# Automatische inhoudsanalyse van Nederlandstalige data

*Een overzicht en onderzoeksagenda*

Trilling, D.; Boumans, J.

**Publication date**
2018
**Document Version**
Final published version
**Published in**
Tijdschrift voor Communicatiewetenschap

Damian Trilling & Jelle Boumans*

# Automatische inhoudsanalyse van Nederlandstalige data

## Een overzicht en onderzoeksagenda

### Inleiding

Steeds meer tekstuele data (nieuwsberichten, persberichten, recensies, reacties van klanten, politieke partijprogramma's) zijn digitaal beschikbaar. Mede als gevolg van de toenemende grootte van dit soort datasets worden geautomatiseerde inhoudsanalyses steeds populairder in communicatiewetenschappelijke studies. Desalniettemin is de automatische inhoudsanalyse nog niet volledig ingeburgerd in de Nederlandse en Vlaamse communicatiewetenschap: een inventarisatie leert dat er sinds 2005 in het *Tijdschrift voor Communicatiewetenschap* 44 artikelen zijn gepubliceerd die gebruik maken van een kwantitatieve inhoudsanalyse; in slechts drie gevallen is dit een geautomatiseerde methode. Hoewel veel analyses van Nederlandstalige data ook in Engelstalige tijdschriften worden gepubliceerd, geeft dit opmerkelijk lage aantal toch een indicatie van de tot nu toe redelijk beperkte toepassing van zulke methoden in de Vlaamse en Nederlandse communicatiewetenschap.

In dit artikel inventariseren we de staat van de automatische inhoudsanalyse in het Nederlandse taalgebied. Eerst geven we een beknopt overzicht van de meest voorkomende automatische inhoudsanalysetechnieken. Vervolgens gaan we in op beperkingen van zulke analyses in een Nederlandstalige context ten opzichte van andere talen, met name het Engels. We sluiten af met een reeks aanbevelingen om de toepassing van automatische inhoudsanalyse van Nederlandstalige teksten te stimuleren. Hiermee proberen we een bijdrage te leveren aan de methodologische discussie over de toepassing van automatische methoden in de context van relatief kleine talen. Naast een overzicht van relevante literatuur proberen we door verwijzingen naar specifieke softwarepakketten praktische handvatten aan te reiken voor onderzoekers die met

* Damian Trilling is als universitair docent verbonden aan de afdeling Communicatiewetenschap van de Universiteit van Amsterdam en de Amsterdam School of Communication Research. Postbus 15791, 1001 NG Amsterdam. E-mail: d.c.trilling@uva.nl.
Jelle Boumans is als universitair docent verbonden aan de afdeling Communicatiewetenschap van de Universiteit van Amsterdam en de Amsterdam School of Communication Research. Postbus 15791, 1001 NG Amsterdam. E-mail: j.w.boumans@uva.nl.

automatische inhoudsanalyses binnen het Nederlandse taaldomein aan de slag willen gaan. Vanwege deze focus zullen we relatief kort ingaan op methoden die grotendeels taalonafhankelijk zijn, en uitgebreider ingaan op methoden waar taalspecifieke aspecten wel een belangrijke rol spelen.

## Een classificatie van automatische-inhoudsanalysetechnieken

In tegenstelling tot de traditionele kwantitatieve inhoudsanalyse (zie bijv. Krippendorf, 2004; Lacy, Watson, Riffe & Lovejoy, 2015) is de automatische inhoudsanalyse als methode minder duidelijk afgebakend. Dit komt vooral omdat een zeer breed scala aan technieken onder de overkoepelende term 'automatische inhoudsanalyse' valt. Deze technieken variëren van door de onderzoeker opgestelde woordenlijsten tot geavanceerde modellen uit de informatica (waarbij niet de onderzoeker, maar een programma bepaalt welke woorden interessant zijn) of uit de *computational linguistics* (waarbij een programma bijvoorbeeld geautomatiseerd zinnen ontleedt). Hoe verschillend deze technieken ook zijn, zij maken allemaal onderdeel uit van de gereedschapskist die onderzoekers tot hun beschikking hebben wanneer ze de hulp van de computer willen inschakelen om media-inhoud te analyseren (Boumans & Trilling, 2016; Grimmer & Stewart, 2013; Günther & Quandt, 2016). Soms wordt de term '*computer assisted content analysis*' (CACA) gebruikt om deze technieken aan te duiden. Maar hoewel de onderzoeker uiteindelijk de regie voert, is de rol van de computer – zeker bij geavanceerdere methoden – vaak meer dan alleen assisterend. We hanteren daarom in het vervolg de term 'geautomatiseerde inhoudsanalyse'.

De beschikbare technieken zijn te rangschikken op een schaal van deductief (of *top-down*) naar inductief (of *bottom-up*) (Boumans & Trilling, 2016). Hoe deductiever de techniek, hoe meer regels de onderzoeker van tevoren opstelt die in de data teruggezocht worden; hoe inductiever, hoe meer patronen uit de data zelf naar voren moeten komen. Aan het deductieve eind van het spectrum bevinden zich regelgebaseerde (*rule based*) en op woordenlijsten gebaseerde (*dictionary based*) methoden. Regelgebaseerde methoden zijn met name geschikt voor het coderen van manifeste kenmerken van een tekst ('politicus P wordt wel/niet genoemd'). Voor variabelen die minder manifest zijn en meer mogelijke verschijningsvormen kennen (zoals het onderwerp van een artikel), is het over het algemeen moeilijker om a priori expliciete regels op te stellen.

Wanneer woordenlijsten tekortschieten kan *supervised machine learning* een uitkomst bieden. Hierbij stelt de onderzoeker van tevoren categorieën op, bijvoorbeeld de onderwerpen (politiek, economie, sport, enz.) waarover een tekst kan gaan. In tegenstelling tot regelgebaseerde methoden worden er echter geen expliciete regels opgesteld aan de hand waarvan het programma zou kunnen bepalen hoe de tekst gecodeerd moet worden. In plaats daarvan probeert een algoritme op basis van een aantal

handmatig gecodeerde artikelen te achterhalen welke gemeenschappelijke karakteristieken de artikelen over bijvoorbeeld politiek hebben, en welke karakteristieken sportartikelen hebben; de computer leert in feite zelf de regels uit het geannoteerde materiaal. Als dit succesvol blijkt, kan vervolgens een onbeperkte hoeveelheid 'nieuwe' data automatisch geannoteerd worden. Daarom kan *supervised machine learning* als een tussenvorm van een top-down- en een bottom-upmethode worden beschouwd: de categorieën worden top-down bepaald, maar de regels bottom-up.

Aan het inductieve, bottom-upeind van het spectrum vinden we *unsupervised machine learning methods*. Deze methoden achterhalen de betekenis van een collectie van teksten door patronen van samen voorkomende woorden in de data te signaleren. In tegenstelling tot *supervised machine learning* en op regels of woordenlijsten gebaseerde methoden is niet van tevoren vastgelegd welke specifieke informatie uit de dataset dient te worden gehaald. *Unsupervised machine learning* is daardoor bij uitstek geschikt voor exploratief en inductief onderzoek, waarbij de patronen in de data een indicatie kunnen zijn van in de tekst aanwezige onderwerpen en/of frames.

## Automatische inhoudsanalyses van Nederlandstalige teksten

In het vervolg van dit artikel zullen we aan de hand van enkele voorbeelden beschrijven hoe de technieken uit het arsenaal van geautomatiseerde inhoudsanalyses tot nu toe zijn toegepast op Nederlandstalige data. Daarbij is specifiek aandacht voor taalafhankelijke aspecten die de toepassing van de techniek kunnen belemmeren. Op technieken die minder gevoelig zijn voor taalkenmerken zullen we daarom korter ingaan dan op methoden die juist zeer afhankelijk hiervan zijn. Voor een algemener bespreking van geautomatiseerde inhoudsanalyses verwijzen we naar bestaande reviews (bijv. Boumans & Trilling, 2016; Grimmer & Stewart, 2013; Günther & Quandt, 2016).

### *Nederlandstalige woordenlijsten*

De meest simpele varianten van woordenlijst- of regelgebaseerde analyses zijn in principe weinig gevoelig voor taalspecifieke bijzonderheden. Zo zijn zichtbaarheidsanalyses, waarin met behulp van een woordenlijst of combinatie van zoektermen bijvoorbeeld gemeten wordt hoe vaak een bepaalde organisatie of politicus in het nieuws voorkomt, niet gevoelig voor de specifieke grammatica van een taal. De zichtbaarheid van politieke partijen kan dan ook eenvoudig worden gemeten door in een digitaal krantenarchief op partijnamen te zoeken (bijv. Vliegenthart & Van Aelst, 2010). Ook kan de interactiviteit van Twittergebruikers gemeten worden door simpelweg het aantal @-tekens in een bestand met tweets te tellen (Kruikemeier, Van Noort, Vliegenthart & De Vreese, 2015).

Complexer wordt het gebruik van een woordenlijstbenadering wanneer de tekstkenmerken minder manifest zijn. Stel dat iemand een lijst van woorden wil opstellen die een artikel classificeren als 'economisch nieuws': hoe uitgebreid moet deze lijst dan zijn? Het ligt voor de hand dat een minimale lijst die slechts de twee woorden 'economie' en 'economisch(e)' bevat, niet alle relevante artikelen zal vinden. En een te uitgebreide lijst met termen die slechts zijdelings of alleen in bepaalde contexten met economie te maken hebben, zal weliswaar nagenoeg alle artikelen over economisch nieuws identificeren (dit wordt een hoge *recall* genoemd), maar tegelijkertijd zal zo'n lijst ook allerlei teksten ten onrechte als economisch nieuws aanmerken, bijvoorbeeld een tekst over een kunsttentoonstelling in een beursgebouw (dit wordt een lage *precision* genoemd).

Het is vaak lastig om vooraf regels en woordenlijsten op te stellen die én een hoge *recall* én een hoge *precision* bieden, al zijn er verschillende mogelijkheden om de *precision* en *recall* te verhogen. Zo maken Van Noije, Kleinnijenhuis en Oegema (2008) in een poging om met behulp van een woordenlijst issues in de parlementaire agenda en media-agenda te coderen, gebruik van een *ramped window*: een indicator voor een bepaald issue (immigratie, agricultuur, drugs, milieu, enz.) wordt alleen geteld als er binnen een bepaalde woordafstand een term met een semantisch vergelijkbare betekenis voorkomt; woorden die dichterbij zitten tellen hierbij zwaarder mee.

Op een soortgelijke manier is het mogelijk om rekening te houden met negaties (woorden die een ontkenning uitdrukken, zoals 'niet' en 'zonder'). Aaldering en Vliegenthart (2016) trachten op basis van woordenlijsten conclusies te trekken over verbanden tussen politieke actoren en karaktereigenschappen in mediaberichtgeving. Door negaties die binnen een bereik van vijf woorden van een eigenschap staan, te betrekken in de analyse, houden de auteurs in zekere mate rekening met de context waarin een eigenschap wordt genoemd. In theorie kan hetzelfde gedaan worden met versterkende woorden ('zeer', 'bijzonder'). Deze benadering is echter niet feilloos: in de zin 'Wilders is niet moedig, maar Rutte wel' is het niet mogelijk om met behulp van zo'n vijfwoordenregel te achterhalen wie van de twee politici al dan niet moedig is. Of dit in de praktijk daadwerkelijk tot problemen leidt of dat het aantal misclassificaties verwaarloosbaar is, is een vraag die – per taal en context – nader zou moeten worden onderzocht.

Een andere benadering om de nauwkeurigheid van woordenlijstmethoden te verhogen, kiezen Jonkman, Trilling, Verhoeven en Vliegenthart (2016): zij gebruiken zogenoemde reguliere uitdrukkingen (*regular expressions*) om variaties in de spelling van namen te corrigeren en ambiguïteiten op te lossen. Zo is het bijvoorbeeld mogelijk om het woord 'Rutte' te vervangen door 'Mark_Rutte' indien de tekst ten minste één keer de woordcombinatie 'Mark Rutte' of de afkorting 'VVD' of het woord 'pre-

mier' of 'minister-president' bevat, omdat men in dat geval vrijwel zeker weet dat geen andere persoon met de achternaam Rutte wordt bedoeld.

Hoewel dit soort uitbreidingen sommige tekortkomingen van de woordenlijstbenadering kunnen verhelpen, zijn woordenlijsten en zoektermen vooral nuttig om zeer manifeste variabelen te coderen. Hoe latenter een concept, hoe groter de kans dat een woordenlijstbenadering tekortschiet. Het blijkt bijvoorbeeld niet mogelijk om met behulp van een woordenlijst alle facetten van het concept 'populisme' op een volwaardige manier te meten: een voornaam kenmerk van populisme – verwijzingen naar het 'wij-gevoel' en 'ons (nationaal) belang' – is nauwelijks in een woordenlijst te vangen (Rooduijn & Pauwels, 2011). Rooduijn en Pauwels stellen dan ook vast dat de met behulp van een woordenlijst geautomatiseerde inhoudsanalyse van het concept minder contentvalide is dan een manuele analyse.

Desalniettemin is er een aantal voorbeelden waarin woordenlijstgebaseerde methoden succesvol zijn toegepast binnen het Nederlandse en Vlaamse taalgebied om latentere constructen te meten. Voorbeelden van het meten van latente constructen met een woordenlijstbenadering zijn een studie waarin frames in krantenberichten over oudere werknemers worden geïdentificeerd (Kroon, 2017), een studie naar frames in artikelen over immigratie (Vliegenthart & Roggeband, 2007), een studie naar de onderwerpen 'immigratie' en 'Europese integratie' (Van der Pas, 2014) en een studie waarin racisme in Vlaamse online *comments* wordt geïdentificeerd (Tulkens, Hilte, Lodewyckx, Verhoeven & Daelemans, 2016). Het opstellen van dergelijke woordenlijsten om latente constructen te identificeren, kan echter problematisch zijn: zo wijzen Tulkens en collega's op de zeer lage intercodeurbetrouwbaarheid van hun manueel samengestelde woordenlijst. Kennelijk is het ook voor menselijke codeurs lastig om het eens te worden over de vraag wanneer een woord een uiting van racisme is.

### Sentimentanalyse in de Nederlandstalige context

Sentimentanalyses zijn een populaire techniek binnen de geautomatiseerde inhoudsanalyse. Door de complexiteit van teksten terug te brengen naar een simpele maat van positiviteit of negativiteit is het mogelijk uitspraken te doen over attitudes ten opzichte van een actor of een onderwerp. Vroege vormen van sentimentanalyse waren puur op lijsten van positieve dan wel negatieve woorden gebaseerde methoden, die een zogenoemde *bag of words*-benadering volgden (dus geen rekening hielden met de volgorde van woorden). Zulke analyses (bijv. Vliegenthart, Boomgaarden & Boumans, 2011; Young & Soroka, 2012) kunnen dan ook onder de eerder besproken woordenlijstbenaderingen geschaard worden. De gebruikte woordenlijsten kunnen in principe vertaald worden vanuit elke willekeurige taal. Recentere methoden houden echter rekening met de woordvolgorde, versterkende of afzwakkende woorden, leestekens en

spellingsvariaties (bijv. Hutto & Gilbert, 2014; Thelwall, Buckley, Paltoglou, Cai & Kappas, 2010), wat een integrale vertaling aanzienlijk bemoeilijkt.

Een voorbeeld van een Nederlandstalige sentimentanalyse is te vinden in een studie van Junqué de Fortuny, De Smedt, Martens en Daelemans (2012). Hun benadering gaat verder dan het uitsluitend tellen van positieve en negatieve woorden. Gebruik makend van de door hen zelf ontworpen Python-module 'Pattern' stellen zij van 68.000 nieuwsartikelen het sentiment ten opzichte van Vlaamse partijen en politici vast. Pattern bevat onder meer een handmatig opgestelde woordenlijst bestaande uit ruim 3.000 Nederlandse bijvoeglijke naamwoorden die zijn voorzien van een score voor polariteit (van negatief tot positief op een schaal van –1,0 tot +1,0) en subjectiviteit (van objectief tot subjectief op een schaal van 0,0 tot 1,0). Nadat er rekening is gehouden met een aantal factoren, waaronder negaties ('niet') en versterkers ('zeer'), wordt een totaalscore voor de tekst berekend. Dit laat zien dat voor het uitvoeren van een sentimentanalyse meer nodig is dan alleen een woordenlijst: de woorden 'niet' en 'zeer' zijn niet inherent positief of negatief: hun betekenis is afhankelijk van de context.

Helaas zijn Nederlandse sentimentanalyse-algoritmes die rekening houden met versterkers en negaties niet alleen schaars, maar vaak ook minder goed gevalideerd dan hun Engelse variant (zoals in het geval van Sentistrength; zie Thelwall et al., 2010) of worden ze niet meer actief onderhouden (zoals in het geval van Pattern; zie De Smedt & Daelemans, 2012).[1] Er is ons op dit moment geen pakket voor Nederlandse sentimentanalyses bekend dat zonder voorbehoud aangeraden kan worden. Ook zijn de in deze pakketten gebruikte woordenlijsten niet zonder meer in alle contexten en onderzoeksdomeinen toe te passen. Vorengenoemde Sentistrength-woordenlijst bijvoorbeeld, is ontwikkeld met het oog op socialemediateksten, terwijl Pattern is gebaseerd op recensies en reviews. Het is onduidelijk hoe accuraat de woordenlijsten presteren wanneer ze bijvoorbeeld worden toegepast op krantenartikelen.

Onderzoekers die voor de Nederlandse taal gebruik willen maken van een woordenlijstbenadering ervaren daarnaast een ander probleem: in tegenstelling tot bijvoorbeeld de Engelse taal zijn er relatief weinig Nederlandstalige woordenlijsten vrij beschikbaar. Het is tekenend dat een recent gepubliceerd onderzoek naar sentiment in krantenkoppen (Kuiken, Schuth, Spitters & Marx, 2017) gebruik maakt van een woordenlijst uit 2008 (Jijkoun & Hofmann, 2008).

Wanneer de lijst een vertaling is vanuit een andere taal kan dit voor problemen zorgen. De veelgebruikte *Linguistic Inquiry and Word Count* (LIWC), een set van woordenlijsten die gerelateerd zijn aan psychologische constructen (zie Tausczik & Pennebaker, 2009) is wel in het Nederlands beschikbaar, maar zoals de vertalers opmerken, zorgen een aantal taalkundige verschillen ervoor dat niet alle categorieën even goed overeenkomen met de Engelse variant (Zijlstra, Van Meerveld, Van Middendorp,

Pennebaker & Geenen, 2004). Desalniettemin is de Nederlandstalige LIWC-variant ook gebruikt in communicatiewetenschappelijk onderzoek, bijvoorbeeld om framing van terroristische aanslagen te analyseren (Ruigrok & Van Atteveldt, 2007) en recenter om de emotionaliteit van economisch nieuws te bepalen (Strauß, Vliegenthart & Verhoeven, 2016). Een ander groot nadeel is dat de LIWC-lijsten niet openbaar toegankelijk zijn, wat verdere ontwikkeling in de weg staat en tot een onwenselijke *black box*-situatie leidt, waarin derden niet zonder meer kunnen achterhalen hoe de resultaten tot stand zijn gekomen.

### Taalkundige benaderingen en zinsstructuur

Zoals we hebben laten zien, houden woordenlijstgebaseerde methoden niet of nauwelijks rekening met de zinsstructuur. Daarom is het – op de slechtere beschikbaarheid van zulke lijsten na – verder nauwelijks relevant of dergelijke analyses in het Nederlands of in het Engels uitgevoerd worden.

Onderzoekers die gebruik maken van woordenlijstmethoden lopen al vrij snel tegen de grenzen aan. In het bijzonder moeten er veel vergaande assumpties worden gemaakt, vooral wat betreft het interpreteren van de betekenis van het al dan niet voorkomen van een woord. Dit geldt in het bijzonder voor het trekken van conclusies over de betekenis van het *samen* voorkomen van woorden: als de woorden 'moslim' en 'terreur' in dezelfde zin voorkomen, is men misschien geneigd om te concluderen dat een moslim een aanslag heeft gepleegd – maar het is even goed mogelijk dat een moslim heeft opgeroepen om terreur te bestrijden. Uitsluitend op basis van het feit dat een woord en een actor in dezelfde tekst, alinea of zin voorkomen kan de aard van het verband tussen de twee niet worden bepaald: daarvoor is inzicht in grammaticale structuren vereist. Daarom zijn er inmiddels geavanceerdere methoden ontwikkeld om meer inzicht in de betekenis van een tekst te krijgen. Door zinnen geautomatiseerd te ontleden, is het bijvoorbeeld mogelijk om te herkennen of het woordje 'haar' in een zin naar 'mevrouw Jansen' in de zin ervoor verwijst. Deze taalkundige taak wordt aangeduid als *anaphora resolution*. Hoewel er geen kant-en-klare softwaremodules voor het Nederlands beschikbaar zijn, hebben Van Atteveldt, Kleinnijenhuis en Ruigrok (2008) in een analyse van Nederlandstalige krantenartikelen laten zien dat het met een aantal heuristische regels mogelijk is om rekening te houden met zulke verwijzingen. Zij geven als voorbeeld de volgende twee zinnen: 'Hirsi Ali verliet de politiek omdat Minister Verdonk haar het Nederlanderschap leek te ontnemen. De minister heeft dit later heroverwogen.' Het algoritme was in staat te achterhalen dat 'haar' op Hirsi Ali slaat, en 'de minister' op Verdonk.

Het achterhalen van de zinsstructuur kan dus nuttige informatie voor geautomatiseerde inhoudsanalyses opleveren. Hoewel het taalkundig gezien lastiger is om een Nederlandse zin automatisch te ontleden (het zogenoemde *part-of-speech* (POS) *tag-*

*ging*) dan een Engelse zin, is er een aantal POS-*taggers* voor het Nederlands beschikbaar, bijvoorbeeld ALPINO (Van Noord, 2006) of het Python-pakket Pattern (De Smedt & Daelemans, 2012; zie hiervoor). Hetzelfde geldt voor het gerelateerde *named entity recognition* (NER), het herkennen van entiteiten, zoals personen, organisaties of locaties.

De vaak gebruikte *Natural Language Toolkit* (NLTK; Bird, Klein & Loper, 2009) ondersteunt *part-of-speech tagging* en *named entity recognition* voor Engelstalige teksten. Deze technieken kunnen van groot belang zijn voor communicatiewetenschappelijke inhoudsanalyses. Zo gebruiken Burggraaff en Trilling (2017) NLTK om actoren in krantenartikelen te identificeren. Burscher (2016) heeft met behulp van NLTK zelfstandige en bijvoeglijke naamwoorden geïdentificeerd en deze vervolgens voor een framinganalyse gebruikt. Maar wie zoals deze onderzoekers NLTK voor POS of NER met Nederlandstalige teksten wil gebruiken, moet hiervoor eerst zelf een *machine learning model* trainen. Hoewel Nederlandstalige trainingdatasets beschikbaar zijn (Tjong Kim Sang, 2002) en het niet per se moeilijk is om deze datasets te gebruiken, zorgt het wel voor een extra drempel. Bovendien zijn *precision* en *recall* vaak acceptabel, maar zeker niet perfect (Tjong Kim Sang, 2002), wat betekent dat een niet te verwaarlozen aantal woorden verkeerd wordt geclassificeerd.

POS kan niet alleen worden gebruikt om de relaties tussen woorden in een zin te achterhalen, maar ook om relevante van minder relevante woorden te scheiden. De meeste geautomatiseerde inhoudsanalyses worden voorafgegaan door een zogenoemde *preprocessing* ofwel voorbewerkingsstap. Zo is het gebruikelijk om vaak voorkomende woorden zonder specifieke betekenis, zogenoemde stopwoorden, te verwijderen. Hieronder vallen bijvoorbeeld lidwoorden en vaak ook voornaamwoorden. Voor het verwijderen ervan voldoet een woordenlijstbenadering, en er zijn voldoende lijsten met Nederlandse stopwoorden beschikbaar. Sommige onderzoekers gaan verder in het 'opschonen' van hun data. Zo heeft Burscher (2016) POS-*tagging* op Nederlandstalige nieuwsberichten toegepast om uitsluitend zelfstandige en bijvoeglijke naamwoorden te gebruiken voor een analyse van frames en onderwerpen, wat tot minder ruis in de data heeft geleid.

Een andere taalkundige techniek die voor het opschonen van datasets en het voorbereiden van geautomatiseerde inhoudsanalyses wordt gebruikt, is het terugbrengen van woorden tot hun woordstam ('stemming') om te voorkomen dat vervoegingen en verbuigingen van hetzelfde woord als verschillende woorden worden aangezien. Helaas zijn verbuigingen en vervoegingen in de Nederlandse taal complexer en onregelmatiger van structuur dan in de Engelse taal. De meeste *packages* voor taalverwerking (zoals NLTK; Bird et al., 2009) ondersteunen het 'stemmen' van Nederlandse teksten wel, maar over het algemeen levert dit meer problemen op dan in het geval van de Engelse taal.

Meer in het algemeen kunnen we stellen dat de gebruikelijke technieken voor taalbewerking weliswaar in het Nederlands beschikbaar zijn, maar onderzoekers minder keuzes hebben. De populaire *tools* van de *Stanford Natural Language Processing Group* (https://nlp.stanford.edu/software/) zijn bijvoorbeeld naast in het Engels vaak ook in het Duits, Spaans en Chinees beschikbaar, maar niet in het Nederlands. Daar staat tegenover dat er ook in Nederland veel onderzoek op het gebied van natuurlijke-taalverwerking wordt gedaan, inclusief een jaarlijkse conferentie met bijbehorend tijdschrift: *Computational Linguistics in the Netherlands* (http://www.clinjournal.org). Een voorbeeld is het *NewsReader*-project, waarin gebeurtenissen in het nieuws automatisch worden geïdentificeerd (Vossen et al., 2016).

### *Supervised machine learning* in de Nederlandstalige context

De tot nu toe besproken technieken waren allemaal in meer of mindere mate taalgebonden, zij het doordat grammaticale regels opgevolgd moeten worden of omdat zoektermen of woordenlijsten nu eenmaal in de juiste taal beschikbaar moeten zijn. Maar bij het gebruik van andere technieken zijn taalspecifieke overwegingen minder belangrijk. Zo een techniek is *supervised machine learning*. *Supervised machine learning* betekent niets anders dan het schatten van een statistisch model met één afhankelijke variabele en (zeer veel) onafhankelijke variabelen (ook *features* genoemd), om vervolgens de onafhankelijke variabele voor teksten waar deze onbekend is, te kunnen voorspellen. Door grofweg tussen de 1.000 en 3.000 artikelen handmatig te coderen (in deze context spreekt men vaak van 'annoteren'), is het mogelijk om een model te trainen dat vervolgens de gewenste variabele voor een willekeurig aantal artikelen zelfstandig codeert.

*Supervised machine learning* is de laatste jaren succesvol toegepast om Nederlandstalige teksten te analyseren. Voorbeelden zijn het identificeren van generieke frames (Burscher, Odijk, Vliegenthart, De Rijke & De Vreese, 2014) en onderwerpen (Burscher, Vliegenthart & De Vreese, 2015) in Nederlandstalig nieuws. De door Burscher en zijn collega's ontwikkelde methode is ook door anderen in vervolgonderzoek toegepast (Bartholomé, Lecheler & De Vreese, 2017; Trilling, Tolochko & Burscher, 2017). Andere onderzoekers hebben laten zien dat *supervised machine learning* gebruikt kan worden om Nederlandstalige tweets als al dan niet werkgerelateerd te classificeren (Van Zoonen & Van der Meer, 2016). Het is wel een beperking dat er weinig geannoteerde trainingdatasets beschikbaar zijn; onderzoekers moeten training- en testdata dus vaak zelf (laten) annoteren.

*Supervised machine learning* kan ook een uitkomst bieden voor het gebrek aan valide sentimentanalysemodules in het Nederlands. Om een sentimentanalyse met behulp van *supervised machine learning* uit te voeren, wordt eerst van een aantal teksten handmatig gecodeerd of ze positief dan wel negatief van aard zijn, waarna de computer de

achterliggende kenmerken van de positieve en negatieve teksten herleidt (Gonzalez-Bailon & Paltoglou, 2015; zie ook Van Atteveldt, Kleinnijenhuis, Ruigrok & Schlobach, 2008). Hoewel deze methode door dit handmatige component relatief hoge opstartkosten met zich meebrengt, kent zij een groot voordeel: de onderzoeker kan de validiteit van de methode beter waarborgen. Anders dan in het geval van een woordenlijstmethode kan namelijk precies berekend worden hoe goed het getrainde model presteert ten opzichte van de handmatige codering.

Bij *supervised machine learning* wordt elke tekst doorgaans gerepresenteerd als de frequentie van de erin voorkomende woorden. Of deze woorden (oftewel *features*) nu Nederlandse, Engelse of Chinese woorden zijn, maakt in principe niet uit. In tegenstelling tot woordenlijst- of regelgebaseerde benaderingen is de beperkte beschikbaarheid van taalspecifieke onderzoeksinstrumenten dan ook een minder groot potentieel probleem. Hoewel er voor het voorbereiden van de data een aantal stappen vereist is, zoals het wegen op basis van de populariteit van woorden of het filteren van bepaalde woorden of woordcategorieën, zijn er aanmerkelijk minder taalspecifieke keuzes nodig dan in het geval van woordenlijstgebaseerde of regelgebaseerde methoden.

### *Unsupervised machine learning* in de Nederlandstalige context

*Unsupervised machine learning* wordt gebruikt om patronen te herkennen en cases te groeperen als er geen afhankelijke variabele is gemeten. Simpele voorbeelden zijn clusteranalyse, factoranalyse en hoofdcomponentenanalyse (*principal component analysis*, PCA). Door elke tekst te representeren als een vector van woordfrequenties en deze vervolgens als input voor een factor- of hoofdcomponentenanalyse te gebruiken, is het mogelijk te achterhalen welke woorden op een factor of component laden (bijv. De Graaf & Van der Vossen, 2013; Vlieger & Leydesdorff, 2012). De factoren of componenten worden vervolgens vaak als frames geïnterpreteerd. Deze benadering is het afgelopen decennium meermalen toegepast in communicatiewetenschappelijk onderzoek, onder andere om crisiscommunicatie te analyseren (Van der Meer, 2014; Van der Meer & Verhoeven, 2013; Van der Meer, Verhoeven, Beentjes & Vliegenthart, 2014).

Enkele kritische stemmen (zoals Leydesdorff & Nerghes, 2016) ten spijt zijn factor- en hoofdcomponentenanalyses op het gebied van onderzoek naar frames en onderwerpen in recenter onderzoek grotendeels verdrongen door zogenoemde *topic models*. Een vaak gebruikte variant, *latent dirichlet allocation* (LDA), is in staat om zowel zogenoemde *topics* (vaak geïnterpreteerd als onderwerpen of frames) in een dataset te identificeren, alsook per tekst aan te geven in welke mate deze *topics* erin voorkomen. In tegenstelling tot wat over het algemeen bij een factor- of hoofdcomponentenanalyse (LDA) wordt verondersteld, is het daarbij een aanname dat woorden onderdeel kunnen uitmaken van meerdere *topics*, wat realistischer lijkt dan te veronderstellen

dat elk woord slechts aan één factor of component toegewezen kan worden. Een voorbeeld van een Nederlandse inhoudsanalyse die gebaseerd is op de LDA-benadering is een studie van Strycharz, Strauß en Trilling (2017) naar de context waarin bedrijven in het nieuws komen.

Het valt buiten het bestek van dit artikel om de verschillende varianten van *topic models* uitgebreid te bespreken, maar het moge duidelijk zijn dat het – net zoals bij *supervised machine learning* – in principe niet uitmaakt in welke taal de teksten geschreven zijn: de enige taalspecifieke keuzes liggen op het gebied van *preprocessing*.

### Overlap en overeenkomsten tussen teksten

De tot nu toe besproken technieken zijn vooral erop gericht om kenmerken van teksten te identificeren – en in die zin zijn ze qua doelstelling vergelijkbaar met veel traditionele inhoudsanalyses. Maar daarnaast is het ook mogelijk om op grote schaal vast te stellen op welke wijze en in welke mate teksten overeenkomen en overlappen, een analyse die handmatig nagenoeg onmogelijk is: menselijke codeurs zouden zich immers de inhoud van duizenden teksten moeten kunnen herinneren om tijdens het lezen van een te coderen artikel vast te kunnen stellen of ze dit artikel al eerder zijn tegengekomen. Geautomatiseerde analyses bieden dan ook unieke mogelijkheden voor onderzoek naar bijvoorbeeld *gatekeeping*- en *agendabuilding*-processen, waarvoor tekstuele overeenkomsten een belangrijke indicatie kunnen zijn. Zo is in meerdere Nederlandse studies onderzocht in hoeverre persberichten, mediaberichten en persbureauberichten overlappen (Boumans, 2017; Welbers, Van Atteveldt, Kleinnijenhuis & Ruigrok, 2016). Hiervoor wordt gebruik gemaakt van maten zoals de *cosine similarity*, een soort correlatiemaat voor woordfrequenties, en de *Levenshtein distance*, een maat voor het aantal bewerkingsstappen (zoals het toevoegen of weghalen van woorden) dat nodig is om tekst A (bijvoorbeeld een persbericht) in tekst B (een krantenbericht) te transformeren. Het ligt voor de hand dat de specifieke taal waarin de teksten zijn geschreven niet relevant is – als het maar dezelfde taal is. Een potentieel probleem voor de Nederlandse context is dan ook vooral dat het – afhankelijk van de specifieke context – mogelijk is dat in een Nederlandstalig corpus enkele Engelstalige documenten staan, bijvoorbeeld omdat een organisatie een jaarverslag of persbericht in het Engels publiceert. Dit onderstreept eens te meer het belang van handmatige validatie van de dataset.

### Wanneer is een geautomatiseerde methode zinvol?

Ondanks de populariteit van inhoudsanalyses in de Nederlandstalige communicatiewetenschap worden er nog steeds betrekkelijk weinig automatische inhoudsanalyses uitgevoerd. Om de mogelijkheden die automatische inhoudsanalyses bieden inzich-

telijker te krijgen, stellen we voor om drie categorieën van inhoudsanalytische vraagstukken met betrekking tot Nederlandstalige teksten te onderscheiden: vraagstukken waarvoor (1) geautomatiseerde technieken voorhanden en zonder meer aan te bevelen zijn, (2) geautomatiseerde technieken waardevol zouden kunnen zijn, en (3) geautomatiseerde technieken (vooralsnog) niet voorhanden zijn.

***Vraagstukken waarvoor (een zekere mate van) automatisering zonder meer aan te bevelen zijn***

Hieronder vallen alle analyses die zich richten op de prominentie van actoren en organisaties. In mindere mate geldt dit ook voor duidelijk afgebakende issues en onderwerpen. Omdat alle gangbare statistiekprogramma's en spreadsheetprogramma's het tellen van zoektermen in *strings* (variabelen die tekst bevatten) ondersteunen, is het automatiseren van deze taken simpel. Voor wie meer mogelijkheden wenst (zoals het kunnen aangeven van een willekeurig getal of bepaalde letters) is het aan te raden om een blik te werpen op zogenoemde reguliere uitdrukkingen (*regular expressions*): deze manier om zoektermen op een gedetailleerde manier te kunnen specificeren, wordt eveneens door zeer veel gangbare programma's ondersteund. Ook andere manifeste kenmerken van een tekst, zoals zijn lengte, kunnen in de bekende standaardprogramma's vastgesteld worden. Er is in feite geen goede reden om zulke kenmerken handmatig te meten. Specifieke technische kennis is niet of nauwelijks nodig.

***Vraagstukken waarvoor geautomatiseerde technieken van toegevoegde waarde zouden kunnen zijn***

Framing is ook voor de Nederlandse taal nog altijd een prominent onderzoeksgebied, zoals een inventarisatie van inhoudsanalytische studies die het afgelopen decennium verschenen zijn in het *Tijdschrift voor Communicatiewetenschap* uitwijst. Het merendeel van de studies gaat daarbij uit van een handmatige benadering (Deprez, Raeymaeckers & Van Leuven, 2011; Joris, d'Haenens, Van Gorp & Vercruysse, 2013; Van Gorp & Van der Goot, 2009). Zoals in dit artikel beschreven, zijn verschillende technieken om frames te identificeren met succes toegepast op Nederlandstalige data. Dit geldt zowel voor top-downbenaderingen, waarin frames van tevoren gespecificeerd worden, als voor bottom-upbenaderingen, waarin frames uit de data zelf naar voren moeten komen. Gezien de voordelen met betrekking tot kosten en schaalvergroting en het feit dat de technieken steeds beter toegankelijk en toepasbaar zijn, is het framingonderzoekers aan te bevelen om een geautomatiseerde benadering te overwegen.

Voor het Vlaams medialandschap worden regelmatig diversiteitskwesties aangekaart. De interesse gaat hierbij bijvoorbeeld uit naar gender (De Swert & Hooghe, 2010; De Vuyst, Vertoont & Van Bauwel, 2016; Vandenberghe, d'Haenens & Van Gorp, 2015) of etniciteit (Vandenberghe et al., 2015). Voor veel van de variabelen die in diversiteitsonderzoeken centraal staan, geldt dat ze in meer of mindere mate manifest zijn en de codering ervan derhalve prima uitbesteed kan worden. Zoals Jonkman en collega's

hebben laten zien, zouden actoren bijvoorbeeld geautomatiseerd kunnen worden gefilterd (Jonkman et al., 2016). Ook kenmerken van actoren als gender, beroep en leeftijd zouden via een combinatie van *machine learning* en woordenlijstbenaderingen herleid kunnen worden, evenals het centrale thema van het artikel. Als het om bekende actoren gaat, kunnen sommige variabelen zelfs automatisch bij websites als Wikipedia worden opgevraagd (Burggraaff & Trilling, 2017).

Desalniettemin vergt het gebruik van deze technieken wel specifieke kennis. Hoewel er enkele *machine learning*-omgevingen beschikbaar zijn die via een grafisch gebruikersinterface bediend kunnen worden (zoals KNIME), is het gebruikelijker om hiervoor een programmeertaal zoals Python of R te gebruiken, wat vooral te maken heeft met de uitstekende beschikbaarheid van *machine learning*-pakketten voor deze programmeertalen. Het gebruik ervan is niet veel ingewikkelder dan het schrijven van syntax voor statistische programma's als SPSS of STATA. Toch is dit een vaardigheid die (nog) minder wijdverspreid is binnen de communicatiewetenschap. Gezien het toenemende aanbod van dergelijke vakken of bijscholingscursussen op Nederlandse en Vlaamse universiteiten is het echter te verwachten dat dit op korte termijn zal verbeteren.

#### *Vraagstukken waarvoor geautomatiseerde technieken (vooralsnog) geen toegevoegde waarde bieden*

Het is belangrijk om te benadrukken dat (vooralsnog) lang niet alle inhoudsanalyses geautomatiseerd kunnen worden. Dit geldt vooral voor analyses waar grote nadruk wordt gelegd op interpretatie of domeinkennis, zoals een discoursanalyse. Het is daarnaast geen toeval dat dit overzichtsartikel uitsluitend voorbeelden van tekstuele analyses heeft aangehaald: geautomatiseerde technieken om visuele beelden te analyseren zijn nog in een vroeg stadium van ontwikkeling. Hoewel op het gebied van *computer vision* in de afgelopen jaren indrukwekkende vorderingen zijn gemaakt, sluiten typische communicatiewetenschappelijke vraagstellingen vaak niet aan bij de beschikbare technieken. Gechargeerd gesteld: hoe indrukwekkend het ook is dat zulke systemen automatisch zeilboten, katten en auto's kunnen herkennen, dit is vaak maar van marginale interesse voor communicatiewetenschappelijk onderzoek. Aan de andere kant zijn er wel degelijk toepassingen te bedenken van *computer vision*-methoden in de communicatiewetenschap. Het is goed voorstelbaar dat in de nabije toekomst visuele kenmerken van bijvoorbeeld personalisatie of sensatie in nieuwsberichten op betrouwbare wijze door software te herkennen is. Dit zou niet alleen de analyse van visuele media aanmerkelijk minder arbeidsintensief maken, maar ook letterlijk grensoverschrijdend onderzoek binnen handbereik brengen: in tegenstelling tot taal zijn beelden tenslotte universeel.

## Hoe verder? Enkele voorstellen voor onderzoek én onderwijs

Welke stappen kunnen genomen worden om de ontwikkeling en het gebruik van geautomatiseerde methoden verder te bevorderen? Ten eerste lijkt het met het oog op de toekomstige generatie onderzoekers noodzakelijk om de curricula van de opleidingen communicatiewetenschap te herzien. Zoals we hebben aangegeven, kunnen veel inhoudsanalytische standaardtaken met minimale kennis en met behulp van bestaande programma's geautomatiseerd worden. Desalniettemin wordt dit – voor zover ons bekend – niet onderwezen in reguliere methodevakken. Naast het aanbieden van specifieke cursussen gewijd aan geautomatiseerde methoden, zou het waardevol zijn om basisvaardigheden op te nemen in reeds bestaande, meer algemene cursussen. In een statistische cursus waarin SPSS wordt geïntroduceerd, zou bijvoorbeeld een bijeenkomst aan *string*-functies binnen SPSS gewijd kunnen worden, en in een inhoudsanalysevak een bijeenkomst aan *regular expressions*. Op deze wijze zouden studenten in staat worden gesteld om eenvoudige geautomatiseerde inhoudsanalyses uit te voeren.

Het is niet reëel om te verwachten dat elke bachelorstudent in de nabije toekomst ingewikkelde technieken zoals *machine learning* beheerst. Maar dergelijke meer geavanceerde technieken zijn wel uitermate geschikte materie als keuzevakken voor masterstudenten en promovendi, wat uiteindelijk tot een bredere verspreiding van kennis en vaardigheden leidt. Naast het bevorderen van kennis en vaardigheden is er ook meer methodologisch onderzoek nodig, waarvoor we hieronder enkele specifieke suggesties willen geven.

### *Validatiestudies*

Mede door het nog vrij geringe aantal Nederlandstalige geautomatiseerde inhoudsanalyses is het niet altijd even goed in te schatten hoe betrouwbaar de uitkomsten zijn. Niet voor niets benoemen Grimmer en Stewart (2013) '*validate, validate, validate*' (p. 3) als een van de principes van automatische inhoudsanalyse: omdat de oorspronkelijke teksten in het analyseproces niet meer door menselijke codeurs gelezen worden, kunnen fouten en tegenstrijdigheden makkelijk over het hoofd worden gezien.

Een voorbeeld van een gebied waar validatie achterloopt, is sentimentanalyse. Terwijl er – mede door het grote aantal beschikbare algoritmes – meerdere studies zijn gedaan die Engelstalige sentimentanalyse-algoritmes met elkaar vergelijken (bijv. Gonzalez-Bailon & Paltoglou, 2015), is ons geen studie bekend die dit op een systematische manier voor Nederlandstalige algoritmes heeft gedaan. Sterker nog: voor sommige in meerdere talen beschikbare algoritmes is de Nederlandse variant – in tegenstelling tot de oorspronkelijke Engelse versie – nooit formeel gevalideerd (bijv. Thelwall et al., 2010). Op dit vlak zouden onderzoek en onderwijs elkaar kunnen ontmoeten: replicatie en validatie van bestaande onderzoeken is een uitstekende wijze om methoden te leren beheersen.

***Methodeontwikkeling voor comparatief onderzoek***

Een ander, niet te onderschatten probleem is het automatiseren van comparatieve, meertalige studies. Terwijl internationaal vergelijkende manuele inhoudsanalyses heel gangbaar zijn en – zeker als de codeurs meerdere talen spreken – slechts beperkt problemen opleveren, zijn geautomatiseerde meertalige inhoudsanalyses nog vrijwel onontgonnen gebied. Woordenlijstgebaseerde methoden kunnen tot op zekere hoogte met vertalingen werken (zoals bijv. Van der Pas, 2014), maar kunnen – afhankelijk van het onderwerp – ook snel tegen hun grenzen aanlopen. Dit geldt met name als uitdrukkingen of concepten in een taal niet bestaan (zo kent het Duits geen gebruikelijke uitdrukking voor '[niet-]westerse allochtoon') of wanneer de betekenis van een woord verschillend is voor twee talen (zo omvat de betekenis van het woord 'overheid' veel meer dan de – volgens de Van Dale correcte – Duitse vertaling '*Behörden*').

Waar voor woordenlijstbenaderingen nog geldt dat zulke problemen beperkt kunnen worden door met zorgvuldig samengestelde zoektermen en -regels te werken, gaat dit voor *machine learning*-methoden niet op: deze zijn tenslotte expliciet gebaseerd op de woorden in de te analyseren teksten zelf. Stel dat een onderzoeker met behulp van *topic models* de verslaggeving over een specifiek thema in twee verschillende landen wil vergelijken. De patronen die in een dataset in taal A zijn vastgesteld, kunnen niet zonder meer gerelateerd worden aan patronen in een dataset in taal B. Sterker nog, het is geenszins zeker dat in beide datasets überhaupt vergelijkbare *topics* worden gevonden. Sommigen hebben voorgesteld om dit probleem te omzeilen door de teksten eerst automatisch te vertalen naar één taal (bijv. het Engels). Voor het schatten van een *topic model* blijkt dit inmiddels goed te werken (Lucas et al., 2015; Schumacher, Schoonvelde, Traber, Dahiya & De Vries, 2016). Desalniettemin blijft meer validerend onderzoek van groot belang om tot standaarden en *best practices* te komen.

***Subtielere concepten meten***

Een derde onderzoekslijn zou het ontwikkelen van betrouwbare methoden voor het geautomatiseerd coderen van impliciete en subtiele constructen kunnen zijn. Veel van de voorbeelden die we hebben laten zien, meten concepten waarbij enigszins duidelijk is waarom 'het werkt'. Zelfs in een geavanceerd voorbeeld dat gebruik maakt van complexe statistische methoden, zoals het herkennen van frames met behulp van *supervised machine learning* (Burscher et al., 2014), is de achterliggende assumptie vrij simpel: men veronderstelt dat het gebruik van frames zich zal vertalen in een bepaalde woordkeuze, en dientengevolge moet het in principe mogelijk zijn om een statistisch model te schatten dat op basis van woordfrequenties frames 'herkent'.

Veel lastiger is het om abstracte concepten als argumenten te herkennen of zelfs een heel discours op een gestructureerde manier weer te geven en te herkennen of een bepaald stuk tekst betrekking heeft op een eerder gegeven argument, en zo ja, of de schrijver het ermee eens of oneens is. Hiervoor dient men met veel meer *features* dan alleen woordfrequenties rekening te houden (zie bijv. Stab & Gurevych, 2014, die ook

rekening houden met onder ander grammaticale *features*) – en dit kan nogal problematisch zijn in talen waarin minder pakketten voor natuurlijke-taalverwerking beschikbaar zijn.

## Conclusie

In dit artikel hebben we geprobeerd een overzicht te geven van de huidige stand van de toepassing van geautomatiseerde inhoudsanalyses op Nederlandstalige teksten. We hebben laten zien dat ondanks enkele taalspecifieke uitdagingen veel technieken ook toepasbaar zijn op Nederlandstalige teksten, maar ook dat deze mogelijkheden nog niet voldoende worden benut.

We hebben een breed scala aan methoden en technieken besproken. Het is echter wel belangrijk te realiseren dat de scheidslijnen ertussen vaak minder sterk zijn dan op basis van dit overzicht wellicht verwacht wordt. Waar vroege automatische inhoudsanalyses vooral gebruik maakten van één techniek (zoals het tellen van woorden), worden tegenwoordig namelijk vaak meerdere technieken gecombineerd. Zo analyseert Jacobi (2016) nieuwsberichten zowel op basis van zoektermen met behulp van het geautomatiseerd ontleden van zinnen (*part-of-speech tagging*), alsook op basis van een *topic model*, en combineren Burggraaff en Trilling (2017) onder meer *named entity recognition*, *supervised machine learning* en sentimentanalyse om nieuwswaardes in Nederlandse nieuwsberichten te identificeren. Welke stukken gereedschap uit de gereedschapskist relevant zijn om gecombineerd te worden, verschilt per onderzoeksvraag en dataset – maar daarnaast ook per taalgebied: zo zijn zinnen in sommige talen moeilijker te ontleden dan in andere talen, en de kwaliteit en beschikbaarheid van algoritmen voor dit doel variëren.

Ten slotte willen we graag benadrukken dat geautomatiseerde inhoudsanalyse als onderzoekstechniek in de Nederlandse taal een vlucht zou nemen wanneer er meer samenwerkingsverbanden en gemeenschappelijke platforms voor de deling van kennis en *tools* worden gecreëerd. Dit kan in de vorm van een digitale infrastructuur, zoals de vakgroep Communicatiewetenschap van de Vrije Universiteit Amsterdam al jaren faciliteert in de vorm van AmCAT. Deze infrastructuur biedt verschillende datasets, uitgebreide zoekfuncties en mogelijkheden om data te visualiseren en analyseren. Ook op de Universiteit van Amsterdam is recent een infrastructuur opgezet die het uitvoeren van geautomatiseerde inhoudsanalyses faciliteert. Daarnaast zou samenwerking en kennisdeling kunnen worden gestimuleerd met de oprichting van een interessegroep en de organisatie van workshops of seminars. Dergelijke initiatieven kunnen eraan bijdragen de Nederlandse en Vlaamse reputatie op het terrein van inhoudsanalyse hoog te houden.

## Noot

1 Er lijkt echter weer beweging te komen in het onderhouden van dit pakket: in augustus 2017 werd Pattern geporteerd van Python 2 naar Python 3, zie https://github.com/clips/pattern/tree/development.

## Literatuur


Aaldering, L., & Vliegenthart, R. (2016). Political leaders and the media: Can we measure political leadership images in newspapers using computer-assisted content analysis? *Quality & Quantity, 50*(5), 1871-1905. doi:10.1007/s11135-015-0242-9

Bartholomé, G., Lecheler, S., & De Vreese, C. (2017). Towards a typology of conflict frames. *Journalism Studies, online first*. doi:10.1080/1461670X.2017.1299033

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly.

Boumans, J. (2017). Subsidizing the news? Organizational press releases' influence on news media's agenda and content. *Journalism Studies, online first*. doi:10.1080/1461670X.2017.1338154

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism, 4*(1), 8-23. doi:10.1080/21670811.2015.1096598

Burggraaff, C., & Trilling, D. (2017). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism, online first*. doi:10.1177/1464884917716699

Burscher, B. (2016). *Machine learning-based content analysis: Automating the analysis of frames and agendas in political communication research*. Proefschrift, Universiteit van Amsterdam. Opgehaald van http://hdl.handle.net/11245/1.542098

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures, 8*(3), 190-206. doi:10.1080/19312458.2014.937527

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The ANNALS of the American Academy of Political and Social Science, 659*(1), 122-131. doi:10.1177/0002716215569441

De Graaf, R., & Van der Vossen, R. (2013). Bits versus brains in content analysis. Comparing the advantages and disadvantages of manual and automated methods for content analysis, *Communications: The European Journal of Communication Research, 38*(4), 433-443. doi:10.1515/commun-2013-0025

Deprez, A., Raeymaeckers, K., & Van Leuven, S. (2011). Framing van de Eerste en Tweede Intifada in de Vlaamse en Nederlandse pers. *Tijdschrift voor Communicatiewetenschap, 39*(1), 21-43.

De Smedt, T., Daelemans, W. (2012). Pattern for Python. *The Journal of Machine Learning Research, 13*, 2063-2067.

De Swert, K., & Hooghe, M. (2010). When do women get a voice? Explaining the presence of female news sources in Belgian news broadcasts. *European Journal of Communication, 25*(1), 69-84. doi:10.1177/0267323109354229

De Vuyst, S., Vertoont, S., & Van Bauwel, S. (2016). Sekse-ongelijkheid in Vlaams nieuws. Een kwantitatieve inhoudsanalyse naar de aanwezigheid en hoedanigheid van vrouwen en mannen in Vlaamse nieuwsverhalen. *Tijdschrift voor Communicatiewetenschap, 44*(3), 253-271.

Gonzalez-Bailon, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 95-107. doi:10.1177/0002716215569192

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297. doi:10.1093/pan/mps028

Günther, E., & Quandt, T. (2016). Word counts and topic models. *Digital Journalism, 4*(1), 75-88. doi:10.1080/21670811.2015.1093270

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International AAAI Conference on Weblogs and Social Media*, 216-225. Opgehaald van http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

Jacobi, C. A. (2016). *The quality of political news in a changing media environment*. Proefschrift, Universiteit van Amsterdam. Opgehaald van http://hdl.handle.net/11245/1.503897

Jijkoun, V., & Hofmann, K. (2008). *Task-based evaluation report: Building a Dutch subjectivity lexicon*. ILPS-ISLA, University of Amsterdam.

Jonkman, J. G., Trilling, D., Verhoeven, P., & Vliegenthart, R. (2016). More or less diverse: An assessment of the effect of attention to media salient company types on media agenda diversity in Dutch newspaper coverage between 2007 and 2013. *Journalism, online first*. doi:10.1177/1464884916680371

Joris, W., d'Haenens, L., Van Gorp, B., & Vercruysse, T. (2013). De eurocrisis in het nieuws. *Tijdschrift voor Communicatiewetenschap, 41*(2), 162.

Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications, 39*(14), 11616-11622. doi:10.1016/j.eswa.2012.04.013

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Kroon, A. C. (2017). Biased media? Explaining age discrimination claims with media stereotypes. Paper gepresenteerd bij de jaarlijkse conferentie van de *International Communication Association*. San Diego, CA.

Kruikemeier, S., Van Noort, G., Vliegenthart, R., & De Vreese, C. (2015). Nederlandse politici op Twitter: wie, waarover, wanneer en met welk effect? *Tijdschrift voor Communicatiewetenschap, 43*(1), 4-22.

Kuiken, J., Schuth, A., Spitters, M., & Marx, M. (2017). Effective headlines of newspaper articles in a digital environment. *Digital Journalism, online first*. doi:10.1080/21670811.2017.1279978

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly, 92*(4), 791-811. doi:10.1177/1077699015607338

Leydesdorff, L., & Nerghes, A. (2016). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N < 1,000). *Journal of the Association for Information Science and Technology, 14*(4), 90-103. doi:10.1002/asi.23740

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254-277. doi:10.1093/pan/mpu019

Rooduijn, M., & Pauwels, T. (2011). Measuring populism: Comparing two methods of content analysis. *West European Politics, 34*(6), 1272-1283. doi:10.1080/01402382.2011.616665

Ruigrok, N., & van Atteveldt, W. (2007). Global angling with a local angle: How U.S., British, and Dutch newspapers frame global and local terrorist attacks. *The Harvard International Journal of Press/Politics, 12* (1), 68-90. doi:10.1177/1081180X06297436

Schumacher, G., Schoonvelde, M., Traber, D., Dahiya, T., & De Vries, E. (2016). EUSpeech: A new dataset of EU elite speeches. doi:10.7910/DVN/XPCVEI

Stab, C., & Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 46-56. Opgehaald van http://www.aclweb.org/anthology/D14-1006

Strauß, N., Vliegenthart, R., & Verhoeven, P. (2016). Lagging behind? Emotions in newspaper articles and stock market prices in the Netherlands. *Public Relations Review, 42*(4), 548-555. doi:10.1016/j.pubrev.2016.03.010

Strycharz, J., & Strauß, N., & Trilling, D. (2017). Media coverage and share price volatility: Is it only attention that matters? *International Journal of Strategic Communication, online first*. doi:10.1080/1553118X.2017.1378220

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized

text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54. doi:10.1177/0261927X09351676

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544-2558. doi:10.1002/asi.21416

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *CONLL '02 Proceedings of the sixth conference on Natural Language Learning* (pp. 155-158). doi:10.3115/1118853.1118877

Trilling, D., Tolochko, P., & Burscher, B. (2017). From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & Mass Communication Quarterly, 94*(1), 38-60. doi:10.1177/1077699016654682

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). A Dictionary-based approach to racism detection in Dutch social media. *Proceedings of the Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)*, 11-17. Opgehaald van http://www.clips.ua.ac.be/bibliography/a-dictionary-based-approach-to-racism-detection-in-dutch-social-media

Van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Political Analysis, 16*(4), 428-446. doi:10.1093/pan/mpn006

Van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (2008). Good news or bad News? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics, 5*(1), 73-94. doi:10.1080/19331680802154145

Vandenberghe, H., d'Haenens, L., & Van Gorp, B. (2015). Demografische diversiteit in het Vlaamse perslandschap. *Tijdschrift voor Communicatiewetenschap, 43*(2), 169-185.

Van der Meer, G. L. A. (2014). Organizational crisis-denial strategy: The effect of denial on public framing. *Public Relations Review, 40*(3), 537-539. doi:10.1016/j.pubrev.2014.02.005

Van der Meer, G. L. A., & Verhoeven, P. (2013). Public framing organizational crisis situations: Social media versus news media. *Public Relations Review, 39*(3), 229-231. doi:10.1016/j.pubrev.2012.12.001

Van der Meer, G. L. A., Verhoeven, P., Beentjes, H., & Vliegenthart, R. (2014). When frames align: The interplay between PR, news media, and the public in times of crisis. *Public Relations Review, 40*(5), 751-761. doi:10.1016/j.pubrev.2014.07.008

Van der Pas, D. (2014). Making hay while the sun shines. *The International Journal of Press/Politics, 19*(1), 42-65. doi:10.1177/1940161213508207

Van Gorp, B., & Van der Goot, M. (2009). Van Frankenstein tot de Goede Moeder: de inzet van frames in de strategische communicatie over duurzaamheid. *Tijdschrift voor Communicatiewetenschap, 37*(4), 303-316.

Van Noije, L., Kleinnijenhuis, J., & Oegema, D. (2008). Loss of parliamentary control due to mediatization and Europeanization: A longitudinal and cross-sectional analysis of agenda building in the United Kingdom and the Netherlands. *British Journal of Political Science, 38*(3), 455-478. doi:10.1017/S0007123408000239

Van Noord, G. (2006). At last parsing is now operational. *Verbum Ex Machina (TALN06). Actes de la 13e Conference sur le Traitement Automatique des Langues Naturelles*, 20-42.

Van Zoonen, W., & Van der Meer, G. L. A. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behavior, 63*, 132-141. doi:10.1016/j.chb.2016.05.028

Vliegenthart, R., Boomgaarden, H. G., & Boumans, J. (2011). Changes in political news coverage: Personalisation, conflict and negativity in British and Dutch newspapers. In K. Brants & K. Voltmer (Eds.), *Challenging the Primacy of Politics* (pp. 92-110). London, UK: Palgrave Macmillan.

Vliegenthart, R., & Roggeband, C. (2007). Framing immigration and integration: Relationships between press and parliament in the Netherlands. *International Communication Gazette, 69*(3), 295-319. doi:10.1177/1748048507076582

Vliegenthart, R., & Van Aelst, P. (2010). Nederlandse en Vlaamse politieke partijen in de krant en in de peilingen: een wederkerige relatie. *Tijdschrift voor Communicatiewetenschap, 38*(4), 338-356.

Vlieger, E., & Leydesdorff, L. (2012). Content analysis and the measurement of meaning: The visualization of frames in collections of messages. In M. Mora, O. Gelman, A. Steenkamp, & M. S. Raisinghani (Eds.), *Research methodologies, innovations and philosophies in systems engineering and information systems* (pp. 322-340). Hershey, PA: Information Science Reference.

Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., … Segers, R. (2016). NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems, 110*, 60-85. doi:10.1016/j.knosys.2016.07.013

Welbers, K., van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2016). A gatekeeper among gatekeepers. *Journalism Studies, online first*. doi:10.1080/1461670X.2016.1190663

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29*(2), 205-231. doi:10.1080/10584609.2012.671234

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J. W., & Geenen, R. (2004). De Nederlandse versie van de 'Linguistic Inquiry and Word Count' (LIWC). Een gecomputeriseerd tekstanalyseprogramma. *Gedrag & Gezondheid, 32*(4), 271-281.