



Onderzoek

Gebruik van sociale media

Verkenkend onderzoek naar de mogelijkheden van het gebruik van sociale media voor statistiek voor bedrijven

Shirley Ortega
Nico Heerschap

April 2019

Inhoud

Samenvatting 3

Dankwoord 5

1. Inleiding 5

1.1 Achtergrond 8

1.2 Uitwerking probleemstelling in deelvragen 8

1.3 Definities 9

1.4 Gebruikte middelen en software 10

1.5 Leeswijzer 10

2. Verwerkingsproces van de bronnen 11

2.1 Gebruikte bronnen 11

2.2 Verwijderen van uitbijters in de Coosto-datasets 14

2.3 Opschonen en aanpassen van de databronnen 15

2.4 Koppelen van de databronnen 16

2.5 Beschikbare statistische eenheden 17

2.6 Omzetting van korte naar lange URL's: afleiding van domeinnamen 18

2.7 Output 18

3. Resultaten 19

3.1 Inleiding 19

3.2 Resultaten 21

4. Conclusies en discussie 52

4.1 Algemeen 52

4.2 Belangrijkste conclusies 53

4.3 Discussie 54

5. Referenties 55

Bijlagen 56

Samenvatting

Deze notitie beschrijft de resultaten van een verkennend onderzoek naar het gebruik van sociale mediaberichten voor het maken van officiële statistiek en dan met name indicatoren voor bedrijven. Het onderzoek is gebaseerd op data die via de API van het bedrijf Coosto zijn verzameld gedurende een periode van zes weken in de maanden juni, juli en augustus 2017. In die periode zijn 1,8 miljoen sociale mediaberichten gedownload. Deze berichten hebben als kenmerk dat ze minstens een URL (internetverwijzing) bevatten.

Dit bestand is via de koppelsleutel 'profielnaam' gekoppeld aan een bestand van het bedrijf Dataprovider met alle websites van het Nederlandse internetdomein. Het bestand van Dataprovider was al eerder gekoppeld aan het Algemeen Bedrijvenregister van het CBS. Dit betekent ten eerste dat het gekoppelde analysebestand alleen nog sociale mediaberichten bevatte van bedrijven en niet meer van personen. Van de 1,8 miljoen sociale mediaberichten bleven er uiteindelijk maar circa 120 duizend berichten over, die betrekking hadden op bedrijven. Daarnaast betekent dit dat aan het gekoppelde analysebestand ook de variabelen bedrijfseenheid, grootteklasse en economische activiteit zijn toegevoegd. Om de genoemde bestanden te kunnen koppelen en gebruiken dienden extra bewerkingen te worden uitgevoerd.

In de verkennende analyses zijn onder meer de volgende onderzoeksvragen onderzocht:

- Zijn sociale mediaberichten en profielnamen relevante statistische eenheden voor de output?
- Is het mogelijk om het bestand met sociale mediaberichten te gebruiken voor het vervangen van vragen van bestaande statistiek, zoals de ICT-enquête bedrijven?
- Welke economische activiteiten en grootteklassen maken gebruik van welke sociale mediatypen?
- Naar welke domeinnamen of websites van bedrijven wordt vooral verwezen in sociale mediaberichten?
- Van welke typen sociale mediaberichten wordt door bedrijven vooral gebruikgemaakt?
- Over welke onderwerpen wordt door bedrijven vooral gepost?
- Wat is de reikwijdte van sociale mediaberichten?

De belangrijkste algemene conclusies van de uitgevoerde verkenningen zijn:

- Het gebruik van sociale mediaberichten heeft tot op zekere hoogte potentie om als basis te worden gebruikt voor het maken van officiële statistiek, zoals bijvoorbeeld de vraag naar het gebruik van sociale mediatypen door bedrijven, eventueel uitgesplitst naar economische activiteit en grootteklasse. Het zou daarmee kunnen bijdragen aan de verlaging van de enquêtedruk bij bedrijven, nieuwe indicatoren en een verbetering van de kwaliteit van de uitkomsten.
- Door een koppeling met andere bestanden kunnen ook resultaten worden geproduceerd op basis van economische activiteit en grootteklasse. Dit heeft veel meerwaarde.
- Op basis van sociale mediaberichten kunnen dieper gravende analyses worden uitgevoerd dan op basis van gegevens, die verzameld zijn via een enquête. Dieper gravend in de zin van meer detail, sneller en met een grotere reikwijdte.

- Door methoden zoals text en process mining kan ook iets worden gezegd over de inhoud van sociale mediaberichten, de richting en intensiteit, maar ook de actoren, die in thema's een rol spelen.
- Vooral Twitter wordt gebruikt door bedrijven en die berichten hebben vooral betrekking op de communicatie rond vacatures.

Naast bovengenoemde positieve punten en de potentie van het gebruik van sociale mediaberichten voor officiële statistiek, is ook geconstateerd dat er nog een aantal obstakels bestaat, dat eerst moeten worden opgelost voordat sociale mediaberichten ook daadwerkelijk kunnen worden ingezet voor het maken van statistiek. Het is geen 'laaghangend fruit'. De moeilijkste opgave ligt bij het beoordelen van de representativiteit van de verzamelde sociale mediaberichten. Voor de hier uitgevoerde analyses is maar voor een korte periode data verzameld. Onduidelijk is ook of alle berichten van alle sociale mediatypen via de API van Coosto zijn binnengehaald? Bovendien zijn in deze analyse alleen sociale mediaberichten gebruikt die minstens een URL bevatten. Dat maakt bijvoorbeeld een vergelijking met de resultaten van traditionele enquêtes niet gemakkelijk. Ook staan sommige methoden, die bij big dataonderzoek worden gebruikt, nog in de kinderschoenen. Dat geldt bijvoorbeeld voor het gebruik van zogenoemde 'woordwolken', die in dit onderzoek zijn ingezet om naar de inhoud van sociale mediaberichten te kijken. Dit vraagt om verdere ontwikkeling en onderzoek.

Kijkend naar de probleemstelling, kan dus gesteld worden dat er enige potentie is in het gebruik van sociale mediaberichten voor het maken van officiële statistiek, maar dat er nog verder onderzoek moet worden gedaan naar hoe en op welke wijze. Dat vraagt ook om een verdere afstemming met de potentiële gebruikers. Deze notitie laat 'slechts' een eerste stap zien.

Cruciaal is ook dat de verzamelde data met sociale mediaberichten kunnen worden gekoppeld aan andere bronnen, die het mogelijk maken om sociale mediaberichten van bedrijven te onderscheiden van die van personen. Anders verliest de data veel van zijn meerwaarde.

In dit verkennende onderzoek is specifiek gekeken naar (ICT-)indicatoren op het terrein van het gebruik van sociale mediaberichten door bedrijven. Mogelijk levert analyse op andere terreinen, zoals toerisme, eerder succes op. Daar kunnen bijvoorbeeld sociale mediaberichten met een geolocatie en een tijdsaanduiding worden gebruikt om stromen toeristen in kaart te brengen.

Dankwoord

Dit verkennend onderzoek zou niet tot stand zijn gekomen zonder een financiële bijdrage van het ministerie van Economische Zaken en Klimaat. Daarbij gaat vooral dank uit naar de heer Jan Julianus.

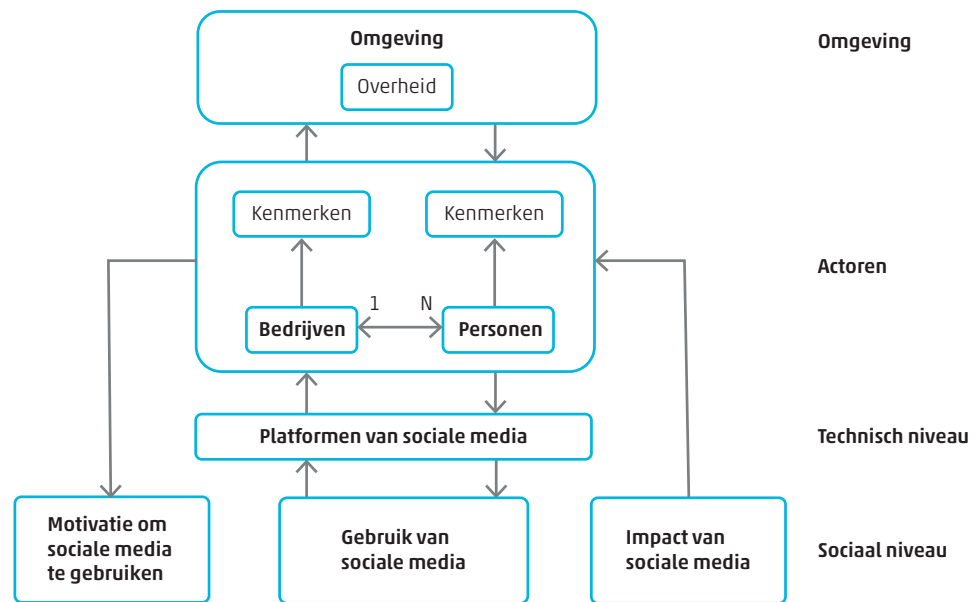
Binnen het CBS danken wij de heer Piet Daas van het Big Datacentrum voor het gebruik van de Coosto-database. We danken de heer Yuri Boskamp voor de verstrekte informatie over de statistiek ICT-gebruik bedrijven.

1. Inleiding

Sociale media, zoals Twitter, LinkedIn, Youtube en Facebook, zijn niet meer weg te denken uit onze samenleving en behoren tot één van de nieuwe en belangrijke digitale ontwikkelingen, die in de laatste 10 jaar snel tot wasdom zijn gekomen¹⁾. In eerste instantie waren sociale media onderdeel van het interpersoonlijke verkeer tussen mensen door het versturen van en het reageren op tekstberichten, foto's, video's en geluid. Door de explosieve groei van het gebruik van sociale media hebben ook bedrijven en de overheid dit fenomeen omarmd. Met deze stap naar 'social business' zijn sociale media gecommercialiseerd (zie figuur 1.0.1). Sociale media zijn nu ook een belangrijk communicatiekanaal geworden van bedrijven en overheid. In het begin lag de aandacht vooral op het gebruik van sociale media voor informatie- en marketingdoeleinden. In de loop van de tijd is de inzet van sociale media door bedrijven echter veel breder geworden. Het gaat nu bijvoorbeeld ook om het onderhouden van netwerken, luisteren naar de omgeving, het genereren van kennis en het opbouwen en behouden van reputaties. Sociale media kunnen in principe een bijdrage leveren aan alle processen van een onderneming, zowel in de relatie met (potentiële) klanten en toeleveranciers als in de relatie met de medewerkers. Sociale media beïnvloeden daarmee de manier en snelheid waarmee wij informatie verzamelen en verspreiden, communiceren en besluiten.

¹⁾ In feite is de basis voor sociale media al eerder gelegd met het introduceren van e-mail.

1.0.1 Sociale media en hun omgeving



Bron: Heerschap en Ortega.

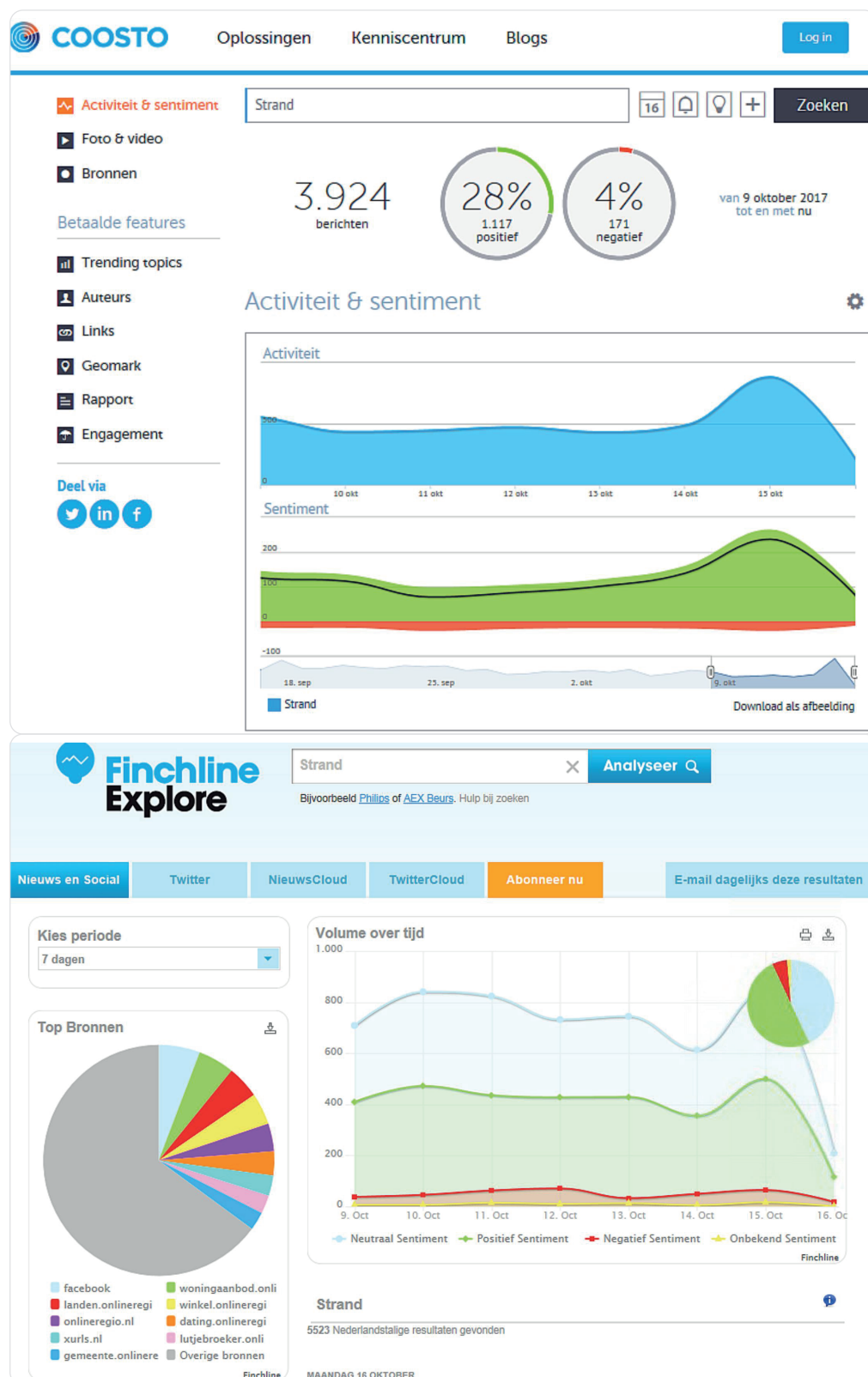
Daarmee genereren niet alleen personen, maar ook bedrijven in toenemende mate grote hoeveelheden zogenoemde sociale media 'posts' of berichten op het internet. De vraag is vervolgens of deze sociale mediaberichten, als digitale voetafdruk, kunnen worden gebruikt om ontwikkelingen bij bedrijven te analyseren via zogenoemde big data-analyses. Dit heeft geleid tot de volgende opdracht en probleemstelling:

Verken in hoeverre sociale mediaberichten gebruikt kunnen worden voor het maken van officiële statistiek. Onderzoek daarbij specifiek het terrein van (ICT-)indicatoren voor bedrijven.

Het gaat hier om een verkennend onderzoek. Omdat met dezelfde data ook andere analyses worden uitgevoerd, betreft het in feite een deelonderzoek.

Om de verkenning te kunnen uitvoeren, moeten de sociale mediaberichten, die in een bepaalde periode op het internet gepost zijn, worden verzameld of 'gescrapet'. Dat wordt gedaan met zogenoemde internetrobots (software scripts). Voor dit onderzoek is er echter voor gekozen om de sociale mediaberichten te gebruiken die al door het bedrijf Coosto (www.coosto.com/nl) zijn gescrapet. Coosto is een bedrijf dat gespecialiseerd is in het scrapen van sociale mediaberichten van het internet. Een vergelijkbaar bedrijf is Finchline HowardsHome (www.howardshome.com). Beide bedrijven verzamelen niet alleen de geposte sociale mediaberichten van het internet, maar faciliteren ook het gebruik ervan door aan onderzoekers [1, 2] gegevens ter beschikking te stellen via monitoringtools (API's) of dashboards. Figuur 1.0.2 is daarvan een voorbeeld. Bij Coosto wordt ook de inhoud van de sociale mediaberichten zelf geanalyseerd. Er wordt namelijk aan de hand van woorden, die in het bericht voorkomen, bepaald of de inhoud van het bericht een positief of negatief sentiment weergeeft. Vooralsnog lijkt deze analyse niet overal even succesvol.

1.0.2 Twee voorbeelden van een dashboard om sociale mediaberichten te analyseren



Bron: Coosto en Finchline.

1.1 Achtergrond

Volgens het CBS maakte in 2017 in Nederland circa 85 procent van de personen van 12 jaar of ouder op een of andere wijze gebruik van sociale media. Vijf jaar eerder was dat nog maar 57 procent. Bij bedrijven ligt het gebruik van sociale media in 2017 iets lager: 69 procent van de bedrijven met 10 of meer werkzame personen. Vijf jaar geleden was dat 41 procent. Deze sterke groei van het gebruik van sociale media door personen en bedrijven maakt het interessant om te onderzoeken of deze zogenoemde big databron kan worden ingezet voor het maken van officiële statistiek. Daarvoor is een eerste verkenning nodig, waarbij hier de focus ligt op het gebruik van sociale mediaberichten door bedrijven. De centrale vraag is in hoeverre met sociale mediaberichten indicatoren kunnen worden gemaakt, die iets zeggen over de situatie bij en het gedrag van bedrijven?

Als sociale mediaberichten kunnen worden ingezet voor het maken of ondersteunen van officiële statistiek, dan kan dat onder meer bijdragen aan snellere en meer gedetailleerde indicatoren en eventueel nieuwe indicatoren, die met traditionele statistieken minder gemakkelijk te maken zijn. Het kan ook leiden tot enquêtedrukvermindering bij bedrijven als vragen van bijvoorbeeld de 'ICT-enquête bedrijven' van het CBS kunnen worden vervangen door indicatoren, die op basis van sociale mediaberichten kunnen worden samengesteld.

Naast dit onderzoek, zijn de data van Coosto ook gebruikt voor onderzoek naar:

- het samenstellen van een register van domeinnamen (RDNB): zie projectaanpak [3];
- indicatoren op het terrein van toerisme in Nederland [4].

Gebruikte bronnen

Om het onderzoek naar het gebruik van sociale mediaberichten door bedrijven te kunnen uitvoeren, is niet alleen gebruikgemaakt van de Coosto-data. De Coosto-data zijn gekoppeld aan een bestand met alle websites van het Nederlandse internetdomein. Deze data zijn afkomstig van het bedrijf Dataprovider (jaar 2015). Alleen op deze manier was het mogelijk om sociale mediaberichten van bedrijven te onderscheiden van die van personen. Zie voor een beschrijving van de koppeling van deze twee bronnen verder hoofdstuk 2.

1.2 Uitwerking probleemstelling in deelvragen

De centrale onderzoeksvraag in dit onderzoek is in hoeverre sociale mediaberichten te gebruiken zijn als databron voor het maken van (ICT-)indicatoren voor bedrijven of meer algemeen: voor het maken van officiële statistiek. Meer specifiek gaat het onder meer om de volgende deelvragen [3]:

- Zijn sociale mediaberichten te gebruiken voor het vervangen van (ICT-)indicatoren van bedrijven van de enquête 'ICT-gebruik bedrijven'?
- Welke bedrijven maken gebruik van welk type sociale mediaberichten?
Te onderscheiden naar economische activiteit (SBI) en grootteklasse.
- In hoeverre is een sociaal mediabericht of -profiel of account een relevante statistische eenheid?
- Naar welke domeinnamen (websites) van bedrijven wordt in sociale mediaberichten het meest verwezen?

- Welke onderwerpen komen vooral voor in sociale mediaberichten, die door bedrijven zijn gepost?
- Van welke typen sociale mediaberichten, posts of reposts, maken bedrijven vooral gebruik?
- Wat is de reikwijdte van sociale mediaberichten, in de vorm van tweets (actief) en retweets (reactief)?

Deze notitie bevat een beschrijving van de aanpak van het onderzoek, de resultaten en, tenslotte, een korte discussie over de uitkomsten.

1.3 Definities

In het rapport zijn de volgende definities gebruikt:

Big Data

Microsoft: 'Big data is de term die wordt gebruikt voor de verwerking van grote en complexe datasets met processen, waarvoor veel rekenkracht nodig is. Daarbij gaat het ook om processen zoals machine learning en kunstmatige intelligentie' [5].

Of met een meer technisch perspectief van de National Institute of Standards and Technology: 'Gegevens die de capaciteit van de huidige conventionele methoden en systemen overschrijden' [5].

Sociale media, gezien vanuit de statistiek ICT-gebruik bedrijven

Het gebruik van internettoepassingen of communicatieplatforms voor het contact met klanten, leveranciers, zakenrelaties of medewerkers van het eigen bedrijf. Daarbij wordt online informatie uitgewisseld. Voorbeelden zijn Twitter, Facebook, YouTube, LinkedIn, Instagram e.d. Bedrijven die sociale media gebruiken hebben een account, een gebruikersprofiel of een gebruikerslicentie.

Sociale media, gezien vanuit de statistiek ICT-gebruik personen

Als vraagstelling: Heeft u internet in de afgelopen 3 maanden gebruikt voor de volgende activiteiten:

- Tekstberichten uitwisselen.
- Deelname aan een sociaal netwerk zoals Facebook en Twitter.
- Deelname aan een professioneel netwerk zoals LinkedIn.

Bedrijfseenheid (BE_ID)

Een bedrijfseenheid is de feitelijke actor in het productieproces die gekenmerkt wordt door autonomie, beschrijfbaarheid en externe gerichtheid. Een bedrijfseenheid heeft in het bedrijvenregister van het CBS (ABR) als unieke identificatie een BE_ID. De BE_ID is de belangrijkste eenheid in het ABR. Op basis hiervan zijn bijvoorbeeld gegevens beschikbaar over de grootteklasse (aantal werkzame personen) en economische activiteit (bedrijfstak) van een bedrijf of BE_ID.

1.4 Gebruikte middelen en software

Om de analyses op de grote hoeveelheid data te kunnen uitvoeren moest extra hard- en software worden ingezet. Daarbij ging het onder meer om:

1. Hardware:

- Een laptop met Windows x86 en een processor core i7, 32-bit.
- Remote Desktop Services (RDS Werk): een omgeving, waarbij extra 'on-demand verwerkingscapaciteit' kan worden opgeroepen en gebruikt. Dat leverde de benodigde rekenkracht op.
Hiermee kon bijvoorbeeld een matrix van 1.8×10^6 rijen (records) x 20 kolommen wordt ingelezen en verwerkt. De verwerkingscapaciteit van de laptop was daarvoor onvoldoende.

2. Software:

- R, versie 3.4.2 en Rstudio, versie 1.0.1.153 (laptop).
- R, versie 3.2.3. en RStudio, versie 0.99 (RDS Werk).

1.5 Leeswijzer

In hoofdstuk 2 is de dataverzameling, de koppeling en het verwerkingsproces van de gebruikte bronnen beschreven. Dit proces heeft geleid tot een basisbestand, waarop de meeste analyses zijn uitgevoerd. De resultaten van de analyses zijn beschreven in hoofdstuk 3. Hoofdstuk 4 bevat de belangrijkste conclusies van het onderzoek en een discussie. De notitie wordt, tenslotte, afgesloten met de belangrijkste referenties.

2. Verwerkingsproces van de bronnen

In dit hoofdstuk worden de dataverzameling, het verwerkingsproces en de koppeling van de gebruikte bronnen beschreven. Daarbij zijn de volgende vijf stappen doorlopen (zie ook bijlage III):

- Het verzamelen van de sociale mediaberichten via de API van Coosto.
- Het verwijderen van uitbijters in de Coosto-data.
- Het aanpassen van de twee databronnen voor zowel de koppeling als de latere analyses.
- De uiteindelijke koppeling van de Coosto-data met sociale mediaberichten aan het bestand met websites van bedrijven van Dataprovider.
- Het definiëren van de statistische eenheden, die bij de analyses kunnen worden betrokken.

Dit proces heeft geresulteerd in een gekoppeld analysebestand, dat de basis is geweest voor de meeste in hoofdstuk 3 beschreven verkenningen.

2.1 Gebruikte bronnen

Voor dit verkennend onderzoek naar mogelijke (ICT-)indicatoren voor bedrijven op basis van sociale mediaberichten zijn twee databronnen gebruikt, namelijk:

- de hoofdbron met alle sociale mediaberichten van het bedrijf Coosto;
- het bestand met data van alle Nederlandse websites van het bedrijf Dataprovider.

Het bestand van Coosto met sociale mediaberichten

Via de API-tool van Coosto zijn twee verschillende databestanden met sociale mediaberichten samengesteld. Deze databestanden moeten, min of meer, gezien worden als een steekproef uit alle beschikbare sociale mediaberichten, die Coosto over een bepaalde periode via het internet heeft verzameld. Deze methode is gekozen, omdat de API van Coosto beperkingen kent. De belangrijkste daarvan is dat niet meer dan 10 duizend sociale mediaberichten (of records) per sessie kunnen worden gedownload. Dat leverde uiteindelijk de volgende twee databestanden op:

1. Sociale mediaberichten, die door de week overdag (van 9:30–17:00) zijn verzameld. Om praktische redenen werd op dinsdag slechts een beperkt aantal sessies uitgevoerd. Deze dataset wordt verder de 'chronologische dataset' genoemd. Deze data zijn verzameld²⁾ in de periode 26-06-2017 tot 04-08-2017 (zes weken). Het totaal aantal verzamelde sociale mediaberichten is 1,6 miljoen.
2. Sociale mediaberichten, die op basis van een randomgenerator door de week in de nacht (van 18:00–9:00) en in het weekeinde van vrijdag 17:00 tot maandagochtend 9:00 zijn verzameld³⁾. Deze dataset wordt verder de 'random dataset' genoemd. Deze dataset is in dezelfde periode samengesteld als de chronologische dataset. Totaal aantal sociale mediaberichten betreft 0,2 miljoen.

²⁾ Na elk uur is er een nieuwe sessie opgestart.

³⁾ Hier geldt een maximum van 10 duizend berichten per nacht en weekeinde.

Van belang is dat bij het downloaden van de sociale mediaberichten via de API van Coosto een filter is gebruikt. Alleen die sociale mediaberichten zijn gedownload, waarvan de tekst een URL bevatte met '.com' of '.nl' aan het einde. Dit heeft te maken met het feit dat de data vooral gebruikt zou worden om een register met domeinnamen op te zetten. Uit de gebruikte URL kan namelijk de domeinnaam worden afgeleid.

Het bestand van Coosto bevat in principe alle typen sociale mediaberichten, die tijdens de waarneemperiode op het internet zijn gepost. In de praktijk blijkt het echter vooral te gaan om berichten van Twitter (tweets en retweets) en in mindere mate om Facebook, LinkedIn, Instagram, Pinterest, blogs en nieuws. Daarom is een deel van de analyses in hoofdstuk 3 alleen uitgevoerd op de sociale mediaberichten van Twitter.

Tabel 2.1.1 en figuur 2.1.2 laten de verdeling zien van het aantal sociale mediaberichten per type sociaal medium van de chronologische en random datasets. Hier is te zien dat de meeste sociale mediaberichten Twitterberichten zijn, namelijk meer dan 90 procent van alle berichten.

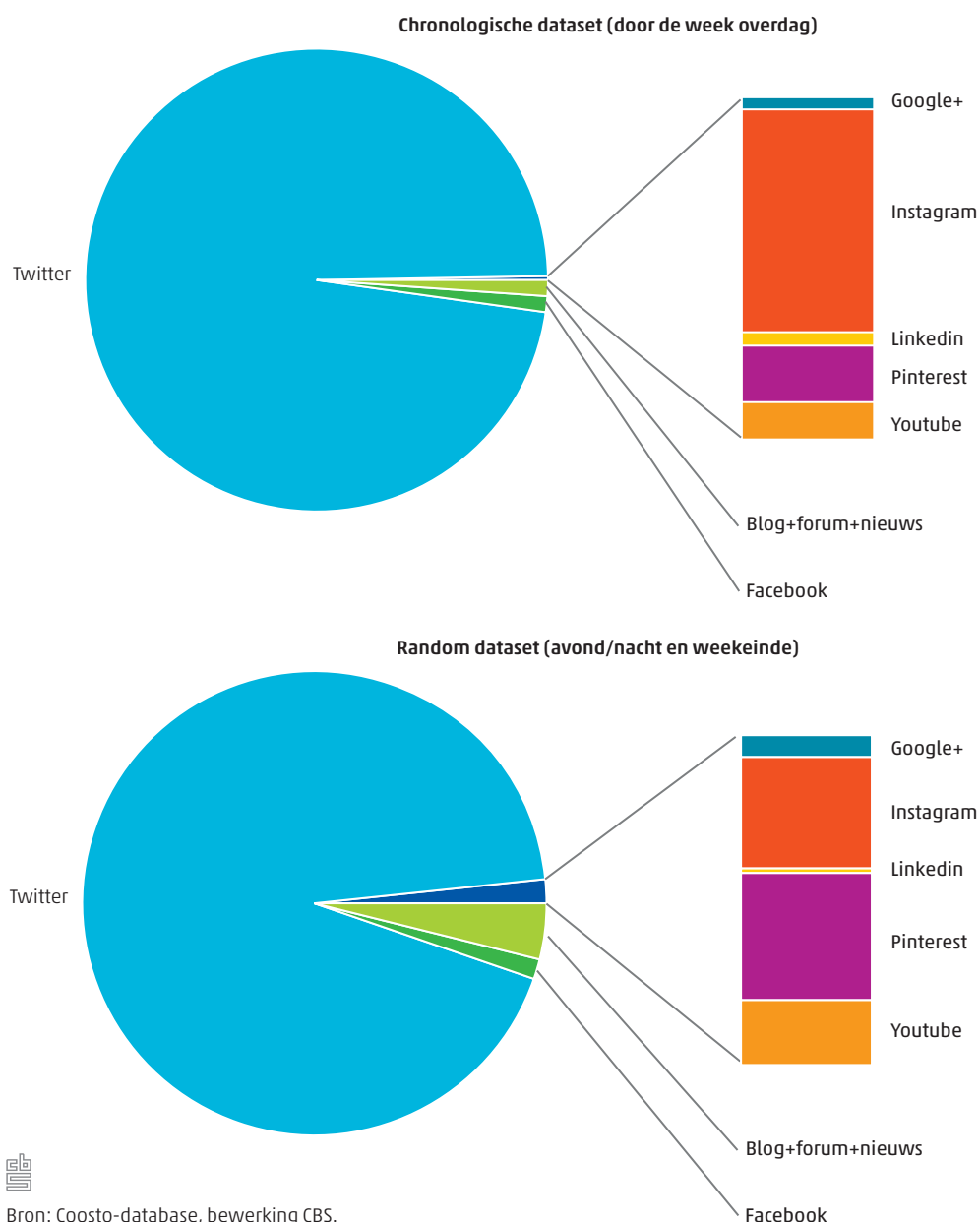
2.1.1 Aantal sociale mediaberichten van Coosto: chronologische- en random- datasets, naar type sociaal medium

Data Chronologisch (week)	<leeg>	Blog	Facebook	Forum	Nieuws	Twitter	Google+	Instagram	LinkedIn	Pinterest	Youtube	Totaal
Aantal	1 906	4 160	15 064	1 743	14 901	1 539 861	82	1 506	91	382	239	1 579 935
Percentage (%)	0,12	0,26	0,95	0,11	0,94	97,46	0,01	0,09	0,01	0,00	0,01	100

Data Random (nacht & weekend)	<leeg>	Blog	Facebook	Forum	Nieuws	Twitter	Google+	Instagram	LinkedIn	Pinterest	Youtube	Totaal
Aantal	57	2 476	3 296	1 642	5 576	225 185	236	1 203	7	1 372	699	241 749
Percentage (%)	0,02	1,02	1,36	0,68	2,31	93,15	0,1	0,5	0,003	0,57	0,3	100

Bron: Coosto-database, bewerking CBS.

2.1.2 Aantal sociale mediaberichten van Coosto: chronologische- en random datasets, naar type sociaal medium



De variabelen die in het Coosto-bestand beschikbaar zijn, zijn onder meer: datum, URL, sentiment-beoordeling, aantal views, aantal volgers, auteur of profielnaam en de tekst van de boodschap. Zie voor een volledig overzicht van de variabelen verder bijlage I.

Op basis van de verzamelde sociale mediaberichten of records kunnen in het Coosto-bestand twee statistische eenheden worden onderscheiden, namelijk:

1. het individuele sociale mediabericht zelf;
2. de profielnaam of unieke auteur van een sociaal mediabericht. Op basis van één profielnaam kunnen meer dan een sociaal mediabericht worden verstuurd.

Bij (de profielnamen van) de 1,8 miljoen sociale mediaberichten, die in deze stap zijn verzameld, kan het gaan om zowel een bedrijf als om een persoon. Pas na de koppeling met de gegevens van Dataprovider kon het onderscheid worden gemaakt tussen (de profielnamen van) sociale mediaberichten van een bedrijf of een persoon.

Het bestand van Dataprovider

Het bestand van Dataprovider bevat, in principe, alle websites van het Nederlandse internetdomein. Het bestand van Dataprovider was al eerder gekoppeld aan het Algemeen Bedrijvenregister van het CBS (ABR). Door deze koppeling richt het bestand zich nog uitsluitend op bedrijven en niet meer op personen. Door de koppeling aan het ABR zijn in het resulterende bestand ook de bedrijfsnaam, bedrijfseenheid (BE_ID), economische activiteit (SBI) en grootteklasse (op basis van aantal werkzame personen) beschikbaar.

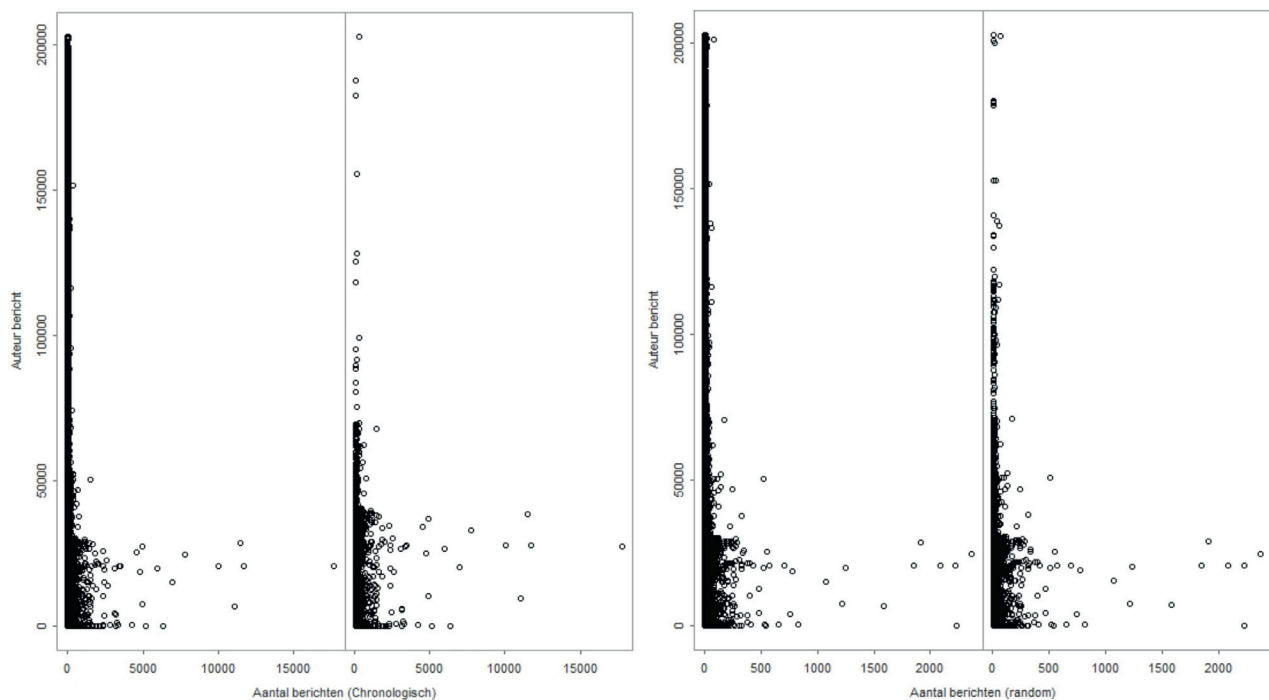
Het bestand van Dataprovider bevat onder meer variabelen, zoals bedrijfsnaam, het nummer van de Kamer van Koophandel, gebruikte sociale media typen, profielnamen, in- en uitgaande links, Alexa-ranking (proxy voor aantal bezoekers van de website) en dus economische activiteit en grootteklasse.

2.2 Verwijderen van uitbijters in de Coosto-datasets

Het doel van deze stap in het verwerkingsproces is om te voorkomen dat uitbijters de uiteindelijke analyses te veel beïnvloeden. Daarbij is aan de onderkant vooral gekeken naar het aantal sociale mediaberichten per auteur of profielnaam. Aan de bovenkant is gekeken naar het aantal volgers van een profielnaam met sociale mediaberichten. Dit heeft geleid tot het verwijderen van een aantal records uit de beide Coosto-datasets.

Figuur 2.2.1 laat de relatie zien tussen unieke auteurs of profielnamen en het aantal sociale mediaberichten dat zo'n auteur in de waargenomen periode heeft gepost. Sommige auteurs posten maar een paar en anderen juist heel veel berichten. De auteurs met maar een paar sociale mediaberichten kunnen de uitkomsten van de analyses als uitbijter tot op zekere hoogte beïnvloeden.

2.2.1 Relatie tussen unieke profielnamen/auteurs en het aantal sociale mediaberichten dat door het profiel is gepost



Bron: Coosto-database, bewerking CBS.

Om een vertekening (bias) door die uitbijters te voorkomen, zijn bij de chronologische dataset alle auteurs met minder dan 90 geposte berichten en bij de random dataset alle auteurs met minder dan 10 geposte berichten verwijderd.

Hoewel hier natuurlijk sprake is van een arbitraire keuze, is het idee dat deze auteurs nauwelijks bijdragen aan het gebruik van sociale mediaberichten. Uiteindelijk zijn circa 3 duizend sociale mediaberichten verwijderd.

Aan de bovenkant zijn profielen met meer dan een miljoen volgers verwijderd. Daarbij gaat het veelal om internationale artiesten, zoals de Rolling Stones en bekende DJ's. Ook deze groep wordt gezien als uitbijters.

De aangepaste datasets zijn vervolgens de input geweest voor de volgende stap in het verwerkingsproces: het opschonen en aanpassen van het Coosto- en Dataprovider-bestand.

2.3 Opschonen en aanpassen van de databronnen

Doel van deze verwerkingsstap is ten eerste om de koppeling tussen de beide Coosto-datasets en het bestand van Dataprovider mogelijk te maken. Daarnaast zijn aanpassingen op de bestanden uitgevoerd ten behoeve van de analyse in een latere fase. Sommige aanpassingen zijn automatisch en sommige handmatig uitgevoerd.

Daarbij gaat het om de volgende aanpassingen:

- Alle bronnen: het gelijktrekken van de kolomnamen in de verschillende bestanden.
- Coosto-datasets: alle tekst is omgezet naar kleine letters. Dit is met name gedaan ten behoeve van de koppeling.

- Coosto-datasets: het verwijderen van alle non-ASCII- en andere punctuatie-tekens in de sociale mediaberichten, zoals emoticons. Dit is met name gedaan ten behoeve van de analyses.
- Coosto-datasets: het verwijderen van Engelse en Nederlandstalige stopwoorden, zoals 'de', 'het' en 'voor'. Deze woorden spelen namelijk geen rol bij de analyses. Dat geldt ook voor het woordje 'tco', dat in de Twitterberichten vaak voorkomt in relatie tot de gebruikte URL's.
- Coosto-datasets: om de lengte van de tekst in sociale mediaberichten hanteerbaar te houden, wordt bijna altijd gewerkt met zogenoemde 'short of korte URL's'. Deze korte URL's moeten worden omgezet in de oorspronkelijke lange URL's. Omdat te kunnen doen zijn de korte URL's in een apart bestand gezet.
- Dataprovider-bestand: om de analyses beter te kunnen uitvoeren, is de variabele 'SocialMediaTypes' opgesplitst in aparte kolommen.
- Dataprovider-bestand: de gegevens over het aantal werkzame personen (proxy voor omvang van een bedrijf) zijn ingedikt naar een kleiner aantal categorieën. Zie schema 2.3.1. Dit is gedaan om de output in hoofdstuk 3 overzichtelijk te houden. In dat kader is bij de indeling van de economische activiteiten niet verder ingezoomd dan het eerste digit.

2.3.1 Indikking van de variabele grootteklasse (het aantal werkzame personen)

2-digit	0	10	21	22	30	40	50	60	71	72	81	82	91	92	93
1-digit			0				1			2				3	
wp			0 - 9			10-49			50-499				500 en meer		

2.4 Koppelen van de databronnen

Een belangrijke en noodzakelijke stap in het verwerkingsproces is de koppeling tussen de datasets van Coosto en het Dataprovider-bestand. Dit is ten eerste van cruciaal belang om onderscheid te kunnen maken tussen sociale mediaberichten van bedrijven en die van personen. Het Dataprovider-bestand bevat namelijk alleen gegevens van bedrijven. De koppeling betekent ook dat aan elk record van de Coosto-datasets variabelen van het Dataprovider-bestand zijn toegevoegd. Dat geldt met name voor de variabelen grootteklasse en economische activiteit. Zie voor alle variabelen verder bijlage II.

De koppeling is uitgevoerd op basis van de koppelsleutel 'profielnaam' (auteur). Deze koppelsleutel 'profielnaam' komt voor in zowel het Coosto-bestand met sociale mediaberichten als in het bestand van Dataprovider met websites van alle Nederlandse bedrijven.

De koppeling heeft het aantal sociale mediaberichten teruggebracht van 1,8 miljoen (1,6 + 0,2 miljoen) naar, in eerste instantie, circa 200 duizend sociale mediaberichten van bedrijven. Deze berichten zijn gepost door circa 10 duizend profielen: gemiddeld dus zo'n 20 geposte sociale mediaberichten per profiel over de periode waarin is waargenomen (juni, juli en augustus 2017). De 200 duizend gekoppelde sociale mediaberichten zijn een bovengrens, aangezien hier alle mogelijke koppelingen zijn meegerekend.

Omdat bedrijven soms dezelfde profielnaam gebruiken bij verschillende sociale media-typen, bijvoorbeeld Twitter en Facebook, is extra ontdebeld en zijn twijfelgevallen ('false positives') buiten de latere analyses gelaten. Dat betekent dat uiteindelijk een bestand met circa 120 duizend sociale mediaberichten van bedrijven is gebruikt voor de verkenningen in hoofdstuk 3.

2.5 Beschikbare statistische eenheden

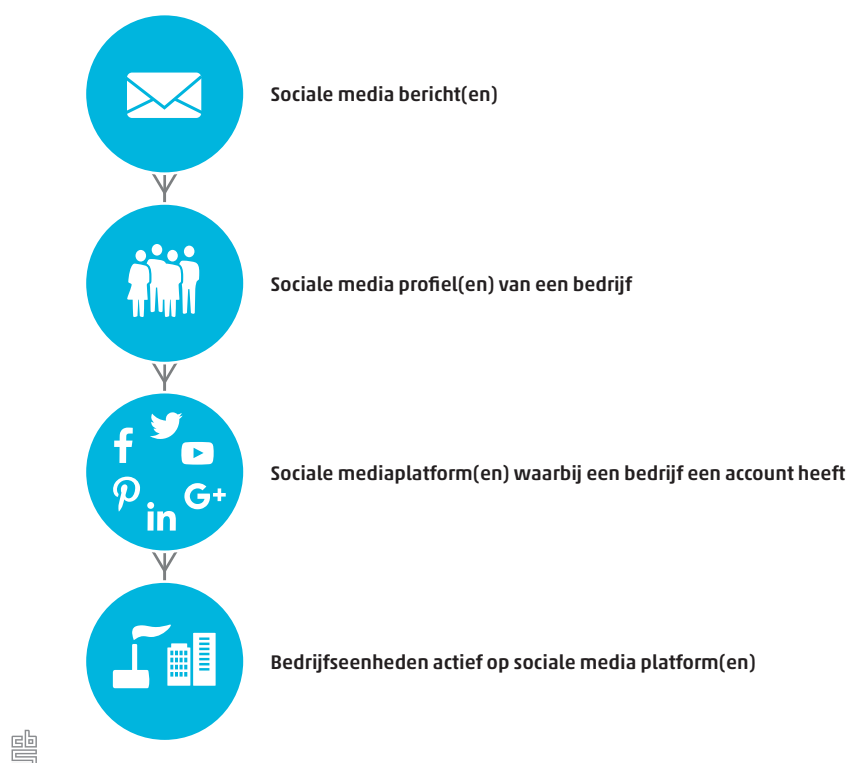
Schema 2.5.1 laat zien op welke verschillende manieren naar het gekoppelde analyse-bestand kan worden gekeken, gezien vanuit de beschikbare statistische eenheden. Ten eerste kunnen we naar de sociale mediaberichten zelf kijken. Deze berichten zijn aan een tijdstip gebonden (= het moment van posten van het bericht). Hierdoor kunnen we analyses maken per tijdseenheid, zoals per uur, per dag of per week. Ten tweede hebben we te maken met profielnamen. Deze profielnamen zijn aan één of meer sociale media-berichten gerelateerd. De profielnamen zijn, ten slotte, gekoppeld aan bedrijfseenheden in het ABR (BE_ID), waarbij een bedrijf meer dan één profielnaam kan hebben, ook bij één en hetzelfde sociale medium.

Na de koppeling gaat het dus niet meer om twee, maar om drie statistische eenheden:

1. Het sociale mediabericht. Dit zijn records met onder meer de tekst van het bericht en een tijdsdimensie.
2. De profielnaam of auteur/account: het sociale mediaprofiel per type sociaal medium.
3. De bedrijfseenheid: BE_ID's kunnen meer dan één sociaal medium gebruiken en daarbinnen meer dan één profielnaam hebben. Bedrijfseenheden kunnen worden geaggregeerd naar statistische eenheden als grootteklasse en economische activiteit.

Bovenstaande betekent dat er vrij complexe analyses kunnen worden uitgevoerd.

2.5.1 De relatie tussen statistische eenheden: sociale mediaberichten, profielnamen en bedrijfseenheden



Naast de hierboven beschreven verwerkingsstappen, wordt hier nog kort ingegaan op twee andere stappen, namelijk de omzetting van korte naar lange URL's en de output.

2.6 Omzetting van korte naar lange URL's: afleiding van domeinnamen

In eerste instantie was het verzamelen van de data van Coosto vooral gericht op het opzetten van een register met domeinnamen, die gekoppeld konden worden aan het ABR. Daardoor zijn vooral sociale mediaberichten met een URL in de tekst verzameld. Om de domeinnamen uit de URL's te halen moeten deze eerst worden omgezet van een verkorte URL naar de oorspronkelijke lange URL. Om ruimte te besparen worden lange URL's bij sociale mediaberichten namelijk automatisch omgezet naar korte URL's. Om te zien welke domeinnamen in sociale mediaberichten worden gebruikt, is het dus nodig om de verkorte URL's weer terug te vertalen naar lange URL's.

2.7 Output

Op basis van de 'master dataset', het gekoppelde analysebestand, zijn de verschillende verkenningen in hoofdstuk 3 uitgevoerd. Daarbij speelt vooral de aggregatiestap naar de variabelen grootteklasse en economische activiteit een belangrijke rol.

Daarnaast is gebruikgemaakt van methoden zoals 'woordwolken' en 'process mining'. Woordwolken zijn visualisaties van de woorden die het meest frequent in sociale mediaberichten voorkomen. Daarmee kan inzicht worden gekregen in de inhoud van de sociale mediaberichten: waar gaat het bericht over? Process mining gaat over het visualiseren van processen en dan vooral de relaties tussen de stappen in het proces. Daarmee kan ook inzicht worden verkregen in de relatie in de tijd van sociale mediaberichten. Bijvoorbeeld als het gaat om het gebruik van tweets of retweets.

3. Resultaten

3.1 Inleiding

In dit hoofdstuk worden de resultaten van het onderzoek beschreven. Nogmaals: het gaat hier om een verkennend onderzoek. Daarbij zijn, samenvattend, de volgende uitgangspunten gehanteerd:

- Gestart is met een bestand met 1,8 miljoen sociale mediaberichten of records, die op basis van een API van Coosto zijn verzameld in de periode juni, juli en augustus 2017 (6 weken). Het bestand is opgedeeld in een chronologische (door de week en overdag verzameld) en een random dataset (in de avond en in het weekeinde verzameld).
- Door de koppeling met het Dataprovider-bestand is het aantal sociale mediaberichten gereduceerd tot circa 120 duizend: 109 453 voor de chronologische dataset en 12 487 voor de random-dataset: respectievelijk 7 en 5 procent van alle verzamelde sociale mediaberichten. Na de koppeling gaat het alleen om berichten die zijn gepost door bedrijven.
- De uitbijters aan de onderkant en de bovenkant zijn verwijderd, dat wil zeggen: profielen met erg weinig geposte sociale mediaberichten en profielen met meer dan 1 miljoen volgers.
- Het Dataprovider-bestand was al eerder gekoppeld aan het ABR van het CBS, waardoor per bericht of profiel ook de grootteklasse en de economische activiteit bekend zijn van het betrokken bedrijf (BE_ID).
- Om de output overzichtelijk te houden zijn de categorieën van de economische activiteit (SBI) en de grootteklasse beperkt gehouden. Bij de SBI gaat het om de 19 hoofdklassen (eerste digit), lopend van A: Landbouw, Bosbouw en Visserij tot U: Extraterritoriale organisaties. Bij sommige analyses worden categorieën van een lager niveau gehanteerd. Men dient er rekening mee te houden dat in sommige SBI's zzp'ers en kleine bedrijven dominant zijn, terwijl in andere bedrijfstakken juist grote bedrijven veel voorkomen. Het aantal grootteklassen is beperkt tot 4.
- Er is een variabele 'Tweet-Retweet' toegevoegd aan het bestand. Deze variabele is bedoeld om een onderscheid te kunnen maken tussen 'oorspronkelijke berichten' en 'doorgestuurde berichten'. Het bereik van een tweet via retweets binnen een SBI wordt onder meer bepaald door het gemiddelde aantal volgers van een profiel dat een retweet post.

- De verkorte URL in een sociaal mediabericht duidt de unieke locatie aan van een webpagina of een document op het internet. De lange URL wordt om praktische redenen automatisch door Twitter verkort. Om domeinnamen te kunnen afleiden, zijn de verkorte URL's terugvertaald naar lange URL's.
- Voor de visualisaties van de resultaten is, naast grafieken en tabellen, ook gebruikgemaakt van twee andere mogelijkheden. Ten eerste is gebruikgemaakt van zogenoemde 'woordwolken'. Deze methode is geschikt om de meest gebruikte woorden in een tekstbericht via een visualisatie te laten zien. Er bestaan verschillende varianten van deze mogelijkheid. Ten tweede is gebruikgemaakt van 'process mining'. Met deze methode kan inzicht worden verkregen in allerlei verbanden tussen variabelen en dus ook stromen. Hier is gebruikgemaakt van het pakket DISCO [6].

Voor de relaties tussen de gebruikte statistische eenheden zijn verwezen naar schema 2.5.1 in hoofdstuk 2. Het gaat om:

1. sociale mediaberichten;
2. profielnamen of accounts;
3. bedrijfseenheden, eventueel geaggregeerd naar bedrijfsgrootte en economische activiteit.

Om richting te geven aan de verkenning zijn de volgende *onderzoeksvragen* gehanteerd, die aansluiten op de probleemstelling:

- Zijn sociale mediaberichten te gebruiken voor het vervangen of aanvullen van bestaande statistiek? Kan bijvoorbeeld de vraag 'Gebruik van sociale media' in de statistiek 'ICT-gebruik bedrijven' door de Coosto-data worden vervangen?
- Naar welke domeinnamen van bedrijven wordt in sociale mediaberichten vooral verwezen?
- In hoeverre is een sociaal mediabericht of een profielnaam een relevante statistische eenheid?
- Van welk type sociale media maken bedrijven, of beter gezegd bedrijfstakken en grootteklassen, vooral gebruik? En welke bedrijfstakken posten meer en welke minder sociale mediaberichten? Bij deze analyse wordt alleen gebruikgemaakt van Twitterberichten.
- Hoe worden sociale mediaberichten door bedrijven gebruikt? Of welke onderwerpen komen vooral voor in de berichten? Om dit te onderzoeken is geëxperimenteerd met zogenoemde 'woordwolken' ('word clouds'). Ook hier is alleen gebruikgemaakt van de Twitterberichten.
- Wat is het bereik van Twitterberichten (tweets en retweets)? In welke mate worden Twitterberichten bekeken (gemiddeld aantal views)?

Doordat er sprake is van verkennend onderzoek, is niet bij alle analyses dezelfde dataset gehanteerd. Doordat zij een belangrijk deel uitmaken van het bestand van Coosto, zijn vaak alleen de Twitterberichten gebruikt.

3.2 Resultaten

Sociale mediaberichten als vervanging van survey-vragen

Een eerste vraag, die in dit onderzoek is verkend, is of het gekoppelde analysebestand met sociale mediaberichten gebruikt kan worden om vragen van de Europees-verplichte enquête 'ICT-gebruik bedrijven' te vervangen? Zijn de resultaten van beide benaderingen vergelijkbaar?

In de enquête 'ICT-gebruik bedrijven' krijgen bedrijven met 10 of meer werkzame personen de vraag of zij gebruikmaken van sociale media en zo ja van welk type? De antwoord-categorieën zijn:

1. Sociale netwerken (bijvoorbeeld Facebook).
2. Blogs of microblogs (bijvoorbeeld Twitter).
3. Multimedia (bijvoorbeeld Youtube).
4. Wiki's.

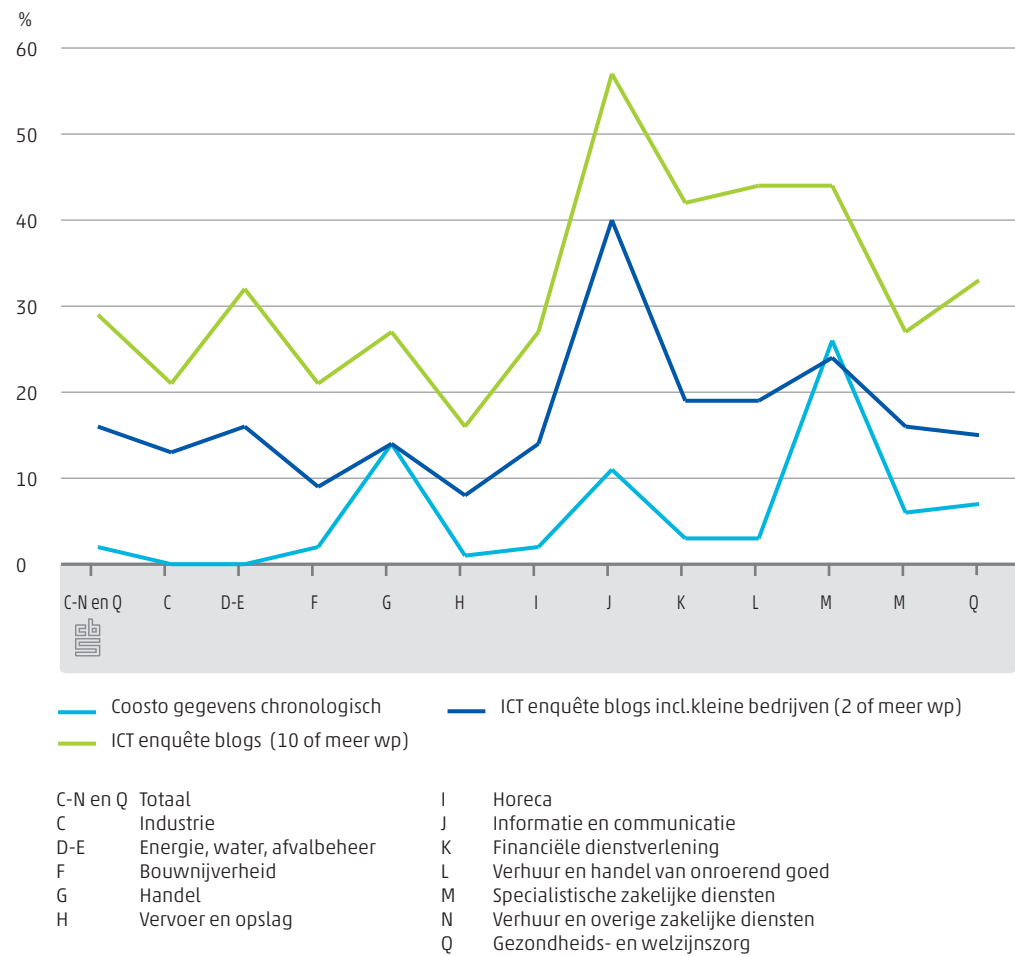
Omdat in het bestand van Coosto de Twitterberichten domineren, is er hier voor gekozen om de vergelijking tussen beide benaderingen uit te voeren op basis van de categorie blogs of microblogs (bijvoorbeeld Twitter).

Sinds 2017 is de populatie van de ICT-enquête bedrijven uitgebreid met de kleinere bedrijven van 2 of meer werkzame personen. De verwachting is dat de resultaten op basis van deze bredere populatie beter vergelijkbaar zijn met de resultaten op basis van de Coosto-gegevens dan de resultaten op basis van alleen de grotere bedrijven.

Figuur 3.2.1 laat de vergelijking per bedrijfstak (SBI 1^e digit) zien tussen het percentage bedrijven dat in 2017 gebruikmaakte van Twitter (blogs) op basis van de Coosto-data (alleen chronologische dataset) en het percentage bedrijven dat volgens de ICT-enquête in 2017 gebruikmaakte van (micro)blogs (bijvoorbeeld Twitter). Bij de data van de ICT-enquête is onderscheid gemaakt tussen de resultaten op basis van het bestand met alleen de bedrijven met 10 werkzame personen of meer en die van het bestand inclusief de kleinere bedrijven met twee of meer werkzame personen.

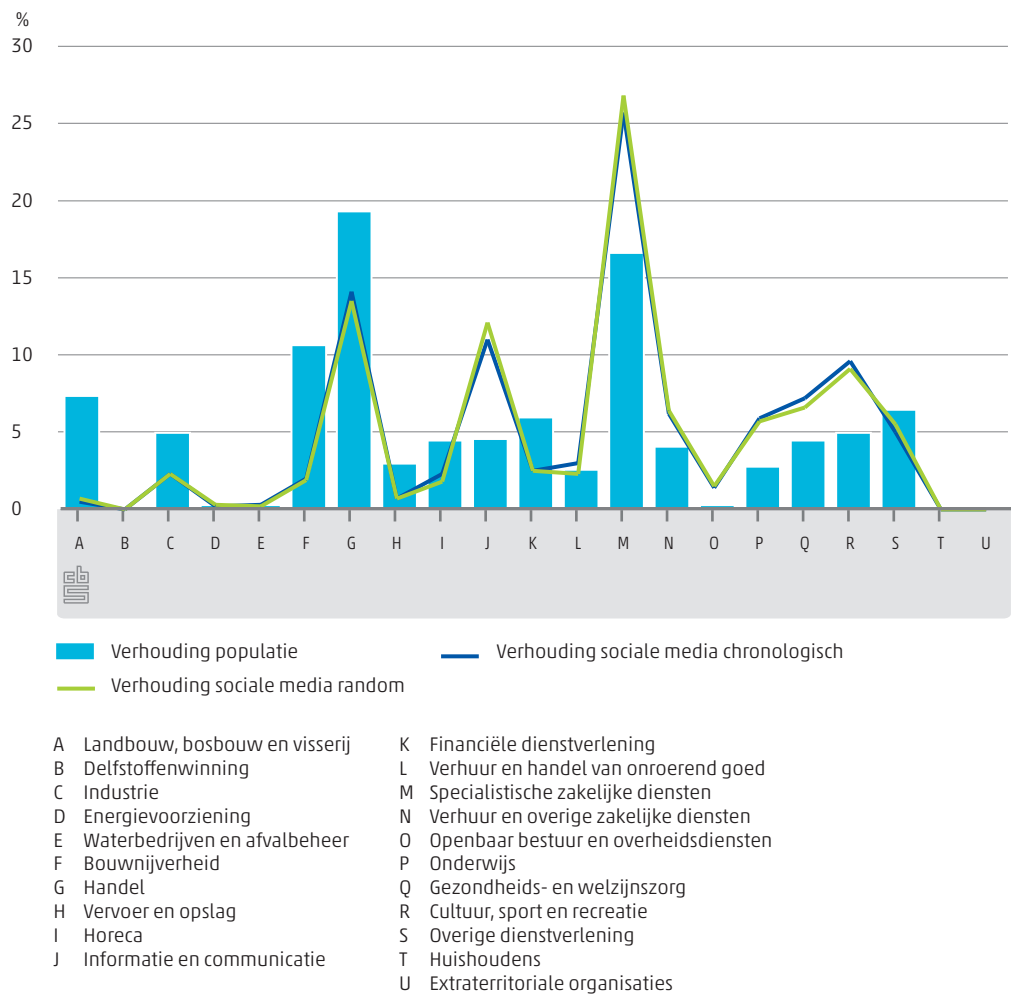
De lineaire samenhang tussen de resultaten op basis van de Coosto-gegevens en de resultaten van de ICT-enquête zijn niet erg sterk. Als de resultaten op basis van de Coosto-gegevens worden vergeleken met de resultaten van de ICT-enquête met alleen de grote bedrijven met 10 werkzame personen of meer dan is de waarde van Pearsons correlatiecoëfficiënt 0,36 (waarden tussen 0-1). Zoals te verwachten was, komt de vergelijking met de ICT-enquête inclusief de kleine bedrijven met 2 of meer werkzame personen iets hoger uit, namelijk 0,48. De resultaten van de ICT-enquête komen bij de meeste bedrijfstakken aanzienlijk hoger uit dan die van de Coosto-data met sociale mediaberichten. Alleen bij de bedrijfstakken G (Handel) en M (Specialistische zakelijke diensten) is het niveau vergelijkbaar.

3.2.1 Relatie van de resultaten van de variabele 'Gebruik van (micro)blogs' van de ICT-enquête bedrijven en het Coosto-bestand, 2017 (alleen Twitterberichten)



Bron: Coosto, bewerking CBS.

3.2.2 Verhoudingen aantal bedrijven per economische activiteit, en het aantal bedrijven dat sociale mediaberichten post ten opzichte van het totaal, 2017



Bron: Coosto, bewerking CBS.

Er kan ook op een andere manier naar de data worden gekeken. Zo geeft figuur 3.2.2 ten eerste de relatieve verhouding weer tussen het aantal bedrijven in de verschillende bedrijfstakken ten opzichte van het totaal aantal bedrijven in Nederland (verhouding populatie). Deze gegevens komen uit de statistiek Bedrijvendemografie, die gebaseerd is op het ABR. Daar overheen is de verhouding per bedrijfstak geplot van het percentage bedrijven dat volgens de Coosto-gegevens actief is met sociale media ten opzichte van het totaal aantal bedrijven dat actief is met sociale media. Daarbij is onderscheid gemaakt tussen de resultaten van de chronologische en random dataset. Een aantal zaken valt op. Ten eerste is er weinig verschil tussen de verhoudingen in de chronologische en random dataset. Ten tweede zijn hier ook bedrijfstakken te zien, zoals Landbouw en Cultuur en Sport en recreatie, die niet in de ICT-enquête worden waargenomen. Daarnaast valt op dat de bedrijfstakken aan de linkerkant (aan het begin van de bedrijfskolom) verhoudingsgewijs relatief weinig gebruikmaken van sociale media. Bedrijfstakken zoals Informatie en communicatie en Specialistische zakelijke dienstverlening maken daarentegen verhoudingsgewijs juist weer veel gebruik van sociale media.

De conclusie van deze eerste verkenning is dat er wel een relatie is tussen de resultaten van het Coosto-bestand met sociale mediaberichten en de resultaten van de ICT-enquête als het gaat om de vraag of een bedrijf een 'blog of microblog (bijvoorbeeld Twitter)' gebruikt. Die relatie is echter niet al te sterk. Het is moeilijk te zeggen waar de verschillen zitten. Het kan zijn dat de waargenomen sociale mediaberichten van Coosto niet representatief zijn. Zo is bijvoorbeeld gefilterd op sociale mediaberichten met een URL erin. Het kan ook zijn dat de respondenten van de ICT-enquête niet alle of onjuiste gegevens hebben verstrekt. Alleen bij de bedrijfstakken G (Handel) en M (Specialistische zakelijke diensten) komen volumes overeen. Bij alle andere sectoren lijkt er sprake te zijn van een onderschatting door de resultaten op basis van de Coosto-data. Meer onderzoek zou hier dus nodig zijn voordat vragen in de ICT-enquête bedrijven kunnen worden vervangen. Uit een latere analyse is gebleken dat bij de andere antwoordcategorieën van de vraag of bedrijven gebruikmaken van sociale media er wel sprake was van een sterker verband tussen de Coosto-data en de data van de ICT-enquête bedrijven.

Een duidelijk voordeel van de Coosto-gegevens is wel dat alle bedrijfstakken in beeld komen, ook degenen die niet door de ICT-enquête bedrijven worden waargenomen. Bovendien kunnen de resultaten veel sneller en gedetailleerder beschikbaar komen.

Verwijzing naar domeinnamen

Een tweede case, waar in dit onderzoek naar is gekeken, betreft de domeinnamen, die afgeleid zijn uit de URL's in sociale mediaberichten van Coosto: naar welke websites of documenten van bedrijven wordt in sociale media vooral verwezen? Een domeinnaam is een onderdeel van een URL. Als de URL bijvoorbeeld '<https://www.cbs.nl/nl-nl/nieuws>' is, dan is de domeinnaam '[cbs.nl](https://www.cbs.nl/nl-nl/nieuws)'.

Deze verkenning is in de eerste plaats uitgevoerd ten behoeve van het opzetten van een register van domeinnamen, dat aansluit op het Algemeen Bedrijvenregister (ABR) van het CBS. De conclusie van dit experiment was echter dat de Coosto-data met sociale mediaberichten relatief weinig domeinnamen opleverde voor zo'n register. De keuze is dan ook gemaakt om met het experiment te stoppen. Op dit moment worden andere methoden gebruikt om zo'n register met domeinnamen op te zetten. Daarbij gaat het onder meer om web scraping en machine learning [16].

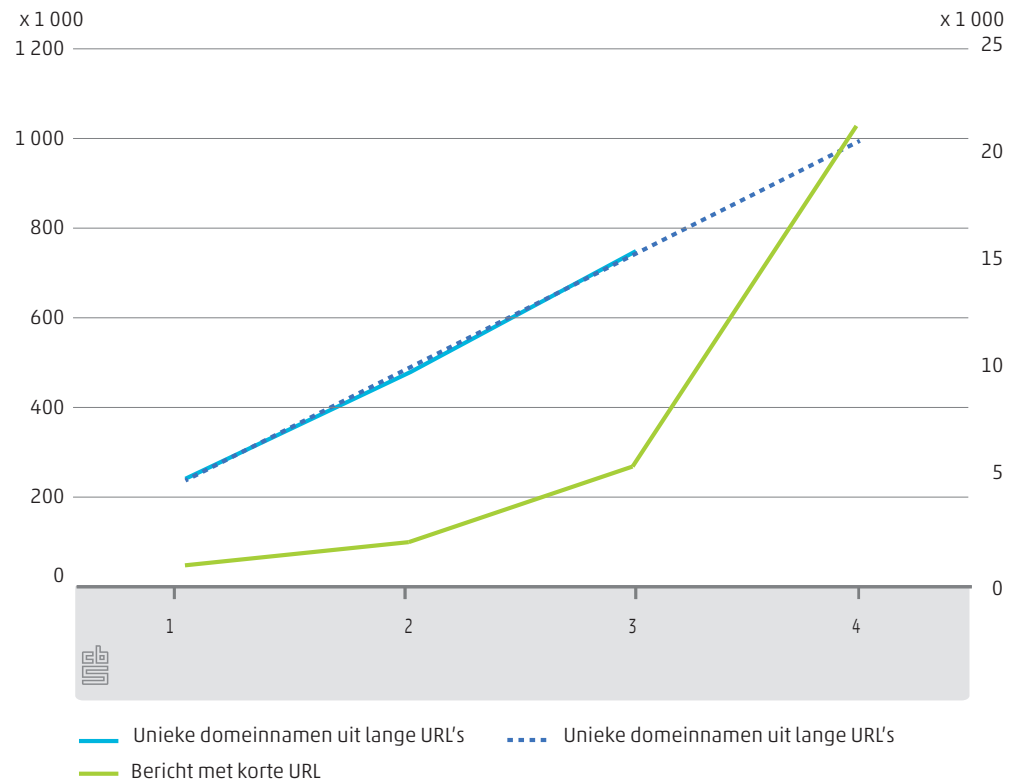
Om een domeinnaam uit een URL af te leiden, is het eerst nodig om de in sociale mediaberichten verkorte URL's te vertalen naar lange URL's. Pas daarna kan de domeinnaam uit de URL worden afgeleid. Om een vertaling van een verkorte URL naar een lange URL mogelijk te maken, is er een procedure geschreven in het softwarepakket R. In eerste instantie is geprobeerd om de vertaling uit te voeren voor alle verkorte URL's van alle 1,8 miljoen sociale mediaberichten, ongeacht of het nu ging om berichten van bedrijven of personen. Dit bleek echter geen succes. De rekenkracht van de beschikbare apparatuur en programmatuur⁴⁾ was onvoldoende om de 1,8 miljoen records op een relatief snelle manier te verwerken. Nadat iets meer dan 20 procent van de sociale mediaberichten waren verwerkt,

⁴⁾ In sommige gevallen werd vanuit de benaderde server doorverwezen naar andere servers ('proxy's'), waardoor het verwerkingsproces veel tijd in beslag nam.

is de procedure gestopt. De circa 360 duizend verwerkte verkorte URL's leverden 286 duizend lange, niet-unieke, URL's op. Na ontdebelling van deze lange URL's leverde dat uiteindelijk 15 600 unieke URL's op, waaruit domeinnamen konden worden afgeleid.

Als het verwerkingsproces was doorgezet, is de schatting dat dit uiteindelijk 20 tot 30 duizend unieke domeinnamen had opgeleverd. Zie figuur 3.2.3.

3.2.3 Schatting van het aantal domeinnamen op basis van een omzetting van 1,8 miljoen korte URLs naar lange URLs



Bron: Coosto, bewerking CBS.

Een vraag, die beter bij de probleemstelling past, is op welke economische activiteiten (SBI's) de gevonden domeinnamen betrekking hebben? Tabellen 3.2.4 en 3.2.5 laten enkele mogelijke resultaten zien.

Zo laat tabel 3.2.4 de verdeling zien van alle sociale mediaberichten naar bedrijfstak (SBI 1^e en 2^e digit), waarin een URL voorkomt die relateert aan één van de top-100 domeinnamen (>300 keer). Uit de tabel is af te leiden dat de bedrijfstakken Informatie en technologie en Specialistische zakelijke dienstverlening samen goed zijn voor 91 procent van alle sociale mediaberichten met een URL, die relateert aan een domeinnaam uit de top-100 van domeinnamen. Op een niveau lager, het 2de digit van de SBI, gaat het vooral om de branches Reclame en marktonderzoek (SBI 73) en in mindere mate Uitgeverijen van kranten (SBI 58) en Dienstverlenende activiteiten op het gebied van de informatie-technologie (SBI 62). Niet onlogisch omdat deze bedrijfstakken veel gebruikmaken van sociale media voor hun communicatie.

Gebruiken we niet de sociale mediaberichten als populatie, maar bedrijfseenheden geaggregeerd naar grootteklassen, dan verschuift het beeld iets. Nog steeds komen de top-100 domeinnamen veel terug in de SBI's J (Informatie en communicatie) en M (Specialistische zakelijke dienstverlening), maar ook in de bedrijfstak N (Verhuur en overige zakelijke dienstverlening). Daarbij valt op dat de meeste sociale mediaberichten in de bedrijfstak Verhuur en overige zakelijke dienstverlening met een domeinnaam uit de top-100 van domeinnamen worden verstuurd door vooral kleine bedrijven. Dat geldt minder voor de Specialistische zakelijke dienstverlening. Zie verder tabel 3.2.5.

3.2.4 Verdeling van alle sociale mediaberichten met een domeinnaam komend uit de top-100 van domeinnamen van bedrijven, naar bedrijfstak

SBI-1	SBI-2-digit	Aantal	Aandeel
-digit			
G	Handel	2 667	2%
J	Informatie en communicatie	66 532	38%
	58 Uitgeverijen van kranten	18 047	10%
	59 Productie en distributie van films en televisieprogramma's	6 692	4%
	60 Verzorgen en uitzenden van radio- en televisieprogramma's	7 656	4%
	62 Dienstverlenende activiteiten op het gebied van informatie technologie	18 847	11%
	63 Web portals	15 290	9%
M	Specialistische zakelijke diensten	92 836	53%
	70 Holdings (geen financiële), conserndiensten binnen eigen concern en managementadvisering	1 112	1%
	73 Reclame en marktonderzoek	91 724	52%
N	Verhuur en handel overige zakelijke diensten	3 696	2%
	78 Arbeidsbemiddeling, uitzendbureaus en personeelsbeheer	3 696	2%
O	Openbaar bestuur en overheidsdiensten	1 227	1%
Q	Gezondheids- en welzijnszorg	6 497	4%
R	Cultuur, sport en recreatie (90 Schrijven en overige scheppende kunsten)	1 157	1%
Totaal aantal sociale mediaberichten met top 100 domeinnaam		174 612	100%

Bron: Coosto en Dataprovider, bewerking CBS.

De conclusie van deze verkenning is dat met behulp van de verkorte URL's in sociale mediaberichten uiteindelijk domeinnamen kunnen worden afgeleid. Dit geeft enig inzicht in de inhoud (websites en documenten van bedrijven) waarop de geposte sociale mediaberichten zich richten. De hoeveelheid domeinnamen die uit de URL's zijn af te leiden is echter beperkt. Hierdoor is niet verder gegaan met het idee om met behulp van verkorte URL's in sociale mediaberichten een register met domeinnamen op te zetten. De resultaten van deze verkenning kunnen hooguit worden gebruikt ter controle van de resultaten van andere methoden.

In meer algemene termen, maakt deze verkenning echter ook duidelijk dat op verschillende manieren naar de data van Coosto kan worden gekeken. Dit kan nieuwe indicatoren opleveren, die niet te produceren zijn op basis van traditionele enquêtegegevens. Om dit verder te ontwikkelen is het van belang dat samen met gebruikers gekeken wordt aan welke indicatoren behoefte is en of in deze behoefte kan worden voorzien met dit soort big datagegevens.

Zijn sociale mediaberichten relevante statistische eenheden?

In het statistiekproces worden eerst gegevens verzameld en opgeslagen. De opgeslagen gegevens worden vervolgens verwerkt tot statistische informatie op basis waarvan conclusies kunnen worden getrokken over bepaalde verschijnselen. De eindresultaten worden, ten slotte, in allerlei vormen, zoals tabellen, grafieken of andere (nieuwe) visualisaties, gepresenteerd. Om dit gehele proces te kunnen uitvoeren, wordt het verschijnsel beschreven op basis van een onderzoekspopulatie. De onderzoekspopulatie is gebaseerd op statistische eenheden, in veel gevallen personen of bedrijven.

3.2.5 Relatieve verdeling van sociale mediaberichten van bedrijven met domeinnamen uit de top-100 van domeinnamen naar economische activiteit (SBI-1-digit) en grootteklasse

	(a) Berichten (%)				(b) Bedrijven (%)			
	GK				GK			
	1	2	3	Alle GK	1	2	3	Alle GK
Bedrijfstak (SBI)								
M: Specialistische zakelijke diensten	11	6	2	18	19	7	6	22
J: Informatie en communicatie	47	39	35	25	13	10	7	16
R: Cultuur, sport en recreatie	5	4	8	5	12	6	1	10
G: Handel	4	3	2	7	10	5	4	9
N: Verhuur en overige zakelijke diensten	6	10	6	21	7	14	9	8
S: Overige dienstverlening	14	7	1	11	11	6	1	7
Q: Gezondheids- en welzijnzorg	6	4	12	3	9	10	25	7
P: Onderwijs	2	7	8	3	5	7	10	6
O: Openbaar bestuur en overheidsdiensten	0	14	18	3	1	22	18	4
L: Verhuur en handel van onroerend goed	1	1	0	1	3	2	0	3
K: Financiële dienstverlening	0	0	4	1	2	1	6	2

Bron: Coosto en Dataprovider, bewerking CBS.

De statistische eenheden aan de waarneemkant hoeven niet gelijk te zijn aan de statistische eenheden aan de outputkant van het proces. Er kan bijvoorbeeld een aggregatiestap tussen zitten, waardoor de statistische eenheid verandert. Een statistische eenheid aan de outputkant kan bijvoorbeeld een wijk zijn, terwijl de statistische eenheid aan de waarneemkant een huishouden is. Uit de waargenomen inkomens van huishoudens in een wijk, kan bijvoorbeeld statistische informatie worden afgeleid over het gemiddelde inkomen per wijk. De wijk is in dit geval de statistische eenheid.

De vraag is nu of sociale mediaberichten en/of -profielen valide en relevante statistische eenheden zijn? Als het gaat om de waarneemkant dan zijn sociale mediaberichten en profielen zeker relevante en bruikbare statistische waarneemeenheden. Zie bijvoorbeeld de vorige paragraaf. Deze eenheden kunnen, al dan niet door koppeling, geaggregeerd worden naar statistische informatie over andere statistische eenheden, zoals het gebruik van verschillende typen sociale media of het sociale media gebruik door bedrijven of personen.

Als we kijken naar de outputkant dan zijn sociale mediaberichten en profielen minder relevante statistische eenheden. Er zal namelijk eerder gerapporteerd worden over sociale media typen, bedrijven of personen of bedrijven die sociale mediaberichten posten of profielen hebben. Bij bedrijven gaat het dan vooral om de aggregatiestap naar statistische eenheden als grootteklasse of economische activiteit. Een sociaal medium profiel zou nog

wel kunnen fungeren als een kenmerk van een statistische eenheid bedrijf, bijvoorbeeld in een register. Dat geldt ook voor het gebruik van de verschillende typen sociale media of bijvoorbeeld de afgeleide domeinnamen.

De conclusie van deze verkenning is, dat sociale mediaberichten en/of -profielen valide statistische eenheden zijn aan de waarneemkant, maar veel minder of niet aan de outputkant.

Gebruik van typen sociale media

Tabel 3.2.6 laat op basis van het aantal sociale mediaberichten zien welke combinaties van sociale media typen het meest worden gebruikt door bedrijven. Niet erg verrassend, is, dat Twitter volgens de Coosto-data het meest wordt gebruikt door bedrijven. Op zich is het grote gebruik van Twitter niet onlogisch, omdat bijvoorbeeld LinkedIn, Pinterest en Facebook vaak meer gericht zijn op personen. Ook kan het zijn dat bedrijven bij sommige sociale media juist passiever zijn dan bij andere sociale media.

De meest gebruikte combinaties van sociale mediatypen door bedrijven in de periode juni-augustus 2017 zijn:

- Twitter en Facebook;
- Twitter, Facebook en LinkedIn;
- Twitter en LinkedIn.

Naast Twitter, gebruiken bedrijven dus vaak nog één of meer sociale mediatypen. Daarbij gaat het vooral om Facebook en LinkedIn. Sociale media, zoals Facebook, Pinterest, LinkedIn en Googleplus worden nooit als enig sociaal medium gebruikt door bedrijven. Zie tabel 2.3.6.

3.2.6 Aantal en relatieve verhouding van alle mogelijke combinaties van gebruikte sociale mediatypen op basis van de verzamelde sociale mediaberichten bij bedrijven

	Chronologische dataset			Random dataset		
	Aantal berichten	Percentage (%)		Aantal berichten	Percentage (%)	
	Aandeel	Cum.		Aandeel	Cum.	
Combinaties van sociale media typen (gekoppeld aan het Dataprovider-bestand)						
Facebook, Twitter	42 342	39%	39%	4 519	36%	36%
Facebook, LinkedIn, Twitter	25 627	23%	62%	2 692	22%	58%
LinkedIn, Twitter	12 647	12%	74%	1 626	13%	71%
Twitter (alléén)	9 163	8%	82%	1 221	10%	81%
Facebook, GooglePlus, LinkedIn, Twitter	5 694	5%	87%	613	5%	86%
Facebook, GooglePlus, Twitter	5 389	5%	92%	757	6%	92%
Facebook, Pinterest, Twitter	2 640	2%	95%	302	2%	94%
Facebook, GooglePlus, Pinterest, Twitter	1 573	1%	96%	215	2%	96%
Facebook, GooglePlus, LinkedIn, Pinterest, Twitter	1 284	1%	97%	181	1%	97%
Facebook, LinkedIn, Pinterest, Twitter	1 224	1%	98%	158	1%	98%
GooglePlus, LinkedIn, Twitter	729	1%	99%	40	0%	98%
Twitter, LinkedIn, Pinterest	598	1%	100%	110	1%	99%
GooglePlus, Twitter	377	0%	100%	40	0%	99%
Pinterest, Twitter	87	0%	100%	8	0%	99%
GooglePlus, LinkedIn, Pinterest, Twitter	52	0%	100%	5	0%	99%
GooglePlus, Pinterest, Twitter	27	0%	100%	0	0%	99%
Facebook (alléén)	0	0%	100%	0	0%	99%
LinkedIn (alléén)	0	0%	100%	0	0%	99%
Pinterest (alléén)	0	0%	100%	0	0%	99%
GooglePlus (alléén)	0	0%	100%	0	0%	99%
Totaal	109 453	100%		12 487	99%	

Bron: Coosto en Dataprovider, bewerking CBS.

Tabel 3.2.7 laat de directe telling en aandelen zien van de sociale mediaberichten en bedrijven, opgedeeld in de chronologische en random datasets. Duidelijk is dat volgens de Coosto-data Twitter het meest wordt gebruikt, gevolgd door Facebook. Daarbij is er nauwelijks verschil te zien tussen de chronologische (overdag door de week) en de random ('s avonds en in het weekeinde) datasets.

3.2.7 Aantal en relatieve verhouding van gebruikte sociale mediaberichten naar type

	Chronologische dataset				Random dataset			
	Aantal berichten	Aandeel	Aantal bedrijven	Aandeel	Aantal berichten	Aandeel	Aantal bedrijven	Aandeel
Sociale media type								
Twitter	109 453	41%	10 024	39%	12 487	41%	2 783	40%
Facebook	85 773	32%	8 341	32%	9 437	31%	2 332	33%
LinkedIn	47 855	18%	4 825	19%	5 425	18%	1 151	16%
GooglePlus	15 125	6%	1 721	7%	1 851	6%	432	6%
Pinterest	7 485	3%	1 037	4%	979	3%	296	4%

Bron: Coosto en Dataprovider, bewerking CBS.

Wordt een verdeling naar economische activiteit (SBI) gemaakt en beperken we ons tot alleen de Twitterberichten (108 863 voor de chronologische dataset en 10 024 voor de random dataset), dan laat tabel 3.2.8 zien dat de bedrijfstakken M (Specialistische zakelijke diensten) en J (Informatie en communicatie) in de waarneemperiode de meeste berichten hebben gepost. Samen met bedrijfstak G (Handel), zijn zij goed voor 50 procent van alle geposte sociale mediaberichten. Een vergelijkbaar beeld komt terug als we kijken vanuit het perspectief van het aantal bedrijven dat sociale mediaberichten post. Ook dan vormen de bedrijfstakken M, J en G de top drie, hoewel in een andere volgorde.

3.2.8 Aantal Twitterberichten en bedrijven naar economische activiteit, periode 26-06-2017 tot 04-08-2017

	Aantal berichten		Aantal bedrijven	
	Chronologisch	Random	Chronologisch	Random
Bedrijfstak (SBI-indeling)				
M: Specialistische zakelijke diensten	31 444	3 831	2 577	746
J: Informatie en communicatie	17 141	2 036	1 107	337
G: Handel	9 840	1 041	1 418	375
N: Verhuur en overige zakelijke diensten	9 695	877	618	179
S: Overige dienstverlening	7 122	891	497	153
R: Cultuur, sport en recreatie	6 655	708	958	254
Q: Gezondheids- en welzijnszorg	4 977	448	721	185
P: Onderwijs	4 807	527	593	158
I: Horeca	3 756	480	230	49
L: Verhuur en handel van onroerend goed	3 593	433	302	64
K: Financiële dienstverlening	2 845	347	255	70
C: Industrie	2 178	275	235	64
F: Bouwnijverheid	1 716	214	202	54
O: Openbaar bestuur en overheidsdiensten	1 591	83	145	41
H: Vervoer en opslag	717	62	70	19
D: Energievoorziening	281	32	22	9
A: Landbouw, bosbouw en visserij	271	37	46	20
E: Waterbedrijven en afvalbeheer	178	5	.	.
B: Delfstoffenwinning	56	20	.	.
U: Extraterritoriale organisaties	0	0	0	0
T: Huishoudens	0	0	0	0
Totaal	108 863	12 347	10 024	2 783

Bron: Coosto en Dataprovider, bewerking CBS.

De tabellen 3.2.9a en b maken voor het aantal geposte Twitter-berichten en bedrijven niet alleen onderscheid naar economische activiteit, maar ook naar tijd. De vraag daarbij is of de toevoeging van de tijdsdimensie nog leidt tot andere inzichten? Ook over de tijd gezien blijven de bedrijfstakken M en J het meest actief. Bij de bedrijfstakken G en N is er echter een verschuiving in de rangorde in onder meer de maand juni. Dat geldt met name voor het perspectief vanuit de sociale mediaberichten (tabel 3.2.9a.) en minder voor het perspectief vanuit het aantal bedrijven dat twittert (tabel 3.2.9b).

3.2.9a Aantal Twitterberichten naar economische activiteit en tijd, periode 26-06-2017 tot 04-08-2017

	Aantal berichten					
	juni		juli		aug	
	Chrono	Random	Chrono	Random	Chrono	Random
Bedrijfstak (SBI-indeling)						
M: Specialistische zakelijke diensten	2 129	0	28 119	3 573	1 196	258
J: Informatie en communicatie	1 152	0	15 339	1 931	650	105
N: Verhuur en overige zakelijke diensten	1 145	0	8 244	828	306	49
G: Handel	652	0	8 744	981	444	60
S: Overige dienstverlening	555	0	6 352	820	215	71
R: Cultuur, sport en recreatie	476	0	5 946	664	233	44
Q: Gezondheids- en welzijnszorg	390	0	4 438	416	149	32
P: Onderwijs	358	0	4 300	493	149	34
L: Verhuur en handel van onroerend goed	290	0	3 173	399	130	34
I: Horeca	241	0	3 316	446	199	34
K: Financiële dienstverlening	211	0	2 488	330	146	17
O: Openbaar bestuur en overheidsdiensten	116	0	1 431	80	44	3
F: Bouwnijverheid	107	0	1 565	198	44	16
C: Industrie	94	0	2 000	258	84	17
H: Vervoer en opslag	46	0	637	58	34	4
D: Energievoorziening	17	0	261	32	3	0
E: Waterbedrijven en afvalbeheer	15	0	160	5	3	0
A: Landbouw, bosbouw en visserij	9	0	251	35	11	2
B: Delfstoffenwinning	5	0	50	18	1	2
U: Extraterritoriale organisaties	0	0	0	0	0	0
T: Huishoudens	0	0	0	0	0	0
Totaal	8 008	0	96 814	11 565	4 041	782

Bron: Coosto en Dataprovider, bewerking CBS.

3.2.9b Aantal bedrijven dat twittert naar economische activiteit en tijd, periode 26-06-2017 tot 04-08-2017

Aantal bedrijven						
	juni		juli		aug	
	Chrono	Random	Chrono	Random	Chrono	Random
Bedrijfstak (SBI-indeling)						
M: Specialistische zakelijke diensten	628	0	2 456	716	378	102
J: Informatie en communicatie	304	0	1 057	325	228	64
G: Handel	241	0	1 338	359	186	40
R: Cultuur, sport en recreatie	230	0	886	239	128	31
Q: Gezondheids- en welzijnszorg	168	0	674	173	90	22
P: Onderwijs	145	0	547	150	70	22
N: Verhuur en overige zakelijke diensten	137	0	585	172	107	27
S: Overige dienstverlening	128	0	472	146	79	21
L: Verhuur en handel van onroerend goed	87	0	285	59	39	12
K: Financiële dienstverlening	75	0	245	69	58	10
O: Openbaar bestuur en overheidsdiensten	62	0	145	40	31	3
I: Horeca	36	0	214	46	27	9
F: Bouwnijverheid	33	0	194	54	15	2
C: Industrie	29	0	229	61	25	6
H: Vervoer en opslag	13	0	67	17	15	3
D: Energievoorziening	8	0	21	9	3	0
E: Waterbedrijven en afvalbeheer	.	0	.	.	.	0
A: Landbouw, bosbouw en visserij	7	0	43	18	9	.
B: Delfstoffenwinning	.	0	.	.	.	.
U: Extraterritoriale organisaties	0	0	0	0	0	0
T: Huishoudens	0	0	0	0	0	0
Totaal	2 340	0	9 486	2 659	1 491	377

Bron: Coosto en Dataprovider, bewerking CBS.

Tenslotte is ook de verdeling naar grootteklasse toegevoegd. In de tabellen 3.2.10a en b blijven natuurlijk de bedrijfstakken M, J, G en N het actiefst als het gaat om het twitteren. Dat geldt zowel voor het aantal sociale mediaberichten als voor het aantal bedrijven dat twittert.

Veruit de meeste Twitterberichten worden verstuurd door de kleinste bedrijven. Dat geeft echter een vertekend beeld. Gemiddeld posten de grotere bedrijven namelijk meer Twitterberichten dan de kleinere bedrijven: gemiddeld 12,8 Twitterberichten in grootteklasse 3 tegenover gemiddeld 9,7 Twitterberichten in grootteklasse 0.

In de periode, waarin is waargenomen, worden gemiddeld de meeste Twitterberichten per bedrijf gepost door de combinatie bedrijfstak Informatie en communicatie en grootteklasse 3: gemiddeld 50,5 berichten per bedrijf en de combinatie bedrijfstak Delfstofwinning en grootteklasse 3⁵⁾: gemiddeld 38 berichten per bedrijf. Ook de combinaties bedrijfstak Horeca en grootteklasse 1 (30,9 Twitterberichten) en bedrijfstak Informatie en technologie en grootteklasse 2 (25,6 Twitterberichten) scoren hoog. Gemiddeld worden weinig Twitterberichten verstuurd door de combinaties bedrijfstak Waterbedrijven en grootteklasse 1 en bedrijfstak Horeca en grootteklasse 3. Dit laatste is opvallend gezien het relatief grote aantal gemiddelde Twitterberichten dat door de grotere bedrijven in de Horeca wordt verzonden.

⁵⁾ Onder andere grote oliemaatschappijen.

3.2.10a Aantal sociale mediaberichten (Twitter) naar economische activiteit en grootteklasse, periode 26-06-2017 tot 04-08-2017

SBI indeling (1-digit)	Grootteklassen				Subtotaal
	0	1	2	3	
A: Landbouw, bosbouw en visserij	155	141	12	0	308
B: Delfstoffenwinning	0	0	0	76	76
C: Industrie	1 539	508	374	32	2 453
D: Energievoorziening	202	45	17	49	313
E: Waterbedrijven en afvalbeheer	36	2	89	56	183
F: Bouwnijverheid	603	1 015	244	68	1 930
G: Handel	7 968	1 947	785	181	10 881
H: Vervoer en opslag	604	112	58	5	779
I : Horeca	1 854	2 314	62	6	4 236
J: Informatie en communicatie	13 720	2 196	2 403	858	19 177
K: Financiële dienstverlening	2 611	360	129	92	3 192
L: Verhuur en handel van onroerend goed	3 691	223	96	16	4 026
M: Specialistische zakelijke diensten	32 551	2 030	618	76	35 275
N: Verhuur en overige zakelijke diensten	7 839	1 094	1 116	523	10 572
O: Openbaar bestuur en overheidsdiensten	49	53	1 177	395	1 674
P: Onderwijs	4 465	393	257	219	5 334
Q: Gezondheids- en welzijnzorg	3 507	733	662	523	5 425
R: Cultuur, sport en recreatie	5 794	883	652	34	7 363
S: Overige dienstverlening	6 420	877	708	8	8 013
Totaal	93 608	14 926	9 459	3 217	121 210

Bron: Coosto en Dataprovider, bewerking CBS.

3.2.10b Aantal bedrijven dat twittert naar economische activiteit en grootteklasse, periode 26-06-2017 tot 04-08-2017

SBI indeling (1-digit)	Grootteklassen				Subtotaal
	0	1	2	3	
A: Landbouw, bosbouw en visserij	46	13	7	0	66
B: Delfstoffenwinning	0	0	0	2	2
C: Industrie	129	90	69	9	297
D: Energievoorziening	16	7	3	5	31
E: Waterbedrijven en afvalbeheer	10	2	14	5	31
F: Bouwnijverheid	129	54	60	12	255
G: Handel	1 299	330	134	21	1 784
H: Vervoer en opslag	45	23	19	2	89
I : Horeca	180	75	17	4	276
J: Informatie en communicatie	1 113	217	94	17	1 441
K: Financiële dienstverlening	227	59	21	16	323
L: Verhuur en handel van onroerend goed	284	45	33	2	364
M: Specialistische zakelijke diensten	2 904	318	86	10	3 318
N: Verhuur en overige zakelijke diensten	539	129	106	18	792
O: Openbaar bestuur en overheidsdiensten	8	6	144	28	186
P: Onderwijs	626	60	49	16	751
Q: Gezondheids- en welzijnzorg	594	123	106	81	904
R: Cultuur, sport en recreatie	999	152	53	2	1 206
S: Overige dienstverlening	497	99	53	1	650
Totaal	9 645	1 802	1 068	251	12 766

Bron: Coosto en Dataprovider, bewerking CBS.

De conclusie van deze verkenning is dat bedrijven vooral Twitter gebruiken, al dan niet in combinatie met andere sociale mediatypen, zoals Facebook en LinkedIn. Sociale media worden vooral gepost in de bedrijfstakken M (Specialistische zakelijke dienstverlening), J (Informatie en communicatie) en N (Verhuur van overige zakelijke diensten).

Ook hier is weer te zien dat big data van sociale mediaberichten en -profielen veel gedetailleerder analyses en outputmogelijkheden bieden dan de data van traditionele enquêtes, zowel in de tijd als naar bijvoorbeeld economische activiteit en grootteklasse.

Onderwerpen van sociale mediaberichten (woordwolken)

Een volgende onderzoeksvraag is hoe sociale mediaberichten worden gebruikt door bedrijven? In hoeverre zegt de inhoud van sociale mediaberichten iets over het gebruik ervan en kunnen op basis daarvan statistische indicatoren worden samengesteld? Om dit te kunnen beoordelen, is geëxperimenteerd met de methode met zogenoemde woordwolken ('word clouds').

3.2.11 Woordwolken van Twitterberichten naar grootteklasse, periode 26-06-2017 tot 04-08-2017

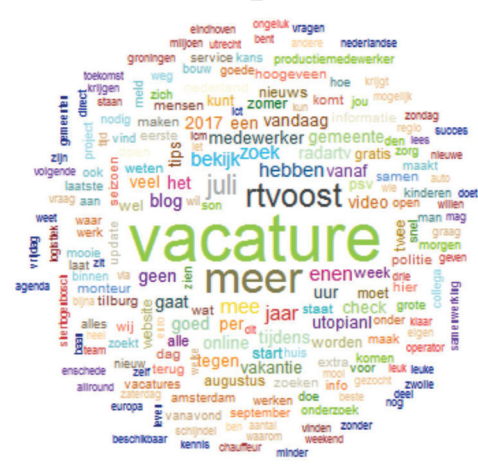
GK_0



GK_1



GK_2



GK_3



Bron: Coosto en Dataprovider, bewerking CBS.

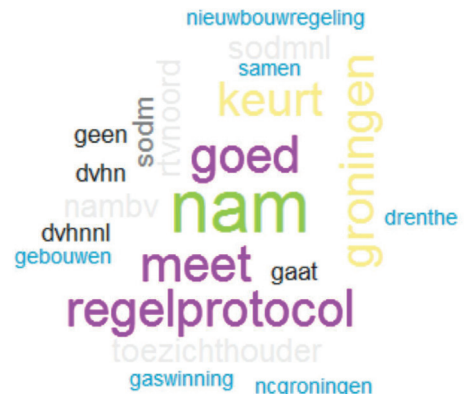
Landbouw, bosbouw en visserij



Industrie



Winning van delfstoffen



Productie en distributie van en handel in
elektriciteit, aardgas, stoom en gekoelde
lucht



CBS | Gebruik van sociale media | April 2019 36

Figuur 3.2.11 laat zien dat veel bedrijven Twitter vooral gebruiken voor communicatie rond 'vacatures'. Daarbij maakt de grootte van het bedrijf eigenlijk weinig uit.

Als gekeken wordt naar de plots op basis van economische activiteit (figuren 3.2.12 en 3.2.13) dan gaat het bij de bedrijfstak Landbouw veel over 'biologisch en duurzaam' en bij de bedrijfstak Winning van delfstoffen veel over de 'NAM, Groningen en regelprotocol'. Bij de bedrijfstak Financiële Instellingen komen logischerwijs termen voor, zoals 'AEX, beurs en trade-idee'. Ten slotte, komt bij de bedrijfstak Informatie en communicatie vaak de woorden 'korting' en 'kortingscode' voor. Dit is waarschijnlijk vooral gerelateerd aan de reclame-uitingen vanuit de telecombedrijven.

3.2.13 Woordwolken van Twitterberichten naar economische activiteit, periode 26-06-2017 tot 04-08-2017

Financiële instellingen

Logies, maaltijd- en drankverstrekking



Bron: Coosto en Dataprovider, bewerking CBS.

Bij andere bedrijfstakken in de plots van de figuren 3.2.12 en 3.2.13 is het minder duidelijk te zien om welk onderwerp het precies gaat. Dat geldt ook voor de meeste andere bedrijfstakken van de SBI, die hier niet zijn geplot. Ook daar springt er geen specifiek onderwerp uit. Het resultaat van de woordwolken hangt ook samen met het feit of veel of weinig sociale mediaberichten worden gepost, of dat het gaat om veel kleine of juist om grote bedrijven. Daardoor kan het zijn dat er juist wel of juist niet een duidelijk onderwerp uitspringt. Wat logischerwijs ook vaak terugkomt in de Twitterberichten zijn aanduidingen voor de tijd en plaats. Bijvoorbeeld 'nederland, utrecht, jaar en 2017'.

Woordwolken kunnen ook op een andere manier gebruikt worden. Zo kunnen woordwolken onderling worden vergeleken om te zien of er overeenkomsten en verschillen zijn tussen bijvoorbeeld bepaalde bedrijfstakken.

Daarnaast kan ook worden onderzocht of specifieke thema's vaak in de sociale mediaberichten terugkomen. Bijvoorbeeld voor promotie-uitingen, het aanbieden van vacatures of bepaalde gebeurtenissen in het nieuws. Zo is hier, als verkenning, gekeken naar thema's zoals Webcare, B2B-relatie, Reclame, Informatie, Alert, Nieuws en HRM. Zie tabel 3.2.14. De rechterkolom van deze tabel laat de trefwoorden zien, die gebruikt zijn om per thema in de sociale mediaberichten te zoeken en deze vervolgens te selecteren.

De kolom 'Aantal records' laat het aantal records of beter Twitterberichten zien, die op basis van de gekozen trefwoorden in de data zijn gevonden. In vergelijking met de andere hier gekozen thema's, worden er relatief veel Twitterberichten gepost over de thema's HRM, Alert en Nieuws. De geselecteerde records of Twitterberichten, waarin de gekozen trefwoorden voorkomen, kunnen vervolgens per thema in een woordwolk worden gepresenteerd. Zie figuur 3.2.16. Dat geeft inzicht in het belang van bepaalde onderwerpen, in relatie tot het gekozen thema. Bij het thema 'Alert' komen woorden voor zoals 'politie en ongeluk' en bij het thema B2B het woord 'zakendoen'.

Een ander perspectief is door te kijken naar bepaalde gebeurtenissen. Tabel 3.2.15 laat daarvan enkele voorbeelden zien, zoals 'Nepnieuws', 'Fipronil' (schandaal met het schoonmaken van kippenstallen), 'MH17' (het neerhalen van een vliegtuig) en 'Nouri' (Ajax-voetballer die een hartstilstand kreeg). Bij deze voorbeelden is eerst gekeken naar woorden, die in krantenkoppen voorkwamen. Daaruit is een aantal trefwoorden gekozen, waarmee in de sociale mediaberichten is gezocht. Van de gevonden Twitterberichten zijn vervolgens per gebeurtenis weer plots met woordwolken gemaakt. Zie figuur 3.2.17. Bij de gebeurtenis rond het fipronil-schandaal zie je bijvoorbeeld de actoren terug, die een rol hebben gespeeld, zoals nwva, pluimveebedrijven en -sector, dierenorganisaties en consument. Maar bijvoorbeeld ook het gevolg van het schandaal: fipronilgehalte, gif, eieren en ruimen. Voor de gebeurtenis 'MH17' geldt hetzelfde. Daar zie je woorden terug zoals 'slachtoffers, verdachten en nabestaanden (actoren) en monument, onthulling en berecht (gevolg)'.

Ten slotte, kan bijvoorbeeld ook nog gekeken worden naar de ontwikkeling van onderwerpen of woorden in de tijd, verschillen tussen de chronologische data (door de week geposte berichten) de random data (in het weekeinde en in de avond geposte berichten) of het verschil tussen tweets en retweets.

3.2.14 Overzicht van het voorkomen van specifieke thema's in Twitterberichten

Thema	Aantal records ¹⁾	Gebruikte trefwoorden
Webcare	51 1 367	"klant", "zorg", "klachten", "care", "webcare", "tevredenheid", "kwaliteit", "product", "kwaliteit", "dienst Voorbeeld: 1 juli/augustus 2017: "wijzigt", "KPN", "Telfort", "T-mobile", "bel- en abonnementstarieven"
B2B-relatie	201 3 292	"b2b", "inkopen", "samenwerking", "projectpartner", "meeting", "evenement", "event", "bedrijven", "samenwerken", "business partner", "zaken doen", "zakelijk dienstverlening" Voorbeeld: 14 juli 2017 "B2B", "beter zicht krijgen", "marketinginspanning"
Reclame	1 990 17 749	"super korting", "aanbieding", "aanbieden", "producten", "offerte", "dienstverlening" ("spam", 2074) Voorbeeld: 22 juli 2017 "Festival experimenteert", "spam", "in het echt"
Informatie	2 068 31 336	"voorlichting", "info"
Nieuws	3 339 197 220	"nieuws", "news"
Alert	3 973 82 209	"alert", "brand", "politie", "ongeluk", "file" ("storing", 4338; NL-alert) Voorbeeld: 3 juli 2017 "NL Alert", "controlebericht", "maandag", "12 uur verzonden"
HRM	7 913 80 280	"vacature", werk, "job", "bedrijf", "werk", "ervaring", "opleiding", "profiel", "solliciteren", "nieuwe collega", "personeel", "recruitment", "gezocht", "medewerker" Voorbeeld: 31 juli 2017 "Buslijnen vallen stil", "tekort aan chauffeurs"

Bron: CBS.

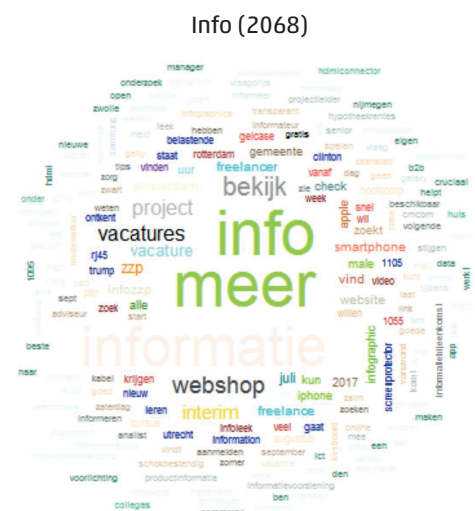
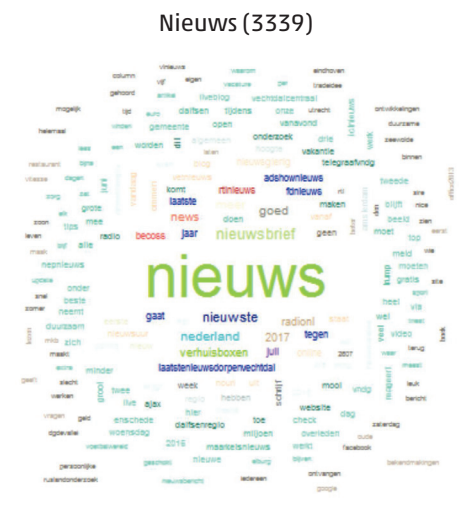
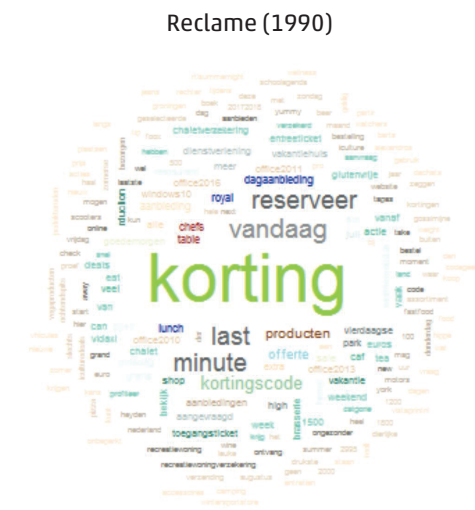
¹⁾ Het aantal records komt uit het gekoppelde analysebestand.

3.2.15 Overzicht van het voorkomen van woorden rond gekozen gebeurtenissen in Twitterberichten

Thema	Tijdstip	Aantal records	Trefwoorden uit krantenkoppen	Gebruikte trefwoorden
Nep-nieuws	30 juni	41	"Twitter", "zoekt", "manieren", "gebruikers", "nepnieuws", "laten", "melden"	"nepnieuws", "fakenews", "fakenieuws", "hoax"
Fipronil	31 juli	49	"miljoenen eieren", "vernietigd", "Fipronil-fipronil", schandaal"	("anti-luizen", "bloedluis"; # 60)
MH17	17 juli	157	"herdenking", "MH17"	"mh17"
Nouri	8 juli	277	"Abdelhak Nouri", "stort in", "oefenduel Ajax"	"nouri", "appie", "staystrong"

Bron: CBS.

3.2.16 Plots van woordwolken van Twitterberichten van specifieke thema's



Bron: Coosto en Dataprovider, bewerking CBS.

De conclusie van deze verkenning is, dat analyses met methoden zoals woordwolken ('word clouds') in potentie bruikbaar zijn om een eerste indruk te krijgen van de inhoud van de sociale mediaberichten. Daaruit kunnen inzichten worden verkregen van wat belangrijk wordt gevonden door de twitteraars, maar bijvoorbeeld ook wie een rol speelt en wat de gevolgen zijn van de gebeurtenis. Het geeft ook de relatie aan tussen woorden of onderwerpen die bij een thema horen.

Men moet wel beseffen dat dit soort analyses nog in de kinderschoenen staat. Extra bewerkingen zijn bijvoorbeeld nodig als men combinaties van woorden en delen van zinnen wil meenemen (bijvoorbeeld via een koppelteken). Verder is het moeilijk om de context in de analyses mee te nemen. Bijvoorbeeld als het gaat om het verschil tussen de betekenis van 'een vacature' of 'geen vacature'. 'Een', 'geen' en 'vacature' zullen in dit voorbeeld los in de woordwolk terugkomen. Verdere ontwikkeling van dit soort technieken is dus nog nodig voordat ze echt goed bruikbaar zijn en ook daadwerkelijk statistische indicatoren kunnen opleveren. Sentiment-analyse is mogelijk een uitzondering. Naar sentiment-analyse is al veel onderzoek gedaan.

Het bereik van sociale mediaberichten

De volgende casestudie, die in dit onderzoek is uitgevoerd, betreft de vraag of er met de beschikbare data indicatoren kunnen worden gemaakt over de reikwijdte van een tweet of set van tweets, als proxy voor het belang ervan? Daarvoor zijn verschillende opties.

Een van die opties is door te kijken naar het totaal aantal views of volgers van een tweet of retweet, eventueel geaggregeerd naar economische activiteit, grootteklasse of onderwerp.

De variabelen 'aantal views' en 'aantal volgers' komen per record (= sociaal mediabericht) voor in het Coosto-bestand. Het maakt weinig uit of nu naar 'volgers' of 'views' wordt gekeken. De overeenkomst tussen deze beide variabelen is erg groot: vaak gaat het om dezelfde aantallen. Voor de duidelijkheid: het gaat hier niet om het aantal retweets, maar het bekijken of volgen van geposte sociale mediaberichten van een bepaald profiel. Dat kan een tweet of een retweet zijn.

Bij de beschrijving van het verwerkingsproces in hoofdstuk 2 is al aangegeven dat de sociale mediaberichten, waarvan het aantal volgers van de profielnaam meer dan 1 miljoen is, uit de populatie verwijderd zijn. Om een vertekening van de resultaten te voorkomen, is er dus gewerkt met een bovengrens. Wel zal het verwijderen van sociale mediaberichten met profielnamen met meer dan 1 miljoen volgers (of views) het gemiddelde iets naar beneden brengen. Voor een onderlinge vergelijking maakt dit echter niet zoveel uit.

3.2.17 Woordwolken op basis van Twitterberichten van recente gebeurtenissen

Nepnieuws (41)



Fipronil (45)



Nouri (277)



MH17 (157)



Bron: Coosto en Dataprovider, bewerking CBS.

Naast de som van het aantal volgers, is een vergelijkbare en betere indicator 'het gemiddelde aantal views' (som aantal views/som aantal berichten). Daar is hiermee geanalyseerd. Daarbij kan het gaan om het gemiddeld aantal views van alle sociale mediaberichten van een profiel, een bedrijf, een economische activiteit, een grootteklasse of een thema. Een andere indicator is de verhouding of ratio van tweets/retweets. Zie verder tabel 3.2.18. Ook bij deze analyses is weer alleen gebruikgemaakt van Twitterberichten.

In de meeste gevallen is het bereik van tweets, met het gemiddeld aantal views als indicator, groter dan die van retweets, hoewel dat niet voor alle SBI's en grootteklassen geldt. Het grootste gemiddelde bereik hebben de tweets in de bedrijfstakken Q (Gezondheidszorg en welzijn) en M (Specialistische zakelijke dienstverlening). Bij de retweets gaat het om de bedrijfstakken L (Verhuur van en handel in onroerend goed) en J (Informatie en communicatie). Zie tabel 3.2.18.

Een analyse van het gemiddeld aantal views (tweets en retweets) naar grootteklasse laat zien dat bedrijven met meer dan 500 werknemers en minder dan 9 werknemers het grootste bereik hebben met hun sociale mediaberichten: respectievelijk gemiddeld 75 duizend en 130 duizend views.

De uitkomsten kunnen ook worden gestandaardiseerd (ten opzichte van het totale gemiddelde), zodat ze makkelijker in één plot kunnen worden gepresenteerd. Zie figuur 3.2.19. In deze plot zijn de verschillende bedrijfstakken (op letter) tegen elkaar uitgezet op basis van enerzijds het gemiddeld aantal views (of volgers) en anderzijds de som van het aantal tweets en retweets. Hierdoor ontstaan vier kwadranten. Dit is gedaan voor een analysebestand met en zonder uitbijters (zie boven). In de plot met uitbijters is bijvoorbeeld te zien dat de bedrijfstak M (Specialistische zakelijke diensten) iets hoger scoort dan het gemiddelde aantal views van alle bedrijfstakken, terwijl de bedrijfstak J (informatie en communicatie) net onder dat gemiddelde zit. De bedrijfstak M scoort ook bijna twee keer zo hoog als het gaat om de som van het aantal tweets en retweets dan de bedrijfstak J. Wordt naar de plot gekeken zonder uitbijters dan zijn er kleine verschuivingen. Dat geldt met name voor de bedrijfstak G (Handel). Het grootste bereik, als het gaat om de som van het aantal tweets en retweets, heeft de bedrijfstak L (Verhuur van en handel in onroerend goed) en in mindere mate bedrijfstak C (Industrie).

3.2.18 Het gemiddeld aantal views van Twitterberichten naar economische activiteit en grootteklasse

	GK	Type bericht		T+RT		Type bericht		Hoogste aantal tweets (T) en retweets (RT)	
		Tweet (T)		Retweet (RT)		T/RT			
		Gem. aantal views	Som berichten	Gem. aantal views	Som berichten	Gem. aantal views	Som berichten	Gem. aantal views	Som berichten
Bedrijfstakken									
M: Specialistische zakelijke diensten		127 046	27 073	45 892	4 579	86 469	31 652	2,8	5,9
	0	249 680	24 884	152 591	4 220	201 136	29 104	1,6	5,9
	1	184 551	1 603	20 307	251	102 429	1 854	9,1	6,4
	2	69 226	525	6 444	98	37 835	623	10,7	5,4
	3	4 726	61	4 227	10	4 476	71	1,1	6,1
B: Bouwnijverheid		62 935	1 235	72 238	482	67 586	1 717	0,9	2,6
	0	1 794	474	2 139	81	1 967	555	0,8	5,9
	1	241 492	553	275 910	325	258 701	878	0,9	1,7
	2	1 867	159	2 445	58	2 156	217	0,8	2,7
	3	6 585	49	8 457	18	7 521	67	0,8	2,7
R: Cultuur, sport en recreatie		93 912	5 042	85 283	1 646	89 597	6 688	1,1	3,1
	0	14 574	3 857	14 928	1 364	14 751	5 221	1	2,8
	1	25 360	600	44 068	203	34 714	803	0,6	3
	2	158 767	560	97 065	70	127 916	630	1,6	8
	3	176 946	25	185 070	9	181 008	34	1	2,8
K: Financiële diensten		4 547	2 665	6 348	197	5 447	2 862	0,7	13,5
	0	12 504	2 202	12 856	123	12 680	2 325	1	17,9
	1	1 381	269	2 048	49	1 714	318	0,7	5,5
	2	2 018	127	1 165	7	1 592	134	1,7	18,1
	3	2 286	67	9 321	18	5 804	85	0,2	3,7

3.2.18 Het gemiddeld aantal views van Twitterberichten naar economische activiteit en grootteklasse (vervolg)

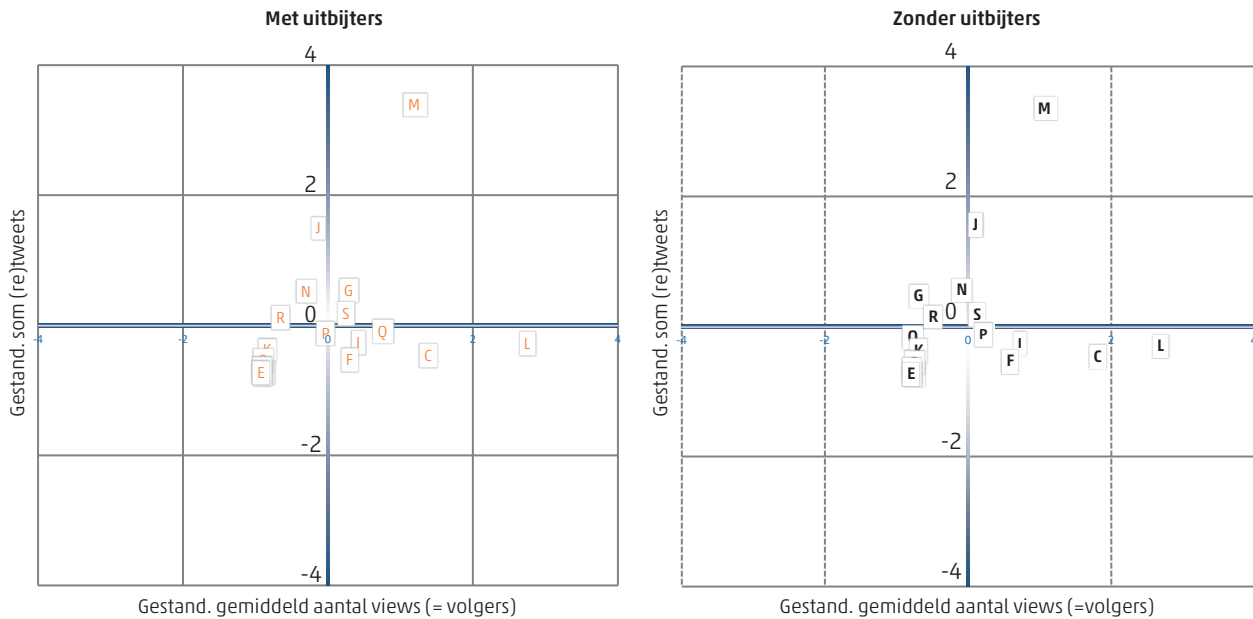
	GK	Type bericht		T+RT		Type bericht		Hoogste aantal tweets (T) en retweets (RT)	
		Tweet (T)		Retweet (RT)		T/RT			
		Gem. aantal views	Som berichten	Gem. aantal views	Som berichten	Gem. aantal views	Som berichten	Gem. aantal views	Som berichten
Q: Gezondheids- en welzijnszorg		180 148	3 974	20 183	1 041	100 165	5 015	8,9	3,8
	0	236 686	2 497	15 297	703	125 992	3 200	15,5	3,6
	1	476 660	615	55 055	85	265 857	700	8,7	7,2
	2	4 329	476	7 040	148	5 685	624	0,6	3,2
	3	2 918	386	3 339	105	3 128	491	0,9	3,7
G: Handel		150 347	9 228	34 955	652	92 651	9 880	4,3	14,2
	0	142 270	6 694	36 143	456	89 207	7 150	3,9	14,7
	1	9 968	1 684	8 841	140	9 405	1 824	1,1	12
	2	436 991	669	91 851	49	264 421	718	4,8	13,7
	3	12 160	181	2 984	7	7 572	188	4,1	25,9
C: Industrie		108 946	1 967	73 552	251	91 249	2 218	1,5	7,8
	0	408 007	1 199	274 620	177	341 314	1 376	1,5	6,8
	1	7 552	406	1 539	37	4 546	443	4,9	11
	2	16 579	338	12 236	32	14 408	370	1,4	10,6
	3	3 647	24	5 814	5	4 730	29	0,6	4,8
J: Informatie en communicatie		108 130	15 638	105 593	1 555	106 862	17 193	1	10,1
	0	90 584	10 945	70 129	1 232	80 357	12 177	1,3	8,9
	1	10 803	1 947	2 426	109	6 615	2 056	4,5	17,9
	2	80 441	2 098	50 072	95	65 257	2 193	1,6	22,1
	3	250 693	648	299 745	119	275 219	767	0,8	5,4
A: Landbouw, bosbouw en visserij		3 032	213	1 389	58	2 210	271	2,2	3,7
	0	3 636	94	1 915	40	2 776	134	1,9	2,4
	1	3 329	113	1 057	14	2 193	127	3,1	8,1
	2	2 131	6	1 194	4	1 663	10	1,8	1,5
I: Horeca		85 701	3 533	73 677	228	79 689	3 761	1,2	15,5
	0	325 443	1 504	282 022	164	303 732	1 668	1,2	9,2
	1	2 720	1 974	5 798	50	4 259	2 024	0,5	39,5
	2	4 258	50	3 271	13	3 765	63	1,3	3,8
	3	10 384	5	3 616	1	7 000	6	2,9	5
P: Onderwijs		33 066	4 061	42 511	746	37 789	4 807	0,8	5,4
	0	112 138	3 362	128 971	616	120 554	3 978	0,9	5,5
	1	1 506	334	3 076	40	2 291	374	0,5	8,4
	2	1 833	219	2 461	30	2 147	249	0,7	7,3
	3	16 789	146	35 537	60	26 163	206	0,5	2,4
O: Openbaar bestuur en overheidsdiensten		5 305	1 264	6 446	327	5 876	1 591	0,8	3,9
	0	6 695	35	9 402	12	8 048	47	0,7	2,9
	1	3 180	38	3 532	13	3 356	51	0,9	2,9
	2	4 966	892	3 377	225	4 171	1 117	1,5	4
	3	6 381	299	9 472	77	7 926	376	0,7	3,9
S: Overige dienstverlening		48 048	6 115	58 618	1 056	52 578	7 171	0,8	5,8
	0	155 653	4 853	158 288	796	156 971	5 649	1	6,1
	1	10 356	669	4 701	141	7 528	810	2,2	4,7
	2	23 432	585	12 866	119	18 149	704	1,8	4,9
	3	2 753	8			2 753	8		
D: Energievoorziening		4 361	213	6 536	68	5 293	281	0,7	3,1
	0	12 814	135	938	51	6 876	186	13,7	2,6
	1	3 698	26	3 609	13	3 653	39	1	2
	2	231	12			231	12		
	3	702	40	15 062	4	7 882	44	0	10

3.2.18 Het gemiddeld aantal views van Twitterberichten naar economische activiteit en grootteklasse (slot)

	Type bericht				T+RT		Type bericht		Hoogste aantal tweets (T) en retweets (RT)	
	Tweet (T)		Retweet (RT)				T/RT			
	GK	Gem. aantal views	Som berichten	Gem. aantal views	Som berichten	Gem. aantal views	Som berichten	Gem. aantal views	Som berichten	
L: Verhuur van en handel in onroerend goed		111 938	3 206	113 510	406	112 611	3 612	1	7,9	Zeer kleine bedrijven bij verhuur/handel in on-roerend goed (0 t/m 9 wp) hebben de meeste vIEWS van T+RT. Geen vIEWS van RT bij zeer grote bedrijven (500 en meer wp).
	0	440 478	2 949	337 392	341	388 935	3 290	1,3	8,6	
	1	1 163	165	1 120	47	1 142	212	1	3,5	
	2	4 172	76	2 017	18	3 095	94	2,1	4,2	
	3	1 937	16			1 937	16			
N: Verhuur en overige zakelijke diensten		26 267	9 045	11 314	724	18 790	9 769	2,3	12,5	Zeer kleine bedrijven bij Verhuur overige zakelijke diensten (0 t/m 9 wp) hebben de meeste vIEWS van T en RT (RT= 4 T).
	0	96 705	6 559	26 349	586	61 527	7 145	3,7	11,2	
	1	4 654	965	4 833	82	4 744	1 047	1	11,8	
	2	2 349	1 008	12 868	51	7 608	1 059	0,2	19,8	
	3	1 359	513	1 206	5	1 282	518	1,1	102,6	
H: Vervoer en opslag		5 842	626	5 664	91	5 753	717	1	6,9	Zeer grote bedrijven (500 en meer wp) bij vervoer en opslag hebben de meeste vIEWS van T en RT (RT= 0,76T).
	0	2 826	482	1 645	74	2 236	556	1,7	6,5	
	1	1 173	93	1 740	6	1 457	99	0,7	15,5	
	2	5 849	47	1 411	10	3 630	57	4,1	4,7	
	3	13 519	4	17 858	1	15 688	5	0,8	4	
E: Waterbedrijven en afvalbeheer		923	164	1 842	28	1 317	192	0,5	5,9	Te weinig bedrijven.
	0	438	30	390	5	414	35	1,1	6	
	1	207	2			207	2			
	2	1 059	83	1 290	4	1 175	87	0,8	20,8	
	3	1 987	49	3 847	19	2 917	68	0,5	2,6	
B: Delfstoffenwinning		4 267	30	4 418	26	4 342	56	1	1,2	Zeer grote bedrijven (500 en meer wp) bij delfstoffenwinning hebben de meeste vIEWS van T en RT (RT= 0.5T).
	3	4 267	30	4 418	26	4 342	56	1	1,2	
Eindtotaal		64 489	95 292	42 422	14 161	53 771	109 453	1,5	6,7	

Bron: Coosto en Dataprovider, bewerking CBS.

3.2.19 Verhouding tussen gemiddeld views (=volgers) en som (re)tweets naar SBI, met/zonder uitbijters



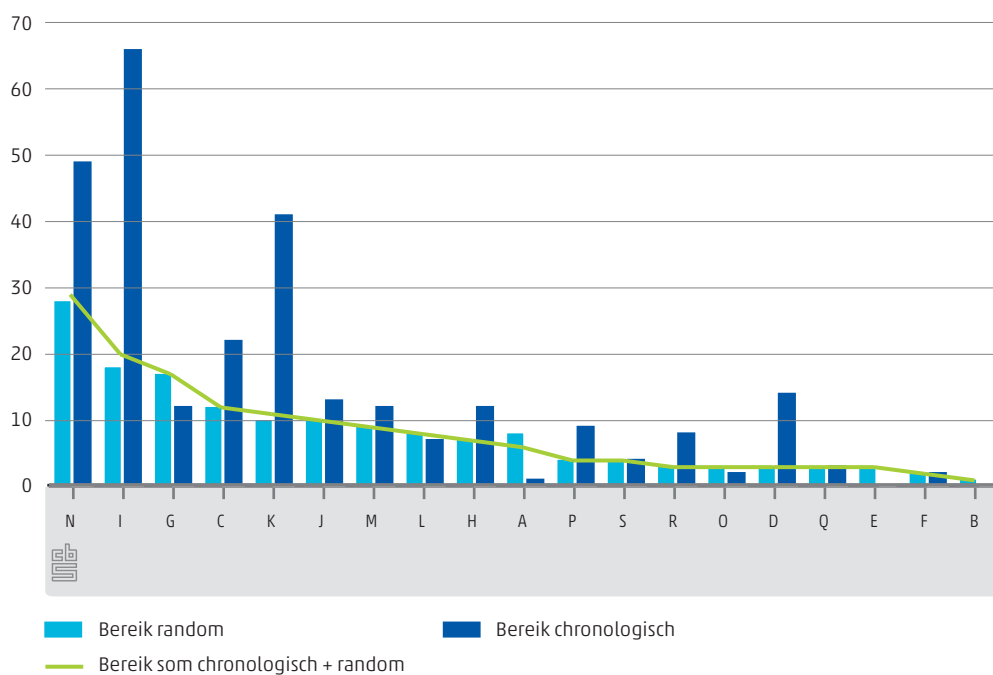
Bron: Coosto en Data provider, bewerking CBS.

Het nadeel van een (gestandaardiseerde) plot is dat het berekeningsproces niet altijd even gemakkelijk te begrijpen is voor de gebruiker. Daarom is gezocht naar een andere indicator, die het bereik per bedrijfstak weergeeft van tweets en retweets. Een mogelijkheid is het aantal tweets x het gemiddeld aantal views per bedrijfstak te delen door het aantal retweets x het gemiddelde aantal views per bedrijfstak. Dus:

$$\text{Ratio bereik tweets versus bereik retweets per bedrijfstak} = \frac{\sum \text{tweets} \times \overline{\text{views}}}{\sum \text{retweets} \times \overline{\text{views}}}$$

3.2.20 Ratio bereik tweets/retweets op basis van aantal views, chronologische, random en som chronologische en random data, naar economische activiteit

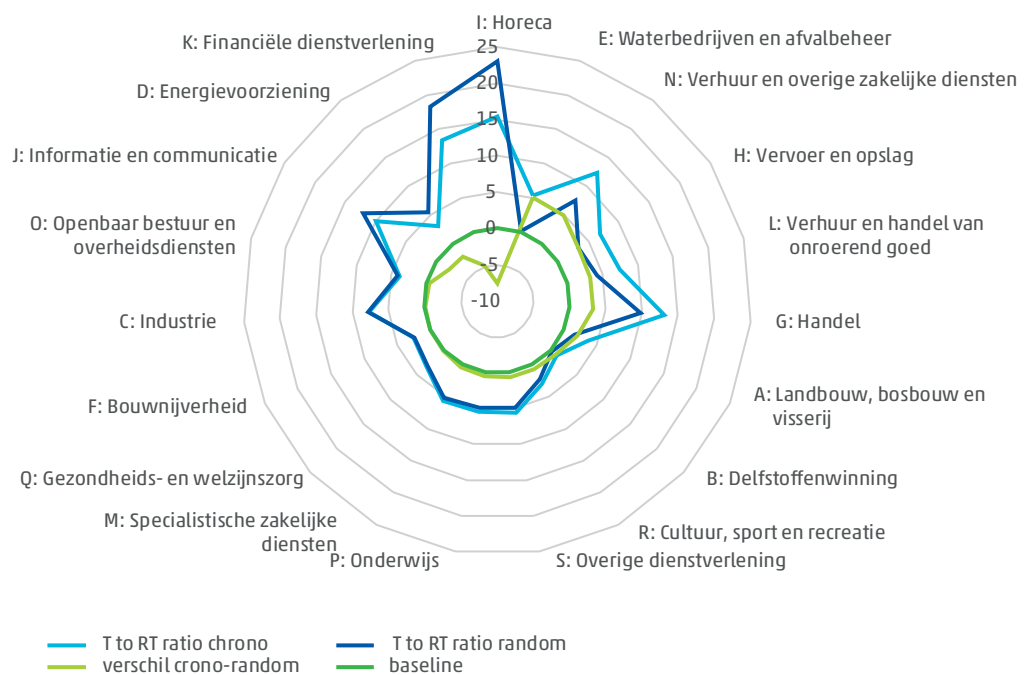
Ratio bereik tweets t.o.v. retweets



N	Verhuur en overige zakelijke diensten	A	Landbouw, bosbouw en visserij
I	Horeca	P	Onderwijs
G	Handel	S	Overige dienstverlening
C	Industrie	R	Cultuur, sport en recreatie
K	Financiële dienstverlening	O	Openbaar bestuur en overheidsdiensten
J	Informatie en communicatie	D	Energievoorziening
M	Specialistische zakelijke diensten	Q	Gezondheids- en welzijnszorg
L	Verhuur en handel van onroerend goed	E	Waterbedrijven en afvalbeheer
H	Vervoer en opslag	F	Bouwnijverheid
		B	Delfstoffenwinning

Bron: Coosto en Dataprovider, bewerking CBS.

3.2.21 Verschillen tussen bereik (som aantal tweets)/(som aantal retweets) van chronologische en random datasets



Bron: Coosto en Dataprovider, bewerking CBS.

Figuur 3.2.20 laat de verschillende uitkomsten zien op basis van de chronologische data, random data en chronologische en random data samen. De bedrijfstakken met het meeste bereik zijn N (Verhuur en overige zakelijke dienstverlening) en I (Horeca). Het kleinste bereik hebben de bedrijfstakken B (Delfstofwinning) en F (Bouwnijverheid). Voor sommige bedrijfstakken is er echter wel een verschil of het gaat om het bereik overdag of in de avond en het weekeinde. Een goed voorbeeld is de bedrijfstak I (Horeca), waar het bereik in de avond en het weekeinde veel groter is dan overdag en door de week. Dit kan ook in een spindigram worden weergegeven: zie figuur 3.2.21.

De conclusie van deze verkenning is, dat de gebruikte indicatoren een goed beeld kunnen geven van het bereik of beter het belang van (groepen van) sociale mediaberichten, in dit geval tweets en retweets van Twitter.

Ook hier wordt weer aangetoond, dat met sociale mediaberichten dieper gravende analyses kunnen worden uitgevoerd op de beschikbare data. Dit is niet mogelijk met data van enquêtes, zoals de ICT-enquête bedrijven. Er kan gemakkelijk onderscheid worden gemaakt naar economische activiteit, grootteklasse of thema.

Process mining

Ten slotte, is als laatste case gekeken naar de mogelijkheden om sociale mediaberichten door de tijd heen te analyseren en te visualiseren. Sociale mediaberichten, in dit geval weer Twitterberichten, bevatten namelijk een tijdsaanduiding. Daarbij worden goede visualisaties van de resultaten steeds belangrijker.

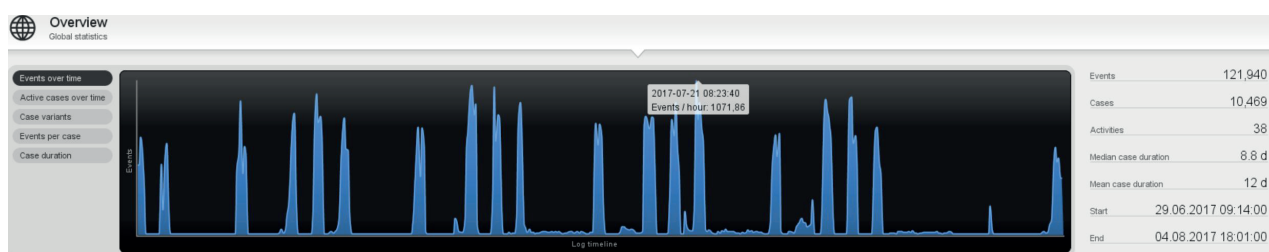
Hier is geëxperimenteerd met het programma DISCO (<https://fluxicon.com/disco>). Dit programma richt zich eigenlijk op het in kaart brengen van processtappen door de tijd heen, bijvoorbeeld om verschillen te kunnen zien tussen de bestaande en de gewenste situatie. Dit wordt gedaan op basis van een grote hoeveelheid data. Het gaat dan om stappen in het proces waar een tijdsindicatie aan vastzit. Daarmee kan op basis van een visualisatie inzicht worden verkregen in wat er volgtijdelijk in het proces gebeurt. Daarom wordt gesproken over 'process mining'.

Dit pakket kan echter ook gebruikt worden om de relaties tussen tweets en retweets per bedrijfstak of bijvoorbeeld stromen van toeristen door de tijd heen inzichtelijk te maken. De vraag zou bijvoorbeeld kunnen zijn: welke bedrijven twitteren veel en welke bedrijven retwitteren veel?

In DISCO is een aantal variabelen als parameter nodig, namelijk de 'case ID', de 'activity' en de 'timestamp'. Bij de sociale mediaberichten zijn de case-id's de bedrijfseenheden (bijvoorbeeld de unieke identificatie van een record). De activity's, als kenmerken van een record of sociaal mediabericht, zijn de bedrijfstakken (19 SBI's) en het soort bericht (tweet of retweet).

Op basis van deze input maakt DISCO de hier gepresenteerde plots. Daarbij wordt gebruikgemaakt van de tijdsdimensie (volgorde).

3.2.22 Tijdlijn van geposte Twitterberichten in de periode 26-06-2017 tot 04-08-2017



Bron: Coosto en Dataprovider, bewerking CBS met DISCO.

Zo laat figuur 3.2.22 bijvoorbeeld de tijdlijn zien van alle sociale mediaberichten van bedrijven, tweets en retweets, die via Coosto zijn waargenomen. Daaruit (zie rechts in figuur 3.2.22) is af te lezen dat het gaat om 121 940 sociale mediaberichten (events), gepost door 10 469 bedrijven (cases) en 38 mogelijke activiteiten (activity's=combinaties SBI x type bericht). De 'gaten' in de tijdlijn laten vooral de weekeinden en de dinsdagen zien: tijdstippen waarop minder sociale mediaberichten door bedrijven zijn gepost en het moment dat er minder is waargenomen. Dit soort plots kunnen bijvoorbeeld ook worden gemaakt per economische activiteit, grootteklasse of per gebeurtenis. Er kan dus ook getoond worden wanneer het twitteren over een thema of gebeurtenis begint en wanneer het eindigt en hoe het tussenliggende verloop is.

Een andere optie van DISCO is om stroomdiagrammen te maken, waarbij de 'activity's' als classificatievariabelen worden gebruikt. In het hier uitgevoerde experiment zijn alleen de sociale mediaberichten of records gebruikt, die geselecteerd zijn op basis van de trefwoorden van de thema's in tabel 3.2.14 met onder meer 'nieuws', 'info', 'reclame' e.d. Er is dus sprake van een deelverzameling van alle verzamelde sociale mediaberichten van bedrijven.

De verdeling van de sociale mediaberichten en hun relaties zijn gebaseerd op de variabelen bedrijfstak, soort bericht (tweet of retweet) en thema (nieuws, info e.d.). Deze parameters worden binnen DISCO activity's genoemd. Daarbij gaat het om de combinaties bedrijfstak x soort bericht x thema. Dit betreft alle mogelijke combinaties. Een paar daarvan komen terug in de blokjes van figuur 3.2.23. Om overzicht te houden wordt hier namelijk slechts een deel van de resultaten getoond: de top 5 combinaties met de meeste geposte sociale mediaberichten. De mate van detail van de visualisatie is in DISCO in te stellen via specifieke knoppen in het programma. Om de tekst van de mogelijke combinaties in de blokjes relatief kort te houden, zijn voor de bedrijfstakken afkortingen gebruikt:

- Verhuur Ovr: N: Verhuur en overige zakelijke diensten
- SpecZDnst: M: Specialistische zakelijke diensten
- ICT: J: Informatie en communicatie
- Handel: G: Handel
- Gezondheid: Q: Gezondheid- en welzijnzorg.

Nadat het bestand met de Coosto-data is in gelezen en alle parameters zijn gezet, kan op basis van het programma inzicht worden verkregen in allerlei stromen en relaties. Bijvoorbeeld welke bedrijfstakken de meeste sociale mediaberichten versturen, twitteren of retwitteren en op basis van welk thema ze dat doen. Figuur 3.2.23 is hier als voorbeeld gepresenteerd.

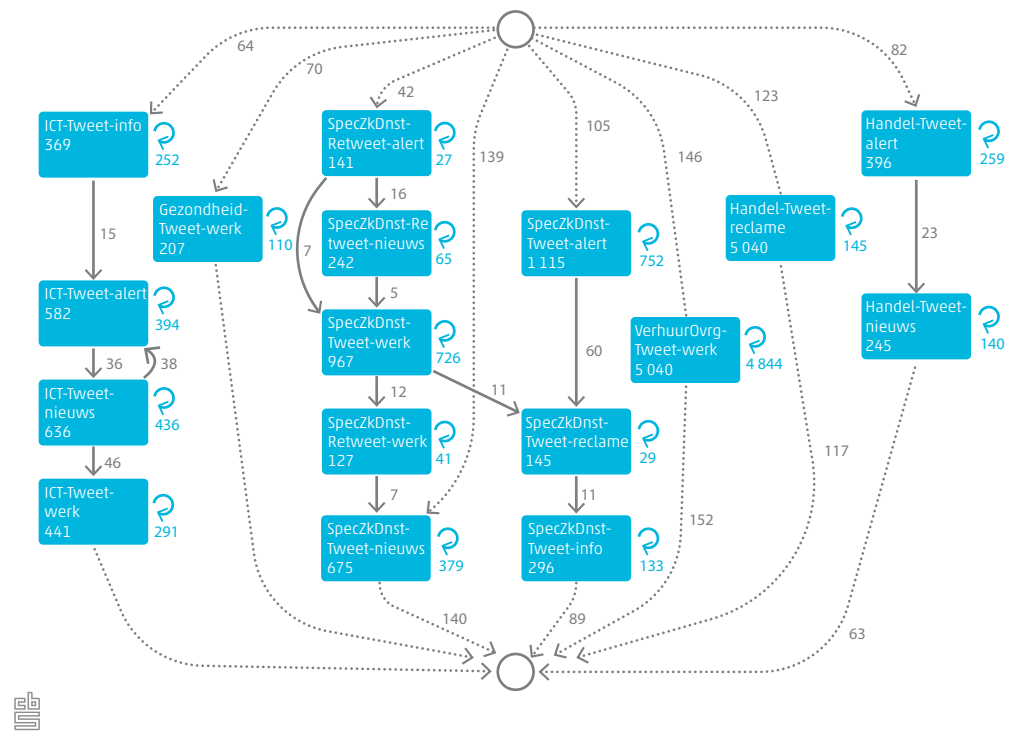
Zo laat figuur 3.2.23 vijf bedrijfstakken zien (van de 19), waarbij door de bedrijfstak Verhuur en zakelijke diensten (VerhuurOvr) vooral wordt getwitterd (5 040 sociale mediaberichten). Een groot deel daarvan bestaat uit het twitteren binnen dezelfde bedrijfstak (4 844 berichten), waarbij het gaat om hetzelfde thema, namelijk 'werk'. Op zich is dit logisch, omdat uitzendbureaus onderdeel zijn van deze bedrijfstak.

De bedrijfstak Specialistische zakelijke diensten laat hetzelfde beeld zien als het gaat om het twitteren (1 115 berichten, waarvan 752 binnen dezelfde bedrijfstak). Maar daar wordt ook getweet op basis van een andere thema's: eerst 'reclame' en daarna 'info'. Het retweeten verloopt via een ander pad en begint met het thema 'alert' (141 berichten).

Op deze wijze kan dus inzicht worden verkregen in welke bedrijfstakken vooral tweeten en welke bedrijfstakken juist retweeten, om hoeveel berichten het gaat en over welk thema ze dat doen. De getallen boven de lijnen geven de absolute frequentie weer van de stromen van sociale mediaberichten.

Met DISCO kan ook een dynamische versie van het stroomdiagram van figuur 3.2.23 wordt gemaakt. DISCO laat namelijk op basis van de tijdsaanduidingen in de records zien wat de stromen van berichten en hun effecten zijn door de tijd heen. De dikte van de lijnen laat de intensiteit (het aantal) sociale mediaberichten zien.

3.2.23 De relatie tussen statistische eenheden: sociale mediaberichten, profielnamen en bedrijfseenheden



Bron: Coosto en Dataprovider, bewerking CBS met DISCO.

De conclusie van deze laatste verkenning sluit aan bij de conclusie van de vorige paragraaf. De toepassing van de hierboven beschreven methoden maken het mogelijk om diepgravende analyses op de data uit te voeren. Op basis daarvan zouden nieuwe indicatoren over het gebruik van sociale media door bedrijven op het internet kunnen worden ontwikkeld. Deze indicatoren geven een beter inzicht in het gedrag van bedrijven op het terrein van sociale media. Ook hier kan dat worden gedaan op basis van economische activiteit, grootteklasse, thema of een andere gewenste indeling.

4. Conclusies en discussie

4.1 Algemeen

In dit rapport zijn de resultaten beschreven van een verkennend onderzoek naar de mogelijkheden om van het internet gescrapete sociale mediaberichten in te zetten voor het maken van officiële statistiek. Daarbij heeft de focus gelegen op (ICT-)indicatoren voor bedrijven. De reden daarvoor is dat het gebruik van sociale media door bedrijven de laatste jaren sterk is toegenomen. Dat heeft de vraag opgeroepen of dit soort big data ook ingezet kan worden voor het maken van officiële statistiek. Dit kan dan mogelijk leiden tot het vervangen van vragen in bestaande traditionele surveys met enquêtedrukvermindering als gevolg en nieuwe indicatoren voor bedrijven.

De uitgevoerde verkenningen hebben duidelijk gemaakt dat sociale mediaberichten en sociale mediaprofielen een zekere potentie hebben om te worden gebruikt voor het produceren van (officiële) statistiek. Veel meer dan de data van enquêtes, maken deze big data het mogelijk om veel dieper gravende analyses uit te voeren. Vraagt de ICT-enquête bijvoorbeeld alleen naar de verschillende typen sociale media die bedrijven gebruiken. Op basis van sociale mediaberichten kan niet alleen deze vraag worden beantwoord, maar kunnen ook analyses worden uitgevoerd naar de intensiteit van het posten van sociale mediaberichten, welke type sociale mediaberichten (bijvoorbeeld tweets of retweets) vooral worden gebruikt en over welke onderwerpen wordt gepost, zoals vacatures of nieuws. De resultaten zijn niet alleen veel gedetailleerder, maar zouden ook sneller beschikbaar kunnen komen. Over het algemeen waren de resultaten van de uitgevoerde verkenningen inhoudelijk plausibel. Kortom: op basis van sociale mediaberichten zou, zeker technisch gezien, een aantal (nieuwe) indicatoren kunnen worden geproduceerd over hoe bedrijven acteren op het internet: over het gebruik, het bereik en de inhoud van sociale mediaberichten, die door bedrijven worden gepost op het internet.

Tegelijkertijd is echter ook duidelijk geworden dat het niet gaat om 'laaghangend fruit': het is niet gemakkelijk om sociale mediaberichten op een betrouwbare en valide wijze te verzamelen, te analyseren en de resultaten goed te visualiseren. Daarvoor is extra onderzoek en verdere afstemming met gebruikers (wat zijn hun behoeften?) nodig. Een belangrijk aspect van de gebruikte data is bovendien expliciet niet in deze verkenning aanbod gekomen, namelijk de representativiteit van de data. De gegevens die hier zijn gebruikt betreffen namelijk 'slechts' data van zes weken van de maanden juni–augustus 2017. Deze data is in eerste instantie verzamelt ten behoeve van het opzetten van een register van domeinnamen, waardoor alleen sociale mediaberichten zijn meegenomen met een URL in de tekst. Dat roept de vraag op hoe representatief deze data zijn? Deze vraag was echter niet het doel van dit onderzoek. Ten slotte, konden de data van Coosto met sociale mediaberichten worden gekoppeld aan die van Dataprovider met informatie over bedrijven. Was dit niet mogelijk geweest, dan had dit onderzoek niet kunnen worden uitgevoerd. Om sociale mediaberichten voor indicatoren voor bedrijven in te zetten, is het namelijk wel nodig dat een onderscheid kan worden gemaakt tussen sociale mediaberichten van personen en die van bedrijven. Het voordeel van de koppeling met het Dataprovider-bestand was ook dat de variabelen bedrijfseenheid (gekoppeld aan het ABR),

grootteklasse en economische activiteit in de data beschikbaar kwamen. Grootteklasse en economische activiteit zijn belangrijke outputvariabelen.

Ook de gebruikte methoden om sociale mediaberichten te analyseren, zoals text mining met woordwolken en het programma DISCO, vragen nog om verdere ontwikkeling. Veel van deze technieken staan nog in de kinderschoenen. Ook goede en begrijpbare visualisaties worden steeds belangrijker.

Bij data van mobiele telefoongegevens of creditcard data is per definitie sprake van discussies over de privacy en of dit soort big data voor onderzoek en statistiek kunnen worden ingezet en gedeeld. Dat geldt ook voor de data die grote partijen, zoals Google, Microsoft en Amazon en sociale media zoals Facebook en Twitter, verzamelen. Voor het gebruik van 'openbare' sociale mediaberichten an sich lijkt de discussie over het gebruik voor onderzoek en statistiek minder te spelen.

4.2 Belangrijkste conclusies

Meer in detail kunnen de volgende conclusies uit de verkenningen met sociale mediaberichten van bedrijven worden getrokken:

1. Er is een relatie tussen de resultaten van de Coosto-data met sociale mediaberichten enerzijds en de resultaten van de ICT-enquête bedrijven anderzijds als het gaat om de vraag naar aantallen bedrijven dat gebruikmaakt van blogs en microblogs, waaronder Twitter. De relatie is echter niet erg sterk. Alleen bij de bedrijfstakken G (Handel) en M (Specialistische zakelijke diensten) kwamen de percentages overeen. Bij alle andere bedrijfstakken was er sprake van een onderschatting door de data van Coosto. Dit kan liggen aan de wijze waarop de sociale mediaberichten zijn verzameld: alleen die sociale mediaberichten met een URL in de tekst. Wel is het zo dat de Coosto-data resultaten opleverden over alle bedrijfstakken, terwijl de ICT-enquête een aantal bedrijfstakken buiten beschouwing laat.
2. De URL's die te vinden zijn in sociale mediaberichten van bedrijven kunnen behulpzaam zijn bij het vinden van domeinnamen, die bij een bedrijfseenheid horen. Het aantal domeinnamen dat via sociale mediaberichten wordt gevonden, is echter relatief klein. Hierdoor is de bijdrage aan een register met domeinnamen beperkt. Hooguit zouden de resultaten gebruikt kunnen worden om de resultaten van andere methoden te valideren. Wel kan snel inzicht worden verkregen naar welke domeinnamen (websites, apps of documenten) het meest wordt verwezen in sociale mediaberichten, die door bedrijven worden gepost.
3. Bedrijven maakten vooral gebruik van Twitter, eventueel in combinatie met Facebook en LinkedIn.
4. Sociale mediaberichten en sociale media profielen zijn bruikbare statistische eenheden aan de waarneemkant. Aan de outputkant spelen ze geen rol. Daar gaat het meer om statistische eenheden, zoals grootteklasse en economische activiteit.
5. (Big) data met sociale mediaberichten en – profielen bieden veel meer mogelijkheden om diepgaande analyses uit te voeren over het internetgedrag van bedrijven, dan enquêtedata. Voorbeelden zijn analyses in de tijd en naar omvang van het aantal geposte berichten naar type, grootteklasse en economische activiteit. Wat dit laatste betreft kwam bijvoorbeeld het beeld naar boven dat bedrijven aan het begin van de waardeketen minder gebruikmaken van sociale mediaberichten dan bedrijven die aan het einde van de waardeketen actief zijn.

6. Met behulp van methoden als 'woordwolken' ('word clouds') is het mogelijk om een eerste beeld te krijgen van de inhoud van sociale mediaberichten. Daarbij is het mogelijk de resultaten te differentiëren naar bijvoorbeeld economische activiteit of een bepaald thema. Ook is het mogelijk om de gerelateerde onderwerpen te vinden of bijvoorbeeld de stakeholders of de gevolgen van een thema of gebeurtenis. Duidelijk is wel dat dit soort technieken nog in de kinderschoenen staat. Zo is het bijvoorbeeld niet eenvoudig om ook de context erbij te betrekken. In dit onderzoek is gekeken naar bijvoorbeeld economische activiteit en grootteklasse, maar ook naar een aantal recente gebeurtenissen, zoals de MH17.
7. Bedrijven gebruiken Twitter vooral voor het communiceren over het onderwerp vacatures;
8. Er is onderzocht of er indicatoren konden worden samengesteld, die inzicht geven in het bereik of beter gezegd het belang van (groepen van) sociale mediaberichten. Er is geëxperimenteerd met de som van het aantal views (of volgers) van sociale mediaberichten, het gemiddelde aantal views en de ratio tweets/retweets. Al deze indicatoren kunnen snel worden samengesteld, ook gedifferentieerd naar economische activiteit of grootteklasse. Zo bleek het belang van sociale mediaberichten voor de bedrijfstak Horeca groter in de avond en in het weekeinde dan overdag door de week.
9. Door middel van de methode 'process mining' kan inzicht worden verkregen in de het gebruik van sociale media door bedrijven door de tijd heen. Daarvoor is het programma DISCO gebruikt.

4.3 Discussie

Op basis van deze verkenning is het duidelijk geworden dat analyses op basis van sociale mediaberichten en -profielen op den duur kan leiden tot het vervangen van vragen in bestaande traditionele enquêtes en tot nieuwe indicatoren. Duidelijk is ook geworden dat dieper gravende analyses mogelijk zijn dan die bij enquêtedata. Duidelijk is echter ook geworden dat dit soort analyses zich nog in de beginfase van ontwikkeling bevinden. Een belangrijke voorwaarde is ook dat sociale mediaberichten van bedrijven kunnen worden onderscheiden van die van personen. Daarvoor is hier nu de data gebruikt van het bedrijf Dataprovider.

Het antwoord op de centrale probleemstelling van dit onderzoek – kunnen sociale mediaberichten worden ingezet voor het maken van (officiële) statistiek voor bedrijven? – kan hier niet met een volmondig nee of ja worden beantwoord. Het antwoord ligt ergens in het midden: de potentie is er, maar ligt niet voor het oprapen.

Dit roep de vraag op in hoeverre dan ook moet worden verdergegaan met dit soort onderzoek door het CBS of dat gewacht moet worden totdat methoden zich wat beter hebben bewezen. Verder onderzoek is mogelijk, maar dat kost geld. Niet alleen voor de arbeidsuren, maar ook voor het verkrijgen van voldoende data. Daarnaast is het van belang om met (potentiële) gebruikers om de tafel te gaan zitten en te bekijken aan welke specifieke indicatoren of analyses behoefte is, naast het eventueel vervangen van vragen in

de ICT-enquête bedrijven. Verder is het noodzakelijk dat over een langere periode dan 6 weken data wordt verzameld en dat alle data wordt verzameld en niet alleen de sociale mediaberichten met een URL in de tekst.

Daarbij kan onder meer aan de volgende Indicatoren worden gedacht:

- Welke typen sociale media worden door bedrijven gebruikt en hoe ontwikkeld dat zich door de tijd heen (intensiteit)? Het lijkt erop dat het gebruik van andere sociale mediatypen door bedrijven hogere correlaties laten zien dan de hier onderzochte blogs en microblogs (bijvoorbeeld Twitter).
- Welke onderwerpen spelen in sociale mediaberichten, die door bedrijven worden gepost, vooral een rol, hoe ontwikkelt zich dat in de tijd en welke actoren spelen daarbij een rol? Is het daarnaast ook mogelijk om iets meer te zeggen over de context?
- Wat is het bereik van geposte sociale mediaberichten door bedrijven?
- Is het mogelijk om op basis van sociale mediaberichten inzicht te krijgen in vacatures naar bedrijfstak?
- Welke sociale mediaprofielen hebben de verschillende bedrijfstakken en grootteklassen?
- Welke bedrijfstakken twitteren vooral en welke retwitteren over welke onderwerpen?
- Op welke wijze, positief of negatief, worden sociale mediaberichten door bedrijven in de verschillende bedrijfstakken gepost en kan dit een beeld geven van de economische ontwikkeling op dat moment (een soort producentenvertrouwen)?
- Indicatoren voor specifieke bedrijfstakken, op een lager niveau dan 1 digit SBI.

5. Referenties

[1] COOSTO (2017), <https://www.COOSTO.com/nl/kenniscentrum/industry-reports/>, 16 oktober 2017.

[2] Finchline HowardsHome (2017), <https://www.howardshome.com/klantcases/>, 16 oktober 2017.

[3] Wassink, E. (2017), *Een register van domeinnamen: projectaanpak*, 30 juni 2017.

[4] Heerschap, N. en S. Ortega. (2017), *Sociale media en toerisme onderzoek*, 1 april 2017.

[5] Technology review, <https://www.technologyreview.com/s/519851/the-big-data-conundrum-how-to-define-it/>, 1 februari 2018.

[6] DISCO (2017), 'Process Mining DISCO', intranet: [CBS-Intern rapport](#)

[7] Heerschap, N. en S. Ortega, *Sociale media en bedrijven: een kader en cijfers, Onderzoeksrapport nr 4*, Impact ICT, april 2013.

Bijlagen

Bijlage I: Overzicht kenmerken Coosto-datasets

```
zoekopdracht
link: .nl OR .com          :1207053
link: .nl                  : 479119
link: .nl -extra:hasmedia   : 75956
link: .nl OR .com extra:ispost: 59556
(Other)                    :      0

datum
2017-07-03 09:00:         805
2017-07-07 11:13:         710
2017-07-07 11:20:         704
2017-07-07 11:40:         685
(Other)                   :1818780

url
https://instagram.com/p/BXA7xA3BKPw/      :      60
http://www.ellegirltalk.nl/topic/het-andere-bloggers-youtubers-
topic.700189/page-1327                     :      18
http://www.wijzijnvoetbal.nl/forum/3e-helft/54447-wzv-sc-2224-trots-op-
polen-altijd-5.html                       :      18
http://www.wijzijnvoetbal.nl/forum/3e-helft/54451-wzv-sc-2227-a-1.html:18
Other)                                     :1821570

sentiment
:1546662
- : 104148
+ : 170874
me:      0

type
comment: 502222
post    :1319462

discussielengte
Length:1821684
Class :character
Mode  :character

views
Min.    :      0
1st Qu.:    172
Median  :    588
Mean    :   6203
3rd Qu.:   1695
Max.    :44568746
NA's    :   54366

auteur
OpenBrug      : 19596
P2000_        : 13840
watkopenwij   : 13413
DutchNewspapers: 12668
(Other)       :1762167
```

```

volgers
Min.      :      0
1st Qu.:    139
Median :    526
Mean      :   5341
3rd Qu.:   1575
Max.      :15525048

invloed
Length:1821684
Class :character
Mode  :character

GPS.breedtegraad
Min.      :-43.5
1st Qu.:  52.0
Median :  52.2
Mean      :  51.0
3rd Qu.:  52.5
Max.      :  61.2
NA's      :1804050

GPS.lengtegraad
Length:1821684
Class :character
Mode  :character

type.bron
twitter :1765046
nieuws  : 20477
facebook: 18360
blog     :  6636
Other)   : 11165

titel
                                                :1793627
Overzicht                                  :    111
Gevoel van de Vierdaagse goed voor 870.000 kijkers:    16
TT: Veel clubs azen op Feyenoorder Woudenberg      :    16
(Other)                                           :  27914

bericht.tekst
RT @3FM: De dappere Tijn is vannacht overleden. Wij wensen de familie en
vrienden heel veel sterkte toe.Lees meer hier: https://t.co/nunAySIm7z
https://t.co/Z3WqAwJPb4 :    1042
RT @Feyenoord: Een bericht waar we ook bij Feyenoord stil van worden...
We leven mee met Abdelhak Nouri, zijn familie en @AFCAjax.
#StayStrongAppie https://t.co/J2fdzAAcCw :    954
:    921
RT @AFCAjax: #StayStrongAppie https://t.co/eNT2Y9050v :    641
(Other)                                           :1818126

datum1 (op basis van variabel datum)
Min.      :2017-06-29 09:14:00
1st Qu.:2017-07-07 17:59:00
Median :2017-07-17 11:57:00
Mean      :2017-07-16 12:01:54
3rd Qu.:2017-07-24 10:20:00
Max.      :2017-08-04 18:01:00

```

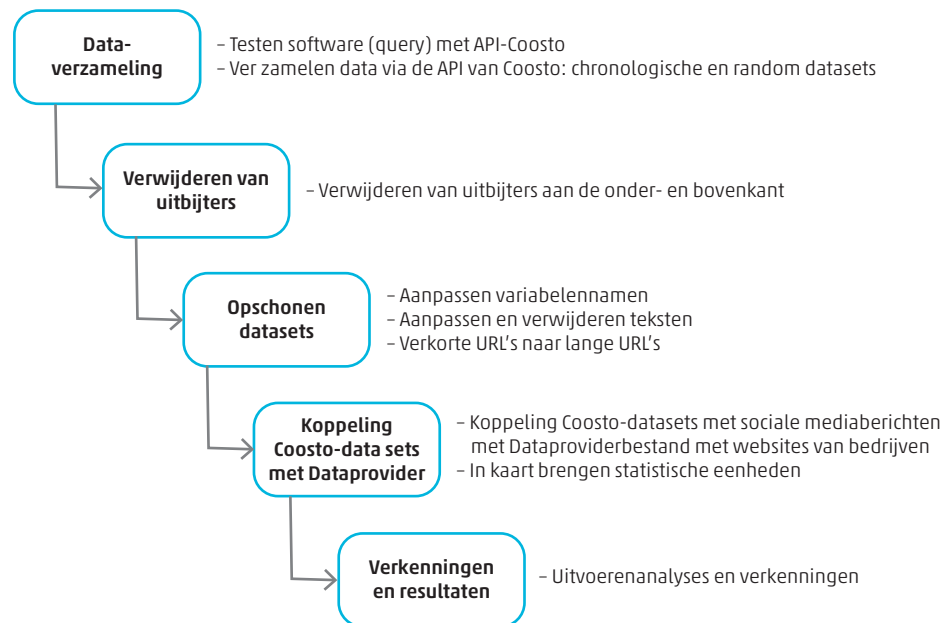
Bijlage II: Overzicht kenmerken analysebestand

Variabelen namen

- [1] "SBI2d"
- [2] "Facebook"
- [3] "BEid"
- [4] "be_handelsnaam"
- [5] "be_juridische_naam"
- [6] "BE_SBICoor"
- [7] "BE_coor"
- [8] "Aant_KvK_BE"
- [9] "aant_web_KvK_BE"
- [10] "OGid"
- [11] "OG_SBIcoor"
- [12] "OG_GKcoor"
- [13] "Topx"
- [14] "Congo"
- [15] "SFGO"
- [16] "Aant_Be_OG"
- [17] "KvKnummer"
- [18] "CbsPersoonid"
- [19] "CBSpersoon_handnaam_handelsnaam"
- [20] "KvK_Meer_BE"
- [21] "KvK_Google"
- [22] "SBI_actueel_KvK"
- [23] "werkz_pers_KvK"
- [24] "Kwal_A_kvnummer_kwaliteit"
- [25] "Kwal_B_kvnummer_kwaliteit"
- [26] "Kwal_C_kvnummer_kwaliteit"
- [27] "Kwal_D_kv nummer_kwaliteit"
- [28] "Kwal_E_kvnummer_kwaliteit"
- [29] "aant_web_kvK"
- [30] "Hostname"
- [31] "Chamberofcommercenummer"
- [32] "Socialmediatypes"
- [33] "Socialmediaprofiles"
- [34] "TwitterNaam"
- [35] "V1"
- [36] "V2"
- [37] "V3"
- [38] "V4"
- [39] "V5"
- [40] "twitter"
- [41] "V22"
- [42] "V33"
- [43] "V44"
- [44] "V55"
- [45] "linkedin"
- [46] "pinterest"

[47] "plus.google.com"
[48] "zoekopdracht"
[49] "datum"
[50] "url"
[51] "sentiment"
[52] "type"
[53] "discussielengte"
[54] "views"
[55] "volgers"
[56] "invloed"
[57] "GPS.breedtegraad"
[58] "GPS.lengtegraad"
[59] "type.bron"
[60] "titel"
[61] "bericht.tekst"
[62] "jaar"
[63] "maand"
[64] "dag"
[65] "tijd"
[66] "waarneeming"
[67] "datum0"
[68] "weekday"
[69] "uur"
[70] "N_berichten"
[71] "datum1"
[72] "indi"
[73] "BE_coor4"
[74] "Beschrijving"
[75] "SBI1dig"
[76] "Beschrijving1dig"
[77] "Beschrijving_1dig"
[78] "weekdays"
[79] "months"
[80] "SMType"
[81] "f_twit"
[82] "f_face"
[83] "f_link"
[84] "f_goo"
[85] "f_pint"
[86] "unit"

Bijlage III: Procesbeschrijving



Verklaring van tekens

Niets (blanco)	Een cijfer kan op logische gronden niet voorkomen
.	Het cijfer is onbekend, onvoldoende betrouwbaar of geheim
*	Voorlopige cijfers
**	Nader voorlopige cijfers
2018–2019	2018 tot en met 2019
2018/2019	Het gemiddelde over de jaren 2018 tot en met 2019
2018/'19	Oogstjaar, boekjaar, schooljaar enz., beginnend in 2018 en eindigend in 2019
2016/'17–2018/'19	Oogstjaar, boekjaar, enz., 2016/'17 tot en met 2018/'19

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever

Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress

Centraal Bureau voor de Statistiek

Ontwerp

Edenspiekermann

Inlichtingen

Tel. 088 570 70 70
Via contactformulier: www.cbs.nl/infoservice

© Centraal Bureau voor de Statistiek, Den Haag/Heerlen/Bonaire, 2019.
Verveelvoudigen is toegestaan, mits het CBS als bron wordt vermeld.