TAALTECHNOLOGIE 2.0: SENTIMENTANALYSE EN NORMALISATIE

Bart Desmet, Orphée De Clercq, Marjan Van de Kauter, Sarah Schulz, Cynthia Van Hee & Véronique Hoste

De opkomst van het internet voor en door iedereen, dat als Web 2.0 bekendstaat, heeft ingrijpende gevolgen gehad voor wat er (online) wordt geschreven. Naarmate het volume tekst in blogposts, tweets, Facebookupdates en andere sociale media toeneemt, slinkt het aandeel geredigeerde tekst. Internetgebruikers schrijven wat ze denken en voelen, en doen dat niet altijd in standaardtaal.

Dat taal in nieuwe media afwijkt van de norm, kan op verschillende manieren worden verklaard: spreektaal, jongerentaal, een klein toetsenbord, weinig tijd, een beperkt aantal karakters… Daarnaast wordt het internet niet enkel gebruikt om objectieve informatie te zoeken en te delen. We uiten ons enthousiasme of ongenoegen over gebeurtenissen, producten of diensten, en gaan ook actief op zoek naar de mening van anderen, bijvoorbeeld over een hotel, een film of een politieke beslissing. Er is ook interesse om in de veelheid aan online tekst bepaalde gevoelens te kunnen vinden, samen te vatten of tendensen te ontdekken.

In het domein van de automatische sentimentanalyse, ook wel *opinion mining* genoemd, wordt onderzoek gedaan naar de detectie en analyse van polariteit, opinies, emoties en andere subjectieve informatie. Een systeem dat sentiment analyseert, heeft diverse toepassingsmogelijkheden, zowel maatschappelijk als commercieel. De pers bericht bijvoorbeeld geregeld over de gevaren van sociale netwerksites. In het HOF-project SubTLe[1] onderzoeken we of suïcidale berichten op zulke sites automatisch gedetecteerd kunnen worden, en in AMiCA[2] wordt schadelijke content opgespoord, zoals online pestgedrag of pedofilie. Het IWT-project SentiFM[3] analyseert

dan weer sentiment over specifieke gebeurtenissen in economische nieuwsberichten. De resultaten van dergelijk onderzoek kunnen onder meer als hulpmiddel gebruikt worden om het effect van specifieke informatie op de financiële markten te achterhalen. Hoe we automatische sentimentanalyse kunnen aanwenden om de voorkeuren van online gebruikers te bepalen om vervolgens gepersonaliseerde advertenties aan te bieden, is dan weer het onderwerp van het PARIS[4]-project.

De eerste en fundamentele stap in al deze projecten is om via diepe tekstuele analyse automatisch een model van het sentiment in een tekst te vormen. Daarvoor baseren we ons op corpora waarin expressie van sentiment manueel geannoteerd wordt. Omdat bij de bestaande annotatieschema's de nadruk vooral ligt op expliciete uitingen van positief of negatief sentiment (Wiebe et al., 2005), stelden we nieuwe richtlijnen op die dieper gaan en ook impliciet sentiment kunnen detecteren en analyseren (Van de Kauter et al., 2013). Dat laatste is met name van belang in objectievere teksttypes zoals nieuwsberichtgeving, waarin expliciet uitgedrukt sentiment vaak ontbreekt. Een paar voorbeelden maken dat duidelijk:

(1) Khaaaat men leve!!! :(

(2) Batterij van nieuwe iPhone 5s gaat 2 uur mee

In (1) wordt expliciet negatief sentiment uitgedrukt (*haten*) door een bron (*ik*) over een doel (*mijn leven*). Het sentiment wordt verder versterkt door zogenaamde modificatoren als *!!!* en *:(*. Ook (2) drukt negativiteit uit, namelijk over het doel *batterij van nieuwe iPhone 5s*, maar dan op een impliciete manier. Ook al geeft de zin objectieve informatie, een batterij met een levensduur van 2 uur wordt (in deze context en met wereldkennis) als een minpunt beschouwd. We spreken in dat geval over een polair feit. Door inferentie willen we ook uit de formulering kunnen afleiden dat een negatief oordeel over de batterij bovendien slecht nieuws betekent voor de *iPhone 5s*, en

bij uitbreiding *Apple*. Dergelijke *deel-van*-relaties worden daarom ook geannoteerd.

Uit de bespreking van (1) wordt al snel duidelijk dat er op sociale media niet noodzakelijk in standaardtaal wordt geschreven. Dit is een groot probleem voor onze huidige taaltechnologische tools, aangezien zij moeilijk overweg kunnen met niet-standaardtaal. Hier zijn twee mogelijke oplossingen voor: de tools robuuster maken voor taalvariatie (adaptatie) of het aangetroffen taalgebruik dichter bij de norm brengen (normalisatie).

Wij hebben gekozen voor de laatste aanpak. We verzamelden een corpus van Nederlandse en Engelse teksten uit sociale media en zetten het manueel om naar standaardtaal. Omdat we voor de automatische normalisatie een robuust systeem willen hebben, werden drie verschillende genres geselecteerd: SMS-berichten, forummateriaal en tweets. In zulke berichten worden bepaalde woorden of letters vaak weggelaten, en worden afkortingen en acroniemen frequent gebruikt om tijd of karakters te sparen. Daarnaast bevatten de teksten regionale woorden en klanken, en laat de spelling soms te wensen over. Emoties worden in nieuwe media ook vaak uitgedrukt door herhalingen, smileys of gebruik van hoofdletters.

Deze dataset diende als basis voor een eerste automatisch systeem voor het Nederlands. We hebben ervoor gekozen om deze omzetting te beschouwen als een vertaling van een eerste naar een tweede variant van dezelfde taal (De Clercq et al., 2013). Een belangrijk en vernieuwend aspect van dit onderzoek was dat we de vertaling in twee rondes hebben uitgevoerd: eerst een vertaling op woordniveau (een woord-voor-woordvertaling) met de bedoeling om veelvoorkomende afwijkingen van standaardtaal op te vangen (zoals afkortingen), daarna een vertaling op karakterniveau. Met die opsplitsing in karakters hoopten we variaties op de standaardtaal te vinden die bepaalde regelmatigheden vertonen, zoals woorden die aan elkaar geplakt zijn (*dak* wordt *dat ik*). Deze aanpak bleek

succesvol: in totaal zijn we erin geslaagd om 63% van de variaties om te zetten in de standaardvorm. Momenteel onderzoeken we volop mogelijkheden om het systeem verder uit te breiden zodat ook typische problemen, zoals spellingsfouten of een schrijfwijze gebaseerd op uitspraak (*sgatjeuh*), genormaliseerd worden.

Kortom, taaltechnologie 2.0 is een nieuwe onderzoekslijn binnen LT3, de afdeling Taaltechnologie van de vakgroep, waarvan we de komende jaren veel interessante resultaten mogen verwachten. We streven er bij dit onderzoek dan ook naar om onze tools zoveel mogelijk in te schakelen voor maatschappelijke en commerciële doeleinden, zonder al te veel de "Big Brother" van de vakgroep te worden.

Referenties

De Clercq, O., Schulz, S., Desmet, B., Lefever, E. & Hoste, V. (2013). Normalization of Dutch User-Generated Content. Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013). Hissar, Bulgaria.

Van de Kauter, M., Desmet, B. & Hoste, V. (2013). Guidelines for the Fine-Grained Analysis of Polar Expressions, version 1.0. LT3 Technical Report - LT3 13-01.

Wiebe, J., T. Wilson & C. Cardie. (2005). Annotating expressions of opinions and emotions in language. *Computer Intelligence*, *39*(2), 165–210.

Projectwebsites

[1] Subjectivity Tagging and Learning, http://lt3.hogent.be/en/projects/subtle/

[2] Automatic Monitoring for Cyberspace Applications, http://lt3.hogent.be/en/projects/amica/

[3] Sentiment mining for Financial Markets, http://lt3.hogent.be/en/projects/sentifm/

[4] Personalized AdveRtisements buIlt from web Sources, http://lt3.hogent.be/en/projects/paris/