van Piet J.H. Daas

onderwerp Ontwikkeling van een Corona sentimentsindicator: Methodologische verantwoording

datum 11 mei 2020

1. Introductie

Social media is een interessante en snel beschikbare bron van informatie. De recente uitbraak van het Corona virus heeft de ontwikkeling van indicatoren op dit terrein, denk aan de verspreiding en de besmettingsgraad van het virus, erg gestimuleerd (Wired 2020). Het Centraal Bureau voor de Statistiek heeft in het recente verleden reeds een aantal social media gebaseerde indicatoren ontwikkeld, zoals de algemene sentimenten indicator (Daas en Puts 2014; Van den Brakel et al. 2017) en de sociale spanningen indicator (CBS 2020a). De eerste meet veranderingen in het sentiment op social media en vertoont een samenhang met het consumentenvertrouwen (Daas en Puts 2014), terwijl de tweede de spanningen en onderwerpen in de samenleving reflecteert (CBS 2020a).

Doordat social media berichten snel beschikbaar zijn, geven dergelijke indicatoren zeer actuele inzichten in het fenomeen waarvoor ze ontwikkeld zijn. Dit leidde tot het idee om een indicator te ontwikkelen om de veranderingen in het sentiment ten aanzien van het Corona virus en de door dit virus veroorzaakte ziekte in Nederland te bepalen. Dit zouden we ook, elke dag, bij een willekeurige geselecteerde groep Nederlanders d.m.v. een gestructureerd interview kunnen doen. Nadeel is dat dit een erg belastende en erg dure methode is. Daarom is voor een alternatieve aanpak gekozen, één die gebruik maakt van reeds bestaande gegevens, namelijk gebruik te maken van social media. Deze keuze heeft een aantal belangrijke consequenties en leidt tot de volgende uitgangspunten:

1) Er worden uitsluitend openbare social media berichten bestudeerd die op de meest populaire social media platforms verschijnen. Dit betreft vooral Twitter en Facebook.
2) Zowel originele berichten als doorgestuurde berichten (zoals retweets) en de reacties daarop worden meegenomen.
3) De onderzoekpopulatie zijn de berichten en reacties die door Nederlandse accounts op de onderzochte social media platforms worden geplaatst.
4) Er worden alleen berichten waarin Corona en Corona gerelateerde woorden voorkomen in de analysis meegenomen.
5) Het sentiment van een bericht, d.w.z. de verhouding tussen de positieve en negatieve woorden in een bericht, wordt als indicatie gebruikt voor de attitude en het gevoel t.a.v. Corona door de schrijver/verzender van het bericht.
6) Het sentiment van de geselecteerde berichten is een afspiegeling van het sentiment van de Nederlandse bevolking op dat moment.

De resultaten van eerdere studies hebben laten zien dat deze werkwijze, ondanks de aannames, tot interessante inzichten leidt. De uitgangspunten geven duidelijk aan dat er een verschil is tussen het concept dat we graag zouden willen meten (d.w.z. de attitude t.a.v. het Coronavirus en de ziekte die het veroorzaakt onder de Nederlandse bevolking) en de uiteindelijke operationalisatie daarvan; de te ontwikkelen indicator. De consequenties hiervan worden in dit document uitgebreid besproken.

2. Materiaal en methode

Voor de toegang tot social media wordt gebruikt gemaakt van de database van Coosto (2020). Dit bedrijf verzamelt routinematig zoveel mogelijk openbare berichten die op verschillende social media platformen geplaatst worden. Coosto streeft hierbij nadrukkelijk naar volledigheid.

Voor het selecteren van berichten wordt gebruik gemaakt van de webinterface van Coosto waarbij gebruik gemaakt is van een combinatie van selectie- en uitsluitwoorden. De woorden die gebruikt worden om berichten over Corona te selecteren zijn: 'Corona', 'Coronavirus', 'COVID-19', 'COVID19', ''COROVID19', SARS-COV-2' en 'pandemie'. Berichten die één of meerdere van deze woorden of hashtags (#) daarvan bevatten worden allemaal geselecteerd. Uitsluitwoorden zijn woorden die gebruikt worden om te voorkomen dat bepaalde berichten in de query worden meegenomen. Dit zijn 'bier', 'beer', 'indicator' en 'index'. De eerste twee zorgen ervoor dat berichten over het biermerk Corona niet worden meegenomen, terwijl de laatste twee ervoor zorgen dat berichten over (andere) Corona indicatoren en indexen worden uitgesloten. Beiden zorgen ervoor dat er zoveel mogelijk berichten die over het Coronavirus gaan worden geselecteerd.

Door de keuze voor de gehanteerde selectie- en uitsluitwoorden worden bepaalde berichten nadrukkelijk uitgesloten. Bijvoorbeeld een (fictief) bericht van een zeer waarschijnlijk eenzaam persoon die het volgende meldt: "Vind het erg jammer dat mijn familie me niet meer mag komen bezoeken. Wat een nare periode", wordt niet geselecteerd. Alles in dit bericht wijst erop dat het een gevolg van de Coronacrisis beschrijft. Dit gevolg is echter niet het concept dat we willen meten. Daarnaast is het zeer lastig om een nog uitgebreidere lijst met woorden op te stellen in een poging om zoveel mogelijk wel relevante berichten te selecteren zonder dat er veel niet/minder relevante berichten worden meegenomen.

Het is tevens belangrijk ervoor te zorgen dat uitsluitend berichten van Nederlandse gebruikers worden verzameld. Eerder ervaringen hebben geleerd dat dit het beste in Coosto kan worden gedaan door berichten met een locatiekenmerk van een aantal Europese landen, met name België, Duitsland, Frankrijk en het Verenigd Koninkrijk, uit te sluiten (CBS 2019). De periode waarover berichten kunnen worden geselecteerd loopt van januari 2009 t/m heden. Voor deze Corona studie is de periode half december 2019 t/m heden uiteraard het meest interessant.

Het is mogelijk dat onbetrouwbare berichten die op social media worden geplaatst, zogenaamd nepnieuws, de resultaten negatief kunnen beïnvloeden. Een recente studie die in maart 2020 is uitgevoerd (Pointer 2020) heeft laten zien dat dit in de bestudeerde periode maximaal 160 Twitter berichten per dag betreft. Gezien de grote hoeveelheden berichten die in deze studie worden meegenomen (zie pagina 4) lijkt dat -op dit moment- geen groot probleem te zijn. Het is echter zeker iets om in de gaten te houden.

Het resultaat van de query is een lijst met per dag de aantallen geselecteerde berichten en hun sentiment. De sentimenten die door Coosto onderscheiden worden zijn neutraal, positief en negatief. Berichten waarin meer positieve woorden dan negatieve woorden voorkomen zijn positief, berichten met meer negatieve woorden dan positieve woorden zijn negatief en berichten met geen positieve en negatieve woorden of berichten met evenveel positieve als negatieve woorden zijn neutraal. De woordenlijsten die Coosto hiervoor gebruikt zijn gebaseerd op een standaard sentiment classificatie van Nederlandstalige woorden, zoals WordNet (2020), aangevuld met een lijst met specifieke social media termen en afkortingen; zie Daas en Puts (2014) voor meer detail.

Na export van de gegevens uit de Coosto omgeving, als CSV-bestand, worden de analyses verder in R uitgevoerd. Het dagelijkse corona sentiment wordt berekend zoals weergegeven in (1):

$$Corona\ sentiment = \begin{cases} \#\ totaal\ aantal\ berichten = 0: & 0 \\ \#\ totaal\ aantal\ berichten > 0: & \frac{(\#\ positieve\ berichten - \#\ negatieve\ berichten)}{\#\ totaal\ aantal\ berichten} \end{cases} \quad (1)$$

Het sentiment zal hierdoor altijd tussen de 1 en -1 liggen. Wanneer er op een dag geen berichten over Corona worden geplaatst zal de indicator 0 zijn. Omdat het dagelijkse sentiment erg volatiel bleek, is er vervolgens een voortschrijdend gemiddelde filter toegepast om dit effect te verminderen en daardoor een duidelijker signaal te krijgen. De optimale breedte van het filter, in dagen, is bepaald door het minimum te bepalen van het Akaike Informatie Criterium voor filters van verschillende breedte. Hiervoor is de sma-functie in het 'smooth' package gebruikt (Svetunkov 2020). Deze functie liet zien dat een tweezijdig filter over een periode van 10 dagen het minste informatieverlies gaf. Nadeel van zo'n filter is dat er, na toepassen ervan, aan het begin en het eind van de reeks gegevens ontbreken. Aan het begin van de reeks is dit opgelost door de selectiequery eerder, op 15 december 2019, te laten beginnen. Hierdoor kan het gefilterde sentiment zonder problemen vanaf 1 januari 2020 worden getoond. Aan het eind van de reeks is dat niet mogelijk. Om te zorgen dat er toch een indruk wordt verkregen van de meest recente situatie worden de wel beschikbare ruw ongefilterde gegevens aan het eind van de gefilterde reeks toegevoegd. Deze aanpak is voor verbetering vatbaar en onderwerp van huidig onderzoek. Omdat de gegevens van de dag waarop de query is uitgevoerd *niet* volledig zijn, die dag is dan immers nog niet voorbij, wordt dat meetpunt nooit in de reeks weergegeven.

3. Resultaten

Tijdens het schrijven van dit verslag is een zoekquery is uitgevoerd voor de periode 1 januari 2020 t/m 29 maart 2020. Hierbij werden 4.707.014 berichten geselecteerd. Het merendeel was afkomstig van Twitter, in totaal 3.238.950. De rest van de berichten zijn op Facebook geplaatst. De verdeling van alle berichten is per dag weergegeven in Figuur 1. De figuur laat zien dat de aantallen Corona gerelateerde berichten vanaf 23 februari langzaam toeneemt en in de 2e week van maart zeer sterk toeneemt. Daarna is er een langzame daling te zien.

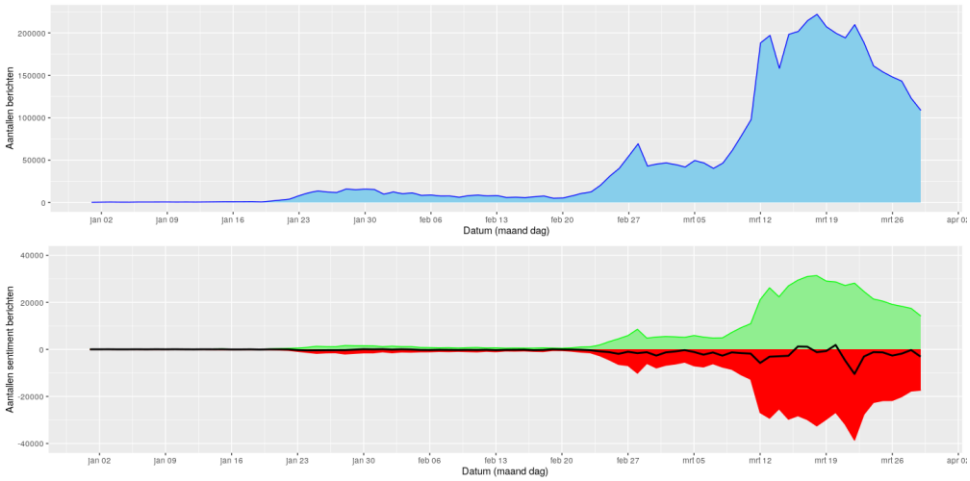
Ter illustratie zijn hieronder enkele voorbeelden van geselecteerde berichten voor de verschillende sentimenten weergegeven. Er is geprobeerd de berichten zoveel mogelijk te anonimiseren.

Neutraal:

1) coronavirus is nr.10 trending hashtag in NL in afgelopen 4 uur. #coronavirus
2) RT Deel svp deze oproep van Brabantse longarts Braunstahl Lever je mondkapjes in Ziekenhuizen hebben ze *HEEL HARD NODIG* #COVID19 #mondkapjes

Positief:

1) Diederik Gommers. Een baken van rust in de onvoorspelbare Coronastorm. Respect. #Coronavirusnl
2) Huisarts uit Loon op Zand heeft goede hoop na hectische week: 'Corona wat kunnen indammen'



Figuur 1. *Dagelijkse verdeling van de aantallen berichten (in blauw) en de berichten met positief (groen) en negatief (rood) sentiment van 1 januari t/m 29 maart 2020. De zwarte lijn geeft de verhouding tussen de berichten met positieve en negatieve sentiment weer.*

Negatief:

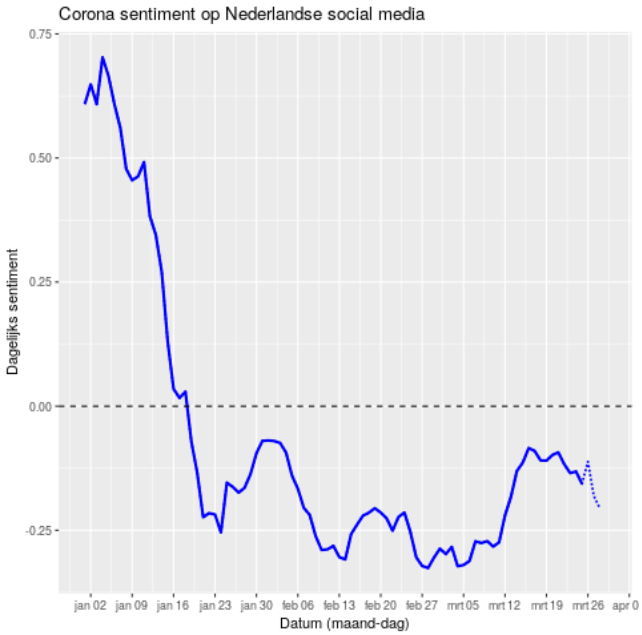
1) RT Zou er totaal geen moeite mee hebben als dit aso-gedrag gelijk wordt gesteld met spugen en roepen "ik heb #corona". Aantal weken de cel in.
2) Zojuist een @Schiphol beveiliger gesproken met 39 graden koorts, maar te bang om zich ziek te melden. ‘Als ik me ziek meld dan zitten daar consequenties aan vast, daarnaast moet ik ook mijn rekeningen betalen’! Het gevolg van doorgeslagen flexibilisering #COVID19

Een nadere controle van steekproeven van de geselecteerd berichten bevestigen dat een zeer groot deel van de berichten over het Coronavirus en de veroorzaakte ziekte gaan. Dit is echter niet 100% het geval omdat soms ook berichten worden geschreven die bijvoorbeeld de hashtag #Corona bevatten en meer over de consequenties van de Coronacrises gaan. Een eerst schatting geeft aan dat dit zo’n 15% van de berichten bedraagt. Tevens is voor elk selectiewoord gekeken wat het effect was van het wel of niet meenemen in de query. Dit liet zien dat het belangrijkste woord Corona is gevolgd door COVID19 (en varianten daarvan). De rest van de woorden zijn veel minder belangrijk. De uitsluitwoorden hadden nauwelijks effect.

Daarnaast zijn een aantal willekeurig selecties van berichten bestudeerd om een indruk te krijgen hoeveel berichten er per auteur per dag worden geplaats. De aantallen berichten vertoonde een maximum van 9 per dag met een mediaan van 1. Heel veel auteurs blijken dus 1 bericht per dag te schrijven. Echter, het bleek ook dat er per dag behoorlijke aantallen berichten zonder auteur verschijnen; soms wel meer dan 1000. Dit blijken allemaal Facebook berichten te zijn. De exacte reden hiervoor is onbekend en wordt bij Coosto nagevraagd; het zal ongetwijfeld met de wijze van verzamelen te maken hebben.

De aantallen berichten dat per dag wordt geselecteerd kan sterk verschillen. Het minimum aantal berichten per dag gedurende de bestudeerde periode is 264 en het maximum 225.476. Nul

berichten zien we niet in deze periode. Het eerste getal laat zien dat er altijd wel een aantal berichten over Corona worden opgesteld. De grote variatie in het aantal berichten leidt ertoe dat het absolute aantal positieve en negatieve berichten per dag niet echt direct te gebruiken is voor het opstellen een dagelijkse sentimenten indicator. Om het effect van die sterk verschillende aantallen per dag te verdisconteren is dan ook besloten het dagelijkse sentiment t.a.v. Corona uit te rekenen door het verschil te nemen tussen de aantallen positieve berichten min de negatieve berichten gedeeld door het totaal aantal berichten van die dag; zie vergelijking (1). Een dergelijke aanpak wordt ook bij het bepalen van het consumentenvertrouwen gebruikt (CBS 2020b) en werkte ook goed bij het bepalen van het dagelijks sentiment in social media (Daas en Puts 2014). Desondanks, bleek het dagelijkse sentiment toch behoorlijk volatiel. Dit is echter niet onverwacht want ook het dagelijkse algemene social media sentiment vertoonde dit probleem (zie figuur 1 op pagina 9 in Daas en Puts 2014). Bij dit laatste onderzoek bleek dat het aggregeren van het sentiment over langere perioden dan één dag een vermindering van de volatiliteit tot gevolg had. Eenzelfde effect wordt verkregen door een voortschrijdend gemiddelde filter toe te passen op het ruwe dagelijkse signaal. Het gefilterde signaal is in Figuur 2 weergegeven als een dikke blauwe lijn.



Figuur 2. *Dagelijks Corona sentiment in Nederlandse social mediaberichten na toepassing van een tweezijdig voortschrijdend gemiddelde filter van 10-dagen. De dikke blauwe lijn geeft de gefilterde reeks weer, terwijl de gestippelde lijn aan het eind het ongefilterde sentiment van de laatste 3 dagen weergeeft.*

Figuur 2 laat zien dat het Corona sentiment in publieke social media berichten aan het begin van het jaar positief begint en negatief wordt op 20 januari. Deze datum komt overeen met de periode waarin bekend werd dat het coronavirus zich verspreidt had vanuit China naar een aantal omliggende landen; o.a. naar Zuid- Korea en Japan. Na die datum blijft het Corona sentiment op Nederlandse sociale media negatief. De dag waarop de eerste Corona patiënt officieel in Nederland

werd geregistreerd was op 27 februari. Omdat dit nieuws rond 23 uur in de avond algemeen bekend werd is te verwachten dat dit nieuws ook nog een effect heeft op het sentiment van de daarop volgende dag. Dit klopt ook, we zien de laagste waarden voor het Corona sentiment op 27 en 28 februari.

Nadat de indicator negatief wordt, blijft het sentiment zich tussen de -0,31 en -0,05 bewegen. Deze beweging is voornamelijk het gevolg van verschuivingen in de verhouding tussen de aantallen positieve en negatieve berichten t.g.v. reacties op berichten en/of media optreden van politici, publieke en overheidsinstellingen, bekende Nederlanders en de koning. Wanneer die berichten vooral positief worden ontvangen is er een lichte toename van het sentiment te zien. Bijvoorbeeld, op 12 maart deelde Minister President Rutte op TV mee dat er drastische maatregelen worden genomen om de verspreiding van het Coronavirus in Nederland te verminderen; dit betrof o.a. het sluiten van alle scholen. De sentimenten indicator nam hier licht na toe, wat erop lijkt te wijzen dat zijn besluit -in het algemeen- goed werd ontvangen.

In de bovenstaande tekst zijn een aantal van de keuzes die gemaakt zijn geverifieerd. Denk aan het gebruik van bepaalde selectiewoorden, de bijvangst van minder relevante berichten en de relatie tussen auteur en de aantallen verstuurde berichten. Een andere belangrijke controle is de keuze om wel of geen reacties op (originele) berichten mee te nemen. Bij deze controle bleek dat het uitsluiten van reacties op originele berichten tot gevolg had dat het berekende Corona sentiment vanaf 15 maart positief werd en dat gedurende de rest van de periode ook bleef. Dit is op zich een opmerkelijk resultaat dat niet goed lijkt te passen bij het te meten concept. De intelligente lockdown was toen immers net begonnen en bij veel social mediagebruikers begonnen de consequenties hiervan door te dringen. Vooral het feit dat het sentiment positief wordt en blijft is niet iets wat je van een Corona sentiment indicator zou verwachtten tijdens een periode dat Nederlanders thuis moeten zitten om een mogelijke besmetting met het virus te voorkomen. Je verwacht dan eerder dat het sentiment negatief is en blijft t.a.v. Corona. Nederlanders worden immers beperkt in hun bewegingsvrijheid door het virus. Deze laatste overweging is een belangrijke reden om het sentiment van reacties op berichten wel mee te nemen in de uiteindelijke indicator.

4. Discussie

Deze memo beschrijft de ontwikkeling van een Corona sentiment indicator gebaseerd op publieke Nederlandse social mediaberichten. Doordat met social media is gewerkt zijn een aantal belangrijke keuzes gemaakt tijdens de operationalisatie van het te meten concept. Niet alle consequenties konden worden geverifieerd, de belangrijkste hiervan zijn de volgende.

i) Er is uitsluitend naar het sentiment gekeken van personen die openbare social media berichten sturen. Personen die dat niet doen ontbreken. Het is bekend dat de achtergrondkenmerken van personen die actief zijn op social media afwijken van die van de totale bevolking (Ortega en Heerschap 2019). Dit zal consequenties hebben voor de indicator. Voor wat betreft de signaalfunctie van de indicator is het echter nog maar de vraag of dat een groot probleem is. Personen die social media berichten plaatsten geven immers een signaal af en dat is net wat we willen meten.

ii) Er is niet naar het sentiment van personen maar naar het sentiment van berichten gekeken. Door de grote hoeveelheden berichten die per dag over Corona werden verstuurd bleek het niet mogelijk de berichten op persoonsniveau (d.w.z. social media username) te aggregeren. Dit is vooral het gevolg van een beperking van de Coosto interface. Een handmatige controle van willekeurig selecties van berichten liet zien dat per dag maximaal 9 berichten per useraccount zijn verstuurd. Het merendeel stuurt slechts 1 bericht. Ander probleem dat werd gevonden was het feit dat er Facebook berichten voorkomen zonder bekende auteur. Dit bleek per dag om behoorlijke aantallen te gaan; soms wel meer dan 1000 per dag. De reden hiervoor wordt nagevraagd.

iii) Voor wat betreft het sentiment van een bericht rest de vraag of dit inderdaad direct gerelateerd is aan het Coronavirus/de ziekte of dat dit het indirect gevolg is van de Coronacrisis. De paar voorbeeldberichten in dit document laten al zien dat dit eerste niet altijd het geval lijkt te zijn. Zo gaat één van de negatieve berichten duidelijk over een persoon. Een nadere studie van willekeurige selecties van berichten liet zien dat het rond de 15% van de berichten betreft. Het aanpassen van de selectiewoorden had hier geen duidelijk effect op. Mogelijk kunnen dergelijke berichten in een vervolgstudie wel beter worden onderscheiden, door bijv. hiervoor een tekst-gebaseerd classificatie model te ontwikkelen. Voor deze initiële studie ging dat echter iets te ver.

De eerste resultaten van de indicator laten zien dat het sentiment licht positief begint en daarna negatief wordt. De momenten waarop dit gebeurt, de verspreiding van Corona vanuit China en de detectie van de eerste officiële Corona-patiënt in Nederland, en de effecten op de indicator lijken er op te wijzen dat de indicator het 'fenomeen' meet waarvoor het is ontwikkeld. Metingen over een langere periode zijn nodig om dit met meer duidelijkheid te kunnen zeggen. Dit ook om te kijken wat het effect zal zijn van kleine hoeveelheden Corona berichten op de indicator.

Het is ook zeker interessant om te zien hoe de indicator zich in de nabije toekomst zal ontwikkelen. Bijvoorbeeld wat gebeurt er als er door de regering maatregelen worden genomen die een (korte) opleving van het aantal besmetting tot gevolg hebben. Er is dan te verwachten dat de indicator gaat dalen. Daarnaast zal, aan het eind van de Coronacrisis, d.w.z. wanneer het aantal besmettingen en doden gaat afnemen en de lockdown wordt opgeheven, de indicator zich vermoedelijk naar nul en daarboven gaan bewegen; uiteraard als alles goed verloopt. Het wordt interessant te zien wanneer dat precies gebeurd. Al met al laten de eerste resultaten zien dat het door deze indicator en de snelle beschikbaarheid van social media mogelijk is snel inzicht te krijgen in de stemming van Nederlandse social media gebruikers t.a.v. het Coronavirus en de door het virus veroorzaakte ziekte. In de bijlage is het meest recente resultaat van de Corona sentimentindicator te vinden.

Referenties:

CBS (2019) Dashboard sociale spanningen. Intern rapport voor CBS and WODC, Version 13 september.

CBS (2020a) Sociale spanningen-indicator: Peilstok samenleving. Beta-publicatie op de Innovatie pagina van het CBS, link at: https://www.cbs.nl/nl-nl/onze-diensten/innovatie/project/sociale-spanningen-indicator-peilstok-samenleving.

CBS (2020b) Twee indicatoren voor het Nederlandse consumentenvertrouwen. CBS-web pagina Link: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/aanvullende%20onderzoeksbeschrijvingen/twee-indicatoren-voor-het-nederlandse-consumentenvertrouwen

Coosto (2020) Social Media Management Software. Webpagina link: https://www.coosto.com/nl

Daas, P.J.H., Puts, M.J.H. (2014) Social Media Sentiment and Consumer Confidence. European Central Bank Statistics Paper Series No. 5, Frankfurt, Germany.

Ortega, S., Heerschap, N. (2019) Gebruik van sociale media: Verkennend onderzoek naar de mogelijkheden van het gebruik van sociale media voor statistiek voor bedrijven. CBS publicatie, april 2019. Located at: https://www.cbs.nl/-/media/_pdf/2019/16/gebruik-van-sociale-media.pdf

Pointer (2020) Flinke stijging van onbetrouwbaar nieuws over coronavirus op Twitter, Meer misinformatie gemaakt en gedeeld sinds uitbraak in Nederland. Maart 20. Webpagina link: https://pointer.kro-ncrv.nl/artikelen/flinke-stijging-van-onbetrouwbaar-nieuws-over-coronavirus-op-twitter

Svetunkov, I. (2020) Vignette package 'smooth': Forecasting Using State Space Models. Versie 2.5.5. Located at: https://cran.r-project.org/web/packages/smooth/smooth.pdf

Van den Brakel, J., Sohler, E., Daas, P., Buelens, B. (2017) Social media as a data source for official statistics; the Dutch Consumer Confidence Index. Survey Methodology 43 (2), pp. 183-210.

Wired (2020). How AI Is Tracking the Coronavirus Outbreak, Machine-learning programs are analyzing websites, news reports, and social media posts for signs of symptoms, such as fever or breathing problems. February 8th. Webpagina link: https://www.wired.com/story/how-ai-tracking-coronavirus-outbreak/

WordNet (2020) Open Dutch Wordnet: a Dutch lexical semantic database. Webpagina link: http://wordpress.let.vupr.nl/odwn/

Bijlage:

Meest recente update. De Corona sentiment plot van 18 mei 2020.