Katona Eszter: RC2S2, ELTE TáTK

Fazekas Mihály: CEU

Kísérlet a Corruption Risk Index továbbfejlesztésére a természetes nyelvfeldolgozás segítségével.

A közbeszerzések a GDP 15 százalékát, valamint a kormányzati kiadások mintegy egyharmadát teszik ki. Európa-szerte elterjedtek a korrupcióval kapcsolatos vádak, ennek ellenére keveset tudunk a korrupcióról, és arról, hogy mi mozgatja azt. Az utóbbi években robbanásszerűen megnőtt a hivatalos, kormányzati nyilvántartáson alapuló, pályázati felhívás szintű adatok elérhetősége. Ez új lehetőséget nyitott a korrupció és a korlátozott verseny tanulmányozására, szövegbányászati módszerek alkalmazásával. Kutatásunk célja éppen ezért a nyílt verseny korlátozásának előrejelzése a közbeszerzési pályázatok tender szintű szöveges információinak felhasználásával.

Szövegbányászati megközelítésünk meghatározó lehet a jelenleg alkalmazott korrupciókockázati mutatók mellett (ilyenek például, ha a pályázati felhívást nem teszik közzé a potenciális ajánlattevők számára, vagy ha egy állami tendert közvetlenül nyílt verseny nélkül ítélnek oda). A megközelítésünk a fenti indikátorokra épít, de túlmutat azokon: a pályázati felhívásoknak azokból a szöveges részeiből indul ki, melyek leírják a beszerezni kívánt árukat és szolgáltatásokat, valamint az ajánlattevőkre vonatkozó feltételeket (például a megkövetelt előzetes tapasztalatokat). Korábbi kvalitatív esettanulmányok azt találták, hogy a pályázati feltételeket sokszor egy előnyben részesített ajánlattevőre szabják, hogy ezáltal kizárják a versenytársakat a pályázatból. Ez gyakran a pályázati feltételekbe rejtett, bonyolult követelmények révén történik. Elemzésünk tehát a korábbi kvalitatív kutatások eredményeire épít, és a gépi tanulás módszereit alkalmazza adminisztratív adatokon. Magyar közbeszerzési szerződéseket elemzünk a 2011 és 2020 közötti időszakból. Először – alapmodellként – a szöveges tartalom felhasználása nélkül replikáltuk a korábbi kutatásokat, melyek célja a verseny korlátozásának előrejelzése, azaz az (egyébként versenyképes tenderre) egyetlen benyújtott ajánlat prediktálása. Az, ha egy felhívásra egyetlen pályázat érkezik, korrupciós szándékot jelezhet. Ezután logisztikus regressziót és Random Forest modelleket illesztettünk n-gramok felhasználásával ugyanúgy az egy ajánlattevő, mint korrupciós kockázat megjóslására. Grid search segítségével kerestük az optimális hiperparaméter-beállításokat. A szöveges adatokon felül kontrollváltozókat is bevontunk a modellekbe (az évet, a termékkódon alapuló felosztást, az ajánlati árat, a helyszínt és a vevőtípust).

Előzetes eredményeink azt mutatják, hogy a szöveges adatokat használó modellek felülmúlják az alapmodelleket az egy ajánlattevős pályázatok előrejelzésében. Kíváncsiak voltunk rá, hogy mely szövegrészek a legfontosabbak a korrupciós kockázat mérése szempontjából, így a szöveg három különböző részét külön-külön használtuk. Külön modellt illesztettünk a szövegek azon részeire, amelyek a részvételi feltételeket, az értékelési szempontokat és a termékleírást tartalmazzák. Azt találtuk, hogy a különböző szövegrészek eltérő hatékonysággal működnek.