

INFORMATIKA A VERSENYKÉPESSÉGÉRT • ALAPÍTVÁ 1969

A KÉTHETI SZAKLAP

COMPUTERWORLD

2022. november 16. • LIII. évfolyam 22. szám

www.computerworld.hu

TRAINING360

PINTÉR
MIKLÓS
marketingigazgató,
Training360

SZABADON TÉRBE és IDŐBEN

E-LEARNING NELSON MANDELÁTÓL
ARNOLD SCHWARZENEGGERIG

2022. november 16. / LIII. évfolyam 22. szám

Ára: 495 Ft



9 770587 151006 22022



MALLÁSZ
JUDIT

TECHNOLÓGIA

Mélytanulással foglalkozó kutatás az MNB-ben

Közös projekten dolgozik a Magyar Nemzeti Bank és a Műegyetem. Céljuk, hogy a gazdasági hírek szövegéből, mélytanulás-alapú mesterséges intelligencia (MI) módszerekkel, a makrogazdasági mutatók alakulásával kapcsolatos szentiment (vélemény, hangulat) elemzést végezzenek.

Léteznek olyan indexek, amelyek megpróbálják leképezni a gazdasági változók pillanatnyi állapotát (nowcast). Ilyenek például a bizalmi indexek, de több szervezetnek, például az MNB-nek (MNB GDP Tracker), az Európai Bizottságnak vagy az OECD-nek is vannak hasonló célú mutatói. Általános törekvés, hogy ezeket az indexeket nyomon követve, további támpontokkal kiegészítve megpróbálják a gazdaság aktuális állapotát leképezni. Az eddig alkalmazott, többnyire kérdőíves felmérésekkel azonban van némi gond. Ezen eljárások során jellemzően kisfrekvenciás, általában havi gyakoriságú adatok keletkeznek, ehhez járul még a kérdőívek feldolgozási ideje, továbbá – mivel többnyire cégeket kérdeznek meg – fennáll a szelekciós torzítás veszélye is.

– Az MNB-nél – megoldást keresve ezen problémákra – más irányba mozdultunk el. Célul tűztük ki nagyfrekvenciás, algoritmus-alapú, minél inkább torzításmentes index előállítását. Ennek mentén érdemes lehet szöveges híreket felhasználni, hiszen azok jó leképezései lehetnek a gazdaság állapotának, a gazdasági makrováltozók dinamikájának – nyúlunk vissza a kezdetekhez Ónozó Lívia, az MNB Digitalizációs Technológia Főosztályának vezetője.

– A problémát elemezve először az merült fel, hogy csináljunk egy predikciós modellt a GDP idősorra. Ezzel azonban az a probléma, hogy nagyon kevés, negyedévente mindössze egy adat áll rendelkezésre. Ilyen mennyiségű adatból a legfejlettebb MI modellek is jellemzően csupán egyszerű extrapolációt tudnak megtanulni, így ennek igencsak korlátozott a hasznossága. Ezért más adatforrás után kellett tehát nézni. Az MNB azt javasolta, hogy támaszkodjunk a gazdasági hírekre – mutat rá Gyíres-Tóth Bálint, a Műegyetem Távközlési és Média-informatikai Tanszékének (TMIT) docense.

Hírek legálisan

A BME-vel 2021-ben kezdődött az együttműködés ezen projekt kapcsán. A fő megoldandó kérdés az volt, hogy miként tudnának a sajtóban megjelent szöveges adatokból olyan mutatót kinyerni, amely bármilyen módon segítheti a gazdaság, a makrogazdasági változók előrejelzését. A kezdeti munka

fókuszában a megfelelő hírek kiválasztása állt. Olyan hírportálokat kerestek, amelyekben gyakorlatilag minden fontos hír szerepel, illetve amelyeket sokan olvasnak. A kutatásban két nagy hírportál cikkei használják fel.

– Az adatok legális megszerzése még akkor sem triviális, ha ingyenesen elérhető anyagokról van szó. Ha valaki elolvassa az általános felhasználási feltételekre vonatkozó megkötéseket, kiderül, hogy a legtöbb helyen még kutatási célra sem engedélyezett a cikkek felhasználása. Ez állandó probléma, megoldása sokszor nem egyszerű. Számunkra is kihívást jelentett, de végül sikerült megoldani a kérdést – hívja fel a figyelmet mindkét kutató.

Motivációk, elvárások a gépi tanulással szemben

A munka első felét, az előzetes fejlesztéseket főként az adatbeszerzés, az algoritmikus módszerek kidolgozása, a cikkek megfelelő tárolása, olvasható formátumra alakítása tette ki. – Különböző motivációink voltak. Azt szerettük volna, hogy nagyfrekvenciás, algoritmus-alapú, fordításoktól minél kevésbé terhelt indexet állítsunk elő. Így születtek meg az algoritmikus módszerek – fogalmaz Ónozó Lívia.

– A továbbiaknál figyelembe kellett venni, hogy a GDP becslésekor – a közgazdaságtanban bevett formulákon túl – mindig van egy tapasztalati, szakértői rész, ami nem kezelhető teljes mértékben analitikusan. Úgy is fogalmazhatunk, hogy jelenleg a gépi tanulás az emberben van. Az ember, tapasztalatai és tudása alapján, meghozza azokat a döntéseket, amiket majd – pusztán a módszertan segítségével – a gépi tanulástól várnánk. Egy egyszerű, hatékony és jól magyarázható módszer a kulcsszóelosztás elemzése, mely során azt vizsgáljuk, hogy adott szavak hányszor jelennek meg a cikkekben, például egy napra aggregálva – teszi hozzá Gyíres-Tóth Bálint.





Tanítás magyar szövegre

– Az volt az ötletünk, hogy meglévő gazdasági szentiment szótárakhoz társítsunk bizonyos szentiment értékeket, majd ezek alapján tegyünk előrejelzéseket. Ha például egy létező személy olvasna el egy cikket, és abban sokszor szerepelnek olyan szavak, amelyekhez negatív érzelmeket társít, akkor annak a cikknek lesz egy negatív szentimentje. Ez a folyamat az úgynevezett vélemény- vagy hangulatbányászat, angolul sentiment analysis vagy sentiment prediction – emeli ki az MNB főosztályvezetője.

A folyamat első részében történik a szövegek részegységekre bontása, például a szótövezés, aztán a stopszó szűrés, azaz a nem releváns szavak kiszűrése.

– A magyar szöveg nagyon különbözik az angolszász vagy latin szövegektől. Ez utóbbiakra már sok beépített, használható könyvtár létezik, a magyar szövegekre rendelkezésre álló forrásoknak azonban korlátozott a felhasználhatósága. Számunkra az egyik nagy kihívás, hogy miként tanítunk valamit magyar szövegre. Alapvető statisztikai módszerek alapján azt láttuk, hogy a mi algoritmus-alapú megoldásaink, például a kulcsszó-alapú algoritmus szépen korrelál akár az üzleti indexekkel, akár a gazdasági mutatószámokkal. Mindazonáltal szeretnénk volna mélyebb, nem csak algoritmus-alapú kutatást végezni. Egyrészt célunk volt a folyamat automatizálása, másrészt meg akartuk tanítani a gépünket arra, hogy kontextusok és más szempontok alapján ő maga pontosabb szentiment predikciót adjon. Nagyjából ezen a ponton kezdtünk el a TMIT-tel együtt dolgozni – tekint vissza a múltba Ónozó Lívia.

Korreláció a GDP és a szentiment között

– Az ilyen jellegű együttműködést úgy kezdjük, hogy végignézzük és egységesítjük az egész modellezési folyamatot. Itt arra gondolok, hogy amikor különböző algoritmusokat dolgozunk ki, arra törekszünk, hogy azok tényleg egyértelműen összehasonlíthatóak legyenek. Esetünkben viszonylag soklépcsős az eljárás. Bejönnek a cikkek, azokat előválogatjuk, ott is van egyfajta gépi tanuló algoritmus, aztán az előválogatott cikkeken szentiment prognosztizálunk, majd a szentiment előrejelzést illesztjük a gazdasági mutatókhoz, egyelőre a GDP-hez, illetve most éppen a BMI-hez (Beszerzési Menedzser Index). Kutatásaink alapján Magyarországon a GDP és a gazdasági hírek szentimentje között van kimutatható korreláció – mondja a BME TMIT docense, majd áttér az MNB-vel folyó közös munka egy másik aspektusára.

– A szentiment modellek javítása és a legújabb MI technológiák bevezetése a másik irány. A neurális

hálózatok a természetes nyelvfeldolgozásban (Natural Language Processing, NLP) bizonyítottan nagyon hatékonyak. Munkánk során elsődlegesen többnyelvű megoldásokkal dolgozunk, így lehetséges többnyelvű adatbázissal együttesen betanítani a neurális szentiment modelleket. Az eredmény egyelőre az MNB berkein belül marad, mivel – megfelelő jogi háttér hiányában – az adatok átadása az MNB és a BME között sem történhet meg.

Fontos feladat a szentiment adatbázis címkézése, azaz az egyes tartalmakhoz pozitív, negatív vagy neutrális címke hozzárendelése. A TMIT többkon-szenzusos adatbázis kialakítását javasolta, tehát ne csak egy ember címkézze fel az adatbázist, hanem többen végezzék el ezt a munkát. Ezzel meg lehet állapítani, hogy mi az az emberi pontosság, aminél jobbat várhatóan az MI-vel sem lehet elérni. A címkézés során bebizonyosodott, hogy a gép is olyan helyeken téveszt, ahol az emberek is nehezen egyeznek meg egymással.

Egyszerűsödik az előrejelzés

Szentiment predikcióval, szentiment elemzéssel sokan foglalkoznak, sok adatbázis létezik. Fontos azonban kiemelni, hogy minden témakör új kihívásokat jelent. Amíg viszonylag egyszerű szentiment előrejelzést végezni például egy webboltot érő termékkritikákra, addig a gazdasági hírekre nem igaz ugyanez. Ezért is használták ki az MNB–BME közös projektben a többnyelvűséget, a többnyelvű szentiment predikciót, mert ilyenkor az angol adatbázisokat is fel tudják használni a tanításhoz. Az angolul fellelhető gazdasági hírek adatbázisát is bevették a kísérletekbe, amelyekben együtt tanítják az angol és a magyar szentiment adatbázist. Így sokszoros adatmennyiséggel tudják a rendszert tanítani.

– Jelenleg ott tartunk, hogy a mély neurálishálózat-alapú megoldás pontosabban tudja az adatbázis szerint a szentimentet előrejelezni, viszont az idősorhoz való illesztés, tehát az, hogy a GDP-t mennyire lehet a szentimentből megmondani, egyelőre a kicsit kevésbé pontos szótár alapon működik jobban. Amikor technológiai oldalról minden kérdést lezártunk, utána tudunk szélesebb körben egyeztetni, hogy a gazdasági híreknél ténylegesen milyen feltételekkel, milyen eredmények érhetők el – tájékoztat a Műegyetem docense.

A mélytanulás-alapú természetes nyelvfeldolgozással végzett szentiment elemzés viszonylag új terület az MNB-nél. – Nagyon pragmatikusan nézve, a projekt kimenete lehet egy idősor, ami megmutatja, hogy jelenleg a gazdasági hírek szentimentje milyen. Hosszú távon ezek az adatok bevonhatók a makrogazdasági prognózisokba, akár érdemben támogatva szakmai szinten az előrejelzési feladatot – mutat rá Ónozó Lívia. **CW**