DOKTORI (PHD) ÉRTEKEZÉS

Barta Ákos

GÖDÖLLŐ

2024

Magyar Agrár- és Élettudományi Egyetem

OLAJÁR ELŐREJELZÉS MESTERSÉGES NEURÁLIS HÁLÓK ALKALMAZÁSÁVAL A WALL STREET JOURNAL CIKKEIRE ALAPOZVA

DOI: 10.54598/004540

DOKTORI (PHD.) ÉRTEKEZÉS

BARTA ÁKOS

GÖDÖLLŐ

2024

A doktori iskola

megnevezése: Gazdaság- és Regionális Tudományi Doktori Iskola

tudományága: gazdálkodás és szervezéstudomány

vezetője: **Dr. Bujdosó Zoltán PhD**
egyetemi tanár
Magyar Agrár- és Élettudományi Egyetem
Fenntartható Fejlesztés és Gazdálkodás Intézet

Témavezető(k): **Dr. habil Molnár Márk PhD**
egyetemi docens
Eötvös Lóránd Tudományegyetem
Gazdaságtudományi Kar
Összehasonlító Gazdaságtan Tanszék

Dr. Naárné Dr. Tóth Zsuzsanna Éva PhD
egyetemi docens
Budapesti Metropolitan Egyetem
Gazdaságtudományi Intézet

.....	.....
Az iskolavezető jóváhagyása	A témavezető(k) jóváhagyása

Tartalomjegyzék

1. Bevezetés

Több mint 40 év telt el az 1973-as első olajár-sokk óta. Ebben az időszakban az olaj iránti globális kereslet drámaian megnőtt, miközben az új energiával kapcsolatos technológiák és új energiaforrások ellenállóbbá tették a globális fogyasztókat az olajsokkokkal szemben. Az 1970-es évek olajsokkjai óta a feltörekvő gazdaságok szerepen jelentősen megnövekedetta globális energiafogyasztásban. A Kínai Népköztársaság részesedése például ötször nagyobb, mint az 1970-es években volt. Másrészt a legnagyobb (USA) és a jelenlegi harmadik legnagyobb (Japán) olajfogyasztók részesedése az 1970-es évek óta csökkent, az Egyesült Államoké 32%-ról 21%-ra, Japáné 10%-ról 5%-ra.

Az 1970-es évek olajválságait és az azt követő gazdasági recessziókat követően több tanulmány is megállapította, hogy az olajársokkok jelentős szerepet játszottak a gazdasági visszaesésekben. Az elmúlt években az olajárak 2001-ben megkezdődött meredek emelkedése és a másodlagos jelzáloghitel-válságot követő 2008-as meredek csökkenés újra felkeltette az érdeklődést az olajárak makrogazdaságra gyakorolt hatásai iránt.

A fosszilis tüzelőanyagok továbbra is a jelenlegi globális energiafelhasználás jelentős részét fedezik, 2020-ban ennek megfelelő 80%-os részesedéssel. (IEA, 2020). Az olaj továbbra is a világ vezető üzemanyaga, 2020-ban a globális energiafogyasztás 31,2%-át tette ki, ami azt jelzi, hogy a kőolaj továbbra is fontos a nemzetközi tényezőpiacokon. Ezért a kőolajár-determinánsok történeti fejlődésének megértése a gazdaságpolitikai tervezés érdekében kiemelten fontos kutatási kérdés. E tekintetben a kőolajpiacon az új évezred első évtizedében tapasztalható ármozgások főként két okból keltették fel a figyelmet: egyrészt az 1980-as és 1990-es évek alatti alacsony szintről több éven keresztül emelkedett az ár, amely rekordotdöntött. Az árak ezen alakulása széleskörű vitát váltott ki egy újabb olajválságról, utalva az 1970-es és 1980-as évek két olajválságára. Másodszor, ami még fontosabb, az első két olajválsággal ellentétben ennek a mostani árcsúcsnak az oka nem egyértelmű: egyszerre több potenciálisan releváns fejlemény zajlott le, ami megnehezíti azok árra gyakorolt hatásának azonosítását. A gazdasági szakirodalomban egyre több cikk tanúskodik erről. Az akadémiai diskurzus általában három magyarázat között ingadozik, tükrözve a kőolaj árára ható piaci erőket: először is azt állítják, hogy az áremelkedés a kőolajkészletek végességét és a termelési kapacitások további bővítésének képtelenségét mutatja (kínálatvezérelt áremelkedés). (Kaufmann, 2011) Másodszor, feltételezik, hogy a feltörekvő országok, például Kína és India váratlanul erős gazdasági növekedése eredményezte a kőolaj iránti kereslet váratlan növekedését, ami a kőolaj azonnali szállításának megszorulásához és az ár emelkedéséhez (keresletvezérelt áremelkedés) vezetett. (Hamilton, 2008) (Kilian, 2009) Harmadszor megállapítják, hogy a spekulánsok növekvő száma a kőolajpiacon jelentősen megerősítette az előretekintő keresleti tevékenységek súlyát, és ezzel megváltoztatta az árdinamikát (várakozásokvezérelte áremelkedés) (Fattouh, Kilian, & Mahadeva, 2013)

Abból kiindulva, hogy a kőolaj az egyik legfontosabb nyersanyag, az áringadozás jelentős világgazdasági hatással bír. Ebből kifolyólag a gazdaság szereplői igyekeznek előrejelezni az árfolyam változásait, a trendeket, a jövőbeli értékeket. Mivel a gazdasági szereplők nagy része nincs piacbefolyásoló pozícióban, így a hozzá beérkező információkból igyekszik ezen adatokat és információkat kinyerni.

A mai felgyorsult világban az információ óriási tömegével szembesülünk. Clickbait cikkek, hitelesnek vélt források ezrei állnak rendelkezésre, amelyekből igyekszünk kigyűjteni a számunkra hasznos információkat. A kulcs tehát nem feltétlenül az információ vagy hír beszerzése, hanem az információáradatból kiszűrni a releváns információtartalmat, majd azt hatékonyan feldolgozni. A probléma összetett és a gazdasági szereplők egyre komplexebb információfeldolgozási és szűrési metódosukat alkalmaznak, egyre bonyolultabb információs hálózatokból és nagy adatmennyiségből igyekszenek kiszűrni a zajt, vagyis a fals információkat, kiegészítő tényeket.

Jelen dolgozat témája az olajárfolyam elemzése egy komplexebb megközelítés alkalmazásával. Mivel mindenkit érint, illetve közvetetten minden iparágra kihatása van, így nagy szereppel bír. Oligopol piacon működik, vagyis a szereplők nagy figyelmet fordítanak a többi szereplő döntéseire. Mivel relatíve nagy a piaci részesedése egy-egy szereplőnek, így nagy hatással bírhatnak. Felhasználóként, végfogyasztóként ráhatás nincs a piacra, de a nagy információtömegből egy speciális szűrőt szükséges alkalmaznunk, hogy tudjuk mely híreket érdemes komolyan venni.

Az elmúlt években egyre nagyobb teret hódítanak a neurális hálók, melyek nem statisztikai alapokon képesek összefüggéseket és következtetéseket, kapcsolatokat felismerni adathalmazok pontjai között, mint például a hírek. A neurális hálók az 1940-es éves környékén indultak fejlődésnek, a számítástechnika fejlődésével párhuzamosan. Eleinte kapacitáskorlátok is voltak, de a jelenlegi technológia fejlettség mellett az ilyen fejlesztések fénykorukat élik. Az információfeldolgozás módszere a neurális hálók esetében nagyban hasonlít az emberi agyban jelenlévő hálózathoz, nem feltétlenül statisztikai alapon működik. Mindamellett jóval gyorsabb és hatékonyabb, mint amire az ember képes lesz/lehet.

Feltételezésem az, hogy a gazdasági és politikai hírekkel kapcsolatos újságcikkek tartalma alapján előre jelezhető az olajár változása, legalábbis bizonyos szinten. Az újságcikkekben szereplő gazdasági és politikai információk segítségével meg lehet válaszolni a kérdést, hogy milyen változás várható az olajárakban. A hírekben szereplő információk alapján előre látható véleményem szerint, hogy mikor várható növekedés vagy csökkenés az olajárakban, valamint a változás mértéke is.

Valószínűsíthető, ha az írott vagy elektronikus sajtóban pozitív, vagy negatív véleményeket közölnek az olajárakra vonatkozóan, akkor ez hatással lehet a vevők és vezetők döntéseire, amelyek végeredményeként befolyásolhatják az olajárak változását. A hírek alapján tehát spekulatív alapokon nyugvó döntéseket hoznak a piaci szereplők, amelyek keresleti vagy kínálati oldali változás miatt tényleges árváltozást fognak eredményezni, úgy, hogy az eredetileg véleményezett és valószínűsített eredmény esetleg be sem következi.

Fontos kijelenteni, hogy gazdasági szereplők alatt olyan háztartásokat, cégeket értek, akik ugyan rendelkeznek információkkal, de nem „tűzközeliek”, vagyis másodkézből tudnak értesülni és alapoznak a szaksajtó hitelességére.

2. Célkitűzések

A jelenlegi információdömpingben a gazdasági aktorok igyekeznek feldolgozni a rendelkezésre álló információkat. Az információfeldolgozás során szükséges kiszűrni a nem valid információt, illetve súlyozni az egyes tények vagy vélemények befolyását. Ezáltal feltételezhetjük azt, hogy a híreknek és információknak manipulációs jellege van. Az olajár vonatkozásában, ami egy oligopol piac jellemző mutatója, ahol a termelési döntéshozatal nem nyilvános, a gazdasági aktorok nagymértékben támaszkodhatnak a hírekre.

Az értekezés módszertanilag alapvetően a mesterséges neurális hálók alkalmazására épül, amik jelenleg fénykorukat élik, egyre szervesebben belefolyva életünkbe. A módszertan hasznossága, felépítése és fejlesztése véleményem szerint az elkövetkezendő időszak egyik fő témaköre lesz. Ebből adódóan a neurális hálók tudományos és napi életbe való alkalmazhatóságát vizsgálom, pontosabban mennyire tudja segíteni napi tevékenykedésünket és döntéshozatali metódusunkat.

Kutatásom során célom, hogy bebizonyítsam, hogy az információáramot elemezve (vagyis híreket, sajtóinformációt, nem pedig fundamentális adatokat) van kapcsolat az olajárfolyam és a sajtóhírek között, ezáltal egyrészt alkalmazható előrejelzésként, másrészt kimutatható a spekuláció alapú árfolyammozgás.

A témakör, illetve a kutatás kapcsán az alábbi hipotéziseket fogalmazom meg, melyeket kutatásom során vizsgálni fogok:

[1] A mesterséges neurális hálózatok képesek hatékony információfeldolgozásra, vagyis nagy adattömegek (Big Data) gyors és hatékony elemzésére is használhatóak.
[2] A vizsgált folyóiratok és az árfolyam között összefüggés mutatható ki, vagyis egyértelműen bizonyítható a spekulációs árfolyammozgás az olajár tekintetében.
[3] A Wall Street Journal (WSJ) újságcikkek tartalmának mesterséges neurális hálóval (ANN) történő elemzésével kellő pontossággal meghatározható a következő napi olajár változás.
[4] Az újságcikkek összefoglalásával, vagyis tömörítésével nagymértékben növelhető a mesterséges neurális hálóval történő előrejelzési hatékonyság.
[5] Az újságcikkek hangulatelemzésével nagymértékben növelhető a mesterséges neurális hálóval történő előrejelzési hatékonyság.
[6] Az olajárfolyam visszacsatolt neurális hálóval (RNN) történő vizsgálatával, tehát csak az árfolyam historikus mozgásának elemzésével kellő pontossággal meghatározható a következő napi olajár változás.
[7] A mesterséges neurális háló rejtett rétegeinek és neuronjainak, vagyis háló-részének nagymértékű növelésével jelentősen növelhető a hatékonyság.
[8] A WSJ újságcikkeinek mesterséges neurális hálóval történő elemzése hatékonyabb, mint az árfolyam visszacsatolt neurális hálóval való elemzése, vagyis az árfolyam mozgásában nagyobb szerepe van a spekulációnak, mint a fundamentumoknak.
[9] A mesterséges neurális háló segítségével kapott előrejelzés hatékonyabb, mint adott matematikai-tőzsdei modellekkel történő árfolyamváltozás-előrejelzés.

3. Irodalmi áttekintés

3.1 Az olaj szerepe a világgazdaságban

A kőolaj egyfajta nélkülözhetetlen alapvető energiaforrás, vegyi anyag és stratégiai erőforrás a társadalmi-gazdasági fejlődésben. A kőolaj árának változása jelentősen befolyásolhatja egy ország gazdasági fejlődését, társadalmi stabilitását, sőt nemzetbiztonságát is. (Wu & Zhang, 2014) Ezért nagy jelentősége van olyan tudományos módszerek kidolgozásának, amelyek a kőolajár-mozgások lehető legpontosabb előrejelzését szolgálják, a kőolajpiaci szélsőséges kockázatok kezelése és a profitszerzési lehetőségek megtalálása érdekében.

A kőolajpiaci kereslet és kínálat, az USA-dollár árfolyamának, a spekulatív kereskedésnek, a geopolitikai konfliktusoknak, a természeti katasztrófáknak stb. összefolyó hatása miatt azonban a kőolaj nemzetközi ára fellendült, és hordónkénti ára 30-150 dollár között mozgott az elmúlt évtizedben magas piaci volatilitással. (Zhang, Fan, Tsai, & Wai, 2008) (Zhang & Wei, 2011) (Zhang Y. J., 2013). A múltbeli adatok azt mutatják, hogy a nemzetközi kőolajárak összetett volatilitási jellemzői, mint például a nemlinearitás, a bizonytalanság és a dinamika megnehezítik a kőolajár előrejelzését, és a várható eredmények nagy kockázatot hordoznak, ami végül jelentős bizonytalanságot okozhat a hozamokban.

Az olajárak kulcstényezők a legtöbb makrogazdasági prognózisban. Az elmúlt évtizedekben a kőolaj valószínűsíthető ára az egyik legfontosabb és legnagyobb kihívást jelentő kérdéssé vált ezen kutatások területén. Az árak trendjének és ingadozásának kiszámíthatósága mindig is kihívást jelentett a befektetők és a kereskedők számára az olajpiacon. Egyrészt a pontos előrejelzés fontos a magánbefektetők és a központi bankok számára, mivel ezek pontos szakszerűségére van szükség ahhoz, hogy megfelelő politikát alakítsanak ki az olajjal kapcsolatos sokkra válaszul. Másrészt az olajárak ingadozása fontos az importőr és termelő országok számára. (Drachal, 2016) Ily módon az olajárak megbízható vélhető adataira van szükség a csoportok széles körében.

A kőolaj nagy jelentőséggel bír a világgazdaságban, az IEA (2020) szerint 2020-ban a világ primer energiájának több mint 31%-át teszi ki, és a legtöbbet a közlekedésben használják fel. Ebben az aspektusban Kilian és Park (2009) azt állítja, hogy az olajkereslet és az olajkínálat sokkjai felelősek az amerikai részvények reálhozamának hosszú távú ingadozásának 22%-ért. Az olajpiaci sokkok okozati összefüggésére és a tőzsdei hozamokra vonatkozó eredményeket Salisu, Raheem és Ndako (2019) is kiemeli.

Hamilton (1996) szerint az olajpiaci sokkok számos csatornán keresztül befolyásolhatják a makrogazdasági változókat, például a szállítási költségek növelését, így az infláción keresztül az egész gazdaságot. Cunado és Gracia (2003) megemlíti továbbá, hogy az olajár volatilitása hatással lehet az árfolyamra, így a nettó kereskedelmi mérlegre. Bár ennek az árucikknek a nagy jelentősége, az olajárak prognózisa nehéz feladat, ha sok változót vagy módszertani megközelítést kell figyelembe venni. Yoshino és Taghizadeh-Hesary (2014) azt írja, hogy a nagy bizonytalanság pillanataiban, mint például a 2008-as másodlagos jelzálogpiaci válság idején, az előrejelzés pontossága veszélybe kerülhet, és a modelleknek figyelembe kell venniük az ilyen eseményeket.

Ebben az értelemben az olajár-előrejelzések szakirodalma folyamatosan fejlődik, mindig megpróbálja az ingadozások forrásait jobban magyarázni, és pontosabb előrejelzést készíteni.

Például 2014-ben, amikor az olaj ára 50,00 USD alá esett. Baumeister és Kilian (2014) szerint ennek oka a líbiai olajtermelés fellendülése és az iraki termeléscsökkentés. Míg Kilian és Murphy (2014) rámutatott, hogy ez a helyettesítő Egyesült Államok palaolaj-termelésének növekedésével is összefügg.

Az olajár változása jelentős hatással van a világgazdaságra több okból kifolyólag:

- **Energiaforrásként való fontosság**: Az olaj a világ legfontosabb energiaforrásai közé tartozik, és számos iparág, közlekedési mód, valamint fűtési és áramtermelési folyamatok alapvető része. Az olajár változása közvetlen hatással van ezeknek az iparágaknak a költségeire.
- **Termelési költségek**: Az olajárak emelkedése növeli a szállítási és gyártási költségeket, mivel sok termék előállítása és szállítása olaj alapú energiát igényel. Ezek finanszírozásagyakran áthárul a fogyasztókra, ami inflációhoz vezethet.
- **Infláció és kamatlábak**: Amikor az olajárak emelkednek, az infláció növekedhet, mivel a termékek és szolgáltatások árai is emelkednek. Az infláció növekedése nyomást gyakorolhat a központi bankokra, hogy növeljék a kamatlábakat az infláció megfékezése érdekében, ami viszont csökkentheti a gazdasági növekedést.
- **Fogyasztói költés**: Az olajárak emelkedése növeli a közlekedési és fűtési költségeket a háztartások számára, ami csökkentheti a fogyasztók rendelkezésre álló jövedelmét. Ez kevesebb költést jelent más termékekre és szolgáltatásokra, ami lassíthatja a gazdasági növekedést.
- **Nemzetközi kereskedelem**: Az olajimportőr és -exportőr országok külgazdasági egyensúlyára is nagy hatással van az olajár változása. Az olajexportáló országok, mint például Szaúd-Arábia, jelentős bevételeket szereznek az olaj eladásából, így az árak csökkenése negatívan befolyásolja a gazdaságukat. Az olajimportáló országok, mint például Japán vagy Németország, viszont profitálhatnak az alacsonyabb olajárakból.
- **Valutapiacok**: Az olajárak változása befolyásolja a valuták árfolyamát is, különösen azokban az országokban, amelyek jelentős mértékben függnek az olajexportból származó bevételektől. Az olajárak esése például gyengítheti ezeknek az országoknak a valutáját.
- **Befektetések és tőzsdék**: Az olajárak ingadozása befolyásolja a részvény- és kötvénypiacokat is. Az olajipari cégek részvényeinek értéke gyakran szorosan követi az olajárakat, így az olajárak esése csökkentheti ezen cégek részvényárfolyamát, míg az emelkedés növelheti.

Az olajnak a világgazdaságban betöltött fontos szerepe miatt nagy mennyiségű kutatás foglalkozott az olajár változásának gazdasági és pénzügyi következményeivel. Számos kutatás utal arra, hogy az olajár-sokkok statisztikailag szignifikáns és negatív hatást gyakorolnak a reálgazdasági tevékenységre. (Hamilton, 1983) (Hamilton, 2011) (Cunado & Garcia, 2003) (Cunado & Perez de Garcia, 2005) (Herrera, Lagalo, & Wada, 2011) (Jo, 2014) (Cunado, Jo, & Perez de Garcia, 2015)

Kevésbé kiterjedt, de még mindig jelentős mennyiségű kutatás vizsgálta az olajár-sokkok nemzetközi részvénypiacokra gyakorolt hatását. Ennek gazdasági hatásait a pénzügyi piacoknak meg kell ragadniuk, mivel ezek befolyással bírnaka pénzáramlásokra, a várható hozamokra és a befektetési döntésekre. Az empirikus vizsgálatok eredményei azt mutatják, hogy az olajár-sokkok,

amelyeket olajár-változásként definiálnak, jelentős hatást fejtenek ki ezekre a piacokra. (Jones & Kaul, 1996) (Sadorsky, 1999) (Cunado & Perez de Gracia, 2013) Egy friss tanulmányában Jo (2014) rámutat arra, hogy a megnövekedett instabilitás ronthat számos reálgazdasági tevékenységet, és a sztochasztikus volatilitású negyedéves vektor-autoregressziós modell (VAR) segítségével megállapítja, hogy az olajár ingadozása negatív hatással van a világ ipari termelésére. Elder és Serletis (2010) egy feltételes heteroszkedaszticitású általánosított autoregressziós modell (GARCH) segítségével tanulmányozza az olajár váltakozásának makrogazdasági hatásait, negatív összefüggést találva a beruházások, a tartós fogyasztási cikkek és a fogyasztás mértékében.

Néhány tanulmány az olajár-emelkedés GDP-re gyakorolt hatását vizsgálta strukturális modellek, különösen általános egyensúlyi (CGE) modellek segítségével. Sanchez (2011) például dinamikus CGE-modell segítségével kimutatja, hogy a 2002–2008-as időszakban az olajár-emelkedés hat olajimportáló országban (Bangladesh, El Salvador, Kenya, Nicaragua, Tanzánia é Thaiföld) évi 2-3%-os (2008-ban) GDP-csökkenést okozott. A CGE modellt használva Aydın és Acar (2011) úgy találja, hogy a magasabb olajárak rövid távon jelentős negatív hatást gyakorolnának a török gazdaságra, bár a gazdaság hosszú távon alkalmazkodna, és a hatások enyhébbek lennének. A 2020-ban hordónkénti 185 USD-t elérő magasabb olajárfolyam a GDP 1,3%-os növekedését okozza éves szinten ahhoz a fogatókönyvhöz képest, amelyben az olajár 2020-ban csak a 108 USD hordónkénti árat éri el. Az elemzés a 2010–2020-as időszakra készült, a GDP-hatások pedig rövid távon (2011 és 2012) 2,3%, illetve 2,3% voltak. Sztochasztikus dinamikus általános egyensúlyi (DSGE) modellt használva Balke et al. (2010) arra jutottak, hogy az olajárak az 1990-es évek óta viszonylag gyengébb hatást gyakoroltak az USA GDP-jére, mint az 1970-es és 1980-as években. Arra a következtetésre jutottak, hogy az Egyesült Államok GDP-jének újabb ingadozásait elsősorban a hazai mozgatórugók magyarázzák, nem pedig az olajár-sokkok.

A klasszikus kínálati oldali közgazdasági elmélet azt sugallja, hogy az olajár-sokk visszafogja a makrogazdasági tevékenységszintet, mivel az olajár emelkedése magasabb termelési költségekhez, a termelékenység csökkenéséhez, végül az egy főre jutó bruttó hazai termék (GDP) csökkenéséhez, ill. egy főre jutó kibocsátás csökkenéséhez vezet. Ez a probléma a munkanélküliségi ráta növekedéséhez és a hazai hitelállomány csökkenéséhez is vezethet a cégek alacsonyabb makrogazdasági aktivitása miatt. Valójában a munkanélküliség az olajár-sokkból fakadhat az olajintenzív iparágakban a foglalkoztatási struktúrák változásán keresztül, ami arra készteti a cégeket, hogy olyan termelési módszereket alkalmazzanak, amelyek kevésbé függenek az olajtól, és ami a munkaerő szektorok közötti széles körű átcsoportosítását eredményezi, ami hosszú távon érinti a munkanélküliséget. (Chang, Jha, Fernandez, & Jam'an, 2011)

Ezek a tényezők összességében magyarázzák, hogy miért van az olajár változásának nagy hatása a világgazdaságra. Az olaj, mint alapvető energiaforrás és gazdasági tényező, szorosan összefonódik számos gazdasági mutatóval és folyamatokkal.

3.2 Olajár előrejelzés

Széles körben elterjedt és vizsgált, hogy az olajár váratlan nagy és tartós ingadozásai mind az olajimportáló, mind az olajtermelő gazdaságok jólétét jelentős mértékben káros módon befolyásolják, ezért számos módon készítenek prognózisokat a kőolaj árának jövőbeli alakulásához kapcsolódóan. A központi bankok, illetve a magánszektor elemzői sok esetben az olaj árát nevezik meg a makrogazdasági előrejelzések készítésének és a makrogazdasági

kockázatok felmérésének egyik kulcsfontosságú, ha nem a legfontosabb változójának. Különösen érdekes az a kérdés, hogy az olaj ára mennyiben segíti a recesszió kellő időben történő prognózisát. Több kutatásban találunk erre bizonyítást, például Hamilton (2009) az Edelstein és Kilian (2009) kutatásaira építve, mely szerint a 2008 végi recesszió felerősödött, és megelőzte az autóipar gazdasági lassulása és a fogyasztói hangulat romlása.

Nem csupán az olaj árának pontosabb előrejelzései valószínűsítik, hogy javítják a makrogazdasági eredményekre vonatkozó előrejelzések precizitását, hanem emellett a gazdaság egyes szegmensei és ágazatai közvetlenül függenek az olajár változásától. Ilyenek például a légitársaságok, akik ezekre az adatokra támaszkodnak a repülőjegy árak meghatározásakor, az autógyártó cégek szintén figyelembe veszikv termékkínálatukvalamint a termékáraik meghatározásánál, a közműcégek pedig az olajár-előrejelzések alapján döntenek a kapacitásbővítésről vagy új üzemek létesítéséről. Hasonlóképpen, a lakástulajdonosok is a fűtőolaj vásárlásának időpontjánál vagy az energiatakarékos lakásfejlesztések befektetésénél veszik figyelembe döntések meghatározó kritériumaként.

Mindezek mellett az olaj árának és származékainak (például a benzin vagy a fűtőolaj) árának valószínűsíthető alakulása fontos az energiaintenzív tartós fogyasztási cikkek, például autók vagy otthoni fűtési rendszerek vásárlásának modellezésében. Szerepet játszanak az energiafelhasználás előrejelzéseinek elkészítésében, az energiaszektorban meghozott befektetési döntések modellezésében, a szén-dioxid-kibocsátás és az éghajlatváltozás előrejelzésében, valamint olyan szabályozási politikák kialakításában, mint például az autóüzemanyag-szabványok vagy a benzinadók kivetése.

Különféle elméleti megközelítések léteznek az olajárak előrejelzési modellezésére. A kőolajár-előrejelzési szakirodalomban a módszereket két fő csoportba sorolhatjuk. Az első csoport a hagyományos statisztikai és ökonometriai technikák, úgy, mint az exponenciális simítási modell (ESM), a lineáris regresszió (LinR), az autoregresszív integrált mozgóátlag (ARIMA), az általánosított autoregresszív feltételes heteroszkedaszticitás (GARCH), a bolyongási folyamat (RW) és a hibakorrekciós modellek (ECM). A hagyományos statisztikai és ökonometriai technikák többnyire csak lineáris folyamatokat képesek adat-idősorokban rögzíteni. Ezek a modellek ebből adódóan nem elegendőek a kőolajárak nemlineáris jellemzőinek figyelembevételéhez. Ennek a korlátnak a leküzdése céljából a mesterséges intelligencia (AI) modellek erőteljes öntanuló képességekkel, például mesterséges neurális hálózatokkal (ANN), támogatott vektorgépekkel (SVM) és intelligens optimalizáló algoritmusokkal, például genetikai algoritmusokkal (GA) egyre népszerűbbek a nyersolajár-előrejelzésben. (Yu, Dai, & Tang, 2016) (Hamdi & Aloui, 2015). Az elmúlt évtizedekben sokan foglalkoztak a témával, számos tanulmány készült az olajár előrejelzéséről. A témában elsőként Amano (1987) kutatott. A szerző egy kis léptékű ökonometriai modellt alkalmazott az olajpiaci előrejelzéshez. Tang és Hammoudeh (2002) nemlineáris regressziót használt az OPEC kosárárának előrejelzésére. Ye et al. (2006) a WTI árak egyszerű ökonometriai modelljét mutatta be magas, illetve alacsony készletű változók felhasználásával. Gori et al. (2007) az adaptív neuro-fuzzy következtetési rendszert (ANFIS) használta a havi olajárak előrejelzésére. Moshiri és Foroutan (2006) az ARIMA és a GARCH modellek alkalmazásával modellezte és előrejelezte a napi határidős kőolajárakat. Xie et al. (2006) pedig a WTI kőolajárakat az ARIMA módszer alkalmazásával.

Összevetették az eredményeket az SVM-ek és az ANN-ok eredményeivel is. Yu et al. (2008) egy empirikus módusz-dekompozíción (EMD) alapuló neurális hálózatok együttes tanulási

modelljét javasolta a nyersolaj azonnali világpiaci árának előrevetítéséhez. Kulkarni és Haidar (2009) prezentált egy többrétegű előrecsatolt neurális hálózatot (FNN – feedforward neural network) a kőolaj azonnali árának valószínű alakulására. Bao et al. (2011) egy wavelet transzformáción és a legkisebb négyzetek támogatási vektoros gépeken (LSSVM) alapuló hibrid modellt javasolt erre a WTI és a Brent kőolajárak esetében. He et al. (2012) bevezetett egy wavelet dekompozíciós ensemble modellt a kőolajár-prognózis pontosságának javítására. Azadeh et al. (2012) mesterséges neurális hálózaton és fuzzy regresszión alapuló rugalmas algoritmust használt. Khashman és Nwulu (2011) összehasonlító elemzést végeztek az SVM-ről és a visszacsatolásról. Ahmed és Shabri (2014) egy technikát javasoltSVM segítségével. Yu et al. (2015) egy dekompozíció-együttes módszertant mutatott be adatok karakterisztikus vezérelt rekonstrukciójával a WTI és a Brent nyersolaj azonnali árának előrejelzésére. Tang et al. (2015) bemutatott egy újszerű ensemble learning paradigmát, amely összekapcsolja a komplementer ensemble empirikus módú dekompozíciót (CEEMD) és a kiterjesztett extrém tanulási gépet (EELM), hogy javítsa a kőolaj valószínűsíthető árának pontosságát. Yu et al. (2016) egy új hibrid tanulási paradigmát mutatott be szintén ezzel a céllal, azaz a hibrid grid-GA-alapú legkisebb négyzetes támogatási vektor regressziós (LSSVR) modellt a West Texas Intermediate és a Brent piacok esetében. Zhao et al. (2017) ugyanakkor az SDAE-B elnevezésű mély tanulási együttes megközelítést tartotta megfelelő eljárásnak e célra. Yu et al. (2017) az LSSVR ensemble learning paradigmáját javasolta bizonytalan paraméterekkel a WTI nyersolaj azonnali árának előrejelzésére.

Az idősoros modellek egyik célja az előrejelzés. Az idősoros modellek történelmi adatok alapján jósolják meg a jövőbeli olajárakat. Ezekben a modellekben a jövőbeni árviselkedés a saját történeti adatokból következik. Ezeket a modelleket leginkább akkor alkalmazzák, amikor az adatok szisztematikus mintát mutatnak, amikor a legtöbb lehetséges magyarázó változó és kölcsönhatásaik olyan strukturális modellt adnak, amelyet nagyon nehéz követni, vagy amikor egy függő változó előrejelzése a magyarázó változók előrejelzésétől függ, amelyek bonyolultabbak lehetnek, mint magát a változót előre jelezni.

Lineáris és nemlineáris modelleket gyakran használtak idősoros modellezéshez. Azonban nehéz diagnosztiálni egy korreláció linearitását vagy nemlinearitását, és nincs teljes lineáris vagy nemlineáris korreláció (gyakran a kettő kombinációja). Ezen kívül általános vélemény az, hogy egy modell nem tudja bevonni az összes tényezőt és összefüggést az idősoros adatokban. Ráadásul az előrejelzési szakirodalomban szinte általánosan elfogadott az a vélemény, hogy egyetlen modell sem a legjobb minden helyzetben, mert egy valós probléma gyakran összetett természetű. (Khashei & Bijari, 2011) Ezért javasolt a rendelkezésre álló típusú előrejelzési modellek kombinációjának használata az egy modell helyett. (Timmermann, 2006) Bates és Granger (1969) voltak az elsők, akik a modellek kombinációját tanulmányozták. Ebben a tekintetben számos tanulmány kombinált lineáris és nemlineáris modelleket. Úgy, mint Stock és Watson (2004) lineáris és nemlineáris előrejelzési modelleket használt a pénzügyi és gazdasági változók tanulmányozására, és arra a következtetésre jutott, hogy a modellek kombinációja jobban teljesít, mint egy egyedi modell. Teräsvirta (2006) eredményei megerősítik, hogy a lineáris és nemlineáris modell kombinációjával történő előrejelzés jobb eredményeket ad, mint egyetlen nemlineáris modell használata. Ezért a különböző modellek kombinálása hatékony módja lehet az egyes modellek előrejelzési teljesítményének javításának. Ezenkívül az eltérőmodellek kiegészíthetik egymást, hogy különböző kapcsolatokat és mintákat rögzítsenek az idősoros adatokban.

A hibrid módszerek gyakran olyan interdiszciplináris módszerek kombinációját jelentik, amelyek egyesítik erősségeiket, és hozzávetőlegesen három kategóriába sorolhatók:

1. soft-computing módszerek kombinációja, mint például az intelligens optimalizálási algoritmusok, GA, SVM;
2. ökonometriai módszerek, például GARCH, ARIMA kombinációja;
3. soft-computing és ökonometriai módszerek kombinációja, például a GARCH és az ANN módszer. (Zhang, Zhang, & Zhang, 2015)

Habár az ökonometriai és a lágy számítási módszerekből származó előrejelzési modellek kombinálása többnyire jobban teljesít, azonban az egyes modellek súlytényezőinek meghatározása kulcsfontosságú lépés a kombinált modell felépítésében. Draper (1995), Leamer (1978), Strachan és Van Dijk (2008) Bayes-féle átlagoló modelleket használt a kombinált modellek kialakításához. Más tanulmányok állandó súlyok használatát javasolják, mások pedig időben változó súlyokat. Terui és Van Dijk (2002) tanulmányaiban állandó koefficiens regressziót és időben változó módszert alkalmazott a lineáris és nemlineáris előrejelző modellek kombinálására. Arra jutottak, hogy a kombinált előrejelzési modellek megfelelően teljesítettek, különösen az időben változó együtthatókkal. Hendry és Clements (2004) kimutatta, hogy egy egyszerű kombináló módszer, az egyszerű átlagolás megfelelő működést mutatott a fejlett kombinálási módszerekkel összehasonlítva (amelyekben a súlyok az előrejelzési hiba kovarianciamátrixától függenek). Guidolin és Timmermann (2007) olyan modellt javasolt, amelyben a súlyok dinamikus rezsimváltásból származnak. Egyes kutatók optimalizáló algoritmusokat is használtak az egyes modellek optimális súlyának meghatározásához. Például Wang et al. (2010) adaptív részecskeraj-optimalizálást alkalmaztak a kombináció optimális tömegének eléréséhez.

A kőolaj-árelőrejelzés számos módja ismert. A többféle tanulmány, számítási és prognosztikai módszer azért alakult ki, mert a befolyásoló tényezők nagyon széleskörűek, különböző súlyokkal szerepelnek és a legtöbb esetben a múltbeli események feldolgozásánál működnek jól, a jövőbeli eseményeknél már nagy hibahatárokkal dolgoznak. Ezért a kutatók folyamatosan újabb és újabb eljárásmóddal próbálkoznak, hogy még hatékonyabban tudják előre jelezni az olajárakat.

3.3 Python

A Python nyelvet jelen viszonylatban csak a későbbi felhasználhatóság miatt vizsgálom, hogy képbe kerüljünk a szerkezettel, illetve a későbbi metodikával.

A Raw Python, azaz önmagában a Python korlátozott képességekkel rendelkező nyelv, de a számos elérhető modul közül egy vagy több importálásával bővíthető. A Python jelenlegi verziói letölthetők az internetről, és mindegyik ingyenes. Mindegyikhez kiváló online dokumentáció tartozik, beleértve egy oktatóanyagot is.

A Python egy nagy teljesítményű, objektumorientált képességekkel rendelkező, magas szintű programozási nyelv, amelyet az 1990-es évek elején Guido van Rossum, az amszterdami Holland Nemzeti Matematikai és Számítástechnikai Kutatóintézet (CWI) akkori programozója tervezett és fejlesztett. Az alap Python disztribúció nyílt forráskódú, és több platformon is elérhető, többek között a Windows, Linux/Unix és Mac OS X rendszereket. Az alapértelmezett CPython implementáció, valamint a szabványos könyvtárak és dokumentáció ingyenesen elérhető a www.python.org webhelyről, melyeket a Python Software Foundation, egy non-profit szervezet

kezel. Van Rossum a továbbiakban is felügyeli a nyelvi fejlődést, amely biztosította a funkciók, a design és a filozófia erős folytonosságát és egy irányba fejlődését. A Python könnyen megtanulható és használható, nagyon világos, tömör és logikus szintaxisáról ismert. Ez a funkció önmagában különösen alkalmassá teszi a gyors szoftverprototípus-készítésre, és nagyban megkönnyíti a későbbi programkarbantartást és hibakeresést, valamint a szerző vagy más felhasználó általi bővítést. (Bilina & Lawford, 2012)

A Python programozási nyelv óriási népszerűségre tett szert a statisztikusok és szoftverfejlesztők körében. (Robinson, 2017.) A főként statisztikai adatelemzésre szánt R programozási nyelvtől eltérően a Python sokkal szélesebb körű alkalmazásokban jelenik meg, mint például az internet- és webhelyfejlesztés, az adatbázis-hozzáférés, az asztali grafikus felhasználói felületek, a tudományos számítások, valamint a szoftver- és játékfejlesztés. Két fő Python-verziósorozat létezik, a 2.x és a 3.x verzió, és ezek nem teljesen kompatibilisek, bár a legtöbb részük hasonló. A 2.x-es verzió egy örökölt verzió, amelynek támogatása és karbantartása a tervek szerint 2020 körül véget ért. A 3.x-es verzió a 2.x-es verzión alapuló újra tervezés és a Python jövőjének tekinthető.

Főbb jellemzők:

- Olvashatóság és egyszerűség: A Python szintaxisát úgy tervezték, hogy intuitív legyen, kódja pedig könnyen olvasható, így könnyen megtanulható és használható.
- Értelmezett nyelv: A Python kódot soronként hajtják végre, ami lehetővé teszi az interaktív tesztelést és hibakeresést.
- Dinamikus gépelés: A Python változóinak nincs szükségük kifejezett deklarációra a memóriaterület lefoglalásához, és a típusok dinamikusan következtetnek futás közben.
- Magas szintű adatstruktúrák: A Python hatékony adatstruktúrákat tartalmaz, például listákat, szótárakat, készleteket és sorokat, amelyek megkönnyítik az adatok kezelését és tárolását.
- Kiterjedt szabványos könyvtár: A Python átfogó szabványos könyvtárral büszkélkedhet, amely számos általános programozási feladatot támogat, például fájl I/O-t, rendszerhívásokat és internetes protokollokat.
- Platformok közötti kompatibilitás: A Python platformfüggetlen, ami azt jelenti, hogy a Pythonban írt kód módosítás nélkül futhat különféle operációs rendszereken.
- Nagy ökoszisztéma és közösség: A Python harmadik féltől származó könyvtárak és keretrendszerek hatalmas gyűjteményével rendelkezik (például NumPy, pandas, Django, Flask és TensorFlow), valamint egy nagy, aktív közösséggel, amely hozzájárul a fejlesztéséhez és támogatásához.

Gyakori felhasználások:

- Webfejlesztés: Az olyan keretrendszerek, mint a Django és a Flask, népszerűek robusztus webalkalmazások készítésére.
- Adattudomány és gépi tanulás: Az olyan könyvtárak, mint a NumPy, a pandas, a Matplotlib és a scikit-learn, a Pythont vezető nyelvvé teszik az adatelemzés, -vizualizáció és gépi tanulás terén.
- Automatizálás és szkriptelés: A Python-t gyakran használják az ismétlődő feladatok automatizálására és a rendszeradminisztrációs szkriptek írására.

- Szoftverfejlesztés: A Python világos szintaxisa és hatékony könyvtárai támogatják a szoftverprototípusok és alkalmazások gyors fejlesztését.
- Oktatás: A Python-t széles körben használják tanítási nyelvként, egyszerűsége és könnyű megtanulása miatt.
- Objektum-orientált és funkcionális: A Python támogatja mind az objektum-orientált, mind a funkcionális programozási paradigmákat, rugalmasságot biztosítva a fejlesztőknek a kódírás során.
- Közösség és támogatás: A Python Software Foundation (PSF) felügyeli a Python fejlesztését, és elősegíti növekedését és elfogadását.

Összességében elmondható, hogy a Python egyszerűségének, sokoldalúságának és hatékony funkcióinak ötvözete az alkalmazások széles skálájához használható nyelvvé teszi, és népszerű választás a fejlesztők körében szerte a világon.

A Python nem lefordított nyelv, ami azt jelenti, hogy nem fordítja le binárisra előre a kódot. Ehelyett egy szoftverkörnyezet, a Python interpreter lefordítja a szkriptet binárisra a kód valós idejű végrehajtása során. Elosztásával a Python néhány alapvető funkcióval rendelkezik, de szinte minden numerikus számítás elvégzéséhez külső csomagokra támaszkodik. Az elmúlt 10 év természetes szelekciós folyamata után néhány alapvető számítási képességet biztosító csomag széles körben elfogadott a Python közösségben. (Hao & Ho, 2019) Értelmezett vagy lefordított: mint értelmezett nyelv, a Python programokat közvetlenül az értelmező hajtja végre, ami lassabb lehet, mint a lefordított nyelvek, de nagyobb rugalmasságot és egyszerűbb használatot tesz lehetővé.

A kutatás során használt főbb python könyvtárak, csomagok a teljesség igénye nélkül:

A. **Torch**: nyílt forráskódú gépi tanulási könyvtár, egy tudományos számítási keretrendszer és egy Lua programozási nyelven alapuló szkriptnyelv. Algoritmusok széles skáláját kínálja a mély tanuláshoz, és a LuaJIT szkriptnyelvet és a mögöttes C implementációt használja. Az EPFL IDIAP-nál hozták létre. 2018-tól a Torch már nincs aktív fejlesztés alatt. A Torch könyvtáron alapuló PyTorch azonban 2021 júniusától aktívan fejlődik. (Torch - A Scientific Computing Framework For LuaJIT, 2020.)
B. **PyTorch**: egyedülálló módon építi fel a neurális hálózatokat: egy „magnót” használ és játszik le. A legtöbb keretrendszer, például a TensorFlow, a Theano, a Caffe és a CNTK statikus világnézettel rendelkezik. Fel kell építeni egy neurális hálózatot, és újra és újra fel kell használni ugyanazt a struktúrát. A hálózat viselkedésének megváltoztatása azt jelenti, hogy a nulláról kell kezdeni. A PyTorch esetében a fordított módú automatikus differenciálásnak nevezett technikát használjuk, amely lehetővé teszi a hálózat viselkedésének tetszőleges megváltoztatását nulla késleltetés vagy többletterhelés nélkül. (Yegulalp, 2017.) (Ketkar, 2017)
C. **Torch.nn**: a PyTorch nn modul magas szintű API-kkal rendelkezik a neurális hálózat felépítéséhez. A Torch.nn modul Tenzorokat és Automatikus differenciálási modulokat használ az olyan rétegek betanításához és felépítéséhez, mint a bemeneti, rejtett és kimeneti rétegek. A Pytorch egy torch.nn alaposztályt használ, amely paraméterek, függvények és rétegek becsomagolására használható a torch.nn modulokba. Bármely mély tanulási modellt a torch.nn modul alosztályával fejlesztenek ki, olyan metódust használ, mint a forward(input), amely

visszaadja a kimenetet. Egy egyszerű neurális hálózat bemenetet vesz fel, hogy súlyokat és torzításokat adjon hozzá, több rejtett rétegen keresztül táplálja a bemenetet, és végül visszaadja a kimenetet. (PyTorch - Torch.NN documents, 2022.)

D. **Pandas**: alkalmas a Python programozási nyelvhez írt szoftverkönyvtár adatkezelésre és elemzésre. Különösen adatstruktúrákat és műveleteket kínál numerikus táblák és idősorok manipulálásához. Három szakaszból álló BSD licenc alatt adtak ki. (Pandas 1.0.0 documentation, 2021.) A név a "paneladatok" kifejezésből származik, amely egy ökonometriai kifejezés olyan adatkészletekre, amelyek ugyanazon személyek több időszakra vonatkozó megfigyeléseit tartalmazzák. (McKinney, 2011.) A neve magával a „Python adatelemzés” („Python data analysis”) kifejezéssel játszik.
E. **Numpy**: egy Python-könyvtár, amely többdimenziós tömbobjektumot, különféle származtatott objektumokat (például maszkolt tömböket és mátrixokat) biztosít, valamint rutinok választékát a tömbök gyors műveleteihez, beleértve a matematikai, logikai, alakmanipulációt, rendezést, kijelölést, I/O-t, diszkrét Fourier transzformációk, alapvető lineáris algebra, alapvető statisztikai műveletek, véletlenszerű szimulációkat és még sok mást. (Numpy documents, version: 1.22, 2022.)
F. **Newspaper3k**: egy Python-könyvtár, amelyet webes cikkek scrapelésére használnak. Használja a requests könyvtárát, és a BeautifulSoup függőséget használja, miközben lxml-re értelmezi. A Newspaper3k nem csak a cikk teljes szövegét képes letölteni, hanem más típusú adatokat is lekérhet, mint például a közzététel dátuma, szerző(k), URL, képek és videó, hogy csak néhányat említsünk. Ha egyszerűen csak tudni szeretnénk, miről szól a cikk anélkül, hogy a teljes cikket el kellene olvasnia, a Newspaper3k összefoglalót is készíthet a cikkről. (Python Package Index - newspaper3k 0.2.8, 2022.)
G. **BeautifulSoup**: egy Python-könyvtár, amellyel adatokat vonhatunk ki HTML- és XML-fájlokból. Egy összetett HTML-dokumentumot Python-objektumok összetett fájává alakít. A létrehozott fa modell az elemzett oldalak számára, mely segítségével adatok kinyerhetők a HTML-ből, ami hasznos a web scrapeléshez. (Beautiful Soup Documentation, 2022.) (Hajda, 2018)
H. **Matplotlib:** a Python programozási nyelv és annak NumPy numerikus matematikai kiterjesztésének ábrázoló könyvtára. Objektum-orientált API-t biztosít a ábrák és kimutatások alkalmazásokba való beágyazásához olyan általános célú grafikus felhasználói felület eszközkészletek használatával, mint a Tkinter, wxPython, Qt vagy GTK. Létezik egy állapotgépen (például az OpenGL-en) alapuló procedurális "pylab" interfész is, amelyet úgy terveztek, hogy nagyon hasonlítson a MATLAB-hoz. (Matplotlib 3.5.1 documentation, 2022.)
I. **Scikit-learn:** egy szoftver gépi tanulási könyvtár a Python programozási nyelvhez. Különféle osztályozási, regressziós és klaszterezési algoritmusokat tartalmaz, beleértve a támogatási vektor gépeket, véletlenszerű erdőket, gradiens-növelést, k-means-t és DBSCAN-t, és úgy tervezték, hogy együttműködjön a Python NumPy és SciPy numerikus és tudományos könyvtárakkal. (scikit-learn - Machine Learning in Python, 2022.)

3.3.1 Scraper

A rengeteg adat jelenléte a világhálón egyszerre lehet áldás és átok. Hatalmas forrásként sok releváns információ kerül bemutatásra, azonban ezeknek az információknak a megszerzése kihívást jelent. A weben a legtöbb információ HTML-dokumentumként jelenik meg, amely egy vázlat, vagyis az adatok strukturálatlan adatok. Az adatok növekedési üteme az interneten évről évre szárnyal, és az adatok túlnyomórészt nem alkotnak rendszert. Mivel a rendszerezetlen adatok semmilyen adatmodellt nem követnek, az információcsere nem egyszerű.

A web scrapelés (webkaparás magyarosan), vagyis a webes adatgyűjtés vagy a webes adatkinyerés olyan adatszerzési módszer, amelyet a webhelyekről származó adatok kinyerésére használnak. A web scraper szoftver közvetlenül hozzáférhet a világhálóhoz a Hypertext Transfer Protocol vagy egy webböngésző használatával. Míg az adatgyűjtést a szoftverhasználó manuálisan is elvégezheti, ez a kifejezés általában bottal vagy webrobottal megvalósított automatizált folyamatokra utal. Ez egy olyan másolási forma, amelyben meghatározott adatokat gyűjtenek össze és másolnak a webről, jellemzően egy központi helyi adatbázisba vagy táblázatba, későbbi visszakeresés vagy elemzés céljából.

Egy weboldal internetes scrapelés magában foglalja annak lekérését és kibontását. A lekérés egy oldal letöltése (amit a böngésző akkor hajt végre, amikor a felhasználó megtekint egy oldalt). Ezért a webes feltérképezés a webscrapelés egyik fő összetevője, amely az oldalakat későbbi feldolgozás céljából lekéri. A lekérés után megtörténhet a kivonás. Egy oldal tartalma elemezhető, kereshető, újraformázható, adatai táblázatba másolhatók vagy adatbázisba tölthetők. A webscraperek általában kivesznek valamit az oldalról, hogy más célra használják fel. Példa erre a nevek és telefonszámok, vagy cégek és URL-címeik vagy e-mail címeik listába másolása (névjegyek összegyűjtése).

A webscrapelést a kapcsolatfelvételhez, valamint a webindexeléshez, webbányászáshoz és adatbányászathoz, online árváltozás figyeléshez és ár-összehasonlításhoz, termékismertető kaparáshoz (a verseny megtekintéséhez), ingatlanhirdetések, időjárási adatok gyűjtéséhez használt alkalmazások összetevőjeként használják (figyelés, webhelyváltozás észlelése, kutatás, online jelenlét és hírnév nyomon követése, webes mashup és webes adatok integrációja).

A weboldalak szöveges jelölőnyelvek (HTML és XHTML) használatával készülnek, és gyakran rengeteg hasznos adatot tartalmaznak szöveges formában. A legtöbb weboldal azonban emberi végfelhasználók számára készült, nem pedig az automatizált használat megkönnyítésére. Ennek eredményeként speciális eszközöket és szoftvereket fejlesztettek ki, amelyek megkönnyítik a weboldalak feltérképezését.

A webscrapelés újabb formái magukban foglalják a webszerverekről érkező adatfolyamok figyelését. Például a JSON-t általában szállítási tárolási mechanizmusként használják az ügyfél és a webszerver között.

Vannak olyan módszerek, amelyeket egyes webhelyek használnak az adatgyűjtés megakadályozására, például észlelik és letiltják, hogy a robotok feltérképezzék (megtekintsék) az oldalaikat. Erre válaszul léteznek olyan webes adatgyűjtő rendszerek, amelyek a DOM-elemzés, a számítógépes látás és a természetes nyelvi feldolgozás technikáira támaszkodnak, hogy szimulálják az emberi böngészést, lehetővé téve a weboldal tartalmának összegyűjtését offline elemzéshez. (Hackinger, 2018) (Fayzrakhmanov, Sallinger, Spencer, Furche, & Gottlob, 2018)

A webscraping, a webhelyekről információk kinyerésére használt technika, amely automatizált scriptek használatát foglalja magában a strukturálatlan webadatok elemzésre és egyéb alkalmazásokra alkalmas egységes formátummá alakítására. Ez a gyakorlat a webtechnológiák és a programozás bonyolultságát kihasználva szisztematikusan gyűjti az adatokat, amelyeket aztán különféle célokra, például kutatásra, piacelemzésre és tartalom-összesítésre lehet felhasználni.

A web scraping alapvetően a Hypertext Transfer Protocol (HTTP) és a Hypertext Markup Language (HTML) elvein alapul. Amikor egy weblehúzó kérést kezdeményez egy weboldalhoz, HTTP-kérést küld, hasonlóan egy webböngészőhöz. A szerver az oldal HTML-tartalmával válaszol, amelyet aztán a lehúzó elemzi a kívánt információ kinyerése érdekében.

A folyamat általában a következő lépéseket tartalmazza:

1. Kérések küldése: A lehúzók HTTP-kérelmeket küldenek a cél URL-ekre olyan könyvtárak használatával, mint a Python kérései.
2. HTML lekérése: A szerver a weboldal HTML-tartalmával válaszol.
3. HTML elemzése: Az olyan könyvtárak, mint a BeautifulSoup vagy az lxml, elemzik a HTML-tartalmat, lehetővé téve a lehúzó számára, hogy navigáljon a dokumentum szerkezetében.
4. Adatok kinyerése: Adott adatpontok kibontása HTML-címkék, attribútumok és hierarchikus kapcsolatok alapján történik.
5. Adatok tárolása: A kivont adatokat ezután strukturált formátumban, például CSV-ben, JSON-ban vagy adatbázisban tárolják további elemzés céljából.

A webscraping számos tudományos és technikai kihívást, valamint megfontolásokat vet fel, többek között:

A HTML szerkezetének összetettsége: A webhelyek HTML összetettsége és felépítése eltérő, ezért a scrapereknek nagymértékben alkalmazkodóképességre van szükségük. A tartalom betöltéséhez JavaScriptet használó dinamikus webhelyek fejlett technikákat tesznek szükségessé, például fej nélküli böngészők (pl. headless selenium) vagy API-k használatát.

Az adatkészlet integritásának megőrzéséhez elengedhetetlen annak biztosítása, hogy a kimásolt adatok pontosak, teljesek és duplikátumoktól mentesek legyenek. Ez gyakran kifinomult hibakezelési és validálási mechanizmusokat igényel.

A web scrapingnek meg kell felelnie a jogi kereteknek és az etikai irányelveknek. A webhelyek szolgáltatási feltételeiről (ToS) kötött szerződések és a szellemi tulajdonjogok gyakran korlátozzák az adatlekopási tevékenységeket. Az etikai megfontolások közé tartozik a felhasználók adatainak tiszteletben tartása és a célszerverek túlzott terhelésének elkerülése is.

A szerverek túlterhelésének megelőzése és a webhelyek sebességkorlátozásainak betartása érdekében a lehúzók gyakran tartalmaznak olyan mechanizmusokat, amelyek késleltetik a kéréseket és szabályozzák a hozzáférés gyakoriságát.

A webhelyek scrapelését gátló technológiákat, például CAPTCHA-t, IP-blokkolást és dinamikus tartalomszolgáltatást alkalmazhatnak az automatikus hozzáférés megakadályozása érdekében. A kaparóknak ellenintézkedéseket kell kidolgozniuk, vagy alternatív adatforrásokat kell keresniük, ha ilyen akadályokkal szembesülnek.

A webscrapingnek sokféle alkalmazása van a különböző területeken:

- A scraping nagy adatkészletekhez biztosít hozzáférést a kutatóknak empirikus elemzés, természetes nyelvi feldolgozás és gépi tanulás céljából.
- A vállalatok a webscraping segítségével versenyinformációkat gyűjtenek, figyelemmel kísérik a piaci trendeket és elemzik a fogyasztói hangulatot.
- A hírgyűjtők és a közösségi média platformok webscrapinget használnak a több forrásból származó információk összeállítására és bemutatására.

Míg a web scraping az adatgyűjtés hatékony eszköze, elengedhetetlen, hogy körültekintően eligazodjon a kapcsolódó technikai, etikai és jogi kihívásokban. A legjobb gyakorlatok betartása, valamint az adatok tulajdonjogának és a magánélet tiszteletben tartása elengedhetetlen a webscraping technológiák felelős használatához.

A folyamat során más weboldalán szereplő adat, információ kerül összegyűjtésre, ami jogi kérdéseket is felvet. A magyar jogszabályok nem kifejezetten részletezik, ezért Európai Unió szinten vizsgálom a kérdéskört.

A webscraping az Európai Unióban (EU) különböző jogi megfontolások tárgyát képezi, elsősorban az adatvédelmi törvények, a szellemi tulajdonjogok és a szolgáltatási szerződések feltételei. A jogi környezetet olyan szabályozások alakítják, mint az Általános adatvédelmi rendelet (GDPR), valamint a szerzői jogokról és az adatbázis-jogokról szóló irányelvek.

A korábban hatályba lépett GDPR (2016) egy átfogó adatvédelmi szabályozás, amely a személyes adatok gyűjtését, feldolgozását és tárolását szabályozza az EU-n belül. Szigorú követelményeket ír elő a személyes adatok gyűjtésére és felhasználására vonatkozóan, ami jelentős hatással van a webscraping tevékenységekre is.

Személyes adat, minden olyan adat, amely közvetlenül vagy közvetve azonosíthatja az egyént, a GDPR hatálya alá tartozik. Az ilyen adatokat gyűjtő kaparóknak biztosítaniuk kell a GDPR követelményeinek való megfelelést.

Az adatfeldolgozásnak, beleértve a webscrapinget is, a GDPR értelmében törvényes alapon kell alapulnia. Ez magában foglalja az érintettek kifejezett hozzájárulásának megszerzését, a szerződés teljesítését vagy a jogos érdek fennállását. Az egyéneknek jogukban áll hozzáférni adataikhoz, helyesbíteni, törölni és korlátozni azok kezelését. A scrapereknek tiszteletben kell tartaniuk ezeket a jogokat, és olyan mechanizmusokat kell alkalmazniuk, amelyek megfelelnek az érintettek kérésének. A szervezeteknek átláthatónak kell lenniük adatgyűjtési gyakorlataik tekintetében, és biztosítaniuk kell az elszámoltathatóságot adatfeldolgozási tevékenységeik során.

Mivel nyilvános adatokat gyűjtünk, így ez a kutatást ezen pontokat nem érinti.

A webhelyek gyakran tartalmaznak szellemi tulajdonjoggal védett tartalmat, beleértve a szerzői jogokat és az adatbázisjogokat.

Szerzői jog az a webhelyen található tartalom, például szöveg, képek és videók, általában szerzői jogvédelem alatt állnak. Az ilyen tartalom engedély nélküli lekaparása és reprodukálása szerzői jogok megsértésének minősülhet. Az EU adatbázis-irányelve jogokat biztosít az adatbázis-készítők számára adatbázisaik tartalmának kinyerésére és újrahasznosítására. Az adatbázis jelentős részének engedély nélküli lekaparása sértheti ezeket a jogokat.

A webhelyek jellemzően szolgáltatási feltételekkel (ToS) rendelkeznek, amelyek felvázolják a webhely és annak tartalmának megengedett használatát. E feltételek megsértése jogi

következményekkel járhat. A webhely Általános Szerződési Feltételeit sértő tevékenységek lemásolása szerződésszegés miatti követeléseket vonhat maga után. A webhelyek kifejezetten megtilthatják az automatikus hozzáférést vagy az adatok kinyerését. E korlátozások figyelmen kívül hagyása jogi lépésekhez vezethet, beleértve a jogi lépéseket végrehajtást és a kártérítést.

Egyértelműen ezt a témakört alaposan körül kell járni, mert adott esetekben súlyos következményekkel is járhat.

Az EU-ban számos bírósági ügy foglalkozott a webscrapinggel, ami jogi precedenst jelent:

Ryanair kontra PR Aviation (2015): Az Európai Unió Bírósága (EUB) kimondta, hogy a Ryanair lekaparást tiltó feltételei a szerződési jog alapján végrehajthatók. Ez az eset rávilágított a webhely általános szerződési feltételeinek betartásának fontosságára.

Svensson kontra Retriever Sverige AB (2014): Az EUB tisztázta, hogy a szerzői jog által védett tartalomra való hivatkozás nem minősül szerzői jogsértésnek, ha a tartalom a nyilvánosság számára szabadon hozzáférhető. A tartalom lemásolása és újbóli közzététele azonban továbbra is sértheti a szerzői jogokat.

Alapvetően az EU-nak nem áll érdekében megakadályozni a webscrapinget. Másrészt bizonyos alapjogokat a folyamat során nem lehet felülírni. Következésképpen a webscraping előnyeit kihasználni kívánó vállalkozásoknak, egyéneknek minden tőlük telhetőt meg kell tenniük a törvények betartása érdekében, és – ami talán még ennél is fontosabb – jóhiszeműen kell eljárniuk a kaparás során bevált gyakorlatokat követve.

Az aggodalmak minimalizálása érdekében a kaparásnak diszkrétnek kell lennie, tiszteletben kell tartania a webhelyek szolgáltatási feltételeit, ellenőriznie kell, hogy a webhelyek a robots.txt protokollt használják-e annak közlésére, hogy a kaparás tilos, kerülje a személyes adatok lekaparását, és ha szükséges, győződjön meg arról, hogy nem sérti meg a GDPR-t. és kerülje a magánjellegű vagy titkos információk lekaparását.

A webscraping, az adatok webhelyekről történő kinyerésének automatizált folyamata, felbecsülhetetlen értékű eszközzé vált a tudományos kutatás és a személyes adatok elemzése számára. A webscraping jogszerűsége azonban, különösen az Európai Unión belül (EU-n belül), összetett környezetet mutat, amelyet többféle jogi keret alakít ki, beleértve a szellemi tulajdonjogokat, az adatvédelmi előírásokat és a webhelyek szolgáltatási feltételeit.

Az EU-ban a webscraping gyakran találkozik a szellemi tulajdonjogokkal, különösen az adatbázisokról szóló irányelvvel. (Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, 1996) Ez az irányelv jogi védelmet biztosít az adatbázisoknak, amelyek kiterjedhetnek a webhelyeken elérhető adatgyűjteményekre is. Ha egy adatbázist jelentős ráfordítással állítottak össze, az a sui generis adatbázisjog alapján védhető, amely kizárólagos jogot biztosít az adatbázis tulajdonosának a tartalma jelentős részének kinyerésére és újrafelhasználására.

Vannak azonban kivételek e védelem alól. Oktatási és tudományos kutatási célból az adatbázis-irányelv 6. cikke (2) bekezdésének b) pontja lehetővé teszi az adatok kinyerését, feltéve, hogy azt a kinyerés célja indokolja, és a forrást feltüntetik. Ez azt jelenti, hogy az akadémiai kutatás céljára végzett webscraping megengedhető az uniós jog szerint, amennyiben tiszteletben tartja ezeket a feltételeket, és nem sérti az adatbázis tulajdonosának jogos érdekeit.

Az Általános Adatvédelmi Rendelet (GDPR) az uniós jog másik kulcsfontosságú vonatkozása, amely hatással van a webes adattárolásra, különösen akkor, ha a kimásolt adatok személyes adatokat is tartalmaznak. A GDPR előírja a személyes adatok jogszerű, tisztességes és átlátható kezelését. Oktatási vagy személyes felhasználás esetén a személyes adatok feldolgozásának jogalapja lehet a jogos érdek, a hozzájárulás vagy a tudományos kutatásra vonatkozó, a 89. cikk szerinti különleges rendelkezések (A közérdekű archiválási, tudományos vagy történelmi-kutatási vagy statisztikai célú adatkezeléssel kapcsolatos biztosítékok és eltérések).

Habár az oktatási és személyes használatra szánt webscraping jogilag megengedett az EU-n belül, ezt a szellemi tulajdonjogok, az adatvédelmi előírások és a szerződéses kötelezettségek árnyalt kölcsönhatása szabályozza. Az akadémiai kutatóknak és magánszemélyeknek körültekintően kell eligazodniuk ezekben a jogi keretekben, hogy biztosítsák a megfelelést. A jogi kockázatok csökkentése érdekében tiszteletben kell tartaniuk az adatbázishoz való jogokat, be kell tartaniuk a GDPR elveit, és be kell tartaniuk a webhely szolgáltatási feltételeit. Jelen esetben a webscrapinget felelősségteljesen és legálisan lehet lebonyolítani, elősegítve az innovációt és a tudást anélkül, hogy jogokat vagy előírásokat sértene.

3.4 A mesterséges neurális hálók

A mesterséges neurális hálózatok (Artificial Neural Networks, vagyis ANN-ok) olyan adatfeldolgozó rendszerek, amelyek az agyban található neurológiai hálózatokon alapulnak, és azokat felépítése és mintája szerint valósulnak meg programozási környezetben. A rendszereket elsősorban mintaazonosításra és -feldolgozásra használják, és a korábbi feladatok elemzési eredményei alapján képesek fokozatosan javítani a teljesítményt. (Jain, Mohiuddin, & Mao, 1996)

A neurális szerveződés és hálózatosodás alapvetően a többrétegű perceptron modellel magyarázható. Ebben a modellben a neurális hálózatok olyan rétegek formájában zajlanak, amelyek egy irányban hoznak létre kapcsolatokat, más néven előrecsatolt neurális hálózatok. A csomópontoknak több rétege van: bemeneti, rejtett és kimeneti. A különböző csomópontok közötti kapcsolatok megváltoztatják a hálózatok viselkedését. A bemeneti rétegek információt kapnak, ekkor a bemeneti és a rejtett rétegek között kapcsolatok jönnek létre. A rejtett rétegek ezt követően feldolgozzák az információt, amely viszont a kimeneti rétegekbe kerül. Végül a kimeneti rétegek a következő réteg bemenetévé válnak, és a sorozat folytatódik. (Xin, 1999)

A mesterséges intelligencia egy olyan számítógépes program, amely az emberi agyhoz hasonló módon képes szervezni az információt. A mesterséges intelligencia, a neurális hálózatok vegyülete a kognitív tehetséggel és a gépek tervezésével kapcsolatos kutatások eredményeként alakult ki. (Kutsurelis, 1998) A mesterséges intelligencia története Aristo-ig nyúlik vissza. Ismeretes, hogy Aristo a gondolkodás algoritmusán dolgozott, és megvitatta annak nehézségeit is. Modern értelemben a mesterséges intelligencia akkor került be a tudományos világba, amikor az 1940-es években üzembe helyezték az első elektronikus számítógépet (megj.: már voltak korábban is számítógépek, pl. a német Konrad Zusé-é a 30-as években, de ezt most már hagyjuk így), és Alan Turing kifejlesztette az első szoftvert. A mesterséges neurális hálózatok, a mesterséges intelligencia legjelentősebb alszegmense, egy statisztikai megközelítés, amelyet előrejelzési modellek fejlesztésére hoztak létre. A mesterséges neurális hálózatok az emberi agy tervezéséhez hasonló feldolgozó eszközökből és adatfeldolgozásból állnak. (Blackard & Dean, 1999)

A mesterséges neurális hálózatok fejlesztési folyamata nagyjából négy szakaszra osztható, nevezetesen az emelkedés szakaszára, az apály szakaszára, az újjáéledés szakaszára, a virágzás szakaszára.

Az 1940-es évek elejétől a kutatók egyre jobban érdeklődnek az agy működése iránt. 1943-ban Warren McCullochand, Walter Pitts logisztikus idegtudós kiadta úttörő tanulmányát, amely először átfogóan leírja, hogyan kommunikálnak az agy neuronjai. (Marsalli, 2006) Céljuk az volt, hogy megértsék, hogyan képes az agy számolni, összetett mintákat létrehozni, érzékelni és sok más bonyolult műveletet elvégezni, pusztán az idegsejtek közötti kapcsolatok felhasználásával. A neurális műveletek elemzése után a logikát és a számítást kombinálták a McCulloch-Pitt-modell (MCP) kifejlesztéséhez. Ez a modell egy neuron alapmodelljévé vált, és végül a mesterséges neurális hálózatok fejlesztésének fontos alapjává. (Abraham, 2002) (McCulloch & Pitts, 1943) Az MCP neuronok korai verziói nem voltak korlátlanok. Az egyik ilyen korlát az volt, hogy az MCP-neuronok nem tudtak tanulni a kapott bemenettől, vagy ahhoz alkalmazkodni. McCulloch és Pitts tanulmányai logikai alapúak, és a konnekcionista mozgalom úttörőinek tartják őket. A kutatók később további funkciókkal is szolgáltak, amelyek lehetővé tették az MCP neuronok számára, hogy elérjék ezt a célt. Az egyik ilyen jellemző a perceptron fogalma volt, amelyet Frank Rosenblatt pszichológus vezetett be. (Rosenblatt, 1958) Az első mesterséges intelligenciával kapcsolatos tanulmányt McCulloch és Pitts végezte 1956-ban egy logikai modellezésen alapuló számítási modellen keresztül, amely mesterséges idegsejteket, fiziológiát és Turing számítási koncepcióját használta fel.

Miközben olyan gépet próbált kifejleszteni, amely képes reprodukálni az agy képességeit, Frank Rosenblatt kifejlesztette a perceptront. Rosenblatt találmányának jelentősége nyilvánvalóvá válik, ha figyelembe vesszük a perceptronban résztvevő algoritmusokat. (Rosenblatt, 1958) Rosenblatt koncepciója más tudományágak kutatóit inspirálta, és további vizsgálatokat végzett az ANN-ok különféle tulajdonságaival kapcsolatban. Gardnera és Dorlinga (1998) alapos leírást adnak a perceptron algoritmikus működéséről a légkörtudományi alkalmazások áttekintéséről szóló cikkükben. Az 1960-as évek végéig a legtöbb kutató elfogadta a Rosenblatt-féle perceptron koncepcióját. Legbefolyásosabb kritikája 1969-ben Minsky és Papert (1969) könyvében olvasható. Ez a könyv nagy matematikai részletességgel elemzi a Rosenblatt találmányában talált számítási hibákat. A könyv népszerűsítését követően bizonytalanság keletkezett, és az egyének nem bíztak többé a perceptronban vagy a neurális hálózatokban. Nem tudja megoldani kétféle lineáris elválaszthatatlan minta osztályozási problémáját. Például az egyszerű lineáris érzékelő nem tudja megvalósítani az XOR logikai kapcsolatát. Ez a következtetés súlyos csapást mért a mesterséges neurális hálózatok akkori kutatására. A neurális hálózatok története innentől kezdve közel 10 évig megállt. (Garson, 1998)

Marvin Minsky kognitív tudós és Seymour Papert matematikus is érdeklődött a mesterséges intelligencia működése iránt. Az elemzést követően számos hiányosságot azonosítottak a neurális hálózatokkal és a számítási gépekkel kapcsolatban. Az egyik ilyen hiányosság az volt, hogy a perceptron nem tudta feldolgozni a két áramkör egyikéből származó információt. A másik az volt, hogy a számítási rendszerek nem tudták biztosítani a nehéz neurális hálózatok működtetéséhez szükséges feldolgozási kapacitást. Ezek a kérdések kritikussá váltak, és ennek eredményeként a számítási rendszerekkel és neurális hálózatokkal kapcsolatos kutatás az 1980-as évekig stagnált. (Garson, 1998)

1972-ben Kohonen finn professzor javasolta az önszerveződő jellemzőtérképet (SOM). A későbbi neurális hálózatok főként Kohonen munkásságán alapultak. A SOM hálózat egyfajta oktatói tanulási hálózat, elsősorban mintafelismerésre, beszédfelismerésre és osztályozási problémákra használták. A „győztes a király" versengő tanulási algoritmusát alkalmazza, amely nagyon különbözik a korábban javasolt perceptrontól. Tanulási és képzési módszere ugyanakkor önszerveződő hálózat, oktatási képzés nélkül. Ezt a fajta tanulási és képzési módszert gyakran egyfajta képzésként használják minősített információk kinyerésére anélkül, hogy tudnák, milyen típusú osztályozás létezik. 1976-ban Grossberg professzor javasolta a híres adaptív rezonanciaelméletet (ART), amely az önszerveződés és az önstabilitás jellemzőivel rendelkezik.

Az 1980-as évek elején egy John Hopfield nevű tudós keltette új életre az ANN területén végzett kutatásokat. Javasolt egy asszociatív modellt a neurális hálózatokhoz, amely az információ tárolását úgy írja le, mint ami a neuronok kapcsolódásai között megy végbe. Hopfield azt javasolta, hogy az adatfeldolgozás úgy valósul meg, hogy egyes neuronokat „be" vagy „kikapcsolnak" külső ingerektől függően. (Hopfield, Neural networks and physical systems with emergent, 1982) Ez a koncepció segített megoldani a Minsky és Papert által eredetileg leírt problémákat. A modell ezt úgy tette, hogy azt javasolta, hogy az egyes neuronok együttműködjenek a körülöttük lévőkkel. Más szóval, ami egy egyedi neuronnal történik, az jellemzően a környező neuronokkal is történik. Ezek a neurális asszociációk biztosítják a mintafelismerés, az asszociatív memória és a hibajavítás alapjait, miközben elegendő feldolgozási kapacitást biztosítanak a nagy neurális hálózatokból származó információk tárolására. (Aiyer, Niranjan, & Fallside, 1990)

Hopfield hálózati modelljének publikálása után a mesterséges neurális hálózatokkal kapcsolatos kutatások nagymértékben megnövekedtek. (Rojas, 1996) Ez előrelépést jelentett a számítási rendszerek és a legkorszerűbb technológia terén. (Rabunal, 2005)

A későbbi kutatók a Ljapunov-függvényt energiafüggvénynek is nevezték, bizonyítva a hálózat stabilitását. 1984-ben Hopfield egy folytonos neurális hálózatot javasolt, hogy a hálózatban lévő neuronok aktivációs funkcióját diszkrétről folyamatosra változtassa. (Hopfield, 1984) Hopfield és Tank (Hopfield & Tank, Neural Computation of Decisions in Optimization Problems, 1985) a Hopfield neurális hálózatot használta a híres Traveling Salesman Problem megoldására. A Hopfield neurális hálózat nemlineáris differenciálegyenletek halmaza. A Hopfield-modell nemcsak nemlineáris matematikai összegzést végez a mesterséges neurális hálózat információtároló és -visszakereső funkciójáról, hanem dinamikus egyenleteket és tanulási egyenleteket is ad. Ezenkívül fontos képleteket és paramétereket ad a hálózati algoritmushoz, aminek köszönhetően a mesterséges neurális hálózatok felépítése és tanulása elmélete a Hopfield-modell hatása alatt áll, számos tudós ösztönzi a neurális hálózatok tanulmányozásának lelkesedését és aktívan részt vesz ezen a tudományos területen. A Hopfield neurális hálózatban rejlő sok szempontból nagy potenciál miatt az emberek nagyobb figyelmet fordítanak a neurális hálózatok kutatására. Egyre többen kezdik el tanulmányozni a neurális hálózatot, és nagymértékben elősegítik a neurális hálózatok fejlődését.

1984-ben Hinton a fiatal tudósokkal, Sejnowskival és munkatársaival együttműködve egy szuperskalár párhuzamos online tanulási gépet javasolt, és kifejezetten javasolta a rejtett egység koncepcióját, amelyet később Boltzmann-gépnek neveztek el. Hinton és Sejnowsky a statisztikai fizika módszertanát és eszköztárát használja, az első javasolt többrétegű hálózati tanulási algoritmust, amelyet Boltzmann gépmodellként ismernek. (Ackley, Hinton, & Sejnowski, 1985) 1986-ban a többrétegű neurális hálózat modellje alapján Rumelhart és munkatársai (1986)

javasolták a visszacsatolásos algoritmust (Error Back Propagation) a többrétegű neurális hálózat súlykorrekciójának megoldására Az előre irányuló neurális hálózat tanulási problémája azt bizonyítja, hogy a többrétegű neurális hálózat erős tanulási képességgel rendelkezik, számos tanulási feladatot képes elvégezni és sok gyakorlati problémát megoldani. 1988-ban Chua és Yang (1988) egy celluláris neurális hálózat (CNN) modellt javasolt, amely egy nagyszabású nemlineáris számítógépes szimulációs rendszer sejtautomaták számára. Kosko (1988) létrehozott egy kétirányú asszociatív tárolási modellt (BAM), amely felügyelet nélküli tanulási képességekkel rendelkezik. 1995-ben Haken és társai (1995) bevezették a szinergiát a neurális hálózatokba. Elméleti keretében Haken és társai úgy vélik, hogy a kognitív folyamat spontán, és azt állítják, hogy a mintafelismerési folyamat a mintaképzés folyamata. A neurális hálózatok aktiválási funkcióosztályainak bővítésével általánosabb késleltetett cellás neurális hálózatok (DCNN), Hopfield neurális hálózatok (HNN) és kétirányú asszociatív memóriahálózatok (BAM) adhatók meg. Több éves fejlesztés után több száz neurális hálózati modellt javasoltak.

C. V. Soumya és Muzameel Ahmed egyik ilyen előrelépése az volt, hogy matematikai algoritmusokat használtak az emberi gesztusok azonosítására. Cikkükben a kifejező testmozdulatok felismerésére és leírására használt innovatív módszert vizsgálnak. (Soumya & Ahmed, 2017) Ezeket a testmozdulatokat főként a Mudrában, az indiai kultúra által gyakorolt klasszikus táncban figyelték meg. Céljuk az volt, hogy olyan mintafelismerést és képfeldolgozást használó rendszert építsenek fel, amely könnyen azonosítani tudja a konkrét emberi gesztusokat, amelyek viszont leírást adnak ezekről a testmozgásokról és egészségügyi előnyeiről.

Az ANN-ok alapvető segédeszközzé váltak a mikrobiális felhalmozódás előrejelzésében és azonosításában különböző környezetekben. Keaton Larson Lesnik (2017) cikkében számos olyan módszert írt le, amelyeket a mikroorganizmusokból származó adatok megszerzésére használtak. Célja az volt, hogy az összegyűjtött információkat az ismeretek bővítésére használja fel, hogy megkönnyítse a mikrobiális üzemanyagcellák teljesítményét a hulladékáramokból származó kémiai potenciális energia elektromos energiává alakításának folyamatában. (Farhat, Psaltis, Prata, & Paek, 1985)

Az ANN-okat nemrégiben alkalmazták a tőzsdei teljesítmény előrejelzésében. Kamran Raza (2017) számos technikát fejlesztett ki az ANN-ok négy különböző változata alapján, amelyeket cikkében ismertet. E technikák kidolgozásának célja elsősorban egy olyan előrejelzési modell megalkotása volt, amely megkönnyíti a tőzsdén érintett befektetők munkáját. A legjobb megtalálása érdekében több technikát külön-külön is összehasonlítottak. Az eredmények azt mutatták, hogy a részvénypiac viselkedése akár 77%-os pontossággal is megjósolható. (Huang, Nakamori, & Wang, 2005)

A mesterséges neurális hálózat az információfeldolgozás teljesen más megközelítése, mint a ma általánosan használatos megközelítések. Ez a számítási technika hatékonyabbnak bizonyult olyan problémák kezelésében, mint a mintafelismerés, a robotvezérlés és a tudásszerzés. (Kohonen, 1988) A problémák e széles skáláján nyújtott nagy teljesítménye miatt az akadémikusok és a gyakorlati szakemberek egyaránt érdeklődnek e technika lehetőségeinek feltárása iránt.

Az ANN koncepciója azon az elméleten alapul, hogy az emberi intelligencia sok idegsejt kölcsönhatásából jön létre, amelyek mindegyike serkentő és gátló jeleket küld más neuronoknak. Az ANN szerkezete számos számítási elemből és a számítási elemek közötti összekapcsolódásból

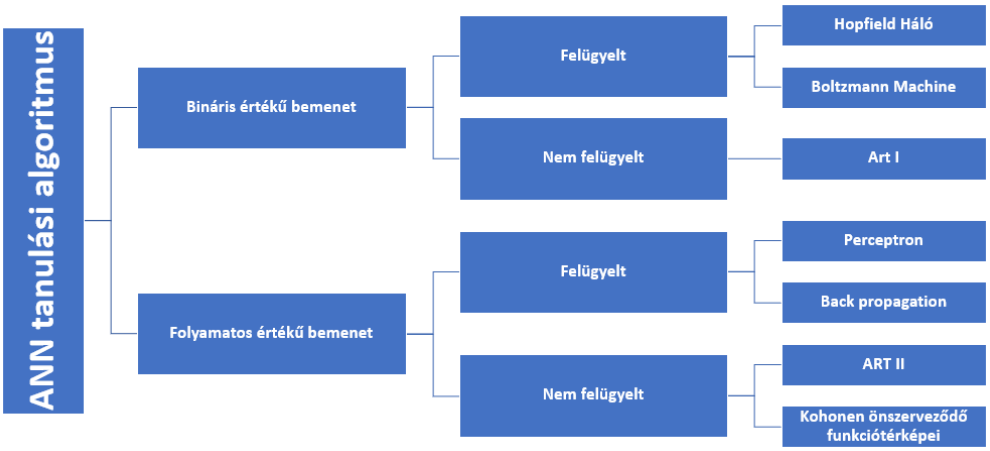
áll, amelyek az emberi agy neuronjaira és szinapszisaira hasonlítanak, hálózatba szervezve. (McClelland, Rumelhart, & Hinton, 1987)

Ez a struktúra az ANN modell számos fontos jellemzőjére rávilágít. Először is, egy ANN-modell megtanulja a bemenet és a kimenet közötti funkcionális kapcsolatot, és azt a hálózatban lévő összekapcsolt súlyok nagyságában kódolja, még akkor is, ha a függvény lehet nemlineáris, hiányos vagy nem egyértelmű. (Grossberg, 1988) Másodszor, a masszívan összekapcsolt elemek párhuzamosan dolgozzák fel az információkat a kialakított kapcsolatokon keresztül. Harmadszor, a modell képes általánosítani az elvi szabályokat, miközben figyelmen kívül hagyja a problémamegoldás viszonylag nagy mennyiségű zaját vagy változását. (Lippman, Review of Neural Networks for Speech Recognition, 1989) Végül egy ANN-modellben a tudás eloszlik a hálózat összes összekapcsolása között úgy, hogy hiba lép fel. A hálózat egy kis része kecsesen rontja a hálózat teljesítményét. (Hinton, McClellan, & Rumelhart, 1986)

A megfelelő súlyok kiszámításának folyamatát az ANN paradigmában "tanulásnak" vagy "képzésnek" nevezik. Számos ANN tanulási algoritmus létezik, amelyek a leírt elveket alkalmazzák. Általánosságban elmondható, hogy az ANN tanulási algoritmusokat vagy az elérendő feladatok, vagy a feladat végrehajtásának módszerei szerint osztályozzák.

A funkcionális specifikációk szerint az algoritmusok négy osztályba sorolhatók: auto-asszociáció, hetero-asszociáció, osztályozás, szabályosság-detektálás. (Rumelhart & Zipser, Feature Discovery by Competitive Learning, 1986) Az autoasszociatív modell megjegyzi a mintákat a tanulási folyamat során. A rendszer ezután visszahívja a teljes mintát egy hiányos vagy zajos minta esetén. A heteroasszociatív modell, az autoasszociáció egy változata, asszociációkat talál két tanulási folyamatot összefogó minta között. A tanulási folyamat befejeztével a modell lekéri a párosított mintát egy bemeneti mintával. Az osztályozási modell megtanulja megfelelően kategorizálni a mintát egy korábbi osztályozási specifikáció szerint. A szabályosság-észlelési modell saját jellemző reprezentációt fejleszt ki a bemeneti sokaság domináns jellemzőinek kódolására anélkül, hogy előzetes osztályozási specifikációkat használna.

Az ANN tanulási algoritmusokat két osztályra osztják a módszertanuk szerint, hogy elérjék feladataikat: felügyelt és nem felügyelt. (Lippmann, 1987) A felügyelt tanulás során a hálózat bemenetet kap a kívánt kimenettel együtt. Egy bemenet minden egyes bemutatásakor a hálózat összehasonlítja saját kimenetét a kívánt kimenettel, és súlyok beállításával megpróbálja csökkenteni a kettő közötti különbséget. Ezt a folyamatot addig ismételjük, amíg a saját kimenete szorosan illeszkedik a kívánt kimenethez. Másrészt a felügyelet nélküli tanulási hálózat csak bemenetet kap. A bemenet minden egyes bemutatása után a teljesítményt mérik, hogy megtudják, hogyan működik a hálózat. A hálózattól elvárás, hogy a teljesítménymérés útmutatásként való felhasználásával önszervezze az információkat. Az e két kategóriába tartozó algoritmusok további két csoportra oszthatók a bemeneti formátumok alapján: bináris vagy folytonos értékű bemenetre. Az ANN algoritmus taxonómiáját az 1. ábra mutatja be.



1. ábra Az ANN tanulási algoritmusok taxonómiája (saját szerkesztés)

Az ANN technikákat számos problémás területen alkalmazták, és bebizonyították, hogy képesek a rendkívül összetett problémák kezelésére. Az ANN technikák alkalmazási területeit két kategóriába sorolom: alacsony szintű és magas szintű kognitív feladatok.

Az alacsony szintű kognitív feladatok azok, amelyeket egy átlagember naponta nehézség nélkül végez, mint például a beszédértés és a tárgyfelismerés. Az alacsony szintű kognitív feladatok számítógépben való megvalósítása magában foglalja a jellemzők mikroszintű ábrázolását, valamint a nyers észlelési bemenet és a kívánt kimenet közötti finom kapcsolat elemzését. A bemeneti adatok nagy változatossága és zaja megnehezíti ezeknek a feladatoknak a számítógépen való megvalósítását. Emiatt szükség volt egy általánosításra alkalmas módszerre, amellyel bizonyos fokú zaj vagy véletlenszerűség elviselhető. Az ilyen típusú problémáknál az ANN-megközelítés nemcsak a látens jellemzők kinyerésére és hálózatba történő kódolására bizonyult hasznosnak, hanem a torz vagy részleges információblokk mintájának visszaállítására is. (Lippman, 1989) (Simpson, 1990) Ez az ígéretes eredmény sok dolgozót arra ösztönzött, hogy további kutatásokat folytassanak az ANN-megközelítés alkalmazásával azzal a kifejezett céllal, hogy az alacsony szintű kognitív információfeldolgozás problémáit kezeljék.

Az alacsony szintű kognitív feladatok sok alkalmazásához képest néhány alkalmazás kihasználta az ANN-modell erejét magas szintű kognitív feladatokban: szakértői rendszerterületeken. Az ANN alkalmazásának tanulmányozása magas szintű kognitív problémákban két irányzatra osztható. Az egyik explicit szabályokat használ tudásbázisként, és egy ANN tanulási algoritmust alkalmaz a következtetési motor munkájának végrehajtására. (Samad, 1988) (Touretzky & Hinton, 1985) Ez a módszer közel azonos a hagyományos szabályalapú rendszerrel, azzal a különbséggel, hogy ANN algoritmust használnak a következtetési folyamatban, így ez a megközelítés nem küszöböli ki a tudásszerzés szűk keresztmetszetét, mivel a tartományi tudás továbbra is explicit szabályokban jelenik meg.

A magas szintű kognitív feladatokban végzett ANN-kutatás másik iránya egy ANN-tanulási algoritmust használ egy olyan tudásbázis létrehozására egy hálózatban, amelynek kapcsolati súlyai implicit döntési kritériumokat határoznak meg, és az algoritmust egy következtetési folyamat végrehajtására használja. (Mulsant, 1988) (Yoon, Brobst, Bergstresser, & Peterson, 1989) Ez a módszer lehetőséget ad a tudásszerzési nehézségek megoldására. Ez azért van így, mert az ANN-megközelítés megtanul egy leképezési függvényt a bemeneti és kimeneti minták között, és kódolja azt a hálózat kapcsolatainak súlyának nagyságában, ahelyett, hogy

tudásbázisként explicit szabályokat követelne meg. Az ANN-alapú rendszerek fő korlátja azonban az, hogy nem tudja megmagyarázni döntéseit vagy a mögöttes tudásbázist úgy, ahogy a hagyományos szabályalapú rendszerek.

Számos ANN-alkalmazás bebizonyította, hogy ez a technika hatékonyabb megoldást kínál az összetett problémák kezelésére, mint a hagyományos technikák. Noha ezen a területen még sok problémát kell megoldani, az ANN megfelelőbb megközelítés egy intelligens modell felépítéséhez számos alkalmazási területen, mint bármely más alkalmazott megközelítés. Ahogy egyre több kutatás folyik, egyértelműnek tűnik, hogy az ANN a számítógépes információfeldolgozó rendszerek új generációját hozza el, és a módszer óriási hatással lesz az üzleti adatfeldolgozásra.

Az ANN elvonatkoztatja az emberi agy neurális hálózatát az információfeldolgozás szemszögéből, egyszerű modellt hoz létre, és különböző kapcsolatok alapján különböző hálózatokat állít össze. (Dong & Hu, 1997) Igyekszik szimulálni az agy neurális hálózatának feldolgozását, a memória információit az információfeldolgozás útján. A mérnöki és tudományos világban gyakran közvetlenül neurális hálózatnak vagy neurális hálózatnak nevezik. A neurális hálózat egy számítástechnikai modell, amely nagyszámú csomópontból (vagy neuronból) kapcsolódik egymáshoz. (Jenkins & Tanguay, 1995) Minden csomópont egy adott kimeneti függvényt képvisel, amelyet aktiválási függvénynek neveznek. A két csomópont közötti kapcsolat egy súlyt jelent a kapcsolaton áthaladó jel számára, amelyet súlynak nevezünk, ami egyenértékű a mesterséges neurális hálózat memóriájával. (Bulsari, 1993) A hálózat kimenete a hálózat csatlakozási módjától, a súlyértéktől és az ösztönző funkciótól függően változik. Maga a hálózat azonban általában valamilyen algoritmus vagy funkció közelítése a természetben, vagy lehet egy logikai stratégia kifejezése. (Luo, Xie, & Zhu, 1997)

Egy mesterséges neurális hálózatban egy neuron feldolgozó egység különböző objektumokat, például jellemzőket, betűket, fogalmakat vagy valamilyen értelmes absztrakciós mintát ábrázolhat. A hálózat feldolgozóegységének típusa három kategóriába sorolható: bemeneti egység, kimeneti egység és rejtett egység. A bemeneti egység jeleket és adatokat fogad a külvilágból. (Balcazar, 1997) A kimeneti egység megvalósítja a rendszer feldolgozási eredményének kimenetét. A rejtett egység egy olyan egység, amely a bemeneti és kimeneti egységek között helyezkedik el, és nem figyelhető meg a rendszeren kívül. (Setiono & Leow, 2000) A neuronok közötti kapcsolati súlyok a sejtek közötti kapcsolat erősségét tükrözik. Az információ megjelenítése és feldolgozása a hálózati feldolgozó egység kapcsolódási viszonyában testesül meg. A mesterséges neurális hálózat egy nem programszerű, adaptív, agystílusú információfeldolgozás, melynek lényege a hálózat átalakulásán és dinamikus viselkedésén keresztül párhuzamosan elosztott információfeldolgozási funkciók, valamint az emberek különböző mértékű és szintű utánzása az agy és az idegrendszer információfeldolgozása. rendszer. (He, Zhu, & Cao, 2004) Részt vesz az idegtudomány, a gondolkodástudomány, a mesterséges intelligencia, a számítástechnika és más interdiszciplináris területek különböző területein.

A mesterséges neurális hálózat egy párhuzamos elosztott rendszer, amely a hagyományos mesterséges intelligencia és információfeldolgozási technológiáktól teljesen eltérő mechanizmust alkalmaz, felülmúlja a hagyományos logikai alapú mesterséges intelligencia hibáit az intuíció és a strukturálatlan információ kezelésében, és rendelkezik az adaptív, önszerveződő előnyökkel. és valós idejű tanulási funkciók. (Kasabov, és mtsai., 2016)

3.4.1 RNN – visszacsatolt neurális hálók

A visszacsatolt neurális háló (RNN) egy speciális neurális hálózat, amely memóriatulajdonságokkal rendelkezik, és képes az adatok által meghatározott múltbeli minták felismerésére és megjóslására. Az RNN-ek általában önkapcsoló rétegekből és hosszú távú memóriából állnak, így képesek figyelembe venni az előzményeket, és következtetéseket levonni azokból. A RNN-ek általában alkalmazzák az olyan feladatok megoldására, mint a szövegfelismerés, a beszédfelismerés, az észlelések értelmezése, a képeknek való szöveghez való hozzárendelés, a mesterséges intelligencia és a prekurzorok előrejelzése. (Goodfellow, Bengio, & Courville, 2016) (Graves A. , 2013)

A visszacsatolt neurális háló (RNN) segítségével a neurális hálók képesek az időbeli dinamikára reagálni, amelyek megadhatják a memória használatát és segíthetnek a különböző jellegű bemenetek és kimenetek kezelésében. RNN egy olyan neurális háló, amelynek tagjai képesek az időbeli változásokkal foglalkozni, a memóriát használni és megőrizni az információkat, valamint a különböző időbeli bemeneteket és kimeneteket kezelni. (Hochreiter & Schmidhuber, 1997)

A visszacsatolt neurális háló (RNN) egy mesterséges neurális háló, amely képes tanulni a múltból. Különlegesek abban, hogy képesek megjegyezni az előző bemenetekből szerzett információkat, és ezt kombinálják a jelenlegi bemenettel, hogy kiértékeljék a jövőbeni kimeneteket. A RNN-ek különböző összetevőkből állnak, beleértve a neurális hálókat, a memóriákat és a számítási egységeket, amelyek mindegyike hozzájárul a tanuláshoz, és különböző módokon működnek együtt. Akkor használhatók, ha a bemenetek és kimenetek egymásra állnak. Különösen hasznosak olyan problémák megoldásában, mint a szövegfolyamok, a beszédfelismerés, a nyelvi feldolgozás és az idősorok elemzése. A RNN-eket különféle alkalmazásokban, beleértve a robotokat, az autóvezérlést, az önvezető autókat, a gépi tanulást, a játékokat és más alkalmazásokat is használják. A memória a RNN-ek másik fontos összetevője. A memória a tanuláshoz szükséges információkat rögzíti, és a tanulási folyamat során segíti. A memória segítségével a RNN-ek képesek megjegyezni az előző bemenetekből szerzett információkat, és ezt kombinálni a jelenlegi bemenettel, hogy kiértékeljék a jövőbeni kimeneteket. Központi egysége a számítási egység, amely a bemeneteket feldolgozza, és a kimeneteket számítja ki. A számítási egység képes kombinálni a neurális hálókat és a memóriát, hogy a bemenetek alapján kiszámítsa a kimeneteket. (Wang & Yang, 2020) (Gers, Schmidhuber, & Cummins, 2000)

A visszacsatolt neurális háló (RNN) egy speciális neurális hálós architektúra, amely képes információt "emlékezni" a korábbi bemenetek alapján. Ezt a tulajdonságát időbeli megfigyelések feldolgozására használják, és ezért jelentősen előrelépést jelentett a neurális hálók fejlesztésében. A RNN-ek az információ feldolgozásának két fő módja közül az egyiket alkalmazzák: a rekurrenciát. A rekurrencia a következő értelemben használható: a hálózat önmagába tér vissza, és minden bemenetet használ az összes előző bemenethez való hozzáféréshez. Egy RNN felépítése egy bemeneti rétegből, egy rekurrens rétegből és egy kimeneti rétegből áll. A bemeneti réteg szakaszban fogadja el a bemeneteket, és továbbítja őket a rekurrens rétegben. A rekurrens réteg a bemeneteket visszatérő összekötésekkel kapcsolja össze, hogy átfogó információkat biztosítson a korábbi bemenetekről. Ezután a kimeneti réteg a feldolgozott információkat a megfelelő kimenetekhez köti. A RNN-ek segítségével a neurális hálók képesek időbeli információkat feldolgozni, és azokat a kimenetekhez kötni. Ez lehetővé teszi a hálózatok számára, hogy időbeli megfigyeléseket végezzenek a környezetükben, és megfelelő reakciókat adhassanak. (Graves A. , 2013)

3.4.2 NLP - Természetes nyelvi feldolgozás

Az NLP (Nyelvi neurális hálózat) olyan mesterségesintelligencia-alkalmazás, amely az emberi nyelv meghatározott területein értelmezésre és elemzésre használja a neurális hálózatot. Az NLP az információfeldolgozás, a számítógépes látás és a beszédfelismerés terén használható módszerek egyike. Az NLP az adatok számítógépes feldolgozását egy olyan neurális hálózat segítségével végzi, amelynek különböző szintjei vannak. Az egyes szintek a nyelvi elemzés alapvető szintjei, amelyek lehetővé teszik a számítógépes rendszer számára, hogy értsen és elemezze a bemenetet. Az NLP segítségével a számítógépes rendszerek olyan feladatokat tudnak megoldani, amelyek korábban csak emberek számára voltak elérhetők. Segítségével a számítógépes rendszerek képesek lehetnek a nyelvi elemzésre, a gondolkodás megértésére és a társalgási játékokra, a számítógép képes lehet a gondolkodás megértésére és a társalgásban való részvételre. Az NLP fejlesztése megelőzte a neurális hálózatok megjelenését. Ma már számos alkalmazása van a számítógépes látás, a beszédfelismerés, a nyelvi elemzés és a társalgás között. Ez a technológia számos számítógépes programozási nyelv használatával dolgozik, beleértve a Python, a Java, a C++, a JavaScript és még sok más. (Chollet, 2018) Ezek a hálózatok általában egy bemeneti szöveget használnak, amelyet több lépésben kezelnek: a szöveg tokenizálása, a szófajok azonosítása, a szöveg értelmezése és a következtetések levonása. (Rish, 2020)

A Nyelvi Neurális Hálózatok (NLP, vagy Natural Language Processing) az informatikában alkalmazott mesterséges intelligencia technológia, amely a számítógépeknek lehetővé teszi, hogy megértsék és feldolgozzák az emberi nyelvet. Az NLP egy olyan algoritmus, amely segít a számítógépeknek abban, hogy megértsék az emberek által használt szavak, kifejezések és mondatok jelentését, és annak megfelelően cselekedjen. A NLP segítségével a számítógépek képesek lesznek olvasni, érteni és reagálni a felhasználók által használt nyelven, felismerni a szövegben lévő jelentéseket, és hogy egy adott mondat milyen jelentést hordoz. A felismerés a szavakon, mondatokon és kifejezéseken keresztül történik. Az NLP-nak köszönhetően a számítógépek képesek lesznek a felhasználók által használt nyelven keresztül beszélgetni, döntéseket hozni és cselekedni. A neurális hálózatok egy olyan mesterséges intelligencia technológiát jelentenek, amelyek hasonlítanak az agyhoz. Egy sor neurális egységből állnak, amelyek között kapcsolatok vannak. A neurális egységek segítségével a számítógép megértheti és megjegyezheti a használt mondatok jelentését. A neurális hálók két alapvető típusra oszthatók: lineáris és konvolúciós. A lineáris neurális háló egy olyan rendszer, amelyet a felhasználó bevitelére használnak az adatok felismeréséhez. A konvolúciós neurális hálót arra tervezték, hogy szöveget értelmezzen és felismerjen. Mindkét típusú neurális hálót használják az NLP-hez. (Srivastava, 2020)

4. Anyag és módszer

Az olajár-előrejelző neurális háló (NN) azokat a Wall Street Journal cikkeket használja az elemzéshez, amelyek az olajárakról szólnak, és ezeket az információkat a neurális háló segítségével dolgozza fel.

Az olaj ár-előrejelzést biztosító Neurális Háló használata segít megérteni a piaci feltételek hatását az olaj árra. A kutatók képesek lehetnek megérteni a spekuláción is alapuló piaci feltételek és a Wall Street Journal cikkei közötti összefüggéseket, és ezáltal jobban megérteni a piacok működését, és pontosítani a jövőbeli olaj árakat.

A mesterséges neurális háló (ANN) tanulja a cikk által összefoglalt információkat, amelyek a piaci mozgásokra vonatkoznak, és alkalmazza a következtetések levonásához. Az ANN figyelembe veszi a piaci trendeket, a piaci feszültségeket, a felső- és alsó határokat, és más hasznos információkat, amelyeket a Wall Street Journal cikkeiből vehet ki. Az ANN előrejelzi az olaj árát, és különböző feltételezéseket használ a különböző piaci szituációk megértéséhez. Az ANN ezenkívül figyelembe veszi a technikai elemzést is, és ezen túlmenően a prognózis során használható egyéb technikákat is.

A korábbiakban ismertettem az ANN és társai működését. Leegyszerűsítve: betöltjük az adatot, majd kijön egy eredmény. Ha nem megfelelő, tudunk finomítani, majd újra futtatni és bízni abban, hogy javul az eredmény.

4.1 Olajárváltozás előrejelzés ANN-nel a WSJ cikkeinek elemzésével

Az olajpiacok hatása a világ egyik legnagyobb mértékű és legkomplexebb kérdése. Az olajárak nagymértékben befolyásolják az iparágakat és az egyén egészségét, és megjósolhatóságuksokkal bonyolultabb, mint más piacok esetében. A mesterséges neurális hálók technológiája képes megerősíteni a jövőbeni olajárak előrejelzését, számos különböző tényező figyelembe vételével. Az alábbi szekció egy olyan kutatást mutat be, mely az újságcikkek elemzését használja a jövőbeli olajár előrejelzéséhez, mesterséges neurális hálókat alkalmazva. A kutatás részletesen elemzi a mesterséges neurális hálós technológiát, és bemutatja, hogyan használható a jövőbeli olajárak előrejelzésére.

Az eredmények remélhetőleg jelentős segítséget nyújtanak majd a gazdasági döntéshozatalhoz.

4.1.1 Adatgyűjtés

Az adatgyűjtési metódus két részre oszthatjuk.

Az egyik az olajár napi kimutatása a vizsgált időszakban, jelen esetben 2000-2020 között. Az adatbeszerzést tekintve ez jóval egyszerűbb, hiszen sok hivatalos forrásból megszerezhető..

A másik feladat pedig a Wall Street Journal, mint az egyik legmeghatározóbb üzleti folyóirat cikkeinek beszerzése és elemzése volt.

4.1.2 Napi olajárak 2000-2020 között

Világszinten nem létezik egységes olajár, hiszen a kitermelt olaj minősége, oligopol piacok külön határoznak meg árakat. Habár az árváltozás iránya, nagysága nagyon hasonló, tehát a különböző árak együtt mozognak.

A valóságban különböző típusú kőolaj létezik – ez a sűrű, feldolgozatlan folyadék, amelyet a fúrók a föld mélyéből vonnak ki –, és némelyik kívánatosabb, mint mások. Például a finomítók könnyebben állítanak elő benzint és gázolajat alacsony kéntartalmú vagy „édes" nyersanyagból, mint a magas kéntartalmú olajból. Az alacsony sűrűségű vagy „könnyű" nyersolaj általában ugyanazon okból kedvez a nagy sűrűségű fajtának. (Jiang, An, Jia, & Sun, 2017)

Az is számít, hogy honnan származik az olaj, ha Ön vevő. Minél olcsóbb a termék szállítása, annál olcsóbb a fogyasztó számára. Szállítási szempontból a tengeren kitermelt olajnak vannak bizonyos előnyei a szárazföldi készletekhez képest, amelyek a csővezetékek kapacitásától függenek.

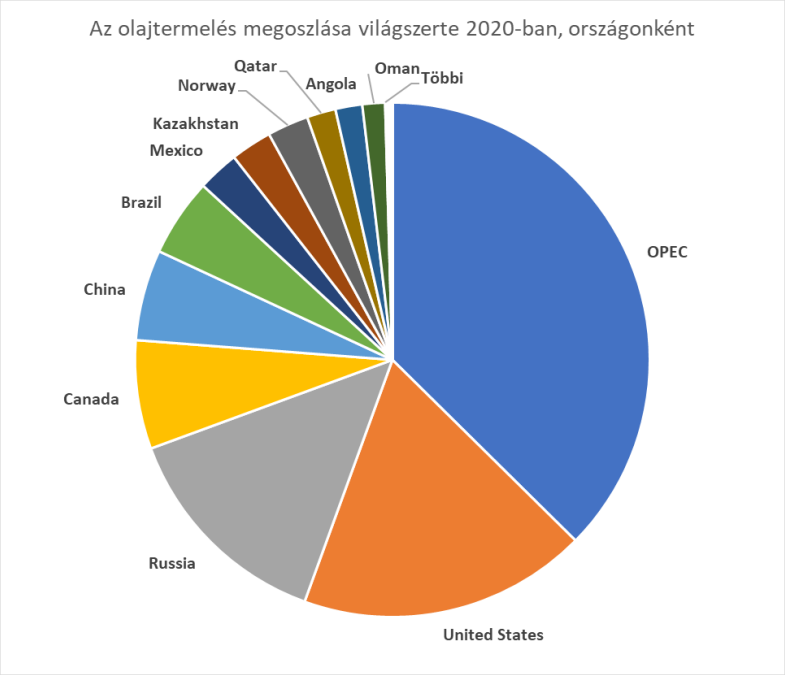
E tényezők miatt a kőolaj vásárlóinak – a spekulánsokkal együtt – egyszerű módszerre van szükségük az áru minősége és elhelyezkedése alapján történő értékelésére. Ezt a fontos célt szolgálják az olyan benchmarkok, mint a Brent, a WTI és a Dubai/Omán. Amikor a finomítók Brent-szerződést vásárolnak, határozott elképzelésük van arról, hogy milyen jó lesz az olaj, és honnan származik. Manapság a globális kereskedés nagy része a határidős piacon zajlik, és minden kontraktus egy bizonyos olajkategóriához kötődik.

A kereslet és kínálat dinamikus természete miatt az egyes benchmarkok értéke folyamatosan változik. Hosszú távon diszkonttal is elérhetővé válhat egy másik indexhez prémiummal értékesített marker. (Arshad, Rizvi, Haroon, Mehmood, & Gong, 2021)

A világ összes nyersolajszerződésének nagyjából kétharmada a Brent Crude-ra vonatkozik, így ez a legszélesebb körben használt jelző. Manapság a „Brent" valójában az Északi-tenger négy különböző mezőjéből származó olajra utal: Brent, Forties, Oseberg és Ekofisk. Az ebből a régióból származó nyersolaj könnyű és édes, így ideális dízel üzemanyag, benzin és más nagy keresletű termékek finomításához. És mivel az ellátás vízi úton történik, könnyen szállítható távoli helyekre.

A WTI az Egyesült Államok kútjaiból kitermelt olajra vonatkozik, amelyet csővezetéken az oklahomai Cushingba küldenek. Maga a termék nagyon könnyű és nagyon édes, így különösen benzinfinomításhoz ideális. A WTI továbbra is az Egyesült Államokban fogyasztott olaj fő etalonja.

Ez a közel-keleti nyersolaj hasznos referencia a WTI-nél vagy a Brentnél valamivel gyengébb olajhoz. Dubajból, Ománból vagy Abu Dhabiból származó nyersanyagból álló „kosár" termék, valamivel nehezebb és magasabb a kéntartalma, így a „savanyú" kategóriába sorolható. Dubai/Omán a fő referencia az ázsiai piacra szállított Perzsa-öböl olaj tekintetében. A 2. ábra Az olajtermelés megoszlása világszerte 2020-ban, országonként, saját szerkesztés mutatja, hogy az olajpiaci kitermelés oligopol piacot képez, vagyis egyes szereplők döntései, nyilatkozatai, előrejelzései vagy épp lokális válságai nagy hatással lehetnek a piaci árakra.



2. ábra Az olajtermelés megoszlása világszerte 2020-ban, országonként, saját szerkesztés (IEA, 2020)

A fentiek értelmében tehát nem lehet egyértelműen meghatározni egy fő benchmarkot.

Értelemszerűen vagy a Brent, vagy a WTI árat szükséges alkalmazni. A hivatalos múltbeli adatokat az Egyesült Államok Energiainformációs Ügynöksége, vagyis az EIA (U.S. Energy Information Administration) adatbázisából töltöttem le. (EIA - Cushing, OK WTI Spot Price FOB (Dollars per Barrel), 2022.) Szükséges vizsgálni az eltéréseket az idősorban, a spreadeket hosszú ideje a normál 2 dollár/hordó körüli tartományban tartják. A 2011 és 2015 közötti időszakban azonban a kettő szakaszosan szétvált, 2012 végén pedig extrém, 24 dolláros felár következett be. Első szinten a WTI és a Brent szórás oka 2011 és 2015 között kínálati szintű tényezők voltak: 2011. január–2011. április, 2012. április–2013. április, 2012. december–2013. április és 2015. január–március a világ nyersolaja a kínálat növelte a Brent olajárakat; 2012. január – 2012. április, 2013. augusztus – 2013. december és 2014. és 2015. július között a Cushing készletténye zője jelentősen csökkentette a WTI olajárakat, ami a kettő közötti különbséget kiszélesítette; Keresleti szintű tényezők: A 2011-től 2015-ig tartó időszakban, amikor a világgazdaság lanyha volt, a kőolaj iránti tényleges fogyasztási kereslet és az olajárakra nehezedő nyomás hatására a WTI-árindex nagyobb mértékben esett, mint a Brent-index, ami drámaibb növekedést eredményezett. a WTI árak csökkenése, és jóval alacsonyabb a Brent mutatórendszernél. Ami a spekulatív tényezőket illeti, mivel a WTI a Cushing-részvény alá tartozik, ára alacsonyabb, mint a Brent olaj ára, így a határidős piaci kereskedők nagyobb valószínűséggel folytatnak arbitrázs kereskedést a New York-i Kereskedelmi Határidős Tőzsdén. A 2015 előtti amerikai kőolajkivitel-tilalom hatása miatt azonban a növekedés csekély mértékű volt. Csak a kőolajkiviteli tilalom feloldása után volt akadálytalan az Egyesült Államok kőolajexportja.

A második szinten a jelenlegi Brent kőolaj határidős árindexe jobban tükrözi a kőolaj világpiacának kínálati szintjének ingadozásait, a WTI pedig a kőolaj világpiaci keresleti szintjének ingadozásait. Mindkettőnek azonban megvan a maga problémája: Például az északi-tengeri olajmező kimerülése a kitermelés gyors csökkenéséhez vezetett, ami a jövőben befolyásolhatja a

Brent-index világszintű befolyását a kőolaj határidős piacán; az amerikai palaolaj kitermelés növekedése után a korlátozott hazai szállítási kapacitás miatt a Cushing készletében bekövetkezett változások jelentősen befolyásolták a WTI árindex rendszer változásait, jobban tükrözve az amerikai belföldi piac változásait. (Tian & Lai, 2019) (Guerrero-Escobar, Hernandez-del-Valle, & Hernandez-Vega, 2019) (Liu, Stevens, & Vedenov, 2018)

Habár a Brent jobban tükrözi a világpiaci olajár változást, viszont a kutatás további részeit figyelembe véve, vagyis, hogy egy amerikai kiadású újságot vizsgálunk, ami erősen foglalkozik a belföldi olajkitermeléssel, így a cikkekben is gyakran jelenhetnek meg erre vonatkozó utalások, ezért a továbbiakban olajár-benchmarkként a WTI árat használjuk.

Az olajár változása kapcsán célunk az áremelkedés és a csökkenő trendek meghatározása, ezen belül pedig a fordulópontok pontos meghatározása és előrejelzése. Csakúgy, mint e fordulópontok kiszámíthatóságának vizsgálata és módszere.

A mozgóátlagos konvergencia-divergencia (MACD) az egyik legmegbízhatóbb és leggyakrabban használt momentummutató. Számítási módja viszonylag egyszerű: két különböző periódusszámú mozgóátlag hányadosa. Több változata ismert, jelenleg a 12 napos exponenciális mozgóátlagot (EMA) osztják a 26 napos exponenciális mozgóátlaggal.

A mutató értéke 1 körül ingadozik. A legtöbb esetben az MACD kiszámítása a két mozgóátlag különbségének figyelembevételével történik, de jelen tőzsdei esetben és kereskedési módszerben a százalékos eltolódás a tényleges piaci mozgásokat mutatja.

Az exponenciális mozgóátlag (EMA) általános képlete a következő:

$$EMA_n = P_{oil}\frac{2}{T+1} + EMA_{n-1}(1-\frac{2}{T+1})$$

A mozgóátlag konvergencia divergenciájának klasszikus képlete:

$$MACD_n = EMA_{12,n} - EMA_{26,n}$$

Ehelyett az aktuális probléma érvényesebb elemzése és a korábban ismertetett tőzsdei dinamika alapján a következő képletet használjuk:

$$MACD_n = EMA_{12,n} / EMA_{26,n}$$

A használt Signal vonal az egy mozgóátlag a MACD vonalra, és a legtöbbször 9 napos exponenciális mozgóátlagként számolják ki a MACD vonalon. Egy késleltetett indikátor, amely segít kiszűrni a rövid távú ingadozásokat, és jobban azonosítani a trendek valódi megfordulását:

$$Signal_n = MACD_n\frac{2}{T+1} + Signal_{n-1}(1-\frac{2}{T+1}) \qquad \text{T=9}$$

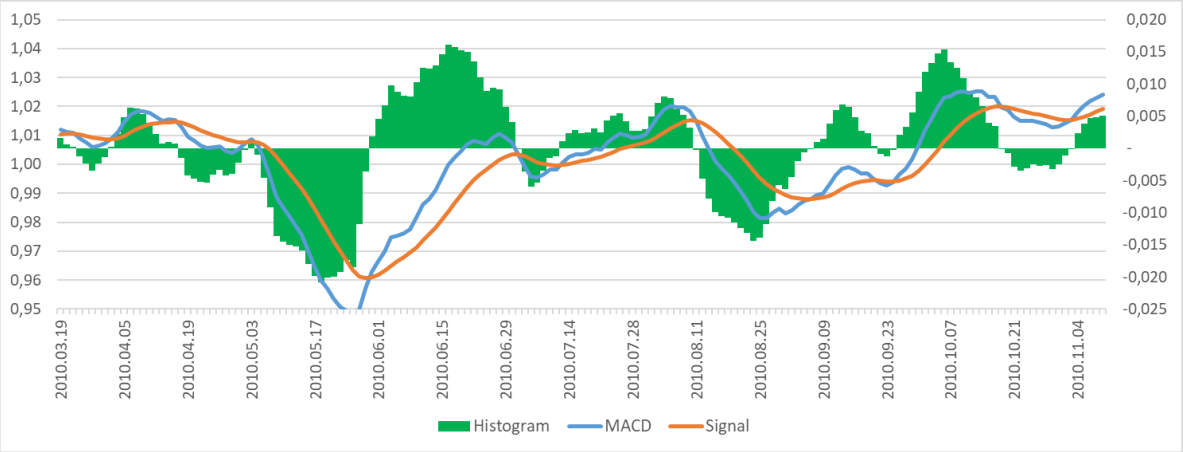
Ezenkívül a vizuális megjelenítéshez használt hisztogram képlet:

$$\text{Histogram}_n = MACD_n - Signal_n$$

A leírt módszerrel megrajzolhatjuk a trendeket, meghatározhatjuk a múltbeli trendfordulópontokat, így vizsgálva a trendfordulók előtti időszakot, összefüggéseket keresve a

cikkek kulcsszavai és a tőzsdei mozgás között. Pontosan a kulcsszavak mennyiségét tekintve a trendfordulók előtt.

A fentebb leírt vizuális ábrázolását a 3. ábra mutatja be.



3. ábra Olajár MACD indikátorjellel (Signal) és hisztogrammal (Histogram) (részlet)

4.1.3 Wall Street Journal

Az elemzett folyóirat esetében szükséges volt olyat találni, ami megfelel az alábbi feltételeknek:

- vezető folyóirat, nagy olvasói számmal, vagyis tényleges ráhatása lehet a piacok működésére, spekulatív döntéshozatalra,
- rendelkezik archív állománnyal, mely hozzáférhető, kellően strukturált,
- scrape-elhető, bár ez inkább programozási kérdés.

A lehetséges folyóiratokat több későbbi felhasználáshoz köthető szempont kapcsán elemeztem, melyeket a 4. ábrában mutatok be.

	saját cikkek	archívum	csak üzleti cikkek	havi oldalmegtekintés	alkalmasság [%]
Forbes	igen	van, de nem egyértelmű	nem	~ 95 millió	75%
Businessinsider	igen	hiányos, csak 2 év visszamenőleg	igen	~ 96 millió	20%
Wall Street Journal	igen	igen	igen	~ 71 millió	100%
Bloomberg	igen	nincs	igen	~ 73 millió	0%
Reuters	igen	hiányos, csak 2 év visszamenőleg	igen	~ 72 millió	20%
Yahoo! finance	nem	nincs	igen	~ 264 millió	0%
Cnbc	igen	nincs	igen	~ 147 millió	0%
NYTimes	igen	igen, strukturálatlan	nem	~ 339 millió	90%

4. ábra Üzleti hír weboldalak összesítése (forrás: havi oldalmegtekintés: (similarweb, 2022.))

A weboldalak elemzése során az Wall Street Journal (továbbiakban: WSJ) bizonyult legalkalmasabbnak az elvégzendő kutatás során az adatgyűjtésre, így itt került lefuttatásra a scraper.

A kutatás során a WSJ 2000-2020 között kiadott cikkeit töltöttem le, pontosabban scrapeltem. A kódsor a M4. Wall street Journal python kód mellékletben olvasható, a következőkben a funkciókat részletezem.

A WSJ token authorizációt alkalmaz. A token egy azonosító, amit bejelentkezéskor vagy egyéb más módon kaphatunk meg, megmarad a munkafolyamat során és ezzel igazoljuk a weboldal felé, amit igazolnunk kell. Vagyis a múltbeli cikkek teljes elolvasása (illetve a jelenbeli cikkek is) csak bejelentkezés és előfizetés esetén lehetséges. A nagy számú oldallekérés miatt szükséges volt session request-et alkalmazni, illetve a login session során a bejelentkezési url is változó kódsort tartalmaz. Vagyis szimulálni a bejelentkezést, az így kapott authorizációs kódot lementeni, majd a további oldallekérésekbe beilleszteni, ezzel biztosítva a weboldalt, hogy nem vagyunk robotok.

A WSJ archívuma könnyen indexelhető a különböző napokra, így az év, hónap és napok külön lettek ütemezve. Az adatbázis indexálása fontos a későbbi feldolgozás hatékonysága szempontjából.

Az adott napon közzétett cikkek linkjét listába gyűjtöttem. Számomra csak a cikkek voltak fontosak, így került bele az „article” megkötés. A videók, ajánlások, reklámok ezáltal nem kerültek bele a listába, csak a későbbiekben elemzendő cikkek.

A mentés folyamatossága, esetleges hibakezelés miatt minden évet külön excel fájlba mentettem. Így egy kritikus hiba következtében csak az adott évet kell újrakezdeni. Hibák közé tartozik az oldal adott pillanatokban való többszöri próbálkozás utáni nem betöltése, az authorizáció kód elhagyása, a böngésző, illetve a weboldal felismerte az automatizált bot tevékenységet, stb.

A cikkek esetében a későbbi esetleges hasznosítás miatt összefoglaltam, vagyis a rövidített verziót is mentettem.

A kódsorba a túl gyors lekérés miatt késleltetést építettem be. Adott időn belül egy IP címről érkező request (lekérés) esetén a weboldal bot-nak minősítheti a felhasználót és letiltja. A késleltetések és szünetek miatt a kódsor lefutása kellően lassú, így humán felhasználó tevékenységnek minősíthető az online tevékenység.

Illetve a scrapelés során az internetre VPN-en keresztül csatlakoztam, így egy esetleges IP cím letiltás esetén manuálisan orvosolható egy újabb, még nem tiltott IP cím generálása. Amennyiben IP cím letiltás történik, úgy az adott weboldal adott ideig (általában 24 óra) nem küld semmilyen adatot lekérés esetén. Így az ilyen letiltások drasztikusan hosszabbíthatják a kódsor futását, vagyis ki kellett küszöbölni. A VPN (Virtual Private Network) a „virtuális magánhálózat” rövidítése – egy olyan szolgáltatás, amely védi az internetkapcsolatot és az online adatvédelmet. A VPN-ek titkosított csatornát hoznak létre az adatok számára, IP-címének elrejtésével védik online személyazonosságát, és lehetővé teszik a nyilvános Wi-Fi hotspotok biztonságos használatát. Tehát a weboldal számára egy kenyai, holland, kínai vagy bármely más országból érkező felhasználó mutatkozik meg.

A kódsor teljes lefutása 1604 órát vett igénybe, vagyis közel 67 napig futott folyamatosan. Visszafejtve egy hónap cikkeit átlagosan 401 perc alatt nézte át és mentette le. A 2000-2020 közötti időszak, vagyis a 21 év összesen 330.435 cikket jelent. Az így készített .xlsx fájl 558 MB méretű lett.

4.1.4 ANN – Mesterséges neurális háló

A Mesterséges neurális hálóval történő vizsgálat során többféle beállítást alkalmazok, bővítem vagy elveszek belőle, ezzel is karcsúsítva az eredményt, aszerint, hogy minél pontosabb, minél hatékonyabban becsülje meg a jövőbeli változásokat. Első körben az árfolyam százalékos változásaira koncentrálok, vagyis a cikkek tartalma alapján mekkora valószínűséggel tudom megmondani, hogy a következő napi árfolyamváltozás milyen irányú és mekkora nagyságú lesz.

4.1.5 Felépítés

Két különböző módszert alkalmaztam. Az első esetben minden egyes dátumot és árfolyamváltozást figyelembe vettem. A második esetben csak azokat, amik elértek egy bizonyos indikátorváltozót, erről a későbbiekben részletesebben lesz szó. Vizsgálva azt, illetve elkerülve a kevés cikkes szavak torzító hatását és nem árfolyam, hanem cikk oldalról vizsgálva csak.

A neurális háló felépítése és működése, egészen pontosan a felkészítési szakasz több részre oszlik. A teljes kódsor megtalálható a M5. Mesterséges Neurális Háló python kód mellékletben. A továbbiakban a kódban található funkciókat részletezem.

Első körben az olajárfolyamot készítettem elő, illetve ott a napi változást számoltam százalékos értékben. Az olajár vonatkozásában egyéb teendő nem volt.

A cikkek esetében indikátor és kulcsszavakat használtam. Az indikátorszavak jelen esetben olyan szavak, amelyek előfordulásával valószínűsítetten, hogy a cikk témája olaj árral kapcsolatos, így javasolt további vizsgálatra. Kulcsszavak azok a szavak, amelyeknek a mennyiségi előfordulását vizsgáltam a cikkekben, az előfordulási gyakoriságukból és a korábbi árfolyamváltozások között kerestem kapcsolatot, ami mintát adhat a jövőbeli változásokra. A

szavak összeállítása saját választáson alapul, mind árfolyam emelkedésre, mind politikai befolyásolásra, termelési szavakra vonatkozik. Az ANN menet közben súlyozza, hogy pontosan melyikre mekkora szükség van, mennyire vegye figyelembe. Valamint a több változat futtatása miatt a hasznosság és a végső formula tervek szerint kialakul.

Indikátorszavaknak jelöltem a következőket: `'opec', 'oil price', 'wti', 'crude oil',`

Kulcsszavaknak pedig az alábbiakat: `'increase', 'rise', 'rising', 'grow', 'optimism', 'enhance', 'expensive', 'climb', 'optimal', 'agreement', 'cooperation', 'solution', 'deal', 'bull', 'gain', 'demand', 'positive', 'decrease', 'bear', 'fall', 'low', 'cut', 'dramatically', 'pessimism', 'emergency', 'emerge', 'recession', 'collapse', 'negative', 'reduce', 'disagree', 'decline', 'cheap'`

Minden cikkben megszámoltam mind az indikátorszavak, mind a kulcsszavak előfordulását. Majd ezt követően egy új, indikátoroszlopban összesítettem az indikátorszavak együttes előfordulását. A későbbiekben ennek jelentősége lesz, a pontosítás céljából emelem a minimális indikátorszó előfordulást, vagyis a cikket csak akkor veszem validnak, ha megvan a minimum indikátor szó előfordulás.

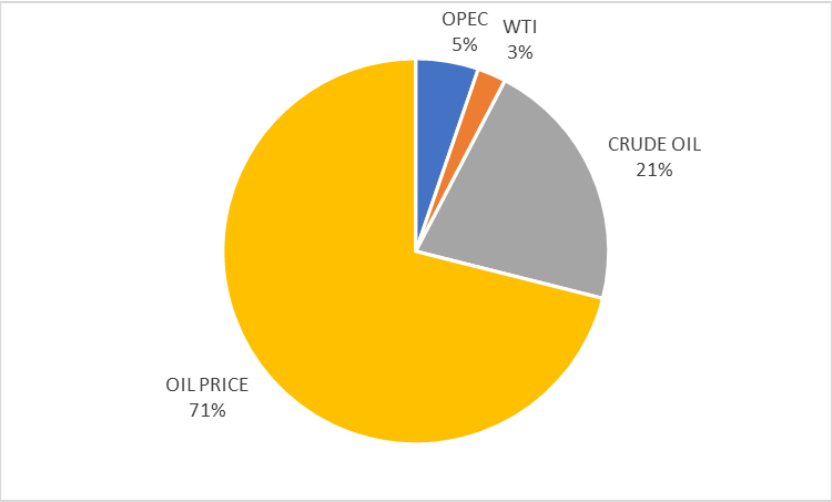
A törléses és törlés nélküli elemzés jelen fázisban vált ketté. Törlés nélkülinél minden napot elemeztem, azon napokat, ahol nem volt megfelelő indikátorszám, ott nullának tekintettem minden szó előfordulását.

Törléses elemzés esetén pedig ignoráltam minden olyan napot, ahol nincs meg a meg a kellő indikátorszám. Így az adatbázis jelentősen kisebb, erre is szükséges volt figyelni a vizsgálatnál.

Valamint mindig ellenőriztem, hogy mekkora darabszám maradt. Kis elemszám esetében a módszer lehet hatékony, de túl ritkán szolgáltathat jó előrejelzést, így a ritkaság miatt nem kellően hasznos.

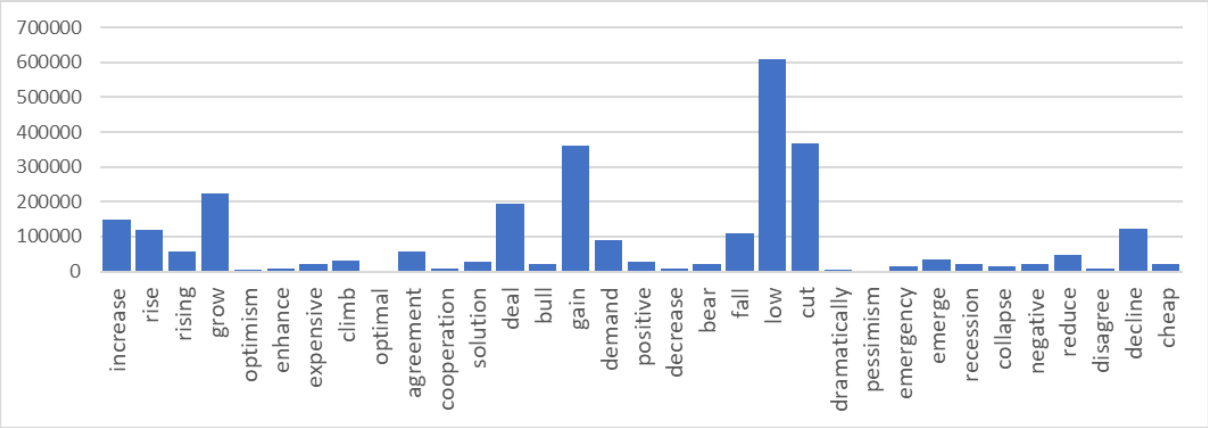
Illetve minden esetben összesítettem napok vonatkozásában az eredményeket. Az indikárotszavak és a kulcsszavak előfordulási volumenét a ~330 ezer cikkben az 5. ábra Indikátorszavak eloszlása a WSJ vizsgált cikkeben előfordulási gyakoriság szerint (saját szerkesztés) és 6. ábra Kulcsszavak eloszlása a WSJ vizsgált cikkeben előfordulási gyakoriság szerint (saját szerkesztés)mutatja be.

Az indikátorszavak előfordulása a következőképpen alakult:



5. ábra Indikátorszavak eloszlása a WSJ vizsgált cikkeben előfordulási gyakoriság szerint (saját szerkesztés)

Valamint a kulcsszavak gyakorisága:



6. ábra Kulcsszavak eloszlása a WSJ vizsgált cikkeben előfordulási gyakoriság szerint (saját szerkesztés)

Miután összegyűjtöttem a kulcsszavakat, illetve dátum szerint rendeztem, egy másik pandas adattáblába az olajárakat, az egyesítés előtt szükséges volt néhány előkészületet elvégezni. A pandas egy Python könyvtár, amelyet adatkezelésre és -elemzésre használnak, és hatékony adatstruktúrákat biztosít, mint a DataFrame és a Series. Segítségével könnyedén lehet adatokat manipulálni, szűrni, aggregálni és vizualizálni. Mivel az olajár folyamatosan emelkedett, értem itt az inflációt, így 20 év távlatában nem hasonlíthatjuk össze a jelenértékét, valamint nem is a pontos olajárat szeretnénk előre jelezni, hanem az árfolyam változását, így egységesítve az előző napi árfolyamhoz mért változást vesszük figyelembe. Illetve, egészen pontosan az aznapi árfolyam függvényében a következő napi változást, aznapra vonatkoztatva. Ez a neurális háló miatt fontos, mivel az aznapi cikkeket veszem alapul a holnapi árfolyamváltozás függvényében, vagyis valahol az egy nap eltérést szükséges korrigálnom, így ezt az olajár változásánál teszem meg.

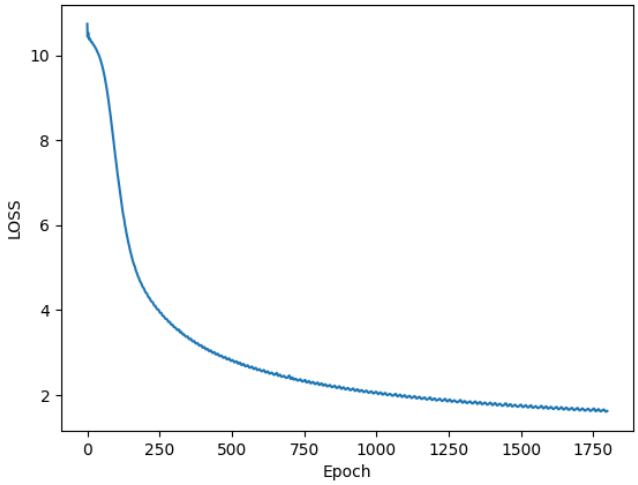
Majd egyesítettem a két pandas adattáblát dátum szerint. Így rendelkezésre állt a közös adattábla, melyet fel tudtam használni a neurális hálóhoz. További probléma, hogy az olajár nem folyamatos, vagyis hétvégén és adott ünnepnapokon nincs kereskedési nap, ezzel szemben az újságcikkek minden nap megjelenhetnek. Vagyis vannak, pontosabban lehetnek olyan

kulcsszavakat tartalmazó napok, melyeken nem ismert olajár változás van. Ennek kiküszöbölésére minden hiányos nap esetében a következő napi árfolyamváltozást vonatkoztattam, ezzel feltöltve a hiányos mezőket.

Így már egy teljes és hiányos mezők nélküli adattábla állt rendelkezésre. Több beállítással futtattam a neurális hálót, ezeket a 4.1.6 Eredmények fejezetben részletezem. Összesen 7632 nap adatát vizsgáltam, ennek 70%-a a tanítóhalmaz, 30%-a pedig a későbbi teszthalmaz.

A disszertáció további részében a teljes neurális hálót nem írom le, csupán az egyes változatok közötti különbséget részletezem.

A neurális háló tesztelése, illetve tanítása során a veszteségeket, vagyis a hibahatárokat csökkentem. Többször lefuttatom és „backward propagation", vagyis visszafejtéssel változtatok a bias értékeken. A backpropagation egy tanulási algoritmus a neurális hálókban, amely visszatáplálja a hiba értékeit a háló rétegei között. Ennek célja, hogy a súlyokat úgy módosítsa, hogy a háló kimeneti hibája csökkenjen. A bias egy hozzáadott érték a neurális háló neuronjaihoz, amely segít az aktivációs függvény eltolásában. Ez lehetővé teszi a modell számára, hogy jobban illeszkedjen az adatokhoz, különösen a nemlineáris kapcsolatok kezelésében. Az így lefutott köröket epoch-nak nevezi a szaknyelv, minden egyes körben vizsgáljuk a veszteséget és addig folytatjuk, amíg érdemben nem sikerül csökkenteni, illetve egy ponton túl a csökkenés mértéke már túl kismértékű, hogy azt érdemben folytatni kelljen. Az epoch tehát egy teljes átfutást jelent a neurális háló tanítási adathalmazán, amely során az összes minta egyszer átmegy a hálón. Több epoch alatt a modell folyamatosan finomítja a súlyait a jobb teljesítmény érdekében. Loss a modell teljesítményét méri az adott kimeneti előrejelzések és a valós értékek közötti különbség alapján. Ez a visszajelzés segíti a hálót abban, hogy a tanulási folyamat során javítsa a pontosságát. 7. ábra Epochs és Loss, futtatásonkénti hibajelölés (saját szerkesztés)jól láthatjuk, ahogy a körök növelésével folyamatosan csökken a hiba, veszteség mennyisége.



7. ábra Epochs és Loss, futtatásonkénti hibajelölés (saját szerkesztés)

Jelen esetben 15.000 kört (epoch) alkalmaztam. Ez a szám volt az, ami növelése után már érdemben nem tudtam csökkenteni a hibahatárt, de eddig még számottevően és folyamatosan csökkent. Az ábrán egy 1800 körös backward propagation-t mutatok be, a kevesebb kör miatt jobban látszik a hibák csökkenése, vagyis a pontosabb eredmény fokozatos elérése. Először a

javulás nagymértékű, majd egyre lassabb. Az ideális körszámot akkor érjük el, amikor a javulás mértékének változása a nullához konvergál. Ebben az esetben célszerű befejezni. Az eredmény ugyan még javul, viszont túlzott energiaráfordítást igényel, ami az optimalizált számítási kapacitás miatt már nem éri meg.

4.1.6 Eredmények

A korábban leírtak szerint a mesterséges neurális hálót több indikátorszámra futtattam, valamint vizsgáltam, hogy az egyes esetszámoknál elfogadott hiba esetén mekkora mértékben ad valid eredményeket.

Első körben törlés nélküli modellt alkalmaztam, vagyis amikor elemeztem azokat a napokat is, ahol olajárfolyamváltozás volt, de a cikkek között nem volt elég indikátor, hogy olajjal kapcsolatos cikket feltételezzen. Vagyis minden kulcsszó darabszám 0-val lett kalkulálva.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,7%	0,6%	0,7%	0,5%	0,5%	0,9%	1,0%	1,0%	0,8%
0,04	1,1%	1,2%	1,5%	1,4%	1,4%	1,5%	1,5%	1,7%	1,5%
0,07	1,9%	2,2%	2,8%	3,1%	3,3%	3,3%	3,4%	3,6%	3,3%
0,1	2,8%	3,7%	4,2%	4,2%	4,8%	5,3%	5,4%	5,2%	5,3%
0,15	3,6%	5,3%	6,2%	6,1%	6,8%	7,0%	7,0%	7,4%	7,1%
0,25	6,3%	9,3%	10,1%	9,9%	11,0%	11,7%	11,9%	12,2%	12,0%
0,4	10,6%	13,9%	16,0%	16,0%	17,6%	18,5%	18,7%	19,1%	18,9%
0,8	21,7%	26,1%	28,8%	30,0%	32,6%	33,7%	34,4%	34,9%	34,6%
1,2	31,3%	36,6%	40,6%	42,1%	45,6%	47,6%	48,3%	48,7%	48,7%
1,6	38,9%	46,3%	50,7%	52,4%	56,1%	58,8%	59,7%	60,2%	59,9%
2	46,2%	54,6%	59,1%	61,0%	65,6%	68,2%	68,9%	69,3%	69,3%
2,5	54,9%	62,5%	67,5%	69,2%	74,6%	77,0%	77,1%	77,8%	77,9%
3	61,5%	68,5%	73,4%	75,5%	80,4%	83,1%	82,9%	83,6%	83,6%
3,5	66,6%	73,8%	78,5%	79,8%	84,6%	86,6%	86,7%	87,3%	87,3%
4	71,7%	77,9%	82,2%	84,3%	88,0%	90,0%	90,0%	90,5%	90,6%
SIGN	60,2%	65,8%	69,0%	69,3%	72,2%	72,7%	73,2%	73,5%	73,6%

1. táblázat: törlés nélküli ANN eredmények

Az 1. táblázat jól mutatja, hogy az indikátorszámok növelésével nő a hatékonyság. Ez előre kalkulálható volt, nem volt meglepetés. Minden sorban sötétebb zölddel jelöltem a magasabb értékeket. Az elfogadott hibahatárok vonatkozásában jól látható, hogy 3-4-7 indikátor alkalmazása esetén éri el azt a számot, ami már érdemben az indikátorszámok növelésével nem növelhető túlzottan, ezt szemlélteti a 2. táblázat. A SIGN sor jelen, illetve minden következő táblázatban az előjel jelzés pontosságát mutatja, vagyis az árfolyam emelkedése vagy csökkenése esetén az ANN által készített becslés mennyire jól és hatékonyan találta el a trendet.

Az elfogadott maximális eltérést, vagyis a fő mérőszámot úgy kell értelmezni, hogy mekkora hibahatáron belül vesszük validnak az értéket. Vagyis, nem tűpontos eredményt, hanem a valós eredmény eltérésen belüli árjelzését pontosnak vesszük. Értelemszerűen minél nagyobb

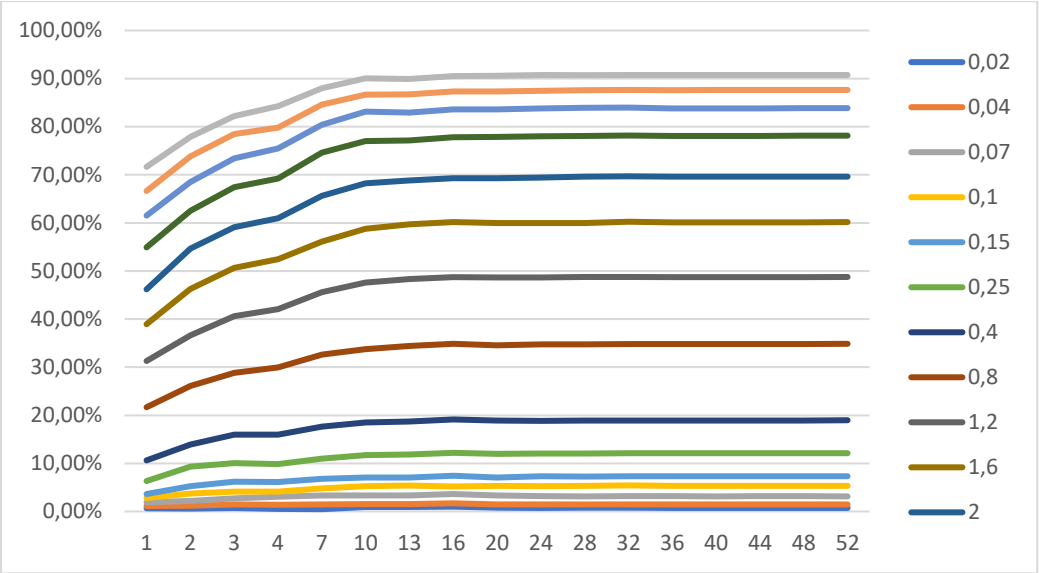
hibát engedünk meg, annál nagyobb lesz a hatékonyság. Az így kapott százalékos eredmények azt mutatják, hogy az összes becslés hány százaléka volt a tényleges árváltozás hibahatárán belül.

A napok vonatkozásában korábban említésre került nap elcsúsztatás is, viszont ez nem növelte a hatékoynságot, így mindig az előrejelzést az előrejelzett napra vonatkoztatom, nem csúsztatom el. Vagyis nem a következő napra vonatkoztatom. Mivel a teljes vizsgálatok így folytak le, illetve mindehol végeztem szimultán vizsgálatot, amelyeket sok esetben nem említek, ezért a táblázatokban bentmarad a *t.* napra vonatkozó becslés.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02		-0,1%	0,2%	-0,2%	0,0%	0,4%	0,0%	0,0%	-0,2%
0,04		0,0%	0,3%	-0,1%	0,0%	0,1%	0,0%	0,2%	-0,2%
0,07		0,3%	0,6%	0,3%	0,3%	0,0%	0,0%	0,3%	-0,3%
0,1		0,9%	0,4%	0,0%	0,7%	0,5%	0,1%	-0,2%	0,2%
0,15		1,7%	0,9%	-0,1%	0,7%	0,2%	0,0%	0,4%	-0,4%
0,25		3,0%	0,7%	-0,2%	1,1%	0,7%	0,2%	0,3%	-0,2%
0,4		3,3%	2,1%	0,0%	1,6%	0,9%	0,2%	0,4%	-0,2%
0,8		4,5%	2,7%	1,2%	2,7%	1,1%	0,7%	0,5%	-0,3%
1,2		5,3%	4,0%	1,5%	3,5%	2,0%	0,7%	0,4%	0,0%
1,6		7,3%	4,4%	1,7%	3,7%	2,7%	0,9%	0,5%	-0,3%
2		8,4%	4,5%	1,9%	4,6%	2,6%	0,6%	0,4%	0,0%
2,5		7,6%	4,9%	1,8%	5,4%	2,4%	0,1%	0,7%	0,0%
3		7,0%	4,9%	2,0%	4,9%	2,8%	-0,2%	0,6%	0,1%
3,5		7,2%	4,7%	1,3%	4,8%	2,1%	0,1%	0,6%	0,0%
4		6,2%	4,3%	2,1%	3,7%	2,1%	-0,1%	0,6%	0,0%
SIGN		5,6%	3,2%	0,3%	2,8%	0,5%	0,6%	0,3%	0,1%

2. táblázat: Indikátorszámok közötti hatékonyság eltérések kimutatása

A változás mértéke a korábbi indikátorszám hatékonyságához viszonyított százalékos hatékonyságnövekedést jelöli. Jól kirajzolódik, hogy az 1 és 2 indikátorszám közötti váltáskor nagyban emelkedik a hatékonyság, majd ez fokozatos 7-ig, onnan kisebb mértékű, valamint a vége felé a hatékonyság stagnál vagy elhanyagolható mértékben változik, vagyis van határa az indikátorszámoknak, így túl sok alkalmazása már szintén nem hatékony. A futtatáskor 52 indikátorszám alkalmazásáig mentem el, az előbbi állítást egyértelműen igazolja.



8. ábra Törlés nélküli ANN hatékonyságának ábrázolása az elfogadott maximális eltérés függvényében

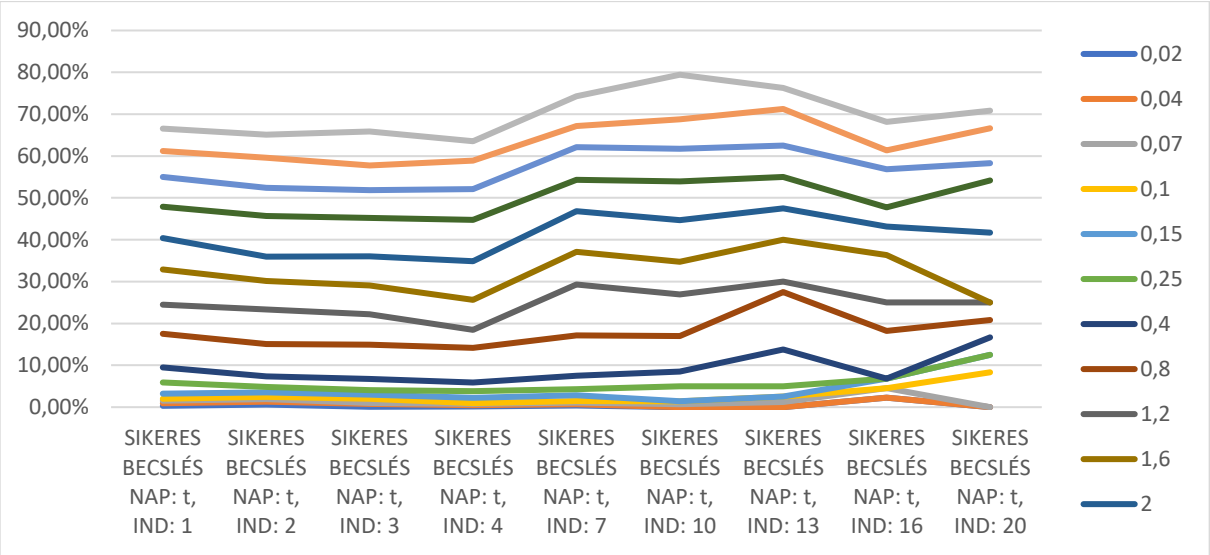
A 8. ábra jól mutatja, 7-10 indikátorszámig dinamikusan növekszik, utána viszont a stagnálás jellemző a hatékonyság vonatkozásában. Maximális eltérés vonatkozásában azt vizsgáltam, hogy az adott előre jelzett százalékos eltérés adott hibahatáron belül az összes becslés vonatkozásában hány százalékban bizonyult helyesnek, vagyis adott hibahatáron belülinek. Ezt azért volt fontos vizsgálni, mivel a változások nem feltétlenül tűpontos mértékére van minden esetben szükség, bizonyos aspektusokban az irány és a mérték nagysága már megfelelő előrejelzés lehet.

A vizsgálat további részénél lefuttattam olyan elemzést, melyben kitöröltem azokat a napokat, melyeknél nem volt meg a kellő indikátorszám. Vagyis azt a hipotézist ellenőriztem, hogy a megfelelő adat nélküli napok félrevezetőek lehetnek, így ignoráltam őket. Ebben az esetben a napok egy jelentős része kiesik, így a ANN-re ható fehér zaj, pontosabban a nagy változások, de kulcsszó nélküli lefutások minimumra vannak redukálva, ezzel is elősegítve a lehető legtisztább elemzést. A későbbiekben szükséges vizsgálni, hogy így mennyiségileg mennyi vizsgálható, úgyszólván elegendő mennyiségű vizsgált napunk marad-e. Az eredményeket a 3. táblázat mutatja be:

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,28%	0,66%	0,12%	0,16%	0,36%	0,00%	0,00%	2,27%	0,00%
0,04	0,96%	1,31%	0,81%	0,48%	0,71%	0,00%	0,00%	2,27%	0,00%
0,07	1,53%	1,89%	1,04%	0,80%	1,07%	0,71%	1,25%	4,55%	0,00%
0,1	1,98%	2,46%	1,97%	0,96%	1,43%	1,42%	2,50%	4,55%	8,33%
0,15	3,23%	3,52%	3,01%	2,23%	2,86%	1,42%	2,50%	6,82%	12,50%
0,25	5,89%	4,84%	4,05%	3,82%	4,29%	4,96%	5,00%	6,82%	12,50%
0,4	9,51%	7,38%	6,71%	5,89%	7,50%	8,51%	13,75%	6,82%	16,67%
0,8	17,55%	15,08%	14,93%	14,17%	17,14%	17,02%	27,50%	18,18%	20,83%
1,2	24,52%	23,36%	22,22%	18,47%	29,29%	26,95%	30,00%	25,00%	25,00%
1,6	32,90%	30,16%	29,05%	25,64%	37,14%	34,75%	40,00%	36,36%	25,00%
2	40,37%	35,98%	36,00%	34,87%	46,79%	44,68%	47,50%	43,18%	41,67%
2,5	47,90%	45,66%	45,25%	44,75%	54,29%	53,90%	55,00%	47,73%	54,17%
3	54,98%	52,38%	51,85%	52,07%	62,14%	61,70%	62,50%	56,82%	58,33%
3,5	61,21%	59,59%	57,75%	58,92%	67,14%	68,79%	71,25%	61,36%	66,67%
4	66,59%	65,08%	65,86%	63,54%	74,29%	79,43%	76,25%	68,18%	70,83%
SIGN	57,13%	56,23%	55,21%	52,71%	63,21%	63,83%	61,25%	54,55%	54,17%
CALC %	77,13%	53,28%	37,74%	27,43%	12,23%	6,16%	3,49%	1,92%	1,05%

3. táblázat: törléses ANN eredmények

A törléses vizsgálatnál a korábbinál magasabb indikátorszámnál van a leghatékonyabb pont, 7-10-13 indikátorszámnál éri el az átlagos csúcspontot. Habár érdekes megfigyelni, hogy 4 indikátor alkalmazásáig valamelyes csökken, 10 indikátornál csúcsosodik, majd ismét csökken. A SIGN érték esetében a változás előre jelzett és a tényleges változás egyező irányának hatékonyságát vizsgáltam, a CALC pedig az összes naphoz képest a vizsgált napokat mutatja. Jól látható, hogy magas indikátorszám esetében nagyon kevés vizsgált napunk marad.



9. ábra Törléses ANN hatékonyság ábrázolása elfogadott maximális eltérés függvényében

Azt mindenképpen fontos figyelembe venni a törléses vizsgálatnál, hogy a leghatékonyabban 10 indikátorszám alkalmazásával működtethető. Viszont ebben az esetben csak a 6,16%-át elemezte, ami viszonyításképpen egy 5 éves ciklust vizsgálva, ahol picivel több, mint 1800 nap van, csak 112 nap ad előrejelzést. Ez kevés, illetve ezekben az esetekben sem tűpontos az eredmény. Fontos volt meghatározni a későbbiekben, hogy szükséges-e megadni minimum vizsgált napszámot, amennyiben törléses ANN-t alkalmazok. A hatékonyság növekedését, stagnálását, majd csökkenését a 9. ábra Törléses ANN hatékonyság ábrázolása elfogadott maximális eltérés függvényébenmutatja be.

Külön-külön már elemeztem a törléses és törlés nélküli futtatást. A törléses esetében látjuk, hogy drasztikusan csökken a vizsgált elemszám, ami azt eredményezi, hogy relatíve kis elemszámot tudtam vizsgálni, ami egy napi szintű előrejelzésre nem alkalmas, vagy legalábbis nem feltétlenül minden esetben. Ennek ellenére szükséges volt megvizsgálni, hogy melyik módszer a hatékonyabb.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,38%	-0,09%	0,62%	0,36%	0,12%	0,92%	0,96%	-1,27%	0,83%
0,04	0,18%	-0,13%	0,68%	0,92%	0,73%	1,53%	1,53%	-0,57%	1,49%
0,07	0,39%	0,29%	1,71%	2,26%	2,25%	2,61%	2,11%	-0,92%	3,32%
0,1	0,82%	1,25%	2,18%	3,19%	3,38%	3,87%	2,87%	0,61%	-3,00%
0,15	0,40%	1,77%	3,19%	3,89%	3,96%	5,61%	4,53%	0,61%	-5,42%
0,25	0,44%	4,47%	6,00%	6,05%	6,68%	6,75%	6,88%	5,37%	-0,49%
0,4	1,11%	6,56%	9,28%	10,10%	10,11%	9,97%	4,95%	12,31%	2,25%
0,8	4,12%	11,04%	13,86%	15,80%	15,49%	16,71%	6,88%	16,68%	13,73%
1,2	6,76%	13,25%	18,37%	23,60%	16,32%	20,67%	18,32%	23,71%	23,67%
1,6	6,03%	16,10%	21,63%	26,78%	19,00%	24,05%	19,68%	23,84%	34,94%
2	5,81%	18,63%	23,07%	26,12%	18,83%	23,52%	21,35%	26,11%	27,62%
2,5	7,01%	16,86%	22,20%	24,45%	20,33%	23,12%	22,11%	30,08%	23,68%
3	6,53%	16,12%	21,59%	23,38%	18,24%	21,44%	20,42%	26,75%	25,29%
3,5	5,41%	14,20%	20,71%	20,85%	17,44%	17,84%	15,47%	25,97%	20,62%
4	5,06%	12,81%	16,36%	20,73%	13,70%	10,61%	13,70%	22,34%	19,73%
SIGN	3,03%	9,56%	13,82%	16,62%	8,96%	8,82%	11,97%	18,93%	19,40%

4. táblázat: törlés nélküli és törléses ANN közti eltérés [törlés nélküli % - törléses %]

A 4. táblázatban a törlés nélküli hatékonyságból kivontuk a törléses hatékonyságát azonos indikátorszám és elfogadott hiba függvényében. Kékkel azok láthatók, ahol a törlés nélküli hatékonyabb, sárga színnel pedig azok, ahol a törléses. Egyértelműen kitűnik, hogy a törlés nélküli sokkal eredményesebb, valamint a korábban említett kisebb elemszám miatt kijelenthetjük, hogy a törléses módszert a jövőben nem érdemes használni.

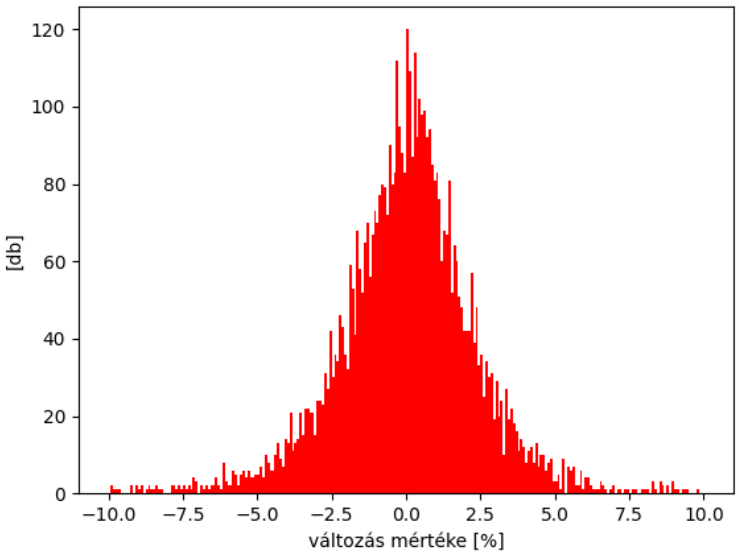
Vagyis több lehetséges ANN-beállítás és vizsgált modell közül az első tesztelésnél már egyértelműen és vitathatatlanul sikerült kijelölnöm a hatékonyabb utat, ezzel több felesleges beállítást és futtatást tudtam megspórolni.

4.1.7 Értékelés

Az eddigi eredményekből egyértelműen látszik, hogy a modell jelenlegi fejlettségével és beállításaival pontos eredményekre nem képes. Vagyis a következő napi árfolyamot csak elnagyolt becsléssel, nagy hibával tudja meghatározni. Fontos tisztázni, hogy mit értünk pontos becslés alatt. Amennyiben elfogadjuk a következő árfolyamváltozás esetében az 1,2%-on belüli eltérést, illetve pontosságot, úgy a modell az esetek ~45%-ban helyesen kalkulált. Ami nem feltétlenül mondható rossznak.

Emellett a SIGN hatékonyság, vagyis a következő napi árfolyamváltozás előjelének meghatározása 60% körüli pontossággal funkcionál, ami egyértelműen jelzi, hogy van kapcsolat a cikkek és az árfolyamelőrejelzés között. A vizsgált időszakban az árfolyam 51,77%-ban emelkedett. Az emelkedési többlet indokolt az inflációt figyelembevéve. Mindazonáltal, ha azt mondjuk, hogy az árfolyam másnap emelkedni fog, versus a Neurális Háló SIGN eredményei, a modell hatékonyabban mutatja ki a következő napi árfolyamot.

Valamint mindenképpen figyelembe kell venni az átlagos árfolyamváltozást, mennyire volatilis, mekkorák a kiugrások. A 10. ábra mutatja be, hogy elég erősen a 0 körül mozog, ahogy a változás mértéke abszolút értékben csökken, úgy egyre kisebb az előfordulási valószínűség.



10. ábra Olaj árfolyamváltozások eloszlása a vizsgált időszakban

Így egy relatíve szenzitív előrejelzés szükségeltetik. A 5. táblázat bemutatja, hogy adott abszolút eltéréseket figyelve mennyire volatilis az árfolyam, a vizsgált időszak változásainak mekkora része esik tartományon kívülre. Amit megfogalmazhatunk, hogy 10% alatti a nulla körüli változás, illetve az esetek nagyobb részében, ~60%-ban 1%-ot meghaladta az árfolyamváltozás. Vagyis azzal, hogy mindig 0-át jelölünk, nem tudunk hatékonyan előre jelezni, így a becslés mindenképp hasznos.

Vizsgált abszolút árfolyamváltozási határérték	4%	3%	2%	1.5%	1%	0.5%	0.2%
Tartományon kívül eső értékek aránya	8.81%	16.53%	32.36%	44.07%	59.32%	78.28%	91.09%

5. táblázat: Árfolyamváltozás abszolút mértékének vizsgálata adott határértékeken belül

A vizsgálat kritikájaként mindenképpen fontos megjegyezni, hogy a kulcsszavak helyes kiválasztása továbbra is kérdéses. Ezek megváltoztatásával akár rontani, akár javítani is lehet a jelenlegi hatékonyságon. Mindazonáltal az eddig elért eredmények nagyon biztatóak. A továbbiakban azt vizsgálom, hogy valamilyen módon lehetséges-e javítani az előrejelzés pontosságán. Továbbá fontos megjegyezni, hogy a kulcsszavak vagy adott szóösszetételek pontos előfordulásának vizsgálata minden bizonnyal növeli a hatékonyságot, viszont emellett a futtatási időt is drasztikusan. Neurális háló építésénél szükséges azt is szem előtt tartani, hogy egy kellően gyors modellről beszéljünk, ami tőle elvárható időn belül képes eredményt produkálni.

4.2 ANN + RNN– Mesterséges Neurális Háló + Visszacsatolt Neurális Háló kombinációja

A következő vizsgálatban lefutattam egy RNN, vagyis visszacsatolt neurális hálózatot. Első körben csak az egy napos árat becsültem meg, majd később a változást is. Illetve az így kapott eredményeket visszahelyeztem a módosított törlés nélküli ANN-be. Valamint vizsgáltam annak lehetőségét, hogy a 12, valamint 26 napos mozgóátlag, vagyis az MACD, így a Signal mekkora mértékben növeli vagy csökkenti a hatékonyságot.

Vagyis egy dupla neurális hálót alkalmaztam az eredmények javítása érdekében. Ez már egy összetettebb Deep Learning, vagyis mélytanulási metódus. Az eredeti ANN layer számai alapján már mélytanulásos módszerről beszélhetünk, ezen bővítés következtében egy bonyolultabb deep learning method jön létre, neurális háló eredményeket ágyazunk a neurális hálóba.

4.2.1 Felépítés

Az RNN esetében első körben felépítettem a neurális hálót. Elvégeztem a beállításokat és felépítettem a programkódot. A kódsor megtekinthető az M6. mellékletben (M6. Visszacsatolt Neurális Háló python kód).

Az így eredményül kapott nap-olajár előrejelzést beépítettem az ANN-be.

Egészen pontosan az idősoros előrejelzést két módszerrel végeztem. Az egyik esetben az olajár előrejelzést építettem be, a másik esetben az RNN az olajár százalékos változását kaptam, így ezt határoztam meg. Vagyis két külön idősoros RNN is beépítésre került, ezeket külön-külön futtattam és vizsgáltam a hatékonyságot.

Mindenképp figyelmet érdemel, hogy amikor az RNN-nel a jövőbeli százalékos árváltozást igyekeztem idősorosan előrejelezni, nagyon lassan változott, vagyis sokkal inkább úgy tűnt, hogy trendet mutat. Végső soron számomra a trend, vagyis a változás iránya és dinamikája fontosabb. De ez a számszaki elemzésből kiderül.

A neurális háló további részében egyéb változtatás nem volt, eggyel több bemeneti adat van, az epoch, layer és egyéb lényeges adatok ugyanazok maradtak, így ezeket nem fejtem ki újfent részletesebben.

4.2.2 Eredmények

Az indikátorszámok növelésével fokozatosan javult a hatékonyság. 20 indikátor alkalmazásánál éri el a csúcspontot, ezt követően egyes helyeken minimálisan még nő, itt 1%-on belüli változást értek, illetve az esetek nagy részében csökken is. A magas indikátorszámot mindig kétségekkel szükséges fogadnunk, mivel ezekben az esetekben csak a cikkek csekély részét elemezzük. Illetve, akár utalhat arra is, hogy nem elég volatilis, így egy folyamatos 0 előrejelzéssel járunk legjobban. Ezeket jelenleg nem kell figyelnünk, a kutatás későbbi szakaszaiban ezeket részletesebben elemezzük, illetve kiküszöböljük.

Az RNN első futtatás esetében az RNN a jövőbeli pontos árat jelezte előre, így az került a Neurális Hálóba plusz bemeneti adatként. Az eredményeket a 6. táblázat: ANN + RNN jövőbeli pontos ár eredmények tartalmazza.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,57%	0,87%	0,66%	0,87%	1,00%	0,74%	1,00%	0,79%	0,92%
0,04	1,18%	1,53%	1,22%	1,92%	1,83%	1,75%	2,10%	1,75%	1,88%
0,07	1,88%	2,66%	2,88%	3,28%	3,06%	3,23%	3,10%	2,93%	3,54%
0,1	2,45%	3,89%	4,06%	4,54%	4,19%	4,59%	4,33%	4,37%	5,11%
0,15	3,80%	5,50%	5,55%	6,73%	6,68%	7,21%	7,12%	6,90%	7,16%
0,25	7,60%	9,39%	9,70%	11,18%	11,84%	11,62%	12,15%	11,40%	12,28%
0,4	12,23%	14,55%	15,47%	17,30%	17,82%	18,57%	19,09%	18,48%	18,92%
0,8	23,24%	26,43%	27,35%	30,84%	32,55%	32,68%	34,73%	34,43%	34,21%
1,2	33,38%	37,53%	39,41%	42,68%	44,56%	46,79%	48,19%	48,01%	47,84%
1,6	40,67%	47,27%	49,37%	52,25%	55,26%	57,14%	58,54%	58,63%	58,41%
2	48,58%	54,83%	57,71%	61,03%	64,57%	66,71%	68,20%	68,81%	68,76%
2,5	56,05%	62,73%	65,92%	69,33%	73,35%	75,93%	76,93%	77,46%	77,24%
3	62,95%	69,72%	73,13%	75,49%	79,60%	81,61%	83,05%	83,40%	83,44%
3,5	68,55%	75,19%	78,37%	80,34%	83,66%	85,58%	86,68%	86,98%	87,42%
4	72,91%	79,69%	82,44%	84,27%	87,37%	89,38%	90,21%	90,43%	90,56%
SIGN	60,38%	66,01%	68,46%	69,86%	70,55%	71,87%	73,13%	73,04%	73,13%

6. táblázat: ANN + RNN jövőbeli pontos ár eredmények

Az eddigi eredményekhez hasonlóan kis hiba engedélyezése esetén nem pontos, illetve az indikátorszám növelésével folyamatosan javul a hatékonyság. Az indikátorszámok esetében egészen 20-ig nő a hatékonyság, utána stagnál, majd csökken. Vagyis kérdés, hogy mennyire az RNN-ből kapott eredmények dominálnak és háttérbe szorítják, vagy csak kis mértékben veszik figyelembe a WSJ cikkeit.

Valamint, szükséges összehasonlítani azokkal az eredményekkel, amikor az RNN nem a pontos árfolyamot, hanem az árfolyam százalékos változását „jósolta meg” az előző napi árfolyamhoz képest.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,61%	0,44%	1,00%	0,96%	0,87%	1,27%	0,66%	0,96%	1,27%
0,04	1,22%	1,18%	1,66%	1,75%	1,62%	2,14%	1,49%	1,62%	1,88%
0,07	1,97%	2,14%	2,75%	2,53%	3,28%	3,76%	2,75%	3,36%	3,23%
0,1	2,45%	3,06%	3,67%	4,06%	4,63%	4,94%	3,89%	4,89%	4,06%
0,15	3,58%	5,11%	6,16%	6,47%	6,60%	7,08%	6,86%	6,99%	6,86%
0,25	6,42%	8,61%	9,48%	10,92%	11,36%	11,45%	11,97%	12,19%	12,10%
0,4	10,66%	14,50%	15,64%	16,99%	18,17%	19,00%	19,00%	19,27%	18,87%
0,8	21,71%	26,26%	28,00%	29,40%	32,15%	33,94%	34,29%	35,12%	34,60%
1,2	31,72%	37,35%	40,02%	41,85%	45,39%	46,75%	47,79%	48,14%	48,80%
1,6	40,45%	46,83%	49,72%	51,59%	56,31%	57,32%	58,45%	59,68%	59,20%
2	48,45%	54,91%	57,45%	59,81%	64,48%	66,54%	67,89%	68,15%	68,15%
2,5	56,01%	63,30%	66,40%	68,33%	72,87%	74,57%	76,37%	76,71%	77,02%
3	62,91%	69,38%	72,61%	74,62%	78,81%	81,17%	82,44%	82,96%	82,66%
3,5	68,50%	74,92%	77,41%	79,64%	82,44%	84,36%	85,93%	86,63%	86,85%
4	73,18%	79,42%	81,70%	83,36%	85,80%	87,68%	89,21%	89,69%	90,13%
SIGN	60,20%	65,57%	68,85%	69,51%	71,17%	72,56%	73,04%	73,26%	73,57%

7. táblázat ANN + RNN jövőbeli százalékos ár változás eredmények

A tematika, illetve a hatékonyság megszokott az eddigiekhez képest, így ezt nem részletezem. Eredmények a 7. táblázat ANN + RNN jövőbeli százalékos ár változás eredményekban találhatóak. A kérdés továbbá az, hogy melyik módszer a hatékonyabb, az eredmény pontosítása érdekében szükséges-e megtartani ezen hatékonyságnövelő módszert.

A két eredménytábla összehasonlítása azonban nem ad perdöntő választ a kérdésre, ez látható a 8. táblázat RNN árváltozás előrejelzés – RNN pontos ár előrejelzés ANN eredmények összehasonlításaban.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,04%	-0,43%	0,34%	0,09%	-0,13%	0,53%	-0,34%	0,17%	0,35%
0,04	0,04%	-0,35%	0,44%	-0,17%	-0,21%	0,39%	-0,61%	-0,13%	0,00%
0,07	0,09%	-0,52%	-0,13%	-0,75%	0,22%	0,53%	-0,35%	0,43%	-0,31%
0,10	0,00%	-0,83%	-0,39%	-0,48%	0,44%	0,35%	-0,44%	0,52%	-1,05%
0,15	-0,22%	-0,39%	0,61%	-0,26%	-0,08%	-0,13%	-0,26%	0,09%	-0,30%
0,25	-1,18%	-0,78%	-0,22%	-0,26%	-0,48%	-0,17%	-0,18%	0,79%	-0,18%
0,40	-1,57%	-0,05%	0,17%	-0,31%	0,35%	0,43%	-0,09%	0,79%	-0,05%
0,80	-1,53%	-0,17%	0,65%	-1,44%	-0,40%	1,26%	-0,44%	0,69%	0,39%
1,20	-1,66%	-0,18%	0,61%	-0,83%	0,83%	-0,04%	-0,40%	0,13%	0,96%
1,60	-0,22%	-0,44%	0,35%	-0,66%	1,05%	0,18%	-0,09%	1,05%	0,79%
2,00	-0,13%	0,08%	-0,26%	-1,22%	-0,09%	-0,17%	-0,31%	-0,66%	-0,61%
2,50	-0,04%	0,57%	0,48%	-1,00%	-0,48%	-1,36%	-0,56%	-0,75%	-0,22%
3,00	-0,04%	-0,34%	-0,52%	-0,87%	-0,79%	-0,44%	-0,61%	-0,44%	-0,78%
3,50	-0,05%	-0,27%	-0,96%	-0,70%	-1,22%	-1,22%	-0,75%	-0,35%	-0,57%
4,00	0,27%	-0,27%	-0,74%	-0,91%	-1,57%	-1,70%	-1,00%	-0,74%	-0,43%
SIGN	-0,18%	-0,44%	0,39%	-0,35%	0,62%	0,69%	-0,09%	0,22%	0,44%
	-0,40%	-0,30%	0,05%	-0,63%	-0,12%	-0,05%	-0,41%	0,11%	-0,10%

8. táblázat RNN árváltozás előrejelzés – RNN pontos ár előrejelzés ANN eredmények összehasonlítása

Az eredmények vizsgálata során bizonyos esetekben és indikátorszám alkalmazásánál hol egyik, hol másik verzió a hatékonyabb. Érdemi különbség nincs. Illetve a teljes vizsgálat minden eredményét összehasonlítva a pontos ár előrejelzése RNN esetében átlagosan 38,53%-os hatékonysággal működött, míg az árváltozás előrejelzés esetében 38,32%-kal, ami elhanyagolható különbségnek tekinthető.

Valamint, ami leszűkíti az értelmezési tartományt, az a SIGN, vagyis a következő napi árfolyam emelkedés vagy csökkenés előrejelzése, valamivel jobb abban az esetben, amikor az RNN a következő napi árfolyamváltozás mértékét becsülte meg. Viszont az eredmények itt is hajszálnyival eltérőek csak, érdemi különbség nem tapasztalható.

4.2.3 Értékelés

A hatékonyság bizonyos esetekben javult, bizonyos esetekben viszont nem.

Jelenleg nem tudjuk kimondani, hogy az idősoros elemzés növeli vagy csökkenti a hatékonyságot. Egyelőre azt tudom megfogalmazni, hogy érdemi változást sem pozitív, sem negatív irányba nem mutat.

Mindamellett fontos megjegyezni, hogy jelenleg a tiszta árakból igyekeztem előrejelezni, semmiféle tőzsdei elemzést nem vizsgáltunk, mint a korábban kifejtett MACD elemzés. Kérdés, hogy ez adott esetben növeli a hatékonyságot vagy érdemben nem tud módosítani.

4.3 ANN – Mesterséges Neurális Háló – Összefoglalt cikkek

Az első verzióban a teljes cikkben kerestem, illetve vizsgáltam a szócikkeket. Második beállításként összefoglaltam a cikkeket és azokon belül vizsgálom a kulcsszavakat. Vizsgálva azt, hogy a tömbösítés, lényegkiemelés növeli-e a hatékonyságot. Így koncentrálódik a mondanivaló, vélhetően csökken a szókincs, ami növeli a releváns szavak előfordulását, így a teljes elemzést.

4.3.1 Felépítés

A spaCy egy ingyenes és nyílt forráskódú könyvtár a Natural Language Processing (NLP) számára Pythonban, számos beépített képességgel. Egyre népszerűbb az adatok NLP-ben történő feldolgozására és elemzésére. A strukturálatlan szöveges adatokat nagy léptékben állítják elő, ezért fontos a strukturálatlan adatok feldolgozása. Ehhez az adatokat számítógépek számára érthető formátumban kell ábrázolni. Az NLP segíthet ebben.

Jelen eseten a spaCy-t arra használtam, hogy a hosszú cikkeket 1-2, maximum 3 mondatban összefoglalja. Ezáltal az információ tömbösítve van, melytől azt vártam, hogy az érdemi mondanivaló koncentráltsága miatt hatékonyabban tud lefutni a Neurális Háló. A kódsor a M7. spaCy cikk összefoglaló python kód mellékletben található.

A teljes Wall Street Journal adatbázison, vagyis a 330.435 darab cikken picivel több, mint 55 órán keresztül dolgozott folyamatosan a program. Ez gyorsnak mondható, egy cikk elemzését és összefoglalását ~1,5 másodperc alatt végezte el.

A program 3144 esetben nem tudott lefutni és nem tudta összefoglalni a cikkeket. Ez a teljes adatállomány 0,95% százaléka, amely elvesztése nem kedvező, de elhanyagolható, így érdemben a metóduson nem változtat. Ezen adatvesztésről a továbbiakban nem beszélünk.

4.3.2 Eredmények

Az eredmények vizsgálatánál az eddigiektől eltérően feltűnően kis indikátorszámnál már eléri a közel végleges hatékonyságot. Habár már 1 indikátorszám, vagyis, ha említésre kerül bármi, ami az olajjal kapcsolatos a cikkben, hatékonyan működik a neurális háló. A hatékonyságot jelen esetben a többi eredmény vonatkozásában értem.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,52%	0,48%	0,74%	0,83%	0,74%	0,65%	0,65%	0,61%	0,61%
0,04	1,96%	2,01%	2,40%	2,27%	2,31%	1,53%	1,48%	1,48%	1,48%
0,07	3,45%	3,49%	3,45%	3,45%	3,45%	3,32%	3,32%	3,32%	3,32%
0,1	4,15%	4,54%	5,24%	5,11%	5,15%	5,19%	5,15%	5,11%	5,11%
0,15	6,42%	7,03%	7,42%	7,38%	7,42%	7,38%	7,42%	7,38%	7,38%
0,25	10,39%	11,92%	12,05%	11,87%	11,87%	11,74%	11,83%	11,83%	11,83%
0,4	16,02%	18,29%	18,51%	18,38%	18,42%	18,38%	18,42%	18,42%	18,42%
0,8	30,25%	33,39%	34,83%	34,83%	34,96%	35,01%	35,01%	35,01%	35,01%
1,2	42,12%	45,79%	47,27%	47,40%	47,36%	47,40%	47,45%	47,45%	47,45%
1,6	52,95%	57,66%	59,54%	59,62%	59,62%	59,76%	59,71%	59,71%	59,71%
2	61,33%	66,70%	68,49%	68,62%	68,62%	68,75%	68,66%	68,70%	68,70%
2,5	70,84%	75,56%	77,48%	77,65%	77,70%	77,87%	77,83%	77,83%	77,83%
3	77,00%	81,62%	83,68%	84,02%	83,98%	83,94%	83,94%	83,94%	83,94%
3,5	81,45%	85,81%	87,78%	88,08%	88,08%	88,08%	88,08%	88,08%	88,08%
4	84,98%	89,13%	90,75%	91,01%	90,96%	90,96%	90,96%	90,96%	90,96%
SIGN	68,44%	70,32%	71,58%	71,76%	71,80%	71,85%	71,85%	71,85%	71,85%

9. táblázat ANN – összefoglalt cikkek elemzés

A 9. táblázat ANN – összefoglalt cikkek elemzésban látható, hogy az összefoglalt cikkek sokkal érzékenyebbé teszik a Neurális Hálót. Így növelve a hatékonyságot. Az eredmények pontosabb kiértékeléséhez szükséges összehasonlítani az eddigi eredményekkel. Hiszen, hiába éri el hamar a maximális pontosságot, ha ez jóval alacsonyabb, mint a korábbi eredmények, akkor a modell használhatatlan.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	-0,14%	-0,09%	0,00%	0,31%	0,26%	-0,27%	-0,31%	-0,39%	-0,22%
0,04	0,82%	0,83%	0,91%	0,87%	0,87%	0,00%	-0,05%	-0,22%	-0,01%
0,07	1,53%	1,31%	0,70%	0,39%	0,13%	0,00%	-0,04%	-0,31%	0,00%
0,1	1,35%	0,83%	1,09%	0,96%	0,34%	-0,10%	-0,22%	-0,05%	-0,22%
0,15	2,79%	1,74%	1,22%	1,26%	0,60%	0,35%	0,39%	-0,05%	0,30%
0,25	4,06%	2,61%	2,00%	2,00%	0,90%	0,03%	-0,05%	-0,36%	-0,18%
0,4	5,40%	4,35%	2,52%	2,39%	0,81%	-0,10%	-0,28%	-0,71%	-0,50%
0,8	8,58%	7,27%	6,04%	4,86%	2,33%	1,28%	0,63%	0,15%	0,45%
1,2	10,84%	9,18%	6,68%	5,33%	1,75%	-0,22%	-0,87%	-1,26%	-1,22%
1,6	14,02%	11,40%	8,86%	7,20%	3,48%	0,96%	0,03%	-0,49%	-0,23%
2	15,15%	12,09%	9,42%	7,63%	3,00%	0,55%	-0,19%	-0,59%	-0,59%
2,5	15,93%	13,04%	10,03%	8,45%	3,08%	0,85%	0,72%	0,02%	-0,02%
3	15,49%	13,12%	10,24%	8,57%	3,60%	0,80%	1,02%	0,37%	0,32%
3,5	14,83%	12,02%	9,32%	8,31%	3,50%	1,45%	1,36%	0,75%	0,79%
4	13,33%	11,24%	8,53%	6,74%	2,97%	0,92%	1,01%	0,44%	0,40%
SIGN	8,28%	4,53%	2,55%	2,43%	-0,37%	-0,80%	-1,37%	-1,63%	-1,72%

10. táblázat ANN összefoglalt cikkek – ANN eredmény különbségek

Magas indikátorszám alkalmazásánál közel ugyanolyan hatékonysággal dolgozott a rendszer, kis indikátorszám alkalmazásánál viszont az eredmények sokkalta pontosabbak. (10. táblázat ANN összefoglalt cikkek – ANN eredmény különbségek)

4.3.3 Értékelés

Az eredmények vizsgálata alapján egyértelműen látszik, hogy a modell érzékenységét nagy mértékben növeli az, hogy nem a teljes cikket, hanem csak egy kivonatot elemzünk. Vagyis, így a cikkek „koncentrációja” növeli a kulcsszókutatási hatékonyságot.

Deklarálhatjuk, hogy a folytatólagos vizsgálatok során szükséges a cikkek kivonatát tovább vizsgálni a későbbi fejlesztési és hatékonyságnövelési tesztelések során.

4.4 ANN – WSJ hangulatelemzés

A kulcsszavak elemzésének kiegészítéseként a cikkek hangulati, avagy sentiment elemzését is elvégeztem az Nltk VADER lexicon segítségével. A hangulat elemzés olyan eljárás, amelynek segítségével az írásos szövegek, általában médiában megjelenő cikkek, hozzászólások vagy tweetek, lelkiállapotát, érzelmeit és véleményét tudjuk megállapítani. Az Nltk (Natural Language Toolkit) egy Python nyelven íródott könyvtár, amely számos eszközt és angol nyelvű korpuszt tartalmaz a számítógépes nyelvészeti kutatásokhoz. Az Nltk rendelkezik olyan eszközökkel, mint a tokenizálás, lemmatizálás és a stop word-ök eltávolítása, amelyek segítségével tisztíthatjuk az adatokat, hogy könnyebben kezelhetők legyenek az elemzések során. A VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon egy kifejezetten közösségi média tartalmak hangulati elemzésére kifejlesztett eszköz. A hangulat elemzés során a szövegekben található kifejezések pozitív, negatív vagy semleges érzelmeket fejeznek-e ki. A VADER lexicon

előnye abban rejlik, hogy képes kezelni az internetes nyelvezet és szleng sajátosságait, így például az iróniát vagy az emotikonokat is. Jelen esetben nem közösségi média szövegekről beszélhetünk, de nem is tudományos művekről. A WSJ cikkei közösségi felhasználásúak, könnyen értelmezhetőek, mindennapi olvasásra javasoltak, így közel állnak a közösségi média nyelvezetéhez. Ezen elemzés eredményeként képesek voltunk meghatározni a cikkekben kifejezett pozitív vagy negatív hozzáállást az olajárral és annak változásaival kapcsolatban. Az összes kulcsszó és sentiment elemzés kombinációjával mélyreható képet kaptunk arról, hogy milyen kontextusban és milyen érzelmekkel tárgyalják az olajárakat a Wall Street Journalban.

4.4.1 Felépítés

A bevezetésben leírtak szerint importáltam az Nltk-t és a szükséges VADER lexicon-t, majd lefuttattam. A pontos metódus megtekinthető a M8. Nltk VADER lexicon szentiment analízis python kód mellékletben.

Alapesetben a pozitív, negatív és semleges értékekhez 3 különböző érték jön, majd ebből számol egy végleges szentiment értéket.

A későbbi NN-ben történő felhasználás miatt a „pozitív”, „negatív” vagy „semleges” jelzőket nemszövegesen jelöltem ki, hanem a későbbi felhasználás miatt ezek helyett +1, -1 és 0 értékeket alkalmaztam. Így a Neurális Háló képes feldolgozni, illetve pontosabban a Neurális Háló felépítése során ezeket szükséges lett volna átalakítani és vélhetően ugyanezeket az értékeket kapta volna.

4.4.2 Eredmények

Az eredményeket a korábbiakhoz hasonlóan 20 indikátorszámig néztem, valamint a hatékonyságot változó megengedett hiba mellett vizsgáltam.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,65%	0,70%	0,79%	0,83%	0,74%	0,65%	0,65%	0,61%	0,61%
0,04	1,88%	2,01%	2,40%	2,27%	2,31%	1,53%	1,48%	1,48%	1,48%
0,07	3,19%	3,67%	3,49%	3,49%	3,45%	3,36%	3,32%	3,32%	3,32%
0,1	4,36%	4,67%	5,33%	5,19%	5,19%	5,19%	5,15%	5,11%	5,11%
0,15	6,42%	7,11%	7,42%	7,38%	7,42%	7,38%	7,42%	7,38%	7,38%
0,25	10,74%	11,70%	11,96%	11,87%	11,87%	11,74%	11,83%	11,83%	11,83%
0,4	16,72%	18,20%	18,33%	18,38%	18,42%	18,42%	18,42%	18,42%	18,42%
0,8	30,86%	33,83%	34,66%	34,83%	34,96%	35,01%	35,01%	35,01%	35,01%
1,2	42,91%	46,05%	47,27%	47,36%	47,36%	47,40%	47,45%	47,45%	47,45%
1,6	53,95%	58,01%	59,45%	59,58%	59,62%	59,76%	59,71%	59,76%	59,76%
2	61,94%	66,65%	68,35%	68,66%	68,62%	68,75%	68,66%	68,70%	68,70%
2,5	71,06%	75,77%	77,30%	77,70%	77,65%	77,87%	77,83%	77,83%	77,83%
3	77,87%	81,84%	83,63%	84,02%	83,98%	83,94%	83,94%	83,94%	83,94%
3,5	81,67%	85,95%	87,69%	88,13%	88,08%	88,08%	88,08%	88,08%	88,08%
4	84,98%	88,96%	90,75%	91,01%	90,96%	90,96%	90,96%	90,96%	90,96%
SIGN	67,74%	70,36%	71,63%	71,72%	71,80%	71,85%	71,85%	71,85%	71,85%

11. táblázat ANN eredmények hangulatelemzéssel bővítve

A 11. táblázat ANN eredmények hangulatelemzéssel bővítve mutatja, hogy a korábbi eredményekhez képest jól látható, hogy már 2, inkább 3 indikátornál eléri a közel maximális értéket. Az indikátorszám további emelésével az eredmények nem növekednek radikálisan. Ez örömteli, mivel ahogy korábban is, az indikátorszám csökkentésével drasztikusan csökken az elemzett cikkek darabszáma. Minél több cikk, annál komplexebb háló, annál pontosabb eredmények, vagyis a kevesebb indikátorszám elméletileg hasznosabb.

Az eredmények pontos értékeléséhez fontos megvizsgálnunk a korábbi, sentiment elemzés nélküli, de összefoglalt cikkek eredményeivel.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,13%	0,22%	0,05%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
0,04	-0,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
0,07	-0,26%	0,18%	0,04%	0,04%	0,00%	0,04%	0,00%	0,00%	0,00%
0,1	0,21%	0,13%	0,09%	0,08%	0,04%	0,00%	0,00%	0,00%	0,00%
0,15	0,00%	0,08%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
0,25	0,35%	-0,22%	-0,09%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
0,4	0,70%	-0,09%	-0,18%	0,00%	0,00%	0,04%	0,00%	0,00%	0,00%
0,8	0,61%	0,44%	-0,17%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1,2	0,79%	0,26%	0,00%	-0,04%	0,00%	0,00%	0,00%	0,00%	0,00%
1,6	1,00%	0,35%	-0,09%	-0,04%	0,00%	0,00%	0,00%	0,05%	0,05%
2	0,61%	-0,05%	-0,14%	0,04%	0,00%	0,00%	0,00%	0,00%	0,00%
2,5	0,22%	0,21%	-0,18%	0,05%	-0,05%	0,00%	0,00%	0,00%	0,00%
3	0,87%	0,22%	-0,05%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3,5	0,22%	0,14%	-0,09%	0,05%	0,00%	0,00%	0,00%	0,00%	0,00%
4	0,00%	-0,17%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
SIGN	-0,70%	0,04%	0,05%	-0,04%	0,00%	0,00%	0,00%	0,00%	0,00%

12. táblázat ANN hangulatelemzéssel bővített eredmények és ANN összefoglalt cikkek eredmények összehasonlítása

Kis indikátorszám alkalmazásával jelenlegi módszer valamivel hatékonyabb, mint a korábban végzett. Az indikátorszám növelésével viszont a különbség eltűnik és a hatékonysági eredmények teljesen megegyeznek. (12. táblázat ANN hangulatelemzéssel bővített eredmények és ANN összefoglalt cikkek eredmények összehasonlítása)

4.4.3 Értékelés

A hangulatelemzéssel való bővítés egyértelműen hatékonyság növekedést eredményezett. Habár az eredmények összességében csak kismértékű növekedést eredményeztek, de egyértelműen hatékonyabbnak mondhatók. Illetve néhány esetben minimális az eredménycsökkenés. Érdekes megfigyelni, hogy 1 indikátorszám esetében jelentősen jobbak az eredmények, valamint 2 indikátorszám esetében is. Három indikátor esetében a korábbi elemzés valamivel hatékonyabb, majd mindkét módszer pontosan ugyanazokat az eredményeket preferálja az indikátorszám növelésével. Viszont, mivel a kevesebb indikátorszám, vagyis a több elemzett cikk eseménysorozatot jobban preferáljuk, így a kevesebb indikátorszám esetében lévő hatékonyságnövekedést mindig hasznosabbnak tekintjük.

5. Neurális hálók kiegészítése mozgóátlag-módszerrel

Az eddigi neurális hálót bővítettem a tőzsdei árfolyamelemzésben használt MACD módszerrel. Ezt korábban már részleteztem. Röviden az MACD egy 12 és 26 napos mozgóátlag változásainak, keresztezéseinek figyelése, ami jó eséllyel jelzi a jelenlegi trendet, tehát az emelkedő vagy csökkenő trendet. Ezt a 4.1.2 fejezetben részletesen kibontottam, így jelenleg nem teszem meg újra.

Az MACD alapján történő befektetés kedvelt és gyakori a tőzsdei kereskedésben. Önmagában nem feltétlenül elég, illetve adhat fals jelzéseket.

Jelen esetben elegendő információt ad, pontosabban egy információs bemeneti adathalmaz egyik eleme, így releváns adathalmaz részének feltételezve beépítésre kerül az ANN-be.

5.1 Felépítés

Az olajár adattáblát bővítettem, a már korábban említett 12 és 26 napos mozgóátlaggal, valamint az MACD index-szel, ami a kettő különbsége, így pozitív vagy negatív trendet feltételez, valamint a signal-lal, ami az MACD index 9 napos mozgóátlaga. A tőzsdei előrejelzésben gyakran használják végső elemzésre a signal értéket.

A beillesztés hasonló a korábbiakhoz, az input adatok számát növeltem, majd a teljes metódust lefuttattam a korábbiakhoz hasonlóan.

5.2 Eredmények

Az eredmények vizsgálatánál a hatékonysági motívumok hasonlóak az eddigi esetekhez. Az indikátorok növelésével gyorsan növekszik, de relatíve kis indikátorszámnál eléri a maximális hatékonyságot és csak akörül ingadozik.

Érdekes módon tapasztaltam azt is, hogy bizonyos esetekben nagyon visszaesett, gondolok itt például a I:4 E:0,02 esetre, ahol eggyel kisebb és nagyobb indikátor esetében is 1% körüli az eredmény, a tárgyalt indikátorszám esetében viszont alig éri el a hatékonyság 0,5%-ot. A hibahatárok növelésével viszont ez a gap eltűnik és a várt eredményeket hozza.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,79%	0,87%	0,96%	0,48%	0,92%	0,65%	1,13%	0,52%	0,52%
0,04	1,31%	1,75%	1,88%	1,40%	1,83%	1,62%	1,92%	1,48%	1,48%
0,07	2,23%	3,14%	3,32%	3,32%	3,32%	3,14%	2,71%	2,88%	2,88%
0,1	3,49%	4,28%	4,36%	4,80%	4,50%	4,10%	4,58%	4,06%	4,06%
0,15	5,54%	5,94%	6,42%	7,03%	6,55%	6,29%	6,24%	5,76%	5,76%
0,25	9,56%	11,22%	11,31%	11,52%	11,22%	10,39%	11,13%	11,09%	11,09%
0,4	16,19%	17,85%	17,85%	18,33%	18,46%	17,59%	17,90%	17,81%	17,81%
0,8	33,39%	34,31%	35,92%	36,67%	35,05%	33,96%	35,22%	35,01%	35,01%
1,2	46,53%	47,45%	48,36%	50,50%	50,02%	49,06%	49,24%	49,80%	49,80%
1,6	57,62%	59,89%	60,15%	61,02%	61,41%	60,98%	61,59%	61,20%	61,20%
2	66,87%	69,18%	69,93%	70,01%	70,76%	70,54%	70,54%	71,06%	71,06%
2,5	74,73%	76,69%	78,52%	78,52%	78,57%	79,00%	78,87%	79,18%	79,18%
3	81,49%	82,67%	84,20%	84,59%	85,16%	84,72%	84,81%	84,55%	84,55%
3,5	86,69%	87,47%	89,09%	89,13%	89,66%	89,57%	89,61%	90,22%	90,22%
4	89,74%	90,44%	92,01%	91,84%	92,58%	92,93%	93,06%	92,58%	92,58%
SIGN	71,15%	71,98%	73,24%	72,94%	73,02%	73,29%	72,94%	73,11%	73,11%

13. táblázat ANN eredmények bővítve hangulatelemzéssel és MACD mutatókkal

A 13. táblázat ANN eredmények bővítve hangulatelemzéssel és MACD mutatókkalban látható, hogy a korábban, „csak” hangulatelemzéssel bővített neurális hálóval, vagyis az eddig legjobb hatékonysággal rendelkező elemzéshez képest hatékonyabb a jelenlegi módszer. Illetve fogalmazzunk úgy, hogy feltételesen hatékonyabb. Láthatóan kis hibatartománynál valamivel hatékonytalanabb, viszont nagyobb hibahatárokat megengedve már számottevően hatékonyabb. Kérdéses, hogy mennyire szigorúan vesszük a hibahatárokat, 0,8 fölött már egyértelműen a jelenlegi módszer jobb. Így a nagyobb kilengéseket valószínűleg jobban mutatja.

Mindazonáltal a Sign eredmények, vagyis az előjel meghatározásnál minden esetben jobb a korábbi verziónál, így jelen modellt összeségében eredményesebbnek tekintem. Ezt bizonyítja a 14. táblázat ANN+Hangulat+MACD összehaonlítása a korábbi, csak ANN+Hangulat eredményekkel.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS NAP: t, IND: 1	SIKERES BECSLÉS NAP: t, IND: 2	SIKERES BECSLÉS NAP: t, IND: 3	SIKERES BECSLÉS NAP: t, IND: 4	SIKERES BECSLÉS NAP: t, IND: 7	SIKERES BECSLÉS NAP: t, IND: 10	SIKERES BECSLÉS NAP: t, IND: 13	SIKERES BECSLÉS NAP: t, IND: 16	SIKERES BECSLÉS NAP: t, IND: 20
0,02	0,14%	0,17%	0,17%	-0,35%	0,18%	0,00%	0,48%	-0,09%	-0,09%
0,04	-0,57%	-0,26%	-0,52%	-0,87%	-0,48%	0,09%	0,44%	0,00%	0,00%
0,07	-0,96%	-0,53%	-0,17%	-0,17%	-0,13%	-0,22%	-0,61%	-0,44%	-0,44%
0,1	-0,87%	-0,39%	-0,97%	-0,39%	-0,69%	-1,09%	-0,57%	-1,05%	-1,05%
0,15	-0,88%	-1,17%	-1,00%	-0,35%	-0,87%	-1,09%	-1,18%	-1,62%	-1,62%
0,25	-1,18%	-0,48%	-0,65%	-0,35%	-0,65%	-1,35%	-0,70%	-0,74%	-0,74%
0,4	-0,53%	-0,35%	-0,48%	-0,05%	0,04%	-0,83%	-0,52%	-0,61%	-0,61%
0,8	2,53%	0,48%	1,26%	1,84%	0,09%	-1,05%	0,21%	0,00%	0,00%
1,2	3,62%	1,40%	1,09%	3,14%	2,66%	1,66%	1,79%	2,35%	2,35%
1,6	3,67%	1,88%	0,70%	1,44%	1,79%	1,22%	1,88%	1,44%	1,44%
2	4,93%	2,53%	1,58%	1,35%	2,14%	1,79%	1,88%	2,36%	2,36%
2,5	3,67%	0,92%	1,22%	0,82%	0,92%	1,13%	1,04%	1,35%	1,35%
3	3,62%	0,83%	0,57%	0,57%	1,18%	0,78%	0,87%	0,61%	0,61%
3,5	5,02%	1,52%	1,40%	1,00%	1,58%	1,49%	1,53%	2,14%	2,14%
4	4,76%	1,48%	1,26%	0,83%	1,62%	1,97%	2,10%	1,62%	1,62%
SIGN	3,41%	1,62%	1,61%	1,22%	1,22%	1,44%	1,09%	1,26%	1,26%

14. táblázat ANN+Hangulat+MACD összehaonlítása a korábbi, csak ANN+Hangulat eredményekkel

5.3 Értékelés

A Neurális Hálót folyamatosan bővítettem, minden egyes plusz bevitt releváns adathalmaz kisebb-nagyobb mértékben növelte a hatékonyságot.

A Sign érték közel 75%-os hatékonysággal működik, ami szerint 4 napból 3 esetben meg tudom mondani, hogy a következő napi árfolyam emelkedni vagy csökkenni fog. Ami elfogadható eredménynek számít.

A leghatékonyabb eredmények összeségében 4 indikátorszámnál lelhetők fel. Ez hatékonynak tűnik, hiszen a cikkek relatíve kevés részét fogja invalidnak nézni. Egészen pontosan kevés olyan cikket nem fog figyelembe venni, amit kellett volna, tehát a releváns cikkek döntő hányadát feldolgozza.

Ha a pontos eredményeket abban az esetben vizsgáljuk, ha a jósolt és valós eredmény között maximum 0,8% eltérés van, akkor így 36,67%-os pontossággal dolgozik, ami azt jelenti, hogy ez esetek picivel több, mint egy harmadában képes volt megjósolni az eredményt. Ez az arány kevésnek tekinthető egy befektetési döntéshozatali algoritmus alkotására. Az esetek felénél jó eredményt mutatott, ha 1,2%-os hibahatárt engedélyeztünk. Szintén megítélés kérdése, hogy ezt mennyire értékeljük jónak.

6. Kulcsszókutatás és árfolyamon elemzés összehasonlítása

Mindenképpen szükséges vizsgálni azt, hogy milyen hatékonysággal működött volna a neurális háló abban az esetben, ha csak árfolyamalapú adatokkal dolgozunk, így kideríthető, hogy jelent-e, illetve mekkora pluszt jelent a kulcsszókutatáson alapuló neurális háló. Pontosabban, érdemes-e ANN-t alkalmazni, vagy ugyanolyan vagy még jobb hatékonyság érhető el, ha csak matematikai modellt alkalmazok.

Így a kulcsszókutatás validitását, illetve erősségét vizsgálva lefutattam a neurális hálót a kulcsszókutatás nélkül, vagyis csak az olajárfolyam változásaiból kinyerhető számszerű adatokat vizsgáltam.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	SIKERES BECSLÉS
0,02	0,91%
0,04	1,52%
0,07	2,86%
0,1	3,65%
0,15	6,27%
0,25	10,28%
0,4	16,67%
0,8	34,06%
1,2	48,30%
1,6	61,25%
2	70,13%
2,5	77,37%
3	84,49%
3,5	88,69%
4	91,91%
SIGN	73,97%

15. táblázat Árfolyam elemzésen alapuló neurális háló előrejelzés eredményei

A 15. táblázat Árfolyam elemzésen alapuló neurális háló előrejelzés eredményei alapján, első ránézésre hasonló eredmények jöttek ki, mint az eddigi kulcsszókutatáson alapuló ANN alapján. Így a pontosabb értékelés mélyebb vizsgálatot igényel.

ELFOGADOTT MAXIMÁLIS ABSZOLÚT ELTÉRÉS	OIL MATH	SIKERES BECSLÉS NAP: t, IND: 4	ANN ind: 4 - OIL MATH különbség		ANN Best Value	ANN Best Value - OIL MATH különbség	
0,02	0,91%	0,48%	-0,43%	-47,25%	1,13%	0,22%	24,18%
0,04	1,52%	1,40%	-0,12%	-7,89%	1,92%	0,40%	26,32%
0,07	2,86%	3,32%	0,46%	16,08%	3,32%	0,46%	16,08%
0,1	3,65%	4,80%	1,15%	31,51%	4,80%	1,15%	31,51%
0,15	6,27%	7,03%	0,76%	12,12%	7,03%	0,76%	12,12%
0,25	10,28%	11,52%	1,24%	12,06%	11,52%	1,24%	12,06%
0,4	16,67%	18,33%	1,66%	9,96%	18,46%	1,79%	10,74%
0,8	34,06%	36,67%	2,61%	7,66%	36,67%	2,61%	7,66%
1,2	48,30%	50,50%	2,20%	4,55%	50,50%	2,20%	4,55%
1,6	61,25%	61,02%	-0,23%	-0,38%	61,59%	0,34%	0,56%
2	70,13%	70,01%	-0,12%	-0,17%	71,06%	0,93%	1,33%
2,5	77,37%	78,52%	1,15%	1,49%	79,18%	1,81%	2,34%
3	84,49%	84,59%	0,10%	0,12%	85,16%	0,67%	0,79%
3,5	88,69%	89,13%	0,44%	0,50%	90,22%	1,53%	1,73%
4	91,91%	91,84%	-0,07%	-0,08%	93,06%	1,15%	1,25%
SIGN	73,97%	72,94%	-1,03%	-1,39%	73,29%	-0,68%	-0,92%

16. táblázat ANN leghatékonyabb eredményeinek és csak olajárfolyamon alapuló

A 16. táblázat jól mutatja az eltéréseket, illetve a valódi különbségeket. A vizsgált hatékonyság az eddig megszokottak szerint az elfogadott maximális eltérés, vagyis az előrejelzett eredmények hány százaléka van a tényleges eredmények esetében a meghatározott hibahatáron belül. Az OIL MATH oszlop a csak olajárfolyam napi zárásainak neurális hálóval történő idősoros elemzéséből előrevetített eredmények pontossága. A korábbi fejezetekben tárgyaltak szerint a 4 indikátorszámos elemzést vettem legérvényesebbnek. Ezen túl csak kismértékben nő a hatékonyság, illetve ennyi indikátor alkalmazásánál nem esik ki túl sok cikk. Túl magas indikátorszám alkalmazásánál fennáll a veszély, hogy túl sok releváns cikket nem veszünk figyelembe. Ezt követő két oszlop a számszerű eltérés a kulcsszókutatáson alapuló ANN javára, illetve az ANN bázisához viszonyított eltérés. Ezt követően minden kulcsszókutatásos ANN esetében kiválasztottam soronként a legjobb eredményt, majd a korábbi módszerrel összehasonlítottam az olajáron alapuló Neurális Háló eredményeivel.

7. Egyéb ANN változatok

A neurális háló hatékonyságának növelése érdekében teszteltem, hogy mennyiben javul az eredmény, ha a rétegszámot növelem, vagyis még inkább a mélytanulás irányába viszem el az elemzési folyamatot. A korábbi kutatások során két köztes réteg volt. Ezeket bővítettem 3, illetve 4 köztes rétegre.

Az eredmény viszont nem hozta a várt eredményt. Szinte ugyanazok az eredmények jöttek, minimális különbséggel, fogalmazhatunk úgy, hogy számottevő változás nélkül. Viszont a futtatási idő drasztikusan megnövekedett. Így ezt a táblázatot és eredményeket nem publikálom.

8. Hat módszer az idősorok adatai közötti szinkron számszerűsítésére

Az idősorok szinkronitásának elemzése napjainkban egyre inkább előtérbe kerül a tudományos kutatásokban, mivel számos interdiszciplináris területen alkalmazható, mint például a gazdaságtan, a meteorológia, a biológia vagy akár a társadalomtudományok (Brockwell & Davis, 2016). Ebben a rövid összefoglalóban négy fő módszert mutatok be, amelyeket az idősorok szinkronitásának elemzésére használnak: a Pearson korrelációt, az időkésleltetett keresztkorrelációt, a dinamikus idővetemítést, az azonnali fázisszinkront, a Wilmott-féle egyezési indexet és a R^2 mutatót.

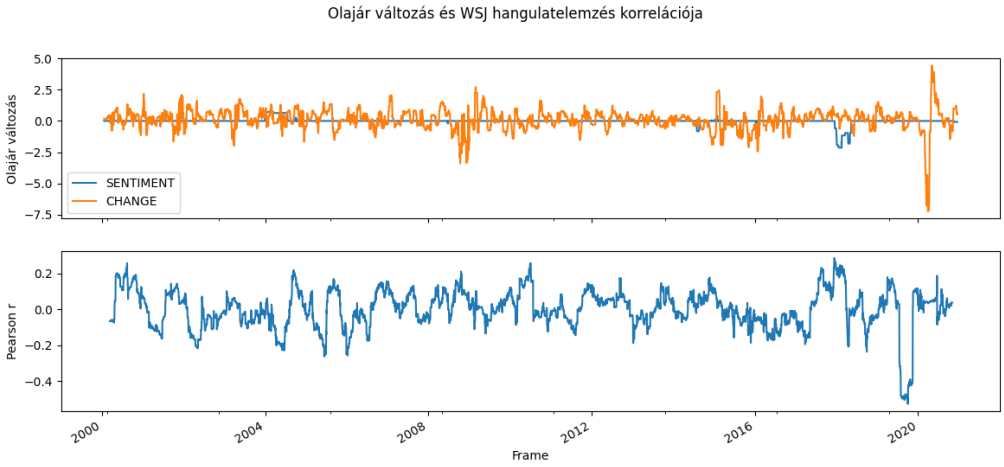
8.1 Pearson korreláció

A Pearson-korreláció azt méri, hogy két folytonos jel hogyan változik az időben, és a lineáris kapcsolatot -1 (negatív korreláció) és 0 (nem korrelált) és 1 (tökéletesen korrelált) közötti számként jelzi. Intuitív, könnyen érthető és könnyen értelmezhető. Két dologra kell vigyázni a Pearson-korreláció használatakor: 1) a kiugró értékek torzíthatják a korrelációs becslés eredményeit, és 2) feltételezi, hogy az adatok homoszkedasztikusak, így az adatok szórása homogén az adattartományban. Általában a korreláció a globális szinkron pillanatképe. Ezért nem ad információt a két jel közötti irányultságról, például arról, hogy melyik jel vezet és melyik következik.

A Pearson korreláció az egyik legáltalánosabb és legnépszerűbb módszer, amelyet az idősorok közötti lineáris kapcsolat osszefüggésének mérésére alkalmaznak (Pearson, 1895). A módszer célja, hogy egy -1 és 1 közötti értéket adjon, ahol az 1 pozitív lineáris kapcsolatot, a -1 negatív lineáris kapcsolatot, míg a 0 pedig semmilyen lineáris kapcsolatot jelent. A Pearson korreláció előnye, hogy könnyen értelmezhető és számítható, ugyanakkor hátránya, hogy csak a lineáris összefüggéseket képes detektálni.

Az idősorok szinkronitásának elemzése során több módszer is alkalmazható, melyek különböző előnyökkel és hátrányokkal rendelkeznek. Az időkésleltetett keresztkorreláció segítségével időbeli késleltetéseket is figyelembe vehetünk, ugyanakkor a jelentős korrelációk értelmezése nehezebb lehet. A dinamikus idővetemítés rugalmasan alkalmazható és nem lineáris összefüggéseket is detektál, de számításigényes lehet. Az azonnali fázisszinkron érzékeny az időbeli fáziskapcsolatokra, de a zajra és a számítások összetettségére is érzékeny lehet. A kutatóknak figyelembe kell venniük ezeket a tulajdonságokat, hogy a legmegfelelőbb módszert válasszák az idősorok szinkronitásának elemzéséhez az adott kutatási kérdések és adathalmazok alapján.

A Pearson korrelációt elvégeztem a WSJ cikkek hangulatelemzésére alapozva. Csak azokat a cikkeket vettem figyelembe, ahol minimum egy indikátorszó előfordult és egyesítettem a napi eredményeket. Az eredményeket a 11. ábra Olajár változás és WSJ Hangulatelemzés Pearson korrelációláthatjuk.



11. ábra Olajár változás és WSJ Hangulatelemzés Pearson korreláció

Az r értéke a Pearson korrelációs együttható, amely azt mutatja, milyen erős a lineáris összefüggés két változó között. Az r értéke -1 és 1 között változik, ahol -1 negatív tökéletes korrelációt, 0 pedig semmilyen lineáris összefüggést, és 1 pozitív tökéletes korrelációt jelent.

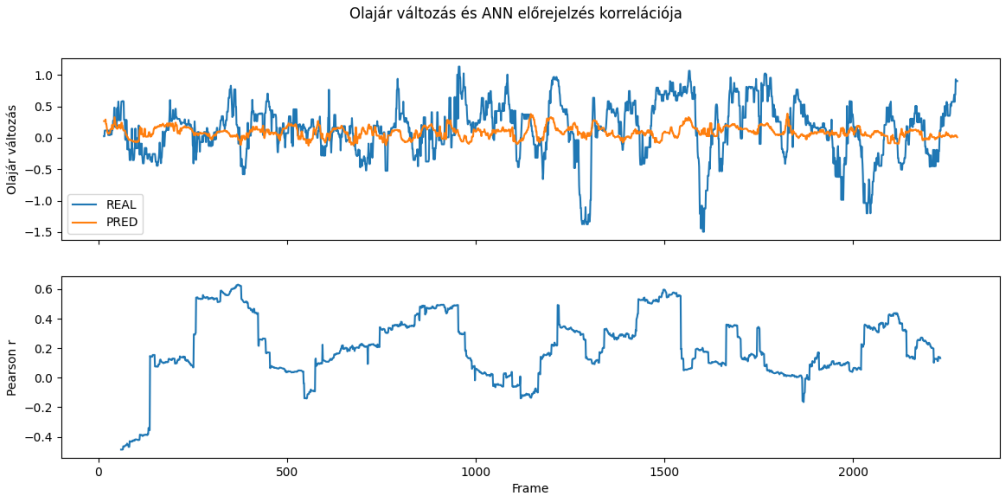
Az 0.0091-es r érték nagyon közel van a 0-hoz, ami azt jelenti, hogy a két változó között gyakorlatilag nincs lineáris összefüggés. Más szóval, a két változó változásai nem követik egymást egyenes vonal mentén. Ez azonban nem jelenti azt, hogy a két változó között nincs semmilyen összefüggés, csak azt, hogy a lineáris összefüggés gyenge vagy nem létezik. Lehet, hogy más típusú összefüggés van a két változó között (pl. nemlineáris), amelyet a Pearson korrelációs együttható nem képes detektálni.

A p-érték a statisztikai hipotézisvizsgálat során használt érték, amely azt mutatja, hogy milyen valószínűséggel fordulhat elő egy adott mintában a megfigyelt eredmény, ha a nullhipotézis igaz. A nullhipotézis általában azt jelenti, hogy nincs összefüggés vagy hatás a vizsgált változók között.

Egy 0.42-es p-érték azt jelenti, hogy a megfigyelt eredmény (például egy korreláció) 42% valószínűséggel fordulhat elő, ha a nullhipotézis igaz. Általában egy előre meghatározott szignifikancia szintet (alfa) használunk a p-érték értelmezéséhez, és ha a p-érték alacsonyabb az alfa értéknél (általában 0.05), akkor elutasítjuk a nullhipotézist, és azt mondjuk, hogy a kapott eredmény szignifikáns.

Ebben az esetben, ha a p-érték 0.42, az azt jelenti, hogy nincs elegendő bizonyíték a nullhipotézis elutasításához, vagyis nem tekinthetjük a kapott eredményt szignifikánsnak. Más szóval, nem állapíthatjuk meg biztosan, hogy a vizsgált változók között van-e valódi összefüggés vagy hatás, mivel a kapott eredmény a véletlen következménye is lehet.

A Pearson korrelációt elvégeztem az ANN legutolsó, vagyis relatíve a legpontosabb előrejelzést biztosító modelljére is. (12. ábra Olajár változás és ANN előrejelzés Pearson korreláció) Ebben az esetben valamivel kisebb időtartamot vizsgáltam, mivel csak a teszthalmaz eredményeit vettem alapul, de így is több, mint 2000 esetet vizsgáltam, ami elegendő esetszámnak definiálható.



12. ábra Olajár változás és ANN előrejelzés Pearson korreláció

A 0,115-ös r érték azt jelenti, hogy a két változó között van egy gyenge pozitív lineáris összefüggés. Más szóval, a két változó közötti kapcsolat nem erős, de amikor az egyik változó értéke növekszik, a másik változó értéke is valamelyest növekedni fog. Azonban ez az összefüggés nem elég erős ahhoz, hogy határozott következtetéseket vonjunk le a két változó közötti kapcsolatról, és lehet, hogy más tényezők is befolyásolják a változók viselkedését.

Fontos megjegyezni, hogy a korreláció nem jelent okozati összefüggést, csak azt mutatja, hogy a két változó hogyan mozog együtt. A kapcsolat mögötti okokat további vizsgálatokkal kell meghatározni.

Ebben az esetben, mivel a p-érték 0.003008, ami alacsonyabb az általában használt 0.05-ös alfa szintnél, elutasíthatjuk a nullhipotézist, és azt mondhatjuk, hogy a kapott eredmény szignifikáns. Ez azt jelenti, hogy a vizsgált változók között valószínűleg van valamilyen összefüggés vagy hatás, ami nem a véletlen következménye. Azonban fontos hangsúlyozni, hogy a szignifikancia csak azt mutatja, hogy valószínűleg van összefüggés a változók között, de nem ad okozati összefüggést. A kapcsolat mögötti okokat további vizsgálatokkal kell meghatározni.

8.2 Időkésleltetett keresztkorreláció (Time Lagged Cross Correlation - TLCC)

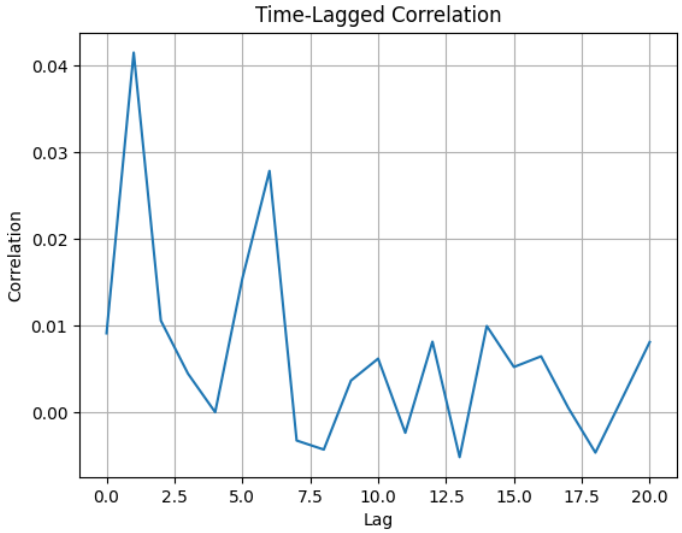
Az időkésleltetett keresztkorreláció (TCC) egy olyan módszer, amely figyelembe veszi az időbeli késleltetéseket az idősorok közötti kapcsolatok elemzésében (Chatfield, 2016). A TCC segítségével megállapítható, hogy az egyik idősor hogyan befolyásolja a másikat bizonyos időbeli késleltetéssel, ami fontos információ lehet például a kauzális kapcsolatok megértésében. A TCC hátránya, hogy nagy számú késleltetés esetén a számítások összetettebbé válnak, és a jelentős korrelációk értelmezése nehezebb lehet.

A TLCC elemzést az előbbihez hasonlóan elvégeztem a WSJ cikk hangulatelemzésére alapozva. Eredményeket a különböző futtatások esetében a 13. ábra TLCC eredmények olajár változás és hangulatelemzés (x=Hangulat), 14. ábra TLCC eredmények olajár változás és hangulatelemzés (x=Olajár)és 15. ábra TLCC eredmények olajár változás és ANN eredmények (x=ANN)mutatja be.

A "max lag" (maximális időbeli eltolás) kifejezés az időbeli eltolású korrelációban azt jelenti, hogy a két idősoros adatsor közötti korreláció kiszámításához mekkora a maximális eltolás az időben.

Az időbeli eltolású korreláció kiszámításakor általában a két idősoros adatsor közötti korrelációt számoljuk különböző eltolásokra, kezdve a 0-tól (nincs eltolás) egészen egy meghatározott maximális eltolásig. Ennek célja, hogy megállapítsuk, van-e bármilyen korreláció a két idősor között különböző eltolásoknál, és ha igen, melyik eltolásnál a legerősebb..

Például, ha napi adatokkal rendelkezünk és a max lag értékét 30-ra állítjuk, a korrelációkat az eredeti idősoron és az 1, 2, 3, ..., 30 nappal eltolódott idősoron számoljuk. Ez segít megérteni, van-e korreláció a két idősor között bármelyik meghatározott késleltetésnél.



13. ábra TLCC eredmények olajár változás és hangulatelemzés (x=Hangulat)

Az x értéknek beállítandó idősor választása attól függ, hogy melyik két idősor közötti korrelációt szeretnénk vizsgálni és melyik idősor eltolását szeretnénk megvizsgálni az időbeli eltolású korreláció során.

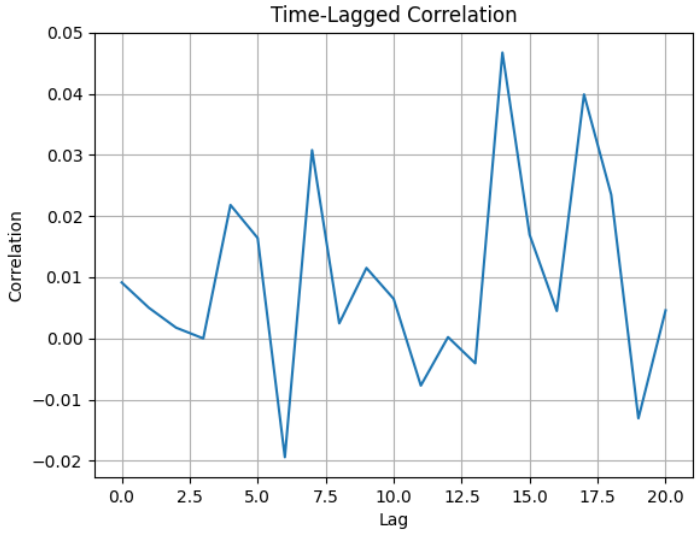
Általában nincs szigorú szabály arra, hogy melyik idősor legyen az x érték, de a legfontosabb szempont az lehet, hogy melyik idősor eseményei vagy hatásai előbb következnek be, mint a másik idősoré. Ha az egyik idősor hatása az időben a másik idősor előtt van, akkor érdemes azt az x értéknek beállítani.

Azonban fontos megjegyezni, hogy az időbeli eltolású korreláció szimmetrikus, tehát az x és y idősorok helyének felcserélése ugyanazt az eredményt adja, csak a kapott korrelációk előjelét változtatja meg. Ez azt jelenti, hogy a két idősor közötti korreláció értéke ugyanaz lesz, függetlenül attól, hogy melyiket választjuk az x értéknek, csak a kapcsolat előjele változik. Ha az egyik idősor eltolásával pozitív korrelációt kapunk, akkor a másik idősor eltolásával negatív korrelációt kapunk, és fordítva.

Az időbeli eltolású korreláció során az értékek a Pearson korrelációs együtthatóra utalnak, amely -1 és 1 között változik. Ha az érték 1-nél a legmagasabb, az azt jelenti, hogy a két idősor között tökéletes pozitív lineáris összefüggés van, amikor nincs időbeli eltolás (lag = 0). Ez azt

sugallja, hogy a két idősor együttesen növekszik vagy csökken, és nagyon hasonló mintázatot mutatnak.

Ha az érték 1-nél 0,04, akkor az azt jelenti, hogy a két idősor között csak gyenge pozitív lineáris összefüggés van, amikor nincs időbeli eltolás (lag = 0). Ez azt jelenti, hogy a két idősor nem igazán mozog együtt, és a közöttük lévő kapcsolat nem erős. Ebben az esetben a két idősor közötti korreláció valószínűleg nem jelentős, és nem lehet biztos következtetéseket levonni a két idősor közötti kapcsolatról. Azonban érdemes megvizsgálni a korrelációkat más időbeli eltolásoknál (nem csak 1-nél), hogy ellenőrizze, van-e erősebb kapcsolat a két idősor között más eltolásoknál.



14. ábra TLCC eredmények olajár változás és hangulatelemzés (x=Olajár)

Ha az időbeli eltolású korreláció eredményei nagyon különbözőek attól függően, hogy melyik idősor van beállítva az x értéknek, akkor ez azt jelezheti, hogy az eltolás iránya fontos a két idősor közötti kapcsolat vizsgálatában.

Ahogy korábban említettem, az időbeli eltolású korreláció szimmetrikus, és az x és y idősorok helyének felcserélése ugyanazt az eredményt adja, csak a kapott korrelációk előjelét változtatja meg. Azonban, ha az eredmények jelentősen eltérnek attól függően, hogy melyik idősor van beállítva az x értéknek, akkor az azt jelentheti, hogy az egyik idősor hatása a másikra aszimmetrikus és az eltolás iránya fontos. Ebben az esetben érdemes további vizsgálatokat végezni és pontosabban megvizsgálni az adatokat, hogy megértsük a különböző eredmények okát és a két idősor közötti kapcsolat jellemzőit.

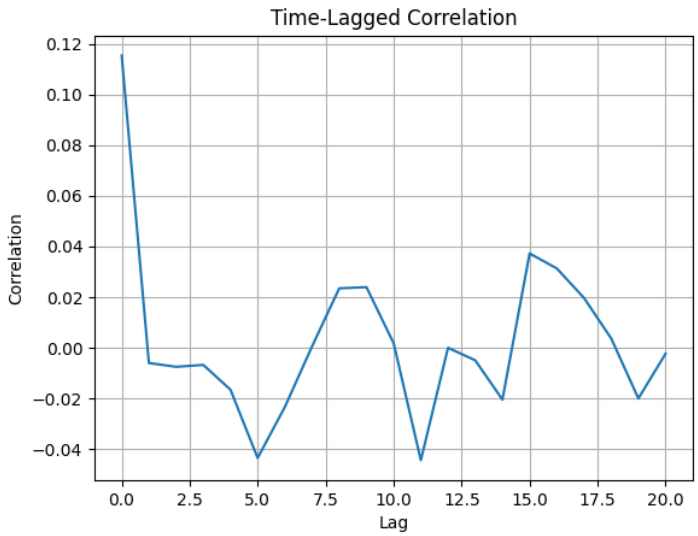
Egy másik lehetőség, hogy a két idősor közötti összefüggés nemlineáris, és az x és y helyének megváltoztatása eltérő eredményeket adhat, attól függően, hogy a két idősor hogyan viszonyul egymáshoz ebben a nemlineáris összefüggésben. Ebben az esetben érdemes megvizsgálni más típusú korrelációkat vagy módszereket, amelyek az ilyen nemlineáris összefüggések értelmezésére alkalmasak.

Ha az időbeli eltolású korreláció során a legmagasabb érték 14-nél található, és ez az érték 0,048, az azt jelenti, hogy a két idősor közötti legnagyobb lineáris összefüggés akkor következik be, amikor az egyik idősor 14 egységnyi eltolással van a másikhoz képest. Ez azt sugallja, hogy a

két idősor közötti kapcsolat valamelyest erősebb, amikor az egyik idősor 14 időegységgel előbb vagy később van, mint a másik.

Azonban fontos megjegyezni, hogy a 0,048-as érték továbbra is gyenge pozitív összefüggésre utal. Ez azt jelenti, hogy a két idősor közötti kapcsolat még mindig nem erős, és lehet, hogy véletlen vagy más tényezők is befolyásolják a változókat.

A TLCC vizsgálatot lefolytattam az ANN által előrejelzett eredményekre is.



15. ábra TLCC eredmények olajár változás és ANN eredmények (x=ANN)

Ha az időbeli eltolású korreláció során a legmagasabb érték 0-nál található, és ez az érték 0,12, akkor ez azt jelenti, hogy a két idősor közötti legnagyobb lineáris összefüggés akkor következik be, amikor nincs időbeli eltolás a két idősor között (lag = 0). Ez azt sugallja, hogy a két idősor egyidejűleg növekszik vagy csökken, és a közöttük lévő kapcsolat erősebb, amikor nincs eltolás.

Azonban a 0,12-es érték továbbra is gyenge pozitív összefüggésre utal. Ez azt jelenti, hogy a két idősor közötti kapcsolat nem erős, de még mindig van némi összefüggés a két idősor között.

8.3 Dinamikus idővetemítés (Dynamic Time Warping - DTW)

A dinamikus idővetemítés (Dynamic Time Warping, DTW) egy olyan idősorok közötti hasonlóság mérő módszer, amely képes az időbeli nyújtások és összehúzódások figyelembevételére (Berndt & Clifford, 1994). A DTW azon alapul, hogy egy költségfüggvény minimalizálásával optimális párosítást talál az idősorok között, ami lehetővé teszi az időfüggő jellemzők összehasonlítását a szinkronitás elemzésében. A DTW előnye, hogy nem lineáris összefüggéseket is képes detektálni, és rugalmasan alkalmazható különböző idősorokra. Ugyanakkor hátránya, hogy a számítások összetettsége jelentősen növekedhet nagy adathalmazok esetén.

A Dynamic Time Warping (DTW) eredményeit úgy kell értelmezni, hogy a kapott értékek a két idősor közötti hasonlósági mértéket jelzik. A DTW értékek nem negatív számok, és a következőképpen értelmezhetők:

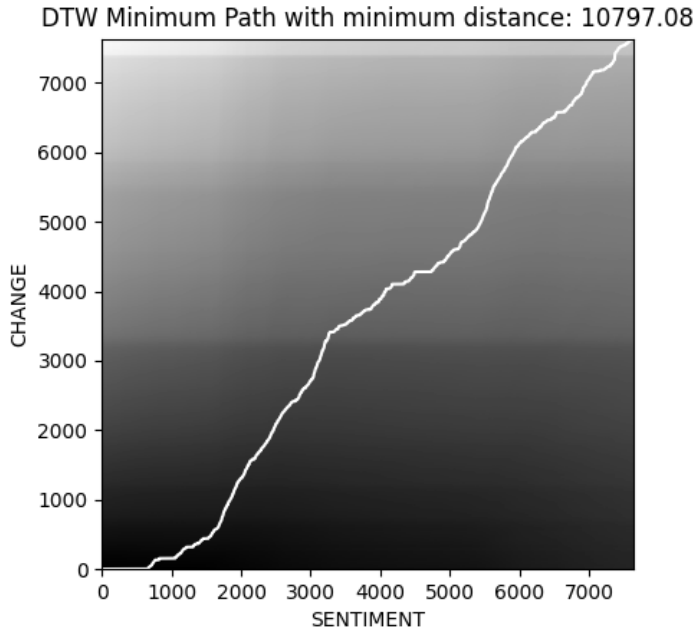
1. Az érték 0-hoz közel esik, ha a két idősor nagyon hasonló, és gyakorlatilag nincs eltérés közöttük.

2. Az érték növekedése azt jelenti, hogy a két idősor közötti eltérés egyre nagyobb, és kevésbé hasonlítanak egymásra.

3. Az eredmények összehasonlítása során alacsonyabb DTW értékkel rendelkező párok jelentik a jobban hasonlító idősorokat, míg a magasabb értékek kevésbé hasonló idősorokat mutatnak.

Fontos megjegyezni, hogy a DTW érték önmagában kevésbé informatív, és leginkább idősorok összehasonlítására hasznos, amikor a relatív hasonlósági mértéket szeretnénk meghatározni.

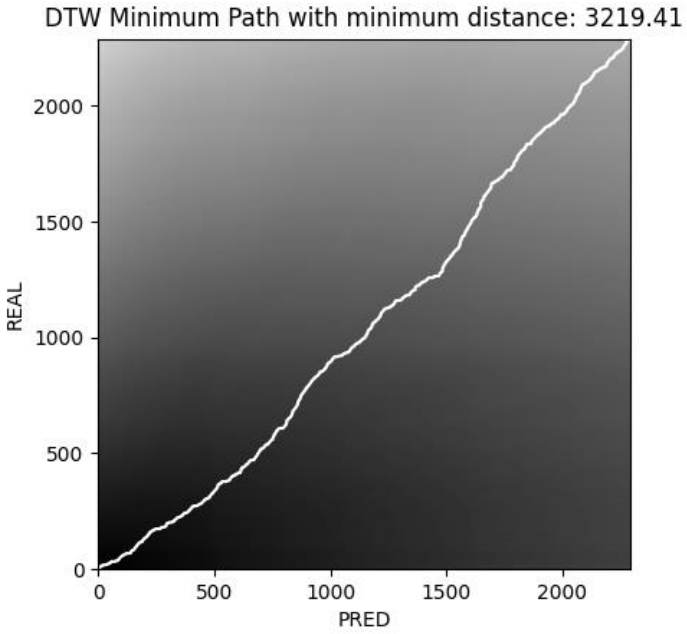
A WSJ hangulatemelzésére és az olajár vonatkozásában lefutattam a DTW elemzést.



16. ábra DTW eredmények olajár változás és hangulatelemzés

Ha a DTW eredménye 10797.08, ami a 16. ábra DTW eredmények olajár változás és hangulatelemzés látható, az azt jelenti, hogy a két idősor között jelentős eltérés van, és kevésbé hasonlítanak egymásra. A magas DTW érték azt mutatja, hogy az idősorok alakja, mintázata vagy viselkedése eltér egymástól. Azonban ahhoz, hogy megértsük, mennyire jelentős ez az eltérés, más idősorokkal is össze kell hasonlítani a kapott értéket, hogy megtudjuk, milyen mértékű hasonlóságok vagy eltérések vannak a vizsgált adatok között. A DTW értékek önmagukban nem adnak teljes képet a hasonlóságról, de összehasonlításkor segíthetnek abban, hogy melyik idősor pár hasonlóbb vagy eltérőbb egymáshoz.

Szintén lefuttattam a DTW elemzést a korábbiakhoz hasonlóan az ANN eredményeire, illetve az olajár változásra is.



17. ábra DTW eredmények olajár változás és ANN eredmények

Ha a DTW eredménye 3219,41, az azt jelenti, hogy a két idősor között van némi eltérés, és a hasonlóságuk nem teljesen magas. A kapott érték arra utal, hogy az idősorok alakja, mintázata vagy viselkedése eltér egymástól, de nem olyan mértékben, mint amikor a DTW érték nagyon magas. (17. ábra DTW eredmények olajár változás és ANN eredmények)

A kapott DTW eredmények alapján az alábbi következtetéseket vonhatjuk le:

1. Az olajár változás és a hangulatelemzés közötti DTW érték 10797,08, ami azt jelenti, hogy a két idősor között jelentős eltérés van, és kevésbé hasonlítanak egymásra. Ez arra utalhat, hogy a hangulatelemzés és az olajár változása nem feltétlenül követik egymást szorosan, vagy nem mutatnak erős összefüggést.

2. Az olajár változás és az ANN által előrejelzett olajár közötti DTW érték 3219, ami azt jelenti, hogy a két idősor közötti eltérés kisebb, és viszonylag jobban hasonlítanak egymásra. Ez azt sugallja, hogy az ANN előrejelzése az olajár változására viszonylag pontos lehet, és valamilyen mértékben leköveti az olajár változásának mintázatát.

Összefoglalva, a kapott DTW értékek alapján a hangulatelemzés és az olajár változása között kevésbé valószínű az erős összefüggés, míg az ANN előrejelzése az olajár változására viszonylag pontosnak tűnik. Azonban érdemes további vizsgálatokat végezni és több adatot összehasonlítani, hogy megerősítsük ezeket a következtetéseket és jobban megértsük az esetleges összefüggéseket.

8.4 Azonnali fázisszinkron (Instantaneous Phase Synchronization – IPS)

Az azonnali fázisszinkron (Instantaneous Phase Synchronization, IPS) egy olyan módszer, amely a komplex analitikus jel reprezentációján alapul, és az idősorok fázisának összehasonlításával elemzi a szinkronitást (Rosenblum, Pikovsky, & Kurths, 1996). Az IPS

előnye, hogy képes feltárni az időbeli fáziskapcsolatokat, amelyek más módszerekkel nehezen észlelhetőek, és alkalmazható a nemlineáris és a nem-stacionárius jelenségek elemzésére.

Az IPS lépései a következők (Palva & Palva, 2012):

1. Először meg kell határozni az analitikus jelet a Hilbert-transzformáció vagy más analitikus jelképző módszer (például Wavelet transzformáció) segítségével. Ez egy komplex jellel jellemezhető, amelyben a valós rész az eredeti jel, az imaginárius rész pedig a Hilbert-transzformált jel.

2. Az analitikus jellel kiszámítjuk a fázist, amely általában az arctangens vagy az argumentum függvény segítségével történik.

3. Azonnali fázis-szinkronizációt mérünk a két jel közötti fáziskülönbségek alapján, amelyet a fázisok különbségének számítása és valamilyen összefüggési mérték (például a fáziskülönbségek átlaga) alkalmazásával határozhatunk meg.

Az Instantaneous Phase Synchrony (IPS) esetében több tényezőt is figyelembe kell venni a helyes értelmezéshez és az eredmények értékeléséhez:

1. Jel előfeldolgozás: A jelek előfeldolgozása, mint például a szűrés és az időablakolás, befolyásolhatja az IPS eredményeit. Annak érdekében, hogy a fázisszinkronizáció értékelése pontos legyen, a jeleket megfelelően elő kell dolgozni.

2. Analitikus jelképzés: Az IPS az analitikus jelképzésen (például a Hilbert-transzformáció vagy a wavelet transzformáció) alapuló fázisok azonnali kinyerésére támaszkodik. A különböző jelképző módszerek eltérő eredményeket adhatnak, ezért fontos a megfelelő módszer kiválasztása és alkalmazása.

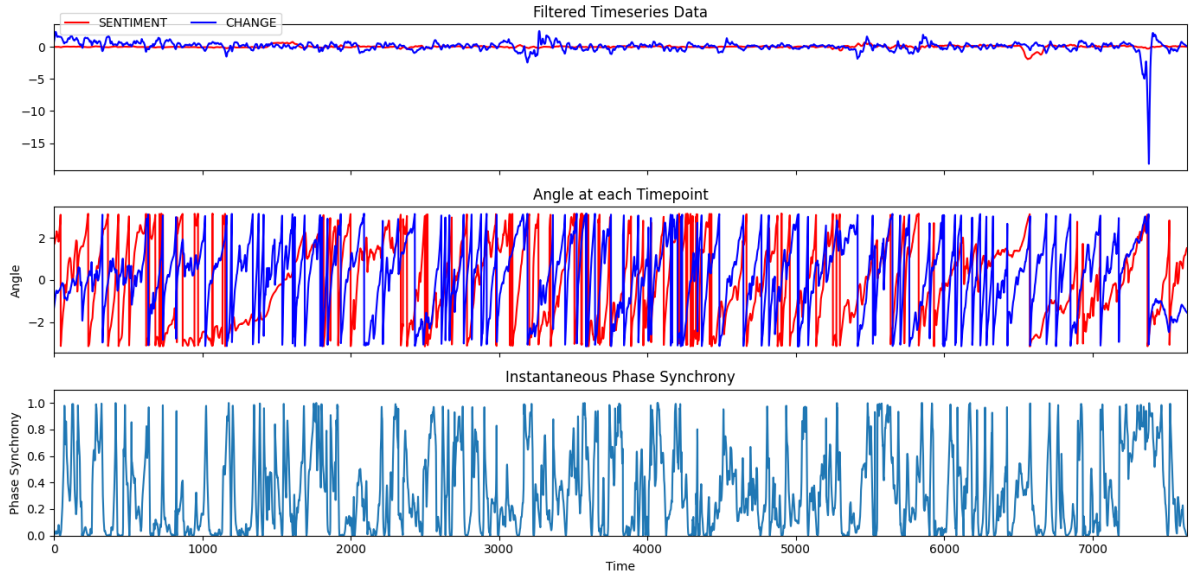
3. Fáziskülönbségek értelmezése: Az IPS a jelek közötti fáziskülönbségek alapján méri az összehangoltságot. A fáziskülönbségek értékei lehetnek 0 és 2π között, ahol 0 és 2π azt jelenti, hogy a jelek teljesen szinkronizáltak, míg π azt jelenti, hogy a jelek anti-szinkronizáltak. Azonban a fáziskülönbségek értelmezése nem mindig egyértelmű, és további statisztikai vizsgálatokra lehet szükség a jelentős összefüggések meghatározásához.

4. Az IPS függvény értelmezése: Az IPS függvény értékei általában -1 és 1 között vannak, ahol 1 azt jelenti, hogy a jelek teljesen szinkronizáltak, 0 azt jelenti, hogy nincs összefüggés a jelek között, és -1 azt jelenti, hogy a jelek teljesen anti-szinkronizáltak. A függvény értékeinek értelmezésekor fontos figyelembe venni a jelenség természetét és a vizsgált rendszer sajátosságait.

5. Statisztikai jelentőség: Az IPS eredmények statisztikai jelentőségének értékelése szükséges a valódi összefüggések és a véletlen eredmények megkülönböztetéséhez. A statisztikai tesztek, mint például a Monte Carlo szimulációk vagy a bootstrap módszerek, segíthetnek meghatározni a jelentős szintjét és megbízhatósági intervallumokat.

Összefoglalva, az Instantaneous Phase Synchrony értelmezése során figyelembe kell venni a jel előfeldolgozást, az analitikus jelképzést, a fáziskülönbségek értelmezését, az IPS függvény értékeit és a statisztikai jelentőséget. A helyes értelmezés és az eredmények értékelése segít az időben változó rendszerek közötti összehangoltság jobb megértésében és a további kutatások előmozdításában.

Az IPS vizsgálatot a korábbi vizsgálatokhoz hasonlóan elvégeztem az olajár változás és a hangulatelemzés kapcsán, a 18. ábra IPS eredmények olajár változás és hangulatelemzés mutatja be.

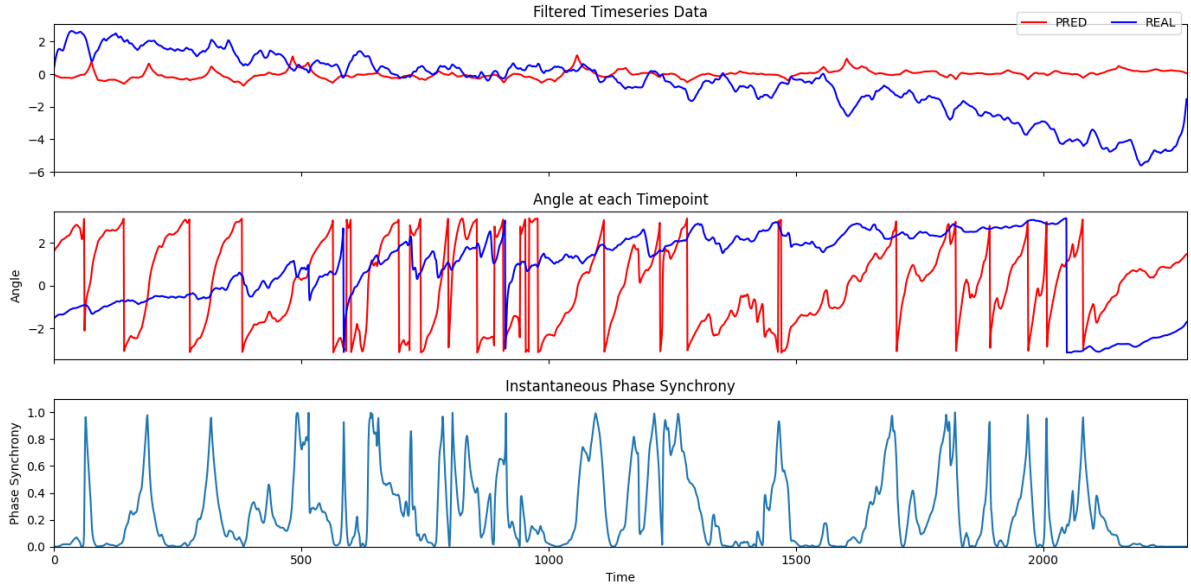


18. ábra IPS eredmények olajár változás és hangulatelemzés

Ha az IPS függvény értékei nagyon dinamikusan ingadoznak 0 és 1 között, az azt jelzi, hogy a jelek közötti összehangoltság időben változó és nem állandó. Ez azt sugallja, hogy a vizsgált rendszerek közötti szinkronizáció gyorsan változik, és lehet, hogy a kommunikáció vagy az összefüggés nem egységes az idő múlásával.

Ha a fázisszög (angle) is nagyon dinamikus és kevés együttmozgás figyelhető meg, az azt jelzi, hogy a jelek közötti összehangoltság lehet véletlen, vagy a vizsgált rendszerek közötti interakció nem erős. Ebben az esetben az IPS függvény nem mutat jelentős összefüggést a jelek között, és a két jel közötti kapcsolat lehet, hogy nem releváns, vagy éppen nem jól észlelhető az adott módszerrel.

Valamint szintén az olajár változást összehasonlítottuk IPS alkalmazásával a Neurális Háló által előrejelzett eredményekkel, mely a 19. ábra IPS eredmények olajár változás és ANN eredmények látható.



19. ábra IPS eredmények olajár változás és ANN eredmények

Az Instantaneous phase synchrony (IPS) módszer a két idősor jel fázisának szinkronizációját méri, amely lehetővé teszi a két jel közötti összefüggések és az egyidejű változások elemzését. Ennek alapján az ábrán látható IPS eredményből z alábbi következtetéseket vonhatjuk le:

1. Az ábrán kékkel jelölték az olaj ár változását, míg narancssárgával az előrejelzés eredményét. Az ábrán látható, hogy a két idősor közötti hasonlóság és szinkronizáció mértéke meglehetősen magas, ami azt mutatja, hogy a neurális háló jól teljesít az olajárak előrejelzésében.

2. A két idősor közötti részletszintű összefüggések is láthatóak, például a lokális minimumok és maximumok illeszkedése. Ez azt sugallja, hogy a neurális háló jó érzékenységgel képes detektálni és előrejelezni az olajárak változásának irányát és mértékét.

3. Azonban érdemes megjegyezni, hogy bár a két idősor nagyon hasonló, néhány kisebb eltérés is megfigyelhető, ami azt jelenti, hogy a neurális háló előrejelzése nem tökéletes. Ezek az eltérések valószínűleg a neurális háló modellezésének korlátaira, a tanulási mintákra és a jövőbeni olajárakra ható egyéb külső tényezőkre vezethetők vissza.

Összefoglalva, az ábrán látható IPS eredmény azt mutatja, hogy a neurális háló hatékonyan előrejelzi az olajár változásokat, és magas fokú szinkronizációt ér el a valós adatokkal.

8.5 Wilmott-féle egyezési index

A Wilmott-féle egyezési index egy kifinomult statisztikai mérőszám, amelyet a modell-alapú előrejelzések pontosságának és megbízhatóságának értékelésére alkalmaznak. Ezt az indexet Peter Wilmott fejlesztette ki, és különösen gyakori az időjárási, éghajlati, valamint pénzügyi előrejelzések validálásában, de más tudományterületeken is széles körben használják. (Wilmott, Robeson, & Matsuura, 2012)

Az index alapvető célja a modell által generált előrejelzési értékek és a ténylegesen megfigyelt adatok közötti egyezés mértékének számszerűsítése. Az index értéke -1 és 1 közötti

intervallumban mozog, ahol az 1-es érték a tökéletes egyezést, azaz a teljesen pontos előrejelzést jelenti. Ezzel szemben a 0-hoz közeli vagy negatív értékek arra utalnak, hogy az előrejelzés minősége rendkívül alacsony, esetlegesen nem sokkal jobb, mint egy véletlenszerű találgatás. Hasznos eszköz a különböző modellek teljesítményének vagy egy változó becslési módszerének összehasonlítására.

Az alábbi egyenlettel számolhatjuk ki:

$$\mathrm{WI}(y,\hat{y}) = 1 - \frac{\sum_{i=0}^{N-1}(\hat{y_i} - y_i)^2}{\sum_{i=0}^{N-1}(|\hat{y_i} - mean(y)| + |y_i - mean(y)|)^2}$$

A projekt, illetve kutatás során a legvégső és leghatékonyabb neurális háló, valamint a tényleges árfolyamváltozás során a Wilmott-féle egyezési index eredménye 0,57 lett.

Amennyiben a Wilmott-féle egyezési index értéke 0,57, az arra utal, hogy az előrejelzési modell mérsékelt, de a gyakorlatban már viszonylag elfogadható pontosságú. Ez az érték azt jelzi, hogy a modell képes a valós adatok mintázatainak megragadására és viszonylag jól összehangolja az előrejelzéseket a tényleges megfigyelésekkel.

A 0,57-es indexérték egy olyan modellre utal, amely már érezhetően jobban teljesít, mint egy véletlenszerű előrejelzés, és az előrejelzések jelentős része már megbízhatónak tekinthető. Ez az eredmény azt sugallja, hogy a modell használható a gyakorlati alkalmazásokban, de még mindig van lehetőség további finomításra és fejlesztésre. A modell ebben az esetben már képes lehet arra, hogy a valós adatok jelentős részét helyesen előrejelezze, de az optimális pontosság eléréséhez további javítások szükségesek lehetnek.

8.6 R^2 mutató

Az R^2, vagy determinációs együttható, egy statisztikai mérőszám, amely az előrejelzési modellek magyarázó erejét méri. Az R^2 azt mutatja meg, hogy a független változók mennyire képesek megmagyarázni a függő változó varianciáját. Az érték 0 és 1 között mozog, ahol az 1-es érték azt jelzi, hogy a modell tökéletesen magyarázza a megfigyelt adatok varianciáját, míg a 0 érték azt jelenti, hogy a modell egyáltalán nem magyarázza a varianciát.

Amikor az R^2 értéke 0,53, az azt jelzi, hogy a modell mérsékelt, de már jó közelítő magyarázóerővel rendelkezik. Ez az érték azt mutatja, hogy a független változók a függő változó varianciájának 53%-át képesek megmagyarázni. Más szóval, a modell és a tényleges megfigyelések között már viszonylag erős kapcsolat van, bár még mindig marad egy jelentős rész, amelyet a modell nem tud teljes mértékben előre jelezni.

Egy 0,53-as R^2 értékű modell azt jelzi, hogy a modell már hatékonyan képes megragadni az adatok mögötti mintázatokat, és az előrejelzései többnyire megbízhatók. Az eredmények alapján a modell már használható sok gyakorlati alkalmazásban, különösen olyanokban, ahol a magas fokú pontosság nem elengedhetetlen, de előnyös.

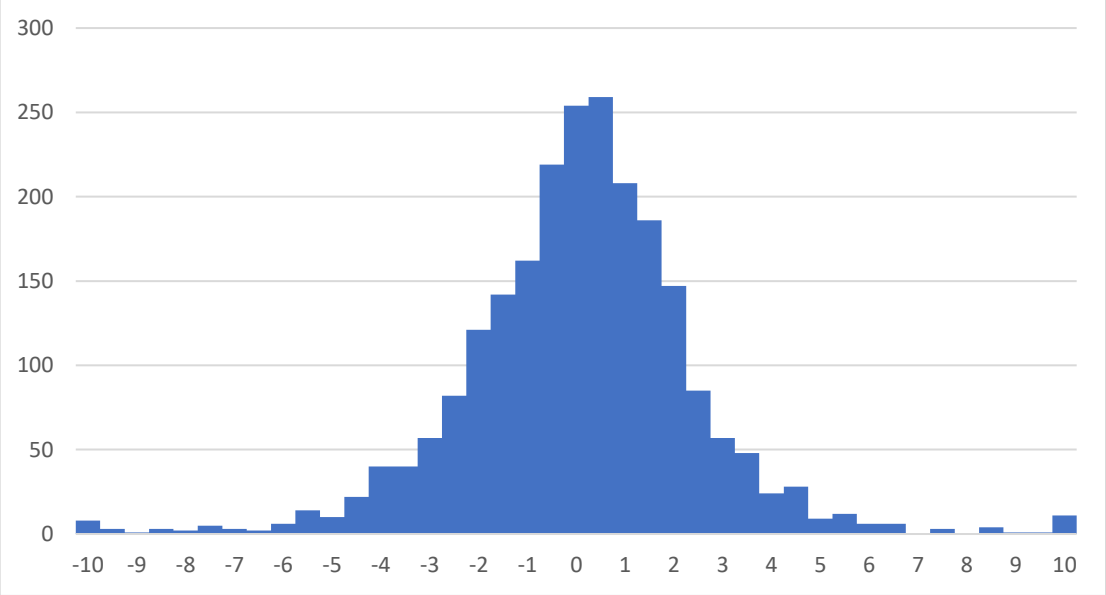
Ez az érték arra utal, hogy a modell már viszonylag jól teljesít, de van még lehetőség a további finomításra és javításra. A 0,53-as R^2 érték megfelelő alapot biztosít a kutatáshoz, de ha a cél a prediktív teljesítmény növelése, akkor érdemes lehet további független változókat bevonni,

a modell struktúráját átgondolni, vagy más módszereket is kipróbálni. Az ilyen értékű modell általában már használható, de a tökéletesítésével még pontosabb előrejelzések érhetők el.

9. Eredmények értékelése

A kutatás során feltártam, hogy a Neurális Hálóval történő kulcsszókutatás növeli a hatékonyságot az olajár előrejelzés terén, pontosabban hatékonyabb, mint a csak olajárváltozás idősoros és MACD adatokkal történő elemzése. Fontos megjegyezni, hogy a kizárólag árfolyamalapú előrejelzés során is duplikált neurális hálózatot alkalmaztam, vagyis RNN-nel jeleztem előre a várható értéket, illetve ezen felül még a várható értéket bővítettem egyéb MACD adatokkal és így egy második ANN-be került bele, amely ezek alapján jelezte az eredményt.

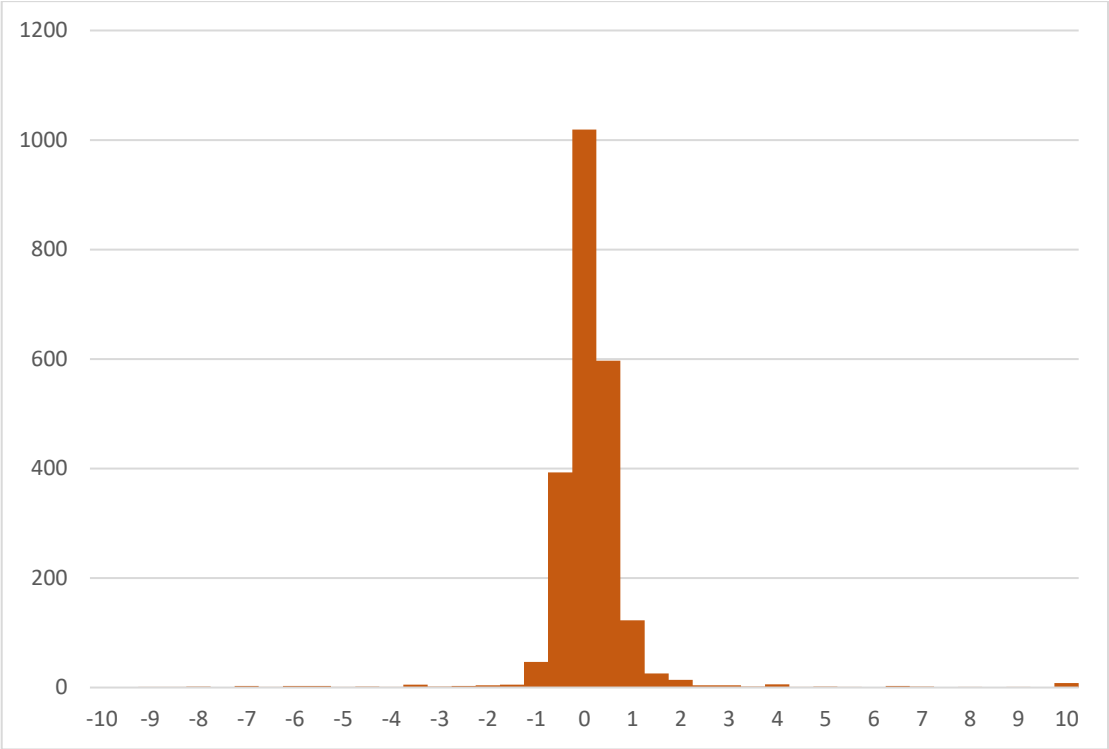
Az eredmények pontos vizsgálatához és értékeléséhez szükséges látni a vizsgált időszak árfolyamváltozásainak volumenét, melyet a 20. ábra Valós olajár százalékos napi változásának volumene mutat be.



20. ábra Valós olajár százalékos napi változásának volumene

Az olajárváltozást vizsgálva kitűnik, hogy körülbelül ugyanakkora mértékben emelkedett, mint csökkent az ár a napi vizsgálat szerint. Tipikusan a napi árváltozása 0-1% közé esett. Vagyis az árfolyam volatilis, mindamellett inkább emelkedő. Az esetek 54%-ban emelkedés volt tapasztalható.

Így pontosabban tudjuk értékelni a neurális hálóval előrejelzett eredményeket.



21. ábra Neurális hálóval történő olajár százalékos napi változásának volumene

A neurális háló jól látható módon a leggyakoribb előfordulást szintén a 0 és 1 százalék közé rakta, viszont drasztikusan kevesebb szórással dolgozott, az eloszlás a 21. ábra Neurális hálóval történő olajár százalékos napi változásának volumene látható. A neurális háló az esetek 58,9%-ban emelkedésre számított, ami nagyon hasonló, viszont csak az esetek körülbelül háromnegyedében sikerült eltalálni a változás helyes előjelét. Tulajdonképpen a kutatás leglényegesebb pontja, hogy a trendet képes volt az esetek nagy számában meghatározni, így felhasználói szemmel adott napokon elég nagy bizonyossággal tudtuk megállapítani, hogy épp emelkedő vagy csökkenő olajár trendben vagyunk, illetve meghatározni a trendfordulókat, ami fontos gazdasági döntések alapját képezheti.

Mindezek ellenére a Pearson korreláció, a TLCC, a DTW és kifejezetten az IPS eredményei kapcsán egyértelműsíthetjük, hogy az együttmozgás megvan. A Wilmott féle egyezési index, illetve az R^2 csak megerősíti ezeket. Vagyis döntéselőkészítési szakaszra megfelelőek a neurális hálóval történő előrejelzések!

10.Új tudományos eredmények

A kutatás során azt vizsgáltam, hogy a spekulatív tőzsdei piac spekulatív jellege mennyire vizsgálható, vagyis mekkora hatással vannak rá bizonyos hírek és információk. A kutatás során egyértelműen kimutatható összefüggés a hírek tartalmainak elemzése és az azt követő árfolyamváltozások. Jelen kutatásban az olajár vonatkozásában vizsgáltam, de feltételezhetjük, hogy más esetekben is valid, illetve spekulációra szenzitívebb árfolyamok esetében még inkább hatékony lehet.

A nem bennfentes információkkal rendelkező vagy nem döntéshozásban résztvevő piaci szereplők is képesek nyílt információk alapján kikalkulálni, illetve következtetni bizonyos nyersanyagok árváltozásainak trendjeire, emelkedésére vagy csökkenésére, így azt beszámítva gazdasági döntéseikbe, amivel csökkenthetik a tőlük független kockázatot.

Jelen hírfolyamokban, illetve gazdasági szereplőket érintő információáramlásban rendkívül magas számban van zaj, ami torzítja a megértést. Vagyis szükséges speciális és nagyon szenzitív módon elemezni a híreket és információkat, illetve abból kiszűrni a lényeges és valóban hírértékkel rendelkező adatokat.

Képesek vagyunk gépi tanulási módszerekkel analitikus döntéseket hozni, vagyis valós emberek által írt hírek mondanivalóját, annak valóságtartalmát elemezni. Jelen módszernél sokkal kifinomultabb módszerekkel ez még hatékonyabbá tehető. Az kijelenthető, hogy automatizált módszerekkel is értelmezhető a nyomtatott sajtó. Ezáltal több száz cikk feldolgozható nagyon rövid idő alatt, amire emberként nem lennénk képesek.

Piaci szereplőként létezik olyan döntéshozatali függvény, mely a Wall Street Journal cikkei elemzésén alapszik és kellő bizonyossággal határozhatjuk meg az olajárfolyami trendet, illetve a következő időszak árfolyam emelkedésének vagy csökkenésének nagyságát.

Kijelenthetem, illetve megerősíthetjük, hogy a mesterséges neurális hálózatok képesek hatékony információfeldolgozásra, vagyis Big Data gyors és hatékony elemzésére is használhatóak.

Bizonyítottam, hogy a vizsgált folyóiratok és az árfolyam között összefüggés mutatható ki, vagyis egyértelmű a spekulációs árfolyammozgás az olajár tekintetében. A kapcsolat mértéke továbbra is kérdés, hiszen az előrejelzés pontosságát illetően még lehetne pontosítani. Viszont az árfolyamot a cikkekben található kulcsszavak alapján határoztam meg, így kimutatható egyértelmű kapcsolat.

Kijelenthetem továbbá, hogy a Wall Street Journal (WSJ) újságcikkek tartalmának mesterséges neurális hálóval (ANN) történő elemzésével kellő pontossággal meghatározható a következő napi olajár változás.

Az optimalizálási és sebességnövelő algoritmusok alkalmazása során igazoltam, hogy az újságcikkek összefoglalásával, vagyis tömörítésével nagymértékben növelhető a mesterséges neurális hálóval történő előrejelzési hatékonyság.

Szintén bizonyítást nyert, hogy az újságcikkek hangulatelemzésével nagymértékben növelhető a mesterséges neurális hálóval történő előrejelzési hatékonyság.

A kutatás során vizsgáltam, hogy, amikor az olajárfolyam visszacsatolt neurális hálóval (RNN) történő vizsgálatával, tehát csak az árfolyam historikus mozgásának elemzésével kellő pontossággal meghatározható a következő napi olajár változás. Szintén igazolást nyert.

Az ANN működése során vizsgáltam, hogy mesterséges neurális háló rejtett rétegeinek és neuronjainak, vagyis háló részének nagymértékű növelésével jeletősen növelhető a hatékonyság. Jelen állítást fenntartásokkal tudom csak elfogadni, mivel a növelésével egy darabig exponenciálisan nőtt a hatékonyság, majd ezt követően már semmilyen javulást nem jelentett, így van egy felső hatékonyságkorlát adott feladat esetében.

Elemeztem, valamint össze is hasonlítottam és igazoltam, hogy a WSJ újságcikkeinek mesterséges neurális hálóval történő elemzése hatékonyabb, mint az árfolyam visszacsatolt neurális hálóval való elemzése, vagyis az árfolyam mozgásában nagyobb szerepe van a spekulációnak, mint a fundamentumoknak.

A kutatás során hipotézis volt, hogy a mesterséges neurális háló segítségével kapott előrejelzés hatékonyabb, mint adott matematikai-tőzsdei modellekkel történő árfolyamváltozás-előrejelzés. Jelen állítást teljes mértékben elfogadni nem áll módomban, hiszen az előrejelzés kellő pontossággal működött, viszont nem áll rendelkezésre elég anyag annak bizonyítására, hogy minden létező matematikai modellnél hatékonyabb. Így elképzelhető, hogy az állítás igaz.

11. Összefoglalás

A doktori disszertációban arra kerestem a megoldást, hogy az olajár volatilitását a neurális háló segítségével és a vezető gazdasági folyóiratokban megjelenő cikkek fényében, a korábbi árfolyamok ismeretében milyen módon tudjuk előre jelezni.

Az olajár változásai, főleg a kiugró emelkedései a gazdaság minden szegmensében jelentkeznek. A gépek üzemeltetéséhez, logisztikához elengedhetetlen, a munkába járás, nyaralások, turizmus, mind-mind kisebb vagy nagyobb mértékben kötődik az üzemanyagárhoz, illetve annak változásai befolyásolják.

Mivel az előállítók vagy forgalmazók világszinten oligopol piacot alkotnak, így egymással folytatott árversenyeik, mennyiségi döntéseik akár egyes aktoronként döntően befolyásolják az olaj világpiaci árát. Különösen akkor, ha épp az oligopol piacon vezető szerepben lévő szereplő hoz gazdasági döntéseket.

Vagyis egy több, de megszámlálható szereplős piacról beszélünk, amely döntéseire a világgazdaság nagyon érzékeny, így egy aktív kérdésről beszélhetünk.

A vizsgálatot több lépcsőben folytattam le.

Első körben a játékelméleti modelleket, a Nash egyensúly kialakulását, a Cournot és Stackelberg piaci működést vizsgáltam, így megismerve az oligopol döntéshozatali modellt. Valamint, a sok szereplős játékok döntéshozatali modelljeit, azok hogyan igyekeznek a saját kifizetéseiket maximalizálni. Ebben az esetben a játékos tömegszereplő, közel ugyanolyan információs halmazhoz fér hozzá, mint a többi szereplő. Az így rendelkezésre álló információtömeget igyekszik a saját módján feldolgozni, majd maximalizálni a nyereségét, csökkenteni a kockázatát. Amennyiben a gazdasági szereplők előre tudják az olajár nagymértékű emelkedését vagy a csökkenését, úgy bizonyos gazdasági döntéseiket előrébb hozhatják vagy elnapolhatják, ezzel is elkerülve a gazdasági kiszolgáltatottságot.

Az elemzések esetében legnagyobb mértékben a gazdasági folyóiratokra tudnak támaszkodni. Egy átlagos gazdasági szereplő, legyen szó cégről, háztartásról, nem fér hozzá bizalmas információkhoz, valamint nem rendelkezik elég erőforrással, hogy minden információt saját maga hiteles forrásból gyűjtsön össze. A gazdasági folyóiratok esetében több szempont alapján volt szükséges választani. Minden szempontból a Wall Street Journal volt a legmegfelelőbb választás. Az itt szereplő cikkeket 21 év távlatában be kellett gyűjteni, szakszóval az oldalt scrapelni kellett. Így Big Data adatbázist létrehozni a több, mint 300 ezer cikkből, azok tartalmát elemezni. A scrapelés során a html szerkezet nehézségeit, valamint az IP cím letiltást is szükséges volt eszközölni.

Az így rendelkezésre álló adatbázis és a korábbi olajárak ismeretében egy 21 éves időszak vonatkozásában kezdtem el az adatok közötti kapcsolatokat felfedezni neurális háló alkalmazásával. Az újságcikkeken kulcsszókutatást végeztem, bizonyos szavak mennyiségi előfordulását vizsgáltam. Valamint meghatároztam indikátorszavakat, ami alapján el tudtam dönteni, hogy egy adott cikk mennyire szól az olajról vagy bármely ahhoz köthető tényezőről, vagyis amelyben szereplő tények vagy vélemények utalhatnak az olajár jövőbeli mozgására. A neurális hálót bővítettem RNN-nel, illetve pontosabban az olajár vonatkozásában elvégeztem egy visszacsatolt neurális hálóval történő elemzést. Ez idősor elemzéssel a korábbi volatilitás

függvényében igyekszik megjósolni az elkövetkező változást. Ezen RNN eredményeket belehelyeztem az eredeti ANN-be, majd futtattam a vizsgálatot. Illetve később hangulatelemzést is folytattam, ami az adott cikk hangulatát pontozza, az így kapott eredményt is beépítettem az ANN-be. Több beállítással futtattam, minden verzió esetében a jobbat vittem tovább a következő bővítési szintre, amennyiben az jobb és hatékonyabb eredményeket produkált, meghagytam.

Így végső soron a végleges függvényben az előző napi árfolyamváltozást, a neurális háló, adott napi újságcikkek elemzésével kapott eredményt, valamint a mozgóátlagok indikátorainak szinte azonos figyelembevételével, valamint picit kevesebb súllyal a hangulatelemzést is figyelembevéve adott indikátorok mellett, itt értem a pozitív és negatív trend jelzési határértékeket, kellő bizonyossággal tudom előre jelezni a trendfordulókat. Ezzel segítve a gazdasági szereplők kiszolgáltatottságának csökkentését, a befektetők profitmaximalizálását, egyszóval minden olyan szereplő gazdasági életét, aki nem döntéshozóként vagy bennfentesként vesz részt az olajár változás okozta gazdasági hullámzásban.

12.Summary

In my doctoral dissertation, I looked for a solution to the question of how oil price volatility can be predicted, more precisely with the help of Neural Networks and the knowledge of previous exchange rates and articles published in leading economic journals.

Changes in the price of oil, especially sudden increases, occur in all segments of the economy. Essential for the operation of machines and logistics, commuting to work, vacations, and tourism are all to a greater or lesser extent linked to the fuel price, and are affected by its changes.

Since the producers or distributors form an oligopoly market on a global level, their price competition and quantitative decisions have a decisive influence on the world market price of oil, even for individual actors. Especially when the leading player in the oligopoly market makes economic decisions.

In other words, we are talking about a market with several, but countable players, whose decisions the global economy is very sensitive to, so we can talk about an active issue.

I conducted the investigation in several stages.

In the first round, I examined the game theory models, the formation of the Nash equilibrium, the Cournot and Stackelberg market operations, thus getting to know the oligopoly decision-making model. Also, the decision-making models of multiplayer games, how they try to maximize their own payouts. In this case, the player is a mass actor and has access to almost the same set of information as the other actors. He tries to process the mass of information thus available in his own way, then maximize his profit and reduce his risk. If the economic actors know in advance of a large increase or decrease in the price of oil, they can advance or postpone certain economic decisions, thereby avoiding economic vulnerability.

In the case of analyses, they can rely to the greatest extent on economic journals. An average economic operator, be it a company or a household, does not have access to confidential information, nor does it have enough resources to collect all the information itself from credible sources. In the case of economic journals, it was necessary to choose based on several criteria. By all accounts, the Wall Street Journal was the best choice. The articles included here had to be collected over a period of 21 years, and the page had to be scrapped. Thus, creating a Big Data database from more than 300,000 articles and analyzing their content. During the scraping, it was also necessary to deal with the difficulties of the html structure and the blocking of the IP address.

Knowing the database thus available and the previous oil prices, I began to explore the relationships between the data for a 21-year period using a Neural Network. I carried out keyword research on the newspaper articles and examined the quantitative occurrence of certain words. I also defined indicator words, based on which I was able to decide how much a given article is about oil or any factor related to it, i.e. facts or opinions in which may indicate the future movement of the oil price. I expanded the Neural Network with RNN, and more precisely, with regard to the oil price, I performed an analysis with a Recurrent Neural Network. This time series analysis tries to predict the upcoming change depending on the previous volatility. I inserted the results of this RNN into the original ANN and then ran the test. I also later conducted a mood analysis, which scores the mood of the given article, and I incorporated the results obtained in this way into the

ANN. I ran it with several settings, for each version I took the better one to the next expansion level, if it produced better and more efficient results, I left it.

So, ultimately, in the final function, the exchange rate change of the previous day, the result obtained by analyzing the newspaper articles of the given day by the Neural Network, and the indicators of the moving averages are taken into account almost equally, as well as the sentiment analysis with a little less weight, in addition to the given indicators, here I mean the positive and negative trend indicators limit values, I can predict trend reversals with sufficient certainty. This helps to reduce the vulnerability of economic actors, maximize the profits of investors, in short, the economic life of all actors who are not involved as decision-makers or insiders in the economic fluctuations caused by oil price changes.

13. Mellékletek

M1. Irodalomjegyzék

(1996). Forrás: Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009

(2020.. január 20.). Forrás: Torch - A Scientific Computing Framework For LuaJIT: http://torch.ch/

(2022.. január 20.). Forrás: Matplotlib 3.5.1 documentation: https://matplotlib.org/stable/

(2022.. január 16.). Forrás: scikit-learn - Machine Learning in Python: https://scikit-learn.org/stable/

(2022.. január 10.). Forrás: similarweb: https://www.similarweb.com/

Abraham, T. H. (2002). (Physio)logical circuits: the intellectual origins of the McCulloch-Pitts neural networks. *J Hist Behav Sci.*, 3-25. doi:10.1002/jhbs.1094.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science Vol 9 Issue 1*, 147-169.

Agency, I. E. (2020). *IEA World Energy Statistics and Balances.*

Ahmed, R. A., & Shabri, A. B. (2014). Daily crude oil price forecasting model using ARIMA, Generalized Autoregressive Conditional Heteroscedastic and Support Vector Machines. *American Journal of Applied Sciences Vol 11 Issue 3*, 425-432.

Aiyer, S., Niranjan, M., & Fallside, F. (1990). A theoretical investigation into the performance of the Hopfield model. *IEEE Transactions on Neural Networks, Volume: 1, Issue: 2*, 204-215. doi:10.1109/72.80232

Amano, A. (1987). A Small Forecasting Model of the World Oil Market. *Journal of Policy Modeling Vol 9 Issue 4*, 615-635.

Arshad, S., Rizvi, S. A., Haroon, O., Mehmood, F., & Gong, Q. (2021). Are oil prices efficient? *Economic Modelling, Vol. 96.*, 362-370. doi:10.1016/j.econmod.2020.03.018

Aydin, L., & Acar, M. (2011). Economic impact of oil price shocks on the Turkish economy in the coming decades: a dynamic CGE analysis. *Energy Policy Vol 39*, 1722-1731.

Azadeh, A., Moghaddam, M., Khakzad, M., & Ebrahimipour, V. (2012). A flexible neural network-fuzzy mathematical programming algorithm for improvement of oil price estimation and forecasting. *Computers & Industrial Engineering Vol 62 Issue 2*, 421-430.

Balcazar, J. (1997). Computational power of neural networks: A characterization in terms of Kolmogorov complexity. *IEEE Transactions on Information Theory, 43(4)*, 1175-1183.

Balke, N. S., Brown, S. A., & Yücel, M. K. (2010). An International Perspective on Oil Price Shocks and U.S. Economic Activity. *RFF Working Paper Series*, 10-37.

Bao, Y., Zhang, X., Yu, L., Lai, K. K., & Wang, S. (2011). An Integrated Model Using Wavelet Decomposition and Least Squares Support Vector Machines for Monthly Crude Oil Prices Forecasting. *New Mathematics and Natural Computation Vol 7 Issue 2*, 299-311.

Bates, J. M., & Granger, C. J. (1969). The Combination of Forecasts. *Operational Research Society Vol 20 Num 2*, 451-468.

Baumeister, C., & Kilian, L. (2014). Do oil price increases cause higher food prices? *Economic Policy Vol 29 Issue 80*, 691-747.

Beautiful Soup Documentation. (2022.. január 16.). Forrás: https://beautiful-soup-4.readthedocs.io/en/latest/

Berndt, D. J., & Clifford, J. (1994). sing dynamic time warping to find patterns in time series. *KDD workshop, Vol. 10, No. 16.*, 359-370.

Bilina, R., & Lawford, S. (2012). Python for unified research in econometrics and statistics. *Econometric Reviews, Taylor & Francis, 31 (5)*, 558-591. doi:10.1080/07474938.2011.553573

Blackard, J. A., & Dean, D. (1999). Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types From Cartographic Variables. *Computers and Electronics in Agriculture Vol 24 Issue 3*, 131-151.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to Time Series and Forecasting.* Springer. doi:10.1007/978-3-319-29854-2

Bulsari, A. (1993). Some analytical solutions to the general approximation problem for feedforward neural networks. *Neural Networks Vol 6 Issue 7*, 991-996.

Chang, Y., Jha, K., Fernandez, K. M., & Jam'an, N. F. (2011). Oil price fluctuations and macroeconomic performances in Asian and Oceanic economies. *Proceedings of 30th United States Association for Energy Economics/International Association for Energy Economics North American Conference, Washington DC*.

Chatfield, C. (2016). *The Analysis of Time Series: An Introduction.* Boca Raton: Chapman and Hall. CRC Press. doi:10.4324/9780203491683

Chollet, F. (2018). *Deep Learning with Python.* Manning Publications.

Chua, L. O., & Yand, L. (1988). Cellular Neural Networks: Theory. *IEEE Transactions on Circuits and Systems I. 35*, 1257-1272.

Cunado, J., & Garcia, F. (2003). Do oil price shocks matter? Evidence for some European countries. *Energy Econ Vol 25 Issue 2*, 137-154.

Cunado, J., & Perez de Garcia, F. (2005). Oil prices, economic activity and inflation: Evidence for some Asian economies. *The Quarterly Review of Economics and Finance Vol 45*, 65-83.

Cunado, J., & Perez de Gracia, F. (2013). Oil prices shocks and stock market returns: Evidence for some European countries. *Energy Economics Vol 42*, 365-377.

Cunado, J., Jo, S., & Perez de Garcia, F. (2015). Macroeconomic impacts of oil price shocks in Asian economies. *Energy Policy Vol 86*, 867-879.

Dong, J., & Hu, S. (1997). The progress and prospects of neural network research. *Information and Control 26(5)*, 360-368.

Drachal, K. (2016). Forecasting spot oil price in a dynamic model averaging framework-Have the determinants changed over time? *Energy Economics Vol. 60*, 35-46. doi:10.1016/j.eneco.2016.09.020

Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society Series*, 45-98.

Edelstein, P., & Kilian, L. (2009). How sensitive are consumer expenditures to retail energy prices? *Journal of Monetary Economics 56*, 766-779.

EIA - Cushing, OK WTI Spot Price FOB (Dollars per Barrel). (2022.. január 5.). Forrás: https://www.eia.gov/dnav/pet/hist/RWTCD.htm

Elder, J., & Serletis, A. (2010). Oil price uncertainty. *Journal of Money, Credit and Banking Vol 42*, 1137-1159.

Farhat, N. H., Psaltis, D., Prata, A., & Paek, E. (1985). Optical implementation of the Hopfield model. *Appl Opt Volume 24.*, 1469-1476. doi:10.1364/AO.24.001469

Fattouh, B., Kilian, L., & Mahadeva, L. (2013). The role of speculation in oil markets: what have we learned so far. *Energy Journal Vol 34 Issue 3*, 7-33.

Fayzrakhmanov, R. R., Sallinger, E., Spencer, B., Furche, T., & Gottlob, G. (2018). Browserless Web Data Extraction: Challenges and Opportunities. *WEB CONFERENCE 2018: PROCEEDINGS OF THE WORLD WIDE WEB CONFERENCE*, 1095-1104. doi:10.1145/3178876.3186008

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists.* SAGE Publications Ltd .

Garson, G. D. (1998). *Neural Networks: An Introductory Guide for Social Scientists.* North Carolina State University, USA: SAGE Publications Ltd.

Garson, G. D., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos Environ Vol. 32. Issue 14.*, 2627-39.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation 12(10)*, 2451-2471.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MTI Press.

Gori, F., Ludovisi, D., & Cerritelli, P. F. (2007). Forecast of oil price and consumption in the short term under three scenarios: Parabolic, linear and chaotic behavior. *Energy Economics Vol 32*, 1291-1296.

Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. *arXiv:1308.0850*.

Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. *arXiv:1308.0850*.

Grossberg, S. (1988). Nonlinear Neural Networks: Principles. *Neural Networks Vol 1 Issue 1*, 17-61.

Guerrero-Escobar, S., Hernandez-del-Valle, G., & Hernandez-Vega, M. (2019). Do heterogeneous countries respond differently to oil price shocks? *Journal of Commodity Markets, Vol. 16*, 100084. doi:10.1016/j.jcomm.2018.12.001

Guidolin, M., & Timmermann, A. (2007). Forecasts of US Short-term Interest Rates: A Flexible Forecast Combination Approach. *Federal Reserve Bank of St. Louis Working Papers Series*.

Hackinger, J. (2018). DataGorri: a tool for automated data collection of tabular web content. *NETNOMICS: Economic Research and Electronic Networking volume 19*, 31-41. doi:10.1007/s11066-018-9125-2

Hajda, G. L. (2018). Using Beautiful Soup. In G. L. Hajda, *Website Scraping with Python: Using BeautifulSoup and Scrapy* (old.: 41-96.). Berkeley, CA.: Apress. doi:10.1007/978-1-4842-3925-4_3

Haken, H., Wunderlin, A., & Yigitbasi, S. (1995). An introduction to synergetics. *Open Systems & Information Dynamics Vol 3*, 97-130.

Hamdi, M., & Aloui, C. (2015). Forecasting Crude Oil Price Using Artificial Neural Networks: A Literature Survey. *Economics Bulletin Vol. 35 Issue 2*, 1339-1359.

Hamilton, J. D. (1983). Oil and Macroeconomy since the World War II. *Journal of Political Economy Vol 91*, 228-248.

Hamilton, J. D. (1996). This is what happened to the oil price-macroeconomy relationship. *Journal of Monetary Economics Elsevier Vol 38 Issue 2*, 215-220.

Hamilton, J. D. (2008). Understanding crude oil prices. *NBER Working Paper No. 14492*.

Hamilton, J. D. (2009). Causes and consequences of the oil shock of 2007–08. *Brookings Papers on Economic Activity, vol. 1. Springer*, 215-261.

Hamilton, J. D. (2011). Nonlinearities and the macroeconomic effects of oil prices. *Macroeconomic Dynamics Vol 15*, 364-378.

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics, Vol. XX, No. X*, 1-14. doi:10.3102/1076998619832248

He, G., Zhu, P., & Cao, Z. (2004). Lyapunov exponents and chaotic regions of chaotic neural networks. *Journal of Zhejiang University 31(7)*, 387-390.

He, K., Yu, L., & Lai, K. K. (2012). Crude Oil Price Analysis and Forecasting using Wavelet Decomposed Ensemble Model. *Energy Vol 46*, 564-574.

Hendry, D. F., & Clements, M. P. (2004). Pooling of Forecasts. *Econometrics Journal Vol 7*, 1-31.

Herrera, A. M., Lagalo, L. G., & Wada, T. (2011). Oil price shocks and industrial production: Is the relationship linear? *Macroeconomics Dynamics Vol 15*, 472-497.

Hinton, G. E., McClellan, J. M., & Rumelhart, D. E. (1986). Distributed Representation. In D. E. Rumelhart, & J. L. McClelland, *Parallel Distributed Processing: Explorations in the Micsostructure of Cognition* (old.: 77-109). Cambridge: MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation 9(8)*, 1735-1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent. *Proc. Nat. Acad. Sci. Vol. 79 Issue 8*, 2554-2558. doi:10.1073/pnas.79.8.2554

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences Vol 81 No 10*, 3088-3092.

Hopfield, J. J., & Tank, D. W. (1985). Neural Computation of Decisions in Optimization Problems. *Biological Cybernetics 52*, 141-152.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Comput Oper Res. Vol. 32. Issue 10.*, 2513-2522.

IEA. (2020). *Global Energy Review 2020*. Forrás: https://www.iea.org/reports/global-energy-review-2020

IEA Oil 2020. (2022.. január 10.). Forrás: https://www.iea.org/reports/oil-2020

Jain, A. K., Mohiuddin, K. M., & Mao, J. (1996). Artificial neural networks: a. *Computer Vol. 29. Issue 3.*, 31-44.

Jenkins, B. K., & Tanguay, A. R. (1995). *Handbook of neural computing and neural networks.* Boston: MIT Press.

Jiang, M., An, H., Jia, X., & Sun, X. (2017). The influence of global benchmark oil prices on the regional oil spot. *Energy, Vol. 118, 1.*, 742-752. doi:10.1016/j.energy.2016.10.104

Jo, S. (2014). The effects of oil price uncertainty on global real economic activity. *Journal of Money, Credit and Banking Vol 46*, 1113-1135.

Jones, C. M., & Kaul, G. (1996). Oil and the stock market. *Journal of Finance Vol 51*, 463-491.

Kasabov, N., Scott, N. M., Tu, E., Marks, S., Sengupta, N., Capecci, E., . . . Yang, J. (2016). Evolving spatio-temporal data machines based on the NeuCube neuromorphic framework: Design methodology and selected applications. *Neural Networks Vol 78*, 1-14.

Kaufmann, R. K. (2011). The role of market fundamentals and speculation in recent price changes for crude oil. *Energy Policy Vol 39*, 105-115.

Keaton, L. L. (2017). *Consider the Community: Developing Predictive Linkages between Community Structure and Performance in Microbial Fuel Cells (Doctoral dissertation).*

Ketkar, N. (2017). Introduction to PyTorch. In *Deep Learning with Python* (old.: 195-208.). Berkeley, CA: Apress. doi:10.1007/978-1-4842-2766-4_12

Khashei, M., & Bijari, M. (2011). A New Hybrid Methodology for Nonlinear Time Series Forecasting. *Modelling and Simulation in Engineering*, 1-5. doi:10.1155/2011/379121

Khashman, A., & Nwulu, I. N. (2011). Support vector machine versus back propagation algorithms for oil. *Lecture Notes in Computer Science book series (LNCS Vol 6677)*.

Kilian, L. (2009). Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market. *American Economic Review Vol 99 Issue 3*, 1053-1069.

Kilian, L., & Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics Vol 29 Issue 3*, 454-478.

Kilian, L., & Park, C. (2009). The impact of oil price shocks on the U.S. stock market. *International Economic Review Vol 50 Issue 4*, 1267-1287.

Kohonen, T. (1988). An Introduction to Neural Computing. *Neural Networks Vol 1 Issue 1*, 3-16.

Kosko, B. (1988). Bidirectional Associative Memories. *IEEE Transactions on Systems, Man and Cybernetics Vol 18 No 1*, 49-60.

Kulkarni, S., & Haidar, I. (2009). Forecasting model for crude oil price using artificial neural networks and commodity futures prices. *International Journal of Computer Science and Information Security Vol 2 Issue 1*.

Kutsurelis, J. (1998). *Forecasting Financial Markets Using Neural Networks: An Analysis Of Methods And Accuracy, Thesis.* Naval Postgraduate School Monterey: California.

Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Non-experimental Data.* New York: John Wiley & Sons.

Lippman, R. P. (1989). Review of Neural Networks for Speech Recognition. *Neural Computation Vol 1 Issue 1*, 1-38.

Lippman, R. P. (1989). Review of Neural Networks for Speech Recognition. *Neural Computation Vol 1 Issue 1*, 1-38.

Lippmann, R. P. (1987). An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine*, 4-42.

Liu, P., Stevens, R. B., & Vedenov, D. (2018). The physical market and the WTI/Brent price spread. *OPEC Energy Review, Volume 42, Issue 1.*, 55-73.

Luo, Z. H., Xie, Y., & Zhu, C. (1997). The study of convergence of CMAC learning process. *Acta Automatic Sinica 23(4)*, 455-461.

Marsalli, M. (2006). McCulloch-Pitts Neurons. *The 2008 Annual Meeting of the consortium on cognitive science instruction (ccsi)*, 172-179.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1987). The appeal of parallel distributed processing. In D. E. Rumelhart, & J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (old.: 3-44). Cambridge: MIT Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent. *The bulletin of mathematical biophysics 5.*, 115-133. doi:10.1007/BF02478259

McKinney, W. (2011.). pandas: a Foundational Python Library for Data. *Python for high performance and scientific computing, 14(9)*, 1-9.

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press. doi:10.7551/mitpress/11301.001.0001

Moshiri, S., & Foroutan, F. (2006). Forecasting Nonlinear Crude Oil Prices. *Journal of Energy Vol 27*, 81-95.

Mulsant, B. H. (1988). A Connectionist Model for Medical Diagnosis. *Proceedings of Symposium : Artificial Intelligence in Medicine*, 63-64.

Nils Svensson and Others v Retriever Sverige AB. (2014). Forrás: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0466&qid=1718571768593

Numpy documents, version: 1.22. (2022.. január 16.). Forrás: https://numpy.org/doc/stable/

Palva, S., & Palva, J. M. (2012). Discovering oscillatory interaction networks with M/EEG: challenges and breakthroughs. *Trends in cognitive sciences, 16(4)*, 219-230. doi:doi.org/10.1016/j.tics.2012.02.004

Pandas 1.0.0 documentation. (2021.. január 16.). Forrás: https://pandas.pydata.org/: https://pandas.pydata.org/pandas-docs/stable/getting_started/overview.html

Pearson, K. (1895). Notes on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London, 58.*, 240-242. doi:10.1098/rspl.1895.0041

Python Package Index - newspaper3k 0.2.8. (2022.. január 16.). Forrás: https://pypi.org/project/newspaper3k/

PyTorch - Torch.NN documents. (2022.. január 16.). Forrás: https://pytorch.org/docs/stable/nn.html

Rabunal, J. R. (2005). *Artificial Neural Networks in Real-Life Applications.* IGI Global.

Raza, K. (2017). Prediction of Stock Market performance by using machine learning techniques. *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, 1-1. doi:10.1109/ICIEECT.2017.7916583

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. (2016). Forrás: https://eur-lex.europa.eu/eli/reg/2016/679/oj

Rish, I. (2020). *An Introduction to Natural Language Processing (NLP).* Cognitive Class.

Robinson, D. (2017.. szeptember 6.). *The incredible growth of Python.* Forrás: Stack Overflow: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Rojas, R. (1996). *Neural Networks - A Systematic Introduction.* Berlin, New-York: Springer-Verlag.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Review Volume 65. Issue 6.*, 368-408. doi:10.1037/h0042519

Rosenblum, M. G., Pikovsky, A. S., & Kurths, J. (1996). Phase synchronization of chaotic oscillators. *Phys Rev Lett Vol. 76, Issue 11.*, 1804-1807. doi:10.1103/PhysRevLett.76.1804

Rumelhart, D. E., & Zipser, D. (1986). Feature Discovery by Competitive Learning. In D. E. Rumelhart, & J. L. McClelland, *Parallel Distributed Processing : Explorations in the Microstructure of Cognition* (old.: 151-193). Cambridge: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature 323*, 533-536.

Ryanair Ltd v PR Aviation BV. (2015). Forrás: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62014CJ0030

Sadorsky, P. (1999). Oil price shocks and stock market activity. *Energy Economics Vol 21*, 449-469.

Salisu, A. A., Raheem, I. D., & Ndaku, U. B. (2019). A sectoral analysis of asymmetric nexus between oil price and stock returns. *International Economic Review Vol 61*, 241-259.

Samad, T. (1988). Towards Connectionist Rule-Based Systems. *Proceedings of the International Conference on Neural Networks*, 525-532.

Sanchez, M. V. (2011). Welfare effects of rising Oil prices in Oil-importing developing countries. *The Developing Economies Vol 49 Issue 3*, 321-346.

Setiono, R., & Leow, W. K. (2000). FERNN: An algorithm for fast extraction of rules from neural networks. *Applied Intelligence, 12(1–2)*, 15-25.

Simpson, P. K. (1990). *Neural Networks : Research and Applications.* New York: Pergamon Press.

Soumya, C. V., & Ahmed, M. (2017). Artificial neural network based identification and classification of images of Bharatanatya gestures. *Innovative Mechanisms for Industry Applications (ICIMIA)*, 162-166.

Srivastava, N. (2020). *Neural Networks and Deep Learning.* Coursera.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven country data set. *Journal of Foracasting Vol 23 Issue 6*, 405-430.

Stock, J. H., & Watson, M. W. (2006). Forecasting With Many Predictors. *Handbook of Economic*, 516-550.

Strachan, R. W., & Van Dijk, H. K. (2008). Bayesian Averaging over Many Dynamic Model Structures with Evidence on the Great Ratios and Liquidity Trap Risk. *SSRN Electronic Journal*.

Tang, L., & Hammoudeh, S. (2002). An empirical exploration of the world oil price under the target zone model. *Energy Economics Vol 24*, 557-596.

Tang, L., Dai, W., Yu, L., & Wang, S. (2015). A Novel CEEMD-Based EELM Ensemble Learning Paradigm for Crude Oil Price Forecasting. *International Journal of Information Technology & Decision Making Vol 14 Issue 1*, 141-169.

Teräsvirta, T. (2006). Univariate time series models. *Palgrave Handbook of Economics Vol 1*, 396-424.

Terui, N., & Van Dijk, H. K. (2002). Combined forecasts from linear and nonlinear time series models. *International Journal of Forecasting Vol 18*, 421-438.

Tian, H.-Z., & Lai, W.-D. (2019). The causes of stage expansion of WTI/Brent spread. *Petroleum Science, Vol. 16.*, 1493-1505.

Timmermann, A. (2006). Forecast Combinations. *Handbook of Economic Forecasting Vol 1.*, 135-196.

Touretzky, D. S., & Hinton, G. E. (1985). Symbols Among the Neurons: Details of a Connectionist Inference Architecture. *Proceedings of the International Joint Conference on Artificial Intelligence*, 238-243.

Wang, J. Z., Zhu, S. L., Zhang, W. Y., & Lu, H. Y. (2010). Combined modeling for electric load forecasting with adaptive particle swarm optimization. *Energy Vol 35 Issue 4*, 1671-1678.

Wang, M., & Yang, Y. (2020). A comprehensive overview of recurrent neural networks. *arXiv: 2001.09981*.

Wilmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology Vol 32 Issue 13*, 2088-2094.

Wu, G., & Zhang, Y. J. (2014). Does China factor matter? An econometric analysis of international crude oil prices. *Energy Policy Vol 72*, 78-86.

Xie, W., Yu, L., Xu, S., & Wang, S. (2006). A new method for crude oil price forecasting based on support vector machines. *International Conference on Computational Science*, 444-451.

Xin, Y. (1999). Evolving artificial neural networks. *Proceedings of the IEEE, vol. 87, no. 9.*, 1423-1447. doi:10.1109/5.784219

Ye, M., Zyren, J., & Shore, J. (2006). Forecasting short-run crude oil price using high and low-inventory variables. *Energy Policy Vol 34*, 2736-2743.

Yegulalp, S. (2017.). Facebook brings GPU-powered machine learning to Python. *InfoWorld*, Article: 3159120.

Yoon, Y., Brobst, R., Bergstresser, P., & Peterson, L. (1989). A Desktop Neural Network for Dermatology Diagnosis. *Journal of Neural Network Computing Vol 1 Issue 1*, 43-52.

Yoshino, N., & Taghizadeh-Hesary, F. (2014). Monetary policy and oil price fluctuations following the subprime mortgage crisis. *International Journal of Monetary Economics and Finance Vol 7 Issue 3*, 157-174.

Yu, L., Dai, W., & Tang, L. (2016). A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. *Engineering Applications of Artificial Intelligence Vol. 47.*, 110-121. doi:10.1016/j.engappai.2015.04.016

Yu, L., Dai, W., Tang, L., & Wu, J. (2016). A hybrid grid-GA-based LSSVR learning paradigm for crude oil price forecasting. *Neural Computing and Applications Vol 27 Issue 8*, 2193-2215.

Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting Crude Oil Price with an EMD-based Neural Network Ensemble Learning Paradigm. *Energy Economics Vol 30 Issue 5*, 2623-2635.

Yu, L., Wang, Z., & Tang, L. (2015). A decomposition-ensemble model with data-characteristic-driven reconstruction for crude oil price forecasting. *Applied Energy Vol 156*, 251-267.

Yu, L., Xu, H., & Tang, L. (2017). LSSVR ensemble learning with uncertain parameters for crude oil price forecasting. *Applied Soft Computing Vol 56*, 692-701.

Zhang, J. L., Zhang, Y. J., & Zhang, L. (2015). A novel hybrid method for crude oil price forecasting. *Energy Economics Vol 49*, 649-659.

Zhang, Y. J. (2013). Speculative trading and WTI crude oil futures price movement: an empirical analysis. *Applied Energy Vol 107*, 394-402.

Zhang, Y. J., & Wei, Y. M. (2011). The dynamic influence of advanced stock market risk on international crude oil return: an empirical analysis. *Quantitative Finance Vol 11 Issue 7*, 967-978.

Zhang, Y. J., Fan, Y., Tsai, H. T., & Wai, Y. M. (2008). Spillover effect of US dollar exchange rate on oil prices. *Journal of Policy Modeling Vol 30 Issue 6*, 973-991.

Zhao, Y., Li, J., & Yu, L. (2017). A deep learning ensemble approach for crude oil price forecasting. *Energy Economics Vol 66*, 9-16.

M2. Ábrajegyzék

M3. Táblázatjegyzék

M4. Wall street Journal python kód

```
            username="**************"
            password="*************************
            base_url="https://accounts.wsj.com"

            headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:46.0)
Gecko/20100101 Firefox/46.0'}

            session = requests.Session()
            session.headers.update(headers)

            r = session.get("{}/login".format(base_url))
            soup = BeautifulSoup(r.text, "html.parser")
            jscript = [
                t.get("src")
                for t in soup.find_all("script")
                if t.get("src") is not None and "app-min" in t.get("src")
            ]

            credentials_search = re.search("Base64\.decode\('(.*)'", r.text, re.IGNORECASE)
            base64_decoded = base64.b64decode(credentials_search.group(1))
            credentials = json.loads(base64_decoded)

            print("client_id : {}".format(credentials["clientID"]))
            print("state     : {}".format(credentials["internalOptions"]["state"]))
            print("nonce     : {}".format(credentials["internalOptions"]["nonce"]))
            print("scope     : {}".format(credentials["internalOptions"]["scope"]))

            connection = 'DJldap'

            r = session.post(
                'https://sso.accounts.dowjones.com/usernamepassword/login',
                data = {
                    "client_id": credentials["clientID"],
                    "connection": connection,
                    "headers": {
                        "X-REMOTE-USER": username
                    },
                    "nonce": credentials["internalOptions"]["nonce"],
                    "ns": credentials["internalOptions"]["ns"],
                    "password": password,
                    "protocol": "oauth2",
                    "redirect_uri": credentials['callbackURL'],
                    "response_type": "code",
                    "scope": credentials["internalOptions"]["scope"],
                    "state": credentials["internalOptions"]["state"],
                    "tenant": "sso",
                    "ui_locales": credentials["internalOptions"]["ui_locales"],
                    "username": username,
                    "_csrf":  credentials["internalOptions"]["_csrf"],
                    "_intstate":  credentials["internalOptions"]["_intstate"]
                })
            soup = BeautifulSoup(r.text, "html.parser")

            login_result = dict([
                (t.get("name"), t.get("value"))
                for t in soup.find_all('input')
                if t.get("name") is not None
            ])

            r = session.post(
                'https://sso.accounts.dowjones.com/login/callback',
                data = login_result)
```

```
            r = session.get("https://www.wsj.com")
            username_search = str(re.search('\"firstName\":\s*\"(\w+)\",', r.text,
re.IGNORECASE))
            print("connected user : " + username_search)

            f=open("index.html","w+", encoding="utf-8")
            f.write(r.text)
            f.close()

for loop in range(**, **):
       page = session.get(********)
       soup = BeautifulSoup(page.text, 'html.parser')
       time.sleep(2)

url_search_list = []

for a in soup.find_all('a', href=True):
       if "/articles/" in a['href']:
              url_search_list.append(a['href'])

            for article in url_search_list:

                html = session.get(article)

                soup = BeautifulSoup(html.text, 'html.parser')
                text = soup.find("div", { "class" : "article-content" }).text
                text = text.replace("\n", "")
                textend = text.find("Copyright")
                text = text[:textend]

                date = str(year)+"."+str(month)+"."+str(day)

                wb = load_workbook('*******\WSJ_DATA_' + str(year) + '.xlsx')
                ws = wb.active
                ws.append([date, article, text])
                wb.save('*******\WSJ_DATA_' + str(year) + '.xlsx')

            USER_AGENT = 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:78.0)
Gecko/20100101 Firefox/78.0'

            config = Config()
            config.browser_user_agent = USER_AGENT
            config.request_timeout = 300

            toi_article = Article(x, config = config)

            toi_article.download()

            toi_article.parse()
            toi_article.nlp()

            KEYS = str(toi_article.keywords)
            TITLE = toi_article.title
            SUMMARY = toi_article.summary
```

M5. Mesterséges Neurális Háló python kód

```
    oil_price_df = pd.read_excel('D:\\***\oil_price.xlsx', index_col=False)

    oil_price_df_change = oil_price_df['PRICE']

    oil_price_df['CHANGE'] = oil_price_df_change.pct_change()*100
    oil_price_df['DATE'] = oil_price_df['DATE']-timedelta(days=1)
    oil_price_df = oil_price_df.drop(['PRICE'],axis=1)
    oil_price_df = oil_price_df.set_index('DATE')

     for key_word in key_word_list:
        wsj_df[key_word] = wsj_df['SUM'].str.count(key_word)

    wsj_key_words = wsj_df.drop(['SUM'],axis=1)
    wsj_key_words['INDICATE'] = wsj_key_words['opec'] + wsj_key_words['oil price'] +
wsj_key_words['wti'] + wsj_key_words['crude oil']

#TÖRLÉS NÉLKÜLI ELEMZÉS
print("TÖRLÉS NÉLÜLI ELEMZÉS")
wsj_key_words['INDICATE'] = np.where(wsj_key_words['INDICATE'] > (indikatorkor-1), True, False
)
for key_word in key_word_list:
        wsj_key_words.loc[wsj_key_words['INDICATE'] == False, key_word] = 0
wsj_key_words = wsj_key_words.groupby(['DATE'], as_index=False).sum()
wsj_key_words = wsj_key_words.set_index('DATE')

# TÖRLÉSES ELEMZÉS
main_length = len(wsj_key_words)
print(f"DataFrame main length:  {main_length}")
wsj_key_words = wsj_key_words[wsj_key_words['INDICATE']>=indikatorkor]
cutted_length = len(wsj_key_words)
print(f"DataFrame cutted length:  {cutted_length}")
left_acc = f"{(cutted_length/main_length*100):2.3f}%"
print(f"LEFT_ACC:  {left_acc}")
wsj_key_words = wsj_key_words.groupby(['DATE'], as_index=False).sum()
wsj_key_words = wsj_key_words.set_index('DATE')
oil_price_df_change = oil_price_df['PRICE']
oil_price_df['CHANGE'] = oil_price_df_change.pct_change()*100
oil_price_df['DATE'] = oil_price_df['DATE']-timedelta(days=1)
oil_price_df = oil_price_df.drop(['PRICE'],axis=1)
oil_price_df = oil_price_df.set_index('DATE')

df['CHANGE'].fillna(method='ffill', inplace=True)


class Model(nn.Module):

        def __init__(self, in_features=len(key_word_list), h1=120, h2=80, out_features=1):
            super().__init__()
            self.fc1 = nn.Linear(in_features,h1)    # input layer
            self.fc2 = nn.Linear(h1, h2)            # hidden layer
            self.out = nn.Linear(h2, out_features)  # output layer

        def forward(self, x):
            x = F.relu(self.fc1(x))
            x = F.relu(self.fc2(x))
            x = self.out(x)
            return x

    torch.manual_seed(32)
    model = Model()

    X = df.drop(['INDICATE', 'CHANGE'],axis=1)
```

```
y = df['CHANGE']

X = X.values
y = y.values

print(f"MIN Y:  {min(y)}")
print(f"MAX Y:  {max(y)}")

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=33)

X_train = torch.FloatTensor(X_train)
X_test = torch.FloatTensor(X_test)
y_train = torch.FloatTensor(y_train).reshape(len(y_train),1)
y_test = torch.FloatTensor(y_test).reshape(len(y_test),1)

criterion = nn.MSELoss()
optimizer = torch.optim.Adam(model.parameters(),lr=0.001)

epochs = 1800
losses = []

for i in range(epochs):
    i+=1
    y_pred = model(X_train)
    loss = criterion(y_pred,y_train)
    losses.append(loss)
    if i%25==0:
        print(f"epoch {i} and loss is {loss}")

    optimizer.zero_grad()
    loss.backward()
    optimizer.step()

    with torch.no_grad():
        y_eval = model.forward(X_test)
        loss = criterion(y_eval,y_test)
    correct = 0
    calc = 0
    with torch.no_grad():
        for i,data in enumerate(X_test):
            y_val = model.forward(data)
            calc += 1
            if abs(y_val-y_test[i])<=ACC:
                    correct += 1
```

M6. Visszacsatolt Neurális Háló python kód

```
test_size = 12
test_start = len(df)-2000
# Create train and test sets
train_set = y[test_start:-test_size]
test_set = y[-test_size:]

from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(-1, 1))

train_norm = scaler.fit_transform(train_set.reshape(-1, 1))

train_norm = torch.FloatTensor(train_norm).view(-1)

window_size = test_size

def input_data(seq,ws):  # ws is the window size
    out = []
    L = len(seq)
    for i in range(L-ws):
        window = seq[i:i+ws]
        label = seq[i+ws:i+ws+1]
        out.append((window,label))
    return out

train_data = input_data(train_norm,window_size)
print(len(train_data))

class LSTMnetwork(nn.Module):
    def __init__(self,input_size=1,hidden_size=100,output_size=1):
        super().__init__()
        self.hidden_size = hidden_size

        self.lstm = nn.LSTM(input_size,hidden_size)

        self.linear = nn.Linear(hidden_size,output_size)

        self.hidden = (torch.zeros(1,1,self.hidden_size),
                       torch.zeros(1,1,self.hidden_size))

    def forward(self,seq):
        lstm_out, self.hidden = self.lstm(
            seq.view(len(seq),1,-1), self.hidden)
        pred = self.linear(lstm_out.view(len(seq),-1))
        return pred[-1]  # we only want the last value

torch.manual_seed(101)
model = LSTMnetwork()

criterion = nn.MSELoss()
optimizer = torch.optim.Adam(model.parameters(), lr=0.001)

def count_parameters(model):
    params = [p.numel() for p in model.parameters() if p.requires_grad]
    for item in params:
        print(f'{item:>6}')
    print(f'______\n{sum(params):>6}')

count_parameters(model)

epochs = 100

for epoch in range(epochs):
        for seq, y_train in train_data:
```

```
        optimizer.zero_grad()
        model.hidden = (torch.zeros(1,1,model.hidden_size),
                        torch.zeros(1,1,model.hidden_size))

        y_pred = model(seq)

        loss = criterion(y_pred, y_train)
        loss.backward()
        optimizer.step()

    print(f'Epoch: {epoch+1:2} Loss: {loss.item():10.8f}')

future = 12

preds = train_norm[-window_size:].tolist()

model.eval()

for i in range(future):
    seq = torch.FloatTensor(preds[-window_size:])
    with torch.no_grad():
        model.hidden = (torch.zeros(1,1,model.hidden_size),
                        torch.zeros(1,1,model.hidden_size))
        preds.append(model(seq).item())

true_predictions = scaler.inverse_transform(np.array(preds[window_size:]).reshape(-1, 1))

true_predictions = true_predictions.tolist()

main_dates = []
i = 1
while i < 13:
    last_date = last_date + timedelta(1)
    input_date = last_date.strftime("%Y-%m-%d")
    excluded = (6,7)
    if last_date.isoweekday() not in excluded:
        main_dates.append(input_date)
        i += 1
```

M7. spaCy cikk összefoglaló python kód

```
per = 0.05

nlp = spacy.load('en_core_web_sm')
doc= nlp(text)
tokens=[token.text for token in doc]
word_frequencies={}
for word in doc:
    if word.text.lower() not in list(STOP_WORDS):
        if word.text.lower() not in punctuation:
            if word.text not in word_frequencies.keys():
                word_frequencies[word.text] = 1
            else:
                word_frequencies[word.text] += 1
max_frequency=max(word_frequencies.values())
for word in word_frequencies.keys():
    word_frequencies[word]=word_frequencies[word]/max_frequency
sentence_tokens= [sent for sent in doc.sents]
sentence_scores = {}
for sent in sentence_tokens:
    for word in sent:
        if word.text.lower() in word_frequencies.keys():
            if sent not in sentence_scores.keys():
                sentence_scores[sent]=word_frequencies[word.text.lower()]
            else:
                sentence_scores[sent]+=word_frequencies[word.text.lower()]
select_length=int(len(sentence_tokens)*per)
summary=nlargest(select_length, sentence_scores,key=sentence_scores.get)
final_summary=[word.text for word in summary]
summary=''.join(final_summary)
```

M8. Nltk VADER lexicon szentiment analízis python kód

```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer

sid = SentimentIntensityAnalyzer()

not_analyzed = 0
#analyzing the sentiment of the article
for rows in range(len(wsj_df)):
    try:
        sentiment = sid.polarity_scores(wsj_df.iloc[rows].loc['SUM'])

        #checking the sentiment score
        if sentiment['compound'] > 0:
            wsj_df.iloc[rows].loc['SENTIMENT'] = 1
        elif sentiment['compound'] < -0:
            wsj_df.iloc[rows].loc['SENTIMENT'] = -1
        else:
            wsj_df.iloc[rows].loc['SENTIMENT'] = 0
        print(f"{rows+1}   |   {sentiment}")
    except:
        wsj_df.iloc[rows].loc['SENTIMENT'] = 0
        not_analyzed += 1
        print(f"{rows+1}   |   not analyzed")
```