Szövegaugmentálási módszerek összehasonlítása politikai szövegek szentimentanalízise során

Üveges István[1,2], Csányi Gergely Márk[1,3], Ring Orsolya[2], Orosz Tamás[1]

[1]MONTANA Tudásmenedzsment Kft.
[2]Társadalomtudományi Kutatóközpont, Politikatudományi Intézet
[3]Budapest Műszaki és Gazdaságtudományi Egyetem
{uvegesi,csanyi.gergely,orosz.tamas}@montana.hu
ring.orsolya@tk.hu

Kivonat Cikkünkben bemutatjuk a gépi tanítási feladatokban gyakran előforduló kiegyensúlyozatlan tanítóhalmaz probléma egy lehetséges megoldását az alacsony elemszámú kategóriák szöveg-augmentálásával. Az összevethetőség érdekében egyszerű szövegaugmentálási technikákkal (EDA) és egy szóvektor alapú módszerrel is kísérletet teszünk. A módszerek hatékonyságát politikai doménbe tartozó szövegek szentimentelemzési feladatán teszteljük, amihez a TK-MILAB szentiment korpusz egy kisebb szeletét használjuk. Az alulreprezentált kategória bővítésével elért eredményeket a kiváltott F-érték változás függvényében értékeljük.
Kulcsszavak: Easy Data Augmentation, kiegyensúlyozatlan tanítóhalmaz, emóció elemzés

1. Bevezetés

A gépi tanuláson alapuló feladatok esetében, amelyek például szövegosztályozást kísérelnek megvalósítani, gyakran előforduló probléma az adatok kiegyensúlyozatlansága a tanítóhalmazban (Kubat és mtsai, 1997; Menardi és Torelli, 2014). Ilyen esetben a különböző címkék aránya erősen aszimmetrikus, egyes osztályok jelentősen alulreprezentáltak, ami megnehezíti az ilyen osztályok jó hatásfokú predikcióját.

Ez a probléma felmerül magyar nyelvű politikai szövegek felügyelt gépi tanuláson alapuló klasszifikálása során is. Az Comparative Agendas Project[1] keretében zajló klasszifikálás során az egyes megfigyeléseket szokásosan 21 közpolitikai osztályba sorolják. A kézzel képzett tanítóhalmaz minden esetben kiegyensúlyozatlan, aminek a gépi osztályozás során (Sebők és Kacsuk, 2021) egy lehetséges megoldása például az ún. bináris hólabda megközelítés (*binary snowball approach*) alkalmazása, melynek során bináris választások sorozatává egyszerűsítjük a többosztályos osztályozást.

Az ilyen kiegyensúlyozatlan adatbázis-struktúrák általában is jellemzik az összehasonlító politikatudományi elemzésekre használt korpuszokat. Ezekben az

[1] https://cap.tk.hu/

esetekben a legnagyobb problémát az osztályozás során a nagy pontosság mellett a magas recall arány elérése jelenti (Kumar és Gopal, 2010) különösen szupport vektor gép (Support Vector Machine, SVM) használatakor. A probléma kezelésének egyik módja, ha teljesen figyelmen kívül hagyjuk az alulreprezentált kategóriát, ezzel azonban a kutatás szempontjából értékes megfigyeléseket veszíthetünk. Hogy ezt elkerüljük, az egyik lehetőség a véletlenszerű mintavételezés (*random sampling*) túl- vagy az alulmintavételezéssel, ezzel kompenzálva a korpusz belső egyensúlytalanságát. A véletlenszerű mintavételezés hátránya azonban, hogy megnöveli a túlillesztés esélyét (Lango és Stefanowski, 2018; Nguyen és mtsai, 2011). A nemzetközi politikatudományban emellett az elmúlt években jelentősen megnőtt a nagyméretű (például parlamenti szövegekből, jogszabályokból vagy politikai hírekből álló) korpuszok elemzésén alapuló kutatások száma is. Mindez magával hozta az igényt a különböző gépi klasszifikálási feladatok hatékonyságának növelésére, ami szükségessé teszi a kiegyensúlyozatlan tanulóhalmazok problémájának megoldását (Hillard és mtsai, 2008; Breeman és mtsai, 2009; Burscher és mtsai, 2015).

Napjainkban ugyancsak kiemelkedő jelentőségű kutatási feladat a különböző szövegek gépi tanuláson alapuló szentiment- illetve emóció-klasszifikálása (Van Atteveldt és mtsai, 2008; Bhowmick és mtsai, 2009; Jia és mtsai, 2009; Young és Soroka, 2012; Dadgar és mtsai, 2016). Mivel a politikai doménbe tartozó szövegek emóciótartalma ugyancsak különösen kiegyenlítetlen, ezen a területen kiemelkedően fontos a tanítóhalmaz kiegyensúlyozása.

A már említetteken kívül az adatok kiegyensúlyozatlanságára megoldást jelenthet a rendelkezésre álló adatok augmentálása is, ami a meglevő példák alapján új példányok készítését jelenti a tanítóhalmazba. Ez a technika a gépi látás területén már régóta bevett eljárásnak számít (Zhang és mtsai, 2015; Fawzi és mtsai, 2016; Taylor és Nitschke, 2018), később néhány alapvető ötletet is innen merített az NLP területe (Fadaee és mtsai, 2017; Wei és Zou, 2019; Csányi és Orosz, 2021).

Tanulmányunk célja, hogy különböző szöveg augmentálási technikákat próbáljunk ki politikai doménbe tartozó mondat szintű szövegek szentiment osztályozhatóságának javítása érdekében. Tekintettel a kutatás pilot jellegére, ezt két kiválasztott szentiment osztály példáján keresztül mutatjuk be. Véleményünk szerint ugyanakkor a bemutatott eredmények hasznos tanulságokkal szolgálhatnak bármely kiegyensúlyozatlan tanítóhalmazzal rendelkező gépi tanítási feladat esetében. A különböző augmentálási technikák hatékonyságát háromféleképpen előfeldolgozott szövegen hasonlítottuk össze. Az itt bemutatott módszereket a `digital-twin-distiller`[2] keretrendszerbe integráltuk. A cikkben bemutatott eszközök és a szentimentanalízis során betanított modellekből készített applikáció is megtalálható a GitHubon[3].

[2] digital-twin-distiller projekt elérhetősége a GitHubon: https://github.com/montana-knowledge-management/digital-twin-distiller

[3] Politikai témájú szövegek mondatszintű szentiment analízise projekt keretében készült szentiment felismerő eszközök: https://github.com/montana-knowledge-management/hungarian-political-sentiment-analysis

2. TK-MILAB szentiment korpusz politikai doménre

2.1. Annotálási elvek

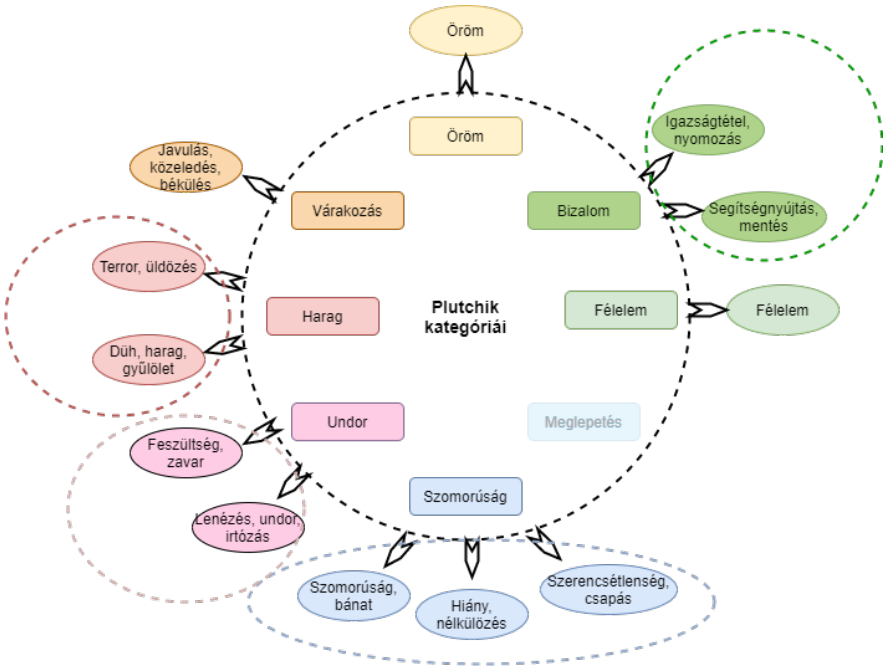
Vizsgálatunk kezdetekor a TK-MILAB „Doménspecifikus szentimentelemzési-eljárás kidolgozása magyar nyelvű szövegek elemzésére" részprojektje keretében készülő[4] korpuszban mintegy 5700 mondatnyi kézzel annotált szöveg volt elérhető. A korpusz 12 emóció-kategória szerint osztályozott mondatokat tartalmaz, és benne minden mondat pontosan egy emóció kategóriára jellemző címkével van ellátva. A kategorizálás során először induktív módon a szövegből kiindulva kerültek meghatározásra az egyes emóciókategóriák, melyek szükség esetén aggregálhatók Plutchik (Kellerman és Plutchik, 1968) emóció-kategória rendszerére, amely nyolc osztályt különböztet meg (lásd 1. ábra).

Erre a kibővített rendszerre azért volt szükség, mert a politikai hírszövegekben található mondatok egyébként nem vagy csak rendkívül alacsony annotátori egyetértés mellett voltak besorolhatóak a Plutchik-féle („hagyományos") kategóriákba. Egy jó példa lehet erre a következő mondat: „Egy ember összeesett az utcán, de a járókelők megmentették az életét.". Itt az első tagmondatnak a „Szomorúság", míg a másodiknak a „Bizalom" az alapérzelme. Az egész mondatot azonban nem lehet egyértelműen besorolni az egyik vagy a másik kategóriába az eredeti, Plutchik-féle érzelemkerék alapján. A bővített, TK-MILAB projekt keretében kidolgozott rendszer segítségével az előbbi mondat egyértelműen besorolható a „Segítségnyújtás, mentés" kategóriába, mely azután Plutchik-féle rendszer „Bizalom" kategóriájára aggregálható. A bővített rendszer segítségével a korpusz magas kódolók közötti egyetértéssel volt annotálható (Ring és mtsai, 2021).

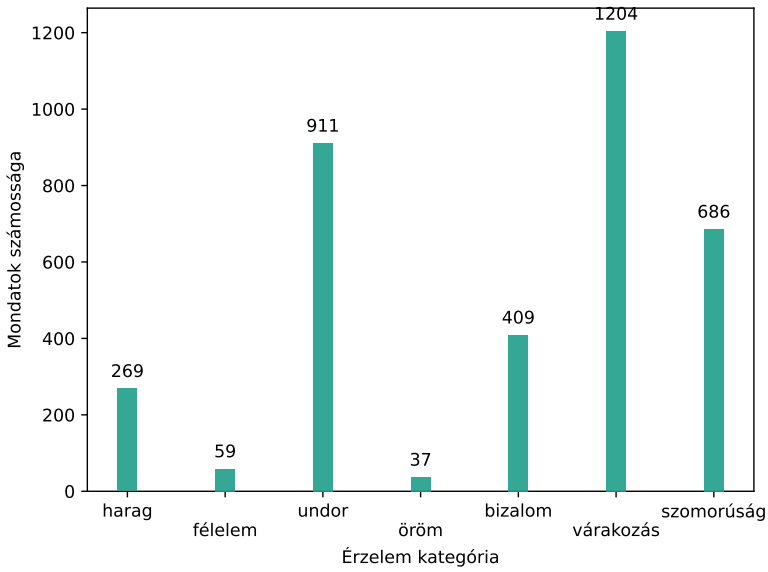
2.2. Annotálási eljárás

A mondatok címkézését három annotátor végezte. Elsőként két független annotátor egy-egy címkét rendelhetett minden mondathoz, majd ezt követően egy szakértő kiválasztotta, illetve validálta a korpuszba kerülő címkét. A kész korpuszban tehát minden egyes mondat pontosan egy emóció kategóriához tartozhatott. Minden érzelmi kategória leírható egy fő -, illetve egy, a TK-MILAB korpuszban feltüntetett mellékkategóriával. Az iménti példamondat eszerint tehát a „Bizalom", azon belül pedig a „Segítségnyújtás/mentés" kategóriába sorolható. A korpuszban szereplő mondatok közül csak azokat használtuk fel az augmentáláshoz, amelyek esetében a két annotátor egyetértésben adta ugyanazt a címkét. Ennek célja az volt, hogy a gépi tanításhoz olyan adatokat használjunk fel, amelyek emberi megítélés szerint a lehető leginkább mentesek a bizonytalanságtól. Az így előállt kategóriacsoportok mérete nagyon eltérően alakult (lásd 2. ábra). A „Várakozás" kategóriába került a legtöbb mondat, szám szerint 1204, míg az „Öröm" -be a legkevesebb, összesen 37 darab.

[4] TK-MILAB projekt elérhetősége: https://milab.tk.hu/domenspecifikus-szentimentelemzesi-eljaras-kidolgozasa-magyar-nyelvu-szovegek-elemzesere



1. ábra. Plutchik kategóriái és a belőlük képzett emóció kategóriák a TK-MILAB korpuszban.



2. ábra. Plutchik kategóriáira aggregált címkéjű mondatok számossága a TK-MILAB korpuszban.

2.3. Összehasonlításhoz használt tanítóadat-halmazok

A szövegaugmentálási technikák segítséget nyújthatnak ahhoz, hogy a kis elemszámmal rendelkező kategóriákat a meglévő mondatokból előállított szintetikus mondatokkal felbővítsük. Ahhoz, hogy a különböző technikákat tesztelni tudjuk, a különböző módszerek összehasonlításához a rendelkezésre álló kategóriák közül azt a kettőt választottuk ki, amelyek a legnagyobb számossággal bírtak a leválogatott korpuszban („Várakozás" és „Undor"). A vizsgálatokhoz bináris osztályozókat készítettünk különböző méretű tanítóadat-halmazokon. Ezek a bináris osztályozók „Várakozás" vagy „Undor" értékeket vehették fel, hiszen csak az ezeknek megfelelő mondatokat tartalmazta a tanításhoz használt korpusz. A célunk ezzel az volt, hogy megvizsgáljuk, hogy az eredeti adatok felhasználásával milyen hatékonyságú osztályozás érhető el. Az így kapott értékeket a későbbiekben kevesebb eredeti adaton, adott mennyiségű augmentált adat hozzáadása mellett készült modellek értékeivel vetettük össze.

Ehhez ugyanazt a modellt különböző méretű tanítóadat-halmazokon tanítottuk föl. Elsőként a két kiválasztott kategória 900 - 900 mondatából kialakított szeletén hajtottunk végre gépi tanítást, a `scikit-learn`[5] lineáris kernelű szupport vektor gép modelljével. A tanításokat - a kiegyensúlyozatlan minták hatását csökkentve - a class_weight="balanced" beállítással végeztük, a többi paramétert alapértelmezettként hagytuk. A tanítást minden esetben tf-idf vektorokon (Luhn, 1957; Jones, 1972) végeztük el, uni- és bigramokat is figyelembe véve a vektorokban, a többi paramétert szintén alapértelmezetten hagytuk.

Az „Undor" kategória mondatait használtuk később augmentálásra. A választott részhalmazok számosságai a következők szerint alakultak: 10, 25, 50, 100, 250, 500 (ezekre az értékekre a későbbiekben az augmentálás *bázisa*-ként hivatkozunk). A „Várakozás" kategória minden tanítás esetén 720 mondatnyi tanító adatot tartalmazott. A referenciaként betanított modellek ennek megfelelően nem kiegyensúlyozott tanító adatot kaptak bemenetként (pl. 10 db „Undor", 720 db „Várakozás" mondatot). Az augmentáláshoz használt modellek vizsgálata során az „Undor"-ban lévő mondatok számosságát mindig 720-ra augmentáltuk. Az augmentálás arányát($\frac{n_{aug}}{n}$, ahol n az augmentálás bázisa, n_{aug} pedig az augmentált adat számossága), valamint a kapott korpuszokba bekerülő augmentált mondatok számosságait az 1. táblázat szemlélteti. Az ilyen módon kiválasztott (összesen minden esetben 720 mondatnyi) „Undor" címkével annotált mondat mellett konstans 720 mondatnyi „Várakozás" címkéjű mondatot választottunk ellenpéldaként. Így kiegyenlített tanítóhalmazon történt a modellek tanítása. A kiértékelést minden modell esetében emóciókategóriánként 180 (összesen tehát 360) mondaton végeztük el. Az azonos méretű tanítóhalmazon augmentált adatok F_1 értékeinek javulását hasonlítottuk az azonos méretű, kiindulási referencia korpuszhoz viszonyítva.

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

n	n_{aug}	$\frac{n_{aug}}{n}$	$\frac{n_{aug}}{N}$
[db]	[db]	[-]	[%]
10	710	71	98,6
25	695	27,8	96,53
50	670	13,4	93,06
100	620	6,2	86,11
250	470	1,88	65,28
500	220	0,44	30,56

1. táblázat

3. Szöveg augmentálás

3.1. Alkalmazott módszerek

A szövegek augmentálásához a `digital-twin-distiller` keretrendszeren belül elérhető augmentáló algoritmusokat (Csányi és Orosz, 2021) alkalmaztuk. Ezeknek két nagyobb csoportja különíthető el, melyek közül az elsőbe az *egyszerű augmentálási módok* tartoztak (*Easy Data Augmentation - EDA*, Wei és Zou (2019)), a másikba pedig a szóbeágyazáson alapuló módszer (Csányi és mtsai (2021)).

Az összehasonlítás során az EDA-ba tartozó technikák közül a következő négy augmentálási megoldást alkalmaztuk:

- Szinonima helyettesítés (Synonym Replacement - SR): adott számú szót egy random választott szinonimájával helyettesít.
- Random beszúrás (Random Insertion - RI): kiválaszt szavakat a mondatból, és azok szinonimáit random pozíciókra helyezi el a mondaton belül.
- Random csere (Random Swap - RS): adott mennyiségű szópár pozíciójának cseréje a mondaton belül.
- Random törlés (Random Deletion - RD): szavak törlése az augmentált szövegből adott valószínűséggel.

A szóbeágyazási módszer esetében 100, illetve 300 dimenziós `fastText`[6] (Bojanowski és mtsai (2017)) modelleket alkalmaztunk. A modellek alapját az augmentálásra kiválasztott mondatok adták, augmentálás során pedig a modellben leginkább hasonló 10 szó közül véletlenszerűen választottunk egyet, amellyel az eredeti tokent az algoritmus helyettesítette. Azt, hogy az augmentálás milyen arányban történjen, az α paraméter változtatásával lehetett beállítani.

3.2. Preprocesszálás

A szövegek normalizálása során közös lépés volt a mondatok stopszó szűrése, a számok és az írásjelek eltávolítása, valamint a kisbetűsítés. A különböző ragozott és képzett alakok normalizálását emellett három különböző eszközzel is elvégeztük:

[6] https://fasttext.cc/docs/en/crawl-vectors.html

– a `spaCy` 2.0 verziójához (Honnibal és Montani, 2017) elérhető, Orosz György által készített nyelvmodellt[7] használva készítettünk a mondatokból egy lemmatizált változatot. A továbbiakban az ilyen módon előkészített szövegekre mint lemmatizált korpuszvariánsra hivatkozunk,
– kétféle stemmer segítségével szótövezést végeztünk; az egyik a `Hunspell` programcsomag beépített szótövezője volt (a továbbiakban: hunspell[8]), a másik eszköz pedig a `Hunspell` keretrendszerhez bővítményként vagy önálló python csomagként is elérhető `hungarian-stemmer`[9] csomag volt, amely egy kevésbé agresszív szótövező.

A preprocesszált szövegeken betanított modellekkel kapott eredményeket a 2. táblázat ismerteti. Az első két sor jelenti az adott kategóriából a modell tanításához felhasznált tanítóadatok számosságát. A táblázat 100 futtatás átlagát tartalmazza, zárójelben a kapott szórások láthatók. A kapott értékeket jól jellemzi, hogy közeledve a két korpusz kiegyensúlyozott arányához az F_1-értékek minden esetben jelentősen javulnak. Azon három esetben, amikor mindkét kategória teljes szöveganyaga felhasználásra került, a legjobb értéket a Hunspellel történt normalizálás mellett értük el.

Kategória	**Undor**	10	25	50	100	250	500	720
	Várakozás	720	720	720	720	720	720	720
spaCy	P	6 (23,75)	22 (41,42)	55,83 (48,44)	84,4 (12,43)	83,18 (3,15)	75,55 (1,51)	70,97 (0)
	R	0,03 (0,13)	0,13 (0,25)	0,48 (0,52)	4,28 (1,47)	35,63 (2,71)	63,12 (1,82)	73,33 (0)
	F_1	0,07 (0,26)	0,25 (0,49)	0,95 (1,02)	8,11 (2,68)	49,83 (2,9)	68,76 (1,35)	72,13 (0)
Hunspell	P	6,5 (24,14)	26,5 (43,28)	67,63 (43,49)	82,45 (12,4)	82,65 (3,11)	76,93 (1,42)	73,51 (0)
	R	0,04 (0,14)	0,17 (0,3)	0,69 (0,63)	5,04 (1,88)	35,05 (2,69)	62,97 (1,98)	75,56 (0)
	F_1	0,08 (0,28)	0,34 (0,6)	1,36 (1,23)	9,44 (3,36)	49,17 (2,94)	69,24 (1,42)	74,52 (0)
hungarian-stemmer	P	8 (27,13)	28,75 (45,05)	76,77 (38,29)	88,13 (8,34)	82,2 (3,23)	74,17 (1,39)	69,15 (0)
	R	0,04 (0,15)	0,21 (0,37)	0,96 (0,78)	6,44 (1,9)	37,99 (2,7)	64,86 (1,66)	72,22 (0)
	F_1	0,09 (0,3)	0,42 (0,73)	1,89 (1,51)	11,95 (3,33)	51,92 (2,82)	69,19 (1,31)	70,65 (0)

2. táblázat. Különböző tanítóadat számosságok mellett mért pontosság (P), fedés (R) és F_1-érték (F_1) százalékosan kifejezve (zárójelben: a 100 futtatás után mért szórás).

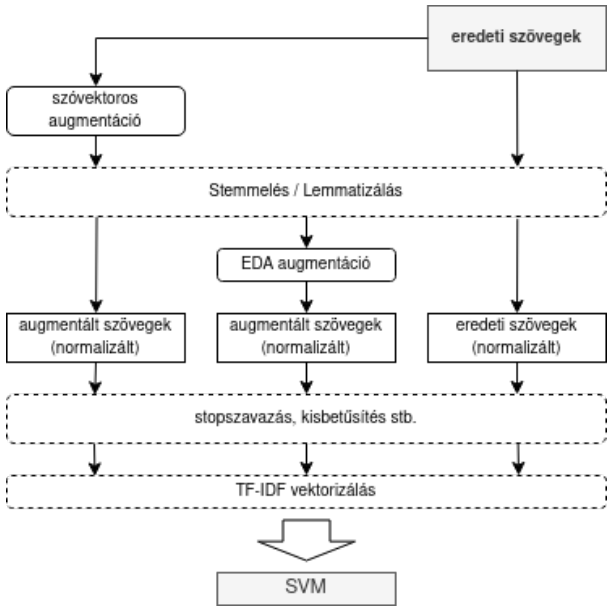
Hasonló a tendencia a köztes számosságok esetén is; a mért F_1 értékek minden tanítóadat összeállítás esetén az azonos számosságú esetekben hasonlóan alakulnak. Az ezen modellekkel kapott eredményeket használtuk a továbbiakban referenciaként a szintetikus mondatokkal felbővített halmazon betanított modellekkel való összehasonlításhoz.

3.3. Az augmentálás folyamata

Az egyes augmentálási módszereket az optimális eredmény elérése érdekében különbözőképpen és az előfeldolgozási lánc eltérő pontjain alkalmaztuk. Az EDA

[7] https://github.com/spacy-hu/spacy-hungarian-models
[8] Hunspell: http://hunspell.github.io/
[9] https://github.com/montana-knowledge-management/hungarian-stemmer

módszereket (SR, RI, RS, RD) a lemmatizált vagy szótövezett szövegen futtattuk, azonban még a stopszó-szűrés és a kisbetűsítés előtt. A szóvektor alapú módszereket ezzel szemben a lemmatizálást / szótövezést megelőzően alkalmaztuk a tanítóadatban szereplő szövegeken. Ennek az volt az oka, hogy a `fastText modell segítségével kapott` leghasonlóbb szavak többször tartalmaztak valamilyen írásjelet vagy nagybetűs szót, és el akartuk kerülni, hogy a preprocesszálás során többször is szükség legyen az írásjelek és számok szűrésére. A folyamatot részletesen a 3. ábra szemlélteti.



3. ábra. Az augmentálási módszerek helye a teljes előfeldolgozási láncban.

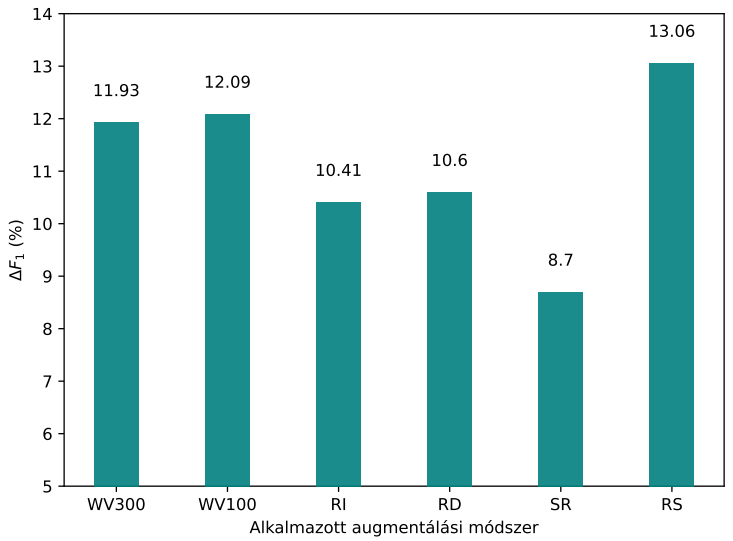
Valamennyi módszer esetében lehetőség van az augmentálást végző algoritmusok további finomhangolására. Az egyik legfontosabb ilyen technika ún. *védett szavak* megadása (ezek esetében a listán szereplő szavakat az augmentálási módszerek nem módosítják). Ez különösen hasznos lehet például szaknyelvek esetében, ahol két köznyelvileg szinonim kifejezés eltérő jelentéssel bír (pl.: „garázdaság” vs. „rongálás” jogi szövegekben). A védett szavak használata további módosításokat nem jelent a kódolás során, alkalmazásuk egy lista megadásával lehetséges, amelynek az összeállításához viszont doménspecifikus szaktudásra van szükség. A vizsgálat során a jelen korpusz egyes kategóriáihoz nem állt rendelkezésre ilyen lista, így ezt a funkciót nem használtuk ki. Az augmentáció vizsgálata során az úgynevezett α-paraméter megválasztásának hatását is vizsgáltuk, amely a szövegben megváltoztatott szavak arányát szabályozza. Vizsgálatunk során ezt 0,1 és 0,5 (tehát 10% és 50%) között változtattuk minden augmentálási mód

esetében, 0,1-es lépésközzel. Az eddig leírtak összes kombinációjaként összesen 540 különböző SVM modellt tanítottunk be, majd az ezekből visszakapott metrikákat értékeltük ki.

4. Eredmények és következtetések

A tanítás során a tesztadatok minden esetben az augmentálásra fel nem használt mondatok közül kerültek ki és a 900 + 900 mondatos korpusz 20 - 20% -át adták.

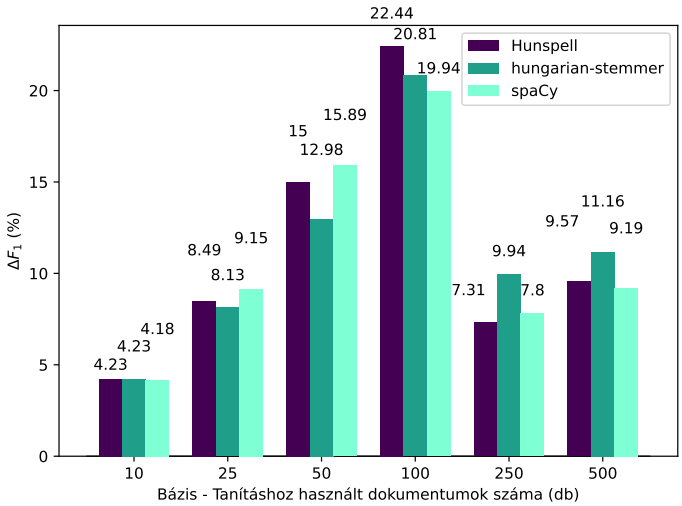
Az összképet tekintve, az augmentálási módszerek szerint összegeztük majd átlagoltuk az eredményeket. Ez minden esetben 90 db különböző modell eredményeinek az átlagolását jelenti. A 4. ábrán az egyes módszerekkel elért átlagos F_1 érték változást tüntettük fel az összes bázis átlagára nézve, az „Undor" kategória felismerése során. Az ábráról leolvasható, hogy az RS algoritmus teljesített a legjobban a vizsgálat során; mintegy 13 %-kal, szignifikánsan növelte az F_1 mérték értéket a referenciahalmazhoz képest. Ezt nagyjából 1%-kal lemaradva a szóvektor alapú módszerek követették (100, illetve 300 dimenziós `fastText` szóbeágyazás használata mellett).



4. ábra. Az augmentálási módszerek összesített hatékonysága a különböző augmentációs módszerek (WV300 és WV100 - 300 illetve 100 dimenziós `fastText` modellel történt augmentálás, RI - Random Insertion, RD - Random deletion, SR - Synonym replacement, RS - Random Swap) esetén.

Más nézőpontból tekintve az eredményeket, ha bázis és szóalak normalizálásra alkalmazott módszerek szerint végezzük az adatok felbontását, akkor

az 5. ábra szerinti F_1 változások adódnak. A felosztás alapján az látszik, hogy a lemmatizált adatok hatékonysága két kisebb bázis esetében haladta meg a többi módszerrel szótövezett változatokét, míg a két legnagyobb bázis esetében a `hungarian-stemmer` használatával volt elérhető a legnagyobb javulás az értékekben. A `Hunspell` szótövezővel pedig a 100-as bázis esetén lehetett átlagosan a legjobb eredményekhez jutni.



5. ábra. Az augmentálási módszerek hatékonysága a tanításhoz használt dokumentumok darabszáma és szóalak normalizáláshoz használt eszköz szerint.

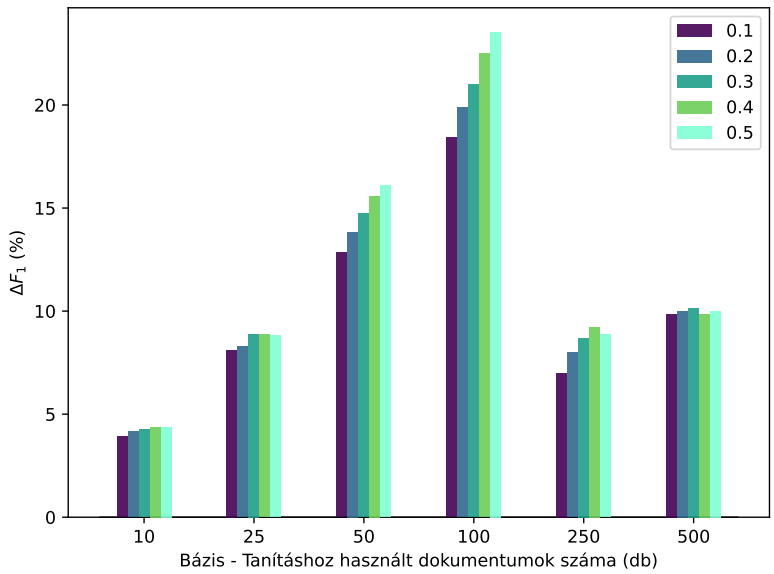
	spaCy	**hunspell**	**hungarian-stemmer**
10	0.04	0.04	0.04
25	0.09	0.08	0.09
50	0.15	0.13	0.16
100	0.24	0.23	0.23
250	0.15	0.20	0.16
500	0.31	0.36	0.30

3. táblázat. Átlagos hibacsökkenés aránya az eredeti és az augmentált adattal feljavított halmazok esetében bázisonként.

A 3. táblázat ezzel szemben az átlagos relatív hibacsökkenést mutatja be, azaz azt a mérőszámot, hogy az eredetihez képest az augmentált adattal feljavított halmazok F_1 értékei mennyivel kerültek közelebb a 100%-os értékhez.

Az adatokból is jól kivehető, hogy az augmentálás hozzáadott értéke egy meredek kezdeti emelkedés után egy alacsonyabb szinten stabilizálódik, ahogyan a korpuszban maradó eredeti adatok számossága és ezáltal feltehetőleg változatossága is megfelelően nagy lesz.

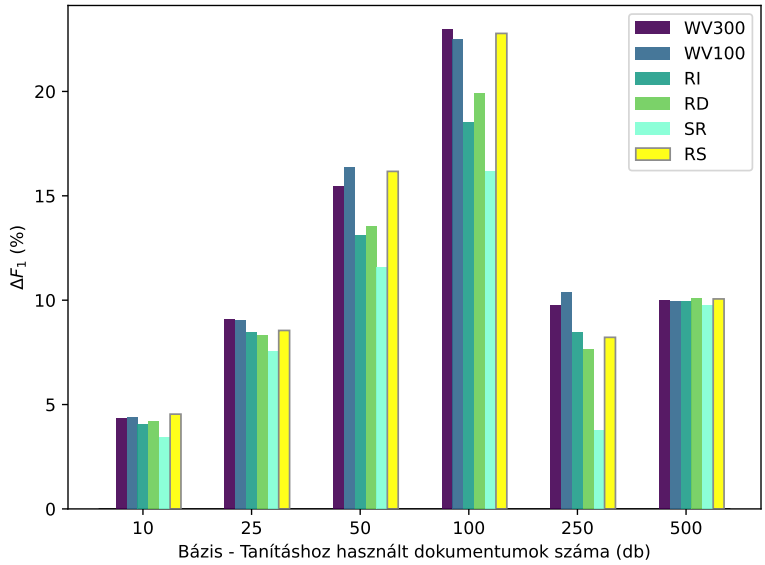
Az α paraméter változtatásának hatását a háromfajta szóalak normalizálási eljárás estén a 6. ábra mutatja be. Az ábrázolt értékek a `Hunspell` normalizálás során kapott konkrét eredményeket mutatják, ugyanakkor a grafikon alakulása tipikusnak volt mondható a másik két normalizálási eljárás esetében is. Az eredményekből látszik, hogy az α paraméter értékének növekedésével párhuzamosan az augmentált adattal nagyobb növekmény érhető el a legtöbb bázis esetében. Ez a hatás 100-as bázis esetében érvényesül a legmarkánsabban, míg az augmentált adatok arányának csökkenésével a tanítóadatban (a bázis növekedésével párhuzamosan) fokozatosan kiegyenlítődni látszik.



6. ábra. Az α paraméter változtatásának hatása az egyes bázisok alapján, Hunspell normalizálás mellett.

A 7.ábrán az egyes augmentálási módok átlagos hatását tüntettük fel bázisok szerint bontva. Az ábrán jól kivehető, hogy közepes méretű bázisok esetében (50–250) a szóvektoros augmentálási módok és a RS magasan a többi módszer felett teljesített, míg a bázisok szélső értékeinek esetében (10 és 500) az eredmé-

nyek sokkal inkább kiegyenlítetten alakultak. Minden esetben kivehető, hogy a legalacsonyabb eredményt a SR megoldás érte el.



7. ábra. Augmentálási módszerek átlagos hatékonysága bázisok szerint.

5. Összefoglalás

Cikkünkben különböző szöveg augmentációs technikák hatását vizsgáltuk a politikai doménre készülő TK-MILAB szentiment korpuszon. Az augmentáláshoz EDA és szóbeágyazás alapú módszereket alkalmaztunk különböző nagyságú tantóadaton betanított SVM modellekkel. Az eredmények azt mutatják, hogy az augmentálási módszerek az 50-100-as bázison betanított modellek esetén növelték a legjobban az F_1 értéket. Az összehasonlítás során az EDA csoportba tartozó random csere (RS) produkálta a legjobb eredményt, amelyet szorosan a szóvektor alapú augmentálási módszerek követtek. Ezeket a módszereket a szótövezett és a lemmatizált szövegen is elvégeztük. Az eredmények összehasonlításából az látszik, hogy a `spaCy`-vel lemmatizált adathalmaz, valamint a `hungarian-stemmer` és a standard `Hunspell`-es szótövezéssel preprocesszált szövegen végzett augmentálás váltakozó eredményt mutatott, egyértelmű trend nem volt kimutatható. A `digital-twin-distiller`-ben készített szemantikai elemző modell szabadon letölthető és kipróbálható a projekt GitHub tárolójából[10].

[10] A projektfájlok és a projekthez tartozó applikáció elérhető a https://github.com/montana-knowledge-management/hungarian-political-sentiment-analysis címen.

Köszönetnyilvánítás

A kutatást az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

Hivatkozások

Bhowmick, P.K., Basu, A., Mitra, P.: Reader perspective emotion analysis in text through ensemble based multi-label classification framework. Comput. Inf. Sci. 2(4), 64–74 (2009)

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146 (2017)

Breeman, G., Then, H., Kleinnijenhuis, J., van Atteveldt, W., Timmermans, A.: Strategies for improving semi-automated topic classification of media and parliamentary documents (2009)

Burscher, B., Vliegenthart, R., De Vreese, C.H.: Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? The ANNALS of the American Academy of Political and Social Science 659(1), 122–131 (2015)

Csányi, G.M., Orosz, T.: Comparison of data augmentation methods for legal document classification. Acta Technica Jaurinensis (2021)

Csányi, G.M., Nagy, D., Vági, R., Vadász, J.P., Orosz, T.: Challenges and open problems of legal document anonymization. Symmetry 13, 1490 (2021)

Dadgar, S.M.H., Araghi, M.S., Farahani, M.M.: A novel text mining approach based on tf-idf and support vector machine for news classification. In: 2016 IEEE International Conference on Engineering and Technology (ICETECH). pp. 112–116. IEEE (2016)

Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. CoRR abs/1705.00440 (2017), http://arxiv.org/abs/1705.00440

Fawzi, A., Samulowitz, H., Turaga, D., Frossard, P.: Adaptive data augmentation for image classification. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 3688–3692 (2016)

Hillard, D., Purpura, S., Wilkerson, J.: Computer-assisted topic classification for mixed-methods social science research. Journal of Information Technology & Politics 4(4), 31–46 (2008)

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

Jia, Y., Chen, Z., Yu, S.: Reader emotion classification of news headlines. In: 2009 International Conference on Natural Language Processing and Knowledge Engineering. pp. 1–6. IEEE (2009)

Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation (1972)

Kellerman, H., Plutchik, R.: Emotion-trait interrelations and the measurement of personality. Psychological Reports 23, 1107–1114 (1968)

Kubat, M., Matwin, S., és mtsai: Addressing the curse of imbalanced training sets: one-sided selection. In: Icml. vol. 97, pp. 179–186. Citeseer (1997)

Kumar, M.A., Gopal, M.: A comparison study on multiple binary-class svm methods for unilabel text categorization. Pattern Recognition Letters 31(11), 1437–1444 (2010)

Lango, M., Stefanowski, J.: Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data. Journal of Intelligent Information Systems 50(1), 97–127 (2018)

Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development 1(4), 309–317 (1957)

Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery 28(1), 92–122 (2014)

Nguyen, H.M., Cooper, E.W., Kamei, K.: Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms 3(1), 4–21 (2011)

Ring, O., Martina Katalin, S., Guba, C., Váradi, B., Üveges, I.: Approaches to sentiment analysis of hungarian political news at sentence level with dictionary-based method and with machine learning. PLoS One (2021), megjelenés alatt

Sebők, M., Kacsuk, Z.: The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach. Political Analysis 29(2), 236–249 (2021)

Taylor, L., Nitschke, G.: Improving deep learning with generic data augmentation. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1542–1547 (2018)

Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., Schlobach, S.: Good news or bad news? conducting sentiment analysis on dutch text to distinguish between positive and negative relations. Journal of Information Technology & Politics 5(1), 73–94 (2008)

Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks (2019)

Young, L., Soroka, S.: Affective news: The automated coding of sentiment in political texts. Political Communication 29(2), 205–231 (2012)

Zhang, C., Zhou, P., Li, C., Liu, L.: A convolutional neural network for leaves recognition using data augmentation. In: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. pp. 2143–2150 (2015)