Neurális entitásorientált szentimentelemző alkalmazás magyar nyelvre

Yang Zijian Győző, Laki László János

Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
{yang.zijian.gyozo, laki.laszlo}@nytud.hu

Kivonat Napjaink egyik kiemelkedő kutatása a szentimentelemzés. Kognitív tudományok számára fontos feladat, hogy mélyebb megértésre jussanak egy személy állapotával kapcsolatban. A nyelvtechnológia fejlődésével lehetőség nyílt a szöveges tartalmakban rejlő érzelmek, emóciók meghatározására. A szociális média elterjedésével és az online platformokon megjelenő tartalmak mennyiségének növekedésével egyre nagyobb igény támad a bennük lévő szentimentek automatikus elemzésére. A szentimentelemzés egyik feladata a szöveg érzelmi töltetét egy adott személyre nézve megvizsgálja, ezt hívjuk entitásorientált szentiment elemzésnek. Nyelvtechnológia szempontjából abban nehezedik a feladat, hogy először magát a személyt kell felismerni a szövegben és utána a hozzátartozó emóciót az adott környezetben. Kutatásunkban különböző neurális nyelvmodelleket finomhangoltunk entitásorientált szentimentelemzésre, majd adatkiterjesztési eljárásokkal növeltük a modellek teljesítményét. Továbbá tanulmányunkban egy olyan pipeline-t mutatunk be, amely először egy adott szövegben megkeresi a személyneveket, majd meghatározza a megtalált személynevekhez tartozó érzelmi töltetet az adott szövegkörnyezetben, majd végül az egész szövegre is megállapítja a szentimentet. A pipeline moduljai mind transzformer-alapú finomhangolt neurális modellek. Összesen három modell dolgozik az alkalmazásban. Egy névelemfelismerő, egy entitásorientált szentimentelemző és egy mondatalapú szentimentelemző modell. A pipeline-hoz készítettünk egy demo felületet is. Végül, de nem utolsó sorban kiértékeltük a pipeline teljesítményét.
Kulcsszavak: entitásorientált szentiment analízis, névelemfelismerés, adatkiterjesztés, gépi fordítás, transzfer tanítás, zeroshot

1. Bevezetés

A mesterséges intelligencia által támogatott megoldások megjelenése a szentiment analízisben megegyezik azokkal a tendenciákkal, amelyek megfigyelhetők a kognitív infokommunikáció területén is. Kiemelt fontossággal bír, hogy jobban érthetővé váljon az adatbányászat hátterében álló elméleti keretrendszer, amely elősegíti a felhasználói érzelmi állapotokat vagy tömegvélemények alakulását prediktáló alkalmazások kifejlesztését (Baranyi és Csapó, 2012). Az online platformokon megjelenő tartalmak mennyiségének növekedésével és a szociális

média térhódításával nagy igény alakult e tartalmak automatikus elemzése. A nyelvtechnológia fejlődésével lehetőség nyílt a szövegek magas minőségű automatikus elemzésére. Az információ világában egyre fontosabbak azok az elemzések, amelyek egy adott termékről vagy személyről adnak adott szempontból összefoglaló információt. Ezt a célt szolgálhatja egy jól működő entitásorientált szentimentelemző, amellyel kimutatást lehet vele készíteni egy adott termékről vagy személyről, hogy hogyan vélekednek róla az emberek. Másik alkalmazási irány, ahol egyre nagyobb igény mutatkozik a szentiment analízisre, az a pszichológiai kutatások, elemzések. A páciensek interjúinak és beszámolóinak szentimentelemzésével következtetni lehet az adott betegség fejlődésére. Illetve adott dokumentumokban bizonyos jelenségek, mint például stigmatizáció jelenségének keresésére is lehet használni egy ilyen eszközt.

Jelen tanulmányban először bemutatjuk a neurális entitásorientált szentimentelemző kutatásunk eredményeit. Ezután bemutatjuk a szentimentelemző alkalmazásunkat és annak teljesítményét. Az alkalmazásunk egy pipeline-t valósít meg, amely egy bemeneti szöveget elemez. Felismeri a benne lévő személyneveket, majd a megtalált személynevekre megállapítja a szentimentet az adott környezetben. Továbbá az alkalmazás egy mondatszintű szentimentet is visszaad eredményként. A legjobban teljesítő modelljeink elérhetőek a Hugging Face oldalunkon[1]: NYTK/sentiment-ohb3-hubert-hungarian, NYTK/sentiment-ohb3-xlm-roberta-hungarian.

2. Kapcsolódó irodalom

A szentiment analízis egy komplex nyelvtechnológiai feladat és számos területen alkalmazható. Többek között szociális média monitorozására és elemzésére (Neri és mtsai, 2012), üzleti döntéshozatalok támogatására, gazdasági hírek szemantikus kontextusának elemzésére (Lutz és mtsai, 2018; Saura és mtsai, 2019) vagy pszichés állapot vizsgálatára (Jo és mtsai, 2018). Jelenleg többféle irányban zajlanak a szentiment analízis kutatások. Kezdetben a szentimentelemzés során osztályozni lehetett dokumentumokat, mondatokat és a szövegeket negatív, pozitív vagy semleges osztályba sorolták. Ezek a kutatások mind a mai napig aktuálisak (Pang és mtsai, 2002; Feldman, 2013; Lutz és mtsai, 2018). Később egy új irány alakult, az aspektus-alapú, vagy entitásorientált módszer, amely kifinomultabb és célja, hogy azonosítsa egy tárgy, személy vagy bármilyen entitás olyan aspektusait, melyek az érzelmek kiváltásáért felelősek (Pontiki és mtsai, 2014), illetve a vizsgált entitásról megállapítsa a szentimentjét az adott környezetben.

Magyar nyelvre kevés szentimentelemző tanítóanyag található. A HuSent (Szabó és mtsai, 2016) egy manuálisan annotált szentiment korpusz, amely a Dívány honlapról vett magyar nyelvű véleményeket tartalmaz a különböző termékekkel kapcsolatban. A korpusz 154 véleményt tartalmaz, mintegy 17 000 mondatból és 251 000 tokenből áll. Entitásorientált szentiment analízis korpusz az OpinHuBank (Miháltz, 2013). Ez egy manuálisan annotált korpusz, amely a véleménykutatást támogatja. 10 000 mondatból áll és személyneveket tartalmaz a

[1] https://huggingface.co/NYTK

főbb magyarországi honlapokról és blogokról. Minden egyes szegmenst 5 annotátor elemzett, a szentiment polaritásra nézve lehet semleges, pozitív vagy negatív. (Steinberger és mtsai, 2011) kutatásai során előállított egy aspektus-alapú szentiment korpuszt többnyelvű párhuzamos korpuszokkal, amely tartalmaz egy magyar nyelvű alkorpuszt is. (Laki és Yang, 2021) kutatásukban mondatszintű neurális szentimentelemző modellekkel kísérleteztek. Különböző magyar és többnyelvű modelleket finomhangoltak. A modellek teljesítményét különböző adatkiterjesztési eljárásokkal, transzfer és gépi fordítás módszerekkel növelték.

Jelen kutatás keretében az entitásorientált szentimentelemzésre fókuszáltunk, valamint a szentimentelemző alkalmazás pipeline-t mutatjuk be.

3. Tanítóanyag

Az entitásorientált szentimentelemző modell tanításához a szabadon elérhető magyar nyelvű OpinHuBank (Miháltz, 2013) (OHB3) korpuszt használtuk fel. A korpusz címkéi: -1 (negatív), 0 (semleges), 1 (pozitív). Minden szegmenst öt annotátor annotált. A feladatunkhoz összevontuk és átkonvertáltuk a címkéket: 1 (negatív), 2 (semleges), 3 (pozitív). Így minden szegmenshez pontosan egy darab címke tartozik. Továbbá végeztünk transzfer és fordítással történő adatkiterjesztési kísérleteket is. Ehhez a feladathoz az acl-14-short-data (Dong és mtsai, 2014) (ACL14) angol nyelvű korpuszt használtuk fel. A fordításhoz a Laki és Yang (2022) által közzétett gépi fordító rendszert használtuk fel.

A 1. táblázatban láthatóak a korpuszok legfőbb tulajdonságai. A duplavonal felett szereplenek a fontosabb statisztikai információk, a duplavonal alatt találhatóak a tanító- és tesztanyagra vonatkozó főbb tulajdonságok. Az OHB3 tanító- és tesztanyag előállításához 90%-10% vágást végeztünk. A korpusz első 1000 szegmense alkotja a tesztanyagunkat. A 1. táblázatban továbbá láthatóak a címkék eloszlása. Jól megfigyelhető, hogy nem kiegyensúlyozott a korpusz, ezért a kutatásunk során végeztünk adatkiegyenlítési kísérleteket is.

Végül, de nem utolsó sorban megmértük az annotátorok egyetértését is, amit a 2. táblázat mutat.

A szentimentelemző modell tanításához különböző transzformer-alapú neurális nyelvi modelleket finomhangoltunk. A transzformer modellek előnye, hogy képesek a bemeneti szövegben lévő viszonyokat megtanulni. Ez az entitásorientált tanítóanyag létrehozását is megkönnyíti, hiszen csak meg kell jelölni, vagy ki kell emelni az elemezni kívánt szövegrészt. A kiemelés úgy történt a mi esetünkben, hogy a vizsgált személynevet a szegmens elejére szúrtuk be, majd egy szeparátor címkével választottuk el a szegmenstől. A különböző modell típusok különböző szeparátort használnak. A modelljeink esetében a tanítóanyag formátumát a következő kettő példa demonstrálja:

- BERT: Tamás [SEP] Milan és Tamás is jól beilleszkedett már.
- RoBERTa: Tamás </s> Milan és Tamás is jól beilleszkedett már.

		ACL14	OHB3	ACL14 fordított
Szegmens		6 940	10 005	6 940
Token		150 904	308 407	146 717
Type		16 405	47 492	23 422
Átlagos szószám		17,64	26,36	17,01
Címkék		1;2;3	1;2;3	1;2;3
Tanító	szegmens	2 468	9 005	2 468
	1 (címke)	1 560	2 253	1 560
	2 (címke)	3 127	5 205	3 127
	3 (címke)	1 561	1 548	1 561
Teszt	szegmens	269	1 000	269
	1 (címke)	173	136	173
	2 (címke)	346	565	346
	3 (címke)	173	299	173

1. táblázat. Korpuszok tulajdonságai

	OpinHuBank
Fleiss Kappa	0,6551
Krippendorff-alfa	0,6548
Scott pi	0,6548
Cohen-féle kappa	0,6549

2. táblázat. Annotátorok közötti egyetértés

4. Kísérletek

A kutatásunkban először finomhangoltunk három magyar és kettő többnyelvű neurális nyelvi modellt az entitásorientált szentimentelemzés feladatára: huBERT (Márk, 2021), HILBERT (Feldmann és mtsai, 2021), HIL-RoBERTa (Yang és Váradi, 2021), XLM-RoBERTa base (XLM-R) (Conneau és mtsai, 2020), Multilingual Cased BERT-Base (mBERT) (Devlin és mtsai, 2019).

Ezután végeztünk adatkiegyenlítés és adatkiterjesztési kísérleteket a modelljeink minőségének javítására.

Az adatkiegyenlítés eljárás során kettő kísérletet végeztünk:

- **bővítés**: vettük a legtöbb szegmenssel bíró osztálycímkéhez tartozó szegmensek számát, ami az OHB3 esetén a 2-es címke 5 205 szegmensszámmal. Majd a többi osztálycímkéhez tartozó szegmensek számát bővítettük addig, amíg el nem érte ezt az 5 205 számot.
- **elvonás**: vettük a legkevesebb szegmenssel bíró osztálycímkéhez tartozó szegmensek számát, ami az OHB3 esetén a 3-as címke 1 548 szegmensszámmal. Majd a többi osztálycímkéhez tartozó szegmensek számát csökkentettük addig, amíg el nem érte ezt az 1 548 számot.

Az adatkiterjesztéses kísérleteink során a Laki és Yang (2021) által végzett eljárásokat végeztük el. Így a kutatásunk során összesen kilenc különböző mérést végeztünk el:

- **eredeti**: minden előtanított nyelvi modell finomhangolásra került az eredeti OHB3 korpuszon. Ezt tekintjük az alapeljárásnak.
- **bővítés**: adatkiegyenlítő eljárás során történt adatbővítéssel kiegyenlített OHB3 korpuszon finomhangoltunk minden előtanított modellt.
- **elvonás**: adatkiegyenlítő eljárás során történt adatelvonással kiegyenlített OHB3 korpuszon finomhangoltunk minden előtanított modellt.
- **zeroshot**: többnyelvű modellek képesek magyar nyelvű NLP feladatok predikciójára. Az angol ACL14 korpuszt használtuk finomhangolásra és a rendszernek magyar nyelvű mondatokra kellet prediktálnia.
- **transzfer**: többnyelvű modellek finomhangolva lettek az ACL14 korpuszon, majd további finomhangolásra kerültek az OHB3 tanító korpuszon.
- **fordított**: minden előtanított modell finomhangolva lett a magyarra fordított ACL14 korpuszon (ACL14_hu), majd közvetlenül tesztelve az OHB3 tesztkorpuszon.
- **fordított+finomhangolás**: minden előtanított modell finomhangolva lett az ACL14_hu korpuszon, majd ismét finomhangolva lett az OHB3 tanító korpuszon.
- **mix**: minden előtanított modell finomhangolva lett a konkatenált és összekevert ACL14_hu és OHB3 tanító korpuszon, majd le lett tesztelve az OHB3 tesztelő korpuszon.
- **mix+finomhangolás**: minden előtanított modell finomhangolva lett a konkatenált és összekevert ACL14_hu és OHB3 tanító korpuszon, majd újabb finomhangolás következett az OHB3 tanító korpuszon

Az elvégzett kísérletek eredményeit az OHB3 tesztelő korpuszán értékeltük ki. A technikai feltételek és beállítások megegyeznek a Laki és Yang (2021) által közzétett cikkben leírtakkal.

5. Eredmények

A 3. táblázatban láthatóak a méréseink eredményei. Első megfigyelés, hogy az adatkiegyenlítési kísérletek nem jártak sikerrel. Az elvonás eljárás sikertelensége várható volt, hiszen drasztikusan lecsökkent vele a tanítóanyag mérete.

Az adatkiterjesztéses mérések eredményei hasonlóak a (Laki és Yang, 2021) kutatásaiban bemutatott mondatalapú szentimentelemzés kísérleteihez. A fordítással hozzáadott adat képes növelni a modell minőségén. Továbbá az látható még, hogy ez a feladat sokkal jobban megosztotta a modelleket, ami azt jelenti, hogy ez a feladat már nehezebbnek bizonyult. A HILBERT eredményeit nézve gyakorlatilag nem tudta megtanulni ezt a feladatot. A huBERT várhatóan kimagaslóan teljesít ebben a feladatban is. Ami még érdekes, hogy a többnyelvű XLM-R modell is igen magas eredményt ért el.

A 4. táblázatban látható három példa a legjobban teljesítő huBERT mix modelltől. Látható, hogy egy mondaton belül, a különböző személynevekhez megfelelő szentimentet prediktált. Külön kiemelendő, hogy a harmadik példában Kovácsné Erzsébet helyesen semleges szentiment értéket kapott, hiszen ő ebben az esetben csak mesélő.

	huBERT	HILBERT	HIL-RoBERTa	XLM-R	mBERT
eredeti	81,99	57,80	68,50	79,69	75,49
bővítés	81,00	57,40	68,00	78,50	75,19
elvonás	77,99	56,99	66,29	75,40	74,09
zeroshot	-	-	-	66,79	61,19
transzfer	-	-	-	79,79	75,70
fordított	63,70	52,10	54,90	63,89	41,60
fordított+finonhangolás	81,69	54,40	69,49	**80,59**	61,40
mix	**82,30**	51,49	**69,80**	79,69	75,49
mix+finonhangolás	81,49	**58,60**	66,69	77,70	**76,39**

3. táblázat. Entitásorientált szentimentelemzés eredménye

Vizsgált személynév	Mondat	Prediktált szentiment
Kovácsné Erzsébet	A Kovácsné Erzsébet nagyon jól érzi magát a Nokiánál, azonban a Németországból érkezett Kovács Péter nehezen boldogul a beilleszkedéssel.	pozitív
Kovács Péter	A Kovácsné Erzsébet nagyon jól érzi magát a Nokiánál, azonban a Németországból érkezett Kovács Péter nehezen boldogul a beilleszkedéssel.	negatív
Kovácsné Erzsébet	A Kovácsné Erzsébet azt mondta, hogy a Németországból érkezett Kovács Péter nehezen boldogul a beilleszkedéssel.	semleges

4. táblázat. Példák a huBERT mix modelltől

A végső alkalmazásunkban a legjobb eredményt elért finomhangolt huBERT mix modellt használtuk fel.

6. Névelem kinyerése

Az entitásorientált szentimentelemzés valós időben való alkalmazásához szükség van egy névelemfelismerő modellre (NER), amellyel a személyneveket kinyerjük a szövegből. A magyar nyelvre elérhető neurális névelemfelismerő modellek korlátolt számossága miatt, a feladathoz a (Yang és Váradi, 2021) által közzétett NER modellt használtuk fel. A használt NER modell az NYTK-NerKor korpuszon (Simon és Vadász, 2021) tanított, amely 90,18% F1 mértékkel tudja felismerni a névelemeket. Az entitásorientált elemző alkalmazásunkhoz csak a személyneveket (PER címkével ellátott) használtuk fel.

A névelemfelismerés pontosságától erősen függ az alkalmazás pontossága. A NER modellek minőségének hatásait mértük a következő kísérletünkben. A kísérlethez vettük a NER feladatra finomhangolt HILBERT és HIL-RoBERTa modelleket (Yang és Váradi, 2021). A 5. táblázatban mutatjuk a kiértékelés arról, hogy a NER modellek hogyan teljesítenek az összes címkére nézve (NerKor-MIND) és csak a személynevekre nézve (NerKor-PER). Továbbá láthatóak a modellek teljesítményei az OHB3 korpuszban lévő személynevek felismerésére.

Az OHB3 korpuszban csak a személynevek vannak megjelölve, ezért a vizsgálatunkban is csak a személynevekre koncentráltunk, vagyis a szervezet (ORG), helynevek (LOC) és a vegyes (MISC) címkékkel ellátott tulajdonneveket nem vettük figyelembe. Az eredményekből tisztán látszik, hogy a NerKor tesztanyagán mért 5% különbség az OHB3 korpuszban 20-40% teljesítményeltérést is mutathat. Ez azt is mutatja, hogy például a HILBERT-en finomhangolt modell a NerKor korpuszon túltanult és egy független korpuszon rendkívül gyengén teljesített.

	NerKor-MIND (F1)	NerKor-PER (F1)	OHB3 (Pontosság)
huBERT	90,18	96,25	87,4
HILBERT	84,81	92,41	46,0
HIL-RoBERTa	85,16	93,18	69,2

5. táblázat. NER modellek kiértékelése

Ezután a következő kísérletünkben a legjobban teljesítő finomgangolt huBERT NER modellt teszteltük a különböző adatkiterjesztéses kísérleteinkben. A 6. táblázatban láthatóak azok az eredmények (lásd második oszlop), ha a tesztanyagban a személyneveket a huBERT NER modellel nyernénk ki és adnánk oda a szentimentelemző modellnek. Az eredményekben meg is látszik a különbség, hiszen megmutatkozik a NER modell hibája, de a hiba mértéke a legtöbb esetben nem szignifikáns.

	Referncia entitás	huBERT NER
huBERT eredeti	81,99	81,49
huBERT fordított	63,70	51,89
huBERT fordított+finomhangolás	81,69	81,30
huBERT mix	82,30	82,20
huBERT mix+finomhangolás	81,49	81,30

6. táblázat. huBERT NER eredmények

Végül megvizsgáltuk a hibásan felismert személynevek hibatípusait (lásd 7. táblázat). A példákból látható, hogy a hibák nagyobb része referenciahiba. A két nagyobb hibakategória ezekből a hiányzó előtag és a helytelen kinyerés. Ebből látszik, hogy automatikus módszerekkel nyerték ki a tulajdonneveket. A modell hibái közül nagy része a hibásan kinyert személynevek, de találunk tokenizálási hibát is.

7. Demó applikáció

A demó applikáció egy pipelinet valósít meg, amelyben van egy NER modell és egy szentimentelemző modell. A pipeline egy mondatot vár, amelyet átad a

Hiba típusa	**Arány (darab)**	**Példa** (név a referenciában — kinyert név)
Referencia hiba	**70,63% (89)**	
Hiányzó előtag	34,83%	Johann Strauss — ifj. Johann Strauss Litavecz Anna — Dr. Litavecz Anna János Pál — II. János Pál
Nem személynév	10,12%	Szent István — Szent István templomban Puskás Ferenc — Puskás Ferenc stadion
Lemmatizált	13,48%	Magyar Kornélia — Magyar Kornéliát Dombóvári István — Dombóvári Istvánnal
Helytelen kinyerés	41,57%	Varga Miklós — Varga Miklósné Joe Bi — Joe Biden Kristin Krohn — Kristin Krohn Devold
Modell hiba	**29,37% (37)**	
Tokenizálás	24,32%	Széles-Kovács Gyula — Széles - Kovács Gyula
Sztemmelt	29,73%	Bogár Zsolttal — Bogár Zsolt Jakab Jánosék — Jakab János
Helytelen kinyerés	45,95%	Varga Miklós — Varga Miklós Európ Bándi Gyula — Bándi Gyul Boross József — ∅

7. táblázat. huBERT NER modell hibaelemzése az OHB3 tesztanyagon

NER modell számára. A NER modell a szövegben felismeri a tulajdonneveket, amelyeket meg is jelöl. Ezután a felismert névelemek közül a személyneveket (PER címkével ellátott) továbbadja a szentimentelemző modell számára, aki a névelem és a bemeneti mondat segítségével prediktálja a személynévhez tartozó szentimentet.

Az 1. ábrán látható a demó alkalmazás felülete. Egy szövegbeviteli mezőben meg lehet adni egy mondatot, majd az Elemzés gombbal az alkalmazás leelemzi a mondatot. A szövegbeviteli mező alatt jelenik meg az elemzett szöveg. A huBERT NER modellje által felismert tulajdonneveket a program különböző színekkel megjelöli: zöld: személynév; kék: szervezet; narancssárga: helynév; lila: vegyes. A NER modell képes nem csak személyneveket felismerni, azonban a jelenlegi entitásorientált szentimentelemző modellünk csak személynevekkel dolgozik.

Az elemzett szöveg alatt látható a mondatszintű szentimentelemzés eredménye a modell által prediktált valószínűséggel. A mondatszintű elemzéshez a Laki és Yang (2021) által tanított huBERT mondatalapú szentimentelemző modellt használtuk. Ezután alatta következik egy felsorolás, ahol a szövegben felismert személyneveket jeleníti meg a hozzátartozó szentimentelemzéssel és a modell által prediktált valószínüséggel. A szentimentelemzéseket különböző színű háttérrel is jelöltük: zöld: pozitív; bézs: semleges; bordó: negatív.

1. ábra: Demó alkalmazás

8. Összegzés

Kutatásunkban különböző magyar és többnyelvű neurális nyelvmodellt finomhangoltunk entitásorientált szentimentelemzésre. A modellek teljesítményét különböző adatkiterjesztési, transzfer és gépi fordítás módszerekkel növeltük. Eredményünkben bemutattuk, hogy adatkiterjesztési módszereinkkel valóban tudtuk növelni a modellek minőségét. Továbbá bemutattunk egy alkalmazást, amely képes szentimentelemezni egy bemeneti szöveget. Az alkalmazás először egy névelemfelismerő modell segítségével kiválasztja a személyneveket, majd az entitásorientált szentimentelemző modell a kiválasztott személynevekre megvizsgálja az érzelmi polaritást az adott környezetben. Végül a rendszer ad egy mondatszintű szentiment értéket is. Az alkalmazáshoz készítettünk egy demó felületet is. Végül, de nem utolsó sorban kiértékeltük az entitásorientált szentimentelemzőnk teljesítményét a névelemfelismerő modell függvényében is.

Hivatkozások

Baranyi, P., Csapó, Á.: Definition and synergies of cognitive infocommunications. Acta Polytechnica Hungarica 9(1), 67–83 (2012)

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. In: Proceedings of the 58th Annual

Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 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)

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., Xu, K.: Adaptive recursive neural network for target-dependent twitter sentiment classification. In: The 52nd Annual Meeting of the Association for Computational Linguistics (ACL). ACL (2014)

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)

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)

Laki, L., Yang, Z.G.: Improving performance of sentence-level sentiment analysis with data augmentation methods. In: Proceedings of 12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021). pp. 417–422. IEEE, Online (2021)

Laki, L., Yang, Z.G.: Jobban fordítunk magyarra, mint a Google! In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 357–372. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2022)

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

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

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)

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

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)

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

Simon, E., Vadász, N.: Introducing nytk-nerkor, A gold standard hungarian named entity annotated corpus. In: Ekstein, K., Pártl, F., Konopík, M. (szerk.) Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12848, pp. 222–234. Springer (2021)

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)

Yang, Z.G., Váradi, T.: Training language models with low resources: Roberta, bart and electra experimental models for hungarian. In: Proceedings of 12th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2021). pp. 279–285. IEEE, Online (2021)