Jogi szövegek tezaurusz alapú osztályozása: egy nyelvfüggetlen modell létrehozásának problémái

Nyéki Bence

Nyelvtudományi Intézet
nyeki.bence@nytud.hu

Kivonat A cikkben jogi szövegek automatikus többcímkés osztályozását vizsgáljuk. A feladat nagy mennyiségű betanító adatot igényel, azonban ha az osztályozás kivitelezhető a többnyelvű EUROVOC tezaurusz terminusai alapján, akkor elméleti lehetőség nyílik arra, hogy egy meghatározott nyelvű korpuszon betanított osztályozó nyelvfüggetlenül működhessen. A bináris relevancia módszerén alapuló osztályozónkat horvát korpuszon tanítottuk be, és bár teljesítménye horvát szövegeken elfogadható, kis méretű annotált magyar mintánkra alkalmazva gyenge eredményt mutatott. Ennek legvalószínűbb oka a horvát és a magyar korpusz közötti különbség a terminus- és címkeeloszlás szempontjából.
Kulcsszavak: osztályozás, többcímkés, tezaurusz, EUROVOC, nyelvfüggetlen

1. Bevezetés

Az Európai Unió számára fontos feladatnak számít az egyes tagállamok nemzeti nyelvű jogi szövegeinek összegyűjtése és egységes feldolgozása. Az ennek megvalósítására irányuló munka eredménye például az EUR-Lex[1] weboldal, mely az Európai Unió 24 nyelvén írt hivatalos dokumentumokhoz biztosít hozzáférést, vagy az EUROVOC[2] többnyelvű tezaurusz. A MARCELL CEF Telecom[3] projekt ugyancsak egy lényeges cél elérésére irányul: jogi szövegek automatikus fordításának fejlesztésére hét nyelv között (horvát, magyar, román, bolgár, szlovák, szlovén, lengyel). E projekt keretein belül már létrehoztak egységesen annotált korpuszokat az érintett nyelvek jogi szövegeiből (Váradi és mtsai, 2020). Bár a korpuszok részletes morfológiai és szintaktikai információt tartalmaznak, valamint a már említett EUROVOC tezaurusz és a IATE[4] adatbázis terminusainak jelölését is kivitelezték, még nem minden érintett nyelv anyagainak annotációja teljes: a következő lépés a dokumentumok szövegszintű annotálása az EUROVOC tezaurusz 21 legfelső fogalmi kategóriájával (doménjével).

A szövegek automatikus címkézésének egyik módja egy osztályozó létrehozása gépi tanulás segítségével. Ehhez azonban már előre megjelölt betanító adatokra

[1] https://eur-lex.europa.eu/homepage.html
[2] https://op.europa.eu/en/web/eu-vocabularies
[3] https://marcell-project.eu/
[4] https://iate.europa.eu/home

van szükség. Ilyen adatokat jelenleg a horvát és a szlovén korpusz tartalmaz. A rendelkezésünkre bocsátott horvát korpusz szövegeinek nagyobb részéhez (pontosan 19.802 dokumentumhoz) már hozzárendeltek egy vagy több EUROVOC domént. Megkíséreltük felhasználni ezeket a horvát nyelvű anyagokat egy nyelvfüggetlen osztályozó betanítására, amely képes lenne a magyar (vagy elméletileg bármely más nyelvű) korpusz szövegszintű annotálására is. Ennek elméleti lehetőségét az EUROVOC terminusai teremtik meg: a tezaurusz alapját a SKOS (*Simple Knowledge Organization System*) séma (Isaac és Summers, 2009) szerint fogalmak alkotják, amelyek egymással alá-fölé rendeltségi viszonyban vannak, továbbá konceptuális sémákba és doménekbe rendeződnek. Ezeknek a fogalmaknak a tezaurusz által lefedett összes nyelven egy-egy (vagy akár több szinonim) terminus felel meg. Ez egyértelmű megfeleltethetőséget eredményez különböző nyelvek egyes kifejezései között. Amennyiben betanítunk egy tetszőleges nyelvű annotált korpuszon egy olyan osztályozót, amely csak a tezaurusz fogalmainak nyelvfüggetlen azonosítóit veszi figyelembe, akkor az alkalmazhatóvá válik bármilyen más nyelvű szövegre is. Ennek feltétele persze, hogy az adott szövegben előzetesen azonosítókkal jelöljük meg az azon fogalmaknak megfelelő terminusokat, amelyeken az osztályozó betanult. Ez az annotáció rendelkezésre is áll a korpuszokban.

A cikk hátralevő része a következőképpen épül fel: a 2. részben röviden tárgyaljuk a kapcsolódó irodalmat, majd a 3. részben áttérünk a horvát és magyar korpusz jellemzésére. A 4. rész a horvát szövegeken betanított modelleknek a horvát validációs és tesztanyagokon, illetve egy kis méretű magyar mintán való kiértékelését közli. Az 5. rész összegzi az eredményeket.

2. Kapcsolódó irodalom

Mivel egy-egy dokumentumot az EUROVOC több felső fogalmi kategóriája is jellemezhet, a fent vázolt feladat többosztályos és többcímkés automatikus osztályozás kivitelezését teszi szükségessé. A gépi tanulásban erre két megoldási stratégiát különítenek el: problématranszformációt és algoritmusadaptációt (Tsoumakas és mtsai, 2010; Dharmadhikari és mtsai, 2011). Az előbbi lényege a feladat olyan lépésekre való lebontása, amelyekre alkalmazhatók egycímkés algoritmusok, míg az utóbbi lényege az egycímkés algoritmusok olyan átalakítása, hogy egy-egy osztályozandó objektumhoz több címkét is képesek legyenek társítani. A két stratégiához tartozó problémákról, módszerekről és értékelésükről ír (Tsoumakas és mtsai, 2010). A különböző módszerek hatékonyságát mérte (Dharmadhikari és mtsai, 2011). A kitűzött dokumentumklasszifikációs feladathoz választott módszer részletesebb ismertetésére a 4. pontban kerül sor.

Tezaurusz alapú gépi tanulást alkalmaztak (Sabbagh és mtsai, 2018) vállalatok gyártási kapacitásának osztályozására. A kutatás anyagául különböző vállalatok weblapjaiból kinyert szövegek szolgáltak. A szerzők szerint a tezaurusz alkalmazásának előnye abban áll, hogy jól szűrhetővé válnak a szövegek azon elemei, amelyek lényeges információt tartalmaznak az osztályozás szempontjából. Ezenkívül az osztályozó betanítása is a tezaurusz segítségével történt (nem

előre annotált korpusz alapján): minden címkéhez félig automatizált módszerrel választottak ki és súlyoztak releváns fogalmi egységeket a tezauruszból, konceptuális tereket hozva létre ezzel az egyes címkékhez. E munka keretein belül egynyelvű tezaurusszal dolgoztak.

Ugyancsak meg kell említeni, hogy az elmúlt években több kísérlet történt az EUROVOC tezaurusz felhasználására jogi szövegek osztályozásának céljából (Steinberger és mtsai, 2012; Chalkidis és mtsai, 2019). E munkák azonban címkékként használták az EUROVOC fogalmait, nem pedig jellemzőkként. Következésképpen az így létrehozott osztályozók óriási (több mint 7.000 elemből álló) címkehalmazzal dolgoztak, míg a jelen cikkben közölt kutatásban mindössze a 21 felső fogalmi kategória alkotta ezt a halmazt. Továbbá a betanítás nem terminusokon alapult, hanem a szövegek szavain (*bag of words*) vagy azok beágyazásain. A JEX szoftver[5] (Steinberger és mtsai, 2012) leginkább a kézi annotálás megkönnyítésére alkalmas eszköz, amely a bemeneti dokumentumok vektorait az egyes címkék profiljaival (azaz a tanító adatok alapján kiszámolt centroidokkal) veti össze, és az n legjobb címkét rendeli az adott dokumentumokhoz (n egy meghatározott szám, a szerzők által beállított alapértelmezett értéke a 6). Az algoritmus így minden dokumentumhoz rangsorolja a címkéket, de mindig n darabot ad ki, ami az esetek jelentős részében nem egyezik a helyes címkék számával. A címkék rangsorolását mások később mélytanulási technológiák segítségével kivitelezték (Chalkidis és mtsai, 2019).

3. A korpuszok

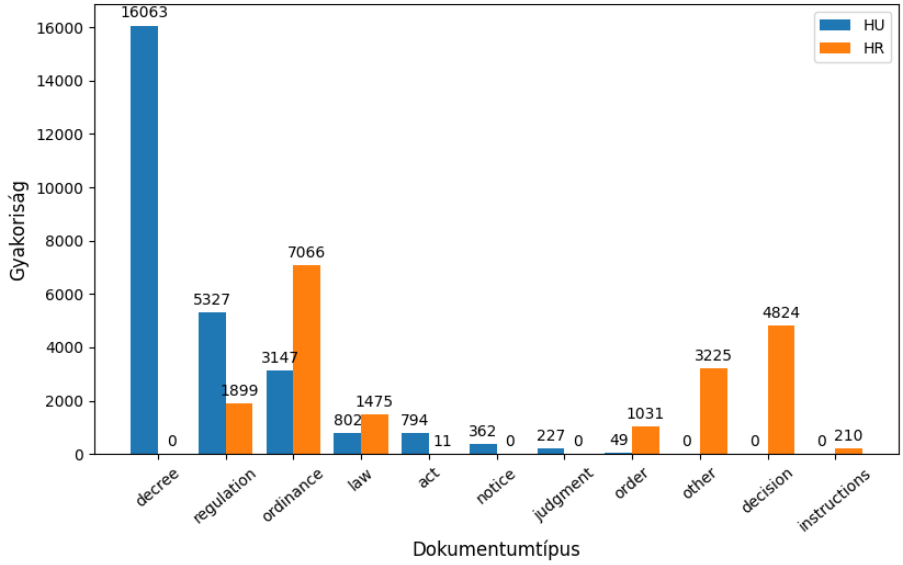
Ez a rész a horvát és a magyar jogi korpusz összevetését tartalmazza. A MARCELL projekt korpuszainak átfogó leírását adja (Váradi és mtsai, 2020).

3.1. Szavak és terminusok gyakoriságai

A magyar korpusz 26.821 dokumentumból áll, melyek döntő többsége határozat (16.063 szöveg tartozik ebbe a kategóriába). Ebből egyelőre mindössze 200 véletlenszerűen kiválasztott dokumentumot annotáltunk EUROVOC doménekkel. A teljes horvát anyag 33.559 dokumentumot foglal magába, a továbbiakban azonban csak annak a 19.802-nek adjuk jellemzését, amelyet kézileg megjelöltek az említett fogalmi kategóriákkal; ezekre fogunk horvát korpuszként hivatkozni, mivel a feladat szempontjából ezek relevánsak. Itt a leggyakoribb dokumentumtípus az utasítás, az idetartozó szövegek megközelítőleg harmadát teszik ki a korpusznak. A továbbiakban is láthatjuk, hogy a horvát minta több szempontból kiegyensúlyozottabb, mint a magyar.

Az 1. ábra a dokumentumtípusok eloszlását mutatja a magyar és horvát szövegekben. Mindkét korpusz dokumentumai tartalmazzák a megfelelő típus angol megnevezését, ezek a megnevezések láthatók az ábrán. Erre csak azok a típusok kerültek fel, amelyek legalább a két korpusz egyikében 100-nál többször fordulnak elő. A két korpusz típuscímkéi között nem teljes az átfedés.

[5] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer

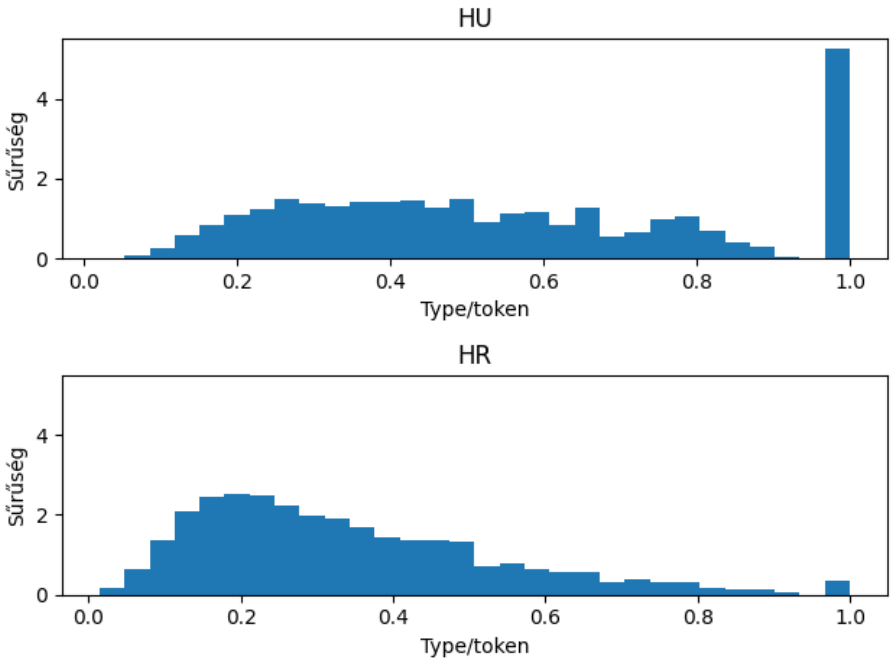


1. ábra: Dokumentumtípusok eloszlása

Az 1. táblázat néhány statisztikai adatot tüntet fel. Bár a horvát korpuszt mind az egyes dokumentumokban számolt szavak, mind az EUROVOC terminusok száma alapján nagyobb szórás jellemzi, az átlagok és mediánok is jóval magasabbak. Ezt a különbséget főként a magyar korpuszban nagy számban jelenlévő rövid (akár csak néhány mondatos) határozatok okozzák. Erre jobban rávilágít a 2. ábra, mely a két korpusz szövegeinek EUROVOC terminusai alapján kiszámolt type-token arányok hisztogramját mutatja (az egy dokumentumhoz tartozó type-token arányon a dokumentumban megtalálható különböző terminusok számának és a terminusok teljes számának hányada értendő). A hisztogram nem az egyes x-tengely menti szakaszokhoz tartozó abszolút gyakoriságokat, hanem az ezekből számolt sűrűséget tünteti fel. Ez azt jelenti, hogy az abszolút gyakoriságokat elosztjuk a gyakoriságok összegével és a hisztogram megfelelő oszlopainak szélességével, hogy az oszlopok területének összege 1 legyen.

1. táblázat. Statisztikai adatok gyakoriságokból

	HU		**HR**	
	szavak	**terminusok**	**szavak**	**terminusok**
Átlag	1164,23	64,62	2581	258,72
Medián	242	17	769	84
Szórás	3531,74	190,79	7411,67	635,55



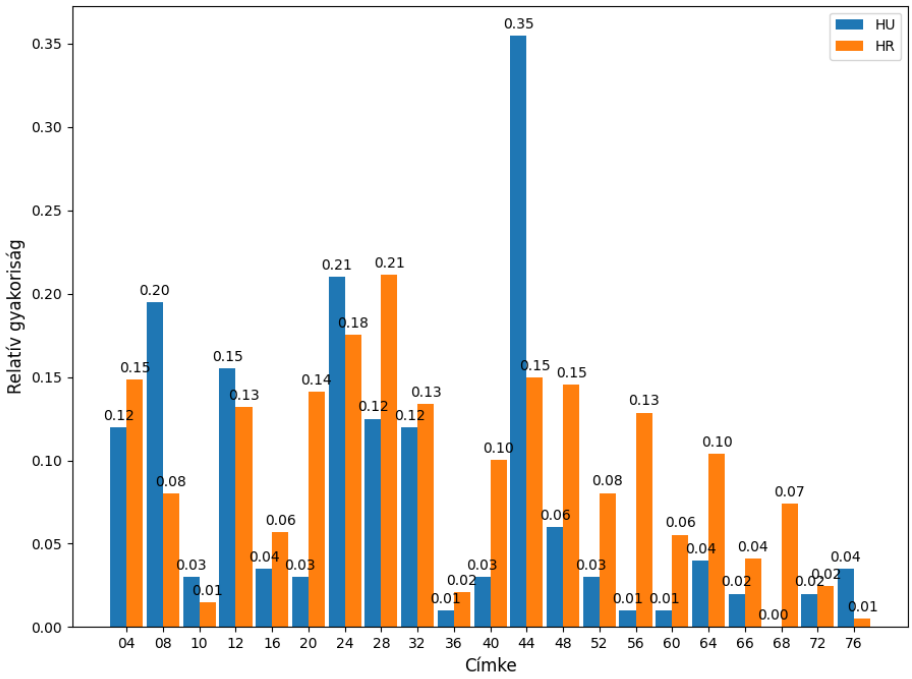
2. ábra: Type-token arányok

3.2. Címkék

Ugyancsak érdemes egy pillantást vetni az egyes címkék gyakoriságaira. A horvát korpuszból és a 200 kézileg megjelölt magyar szövegből kinyert relatív címkegyakoriságok[6] a 3. ábrán láthatók. A relatív gyakoriságot úgy számolhatjuk, hogy elosztjuk az adott címke abszolút gyakoriságát a mintában lévő összes dokumentum számával, így jelen esetben a relatív gyakoriságok összege nagyobb, mint 1, hiszen egy dokumentumhoz több címke is tartozhat.

A horvát korpuszban a két leggyakoribb címke a 28 (társadalmi kérdések) és a 24 (pénzügyek): relatív gyakoriságuk 0,21, illetve 0,18 (mindegyik több mint

[6] A feltüntetett kódok az EUROVOC doméneket jelzik. Feloldásuk a következő: 04: Politika, 08: Nemzetközi kapcsolatok, 10: Európai Unió, 12: Jog, 16: Közgazdaságtan, 20: Kereskedelem, 24: Pénzügyek, 28: Társadalmi kérdések, 32: Oktatás és kommunikáció, 36: Tudomány, 40: Vállalkozások és verseny, 44: Foglalkoztatás és munkakörülmények, 48: Közlekedés, 52: Környezet, 56: Mezőgazdaság, erdészet és halászat, 60: Agrárélelmiszer-ipar, 64: Termelés, technológia és kutatás, 66: Energia, 68: Ipar, 72: Földrajz, 76: Nemzetközi szervezetek. Lásd https://op.europa.eu/en/web/eu-vocabularies/th-top-concept-scheme/-/resource/eurovoc/100141?target=Browse.



3. ábra: Relatív címkegyakoriságok a korpuszokban

3.000 dokumentumban jelenik meg). Érdekes, hogy a 200 megjelölt dokumentumot tartalmazó magyar mintában a legtöbbször megfigyelhető címke gyakorisága sokkal inkább kiugró értéknek látszik: a 44-es EUROVOC domén (foglalkoztatás és munkakörülmények) relatív gyakorisága 0,35 (a címke 71-szer fordul elő), a második helyen álló 24-esé (pénzügyek) pedig 0,21 (ez 42-szeres előfordulást jelent). A legtöbb címke relatív gyakorisága a mintában nem éri el a 0,1-et, a 68-as (ipar) pedig teljesen hiányzik. Hangsúlyozni kell, hogy ezt a 200 dokumentumot (ami természetesen nem reprezentatív minta) egyetlen ember jelölte meg. Másik szakértő bizonyára részben más címkéket társított volna a dokumentumokhoz. Mindazonáltal ezek a számok tükrözik a magyar szövegek osztályozásának problémáját: a 44-es címke általában rövid kinevezéseket (pl. bírói tisztségre) jellemez. Ezekben a szövegekben ritkán fordul elő annyi terminus, hogy azok jól mutassák, melyik osztályba tartozik a dokumentum.

A type-token arányok és a címkegyakoriságok alapján arra következtethetünk, hogy a két korpusz összetétele erősen különbözik. Ezért a következőkben két problémát vizsgálunk:

1. Ha létrehozunk egy osztályozót, amely a horvát szövegeken tanult be, az mennyire hatékonyan fog további horvát szövegeket klasszifikálni?
2. Ugyanez az osztályozó használható lesz-e magyar szövegek feldolgozására is?

Egy horvát szövegeket hatékonyan osztályozó modell létrehozása tehát nem jelenti azt, hogy ugyanaz a modell a magyar szövegeket is jól fogja annotálni. Az alább bemutatott kísérletek éppen ezt támasztják alá.

4. Osztályozás

A következőkben az automatikus osztályozók kiértékelésére térünk át. Ehhez a horvát korpuszt dokumentumok véletlen kiválasztásával három részre osztottuk: tanító halmazra (15.842 dokumentum, kb. 80%), valamint validációs és teszthalmazra (mindegyik 1.980 dokumentumot tartalmaz, ez megközelítőleg 10-10%).

4.1. A címkék rangsorolása, egycímkés osztályozás

Ha ismert az egyes dokumentumokhoz társítandó címkék száma, a többcímkés osztályozás megoldható a lehetséges címkék rangsorolásával aszerint, hogy mennyire jól illenek az adott osztályozandó dokumentumhoz. Az eddig tárgyalt feladat nem ehhez az esethez tartozik: egy szöveget a címkék tetszőleges nem üres részhalmaza jellemezhet, amelynek elemei nem rendezettek relevancia szerint. A címkék rangsorolása tehát nem tartozik a munka fő céljai közé. Mindazonáltal célszerű lehet néhány egyszerű, címkerangsorolásra irányuló kísérlet elvégzése. Ha ugyanis a nyelvfüggetlen többcímkés osztályozást nem sikerül kielégítően kivitelezni, a rangsorolást viszont igen, a címkéket az egyes dokumentumokhoz relevancia szerint rendező algoritmus felhasználható egycímkés osztályozóként, ami voltaképpen félmegoldást jelentene. Ez nem valósult meg, de megfigyelhetjük, hogy a rangsoroláson alapuló egycímkés osztályozáskor és a többcímkés klasszifikációkor ugyanaz a probléma merült fel: a rendelkezésre álló nyelvi anyagok különbözősége.

A címkék rangsorolásával kapcsolatos kísérlet lényege a következő: minden dokumentumhoz rangsoroljuk a címkéket, majd kiválasztjuk a legrelevánsabbat, és ellenőrizzük, az valóban eleme-e a valódi címkék halmazának (ha igen, az eredményt elfogadjuk).

Ennek megoldására két algoritmust dolgoztunk ki. Az egyik naiv, nem gépi tanulásra épülő módszer az EUROVOC tezaurusz hierarchiáját igyekszik kihasználni. A SKOS struktúrában a fogalmak egyfelől hierarchikus viszonyban vannak egymáshoz képest (bővebb és szűkebb fogalmak szerint), másfelől elvontabb fogalmi sémák alá tartoznak. Utóbbiak az EUROVOC tezauruszban közvetlenül a legfelső kategóriák, a domének alatt helyezkednek el. Ha tehát minden dokumentumban megszámoljuk, hány terminus (vagyis ezeknek megfelelő fogalom) tartozik az egyes doménekhez, úgy azt a domént tekinthetjük a legjobb címkének, amelyet a legtöbb terminus reprezentál.

A másik lehetőségként egy naiv Bayes osztályozót vizsgáltunk, amely a nyers terminusgyakoriságokon tanult be (tehát egy olyan mátrixon, melynek sorai dokumentumoknak, oszlopai fogalmaknak felelnek meg, egyes elemei pedig egy adott fogalomnak megfelelő terminus dokumentumonkénti abszolút gyakoriságait mutatják). Az ehhez szükséges számítások elvégzéséhez úgy tekinthető, hogy

egy dokumentum beletartozik az összes olyan osztályba, amelynek címkéjével annotálták.

A hierarchia alapú osztályozó a teljes horvát korpuszon így 24,05%-os teljesítményt ért el, a 200 megjelölt magyar dokumentumon pedig 30%-ot. A feladat egyszerűségét figyelembe véve ezt az eredményt elégtelennek kell tekintenünk: nyilvánvaló, hogy egy gyakorlati alkalmazásban ennél jobb teljesítményt várnánk. A naiv Bayes elfogadhatóbb eredményeket adott a horvát tesztanyagokon (mivel itt csak egy modellel kísérletezünk, a validációs anyagok egyesíthetők voltak a teszthalmazzal): 77,88%-ot, a tanító halmazon pedig 79,72%-ot. A magyar mintán egészen más eredmény született: mindössze 24%. A naiv Bayes teljesítményére erősen negatív hatással lehet, ha a tanítóhalmazban az osztályok reprezentációja kiegyensúlyozatlan, ami magyarázhatja a horvát anyagon mért kb. 20% hibát. A magyar szövegeken kapott alacsony érték azt támasztja alá (bár nem bizonyítja), hogy a magyar minta a terminusok eloszlását tekintve más jellegű dokumentumokból áll, mint a horvát korpusz.

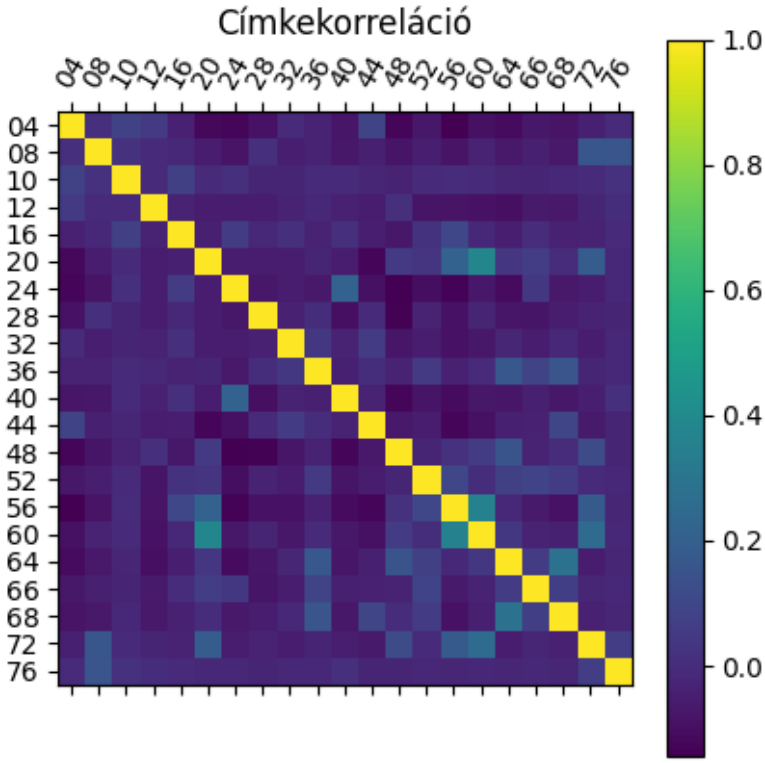
4.2. Többcímkés osztályozás

A relevánsabb feladat egy olyan osztályozó létrehozása volt, amely minden dokumentumhoz egy vagy több nem rangsorolt címkét társít. Ennek kivitelezésére a BR (*Binary Relevance*) vagy OvR (*One vs Rest*) módszert használtuk. Ez egy problématranszformációs módszer, melynek lényege, hogy minden egyes lehetséges címkéhez felállít egy-egy bináris osztályozót, amely a többi címkétől függetlenül abban hoz döntést, hogy az adott címke jól jellemzi-e az osztályozandó dokumentumot vagy sem (Tsoumakas és mtsai, 2010; Dharmadhikari és mtsai, 2011). A bináris osztályzók bármilyen algoritmus alapján működhetnek. A természetes nyelvek feldolgozására gyakran alkalmazott naiv Bayes, SVM (*Support Vector Machine*) és k-legközelebbi szomszéd (KNN, *K Nearest Neighbors*) algoritmusokkal dolgoztunk. A modellek létrehozására a Python *scikit-learn* könyvtárát használtuk [7] (Pedregosa és mtsai, 2011). A BR előnye, hogy konceptuálisan egyszerű és nem igényel nagy számítástechnikai kapacitást, hátránya viszont az, hogy nem veszi figyelembe az egyes címkék közötti korrelációt. Ez komoly leegyszerűsítést jelenthet, bár ha vetünk egy pillantást a címkék korrelációs mátrixára (lásd 4. ábra), amelyet a teljes horvát korpusz alapján számoltunk ki, akkor csak viszonylag alacsony együtthatókat látunk. A legnagyobb korrelációs koefficienst a 60-as (agrárélelmiszer-ipar) és a 20-as (kereskedelem) címkék között számoltuk ki, de ennek értéke is kisebb, mint 0,4.

A választott módszer hátránya ellenére a horvát validációs és tesztanyagokon mért eredmények azt mutatják, hogy a BR alkalmazása elfogadható eredményekhez vezet.

Első lépésként a horvát szövegekből gyakorisági mátrixokat nyertünk ki, melyekben minden sor egy-egy dokumentumot, az oszlopok pedig egy-egy EUROVOC fogalmat reprezentáltak: az egyes elemek így értelemszerűen azt mutatják, hányszor fordul elő valamely fogalom (azaz a neki megfelelő terminus) egy adott

[7] https://scikit-learn.org/stable/



4. ábra: A horvát szövegek címkéi közötti korrelációs együtthatók

dokumentumban. Ezeket az adatokat előzetesen átalakítottuk. Ez az osztályozó létrehozásának rendkívül fontos lépése, mely erősen befolyásolhatja a végeredményt. A különböző paraméterű SVM modelleknek német nyelvű szövegeken való betanítása után (Leopold és Kindermann, 2002) például arra a következtetésre jutottak, hogy az előfeldolgozás (TF-IDF, lemmatizálás stb.) fontosabbak az eredmények szempontjából, mint a modell magfüggvényének választása.

Két előfeldolgozási módszert alkalmaztunk: TF-IDF-et L2 normalizálással és főkomponens-analízist. A TF-IDF lényege, hogy nem csak abszolút gyakoriságaik szerint súlyozza a terminusokat, hanem azt is számításba veszi, hogy hány különböző dokumentumban fordulnak elő: nagyobb súlyt kapnak azok, amelyek kevesebb dokumentumban jelennek meg. A főkomponens-analízis pedig olyan technika, mely a jellemzőteret kisebb dimenziójú térbe képzi le kismértékű in-

formációvesztéssel (magas megmagyarázott varianciával a jellemzőket tekintve). Az utóbbi eljárás hátránya, hogy az új, csökkentett dimenziójú mátrix néhány eleme 0-nál kisebb értéket fog felvenni, így a főkomponens-analízis naiv Bayes osztályozóval nem használható.

Kétféleképpen értékeltük a modelleket. A (Tsoumakas és mtsai, 2010) által megadott képletek közül a klasszifikációs pontosságot (1. képlet) és az egyszerű pontosságot (*Accuracy*, 2. képlet) választottuk.[8] Az előbbi nagyon szigorú, mivel csak azokat az eseteket fogadja el, amikor a modell által jósolt címkék halmaza pontosan megegyezik az adott dokumentumhoz tartozó helyes címkehalmazzal (ezért a továbbiakban az "egyezés" megnevezéssel fogunk rá utalni). Pontos egyezés gyakran két emberi szakértő annotációi között sem várható. Ezért a második érték jobban jellemzi a modelleket.

A megadott képletekben Z_i és Y_i az i-edik megfigyeléshez tartozó valós, illetve jósolt címkehalmazt jelenti, I pedig olyan függvény, melynek értéke 1, ha az argumentuma igaz, egyébként pedig 0.

$$\frac{1}{m}\sum_{i=1}^{m} I(Z_i = Y_i) \tag{1}$$

$$\frac{1}{m}\sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{2}$$

A különböző előfeldolgozási eljárások után kapott legfontosabb eredményeket a 2. táblázatban közöljük. A feltüntetett SVM modellek magfüggvénye a radiális bázisfüggvény (*rbf*). Lineáris magfüggvénnyel is kísérleteztünk, de az gyengébb eredményeket hozott, amiket itt nem közlünk. Hasonlóképpen a k-legközelebbi szomszéd algoritmust magasabb k értékekkel is teszteltük (ezt az értéket a táblázatban a KNN rövidítés melletti szám jelzi), amelyek kevésbé voltak hatékonyak, mint az alább megadott KNN modellek. Ehhez az algoritmushoz a dokumentumok közelségét az euklideszi távolsággal mértük.

A következő táblázattal kapcsolatban meg kell még jegyeznünk, hogy azokban az esetekben, amikor főkomponens-analízist használtunk (a megfelelő oszlopot az angol *Principal Component Analysis* terminus PCA rövidítése jelzi), csak annyira redukáltuk a jellemzőteret, hogy az átalakítás után legalább 95%-os megmagyarázott varianciát kapjunk. Ehhez 956 dimenzióra volt szükség.[9]

A naiv Bayes eredményei gyengék, de a másik két osztályozónak sikerült olyan paramétereket találni, amelyek mellett viszonylag jól működnek. Ami az előfeldolgozást illeti, a TF-IDF határozottan javította az eredményeket, a főkomponens-analízis viszont többnyire rontotta. A jellemzők súlyozása tehát kulcsfontosságú,

[8] Ez nem összetévesztendő azzal a szintén pontosságként vagy precízióként ismert mértékkel, amely azt mutatja, hogy mekkora része helyes azoknak az értékeknek, amelyeket egy modell megjósol.

[9] Az EUROVOC összesen 7214 fogalmat tartalmaz, ám ezek jelentős része nem szerepelt a betanító adatokban. Így főkomponens-analízis nélkül is mindössze 4.722 dimenziója volt a jellemzőtérnek.

2. táblázat. A BR egyes algoritmusainak eredményei a validációs halmazon

Algoritmus	**TF-IDF**	**PCA**	**Egyezés**	**Pontosság**
NB	+	-	0,2676	0,5561
NB	-	-	0,0561	0,3898
KNN 1	+	+	0,5727	0,7291
KNN 1	+	-	**0,5803**	**0,7317**
KNN 1	-	+	0,4581	0,6008
KNN 1	-	-	0,4753	0,6536
KNN 5	+	+	0,4753	0,6536
KNN 5	+	-	0,4672	0,6467
KNN 5	-	+	0,3732	0,5141
KNN 5	-	-	0,4040	0,5521
SVM	+	+	**0,5096**	**0,6880**
SVM	+	-	0,5051	0,6844
SVM	-	+	0,1000	0,1623
SVM	-	-	0,1384	0,2148

dimenziócsökkentésre azonban nem volt szükség. Érdekes, hogy a legjobb teljesítményt a legegyszerűbb modell érte el: a k-legközelebbi szomszéd $k = 1$ paraméterrel. Ez azt jelenti, hogy a modell minden új dokumentumhoz megkeresi a jellemzőtérben a hozzá legközelebbi dokumentumot a tanító halmazból, és annak címkéit rendeli hozzá. Emellett az SVM sem teljesített rosszul a TF-IDF-fel és főkomponens-analízissel feldolgozott dokumentumokon (50% fölötti egyezés semmiképpen sem tekinthető gyengének). A tanító halmazon történő kiértékeléskor ugyanezek a modellek bizonyultak a legsikeresebbeknek: magától értetődő, hogy ebben a halmazban minden dokumentum legközelebbi szomszédja önmaga, ezért a KNN sosem hibázik. A jellemzők TF-IDF szerinti súlyozása és dimenziócsökkentés után az SVM által elért egyezés értéke 0,6458 volt, a pontosságé pedig 0,8065.

A táblázatban kiemelt két modellt a horvát tesztanyagokon és a magyar mintán is kipróbáltuk. Az így kapott értékeket a 3. táblázatban közöljük.

3. táblázat. A két legjobb modell kiértékelése a teszthalmazokon

	HR		**HU**	
Modell	**Egyezés**	**Pontosság**	**Egyezés**	**Pontosság**
KNN	0,5823	0,7372	0,0500	0,1628
SVM	0,5010	0,6894	0,0600	0,1446

A horvát teszthalmazon majdnem ugyanazokat az eredményeket kaptuk, mint a validáción, a magyar mintán azonban mindkét modell pontatlan, akárcsak a címkék rangsorolása esetében. Érdekes, hogy az SVM a magyar dokumentumok feléhez egyáltalán nem tudott címkét társítani: ez azt jelenti, hogy a BR összes

bináris osztályozója negatív döntést hozott, azaz irrelevánsnak értékelte azt a címkét, amelyikre betanult. Az SVM legtöbbször a 24-es (pénzügyek) címkét osztotta ki: 30-szor, míg a kézi annotációban ez többször (42-szer) szerepel. A KNN minden dokumentumhoz társított legalább egy címkét, de az eredményekből látható, hogy ezzel nem ért el jelentősen jobb teljesítményt. A KNN által a magyar szövegekhez leggyakrabban (62-szer) hozzárendelt címke a 04-es (politika), míg a kézi annotációban ez jóval ritkábban, 24-szer figyelhető meg. Az emberi jelölés szerint legtöbbször előforduló 44-es (foglalkoztatás és munkakörülmények) csak 20-szor jelent meg a KNN által jósolt címkék között.

Ugyancsak érdemes megfigyelni, hogyan változnak az eredmények, ha a magyar mintából csak a legtöbb EUROVOC terminust tartalmazó dokumentumokat vizsgáljuk. A csökkenő terminusgyakoriság szerint sorba rendezett magyar annotált szövegek közül az első 50 mindegyike legalább 49 terminust tartalmaz, a terminusgyakoriságok átlaga pedig 168,14 erre az 50 dokumentumra számolva. Csak ezeket figyelembe véve valamivel jobb eredményeket kapunk: a KNN 0,1 értékű egyezést és 0,2457-es pontosságot ad, ugyanezen mutatók értéke az SVM kimenetén pedig 0,14 és 0,2217. Bár ezek sem nagy számok, azt mutatják, hogy a rövidebb, kevés terminust tartalmazó magyar dokumentumokat (ahogy ez várható is) nehezebben kezelik a modellek.

Eddigi kísérleteink leírásának ezzel végére értünk. A következő feladatok között szerepel egy nagyobb magyar minta kézi annotálása, hogy így lehető legyen pontosabb képet kapni a létrehozott osztályozók működéséről magyar szövegeken.

5. Összegzés

Sikerült tehát egy egyszerű és megbízható (bár még fejleszthető) modellt létrehozni horvát jogi szövegek osztályzására. Az annotált magyar mintánkon azonban nem sikerült kielégítő eredményt elérni. A minta kis mérete miatt ezt nem tudjuk pontosan megmagyarázni, mindazonáltal a két korpusznak a 3. pontban leírt összevetését is figyelembe véve nagyon valószínű, hogy a dokumentumgyűjtemények közötti különbségek akadályai lehetnek egy nyelvfüggetlen osztályozó megalkotásának. Amennyiben a már rendelkezésre álló korpuszok valóban jól tükrözik a terminus- és címkeeloszlásokat az egyes nyelvekben, úgy a horvát jogi szövegeken aligha tanítható be olyan tezaurusz alapú osztályozó, amely a magyar szövegeket is helyesen kezeli. Mindazonáltal ahhoz, hogy általánosan értékelhessük a tezaurusz alapú nyelvfüggetlen osztályozókat, a megkezdett munka folytatásaként több különböző nyelv korpuszára és több nyelvpár anyagain végzett kísérletekre lesz szükség.

Köszönetnyilvánítás

A szerző köszönetét fejezi ki Marko Tadićnak és Vanja Štefanecnek a horvát korpusz megosztásáért, valamit Sass Bálintnak és Héja Enikőnek szakmai támogatásukért, javaslataikért.

Hivatkozások

Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Androutsopoulos, I.: Large-Scale Multi-Label Text Classification on EU Legislation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6314–6322. Association for Computational Linguistics, Florence, Italy (07 2019), https://www.aclweb.org/anthology/P19-1636

Dharmadhikari, C., Ingle, M., Kulkarni, P., Dharmadhikari, S.: A Comparative Analysis of Supervised Multi-label Text Classification Methods. International Journal of Engineering Research and Applications 1, 1952–1961 (2011)

Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C (2009)

Leopold, E., Kindermann, J.: Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? Machine Learning 46, 423–444 (01 2002)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

Sabbagh, R., Ameri, F., Yoder, R.: Thesaurus-Guided Text Analytics Technique for Capability-Based Classification of Manufacturing Suppliers. Journal of Computing and Information Science in Engineering 18 (03 2018)

Steinberger, R., Ebrahim, M., Turchi, M.: JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool. In: LREC'2012 (2012)

Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data. In: Maimon, O., Rokach, L. (szerk.) Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer US, Boston, MA (2010)

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Niton, B., Ogrodniczuk, M., Pkezik, P., Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Pais, V.F., Tufis, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J.: The MARCELL Legislative Corpus. In: LREC (2020)