Projectvoorstel voor het Publiek-Privaat Center for Digital Humanities

Aanvragende organisatie:

Meertens Instituut

Private partner:

Teezir B.V. Utrecht
(http://teezir.com/nl/)

Titel projectaanvraag

TINPOT: Taal, Identiteit, Netwerken en Produktgeruchten Op Twitter
(aanvang 1 september 2012)

Keywords
Leeftijd, geslacht, produktgerucht, taal, netwerk, Twitter

Project omschrijving

Veel onderzoek op het Meertens Instituut staat van oudsher in het teken van veranderlijkheid, zoals onderzoek naar dialectvariatie, jongerentaal en mondelinge overlevering van verhalen en liederen. Met de digitalisering van de samenleving wordt het steeds aantrekkelijker om analyses toe te passen op grote hoeveelheden digitale data die al voorhanden zijn. Twitter als sociaal medium herbergt een enorm reservoir aan talige data en informeert van seconde tot seconde over sociaal gedrag en meningen/oordelen. Het bureau Teezir doet in opdracht van bedrijven reeds onderzoek naar produktreputaties in sociale media.
Er kan aan de hand van een grote hoeveelheid trainingsdata een module ontwikkeld worden die in staat is om op basis van taalgebruik uit te maken in welke leeftijdscategorie iemand valt, of men een man of een vrouw is, hoe groot iemands netwerk is, en of iemand een opinieleider of een trendvolger is. Vervolgens kunnen produktgeruchten gevolgd worden van begin tot eind. Produktgeruchten behoren ten dele tot de volksverhalen die we broodjeaapverhalen noemen (bijv. over beton in eikenhouten meubelen, kankerverwekkende toegevoegde E-stoffen, slachtafval in fricadellen, regenwormen in hamburgers van McDonald's, schadelijke straling van mobiele telefoons of zendmasten etc.). De module kan dan analyseren in welke groepen bepaalde produktgeruchten circuleren (leeftijdsgroepen, mannen, vrouwen) en langs welke netwerken en kanalen. Er kan een sentimentanalyse worden losgelaten op de produktgeruchten: welke positieve of negatieve oordelen worden er geveld? Hoe stabiel blijven de geruchten? Is het steeds een kwestie van letterlijk retweeten, of veranderen mensen de boodschap of voegen ze een commentaar toe (bijv. lol of een mededeling met een #)? De variatie van mededelingen op Twitter vertoont zekere overeenkomsten met variatie in de mondelinge overlevering van geruchten.

De analyses bieden meer zicht op hoe produktgeruchten zich verspreiden, in welke groepen, wiens mening belangrijk gevonden wordt, welke imago's er aan produkten verbonden zijn en hoe de geruchten kunnen veranderen. Het onderzoek kan leiden tot aanbevelingen hoe bedrijven zich kunnen/moeten wapenen tegen bepaalde negatieve geruchten.
Voor de module is nog geen voorwerk gedaan, maar er bestaan bij Teezir al wel technieken om produktgeruchten te monitoren.

Planning (6-9 maanden)

Het project staat gepland als een stage en afstudeerproject, dat doorgaans 6 maanden duurt. Er wordt een module getraind en ontwikkeld, die de volgende stappen kan maken op basis van een grote hoeveelheid Nederlandstalige tweets:
- analyse taal op leeftijdscategorie
- analyse taal op geslacht
- analyse op attitude: opinieleider of trendvolger
- analyse van het netwerk
- volgen van produktgeruchten (langs welke groepen en kanalen gaan de verhalen, welke sentimenten spelen er, welke variatie ontstaat er?)

Begroting

Het betreft een betaald afstudeerproject met twee master studenten/stagiairs (één met specialisatie op het technische terrein van Human Media Interaction, één met specialisatie taalkunde/etnologie). Daarnaast is er behoefte aan een student-assistent die mee kan helpen om testcorpora te annoteren en testresultaten te evalueren.

stagiair/master student HMI	8.000
stagiair/master student taalkunde/etnologie	8.000
student-assistent/annotator/evaluator testgegevens	16.000
3 laptops	5.100
reiskostenvergoeding	4.000
TOTAAL	**41.100**

Het gaat bij de bovenste drie posten om salariskosten (bruto salaris plus werkgeverslasten) en de looptijd is 6 maanden. De twee master studenten/ stagiairs ontvangen een stagevergoeding van circa 1000 euro bruto per maand voor de duur van een half jaar (wat neerkomt op circa 850 euro netto). De student-assistent/annotator/evaluator komt voor een half jaar in dienst van het Meertens en verdient 1801 bruto per maand.
Voor de drie werknemers zijn drie laptops (MacBookPro) voorzien. Omdat niet iedereen in Amsterdam zal wonen, zijn ook wat reiskosten ingecalculeerd.

Maatschappelijke waarde

Tot de doelgroepen voor de module behoren: onderzoekers van computeranalyses van natuurlijke taal, onderzoekers van taal en cultuur, media-onderzoekers, Teezir, bedrijven die inzicht willen in de achtergronden van

produktgeruchten, en op zoek zijn naar strategieën om te reageren op produktgeruchten.

Risico's en afhankelijkheden

Het project is voornamelijk afhankelijk van beschikbare uitvoerders. Verder zijn er geen belemmeringen.

Haalbaarheid

Het project wordt ingepland als een gecombineerd afstudeerproject aan het Meertens Instituut i.s.m. de Universiteit Twente. Een gemiddeld afstudeerproject duurt 6 maanden. Het betreft twee stageplaatsen met een stagevergoeding die ruim vier maal hoger is dan gebruikelijk. De annotator/evaluator krijgt een aanstelling bij het Meertens als student-assistent.

Organisatie:

Coördinatie: Theo Meder

Uitvoerder(s): 1 master student/stagiair van Human Media Interaction, 1 master student/stagiair taalkunde/etnologie (afstudeerproject) & 1 student-assistent/annotator/evaluator, en als begeleiders: Dolf Trieschnigg, Dong Nguyen, Theo Meder, Marc van Oostendorp, Leonie Cornips (UTwente & Meertens)

Private partner: Teezir (Utrecht)

Relatie met bestaand onderzoeksprogramma?

Het onderzoek past binnen programma's zoals die worden verricht aan de Universiteit Twente en het Meertens Instituut (in het laatste geval vooral op de terreinen van CATCH en e-Humanities). De UTwente is vooral gespecialiseerd in data-analyse van natuurlijke taal en het programmeren van tools en modules. Op het Meertens Instituut wordt onderzoek gedaan naar variabiliteit in taal en cultuur, ook met gebruikmaking van digitale technieken. Het voorgestelde taalvariatie-onderzoek heeft met name betrekking op leeftijd, geslacht, attitudes en netwerken. Het voorgestelde cultuurvariatie-onderzoek ligt op het terrein van schriftelijke en mondelinge overlevering, geruchtenverspreiding en variabiliteit in netwerken.

Rol en insteek van private partner

De private partner Teezir levert de trainingsdata waarop de module getraind wordt. Na ontwikkeling van de tool zal Teezir deze ook inzetten bij stemmings-onderzoek op sociale media.

Deliverables/concreet product

Teezir doet al onderzoek op sociale media om in opdracht de reputatie van produkten te monitoren. Bij de reeds bestaande software wordt een verrijkende

module ontwikkeld. De module is in staat om op basis van taalgebruik op Twitter te bepalen of het om een man of een vrouw gaat, en welke leeftijd de taalgebruikers bij benadering hebben. Daarnaast is de module in staat om te onderscheiden of het om opinieleiders of trendvolgers gaat, en of mensen grote of kleine netwerken hebben. Meer specifiek wordt hierna gekeken naar op Twitter circulerende produktgeruchten: wie doen hier overwegend aan mee? Mannen of vrouwen, jongeren of ouderen? Wat is de invloed van opinieleiders en trendvolgers? In hoeverre variëren de produktgeruchten? Gaat het steeds om letterlijke retweets? Worden er commentaartjes aan de tekst toegevoegd? Of begint het produktgerucht na verloop van tijd nog sterker te variëren? Welke invloeden zijn hierbij van betekenis?
De deliverable is een module bij een bestaande commerciële tool. De module kan gebruikt worden voor onderzoek aan de UTwente, het Meertens Instituut en Teezir.

Techniek

Html, Open NLP, sentiment-analyse, Twython.