TEKSTANALYSEMETHODEN

Toepassingen in de officiële statistiek

Tekstanalyse, ook wel *text mining* genoemd, is het proces waarmee geautomatiseerd waardevolle informatie wordt afgeleid uit teksten. Een voorbeeld is het automatisch classificeren van e-mails naar spam en niet-spam. Tekstanalysemethoden zijn in de officiële statistiek tot nu toe vooral gebruikt om antwoorden van respondenten op open enquêtevragen in te delen naar een vooraf vastgestelde classificatie. Bijvoorbeeld het toekennen van een tekstomschrijving voor 'beroep' naar een categorie van de beroepenclassificatie. Momenteel worden binnen de officiële statistiek ook nieuwe toepassingen onderzocht. Aan welke toepassingen kun je dan denken? En hoe werkt dat? In het artikel gaan we op deze vragen in.

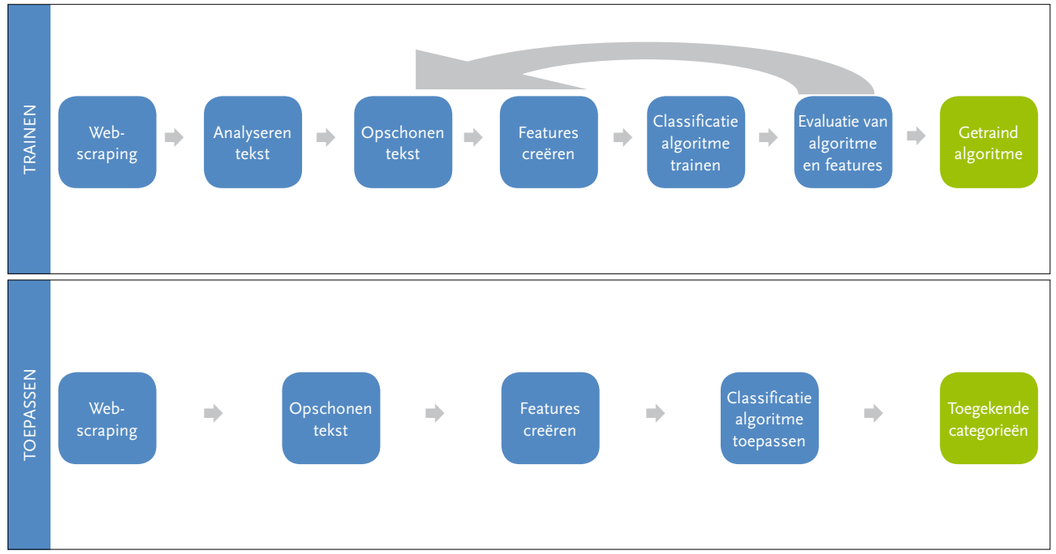
Arnout van Delden, Piet Daas, Olav ten Bosch & Dick Windmeijer

Tekstanalyse kan op meerdere manieren nuttig zijn voor de officiële statistiek. We kunnen het gebruiken om bedrijfsgegevens af te leiden uit teksten van bedrijfswebsites, zoals contactgegevens van bedrijven. Website-gebaseerde tekstanalyse kunnen we ook gebruiken voor het afbakenen van populaties, bijvoorbeeld voor het vinden van duurzame bedrijven. We kunnen het ook gebruiken om objecten automatisch, in plaats van handmatig in te delen naar standaardclassificaties. Zo gebruikt het CBS websites om prijs en andere informatie van kleding te verzamelen voor het berekenen van de prijsontwikkeling van kleding. Een ander soort toepassing is informatie extractie, bijvoorbeeld uit een bouwvergunningstekst automatisch het soort object, de locatie, het bedrag en de start van de bouwperiode afleiden. Ten slotte kunnen we het toepassen voor het afleiden van sentimenten in teksten. Deze techniek heeft het CBS gebruikt om een sociale spanningen-indicator te ontwikkelen, zie < https://www.cbs.nl/nl-nl/onze-diensten/innovatie/ > .

Classificatiemethoden

We beperken ons hier tot tekstanalysemethoden die zich richten op automatisch classificeren. Daarvan zijn er drie soorten. De eerste zijn regel-gebaseerde methoden. Deze werken met ALS–DAN regels, bijvoorbeeld ALS 'meester' DAN 'leerkracht basisonderwijs'. Deze methode gebruikt het CBS bijvoorbeeld om beroepen uit een open enquêtevraag te classificeren en om doodsoorzaken op basis van informatie van artsen te classificeren. Hoewel er met ALS–DAN regels een behoorlijke precisie bereikt kan worden is het nadeel dat er vaak veel regels nodig zijn, die bovendien nogal onderhoudsintensief zijn. De andere twee methoden werken met statistische modellen. Hierbij onderscheiden we *supervised* en *unsupervised learning*. Bij *supervised learning* worden modellen getraind op basis van een flink aantal voorbeelden die vooraf met de hand zijn geclassificeerd. Het getrainde model wordt vervolgens gebruikt om voor nieuwe teksten de categorie te kunnen voorspellen. Veel gebruikte methoden voor *supervised learning* zijn Naive Bayes, logistische regressiemodellen, Support Vector Machines, beslisbomen en neurale netwerken (Hastie et al., 2016).

Bij de methoden voor *unsupervised learning* wordt niet gewerkt met voorbeelden, maar worden de teksten in clusters ingedeeld. Hierbij wordt vaak *k-means* gebruikt maar ook zogenaamde 'topic modellen' (Blei, 2012) en



Figuur 1. *Supervised learning* bij tekstanalyse op basis van websites

word embeddings (Ruder, 2016) om (semantische) structuren in teksten te vinden.

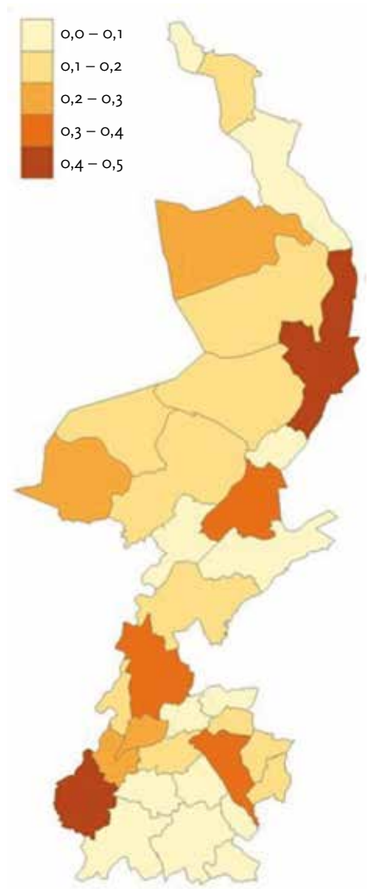
In het vervolg van de tekst gaan we nader in op *supervised learning* methoden (zie ook figuur 1).

Features

Voor veel nieuwe toepassingen is het belangrijk om de juiste, meest voorspellende, eigenschappen, *features* genoemd, uit de tekst te halen voordat een *machine learning* model kan worden ingezet. Het begint er mee, dat de teksten van het internet gehaald moeten worden, bijvoorbeeld via *web scraping*. Vervolgens worden de teksten geanalyseerd: welke woorden en tekens komen voor en hoe vaak? Een grafische weergave van de teksten kan daarbij heel nuttig zijn. Uit de analyse kan naar voren komen dat bepaalde woorden vaak voorkomen, maar dat deze niet bijdragen aan de informatie die je uit de tekst wilt halen. De volgende stap is dan dat een tekst wordt opgeschoond. Daarbij worden bijvoorbeeld spaties, leestekens en emoticons verwijderd en woorden (of delen van woorden) geïdentificeerd. Vervolgens wordt vaak een automatische taalherkenning toegepast, mede omdat de vervolgstappen taalgevoelig zijn. Daarna kan het nuttig zijn de woorden te standaardiseren naar de 'stam' van het woord en stopwoorden te verwijderen. Ook worden heel vaak voorkomende woorden soms ook nog verwijderd. De woorden die overblijven vormen de basis voor de *features* die in de modellen ingezet worden. Daarbij worden ook varianten gebruikt, bijvoorbeeld 'n-grammen' waarbij groepen van woorden die vaak in combinatie voorkomen als één *feature* worden gekozen. Ten slotte worden de *features* omgezet naar getallen. De twee bekendste methoden hiervoor zijn het bepalen van de relatieve frequentie van een woord in een document, rekening houdend met het voorkomen ervan in alle andere documenten (TF-IDF) en het omzetten naar vectoren, de zogenaamde *word embeddings*. De laatste methode houdt rekening met de context van een woord in de tekst (Ruder, 2016).

Modelselectie

In praktijk worden meerdere modellen uitgeprobeerd, worden per model meerdere waarden van de modelparameters getest en wordt een *feature*-selectie uitgevoerd. De beste manier om een model te ontwikkelen bij *supervised learning* is om de dataset in tweeën te verdelen. Eén deel wordt gebruikt voor eenvoudige kruisvalidatie waarbij modelparameters worden getuned.



Figuur 2. Relatieve aantal innovatieve bedrijven in Limburg per gemeente, gecorrigeerd voor de bevolkingsdichtheid

Het tweede deel van de data wordt gebruikt om een model waarvan alle parameters getuned zijn te testen. Bij het valideren en testen kan de kwaliteit van de voorspelde categorieën op verschillende manieren gemeten worden. De precisie van een voorspelde klasse meet welk aandeel daarvan correct is. De *recall* meet welk deel van een werkelijke categorie goed voorspeld is. De nauwkeurigheid meet de totale fractie goed voorspelde categorieën. Afhankelijk van het doel van het classificeren wordt bepaald welke maat het best kan worden gekozen. Als eenmaal een getraind model is verkregen, kan dat gebruikt worden om voor nieuwe teksten de categorie te voorspellen.

Innoverende bedrijven

Ten slotte geven we drie voorbeelden van recent onderzoek op het CBS waarbij tekstanalysemethoden zijn gebruikt. Het eerste voorbeeld betreft het vinden van innovatieve bedrijven. Het CBS houdt tweejaarlijks een innovatie enquête onder bedrijven met 10 of meer werknemers. De vraag was of innoverende bedrijven ook op een andere manier kunnen worden geïdentificeerd. Besloten werd hiervoor naar de tekst op de hoofdpagina van de website van bedrijven te kijken. De websites van 3000 innovatieve en 3000 niet-innovatieve bedrijven volgens de innovatie-enquête uit 2014 en die van bedrijven in de innovatietop 100 van het midden- en kleinbedrijf van 2009-2017 waren hierbij de training- en testset. Na opschoning van de tekst en trainen van het model, werden met een logistisch regressiemodel op een onafhankelijke testset de volgende kwaliteitsscores gehaald: precisie 100%, recall 87% en nauwkeurigheid 92% (van der Doef et al., 2018). Met behulp van dit model zijn veel meer bedrijvenwebsites geclassificeerd en zijn kaarten gemaakt waarbij een schatting is gemaakt van de dichtheid aan innovatieve bedrijven per provincie en gemeente, gecorrigeerd voor de bevolkingsdichtheid (zie figuur 2). Op basis van de enquêtegegevens kunnen we dit niet per gemeente schatten.

Verhuiswens

Een tweede voorbeeld is de vraag van een externe klant of verhuiswensen van personen afgeleid kunnen worden uit sociale media berichten. Daarbij zijn uit publieke sociale media berichten tussen 2014 en 2017 eerst alle berichten geselecteerd met de woorden verhuis* of verhuiz*. Als training- en testset is vervolgens uit deze berichten een random steekproef van 1000 berichten getrokken die door meerdere personen, onafhankelijk van elkaar, handmatig zijn getypeerd. Hierbij werd gekeken of het bericht

van een persoon afkomstig was die duidelijk aangaf te willen verhuizen. Uiteindelijk waren er bijna 120 berichten met een verhuiswens gevonden. Bij het opschonen bleek dat leestekens en emoticons verwijderd moeten worden, maar dat het standaardiseren van de woorden naar stam juist niet verstandig is. Vervolgens is een aantal modellen getest, waarvan er drie een vergelijkbare nauwkeurigheid bleken op te leveren: Logistische regressie, *Gradient Boosting* en een neuraal netwerk (allen 85%). *Gradient Boosting* had daarnaast een precisie van 90% en een recall van 89%. In het vervolgonderzoek wil het CBS nagaan in hoeverre deze berichten te gebruiken zijn als voorspeller van verhuizingen, bijvoorbeeld door dit te vergelijken met de verdeling van verhuizingen in de populatie.

Economische activiteit

Een laatste voorbeeld betreft het afleiden van de economische activiteit van een bedrijf op basis van bedrijfswebsites (Roelands et al., 2017). Dit gegeven wordt geregistreerd als een bedrijf zich bij de Kamer van Koophandel inschrijft, maar wijzigingen in activiteit worden zelden doorgegeven. Met website-informatie kan mogelijk een actuelere code verkregen worden. Het onderzoek is beperkt tot het voorspellen van de negen zogeheten topsectoren van de economie, die weer onderverdeeld zijn in 29 subsectoren. Uit elke subsector is een steekproef van 70 bedrijven met URL getrokken. Van elke URL is de hoofdpagina plus één onderliggende pagina *gescraped*, en met taalherkenning zijn pagina's met de Nederlandse taal geselecteerd. De teksten zijn geschoond, en een trefwoordenlijst voor economische activiteit is gebruikt als de basis voor de voorspellende *features*. Met het best passende model is uiteindelijk op topsectorniveau, op een onafhankelijke testset, een kwaliteit behaald van 80% precisie, 58% recall en 51% nauwkeurigheid. Op subsector niveau waren de kwaliteitsscores een stuk lager. Hier zijn nog talloze punten waarop de tekstanalyse verbeterd kan worden, maar met name het voorspellen van een grote range verschillende categorieën lijkt lastig te zijn.

Toekomst

In de toekomst wil het CBS steeds meer gebruik maken van informatie die in de vorm van teksten, zoals op websites, beschikbaar is. Er is een aantal onderwerpen waar we verder onderzoek naar willen doen. Ten eerste willen we graag populaties van personen en bedrijven flexibel en snel kunnen indelen naar allerlei kenmerken. Behalve het toekennen van de categorie op basis van de tekst, speelt daarbij ook de vraag hoe nauwkeurig we die populaties kunnen toekennen, en hoe we kunnen corrigeren voor selectiviteit in de verschillende categorieën waarvoor informatie beschikbaar is. Een tweede onderwerp vormt het combineren van informatie uit teksten met de al beschikbare bronnen op het CBS. Bij koppelen op eenheidenniveau is het probleem dat de direct identificerende kenmerken ontbreken (twitterberichten), dat ze nog allerlei fouten bevatten, of dat ze niet uniek zijn (websites). Een laatste onderwerp heeft te maken het schatten van de kwaliteit van CBS-publicaties. Stel we hebben van een categoriale variabele twee reeksen: één op basis van een steekproef of een administratieve bron en één op basis van tekstanalyse. Dan kunnen we vervolgens die twee reeksen combineren in een zogeheten latente variabele model om vervolgens de kwaliteit van beide bronnen, dus ook die van de CBS-publicatie, te schatten.

Literatuur

Blei, D. (2012) Introduction to Probabilistic Topic Models. *Comm. ACM*, *55*(4), 77–84.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The elements of statistical learning: Data mining, inference, and prediction* (second edition). New York: Springer VerlagSpringer.

Roelands, M., Delden, A. van, & Windmeijer, D. (2017). Classifying businesses by economic activity using web-based text mining. CBS discussion paper 2017-18.

Ruder, S. (2016) On word embeddings – part 1. Blog: http://ruder.io/word-embeddings-1/

Van der Doef, S., Daas, P., & Windmeijer, W. (2018). Identifying innovative companies from their website. Abstract for BigSurv18 conference (ingediend).

Arnout van Delden is senior methodoloog bij het Centraal Bureau voor de Statistiek in Den Haag
E-mail: a.vandelden@cbs.nl

Piet Daas is senior methodoloog en lead data scientist bij het Centraal Bureau voor de Statistiek in Heerlen
E-mail: pjh.daas@cbs.nl

Olav ten Bosch is project manager bij het Centraal Bureau voor de Statistiek in Den Haag
E-mail: o.tenbosch@cbs.nl

Dick Windmeijer is data scientist bij het Centraal Bureau voor de Statistiek in Den Haag
E-mail: hjm.windmeijer@cbs.nl


Stephanie van der Pas

Jury report Willem R. van Zwet Award 2017

This year we received five nominations for the Van Zwet award. All five were high quality theses, and together they form a nice profile of our society by their variety of topics on Statistics and Operations Research. It was going to be a difficult choice for the jury. Nevertheless, the jury saw two theses standing out.

One of them is awarded as the runner-up: Krzysztof Postek, title of the thesis is *Distributionally and Integer Adjustable Robust Optimization*, supervised by professor Dick den Hertog from the Universiteit Tilburg. He defended his thesis in February 2017. Krzysztof's thesis is about new optimization models and algorithms that are robust with respect to all kinds of uncertainty in the data. The thesis is based on five papers, most of these are now published in high-ranked Operations Research journals. The jury was impressed by the scientific quality both theoretically and practically. And specifically the chapter where robust optimization is applied to stochastic programming problems is remarkable because it combines two major but rather different fields in OR.

The winner is Stephanie van der Pas, title of the thesis is *Topics in Mathematical and Applied Statistics*, supervised by professor Aad van der Vaart from the Universiteit Leiden. She defended her thesis also in February 2017. Stephanie's thesis is based on 7 papers that are spread over 4 topics in statistics: frequentist properties of Bayesian horseshoe prior; network analysis; sequentially collected data; and the last part is the application to survival analysis of hip replacements. All these papers found their way to international scientific journals. Specifically, the jury would like to mention chapter 3 which is based on a paper in which Stephanie shows that Bayesian methods do not correct automatically for multiplicity. This paper together with discussions of experts is published in *Bayesian Analysis*, and has already hundred plus citations. It will likely become an influential paper.

Jelle Goeman, chairman of the jury
e-mail: J.J.Goeman@lumc.nl

Jury rapport voor de Jan Hemelrijk Award 2017

Er waren dit jaar 4 aanmeldingen voor de Hemelrijk Award, gericht op de beste masterscriptie in Statistiek en Operation Research. We zien al een aantal jaar eenzelfde patroon hierbij: het gemiddelde cijfer van de 4 scripties was een 9,5, dus de eisen die de begeleiders stellen voordat ze een scriptie insturen zijn zeer hoog. Dit maakt het werk van de jury zowel lichter als plezieriger: hiervoor dank!

Een ander opvallend patroon is het volgende: hoewel de inzendingen allemaal van zeer hoog niveau zijn, is er ook dit jaar weer duidelijke overeenstemming binnen de jury over wie de Award verdient: een scriptie die op bijna alle punten beter scoort. Dit jaar is dat de scriptie van de winnaar van de Hemelrijk Award 2017: Stefan ten Eikelder!

Stefan heeft onderzocht hoe bij *radiation therapy* gekozen moet worden tussen fotonen- en protonenstraling. Deze worden op verschillende manieren opgenomen door het lichaam, waardoor ze ook op verschillende manieren schade aan gezond weefsel veroorzaken. U kunt zich voorstellen dat dit zeer relevant onderzoek is, en Stefans werk wordt dan ook nu al gebruikt in sommige ziekenhuizen om stralingstherapie te optimaliseren. Maar deze scriptie is niet alleen zeer relevant: om de gebruikte stralingsmethoden te kunnen optimaliseren binnen redelijke tijd, moest Stefan een nieuwe op maat gemaakte optimaliseringmethode ontwikkelen, gebaseerde op slimme heuristiek, maar ook getest op kwaliteit. Zo draagt Stefan ook bij aan de wiskunde in het algemeen. Al met al een zeer indrukwekkend resultaat en hij is dan ook een zeer verdiende winnaar van de Hemelrijk Award 2017!

Eric Cator, voorzitter van de jury
e-mail: e.cator@science.ru.nl


Stefan Eikelder