterek száma a dokumentumban | nchd | 504 | 100 |
| Nagybetűk száma a dokumentumban | ncchd | 21 | 1 |
| Nagybetűk aránya (ncchd/nchd) | rcchd | 0,04 | 0,01 |
| Kisbetűk száma a dokumentumban | nlchd | 462 | 93 |
| Kisbetűk aránya (nlchd/nchd) | rlchd | 0,96 | 0,99 |
| Nagybetű-kisbetű arány (ncchd/nlchd) | rclchd | 0,04 | 0,01 |
| *Lexikális jellemzők* | | | |
| Unigram jellemzők | | | |
| Szógyakoriságokat tartalmazó szó-dokumentum mátrix | utdf | | |
| *Szintaktikai jellemzők* | | | |
| Nyelvtani jellemzők | | | |
| Alfabetikus karakterek száma (ncchd+nlchd) | nachd | 483 | 94 |
| Felkiáltójelek száma a dokumentumban | nemd | 0 | 0 |
| Kérdőjelek száma a dokumentumban | nqmd | 0 | 0 |
| Idézőjelek száma a dokumentumban | nqqd | 0 | 0 |
| Írásjelek száma a dokumentumban | npchd | 17 | 6 |
| Számjegyek száma a dokumentumban | nnchd | 4 | 0 |
| Mondatonkénti átlagos numerikus információtartalom (rnchd=nnchd/nsnd) | rnchd | 1,33 | 0 |
| Mosolykódok (smiley-k) száma a dokumentumban | nsmd | 3 | 0 |
| Helyesírásilag elfogadható szavak száma | nspwd | 91 | 16 |
| Helyesírásilag elfogadható szavak aránya (nspwd/nwrd) | rspwd | 1,11 | 1 |
| *Szemantikai jellemzők* | | | |
| Tájolás | | | |
| A dokumentum pozitív, negatív vagy semleges tájolása | lopd | 1 | 4 |

*1. táblázat.* A hasznosság megállapítására felhasználható attribútumok[2]

[2]Fontos megemlíteni, hogy az UTDF indikátor – lévén egy mátrix – valójában nem egy, hanem az oszlopszámnak (a dokumentumokban előforduló különböző szavak száma) megfelelő számú mutatót jelöl. Hasonló mátrix készíthető a szópárok (BTDF) illetve szó n-esek (NTDF) dokumentumbeli előfordulásáról is. A továbbiakban a szógyakorisági UTDF mutatókra a szakirodalomban jobban elterjedt szó-dokumentum mátrix (TDM) jelölést használjuk. Mivel a TDM mérete és jellege alapján is eltér a többi attribútumtól, ezért a későbbi elemzések során szeparáltan vizsgáljuk annak osztályozó erejét.

Bizonyos mutatók előállítása történhet egyszerű leszámlálással, míg mások képzése komplexebb módszereket kíván meg, melyek lehetnek például

- skálatorzítás: az alapértékeket valamilyen függvény szerint transzformáljuk azért, hogy az alkalmazandó módszerekben a nagyságrendi eltérések ne okozzanak problémákat,
- aggregálás: egy mutatóhoz tartozó részsokaság numerikus jellemzése (például átlag, szórás),
- relativizálás: a mutató értékeit egy konstanshoz vagy másik mutatóhoz viszonyítjuk.

A szövegjellemzők mindegyike képezhető bekezdésekre vagy mondatokra is, de jól strukturált dokumentumok esetén akár más részegységek (például cím, bevezetés, törzs, összefoglaló) bontásában is. A lehetséges mutatók széles skálájából az *1. táblázatban* felsorolt indikátorok bizonyultak meghatározhatónak a rendelkezésre álló dokumentumok jellemzésére.

A számítások során, ha egy relativizált mutató esetében a nevező nulla volt, úgy a származtatott érték nem került definiálásra. A karakterek leszámlálásakor a magyar nyelv kettős betűit (digráfok) két betűnek, de csak egy hangnak tekintettük. Amikor a szóismétlések összeszámlálására került sor, akkor azok kis- és nagybetűs írásmódjait nem különböztettük meg egymástól. Néhány speciális írásjel (például az aláhúzás vagy kötőjel) esetén azokat a nem szóalkotó betűk közé soroltuk.

A fent bemutatott és számszerűsített indikátorok körét több okból is érdemes szűkíteni. Először is, a dokumentumminta elemszámának jelentősen meg kell haladnia az indikátorok számát, ellenkező esetben ugyanis elkerülhetetlen az adatsorok egyedi felismerését eredményező túltanulás jelensége. Másodszor, a kevesebb indikátor előállítása jelentősen csökkenti az osztályozó algoritmusok futásidejét. Harmadszor, az adatok mögött zajló folyamatok felismerése, a lényegi jellemzők megragadása csökkenti az egyéb forrásból származó zajok hatását. A fentiek értelmében egyrészt elfogadható méretűre redukáltuk a TDM-et, másrészt megvizsgáltuk a többi attribútum (ezekre a későbbiekben statisztikai szövegjellemzőkként (STAT) hivatkozunk) között fennálló redundanciát, és meghatároztuk a tanuló algoritmusok futtatása során felhasználandó szűkített mutatóhalmazt. Elsőként ez utóbbi indikátorszűkítési folyamatot mutatjuk be.

# 4 A statisztikai szövegjellemzők körének szűkítése

Guyon és Elisseeff (2003) munkája alapján ismert, hogy tökéletesen korrelált attribútumok esetén azok bármelyike ugyanazzal az osztályozó erővel bír, mint maga a teljesen korrelált halmaz, így a korrelációs osztályokból elegendő egyetlen tetszőleges attribútumot meghagyni a modellezés során. Nem tökéletes korreláció esetén azonban létezhetnek olyan mutatók, melyek

önmagukban nem hordoznak információt, de más attribútumokkal együtt jelentősen megnő az osztályozó képességük. Ennek megfelelően nagyon erős (anti)korreláció esetén már nem állítható biztonsággal, hogy bármelyik indikátor helyettesíthető lenne a többi segítségével, de a gyakorlatban ez mégis elfogadott kompromisszum. A fentiek alapján a dokumentumonként számszerűsített mutatók értékeire kiszámítottuk a Pearson-féle korrelációs mátrixot (*3. ábra*), melynek segítségével meghatároztuk az együttmozgó indikátorok halmazait (ld. később). A mátrix sorait és oszlopait úgy rendeztük, hogy az egymással nagyon erős (anti)korrelációt mutató indikátorok egymás mellé kerüljenek, a főátló mentén pedig bekereteztük az egy halmazba kerülő mutatók értékeit.

| | lopd | nsnd | npchd | nwrd | dwrd | nsyd | nchd | nlchd | nachd | ncwd | nspwd | n4wrd | lxdn | rcwd | awly | aslw | aslc | rspwd | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| lopd | 1 | -0,13 | -0,15 | -0,14 | -0,15 | -0,14 | -0,14 | -0,13 | -0,14 | -0,14 | -0,14 | -0,16 | 0,1 | -0,06 | 0 | -0,05 | -0,05 | -0,11 | |
| nsnd | -0,13 | 1 | 0,82 | 0,85 | 0,84 | 0,85 | 0,85 | 0,84 | 0,84 | 0,81 | 0,87 | 0,75 | -0,56 | 0,11 | 0,11 | -0,13 | -0,1 | 0,33 | |
| npchd | -0,15 | 0,82 | 1 | 0,88 | 0,88 | 0,87 | 0,89 | 0,86 | 0,87 | 0,84 | 0,9 | 0,79 | -0,57 | 0,12 | 0,11 | 0,09 | 0,14 | 0,39 | |
| nwrd | -0,14 | 0,85 | 0,88 | 1 | 0,99 | 0,99 | 0,99 | 0,98 | 0,99 | 0,95 | 0,99 | 0,91 | -0,71 | 0,11 | 0,1 | 0,26 | 0,27 | 0,31 | |
| dwrd | -0,15 | 0,84 | 0,88 | 0,99 | 1 | 0,99 | 0,99 | 0,98 | 0,98 | 0,95 | 0,99 | 0,93 | -0,68 | 0,12 | 0,11 | 0,27 | 0,28 | 0,33 | |
| nsyd | -0,14 | 0,85 | 0,87 | 0,99 | 0,99 | 1 | 1 | 0,99 | 1 | 0,98 | 0,99 | 0,89 | -0,68 | 0,18 | 0,18 | 0,24 | 0,27 | 0,32 | |
| nchd | -0,14 | 0,85 | 0,89 | 0,99 | 0,99 | 1 | 1 | 0,99 | 1 | 0,97 | 0,98 | 0,89 | -0,68 | 0,16 | 0,17 | 0,23 | 0,29 | 0,32 | |
| nlchd | -0,13 | 0,84 | 0,86 | 0,98 | 0,98 | 0,99 | 0,99 | 1 | 0,99 | 0,97 | 0,98 | 0,89 | -0,68 | 0,17 | 0,17 | 0,24 | 0,29 | 0,31 | |
| nachd | -0,14 | 0,84 | 0,87 | 0,99 | 0,98 | 1 | 1 | 0,99 | 1 | 0,97 | 0,98 | 0,89 | -0,68 | 0,17 | 0,17 | 0,24 | 0,29 | 0,31 | |
| ncwd | -0,14 | 0,81 | 0,84 | 0,95 | 0,95 | 0,98 | 0,97 | 0,97 | 0,97 | 1 | 0,95 | 0,85 | -0,64 | 0,28 | 0,26 | 0,21 | 0,26 | 0,31 | |
| nspwd | -0,14 | 0,87 | 0,9 | 0,99 | 0,99 | 0,99 | 0,98 | 0,98 | 0,98 | 0,95 | 1 | 0,9 | -0,7 | 0,12 | 0,11 | 0,22 | 0,24 | 0,36 | |
| n4wrd | -0,16 | 0,75 | 0,79 | 0,91 | 0,93 | 0,89 | 0,89 | 0,89 | 0,89 | 0,85 | 0,9 | 1 | -0,77 | 0,11 | 0,1 | 0,38 | 0,39 | 0,46 | |
| lxdn | 0,1 | -0,56 | -0,57 | -0,71 | -0,68 | -0,68 | -0,68 | -0,68 | -0,68 | -0,64 | -0,7 | -0,77 | 1 | -0,02 | 0,02 | -0,33 | -0,31 | -0,29 | |
| rcwd | -0,06 | 0,11 | 0,12 | 0,11 | 0,12 | 0,18 | 0,16 | 0,17 | 0,17 | 0,28 | 0,12 | 0,11 | -0,02 | 1 | 0,81 | -0,05 | 0,09 | 0,19 | |
| awly | 0 | 0,11 | 0,11 | 0,1 | 0,11 | 0,18 | 0,17 | 0,17 | 0,17 | 0,26 | 0,11 | 0,1 | 0,02 | 0,81 | 1 | -0,06 | 0,1 | 0,19 | |
| aslw | -0,05 | -0,13 | 0,09 | 0,26 | 0,27 | 0,24 | 0,23 | 0,24 | 0,24 | 0,21 | 0,22 | 0,38 | -0,33 | -0,05 | -0,06 | 1 | 0,94 | 0,05 | |
| aslc | -0,05 | -0,1 | 0,14 | 0,27 | 0,28 | 0,27 | 0,29 | 0,29 | 0,29 | 0,26 | 0,24 | 0,39 | -0,31 | 0,09 | 0,1 | 0,94 | 1 | 0,08 | |
| rspwd | -0,11 | 0,33 | 0,39 | 0,31 | 0,33 | 0,32 | 0,32 | 0,31 | 0,31 | 0,31 | 0,36 | 0,46 | -0,29 | 0,19 | 0,19 | 0,05 | 0,08 | 1 | |
| ... | | | | | | | | | | | | | | | | | | | |

*3. ábra.* Attribútumok Pearson-féle korrelációs mátrixa

Az így nyert halmazokból már csak egy-egy mutatót érdemes a további elemzésekhez meghagyni. Ehhez többféle attribútum-sorrendező vagy attribútum kiválasztó módszert használtunk fel. Az előbbiek esetén a korrelációs halmazonként legmagasabb pontszámmal rendelkező mutató nyert egy szavazatot, míg az utóbbiak során a korrelációs halmazonként kiválasztott legelső elem. Az attribútum szűkítő módszerek (döntési bizottság tagjai) által leadott szavazatok összesítése után a korrelációs halmazonként legtöbb szavazatot nyert mutató került kiválasztásra.

A kiválasztó bizottság 11 tagját a Weka[3] nyílt forrású adatbányászati alkalmazás alábbi módszerei alkották (az első módszer attribútum kiválasztó, a többi attribútum sorrendező):

[3] www.cs.waikato.ac.nz/ml/weka

| Módszer neve | Weka elnevezés |
|---|---|
| Correlation-based Feature Subset Evaluation | CfsSubsetEval |
| Consistency Attribute Subset Evaluation | ConsistencySubsetEval |
| Latent Semantic Ananlysis | LatentSemanticAnalysis* |
| Chi-squared Attribute Evaluation | ChiSquaredAttributeEval |
| Filteres Attribute Evaluation | FilteredAttributeEval |
| Gain Ration Attribute Evaluation | GainRatioAttributeEval |
| Information Gain Attribute Evaluation | InfoGainAttributeEval |
| OneR Attribute Evaluation | OneRAttributeEval |
| Relief f Attribute Evaluation | ReliefFAttributeEval |
| Support Vector Machine Attribute Evaluation | SVMAttributeEval |
| Symmetrical Uncertainty Attribute Evaluation | SymmetricalUncertAttributeEval |

*2. táblázat.* Attribútum-sorrendező és kiválasztó módszerek a döntési bizottságban

A mutatónkénti szavazatok számát tartalmazza a *3. táblázat*, ahol csillaggal jelöltük a korrelációs halmazonként legtöbb szavazatot szerzett statisztikai szövegjellemzőket.

| Halmaz | Indikátor | Szavazat | Halmaz | Indikátor | Szavazat |
|---|---|---|---|---|---|
| 1 | nsnd | 1 | 4 | rspwd* | 11 |
| 1 | npchd | 1 | 5 | nqqd* | 10 |
| 1 | nwrd | 0 | 6 | nfcwd* | 7 |
| 1 | n4wrd | 2 | 6 | nacwd | 2 |
| 1 | dwrd | 1 | 6 | ncchd | 2 |
| 1 | nsyd | 0 | 7 | nsmd* | 10 |
| 1 | nchd* | 3 | 8 | rnchd* | 9 |
| 1 | nlchd | 0 | 9 | rcchd | 3 |
| 1 | nachd | 0 | 9 | rlchd* | 6 |
| 1 | ncwd | 2 | 10 | rclchd* | 9 |
| 1 | nspwd | 1 | 11 | awlc* | 9 |
| 1 | lxdn | 0 | 12 | nemd* | 9 |
| 2 | rcwd* | 6 | 13 | nqmd* | 9 |
| 2 | awly | 5 | 14 | nnchd* | 10 |
| 3 | aslw | 2 | 15 | lopd* | 9 |
| 3 | aslc* | 9 | | | |

*3. táblázat.* Az attribútumok által nyert szavazatok száma

A szavazás kapcsán érdemes megjegyezni, hogy annak eredményei alátámasztják Yang és Pedersen (1997) eredményeit, miszerint általában már az információnyereség (IG) és a khi-négyzet (CHI) alapú attribútum sorrendező módszerek egyedüli alkalmazása is elegendően helyes kiválasztást ad. Az is bíztató, hogy a későbbiekben az osztályozáshoz használandó SVM-re épülő kiválasztási módszer sem adott a közös döntéstől jelentősen eltérő eredményt, azaz az elhagyhatónak ítélt indikátorok ezen módszer szerint is kevés addicionális információt hordoznak. A továbbiakban a TDM redukálásának folyamatát mutatjuk be.

# 5 TDM redukálása a vektortér-modellhez

A TDM alkalmazásának hátterében az a feltevés húzódik meg, hogy a kommentek hasznosságának megítélése során jelentős szerep juthat bizonyos in-

dikátor értékű szavaknak, amelyek a mondatban az alany, a tárgy vagy az állítmány szerepét töltik be. (Megjegyezzük, hogy az elemzések során a szavak mondatban betöltött szerepét nem vesszük figyelembe, mert e nélkül is kellő információt kapunk a probléma hatékony megoldásához.) A szavak attribútumként való figyelembe vételével a vektortér-modellben használt szó-dokumentum mátrixhoz jutunk (TDM), amely azonban a megfigyelésekhez képest aránytalanul sok dimenzióval rendelkezik. Ez a tény önmagában kedvezően hat a hasznos és haszontalan dokumentumok szeparálhatóságára, azonban fölösleges számolási kapacitásokat emészt fel, ezért kívánatos a mátrix méretének csökkentése. Ebben a szakaszban bemutatjuk, hogy milyen módszerrel csökkentettük a problématér dimenziójának számát, és hogyan állt elő a végső adatbázis, amelyen a szövegosztályozási algoritmusokat futtattuk.

A 4. táblázat mutatja be a TDM szerkezetét, melynek sorai képezik a dokumentumok reprezentációit a vektortér-modellben. Az oszlopok – a dokumentumvektorok dimenziói – pedig a korpuszban előforduló szavak, amelyeket nyelvtechnikai módszerekkel előzetesen redukáltunk, így az azonos szótövű szavakhoz azonos dimenzió tartozik (a szótővel azonosítjuk az oszlopokat). A vektorok koordinátái az 1. táblázatban látható reprezentációban megmutatják, hogy hányszor fordult elő az adott szótőnek valamilyen változata az adott dokumentumban.

| Dokumentum | fog | kép | kér | mer | optika | szerinte | ... |
|---|---|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... | ... | |
| Egy hasznos hozzászólás | 0 | 3 | 1 | 1 | 3 | 1 | ... |
| Egy nem hasznos hozzászólás | 1 | 0 | 1 | 0 | 0 | 1 | ... |
| ... | ... | ... | ... | ... | ... | ... | |

*4. táblázat.* Minta TDM mátrix a példa hozzászólások alapján

Az eredeti adatbázisban nyelvtechnikai dimenziócsökkentés (szótövezés, stopszavazás) után kb. 6500 kifejezés szerepelt. A szótár méretének további redukálását matematikai-statisztikai módszerekkel történő dimenziócsökkentés révén értük el. Jellemzőkinyerő módszereket (például látens szemantikus analízis – LSA) az elemzés során nem alkalmaztunk, mert a kinyert szintetikus jellemzők értelmezéséhez jelentősen el kellett volna vonatkoztatni az eredeti szavak jelentésétől. A jellemzőkiválasztó módszerek közül négyet vizsgáltunk meg, melyek

- a gyűjteménytámogatottság (collection frequency),
- a kölcsönös információ (MI/PMI, (pointwise) mutual information),
- az információnyereség (IG, information gain), és
- a khi-négyzet

mutatók. Ezen mutatók segítségével a szótár szavaihoz valós számokat rendelünk, melyekkel a szavak korpuszon belüli fontosságát mérjük. Az így becsült fontosság ismeretében hozhatunk döntést arról, hogy mely szavakat tartsuk meg a további elemzésekhez. Az alábbiakban a mutatók jelentéséről adunk részletesebb leírást.

*Alkalmazott jelölések*

$c_j$ $j$ kategória, ahol $j \in \{ h, n \}$, $h$ a hasznos és $n$ a nem hasznos kategória azonosítója

$t_k$ a $k$-adik szó tartalmazása

$\overline{t}_k$ a $k$-adik szó komplementere, azaz a $k$-adik szó nem tartalmazása

$n(c_j)$ a $j$ kategóriába tartozó dokumentumok száma

$n(t_k)$ a $k$-adik szót tartalmazó dokumentumok száma

$n(t_k, c_j)$ a $k$-adik szót tartalmazó és a $j$ kategóriába tartozó dokumentumok száma

$N$ a dokumentumok száma a korpuszon belül

$P(c_j)$ egy dokumentum $j$ kategóriába esésének valószínűsége

$P(t_k)$ a $k$-adik szó felbukkanásának valószínűsége, a dokumentumgyakorisággal ($d_f = n(t_k)/N$) becsülhető

$P(t_k, c_j)$ annak valószínűsége, hogy egy dokumentum tartalmazza a $k$-adik szót és $j$ kategóriájú

A gyűjteménytámogatottság (collection frequency) lényegében a szó adott korpuszon belüli előfordulásainak számát jelenti.

A kölcsönös információ ((pointwise) mutual information) azt méri, hogy mennyi az adott szó adott kategóriában („hasznos” illetve „nem hasznos”) való előfordulásának tényleges és függetlenség esetén várható információtartalmának különbsége:

$$MI_k = \log \frac{P(t_k, c_h)}{P(t_k)P(c_h)}$$

Az információnyereség (information gain) ezzel szemben nem csak ezt a különbséget veszi figyelembe: a szónak, a szó hiányának, a kategóriának, valamint a kategória komplementerének Descartes-szorzataként előálló négy halmazra számolt különbség várható értékét adja meg:

$$IG_k = \sum_{j \in \{h,n\}} \sum_{t \in \{t_k, \overline{t}_k\}} P(t, c_j) \log \frac{P(t, c_j)}{P(t)P(c_j)}$$

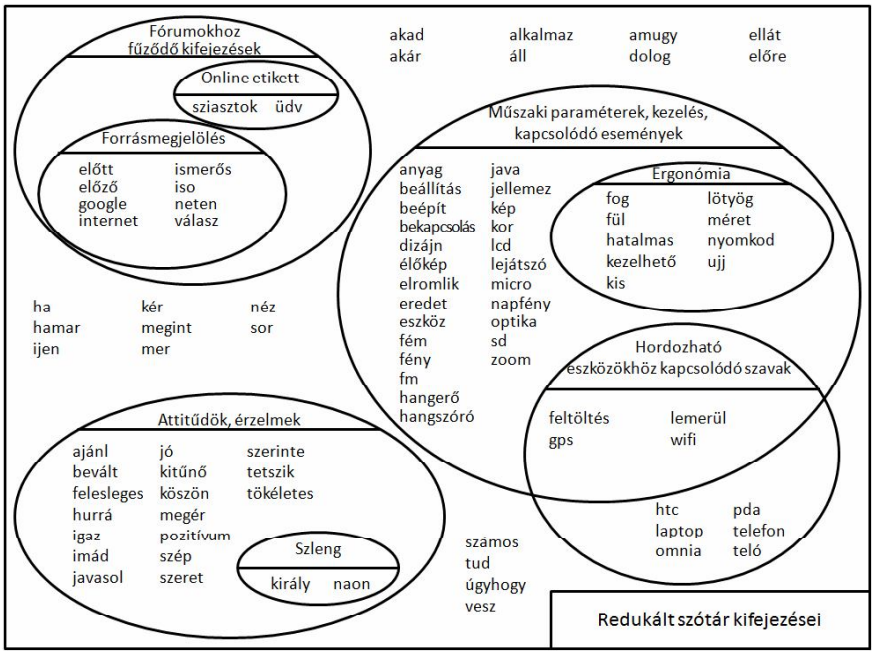
A khi-négyzet mutató is hasonló elven működik, hiszen ez is a szó-kategória halmazok Descartes-szorzatain belüli tényleges és függetlenség esetén várható közös előfordulási valószínűségekkel számol.

$$\chi^2_k = \frac{N - \big(n(t_k, c_h)n(\overline{t}_k, c_n) - n(t_k, c_n)n(\overline{t}_k, c_h)\big)^2}{n(c_h)n(c_n)n(t_k)n(\overline{t}_k)} \,.$$

Eredményeink szerint, a szavakra kiszámolt $IG$ és a khi-négyzet mutatók a vizsgált korpuszon erősen korreláltak. A lineáris korrelációs együttható a két adatsor között 98,65% volt, ezért a továbbiakban a khi-négyzet mutatót nem vizsgáltuk külön. Az $IG$ mutatóval pozitívan korrelált továbbá a

gyűjteménytámogatottság, 36,79%-os lineáris korrelációs együtthatóval. Az $MI$ mutató értékei nem mutattak erős korrelációt sem az $IG$ sem a gyűjteménytámogatottság adatsorával (a lineáris korrelációs együtthatók rendre -6,12% és -14,2%). A fenti mutatók közüli választás során szakirodalmi forrásokra támaszkodtunk.

Yang és Pedersen (1997) nyomán tudjuk, hogy a kölcsönös információn alapuló dimenzióredukció teljesítménye nem közelíti meg a legeredményesebb módszerek közé tartozó információnyereség mutatóét. Tekintve, hogy az $IG$ mutató számítása során az $MI$ értékét is figyelembe veszi, nem meglepő, hogy a kifinomultabb mutató jobb eredményre képes. A gyűjteménytámogatottság a legjobb teljesítményű nem felügyelt dimenzióredukciós módszerek közé tartozik (Garnes, 2009). Ez azt jelenti, hogy amennyiben nem áll rendelkezésre előzetesen felcímkézett adatbázis – és így nem használhatók a felügyelt módszerek, mint például az $IG$, akkor ezzel a módszerrel hatékonyan lehet csökkenteni a szótár méretét. Garnes (2009) vizsgálata során azonban az mutatkozott, hogy a gyűjteménytámogatottság alulmarad az információnyereség mutatóval szemben osztályozási feladatok pontosságának javítása szempontjából – összhangban azzal, amit Yang és Pedersen (1997) is állított, hogy az $IG$ mutató az egyik leghatékonyabb az osztályozási problémák dimenziószámának csökkentésére. A korábbi kutatási eredmények áttekintése alapján tehát az információnyereség mutató bizonyult a legalkalmasabbnak a redukálandó dimenziók kiválasztására, ezért mi is ezt alkalmaztuk.




*4. ábra.* A tanítási folyamatba bevont redukált szólista elemei

Kutatásunkban az *IG* mutató alapján rangsorolt szólista legkisebb értékű elemeit hagytuk el, 100 eleműre csökkentettük a szótár méretét (ezek egyfajta kategorizálását mutatja a *4. ábra*).

Próbafuttatások során nem kaptunk jelentősen jobb eredményt a több szóból álló TDM-ek segítségével, az agresszívebb redukció viszont már jelentősen rontotta az eredményeket. Ennek oka nem csupán közvetlenül a dimenzió-szám csökkentése volt, hanem – abból következően – a nemnulla koordinátájú dokumentumok számának a csökkenése is. A TDM ritka mátrix, ezért várható, hogy a dimenziók számának csökkentése maga után vonja a nullvektorral jellemezhető dokumentumok számának növekedését. Mivel mind az ANN, mind az SVM szeparálósíkok révén osztályozza a dokumentumokat, ezért az origó – a határesettől eltekintve – csak az egyik osztálytérnek lehet eleme. Ha az origóban elhelyezkedő dokumentumokban az osztálycímkék koncentrációja alacsony – azaz a manuális osztályozással sok hasznosnak, de sok nem hasznosnak ítélt dokumentum is található köztük – akkor az origóban lévő elemeknek bármely osztályhoz rendelése magas hibát eredményezhet.

A részletes elemzés előtt, a fenti okok miatt megvizsgáltuk a mintában található azon dokumentumokat, amelyeknek reprezentációja nullvektor volt a redukált TDM-ben. Az eredetileg 991 dokumentumot tartalmazó gyűjteményből a jellemző-kiválasztás után 93 dokumentumvektornak minden koordinátája nulla volt. Ebben a 93 elemű részmintában 4 hasznos és 89 nem hasznos kommentet találtunk, ebből következik, hogy nem vétünk nagy hibát, ha nem hasznosnak ítéljük azokat a hozzászólásokat, amelyek nem tartalmaznak egyet sem a redukált szótár szavai közül. Ebből kifolyólag a nullvektorokat kizártuk a mintából, az osztályozó algoritmusok jóságát csak a többi dokumentumvektoron elért teljesítményre állapítottuk meg.

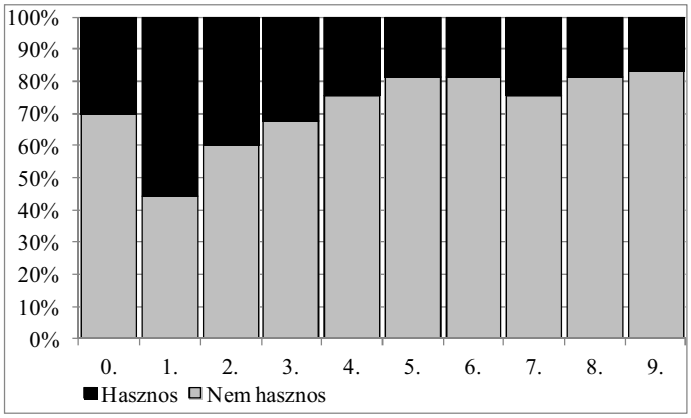
# 6 A felügyelt tanulás folyamata

Az attribútum szűkítési és TDM redukálási folyamat eredményeként egy 898 elemű dokumentumgyűjtemény állt elő. Ezt tíz, közel egyenlő méretű részhalmazra osztottuk, hogy rajtuk tízszeres keresztvalidációt hajtsunk végre. A tízszeres keresztvalidáció során tíz független futtatást végzünk, ahol az egyes futtatások során kilenc részhalmaz dokumentumai szolgálnak tanítómintaként, míg a maradék részhalmaz játssza a tesztminta szerepét. Az osztályozó módszerek jóságát a tesztmintákon elért találati aránnyal mérjük. Az angol kifejezéssel accuracy measure-nek nevezett mutatószám képlete a következő (Powers 2011):

$$\text{Accuracy} = \frac{\sum_{j\in\{h,n\}} n(\hat{c}_j, c_j)}{N} \,,$$

ahol $\hat{c}_j$ a dokumentumok becsült kategóriáját jelöli ($\hat{c}_h$ a hasznos, $\hat{c}_n$ a nem hasznos becsült kategória jele), $n(\hat{c}_j, c_j)$ pedig a manuálisan $j$ kategóriába ($c_j$) sorolt és $j$ kategóriájúnak is becsült ($\hat{c}_j$) dokumentumok száma. A számláló tehát azon dokumentumok számát adja meg, melyek manuális és gépi címkéje (kategóriába sorolása) megegyezik.

Az *5. ábrán* látható a – manuális címkézés szerint – hasznos és nem hasznos dokumentumok megoszlása az egyes részhalmazokban, amelyeket 0-tól 9-ig számoztunk. Mint látható, egyedül az 1. tesztmintában haladta meg a hasznos dokumentumok aránya az 50%-ot (a redukált korpuszon belül a hasznosnak ítélt dokumentumok aránya 28%).




*5. ábra.* A hasznos és nem hasznos dokumentumok megoszlása az egyes részmintákban

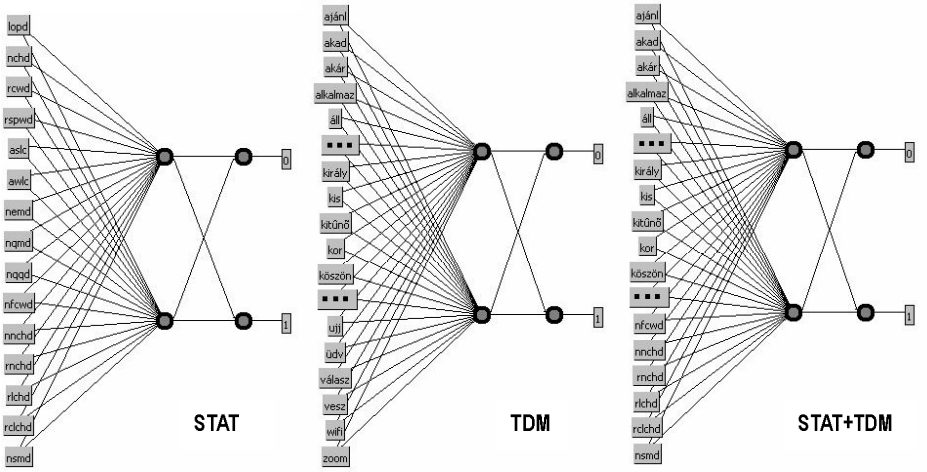
A hasznosság automatikus megállapításához két módszert alkalmaztunk: mesterséges neurális hálózatot (Artificial Neural Network – ANN) és Support Vector Machine-t (SVM). A felügyelt tanítást mindkét módszer esetén három verzióban hajtottuk végre:

- csak a TDM-en kívüli, szűkített attribútum halmaz, tehát a statisztikai szövegjellemzők (STAT),
- csak a redukált szó-dokumentum mátrix (TDM), valamint
- a két halmaz uniójának (TDM+STAT) felhasználásával.

A neurális hálózatok topológiájáról elmondhatjuk, hogy kismértékben eltérnek egymástól a három verzióban, viszont az összehasonlíthatóság érdekében csak a feltétlenül szükséges módosítási lehetőségekkel éltünk. A modellezéshez a Weka 3.6 szoftver MultiLayer Perceptron (MLP) nevű eszközét használtuk.

Mindhárom többrétegű perceptron topológia egy input rétegből, egy rejtett rétegből és egy output rétegből áll. Az input réteg a STAT inputhalmaz esetében 15 neuronos, tehát megegyezik a statisztikai szövegjellemzők számával. A TDM inputhalmazhoz tartozó topológia esetén – a magyarázó attribútumok számának megfelelően – 100, míg a TDM+STAT inputhalmaz esetén 115 neuron széles az input réteg. Az input neuronok aktivációs függvénye a szokásos lineáris jelzési függvény. A rejtett réteg mindhárom esetben két neuron széles, és szigmoid jelzési függvényekkel rendelkezik. A rejtett réteg szélességét nincs értelme MLP topológia esetén alacsonyabbra állítani, mivel 1 rejtett neuron esetén a Perceptron topológiával ekvivalens modellt kapunk. Két rejtett neuron viszont már nem csak lineáris szeparációkra

ad lehetőséget: az OR, AND és NOT műveletek mellett az XOR kapcsolat modellezésére is képessé teszi a hálózatot. A kettőnél több rejtett neuronnal rendelkező hálózatok pontosságát is megvizsgáltuk, de azok nem adtak jobb becslést a kategóriák címkéire. Ez alól nem volt kivétel az input- és output neuronok számán alapuló hüvelykujj szabály által javasolt neuronszám sem. Hasonlóképpen, a mélyebb topológiák sem eredményezték a pontosság jelentős javulását, tehát nem alkalmaztunk egynél több rejtett réteget. Az output réteg tulajdonságai igazodnak a becsülendő változó lehetséges értékeinek számához, tehát mindhárom esetben két output neuronra van szükség – egyik a hasznos kategóriához, másik a nem hasznos kategóriához. Az output jelzési függvények szigmoid típusúak. A három réteg között teljes előrecsatolást létesítettünk, ahogy a Multilayer Perceptron topológiánál szokás. A hálózat szinaptikus súlyait backpropagation algoritmussal állítottuk be, 500 epochos tanulás során. A tanuláshoz 0,3-as tanulási rátát és 0,2-es momentum paramétert használtunk. Ezek az értékek alapbeállítások a Weka 3.6-ban, és a további, nem említett paramétereket is a szoftver által ajánlott értéken hagytuk. A három topológia sematikus illusztrációi láthatók a *6. ábrán.*




*6. ábra.* A mesterséges neurális hálózatok topológiája

A Support Vector Machine alkalmazásához a Weka 3.6 szoftver libSVM modulját használtuk. Mindhárom attribútumhalmaz esetén $\nu$-SVM (Chen és sztsai (2005)) tanulási módszert és radiális bázis kernel függvényt használtunk. A három verzióban rács-keresési (grid search) algoritmust alkalmazva határoztuk meg $\gamma$ és $\nu$ azon értékeit, melyek mellett a legjobb becslési eredményeket nyertük (*5. táblázat*). A többi paraméter értékét alapbeállításon hagytuk.

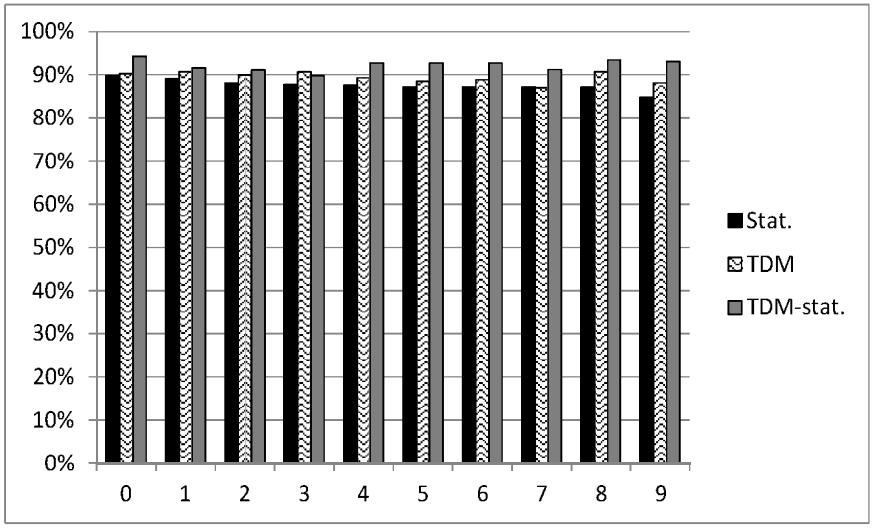
| Attribútumok köre | $\gamma$ | $\nu$ |
|---|---|---|
| TDM | 0,04 | 0,028 |
| STAT | 0,0001 | 0,3 |
| TDM + STAT | 0,0001 | 0,3 |

*5. táblázat.* $\gamma$ és $\nu$ értékei az SVM különböző verzióiban

# 7 A tanulás eredményességének elemzése

Ebben a szakaszban bemutatjuk, hogy milyen pontossággal állapítják meg a hozzászólások hasznosságát a vizsgált módszerek a különböző attribútumhalmazok esetén. A mesterséges neurális hálózattal osztályozott korpusz esetén kiszámított pontosság mutatók (accuracy measure) értékei a *7. és 8. ábrán* láthatók. A tízszeres keresztvalidáció tíz tanulási futtatásának jelölése az aktuális tesztminta sorszámával egyezik meg, azaz a diagram „0" jelzésű oszlopai azt az esetet mutatják, amikor a tesztmintát a 0. sorszámú, míg a tanítómintát az 1-9. sorszámú részhalmazok alkották.

A 7. ábrán a tanítómintán történt tanulás jósága látható. A tíz tanítómintán hasonló sorrendben követték egymást a háromféle attribútumhalmaz esetén számított pontossági mutatók. A legjobb eredményt a 3-as jelű futtatás kivételével mindig a kombinált attribútumhalmazhoz (TDM+STAT) tartozó osztályozás adta, általában 90% feletti értékekkel. Ezt követte a csupán a TDM információit használó címkézés, azonban ennek pontossága már nem mindenhol érte el a 90%-ot. Az előbbiekből pedig az következik, hogy a tanítómintákon a leggyengébb eredményt a tisztán szövegstatisztikai attribútumokon (STAT) alapuló klasszifikáció szolgáltatta, melynek pontossága csupán egy esetben érte el a 90%-ot. Az oszlopdiagramhoz tartozó számadatokat a Függelék tartalmazza.

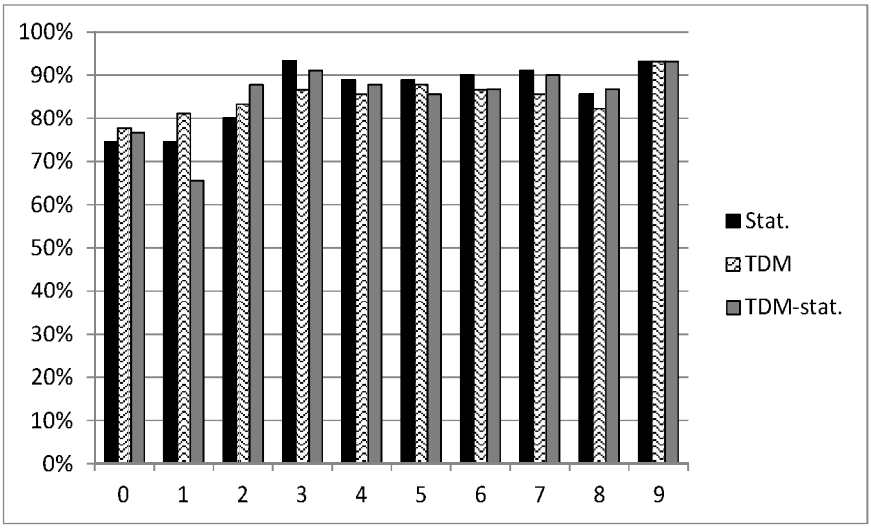



*7. ábra.* Az ANN által elért pontossági értékek a tanítómintákon

Minél magasabb dimenziójú térben szeretnénk pontokat elhatárolni egymástól, annál kevésbé kell nemlineáris szeparációkhoz folyamodnunk, így a többrétegű perceptron topológia súlyparaméterei gyorsabban konvergálnak. Mivel a három inputminta közül a szövegstatisztikai jellemzők vannak a legkevesebben (15), ezért várható volt, hogy ebben a térben nehezebben szeparál az ANN. A 100, attribútumként szolgáló szóval kibővített input-adathalmaz

(TDM+STAT) viszont – a várakozásoknak megfelelően – a legjobban szétválasztható input-teret szolgáltatta.

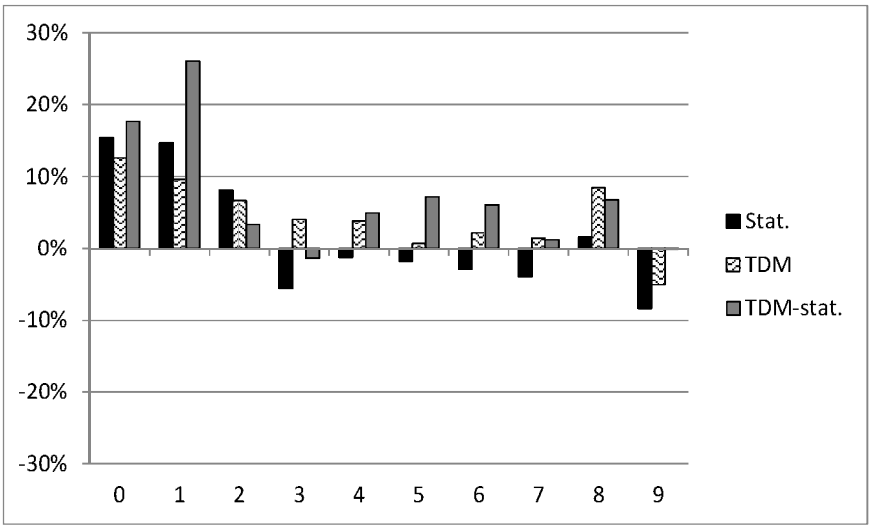
Az 8. ábra a tesztmintákon történő osztályozás pontosságát mutatja. Azt láthatjuk, hogy a tesztmintákon jelentősen romlott a pontosság, és a korábbi sorrendet sem őrizték meg a modellek. Meglepő módon a 90%-os határt most a STAT halmazt használó modell érte el a leggyakrabban, a kombinált (TDM+STAT) inputhalmaz alapján történő klasszifikációhoz képest eggyel többször.




*8. ábra.* Az ANN által elért pontossági értékek a tesztmintákon

Árnyaltabb képet kaphatunk a bekövetkezett változásokról, ha megvizsgáljuk a tanítómintákra és a tesztmintákra számított pontosság mutatók különbségét (*9. ábra*).

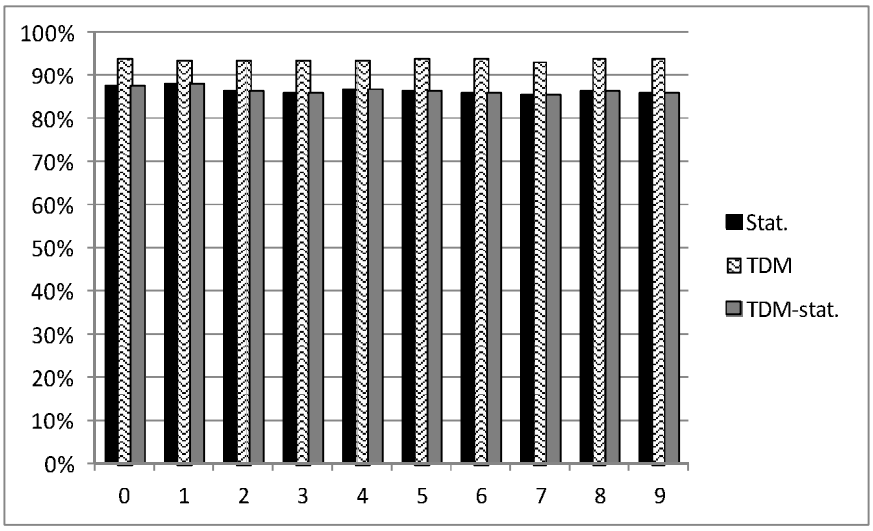
Az ábra alapján elmondható, hogy a „0", „1", „2" és „8" jelű tesztminták esetén mindegyik módszer jelentősen rosszabb teszteredményeket szolgáltatott a tanítómintákra számítottnál. Ez a pontosság mutatók pozitív előjelű különbségéből látszik, ami azt jelenti, hogy a tanítómintán magasabb pontosságarányt sikerült elérni, mint a tesztmintán. A többi tesztmintán viszont a kizárólag szövegstatisztikai attribútumokon (STAT) végzett osztályozás még pontosabbnak is mutatkozik a tanítómintához képest. Mivel a tanítóminta segítségével lettek beállítva a neurális hálózatok paraméterei, ráadásul az összesített négyzetes hiba lokális minimuma közelében, ezért egy ettől a tanítómintától eltérő adathalmaz esetén nem várható jelentősen jobb eredmény. Megfigyelhető továbbá, hogy a két másik input-adathalmazra épülő modell közül csak egy-egy esetben kaptunk a tanítómintánál jobb pontosságmutatót a hozzá tartozó tesztmintán. A pontos adatok a Függelékben megtekinthetők.




*9. ábra.* Az ANN által osztályozott tanító- és tesztminták pontossági értékének különbsége

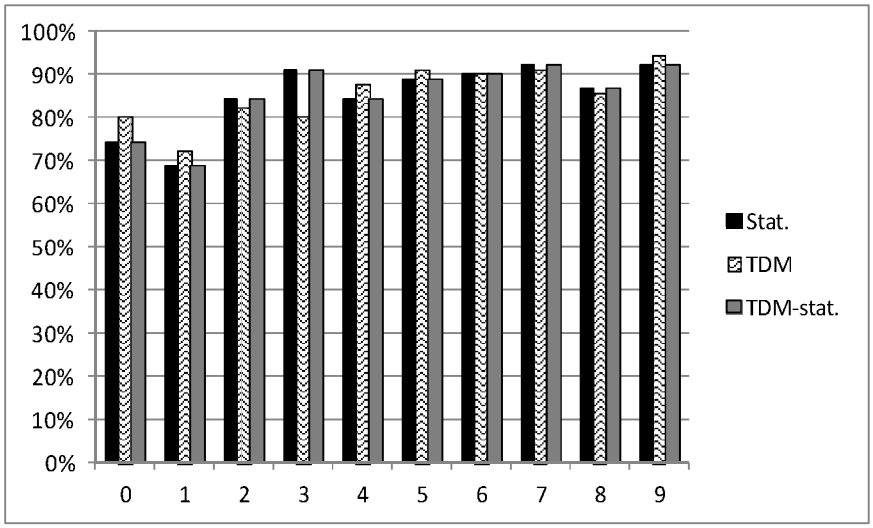
A fentiek arra engednek következtetni, hogy a magas dimenziójú terekben a hálózatok olyan szeparációkat létesítenek, amelyek a tanítómintában meglévő – csupán az adott mintára jellemző – térrészek felismerését teszik lehetővé, és ez a tesztmintákon hibás szeparációkhoz vezet. Ezt nevezzük a neurális hálózat túltanulásának is. A kis dimenziójú döntési terekben tehát sokkal jobb általánosító-képességű modelleket kaphatunk, feltéve, hogy megfelelő információt szolgáltatnak ezek a dimenziók. (Így például hiába próbálkoznánk a TDM dimenzióinak további csökkentésével, mert az már jelentős információveszteséghez vezetne.)

Az SVM módszerrel elvégzett tanítás pontosság mutató értékei láthatók a *10. és 11. ábrán*. A részminták jelölései megegyeznek az ANN eredményeinek bemutatásakor alkalmazottal. A 10. ábrán a tanítómintára való rátanulás jósága látható. Mind a tíz tanítómintán a TDM attribútumhalmaz segítségével érte el a legnagyobb pontossági értéket az SVM algoritmus, minden esetben 90% fölötti pontossággal. A STAT, illetve a TDM+STAT attribútumhalmazon minden esetben 90% alatt maradt (86-88%) a pontosság. További érdekesség, hogy az utóbbi két attribútumhalmaz esetén pontosan ugyanazokat az eredményeket kaptuk. Ennek magyarázata az lehet, hogy a kommentek hasznosságát elsősorban a komment szerkezete befolyásolja, és csak másodsorban a kulcsszavak: a STAT dimenziók nagyobb súlyt kapnak a szeparálósíkok illesztésében, és ezek a szeparációk a TDM dimenziók mentén is pontosan bontják szét a megfigyeléseket hasznos és nem hasznos dokumentumokra. Az oszlopdiagramhoz tartozó számadatokat a Függelék tartalmazza.




*10. ábra.* Az SVM által elért pontossági értékek a tanítómintákon

A *11. ábra* a tesztmintákon történő SVM-osztályozás pontosságát mutatja. Azt láthatjuk, hogy a tesztmintákon elért pontosság szórása nagyobb a tíz tesztminta esetén, mint a hozzájuk tartozó tanítómintáknál. A 90%-os pontosságot ritkán érte el a TDM terében osztályozó SVM, viszont a másik két attribútumhalmaz – a tanítómintákkal ellentétben – több tesztmintán is meghaladta ezt a szintet.

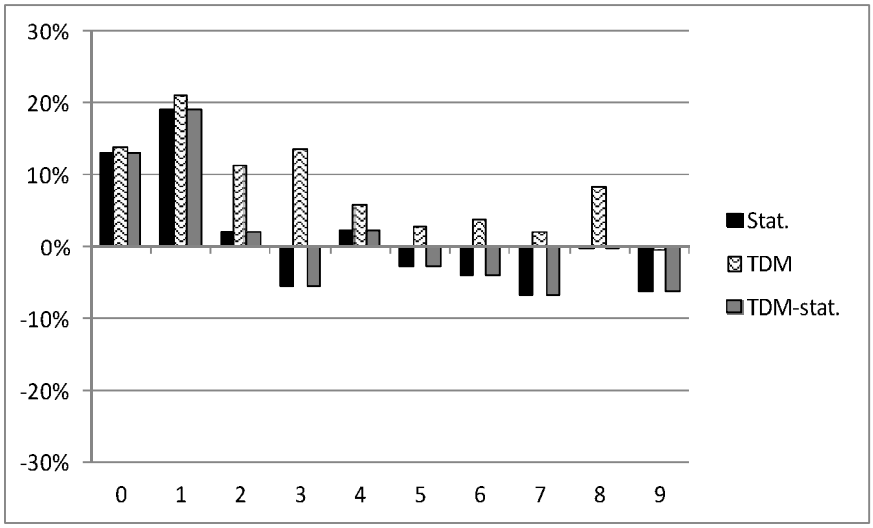



*11. ábra.* Az SVM által elért pontossági értékek a tesztmintákon

A Weka adatbányászati alkalmazás – ahogy korábban láttuk – lehetőséget nyújt arra, hogy sorrendbe állítsuk az attribútumokat aszerint, hogy azok

egy SVM során milyen osztályozó erővel bírnak. A korábban kiválasztott STAT attribútumok közül a karakterek száma (nchd), a nagybetűvel kezdődő szavak száma (nfcwd) és a helyesírásilag elfogadható szavak aránya (rspwd) jellemzők kerültek az első, míg az átlagos szóhossz (awlc), az idézőjelek száma (nqqd) és a smiley-k száma (nsmd) attribútumok az utolsó három helyre. A TDM elemei kapcsán – nem meglepő módon – a műszaki paraméterekkel kapcsolatos szavak kerültek nagyobb arányban a lista elejére. Az együttes attribútumhalmazra végzett sorbarendezés kapcsán azt a megállapítást tehetjük, hogy egyik jellemzőtípus sem dominálja a másikat, az abszolút első helyre pedig a karakterek száma attribútum került. Megjegyezzük, hogy ilyen sorbarendezést a mesterséges neurális hálózat esetén nem tudunk adni.

Vizsgáljuk meg az SVM esetén is a tanítómintákra és a tesztmintákra számított pontosság mutatók különbségét (*12. ábra*).




*12. ábra.* Az SVM által osztályozott tanító- és tesztminták pontossági értékének különbsége

Az ANN vizsgálatánál leírtakhoz hasonlóan a pontosság mutatók pozitív előjelű különbsége azt jelenti, hogy a tanítómintán magasabb pontosságarányt sikerült elérni, mint a tesztmintán. A 12. ábra alapján elmondható, hogy a ,,0", ,,1", ,,2" és ,,4" jelű tesztminták esetén mindegyik módszer rosszabb teszteredményeket szolgáltatott a tanítómintákra számítottnál. A többi tesztmintán viszont a STAT és TDM+STAT attribútumokon végzett osztályozás pontosabbnak (vagy legalább ugyanolyan pontosnak) mutatkozott, mint a tanítómintákon végrehajtott. A pontos adatok a Függelékben megtekinthetők.

A fent nyert eredmények megítéléséhez segítséget nyújt, ha felidézzük, hogy a megfigyeléseknek kb. 28 százaléka bizonyult hasznosnak a manuális címkézés után. Ezt kihasználva konstruálhatunk olyan triviális modellt, amely kb. 72%-os találati arányt képes elérni úgy, hogy minden dokumentum esetén a ,,nem hasznos" címkét alkalmazza. Ebben az esetben az összes

nem hasznos dokumentumot helyesen becsülnénk, de lemondanánk a 28%-nyi hasznos hozzászólás elkülönítéséről. Az általunk használt modellek kevésbé biztosan címkézik a nem hasznos hozzászólásokat, cserében viszont a hasznos hozzászólások egy jelentős részét felismerik. A teszteken elért 80-90% közötti pontosság abból tevődik össze, hogy egyrészt a triviális modell által a hasznos dokumentumok osztályozása kapcsán elkövetett 28%-nyi hibát az ANN és az SVM 10% alá szorította, másrészt a hibásan osztályozott negatív dokumentumok aránya 0%-ról kb. 5%-ra emelkedett átlagosan.

# 8 Összefoglalás

Cikkünkben ismertettük, hogy az irodalomban miként értelmezik a termékekre vonatkozó internetes hozzászólások hasznosságát, valamint hogy milyen szövegjellemzőkkel modellezik azt. Az irodalmi tapasztalatokat is felhasználva mutattunk be egy mesterséges neurális hálózatra (ANN) és egy support vector machine-re (SVM) épülő módszert, amikkel lehetőség nyílik a hozzászólások hasznosságának automatikus megállapítására.

A kapott eredmények azt mutatják, hogy a TDM attribútumok mentén mind az ANN, mind az SVM szeparáló módszerek nagy pontosságot (accuracy) képesek elérni a tanítómintán, azonban a független tesztmintákon számottevően gyengébb eredményt produkálnak. Ezzel ellentétben, a statisztikai szövegjellemzők (STAT) segítségével végzett osztályozás esetén, a tesztmintán történő validáció során még pontosságjavulást is ki tudtunk mutatni a tanítómintán elért értékekhez képest. Az ANN és SVM módszerek egymáshoz hasonló eredményeket szolgáltattak, ezért azokat egyformán alkalmasnak tartjuk a szövegosztályozási feladat elvégzésére.

Mindeközben azt is megfigyelhettük, hogy a statisztikai szövegjellemzők (STAT) osztályozásra való alkalmasságuk szerint jól meghatározott csoportokba (strukturális, nyelvtani, tájolás stb.) rendelhetők. Ezek közül nyelvfüggetlen módon egyedül a strukturális attribútumok kezelhetők, de a többi jellemző értékének meghatározásához sem alkalmaztunk speciális nyelvészeti eszközöket. Ilyen módszerek felhasználásával bizonyos attribútumok precízebb tartalmat kaphatnak, ami jelentősen javíthatja az osztályozás pontosságát.

A kidolgozott eljárás korlátai között elsőként azt említhetjük, hogy módszerünk statikus korpuszon alapul. A korpusz változása megkövetelheti a teljes újrafuttatást, egyes paraméterek ismételt kalibrálását. Amíg az SVM esetén az attribútumok sorbarendezhetők osztályozó erejük szerint, addig az alkalmazott mesterséges neurális hálózat kapcsán semmilyen elképzeléssel nem rendelkezünk a jellemzők osztályozáshoz való hozzájárulásáról. Ennek legfőbb oka az, hogy a nem lineáris topológia nem támogatja az optimális paraméterekből történő információkinyerést. Tanulmányunk meghatározó részét tette ki az attribútumok kiválasztására vonatkozó rész, aminek folyamata teljes mértékben az aktuális korpusz jellemzőin alapult: az attribútumok optimális részhalmazának meghatározására nem rendelkezünk általános módszerrel, ennek kidolgozása további kutatási irányt jelöl ki.

További fejlesztési lehetőségként merül fel a hozzászólások gyűjtése során megszerezhető metaadatok felhasználása (például szerző, tetszés index), de egyéb attribútumokkal is bővíthető a magyarázó változók köre. Érdekes kérdésként merül fel a kommentek manuális címkézési módszerének módosítása, ugyanis több megkérdezett bevonása lehetővé tenné a szubjektum szerepének mélyebb vizsgálatát.

# Függelék

A 7. szakaszban bemutatott ábrákhoz tartozó pontos számadatok az alábbiakban olvashatók:

| | ANN pontossági értékek, % | | | SVM pontossági értékek, % | | |
|---|---|---|---|---|---|---|
| | STAT | TDM | TDM+STAT | STAT | TDM | TDM+STAT |
| Tanítóminta | | | | | | |
| 0 | 89,85 | 90,35 | 94,31 | 87,50 | 93,94 | 87,50 |
| 1 | 89,11 | 90,72 | 91,58 | 87,87 | 93,44 | 87,87 |
| 2 | 88,12 | 89,98 | 91,09 | 86,39 | 93,56 | 86,39 |
| 3 | 87,75 | 90,72 | 89,73 | 85,77 | 93,56 | 85,77 |
| 4 | 87,62 | 89,36 | 92,70 | 86,63 | 93,56 | 86,63 |
| 5 | 87,13 | 88,49 | 92,70 | 86,26 | 93,94 | 86,26 |
| 6 | 87,13 | 88,86 | 92,70 | 86,01 | 93,81 | 86,01 |
| 7 | 87,13 | 87,00 | 91,21 | 85,52 | 93,32 | 85,52 |
| 8 | 87,13 | 90,70 | 93,44 | 86,39 | 93,81 | 86,39 |
| 9 | 84,81 | 88,15 | 93,09 | 85,93 | 93,83 | 85,93 |
| Tesztminta | | | | | | |
| 0 | 74,44 | 77,78 | 76,67 | 74,44 | 80,00 | 74,44 |
| 1 | 74,44 | 81,11 | 65,56 | 68,89 | 72,22 | 68,89 |
| 2 | 80,00 | 83,33 | 87,78 | 84,44 | 82,22 | 84,44 |
| 3 | 93,33 | 86,67 | 91,11 | 91,11 | 80,00 | 91,11 |
| 4 | 88,89 | 85,56 | 87,78 | 84,44 | 87,78 | 84,44 |
| 5 | 88,89 | 87,78 | 85,56 | 88,89 | 91,11 | 88,89 |
| 6 | 90,00 | 86,67 | 86,67 | 90,00 | 90,00 | 90,00 |
| 7 | 91,11 | 85,56 | 90,00 | 92,22 | 91,11 | 92,22 |
| 8 | 85,56 | 82,22 | 86,67 | 86,67 | 85,56 | 86,67 |
| 9 | 93,18 | 93,18 | 93,18 | 92,05 | 94,32 | 92,05 |
| Különbség | | | | | | |
| 0 | 15,41 | 12,57 | 17,64 | 13,06 | 13,94 | 13,06 |
| 1 | 14,66 | 9,61 | 26,03 | 18,98 | 21,22 | 18,98 |
| 2 | 8,12 | 6,64 | 3,31 | 1,94 | 11,34 | 1,94 |
| 3 | -5,59 | 4,05 | -1,38 | -5,34 | 13,56 | -5,34 |
| 4 | -1,27 | 3,80 | 4,92 | 2,19 | 5,79 | 2,19 |
| 5 | -1,76 | 0,71 | 7,14 | -2,63 | 2,82 | -2,63 |
| 6 | -2,87 | 2,19 | 6,03 | -3,99 | 3,81 | -3,99 |
| 7 | -3,98 | 1,45 | 1,21 | -6,70 | 2,21 | -6,70 |
| 8 | 1,57 | 8,48 | 6,77 | -0,28 | 8,26 | -0,28 |
| 9 | -8,37 | -5,03 | -0,10 | -6,12 | -0,49 | -6,12 |

## Irodalom


1. Burk, S. (2007): An automated scoring system for measuring email emotion. *Marketing Bulletin*, 18, 1–12.
2. Cao, Q., Duan, W., Gan, Q. (2011): Exploring determinants of voting for the „helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, 50 (2), 511–521.
3. Chen, C. C., Tseng Y. (2011): Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, 50 (4), 755–768.
4. Chen, P., Lin, C., Schölkopf, B. (2005): A tutorial on n-support vector machines. *Applied Stochastic Models In Business and Industry*, 21, 111–136.
5. Cheung, K., Kwok, J. T., Law, M. H., Tsui, K. (2003): Mining customer product ratings for personalized marketing. *Decision Support Systems*, 35 (2), 231–243.
6. Coussement, K., Van den Poel, D. (2008): Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44 (4), 870–882.
7. Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., Lee, L. (2009). How opinions are received by online communities: a case study on amazon.com helpfulness votes. *WWW '09 Proceedings of the 18th international conference on World Wide Web*, 141–150.
8. Decker, R., Trusov, M. (2010): Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27, 293–307.
9. Dellarocas, C. (2003): The digitization of word of mouth: promise and challenges of online feedback mechanisms. *Management Science*, 49 (10), 1407–24.
10. Duan, H., Zirn, C. (2012): Can we identify manipulative behavior and the corresponding suspects on review websites using supervised learning?, In *Proceedings of NordSec'12*, Berlin, Heidelberg.
11. Duan, W., Gu, B., Whinston, A. B. (2008): The dynamics of online word-of-mouth and product sales – an empirical envestigation of the movie industry. *Journal of Retailing*, 84 (2), 233–242.
12. Garnes, Ø. L. (2009): Feature selection for text categorisation, master's thesis, Norwegian University of Science and Technology, http://ntnu.diva-portal.org/smash/get/diva2:347827/FULLTEXT01, Letöltve: 2011.12.13.
13. Guyon, I., Elisseeff, A. (2003): An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
14. Kim, S., Pantel, P., Chklovski, T., Pennacchiotti, M. (2006): Automatically assessing review helpfulness. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 423–430.
15. Li, N., Wu, D. D. (2010): Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48, 354–368.
16. Liu, J., Cao, Y., Lin, C., Huang, Y., Zhou, M. (2007): Low-quality product review detection in opinion summarization. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 334–342.
17. Lu, Y., Tsaparas, P., Ntoulas, A., Polanyi, L. (2010): Exploiting social context for review quality prediction. *WWW '10 Proceedings of the 19th international conference on World Wide Web*, 691–700.

18. O'Mahony, M. P., Smyth, B. (2010): Using readability tests to predict helpful product reviews. In *Proceedings of RIAO'2010*, 164–167.
19. Pan, Y., Zhang, J. Q. (2011): Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of Retailing*, 87 (4), 598–612.
20. Pon, R. K., Cárdenas, A. F., Buttler, D. J., Critchlow, T. J. (2011): Measuring the interestingness of articles in a limited user environment. *Information Processing and Management*, 47, 97–116.
21. Powers, D. M. W. (2011): Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2 (1), 37–63.
22. Siersdorfer, S., Chelaru, S., Nejdl, W., Pedro, J. S. (2010): How useful are your comments?: analyzing and predicting youtube comments and comment ratings. *WWW '10 Proceedings of the 19th international conference on World Wide Web*, 891–900.
23. Tsur, O., Rappoport, A. (2009): RevRank: A fully unsupervised algorithm for selecting the most helpful book reviews. *Proceedings of the Third International ICWSM Conference*, 154–161.
24. Wu, G., Greene, D., Cunningham, P. (2010): Merging multiple criteria to identify suspicious reviews. In *Proceedings of RecSys'2010*, 241–244.
25. Xie, S., Wang, G., Lin, S., Yu, P. S. (2012): Review spam detection via temporal pattern discovery, In *Proceedings of the 18th ACM SIGKDD*, New York, 823–831.
26. Yang, Y., Pedersen, J. O. (1997): A comparative study on feature selection in text categorization. *CML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, 412–420.
27. Zhu, F., Zhang, X. M. (2010): Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74 (2), 133–148.

## ASSESSING THE HELPFULNESS OF ONLINE PRODUCT REVIEWS WITH SUPERVISED MACHINE LEARNING TECHNIQUES

In recent years Internet became a major source of information for the corporate marketing function. More and more articles study the opportunities to utilize user-generated web documents. Concept Extraction (Concept Mining) is a potential research direction of extracting information from customer reviews on products. Concept Extraction explores and analyzes customers' opinions on products and focuses on the content, quality or helpfulness of their reviews. In this paper, first we collect and systematize the different approaches of customer review helpfulness, then we present an Artificial Neural Network (ANN) and a Support Vector Machine (SVM) supervised learning method based on three different sets of text features to automatically determine the helpfulness of customer reviews.