View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by SZTE Doktori Értekezések Repozitórium (SZTE Repository of Dissertations)

Célorientált szentimentelemzés különböző műfajú szövegeken

Hangya Viktor

Témavezető: Dr. Farkas Richárd

Ph.D. értekezés tézisei

Szegedi Tudományegyetem
Informatika Doktori Iskola

2019. június

1. Bevezetés

A Web 2.0 kialakulásának köszönhetően nagy mennyiségű felhasználói adat keletkezik nap mint nap. Ennek az adatnak egy jelentős hányada szöveges formátumú, melyek a szerzők véleményeit és érzelmeit tartalmazzák. Különböző szervezetek kezdték alkalmazásaikban felhasználni ezen adatokat, mivel segítségükkel bepillantást nyerhetünk adott entitásokról alkotott véleményekbe. Ilyen alkalmazások például a termékek és márkák megítélésének követése, választások eredményének megbecslése vagy akár katasztrófahelyzetek könnyebb irányítása. Az adatok kézi feldolgozása azonban azok nagy mennyisége miatt szinte lehetetlen, ezért automatikus módszerekre van szükség.

A szentimentelemzés feladata tehát, hogy adott szövegek érzelmi (szentiment) polaritását meghatározza, mely a legtöbb alkalmazás esetében a pozitív, negatív vagy semleges polaritások egyike. A legtöbb módszer célja a szövegek globális szentimentjének meghatározása, annak ellenére, hogy számos esetben inkább adott entitások különböző tulajdonságaival kapcsolatban osztjuk meg véleményünket. Ezen felül sokszor különböző entitásokat hasonlítunk össze egy mondaton belül, mely esetben ugyanaz a véleménykifejezés pozitív lehet az egyik entitásra nézve, míg negatív a másikra. Ebből kifolyólag a globális *dokumentumszintű* elemzés nem megfelelő, a kidolgozott rendszereknek meg kell különböztetnie az egyes célentitásokra vonatkozó érzelmeket. A *célorientált* szentimentelemzés célja, hogy adott szöveg és célentitás páros esetében meghatározza a célra vonatkozó szentimentek polaritását. A disszertáció első fele foglalkozik ezzel a feladattal, melyben bemutatjuk a feladat nehézségeit a hagyományos módszerekhez képest, valamint ismertetjük kifejlesztett módszereinket a szövegek releváns tartalmának és azok szentimentjének pontosabb meghatározására.

A szentimentelemzés és általánosságban véve a természetesnyelv-feldolgozás egyik sokat kutatott területe a különböző műfajokból és doménekből származó szövegek kezelése. A közösségi médiából származó szövegek számos stílust és nyelvet ölelnek fel, ami nagy nyelvhasználati változatosságot jelent, megnehezítve ezzel széles körben is használható rendszerek létrehozását. Vegyük például a blogokat és újságcikkeket, melyek a közösségi médiáról származó szövegek közül az egyik legsztenderdebbnek tekinthetőek. Helyesírásilag magas színvonalúak, valamint a szövegek hosszát tekintve legtöbbször nincs megkötés, ezért az adott témák részletesen ki vannak fejtve. Ezzel szemben a mikroblogokról, mint például Twitterről származó szövegek hossza korlátozott, ami tartalmuk tömörségét eredményezi. Ezen felül mivel az ilyen jellegű médiumok gyors üzenetváltásokra használatosak, a szövegek helyesírásilag rosszabb

minőségűek és számos szleng szót tartalmaznak. A disszertáció második része a szövegek doménbeli, műfajbeli és nyelvi különbségei által okozott problémákkal foglalkozik, melyek a szentimentelemzés során felmerülhetnek. Bemutatjuk a szentimentlexikonok doménadaptációjára kidolgozott módszereinket, melyek fontos alkotóelemei az elemzőrendszereknek. Ezen felül kétnyelvű rendszerek adaptációját célzó módszereket is kifejlesztettünk, melyek erőforrásszegény nyelvek esetében különösen fontosak.

2. A disszertáció felépítése

A disszertáció célja az aktuális szentimentelemző technikák határainak kiterjesztése volt. Különböző természetesnyelv-feldolgozó technológiákat ismertettünk kezdve a szentimentspecifikus előfeldolgozási lépésekkel, jellemzőkinyerési technikákon át egészen a szentimentelemzés különböző szintjeiig. A legtöbb bemutatott módszer a hagyományosabbnak tekinthető jellemzőkinyerés alapú lineáris osztályozókra épül, de neurális hálózat alapú technikákkal is találkozunk.

Ahogy az fentebb is említésre került, a disszertáció két fő témát dolgoz fel, 4 tézisponton és egy összehasonlító jellegű fejezeten keresztül. Az első téma a célorientált szentimentelemzéssel, míg a második a doménbeli, műfajbeli és nyelvi különbségekkel foglalkozik. Az alábbiakban a disszertáció felépítésére és a fejezetek rövid bemutatására, valamint a szerző fontosabb publikációinak ismertetésére kerül sor. A fejezetek részletesebb leírása és az elért eredmények ismertetése a következő fejezetekben olvasható.

A 3. fejezetben bemutatjuk a célorientált szentimentelemzést. A szövegek adott célra vonatkozó részeinek detektálásához felszíni jellemzőkön és szintaxison alapuló módszereket fejlesztettünk. A kidolgozott módszerek magas hatékonyságát különböző adatbázisokon igazoljuk. Továbbá, a kidolgozott rendszerek nemzetközi versenyeken is előkelő helyezést értek el.

A legtöbb szentimentelemző rendszer szövegenként egy szentimentpolaritást határoz meg, függetlenül attól, hogy dokumentumszintű vagy célorientált elemzést végez-e. A 4. fejezet részletes elemzéssel foglalkozik, mely során a mondatok minden konstituenséhez egy szentimentértéket rendelünk. Más részletes elemzőrendszerek nehezen hozzáférhető erőforrásokon alapulnak, ami megnehezíti azok széles körű alkalmazását. A probléma javítása érdekében egy félig felügyelt módszert dolgoztunk ki, mely az egyes konstituensek szentimentértékét rejtett változóként kezelve képes az eredmények javítására. A módszer pontosságát dokumentumszintű és célorientált feladatokon illusztráljuk.

A disszertáció második témakörét az 5. fejezetben vezetjük be, mely a szövegek műfajbeli, doménbeli és nyelvi különbségeinek szentimentelemzés során felmerülő problémáival foglalkozik. A fejezetben ezen különbségeket ismertetjük, megalapozván a disszertáció utolsó két fejezetét.

A 6. fejezet szentimentlexikonok különböző doménekben való használatának nehézségeit mutatja be. Bizonyos szavak szentimentpolaritása ellentétes lehet más-más doménekben, mely megnehezíti a lexikonok használatát. A probléma kiküszöbölésére olyan módszereket fejlesztettünk ki, amelyek jelölt és jelöletlen nyelvi erőforrásokra támaszkodva képesek lexikonokat létrehozni, illetve adaptálni. A módszerek eredményeit magyar nyelvű adatbázisokon mutatjuk be.

Végezetül, a 7. fejezet keresztnyelvi szentimentelemzés doménkülönbségek okozta problémáival foglalkozik. Bemutatjuk a kidolgozott félig felügyelt doménadaptációs módszerünket, mely csak jelöletlen adatokra támaszkodik. Az elvégzett kísérletek alátámasztják, hogy a feladat során jó eredmények eléréséhez mind általános doménekből származó tudásra, mind pedig doménspecifikus információra szükség van.

Az 1. táblázatban találhatók a szerző disszertációjához kapcsolódó publikációi.

			Fejezetek				
			3	4	5	6	7
CLEF	2013	(Hangya és Farkas, 2013a)	•				
CogInfoCom	2013	(Hangya és Farkas, 2013b)	•				
ICCIA	2017	(Hangya és mtsai., 2017)		•			
SemEval	2013	(Hangya és mtsai., 2013)			•		
AIRE	2017	(Hangya és Farkas, 2017)	•		•		
TSD	2015	(Hangya, 2015)				•	
ACL	2018	(Hangya és mtsai., 2018)					•

1. táblázat. A disszertáció fejezetei és a hivatkozott saját publikációk közötti kapcsolat. A 3., 5. és 6. fejezetekben a szerző önálló hozzájárulása volt a meghatározó, míg a 4. és 7. fejezetek más kutatókkal való együttműködés eredményei.

3. Célorientált szentimentelemzés

A 3. fejezetben a célorientált szentimentelemzést ismertettük, mely esetében a globális érzelmek felderítése helyett adott célra vonatkozó szentimentek osztályozása a feladat. A módszert számos szervezet alkalmazza, mint például termékek gyártója vagy ismert személyiségek annak érdekében, hogy a felhasználóik vagy követőik véleményét könnyebben megismerhessék. A feladat megoldása során számos nehézséggel kell szembenézni, mint például a kérdéses célra vonatkozó szövegrészek detektálása

több célt tartalmazó mondatok esetében.

A szerző egy versenyképes célorientált szentimentelemző rendszert dolgozott ki, melyet különböző stílusú és nyelvű szövegeken értékelt ki. Annak érdekében, hogy magas pontosságot érjen el tweeteken, egy Twitter-specifikus előfeldolgozó módszert mutatott be. Különböző célorientált jellemzőket dolgozott ki, melyekkel dokumentumszintű rendszerek célorientáltá tehetők. A szövegek releváns részeinek detektálásához felszíni jellemzőkön és szintaxison alapuló módszereket hozott létre. Felszíni jellemzőkön alapuló módszerek például a szövegeket leíró jellemzők fontosságának súlyozása a célkifejezés mondatban elfoglalt pozíciója alapján. Ezen felül további, a lehetséges célentitásokkal, illetve témakörökkel kapcsolatos jellemzőket is bemutattunk. Ezzel szemben, a szövegek szintaxisán alapuló módszerek a releváns információk előtérbe helyezését célozzák konstituens és függőségi fák segítségével.

A rendszer eredményeit Twitterről származó és sztenderdebb angol és magyar nyelvű szövegeken illusztráljuk. Az eredmények megmutatták, hogy a felszíni jellemzőkön alapuló módszerek zajosabb szövegeken, mint például tweeteken teljesítenek jobban, míg a szintaxisalapú módszerek jól strukturált mondatokon. Továbbá, a legjobb eredményt a két módszer kombinációja adta. Ezen felül nemzetközi versenyfeladatokra fejlesztett specifikus rendszereket is bemutattunk, melyek a legjobb rendszerek között szerepeltek. A RepLab 2013 versenyre kidolgozott rendszer első helyezést ért el. Korábbi publikációi (Hangya és Farkas, 2013a,b, 2017) alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- Célorientált jellemzők kidolgozása felszíni jegyek alapján kevésbé jól formált szövegek esetében a releváns tartalom hangsúlyozása érdekében;
- Szintaxisalapú célorientált jellemzők kidolgozása jól formált szövegek esetében dependencia és konstituens elemzők felhasználásával;
- Hibrid rendszer kidolgozása a fent említett technikák kombinálásával és a rendszer hibáinak elemzése különböző adatbázisokon.

4. Aprólékos szentimentelemzés

A 4. fejezetben az előbbiekhez képest aprólékosabb szentimentelemzést mutatunk be. A legtöbb elemzőrendszer csak egy szentimentpolaritást határoz meg egy adott szöveg vagy dokumentum tartalma alapján, függetlenül attól, hogy a feladat dokumentumszintű vagy célorientált. Az elemzés aprólékosságát tekintve a következő lépés a

frázisszintű elemzés, mely során a mondatok minden szavához, frázisához és tagmondatához, valamint az egész mondathoz egy-egy szentiment értéket rendelünk alulról felfelé haladva. Az ilyen jellegű elemzésnek több előnye is van. Először is azzal, hogy a mondatok egyes alkotóelemeinek szentimentértékét, valamint azok kapcsolatát is meghatározzuk – melyek lehetnek például összehasonlítóak vagy fokozók – a rendszer képes a mondatok tartalmának jobb megértésére és ezáltal a dokumentumszintű szentimentérték pontosabb meghatározására. Továbbá, az ilyen aprólékos elemzés eredménye felhasználható más feladatokhoz, mint például célorientált szentimentelemzéshez azzal, hogy jellemzőket nyerünk ki belőle.

Korábbi munkák megmutatták, hogy az aprólékos elemzés javítja a dokumentumszintű szentimentelemzés minőségét, azonban ezeknek a módszereknek a működéséhez vagy nehezen előállítható adatokra van szükség, melyek csak bizonyos nyelvek és domének esetében érhetőek el, vagy nyelvfüggő szabályokra épülnek, melyek nehezen adaptálhatóak más felhasználási esetekhez. A szerző és munkatársai egy rejtett szintaktikai struktúrán alapuló módszert dolgoztak ki, mely széles körben alkalmazható, mivel csak mondatszinten annotált szövegekre épül. A rendszer tanítása során csak a teljes mondatok szentimentpolaritása adott, míg a mondatok egyes elemeinek szentimentértéke rejtett változóként van jelen, melyek valószínűségi eloszlása egy iteratív folyamat során számítódik ki. A módszer általános jellemzőkészletre épül, melyek az egyes frázisok kapcsolatának és szentimentjének meghatározását hivatottak elősegíteni. A jellemzők általánosságának köszönhetően a kidolgozott módszer széles körben alkalmazható.

Kísérleteken keresztül megmutattuk, hogy a kidolgozott módszer javítja mind a dokumentumszintű, mind a célorientált osztályozók pontosságát. Annak ellenére, hogy más aprólékosan annotált adatokat használó rendszerek jobban teljesítenek a megfelelő adatok rendelkezésre állása esetében, az itt kidolgozott módszer jobb eredményt ért el, amikor csak mondatszintű jelölés érhető el, ami egyébként az esetek többségére igaz. Ezen felül a rendszer által generált szentimentfákat felhasználtuk célorientált elemzőrendszerekben is a fákból kinyert jellemzők segítségével, melyek tovább növelték a célorientált rendszer minőségét. Korábbi publikációja (Hangya és mtsai., 2017) alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- Rejtett szintaktikai struktúrán alapuló aprólékos szentimentelemző rendszer bevezetése csak mondatszinten annotált szövegek felhasználásával;

- Dokumentumszintű szentimentelemzés pontosságának javítása mondatok összetevős elemzésével;
- Az aprólékos szentimentelemző rendszer integrálása célorientált elemzésbe szentimentfa alapú jellemzők kidolgozásával.

5. Különböző stílusú szövegek szentimentelemzése

Az 5. fejezet a disszertáció második fő témakörét, a doménkülönbségeket vezeti be, mely általánosságban véve a gépi tanulás egy nagy feladata. A probléma lényege, hogy egy adott doménbe sorolt szöveghalmazon betanított rendszer nem teljesít jól más domének esetében. A szerző egy dokumentumszintű szentimentelemző rendszer működését hasonlította össze angol és magyar nyelvű szövegeken, melyek különböző doménekből és műfajokból származnak. A következő fejezeteket megalapozva, a szerző rámutatott a különböző doménekből származó szövegek használata esetén előforduló főbb problémákra.

Az 5. fejezet nem tartozik a disszertáció fő tézispontjai közé. Korábbi publikációi (Hangya és mtsai., 2013; Hangya és Farkas, 2017) alapján a szerző a legfontosabb saját eredményeinek a következőket tekinti:

- Dokumentumszintű szentimentelemzés kvantitatív és kvalitatív összehasonlítása különböző doméneken, műfajokon és nyelveken.

6. Doménspecifikus szentimentlexikonok

A 6. fejezet középpontjában szentimentlexikonokat létrehozó és doménadaptáló módszerek állnak, mely lexikonok kulcsfontosságú információforrások szentimentelemzéshez. Ezek a lexikonok szavakat és azok szentimentértékét tartalmazzák, melyek külső nyelvi erőforrásként alkalmazhatók szövegek jellemzőkinyeréséhez szentimentelemző rendszerekben. Ahogy az az előző fejezetekben elhangzik, egyes kifejezések más-más szentimentpolaritással rendelkezhetnek különböző doménekben. Vegyük például a *hangos* szót, mely pozitív jelentéssel bír hangszórók leírása esetén, de negatívval, ha konyhai eszközökről fejezzük ki véleményünket. Mivel ilyen lexikonok előállítása időigényes és drága folyamat, a legtöbb rendelkezésre álló lexikon általános célú, melyekből hiányoznak az adott doménekre specifikus információk.

A szerző megmutatta, hogy még a jó minőségű lexikonok is ronthatják a rendszerek pontosságát, ha nem megfelelő domén esetében alkalmazzuk őket. A problémára

olyan automatikus módszereket mutattunk be, amelyek doménspecifikus lexikonok előállítására, illetve adaptációjára képesek. A módszerek egy része kisméretű alaplexikonok kiegészítését teszi lehetővé ismert szentimentértékek propagálásával szavak közti lexikai kapcsolatokon keresztül. Továbbá, egy annotált adatokra épülő módszert is bemutattunk, mely új lexikonok létrehozását valósítja meg.

Mivel a legtöbb rendelkezésre álló lexikon angol nyelvű, a szerző magyar nyelvű szövegeken hajtott végre kísérleteket. A kísérletek során a különböző lexikonok minőségét vetettük össze azok szentimentelemzés során mért eredményességén keresztül. Bemutattuk, hogy jobb eredmények érhetőek el mind doménspecifikus alaplexikonok kiterjesztésével, mind pedig új doménspecifikus lexikonok építésével annotált szövegek segítségével, mint kézzel előállított jó minőségű lexikon doménen kívüli használata esetén. Korábbi publikációja (Hangya, 2015) alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- Doménspecifikus szentimentlexikonokat előállító nyelvfüggetlen automatikus módszerek kidolgozása: alaplexikonok kiterjesztése és újak létrehozása annotált szövegek segítségével;

- Doménspecifikus lexikonok hasznosságának bemutatása különböző magyar korpuszokon a minél nagyobb pontosság elérésének érdekében.

7. Keresztnyelvi szentimentelemzés doménadaptációja

A 7. fejezet a keresztnyelvi szentimentelemzéshez kidolgozott doménadaptációs módszereket mutatja be. Számos nyelv esetén nem áll rendelkezésre megfelelő mennyiségű annotált szöveg ahhoz, hogy jó minőségű szentimentelemző rendszereket hozhassunk létre. Szóbeágyazások és neurális hálózatok együttese jó minőségű szentimentelemző rendszerek létrehozását tette lehetővé. Továbbá, az ezekre épülő keresztnyelvi módszerek lehetővé tették, hogy egy erőforrásokban gazdag nyelv annotált szövegeit felhasználva olyan modellt hozzunk létre, mely alkalmazható olyan nyelvek esetében, melyekhez nem áll rendelkezésre annotált szöveg.

Ezen módszerek sok esetben jól teljesítenek, de a pontosságuk jelentősen csökkenhet, ha a tanító és kiértékelő adatok más doménből származnak. A szerző és munkatársai bemutatták, hogy általános célú erőforrásokat használó kétnyelvű szóbeágyazási modellek számos szó helytelen (a doménbe nem illeszkedő) jelentését tartalmazzák, valamint számos szó hiányzik belőlük. A probléma javítására egy olyan

kétlépcsős doménadaptációs módszert ismertettünk, mely korábbi munkákkal ellentétben csak egynyelvű jelöletlen szövegeket igényel. Első lépésként a módszer kétnyelvű szóbeágyazási modelleket adaptál azok doménspecifikus információkkal való gazdagításával. Második lépésként, egy félig felügyelt módszer hatékonyabb keresztnyelvi információátvitelt tesz lehetővé további doménspecifikus jelöletlen adatokra támaszkodva.

A végrehajtott kísérletek megmutatják, hogy kétnyelvű szóbeágyazási modellek doménadaptációjával nagyobb pontosságú keresztnyelvi szentimentelemző rendszerek hozhatóak létre. Továbbá bemutattuk, hogy a kidolgozott módszer könnyen alkalmazható más nyelvek és feladatok esetén is, mint például kétnyelvű lexikonok indukciójához. Korábbi publikációja (Hangya és mtsai., 2018) alapján a szerző a tézis legfontosabb saját eredményeinek a következőket tekinti:

- Egyszerű, de hatékony módszer kidolgozása kétnyelvű szóbeágyazások doménadaptációja céljából;
- Feladatspecifikus célnyelvi/doménbeli adatok integrációját lehetővé tevő félig felügyelt módszer kidolgozása hatékonyabb keresztnyelvi információátvitel érdekében;
- Keresztnyelvi problémák doménadaptációja jelöletlen szövegek felhasználásával.

Irodalomjegyzék

Hangya Viktor. 2015. Automatic Construction of Domain Specific Sentiment Lexicons for Hungarian. In *Proceedings of the 18th International Conference on Text, Speech and Dialogue*, pp. 201–208.

Hangya Viktor, Berend Gábor és Farkas Richárd. 2013. SZTE-NLP: Sentiment Detection on Twitter Messages. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, pp. 549–553.

Hangya Viktor, Braune Fabienne, Fraser Alexander és Schütze Hinrich. 2018. Two Methods for Domain Adaptation of Bilingual Tasks: Delightfully Simple and Broadly Applicable. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 810–820.

Hangya Viktor és Farkas Richárd. 2013a. Filtering and Polarity Detection for Reputation Management on Tweets. In *Working Notes of the 2nd Conference and Labs of the Evaluation Forum.*

Hangya Viktor és Farkas Richárd. 2013b. Target-oriented opinion mining from tweets. In *Proceedings of the 4th IEEE International Conference on Cognitive Infocommunications*, pp. 251–254.

Hangya Viktor és Farkas Richárd. 2017. A comparative empirical study on social media sentiment analysis over various genres and languages. *Artificial Intelligence Review*, 47(4):485–505.

Hangya Viktor, Szántó Zsolt és Farkas Richárd. 2017. Latent Syntactic Structure-Based Sentiment Analysis. In *Proceeding of the 2nd IEEE International Conference on Computational Intelligence and Applications.*