Inzulinrezisztencia betegség jelenségének felismerése és osztályozása orvosi dokumentumokban

Yang Zijian Győző

Sightspot Network Kft.
4034 Debrecen, Vágóhíd u. 2.
Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
yang.zijian.gyozo@nytud.hu

Kivonat A jelen cikkben egy kutatás-fejlesztés projekt első fázisának részleteit mutatjuk be, amelynek keretében az inzulinrezisztencia betegség kialakulásának veszélyét szeretnénk előre jelezni a nyelvtechnológia eszközeivel. A kutatásunk kétmillió magyar nyelvű kórházi kórlap feldolgozásával történt a modern neurális nyelvtechnológia segítségével. A feladatot osztályozási feladatként értelmeztük, amelyben három különböző esetet különböztettünk meg: inzulinrezisztenciás betegek, nem inzulinrezisztenciás páciensek és gyanús esetek. A gyanús esetek közé azokat a pácienseket soroltuk, akik a kórlapjuk alapján nem inzulinrezisztenciások, de közben tudjuk, hogy azok. A feladat nehézsége, hogy a programunknak fel kell ismernie a gyanús eseteket úgy, hogy a kórlapon nem szerepel az inzulinrezisztencia betegség. A probléma ily módon egy háromosztályos klasszifikáció feladatként oldható meg. A kórlapok zajossága és félig strukturáltsága miatt, rendkívül nehéz belőle egységes releváns tulajdonság jegyeket kinyerni, ezért a probléma megoldására egyedül a modern nyelvi modellek jöhettek csak számításba, amelyek automatikusan nyerik ki a számukra relevánsnak számító nyelvi jegyeket. A kutatásunkban felhasználtunk egy statikus és egy környezetfüggő neurális nyelvi modellt. Az eredményeink alapján, a modelljeink közel 80%-os pontossággal tudta megbecsülni, hogy az adott kórlap a fent említett három kategóriából melyikbe tartozott. Az általunk létrehozott osztályozási modellekkel orvosi támogatást tudunk nyújtani, amelynek során a gép jelezni tudja azon eseteket, ahol, bár a beteg másféle kivizsgáláson vesz részt, a kórlap alapján az adott páciensnél felmerülhet az inzulinrezisztencia betegségének veszélye.
Kulcsszavak: inzulinrezisztencia, neurális szövegosztályozás, fastText, huBERT

1. Bevezetés

Az inzulinrezisztencia korunk egyik jelentős, sok embert érintő, de annál nehezebben előre jelezhető betegsége. A kórházba látogató páciensek kórlapjai az évek során temérdek mennyiségben halmozódtak fel. Ezek feldolgozása nem kis

feladat. A digitalizálás fejlődésével az egészségügyi intézetek elkezdték adatbázisokba menteni a dokumentumaikat. Ez a folyamat lehetőséget ad arra, hogy automatikus módszerekkel elemezzük őket. Azonban jelentős probléma, hogy ezek a dokumentumok kevésbé strukturáltak. Ez nagyban megnehezíti a szövegfeldolgozást. Továbbá az a tény is nehezíti a feladatot, hogy a különböző páciensek különböző betegségei is szerepelnek a kórlapokon. Ez megnehezíti az egységes feldolgozást. A nyelvtechnológia fejlődésével, különösen a neurális nyelvmodellek megjelenésével áthidalhatóak ezek a problémák. Az új generációs transzformer alapú nyelvmodellek képesek egy adott szövegben lévő összefüggéseket felismerni, független azok strukturális felépítésétől. Ezeknek a modelleknek a finomhangolásával képesek vagyunk a nyers strukturálatlan szöveget elemezni és számunkra hasznos információkat kinyerni.

Kutatásunkban kétféle neurális modellel végeztünk kísérletet. Egy statikus szóbeágyazással tanított fastText modellel, illetve egy kontextuális finomhangolt magyar nyelvű BERT modellel.

2. Kapcsolódó irodalom

Az utóbbi években a nyelvtechnológia fejlődésével számos kutatás célozta meg az orvosok támogatását a nyelvtechnológia segítségével. Chen és mtsai (2017) kutatásukban konvolúciós hálózattal prediktáltak betegségeket. Nyelvi osztályozó modellük tanításához 2013-2015 között gyűjtött kórházi elektronikus egészségügyi dokumentációkat használtak. Nori és mtsai (2015) a betegségek közötti hasonlóságot vetették össze az elektronikus egészségügyi dokumentumok közötti hasonlósággal, ezzel betegség specifikus összefüggéseket tudtak beépíteni a halálozási modellezésbe, amellyel pontosabb prediktív modelleket tudtak létrehozni. Szintén konvolúciós hálózatot használt Yao és mtsai (2018), akik klinikai szövegeket osztályoztak szabályalapú jegyek hozzáadásával. Geraci és mtsai (2017) rekurrens hálózat segítségével kerestek alkalmas jelölteket a fiatalkori depresszió kutatásához, ehhez strukturálatlan szövegekből azonosították a fiatalkori depresszió tüneteit.

Magyar nyelvre Siklósi és Novák (2014); Orosz és Prószéky (2014) végeztek kutatásokat klinikai szövegek normalizálására. Papp és mtsai (2014) az Alzheimer-kórban szenvedő páciensek beszédeit elemezték, hogy a korai Alzheimer-kórra jellemző nyelvi tüneteket detektálják. Bagi és mtsai (2019) kísérleteiben a szkizofréniát azonosították spontán beszéd temporális paraméterei alapján. Kicsi és mtsai (2020) kutatásukban radiológiai leletek szövegében azonosítottak testrészeket, elváltozásokat és azok kapcsolatait. Jenei és Kiss (2020) a depressziós állapot automatikus detektálását tűzték ki célul konvolúciós neurális hálózatok segítségével. Vetráb és mtsai (2022) szekvenciális autoenkódert használtak enyhe kognitív zavar automatikus felismerésére.

3. Osztályozó modelleink

A kutatásunkban felhasználtunk egy statikus és egy környezetfüggő, vagy másnéven kontextuális neurális nyelvi modellt.

A **fastText** (Joulin és mtsai, 2017, 2016) a Meta Research (korábban Facebook Research) csapatának fejlesztése, melynek célja a szóreprezentációs és a szövegosztályozó modellek effektív tanítása. A módszerrel szóalapú és n-gramm karaktereken alapúló 'skip-gram' és 'cbow 'modellt lehet tanítani. A rendszer legnagyobb előnye, hogy C++ nyelven implementált, ezért gyors és hatékony megoldást kínál anélkül, hogy előfeldolgozásra vagy felügyeletre lenne szükség (Bojanowski és mtsai, 2017). Nincsen szüksége videókártyára sem a tanításhoz. Szövegosztályozás szempontjából más mély tanulás alapú megoldásokkal összevethető a teljesítménye, és egy lényegesen gyorsabb megoldás tanítás és kiértékelés szempontjából (Joulin és mtsai, 2017). A platformon előre tanított szóvektorok érhetőek el 158 különböző nyelvre, ezáltal egy nagyon kézenfekvő és lehetőségekkel teli eszköznek számít a többnyelvű nyelvfeldolgozás terén is.

A **BERT** (Devlin és mtsai, 2019) (Bidirectional Encoder Representations from Transformer) egy kétirányú transzformer enkóder (Vaswani és mtsai, 2017). A BERT modellt két nyelvmodellezési feladaton tanították elő: szómaszkolás és következő mondat perdikciója. A szómaszkolás során a tanításhoz használt korpuszban a szavak 15%-a véletlenszerűen maszkolásra kerül, a rendszernek pedig ki kell találnia a kimaszkolt szavakat. A következő mondat predikciója során pedig a feladat annak kitalálása, hogy két kiválasztott mondat a szövegben egymást követő mondatok-e vagy csak két véletlenül kiválasztott mondat. A szótár méretének csökkentése érdekében a BERT modell szóelem (word pieces) tokenizáló algoritmust (Schuster és Nakajima, 2012) használ. A BERT előtanítása során általános nyelvi tudásra tesz szert, ezért is hívjuk nyelvmodellnek, majd ezt követően, a modellt finomhangolással további specifikus feladatokra tanítható. A kutatásunkhoz a magyar nyelvű **huBERT** (Nemeskey, 2021) modellt használtuk, amely a 9 milliárd szavas Webkorpusz 2.0 korpuszon (Nemeskey, 2020) lett előtanítva. A huBERT jelenleg a legjobban teljesítő magyar nyelvű BERT modell.

4. A Korpusz

Kutatásunkban összesen 2 000 000 orvosi dokumentummal dolgoztunk, amelyekből 1 000 000 dokumentum az inzulinrezisztencia kategóriába tartozott és 1 000 000 a nem inzulinrezisztencia kategóriába. Az orvosi dokumentumok a következő csoportba tartoznak: kórlap, zárójelentés, xamba, ambuláns lap, egyéb rövid jelentések. A 1. ábrán mutatunk egy példát egy nyers ambuláns lapról, ami az adatbázisban található. Látható, hogy a dokumentum félig strukturált, amelyben a strukturált rész inkább személyes adatokból áll. Továbbá az is látható, hogy az ambuláns lap szabad szöveges részében találhatóak a fontosabb információk a páciensről. Végül az is megfigyelhető, hogy a szöveg anonimizált.

```
VESZ Egészségügyi Szolgáltató Kht.
Gyermek Rehabilitációs Központ
#CIM                    .
Telefon:#T
0921P7101

                                AMBULÁNS LAP

Név:    #AEXTRA
Leánykori név:
TAJ:    #AEXTRA
Szül.dátum: #AEXTRA
Születési hely:     #AEXTRA
Anyja neve: #AEXTRA
Lakcím: #AEXTRA
A megjelenés ideje: 2007#      09:32
Naplószám:      02000449

PSZICHOLÓGIAI VÉLEMÉNY

Előzmények

    Az elmúlt tanévben havi 1 alkalommal vett részt nálam a kisfiú játékterápiás foglalkozáson.
    2007. júliusban kizárták izombetegség meglétét a #NEV, mely miatt születése óta (az anya
    betegsége miatt) rendszeres orvosi kontrollokra járt ill. gyógytornán is részt vett.

Vizsgálatok

    Viselkedésmegfigyelés és kontroll tesztvizsgálat történt (B. Binet, figyelemvizsgálat).

    Eredmény: #NEV viselkedése különösen a 2006. októberi státuszhoz képest sokat változott.
    Csapongása sokat mérséklődött, figyelme, kitartása életkorának megfelelő szintű. B. Binet
    tesztben teljesítménye életkorát 3 hónappal meghaladja, IQ=105. Általános ismeretei
    korának megfelelőek, az analógiás gondolkodás alakulóban, rövid távú verbális memóriája 3
    elemű, konstrukciós készségei, téri-vizuális észlelése jó. Enyhe grafomotoros ügyetlenség
    jellemzi, hangzóhibái még vannak.
    Szomatizációs problémái mérséklődtek; viselkedése érettebb.

Vélemény, javaslat

    A grafomotoros készségek javítását próbálják otthon formamásolással, közös rajzolással
    fejleszteni; a hangzóhibákat óvodai logopédia keretében lehetne korrigálni. Javaslom a
    Színezd ki és rajzolj te is...; és Színezd ki és számolj te is .. feladatlapok otthoni
    használatát.
    Kontroll: 1 év múlva, friss háziorvosi beutalóval.

Dátum:  Debrecen,  2007#      09:32

#NEV             gy.klin. mentalhig.szakpszich.

#NEV           !

Szíveskedjen minden további jelentkezéskor korábbi orvosi dokumentációját magával hozni!
Köszönjük, hogy megtisztelt bizalmával!
```

1. ábra: Példa egy nyers ambuláns lapról

A 2. ábra egy rövid dokumentumot mutat be. Rövid felsorolás jellegű, rövidítésekkel és kulcsszavakkal. A két példából az figyelhető meg, hogy nagy a különbség a dokumentumok között, ezért nehéz egységesen feldolgozni őket.

```
Víztiszta cervixnyák. Váladékvizsgálat: I. tiszt.fok.
Mikrobiológiai mintavétel. Gombavizsgálat: negatív.
TVS: Uterus afv-ben. Kismedencében kóros nem ábr.
```

2. ábra: Példa egy rövid orvosi dokumentumról

Az adatbázisban az orvosi dokumentumok egy része további meta adatokat tartalmaztak, amelyekből az egyik információ a betegségek nemzetközi osztályozására használt azonosító, az úgynevezett BNO kód. Az inzulinrezisztencia betegség kódjai 'E10' karaktersorozattal kezdődnek és 0-9 számokkal folytatódnak: E100, E101, E102, E103, E104, E105, E106, E107, E108, E109. Maga a kórházi adatbázishoz sajnos nem volt hozzáférésünk, mi már szűrt adatokat kaptunk kézhez. Tudomásunk szerint a szűrés a fent említett BNO kódok alapján történt. Az első egymillió kórlap tartalmazza valamelyik inzulinrezisztenciával kapcsolatos BNO kódot. A második egymillió kórlap a fent említett inzulinrezisztenciával kapcsolatos BNO kódokat nem tartalmazza. Miután a szűrés megtörtént, adatbiztonsági okokból a személyes adatokat törölték a kórlapokból. A törlés szabály alapon történt, a személyes adatok megjelenése a dokumentumokban valamennyire szabályosságot mutatott, például a nevek előtt megjelent a 'név:' vagy a születési dátumok előtt a 'szül. dátum' kifejezések. A személyes adatok törlésében szintén nem vettünk részt és nem is kaptunk részletes információt arról, hogy pontosan milyen adatokat töröltek ezért csak a kapott szövegek alapján tudtunk következtetéseket levonni. A kapott szövegek elemzése alapján a következő adatokat törölték:

- Beteg személyes adatai: név, születési hely és dátum, születési név, anyja neve, lakcím, TAJ szám, munkahely, halál dátuma
- Kezelő orvos személyes adatai: kezelőorvos neve, vezető főorvos neve
- Általános információk (beteg, orvos vagy intézmény): telefonszám, fax

A szabály nem terjed ki minden esetre, ha egy adat nem illeszkedett pontosan a megadott szabályra, akkor az a személyes adat benne maradt. Ilyen például, amikor a dátumba '#' jel került: 2018#, vagy amikor szabad megfogalmazásban jelent meg a személyes adat.

Az így kapott adatokat további szűrések alá vetettük, végeztünk rajta tokenizálást a huSpacy (Orosz és mtsai, 2022) eszközzel és a szövegből szabály alapján kitöröltünk minden inzulinnal kapcsolatos BNO kódot: E10, E100, E101, E102, E103, E104, E105, E106, E107, E108, E109, E11, E110, E111, E112, E113, E114, E115, E116, E117, E118, E119.

Így első feladatként a kétmillió kórlapot konkatenáltuk és létrehoztunk egy bináris osztályozó modellhez egy korpuszt. A korpuszt véletlenszerűen megkevertük és kivettünk belőle 10%-ot tesztelő anyagnak. Így létrejött egy 90%-10% vágás a tanító és a teszt halmaz létrehozásához. A korpusz részletesebb tulajdonságai az 1. táblázatban láthatóak.

Következő feladatként a kapott adathalmazból kiválogattuk azokat a kórlapokat, amelyek bár nem lettek ellátva inzulinrezisztencia BNO kóddal, mégis tudjuk a paciensről, hogy inzulin rezisztenciával rendelkezik. Ehhez a feladathoz kigyűjtöttük az összes kórlaphoz tartozó páciens azonosítót, majd megkerestük a metszetet. A metszet alkotja azokat a betegek azonosítóját, akiknek a kórlapjai szerepelnek mind az inzulinrezisztenciával rendelkező, mind a nem inzulinrezisztenciával rendelkező adathalmazban. Ezt a halmazt elneveztük 'GYANÚS' halmaznak. Az adathalmazok további kialakításában a metszet/GYANÚS halmaz (szűkössége miatt) mennyisége lett a mérvadó. Így összesen igazodva a GYANÚS

halmazhoz, kerekítve mind a három osztályhoz 400 000 kórlapot választottunk ki véletlenszerűen a tanítóanyaghoz és 15 000 kórlapot a tesztanyaghoz. Az 1 táblázatban találhatóak a korpusz részletes kvantitatív tulajdonságai.

	Szegmens	Token	Type	Átlagos tokenszám / kórlap	Átlagos mondatszám / kórlap
2-osztályos tanító	1 800 000	290 254 939	730 914	161,25	20,69
2-osztályos teszt	200 000	32 423 375	270 045	162,12	20,79
3-osztályos tanító	1 200 000	181 306 262	631 334	151,09	19,33
3-osztályos teszt	45 000	6 853 141	132 040	152,29	19,45

1. táblázat. Korpuszok tulajdonságai

Végül utolsó feladatként, mielőtt a modell tanítása történt, minimális tisztítást végeztünk az adatokon. A következő lépéseket végeztük el:

- Ismétlődő írásjelek normalizálása: A szövegben számtalan helyen díszítősor jellegű szövegrészek találhatóak, ilyen a sok pont, gondolatjel és kérdőjel egymás után. Ezekből egyet-egyet hagytunk meg.
- Néhány szó és rövidítés normalizálása: Az Ambuláns lap cím számtalan esetben a következőképpen szerepelt: A M B U L Á N S L A P. Ezt konvertáltuk "AMBULÁNS LAP" formába, illetve: 'k . m . n .', 'K f t .'.
- Személyes adatok: Mivel a személyes adatokat törölték, ezért töröltük a hozzátartozó helyőrzőket és megnevezéseket is: #NEV, #EMAIL, szül.ideje:, TAJ száma, #CIM, #AEXTRA, #DOC, #T.
- Whitespacek normalizálása: Ahol több whitespace állt egymás után, azokat lecseréltük egy szóközre.

5. Mérések és eredmények

A kutatásunk során felhasználtuk a fastText statikus és a BERT környezetfüggő neurális nyelvi modellt az osztályozó modellek tanítására. Mind a fastText keretrendszerrel, mind a BERT modellel két predikciós modellt tanítottunk: egy 2-osztályos és egy 3-osztályos modellt. A modellek tanításához a következő hiperparamétereket használtuk:

- fastText: wordNgrams: 3; tanulási ráta: 0,8; bucket: 200000; dim: 200; epoch: 10.
- huBERT: tanulási ráta: 5e-5; batch méret: 64; epoch: 10.

A BERT modell tanításához a Hugging Face által közzétett szövegosztályozó szkriptet[1] alkalmaztuk.

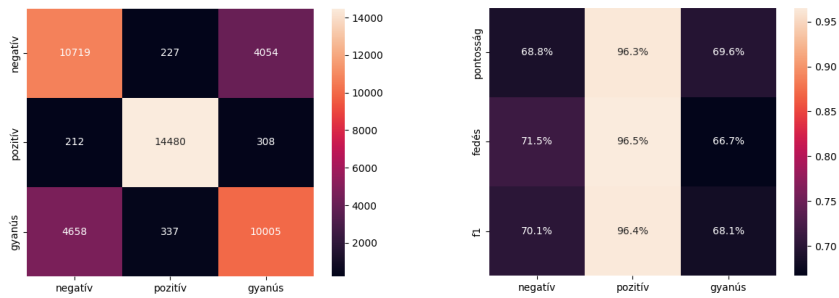
A 2. táblázatban láthatóak a modellek eredményei. A kiértékeléshez a pontosság (accuracy) metrikát használtuk.

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

	2-osztályos	3-osztályos
fastText	97,58%	78,23%
huBERT	**98,19%**	**81,34%**

2. táblázat. Az osztályozó modellek eredményei

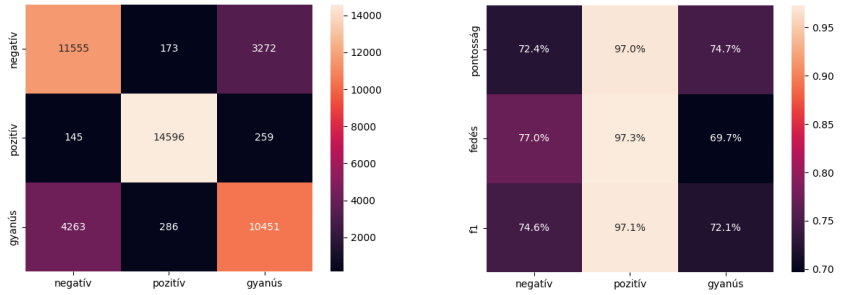
A 3. és a 4. képen láthatóak a 3-osztályos fastText és huBERT modellek teljesítményei címkékre bontva. A képek bal oldalán van a tévesztési mátrix, míg a jobb oldali ábrákon a pontosság, fedés és F1 mértékek találhatóak. A modellek a pozitív eseteket tudták könnyen megállapítani. Ez nem meglepő, hiszen a pozitív eseteknél csak a BNO kódokat töröltük a szövegből, de egy ilyen esetben többféle információ is utalhat magára a betegségre. Leggyengébben a gyanús esetek osztályozásánál teljesítettek a modellek, ez is várt eredmény, hiszen ezekben az esetekben kevés vagy semmilyen jel nem utal az inzulinrezisztenciára, azonban mégis azok lehetnek. A hőtérképek alapján a két modell rendkívül hasonlóan teljesített, ami a számokban is látszik. A huBERT fedése magasabb a negatív eseteknél, vagyis több tényleges negatív esetet tudott detektálni, ami a gyanús esetek predikciójának bizonytalanságát csökkenti. Az ábrán is látszik, hogy a gyanús eseteknél a pontosság magasabb, mint a fedés, ami a mi esetünkben azt jelenti, hogy amikor gyanús esetet mond a modell, azt magabiztosabban teszi.



3. ábra: A 3-osztályos fastText modell teljesítményének hőtérképei

6. Összegzés

Kutatásunkban az inzulinrezisztencia betegség kialakulásának kockázatát szeretnénk előre jelezni orvosi dokumentumok alapján. A feladat megoldásához az új neurális nyelvtechnológia eszközeit használtuk fel. Kísérleteinkben a szövegek osztályozásához a statikus fastText és a kontextuális huBERT modelleket tanítottuk be. A modellek tanításához kétmillió orvosi dokumentumot használtunk



4. ábra: A 3-osztályos huBERT modell teljesítményének hőtérképei

fel. Három esetre bontottuk a dokumentumokat: negatív, pozitív és gyanús. A gyanús esetek azok, akikről tudjuk, hogy kialakult nála az inzulinrezisztencia betegség, de az adott kórlapon ez nem volt feltüntetve, mivel lehet, hogy másféle vizsgálaton volt az illető. Eredményeink azt mutatják, hogy a huBERT modellel 81,34% pontossággal tudta megoldani ezt a 3-osztályos klasszifikációs feladatot. Rendelkezésünkre áll további hatmillió orvosi dokumentum, amelyeket következő lépésként szeretnénk feldolgozni.

Hivatkozások

Bagi, A., Gosztolya, G., Szalóki, S., Szendi, I., Ildikó, H.: Szkizofrénia azonosítása spontán beszéd temporális paraméterei alapján – egy pilot kutatás eredménye. In: XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 189–202. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2019)

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, M., Hao, Y., Hwang, K., Wang, L., Wang, L.: Disease prediction by machine learning over big data from healthcare communities. IEEE Access 5, 8869–8879 (2017)

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)

Geraci, J., Wilansky, P., de Luca, V., Roy, A., Kennedy, J.L., Strauss, J.: Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evidence-Based Mental Health 20(3), 83–87 (2017)

Jenei, A.Z., Kiss, G.: Depresszió detektálása korrelációs struktúrán alkalmazott konvolúciós hálók segítségével. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 59–71. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2020)

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

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)

Kicsi, A., Szabó, L.K., Pusztai, P., Németh, P., László, V.: Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 15–27. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2020)

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)

Nori, N., Kashima, H., Yamashita, K., Ikai, H., Imanaka, Y.: Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 855–864. KDD '15, Association for Computing Machinery, New York, NY, USA (2015)

Orosz, G., Prószéky, G.: Hol a határ? Mondatok, szavak, klinikák. In: X. Magyar Számítógépes Nyelvészeti Konferencia. pp. 177–187. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2014)

Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., Farkas, R.: Huspacy: an industrial-strength hungarian natural language processing toolkit. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 59–73. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2022)

Papp, P.A., Rácz, A., Vincze, V.: Automatikus morfológiai elemzés a korai Alzheimer-kór felismerésében. In: X. Magyar Számítógépes Nyelvészeti Konferencia. pp. 199–207. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2014)

Schuster, M., Nakajima, K.: Japanese and korean voice search. In: ICASSP. pp. 5149–5152. IEEE (2012)

Siklósi, B., Novák, A.: Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben. In: X. Magyar Számítógépes Nyelvészeti Konferencia. pp. 167–176. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2014)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett,

R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)

Vetráb, M., José, V.E.L., Balogh, R., Imre, N., Hoffmann, I., Tóth, L., Pákáski, M., Kálmán, J., Gosztolya, G.: Enyhe kognitív zavar automatikus felismerése szekvenciális autoenkóder használatával. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 175–184. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2022)

Yao, L., Mao, C., Luo, Y.: Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. In: 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W). pp. 70–71 (2018)