SZÖVEGANALÍZIS ÉS MESTERSÉGES INTELLIGENCIA: BEVEZETÉS A GÉPI TANULÁS ÉS A MINTAKERESÉS ÁLTAL NYÚJTOTT LEHETŐSÉGEKBE

Szerző:

Mező Péter Dániel
Forward University

Első szerző e-mail címe:
peter.mezo1@gmail.com

Lektorok:

Szabóné Balogh Ágota (PhD)
Gál Ferenc Egyetem (Magyarország)

Borbélyné Bacsó Viktória (Ph.D.)
Medgyessy Ferenc Gimnázium,
Művészeti Szakgimnázium és Technikum

és további két anonim lektor…

Absztrakt

A tanulmány a mesterséges intelligencia szövegelemzésbeli felhasználási lehetőségét vázolja fel. A következőkben szó lesz a szövegelemzés és mesterséges intelligencia kapcsolatáról, valamint a felügyelt és nem felügyelt gépi tanulásról és azokat használó algoritmusokról.

Kulcsszavak: szövegelemzés, mesterséges intelligencia, gépi tanulás, algoritmus
Diszciplína: informatika

Abstract

TEXT ANALYSIS AND ARTIFICIAL INTELLIGENCE: INTRODUCTION TO THE POTENTIAL PROVIDED BY MACHINE LEARNING AND PATTERN SEARCHING

This paper drafts the possibilities of Artificial Intelligence (AI) for text analysis. In the following, the paper will discuss the relationship between text analysis and artificial intelligence; supervised and unsupervised machine learning; and the algorithms that use them.

Keywords: text analysis, artificial intelligence, machine learning, algorithm
Discipline: computer science

Mező Péter Dániel (2023): Szöveganalízis és mesterséges intelligencia: bevezetés a gépi tanulás és a mintakeresés által nyújtott lehetőségekbe. *OxIPO – interdiszciplináris tudományos folyóirat, 2023/2.* 67-72. DOI 10.35405/OXIPO.2023.2.67

Jelen tanulmány célja, hogy betekintést nyújtson mesterséges intelligencia (MI) szereplehetőségébe a természetes nyelv területén végzett szövegelemzés terén. Az infokommunikáció különböző területein jelenlévő szöveges adatok bővülése miatt napjainkban megnőtt az igény a szövegelemzésre szakosodott programokra. A szövegelemző feladatok emberi aspektusa hajlamos a hibákra és kevésbé hatékony az automatizált algoritmusokhoz képest, amelyek gépi tanulást alkalmaznak önfejlesztésre és a korábbi modellekből való tanulásra. Ezek az algoritmusok felügyelt vagy nem felügyelt tanulási módszereket használnak a szövegelemzés különböző akadályainak leküzdése érdekében.

A szövegelemzésről

A szövegelemzés, a szövegbányászat (text mining) vagy természetes nyelvi feldolgozás (natural language processing) informatikai manifesztációja számítástechnikai eszközök segítségével von ki értékes információkat strukturálatlan szövegekből. Több területen fontos szerepet játszanak az effajta algoritmusok – például a gazdasági, az egészségügyi és közösségi média analízis területén (Aggarwal, 2012). A manuális, ember által végzett szövegelemzés idő- és munkaigényes folyamat, emellett az emberi tényező növeli a figyelmetlenségből, fáradékonyságból, motiválatlanságból eredő hibák lehetőségét. A digitális tartalmak növekedésével az érdekelt egyéneknek és szervezeteknek meg kell küzdenie a számukra lényeges információk megszerzésével. Nehézséget jelenthet például az adott érdekekhez (például projektekhez, kutatásokhoz, vállalkozásokhoz) megfelelő adatok kivonása az internet folyamatosan változó, növekedő fokozódó adathalmazból. A szövegelemző algoritmusok megoldást nyújthatnak e problémák kezelésében, s az automatizáció segítségével a humán információfeldolgozáshoz képest effektívebb módszert biztosíthatnak a nagy volumenű, szöveg alapú adatok analizálásához (Bird, Klein és Loper, 2009) – bár az adatok értelmezésében az emberi intelligencia még mindig szükséges lehet.

A mesterséges intelligencia és a szövegelemzés

A mesterséges intelligencia (MI, angolul: artificial intelligence; AI) gyűjtőfogalom olyan algoritmusokra utal, amelyek az emberi elme működésének részét vagy egészét modellezve végeznek el feladatokat.

A mesterséges intelligencia három átfogó típusára, az ANI, az AGI és az ASI mozaik szavak utalnak.

Az ANI (Artificial Narrow Intelligence) algoritmusok feladat-specifikusak és egyetlen specifikus feladat ellátására szolgálnak. Jelenleg a legfejlettebb AI algoritmusok többsége ebbe a kategóriába tartozik.

Az AGI (Artificial General Intelligence) el tudja érni és egyes részfeladatokban meg tudja haladni az emberi intelligenciát. Tud érvelni, tervezni, problémákat megoldani, elvontan gondolkodni, gyorsan tanulni és tapasztalatokat szerezni (Gottfredson, 1997).

Az ASI (Artificial Superintelligence) jelenleg még csak a sci-fi műfajban létezik. E teoretikus algoritmus minden területen meghaladja az emberi képességeket – például a kreativitás, az általános tudás és a szociális készségek terén is (Bostrom, 2006; Strelkova, 2017).

A mesterséges intelligencia felhasználási lehetőségei – a probléma-orientált fejlesztés lehetősége miatt – igen nagyok a szövegelemzés területén is. Az MI révén lehetőség van akár arra is, hogy nagy mennyiségű, strukturálatlan szövegben is viszonylag gyorsan találjunk indirekt (nem nyilvánvaló, közvetett) információkat, sőt: e találatokból tanulva az algoritmusok akár saját teljesítményüket is fokozni tudják.

A gépi tanulást alkalmazó algoritmusok két fő csoportra oszthatók az alapján, hogy felügyelt (supervised') vagy nem felügyelt (unsupervised) tanulást alkalmaznak-e.

Felügyelt gépi tanulás

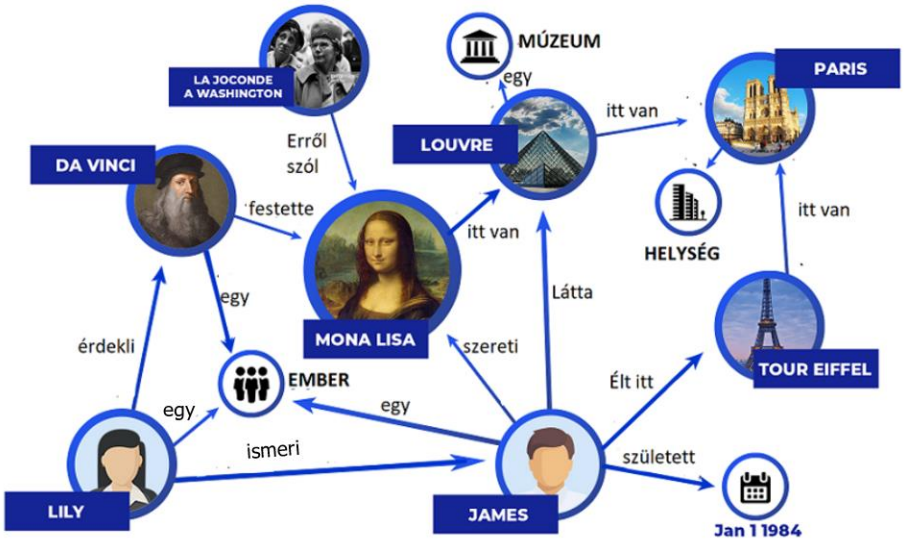
A felügyelt tanulással működő algoritmus ember által megadott és összekapcsolt bemeneti és kimeneti adatok bázisán állít fel olyan modellt, ami alapján a jövőben a korábban nem ismert bemeneti adatokhoz hozzá tudjon társítani valamilyen kimeneti adatot (Mitchell, 2020; Net2). Ezeket, a kezdetben megadott adathalmazban egymáshoz kapcsolt be- és kimeneti adatokat „labelled data"-nak (címkézett adatoknak) nevezzük.

Ennek a metódusnak egyik gyakori felhasználási területe a hangulat analízis (Sentiment analysis – v.ö.: Cambria, Schuller, Xia, és Havasi, 2013). Ebben a feladatkörben a felügyelt gépi tanulást alkalmazó algoritmus egy szöveg elemzése során a felhasználó lehetséges érzelmeit azonosítja. Ez például a közösségi média vagy a felhasználói visszajelzés-értékelés esetében lehet lényeges segítség, amellyel az érdekelt fél fel tudja mérni a közvéleményt és a felhasználói elégedettség mértékét, optimalizálhatja a marketing és branding folyamatokat; esetleg predikciókat tud készíteni a következő trendekre, annak érdekében, hogy a versenytársak előtt érje el a potenciális piacot. Egy másik felhasználási módja a felügyelt tanulásnak az úgynevezett named entity recognition (NER). A NER algoritmus célja ilyenkor a nevekkel rendelkező entitások felkeresése a szövegben (legyenek ezek a nevek emberi nevek, szervezetek, helyszínek stb. megnevezései). Ez teret ad az entity linking metódusnak is, amely alapján például a Wikipédia működik, és aminek lényege: felfedezett entitások hozzárendelése a már egy létező adatbázisban fellelhető entitásokhoz – lásd: Net3; Mihalcea és Csomai, 2007). Ezek az algoritmusok fontos szerepet játszanak az olyan alkalmazásokban, amelyek célja például az információkeresés és a tudásgráf készítés (knowledge graph construction – lásd: Ratinov és Roth (2009), illetve 1. ábra.

Nem felügyelt gépi tanulás

A nem felügyelt tanulás technikák, mint például a clustering (adatok csoportokba helyezése hasonlóság alapján – Net4), vagy a topic modelling (hasonló tartalmú szöveg csoportosítása – v.ö.:Net5), nem szorulnak

1. ábra: Entitások közötti kodependencia ábrázolása tudásgráf használatával. Forrás: Net5



külső, kezdeti, emberi beavatkozásra (például labelled adatokra) ahhoz, hogy felállítsanak egy modellt a szöveg alapján. Ez a megközelítés kutató jellegű analízist és információkeresést folytat csak az adott szöveg segítségével, korábbi „betanulási"/ „betanítási" fázis szüksége nélkül. Konkrét példa a topic modelling megközelítésre a Latent Dirichlet Allocation (LDA), amely automatikusan felfedezi a szavak elhelyezkedésének vizsgálatával végzett tartalomelemzés során történik (Blei, Ng, és Jordan, 2003).

Az MI alapú szövegelemzés és a tanulás OxIPO-modellje

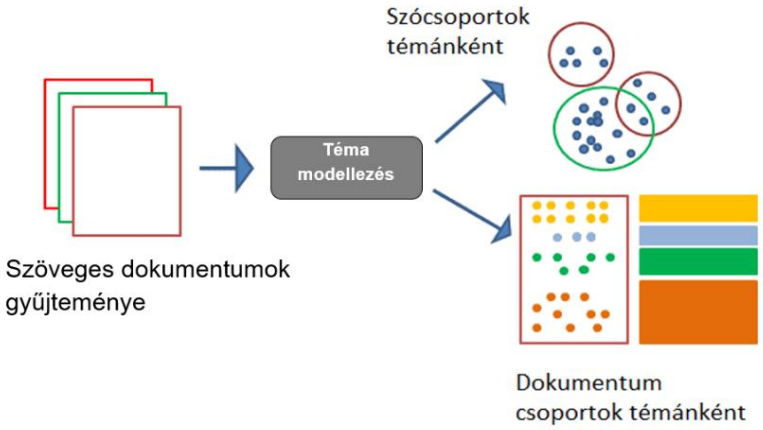
Az emberek által végzett tanuláshoz hasonlóan a mesterséges intelligencia alapú gépi tanulás is elemezhető, tervezhető a tanulást informcáiófeldolgozási folyamatnak tekintő OxIPO modell segítségével –a modell részletes bemutatását lásd Mező és Mező (2019) tanulmányában. A tanulás OxIPO-modellen alapuló formulája:

Tanulás=Organizáció*(Input*Process*Output)

Az OxIPO-formula komponenseit a gépi tanulás során például a követ-kezőképpen lehet értelmezni:

- Input: az elemzésre szolgáló szöveg, illetve a szöveget jellemző érték;
- Process: a módszer, illetve folyamat amivel hozzáköt egy értéket az input szöveg adott szakaszához az algoritmus;

2. ábra: az LDA típusú algoritmus működése. Forrás: Net6



- Output: az input szöveghez rendelt értékekkel jellemezhető adatbázis, visszajelzés, jelentés stb.
- Az Organizáció a felügyelt gépi tanulási folyamat szervezését – az elemzésre szolgáló szöveg kiválasztását, az elemzési szempontok meghatározását, a felügyeletett, a folyamathoz szükséges hardverkörnyezet biztosítását stb. – jelenti. Nem felügyelt tanulás esetén a humán erőforrással kapcsolatos szervezési feladatok egyszerűsödnek.

Mivel az Organizáció (tanulásszervezés) szorzó szoros kapcsolatban áll a zárójelen belüli értékekkel, egyértelmű, hogy a teljesítményt alapvetően befolyásolja a gépi tanulás szervezettsége.

Összegzés

Összegzésként megfogalmazható, hogy a mesterséges intelligencia sokrétű feladatot képes ellátni a szövegelemzési feladatok esetében. Az MI technológia képes nagy mennyiségű strukturálatlan szöveges adatok feldolgozására is, annak érdekében, hogy indirekt információt nyerjen ki azokból. Az automatizáció segítségével azok a feladatok, amelyek az emberi kapacitást nagy mértékben igénybe vették, egyre hatékonyabban végeztethetők el gépi tanulásra is képes algoritmusok segítségével. A felügyelt és nem felügyelt gépi tanulási technikák tehát fontos szerepet játszhatnak a szöveganalízis feladatokban. Ahogy a technológia fejlődik, úgy a mesterséges intelligencia szövegelemzés területén elért eredményei és lehetőségei is egyre magasabbra fognak törni.

Köszönetnyilvánítás

A tanulmány a Kulturális és Innovációs Minisztérium és a Nemzeti Tehetség Program NTP-NFTÖ-22-A2-0249 pályázati azonosítószámú támogatásával valósult meg. A támogatást ezúton is tisztelettel köszönöm!

Kulturális és Innovációs Minisztérium
Nemzeti Tehetség Program

Irodalom

Aggarwal, C. C. (2012). *Text Mining: The state of the art and the challenges*. Springer Science & Business Media.

Bird, S., Klein, E., és Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Beijing.

Blei, D. M., Ng, A. Y., és Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Cambria, E., Schuller, B., Xia, Y.,& Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), 15-21. DOI 10.1109/MIS.2013.30

Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing* (3rd ed.). Pearson, California.

Gottfredson L.S. (1997). *Mainstream Science on Intelligence : An Editorial With 52 Signatories, History and Bibliography*. Linda S. Gottfredson. - Ablex Publishing Corporation, USA.

Mező Ferenc és Mező Katalin (2019): Az OxIPO-modell – az interdiszciplináris kutatások egy lehetséges értelmezési kerete. *OxIPO – interdiszciplináris tudományos folyóirat*, 2019/1, 9–21. doi: 10.35405/OXIPO.2019.1.9

Mihalcea R és Csomai A (2007). Wikify! linking documents to encyclopedic knowledge. *In CIKM '07 Proceedings of the sixteenth ACM conference on conference on information and knowledge management*, 233–242. DOI 10.1145/1321440.1321475

Mitchell, T. (2020). *Machine Learning*. McGraw Hill, USA.

Net1: Lexiq (2023). Supervised learning. Utolsó megtekintés: 2023.06.14. URL: https://lexiq.hu/supervised-learning

Net2: Wikipedia (2023). Entity linking. Utolsó megtekintés: 2023.06.15. URL: https://en.wikipedia.org/wiki/Entity_linking

Net3: Google for Developers (2022). What is Clustering?. Utolsó megtekintés: 2023.06.15 URL: https://developers.google.com/machine-learning/clustering/overview

Net4: MonkeyLearn (2019) Topic Modeling: An Introduction. Utolsó megtekintés: 2023.06.15. URL: https://monkeylearn.com/blog/introduction-to-topic-modeling/

Net5: Megnyitás: 2023.06.15. URL: https://yashuseth.files.wordpress.com/2019/10/knowledge-graph.jpg

Net6: Megnyitás: 2023.06.15. URL: https://editor.analyticsvidhya.com/uploads/4519707623a1e07ae153046bc387e0136a65f.image001-min.png

Ratinov, L., és Roth, D. (2009). *Design Challenges and Misconceptions in Named Entity Recognition*. Proceedings of the Thirteenth Conference on Computational Natural Language Learning, 147-155.

Strelkova, O. (2017). *Three types of Artificial Intelligences*. Khmelnitsky National University.