Mondatszintű szentiment analízis teljesítményének javítása adatkiterjesztéses eljárásokkal

Laki László János, Yang Zijian Győző

Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
{laki.laszlo, yang.zijian.gyozo}@nytud.hu
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.
{laki.laszlo, yang.zijian.gyozo}@itk.ppke.hu

Kivonat A szentiment analízis egy olyan eljárás, amelynek segítségével információkat nyerhetünk az írott tartalmak emocionális töltetét illetően. Számítógépes algoritmusok révén olyan újgenerációs modellek kifejlesztése válik lehetségessé, amelyek korábban nem tapasztalt mennyiségű és minőségű adatot képesek feldolgozni. Ugyanakkor, ezen modellek gyakran rendkívül nagy mennyiségű erőforrást igényelnek, hogy a kívánt teljesítményt elérjék. Ennek megfelelően kiemelt szerep jut azon kutatásoknak, amelyek a modellek struktúrájának és funkciójának javításával magas minőségű eredményeket tudnak generálni amellett, hogy a működésükhöz szükséges erőforrásigényt csökkenteni képesek. A kognitív tudományok szemszögéből fontos célkitűzés, hogy tanulmányozzuk és mélyebb megértésére jussunk egy adott személy mentális állapotának, illetve annak valamilyen aktivitás formájában történő kivetülésének, és ezek potenciális felhasználhatóságát a mögöttes érzések és érzelmek meghatározására. A digitális eszközök egyre elterjedtebb használatával nagy mértékben növekszik a szociális médiában és az egyéb internetes platformokon megjelenő írott tartalmak mennyisége, amely fontos forrásként használható a tartalmakat generáló személyek érzelmeinek meghatározására. Ebből kifolyólag ezek az internetes tartalmak kifejezetten alkalmas lehetőséget nyújtanak szentiment analízis elvégzésére. Az elmúlt években megfelelően finomhangolt nyelvi modellek egyre nagyobb mértékben lettek alkalmazva olyan természetes nyelvelemzési feladatokra, mint a szentiment analízis. Számos hátráltató tényező nehezíti azonban a finomhangolás folyamtát, ilyen például a betanításhoz használható megfelelő méretű korpusz hiánya, vagy az ilyen irányú felhasználásra alkalmas korpuszok teljes hiánya. Jelen kísérletes megközelítésünk során olyan adatnövelő módszereket alkalmazunk, mint a gépi fordítás és a nyelvek közötti transzfer tanítás, és ezek segítségével növeljük a betanító korpuszok méretét. 9 különböző nyelvi modellel végzett kísérleteink eredményét mutatjuk be, melyeket a Hungarian Twitter Sentiment Corpus-on tanítottunk be. Kutatásunk rávilágít arra, hogy a betanított modellek teljesítménye növelhető, ha géppel lefordított szöveget adunk a betanító korpuszhoz.

Továbbá, több általunk használt modell is képes volt jobb teljesítményre a jelenlegi magyar state-of-the-art modelleknél.
Kulcsszavak: szentiment analízis, mondatszintű osztályozás, adatkiterjesztés, gépi fordítás, transzfer tanítás, zeroshot

1. Bevezetés

A szentiment analízis az érzelmek automatizált azonosítása egy adott szövegben és ezek osztályozása olyan kategóriákba, mint negatív, semleges vagy pozitív. A szociális médiát használók köre egyre növekszik, ezáltal óriási mennyiségű szöveges információ keletkezik és áll rendelkezésre az interneten, amely rendkívül hasznosnak bizonyul a tartalmakat szerző személyek érzelmeinek meghatározásában. Mivel egyedülálló lehetőség nyílik erre számos potenciális alkalmazhatósággal karöltve, mind az akadémia mind az ipar képviselői komoly és egyre növekvő érdeklődést mutatnak a szövegekből történő szentimentális információ kinyerése iránt (Hoang és mtsai, 2019).

A neurális háló alapú nyelvi modellezés áttörést eredményezett a legtöbb természetes nyelvfeldolgozási feladatban. Szinte egyik napról a másikra jelennek meg újabb és újabb modellek, amelyek túlszárnyalják a korábbi modellek teljesítményét. A nyelvi modellek nemcsak a betanításhoz használt adat tekintetében különböznek, hanem a használt neurális hálók belső struktúrájában és a tanítási módszerekben is. Következésképpen, egy adott természetes nyelvfeldolgozási feladat megoldható egy megfelelően választott nyelvi modellel. Továbbá, érdekes felvetés, hogy egy nyelvspecifikus modell túlteljesíthet-e egy többnyelvű modellt, amely egy kulcskérdés a magyar nyelvű szentiment analízis esetében is. A jelenlegi legjobb eredményt hozó megoldás egy adott természetes nyelvfeldolgozási feladat megoldására egy már előtanított nyelvi modell további finomhangolása egy alkalmazási területre specifikus feladatra. Az ilyen rendszerek minősége nemcsak az előtanított modellektől függ, hanem a finomhangoló adathalmaz méretétől is. Az olyan természetes nyelvfeldolgozási feladatok, mint a magyar nyelvre kifejlesztett szentiment analízis komoly érdeklődésre tett szert az ipari szegmens részéről, ellenben limitálva van a szabadon elérhető adatok szempontjából. Jelen állás szerint nem találtunk olyan korábban publikált megoldást a magyar nyelvű szentiment analízisre, ami azt jelenti, hogy az általunk kínált alkalmazás a legelső ilyen megjelenés a témában.

Kutatásunk során egy gépi fordító rendszert alkalmaztunk angol nyelvű szentiment analízis adathalmaz magyar nyelvre történő lefordításához. A gépi fordító rendszer betanítását magunk végeztük. A lefordított korpuszt integráltuk a rendszerünkbe.

A 2. Rész bemutatja az eddig leírt megoldásokat a témában, a 3. Rész a gépi fordító rendszert, a 4. Részben bemutatásra kerülnek a korpuszok és a modellek, majd az 5. Részben ismertetjük az elvégzett kísérleteket. Az eredmények az 6. Részben találhatók.

Modelljeink és szkriptjeink megtalálhatóak a Github[1] és Hugging Face[2] oldalainkon.

2. Kapcsolódó irodalom

A szentiment analízis egy rendkívül komplex természetes nyelvfeldolgozási feladat és számos területen alkalmazható, így például szociális média monitorozása során (Neri és mtsai, 2012), befektetők döntéshozatali folyamatának támogatására a gazdasági hírek szemantikus kontextusának elemzésével (Lutz és mtsai, 2018; Saura és mtsai, 2019), digitális marketing esetében (Kinholkar és Waghmare, 2016), pszichés állapot vizsgálatánál (Jo és mtsai, 2018), valamint számos további területen, ahol az ilyen magas szintű szövegbányászati eljárások nagyban megnövelhetik a hatékonyságot. Jelenleg többféle irányban zajlanak a szentiment analízis célú fejlesztések. Kezdetben a cél az volt, hogy osztályozni lehessen dokumentumokat és szövegeket az általános polaritásuk alapján (negatív, pozitív vagy semleges) (Pang és mtsai, 2002). Egy másik fő irány az aspektus-alapú módszer, amely kifinomultabb és célja, hogy azonosítsa egy tárgy vagy bármilyen entitás olyan aspektusait, melyek az érzelmek kiváltásáért felelősek (Pontiki és mtsai, 2014). Egy alternatív stratégia pedig a mondatszintű szentiment analízis, mely a mondatot állítja vizsgálódása középpontjába, tehát feladata egy szövegben szereplő mondat érzelmi polarizáltságát (Feldman, 2013; Lutz és mtsai, 2018).

A Kahla és mtsai (2021) által publikált nyelvek közötti transzfer kísérletek (Kahla és mtsai, 2021) bebizonyították, hogy egy kétlépcsős finomhangolási megközelítés jelentősen javíthatja a természetes nyelvfeldolgozási feladat kimenetelét. Az első finomhangolás angol nyelven történt, melyet a második arab nyelvű finomhangolási lépés követett.

Magyar nyelvre elsősorban aspektus-alapú szentiment analízis korpuszok és eszközök léteznek. OpinHuBank (Miháltz, 2013) egy manuálisan annotált korpusz, mely a véleménykutatást és a szentiment analízist támogatja. 10000 mondatból áll és személyneveket tartalmaz a főbb magyarországi honlapokról és blogokról. Minden egyes entitás értékelve volt 5 különböző emberi annotátor által a szentiment polaritásra nézve (semleges, pozitív vagy negatív). A HuSent (Szabó és mtsai, 2016) egy manuálisan annotált szentiment korpusz, mely a Dívány honlapról[3] vett magyar nyelvű véleményeket tartalmaz a különböző termékekkel kapcsolatban. A korpusz 154 véleményt tartalmaz, mintegy 17.000 mondatból és 251.000 tokenből áll. Steinberger és mtsai (Steinberger és mtsai, 2011) kutatásai során előállított egy aspektus-alapú szentiment korpuszt többnyelvű párhuzamos korpuszokkal, amely tartalmaz egy magyar nyelvű alkorpuszt is.

Jelen kutatás keretében a mondatszintű szentiment analízisre fókuszáltunk, emellett aspektus-alapú szentiment analízis kísérleteket is tervezünk a jövőben.

[1] https://github.com/nytud/sentiment-analysis
[2] https://huggingface.co/NYTK
[3] http://divany.hu

3. Adatkiterjesztés gépi fordítással és nyelvek közötti transzferrel

A bevezetésben említettek alapján egyértelműen kirajzolódik, hogy a betanításhoz szükséges adathalmaz mérete kiemelkedő fontosságú a neurális háló alapú modellek estében. Sajnos jelenleg nem áll rendelkezésre megfelelő minőségű adat, a manuális módon történő adatgenerálás pedig rendkívül költséges. Kutatásunk során ezen hiány áthidalására gépi fordítást és nyelvek közötti transzfert alkalmaztunk az adathalmazunk méretének növelése érdekében (Fadaee és mtsai, 2017). Azzal az ötlettel álltunk elő, hogy már meglévő angol nyelvű korpuszokat használunk és lefordítva azokat kiegészítő betanító adatként alkalmazzuk. Az ötlet a gépi fordítás területéről származik, amikor visszafordított korpuszokat használnak annak érdekében, hogy javítsák a fordítás minőségét egy alacsony ellátottságú nyelvpár esetén (Poncelas és mtsai, 2018).

A lefordított korpuszok felhasználására két lehetőség adódik. Az első esetben a nyelvek közötti adattranszfer során az angol nyelvű korpusz úgy van alkalmazva, mint egy első körös finomhangolásos adathalmaz mielőtt felhasználásra kerülne, mint doménen belüli magas minőségű változat (későbbiekben fordított+finomhangolás megjelöléssel utalunk rá). Másodsorban, a segédkorpusz konkatenálható a doménen belülivel (erre mix megjelöléssel utalunk). Az első finomhangolási lépés a konkatenált adattal (mix) történik, majd a második lépésben a doménen belülivel (mix+finomhangolás).

Kísérletes munkánk során a MarianNMT (Junczys-Dowmunt és mtsai, 2018) szoftvercsomagot alkalmaztuk, amely egy szabad forráskódú C++ programnyelven írott alkalmazás. Ez egy könnyen installálható, alaposan ledokumentált, memória- és erőforráskímélő implementáció, amely gyakori használatnak örvend az akadémiai és a fejlesztői körökben egyaránt (Barrault és mtsai, 2019). Egy transzformeren alapuló enkóder-dekóder struktúra lett alkalmazva SentencePiece tokenizáláshoz. A tokenizáló algoritmus egy általános szószedetet használt mindkét nyelv irányába 32000-es szótárméret mellett. Az alapbeállítások mellett használtuk az alkalmazást a rejtett rétegek méretét és az optimalizációs metrikákat illetően. A betanító adathalmazként a ParaCraw16 és az OpenSubtitles (Lison és Tiedemann, 2016) platformokról származó angol-magyar nyelvpárok kerültek felhasználásra. A teljes betanító adathalmaz 45,5 millió szegmenst és 573 millió angol nyelvű tokent tartalmaz. A rendszer 36,873 százalékos BLEU (Papineni és mtsai, 2002) értéket ért el a tesztelés során (3000 teszt adathalmazból random módon kiválasztott szegmenssel).

4. Korpuszok és Modellek

A mondatszintű szentiment analízis betanításhoz a Precognox Kft.[4] Által készített Hungarian Twitter Sentiment[5] (MTS) korpuszt használtuk. Jelenleg ez az

[4] https://www.precognox.hu

[5] http://opendata.hu/dataset/hungarian-twitter-sentiment-corpus

egyetlen szabadon elérhető korpusz, amely magyar nyelvű szentiment analízisre használható. Az 1. táblázat foglalja össze a MTS korpusz főbb jellegzetességeit. A skálázás 1-től 5-ig a következő módon történik: 1 – nagyon negatív, 2 – negatív, 3 – semleges, 4 – pozitív, 5 – nagyon pozitív. Erre a korpuszra a MTS5 megjelöléssel utalunk. Egy másik esetben a 0 és 1 értékeket negatívnak, a 4 és 5 értékeket pedig pozitívnak értékeltük, a 3-as értékeket pedig kihagytuk. Ez utóbbi korpuszra MTS2 néven utalunk.

Adataink kiterjesztéséhez az SST2 és SST5 korpuszokat választottuk (Wang és mtsai, 2018). Mindkét korpusz angol nyelvű mondatokat tartalmaz. Gépi fordítás segítségével ezeket a korpuszokat felhasználtuk hozzáadott adatként (SST2_hu és SST5_hu megjelöléssel utalunk ezekre a magyarra fordított korpuszokra). Az 1. táblázat alapján látható, hogy az SST korpuszok lényegesen nagyobbak, mint a MTS korpuszok.

	SST2	SST5	MTS2	MTS5
Segments	70.045	11.855	2.737	4.000
Token	652.594	227.245	33.279	46.683
Type	17.516	21.699	15.900	21.689
Tanító anyag	67.350	8.544	2.193	3.200
Validációs anyag	-	-	273	400
Teszt anyag	873	1.101	273	400
Osztályok	0;1	1;2;3;4;5	0;1	1;2;3;4;5

1. táblázat. A korpuszok tulajdonságai.

Kutatásunk során a MTS5 korpuszt felosztottuk 90-10 százalékos arányban betanító és tesztelő korpuszokra. Az első 400 tweet képezi a tesztelő korpuszunkat. A MTS2 esetében a 3-as értéket tartalmazó dokumentumok kihagyásra kerültek. A 2. táblázat mutatja korpuszok felépítését és az adatok megoszlását. Kísérleteink során 6 különböző egynyelvű (magyar) kontextuális nyelvi modellt, 2 többnyelvű modellt és egy klasszikus szóbeágyazáson alapuló modellt használtunk.

	train	test	train	test
label	SST2		MTS2	
0	29.755	428	1.021	108
1	37.539	444	1.448	162
	SST5		MTS5	
1	1.089	139	93	12
2	2.200	289	936	88
3	1.594	229	1.111	150
4	2.259	279	1.349	141
5	1.266	165	111	9

2. táblázat. Címkék eloszlása a korpuszokban.

A fejezet második részét e modellek rövid bemutatására szánjuk.

huBERT (Nemeskey, 2021): egy magyar BERT (Devlin és mtsai, 2019) base modell, mely a Webcorpus 2.0 korpuszon (Nemeskey, 2020) lett betanítva, ez utóbbi a Common Crwal webarchívumból és a magyar nyelvű Wikipédiából tevődik össze. Kisbetűsített és kisbetűsítés nélküli verzió is készült a huBERT-hez. Kiemelendő, hogy a huBERT túlszárnyalja a többnyelvű BERT modellt számos feladatban, így például a maszkolt nyelvi modellezésben, névelemfelismerésben vagy névcsoport azonosításban. Jelenleg a state-of-the-art megoldásnak számít a névelemfelismerés azonosítás területén.

HILBERT (Feldmann és mtsai, 2021): egy BERT large modell magyar nyelvre, amely kiemelkedő teljesítményt nyújt nyelvfeldolgozási feladatokban. A HILBERT a NYTK-BERT (Feldmann és mtsai, 2021) korpuszon lett betanítva. A modell számos feladat esetén figyelemreméltó eredményeket ér el, így például névelemfelismerésben és összefoglaló generálásban (Yang és mtsai, 2021). Az egyik előnye ennek a modellnek a huBERT-tel szemben, hogy több paramétert tartalmaz, ellenben kevesebb betanító adattal.

HIL-RoBERTa[6]: az egyik legfőbb kihívás a nyelvi modellek optimalizálása során az előtanításban mutatkozik. Mivel az előtanítás egy rendkívül erőforrásigényes folyamat, ezért kiemelten fontos az új módszerek kutatása és kifejlesztése, amelyek szignifikáns javulást tudnak indukálni ezen a területen. A RoBERTa (Liu és mtsai, 2019) a Robustly optimized BERT pre-training approach angol nyelvű szakirodalmi megjelölés rövidítéséből származik. A RoBERTa kiemelkedő eredményeket ér el olyan sztenderd feladatokban, mint a GLUE (Wang és mtsai, 2018), a RACE (Lai és mtsai, 2017) vagy a SQuAD (Rajpurkar és mtsai, 2016), amelyet, hogy lényegesen kevesebb erőforrást használ optimalizált előtanítási paradigmájának köszönhetően. A HIL-RoBERTa egy RoBERTa small modell, amely a magyar nyelvű Wikipédián lett betanítva. HIL-RoBERTa kiváló eredményeket ér el névelemfelismerésben és névcsoport azonosításban, megközelítve a huBERT teljesítményét.

HIL-ALBERT[7]: számos erőfeszítés irányul arra, hogy emeljük a nyelvi modellek teljesítményét a célfeladatra történő előtanítás erőforrásigényének csökkentése mellett. Az ALBERT (A Lite BERT rövidítése) egy olyan modell, amely paraméter csökkentő technikákat foglal magába (Lan és mtsai, 2020). A magyar nyelvre történő implementáció során két előtanított, kisbetűsítés nélküli ALBERT modell készült el: az egyik a magyar nyelvű Wikipédián lett betanítva (a Webcorpus 2.0 korpusz része), a másik pedig a NYTK-BERT korpusz egy részén. Kutatásunk során a HIL-ALBERT NYTK modellt használtuk.

HIL-ELECTRA[8]: az ELECTRA (Efficiently Learning an Encoder that classifies Token Representation Accurately) (Clark és mtsai, 2020) modellek egy sikeres alternatív megoldást nyújtanak a maszkolt nyelvi modellezés (MLM) mellett azáltal, hogy felcserélt token detektálást alkalmaznak, amely egy önfelügyelő előtanítási feladat, melynek során a modell megtanulja megkülönböztetni az ere-

[6] https://hilanco.github.io/models/roberta.html
[7] https://hilanco.github.io/models/albert.html
[8] https://hilanco.github.io/models/electra.html

deti bemenetet a mesterségesen generáltalt behelyettesítésektől. Az ELECTRA modellek a GAN (Generative Adversarial Network) módszeren alapulnak. Kísérletes eredmények azt mutatják, hogy ez a módszer hatékony és nagy teljesítményű más módszerekkel összehasonlítva. Az ELECTRA magyar nyelvű implementációja során két verzió született, az ELECTRA wiki és az ELECTRA NYTK-BERT. Az előbbi a magyar Wikipédián lett betanítva, míg az utóbbi az NYTK-BERT korpuszon. Kutatásunk során a HIL-ELECTRA NYTK modellt alkalmaztuk.

HILBART[9]: a BART (a Bidirectional and Auto-regressive Transformers) (Lewis és mtsai, 2020) alapú megközelítések komoly potenciállal rendelkező eszközök a seq2seq (sequence to sequence) előtanítás tekintetében. A BART gyakorlatilag ötvöz egy BERT (Devlin és mtsai, 2019) és egy GPT (Radford és Narasimhan, 2018) típusú modellt. A BART a szöveggenerálási feladatokban teljesít a legjobban de kiemelkedő eredményeket ér el diszkriminatív és összefoglaló feladatokban is. A BART magyar nyelvű implementációjának eredményeként jöttek létre a HILBART modellek. Ezek a HILBART large web (Webcorpus 2.0 1 százalékán tanítva), a HILBART base web (Webcorpus 2.0 10 százalékán tanítva) és a HILBART base wiki (magyar nyelvű Wikipédián tanítva). Kutatásunk során a HILBART base web modellt használtuk.

mBERT (Devlin és mtsai, 2019): az mBERT (multilingual BERT) struktúráját tekintve a BERT-en alapul, ugyanazon betanítási paradigmát is használja azzal a fontos különbséggel, hogy az előtanítás során 104 különböző nyelv Wikipédia cikkek szövegeit használták. Az mBERT modell alkalmazása különösen előnyös akkor, ha alacsonyan ellátott nyelvekről van szó, vagyis amikor kevés annotált mondat áll rendelkezésre az adott nyelven. Nyelvek közötti előtanító modelleket (mBERT-et is beleértve) alkalmaztak például egy névelemfelismerési feladat során magyar és ujgur nyelvekre (Chen és mtsai, 2021). Kutatásaink során a kisbetűsítés nélküli mBERT base modellt alkalmaztunk.

XLM-RoBERTa (Conneau és mtsai, 2020): a nyelvek közötti érthetőség (angolul Cross-Language Understanding, rövidítve XLU) elérése egy komoly kihívás és egy innovációs gyorsítóként szolgál a többnyelvű modellek esetében. 2020-ban a Facebook mesterséges intelligenciával foglalkozó csapata előállt az XLM-RoBERTa (XLM-R-ként is rövidítve) modellel, amely egy transzformeren alapuló többnyelvű maszkolt nyelvi modell. A modell előtanítása során a CC-100 korpuszt használták, amely 100 különböző nyelv szövegeit tartalmazza, köztük magyar nyelvű szövegeket is (magyar nyelvű tokenek száma: 7807 millió, magyar nyelvű korpusz mérete: 58,4 GiB). A szerzők publikációja alapján az XLM-RoBERTa versenyképes eredményeket ért el számos sztenderd feladat elvégzése során olyan egynyelvű modellekkel összevetve, mint például a RoBERTa. Továbbá, XLM-R képes volt túlteljesíteni az mBERT-et nyelvek közötti osztályozásban olyan nyelvek esetében, ahol relatíve kevés a rendelkezésre álló forrásanyag. Figyelemre méltó, hogy state-of-the-art eredményeket ért el az XLM-RoBERTa XNLI, NER és nyelvek közötti válaszadási feladatokban. Kutatásunk során az XLM-RoBERTa base modellt használtunk.

[9] https://hilanco.github.io/models/hilbart.html

fastText (Joulin és mtsai, 2016b,a): a fastText szintén a Facebook mesterséges intelligenciával foglalkozó csapatának fejlesztése, melynek célja a szövegosztályozás és a reprezentációs tanulás elősegítése. A létrehozott paradigma azon alapul, hogy karakter n-grammokat foglal az úgynevezett skipgram modellbe, amely egy gyors és hatékony megoldást kínál anélkül, hogy előfeldolgozás vagy felügyelet szükséges lenne (Bojanowski és mtsai, 2017). Szövegosztályozás szempontjából más deep learning-alapú megoldásokkal összevethető a teljesítménye a pontosság tekintetében, és egy lényegesen gyorsabb lehetőség tanítás és kiértékelés szempontjából (Joulin és mtsai, 2017). A platformon elérhetőek szóvektorok angolra és 157 másik nyelvre, ezáltal egy nagyon kézenfekvő és lehetőségekkel teli eszköznek számít a többnyelvű nyelvfeldolgozás terén.

5. Kísérletek

Kutatásunk során 7 különböző kísérletet végeztünk el:

- **eredeti**: minden előtanított modell finomhangolásra került az eredeti MTS korpuszon. Ezt tekintjük az alap eljárásnak.
- **zeroshot**: többnyelvű modellek képesek magyar nyelvű NLP feladatok predikciójára. Ez esetben angol nyelvű korpuszokat használtunk finomhangolásra és a rendszernek magyar nyelvű mondatokra kellet prediktálnia.
- **transzfer**: többnyelvű modellek finomhangolva lettek az SST korpuszon, majd további finomhangolásra kerültek a MTS tanító korpuszon.
- **fordított**: minden előtanított modell finomhangolva lett a magyarra fordított SST korpuszon (SST_hu).
- **fordított+finomhangolás**: minden előtanított modell finomhangolva lett az SST_hu korpuszon, majd ismét finomhangolva lett a MTS tanító korpuszon.
- **mix**: minden előtanított modell finomhangolva lett a konkatenált és összekevert SST_hu és MTS tanító korpuszon, majd le lett tesztelve a MTS tesztelő korpuszon.
- **mix+finomhangolás**: minden előtanított modell finomhangolva lett a konkatenált és összekevert SST_hu és MTS tanító korpuszon, majd újabb finomhangolás következett a MTS tanító korpuszon

Az elvégzett kísérletek eredményei a MTS tesztelő korpuszon lettek kiértékelve.

A jobb összehasonlíthatóság érdekében ugyanazokat a hiperparaméter beállításokat alkalmaztuk gyakorlatilag az összes modellre. A hiperparaméterek a következők: learning rate: 2e-5, batch méret: 32 eszközönként (4 x GPU), epoch érték: 4, maximális szekvenciahossz: 128. A HILBERT modell esetében a CUDA memóriatúllépés elkerülése végett a batch méretet eszközönkénti 8-as értékre módosítottuk. Az ELECTRA modellek egyetlen GPU felhasználása mellett futottak. Végül, fastText esetében GPU nem került felhasználásra, csak CPU, a batch méret pedig 1-es értékre lett állítva. Minden kísérletnél 4 darab GeForce RTX 2080 Ti típusú videókártyát és 40 darab Intel(R) Xeon(R) Silver 4114 típusú CPU-t használtunk.

A transzformer modellek (kivétel az ELECTRA) finomhangoláshoz a Hugging Face által rendelkezésre bocsátott „transformers text classification library"-t[10], az ELECTRA finomhangolásához a Google által implementált kódot[11], míg a fastText esetében a Facebook által közzétett eszközt[12] használtuk. Minden általunk használt és e publikáció keretében bemutatott modell és implementáció megtalálható a projekt honlapunkon[13].

6. Eredmények

A 3. táblázat foglalja össze kísérleteink eredményeit. Általánosan elmondható, hogy lefordított szöveg hozzáadása a tanító korpuszhoz javítja a mondatszintű szentiment analízis osztályozás teljesítményét. Minden egyes általunk tesztelt modell esetében valamely fordítási metódus jobb eredményeket hozott, mint az alap eljárás.

Három minőségi sáv definiálható az általunk használt modellek teljesítménye alapján. A leggyengébb rendszerek a HILBART és a fastText révén keletkeztek. Ezek a várakozásunknak megfelelő eredmények, hiszen a HILBART elsősorban szöveggenerálásra használatos, a fastText pedig egy bizonyos tekintetben elavult, statikus, nem kontextuális szóreprezentációs eljárás, mely rosszabbul teljesít a kontextuális nyelvi modellekhez képest. Ugyanakkor hozzá kell tennünk, hogy a fastText alkalmazás lényegesen kevesebb erőforrást igényel a rendszer betanításához és online predikció esetén csak CPU-t használ.

A második minőségi sáv tartalmazza azokat a rendszereket, amelyek 77-80 százalékos pontossági értéket mutattak a bináris osztályozásban és 58-63 százalékos pontosságot az 5-osztályos besorolás alkalmával.

Végül három rendszer található a legfelső minőségi sávban (huBERT, XLM-RoBERTa és HILBERT modellek) 85,5 százalékos pontossággal a bináris osztályozásban és 66-69 százalékos pontossággal az 5-osztályos besorolásban. Érdekes eredményként könyvelhető el, hogy az XLM-RoBERTa többnyelvű modell jobban teljesített a MTS2 feladatban, mint a magyar nyelvre specifikus huBERT modell, amely a state-of-the-art nyelvi modell a legtöbb természetes nyelvfeldolgozási feladatban. Továbbá, a HILBERT modell szintén túlszárnyalta a huBERT modellt a MTS2 feladatban, amely szintén várt eredmény, hiszen a kevesebb betanító adatot ellensúlyozni tudta a modell nagyobb mérete, a több paraméter és a hozzáadott finomhangolási adat.

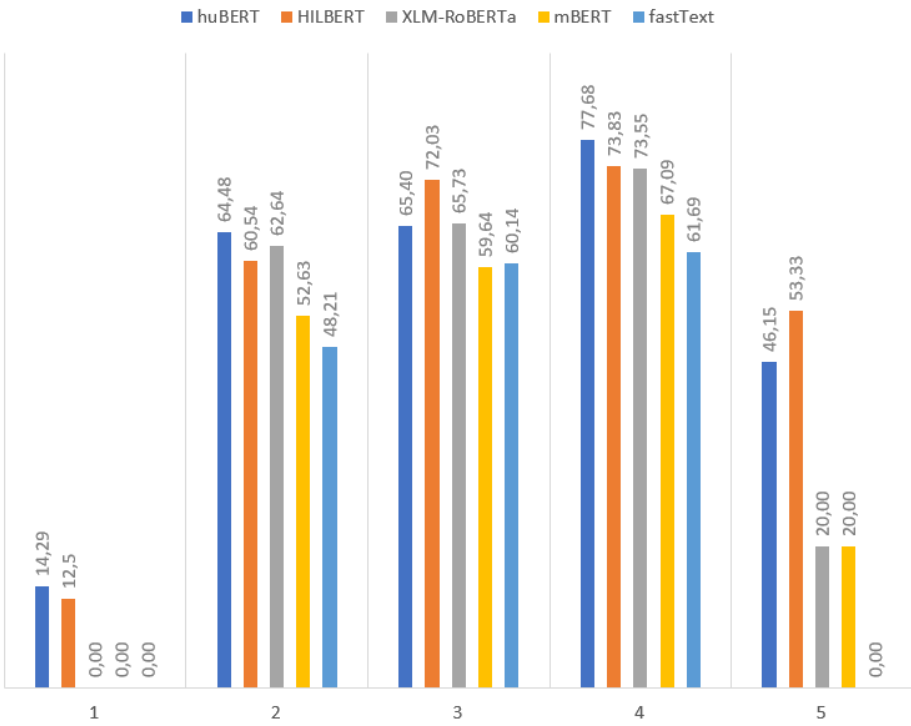
Az 1. ábrán összehasonlítottunk 5 különböző modell teljesítményét az F-mértékek tekintetében. A magyar nyelvű state-of-the-art huBERT modell, a HILBERT large modell, a nem-kontextuális fastText és két többnyelvű modell került összehasonlításra. Az egyetlen szignifikáns eredmény abban mutatkozott, hogy a huBERT és a HILBERT modellek több 1-es és 5-ös értéket prediktáltak, mint a

[10] https://github.com/huggingface/transformers/tree/master/examples/pytorch/text-classification
[11] https://github.com/google-research/electra
[12] https://fasttext.cc
[13] https://github.com/nytud/sentiment-analysis

		MTS2	MTS5
huBERT	eredeti	84.07	66.00
	fordított	73.33	29.25
	fordított+finomhangolás	85.55	66.50
	mix	**85.55**	**68.99**
	mix+finomhangolás	84.81	68.00
HILBERT	eredeti	83.33	68.00
	fordított	74.07	34.75
	fordított+finomhangolás	82.59	67.75
	mix	82.22	**68.50**
	mix+finomhangolás	**85.56**	68.00
HIL-RoBERTa	eredeti	75.92	59.15
	fordított	48.89	29.75
	fordított+finomhangolás	**79.63**	56.75
	mix	76.66	**59.25**
	mix+finomhangolás	77.78	57.99
HIL-ALBERT	eredeti	75.56	55.49
	fordított	52.59	28.75
	fordított+finomhangolás	77.03	56.75
	mix	72.22	60.50
	mix+finomhangolás	**77.41**	**60.75**
HIL-ELECTRA	eredeti	78.89	59.11
	fordított	55.02	37.34
	fordított+finomhangolás	**79.93**	61.15
	mix	76.58	60.90
	mix+finomhangolás	79.18	**62.66**
HILBART	eredeti	71.11	51.25
	fordított	47.77	31.00
	fordított+finomhangolás	74.07	53.25
	mix	71.48	52.50
	mix+finomhangolás	**76.66**	**54.75**
mBERT	eredeti	78.51	57.74
	zeroshot	47.41	30.50
	transzfer	78.52	57.99
	fordított	48.88	28.75
	fordított+finomhangolás	**79.25**	56.75
	mix	77.77	56.99
	mix+finomhangolás	78.89	**59.75**
XLM-RoBERTa	eredeti	83.33	63.49
	zeroshot	68.88	40.99
	transzfer	84.81	66.25
	fordított	68.51	35.25
	fordított+finomhangolás	85.18	66.00
	mix	85.18	66.25
	mix+finomhangolás	**85.56**	**66.50**
fastText	eredeti	71.9	53.2
	fordított	62.2	32.0
	fordított+finomhangolás	73.3	**56.2**
	mix	74.1	51.7
	mix+finomhangolás	**75.6**	53.5

3. táblázat. Mondatszintű szentiment analízis eredmények.

többnyelvű modellek vagy a fastText. A fastText nem prediktált sem 1-es, sem pedig 5-ös értékeket. Ez azt jelenti, hogy az 1-es és 5-ös értékek ritkán fordulnak elő a betanító korpuszban (lásd 2. táblázatban). A huBERT és a HILBERT magyar nyelvű modellek, az előbbi egy 9 milliárd tokent tartalmazó korpuszon lett betanítva, míg az utóbbi egy large modell 340 millió paraméterrel, aminek a segítségével még kifinomultabb részletek megtanulására is képesek.



1. ábra: MTS5 F-mértékek összehasonlítása az osztálycímkék függvényében.

A legjobban teljesítő magyar (huBERT[14,15]) és többnyelvű (XLM-RoBERTa[16,17]) modellek megtalálhatóak a Hugging Face oldalunkon.

[14] https://huggingface.co/NYTK/sentiment-hts2-hubert-hungarian
[15] https://huggingface.co/NYTK/sentiment-hts5-hubert-hungarian
[16] https://huggingface.co/NYTK/sentiment-hts2-xlm-roberta-hungarian
[17] https://huggingface.co/NYTK/sentiment-hts5-xlm-roberta-hungarian

7. Összegzés

Jelen tanulmányunk új megközelítéseket alkalmaz a Twitter közösségi média platformról származó magyar nyelvű szövegek mondatszintű szentiment analízisében. Kísérletes eredményeink alapján az a következtetés vonható le, hogy magyarra fordított szövegek hozzáadása a korpuszokhoz képes javítani a modellek teljesítményét, amely egy rendkívül fontos előrelépés a szentiment analízis folyamatának optimalizálásában. Kiemelkedő jelentőségű, hogy az általunk használt modellek jobban teljesítettek több feladatban is a jelenlegi state-of-the-art-nak számító huBERT modellnél, amely ígéretes új utakat nyit meg a jelen publikáció keretében bemutatott eredmények szélesebb körű alkalmazhatóságát illetően, elősegítve ezzel a szentiment analízishez kapcsolódó területek haladását. Mindamellett, kutatási eredményeink kifejezetten relevánsak az olyan új stratégiák kialakításában, ahol a rendelkezésre álló tanítóanyag mennyisége nem elegendő egy osztályozási modell tanítására.

Hivatkozások

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., MĂźller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019), http://www.aclweb.org/anthology/W19-5301

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)

Chen, S., Pei, Y., Ke, Z., Silamu, W.: Low-resource named entity recognition via the pre-training model. Symmetry 13(5) (2021)

Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. In: ICLR (2020)

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2020)

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

Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 567–573. Association for Computational Linguistics, Vancouver, Canada (Jul 2017), https://aclanthology.org/P17-2090

Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM 56(4), 82–89 (Apr 2013)

Feldmann, Á., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Yang, Z.G., Váradi, T.: Hilbert, magyar nyelvű bert-large modell tanítása felhő környezetben. XVII. Magyar Számítógépes Nyelvészeti Konferencia pp. 29–36 (2021)

Hoang, M., Bihorac, O.A., Rouces, J.: Aspect-based sentiment analysis using BERT. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics. pp. 187–196. Linköping University Electronic Press, Turku, Finland (Sep–Oct 2019)

Jo, H., Kim, S.M., Ryu, J.: What we really want to find by sentiment analysis: The relationship between computational models and psychological state (2018)

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016a)

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016b)

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (Apr 2017)

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Kahla, M., Yang, Z.G., Novák, A.: Cross-lingual fine-tuning for abstractive arabic text summarization. In: Proceedings of International Conference Recent Advances In Natural Language Processing (RANLP 2021). pp. 660–668. INCOMA Ltd., Shoumen, Bulgaria (2021)

Kinholkar, S.A., Waghmare, P.K.C.: Enhance digital marketing using sentiment analysis and end user behavior. In: International Research Journal of Engineering and Technology (IRJET). vol. 3 (2016)

Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: Large-scale ReAding comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 785–794. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: Proceedings of the Eighth International Conference on Learning Representations (2020)

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020)

Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 923–929. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1147

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)

Lutz, B., Pröllochs, N., Neumann, D.: Sentence-level sentiment analysis of financial news using distributed text representations and multi-instance learning (2018)

Miháltz, M.: OpinHuBank: szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 343–345. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2013)

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

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

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., By, T.: Sentiment analysis on social media. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 919–926 (2012)

Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). pp. 79–86. Association for Computational Linguistics (Jul 2002), https://www.aclweb.org/anthology/W02-1011

Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002), https://www.aclweb.org/anthology/P02-1040

Poncelas, A., Shterionov, D.S., Way, A., de Buy Wenniger, G.M., Passban, P.: Investigating backtranslation in neural machine translation. In: Proceedings of the 21st Annual Conference of the European Association for Machine Translation. pp. 249–258. Alacant, Spain (2018)

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 27–35. Association for Computational Linguistics, Dublin, Ireland (Aug 2014)

Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016)

Saura, J.R., Palos-Sanchez, P., Grilo, A.: Detecting indicators for startup business success: Sentiment analysis using text data mining. Sustainability 11(3), 917 (Feb 2019), http://dx.doi.org/10.3390/su11030917

Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R., van der Goot, E.: Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. pp. 770–775. Association for Computational Linguistics, Hissar, Bulgaria (Sep 2011), https://www.aclweb.org/anthology/R11-1113

Szabó, M.K., Vincze, V., Simkó, K.I., Varga, V., Hangya, V.: A Hungarian sentiment corpus manually annotated at aspect level. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2873–2878. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016)

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 353–355. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)

Yang, Z.G., Agócs, Á., Kusper, G., Váradi, T.: Abstractive text summarization for hungarian. Annales Mathematicae et Informaticae 53, 299–316 (2021)