Közérthetőség mint osztályozási probléma (?) - gépi tanulási kísérlet kézzel címkézett korpuszon

Üveges István[1,2]

[1]Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
[2]MONTANA Tudásmenedzsment Kft.
uvegesistvan898@gmail.com, uvegesi@montana.hu

Kivonat Cikkünkben bemutatjuk a laikusoknak címzett hivatalos szövegek osztályozási kísérletét felügyelt gépi tanuló algoritmusok segítségével. Vizsgálatunkhoz szakértők által, kézzel készített korpuszt használunk, amely közérthetőre fogalmazott és még átfogalmazás előtt álló mondatokat tartalmazott. Célunk ezzel egy olyan gépi tanult modell készítése, amely alkalmas lehet a szakértők figyelmét felhívni azon mondatokra egy-egy hivatalos szövegben, amelyek további megfontolást érdemelnek a szöveg közérthetőbbre alakítása során, ezzel gyorsítva a szakértői munkát. A kísérletet pilot jelleggel végezzük, az eredmények függvényében korszerűbb módszerek (pl. LSTM, BERT) esetleges kipróbálása előtt, a tapasztalatokat pedig a fentiek szerinti bináris klasszifikációs problémára jellemzően értékeljük.
Kulcsszavak: közérthetőség, Plain Language Movement (PLM), Support Vector Machine (SVM), kézzel címkézett korpusz

1. Bevezetés

A jogállamiság egyik követelménye a jog átláthatósága és kiszámíthatósága. Ez a fajta kiszámíthatóság különösen hangsúlyosan jelenik meg az olyan helyzetekben, amikor laikus, az adott szakma nyelvét "nem beszélő" célközönség találkozik állami szervek által neki címzett hivatalos / gyakran jogi tartalmú dokumentumokkal (Dobos, 2015; Vinnai, 2018; Tóth, 2019). Az USÁ-ból a múlt században indult Plain Language Movement célja ezen átláthatóság előmozdítása a nem szakértő címzettek számára azáltal, hogy a hivatalos dokumentumok megszövegezéséhez a közérthetőséget segítő (nyelvi) javaslatokat tesz.

A gyakorlatban azonban egy effajta kontroll rendkívül nehéz feladat, amelyet csak az adott területen kompetens szakértők képesek végrehajtani. A jelen tanulmány célja kettős; egyrészről kísérleti jelleggel választ keres arra a kérdésre, hogy a megfelelő tanító adatok birtokában értelmezhetjük-e a közérthető megfogalmazást osztályozási problémaként; erre vonatkozóan az elért eredmények adhatnak támpontot. Másrészről, amennyiben egy ilyen értelmezés kivitelezhetőnek bizonyul, olyan gépi tanult modell készítése, amellyel lehetőség nyílik a szakértői munka támogatására a különös figyelmet igénylő mondatok címkézésével az átfogalmazás előtt álló szövegekben.

A tanulmány a következők szerint épül fel; a 2 fejezetben röviden áttekintjük a "közérthetőség" főbb történeti megközelítéseivel foglalkozó fontosabb szakirodalmat. A 3 és a 4. fejezet ismerteti a kísérletünk alapjául szolgáló korpusz tulajdonságait keletkezését. Ezt követően az 5. fejezet ismerteti az első gépi tanulási eredményeket, amelyeken a 6. fejezet során hiperparaméter hangolással végzünk optimalizálást. A tanulmányt ezt követően rövid konklúzió zárja.

2. Kapcsolódó irodalom

A jogi / hivatalos szövegek közérthetőségét az elmúlt évtizedekben sokan sok különböző nézőpontból vizsgálták. A teljesség igénye nélkül, a történetileg legkorábbi módszerek közül érdemes megemlíteni például az olvashatósági formulák (*readability formulas*) alkalmazását, amelyek a szavak átlagos szótagszámának, valamint a mondatok átlagos szószámának arányából vonnak le következtetéseket a szöveg befogadhatóságára vonatkozóan (Edgar és Jeanne S., 1948; Dubay, 2004; Üveges, 2020). Ezek virágkorukat a múlt század közepén élték, de elterjedtségüket jól jelzi, hogy a MS Word szövegszerkesztőjében (az angol verzió esetében) a mai napig megtalálhatók a szöveget minősítő opcionális indikátorként.

Kognitív szempontból az érthetőség kérdésével a nyelvtudományon belül a pszicholingvisztika foglalkozik behatóan. A pszicholingvisztikai szakirodalom (Pléh és Lukács, 2014; Kas és Lukács, 2012; Pléh, 2013) állásfoglalása sokkal kevésbé mechanikus, egyben nem is feltétlenül jól automatizálható megoldást kínál. Az érthetőséget gátló tényezőket nem a mondat, hanem sokkal inkább a pragmatikai kontextus és a szöveg, mint egység "globális" szintjén helyezi el, és csak kevesebb olyan aspektust azonosít, amelyek a mondat határán belül detektálhatók (ilyen például a főmondat megszakítottsága).

Ebbe a sodorba illeszkedik be a jelen tanulmány keretét adó irányvonal, az USÁ-ban a '70-es években indult Plain Language Movement (PLM) is, amely főleg a kommunikáció hatékonyságának javítása felől közelíti meg a kérdést, és amelynek nyomán az Egyesült Államokban több mint egy évtizede törvény[1] is szabályozza az ottani hivatalok tájékoztató anyagaiban érvényesítendő közérthetőségi szempont betartását (Felsenfeld és mtsai, 1981; Cutts, 1999; Garner, 2001; Willerton, 2015). A dokumentum a közérthetőség feltételeit a következők szerint definiálja: "Világos, tömör, jól szervezett és a témának vagy területnek és a célközönségnek megfelelő egyéb bevált gyakorlatokat követő írás"[2].

Egy másik közkeletű megfogalmazásban egy szöveg akkor közérthető, ha az olvasóját képessé teszi arra, hogy:

- megtalálja, amire szüksége van,
- megértse, amit talál, amikor először olvassa vagy hallja a szöveget,

[1] Public Law 111 - 274 - Plain Writing Act of 2010

[2] Writing that is clear, concise, well-organized, and follows other best practices appropriate to the subject or field and intended audience.

– felhasználja azt, amit talál a szükségletei érdekében[3].

A dominánsan pragmatikai központú meghatározás ellenére a PLM számos gyakorlati megoldást is szorgalmaz, amelyek stiláris-, szintaktikai vagy lexikai átalakítások útján segít(het)ik az olvasót a szöveg könnyebb befogadásában.

Az amerikai kormányzat 2011-ben kiadott szövetségi közérthetőségi iránymutatása, a Federal Plain Language Guidelines[4] talán a legteljesebb összefoglalója ezen javaslatoknak, amely többek között ajánlja cselekvő igék használatát passzív szerkesztés helyett, a nominalizáció kerülését, a funkcióigék túlhasználatának mellőzését vagy éppen az átlagos mondathossz lehetőség szerinti csökkentését.

A magyar szakirodalomban az általános értelemben laikusok nézőpontja főként a "jog és nyelv" kutatások kapcsán került előtérbe. Középpontjában jellemzően az a gondolat áll, miszerint a joghoz való egyenlő hozzáférés alappillére (egyebek mellett) a mindenki számára egyenlően érhető megfogalmazás követelményének érvényre jutása is (Vinnai, 2014; Szabó és Vinnai, 2018a; Minya és Vinnai, 2018).

A jelen tanulmány keretében a vizsgálódás elvi alapját főként a PLM keretrendszerében lefektetett elveknek és ajánlásoknak a széles körben elérhetővé és alkalmazhatóvá tétele motiválja.

3. Gépi reprodukálhatóság

A Plain Language által elvárt stiláris, lexikai jellegzetességeknek, preferált és diszpreferált fogalmazási módoknak ugyanakkor csak egy korlátozott része lehet az, amely a mondat szintjén belül szintaktikai és / vagy lexikai jegyek összességére könnyen lefordítható. Nagyobb részük olyan, a teljes szöveget, vagy éppen az olvasó szempontjait tekintetbe vevő elveket fogalmaz meg, amelyek a hagyományos NLP eszközökkel nem, vagy csak nehezen megragadhatók.

A 2. fejezetben említett elvek az algoritmikus megvalósítás szempontjából sok esetben közvetlenül túlságosan absztraktak, ezek alkalmazása szakértői feladat, amelyhez szükséges a címzettek nézőpontjának és vélhetőleges kontextuális tudásának beható ismerete is. Különösen igaz ez, ha tekintetbe vesszük, hogy más-más hivatali közegben a gyakorlatban eltérő elvek vezethetnek érthetőbb megfogalmazáshoz.

3.1. Felügyelt gépi tanulási módszerek

A felügyelt gépi tanulási módszerek, mint amilyen a Naive Bayes modell (NB), a Support Vector Machine (SVM) vagy éppen a Logistic Regression (LR) számos

[3] "Material is in plain language if your audience can: find what they need, understand what they find the first time they read or hear it, use what they find to meet their needs." Forrás: https://www.plainlanguage.gov/about/definitions/ (Elérés: 2021.12.27.)

[4] Federal Plain Language Guidelines, US Government, March 2011, online: https://www.plainlanguage.gov/media/FederalPLGuidelines.pdf (Elérés: 2021.12.27.)

nyelvtudományi probléma esetében szolgáltattak már hatékony megoldásokat, amennyiben például az adott kérdés lefordítható volt osztályozási problémára például a szentimentelemzésben (Chauhan, 2017) vagy a spam-detekció területén (Sun és mtsai, 2020).

Bár szükségszerű korlátokkal, de hasonló osztályozás elképzelhető lehet a közérthetőség kérdésében is, amennyiben rendelkezünk elegendő mennyiségű tanítóadattal, amelyet szakértők címkéztek fel az adott szakterületnek és címzetti körnek megfelelően "közérthető" és "nem közérthető" címkékkel.

3.2. Nemzeti Adó- és Vámhivatal - Közérthetőségi Program

A Nemzeti Adó- és Vámhivatal Kommunikációs Főosztályának Médiaosztálya immár három éve működtet Közérthetőségi Programot, amelynek keretében három szakértő a hivatal valamennyi kommunikációs anyagát ellenőrzi, és azt javaslatokkal korrektúrázza. Az átvizsgált dokumentumok ezt követően még egy végső szakmai ellenőrzés után kerülnek publikálásra. Ezeken a változatokon ismételt ellenőrzés a Közérthetőségi Program munkatársai részéről már nem történik.

A hivatal munkatársai a jelen kutatáshoz rendelkezésre bocsátották a teljes 2021 -ben korrektúrázott anyagot, amelyben követhetők az újra fogalmazás előtti és a már átírt anyagok. Ennek köszönhetően a fent említett gépi tanulási kísérlet elvben megvalósíthatóvá vált. Fontos kiemelni, hogy a kísérletben felhasznált dokumentumverziók tehát a záró szakmai szempontú ellenőrzés előtti szöveget tartalmazzák.

4. A korpusz rövid bemutatása

Az átírt dokumentumokból minden esetben rendelkezésre állt tehát két változat; egy eredeti és egy szakértők által korrektúrázott változat; előbbire a továbbiakban `eredeti` -ként, míg utóbbira `átfogalmazott`-ként utalunk. Mivel a jelen kutatás nagyban kapcsolódik a PLM által propagált elvekhez, így érdemes megemlíteni, hogy a szakértőkkel folytatott beszélgetés, valamint az általuk a NAV munkatársai számára készített segédletek, belső kommunikációs anyagok alapján az átfogalmazáskor alkalmazott elvek túlnyomó többsége metszetet képez a PLM elvárásaival (pl. az olvasó szempontjainak középpontba helyezése, a funkcióigék kerülése stb.).

A fentieket alapul véve a keletkezett korpusz ideális választásnak tűnik annak vizsgálatára, hogy egy olyan absztrakt és nehezen körülhatárolható fogalom, mint a "közérthetőség" megragadható-e valamilyen szinten korpusznyelvészeti és gépi tanulási módszerek alkalmazásával is?

Az eredeti korpuszt a következők szerint osztottuk részekre:

– minden dokumentumból rendelkezésre állt `eredeti`, és `átfogalmazott` verzió is. Az összetartozást a fájlok elnevezési konvenciója kódolta; a közös, korpuszban egyedi prefix után "A" jelölte az `eredeti`, "B" pedig az `átfogalmazott` szöveget (pl.: A44**A** és A44**B**),

– a dokumentum-párokat a spaCy (Honnibal és Montani, 2017) természetesnyelvi elemző segítségével és az Orosz György által készített nyelvmodellt[5] alkalmazva mondatokra szegmentáltuk,
– az így kapott mondatok közül eltávolítottuk azokat, amelyek a dokumentumpár mindkét tagjában azonosan fellelhetők voltak.

A megmaradt mondatok a szülő fájlok fájlneveinek posztfixei ("A" vagy "B") alapján kerültek besorolásra vagy az `eredeti`, vagy az `átfogalmazott` szövegek részkorpuszába.

A válogatás eredményképpen 5710 mondat került az `eredeti`, és 3010 mondat az `átfogalmazott` részkorpuszba, összesen több mint 270 ezer token terjedelemben (az adatokat részletesen az 1. táblázat szemlélteti[6]).

Szófaj	**Átfogalmazott**	**Átfogalmazás előtti**	**Arány**	
token	85754	186900	**0,3145**	**0,6854**
NOUN	25649	57792	0,2991	0,3092
ADJ	14293	32716	0,1667	0,175
PRON	2634	5929	0,0307	0,0317
CONJ	3386	6784	0,0395	0,0363
NUM	1502	3763	0,0175	0,0201
VERB	5543	11579	0,0646	0,062
ADV	3521	7289	0,0411	0,039
PROPN	2826	5245	0,033	0,0281
ADP	1238	3086	0,0144	0,0165
AUX	0	0	0	0
DET	9874	20578	0,1151	0,1101
INTJ	13	42	0,0002	0,0002
PART	241	558	0,0028	0,003
PUNCT	13681	28545	0,1595	0,1527
SCONJ	1151	2566	0,0134	0,0137
SYM	60	97	0,0007	0,0005
X	142	331	0,0017	0,0018

1. táblázat. Az egyes részkorpuszok szófaji statisztikái

A két részkorpusz méretének eltérését főként az a jelenség okozza, hogy az `eredeti` szövegek mondatai első lépésben sok esetben jelentősen rövidültek (például a jogszabály hivatkozások lábjegyzetbe utalása, vagy a nem elengedhetetlen kifejezések eltávolítása miatt), amely által több, korábban különálló mondat

[5] `https://github.com/spacy-hu/spacy-hungarian-models`

[6] token - tokenszám, ADJ – melléknév, ADV – határozószó, ADP – névutó, AUX – segédige, CONJ – mellérendelő kötőszó, DET – névelő, INTJ – indulatszó, NOUN – főnév, NUM – számnév, PART – igekötő, PRON – névmás, PROPN – tulajdonnév, PUNCT – központozás, SCONJ – alárendelő kötőszó, SYM – szimbólum, VERB – ige, X – egyéb

összeolvasztására volt lehetőség az átalakítás során. Érdemes megemlíteni, hogy az `átfogalmazott` mondatok átlagos tokenszáma még ezzel együtt is alacsonyabb (28,48), mint az `eredeti` mondatoké (32,73) amely jól harmonizál a 2. fejezetben megfogalmazottakkal.

Az átalakításokra néhány példát az (1) - (2), valamint a (3) - (4) párok szolgáltatnak, melyek esetében a pár első tagja az `eredeti`, a második pedig az `átfogalmazott` változat.

(1) A Bevezető felületen található tájékoztató szövegek segítik a felhasználót az alkalmazás megismerésében. Az alkalmazás megismerését a felhasználó a Bevezető felületen található Tovább funkciógombra való koppintással tudja elkezdeni, valamint folytatni.

(2) A Bevezető tájékoztató szövegei segítik a felhasználót az alkalmazás megismerésében. Ezt a felhasználó a Bevezető felületen található Tovább funkciógombra koppintva tudja elkezdeni.

A (3) - (4) jó példa arra, amikor a szövegből a funkcióige ("kerül") eltávolításával a szerkezet egyszerűsíthető. A funkcióigés konstrukciók esetében a jelentést voltaképpen a kifejezés névszói tagja hordozza, a szerkezet egésze pedig (kivéve terminus technicusok esetében) helyettesíthető egyetlen igével, mint a "bemutatásra kerül" (=bemutatják), vagy a "ellenállást tanúsít" (=ellenáll) esetében (Lanstyák, 2020).

(3) A felugró ablakban a Mégsem funkciógombra való koppintás hatására az eredetileg elmentett összekapcsolás megmarad és **nem kerül felülírásra** az új összerendeléssel.

(4) A felugró ablakban a Mégsem funkciógombra koppintva az eredetileg elmentett összekapcsolás megmarad és az új összerendelés **nem írja felül**.

A korpusz jellemzői kapcsán lényeges információ lehet annak hasonlósága más jogi / hivatalos doménbeli szövegekhez. Ehhez az összevetéshez a Miskolc Jogi Korpusz (MJK) részkorpuszainak szöveganyagát használtuk fel, a hasonlóság kifejezésére pedig a szókincs alapú Jaccard-távolságot (Hancock, 2004) alkalmaztuk. A részkorpuszonkénti összevetés amiatt lehet szükséges, mivel a MJK, bár domén tekintetében közös szövegeket tartalmaz, mégsem tekinthető egységesnek; a jogi fórumok szövegei például inkább a köznyelv, míg a törvényszövegek a formális jogi nyelvhasználat prototipikus esetei felé tendálnak (Szabó és Vinnai, 2018b).

Jaccard távolság szerint az `eredeti`, és a már `átfogalmazott` szövegek közötti különbség azonnal szembeötlő; a metrika 0,61-es értéke a szókincs jelentős eltérését mutatja. A MJK részkorpuszaival vett hasonlóságokat a 2. táblázat mutatja be (a feltüntetett összevetések: fórum - jogi fórumok szövegei, átirat - bírósági tárgyalásokon és rendőrségi kihallgatásokon készített átiratok, jogszabályok - jogszabályok szövegrészletei, ítéletek - bírósági és törvényszéki ítéletek, metanyelv - jogi tankönyvek, miniszteri indoklások szövege, kódexjog - törvények szövegei).

	Átfogalmazott	Eredeti
fórum	**0,85**	0,83
átirat	0,83	0,83
jogszabályok	0,82	0,79
ítéletek	0,82	0,79
metanyelv	0,81	0,77
kódexjog	0,76	0,74

2. táblázat. Az `átfogalmazott` / `eredeti` részkorpuszok lexikai hasonlósága a MJK egyes részkorpuszaival.

Az eredmények alapján mindamellett, hogy az `átfogalmazott` szövegek szókincse jelentősen eltávolodott az `eredeti` verziótól, eközben egy kivétellel a MJK valamennyi részkorpuszához minimálisan közeledett (kivételt ez alól az átiratok szövegei képeznek, amelyek azonban beszélt nyelvi közlések leiratai, ilyenformán maguk is valamelyest önálló csoportot alkotnak).

Fontos azonban kiemelni, hogy mind az `eredeti`, mind az `átfogalmazott` szövegek a MJK leginkább "köznyelvi"-nek tekinthető részeihez a leghasonlóbbak, vagyis a fórumok, és az átiratok szövegeihez.

5. Gépi tanítási kísérlet

Az előfeldolgozás standard lépéseket foglalt magában: kiszűrtük a stopszavakat, eltávolítottuk a mondatokból a számjeggyel írt számokat, kisbetűsítettük az egyes szavakat, valamint a spaCy segítségével lemmatizáltuk azokat, majd az így kapott, normalizált mondatokat (amely a szótárat és az IDF súlyokat a tanító / validációs halmazon tanulta) TF-IDF vektorizáltuk (uni- és bigramokat is megengedve).

A továbbiakban három gépi tanulási algoritmust alkalmaztunk; a korábban már említett (Bernoulli) Naive Bayes (NB), Support Vector Machine (SVM) és Logistic Regression (LR) módszereket. Első közelítésben alapértelmezett paraméterezéssel kíséreltük meg a tanítást. A kapott eredményeket a 3. táblázat szemlélteti, ahol az első oszlop a választott módszert, a második pedig a szövegosztályt mutatja. A tanító és teszthalmazok felosztásakor a 90% - 10% arányt választottuk.

Mindhárom módszer esetén az `eredeti` mondatok felismerése volt hatékonyabb, a legjobb eredményt e tekintetben a LR hozta; itt az F értéke 0,77 volt. A leginkább kiegyensúlyozott eredmény ezzel szemben a SVM esetében volt megfigyelhető; itt volt a legkisebb eltérés a két kategória F-értékei között.

6. Hiperparaméter-optimalizálás

A fenti eredmények arra engedtek következtetni, hogy a korpuszban rendelkezésre állhat elegendő tanítóadat jobb eredmények eléréséhez is. A hatékonyság

		Pontosság	Fedés	F-érték
NB	Eredeti	0,64	0,81	0,71
	Átfogalmazott	0,25	0,12	0,16
SVM	Eredeti	0,67	0,63	0,65
	Átfogalmazott	0,37	0,42	0,39
LR	Eredeti	0,68	0,87	0,77
	Átfogalmazott	0,49	0,23	0,31

3. táblázat. Az kipróbált gépi tanuló algoritmusok teljesítménye alapbeállításokkal.

maximalizálása érdekében kísérletet tettünk rá, hogy az egyes modellek hiperparamétereit hangoljuk.

A rendelkezésre álló korpuszt a későbbiekben tanító-, validációs- és teszthalmazra bontottuk 80%-10%-10% arányban. Az optimalizálást a validációs halmazon végeztük, illetve az adatokon 10-szeres keresztvalidációt alkalmaztunk. A korpusz címkéinek kiegyensúlyozatlansága miatt a modelleken minden esetben a `class_weight='balanced'` beállítást használtuk.

6.1. SVM

A gépi tanuló algoritmusok általánosan kétféle paraméter készlettel rendelkeznek; a modell paraméterek tanulása (vagy becslése) a gépi tanulási folyamat része, a hiperparaméterek viszont közvetlenül nem tanulhatók a rendelkezésre álló adatokból, emiatt azok optimalizálása kézzel történhet meg. Ez utóbbi tulajdonságuk ellenére optimális beállításuk nagyban befolyásolhatja az algoritmus teljesítményét.

SVM esetében ilyen hiperparaméterek lehetnek például a C (regularizáció - amely az egyes osztályokat elválasztó hipersík margin-jának nagyságát szabályozza), a Gamma (amely arra hat ki, hogy a potenciális szeparáló határ megválasztásakor attól mennyire távoli pontok játszanak még szerepet a döntésben) illetve a kernel, avagy magfüggvény (amely a hipersík kiszámítási módjának egyenletét határozza meg). A jelen kísérletben a következő lehetséges értékek valamennyi (összesen $4^3 = 64$) kombinációjával tanítottunk modelleket, majd ezeket értékeltük ki:

- C: [0,1, 1, 10, 100]
- *gamma*: [1, 0,1, 0,01, 0,001]
- *kernel*: ['rbf', 'poly', 'sigmoid', 'linear'].

Tekintettel arra, hogy az esetleges gyakorlati alkalmazás során fontosabbnak tartottuk biztosítani, hogy a modell által problémásnak ítélt mondatok valóban megvizsgálandók legyenek a szakértők által, mint azt, hogy nagy számú mondatot jelöljünk ellenőrzésre, ezért az optimalizálás során az `eredeti` szövegosztály predikcióinak pontosság (Precision) értékét igyekeztünk maximalizálni.

Mindezen feltételek mellett a legjobb eredményt {C=100, $Gamma$=0.001, $kernel$='rbf'} választással értük el (a legjobb 5 modell eredményeit a 4. táblázat mutatja be).

Paraméterek: [C,Gamma,kernel]	**(Átlagos-) pontosság**	**Szórás**
100, 0,001, rbf	**0,767**	0,02
10, 0,01, rbf	0,765	0,03
1, 1, sigmoid	0,755	0,02
1, 0,1, rbf	0,752	0,03
0,1, 1, sigmoid	0,736	0,03

4. táblázat. A legjobb eredményt (átfogalmazandó mondatok - pontosság) adó beállítások (SVM).

6.2. LR

A SVM-mel ellentétben Logisztikus Regresszió esetén kevéssé ismert, hogy a lehetséges hiperparaméterek változtatása garantáltan javítaná a predikciós teljesítményt. Ennek ellenére (tekintettel arra, hogy a kiinduló modellek közül alapbeállításokkal ez érte el a legígéretesebb eredményt) itt is kísérletet tettünk a következő beállítások használatával:

- C: [0,01, 0,1, 1, 10, 100, 1000]
- *solver*: ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'].

ahol C ismét a regularizációért felelt, a *solver* pedig a probléma megoldásáért felelős algoritmus kiválasztását tette lehetővé.

Az eredmények értékelése során megfigyeltük, hogy a C érték növelésével (ahogy az várható) az átfogalmazandó szövegosztály predikciói esetében a pontosság valamelyest növekedett. A különböző solverek közül a tapasztalatok alapján, C bármely választása mellett teljesítmény tekintetében az alábbi (az 5. táblázatban is megfigyelhető) sorrend alakult ki:

'lbfgs' > 'newton-cg' > 'liblinear' > 'sag' > 'saga'

A LR esetében, továbbra is az `eredeti` szövegosztály predikcióinak pontosságát tekintve elsődlegesnek a {C=100, *solver*='newton-cg'} paraméter választás bizonyult a leghatékonyabbnak.

A kapott adatsor alapján általánosan is igaz, hogy a C változása a LR esetében függött össze jelentősen a magasabb pontosság értékek elérésével, míg a SVM esetében inkább a megfelelő kernel megválasztása bizonyult döntő fontosságúnak. Az adatokból emellett az látszik, hogy a SVM esetében az 'rbf', valamint a 'sigmoid' kernelek teljesítettek a legjobban.

Paraméterek: [C,solver]	**(Átlagos-) pontosság**	**Szórás**
1000, lbfgs	**0,705**	0,02
1000, newton-cg	0,703	0,02
1000, liblinear	0,703	0,02
1000, sag	0,701	0,02
1000, saga	0,7	0,02

5. táblázat. Metrikák az optimálisnak tekintett paraméterek mellett (LR).

6.3. Kiértékelés

Az ilyen módon optimalizált modell-paramétereket ezt követően az eredeti teszt halmazon ismét kiértékeltük. Ennek kapcsán a következő eredmények születtek:

		Pontosság	**Fedés**	**F-érték**
SVM	Eredeti	**0,78**	0,60	0,68
	Átfogalmazott	0,48	0,68	0,56
LR	Eredeti	0,70	0,76	0,73
	Átfogalmazott	0,46	0,39	0,42

6. táblázat. Az optimalizált paraméterek mellett tanított modellek teljesítménye az eredeti teszt halmazon kiértékelve.

Összességében elmondható, hogy a megfelelő beállításokkal mintegy 11% nyerhető az `eredeti` szövegosztály optimalizálás nélküli pontosság értékeihez képest SVM használatával. A LR a teszthalmazon 2%-kal jobban teljesített mint eredetileg, ez azonban még a keresztvalidáció során megfigyelt szóráson belül van. Érdekes tendencia, hogy míg a hiperparaméter hangolás SVM esetében egyértelműen pozitívan hatott a teljes osztályozás sikerességére (mindkét szövegosztály F-értéke nőtt), addig a LR esetében ez a hatás aszimmetrikusan jelent meg; az `eredeti` szövegosztály predikcióinak pontossága ugyan nőtt, de F-értéke csökkent, míg a már `átfogalmazott` szövegek esetében a fedés nagy mértékű javulása magával húzta az F-érték pozitív elmozdulását is.

A tévesen klasszifikált mondatok esetében még további vizsgálatot igényel, hogy milyen jellemző okok állnak azok hátterében. Néhány jellemző példa a tévesen átfogalmazandónak ítélt mondatokra[7]:

(5) Az Ön által képviselt adózótól 2021. augusztus 26. és szeptember 1. között beérkezett adatok azt valószínűsítik, hogy egyes számlával (egyszerűsített

[7] Tekintettel arra, hogy az algoritmus által kapott normalizált és lemmatizált alak kevéssé jól olvasható, itt a megfelelő korpuszbeli "eredeti" mondatot közöljük.

számlával) bizonylatolt értékesítéseiről pénztárgépes nyugtát is állított ki.

(6) Ha nem indul el a telepítőprogram, akkor a JAR állományok futtatása és hozzárendelése a java futtatási környezethez című dokumentáció nyújthat segítséget a hiba megoldásában.

A második mondat esetében megjegyzendő, hogy az instrukció az ÁNYK (Általános Nyomtatványkitöltő) alkalmazás telepítéséhez ad segítséget.

A következő példák két (tévesen) `átfogalmazott`-nak ítélt mondatot szemléltetnek:

(7) Abban az esetben, ha az utas nem nyilatkozik illetve hamis, pontatlan vagy hiányos információkat ad meg, a vámhatóság a készpénzt lefoglalhatja vagy visszatarthatja illetve az utas büntetéssel sújtható.

(8) A személyes közreműködés módját és ellentételezését a szövetkezet tagjának a szövetkezettel kötött tagsági megállapodása tartalmazza.

A tévesen osztályozott mondatok jól rávilágít(hat)anak a tanulmányban tárgyalt módszerek korlátaira és arra a körülményre, hogy a "közérthetőség" jelentős mértékben nem csak a dokumentum szóhasználatán keresztül, de a szakértők által ismert és tekintetbe vett pragmatikai kontextusban érhető tetten. Ezekben a konkrét esetekben például a téves osztályozás magyarázata az lehet, hogy a mondatok stílusukban valóban párhuzamba állíthatók a már `átfogalmazott` dokumentumokkal.

7. Konklúzió

Cikkünkben kísérletet tettünk arra, hogy kézzel készített korpusz felhasználása mellett felügyelt gépi tanulással kíséreljünk meg szövegeket osztályozni azok közérthető jellege szerint. A szakértők által korábban átfogalmazott mondatok jellegzetességeit a TF-IDF vektorizálással alakítottuk a gépi tanuló algoritmusok számára értelmezhető bemenetté. A cél egy olyan modell készítése volt, amely alkalmas lehet a szakértők figyelmét felhívni azon mondatokra egy-egy hivatalos szövegben, amely további megfontolást érdemel a szöveg közérthetőre alakítása során, ezzel gyorsítva a szakértői munkát.

A közérthetőség nehezen definiálható mibenléte ellenére az eredmények alapján úgy tűnik, hogy megfelelő mennyiségű és minőségű tanítóadattal a probléma valamelyest kezelhető gépi tanulás segítségével.

További kutatást igényel, hogy az elért eredmények csak egy szűk szövegrétegen belül (a NAV tájékoztató anyagai) vagy általánosságban, a laikusoknak címzett hivatalos szövegek esetében is alkalmazható-e. A jelen cikkben (annak pilot jellege miatt) nem tettünk kísérletet fejlettebb módszerek (LSTM, BERT stb.) alkalmazására, azonban az eredmények tükrében ilyen kísérlet is indokolt lehet.

Szintén megoldandó feladat, hogy a TF-IDF vektorizáció korlátain kívül milyen egyéb jegyek figyelembevétele (szórend, szintaktikai sajátosságok stb.) lehet fontos egy ilyen jellegű osztályozási probléma esetén, illetve egy ehhez illeszkedő, összetettebb, a nyelvi megformáltságot jobban tekintetbe vevő vektorizálási forma kidolgozása is indokolt lehet.

Köszönetnyilvánítás

Az Innovációs és Technológiai Minisztérium ÚNKP-21-3 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

Hivatkozások

Chauhan, P.: Sentiment analysis: A comparative study of supervised machine learning algorithms using rapid miner. International Journal for Research in Applied Science and Engineering Technology pp. 80–89 (2017)

Cutts, M.: The Plain English Guide. Oxford University Press (1999)

Dobos, C.: Nyelven belüli fordítás és tisztességes jogi eljárás. In: Szabó, M. (szerk.) A jog nyelvi dimenziója, pp. 215–226. Bíbor Kiadó, Miskolc, Magyarország (2015)

Dubay, W.H.: The Principles of Readability. Costa Mesta: Impact Information (2004)

Edgar, D., Jeanne S., C.: A formula for predicting readability. Educational Research Bulletin 27, 11–20 (1948)

Felsenfeld, C., Cohen, D.S., Fingerhut, M.: The plain english movement in the united states: Comments. Canadian Business Law Journal 6, 408–452 (1981)

Garner, B.A.: Legal Writing in Plain English. University of Chicago Press: Chicago (2001)

Hancock, J.: Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient) (10 2004)

Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear

Kas, B., Lukács, A.: Processing relative clauses by hungarian typically developing children. Language and Cognitive Processes 27, 500–538 (2012)

Lanstyák, I.: LanstyÁk istvÁn a funkcióigés szerkezetek néhány általános kérdéséről istvÁn lanstyÁk on some general questions of light verb constructions 21, 61–91 (3 2020)

Minya, K., Vinnai, E.: Hogyan írjunk érthetően? kilendülés a jogi szaknyelv komfortzónájából. Magyar Jogi Nyelv pp. 13–18 (2018)

Pléh, C.: A lélek és a nyelv. Akadémiai Kiadó: Budapest (2013)

Pléh, C., Lukács, A.: Pszicholingvisztika. Akadémiai Kiadó: Budapest (2014)

Sun, N., Lin, G., Qiu, J., Rimba, P.: Near real-time twitter spam detection with machine learning techniques. International Journal of Computers and Applications 0(0), 1–11 (2020)

Szabó, M., Vinnai, E.: A törvény szavai. Miskolc, Bíbor Kiadó (2018a)
Szabó, M., Vinnai, E.: A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25. Miskolc, Magyarország : Bíbor Kiadó (2018b)
Tóth, J.: Tudnak-e a jogászok érthetően fogalmazni, avagy nem is kell azt tudni? Magyar Jogi Nyelv pp. 31–37 (2019)
Vinnai, E.: A magyar jogi nyelv kutatása. Glossa Iuridica p. 29–48 (2014)
Vinnai, E.: Megértette a figyelmeztetést? a figyelmeztetések és tájékoztatások közlése a büntetőeljárásokban. In: Szabó, M., Vinnai, E. (szerk.) A törvény szavai : Az OTKA-112172 kutatási zárókonferencia anyaga Miskolc, ME – MAB, 2018. május 25., pp. 281–295. Bíbor Kiadó, Miskolc, Magyarország (2018)
Willerton, R.: Plain Language and Ethical Action - A Dialogic Approach to Technical Content in the 21st Century. Routledge: New York (2015)
Üveges, I.: Automatizálható a közérthető megfogalmazás? Jog, számítógépes nyelvészet és pszicholingvisztika találkozása. Magyar Jogi Nyelv 4, 1–8 (2020)