

Doktori (PhD) értekezés tézisei

Gépi tanuló modellek tervezése és optimalizációja

Lakatos Róbert

Témavezető: Dr. Hajdu András



DEBRECENI EGYETEM
Informatikai Tudományok Doktori Iskola

Debrecen, 2025

Tartalom

1. Háttér	4
2. Az értekezés fő eredményei	7
2.1. Az olajár-trend irányának előrejelzése hangulatelemzés segítségével	7
2.2. Gépi tanuláson alapuló rendszer a vásárlói véleményekből származó információk kinyerésére	10
2.3. Limitált adatmennyiség esetén alkalmazható mély tanuló architektúra a dohányzás felismeréséhez	13
2.4. Domaintranszfer technikák vizsgálata LLM-vezérelt tudásalapú rendszerek fejlesztéséhez	17
2.5. Elméleti modellek optimalizálása	20
Hivatkozások	21
Publikációs lista	26

1. fejezet

Háttér

A mesterséges intelligencia kutatása átalakuló evolúción ment keresztül, amelyet elsősorban a gépi tanuló modellek hatékonyságának és teljesítményének fokozása vezérel. Azonban ezekben a modellekben rejlő valódi potenciál az optimalizálásukon múlik.

Az optimalizálás azonban sokrétű fogalom. Magába foglalja különböző optimalizáló algoritmusokat és stratégiákat egyaránt. Például a Kingma és Ba [1] által bevezetett Adam optimalizáló széles körben népszerűvé vált a neurális hálózatok betanításában, köszönhetően az adaptív tanulási képességének. Hasonlóképpen, a generatív modellezési feladatokban a generatív ellenséges hálózatokban használt specifikus optimalizálási stratégiák kulcsfontosságúak [2].

Az optimalizálási folyamat eleve függ a modell típusától és a feladat természetétől, a neurális hálózat alapú architektúrák pedig érzékenyek a konkrét optimalizálási stratégiákra. Ezért ebben az összefüggésben az optimalizálási technika megválasztását a célfeladat és a rendelkezésre álló adatok minősége határozza meg.

A gépi tanulási modellek optimalizálását meghatározó kritikus kérdések a következők:

- Hogyan válasszuk ki a megfelelő betanítási módszert?
- Hogyan optimalizáljuk a hiperparamétereket?
- Hogyan építsünk komplex modell architektúrákat optimálisan?

Az adott probléma természete jellemzően meghatározza a képzési módszereket. A gépi tanulási feladatokat általában felügyelt, félig felügyelt vagy

nem felügyelt tanulási kategóriákba sorolják [3], amelyek külön optimalizálási stratégiákat igényelnek.

A felügyelt tanulás során a címkézett adatok elérhetősége irányítja az optimalizálási folyamatot. A félig felügyelt tanulás ezzel szemben olyan stratégiákat használ, amelyek címkézett és címkézetlen adatokat használnak fel, míg a felügyelet nélküli tanulás célja rejtett minták felfedezése az adatokban.

Ezenkívül az olyan haladó képzési stratégiák, mint például az átviteli tanulás, a [3] finomhangolás és az előképzés is elengedhetetlenné váltak [4], és ezek a nagy nyelvi modellek (LLM-ek) megjelenésével különösen nagy hangsúlyt kaptak. Ugyanis a [5] transzformátor-architektúrára épülő modellek kiterjedt adatkészletekre vannak kiképezve, és figyelemre méltóan teljesítettek a gépi fordítási, szövegösszegzési és kérdésmegválaszoló feladatokban. A kutatások kimutatták, hogy az általános feladatokra előképzett és a konkrét feladatokra finomhangolt LLM-ek jobban teljesítenek, mint a kizárólag a feladatspecifikus adatokra kiképzett [4] modellek. Az előképzés tehát javítja a feladatspecifikus teljesítményt és javítja a modell általánosítási képességét.

A hiperparaméterek optimalizálása kulcsfontosságú a gépi tanulási modellek sikeréhez. A hiperparaméter optimalizálás szabványos módszerei közé tartozik a Grid Search, a Random Search [6], a Bayes optimalizálás [7] és a genetikai algoritmusok [8]. Ezenkívül az ensemble technikák, mint például a Bagging, Boosting és Stacking [9] robusztus megoldásokat kínálnak az összetett modellek architektúrájának optimalizálására.

A következő fejezetekben röviden áttekintjük a dolgozat főbb eredményeit öt kidolgozott módszerre vonatkozóan.

A 2.1 fejezet az LLM-eken és a visszatérő neurális hálózatokon alapuló ensemble megoldás fejlesztését vizsgálja, amely lehetővé teszi az olajármozgások előrejelzését. Szöveges adatok, hangulatalapú LLM-ek és idősoros modellek alkalmazásával bemutatom a különböző adatforrások kombinálásának lehetőségét egy predikciós modell felépítéséhez.

A 2.2 fejezetben egy olyan Pipeline-t mutatok be, amely a gépi tanulás és a klasszikus módszerek együttesén alapul. Mérési eredményeink azt bizonyítják, hogy a jelenleg népszerű módszereknél kifinomultabb és egységesebb megoldást nyújt a vásárlói vélemények releváns megállapításainak kinyerésére.

A 2.3 fejezet a multimodális architektúrák teljesítménypotenciáljára összpontosít, különösen a dohányzásra utaló viselkedések szövegből és képekből történő kimutatásában. Egy olyan modellt mutatok be, amely korlátozott edzésadatok mellett is jól teljesít.

A 2.4 fejezetben összehasonlítom a RAG és a tartomány-specifikus finomhangolás teljesítményét, és megmutatom, hogy a RAG-nak előnye van az LLM-en alapuló új tudásbázisrendszerek fejlesztésében.

Végül a 2.5 fejezetben megmutatom, hogyan vezethetjük vissza az elméleti modellek, például a SIR optimalizálását olyan algoritmusok alkalmazásáig, amelyek megoldják a legkisebb négyzetek görbe illesztési problémáit. Ez lehetővé teszi, hogy olyan optimális elméleti modelleket állítsunk elő, amelyek alkalmasak a neurális hálózatok előképzésére.

2. fejezet

Az értekezés fő eredményei

2.1. Az olajár-trend irányának előrejelzése hangulatelemzés segítségével

1. tézispont. *Kidolgoztam, egy gépi tanuláson alapuló architektúrát, amely figyelembe veszi a természetes emberi nyelven írt hírek hangulatát az olajárának előrejelzéséhez. A hírek hangulatának meghatározásához klasszikus statisztikán alapuló algoritmust és nagy nyelvi modelleket használtam. Ezzel pedig egy olyan adatforrást hoztam létre, amelyben a hangulati értékek nagy pontosságú, 84,08%-os korrelációt mutattak az olajár valós irányával. Az eljárás javítani tudta az LSTM és GRU modellek előrejelző képességét. Ezen eredmények alapján megmutatom, hogy a hangulat elemzésének vannak potenciális lehetőségei az olajármozgások és ezzel együtt akár más tőzsdei termékek előrejelzésében.*

Kapcsolódó publikáció: [P12]

Az értekezés ezen fejezetében bemutatott módszer a szöveges információk és az olajármozgások közötti összefüggések feltárásával foglalkozik. A fejezet középpontjában a Twitter (új nevén X) [10] közösségi média platformon megjelenő hírek által kifejtett érzelmek és az egyik legfontosabb energia-hordozó az olajárának irányváltozásai közötti korreláció elemzése áll, melyhez természetes nyelvi feldolgozás technikáit, valamint gépi és mély-tanuló módszereket alkalmaztunk.

Az előrejelzés javítása érdekében mind a szöveges, mind pedig az áradatoknak előzetes feldolgozáson kellett átesniük. Az áradatok esetében a következő műveleteket alkalmaztuk:

- A zajos árváltozások kiszűrése érdekében mozgóátlagot használtunk 10-es ablakmérettel.
- A *Brent* és a *Crude* olajárak közötti korreláció elemzésével egy egységes olajár mutatót hoztunk létre.
- Az ármozgásokat előző naphoz vett irányuknak megfelelően -1 (csökkenés) vagy +1 (növekedés) értékekre transzformáltuk.

A szöveges híradatok esetében a következő műveleteket alkalmaztuk:

- A nem releváns híreket kiszűrtük és csak azokat tartottuk meg amelyek az "*olaj*" szót megemlítik.
- A speciális karakterek vagy szöveg részeket, mint például az URL linkek, eltávolítottuk.
- Szentiment analízist alkalmazva detektáltuk az egyes hírek érzelmi állapotát.
- A híreket naponként aggregáltuk és a hangulati értékeiket ennek megfelelően átlagoltuk.

Az érzelmek detektálására a Vader [11], a Sentiment Roberta Large English [12] és a Twitter Roberta Base Sentiment [13] szentiment modelleket alkalmaztunk. A hangulat elemzés hatékonyságát az olajármozgások előrejelzésében a tényleges árváltozásokkal való korreláció alapján értékeltük. Ehhez az érzelmeket az árváltozások irányára képeztük le. Ez azt jelentette, hogy a hangulat elemző modellekkel a negatív híreket -1 a pozitív híreket +1 értékekre képeztük le, ami így már megfeleltethetővé vált az ármozgások irányának. Ennek eredményeképpen meg tudtuk határozni azok pontosságát az eredeti árváltozások irányához képest.

Felismerve a hírforrások időbeli változékonyságát, havi elemzéseket végeztünk annak érdekében, hogy az optimális forrásokat minden időszakban azonosítsuk. Ez a megközelítés sokoldalúbb és pontosabb adatkészlet létrehozását tette lehetővé a modellek tanításához. Továbbá, a hangulat elemzést torzító zajos hírek azonosítása és kiszűrése érdekében egy zajszótárt

építettünk. A zajos tételek kiszűrése javította az hangulat elemző modellek előrejelzésének pontosságát.

A továbbiakban a hírek hangulat elemzésével előállított adatokat idősoros adatokként kezeltük, amelyekre így idősoros előrejelzésekre alkalmas modelleket tudtunk illeszteni. Az idősoros előrejelzésekhez GRU [14] és LSTM [15] ismétlődő neurális hálózatokat alkalmaztunk, kimeneteiket az ármozgás irányára szabva. Ezt úgy értük el, hogy tangens hiperbolikus aktivációs függvényt használtunk, hogy az előrejelzések a kívánt $(-1, +1)$ tartományban maradjanak. A legjobb súlyok megtalálásához Adam optimalizációs algoritmust és Early Stopping módszerét alkalmaztuk. A GRU és LSTM hálózatok teljesítményét a 5, 15, 20, 25, 30, 35, 40, 45 és 50-es Unit paraméterekkel is megmértük.

Bemutatott, megoldásunk adatközpontú megközelítést alkalmaz a hatékony architektúra felépítéséhez, amely mind a megfelelő betanítási módszer megválasztása mind az optimális komplex architektúra építésének kérdéseire kiemelt hatással van.

2.2. Gépi tanuláson alapuló rendszer a vásárlói véleményekből származó információk kinyerésére

2. tézispont. *Kidolgoztam egy keretrendszert, amely képes a vásárlói véleményekből kinyerni a releváns információkat. A keretrendszer létrehozásához megvizsgáltam a hatékony információkinyerés tudományterülete által megjelölt legjobb megoldásokat. A szakterület által elfogadott mérési módszereknek megfelelően összehasonlítottam őket saját megoldásommal. A saját megoldásom előnyeit tapasztalati eredményekkel alátámasztottam. A keretrendszert a hatékony működés és a jó átlátható struktúra érdekében egy szekvenciális adatorientált folyamatba csomagoltam.*

Kapcsolódó publikáció: [P3]

Ebben a fejezetben részletesen bemutatjuk azt a módszert, amelyet kidolgoztunk az online kereskedelemben található szöveges adatokból történő információk kinyerésére. Ismertetjük a kiértékelési stratégiánkat, és leírjuk a különböző kulcsszókinyerési technikákat, amelyekből a leghatékonyabb végül egy egységes rendszer (pipeline) tervezéséhez vezetett el bennünket. Három népszerű kulcsszó- és kulcskifejezés-kinyerési megközelítést vizsgáltunk: n-gram, függőségi elemzés és szemantikai beágyazás. Vizsgálatunk eredményeként a szemantikai beágyazást integráltuk modellünkbe (ld. 2.1 ábra) az információkinyerés céljából. A pipeline architektúrája a skálázhatóságot, az adaptálhatóságot és a stabilitást helyezte előtérbe. A rendszer működése a következő lépéseket foglalja magába:

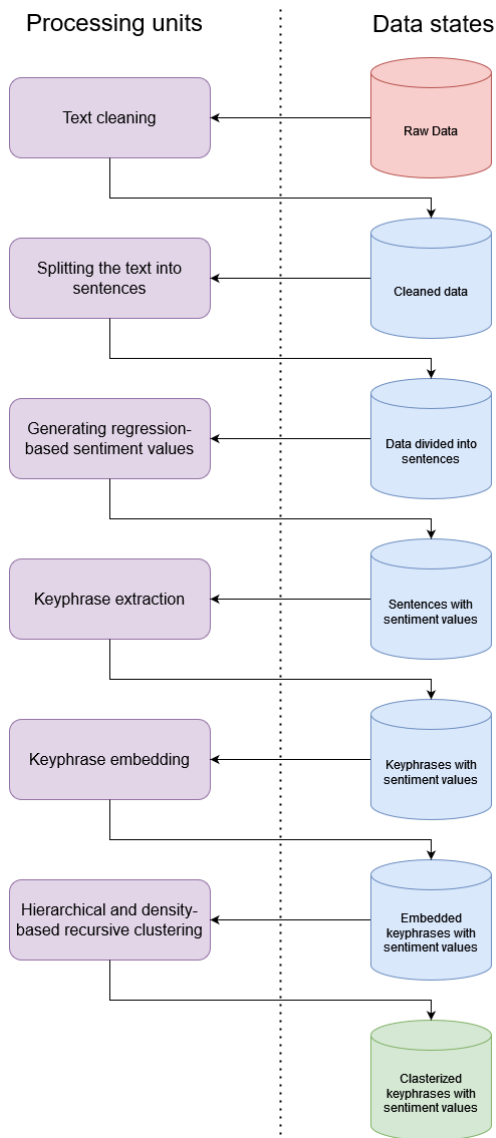
- **Szöveg tisztítása:** A folyamat során eltávolítottuk az irreleváns karaktereket és elemeket, amelyek torzíthatják a modell eredményeit. Az írásjelek és felesleges szóközök eltávolításra kerültek, hogy biztosítsuk a saját tervezésű BERT+R modell tokenizációs követelményeinek való megfelelést.
- **A szöveg felosztása mondatokra:** Tekintettel arra, hogy egy vásárlói véleményben több téma is szerepelhet ezért azokat mondatokra bontottuk. Ezáltal az egyes állításokat szét választottuk. Ez a felbontás lehetővé tette a vásárlói vélemények finomabb elemzését.

- **Érzelmek előrejelzése:** Annak érdekében, hogy javítsuk az elemzést, beépítettünk egy hangulat elemző modellt. Ehhez egy saját BERT+R modellt terveztünk. Megközelítésük az volt, hogy egy BERT [16] nagy nyelvi modellt úgy módosítsuk, hogy annak kimenete egy regressziós (R) réteg legyen. Ennek eredményeképpen pedig 1-től 5-ig terjedő skálán sokkal érzékenyebben tudtuk előrejelezni a különböző vásárlói vélemények hangulatát.
- **Kulcsszókinyerés:** A kulcsszavak kinyerése során a Sentence-BERT [17] modellt alkalmaztuk, amely lehetővé teszi a mondatok beágyazását. A mondatok beágyazásával pedig a szavak kontextusfüggő vektorjait is előállítottuk. Ezáltal a mondatok a szavainak és azok bigram és trigram kombinációból meg tudtuk határozni egy adott állítás kulcsszavait.
- **Hierarchikus és sűrűség-alapú rekurzív klaszterezés:** A beágyazott vektorok alapján létrehozott vektorterekben a közeli vektorok között a koszinusz hasonlóság formulájával mérhető egyezések vannak. Ennek eredményeként a hasonló témákat leíró szövegek sűrűbb csoportonként jelennek meg a vektorterekben. Kulcsfontosságú tulajdonsága megközelítésünknek, hogy a sűrű klaszterek kinyerése érdekében a kiugró szövegek eltávolításra kerülnek. Az azonosított sűrűbb csoportok hierarchikus klaszterezéssel azaz mélységi lefűréssel, tovább bonthatók, így ennek révén a témák szofisztikáltabb módon kerülhetnek előállításra. A rekurzív lefűrés mindaddig ismételhető amíg el nem érjük a minimális elemszámot vagy az általunk empirikus úton meghatározott 0,7-es koszinusz hasonlósági értéket.

Ahhoz, hogy megmérjük rendszerünk hatékonyságát más témamodellező rendszerekhez képest leképeztük a nem felügyelt tanulású kulcsszókinyerési módszerek működését egy felügyelt szövegkategorizálási (bináris osztályozási) feladattá. A mérési megközelítésünk alapját az LDA [18] témamodellező algoritmus fejlesztői fektették. A megközelítés lényege, hogy ha feltételezzük, hogy a modellek képesek jelentős kulcsszavakat kinyerni, akkor egy független kategorizáló modell pontosságának korrelálnia kell a kulcsszavak információtartalmával.

Ennek értékelése során a különböző alapmodellekhez (LDA, Top2Vec [19], BertTopic [20]) meghatároztuk azok legjobb beállítási paramétereit, majd a módszerek hatékonyságát összehasonlítottuk saját rendszerünkkel. Mérési eredményeink alapján az saját megoldásunk pontossága átlagosan

7%-kal multa felul az alapmodelleket az osztályozási feladatban. Sikertult tovabbá a témák szempontjából is szofisztikáltabb csoportosításokat létrehozunk. A bemutatott rendszer rávilágít a komplex architektúrák optimális tervezésének nehézségeire. Beleértve a rendszer építő elemeinek megválasztásából fakadó nehézségeket.



2.1. Ábra. A kulcskifejezés kinyerésének folyamata.

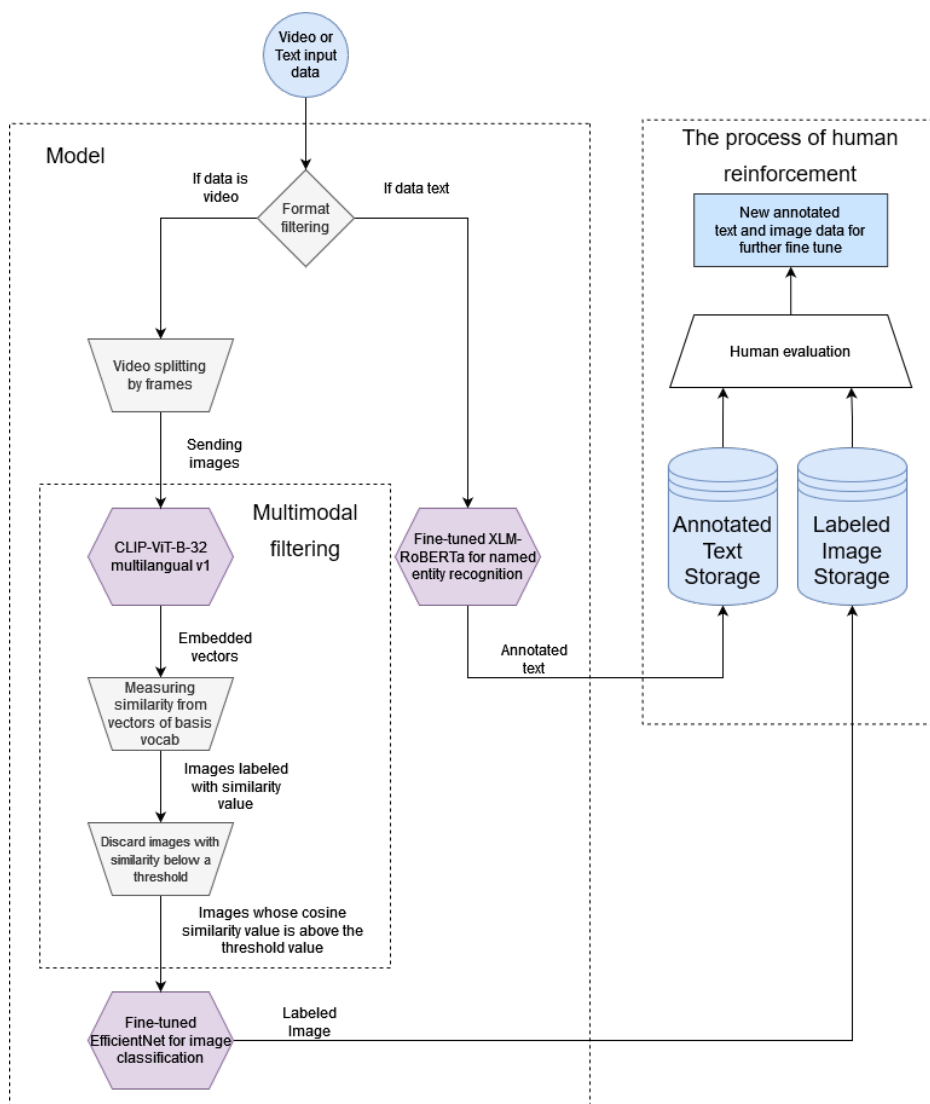
2.3. Limitált adatmennyiség esetén alkalmazható mély tanuló architektúra a dohányzás felismeréséhez

3. tézispont. *Kidolgoztam egy mély-tanuláson, generatív módszereken és humán megerősítésen alapuló integrált szöveg- és képfeldolgozó modellt, amely limitált mennyiségű tanulási adat mellett is képes a dohányzás detektálásra a szöveges és képi anyagok esetében egyaránt. A modell teljesítményét a szakterület elfogadott kiértékelési módszertanával ellenőriztem. Az eljárást más korszerű módszerekkel is összehasonlítottam, amivel bizonyítani tudtam annak hatékonyságát.*

Kapcsolódó publikáció: [P2]

Ebben a fejezetben átfogó áttekintést adunk egy új modellarchitektúra működéséről, amely képes a dohányzással kapcsolatos tartalmak szöveges és képi detektálására. Az architektúra előre betanított nyelvi és képfeldolgozó modelleket ötvözt az adathiány problémájának feloldása, valamint a hatékony és pontos észlelés érdekében.

A bemutatott rendszer amellett, hogy az előre betanított nyelvi és képfeldolgozó modellek kombinációján alapul, integrálja magába az adat-előkészítés lépéseit is. Az adat-előkészítési lépések kiemelten fontosak a gépi tanuló modellek esetében, még akkor is, ha mély-tanuló modelleket alkalmazunk. Gyakori probléma hogy, kevés tanulási adat áll rendelkezésünkre. Ez kiváltképpen igaz ez a dohányzással kapcsolatos adatokra. Mivel az ilyen vonatkozó nyilvánosan elérhető adatkészletek ritkák. A dohányzással kapcsolatos tartalmak felismerésével kapcsolatban végzett kutatómunkánk ideje alatt nem állt rendelkezésre olyan nyilvánosan elérhető szövegekből vagy képekből álló adatkészlet, amelynek mérete lehetővé tette volna, hogy egy modellt véletlen súlykészletből kiindulva hatékonyan betanítsunk.



2.2. Ábra. A dohányzást felismerő architektúra sematikus leírása.

Architektúránk módszertana (lásd a 2.2 ábrát) a következő lépéseket tartalmazza:

1. Válasszuk el a szövegek és a videók detektálását.
2. A videófelvételeket bontsuk képkockákra.

3. A képkockákat szűrjük meg egy előképzett multimodális modell segítségével.
4. Hangoljunk egy előképzett képosztályozót egy limitált méretű de nyilvánosan elérhető dohányzással kapcsolatos adatbázis képeire.
5. A szűrt képeket osztályozzuk tovább a finomhangolt képfeldolgozó modell segítségével.
6. Generáljunk szintetikus szöveges adatokat.
7. Hangoljunk egy nevesített entitások felismerésének feladatára kialakított nyelvi modellt a generált szöveges adatokra.

A fenti szempontok alapján elkészült architektúra elemei a következők:

- **Formátumellenőrzés:** A kezdeti lépésben meg kell határozni a bejövő adatok formátumát, megkülönböztetve a videófelveteleket és a szöveges adatokat.
- **Videók és képek feldolgozása:** A rendszer a videófelveteleket képkockákra bontja, másodpercenkénti mintavételezéssel.
- **Multimodális szűrés:**
 - További információt nyerünk ki azzal, hogy a CLIP-ViT-B-32 [21, 22] modellel beágyazott vektorokat generálunk a képekből. Mivel a CLIP egy multimodális modell ezért egyszerre integrálja a képekből és szövegből kinyerhető információkat, így robusztus általánosítási képességgel rendelkezik. A CLIP képes szövegekhez és képekhez is beágyazott vektorokat generálni, aminek előnye, hogy a generált képekből képzett vektorok összehasonlíthatóak egy szövegvektorral, ami jelen esetünkben a „*dohányzás*” kifejezés adott.
 - Mivel minden egyes képvektor koszinusz hasonlósága mérhetővé válik a szövegvektorhoz képest ezért a képek és a szövegek viszonyából olyan karakterisztikus függvény adódik, ami megmutatja a videó felvétel képkockáinak hasonlóságát a dohányzást leíró kifejezéshez képest.

- A leíró függvény megmutatja, hogy meghatározható olyan koszinusz hasonlóságon alapuló küszöbérték, ami megadja a dohányzás kifejezésből képzett szövegvektor szempontjából nem releváns képkockákat.

- **Finomhangolt képosztályozás:** A szűrést követően a fennmaradó képeket előképzett és dohányzásra finomhangolt képfeldolgozó modell (EfficientNet [23], Inception [24], ResNet [25], VGG [26] vagy Xception [27]) segítségével tovább osztályozzuk. A finomhangolásra ugyan limitált mennyiségű adat állt rendelkezésre, de a multimodális szűrő bevezetésével a finomhangolt és CLIP modell közösen egy Boosting típusú összekapcsolást alkot, amelynek hatékonyabb működését mérésekkel támasztottuk alá.
- **Szövegfeldolgozás:** A szöveges adatok dohányzással kapcsolatos magyar kifejezéseinek detektálásához a NER technikáját alkalmaztunk. Ehhez magyar nyelvű szinonima szótár és a ChatGPT [28] generatív nyelvi modellen alapuló szolgáltatása segítségével szintetikus szöveges adatokat generáltunk. A szintetikus adatokat így automatikusan annotáltuk, majd tanulóadatokat képeztünk belőlük a NER modell számára.

A különböző modellekből felépülő komplex multimodális architektúránk a képeknél 74%-os, a szövegnél 98%-os pontosságot tudott elérni. Eredményeink rávilágítanak arra, hogy a mesterséges intelligencia, különösen a mély tanulás nagy lehetőségeket rejt magában a rejtett reklámok felderítésére, és lehetővé teszi a dohányzással kapcsolatos médiatartalmak elfogulatlan, reprodukálható és tisztességes számszerűsítését. Ez támogathatja az esetleges szabályzó intézkedések hatékony meghozatalát is. Javasoltunk egy mélytanuláson, generatív módszereken alapuló integrált szöveg- és képfeldolgozó modellt, amely limitált mennyiségű tanuló adat mellett is képes a dohányzás detektálásra szöveges és képi anyagok esetében egyaránt. A kialakított megoldásunkkal olyan kérdésekre válaszolunk, amelyek érintik mind a megfelelő betanítási módszerek kiválasztását, mind a komplex modell architektúrák és hiperparamétereik optimalizálását egyaránt.

2.4. Domaintranszfer technikák vizsgálata LLM-vezérelt tudásalapú rendszerek fejlesztéséhez

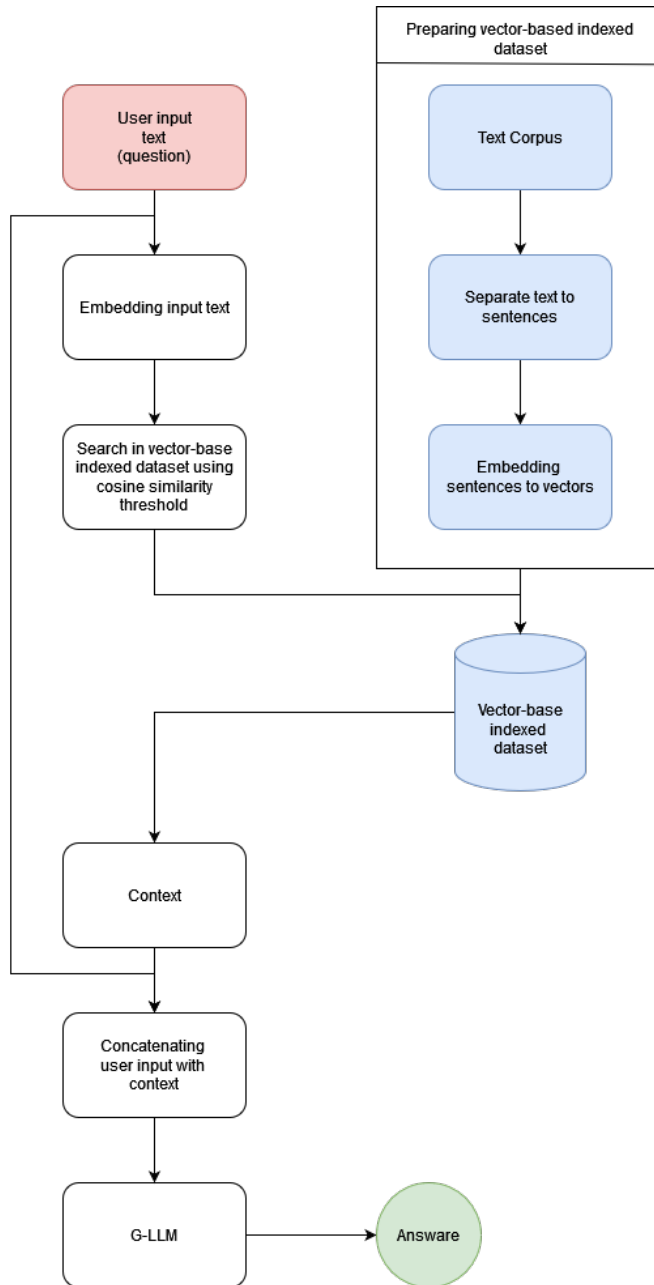
4. tézispont *A szakterület elfogadott kiértékelési módszertanával, valamint publikusan elérhető és általunk készített adatbázisok alkalmazásával empirikusan igazoltam, hogy a RAG a hatékonyabb megközelítés, ha a generatív nagy nyelvi modellek (LLM) alapú, új típusú, mesterséges intelligencia által vezérelt tudásbázist szeretnénk felépíteni. Továbbá megmutattam, hogy a finomhangolás (DFT) és a Retrieval-Augmented Generation (RAG) integrációja milyen új típusú kihívásokat jelent.*

Kapcsolódó publikáció: [P1]

Az LLM-ek jelentős áttörést értek el az NLP területén. A LLM-ek közül például a GPT [4], a PaLM [29] és a LLaMA [30] figyelemre méltó képességeket mutattak a természetes nyelvű szövegek generálásban, és ezzel párhuzamosan aktívabb tudásbázis rendszerek létrehozásában. Ebben a fejezetben átfogó elemzést mutatunk be az LLM alapú tudásbázis építési stratégiákról és azok hatékonyságáról.

Az DFT és a RAG hatékonyságának értékeléséhez a következő módszertant dolgoztuk ki:

- **Modellválasztás:** A GPT-J-6B [31], OPT6.7B [32], LLaMA-7B [30] és LLaMA2-7B [30] modelleket előre meghatározott kritériumok alapján választottuk ki.
- **Mérőszámok:** Az modellek kiértékelését BLEU [33], ROUGE [34], METEOR [35] pontszámok és koszinusz hasonlóság alapján végeztük.
- **Adatelőkészítés:** Tudományos publikációkból és nyilvánosan elérhető forrásokból gyűjtöttük össze az adatsorokat, biztosítva a reprodukálhatóságot és az ellenőrizhetőséget.
- **Finomhangolás:** A modellek finomhangolása tartományspecifikus adatkészleteken történt.



2.3. Ábra. A mondatok vektoriális beágyazásán alapuló keresőmotort használó RAG-modell sematikus leírása.

Módszertanunk alapján elvégzett kísérleteink eredményei alapján a következő megállapításokat tehetjük:

- A célfeladatra történő DFT és a RAG alkalmazása egyaránt javítja az egyes modellek teljesítményét.
- A RAG jobb teljesítményt mutatott a DFT-hez képest a BLEU és ROUGE pontszámok valamint a koszinusz hasonlóság tekintetében. Bár a DFT modellek jobb METEOR pontszámot mutattak, hajlamosabbak voltak a hallucinációra a RAG-alapú rendszerekhez képest.
- A DFT és a RAG kombinálása nem eredményezett teljesítményjavulást.

Mérési eredményeinket alapjául véve felvázoltunk egy egyszerű RAG alapú architektúrát (lásd 2.3 ábra).

Megközelítésünk átlagosan 17%-kal felülmúlja a DFT modelleket a ROUGE pontszám tekintetében, 13%-kal a BLEU pontszám esetén és 36%-kal a koszinusz hasonlóság alapján. A RAG jelentős előnnyel rendelkezik a finomhangolással szemben a hallucináció tekintetében, amit nem ellensúlyoz az a tény, hogy az DFT modellek átlagosan 8%-kal jobb METEOR pontszáma nagyobb kreativitást jelez a RAG-hoz képest. Ezzel megmutattuk, hogyan adhatunk meg olyan elsősorban mérőszámokra alapuló módszertant, amely meghatározza a megfelelő betanítási módszert és ezáltal a végleges egy hatékony architektúra felépítését.

2.5. Elméleti modellek optimalizálása

5. tézispont *Az értekezés utolsó fejezete rávilágít az előképzés előnyére, és bemutatja az egyszerűbb elméleti funkciók lehetőségét a neurális hálózatok inicializálásában. Ennek demonstrálására a fejezet klasszikus eljárásokon alapuló megközelítést mutat be optimalizált SIR [36] elméleti modellek generálására, amelyek azután felhasználhatók egy neurális hálózat előképzésére.*

Kapcsolódó publikáció: [P4]

A SIR elméleti modell optimalizálását a legkisebb négyzetek görbe illesztési problémájára vezettük vissza. Felvázoltunk egy megközelítést, amelyben a Levenberg-Marquardt algoritmust (LMA) [37] [38] és differenciális evolúciós (DEA) [39] algoritmust használtuk a SIR modell optimalizálására.

Tapasztalataink azt mutatják, hogy mindkét módszer jól alkalmazható a görbeillesztési problémára, és felhasználható a tényleges adatokhoz legjobban illeszkedő optimális SIR modell megtalálásához. Az LMA különösen előnyös a sebessége miatt, és a DEA hatékonyan finomítja a meglévő, közel optimális paramétereket. Ezért komplex optimalizálást javasoltunk, hogy megtaláljuk a problémát legjobban leíró elméleti modellt. Az LMA az első lépésben gyorsan megtalálja a legjobb kezdeti paramétereket, míg a DEA a második lépésben finomítja azokat. Az így előállított optimális SIR modellek hatékonyabban közelíthetők [40] [P4] neurális hálókkal és ideálisak azok előképzéséhez.

Hivatkozások

- [1] D. P. Kingma and J. Ba, „Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, „Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [3] I. Goodfellow, „Deep learning,” 2016.
- [4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, „Improving language understanding by generative pre-training,” 2018.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, „Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [6] J. Bergstra and Y. Bengio, „Random search for hyper-parameter optimization,” *Journal of machine learning research*, vol. 13, no. 2, 2012.
- [7] J. Snoek, H. Larochelle, and R. P. Adams, „Practical bayesian optimization of machine learning algorithms,” *Advances in neural information processing systems*, vol. 25, 2012.
- [8] A. E. Eiben and J. E. Smith, *Introduction to evolutionary computing*. Springer, 2015.
- [9] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.
- [10] Twitter, „Twitter,” 2006. Online platform for social networking and microblogging.

- [11] C. Hutto and E. Gilbert, „Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the international AAAI conference on web and social media*, vol. 8, pp. 216–225, 2014.
- [12] M. Heitmann, C. Siebert, J. Hartmann, and C. Schamp, *More than a feeling: Benchmarks for sentiment analysis accuracy*. Ssrn, 2020.
- [13] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, „Tweeteval: Unified benchmark and comparative evaluation for tweet classification,” 2020.
- [14] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, „On the properties of neural machine translation: Encoder-decoder approaches,” *arXiv preprint arXiv:1409.1259*, 2014.
- [15] S. Hochreiter and J. Schmidhuber, „Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, „Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [17] N. Reimers and I. Gurevych, „Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [18] J. K. Pritchard, M. Stephens, and P. Donnelly, „Inference of population structure using multilocus genotype data,” *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.
- [19] D. Angelov, „Top2vec: Distributed representations of topics,” *arXiv preprint arXiv:2008.09470*, 2020.
- [20] M. Grootendorst, „Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics,” 2020.
- [21] N. Reimers and I. Gurevych, „Making monolingual sentence embeddings multilingual using knowledge distillation,” *arXiv preprint arXiv:2004.09813*, 2020.
- [22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, „Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

- [23] M. Tan and Q. Le, „Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.
- [24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, „Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, „Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [26] K. Simonyan and A. Zisserman, „Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [27] F. Chollet, „Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.
- [28] OpenAI, „Chatgpt,” 2024. Accessed on March 12, 2024.
- [29] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, „Palm: Scaling language modeling with pathways,” 2022.
- [30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, „Llama: Open and efficient foundation language models,” 2023.
- [31] B. Wang and A. Komatsuzaki, „GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.” <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.

- [32] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, „Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, „Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (P. Isabelle, E. Charniak, and D. Lin, eds.), (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.
- [34] C.-Y. Lin, „ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.
- [35] S. Banerjee and A. Lavie, „METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, June 2005.
- [36] T. Harko, F. S. Lobo, and M. Mak, „Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates,” *Applied Mathematics and Computation*, vol. 236, pp. 184–194, 2014.
- [37] K. Levenberg, „A method for the solution of certain non-linear problems in least squares,” *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168, 1944.
- [38] D. W. Marquardt, „An algorithm for least-squares estimation of non-linear parameters,” *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.
- [39] R. Storn, „On the usage of differential evolution for function optimization,” in *Proceedings of North American Fuzzy Information Processing*, pp. 519–523, IEEE, 1996.
- [40] G. Bogacsovics, A. Hajdu, R. Lakatos, M. Beregi-Kovács, A. Tiba, and H. Tomán, „Replacing the sir epidemic model with a neural network and

training it further to increase prediction accuracy,” in *Annales Mathematicae et Informaticae*, vol. 53, pp. 73–91, Eszterházy Károly Egyetem Líceum Kiadó, 2021.

Publikációs lista

Folyóirat publikációk Angol nyelven

- [P1] Lakatos, R., Pollner, P., Hajdu, A., and Joó, T. (2025). Investigating the Performance of Retrieval-Augmented Generation and Domain-Specific Fine-Tuning for the Development of AI-Driven Knowledge-Based Systems. *Machine Learning and Knowledge Extraction*, 7(1), 15. <https://doi.org/10.3390/make7010015>
- [P2] Róbert Lakatos , Péter Pollner, András Hajdu, and Tamás Joó. 2024. A multimodal deep learning architecture for smoking detection with a small data approach. In *Frontiers in Artificial Intelligence* (Vol. 7). Frontiers Media SA. <https://doi.org/10.3389/frai.2024.1326050>
- [P3] Róbert Lakatos, Gergő Bogacsovics, Balázs Harangi, István Lakatos, Attila Tiba, János Tóth, Marianna Szabó, and András Hajdu. 2024. A Machine Learning-Based Pipeline for the Extraction of Insights from Customer Reviews. *Big Data and Cognitive Computing* 8, no. 3: 20. <https://doi.org/10.3390/bdcc8030020>
- [P4] Gergo Bogacsovics, András Hajdu, Róbert Lakatos, Marcell Beregi-Kovács, Attila Tiba, Henrietta Tomán. *Replacing the SIR epidemic model with a neural network and training it further to increase prediction accuracy*. In: *Annales Mathematicae et Informaticae*. Eszterházy Károly Egyetem Líceum Kiadó, 2021. p. 73-91. doi.org/10.33039/ami.2021.02.003

Folyóirat publikációk Magyar nyelven

- [P5] Nagy János, Hajdu András, Bogacsovics Gergő, Bojtor Csaba, Illés Árpád, Lakatos Róbert, Mészáros László Precíziós gazdálkodásban használható adatelemzés alapú növénytermesztési döntéstámogató rendszer fejlesztés In: Pőr, Csilla; Szabó-Soós, Adrienn; Szabó, Péter (szerk.) LXV. Georgikon Napok Tudományos Konferencia [65th Georgikon Days Scientific Conference] Keszthely, 2024. május 17–18. : Kivonatkiötet Keszthely, Magyarország : Magyar Agrár- és Élettudományi Egyetem, Georgikon Campus (2024) 206 p. pp. 133-134., 2 p.
- [P6] Bogacsovics Gergo, Hajdu András, Harangi Balázs, Lakatos István, Lakatos Róbert, Szabó Marianna, Tiba Attila, Tóth János. Napelemfarmok Magyarország területén történő elhelyezését segítő döntéstámogató rendszer fejlesztése. KözigazgatásTudomány, 2021, 1.2: 134-145. doi.org/10.54200/kt.v1i2.23
- [P7] Bogacsovics Gergo, Hajdu András, Harangi Balázs, Lakatos István, Lakatos Róbert, Szabó Marianna, Tiba Attila, Tóth János, Tarcsi Ádám. Adatelemzési folyamat és keretrendszer a közigazgatás számára. KözigazgatásTudomány, 2021, 1.2: 146-158. doi.org/10.54200/kt.v1i2.24
- [P8] Bátfai, M., Lakatos, R., Molnár, G., Takácsné Bubnó, K. BPMN 2.0 a könyvtári minőségbiztosításban, Tudományos és Műszaki Tájékoztatás, 63(8), p. 305–316, 2016.

Konferencia publikációk Angol nyelven

- [P9] Róbert Lakatos, Eszter Klára Urbán, Zoltán János Szabó, János Pozsga, Eszter Csernai and András Hajdu. 2024. Designing Prompts and Creating Cleaned Scientific Text for Retrieval Augmented Generation for More Precise Responses from Generative Large Language Models. In 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) (pp. 1-6). IEEE. 10.1109/CITDS62610.2024.10791382
- [P10] Árpád Pandya, Róbert Lakatos Péter Pollner, Eszter Csernai and András Hajdu. 2024. Investigating the Influence of Hyperparameters on the Optimal Time-Series Prediction Ability of Generative Large Language Models. In 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) (pp. 1-6). IEEE. 10.1109/CITDS62610.2024.10791385
- [P11] Gergő Bogacsovics, Balázs Harangi, Marcell Beregi-Kovács, Dávid Kupás, Róbert Lakatos, Norbert Dávid Serbán, Tiba Attila, János Tóth. 2024. Assessing Conventional and Deep Learning-Based Approaches for Named Entity Recognition in Unstructured Hungarian Medical Reports. Presented at the 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI). <https://doi.org/10.1109/sami60510.2024.10432795>
- [P12] Róbert Lakatos, Gergo Bogacsovics, András Hajdu. Predicting the direction of the oil price trend using sentiment analysis. In: 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS). IEEE, 2022. p. 177-182. doi.org/10.1109/CITDS54976.2022.9914158



Nyilvántartási szám: DEENK/140/2025.PL
Tárgy: PhD Publikációs Lista

Jelölt: Lakatos Róbert
Doktori Iskola: Informatikai Tudományok Doktori Iskola
MTMT azonosító: 10079387

A PhD értekezés alapjául szolgáló közlemények

Idegen nyelvű tudományos közlemények hazai folyóiratban (1)

1. Bogacsovics, G., Hajdu, A., **Lakatos, R.**, Beregi-Kovács, M., Tiba, A., Tomán, H.: Replacing the SIR epidemic model with a neural network and training it further to increase prediction accuracy.
Ann. Math. Inform. 53, 73-91, 2021. ISSN: 1787-5021.
DOI: <http://dx.doi.org/10.33039/ami.2021.02.003>

Idegen nyelvű tudományos közlemények külföldi folyóiratban (3)

2. **Lakatos, R.**, Pollner, P., Hajdu, A., Joó, T.: Investigating the Performance of Retrieval-Augmented Generation and Domain-Specific Fine-Tuning for the Development of AI-Driven Knowledge-Based Systems.
Mach. Learn. Knowl. Extr. 7 (1), 1-18, 2025. EISSN: 2504-4990.
DOI: <http://dx.doi.org/doi.org/10.3390/make7010015>
IF: 4 (2023)
3. **Lakatos, R.**, Bogacsovics, G., Harangi, B., Lakatos, I., Tiba, A., Tóth, J., Szabó, M., Hajdu, A.: A Machine Learning-Based Pipeline for the Extraction of Insights from Customer Reviews.
Big Data Cogn. Comput. 8 (3), 1-24, 2024. EISSN: 2504-2289.
DOI: <http://dx.doi.org/10.3390/bdcc8030020>
IF: 3.7 (2023)
4. **Lakatos, R.**, Pollner, P., Hajdu, A., Joó, T.: A multimodal deep learning architecture for smoking detection with a small data approach.
Front. Artif. Intell. 7, 1-8, 2024. EISSN: 2624-8212.
DOI: <http://dx.doi.org/10.3389/frai.2024.1326050>
IF: 3 (2023)





Idegen nyelvű konferencia közlemények (1)

5. **Lakatos, R.**, Bogacsovics, G., Hajdu, A.: Predicting the direction of the oil price trend using sentiment analysis.
In: IEEE 2nd Conference on Information Technology and Data Science (CITDS) :
Proceedings. Ed.: Fazekas István, Institute of Electrical and Electronics Engineers (IEEE),
Piscataway, 177-182, 2022.

További közlemények

Magyar nyelvű tudományos közlemények hazai folyóiratban (3)

6. Bogacsovics, G., Hajdu, A., Harangi, B., Lakatos, I., **Lakatos, R.**, Szabó, M., Tiba, A., Tóth, J.,
Tarcsi, Á.: Adatelemzési folyamat és keretrendszer a közigazgatás számára.
Közigazgatástudomány. 1 (2), 146-158, 2021. ISSN: 2786-1910.
DOI: <http://dx.doi.org/10.54200/kt.v1i2.24>
7. Bogacsovics, G., Hajdu, A., Harangi, B., Lakatos, I., **Lakatos, R.**, Szabó, M., Tiba, A., Tóth, J.:
Napelemfarmok Magyarország területén történő elhelyezését segítő döntéstámogató
rendszer fejlesztése.
Közigazgatástudomány. 1 (2), 134-145, 2021. ISSN: 2786-1910.
DOI: <http://dx.doi.org/10.54200/kt.v1i2.23>
8. Soós, M. E., **Lakatos, R.**, Molnár, G., Takácsné Bubnó, K.: BPMN 2.0 a könyvtári
minőségbiztosításban.
Tud. Műsz. Tájs. 63 (8), 305-316, 2016. ISSN: 0041-3917.

Magyar nyelvű konferencia közlemények (1)

9. Nagy, J., Hajdu, A., Bogacsovics, G., Bojtor, C., Illés, Á., **Lakatos, R.**, Mészáros, L.: Precíziós
gazdálkodásban használható adatelemzés alapú növénytermesztési döntéstámogató
rendszer fejlesztés.
In: LXV. Georgikon Napok Tudományos Konferencia = 65th Georgikon Days Scientific
Conference /szerk. Pőr Csilla, Szabó-Soós Adrienn, Szabó Péter, Magyar Agrár- és
Élettudományi Egyetem Georgikon Campus, Keszthely, 133-134, 2024. ISBN:
9786156338105





Idegen nyelvű konferencia közlemények (3)

10. Bogacsovics, G., Harangi, B., Beregi-Kovács, M., Kupás, D., **Lakatos, R.**, Serbán, N. D., Tiba, A., Tóth, J.: Assessing Conventional and Deep Learning-Based Approaches for Named Entity Recognition in Unstructured Hungarian Medical Reports.
In: 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI). Ed.: Kovács Levente, Liberios Vokorokos, IEEE, Piscataway, 77-82, 2024. ISBN: 9798350317206
11. **Lakatos, R.**, Urbán, E. K., Szabó, Z. J., Pozsga, J., Csernai, E., Hajdu, A.: Designing Prompts and Creating Cleaned Scientific Text for Retrieval Augmented Generation for More Precise Responses from Generative Large Language Models.
In: 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) / Hajdu Andras, Institute of Electrical and Electronics Engineers, Piscataway (NJ), 1-6, 2024.
12. Pándy, Á., **Lakatos, R.**, Pollner, P., Csernai, E., Hajdu, A.: Investigating the Influence of Hyperparameters on the Optimal Time-Series Prediction Ability of Generative Large Language Models.
In: 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS), Institute of Electrical and Electronics Engineers (IEEE), Piscataway, 158-163, 2024. ISBN: 9798350387889

Magyar nyelvű absztrakt kiadványok (1)

13. Soós, M. E., **Lakatos, R.**, Molnár, G., Takácsné Bubnó, K.: BPMN 2.0 a könyvtári minőségbiztosításban.
In: Networkshop 2016. Szerk.: Mikusné Sárvári Klára, NIIF, Budapest, [8], 2016. ISBN: 9789638833563

A közlő folyóiratok összesített impakt faktora: 10,7

**A közlő folyóiratok összesített impakt faktora (az értekezés alapjául szolgáló közleményekre):
10,7**

A DEENK a Jelölt által a Tudóstérbe feltöltött adatok bibliográfiai és tudománymetriai ellenőrzését a tudományos adatbázisok és a Journal Citation Reports Impact Factor lista alapján elvégezte.

Debrecen, 2025.04.08.



Short thesis for the degree of doctor of philosophy (PhD)

Design and Optimization of Machine Learning Models

by Róbert Lakatos

Supervisor: Dr. András Hajdu



UNIVERSITY OF DEBRECEN
Doctoral School of Informatics

Debrecen, 2025

Contents

1	Backgorund	4
2	The main results of the thesis	6
2.1	Predicting the direction of the oil price trend using sentiment analysis	6
2.2	A Machine Learning-based Pipeline for the Extraction of Insights from Customer Reviews	9
2.3	A multimodal deep learning architecture for smoking detection with a small data approach	12
2.4	Investigating the domain transfer techniques for the development of LLM-driven knowledge-based systems	16
2.5	Optimization of theoretical models	19
	References	20
	List of Publications	25

Backgorund

Artificial intelligence research has undergone a transformative evolution, primarily driven by the imperative to enhance the efficiency and performance of machine learning models. However, the true potential of these models hinges on their optimization.

Optimization is a multifaceted concept. On the one hand, it involves algorithms designed to fine-tune models, and on the other hand, it encompasses strategies aimed at improving model effectiveness. For example, the Adam optimizer, introduced by Kingma and Ba [1], has become widely popular for training neural networks due to its adaptive learning rates and ability to handle noisy gradients efficiently. Similarly, in generative modeling tasks, specific optimization strategies used in generative adversarial networks have been crucial [2].

The optimization process inherently depends on the model type and the task's nature, with neural network-based architectures susceptible to specific optimization strategies. Therefore, in this context, the choice of optimization technique is shaped by the target task and the quality of the available data.

Critical questions that drive the optimization of machine learning models include:

- What is the most effective training method?
- How can hyperparameters be fine-tuned for peak performance?
- How can complex model architectures be optimally designed?

The nature of the problem at hand typically determines training methodologies. Machine learning tasks are commonly categorized as supervised, semi-supervised, or unsupervised learning [3], requiring distinct optimization strategies.

In supervised learning, the availability of labeled data guides the optimization process. Semi-supervised learning, in contrast, utilizes strategies that leverage labeled and unlabeled data, while unsupervised learning aims to discover hidden patterns in the data.

Additionally, advanced training strategies such as transfer learning, fine-tuning [3], and pre-training have also become essential [4], and these came into a particular massive focus with the appearance of large language models (LLMs). These models, built on transformer architectures [5], are trained on extensive datasets and have performed remarkably in machine translation, text summarization, and question-answering tasks. Research has shown that LLMs pre-trained for general tasks and fine-tuned for specific tasks outperform models trained solely on task-specific data [4]. This approach improves task-specific performance and enhances the model’s generalization capabilities.

Optimizing hyperparameters is crucial for the success of machine learning models. Standard methods for hyperparameter tuning include Grid Search, Random Search [6], Bayesian optimization [7], and genetic algorithms [8]. Additionally, ensemble techniques such as Bagging, Boosting, and Stacking [9] provide robust solutions for optimizing the architecture of complex models.

In the next chapters, we briefly review the main results of the thesis concerning the five developed methods. Chapter 2.1 explores the development of an ensemble solution based on LLMs and recurrent neural networks, enabling forecasting of oil price movements. By applying textual data, sentiment-based LLMs, and time series prediction models, I demonstrate the potential of combining diverse data sources to build a forecasting model. In Chapter 2.2, I present a Pipeline based on an ensemble of machine learning and classical methods. Our measurement results prove that it provides a more sophisticated and uniform solution for extracting relevant statements from customer reviews than the currently popular methods. Chapter 2.3 focuses on the performance potential of multimodal architectures, specifically in detecting smoking behaviors from text and images. I present a model that performs well, even with limited training data. In Chapter 2.4, I compare the performance of RAG and domain-specific fine-tuning and show that RAG has an advantage in developing new knowledge base systems based on LLM. Finally, in Chapter 2.5, I show how to trace the optimization of theoretical models such as SIR back to applying algorithms that solve least squares curve fitting problems. This allows us to produce optimal theoretical models suitable for neural network pre-training.

The main results of the thesis

2.1 Predicting the direction of the oil price trend using sentiment analysis

Thesis 1. *I developed a machine learning architecture that considers the sentiment of news written in natural human language to predict the price of oil. To determine the sentiment of the news, I used a classical statistics-based algorithm and large language models. In doing so, I created a data source in which sentiment values showed a high accuracy correlation of 84.08% with the oil price's natural direction. This one improved the predictive ability of the LSTM and GRU models. Based on these results, we demonstrated that sentiment analysis has the potential to support the prediction of oil price movements and, in turn, other stock market products. I provided that the temporal variability of news sources and noise reduction techniques are considered.*

Related publication: [P12]

The method presented in this thesis chapter explores the relationship between textual information and oil price movements. The chapter analyzes the correlation between the sentiments expressed by news on the social media platform Twitter (renamed X) [10] and the directional changes in oil price, one of the most critical energy carriers, using NLP techniques and machine and deep learning methods.

Both text and price data had to be pre-processed to improve the forecast. Therefore, the following operations were applied to the price data:

- The moving average with 10 window sizes was used to filter out noisy price changes.
- A single oil price indicator was created by analyzing the correlation between *Brent* and *Crude* oil prices.
- The price movements were transformed from -1 (decrease) to +1 (increase) according to their direction from the previous day.

For text news data, the following operations were applied:

- We filtered out irrelevant news and kept only those that mentioned the word "*oil*".
- Special characters or parts of the text, such as URL links, were removed.
- Sentiment analysis was used to detect the emotional state of each news item.
- The news was aggregated by day, and their sentiment values were averaged accordingly.

To detect sentiment, we used the Vader [11], Sentiment Roberta Large English [12], and Twitter Roberta Base Sentiment [13] sentiment models. We evaluated the effectiveness of sentiment analysis in predicting oil price movements based on correlation with actual price changes. For this purpose, sentiment was mapped to the direction of price changes. The sentiment models were trained from negative news with -1 to positive news with +1, which is then mapped to the direction of price movements. As a result, we could determine their accuracy relative to the direction of the original price movements.

Recognizing the temporal variability of news sources, monthly analyses were carried out to identify the optimal sources for each period. This approach created a more versatile and accurate dataset for training the models. In addition, we built a noise dictionary to identify and filter out noisy news that could distort sentiment analysis.

In the following, sentiment analysis data of news items was managed as time series data. On which we were able to fit models suitable for time-series

predictions. We used GRU [14] and LSTM [15] recurrent neural networks for time series forecasts, with their outputs fitted to the direction of price movements. This was achieved using a tangent hyperbolic activation function to keep the forecasts within the desired $(-1, +1)$ range. We used Adam’s optimization algorithm and Early Stopping method to find the best weights, and we measured the performance of the GRU and LSTM with Unit parameters of 5, 15, 20, 25, 30, 35, 40, 45, and 50.

In the following, data generated by sentiment analysis of news items were managed as time series data. On which we were able to fit models suitable for time-series predictions. For time series forecasts, GRU [14] and LSTM [15] recurrent neural networks were used, with their outputs fitted to the direction of price movements. This was achieved by using a tangent hyperbolic activation function to keep the forecasts within the desired $(-1, +1)$ range. We used Adam’s optimization algorithm and his Early Stopping method to find the best weights. The performance of the GRU and LSTM was also measured with Unit parameters of 5, 15, 20, 25, 30, 35, 40, 45 and 50.

Presented by our solution adopts a data-centric approach to building a six-tier architecture, which has important implications for both the choice of the appropriate paralysis method and the optimal complex architecture construction.

2.2 A Machine Learning-based Pipeline for the Extraction of Insights from Customer Reviews

Thesis 2. *I developed a framework for extracting relevant information from customer reviews. To create the framework, I examined the best solutions identified by the field of expertise in information extraction science. I then compared them with my solution according to the measurement methods accepted in the field of expertise. I supported the benefits of my solution with experimental results. Furthermore, I wrapped the framework in a sequential data-oriented process for efficient operation and transparent structure.*

Related publication: [P3]

In this chapter, we describe in detail the method we have developed to extract information from textual data in online commerce. We describe our evaluation strategy and the different keyword extraction techniques, the most effective of which eventually led us to design a unified system (pipeline).

We investigated three popular keyword and key phrase extraction approaches n-gram, dependency analysis, and semantic embedding. As a result of our investigation, we integrated semantic embedding into our system (see Figure [1]) for information extraction.

Our designed pipeline is focused on scalability, adaptability, and stability and includes the following steps:

- **Text cleaning:** The pipeline removes irrelevant characters and elements that could distort the model results. Moreover, removing punctuation and unnecessary spaces ensures compliance with the tokenization requirements of our designed BERT+R model.
- **Splitting the text into sentences:** Given that a customer review may contain several topics, it has been split into sentences. In this way, each statement is separated. This split allowed a more refined analysis of the customer reviews.
- **Sentiment prediction:** We designed a proprietary BERT+R model to improve sentiment estimation. Their approach was to fine-tune a

BERT [16] high-level language model so that its output is a regression layer. As a result, we could predict more sensitively the sentiment of different customer opinions on a scale of 1 to 5.

- **Keyword extraction:** The Sentence-BERT [17] model was used to extract keywords, allowing sentence embedding. By embedding the sentences, we also generated context-dependent vectors of the words. This allowed us to determine the keywords of a given statement from the words of the sentences and their bigram and trigram combinations.
- **Hierarchical and density-based recursive clustering:** After cleaning, sentiment analysis, and extraction of keywords from the texts, we applied our hierarchical and density-based recursive clustering approach to topic retrieval. Similarities between nearby vectors are measured using a cosine formula in the vector spaces generated from the nested vectors. As a result, texts describing similar topics appear as denser clusters in the vector spaces. A key feature of our approach is that outliers are removed to extract dense clusters. Furthermore, hierarchical clustering can further decompose the identified denser clusters, i.e., depth drilling. That is, it can generate topics in a more sophisticated way. The recursive drilling can be repeated until a minimum number of elements or a cosine similarity value of 0.7, which we have determined empirically, is reached.

To measure the performance of our system compared to other topic modeling systems, we mapped the unsupervised learning keyword extraction methods into a supervised text categorization (binary classification) task. The developers of the LDA laid the basis of our measurement approach [18] in their topic modeling algorithm. They also used this method to measure the effectiveness of LDA. The essence of the approach is that if we assume that the models can extract meaningful keywords, then the accuracy of an independent categorization model should correlate with the information content of the keywords. To evaluate this, we determined the best-fitting parameters for the different basic models (LDA, Top2Vec [19], BertTopic [20]) and compared the efficiency of the methods with our system. Based on our measurement results, the accuracy of our solution outperformed the base models by an average of 7% in the classification task. We also found that it produced more sophisticated clustering in terms of topics.

The presented pipeline highlights the problems of designing complex architectures, particularly the difficulties of choosing the pipeline’s building blocks.

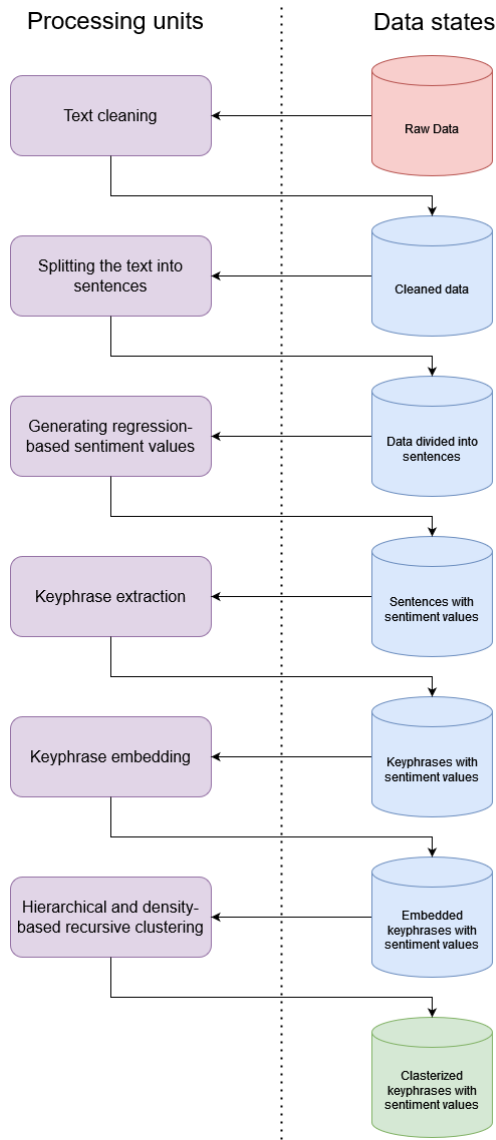


Figure 2.1: Flow diagram of key phrase extraction pipeline.

2.3 A multimodal deep learning architecture for smoking detection with a small data approach

Thesis 3. *I have created an integrated text and image processing model based on deep learning, generative methods, and human reinforcement. With limited training data, this model can also detect smoking in text and image material. The model's performance was verified using the field's accepted evaluation methodology. I also compared the method with other state-of-the-art methods to demonstrate its effectiveness.*

Related publication: [P2]

This chapter comprehensively overviews a proposed model architecture that detects smoking-related content in text and images.

The architecture combines pre-trained linguistic and image processing models to solve limited data problems and provide efficient and accurate detection.

The presented system is based on a combination of pre-trained language and image processing models and integrates the data preparation steps. This is because the data preparation steps are essential for machine learning models, even when deep learning models are used. This is particularly true when little training data is available, which is particularly true for smoking data. Indeed, publicly available datasets on smoking are scarce.

In support of this, during our research on detecting smoking-related content, there was no publicly available dataset of texts or images of a size that would have allowed us to train a model efficiently starting from a random weight set.

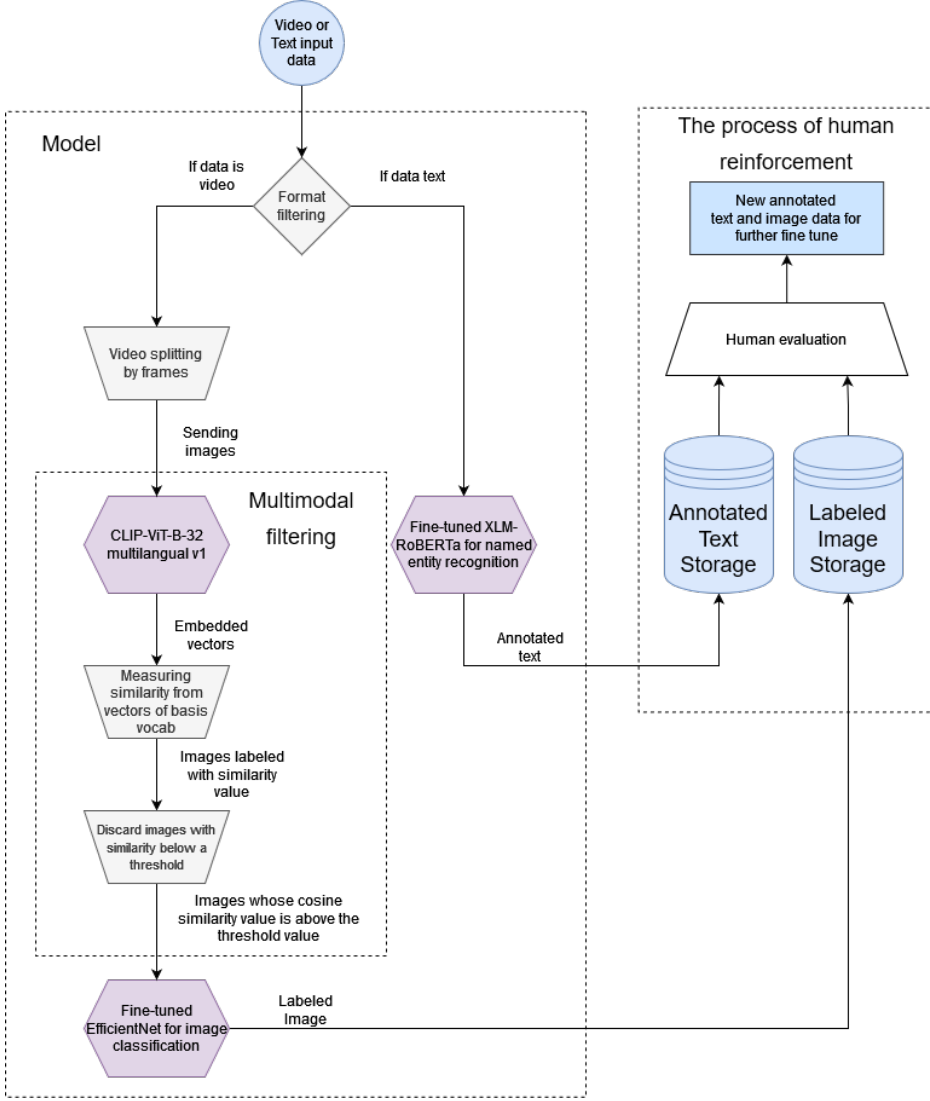


Figure 2.2: Schematic flow diagram of the smoking detector architecture.

The methodology of our architecture (see Figure 2.2) includes the following steps:

1. Separate the detection of text and video material.
2. Break the video footage into frames.
3. Filter the frames using a pre-trained multimodal system.

4. Fine-tune a pre-trained image classifier on images from a limited but openly available smoking-related database.
5. Further, the filtered images are classified using a fine-tuned image processing model.
6. Generate synthetic text data.
7. Fine-tune a language model developed for a named entity recognition (NER) task on the generated text data.

The elements of the architecture based on the proposed design are as follows:

- **Format checking:** The initial step is to determine the format of the incoming data, distinguishing between video and text data.
- **Processing of videos and images:** The system will break down the footage into frames, sampling every second.
- **Multimodal filtering:**
 - Additional information is extracted by generating embedded vectors from the images using the CLIP-ViT-B-32 [21, 22] model. Since the CLIP model is multimodal, it integrates information extracted from images and text simultaneously. Thus, it has a robust generalization capability. It can generate embedded vectors for text and images, with the advantage that the vectors generated from images become comparable to a text vector. In our case, this vector was from the "*smoking*" term.
 - Since the cosine similarity of each image vector is measured against the text vector, a characteristic function is derived from the relationship between the images and the text that shows the similarity of the video recording frames to the phrase describing smoking.
 - The characteristic function shows that a threshold value based on cosine similarity can be determined, making the frames irrelevant to the text vector formed from the smoking expression.
- **Fine-tuned image classification:** After filtering, the remaining images are further classified using a pre-trained and fine-tuned smoking image processing model, for example, EfficientNet [23], Inception [24],

ResNet [25], VGG [26], or Xception [27]. Although a small amount of data is available for fine-tuning, the multimodal filtering helps the fine-tuned and CLIP models form a Boosting-type coupling. Its more efficient performance has been demonstrated by measurements as well.

- **Text processing:** The NER technique detected smoking-related Hungarian terms in the text data. To do this, we synthetically generated text data using a Hungarian synonym dictionary and the generative language modeling service ChatGPT [28]. This synthetic data was then automatically annotated and used as learning data for the NER model.

Our complex multimodal architecture, built from different models, achieved 74% accuracy for images and 98% accuracy for text.

Our results highlight that artificial intelligence has great potential for detecting hidden advertising and enabling unbiased, reproducible, and fair quantification of smoking-related media content. This will support the effective implementation of possible regulatory measures. We proposed an integrated text and image processing model based on deep learning and generative methods, which can detect smoking in text and image material with small training data. Our developed solution addresses questions concerning the choice of appropriate training methods and the optimization of complex model architectures and their hyperparameters.

2.4 Investigating the domain transfer techniques for the development of LLM-driven knowledge-based systems

Thesis 4. *Using standard assessment methods applied in the field and the databases we have created and are publicly available, I compare Retrieval Augmented Generation (RAG) with Domain-Specific Fine-Tuning (DFT) in terms of creating an LLM-driven knowledge base system. I empirically demonstrated that RAG is a more efficient approach to building a state-of-the-art Knowledge base system that applies Generative large language models (LLM). I also show how the integration of DFT and RAG presents new types of challenges.*

Related publication: [P1]

LLMs have made a breakthrough in NLP. GPT [4], PaLM [29], and LLaMA [30], for example, have shown such remarkable capabilities in generating natural language texts that they have opened up new possibilities for the creation of more active knowledge-based systems. However, using LLMs to build knowledge-based systems requires careful consideration of training strategies, such as DFT and RAG, because of the lack of best practices and standardizations. Therefore, I comprehensively analyze these strategies and their effectiveness in this chapter.

To evaluate the effectiveness of FN and RAG, we have developed the following methodology:

- **Model selection:** The GPT-J-6B [31], OPT6.7B [32], LLaMA-7B [30] and LLaMA2-7B [30] models were selected based on some predefined criteria.
- **Metrics:** Models were evaluated based on BLEU [33], ROUGE [34], METEOR [35] scores, and cosine similarity.
- **Data preparation:** Datasets were collected from scientific publications and publicly available sources, ensuring reproducibility and verifiability.
- **Fine-tuning:** Models were fine-tuned on domain-specific datasets.

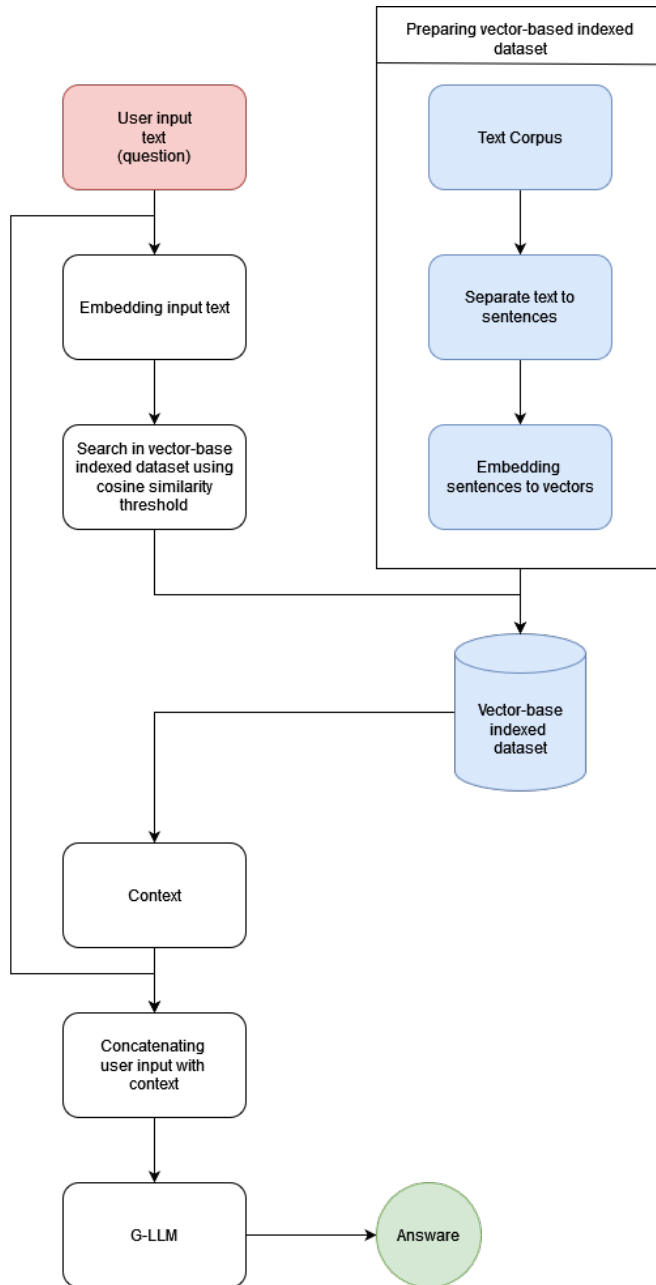


Figure 2.3: Flow diagram of the RAG model that uses a search engine based on the vectorial embedding of sentences.

Based on the results of our experiments using our methodology, the following conclusions can be drawn:

- Applying both DFT and RAG to the target task improves the performance of each model.
- RAG showed better performance compared to DFT in terms of ROUGE, BLEU scores, and cosine similarity. Although DFT models showed better METEOR scores, they were more prone to hallucination compared to RAG-based systems.
- Combining DFT and RAG did not result in performance improvements.

Based on our measurement results, we outlined a simple RAG-based architecture (see Figure 2.3) that outperforms DFT models by an average of 17% for the ROGUE score, 13% for the BLEU score, and 36% for our cosine similarity-based Coverage Score. The RAG has a significant advantage over DFT in terms of hallucination, which is not offset by the fact that the METEOR score of the DFT models, which is, on average, 8% better, indicates greater creativity compared to the RAG. In our investigation, we can create a methodology, based primarily on metrics, that determines the appropriate training method and, hence, the final optimal architecture to build.

2.5 Optimization of theoretical models

Thesis 5. *The final chapter of the thesis highlights the advantages of pre-training and shows the potential of more straightforward theoretical functions in initializing neural networks. To demonstrate this, the chapter presents an approach based on classical procedures to generate optimized SIR [36] theory models, which can then be used to pre-train a neural network.*

Related publication: [P4]

We traced the optimization of the SIR theoretical model back to the least squares curve fitting problem. We outlined an approach in which we used the Levenberg-Marquardt algorithm (LMA) [37] [38] and differential evolution (DEA) [39] algorithms to optimize the SIR model.

Our experience shows that both methods are well suited to the curve fitting problem and can be used to find the optimal SIR model that best fits the actual data. LMA is particularly advantageous due to its speed, and DEA efficiently refines the existing near-optimal parameters. Therefore, we proposed a complex optimization to find the theoretical model that best describes the problem. LMA quickly finds the best initial parameters in the first step, while DEA refines them in the second step. The optimal SIR models produced in this way can be more effectively approximated with neural networks [40] [P4], so they are ideal for their pre-training in a way that can be proven with the measurement methods accepted in the field.

References

- [1] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [3] I. Goodfellow, “Deep learning,” 2016.
- [4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [6] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization.,” *Journal of machine learning research*, vol. 13, no. 2, 2012.
- [7] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” *Advances in neural information processing systems*, vol. 25, 2012.
- [8] A. E. Eiben and J. E. Smith, *Introduction to evolutionary computing*. Springer, 2015.
- [9] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.
- [10] Twitter, “Twitter,” 2006. Online platform for social networking and microblogging.

- [11] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the international AAAI conference on web and social media*, vol. 8, pp. 216–225, 2014.
- [12] M. Heitmann, C. Siebert, J. Hartmann, and C. Schamp, *More than a feeling: Benchmarks for sentiment analysis accuracy*. Ssrn, 2020.
- [13] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, “Tweeteval: Unified benchmark and comparative evaluation for tweet classification,” 2020.
- [14] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *arXiv preprint arXiv:1409.1259*, 2014.
- [15] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [17] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [18] J. K. Pritchard, M. Stephens, and P. Donnelly, “Inference of population structure using multilocus genotype data,” *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.
- [19] D. Angelov, “Top2vec: Distributed representations of topics,” *arXiv preprint arXiv:2008.09470*, 2020.
- [20] M. Grootendorst, “Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics,” 2020.
- [21] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” *arXiv preprint arXiv:2004.09813*, 2020.
- [22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

- [23] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.
- [24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [27] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.
- [28] OpenAI, “Chatgpt,” 2024. Accessed on March 12, 2024.
- [29] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, “Palm: Scaling language modeling with pathways,” 2022.
- [30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” 2023.
- [31] B. Wang and A. Komatsuzaki, “GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.” <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.

- [32] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (P. Isabelle, E. Charniak, and D. Lin, eds.), (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.
- [34] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.
- [35] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, June 2005.
- [36] T. Harko, F. S. Lobo, and M. Mak, “Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates,” *Applied Mathematics and Computation*, vol. 236, pp. 184–194, 2014.
- [37] K. Levenberg, “A method for the solution of certain non-linear problems in least squares,” *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168, 1944.
- [38] D. W. Marquardt, “An algorithm for least-squares estimation of non-linear parameters,” *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.
- [39] R. Storn, “On the usage of differential evolution for function optimization,” in *Proceedings of North American Fuzzy Information Processing*, pp. 519–523, IEEE, 1996.
- [40] G. Bogacsovics, A. Hajdu, R. Lakatos, M. Beregi-Kovács, A. Tiba, and H. Tomán, “Replacing the sir epidemic model with a neural network and

training it further to increase prediction accuracy,” in *Annales Mathematicae et Informaticae*, vol. 53, pp. 73–91, Eszterházy Károly Egyetem Líceum Kiadó, 2021.

List of Publications

Journal articles in English

- [P1] Lakatos, R., Pollner, P., Hajdu, A., and Joó, T. (2025). Investigating the Performance of Retrieval-Augmented Generation and Domain-Specific Fine-Tuning for the Development of AI-Driven Knowledge-Based Systems. *Machine Learning and Knowledge Extraction*, 7(1), 15. <https://doi.org/10.3390/make7010015>
- [P2] Róbert Lakatos , Péter Pollner, András Hajdu, and Tamás Joó. 2024. A multimodal deep learning architecture for smoking detection with a small data approach. In *Frontiers in Artificial Intelligence* (Vol. 7). Frontiers Media SA. <https://doi.org/10.3389/frai.2024.1326050>
- [P3] Róbert Lakatos, Gergő Bogacsovics, Balázs Harangi, István Lakatos, Attila Tiba, János Tóth, Marianna Szabó, and András Hajdu. 2024. A Machine Learning-Based Pipeline for the Extraction of Insights from Customer Reviews. *Big Data and Cognitive Computing* 8, no. 3: 20. <https://doi.org/10.3390/bdcc8030020>
- [P4] Gergo Bogacsovics, András Hajdu, Róbert Lakatos, Marcell Beregi-Kovács, Attila Tiba, Henrietta Tomán. *Replacing the SIR epidemic model with a neural network and training it further to increase prediction accuracy*. In: *Annales Mathematicae et Informaticae*. Eszterházy Károly Egyetem Líceum Kiadó, 2021. p. 73-91. doi.org/10.33039/ami.2021.02.003

Journal articles in Hungarian

- [P5] Nagy János, Hajdu András, Bogacsovics Gergő, Bojtor Csaba, Illés Árpád, Lakatos Róbert, Mészáros László Precíziós gazdálkodásban használható adatelemzés alapú növénytermesztési döntéstámogató rendszer fejlesztés In: Pőr, Csilla; Szabó-Soós, Adrienn; Szabó, Péter (szerk.) LXV. Georgikon Napok Tudományos Konferencia [65th Georgikon Days Scientific Conference] Keszthely, 2024. május 17–18. : Kivonatkötet Keszthely, Magyarország : Magyar Agrár- és Élettudományi Egyetem, Georgikon Campus (2024) 206 p. pp. 133-134., 2 p.
- [P6] Bogacsovics Gergo, Hajdu András, Harangi Balázs, Lakatos István, Lakatos Róbert, Szabó Marianna, Tiba Attila, Tóth János. Napelemfarmok Magyarország területén történő elhelyezését segítő döntéstámogató rendszer fejlesztése. KözigazgatásTudomány, 2021, 1.2: 134-145. doi.org/10.54200/kt.v1i2.23
- [P7] Bogacsovics Gergo, Hajdu András, Harangi Balázs, Lakatos István, Lakatos Róbert, Szabó Marianna, Tiba Attila, Tóth János, Tarcsi Ádám. Adatelemzési folyamat és keretrendszer a közigazgatás számára. KözigazgatásTudomány, 2021, 1.2: 146-158. doi.org/10.54200/kt.v1i2.24
- [P8] Bátfai, M., Lakatos, R., Molnár, G., Takácsné Bubnó, K. BPMN 2.0 a könyvtári minőségbiztosításban, Tudományos és Műszaki Tájékoztatás, 63(8), p. 305–316, 2016.

Conference papers

- [P9] Róbert Lakatos, Eszter Klára Urbán, Zoltán János Szabó, János Pozsga, Eszter Csernai and András Hajdu. 2024. Designing Prompts and Creating Cleaned Scientific Text for Retrieval Augmented Generation for More Precise Responses from Generative Large Language Models. In 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) (pp. 1-6). IEEE. 10.1109/CITDS62610.2024.10791382
- [P10] Árpád Pandy, Róbert Lakatos Péter Pollner, Eszter Csernai and András Hajdu. 2024. Investigating the Influence of Hyperparameters on the Optimal Time-Series Prediction Ability of Generative Large Language Models. In 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) (pp. 1-6). IEEE. 10.1109/CITDS62610.2024.10791385
- [P11] Gergő Bogacsovics, Balázs Harangi, Marcell Beregi-Kovács, Dávid Kupás, Róbert Lakatos, Norbert Dávid Serbán, Tiba Attila, János Tóth. 2024. Assessing Conventional and Deep Learning-Based Approaches for Named Entity Recognition in Unstructured Hungarian Medical Reports. Presented at the 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI). <https://doi.org/10.1109/sami60510.2024.10432795>
- [P12] Róbert Lakatos, Gergo Bogacsovics, András Hajdu. Predicting the direction of the oil price trend using sentiment analysis. In: 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS). IEEE, 2022. p. 177-182. doi.org/10.1109/CITDS54976.2022.9914158



Registry number:
Subject:

DEENK/140/2025.PL
PhD Publication List

Candidate: Róbert Lakatos

Doctoral School: Doctoral School of Informatics

MTMT ID: 10079387

List of publications related to the dissertation

Foreign language scientific articles in Hungarian journals (1)

1. Bogacsovics, G., Hajdu, A., **Lakatos, R.**, Beregi-Kovács, M., Tiba, A., Tomán, H.: Replacing the SIR epidemic model with a neural network and training it further to increase prediction accuracy.
Ann. Math. Inform. 53, 73-91, 2021. ISSN: 1787-5021.
DOI: <http://dx.doi.org/10.33039/ami.2021.02.003>

Foreign language scientific articles in international journals (3)

2. **Lakatos, R.**, Pollner, P., Hajdu, A., Joó, T.: Investigating the Performance of Retrieval-Augmented Generation and Domain-Specific Fine-Tuning for the Development of AI-Driven Knowledge-Based Systems.
Mach. Learn. Knowl. Extr. 7 (1), 1-18, 2025. EISSN: 2504-4990.
DOI: <http://dx.doi.org/doi.org/10.3390/make7010015>
IF: 4 (2023)
3. **Lakatos, R.**, Bogacsovics, G., Harangi, B., Lakatos, I., Tiba, A., Tóth, J., Szabó, M., Hajdu, A.: A Machine Learning-Based Pipeline for the Extraction of Insights from Customer Reviews.
Big Data Cogn. Comput. 8 (3), 1-24, 2024. EISSN: 2504-2289.
DOI: <http://dx.doi.org/10.3390/bdcc8030020>
IF: 3.7 (2023)
4. **Lakatos, R.**, Pollner, P., Hajdu, A., Joó, T.: A multimodal deep learning architecture for smoking detection with a small data approach.
Front. Artif. Intell. 7, 1-8, 2024. EISSN: 2624-8212.
DOI: <http://dx.doi.org/10.3389/frai.2024.1326050>
IF: 3 (2023)





Foreign language conference proceedings (1)

5. **Lakatos, R.**, Bogacsovics, G., Hajdu, A.: Predicting the direction of the oil price trend using sentiment analysis.
In: IEEE 2nd Conference on Information Technology and Data Science (CITDS) :
Proceedings. Ed.: Fazekas István, Institute of Electrical and Electronics Engineers (IEEE),
Piscataway, 177-182, 2022.

List of other publications

Hungarian scientific articles in Hungarian journals (3)

6. Bogacsovics, G., Hajdu, A., Harangi, B., Lakatos, I., **Lakatos, R.**, Szabó, M., Tiba, A., Tóth, J.,
Tarcsi, Á.: Adatelemzési folyamat és keretrendszer a közigazgatás számára.
Közigazgatástudomány. 1 (2), 146-158, 2021. ISSN: 2786-1910.
DOI: <http://dx.doi.org/10.54200/kt.v1i2.24>
7. Bogacsovics, G., Hajdu, A., Harangi, B., Lakatos, I., **Lakatos, R.**, Szabó, M., Tiba, A., Tóth, J.:
Napelemfarmok Magyarország területén történő elhelyezését segítő döntéstámogató
rendszer fejlesztése.
Közigazgatástudomány. 1 (2), 134-145, 2021. ISSN: 2786-1910.
DOI: <http://dx.doi.org/10.54200/kt.v1i2.23>
8. Soós, M. E., **Lakatos, R.**, Molnár, G., Takácsné Bubnó, K.: BPMN 2.0 a könyvtári
minőségbiztosításban.
Tud. Műsz. Tájs. 63 (8), 305-316, 2016. ISSN: 0041-3917.

Hungarian conference proceedings (1)

9. Nagy, J., Hajdu, A., Bogacsovics, G., Bojtó, C., Illés, Á., **Lakatos, R.**, Mészáros, L.: Precíziós
gazdálkodásban használható adatelemzés alapú növénytermesztési döntéstámogató
rendszer fejlesztés.
In: LXV. Georgikon Napok Tudományos Konferencia = 65th Georgikon Days Scientific
Conference /szerk. Pőr Csilla, Szabó-Soós Adrienn, Szabó Péter, Magyar Agrár- és
Élettudományi Egyetem Georgikon Campus, Keszthely, 133-134, 2024. ISBN:
9786156338105





Foreign language conference proceedings (3)

10. Bogacsovics, G., Harangi, B., Beregi-Kovács, M., Kupás, D., **Lakatos, R.**, Serbán, N. D., Tiba, A., Tóth, J.: Assessing Conventional and Deep Learning-Based Approaches for Named Entity Recognition in Unstructured Hungarian Medical Reports.
In: 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI). Ed.: Kovács Levente, Liberios Vokorokos, IEEE, Piscataway, 77-82, 2024. ISBN: 9798350317206
11. **Lakatos, R.**, Urbán, E. K., Szabó, Z. J., Pozsga, J., Csernai, E., Hajdu, A.: Designing Prompts and Creating Cleaned Scientific Text for Retrieval Augmented Generation for More Precise Responses from Generative Large Language Models.
In: 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) / Hajdu Andras, Institute of Electrical and Electronics Engineers, Piscataway (NJ), 1-6, 2024.
12. Pándy, Á., **Lakatos, R.**, Pollner, P., Csernai, E., Hajdu, A.: Investigating the Influence of Hyperparameters on the Optimal Time-Series Prediction Ability of Generative Large Language Models.
In: 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS), Institute of Electrical and Electronics Engineers (IEEE), Piscataway, 158-163, 2024. ISBN: 9798350387889

Hungarian abstracts (1)

13. Soós, M. E., **Lakatos, R.**, Molnár, G., Takácsné Bubnó, K.: BPMN 2.0 a könyvtári minőségbiztosításban.
In: Networkshop 2016. Szerk.: Mikusné Sárvári Klára, NIIF, Budapest, [8], 2016. ISBN: 9789638833563

Total IF of journals (all publications): 10,7

Total IF of journals (publications related to the dissertation): 10,7

The Candidate's publication data submitted to the Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

08 April, 2025

