dataSTREAM 2023
Networking

Közép-európai Egyetem
Budapest
2023. május 18.

Statistical Products Hungary Kft.
ISBN 978-615-01-7510-2

A dataSTREAM 2023 konferencia szervezője:

Clementine/Statistical Products Hungary Kft.

1115 Budapest, Bartók Béla út 105-113. 1/b.

A konferencia házigazdája:

Körmendi György Olivér

Clementine

ügyvezető igazgató

Szerkesztette[1]:

Izsán Orsolya, Keresztesi Ildikó, Pancza Judit

[1]oktatas@clementine.hu

Tartalomjegyzék

Neurális hálózat alapú megoldások az iparban - a sikeres alkalmazások feltételei

Abonyi János

Pannon Egyetem, ELKH-PE Komplex rendszerek figyelemmel kísérése kutatócsoport, Veszprém, Magyarország, janos@abonyilab.com

Kulcsszavak: Gépi tanulási modellek életciklusa, Szoftver szenzorok, Hibadiagnosztika, Üzemeltetés

Tekintettel arra, hogy egyre növekvő figyelem irányul a gépi tanulás alkalmazásában rejlő lehetőségek feltárására, az előadás azokat az adat-alapú folyamattámogatási és hibadiagnosztikai megoldásokat tekinti át, amelyek automatikusan feltárják az összetett rendszerek belső összefüggéseit és alkalmasak a rendszer optimálására és hibadiagnosztikájára. Az előadás középpontjában a gépi tanulási megoldások költséghatékony bevezetése és karbantartása áll. A modellek teljes életciklusának szempontjából sikeres alkalmazások olyan öntanuló rendszerek kialakítását követelik meg, melyek a rendszer működéséből származó adatokra alapozva képesek az alkalmazott modellek és döntési algoritmusok folyamatos fejlesztésére, és ezáltal a rendszer teljesítményének javítására. Ezen öntanuló diagnosztikai megoldások kialakításának alapjaként az ML-Ops fejlesztői keretrendszer alkalmazását javasoljuk. Az ML-Ops, azaz a Machine Learning Operations, egy olyan keretrendszer, amely segíti a fejlesztőket a gépi tanulás alapú rendszerek fejlesztésében és fenntartásában.
Az ML-Ops keretrendszer fontosságára azért hívjuk fel a figyelmet, mert a koncepcióhoz kapcsolódó megoldások a fejlesztők és a felhasználók számára egyaránt lehetővé teszik, hogy könnyen monitorozzák és menedzseljék az ML-alapú rendszereket, valamint javítsák azok teljesítményét.
Az előadás ennek tükrében nem csupán bevezetést ad a neurális hálózatok alkalmazási lehetőségeibe, hanem a sikeres és fenntartható alkalmazások érdekében felhívja a figyelmet arra, hogy

- Gépi tanulási funkciók kialakításának jelentős az adatigénye, így az adatok gyűjtése, a kapcsolódó kísérletek tervezése kritikus fontosságú feladat.
- A hibadiagnosztikai feladatok általában úgynevezett unbalanced osztályozási problémák, mely a felügyelt tanuláshoz címkézett mintákat igényelnek. Az unbalanced osztályozási feladat kezelésére javasoljuk a SMOTE technika alkalmazását.
- Javasoljuk transzfer learning megoldások alkalmazását és ehhez más működő rendszerekből és szimulációs vizsgálatokból gyűjtött adatok alkalmazását.
- Javasoljuk célirányos változószelekciós és változó transzformációs megoldások fejlesztését.
- Javasoljuk az MLOps módszertan pontról pontra történő alkalmazását.

Neurális hálók alkalmazása az élelmiszervizsgálatokban

Sipos László[1,2], Szabó Dániel[1], Nyitrai Ákos[1]

[1]Magyar Agrár- és Élettudományi Egyetem, Élelmiszertudományi és Technológiai Intézet, Árukezelési, Kereskedelmi, Ellátási Lánc és Érzékszervi Minősítési Tanszék, Budapest, Magyarország, sipos.laszlo@uni-mate.hu
[2]Közgazdaság- és Regionális Tudományi Kutatóközpont, Közgazdaságtudományi Intézet, Budapest, Magyarország, sipos.laszlo@krtk.hu

Kulcsszavak: hálózatarchitektúra, hálózattípusok, felderítő elemzés/osztályozás/optimalizálás, predikció, élelmiszertudomány, szenzometria

A mesterséges neurális hálózatok számos előnyös tulajdonsággal rendelkeznek, amelyek alkalmassá teszik őket különböző célfeladatok megoldására. Az adatokban rejlő komplex kapcsolatok és mintázatok felismerésével olyan problémák megoldására is alkalmasak, amely esetekben a hagyományos módszerek nem adnának megfelelő eredményt. A mesterséges neurális hálózatok sikeressége abban rejlik, hogy képesek az adatokban rejlő komplex kapcsolatok és mintázatok felismerésére, valamint az ismeretlen minták előrejelzésére is.
A mesterséges neuronhálózatok rendszere elsősorban a nemlineáris trendek modellezésében ér el magas hatékonyságot, ezért a különböző komplex kapcsolatok modellezésére, osztályozásra, kategóriába sorolásra, vagy nemlineáris változók közötti regresszióra alkalmazzák. Napjainkban számos kifinomult algoritmus áll rendelkezésre a neurális hálók tréningezéséhez, amelyek alternatívát jelentenek a hagyományos, bevett módszerekkel szemben (lineáris diszkriminancia analízis, regresszió, stb.).
A kutatásinkban bemutatjuk a mesterséges neurális hálózatok alkalmazásának kritikus pontjait: feladat célja (feltáró elemzés, osztályozás, optimalizálás, predikció), adathalmaz megbízhatósága és nagysága, adathalmazok felosztása, elvárt pontosság, teljesítményjellemzők, stb. A validálás során a modellek tesztelését, azok kompexitásfüggő előrejelző képességét tudjuk ellenőrizni. Két alapvető formája a kereszt-ellenőrzés (validálás és modell tanítása ugyanazon mintán) és a teszt adatkészlettel végzett validálás (validálás és modell tanítása külön független mintán) történik. A leggyakrabban alkalmazott kereszt-ellenőrzéskor az adathalmaz egyik részén a modellt tanítjuk, a másik részén pedig a modellt teszteljük, és ezt addig ismételjük, amíg minden alcsoportunk megjelent egyszer tesztként is felhasználva. A kereszt-ellenőrzés formáját az alcsoportok mennyisége és az alcsoportok kiválasztásának módja határozza meg (egy elem kihagyásos, blokkonkénti, véletlenszerű stb.). A validálás a modellek komplexitásának meghatározásában is lényeges. Ha túl sok, vagy túl kevés látens változót veszünk figyelembe a modell építésekor, akkor a modellünk túlillesztetté, vagy alulillesztetté válik, így elveszti a stabilitását. Összefoglalóan megállapítható, hogy a modell komplexitásával az előrebecslési hiba a teszt készleten növekszik, a modell hibája a modell tanítása során csökken. A gépi tanuláson és mesterséges intelligencián alapuló algoritmusok – random forest, support vectore machine, mesterséges neurális hálózatok, stb. – használata az élelmiszertudományokban egyre inkább elterjedt az osztályozási és regressziós modellezésre, ugyanakkor egyre fontosabb a külső validálás, mert ezek a modellek kifejezetten hajlamosak a túlillesztésre. A kutatásunkban élelmiszervizsgálati és szenzometriai példákon keresztül mutatjuk be a neurális hálózatok neuralgikus pontjait és jellegzetességeit.

Multi-ágens predikció az önvezetésben

Kis Kornél István

Robert Bosch Kft., Budapest, Magyarország, kornelistvan.kis@hu.bosch.com

Kulcsszavak: multi-ágens, önvezető autó, mesterséges intelligencia, predikció

Az önvezető autók szoftver architektúrájának leggyakoribb megvalósítása - nagyon leegyszerűsítve - a 'sense-think-act' lépéssor valós idejű, iteratív, tehát ismétlődő végrehajtását jelenti. Az önvezető járműnek első lépésként szenzoraival (radar, LIDAR, kamera stb.) azonosítania kell az összes ún. statikus objektumot (pl.: útpadka, közlekedési táblák stb.) és a dinamikus, más megfogalmazásban saját akarattal rendelkező ágenseket a jármű közelében (pl.: gyalogosok, többi közlekedő stb.). Az észlelési lépés után egy értelmező és előre jelző lépés következik, amely során a jármű eldönti, mi a releváns információ a tovább haladáshoz a kiinduló információmennyiségből. A harmadik lépésben a környezetre adott megfelelő reakció kiszámítása történik, amely során egy pontos terv készül az autó jövőbeli pályájára vonatkozóan.

A kutatások egyik központi, sok részletében máig megoldatlan kérdése ez az előre jelző lépés, más szóval: Milyen algoritmus tudja megbecsülni, hogy a dinamikus objektumok hol fognak tartózkodni a következő néhány[1] másodpercben?

A témában sok éve folyik intenzív kutatás-fejlesztési tevékenység, az utóbbi 6-8 évben már egyértelműen az AI alapú megközelítések vannak fókuszban. A feladat igen komplex, hiszen a jövőre vonatkozó előrejelzés mindenképpen valószínűség alapú, és ahogy növeljük az előre jelzési horizontot, a lehetséges kimenetelek száma exponenciálisan nő. Az utóbbi 2-3 évben a GNN (gráf neurális hálózat) alapú megoldások terjedtek el, mint új, legfejlettebb módszer, felváltva a korábbi, főleg CNN (konvolúciós neurális hálózat) alapú algoritmusokat, amelyek jobbára túl nagyok és túl lassúak voltak a gyakorlati felhasználhatósághoz. A gráf típusú adatszerkezeteket feldolgozni képes neurális hálózatokat korábban sikerrel alkalmazták például gyógyszeripari, illetve számos kombinatorikai optimalizálási probléma megoldásában is.

A termékfejlesztés szempontjából kulcskérdés, hogy a betanított neurális hálózat teljesítmények mennyire lesz robosztus, mennyire kiszámítható (átlátható) a működése, és hogy mennyi adat és számítási erőforrás szükséges a létrehozásához, illetve a futtatásához. A GNN alapú megoldások szerencsére ebben is előrelépést jelentenek, hiszen adathatékonyságuk sokkal jobb, mint a CNN alapú megoldásoké, a kevesebb, de célzottabban definiált rejtett rétegek miatt.

Napjainkban a fentebb bemutatott GNN alapú megoldások továbbfejlesztése, illetve integrációjuk zajlik a meglévő autóipari szoftverkörnyezetekbe. Az ilyen tesztekből levont tanulságok is hozzájárulnak ahhoz, hogy a jövő önvezető autói még természetesebben és intelligensebben reagáljanak a komplex forgalmi szituációkra.

[1] A követelmény 2-3 másodperc egy egyszerű funkció esetén, bonyolult, városi önvezetésnél 4-6 másodperc vagy akár több is szükséges lehet.

Deep Learning alapú beszédszintézis

Zainkó Csaba[1,2]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem (BME), Villamosmérnöki és Informatikai Kar (VIK), Távközlési és Médiainformatikai Tanszék (TMIT),
[2]Beszédtechnológia és Intelligens Interakciók Laboratórium (Smartlab), Budapest, Magyarország, zainko@tmit.bme.hu

Kulcsszavak: Beszédszintézis, TTS, AI, mesterséges intelligencia

A gépi beszédszintézis a számítógépen rögzített vagy éppen mostanában a chatGPT által elállított szöveget alakítja át hangzó beszéddé. A generált beszéd a jelenlegi deep learning alapú megoldásokkal már közel emberi minőségű, sok esetben összetéveszthető az emberek által felolvasott beszéddel.
A beszédtechnológia folyamatos fejlődésben van, a korábbi emberi beszédkeltést modellező, vagy emberi beszéd elemekből álló technológiákat felváltotta a mély neurális hálót (deep learninget) használó megoldások. Ezen megoldások jellemzője, hogy nagyobb mennyiségű emberi beszédet felhasználva, olyan modellt alkotnak, amely igen jó minőségű hangzást eredményez.
Az új generációs deep learning alapú modellek 2016-ban jelentek meg. Az első modellek jó minőséget állítottak elő, de sok szempontból korlátozottak voltak a képességei. Hatalmas számítási igényük volt, a beszéd pedig alig volt paraméterezhető, az eredeti beszélő hangján, stílusában és sebességében tudtak beszédet előállítani.
Azóta újabb modellek, megoldások jelentek meg, a tudományos életben kidolgozott módszereket már hétköznapi és ipari megoldásokban is folyamatosan használják, a modellek rugalmasak és lényegesen kisebb erőforrás szükséglettel bírnak.
A modellekkel szemben korábban csak a közeli emberi minőség elérése volt az elvárás, de manapság újabb területek kerültek fókuszba. Felmerültek olyan igények is, hogy többféle hangszínt, stílust tudjon előállítani, esetleg érzelmeket is jelenítsen meg a beszédben. A szöveget természetesen megfelelően értelmezze, de legyen lehetőség a sebesség és a hangmagasság tetszőleges állítására is, amennyiben speciális alkalmazási terület merül fel. A testreszabhatóság, az egyedi hang megalkotása, a mesterséges média tartalmak generáláshoz szükséges szinkronizáció (pl. szájmozgáshoz szükséges időzített adatok generálása) szintén egyre gyakrabban elvárt.
A fejlődés nem áll meg, még vannak bőven olyan területek, ahol továbblépésre van lehetőség. A mai technológiák még nem „értik”, hogy mit olvasnak fel, sok esetben a helyes kiejtéshez, dallammenethez a felszíni értelmezésen túl, egy mélyebb elemzés lenne szükséges. Az NLP megoldások (pl. chatGPT) beszédszintézissel való kombinálása megoldást adhat erre a problémára is.

Felügyelt, önfelügyelt és gyengén felügyelt neurális technikák a magyar nyelvű beszédleiratozásban

Mihajlik Péter[1,2], Kádár Máté[1,2], Dobsinszki Gergely[1,2], Yan Meng[1], Meng Kedalai[1], Mády Katalin[2]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem, Villamosmérnöki és Informatikai Kar, Távközlési és Médiainformatikai Tanszék, Budapest, Magyarország;
[2]Nyelvtudományi Kutatóközpont, Fonetikai Kutatócsoport, Budapest, Magyarország,
mihajlik@tmit.bme.hu

Kulcsszavak: beszédfelismerés, mélytanulás, neuronháló, önfelügyelt tanulás, transfer learning, gyengén felügyelt tanulás

Tanulmányunkban a legújabb mélytanulási (mesterséges intelligencia) technikákat alkalmazzuk magyar nyelvű gépi beszédleiratozásra. A beszédfelismerési kimenet pontossága nagyban függ a beadott hanganyagok jellegétől, ezért mi a valós alkalmazások szempontjából kiemelt jelentőségű magyar nyelvű spontán beszédre fókuszálunk, melyet a standardizált BEA-Base adatbázisból veszünk. Emellett azonos beszélőktől, azonos körülmények között felvett olvasott szöveget is vizsgálunk kontrasztív jelleggel, valamint a független közösségi „CommonVoice" magyar nyelvű részhalmazán is mérünk.
Elsőként a méltán népszerű Transformer neurális architektúra konvolúciós modulokkal kiegészített, ún. Conformer változatát használjuk enkódolásra (belső reprezentáció kinyerésére) és az egyszerű kimeneti réteg a beszédfelismerésben leggyakrabban alkalmazott CTC (Connectionist Temporal Classification) költségfüggvényt használja. Karakter és szótöredék kimeneti egységeket is vizsgálunk, ahol az utóbbit találjuk alkalmasabbnak. Az angol nyelven, az NVIDIA által felügyelten előtanított modellekkel lényegesen jobb eredményeket kapunk, mintha csak a magyar nyelvű adatokra támaszkodnánk és nem alkalmaznánk „transfer learning"-et.
Másodikként az OpenAI Whisper rendszerét teszteltük. Ez teljes (enkóder + dekóder) Transzformer struktúrát használ és gyengén felügyelt módon tanították. Azaz a vélhetően emberi erővel előállított, nem feltétlen egzakt leiratokkal készült többnyelvű (zömmel angol és spanyol + közel száz nyelvű) beszédanyagokon, összesen 680 ezer órányi mintán tanult. Érdekes tapasztalat volt, hogy noha magyar nyelvű hanganyagokat is látott a rendszer, a pontossága spontán magyar beszéden finomhangolás nélkül igen alacsony, alig 50%. A dekóder modul finomhangolásával ez javítható volt az előző rendszer szintjéig, azonban attól körülbelül 100-szor lassabb és jóval több memóriát igényelt a legjobb large-v2 verzió, de a medium méretű modell is rendkívül erőforrásigényes (csökkent pontosság mellett).
Végül, az önfelügyelt – azaz az előtanuláshoz (reprezentáció-tanuláshoz) leiratot egyáltalán nem, csupán hanganyagot igénylő –, a META által fejlesztett wav2vec2 rendszert hangoltuk magyar nyelvű beszédfelismerésre. Meglepő módon nem a legnépszerűbb és legnagyobb, többnyelvű hanganyagokon előtanított modellek finomhangolásával kaptuk a legjobb eredményeket, hanem egy finn-észt-magyar Európai Parlamenti felszólalásokkal előtanított, majd a BEA-Base-en finomhangolt modell végzett első helyen nagy fölénnyel minden adatbázison és minden tekintetben. Transzformer nyelvmodell hozzáadásával tovább javítottuk az eredményeket, amellett, hogy a számítási erőforrás-igény kézben tartható volt, nagyságrendileg alacsonyabb, mint a Whisper esetén.

Következtetésünk, hogy jelenleg az önfelügyelt előtanítású wav2vec (Transzformer) struktúra a megfelelő szakértelem mellett adja a legjobb pontosságot magyar nyelvre. Egy tisztán magyar nyelvű, változatos, spontán beszéddel történő önfelügyelt előtanítás további érdemi javulással kecsegtetne, és ehhez csak nyers (lejegyzetlen) hanganyagra, illetve jelentős számítási erőforrásra lenne szükség.

Sok a szöveg?! Olvass inkább a sorok közt!

Molnár Anna Enikő[1], Tamási-Mészáros Evelin[1]

[1]Statistical Products Hungary Kft. (Clementine), Budapest, Magyarország,
amolnar@clementine.hu, emeszaros@clementine.hu

Kulcsszavak: szóbeágyazási modell, word embeddings, többnyelvű szövegelemzés, Python, duplikációkezelés, hálózat, entitáskinyerés

Az egyre szélesebb körben hozzáférhető szöveges adatok térhódításának köszönhetően egyre többször szembesülhet egy elemző azzal a kihívással, hogy az egyezőnek számító tartalom figyelmen kívül hagyása torzíthatja a statisztikát. A statisztikai adatszolgáltatás egy univerzális elvként értelmezhető a különböző tudományterületeken, amely a nemzetközi és nemzeti jogi elvekben is tükröződik. Az adatok elemezhetősége és elérhetősége egy olyan elvárás, amelyet egy az adott tudományterület iránt érdeklődő ember támaszt a publikált statisztikákkal szemben. Azonban nagyon nehéz eligazodni az információk tengerében. Az EUROSTAT 2022 decemberében indult pályázatának az egyik célja is az volt, hogy felhívja a figyelmet az adatok sokszínűségére, ezzel lehetőséget biztosítva a lelkes elemzők számára, hogy szöveganalitikai képességeiket összemérjék.

A felhívás keretén belül az interneten fellelhető álláshirdetések között található egymást átfedő hirdetéseket kellett feltárni. A verseny célja összefügg azzal a törekvéssel is - amely az Európai Unió jogrendjében is megtalálható -, miszerint „Az (EU) 2016/589 rendelet 17. cikke elrendeli egy egységes rendszer létrehozását a tagállamokból származó állásajánlatok, álláspályázatok és önéletrajzok EURES-portálon történő összegyűjtése érdekében.”[2] Az azonos tartalmat tükröző álláshirdetések torzítják a statisztikát, hiszen ha ugyanazon pozíció több helyen is publikálásra kerül, akkor felülreprezentáltnak tűnhet az adatbázisban.

A szöveganalitikai kihívás első nehézségét az jelentette, hogy a vizsgálandó szövegek között 32 különböző nyelvet tudtunk detektálni, melyekből az alábbi nyelvek szerepeltek leggyakrabban: angol, német, francia, holland, spanyol, olasz, portugál, svéd, észt, lengyel, magyar. A módszertan kialakításakor ezt az aspektust is figyelembe kellett venni, hiszen a különböző nyelvű szövegek előkészítéséhez eltérő eszközök szükségesek. A nyelvek sajátosságai meghatározzák a felhasználható elemzési eszköztárat, hiszen a korpusz előállításához szükséges lépések - mint a tokenizálás vagy a ragok eltávolítása - különböző nyelveken eltérő módon valósíthatóak meg.

A kivitelezés Python nyílt forráskódú programnyelven zajlott szóbeágyazási modellek alkalmazásával, de különféle technikai és módszertani kihívás is felmerült. A többnyelvű szövegek ténye az elérhető szakirodalmakban való további tájékozódást indikált. Jelenleg kevés olyan szövegek közötti hasonlóságot mérő modell létezik, amely áthidalja ezt a problémát, jelen esetben a sentence-transformers csomagra esett a választás, mivel korábbi tanulmányok alapján jól teljesít többnyelvű adatokon is. A csomagon belül különféle előre tanított modellek állnak rendelkezésre, amelyeket saját adattal tovább lehet fejleszteni. A leírás szerint közel 50 nyelvű adatot használtak fel a tanításhoz, amelyek között a magyar is szerepel, így esett a választás a *paraphrase-multilingual-MiniLM-L12-v2* nevű alapmodellre a többnyelvű szövegösszehasonlításhoz, az angol nyelvűek esetén pedig az *all-MiniLM-L6-v2* nevű modellre. A tapasztalat azt mutatta, hogy a modellek segítségével különböző nyelvű szövegek között

[2] https://eur-lex.europa.eu/legal-content/HU/TXT/?uri=CELEX%3A32017D1257&qid=1682459328615

kevésbé azonosítható a hasonló tartalom, mint az azonos nyelvűek között, ezért az előkészítés első lépését a nyelvi egységesítés jelentette, miszerint a szövegek angol nyelvre kerültek lefordításra, majd ez képezte a további elemzés alapját. A feladat magában foglalt egy matematikai problémát is, miszerint minden lehetséges hirdetéspárt szükséges lett volna megvizsgálni, de ez a több mint 112000 álláshirdetésből álló adatbázis esetén a rendelkezésre álló hardverekkel nem volt kivitelezhető, ezért egy iteratív folyamat segítségével fokozatosan kerültek meghatározásra a lehetséges párok, amelyek utána egy manuálisan meghatározott szabályrendszer segítségével az előre meghatározott kategóriákba kerültek besorolásra. Ezzel a módszerrel 575190 potenciális pár került feltárásra, amelyek részben, szemantikailag vagy akár teljesen egymás variánsai voltak.
A verseny beküldési időszaka 2023. március 31-ével véget ért ugyan, de a lelkesedés nem csökkent.
A koronavírus járvány 2020-as megjelenését követően a sajtó nagy érdeklődéssel követte az eseményeket, és az egész médiát áthatotta a téma. A téma nagysága felvetette a probléma aktualitását, hiszen a média különböző forrásai eltérő stílusban és formában teszik közzé a nyilvános híreket. A korábban részletezett módszertan alkalmasnak tűnt arra, hogy az álláshirdetések duplikációmentesítéséhez hasonlóan definiálásra kerüljenek azonos információtartalmak mint például a napi koronavírus statisztika, oltások körüli hírek, közéleti szereplők a járvány szemszögéből, vagy a koronavírussal kapcsolatos intézkedések. A kialakított módszertan finomítása volt szükséges, hiszen a szöveges adatok nyelve magyar volt. Az évek során gyűjtött hírekből álló adatbázis több mint 114000 hírt tartalmaz, ahol a vizsgálandó adatok tartalmazzák a hír eredeti szövegét, a megjelenés dátumát és a hír forrását.
A tapasztalat azt mutatta, hogy a több nyelvre alkalmazott módszertanhoz képest az egynyelvű szövegek kissé eltérő megközelítést igényelnek.
Elméleti szempontból megközelítve a vizsgált adatok angol nyelvre való lefordítása javított a hasonló cikkek azonosításában, hiszen angol nyelvre illesztett előre tanított szóbeágyazási modellek számos helyen elérhetőek, és ezek a rendelkezésre álló adatokkal tovább finomíthatóak. Az álláshirdetések szövegei esetén relevánsnak tűnt a szakmák beazonosítása a hasonlóság típusának meghatározásánál, és ez a szótár alapú megközelítés és egy hasonlósági arányszám együttes kezelésével hatékonynak is bizonyult. A cikkek hasonlóságának feltárása összekapcsolódik egyben egy témameghatározással is, amelyben nagy szerepet kaphat az entitáskinyerés a szakmák definiálásával analóg módon. A nyílt forráskódú entitáskinyerés a Python alapú Spacy csomag segítségével került kivitelezésre, amelyből a későbbiekben a hasonlósági arányszám kiegészítésére szolgáló szótáralapú megoldás megszületett. A gyakorlati alkalmazás azt mutatta, hogy a Spacy magyar nyelven ugyan használható, viszont nagyon érzékeny olyan apró részletekre is mint a kis- és nagybetűk különbsége, egybe-különírás vagy akár a betűszintű elírások. Az entitáskinyerése azonban segítséget nyújtott egy komplett szótár kialakításában, amelyben a szakértők által beazonosított tévesztések korrigálhatóak, de az eredményként kapott helyek, személyek, dátumok és különféle közösségi médiában használatos hivatkozások (@-ok, #-ek) hozzájárultak a hasonlóság pontosabb meghatározásához, még ha azokat fenntartással is kellett kezelni.
A megvalósítás technikai hátterét tekintve az elemzés hardverigénye nem indokolta GPU használatát, elegendő volt CPU segítségével végezni a számításokat. Értelemszerűen nagyobb adatmennyiség esetén szükség lehet a futtatás optimalizálására, amely megnyilvánulhat akár a különféle elemzési lépések párhuzamosításában vagy akár a rendelkezésre álló hardverek bővítésében.
Összességében elmondható, hogy a módszertan finomítása után a tartalmi végeredmény mutatta a várt hipotézist, miszerint a rendelkezésre álló szóbeágyazási modellek saját adatokkal tovább tanított változatából kinyert hasonlósági arányszámok, a megjelenés dátuma és a közös entitások segítségével feltárhatóak voltak lényegi tartalmi összefüggések a szövegek között. A

cikkek nagy hálózata a különféle beazonosított tartalmak mentén olyan alhálózatokra bontható, ahol a szövegek szorosabb kapcsolatban állnak egymással, mint a hálózat többi tagjával. Ezáltal egy hasonló módszertan kivitelezése támpontokat nyújthat egy klaszterezési feladathoz is, de mindig fontos szem előtt tartani az elemzendő adatok sajátosságait.

Árnyék a gépezetben: mit kezdjünk a mesterséges intelligencia fekete dobozával?

Huszti Dániel

IBM Magyarországi Kft., Budapest, Magyarország, daniel.huszti1@ibm.com

Kulcsszavak: mesterséges intelligencia, AI Governance, etikus MI, átlátható MI, fekete doboz

Manapság mindenki a mesterséges intelligenciáról beszél, azonban a legtöbb vállalat bizonytalan milyen területen és miként használja ki a technológia adta lehetőségeket. Ennek főbb okai a technológiába vetett bizalom hiánya, a szigorú törvényi szabályozások és az Európai Unió által bevezetendő mesterséges intelligencia felelősségi elvek. Például a pénzügyi szervezetek gyakran hagyományos szabály alapú elemzéseket végeznek a kockázatelemzési területen, mivel nem bíznak a modellekben, nem látnak bele azok milyen szempontok alapján döntenek, s félnek az esetleges büntetésektől.

AI Governance és a fekete doboz szerű működés feloldása

A mesterséges intelligencia modellek felhasználásának hagyományos eszközkészlettel és manuális támogatással való megközelítésének hátránya, hogy a modell üzleti validálása akár több hónapba is telhet, az adattudósok nem tudják visszamérni az élesben futtatott modellek pontosságát és nem látnak bele azok fekete doboz szerű működésébe.

Az IBM adatelemzési platformja nagy hangsúlyt fektet az AI Governance támogatására, azaz a mesterséges intelligencia etikus és átlátható használatára. Az IBM Cloud Pak for Data megoldás segítségével a teljes a modell életciklus lefedhető: az adatvagyon-gazdálkodás (Data Governance), adatelőkészítési, a modell építési és futtatási (MLOps) folyamat támogatáson át egészen az éles üzembe vett/helyezett döntési folyamat monitorozásáig és üzleti kockázat kezeléséig (AI Governance).

Az IBM platformja olyan eszközkészletet kínál, mely képes a modellek fejlesztését és futtatását felgyorsítani MLOps támogatással és integrált megoldásként képes lefedni az AI Governance három fő pillérjét:

1. **Modell életciklus támogatás**: Egyszerűsíti a modellek menedzselését és üzemeltetését, automatikus dokumentációt biztosít a teljes modell életciklus folyamán, segít feloldani azok fekete doboz szerű működését.
2. **Modell kockázat menedzsment**: Automatizált tények és munkafolyamatok kezelése révén kezeli a kockázatokat és az üzleti irányelveknek való megfelelést.
3. **Törvényi szabályozási megfelelőség**: Külső szabályozások betartása az auditálhatóság és a törvényi megfelelés érdekében.

Az IBM AI Governance keretrendszere átfogó, egyedüli gyártóként támogatja mindhárom pillért, automatizált, egységes, szabályozás vezérelt és mindemellett nyitott, mivel külső gyártók modell építési és futtatási megoldásához jól integrálható. A keretrendszer több szorosan integrált komponensből áll:

1. Modell életciklus menedzsment - „Centralized facts”

2. Automatizált modell validáció, monitorozás és elmagyarázható döntések - „Monitor and explain”
3. Jogszabályi megfelelések teljesítése a modell életciklus során - „Governance control”

Modell életciklus menedzsment

Minél inkább nő a mesterséges intelligencia modellek számossága, annál nehézkesebb a modellek életciklusának és verziójának nyilvántartása. Az AI Governance keretrendszer „Centralized facts” komponense, az IBM Cloud Pak for Data platform AI Factsheets szolgáltatása automatikus modell leltárt biztosít az egyes gépi tanulási modellek teljes életciklusának nyomon követésére. Könnyen áttekinthetik a teljes szervezet (adattudósok, fejlesztők, üzleti döntéshozók), hogy mely modellek vannak fejlesztés alatt vagy élesbe állítva, milyen a validált modell teljesítménye, és milyen tanító adathalmazt használnak.

Automatizált modell validáció, monitorozás és elmagyarázható döntések

Az üzleti felhasználók és az iparági szabályozások megkövetelik, hogy a modellek jól teljesítsenek a gyakorlatban is. A modell validáció, tesztelés és az éles környezetben történő működés közbeni felügyelet nagyon komplex, időigényes és nehezen menedzselhető folyamat. Az AI Governance keretrendszer „Monitor and explain” komponense, az IBM Cloud Pak for Data platform Watson Openscale szolgáltatása ezekre a problémákra nyújt kézenfekvő megoldást a következő főbb képességei által:

- Modellek validálása az élesbe állás előtt. Egyedi elemzések készítése mint részlehajlás vizsgálat (fairness, bias), minőségi kritériumok, döntési körülmények részletes elemzése.
- Modell teszek futtatása és tesztriportok készítése.
- Modellek teljesítményének összehasonlítása (például verziók között).
- Modell éles üzembe helyezés utána folyamatos monitorozása, döntési körülmények részletes elemzése. Modellek elmászásának detektálása és jelzése.
- Üzleti felhasználók számára a modellen átfutó tranzakciók részletes elmagyarázása, ezáltal a fekete doboz szerű működés feloldása.

Jogszabályi megfelelések teljesítése a modell életciklus során

Az elkészült és élesbe helyezett modellek felügyelete mellett a jogszabályi megfelelőség és az üzleti döntéshozatal dokumentálása is fontos. Az AI Governance keretrendszer „Governance control” komponense, az IBM OpenPages Model Risk Governance olyan képességeket kínál, mint a modellek tesztjeinek nyilvántartása, üzleti kockázatok nyilvántartása, testreszabható üzleti munkafolyamatok támogatása és a modell életciklusának teljes dokumentálása. Ez a központi nyilvántartás segít a modellek jól dokumentált, automatizált nyilvántartásában és azokhoz tartozó üzleti kockázat hatékony kezelésében.

Összefoglalva az IBM AI Governance keretrendszere egy olyan átfogó, integrált és nyitott megoldást képes nyújtani, mellyel a meglévő és új mesterséges intelligencia modellek egyaránt etikusak, átláthatók, jól nyomon követhetők és törvényi szabályozásoknak megfelelően kezelhetők. A 2023 májusi IBM Think konferencián bejelentett IBM Watsonx platform további fejlesztéseket hoz a mesterséges intelligencia életciklus kezelés és átláthatóság terén, és már nagy nyelvi modellek (LLM) építésére, megbízható alkalmazhatóságára és felügyeletére is kiterjed.

A mesterséges intelligencia kiberbiztonsági kockázatai

Bányász Péter

Nemzeti Közszolgálati Egyetem Államtudományi és Nemzetközi Tanulmányok Kar
Kiberbiztonsági Tanszék, Budapest, Magyarország, banyasz.peter@uni-nke.hu

Kulcsszavak: mesterséges intelligencia, kiberbiztonság, adathalászat, Chat GPT, dezinformáció, sérülékenység vizsgálat

2022. év végén a main stream médiában szinte berobbantak az Open AI Chat GPT-je okán a mesterséges intelligencia platformok.
„A mesterséges intelligencia (MI) alkalmazása a kiberbiztonság terén komoly lehetőséget kínál a szervezetek számára a biztonsági védelmi megoldások hatékonyságának növelésére. Azonban, miközben az MI képes a kibertámadásokat azonosítani és megelőzni, az MI maga is potenciális sebezhetőségekkel rendelkezik, amelyeket a támadók kihasználhatnak. Az előadás során megvizsgáljuk az MI és a kiberbiztonság közötti kapcsolatot, és bemutatjuk a legfontosabb kihívásokat, amelyekkel a szervezeteknek szembe kell nézniük az MI alkalmazásakor a kiberbiztonság terén. Kitérünk az MI biztonsági vonzataira, valamint a technológiai és társadalmi kihívásokra, amelyekkel az MI alkalmazása járhat a kiberbiztonsági iparban. A konkrét példákon keresztül bemutatjuk, hogy hogyan lehet hatékonyan alkalmazni az MI-t a kiberbiztonság terén, és miként lehet minimalizálni a kockázatokat, amelyekkel az MI alkalmazása járhat. Vizsgáljuk továbbá, hogy milyen intézkedéseket kell meghozni a szervezeteknek annak érdekében, hogy felkészüljenek az MI által jelentett kihívásokra és az MI alkalmazásának következményeire a kiberbiztonsági területen. Az előadás célja, hogy felhívja a figyelmet a MI és a kiberbiztonság közötti kapcsolatra, és arra ösztönözze a szervezeteket, hogy alaposan megvizsgálják az MI alkalmazásának lehetséges előnyeit és kockázatait a kiberbiztonság terén, és megfelelő intézkedéseket tegyenek a kibertámadásokkal szembeni védelem érdekében.” - írta a Chat GPT arra a kérdésre, milyen absztraktot írna egy konferencia előadásra, amin a mesterséges intelligencia és kiberbiztonság kapcsolatát vizsgálná. Az előadás a fentiek mellett kitér az egyéb MI alapú platformok jelentette kiberbiztonsági kockázatok ismertetésére, amelyek kombinált alkalmazása paradigmaváltást fog jelenteni nem csupán az adathalászat, sérülékenységvizsgálat, de a lélektani műveletek aspektusából is.

Biztosítási csalások és hálózatelemzés

Hans Zoltán[1], Hegedüs Pál[1], Pancza Judit[1]

[1]Statistical Products Hungary Kft. (Clementine), Budapest, Magyarország,
zhans@clementine.hu, phegedus@clementine.hu, jpancza@clementine.hu

Kulcsszavak: biztosítás, csalás, hálózat, i2

A biztosító társaságok életében állandó gondot okoz a csalások felderítése. Ez fokozottan érvényes a jelenlegi, válságokkal tarkított időszakra. Megvizsgáltuk, hogyan modellezhető a probléma az i2 szoftverek (ezen belül az i2 Analyst's Notebook (hálózatelemzés és vizualizáció) és az i2 iBase (entitásokat és linkeket tartalmazó adatbázis)) alkalmazásával, és mutatunk néhány megoldási javaslatot a csaló hálózatok leleplezésére. Mivel Magyarországon a gépjármű biztosítási csalások a leggyakoribbak, ezért elsősorban erre koncentrálunk. Elemzés és vizualizáció szempontjából alapvető fontosságú az iBase séma definiálása (azaz: mely adatokból lesznek entitások, linkek és ezek attribútumai). A különböző sémák és a különböző ábrázolásmódok más-más elemzéseket tesznek lehetővé. A „hagyományos" kapcsolati háló diagram mellett kitérünk az idővonalas ábrázolás előnyeire is. A kapcsolati háló ábrázolást főleg akkor használjuk, amikor a résztvevők hálózatán belül keresünk lehetséges csaló részcsoportokat. Az idővonalas ábrázolás viszont akkor lehet hasznos, amikor adott szereplők tevékenységeiben keresünk valamilyen (jellemzően ismétlődő) mintázatot. Mindkét ábrázolásmódra mutatunk példát. Megjegyezzük, hogy az i2 Analyst's Notebook lehetővé teszi ezen két ábrázolásmód egyszerű egymásba alakítását. A csalók sok esetben folyamodnak ahhoz a trükkhöz, hogy szándékosan hibásan adják meg az adataikat, de ettől függetlenül is előfordulhat az egyes személyek adatainak hibás rögzítése. Emiatt nem ritka, hogy ugyanaz a személy vagy cég duplán (vagy még több példányban) fordul elő a biztosítók adatbázisában. Ez természetesen a hálózat elemzése szempontjából sem előnyös. Azt szeretnénk, ha a több példányban létező szereplők a diagramon (és az iBase adatbázisban is) összevontan, egy entitásként jelenjenek meg. Az entitások azonosítására és a nem triviális kapcsolatok feltárására jól használható az IBM Infosphere Identity Insight (ISII) szoftver. Az i2 szoftverek és az ISII együttes alkalmazásával olyan komplex rendszer hozható létre, amely hatékonyan segíti a biztosítási csalások leleplezését.