TrendMiner: politikai témájú Facebook-üzenetek feldolgozása és szociálpszichológiai elemzése

Miháltz Márton[1], Váradi Tamás[1]

[1] MTA Nyelvtudományi Intézet Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály,
Nyelvtechnológiai Kutatócsoport
1068 Budapest, Benczúr utca 33.
{mihaltz.marton, varadi.tamas}@nytud.mta.hu

Kivonat

Az előadásban bemutatjuk a *Trendminer* projekt[1] módszereit és eredményeit. Az FP7-es finanszírozású, európai kooperációban megvalósuló munkálatok célja közösségi média stream-ek valós idejű figyelése, trendkövetése és összegzése. Ezen belül az MTA NYTI által fejlesztett, MTA TTK KPI Narratív Pszichológiai Kutatócsoportjának munkatársaival együttműködésben készített eszközök célkitűzése a magyar Facebook-felhasználók politikai tartalmú posztokra adott nyilvános kommentjeinek nagymennyiségű gyűjtése, elemzése és szociálpszichológiai aspektusokból történő automatikus kiértékelése volt. Ezzel a kutatással reményeink szerint támogatást nyújtunk annak vizsgálatához, hogy milyen érzelmi és társas pszichológiai jelenségek, trendek figyelhetők meg a magyar közösségimédia-hozzászólók politikai témákra reagáló üzeneteiben.

A projektben a Facebook Graph API segítségével 12 hónap alatt mintegy 140 ezer publikus posztot és az ezekre érkezett mintegy 2 millió publikus kommentet gyűjtöttünk össze több mint 1300 Facebook-oldalról, melyek Magyarországon bejegyzett politikai pártokhoz, azok tagszervezeihez, képviselőihez és –jelöltjeihez kötődnek. Figyelmet fordítottunk a 2014-es évben lezajlott országgyűlési és európai parlamenti választások jelöltjeinek és tisztséget elnyerő képviselőinek oldalain megjelent üzenetek gyűjtésére is.

A begyűjtött üzeneteket adatbázisban tárolás után az alábbi pipeline segítségével dolgoztuk fel: mondatszegmentálás, tokenizálás (*huntoken* eszköz), morfológiai elemzés és szófaji egyértelműsítés (*hunpos* és *hunmorph* eszközök), szótő és morfológiai elemzés egyértelműsítése (saját eszköz). Ezt követte a domain számára releváns entitások azonosítása, illetve a tartalomelemzés a *NooJ* eszköz nyílt forrású változatával, melyhez parancssori változatot készítettünk[2]. A tartalomelemzés az üzenetek érzelmi polaritásának (sentiment) vizsgálatán túl további 5 pszichológiai dimenziót érintett [2] (elsődleges-másodlagos gondolkodási szint, közösségiség-ágencia, optimizmus-pesszimizmus, individualizmus-kollektivizmus), melyek ilyen célú felhasználására tudomásunk szerint ez az első példa.

1 http://www.trendminer-project.eu
2 https://bitbucket.org/tkb-/nooj-cmd

Mivel az általunk alkalmazott *hun** nyelvi feldolgozó eszközöket[3] a sztenderd nyelvváltozatot reprezentáló, többnyire híroldalak anyagát tartalmazó szövegek felhasználásával fejlesztették ki, teljesítményük elmaradt a várt szinttől a projektben vizsgált közösségimédia-domain speciális nyelvezetű szövegein. Emiatt szükség volt az eszközök vizsgált nyelvterülethez való adaptációjára. Ehhez készítettünk egy 1,25 millió Facebook üzenetet (29M token) tartalmazó korpuszt és az ismeretlen szavak gyakorisági listáján 15-ször vagy annál gyakrabban szereplő tételeket vizsgáltuk manuálisan. Ennek segítségével azonosítottuk az ismétlődően előforduló reguláris problémákat, valamint a gyakori és releváns ismeretlen szavakat, rövidítéseket, szleng kifejezéseket, emotikonokat stb. Ezek alapján a tokenizáló eszközt elő- és utófeldolgozó szintekkel egészítettük ki a gyakori írásjel-használati és egyéb reguláris problémák kezelésére. Az ismeretlen szavak és rövidítések egy részét sztenderd nyelvi (az elemzők számára ismert) alakra javítottuk a szövegben, másik részüket pedig analóg ismert szavak felhasználásával hozzáadtuk a morfológiai elemző szótárának megfelelő paradigmáihoz annak érdekében, hogy ezután azok különböző inflexiós alakjaikat is fel tudjuk ismertetni.

Az üzenetekben szereplő névelemek azonosítására a *hunNER* statisztikai gépi tanuló eszközzel [3] – a vizsgált nyelvi domain sajátosságai miatt – szerzett kedvezőtlen tapasztalatok után lexikon alapú felismerőt fejlesztettünk a *Java NooJ* eszköz segítségével. Az 5 szociálpszichológiai dimenzió nyelvi kifejezéseinek felismerését – a szótő, szófaj és morfológiai annotáció eredményeire támaszkodva – szintén *NooJ* nyelvtanok fejlesztése segítségével végeztük el. Az érzelmi valencia és a közösségiség-ágencia dimenziók lexikonjainak fejlesztői felhasználtak egy 176K kommentből (4.9M token) álló korpuszt, melynek 100-szor vagy gyakrabban szereplő kifejezéseit 7 független humán annotátor kódolta a megfelelő kategóriákra.

Az eredmények kiértékelésére három, az egyes pártok oldalain érkezett hozzászólásokat a teljes 2M szavas korpusz arányában tartalmazó gold standard korpuszt állítottunk össze, melyek 337, 336 és 672 kommentet tartalmaztak (7.6K, 7.5K, 17.9K token). A korpuszokat egy egyszerű algoritmussal tagmondatokra bontottuk, majd ezeket 3-3 független annotátor látta el jelölésekkel a különböző annotációs kategóriákra. Az egyes modulok annotációs pontossága (precision) a gold standardhoz képest 65.75%-98.36% között változott, fedése (recall) 13.8%-82.05% között. Az egyes kommentek érzelmi (sentiment) polaritásának felismerési pontossága (accuracy) 84.63% volt [1].

A projekt során elkészítettük a magyar politikai élet vizsgált időszakban releváns entitásait, fogalmait és azok tulajdonságait modellező, OWL nyelven írt politikai ontológiát is. Az ontológiát felhasználtuk többek között a szövegekben azonosított entitások linkelésére a projekt számára készült, időszakok, kulcsszavak, entitások, együtt-előfordulások, érzelmi polaritás grafikonok stb. megjelenítésére alkalmas vizualizációs felületben is.

A projekt zárultával szabadon hozzáférhetővé tettük az előállított eszközök és erőforrások egy részét: a *hun** eszközök domain-adaptációját megvalósító kiegészítéseink és a Java NooJ eszköz parancssori változatának forráskódját, a

3 http://mokk.bme.hu/en/eszkozok/

politikai ontológiát, valamint a projektben vizsgált 2M kommentet (az összes metaadattal, NLP- és tartalomelemzési annotációival együtt) tartalmazó korpuszt is[4].

Hivatkozások

1. Miháltz, M.: Socio-psychological Analysis of Social Media Messages in Politics. In: J. L. Martínez Fernández, P. Martínez, M. Ogrodniczuk, M. Miháltz: Newly generated domain-specific language data and tools. TrendMiner Project Public Deliverable D10.1., 46-63 (2014) http://www.trendminer-project.eu/images/d10.1_final_version.pdf
2. Pólya T., Csertő I., Fülöp É., Kővágó P., Miháltz M., Váradi T.: A véleményváltozás azonosítása politikai témájú közösségi médiában megjelenő szövegekben. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (2015), ld. jelen kötetben.
3. Simon, E.: Approaches to Hungarian Named Entity Recognition. PhD dissertation. Budapest University of Technology and Economics, Budapest (2013).

4 Letöltések, további információ: http://corpus.nytud.hu/trendminer/