Acta Agraria Kaposváriensis (2011) Vol 15 No 3, 103-113
Kaposvári Egyetem, Állattudományi Kar, Kaposvár
Kaposvár University, Faculty of Animal Science, Kaposvár

Tájolás és hasznosság mérése rövid szöveges üzenetekben

Kovács B., Kruzslicz F.
Pécsi Tudományegyetem, Közgazdaságtudományi Kar, Gazdaságmódszertani Intézet, 7622 Pécs, Rákóczi út 80.

ÖSSZEFOGLALÁS

A tanulmányban a szövegbányászat egy speciális ágával, a véleménybányászattal foglalkozunk. A vélemény tulajdonosa hisz egy témával kapcsolatos állításban, és ehhez általában jó vagy rossz érzést is társít. A vélemény negatív, pozitív vagy semleges tájolását a tulajdonos szöveges hozzászólásaiból (üzeneteiből) lehet kideríteni. A hozzászólás hasznossága az üzenetet olvasók számára azt mutatja meg, hogy az adott üzenet mekkora hatással lehet az olvasó véleményére. Tehát minél hasznosabb egy üzenet, annál nagyobb eséllyel befolyásolja a fizetőképes keresletet egy termék iránt. Így tehát a hasznosság annak mértéke is, hogy mekkora súllyal kell kezelni az üzenetet a számítások során. A rövid szöveges értékelések tájolási és hasznossági értékeinek mérési és előrejelzési lehetőségeit kétféle módszertani megközelítéssel is megvizsgáltuk: a szupport vektor gépek (SVM), és a mesterséges neurális hálózatok (NN) tanulóalgoritmusok segítségével. Elemzéseink során azt tapasztaltuk, hogy a tájolás és a hasznosság szerinti osztályozás hatékonysága eltér egymástól, ami arra utal, hogy a kétfajta jellemző más-más jellegű kapcsolatban áll a szöveges tartalommal. Ezért annak lehetőségét is megvizsgáltuk, hogy a szövegelemzés során használt szokásos adatok mellett milyen egyéb jellemzők bevonásával lehet javítani az osztályozás pontosságát.
(Kulcsszavak: nyelvfüggetlen, szövegelemzés, osztályozás, véleménybányászat)

ABSTRACT

Measuring sentiment orientation and utility in short text messages

B. Kovács, F. Kruzslicz
University of Pécs, Faculty of Economics, Institute of Applied Business Studies, H-7622 Pécs, Rákóczi út 80.

The main framework of this paper is the opinion mining, a special area of text mining. The holder of an opinion believes a claim about a topic, and he associates a good or a bad sentiment with his belief. This negative, positive or neutral orientation of the opinion can be detected based on the text messages of the holder. The utility of a comment can be described as the magnitude of the effect of the message on the readers' opinion. E. g. if a message has higher utility, it has a greater chance to influence the effective demand for a product. So utility is a measure of weighting of messages in calculations. We examined the orientation and utility valuation of short text messages through two methodological approaches: Support Vector Machines (SVM), and Artificial Neural Networks (NN) learning algorithms. We find that the efficiency of the automatic classification of orientation and utility values differs from each other using both methodologies. So we concluded that both of these measures represent different type of connection with the content of the texts. So we examine the opportunity of improving accuracy of classification by including additional measures.
(Keywords: language-independent, text analysis, classification, opinion mining)

BEVEZETÉS

Napjainkban az egyénre szabott kínálat nyújtása jelenti az igazi versenyelőnyt. A marketingkoncepció e szintjén már a valószínűségi modelleken alapuló statisztikai módszerek sem képesek mindig elegendő információt szolgáltatni a marketing mix kialakításához. Magyarországon is ezért egyre nagyobb szerep jut a különböző adat- és szövegbányászati módszereknek a piackutatásban.

A véleménybányászat a szövegbányászatnak egy szelete, melynek során a szövegekhez, vagy azok részeihez szubjektív érzéseket kifejező címkéket rendelünk. Ennek megfelelően ez a problémakör a szövegosztályozási feladatok körébe tartozik, melynek két fő típusát különböztetjük meg: (1) a dokumentum tartalmához rendelhető pozitív vagy negatív érzelmi megnyilvánulás, (2) a dokumentum tartalmához rendelhető szubjektivitás vagy objektivitás szintje. Az első típus tájolási, míg a második pártatlansági problémaként ismert az irodalomban. További érdekes kapcsolódó kutatási irányokat ismerhetünk meg *Liu* (2010) tanulmánykötetének 26. fejezetéből. Munkánk során az eddigi kutatási palettát egy olyan új elemmel bővítettük, amelyik a dokumentumokhoz az előzőekhez képest egy újfajta jellemzőt, a hasznosságot rendeli hozzá (*1. táblázat*).

Egy dokumentum hasznossága egy olyan aggregált mutató, ami egyszerre jelenti a dokumentum tartalmának újdonságértékét, a benne foglalt információk aktualitását, valamint azt, hogy egy majdani visszakereséskor a dokumentum hányadik helyre kerüljön a találati listában a relevanciája szerint. Míg a tájolás és pártatlanság fogalma inkább a tartalomhoz (objektum) kötődik, addig a hasznosság inkább az olvasóhoz (szubjektum) köthető. Ahogyan viszont ugyanaz az információ más-más felhasználó számára lehet pozitív és negatív is, addig hasonló kettősségre számíthatunk a hasznosság tekintetében.

1. táblázat

Néhány takarmány premixhez kötődő szöveges üzenet és annak besorolása

Termék (1)	**Hozzászólás (2)**	**Tájolás (3)** 1 – n egatív (5), 5 – pozitív (6)	**Hasznosság (4)** 1 – haszontalan (7), 5 – hasznos (8)
Sano	Beltartalmszerint jó! De Ár/érték aránya az rossz! Oltsobban lehet ugyan olyan jót, vagy jobbat is kapni!	1	4
Vitafort	A Vitafortnak is van jobb és rosszab terméke is. Beltartalmi paraméterei és ára alapján tudnék pontos véleményt mondani. ÁR/ÉRTÉK arány.	3	1
Kaluterm	Régebben az 1ik ismerősöm már próbálta -és azt mondta , h nem rossz.Igaz a hízóknak már nagyon jót adni szinte vétek. :D Amugy pedig van aki már 70 kg körül visszaveszi a fehérjét(szóját) B=)	4	3

Table 1: Sample text messages with classification values related to feed premixes

Product(1), Comment(2), Orientation(3), Utility(4), Negative(5), Positive(6), Useless(7), Useful(8)

A véleménybányászat során tájolandó dokumentumokkal kapcsolatosan általában semmilyen megkötést nem teszünk fel. A vizsgálataink során kizárólag fórumokhoz, blogokhoz, hírekhez vagy termékismertetőkhöz fűzött megjegyzéseket használtunk, de a módszerünk tetszőleges egyéb dokumentumtípusra alkalmazható, nemcsak ilyen rövid szöveges hozzászólások esetén. Az Internetről származó, angol nyelvű rövid szövegek ilyen jellegű elemzésével először *Kushal et al.* (2003) cikkében találkozhatunk, ahonnan egyébként a véleménybányászat kifejezés is származik. Aszerint, hogy a szöveg egészéhez, vagy annak csak egyes részeihez rendelünk tájolás és/vagy hasznossági értékeket, megkülönböztetünk dokumentum szintű, bekezdés szintű, mondat szintű és kifejezés szintű hozzárendelési módszereket. A rövid szöveges üzenetek feldolgozása esetén értelemszerűen elegendő a dokumentum szintű hozzárendeléssel foglalkozni.

A véleménybányászathoz kapcsolódó, ezredforduló előtti eredmények kitűnő összefoglalóját találjuk *Pang és Lee* (2008) munkájában, ahol tájolási osztályozáshoz számos felügyelt, illetve felügyelet nélküli módszert is bemutatnak a szerzők. Vizsgálataink során a felügyelt tanulási módszereket alkalmaztuk, ezért előzetes feladatként egy jó minőségű tanító és teszt adathalmaz előállítását kellett megoldani. Bár a futtatások és számítások során magyar nyelvű adatokat használtunk fel, egyik módszer sem feltételezi a nyelv ismeretét, és kellő méretű tanító halmaz előállításával akár nyelvfüggetlen tájolás és hasznosság predikció is megvalósítható. Elsősorban tehát arra voltunk kíváncsiak, hogy a tájolás és a hasznosság fogalmában rejlő különbség kimutatható-e valamiképpen a kiválasztott tanuló algoritmusok tanulási és előrejelzési képességeinek vizsgálatával, illetve hogyan teljesítenek ugyanezen eljárások, ha használatuk előtt a szövegeken különböző előfeldolgozást (szavakra bontást, stopszavazást, szótövezést stb.) végzünk.

ANYAG ÉS MÓDSZER

A tanító adathalmazt különféle termékismertető oldalak hozzászólásainak begyűjtésével (crawling) állítottuk elő. A webes tartalomnál elérhető egyéb attribútumok (például szerző, dátum, blokk, formázás stb.) ugyan kigyűjtésre kerültek, de a kutatás jelen fázisában még nem kerültek felhasználásra. A dokumentumokhoz a tájolási és hasznossági értékek manuálisan lettek hozzárendelve, mindkettő esetében egy 5 fokozatú skálán. Tájolás esetében az egyes érték osztályába kerültek az abszolút negatív hozzászólások, és az ötös értékbe az abszolút pozitívak. Így a hármas értéknek a semleges besorolás feleltethető meg. Hasznosság esetében hasonlóan az egyes értékkel jellemeztük a teljesen hasznavehetetlen hozzászólásokat, míg az ötös osztályzatot a leginkább informatív megjegyzések kapták.

A minél pontosabb besorolás érdekében ezt a műveletet több felhasználó is végezte, így ugyanazon dokumentum besorolása több szempontból is a rendelkezésünkre állt. Mivel ez felettébb időigényes és költséges feladat, ezért a termékek körét egyetlen fajtára korlátoztuk, és azon belül is csak pár terméktípushoz kapcsolódó hozzászólásokat vettünk bele a feldolgozásba. Az ezer hozzászólás kézi besorolásával eredményül kapott adathalmaz vizsgálatakor azt figyeltük meg, hogy az egyes osztályok elemszámai jelentősen eltérnek egymástól. Tájolás és hasznosság tekintetében is túlnyomó többségben a négyes osztályba történő besorolások voltak dominánsak.

A predikciós modelljeinket minden esetben ugyanezen a halmazon végzett háromszoros keresztvalidációval ellenőriztük oly módon, hogy az algoritmusokat az adatok kétharmadán tanítva, a kapott modell eredményességét a tanítás során fel nem használt egyharmadon ellenőriztük. A modellek előrejelzésének pontosságát a hibaaránnyal mértük, amely a félrekategorizált mintaelemek számának és a teljes mintaelemszám hányadosaként adódik. A mérőszám komplementer viszonyban áll az

ún. találati aránnyal, amelyet idősorok előrejelzésénél is szoktak használni az előjel-predikció eredményességének mérésére is.

Walczak (2001) pénzügyi jellegű tanulmányának tapasztalatai alapján a találati arányt akár 60%-ra is lehetett növelni, ami a jelen tanulmányban alkalmazott hibaarány szempontjából a 40%-os célkitűzést jelenti. Annyival mégis ki kell egészíteni az előbb megfogalmazott célt, hogy az előjel-predikció során 3 lehetséges kategóriát különböztettek meg, jelen tanulmány viszont 5 kategóriás osztályozásra vállalkozik, ami nehezíti a cél elérését. Minél kevesebb kategóriával kell ugyanis dolgozni, annál nagyobb találati arány érhető el – például 2 kategória esetén *Kryzanowski et al.* (1993) 72%-os találati arányt értek el (szintén pénzügyi idősoron).

Az algoritmusok implementálására a Python nyelvű fejlesztő környezet *Fan et al.* (2005) LIBSVM szupport vektor gép, *Nissen* (2003) FANN neurális hálózatok, valamint *Hanke at al.* (2009) PyMVPA automatikus validálási moduljait használtuk.

Modellezés mesterséges neurális hálózatokkal

Ebben a részben bemutatjuk az általunk alkalmazott neurális topológiák felépítését. A bemutatás során használt fogalmak értelmezéséhez ajánljuk *Borgulya* (1998) monográfiáját.

A neurális hálózatok vizsgálatánál két topológiát vizsgáltunk. Az egyik topológia lineáris hipersíkokkal történő osztály-szeparációkra képes, a másikkal pedig nemlineáris szeparációkat is végre lehet hajtani. Az előbbi topológiát a tanulmányban lineáris osztályozó topológiának nevezzük (linear classifier, LC). Az utóbbit nemlineáris osztályozó topológiának (non-linear classifier, NLC) nevezzük a tanulmányban, illetve helyenként használjuk a topológiaosztály megnevezésének rövidítését: MLP (multilayer perceptron, többrétegű perceptron).

Az általunk használt lineáris osztályozó topológia egy lineáris jelzési függvényű inputréteget (102 neuron) és egy tangens-hiperbolikus jelzési függvényű outputréteget (5 neuron) tartalmaz. *Tikk* (2007) könyvében bemutatott bináris osztályozó topológiájának továbbfejlesztett modelljéről van tulajdonképpen szó, amely már az életlen (fuzzy) halmazok elméletéhez közelebb áll.

Az alkalmazott MLP topológia pedig egy lineáris jelzési függvényű inputrétegből (102 neuron), két tangens-hiperbolikus jelzési függvényű rejtett rétegből (aktiválási sorrendben: 10 neuron, ill. 5 neuron), valamint egy tangens-hiperbolikus jelzési függvényű output rétegből (5 neuron) épül fel.

A korpuszban található szótövezett szavak közül a reprezentáció csak a 30-nál gyakoribbakat foglalja magában. (Tulajdonképpen tehát a stopszavazást kibővítettük ezekre a szavakra is). Ezzel azt kívántuk elérni, hogy a modell elegendő számú megfigyelés birtokában rendeljen csak tájolási értékeket a korpusz szavaihoz. A kritériumnak 101 szó tett eleget (szótövezett, stopszavazott).

A lineáris jelzési függvényű inputrétegekben tehát 101 neuron feleltethető meg ennek a 101 szónak, a maradék egy neuron pedig minden dokumentumban 1-es súlyt kap – ez technikai jellegű változó, amelynek a feladata, hogy a membrán-aggregációk értékéhez konstans értékkel járuljanak hozzá (bias). Ilyen módon a tangens-hiperbolikus jelzési függvény vízszintes (abszcisszával párhuzamos irányú) eltolása helyett a fent említett technikai változóból kiinduló szinapszis-súlyok értékének változtatásával befolyásolhatjuk az tangens-hiperbolikus jelzési függvény aktivációs értékét.

A topológiák output-rétegeiben a neuronokhoz kategóriák rendelődnek, melyek száma mind a tájolás, mind a hasznosság esetében öt-öt.

A 30-nál kevesebbszer előforduló szavak elhagyásával viszont el kell hagynunk olyan dokumentumokat, amelyek kizárólag ilyen szavakat használtak, így a neurális

hálózat tanítására felhasználható minta mérte 657 darab dokumentum lett, a tesztminta mérete pedig 330 dokumentum. (Így tartva a körülbelül kétharmad-egyharmad arányt.)

Az output neuronok a tangens-hiperbolikus jelzési függvényük miatt -1 és 1 közötti értéket képesek visszaadni, így tanítható a hálózat 1 (kategóriához tartozás) és -1 (a kategória komplementeréhez való tartozás) tanítóadatok segítségével. Ennek alternatívája az előjel jelzési függvény, amely azonban a tangens-hiperbolikussal szemben szakadásos függvény. A tangens-hiperbolikus függvény segítségével azonban életlen hozzátartozások könnyebben modellezhetők. A tangens-hiperbolikus jelzési függvény nemlineáris számításokban való alkalmazására mutat példát *Cotton és Wilamowski* (2011) friss tanulmánya is.

A jelen tanulmányban bemutatott két neurális hálózat algoritmusa azt tanulja, hogyan lehet megállapítani egy dokumentum (hozzászólás) tájolását, illetve hasznosságát.

A tanítóadatoknak és a jelzési függvényeknek a fenti módon történő megválasztásának a következménye, hogy az output neuron membránaggregációjának pozitivitása – és így a kategóriához való tartozás – a bemenő szinapszisok súlyainak pozitivitásától függ.

A megfelelő topológiák kiválasztásához előzetesen 4, egymástól a rejtett rétegek számában eltérő topológiát különböztettünk meg. Az első topológia a lineáris osztályozó volt, amelynek tulajdonképpen nincs rejtett rétege (inputréteg szélessége: 102; ouptutréteg szélessége: 5). A következő topológia az 1 rejtett réteget tartalmazó MLP topológia (továbbiakban: MLP1; inputréteg szélessége: 102; rejtett réteg szélessége: 5; outputréteg szélessége: 5). A 2 rejtett rétegű MLP topológia (MLP2) paraméterei: inputréteg szélessége: 102; első rejtett réteg szélessége: 10; második rejtett réteg szélessége: 5; outputréteg szélessége: 5. A 3 rejtett réteggel rendelkező topológia (MLP3) paraméterei: inputréteg szélessége: 102; első rejtett réteg szélessége: 10; második rejtett réteg szélessége: 5; harmadik rejtett réteg szélessége: 5; outputréteg szélessége: 5.

A topológiákat sztenderd módon hasonlítottuk össze előzetes futtatások során. A tanulási ráta 0,01, az epochszám 150-es paraméterbeállításai mellett a háromszoros keresztvalidáció során kapott minimális tájolási hibaarányú topológiát kerestük további elemzés céljából, az így megjelölt topológiának és a lineáris osztályozó topológiának az összehasonlítására szűkítve le az elemzést. (Amennyiben a minimális hibaarányú topológiának éppen a lineáris osztályozó bizonyult volna, akkor a második legkisebb hibaarányú topológiával dolgoztunk volna tovább.)

Az előzetes vizsgálat eredményét a *2. táblázatban* foglaltuk össze.

2. táblázat

A tájolási értékek becslésének hibaaránya különböző topológiák esetén

	MLP3	**MLP2**	**MLP1**	**LC**
Tanítóminta (1)	10.83%	7.93%	14.19%	24.42%
Tesztminta (2)	57.44%	54.71%	56.63%	56.83%

Table 2: Prediction error ratios of sentiment orientation values using several topologies

Training sample(1), Test sample(2)

A *2. táblázatban* látható eredmények miatt tehát az MLP2 topológia került kiválasztásra a lineáris osztályozó topológia mellé.

Az Eredmény és értékelés részben bemutatjuk, hogyan teljesített a kétféle topológia a dokumentumok tájolási és hasznossági értékeinek becslésében. A becslés jóságát a

tesztmintán elért (keresztvalidáció) hibaaránnyal mértük. Az eredményeket 0,01 tanulási rátával kaptuk. A vizsgálat az epochszám megválasztásának hatását úgy küszöböli ki, hogy az epochszám viszonylag széles tartományában figyeltük meg a topológiák tesztmintán elért hibaarányát.

Modellezés szupport vektor gépekkel

Szövegbányászat esetén a szupport vektor gépek alkalmazása igen kedvelt eljárás, hiszen ez az eljárás sokdimenziós terekben is gyorsan képes az egyes osztályokat elválasztó hipersíkok megtalálására. Mind a tájolás, mind pedig a hasznosság esetében a megfelelő öt osztály egyikébe történő besorolással jellemeztük az adott dokumentumot, ami többcímkés osztályozási feladatot jelent. Az eredeti SVM eljárás azonban csak egycímkés (bináris) osztályozásra alkalmas. Az egyik megoldást a bináris osztályozók kombinációja jelentheti, például az egyszerű többségi szavazás alkalmazásával. A másik megoldás az, ha az SVM gépek működését leíró kvadratikus optimalizálási modellt módosítjuk úgy, hogy a szeparáló hipersíkokat úgy próbáljuk meg meghatározni, hogy az egyes osztályoktól vett távolságokat egyszerre próbáljuk meg minimalizálni. *Qin és Wang* (2009) munkája nyomán több lehetséges megvalósítást hasonlíthatunk össze, melyek közül az egyik leggyorsabb, lineáris magfüggvényű változatot választottuk ki. Az általunk használt Crammer-féle M-osztályos módosított modell leírása tehát az alábbi *Crammer et al.* (2001):

$$\text{Célfüggvény:} \quad \min(\mathbf{w},\xi) = \tfrac{1}{2} \cdot \sum_{m=1..M} \mathbf{w_m}^T \cdot \mathbf{w_m} + C \cdot \sum_{i=1..l} \xi_i \qquad (1)$$

$$\text{Korlátozó feltételek:} \quad \mathbf{w_{y_i}}^T \Phi(\mathbf{x}_i) - \mathbf{w_m}^T \Phi(\mathbf{x}_i) \geq 1 - \delta(y_i, m) - \xi_i \quad i=1..l, \text{ és } \xi_i \geq 0 \qquad (2)$$

Ahol l a dokumentumok, M pedig az osztályok számát jelöli, $\mathbf{w}$ pedig a szeparáló hipersíkok normálvektora. A Φ tetszőleges kernel függvény lehet, melyek közül mi a lineáris változatot választottuk. A δ az úgynevezett Kronecker delta függvény, amelyik az argumentumok egyenlősége esetén 1-et, különben pedig 0-át ad eredményül.

Az $\mathbf{x}_i$ tanító vektorhoz tartozó címkét y_i-vel jelölve egy ismeretlen címkéjű x vektor esetén az alábbi döntési függvénnyel rendeljük hozzá a fenti modell által becsült osztálycímkét:

$$\text{Predikció:} \quad \text{argmax}_{m=1..M} \left(\mathbf{w_m}^T \Phi(\mathbf{x}) \right) \qquad (3)$$

EREDMÉNY ÉS ÉRTÉKELÉS

A minta korpuszban található dokumentumból az előfeldolgozottsági állapotuknak megfelelően négyféle változat készült az elemzésekhez. Az első változat az eredeti szöveg szóközöknél tagolt (nyers) felbontása volt. A második az írásjelek figyelembevételével vett szóelemekre bontás (tokenizált) változat, a harmadik lépésben már a töltelék és lényegtelen szavaktól mentes (stopszavazott) változat volt, melyet a végén a szótövezett változat zárt. A stopszavazásnak és a szótövezésnek létezik nyelvfüggetlen megvalósítása is, de ezen lépések egy automatikus nyelv detektálás után hatékonyabban is elvégezhetőek.

Minden szövegváltozatot miden modell esetében a szózsák modell alkalmazásával transzformáltunk a tanulási módszerekhez alkalmas a szó-dokumentum mátrix (term-document matrix, TDM) adatszerkezetté. A neurális hálózatok esetében a TDM mátrix elemei a bináris súlyozásnak megfelelően az adott szó dokumentumba való beletartozását jelölik, míg a szupport vektor gépek esetén az adott dokumentumba való tartalmazás információ tartalmát mérő entrópia súlyozást használtuk.

A tájolási értékek becsülhetősége mesterséges neurális hálózatokkal

Az optimálisan betanított lineáris osztályozó esetében az inputréteg és az outputréteg közötti szinapszisok (kivéve a technikai változó) súlyai azt mutatják meg, hogy az adott

szó milyen hatással van a hozzászólás tájolására (tájolási adatok esetén). Amennyiben egy szónak megfelelő input neuron és egy kategóriának (tájolási mértéknek) megfelelő output neuron közötti szinapszis negatív, akkor az adott szó tájolása nem áll közel az adott kategóriának megfelelő tájolási értékhez. Amennyiben a szóban forgó szinapszis súlya pozitív, akkor a szó tájolása közel áll az adott kategóriához. Az elmondottak alapján tehát előfordulhat, hogy egy szónak nincs igazán tájolása, illetve az is, hogy többféle tájolása is lehet – nyilván a kontextustól függően.

A hibaarány változása az epochszám függvényében eltérően alakul a két vizsgált topológián a tájolási értékek esetében. Az *1. ábrán* a folytonos vonal mutatja a lineáris osztályozó hibaarányát a keresztvalidáció során, az abszcissza tengelyen az epochszám látható logaritmikus skálán. A szaggatott vonal ugyanezen mértéknek az MLP topológiájú neurális háló által a tesztmintán becsült kategóriák hibaarányát mutatja az epochszám függvényében.

1. ábra

A tájolási értékek becslésének hibaaránya az epochszám függvényében

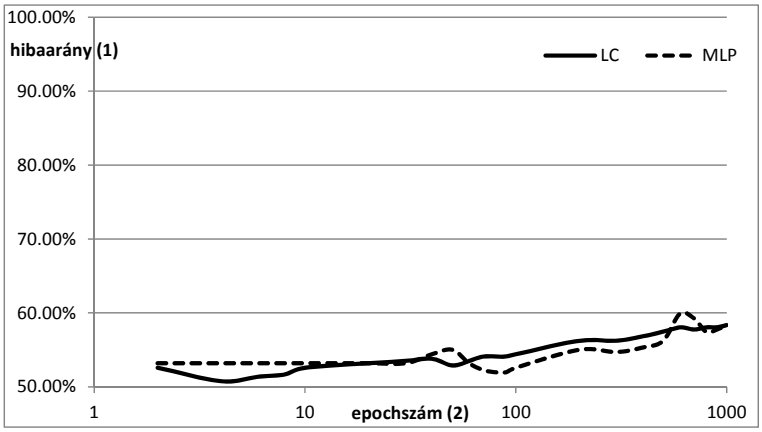

Figure 1: Prediction error ratios of sentiment orientation values according to number of epochs

Error ratio(1), Number of epochs(2)

Az *1. ábráról* leolvasható, hogy a lineáris osztályozó gyorsan rátanul a tanítóminta segítségével az egyes kategóriák között húzható hipersíkok helyzetére: a keresztvalidáció hibaaránya 4 epoch alatt eléri a minimumát, az 50,76%-ot. Ezután pedig lassulva növekszik, ez az oka, hogy logaritmikus skálán mértük fel az epochszámot. Az MLP viszont viszonylag sokáig nem képes felismerni a minta struktúráját, majd rövid ingadozás után általában ugyanúgy lassuló növekedésbe kezd a hibaarány, mint a lineáris osztályozó esetében. Ennek minimuma a századik epoch környékén tapasztalható (51,98%), amely valamivel felette van a lineáris osztályozó minimális hibaarányánál.

A hasznossági értékek becsülhetősége mesterséges neurális hálózatokkal

A tájolási értékek becsülhetőségét leíró rész nyitó gondolatmenetét alkalmazva a hasznossági értékekre, előfordulhat, hogy egy adott szó hasznosságát nem tudjuk

megállapítani, vagy pedig többféle hasznosságot is képviselhet – függően attól, hogy milyen más szavak mellett található.

A *2. ábra* azt mutatja meg, hogyan változik a hasznossági értékek becslésének hibaaránya az epochszám függvényében a két vizsgált topológián. A folytonos vonal a lineáris osztályozó hibáját mutatja keresztvalidáció során, az epochszám függvényében. A szaggatott vonal ugyanezen mértéknek az MLP topológiára vonatkoztatott értékeit ábrázolja.

2. ábra

A hasznossági értékek becslésének hibaaránya az epochszám függvényében

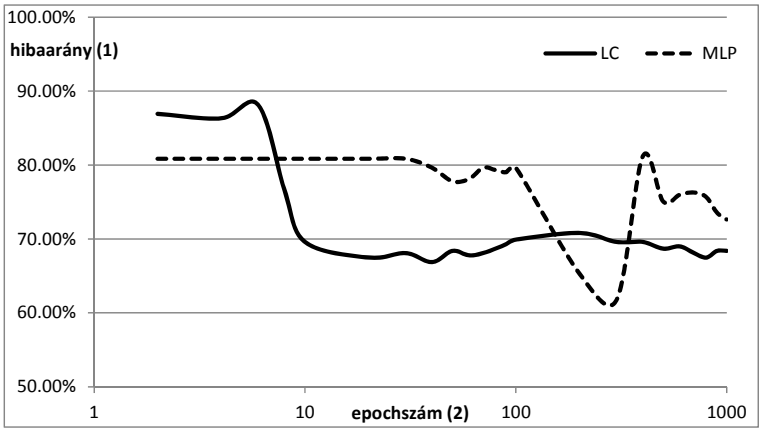

Figure 2: Prediction error ratios of utility values according to number of epochs

Error ratio(1), Number of epochs(2)

A lineáris osztályozó ismét gyorsan képes olyan hipersíkok felismerésére, amelyekkel nem csak a tanítómintán, hanem a tesztmintán is egyre jobb szeparációt lehet megvalósítani. A hatodik epoch után elkezd csökkenni a keresztvalidáció hibaaránya, amely a tízedik epoch után stabilan 70% környékén marad. Ez a hibaarány persze messze elmarad a tájolási értékeknél megfigyelttől, azonban a véletlenhez képest jobb eredményt ad. Az MLP viszont ismét viszonylag sokáig nem képes felismerni a minta struktúráját, majd nagyon enyhe ingadozás után gyorsan csökkenni kezd (200-300. periódus környékén 65-61% körül alakul az értéke). Ezután általában erős túltanulás figyelhető meg, a hibaarány megugrott, majd lassan csökkent.

Látható tehát, hogy a hasznossági értékek becslése az alkalmazott modell alapján korántsem ad elfogadható eredményt, és az is látszik, hogy a tájolási értékekhez képest is jóval nehezebben becsülhető a hasznosság a rövid szöveges üzenetek esetében.

A tájolási értékek becsülhetősége szupport vektor gépekkel

A tájolási címkékre háromszoros validációval kapott összesítő eredmények táblázatát a *3. táblázat* mutatja. A C büntető költség különböző értékeire kapott modellek közül a legkisebb hibát (45,05%) a C=1000 érték mellett kaptuk. Ráadásul ez az érték a stopszavazott előfeldolgozottsági szint mellett adódik. Vagyis a szótövezés – mint a leginkább nyelvfüggő eljárás – kiküszöbölhető. Ekkora hiba mellett ugyan a módszer gyakorlati alkalmazhatósága

(legalábbis az adott tanító mintán) megkérdőjelezhető. De mint látni fogjuk, számunkra inkább az a tény fontosabb, hogy a hiba 50% alá csökkenthető.

3. táblázat

Tájolási értékek osztályozási hibája lineáris szupport vektor módszer mellett

SVM hiba (1)	**Nyers szöveg (2)**		**Tokenizált szöveg (3)**		**Stopszavazott szöveg (4)**		**Szótövezett szöveg (5)**	
C	**Átlag (6)**	**Szórás (7)**	**Átlag**	**Szórás**	**Átlag**	**Szórás**	**Átlag**	**Szórás**
0.001	68,23%	7,07%	48,36%	4,46%	48,36%	4,46%	48,06%	4,43%
1	48,17%	4,88%	48,06%	4,19%	47,97%	4,02%	47,66%	4,01%
1000	48,97%	5,60%	45,05%	2,63%	**45,05%**	**2,63%**	45,23%	4,72%
3000	48,86%	3,27%	47,05%	1,60%	46,76%	4,50%	46,76%	4,50%

Table 3: Prediction error ratios of sentiment orientation values using Linear SVM

SVM error(1), Raw text(2), Tokenized text(3), Text after removing stop words(4), Stemmed text(5), Mean(6), St. dev.(7)

A hasznossági értékek becsülhetősége szupport vektor gépekkel

A *4. táblázat* hasonló futtatások eredményeit mutatja, de nem a tájolási, hanem a hasznossági címkék esetén. A kapott eredmény abban is hasonlít az előzőekhez, hogy itt is a C=1000 érték mellett sikerült a legkisebb hibát elérni, és ráadásul ugyanúgy a stopszavazott előfeldolgozás mellett. Az 59,13%-os hibaarány mellett azonban egyáltalán nem beszélhetünk érdemi tanulásról.

4. táblázat

Hasznossági értékek osztályozási hibája lineáris szupport vektor módszer mellett

SVM hiba (1)	**Nyers szöveg (2)**		**Tokenizált szöveg (3)**		**Stopszavazott szöveg (4)**		**Szótövezett szöveg (5)**	
C	**Átlag (6)**	**Szórás (7)**	**Átlag**	**Szórás**	**Átlag**	**Szórás**	**Átlag**	**Szórás**
0.001	84,22%	5,19%	84,02%	5,80%	84,62%	5,54%	85,22%	5,39%
1	68,23%	7,07%	66,20%	6,50%	66,13%	5,50%	65,23%	4,77%
1000	65,43%	6,16%	59,14%	4,50%	**59,13%**	**4,61%**	61,20%	4,12%
3000	65,37%	5,72%	60,54%	5,37%	61,34%	4,97%	62,24%	4,69%

Table 3: Prediction error ratios of utility values using Linear SVM

SVM error(1), Raw text(2), Tokenized text(3), Text after removing stop words(4), Stemmed text(5), Mean(6), St. dev.(7)

KÖVETKEZTETÉSEK

Az eredmények alapján egyértelműen megállapíthatjuk, hogy a tájolás és a hasznosság lényegükben különböző fogalmak. Hipotézisünk, miszerint a tájolás erősebben kötődik a dokumentum objektív tartalmához, míg a hasznosságot inkább az olvasó szubjektuma

határozza meg, kétszeresen is igazolást nyert. Mind a mesterséges neurális hálózatok, mind pedig a szupport vektor gépek alkalmazásakor azt tapasztaltuk, hogy a tájoláshoz található tanuló algoritmushoz képest (noha az nem is túl hatékony), a hasznossági értékek esetében a felügyelt tanulási módszereknek még az optimalizált változatai sem adtak értékelhető eredményt.

Mindez abba az irányba tereli a további kutatásainkat, hogy ha a hasznosságnak létezik egyáltalán dokumentumokra épülő tanuló modellje, abban vagy a szózsák modellnél pontosabb ábrázolási technikára van szükség, vagy a dokumentum egyéb tulajdonságait (hossz, helyesírás, írásjel-szám-betű arány stb.) is be kell vonni a modellbe. Érdemes továbbá magán a szövegen kívül a webes dokumentumok egyéb attribútumait is (szerző, dátum, menü stb.) bevonni a jellemzők közé.

Eredményül megfogalmazhatjuk, hogy az itt bemutatott módszerek bármelyike alkalmas arra, hogy a tanulhatóság megvizsgálásával tetszőleges dokumentum címkéről eldönthessük, hogy az inkább az objektumhoz köthető címke, vagy a címkéző felhasználó szubjektív véleménye.

KÖSZÖNETNYILVÁNÍTÁS

A kutatás technológiai hátterét az Új Magyarország fejlesztési terv pályázatának GOP-1.1.1-09/1-2009-0019 projektjében résztvevő partnereink biztosították.

IRODALOM

Borgulya, I. (1998): Neurális hálók és fuzzy-rendszerek. Dialóg Campus : Pécs, 226 p.

Cotton, N.J., Wilamowski, B.M. (2011): Compensation of Nonlinearities Using Neural Networks Implemented on Inexpensive Microcontrollers. IEEE Transactions on Industrial Electronics, 58. 3. 733-740. p.

Crammer K., Singer Y., Cristianini N., Shawe-taylor J., Williamson B. (2001): On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2. 265-292.

Fan R.E., Chen P.H., Lin C.J. (2005): Working set selection using second order information for training SVM. Journal of Machine Learning Research 6. 1889-1918. p.

Hanke M., Halchenko Y.O., Sederberg P.B., Hanson S.J., Haxby J.V., Pollmann S. (2009): PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics 7. 37-53. p.

Kryzanowski, L., Galler, M., Wright, D.W. (1993): Using Artificial Neural Networks to Pick Stocks. Financial Analysts Journal 49. 4. 21 p.

Kushal D., Lawrence S., Pennock D.M. (2003): Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of WWW-03, 519-528. p.

Liu B. (2010): Handbook of Natural Language Processing. 2nd Edition, CRC Press LCC, ISBN13 9781420085938, 627-660. p.

Nissen S. (2003): Implementation of a Fast Artificial Neural Network Library, Report 31. Department of Computer Science University of Copenhagen (DIKU) 1-88. p.

Pang B., Lee L. (2008): Opinion mining and sentiment analysis. In Foundations and Trends in Information Retrieval 2. 1-2. 1-135. p.

Qin Y., Wang X. (2009): Study on Multi-label Text Classification Based on SVM. Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on, ISBN: 978-0-7695-3735-1, 300-304.

Tikk D. (szerk.) (2007): Szövegbányászat. Typotex : Budapest, 294. p.
Walczak, S. (2001): An Empirical Analysis of Data Requirements for Financial Forecasting with Neural Networks. Journal of Management Information Systems, 17. 4. 203-222. p.

Levelezési cím (*Corresponding author*):

Kovács Balázs
Pécsi Tudományegyetem, Közgazdaságtudományi Kar
7622 Pécs, Rákóczi út 80.
University of Pécs, Faculty of Economics
H-7622 Pécs, Rákóczi út 80.
Tel.: 36-72-501-599/3113, Fax: 36-72-501-553
e-mail: kovacs.balazs.ktk@gmail.com