Magyar szövegek véleményanalízise[1]

Szaszkó Sándor[1], Sebők Péter[1], Kóczy T. László[1, 2]

[1] Budapesti Műszaki Egyetem, Távközlési és Média Informatikai Tanszék
1117, Budapest, Magyar tudósok körútja 2.
{Szaszko, Sebok, Koczy}@tmit.bme.hu
2 Széchenyi István Egyetem, Jedlik Ányos Gépész-,
Informatikai és Villamosmérnöki Intézet
9026, Győr, Egyetem tér 1.

Kivonat: A témaalapú osztályozásokban ismert módszerek hatékonyságát mutatjuk be a dokumentumok orientációjának eldöntésére. Ehhez összeállítottunk 240 dokumentumos tanító korpuszt. Az angol eredményekhez hasonlóan a klasszikus megoldások közül az SVM a leghatékonyabb, de ennek a teljesítményén is javít az eddig e célra nem használt RRM osztályozó. A Fuzzy-IDF súlyozás bevezetésével a kis felidézésű régióban a pontosságot tovább javítottuk.

1 Bevezetés

A dokumentumosztályozási feladat az egyik legismertebb szövegbányászati kutatási terület. Megoldásával hosszú idő óta foglalkozik a tudományos közösség, mára számos ipari alkalmazás igen jó hatékonyságú eredményt ad. A megoldások hátterében a legerősebb faktor, hogy az osztályokra jellemző szó halmazt gépi tanulási módszerrel közelítjük.

A véleményanalízis egy olyan két kategóriás osztályozási feladat, ahol a dokumentumok témája azonos. A különbséget a szöveg és a téma viszonyában keressük. Kutatásunk során filmekről szóló kritikákat elemeztünk, célunk a kritika pozitív vagy negatív beállítottságának eldöntése volt. A feladat nehézségét jól mutatja, hogy – bár jelentős előnnyel járna – a pozitív vagy negatív minősítés számszerűsítésére a szakirodalomban nem találtunk példát.

A téma a legújabb kutatási területekhez tartozik. Magyar nyelvű szövegek véleményanalízisével – legjobb tudásunk szerint – eddig csak egy szerzőpáros foglalkozott, Berend és Farkas jellemzően rövid, egymásra reagáló fórumbejegyzések alapján eredményesen jósolta a részvevők véleményét egy választási referendumról [9]. A mi vizsgálatunk tárgyát képező, egymástól független, hosszabb szövegek vizsgálata egészen más eszközök bevetését igénylik.

[1] OTKA K75711 számú támogatási szerződés keretében végzett kutatás.

1.1 Irodalmi eredmények

Hatzivassiloglou és McKeown (1997) melléknevek orientációjának meghatározását végezték el, majd ezek előfordulásának függvényében döntöttek [1]. Az általuk javasolt módszer képes arra, hogy a dokumentumokból kinyert melléknevek orientációját bizonyos halmazon 78%-ot meghaladó pontossággal becsülje meg. Szavak orientációját használja még [2] és [3] is.

Pang és társai sok további kutatásnak adtak irányt, amikor bebizonyították, hogy a gépi tanulási módszerekkel jobb eredmény érhető el, mint a priori módszerrel [4]. Naive Bayes, Maximum Entrópia és szupport vektor gép (SVM) módszerek teljesítményét hasonlították össze, ahol az SVM-et találták a leghatékonyabbnak. Ehhez hasonlóan a termékkritikák minősítésével foglalkozó [5] eredményei is az SVM (76%) elsőbbségéről tanúskodik.

A legjobb eredményeket a kivonatolás és gépi tanulási módszerek kombinációjával érték el. A módszer lényege, hogy a szövegeknek csak a szubjektív tartalmú mondatait használjuk fel, ezek alapján építjük a szeparációt végző modellt. A kétlépéses metódussal 86,4%-os eredményt értek el a korábban is említett angol mozikritika adatbázison [6]. Sajnos jelentős hátrány jelent, hogy a szubjektív mondatok kereséséhez nagyméretű példa mondatbázist kell felépíteni.

2 A korpusz

Véleményelemzés mindig egy jól behatárolható központi témakör köré épülő szövegvilág alapján történhet (témák pl.: politika világa, banki szolgáltatások, színházi előadások stb.). Külföldi gyakorlatot követve központi témakörnek a mozifilmek világát választottuk.

Az általunk épített polaritás adatbázis olyan magyar nyelvű kritikákat tartalmaz, amelyeknek a témája a műsorra tűzött mozifilmek tisztán szöveges tartalmi minősítése, nem pedig valamilyen meghatározott skála alapján vett kategorizálása (pl. „ötcsillagos” értékelés stb.). A szöveges értékelések döntő részét a port.hu, illetve az index.hu gyűjtőportál témát érintő moduljairól válogattuk össze. Az elemzési célnak megfelelően az összeállított tanuló-tesztelő polaritás adatbázis két kategóriából tevődik össze: egyik osztály a negatív (NEG), míg a másik a pozitív (POS) kritikákat tartalmazza.

A megépítendő korpusznak mennyiségi és minőségi kritériumoknak is eleget kellett tennie. A végső cél a korpusz méretét illetően az volt, hogy legalább 120 pozitív és ugyanennyi negatív kritikát fel tudjunk használni a módszerek vizsgálatához. A későbbi összehasonlítás reményében a szükséges anyagok összeválogatásánál törekedtünk a külföldi kutatásokban felhasznált angol nyelvű kritika-gyűjteményhez[2] hasonlatos korpusz felépítéséhez. Az általunk megfogalmazott minőségi kritérium szerint próbáltunk eleget tenni annak az elvárásnak, miszerint stílusában, méretében is olyan kritikákat válogassunk össze, mint amilyenek az angol nyelvű korpuszban is találhatók.

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/

Az összeválogatott korpusz összesen 240 kritikát tartalmaz. A korpusz kiegyenlített a két kategória méretét illetően, továbbá az egyes kritikák hossza átlagosan 250 szó. A korpuszt a stopszó[3] szűrés révén 29181 darab szóból álló szótár jellemzi.

A polaritás korpusz építése igen időigényes és fáradságos munkát igénylő folyamat. A kritikák kézzel történő annotációja mellett szűk keresztmetszetet jelentett számunkra, hogy viszonylag csekély számú forrásból gyűjthettünk mintákat, ugyanis kevés magyar nyelvű kritika hozzáférhető az interneten. A külföldi kutatóknak több lehetőségük volt korpuszépítésre, mivel jóval nagyobb angol nyelvű adatbázis állt rendelkezésükre, mint amilyen az imdb.com is.

3 Szózsák modell, fuzzy-IDF bevezetése

A jelenlegi számítási kapacitások szövegbányászati feladatok megoldását leginkább csak szózsák alapú dokumentum reprezentáció esetén teszik lehetővé. A vélemény analízis esetén fel kell vállalnunk azt a tetemes információ veszteségét, amit a szavak sorrendje tartalmaz.

A veszteségek mérséklésére alakították a különböző súlyozási sémákat, melyek különböző módon veszik figyelembe

- szó előfordulásának számát
- dokumentum méretét
- dokumentum csoport, korpusz számosságát

A legáltalánosabban használt mérték a TF-IDF, amely a szó-dokumentum mátrix egyes értékét a következő módon állítja elő:

A *TF(t,d)* kifejezés egy adott szó (*t*) előfordulási gyakoriságát adja meg a vizsgált dokumentumban (*d*):

$$TF(t,d) = \frac{c_{t,d}}{\sum_i c_{i,d}}$$

ahol $c_{i,d}$ (count) az i-edik szó előfordulásának száma a *d* dokumentumban. A kifejezésből adódik, hogy a súlyozás a dokumentumvektorokat egységnyi hosszúságra normálja.

Az IDF súlyozás csökkenti a korpuszban a nagyobb támogatottságú szavak súlyát, míg a kevesebb dokumentumban előforduló szavak súlyát növeli:

$$IDF(j) = \log\left(\frac{N}{DF(j)}\right)$$

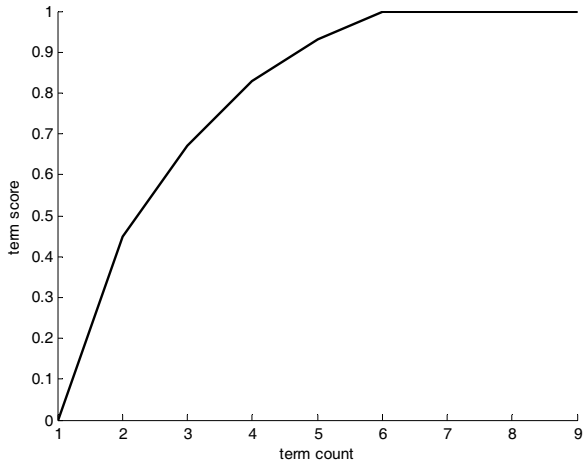
ahol *N a* korpusz dokumentumainak száma, míg *DF(j)* a j-edik szó *támogatottsága* a korpuszban (megadja, hogy a vizsgált szó hány dokumentumban szerepel).

Könnyen belátható, hogy más mértékben módosítja a dokumentum és a szó összetartozását, ha egyről kettőre nől az előfordulás, mint ha 5-ről 10-re. A TF súlyozás esetén ugyan az a hatás jön létre (duplázódik az érték). Ennek a megközelítésnek a logikájában vezettük be a fuzzy súlyozási sémát.

[3] A stopszó lista 743 darab szóból áll.

$$FS(j,d) = sigm(c_{j,d})$$

Az *FS(j,d)* kifejezés a j-edik szóhoz rendel értéket az ábrán látható módon.



1. ábra. Fuzzy súlyozási séma.

A sigmoid függvény telítődő jellege világosan kifejezi, hogy a vizsgált *d* dokumentumban az adott *j* szóhoz rendelt fontosság mértéke nem egyenesen arányos a szó dokumentumbeli előfordulásával (term count). A transzformáció praktikusan a kérdéses szó dokumentumhoz tartozásának erősségét fejezi ki a szóhoz rendelt fuzzy tagsági értékkel. Fuzzy reprezentációval szélsőséges esetben - sigmoid helyett lépésfüggvényt alkalmazva - a szavak *presence* információját kaphatjuk meg.

4 Robosztus Kockázat Minimalizáló alkalmazása

Újítással éltünk az osztályozó kiválasztásakor. A véleményelemzés témakörben ismereteink szerint eddig nem vizsgált megközelítést, a robosztus kockázat minimalizáció (RRM, Robust Risk Minimization) elvét alkalmaztuk a dokumentumok polaritás alapú osztályozásának megvalósítására [9][10][11][12].

Az elv gyakorlati jelentősége röviden összefoglalva abban áll, hogy az osztályozást meghatározó hipersík paramétereit egy *redukált keresési térre* korlátozva határozza meg az eljárás. Egy regularizációs paraméter révén szűkíthetjük a keresési tér méretét, ami egyúttal szabályozza a túltanulásra való hajlam érvényesülését is. Robosztusabbá válik tehát az osztályozó a tanulómintákon jelentkező túlilleszkedéssel szemben, amellett, hogy egyidejűleg minimalizálja a minták rossz osztályba sorolásának kockázatát.

Az általunk implementált algoritmus eltéréseket mutat a [11]-ben leírtaktól, mivel az értelmezése során az említett publikációban néhány cikkben – valószínűsíthetően elírást – fedeztünk fel. Sajnos a szerzőkkel nem sikerült felvennünk a kapcsolatot, de

az általunk megvalósított algoritmus eredményei igazolják az eljárásról alkotott elképzelésünk helyességét.

A fejezetben RRM rövid bemutatása mellett vázoljuk azt az értelmezést, amelyre támaszkodunk a dokumentumok polaritás alapú osztályozásában, illetve amely későbbi módosításaink alapját is képezi. A dolgozatban később ismertetésre kerülő kísérleteinkhez is a fejezetben ismertetett algoritmust és módosított változatait alkalmaztuk.

4.1 Az osztályozási feladat modellje

A kiindulási feladat megegyezik egy szokványosnak tekinthető szövegosztályozási feladattal.

A dokumentumainkat n elemű bináris feature (szó) vektorokkal ($\underline{x}=[x_1,..,x_n]$) reprezentáljuk, ahol n a szótár mérete, a szótárban lévő j-edik terminushoz (x_j) *bináris* értéket rendelünk aszerint, hogy a kérdéses szó előfordul-e a vizsgált dokumentumban vagy sem.

Feladatunk eldönteni azt, hogy a dokumentumvektor mely kategóriába tartozik. A becslést kizárólag annak alapján végezzük el, hogy a vizsgált dokumentum mely terminusokat tartalmazza, illetve melyeket nem. Az egyes osztályokat ($c \in C$) relevanciájuk szerint rangsoroljuk. A legrelevánsabb kategória (c^*) lesz a vizsgált dokumentum osztálya. A megoldást a maximum a posteriori hipotézis adja:

$$c^* = \arg\max_{c \in C} \{P(c \mid \underline{x})\}$$

A becslési feladat matematikailag a Naive-Bayes formulával fogalmazható meg, amely a dokumentumok szózsák alapú reprezentációjára épül. A terminusok előfordulásának bináris ábrázolásával a formulát átírhatjuk a következő formára:

$$\Pr(c \mid \underline{x}) = \frac{\Pr(c)\prod_j \Pr(x_j = 0 \mid c)}{\Pr(\underline{x})} \prod_j \left(\frac{\Pr(x_j = 1 \mid c)}{\Pr(x_j = 0 \mid c)} \right)^{x_j}, \quad c \in C, \quad x_j \in \{0,1\},$$

Ha vesszük a jobb oldali kifejezés természetes alapú logaritmusát, a következő formában is felírhatjuk az osztályozási problémát:

$$\Pr(c \mid \underline{x}) = \frac{1}{\Pr(\underline{x})} \exp\left(\overbrace{\sum_j w_j x_j + b}^{hipersík} \right)$$

A formulában felfedezhető a hipersík egyenlete, ahol a sík paramétereit a következő összefüggések adják:

$$w_j = \ln \frac{\Pr(x_j = 1 \mid c)}{\Pr(x_j = 0 \mid c)}$$

$$b = \ln \Pr(c) + \sum_j \ln \Pr(x_j = 0 \mid c)$$

A $\underline{w}=[w_1,...w_n]$ súlyvektor a tanulómintákból becsülhető, és az adott osztályozási problémát jellemző leíró. A súlyvektor w_j együtthatója azt fejezi ki, hogy az x_j szó

mennyire jellemző a vizsgált kategóriára. Annak az esélyét (odds[4]) fejezi ki, hogy a kérdéses szó a *c* osztályhoz tartozik. A fenti kifejezésekből lehetőségünk van visszavezetni a becslési feladatot a súlyvektor közvetlen meghatározásának problémájára. A szakirodalom arról a tapasztalatról számol be, miszerint Naive Bayes esetében jobb eredményeket lehet elérni a becslésben, ha nem max likelihood alapján számoljuk az osztályok relevanciáját, hanem közvetlenül a szavak súlyait próbáljuk meghatározni valamilyen lineáris döntési modellel [12]. A becslési feladat ilyen megközelítése a lineáris osztályozó módszerek egyik másik értelmezéséhez vezet: a lineáris súlyozás alapú módszerekhez.

4.2 Lineáris súlyozáson alapuló osztályozás – a dokumentum polaritása

Az osztályozási probléma újszerű megközelítésével tehát a szavak súlyozása révén kifejezhetjük, hogy a kategorizálás szempontjából milyen mértékben meghatározóak az egyes szavak. A stratégiát alkalmazhatjuk véleményanalízisre is.

Polaritás alapú osztályozás esetén koncepcionálisan két osztályt alakítunk ki: egyik halmazban a negatív (C_{neg}), míg másikban a pozitív (C_{pos}) véleményt hordozó kritikákat tároljuk. A koncepcióból eredően és (36) alapján tehát a szavakhoz rendelt súlyokra úgy tekintünk, hogy azok a szó által kifejezett vélemény orientációjának a mértékét fejezik ki. Az elgondolásunk alapján tehát a szóhoz rendelt súlyt megkapjuk:

$$w_j = \ln \frac{\Pr(x_j = 1 \mid C_{pos})}{\Pr(x_j = 0 \mid C_{pos})} \quad ahol \quad \Pr(x_j = 1 \mid C_{pos}) = \frac{\#d}{\left|C_{pos}\right|}$$

A kifejezésben $\#d$ azon pozitív kritikák számát adja meg, amelyekben az x_j szó előfordul. A súlyok előjele adott polaritású orientációt kapcsol a szóhoz, a súly nagysága a szó által kifejezett vélemény polaritásának erősségét fejezi ki. Az osztályozás során az ismeretlen polaritású dokumentum által képviselt eredő orientációt a dokumentum szövegében előforduló szavakhoz rendelt súlyok összegeként határozzuk meg:

$$polarity\_score(\underline{x}) = \sum_j w_j x_j + b = \underline{w}^T \underline{x} + b$$

A dokumentumra adott eredő súly polaritása határozza meg a teljes szöveg orientációját.

A fenti kifejezés rávilágít az osztályozási feladat egy más megközelítésű értelmezési lehetőségére, miszerint a vizsgált dokumentum alapján az egyes osztály címkék rangsorolása közvetlenül a dokumentumban lévő szavakhoz rendelt súlyok lineáris kombinációjával is meghatározható egy helyesen becsült $\underline{w}$ súlyvektor ismeretében. A feladatunk tehát az, hogy a tanulóminták alapján megbecsüljük a helyes döntéshez szükséges súlyvektort.

Korábbi fejezetben már ismertetésre került néhány lineáris döntési modellt megvalósító algoritmus. Korábbi kísérleteink azt támasztották alá, hogy érdemes regularizált

4 Odds: angol szakirodalomban terjedt el, jelentése: p/(1-p).

osztályozókkal kísérletezni a modell paramétereinek meghatározásában. A súlyvektor meghatározásához alkalmazott RRM algoritmus alapját T. Zhang és F. J. Oles által kidolgozott keretrendszer alkotja [9].

Az Information Retrieval folyóiratban megjelent tanulmányuk arra keresi a választ, hogy az SVM dokumentumok osztályozásában nyújtott teljesítménye vajon csak az SVM tervezés sajátossága-e, vagy talán alkotható egy olyan egységes matematikai keretrendszer („Regularized Linear Systems"), amelyet alkalmazva más lineáris osztályozók esetében is jó teljesítmény lenne elérhető. A keretrendszer meghatározza a regularizált osztályozási feladatok megoldásához vezető utat a feladat megfogalmazásától kiindulva. A megoldáshoz numerikus módszereket is ajánlanak, amelyek a szövegbányászat nagydimenziós terében képesek hatékonyan megoldani a feladatot.

4.3 Robosztus kockázat minimalizálás elve

Az RRM algoritmus a regularizált lineáris osztályozók csoportjába tartozik. A következőkben az algoritmus ismertetése mellett egyúttal betekintést nyújtunk a regularizált osztályozók alapjaiba is.

Az algoritmussal való ismerkedésünk kiindulási pontja az osztályozási feladat megfogalmazása. Felügyelt tanulási módszerről lévén szó tanulómintákat $\{(x, y)\}_{i=1}^{N}$ alkalmazunk a modellépítési fázisban. A tanulási feladatban a korpuszt alkotó dokumentumok vektoros reprezentációja ($\underline{x}_i$) bemeneti, míg a dokumentumokhoz rendelt osztály azonosítója ($y_i \in \{-1,+1\}$) kimeneti változóként jelenik meg. Az osztályozási feladat koncepcionálisan a következő kényszerekkel fogalmazható meg:

$$\left(\underline{w}^T \underline{x}_i + w_0\right) y_i > 0 \quad \forall i \quad (39)$$

$$\left\|\underline{w}\right\|^2 + w_0^2 \leq A \quad (40)$$

Ahol a (39) feltétel biztosítja, hogy minden $\underline{x}_i$ dokumentum megfelelően legyen osztályozva, míg a (40) regularizációból adódó kényszer korlátozza a lehetséges hipersík paraméterek keresési terének a méretét. A megfogalmazott feladat értelmében tehát keressük a lineáris döntési modell azon paramétereit ($\underline{w}, w_0$), amelyek kielégítik az előírt feltételeket.

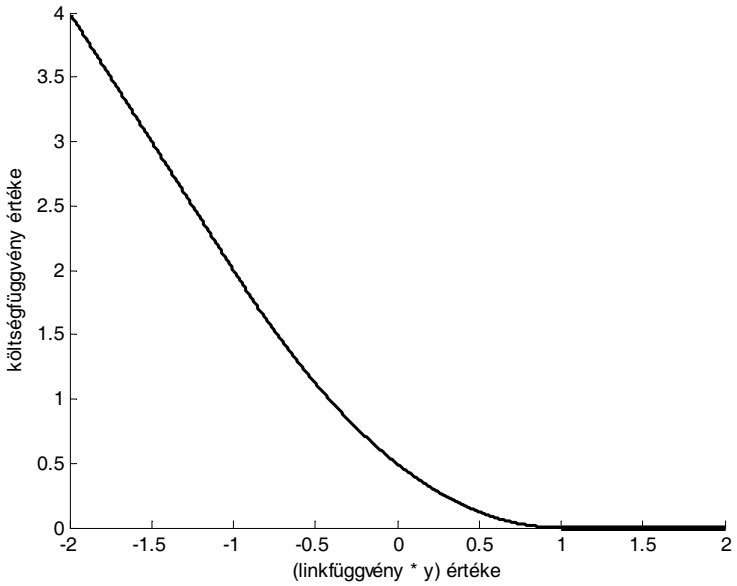
A keresési térben a legjobb modell megtalálásához költségfüggvényt alkalmazunk. A költségfüggvény révén matematikailag kezelni tudjuk az osztályozási problémát, ellenőrizhetjük a modellünk illeszkedését[5], azaz mérhetjük, hogy a modell milyen jól írja le a mintáinkat. Célunk a modell paramétereinek meghatározása úgy, hogy közben az általunk használt költségfüggvényt minimalizáljuk. A függvény optimalizálása általában valamilyen iteratív numerikus módszerrel végezhető, ahol a függvény mentén történő minimalizálás során történik a modellünk lépésenkénti finomítása. Az optimumot eredményező pontban kapjuk azt a modellt, amely a legjobban megfelel a függvény által kifejezett elvárásainknak (osztályozási hiba legyen minimális).

[5] Büntetjük a modell pontatlanságát.

Célunk meghatározni azokat a $(\underline{w}, w_0)$ paramétereket, amelyekre $y_i = \phi(\underline{w}, w_0, \underline{x}_i) \quad \forall i$, ahol $\phi(.)$ az úgynevezett „link függvény” (esetünkben a hipersík matematikai kifejezése). Adott modellparaméterek mellett az illeszkedés mértékét a $L(.)$ költségfüggvénnyel (loss function)[6] határozzuk meg. Numerikus okokból kifolyólag a link függvényt úgy választjuk meg, hogy a keletkező költségfüggvény $L\left(\phi(\underline{w}^T, w_0, \underline{x}_i), y_i\right)$ konvex legyen.

A költségfüggvény révén büntetjük az osztályozás hibáját. A súlyvektort a bemeneti mintákból határozzuk meg az illeszkedés *várható* költségének minimalizálása mentén. A megoldáshoz a következő költségfüggvényt használjuk RRM esetén [12]:

$$L\left(\phi(\underline{w}^T, w_0, \underline{x}_i), y_i\right) = \begin{Bmatrix} -2\phi(\underline{w}^T, w_0, \underline{x}_i) y_i & ha & \phi(\underline{w}^T, w_0, \underline{x}_i) y_i < -1 \\ \frac{1}{2}\left(\phi(\underline{w}^T, w_0, \underline{x}_i) y_i - 1\right)^2 & ha & \phi(\underline{w}^T, w_0, \underline{x}_i) y_i \in [-1,+1] \\ 0 & ha & \phi(\underline{w}^T, w_0, \underline{x}_i) y_i > 1 \end{Bmatrix} \quad (41)$$



2. ábra. Robosztus költségfüggvény.

A mintákra illeszkedő optimális döntési modell paramétereit a robusztus költségfüggvény minimumában kapjuk meg.

6 Magyar szakirodalomban a veszteségfüggvény elnevezés is használatos.

4.4 RM algoritmus pszeudokódja

Előfeldolgozás: bináris szó-dokumentum mátrix generálása a korpuszból

$$\underline{x}^j = [x_1^j,...,x_m^j] \quad x_k^j = \begin{cases} 0 & \text{k. szó} \notin \text{j. dokumentum} \\ 1 & \text{k. szó} \in \text{j. dokumentum} \end{cases}$$

Bemenet: tanuló minták $(\underline{x}^1, y^1),.....,(\underline{x}^N, y^N)$

Paraméterek: K, A, η

Kimenet: súlyvektor $\underline{w} = [w_1,....,w_m]$, w_0

Inicializálás: $\alpha_i = 0$ (i = 1...N), $\underline{w} = \underline{0}$, $w_0 = 0$

for k = 1 to K **do**
  for i = 1 to N **do**
    p = $(\underline{w}^T \underline{x}^i + b) y^i$
    gradient$_i$ =
      $\max(\min(2A - \alpha_i, \eta((A - \alpha_i)/A - p)), -\alpha_i)$ (*)
    $\underline{w} = \underline{w} + \text{gradient}_i \underline{x}^i y^i$
    $w_0 = w_0 + \text{gradient}_i y^i$
    $\alpha_i = \alpha_i + \text{gradient}_i$
  end for
end for

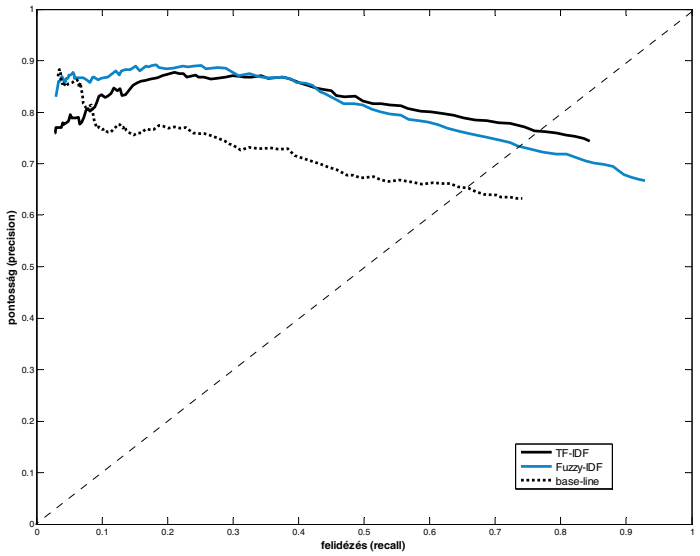
3. ábra. Az RRM algoritmus pszeudokódja.

A pszeudokódban előforduló változók jelentése a következő: *K* paraméter az iteratív algoritmus maximális lépésszáma, *A* paraméter a keresési tér méretét korlátozza ($A = \frac{1}{\lambda N}$), a η paraméter a tanulási ráta, amely az iteráció során a gradiens irányába tett lépésünk nagyságát határozza meg. Az online módszer értelmében a duális α_i változók egy-egy tanulóminta párashoz kapcsolódnak. A gradiens kifejezésénél (*) a maximálás, illetve a minimálás biztosítja, hogy a duális változó az előírt $\left[0, \frac{2}{\lambda N}\right]$ intervallumban maradjon. Ha egy bemeneti $\underline{x}_i$ mintához tartozó duális változó értéke meghaladja az előírt intervallumot, akkor a változót nullázzuk.

5 Eredmények

Méréseink elsősorban arra irányulnak, hogy megvizsgáljuk a szó-dokumentum mátrixon alkalmazható különböző súlyozási séma osztályozás pontosságára gyakorolt hatását az RRM esetén. A tanuló modellt a korábbi fejezetben felsorolt sémák alapján súlyoztuk, majd az algoritmus korpuszra hangolását és tanítását követően 50 mérés-

ből álló tesztsorozat átlagából meghatároztuk az alábbi ábrán látható felidézés – pontosság görbéket.



4. ábra. Felidézés - pontossággörbék magyar korpuszon

Vizsgálataink fókuszában elsősorban a zérus döntési küszöbhöz tartozó felidézés - pontosság értékpárok álltak, vagyis az osztályozás kiértékelését az *összes dokumentumra* hozott döntés figyelembevételével végeztük. Ezen értékpárokat mindig a vizsgált súlyozási sémára vonatkozó görbe maximális felidézés értékéhez tartozó pontosság értékének kettőse alkotja.

Az eredményül kapott görbékből leolvasható, hogy a base-line módszer (pontozott vonal) alkalmazása esetén közel 64%-os pontosság mellett mindössze a POS kritikák 74%-át találtuk meg. Az alacsonyabb felidézési érték mellett az elért pontosság kedvezőtlen hatású az osztályozás találati arányára nézve. Gyakorlatilag azt jelenti, hogy a becslés során a negatív kritikák egy jelentős részét is hibásan becsülte az algoritmus, miközben a pozitív kritikák közel háromnegyedét becsülte csak helyesen.

Az ábrán az is látható, hogy a base-line módszer felidézésén a szó-dokumentum mátrix Fuzzy-IDF súlyozásával nagymértékben sikerült javítani. Az ábrázolt görbe alapján megállapítható, hogy a tesztmintákban a POS kritikák 92,8%-át megtaláljuk közel 67%-os pontosság mellett. Fuzzy-IDF súlyozás alkalmazásával a tesztdokumentumokon elérhető pontosság értéke szinte változatlan maradt ugyan, de a nagyobb felidézés érték azt igazolja, hogy képesek vagyunk szinte az összes pozitív kritikát megtalálni a mintahalmazban.

Az ábra tanulsága szerint a szó-dokumentum mátrixon kipróbált súlyozási módszerek közül a TF-IDF súlyozás bizonyul a leghatékonyabbnak. TF-IDF súlyozású mátrixon tanított algoritmus képes arra, hogy 74,34%-os pontosság mellett megtalálja a

teszthalmazban a POS kritikák több mint 84%-át. Megfigyelhető azonban, hogy az osztályozó pontossága alacsony felidézés mellett csökkent, ami arra enged következtetni, hogy a szeparáló felülettől távol lévő dokumentumok címkéjét pontatlanabbul becsüli az algoritmus.

Az eredmények alapján megállapítható, hogy a tanuló modell különböző súlyozásai révén sikerült javítani az eredeti base-line módszer magyar korpuszon elérhető hatékonyságán. Azt a következtetést vonhatjuk le, miszerint a különböző súlyozási konvenciókkal minden esetben pozitív irányban befolyásoltuk az algoritmust: a Fuzzy-IDF súlyozás hatására hasonló pontosság mellett jobb felidézést értünk el mint a base-line módszer. A tanuló modell TF-IDF súlyozása nagyban javít mind az elérhető pontosságon, mind a felidézésen, továbbá egyben a legnagyobb BEP értékű osztályozót eredményezi.

6 Összefoglalás

Korábbi munkák során az általunk készített magyar filmkritika korpuszon megvizsgáltunk több osztályozó módszert. Ezek illetve a legjobban teljesítő, az e tanulmányban ismertetett RRM módszer eredményeit mutatja az 1. táblázat.

Külön vizsgáltuk a „nem" jelentésmódosító hatását. Pl. „nem jó" szereplése esetén a „nem" stop szót figyelmen kívül hagyjuk és csak a „jó" kerül be a szózsák modellbe. A táblázat utolsó sorában szereplő „NOT TAGGING" esetben a „nem" szó és az általa módosított szó együtt képez tókent.

1. táblázat: Eredmények magyar korpuszon.

	Naive Bayes	Percep-tron	neurális hálózat	SVM	RRM
helyes osztályozási arány	0.63	0.65	0.697	**0.715**	**0.76**
helyes osztályozási arány (NOT-TAGGING)	–	0.645	0.662	0.703	–

A NOT-tagging módszer láthatóan nem segíti a magyar filmkritika korpusz véleményanalízisét.

A magyar nyelvű véleményanalízisre újszerűen alkalmazott RRM az általunk javasolt fuzzy-IDF súlyozással jelentős javulást hozott az eddigi legjobb SVM-mel szemben is.

Módszerünk az angol korpuszon 78,8%-ot ér el, ami hasonló az 1.1-ben olvasható eredményekhez.

Hivatkozások

1. Hatzivassiloglou, V., McKeown, K.R.: Predicting the Semantic Orientation of Adjectives. In Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL. Madrid, Spain, July 1997. Association for Computational Linguistics (1997) 174–181
2. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS) 2003, 21 (4) (2003) 315-346
3. Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss analysis. In: Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, Bremen, DE (2005) 617-624
4. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002 (2002)
5. Na, J-C., Khoo, C., Horng Jyh Wu, P.: Use of negation phrases in automatic sentiment classification of product reviews. Library Collections, Acquistions & Technical Services, 29 (2005) 180-191
6. Pang, L., Lee, A.: Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of ACL 2004 (2004)
7. Damerau, F. J., Zhang, T., Weiss, S. M., Indurkhya, N.: Text categorization for a comprehensive time-dependent benchmark. Information Processing & Management, 40 (2004) 209-221
8. Berend, G., Farkas, R.: Opinion mining in Hungarian based on textual and graphical clues. In: Proceedings of the 4th Intern. Symposium on Data Mining and Intelligent Information Processing, Santander (2008)
9. Zhang, T., Oles. F. J.: Text categorization based on regularized linear classification methods. Information Retrieval, 4 (2001) 5-31
   http://www-cs-students.stanford.edu/~tzhang/papers/ir01_textcat.pdf
10. Zhang, T.: On the dual formulation of regularized linear systems. Machine Learning 46 (2002) 91-129 http://www-cs-students.stanford.edu/~tzhang/papers/ml02_dual.pdf
11. Damerau, F. J., Zhang, T., Weiss, S. M., Indurkhya, N.: Text categorization for a comprehensive time-dependent benchmark. Information Processing & Management, 40 (2004) 209-221 http://www-cs-students.stanford.edu/~tzhang/papers/ipm04-new_reuters.pdf
12. Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.: Text Mining - Predictive Methods for Analyzing Unstructured Information. Springer, ISBN: 978-0-387-95433-2 (2005)
13. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publisher Inc., ISBN: 978-1-60198-150-9 (2008)