Nieuwe dilemma's met oude en nieuwe online data

Sinds een aantal jaren ben ik lid van de Ethische Toetsingscommissie voor de Geesteswetenschappen van de Radboud Universiteit. Deze commissie toetst of voorgenomen onderzoeksprojecten op een verantwoorde manier omgaan met mensen en de gegevens van mensen. Onderzoeksprojecten mogen proefpersonen niet hinderen of schaden in hun privacy. Voor allerlei soorten onderzoek, zoals experiment en interview, hebben we beschreven hoe deze standaard plaatsvinden. Een standaard voor onderzoek met grootschalige online data in het kader van AI en taal/spraaktechnologie is er nog niet.

José Sanders
CLST, Radboud Universiteit

Afbeeldingen: Pixabay

Die beschrijving ontbreekt niet omdat het niet voorkomt – onderzoek met grootschalige online data komt juist steeds vaker voor – maar omdat het helemaal nog niet standaard is. Bij elk voorkomend project bekijken we wat relevante aspecten zijn en zo worden 'werkendeweg' standaardeisen ontwikkeld.

Eigenaar van gegevens

Aanvankelijk was bij AI-onderzoekers de grootste zorg meer juridisch dan ethisch van aard. Het ging om de vraag van wie online data eigenlijk zijn. Bij gegevens die op het internet staan en voor onderzoek worden gebruikt zijn zowel het auteurs- als het reproductierecht van belang. Voor corpusopbouw moesten overeenkomsten worden ontwikkeld (zoals in het SoNaR-corpus).

Men vroeg zich af of ook nog eens afgewogen moest worden of ongerief voor personen, of de bescherming van privacy van personen, bij onderzoek naar internetdata een punt van aandacht moest zijn. De tot de persoon van de auteur herleidbare gegevens uit online data zijn immers via het internet algemeen beschikbaar gesteld door of namens deze persoon. Mogelijk ongerief is dan toch al veroorzaakt door het aanwezig zijn op het internet en niet door het onderzoek.

Identiteit achterhalen

Gaandeweg zijn we daar wel wat genuanceerder over gaan denken. Al is het juridisch geen probleem, het kan toch moreel lastig zijn om gegevens die gebruikers in het verleden geplaatst hebben, voor onderzoek te gebruiken, zonder zich rekenschap te geven van de gevolgen die dit zou kunnen hebben. Dit geldt vooral voor het grootschalig openbaar maken in publicaties en publiek beschikbare corpora, met name als de gegevens achteraf gezien schadelijk kunnen zijn voor mensen (de gebruiker zelf of de mensen over wie de gegevens gaan). Dan moeten tot deze personen herleidbare details worden verwijderd. Dat zijn niet alleen namen en plaatsen maar ook andere elementen die het mogelijk maken om de identiteit van mensen te achterhalen.

Zo'n eis geldt in mindere mate als het gaat om mensen die toch al gekozen hadden voor een openbaar leven, zoals politici en andere in massamedia optredende personen. Verondersteld mag worden dat zij zich ervan bewust zijn dat wat zij in media uiten, beschikbaar blijft. Dit geldt ook voor mensen die doelbewust een groot online publiek nastreven, zoals bloggers en vloggers. Maar overige gebruikers van online media zijn zich daar vaak minder van bewust en gebruiken online media als onderlinge interactiekanalen. Dat is zeker aan de orde als gebruikers ten tijde van het plaatsen van informatie minderjarig waren.

Minderjarigen

Bij het actief verzamelen van online gegevens, zoals geluidsfragmenten van dialecten of tekstfragmenten van Whatsapp, staan onderzoekers voor een dilemma. Hoe kun je online controleren a) wat de leeftijd van deelnemers is en b) of hun ouders of wettelijke vertegenwoordigers toestemming geven voor deelname? Er wordt immers geen identiteit van proefpersonen opslagen. De gelijktijdige eisen om zowel anonimiteit van opgeslagen data te garanderen als deze te controleren voor toestemming in geval van minderjarigheid, zorgen voor een paradox. Bij het online werven van data voor corpora en in surveys wordt daarom geëist dat:

1. in het online protocol de leeftijd wordt ingevoerd en
2. dat bevestigd wordt dat één van de ouders/vertegenwoordigers bij het invullen aanwezig is, meeleest en naleest, en
3. die ouder/vertegenwoordiger expliciet zelf toestemming geeft voor onderzoekgebruik en opslag van wat de jongere aanlevert.

Omdat deze zaken niet achteraf gerepareerd kunnen worden, is het belangrijk om ze voorafgaand aan het plaatsen van oproepen om online informatie goed te regelen. Leeftijdsinformatie en toestemming moeten eerst binnengehaald worden. Waterdicht zijn deze maatregelen niet, maar ze maken aan gebruikers wel expliciet duidelijk wat wel en niet de bedoeling is. Zo krijgen zij daarin een grotere alertheid; in die zin werkt het in elk geval preventief en dat is misschien op dit moment wel het hoogst haalbare.

Gegevens uit het verleden

Het komt ook voor dat in het verleden verzamelde mediagegevens, bijvoorbeeld uit radio of krant, nu in grote doorzoekbare corpora op het internet beschikbaar worden gesteld voor verder onderzoek. Dat levert een extra dilemma op. In de periode vóór het internet zullen mensen zich niet hebben gerealiseerd dat hun ingezonden brief of radio-interview in plaats van éénmalig, eindeloos beschikbaar zou worden gesteld. Achteraf om toestemming vragen is ondoenlijk. Met zulke gegevens moeten we daarom voorzichtig zijn. Openbaarmaking van gegevens die achteraf gezien ongerief kunnen veroorzaken bij nog levende bronnen of bij nabestaanden, moet net als geldt bij historisch onderzoek, voorkomen worden door anonimisering.

De vraag is vervolgens wie moet inschatten wat ongerief kan veroorzaken: de onderzoeker? En is het corpus niet te groot om dit voor alle fragmenten vast te stellen? Als dat niet mogelijk is, zijn er goede argumenten om zo'n corpus niet openbaar te maken, maar alleen ter beschikking te stellen aan onderzoekers. Als die er een deel van willen gebruiken in een publicatie, kan dat stuk in elk geval van herleidbare gegevens worden ontdaan. Ook hier kunnen gegevens van destijds reeds openbaar bekende mensen uitgezonderd worden.

Relativering

Onmiskenbaar veranderen internet en online media de samenleving in de manier waarop wij omgaan met media-informatie. Er zijn zóveel online data dat het misschien gaandeweg minder belangrijk wordt hoeveel en wat van ons ergens online staat. Jongeren staan daar nu vaak al relativerender in dan ouderen. Dat kan betekenen dat we over een aantal jaren minder eisen zullen stellen aan anonimiteit en toestemming, simpelweg omdat van iedereen wel ongerief veroorzakende data te vinden zijn en we het daarom collectief opgeven om ons daar druk over te maken. Maar zo ver zijn we nog niet, en daarom blijft ethiek een extra uitdaging voor iedereen die zich met de verzameling en opslag van online data bezighoudt.