Neme?
Nemek közötti nyelvhasználati különbségek a Facebook bejegyzésekben

Kelemen Sára Bernadett[1], Yang Zijian Győző[2], Kmetty Zoltán[3,4]

[1] Clementine
1115 Budapest, Bartók Béla út 105-113 I/B.
skelemen@clementine.hu
[2] Nyelvtudományi Kutatóközpont,
1068 Budapest, Benczúr u. 33.
yang.zijian.gyozo@nytud.hu
[3] Eötvös Loránd Tudományegyetem, Társadalomtudományi Kar, Szociológia Tanszék,
1117 Budapest, Pázmány Péter sétány 1/C.
kmetty.zoltan@tatk.elte.hu
[4] Társadalomtudományi Kutatóközpont, CSS-Recens kutatócsoport
1097 Budapest, Tóth Kálmán utca 4.

Kivonat: Online szövegelemzések esetében gyakran elhangzó kritika, hogy a szövegek íróiról nincs háttérinformációnk. Ha nyelvhasználatuk alapján képesek lennénk nemet becsülni, értékes plusz tudást szerezhetnénk. Jelen kutatásunk célja 110 nő és 37 férfi magyar nyelvű Facebook posztjainak felhasználásával annak a kérdésnek a megválaszolása volt, hogy az alkalmazott modellek közül melyik az, amelyik a legsikersebben külön tudja választani a nemeket nyelvhasználatuk alapján. Ennek keretén belül nem csak több, hagyományosnak tekinthető felügyelt gépi tanulási módszert vizsgáltunk meg, hanem egy transzformer modellt is finomhangoltunk szövegosztályozásra. A modellek teljesítményét a tesztalmazon összetett mutatók segítségével értékeltük.

Kulcsszavak: közösségi média, Facebook, nyelvhasználat, nem, felügyelt gépi tanulás, szövegosztályozás, huBERT

1 Bevezetés

A szerzők tetszőleges demográfiai változó mentén való kategorizálása gyakori klasszifikációs probléma napjainkban. A nem, életkor vagy akár nyelvhasználat szerinti osztályozás az üzleti alkalmazásokban segítheti a fogyasztók hatékonyabb megismerését.

Az internet térhódításával, valamint a digitális forradalomnak köszönhetően egyre nagyobb mennyiségű elektronikus szöveg válik hozzáférhetővé. A közösségi oldalak, mint például Facebook vagy Twitter, felhasználói által hagyott „digitális lábnyomok”[1]

[1] A 20. század végétől kezdő, számítógépek és a digitalizálás által kiváltott áttörést értjük digitális forradalom alatt. A kifejezés a számítástechnikai és távközlési eszközök, a számítógép és a telefon elterjedésével járó hatások leírásával fejezhető ki. (Karvalics, 2012)

elemzése egyre népszerűbbé válik, hiszen relatív egyszerűen elérhető, hatalmas adatmennyiségről van szó. A közösségi média oldalakról elérhető strukturálatlan adatok elemzése azért is „kifizetődő", mert társadalmunk nagy része megtalálható ezeken a platformokon.

A felhasználók posztjaiból kiinduló, az ő nyelvhasználati szokásaikat kutató, klaszszifikációt alkalmazó cikkek száma is megnövekedett. Ezekben rendszerint Twitter-ről, vagy különböző blogokról származó szövegeket elemeznek (Fosch-Villaronga és mtsai, 2021; Vashisth és mtsai, 2020; Zhang és Zhang 2010), Facebook-on megjelenő szöveges posztokat nagyon ritkán elemeznek ilyen szempontból, elsősorban az adathozzáférés nehézségei miatt. Kutatásunkban azt vizsgáltuk vajon mennyire alkalmasak a Facebookról származó bejegyzések ennek a témakörnek a tanulmányozására. A fellelhető szakirodalmak elsősorban angol nyelvűek, és jelentős részük angol vagy más, magyartól eltérő, idegen nyelvű posztokon keresztül vizsgálja a felhasználók nyelvhasználati szokásait. A netnyelvről, valamint arról, hogy a fiatalok csevegéssel töltött ideje, hogyan befolyásolja élőbeszédüket számos értekezlet elérhető magyar nyelven is. Előbbiek azonban inkább csak történeti összefoglalók, sokszor mindenféle saját kutatást nélkülözve. Utóbbiak pedig többnyire valamilyen kvantitatív interjús megkérdezésre és különböző iskolai feladatként írt fogalmazások analizálásra korlátozódnak. Az utóbbi években azonban egyre inkább találkozhatunk beszéltnyelvi leiratozásokra alapozott analízisekkel (Vincze és mtsai, 2021).

2 Kapcsolódó munkák

Azzal, hogy miben különbözik a nemek közötti nyelvhasználat, a kutatók már jóval azelőtt elkezdtek foglalkozni, minthogy a közösségi médiák platformjai megjelentek volna. Később, miután az interneten egyre inkább elszaporodtak a különböző blogok, megjelentek a közösségi média platformjai, folyamatosan nőtt az igény és az érdeklődés az előbb említett felületekről származó hatalmas adatmennyiség felhasználhatóságának vizsgálatára a témában. A leggyakoribbak a különféle blogbejegyzésekből kiinduló kutatások (Argamon és mtsai, 2003; Schler és mtsai, 2006; Zhang és Zhang 2010), de elterjedtek a Twitter-ről származó tweet elemzések (Markov és mtsai, 2017; Aragón és López, 2018; Vashisth és mtsai, 2020) is. Facebook posztokra alapozott vizsgálatokkal, sokkal ritkábban[2], de szintén találkozhatunk (Rangel és Rosso, 2013; Sap és mtsai, 2014).

Newman és munkatársai 2008-as tanulmányukban összegyűjtötték, hogy milyen, gyakran ellentmondásos eredmények születtek a korábbi kutatások során. Kihangsúlyozták, hogy a szavakra alapozott szöveganalitikák természetüknél fogva nem képesek megragadni azt a kontextust, amelyben a szavakat használják. A nemek közötti különbségek értelmezése árnyalt ügy, a társadalmi célok, szituációs igények és szocializáció komplex kombinációja. A tanulmányban saját kutatásukat is ismertetik. Elemzéseikben kicsi, de szisztematikus különbségeket találtak női és férfi nyelvhasználat között. Ezt négy aspektusból vizsgálták: mikor, hol, miért és hogyan választható szét a

[2] A Facebook esetében az adathozzáférés sokkal nehezebb, mint a Twitter vagy egyes blogok esetében, különösen amióta a Facebook elzárta az API hozzáférést a kutatók elől. Elsősorban ezért a Twitter az elsődleges platformja a nyelvhasználati kutatásoknak.

nyelvhasznált mind abban a tekintetben, hogy mit mondanak, mind abban, hogy hogyan fejezik ki magukat a nők és férfiak. Esetükben a hangsúly funkcionális szavakon volt. Eredményeik segítségével arra mutattak rá, hogy a szó-számolási stratégiák életképesek, nagyhatékonyságú alternatívái az emberi kódokon alapuló nyelvi elemzésnek (Newman és mtsai, 2008).

Huffaker és Calvert 2005-ös írásukban tinédzserek webblogjain keresztül vizsgálták, hogy hogyan fejezik ki a serdülők magukat nyelvileg, milyen érzelmi kódokat használnak. A szerzőpáros nem talált nemi különbségeket a hangulatjelek használatának gyakoriságában. Sőt meglepő módon azt tapasztalták, hogy azok közt, akik használnak emotikonokat több a fiú. Nem tapasztaltak több agressziót a fiúk esetében, vagy nagyobb passzivitást a lányoknál. A korábbi eredményekkel ellentétben nem találtak a nemek között különbséget törődés, illetve együttműködés terén. Arra az eredményre jutottak, hogy a fiúk aktívabb, határozottabb és rugalmatlanabb nyelvet használtak. A lányok ellenben nem használtak passzívabb, kooperatívabb, vagy alkalmazkodóbb nyelvet. E mögött meghúzódó lehetséges ok szerintük, hogy a nyelv és az interneten történő szociális interakciók változnak, talán pont amiatt, mivel a résztvevők is változnak.

Huffaker és szerzőtársa egy tartalomelemző szoftvercsomag segítségével értékelte ki a dokumentumokat a szavak száma, tartalomtípus és nyelvi hangnem szempontjából. A használt nyelv értékelését is a tartalomelemző szoftverre bízták, amely figyelembe vette a nyelvi kontextust, valamint a szavak gyakoriságát. A blogok összehasonlítására Khi-négyzet próbákat és független T-próbákat futtattak (Huffaker és Calvert, 2015).

Rangel és Rosso (2013) spanyol anyanyelvű emberek nyelvhasználatát vizsgálta, hogy milyen grammatikai kategóriákat használnak Facebook posztjaikban. Azokra a kognitív tulajdonságokra összpontosítottak, amik nem és kor szerint különbözővé tesznek minket. Nemekre lebontva eredményeik alapján a férfiak több előszót használnak, talán amiatt, hogy megpróbálják a környezetükben lévő dolgokat hierarchikusan besorolni. A nők ezzel szemben több determinánst, névmást és közbeiktatást alkalmaznak, valószínűleg amiatt, mert több érdeklődéssel vannak a társas kapcsolatok iránt.

Adataik nem szerint igen, de kor szerint nem voltak kiegyensúlyozottak. Szupport Vektor Gép osztályozási eljárást alkalmaztak. Végeredményben pedig arra a megállapításra jutottak, hogy az általuk alkalmazott stilisztikai dimenziók jobban teljesítenek a kor, mint a nem azonosítására, esetleg annak köszönhetően, hogy az írásstílust inkább a szerző korától és nem a nemétől függ (Rangel és Rosso, 2013).

Sap és munkatársai (2014) abból indultak ki, hogy a férfiak és nők jelentősen különböznek érdeklődési és munkabéli preferenciáikban. Az életkor előrehaladtával a személyiség fokozatosan alakul. Emellett a közösségi média nyelve a kor és nem függvényében változik. Facebook posztokat elemeztek, súlyozott szótárt alkalmaztak, amelyet a lineáris többváltozós regressziós és osztályozási modellek együtthatóinak felhasználásával hoztak létre. A nyelv kollinearitás miatt többváltozós lexikon fejlesztési megközelítést alkalmaztak, mely figyelembe veszi a kovarianciát. A nem előrejelzéséhez Szupport Vektor Gép osztályozó eljárást használtak. A tanító adathalmaz felhasználásával több modellt is teszteltek a legjobb modell megtalálásának érdekében. Mind a kor, mind a nem esetében az általuk elért pontosság lényegesen magasabb, mint az alapérték. Azonban azt tapasztalták, hogy minél kevesebb bejegyzés áll rendelkezésre egy-egy felhasználótól, annál kevésbé pontos a nem és az életkor előrejelzése (Sap és mtsai, 2014).

3 Alkalmazott módszerek bemutatása

Többféle osztályozó teljesítményét vizsgáltuk kutatásunk során: Szupport Vektor Gép (Support Vector Machine, továbbiakban SZVG), Bináris Logisztikus Regresszió (továbbiakban LR), Véletlen Erdő (Random Forest, továbbiakban VE), Naiv Bayes (továbbiakban NB) és transzformer modellek. A különféle gépi tanulási modellek tesztelése segít, hogy kiderítsük melyik illeszkedik jobban az adatokra, melyikkel tudjuk a legmegfelelőbben megfogni a klasszifikált változó értékeit reprezentáló pontok és osztálycímkéjük közötti kapcsolatot. Emellett annak érdekében, hogy minél pontosabb klasszifikációt kapjunk érdemes a modellek és hiperparamétereik kombinációjával kísérletezni.

A hagyományosnak tekinthető modellek mellett kíváncsiak voltunk arra is, hogy egy újabb, manapság nagy népszerűségnek örvendő mélyneurális hálózaton alapuló transzformer modell milyen eredményeket produkál ugyanazon az adatbázison. Ebből kifolyólag erre a feladatra egy magyar nyelvű transzformer modellt választottuk, a huBERT modellt.

SZVG. A Szupport Vektor Gép szövegbányászati feladatok során (is) nagy népszerűségnek örvendő modell stabil teljesítmény nyújtása miatt. A nemek osztályozásakor képes valamilyen fajta választ adni a címkéket meghatározó kategória sajátosságaira. Azokban az esetekben hatékony, mikor a dimenziók száma nagyobb, mint a minták száma. Nagy adathalmazok esetén a tanításhoz szükséges idő nagysága miatt nem teljesít túl jól, valamint a zajjal rendelkező adatkészletek is problémát jelentenek számára a target osztályok átfedése miatt (Rakovics, 2016).

LR. A Logisztikus Regresszió olyan klasszifikációs eljárás, mely során előre definiált, egymást kölcsönösen kizáró csoportok egyikébe soroljuk a megfigyeléseket a magyarázó változó(k)ból kinyert információ alapján. Bináris Logisztikus Regresszió esetében a magyarázó változóval/változókkal annak a bekövetkezési esélyét szeretnénk prediktálni, hogy a kimenet 1 lesz (Kovács, 2014).

VE. A Véletlen Erdő előnyei közé sorolható, hogy pontos klasszifikációra képes, nagy adatok esetén is gyorsan lefut, valamint, hogy becsléseket ad arra, melyik változók bírnak nagy jelentőséggel. Nagy mértékben a döntési fák elvén alapul, a zsákolás (bagging) egy speciális alesete, melynél az egyedi döntési fák kvázi korrelálatlanok. Több gyenge osztályozó átlagos teljesítménye alapján klasszifikál, több kisebb méretű döntési fát is épít. Addig folytatja a fák építését, míg az előre rögzített mélységet el nem éri. Az egyes erdők hatékonysága függ a generált fák számától és minőségétől, valamint a fák közötti korrelációtól (ha nő, az eredmény romlik) (James és mtsai, 2013).

NB. A Naiv Bayes féle osztályozók előnye, hogy robusztusok izolált zajos pontokra, illetve az irreleváns attribútumokra (feature). Az osztályozás lényege a Bayes tételen alapszik. Az alapvető feltevés, hogy minden attribútum függetlenül és egyenlően (minden tulajdonság azonos súlyt vagy fontosságot kap) járul hozzá az eredményhez. A szükséges paraméterek becsléséhez kisméretű tanító halmaz is elegendő, valamint a kifinomultabb osztályozókhoz képest sokkal gyorsabban képesek eredményt adni, különösen hasznos nagyon nagy adathalmazok esetén (Tan és mtsai, 2006).

GNB. Gauss Naiv Bayes esetén feltételezzük, hogy az egyes attribútumokhoz tartozó folyamatos értékek egy Gauss-eloszlás szerint oszlanak el, abból kifolyólag, hogy a prediktorok folytonos értéket vesznek fel, és nem diszkrétek.

BNB. Bernoulli Naiv Bayes során az osztályváltozó megjóslásához használt paraméterek bináris változók. Kifejezetten bünteti az olyan attribútum elő nem fordulását, amely egy osztály indikátora.

KNB. A Kiegészítő Naiv Bayes (Complement Naiv Bayes) különösen alkalmas kiegyensúlyozatlan adathalmazok esetében. Az egyes osztályok kiegészítéséből származó statisztikákat használja a modell súlyok kiszámításához. Egy dokumentumot ahhoz az osztályhoz rendeli, amelynek a legszegényebb a bővítmény egyezése (Zhang, 2004).

huBERT. Egy magyar „BERT base" (Devlin és mtsai, 2019) modell, mely a Webcorpus 2.0 korpuszon (Nemeskey, 2020) lett betanítva, ami közel 9 milliárd tokennel rendelkezik. A BERT egy csak enkóderrel rendelkező transzformer architektúrájú neurális nyelvi modell. Jelenleg a huBERT a legjobb teljesítményű BERT modell magyar nyelvre. Több feladatban „state of the art" eredményt ér el, mint például a maszkolt nyelvi modellezésben, névelem-felismerésben vagy főnévi csoport felismerésben.

4 Kutatás

Az elvégzett elemzés célja volt választ kapni azokra a kérdésekre, hogy el lehet-e különíteni a nemeket nyelvhasználatuk alapján, és ha igen, akkor hogyan, milyen mértékben lehet megkülönböztetni őket? Többféle klasszifikációs modell futtatásának segítségével szerettük volna kideríteni, hogy melyik algoritmussal lehet hatékonyabban, helyesen besorolni a saját nemüknek megfelelő osztályba a férfiakat és a nőket.

4.1 Felhasznált adatok

A kutatáshoz nagy időintervallumra kiterjedő Facebook adatokat használtunk fel. A Magyarországon zajló 2019-es kutatás[3] során a résztvevőket egy önkitöltős kérdőív kitöltése után megkérték, hogy töltsék le Facebook adataikat. A minta nem valószínűségi, kvótás (nem és kor), kényelmi minta volt. Ez korlátozza a kutatás megállapításainak általánosíthatóságát, ugyanis a minta fiatalabb volt, mint a magyarországi Facebook-felhasználók átlagos életkora (35-37 év), valamint nagyobb volt a női résztvevők aránya (75%).

Egy fontos kritériuma volt a mintába kerülésnek, hogy adott résztvevőnek rendszeres Facebook felhasználónak kellett lennie. A rendszerességet ebben az esetben úgy definiálták, hogy legalább heti gyakorisággal használja a platformot. Az adatok a résztvevők Facebook használatának teljes időtartamát lefedik, azaz a regisztrációjuktól kezdve a letöltés pillanatáig a felületen végrehajtott összes tevékenységüket, beleértve azokat is, amiket később eltávolítottak. Néhány esetben ez több mint tíz évnyi adatot jelent. A letöltött adatprofilok nem tartalmaznak privát üzeneteket, sem keresési előzményeket vagy audiovizuális tartalmakat (fotók és videók). Valamint kizárták az olyan

3 A kutatás az NKFI-től nyert támogatást a Fiatal Kutató Témapályázaton. A kutatás azonosítója: FK: 128981

tevékenységeket - mint például a Marketplace - melyeket a felhasználók ritkán, vagy szinte soha nem használtak (Breuer és mtsai, 2021).

Az elemzett posztok 20 típusba sorolhatóak. Gyakoriságukat tekintve az első három legnagyobb mértékben előforduló cselekedet típusok a posztok/bejegyzések valakinek az idővonalára, az állapotfrissítések és a csoportba írt bejegyzések voltak. De ezek mellett még sűrűn előforduló posztok voltak a fotók, illetve videók feltöltése, (csak a feltöltött tartalomhoz írt szöveg, egyes esetekben a videó/kép url-je) különböző linkek megosztása, saját idővonalon bejegyzések közzététele. Ritkábban, de az adatbázisban találhatunk többek között emlékmegosztásokat, vagy akár „geotag"-eket (hol járt a felhasználó) is.

A mintát 110 nő és 37 férfi teszi ki. Az átlagos életkor 30 év, a legfiatalabb egyén 18 éves, a legidősebb pedig 71. Az adatbázisban 149.471 poszt szerepel, egy sor egy bejegyzésnek felel meg. Minden poszt szövege mellett szerepel az egyoldali kulccsal anonimizált szerző. Mindegyik résztvevőt egy egyéni azonosítóval láttak el. Ebből a poszt íróját nem lehet visszakövetni, de segítségével tudjuk, hogy melyik posztok származnak ugyanattól a felhasználótól. Emellett fel van tüntetve a felhasználó neme, és születési éve. Valamint szerepel a bejegyzések típuskódja is, hogy hova lett bejegyezve a poszt (saját idővonalra, egy ismerőse idővonalára, csoportba, eseményhez). Ezek mellett a posztok időbélyeggel (timestamp) is el lettek látva, ami alapján másodpercre pontosan meg lehet mondani, hogy mikor született a bejegyzés. A posztok átlagosan 118 szóból állnak. A leghosszabb poszt 13.610 mondatalkotó elemből áll. A nők jelentősebb része 50 és 90 közötti szóból álló bejegyzéseket tett közzé, velük szemben a férfiak nagyobb része 45 és 125 közötti szót használt. A leghosszabb bejegyzések valamilyen csoportba történő posztoláskor születtek.

	Eredeti		Tisztított	
	Tanító	Teszt	Tanító	Teszt
Poszt	38.115	19.334	30.376	16.144
Token	610.574	332.651	523.058	311.243
Type	146.348	91.111	83.305	51.026
Poszt átlagos hossza	16,01 medián: 6	17,21 medián: 8	17,22 medián: 7	19,27 medián: 10
Osztálycímkék	Nő: 26.6204 Férfi: 11.511	Nő: 14.239 Férfi: 5.095	Nő: 21.000 Férfi: 9,376	Nő: 11.365 Férfi: 4.788

1. táblázat. A korpuszra jellemző nyelvtechnológiai értékek.

A posztokban a négy leggyakoribb szófaj a főnevek, melléknevek, igék és határozószók voltak. Nemek szerint bontva nincs jelentős különbség, közel azonos arányban használják mind a négy (és az összes többi) szófajt, bár a nők minden esetben egy-két százalékponttal ugyan, de magasabb használati arányt mutattak, mint a férfiak.

Az elemzett adatbázisban a személyes névmások használata igen gyakori volt. Az egyes szám második és harmadik személyt kifejező „te" és „ő" szavak a 15 leggyakoribb kifejezés közt szerepeltek mind a két nem esetében. Az E/3-t kifejező személyes névmás használata férfiak esetében magasabb volt, és az „ők" személyes névmást is –

bár nagyon elhanyagolható mértékben – de ők alkalmazták többször. Az egyes szám első személyt kifejező „én” szó gyakrabban szerepelt a nők által írt bejegyzésekben, ahogy a „te” és „mi” kifejezések is.

4.2 Adatelőkészítési eljárások

Első lépésként megtisztítottuk a korpuszt minden olyan tartalmi elemtől, melyeknek a szöveg mondanivalójára nézve nincs hozzáadott értéke, nem erősítik a műveletek eredményességét.

Az egyes típus- poszt/bejegyzés valakinek az idővonalára - elhanyagolható részt tekintve csupán születésnapi és névnapi köszöntések különböző formáit, valamint egyéb ünnepekkel kapcsolatos jókívánságokat tartalmazott. Mivel úgy véltük a számtalan féle, gyakran ismétlődő gratulációk magához a klasszifikáció folyamatához nem tesznek hozzá semmilyen plusz értéket, ezért következő lépésként ezt a típust eltávolítottuk az adatbázisból 61.142 sorosra csökkentve ezáltal a korpuszt.

Ezt követően töröltünk minden linket, és e-mail címet a posztokból. Az adatbázis ekkor 57.479 sort számlált. Abból kifolyólag, hogy születésnapi köszöntések (például saját üzenőfalán tette közzé adott egyén a születésnapi felköszöntést, és az érintett embert csak megjelölte a posztban, vagy saját üzenőfalán megköszönte azoknak, aki gondoltak rá születésnapján), valamint egyéb jókívánságok, mint “boldog karácsonyt”, “kellemes ünnepeket”, “sikeres új évet”, “boldog húsvétot” nem csak azok között a posztok között voltak megtalálhatóak, amiket valaki másnak címeztek, hanem a mintában szereplők saját üzenőfalán is rengeteg közzé lett téve, ezért ezeket a többi bennmaradt típusból is kiszűrtük, ami 3.364 sornyi csökkenést eredményezett az adathalmazon.

A posztokban szereplő személyek anonimizálásakor a neveket egy „@”-jelet követő szám és betűkombinációval helyettesítették. Mivel plusz információval ezek az egységek sem szolgáltak, ezért szintén eltávolítottuk őket.

Mind az emotikonok, mind az emojik használata információt tartalmaz(hat) a szerzőről, ezért nem töröltük őket. Azonban a modellek futtatásakor a speciális karakterek problémát okozhatnak, ezért mind az emotikonokat, mind az emojikat átkódoltuk, egy-egy egyedi névvel láttuk el őket.

Orosz György spaCy fejlesztését[4] felhasználva, ami kifejezetten a magyar nyelvű szövegeken való alkalmazásra lett készítve, 46.514 sornyi szöveget tokenizáltunk, majd lemmatizáltunk. Ezt követően „Part of Speech” (Pos) tag-gel is elláttuk őket, azaz szófaj szerint azonosítottuk azokat.

A tiltólistás szavak eltávolítására a Python Natural Language Toolkit (NLTK)[5] nevű csomagját választottuk, azon okból kifolyólag, hogy már magyar nyelvű stopszólistával is rendelkezik. A módszer nem kezeli a szavak sorrendjét, csupán a gyakoriságát. Nem ismeri fel a hétköznapi nyelvhasználatot sem, így a helyesírási hibákat, elgépeléseket nem tudja kezelni, ezek problémát jelenthetnek. A sokat használt, illetve nyelvtani funkciót betöltő szavak törlése javíthatja az elemzésünk végeredményét, hiszen a leggyakrabban használt szavak vizsgálatakor így nem az „a”, „az” vagy épp az „egy” szavakat fogjuk visszakapni. Feltételezhetőleg információvesztéssel járt volna minden

[4] https://github.com/oroszgy/spacy-hungarian-models

[5] https://www.nltk.org/

stopszó eltávolítása, amit az NLTK stopszó listája tartalmaz, hiszen lehet, hogy a személyes névmások használata is olyan sajátosságot jelent, amely alapján el lehet különíteni a férfiakat és nőket nyelvhasználatuk alapján. Ezért a fentebb említett listát módosítva alkalmaztuk, és csak a kötőszókat, névelőket, határozószókat, valamint néhány olyan általunk megadott szót távolítottunk el, melyekről a leggyakrabban használt unigramok vizsgálatát követően kiderült, hogy a posztokban gyakran alkalmazták, de érdemi információt nem hordoznak, nem adnak többletinformációt a kutatáshoz (pl.: debrecen, delon).

4.3 Vektorizálás

Vektorizálás során a tanító halmazból kinyert n-grammok határozzák meg a magyarázó változókat, a későbbiekben teszteléshez és validáláshoz is ezeket kell kinyerni a korpuszból. Emiatt még vektorizálás előtt felosztottuk az adathalmazt tanító és teszthalmazra.

A 46.514 bejegyzésből álló adatbázist véletlenszerűen ⅔ és ⅓ arányban osztottuk fel. A modellek futtatásakor azonban ez túl soknak bizonyult egyes algoritmusok esetében (kifejezetten SZVG, de a Logisztikus Regresszió is lassan futott le), ezért, hogy ne ennyi emberről (egy sor egy egyénnek feleltethető meg, ha nem vesszük figyelembe, hogy egy id-hoz hány darab poszt tartozik) kelljen eldöntenie a modelleknek, hogy női vagy férfi szerzőről van szó, a bejegyzéseket összevontuk. Későbbiekben az adatbázis egy sorában az összes, adott azonosítóval rendelkező személy által írt poszt egyben, egy cellában szerepelt. A bejegyzések összevonását követően az adatbázis már csak 147 sort számlált. Így végül a tanító halmazba 98, míg a teszthalmazba 49 egyén került.

A futtatott modellek a TF-IDF vektorizálást követően produkálták a legjobb értékeket. A modellek a legjobb előre jelző erővel abban az esetben bírtak, mikor unigramokat és bigramokat vizsgáltunk, az 5 és afeletti előfordulási gyakoriságú és a dokumentumok 70%-nál kevesebben előforduló szavakra korlátoztuk a szókincs összeállításakor felhasznált kifejezéseket, valamint a kifejezésgyakoriság szerint rendezett legfelső 500 attribútumot vettük figyelembe.

4.4 A hagyományos modellek hiperparamétereinek hangolása

Miután kiválasztottuk, hogy mely paramétereket fogjuk hangolni, definiáltunk a lehetséges értékekből egy rácsot és „Randomized Search”-öt hajtottunk végre ötszörös keresztvalidálással, mivel így az egyes hiperparaméterek szélesebb értéktartományát tudtuk lefedni nagy végrehajtási idő nélkül. Miután megkaptuk a legjobb hiperparaméterekkel rendelkező modellt - úgy, hogy leszűkítettük az egyes tartományokat – „Grid Search”-öt végeztünk szintén ötszörös keresztvalidálással, kifejezetten meghatározva a kipróbálandó beállítások minden kombinációját, hogy a hiperparaméter térben megtaláljuk a legjobban teljesítő kombinációt.

Ezt követően, mikor megtaláltuk a legjobb kombinációját a hiperparamétereknek, elvégeztük a hiperparaméter hangolást a tanító adatokkal, és ráillesztettük a modellt a tanítóadatokra, értékeltük a teljesítményét a teszthalmazon. Ha az osztályeloszlás kiegyensúlyozatlan, az abszolút pontosság (accuracy) rossz választásnak számít, mivel

minden osztályt egyforma fontosságúként kezel, magas pontszámot ad azoknak a modelleknek, amelyek csak a leggyakoribb osztályt jósolják. Emiatt összetett mutatókat választottunk az abszlút pontosság mellett a modellek teljesítményének értékelésére, mint fedés (recall), relatív pontosság (precision), f-mérték (f1-score).

4.5 huBERT finomhangolása

Kutatásunk során finomhangoltuk a huBERT nyelvi modellt szövegosztályozási feladatra. A tanításhoz a Huggingface által közzétett „transformers text classification" könyvtárat[6] használtuk. A finomhangolást az alábbi módosított hiperparaméterekkel végeztük: maximum bemeneti hossz: 128; batch méret: 8 / GPU (4 darab GeForce GTX 1080 Ti); tanulási ráta: 2e-5; epoch: 10.

Kétféle modellt tanítottunk. Eredetiként arra a modellre utaltunk, mely esetén az eredeti adatbázissal dolgoztunk, amelyből eltávolítottuk az egyes típusú posztokat. A bent hagyott posztokból nem lettek kiszűrve a felköszöntések. A bejegyzésekből nem töröltük a linkeket, e-mail címeket, maszkolásokat, sem a speciális karaktereket. Ez a szöveg nem lett tokenizálva. Tisztítottként neveztük el azt a modell, amely a teljesen megtisztított 46.514 sorból álló adatbázisra lett futtatva.

Mindkét modellt a posztok szintjén tanítottuk és 60 lépésenként készítettünk mentési pontokat. Minden egyes mentéskor végeztünk egy posztszintű kiértékelést, majd a végén kiválasztottuk a legjobb eredményt (huBERT tisztított: 76,01%; huBERT eredeti: 76,45%) elérő mentési pontot. Ezután a kiválasztott mentési pontok segítségével elvégeztük a posztszintű prediktálást, majd összevontuk őket emberszintre. Az emberszintű összevonáskor megszámoltuk az adott emberhez tartozó prediktált osztálycímkéket és amelyikből több volt, az lett az adott emberhez hozzárendelt végső osztálycímke. Azonban azt vettük észre, hogy ezen modellek alkalmazásakor, az emberszintű összevonás után, eredményként azt kaptuk, hogy mindegyik ember nő lett. Ezért az összevonás szempontjából, ezek a modellek nem használhatóak számunkra. Hogy megtaláljuk az összevonás tekintetében a legjobban teljesítő modellt, minden egyes elmentett modellre elvégeztük az emberszintű kiértékelést is. Végül a tisztított modellünk esetében az 1900-as mentési pont, míg az eredeti modellünk esetében a 2160-as mentési pont került kiválasztásra. Ezen modellek posztszintű pontossága: huBERT tisztított: 70,76%; huBERT eredeti: 71,36%.

5 Eredmények

A Véletlen Erdő módszer nem hozott értékelhető eredményt, ezért ennek további közlésétől eltekintünk. Az algoritmus a hiperparaméterek semelyik kombinációjában sem volt képes elkülöníteni a férfiakat és nőket. Csak nőket prediktált a tanító halmaz méretének növelését követően, ahogy keresztvalidáció végrehajtása után is. A modell előtt futtatott SZVG és VE kapcsán azt tapasztaltuk, hogy a két nem elkülönítése nem lehetetlen feladat. A VE „kudarca" mögött meghúzódó okot valószínűsíthetően a kis férfi mintaelemszámban kell keresni.

[6] https://github.com/huggingface/transformers/tree/master/examples/pytorch/text-classification

	Abszolút pontosság	Relatív Pontosság		Fedés		F-mérték	
		Férfi	Nő	Férfi	Nő	Férfi	Nő
SZVG	90%	70%	95%	78%	93%	74%	94%
LR	80%	47%	97%	89%	78%	62%	86%
KNB	79%	47%	94%	78%	80%	58%	86%
BNB	82%	50%	97%	89%	80%	64%	88%
GNB	87%	64%	95%	78%	90%	70%	92%
huBERT (tisztított)	94%	88%	95%	78%	98%	82%	96%
huBERT (eredeti)	92%	86%	93%	67%	98%	75%	95%

2. táblázat. Az egyes modellek által elért eredmények.

	Abszolút pontosság		Relatív Pontosság		Fedés		F-mérték	
	Tanító	Teszt	Tanító	Teszt	Tanító	Teszt	Tanító	Teszt
SZVG	100%	90%	100%	70%	100%	78%	100%	74%
LR	93%	80%	84%	47%	93%	89%	89%	62%
KNB	89%	79%	74%	47%	93%	78%	83%	58%
BNB	87%	82%	70%	50%	93%	89%	80%	64%
GNB	94%	87%	84%	64%	97%	78%	90%	70%

3. táblázat. A hagyományos modellek által elért mutató értékek a tanító és a teszthalmazon.

A futtatott modellek által elért abszolút pontosság, relatív pontosság, fedés és f-mérték értékeket a 2. táblázatban foglaltuk össze.

A hagyományosnak tekinthető, jellemzőkkel dolgozó modellek tekintetében, összetett mutatók szerint sorba rendezve a futtatott osztályozókat az SZVG modell áll az első helyen. Figyelembe véve azonban az SZVG által a tanító halmazon elért nagyon magas eredményeket, és az ezzel szembeni teszthalmazon elért alacsonyabb összetett mutató értékeket (3. táblázat első sora) a Bernoulli Naiv Bayes modellt választottuk, mint az az osztályozási módszer, mely az adatokat a legjobban/legpontosabban el tudta különíteni. Az SZVG esetében a tanító halmazon elért 100%-os érték tökéletes modellt jelent, azonban ez az optimális érték túlillesztésre utal(hat). Egy tanító adatokra túlontúl jól illeszkedő modell pedig rosszabb általánosítási hibával rendelkezhet, mint egy nagyobb tanítási hibával rendelkező. A Bernoulli Naiv Bayes modell abban a tekintetben is jó helyen végzett, hogy összesen hány egyént sorolt rossz osztályba. Noha az SZVG összességében nézve csak 5 embert „rontott el”, de a fentebb is említett valószínűsíthető túlillesztés miatt, új, eddig nem látotta adatokon nem biztos, hogy ilyen jól teljesítene.

A BNB modell mikor egy személyt férfiként jósol meg az esetek 50 százalékában helyesen jár el. A nőként való jóslás esetében 3 százalékban hibázik. A férfi osztály elemeinek 89 százalékát, míg a nők esetében 80 százalékát jósolja a megfelelő egységbe. Az algoritmus a teszthalmazba került 9 férfi esetében 8-at jósol helyesen férfinak, és 1-et tévesen nőnek. Míg a 40 nő esetében 32-t prediktált nőnek, 8-at pedig férfinak (2. táblázat BNB sora).

Nem csak a BNB modell, hanem az összes többi futtatott módszer esetében általánosságában elmondható, hogy a férfiak pontosabb prediktálása a nők helyes jóslási eredményének csökkenésével - és fordítva - járt együtt.

A kontextuális beágyazáson alapuló huBERT esetén a 2. táblázat alján, dupla elválasztóvonal alatt láthatóak a modelljeink által elért eredmények. A huBERT összességében hasonló (kissé jobb) eredményeket produkált a tisztított adatainkon ember szinten összevonva, mint a felügyelt tanulási algoritmusok. Bár esetében - ahogy SZVG kapcsán is - azt tapasztaltuk, hogy nagyon könnyű a modellt túltanítani.

A huBERT_2 modell teljesen tisztított, emberszinten összevont adatok esetén mikor egy személyt férfiként jósol meg az esetek 12 százalékában jár el helytelenül. A nőként való jóslás esetében 5 százalékban hibázik. A férfi osztály elemeinek 78 százalékát, míg a nők esetében 98 százalékát jósolja a megfelelő egységbe.

6 Összegzés

A futtatott modellek által elért eredmények alapján megállapíthatjuk, hogy a nemek közötti nyelvhasználat különbség a Facebookon is megjelenik, a férfiak és nők elkülöníthetőek egymástól nyelvhasználatuk alapján. A korábbi strukturálatlan szövegeken végzett azonos témájú kutatásokkal való teljeskörű összehasonlítás nem igazán lehetséges. Legfőképp amiatt, hogy korábban angol nyelvű (Facebook) szövegeket felhasználva végeztek felméréseket, tudomásunk szerint magyar Facebook posztokon alapuló nyelvhasználati kutatást még nem hajtottak végre.

A kapott eredmények önmagukban is hasznosíthatóak, ám a nem valószínűségi kényelmi minta miatt általánosíthatóságuk korlátozott. Önálló felhasználásuk mellett ideális kiindulópontot jelenthetnek egy olyan gyakorlatban hasznosítható kutatáshoz, melyben esetlegesen a szerzőket más demográfiai változók mentén is kategorizálják (pl.: iskolai végzettség, életkor, társadalmi státusz) és amely aztán a személyre szabott marketing, vagy akár a digitális bűnüldözés és kiberbiztonság területén alkalmazható.

Jövőbeni további munkálatokként érdemes megnézni a legjobbnak választott modell univerzalitását, hogy lehetséges-e a kapott eredményeket a konkrét mintán kívül adott műfajon belül más mintára is általánosítani. Esetlegesen – ha a későbbiekben ilyen adatok is elérhetővé válnak – megvizsgálni, hogy a társadalmi státusz vagy foglalkozás bevonásával lehetséges lenne-e választékosabb modell kiépítésére, mely segítené a nem és a nyelvhasználat kapcsolatának mélyebb megértését. Ezen kívül a keletkezett modelleket érdekes lehetne olyan szempontból is megvizsgálni, hogy attribútum vektorokkal kiegészítve – mint például használt emotikonok halmozódása, a leggyakoribb szófajok, vagy a személyes névmások aránya – az illeszkedésük javítható lenne-e.

Hivatkozások

Aragón, M. E., López-Monroy, A. P.: A Straightforward Multimodal Approach for Author Profiling. Notebook for PAN at CLEF 2018. In: CLEF 2018 Evaluation Labs and Workshop -- Working Notes Papers. CEUR-WS.org, Padova (2018)

Argamon, S., Koppel, M., Fine, J., Shimoni, A. R.: Gender, genre, and writing style in formal written texts. In: Text 23(3). pp. 321–346. (2003)

Breuer, J., Kmetty, Z., Haim, M., Stier, S.: User-focused approaches for collecting Facebook data in the "post-API" age. Kézirat (2021)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deepbidirectional transformers for language understanding. In: Proceedings ofthe 2019 Conference of the North American Chapter of the Association forComputational Linguistics: Human Language Technologies, Volume 1 (Longand Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)

Fosch-Villaronga, E., Poulsen, A., Soraa, R. A., Custers, B. H. M.: A little bird told me your gender: Gender inferences in social media. In: Information Processing & Management 58(3). pp. 1-13. (2021)

Huffaker, D., Calvert S.: Gender, Identity, and Language Use in Teenage Blogs. In: Journal of Computer-Mediated Communication, 10(2.) (2005)

James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning. pp. 303-332. Springer, New York (2013)

Karvalics, L. Z.: Információs kultúra, információs műveltség - egy fogalomcsalád értelme, terjedelme, tipológiája és története. In: Információs Társadalom – 12(1.) pp. 7-43. (2012)

Kovács, E.: Többváltozós adatelemzés. pp. 126-146. Typotex Könyvkiadó, Budapest (2014)

Markov, I., Gómez-Adorno, H., Sidorov, G.: Language- and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling. In: CLEF 2017 Evaluation Labs and Workshop -- Working Notes Papers, 11-14 September, Dublin, Ireland. CEUR-WS.org, Dublin (2017)

Nemeskey, D. M.: Introducing huBERT. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 3–14. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2021)

Nemeskey, D. M.: Natural Language Processing Methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd Tudományegyetem, Budapest (2020)

Newman, M. L., Groom, C. J., Handelman, L. D., Pennebroke, J. W.: Gender Differences in Language Use: An Analysis of 14,000 Text Samples. In: Discourse Process. pp. 211-236. (2008)

Rakovics, M.: Adattudomány jegyzet. pp. 73-76. Eötvös Loránd Tudományegyetem, Budapest (2018)

Rangel, F., Rosso, P.: Use of Language and Author Profiling: Identification of Gender and Age. In: Workshop on Natural Language Processing and Cognitive Science (NLPCS-2013). Marseille, Franciaország (2013)

Sap M., Park G., Eichstaedt J. C., Kern M. L., Stillwell D., Kosinski M., Ungar L. H., Schwartz H. A.: Developing Age and Gender Predictive Lexica over Social Media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1146-1151. Doha, Katar (2014)

Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. pp. 199-205. Stanford, California, USA (2006)

Tan, Pang-Ning, Steinbach, M., Kumar, V.: Bevezetés az adatbányászatba. Panem Kft, Budapest (2011)

Vashisth, P., Meehan K.: Gender Classification using Twitter Text Data. In: 31st Irish Signals and System Conference (ISSC). pp. 1-6. Letterkenny, Írország (2020)

Vincze, V., Üveges, I., Szabó, M. K., Takács, K.: A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)

Zhang, C., Zhang, P.: Predicting gender from blog posts. Technical Report. University of Massachusetts Amherst, USA (2010)

Zhang H.: The Optimality of Naive Bayes. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004). pp. 562-567. Miami Beach, Florida, USA (2004)