

Eindrapportage

Pilot-onderzoek AI-aanbod bij bedrijven

Versie 1.1
10-02-2022

In opdracht van

Ministerie van Economische Zaken en Klimaat

Uitgevoerd door

Innovatiespotter

Protonstraat 3A,
9743 AL Groningen
(050) 313 48 98
info@innovatiespotter.nl

Centraal Bureau voor de Statistiek

Henri Faasdreef 312
2492 JP Den Haag
(070) 337 38 00
info@cbs.nl

Rijksdienst voor Ondernemend Nederland

Prinses Beatrixlaan 2
2595 AL Den Haag
(088) 042 4242
info@rvo.nl

INHOUD

INLEIDING	3
1.1 ACHTERGROND	3
1.2 DOELSTELLING	3
1.3 BEGRIPPENLIJST	3
1.4 LEESWIJZER	4
2 INSCHATTEN VAN AI MET AI	5
2.1 FASE 1 – DEFINITIE EN OPERATIONALISATIE	5
2.2 FASE 2 – SAMENSTELLEN TRAININGSSET EN ONTWIKKELEN MODEL	7
2.3 FASE 3 – VALIDATIE EN TWEEDE ITERATIE	9
2.4 FASE 4 – DERDE ITERATIE	10
2.5 CONCLUSIE	10
3 INSCHATTING OP BASIS VAN AANVULLEND ONDERZOEK	11
3.1 INLEIDING	11
3.2 INSCHATTING OP BASIS VAN ALTERNATIEVE METHODEN	11
3.3 VERGELIJKING VAN DE RESULTATEN MET DE WBSO	12
3.4 INTERNATIONALE CONTEXT	14
4 CONCLUSIE	16
5 SUGGESTIES VOOR VERVOLG	17
5.1 VERVOLGSTAPPEN	17
5.2 ONDSCHIED TUSSEN ONTWIKKELAAR EN TOEPASSER	17
BIJLAGE 1 DETAILRESULTATEN PER UITGEVOERDE STAP	19
STAP 1: DEFINITIE	19
STAP 2: OPERATIONALISATIE	20
STAP 3: PERSONA'S	22
STAP 4. SAMENSTELLEN TRAININGSSET	22
STAP 5. TRAINEN EN TESTEN VAN MODEL	25
STAP 6 VALIDATIE	28
HERHALING STAPPEN 5 EN 6	31
BIJLAGE 2 VERGELIJKING MET AI-PROJECTEN IN DE WBSO NADER BEKEKEN	33
BIJLAGE 3 PERSONAS	35
BIJLAGE 4 AI NETWERKEN	40
BIJLAGE 5 TREFWOORDENLIJST ELSEVIER, RVO, INNOVATIESPOTTER	41
BIJLAGE 6 TREFWOORDENLIJST JRC	45
BIJLAGE 7 FEATURES (WOORDEN) PER MODEL	49
BIJLAGE 8 BEGRIPPENLIJST	54
BIJLAGE 9 INTERNATIONALE CONTEXT	56

INLEIDING

1.1 ACHTERGROND

Bij het Ministerie van Economische Zaken en Klimaat (EZK) bestaat de behoefte om een goed beeld te krijgen van het aantal bedrijven in Nederland dat werkt aan *Artificial Intelligence* (AI)-toepassingen.

Er kan onderscheid worden gemaakt tussen het ontwikkelen en produceren van AI-producten en -diensten (aanbod) en het toepassen van AI (gebruik). Om het ontwikkelen van AI door bedrijven inzichtelijk te maken, zijn medio juni 2020 verkennende gesprekken gestart tussen het ministerie van EZK, de Rijksdienst voor Ondernemend Nederland (RVO), het CBS en Innovatiespotter. Op basis van deze gesprekken is een aanpak vormgegeven die voortborduurt op de eerder opgedane ervaringen voor het identificeren van innovatieve bedrijven¹. Op verzoek van het ministerie van EZK is deze aanpak uitgewerkt tot een nieuw projectvoorstel. De resultaten van dit project worden gepresenteerd in dit document.

1.2 DOELSTELLING

De doelstelling van dit project is het identificeren van het aanbod van AI, dat willen zeggen de populatie bedrijven (in omvang, uitgesplitst naar sector, grootteklasse en eventueel leeftijd en regio) die AI-producten en -diensten ontwikkelen.

1.3 BEGRIPPENLIJST

In de bijlage is een uitgebreide begrippenlijst opgenomen. De belangrijkste begrippen worden hier weergegeven.

Algoritme	Een reeks stappen om een beoogd doel te bereiken. In de context van AI betreft dit een reeks instructies in een bepaalde programmeertaal om de computer een probleem op te laten lossen. Algoritmes worden vaak gebruikt om bepaalde beslissingen (deels) te automatiseren.
Logistische regressie	Een voorspellend model met categorische, vaak dichotome, variabelen. Voorbeelden van dichotome variabelen zijn: waar of niet waar, man of vrouw, etc.
Machine learning	De ontwikkeling van algoritmes waarmee systemen automatisch kunnen leren van data, patronen kunnen identificeren en beslissingen kunnen maken. - Supervised machine learning: hierbij wordt de input en de gewenste output aan de computer meegegeven om patronen in de data te vinden waarmee de input om kan worden gezet in de gewenste output - Unsupervised machine learning: hierbij wordt geen gewenste output meegegeven en wordt het aan de computer zelf overgelaten om patronen in de data te ontdekken
Precision en Recall	Precision (precisie of nauwkeurigheid): als een model positief <i>voorspelt</i> (bv.: dit is een AI-bedrijf), hoe vaak is dit dan correct? Een hoge precision betekent dat er weinig false positives zijn. De formule voor precision is:

¹ Eindrapportage pilot studie Innovatieve bedrijven, CBS en Innovatiespotter, 11-01-2020

	$\frac{\text{True positive}}{\text{True positive} + \text{False positive}}$ Recall: hoe veel van de <i>werkelijke</i> positieve gevallen (bv. werkelijke AI-bedrijven) worden als zodanig door het model herkend. Een hoge recall betekent dat er weinig false negatives zijn. De formule voor recall is: $\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$
Scrapen	Ook wel <i>webscrapen</i> . Het gebruik van software om informatie van webpagina's te extraheren om deze vervolgens te analyseren.
Testset	De testset zorgt voor de laatste evaluatie van het model na de trainingsfase. De testset is nodig om de nauwkeurigheid (zie precision en recall) van het model te bepalen.
Topic modeling	Met machine learning en natuurlijke taalverwerking een statistisch model ontwikkelen voor het ontdekken van de abstracte "onderwerpen" die voorkomen in een verzameling documenten.
Trainingsset	Een deel van de dataset dat gebruikt wordt om het model te trainen. Het model leert hierbij van een fractie van de werkelijke data.
Validatieset	Om ervoor te zorgen dat het model getraind kan worden met behulp van de trainingsset, moet het regelmatig geëvalueerd worden. De validatieset zorgt voor deze evaluatie, door het berekenen van de regelmaat waarop fouten worden gemaakt door het model. Uiteindelijk zal het model haar variabelen op basis van de validatie afstemmen.

1.4 LEESWIJZER

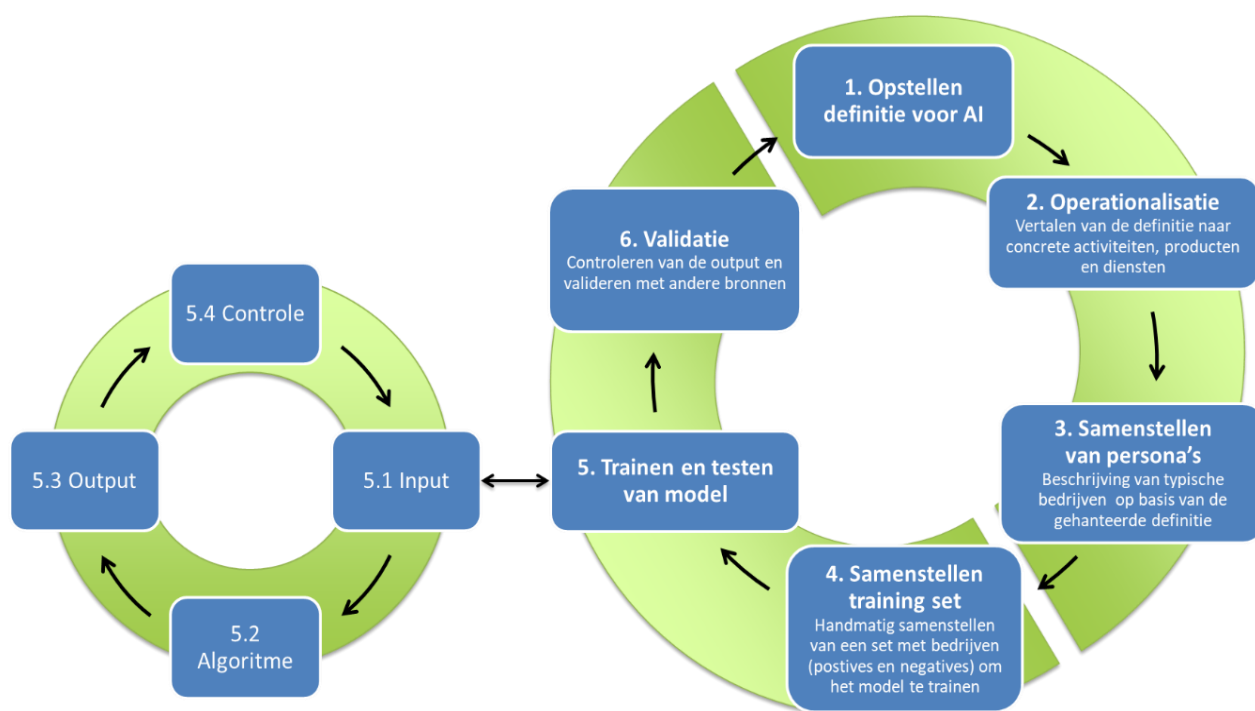
Hoofdstuk 2 beschrijft de belangrijkste resultaten van het onderzoek waarmee met supervised machine learning is getracht een inschatting te maken van het aanbod van AI. In hoofdstuk 3 worden de resultaten van alternatieve aanpakken besproken. De laatste twee hoofdstukken behandelen de conclusies en suggesties voor mogelijk vervolgonderzoek.

2 INSCHATTEN VAN AI MET AI

Als eerste stap in dit project is beoogd om door middel van *supervised machine learning* een model te trainen waarmee automatisch – op basis van de inhoud van de website – kan worden vastgesteld of het bij deze website behorende bedrijf een ontwikkelaar is van AI-diensten of -producten.

Om dit doel te bereiken waren 4 fasen gedefinieerd (zie figuur 1).

- Fase 1 – Definitie en operationalisatie: stap 1-3
- Fase 2 – Samenstellen trainingset en ontwikkelen model 1^e iteratie: stap 4-5
- Fase 3 – Validatie resultaten 1^e iteratie en uitvoering 2^e iteratie: stap 6
- Fase 4 – Derde iteratie: nogmaals stap 4-6 op basis van feedback op de resultaten van fase 2 en 3



Figuur 1 Schematische weergave van de projectaanpak

In het vervolg van dit hoofdstuk worden de belangrijkste resultaten per fase behandeld.

In BIJLAGE 1 Detailresultaten per uitgevoerde stap zijn de uitgevoerde stappen in meer detail beschreven.

2.1 FASE 1 – DEFINITIE EN OPERATIONALISATIE

In de eerste fase van het project is vastgesteld welke definitie gehanteerd diende te worden voor het begrip AI. Er is gezocht naar definities van AI die worden gebruikt door overheid, kennisinstellingen en betrokken partijen uit het bedrijfsleven. Er blijkt geen gangbare definitie te zijn die door iedereen wordt gedeeld. Het Strategisch

Actieplan voor Artificiële Intelligentie (SAPAI)² en ook de Nederlandse Digitaliserings-strategie³ baseren zich op de definitie van de Europese Commissie⁴, die is opgesteld door de High-Level Expert Group on AI (AI HLEG)⁵.

“Er is geen algemeen geldige definitie van AI die consistent wordt gebruikt door alle belanghebbenden. Wij gebruiken de omschrijving van AI door de Europese Commissie: “AI verwijst naar systemen die intelligent gedrag vertonen door hun omgeving te analyseren en - met een zekere mate van zelfstandigheid - actie ondernemen om specifieke doelen te bereiken.” (SAPAI, 2019)

Om de AI-definitie in de praktijk te kunnen toepassen op concrete bedrijven, is een operationalisatie van de definitie gemaakt. Daarvoor moet worden gekeken naar praktische activiteiten, producten en diensten. De NL AI Coalitie gebruikt bijvoorbeeld geen strikte definitie van AI, maar richt zich op specifieke, voor Nederland relevante sectoren. *“Artificial Intelligence (AI) is een generieke technologie die uiteindelijk haar toepassing vindt in alle bedrijfssectoren en maatschappelijke uitdagingen.”*⁶ Ook de AI Index Report van de Stanford University stelt dat AI een technologie is die breed wordt toegepast door bedrijven, *“AI appears to be transforming into a general purpose technology (GPT)”*⁷. Er worden voorbeelden genoemd in sectoren als vastgoed, game-industrie, financiële sector, gezondheidszorg of beveiliging. De OECD stelt ook dat het monitoren van ontwikkelingen in AI lastig is: *“Measuring the development of AI is challenging as the boundaries between AI and other innovations are blurred and change over time.”*⁸

Tabel 1 geeft de uitgewerkte operationalisatie weer met de belangrijkste aspecten voor het evalueren van AI bedrijven. In BIJLAGE 1 Detailresultaten per uitgevoerde stap worden deze elementen in detail beschreven.

Tabel 1. Operationalisatie van AI bedrijven.

Operationalisatie van AI-bedrijven
A) AI kan worden gedefinieerd als:
<i>Systemen die intelligent gedrag vertonen door hun omgeving te analyseren en - met een zekere mate van zelfstandigheid - actie ondernemen om specifieke doelen te bereiken. (SAPAI, 2019)</i>
B) Bij een AI-bedrijf is een of meer AI-(sub)domeinen in meer of mindere mate aanwezig in het product of dienst.
<i>Reasoning, planning, machine learning, natural language processing, computer vision, audio processing, multi-agent systems, robotics and automation, connected and automated vehicles, AI services, AI ethics and philosophy. (Joint Research Center, 2020)</i>

² Strategisch Actieplan voor Artificiële Intelligentie, 2019, p11

<https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie/Rapport+SAPAI.pdf>

³ Nederlandse Digitaliseringsstrategie, 2020, p11 <https://www.nederlanddigitaal.nl/nederlandse-digitaliseringsstrategie/documenten/rapporten/2020/06/25/nederlandse-digitaliseringsstrategie-2020>

⁴ Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM/2018/237 final. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN>

⁵ A definition of AI: main capabilities and scientific disciplines, AI HLEG, April 2019. http://www.pcci.gr/evepimages/0101_F483.pdf

⁶ De Nederlandse AI Coalitie, toepassingsgebieden <https://nlaic.com/toepassingsgebied/>

⁷ Stanford University (2019), “The AI Index 2019 Annual Report”, p.91 <https://hai.stanford.edu/research/ai-index-2019>

⁸ Paragraaf: How can AI-related developments in science and technology be tracked? OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264311992-en>

C) Er is een duidelijk onderscheid in verschillende eigenschappen.	
Software / Hardware	AI kan puur bestaan uit software, of <i>embedded</i> in hardware.
Fysiek / Virtueel	Er kan interactie zijn in de fysieke wereld, of uitsluitend plaatsvinden in de virtuele wereld.
Consument / Zakelijk / Industrie	De toepassing van AI is gericht op een markt. Er is een duidelijk verschil tussen consument, zakelijke en industriële toepassingen.
D) Er worden bedrijven geclassificeerd op basis van informatie op de websites. Bepaalde aspecten kunnen een rol spelen in het onterecht classificeren.	
Achter de schermen (<i>false negative</i>)	De applicatie wordt simpel gepresenteerd, maar gebruikt AI-technologie op de achtergrond.
Foute context (<i>false positive</i>)	Een woord dat is gerelateerd aan AI komt ook in een andere context voor.
Window dressing	Gebruik van relevante woorden terwijl de onderliggende technologie simpel is.
Blogs & nieuws	Bevat blogs over AI, of publiceert/verzamelt nieuws of vacatures over het onderwerp.
Gebruiker	Er wordt AI gebruikt in (een deel van) het bedrijfsproces, maar dit wordt niet zelf ontwikkeld en is niet cruciaal voor het product of dienst.
Enabler	Investeerders, accelerators of andere ondersteuning gericht op AI-bedrijven.

2.2 FASE 2 – SAMENSTELLEN TRAININGSSET EN ONTWIKKELEN MODEL

Samenstellen trainingsset

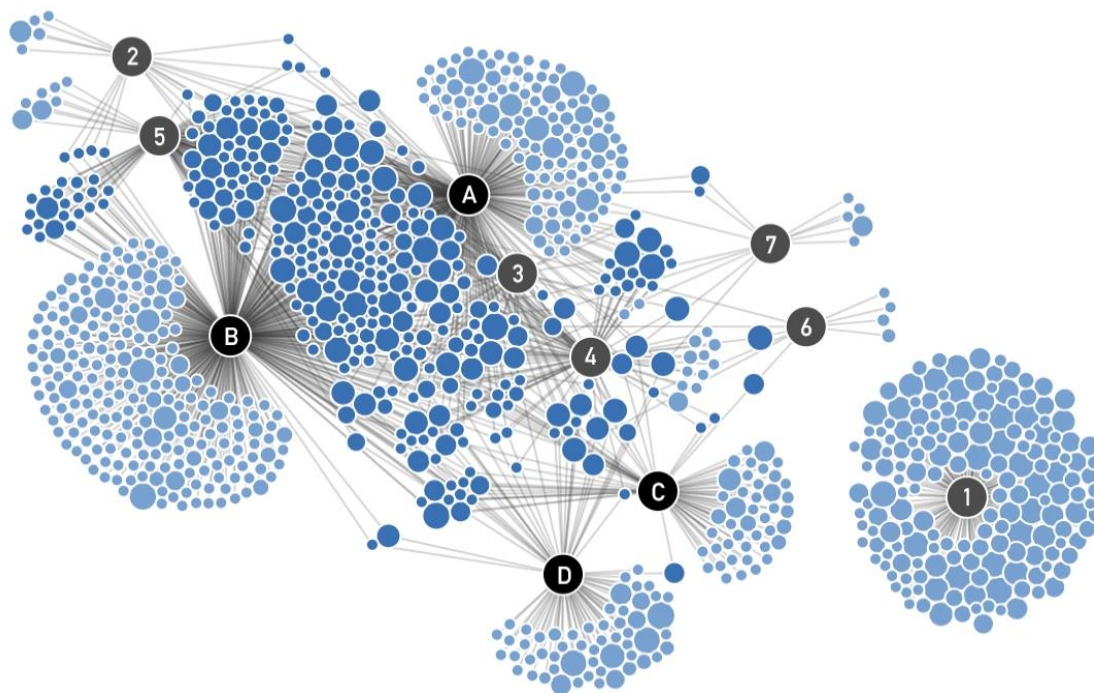
Voor de trainingsset wordt de inhoud van bedrijvenwebsites gebruikt. Bedrijven zijn geselecteerd op basis van hun deelname aan de verschillende AI-netwerken in Nederland. Daarnaast zijn bedrijven handmatig toegevoegd door Innovatiespotter door met trefwoorden te zoeken in de website-inhoud van alle Nederlandse bedrijven met een website. Tabel 2 geeft deze bedrijven en hun samenhang weer. De bedrijvenlijsten zijn afkomstig van de NL AI Coalitie en de AI-hubs. De bedrijvenlijsten zijn geschoond van grote organisaties, adviesbureaus, kennisinstellingen, overheden en andere niet relevante organisaties. Van de overgebleven bedrijven is gekeken hoe de websites van deze bedrijven scoren op beschikbare trefwoordlijsten (zie BIJLAGE 5 Trefwoordenlijst en BIJLAGE 6 Trefwoordenlijst JRC). AI-bedrijven die niet of laag scoren op trefwoorden zijn eveneens verwijderd. Dit leverde een trainingsset op van 951 unieke bedrijven (waarvan 782 afkomstig zijn uit de netwerken, en 169 zijn geselecteerd op basis van de zoekactie met gevalideerde trefwoordlijsten in de Nederlandse bedrijvenpopulatie van Innovatiespotter).

In Figuur 2 en Tabel 2 zijn de relaties te zien tussen de netwerken en bedrijven. Elk bedrijf is een blauwe stip, en de lijnen laten zien uit welke netwerken het afkomstig is. Een donkere kleur betekent dat het bedrijf in meerdere netwerken voorkomt. Een grotere punt betekent dat de website van het bedrijf beter scoort op de trefwoordlijsten.

- AI-hub
- Andere bronnen
- 1 netwerk
- 2 of meer netwerken
- Trefwoord score

- AI-hubs:
A) Techleap
B) InnovationQuarter
C) BOM
D) NL AI Coalitie

- Andere bronnen:
1) Innovatiespotter
2) AI in Lifeport
3) AI Index 2019
4) AI Startup Landscape 2020
5) StartupDelta 2018
6) SBIR oproep AI
7) Rockstart



Figuur 2 Relaties tussen de netwerken en bedrijven uit de trainingsset.

Tabel 2 Aantal bedrijven per netwerk.

Netwerk	Aantal bedrijven
InnovationQuarter	485
Techleap	361
Innovatiespotter*	169
StartupDelta	116
AI Startup Landscape	87
NL AI Coalitie	84
BOM	79
AI Index 2019	38
AI in Lifeport	20
Rockstart AI	13
SBIR AI	10
Totaal	951

*Gevonden op basis van trefwoorden en handmatig beoordeeld.

Het grootste aantal bedrijven is afkomstig uit de netwerken van InnovationQuarter en Techleap, of beoordeeld op basis van gevalideerde trefwoorden door Innovatiespotter. Een derde van de bedrijven komt in 2 of meer netwerken voor (40% als de groep van Innovatiespotter niet wordt meegerekend). Veel van deze bedrijven hebben ook een hoge trefwoordscore.

Trainer en testen van model

Innovatiespotter heeft deze 951 bedrijven overgedragen aan CBS voor de positieve trainingsset. Deze positieve trainingsset is door het CBS aangevuld met (random) negatieve bedrijven voor de negatieve trainingsset. In totaal zijn er 1.779 unieke URL's gebruikt die gescraped konden worden: 80% voor trainingsdoeleinden, 20% voor validatie.

Er zijn vervolgens 3 verschillende modellen getraind en getest. Deze modellen zijn in eerste instantie alleen getraind op teksten van de homepage van de bedrijven. Tijdens de training ligt de nauwkeurigheid van de modellen tussen de 91% en 93%. Dat betekent dat de voorspellingen die het model doet (wel of niet AI) voor een groot deel overeenkomen met de classificatie die wij aan deze bedrijven hebben gegeven. De modellen maken duidelijk onderscheid tussen bedrijven die als AI worden aangemerkt (hoge score) en bedrijven die duidelijk als niet-AI worden aangemerkt (lage score). Dezelfde getrainde modellen zijn vervolgens toegepast op de openbare data van bedrijvenwebsites van 600.000+ bedrijven uit het ABR van het CBS. Dit levert als resultaat – afhankelijk van het model – een classificatie van tussen de 57.322 en 120.137 bedrijven die AI-producten en -diensten ontwikkelen. Dat zijn, gezien schattingen uit ander onderzoek en vanuit gezond verstand, onwaarschijnlijk hoge waarden. In de volgende fase worden deze uitkomsten nader bekeken.

2.3 FASE 3 – VALIDATIE EN EERSTE ITERATIE

Uit de resultaten is een willekeurige steekproef van bedrijven geselecteerd om handmatig te valideren. Voor de validatie is gebruik gemaakt van Q-Necton, de validatiemodule van Innovatiespotter (ISP)⁹. Hiermee wordt de website en alle relevante informatie van een bedrijf getoond om tot een goede afweging te komen en kunnen classificaties handmatig worden aangepast. De validatie is uitgevoerd door CBS, ISP en RVO.

De eerste validatieronde is voortijdig beëindigd omdat veruit de meeste gevonden bedrijven negatief (niet-AI) worden beoordeeld. Er zijn slecht 2 bedrijven (van de 89) die door de meerderheid van de beoordelaars daadwerkelijk als AI ontwikkelaar wordt beschouwd.

Wat daarbij opvalt is dat er overwegend ICT-bedrijven positief geïdentificeerd worden door het model. De SBI is geen feature in het model, dus deze classificatie vindt alleen plaats op basis van de inhoud van de website. Concluderend levert de 1^e iteratie training van het model onvoldoende resultaat voor een betrouwbare schatting.

Op basis van de eerste validatieronde worden de positieve trainingsset en negatieve trainingsset uitgebreid met extra bedrijven, deels op basis van de geannoteerde websites uit de vorige ronde. De verdeling positief-negatief is aangepast van 50-50 naar 33-66. Na hertrainen van de modellen daalde het aantal positief geïdentificeerde bedrijven van 57.322 naar 8.442 of 8.802, afhankelijk van of alleen de voorpagina of de gehele website wordt gebruikt voor de classificatie. Toch blijkt ook in de tweede handmatige validatieronde dat de meeste bedrijven door alle beoordelaars als negatief (niet-AI) worden beoordeeld.

Tijdens de tweede validatieronde viel op dat er minder overduidelijk foute resultaten waren en veel meer twijfelgevallen. De twijfel zit vooral bij bedrijven waarbij de AI-producten of -diensten geen core business lijken te zijn of waarbij er wordt geïmpliceerd dat er AI wordt gebruikt, maar de manier waarop niet duidelijk is. Het meest subtiele verschil is zichtbaar bij websites met Business Intelligence (BI)-achtige applicaties, die wel data verwerken en inzicht bieden, maar het probleem op laten lossen door mensen in plaats van door algoritmes.

⁹ [Qnection.innovatiespotter.nl](https://qnection.innovatiespotter.nl)

2.4 FASE 4 – LAATSTE ITERATIE

Op basis van de resultaten van fase 2 en 3 is nog een laatste iteratie uitgevoerd om de resultaten te verbeteren. Er zijn in deze ronde nog meer niet-AI bedrijven toegevoegd aan de trainingsset om een grotere set (met grotere spreiding) aan negatieve voorbeelden te hebben; in deze set zaten ook iets meer IT-bedrijven. De uiteindelijke trainingsset bevat nu 3.895 niet-AI bedrijven en 898 AI-bedrijven.

Van de steekproef van 370 bedrijven die in deze ronde worden beoordeeld, worden er 36 (een kleine 10%) daadwerkelijk als AI beoordeeld. Veruit de meeste bedrijven (72%) worden echter ook nu alsnog negatief (niet-AI) beoordeeld. De resterende 18% wordt beoordeeld als neutraal/twijfel, waarbij niet met zekerheid vastgesteld kan worden dat het bij AI past, maar wel enkele indicatoren op de website aanwezig zijn.

2.5 CONCLUSIE

Samenvattend kan het volgende worden geconstateerd:

- De eerste versies van het machinelearning-model classificeerden vooral ICT bedrijven. Door de negatieve set uit te breiden kon dit enigszins worden verholpen.
- Hoewel de nauwkeurigheid tijdens het trainen goed leek te zijn, werd een groot deel van de classificaties tijdens de handmatige validatieronden afgekeurd.
- Ook als er op de website indicaties zijn van AI, is vaak niet eenduidig vast te stellen in welke mate er AI wordt ontwikkeld of gebruikt.

In het eerder aangehaalde rapport van de OECD¹⁰ wordt vastgesteld dat het meten van AI lastig is, omdat de scheiding tussen AI en andere innovaties onscherp is en voortdurend wijzigt. Het resultaat van ons onderzoek lijkt dit te onderschrijven. De conclusie op basis van deze resultaten is dat er met de gehanteerde aanpak geen betrouwbare schatting af te geven is van het aantal bedrijven dat AI-producten en -diensten ontwikkelt. We vermoeden dat een extra complicerende factor is dat de populatie bedrijven die AI-producten of -diensten aanbieden vrij klein is. Een machinelearning-model zal in dit geval dus uiterst precies moeten zijn om geen AI bedrijven fout te classificeren.

¹⁰ Paragraaf: *How can AI-related developments in science and technology be tracked?* OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264311992-en>

3 INSCHATTING OP BASIS VAN AANVULLEND ONDERZOEK

3.1 INLEIDING

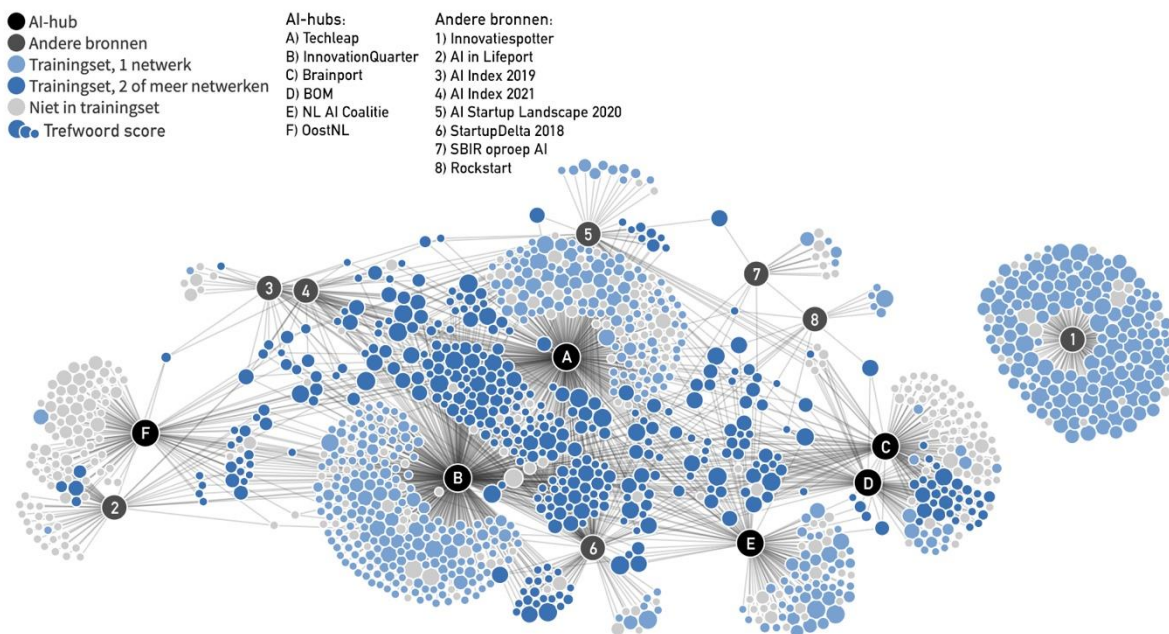
In het vorige hoofdstuk bleek dat het met supervised machine learning niet mogelijk is om een betrouwbare schatting af te geven van het aantal bedrijven dat AI-producten en -diensten ontwikkelt. Daarom is gezocht naar alternatieve methoden en andere onderzoeken om toch een beter beeld te krijgen van de populatie. De resultaten van deze exercities worden in dit hoofdstuk beschreven.

Allereerst is geprobeerd om de initiële trainingsset uit te breiden door extra netwerken en bedrijven handmatig toe te voegen. Vervolgens is gekeken naar de populatie AI-bedrijven in de fiscale stimuleringsregeling WBSO. De twee genoemde populaties zijn vergeleken om te kijken in hoeverre er overlap is tussen deze groepen. Dit kan een aanvullend inzicht geven in de omvang van de AI-populatie in Nederland.

Als laatste wordt gekeken naar internationale onderzoeken, om te kijken welke populaties AI-bedrijven zij vinden en op welke manieren dit is gedaan. Deze exercities helpen om beter in te kunnen schatten hoe groot de groep AI-ontwikkelaars in Nederland is en wat de kwaliteit is van de gevonden populatie.

3.2 INSCHATTING OP BASIS VAN ALTERNATIEVE METHODEN

Zoals beschreven in paragraaf 2.2, is de initiële trainingsset van bedrijven samengesteld op basis van deelname aan de verschillende AI-netwerken in Nederland, aangevuld met 169 bedrijven die door ISP middels trefwoorden zijn gevonden. Gedurende het verloop van het project zijn er ook nog aanvullende bedrijvenlijsten ontvangen van AI-hubs zoals Brainport, OostNL, en een aanvullende ledenlijst van de NL AI Coalitie. Dit heeft geresulteerd in een verzameling van 1.258 bedrijven. De relaties van deze netwerken en bedrijven zijn te zien in Figuur 3.



Figuur 3. Aanvulling op de trainingsset (in grijs).

De 1.258 bedrijven die op deze manier zijn gevonden, zijn vervolgens nog aangevuld met bedrijven uit een laatste trefwoord-exercitie. Door de gehele Nederlandse bedrijvenpopulatie (Q2 2021) te doorzoeken met trefwoorden is gezocht naar bedrijven met indicatoren voor AI-ontwikkelactiviteiten. Beide methodes (nieuwe netwerken zoeken en met trefwoorden zoeken in de gehele Nederlandse bedrijvenpopulatie) zijn herhaald tot vrijwel geen nieuwe bedrijven meer werden gevonden. Op deze manier zijn er nog eens 332 bedrijven gevonden, wat het totaal op 1.590 bedrijven heeft gebracht.

Verskil in aantal KVK-inschrijvingen en aantal bedrijfswebsites

Voor een goede vergelijking met het aantal gevonden AI-bedrijven in de WBSO (zie de volgende paragraaf, 3.3), is gerekend met KvK-nummers, waar in andere delen van dit rapport het aantal bedrijven wordt geteld aan de hand van de websites. Het gaat om 1.590 bedrijven en 2.387 KvK-nummers.

Kanttekeningen

De gebruikte trefwoorden zijn (gevalideerde) indicatoren voor AI-ontwikkelaars (zie bijlage 5), maar een trefwoord als “machine learning” bevat geen context of dit door een bedrijf wordt ontwikkeld of toegepast. Ook presteert een trefwoordenlijst die is opgesteld op basis van projectbeschrijvingen of literatuur niet hetzelfde op websiteteksten. Daardoor kunnen bedrijven die AI alleen toepassen ook in deze selectie voorkomen (false positives). Bedrijven die wel AI ontwikkelen maar hiermee niet via hun website in de openbaarheid treden, missen in deze populatie (false negatives). Daarom is het niet mogelijk harde uitspraken te doen over de totale populatie AI-ontwikkelaars.

Conclusie

Via netwerken en zoeken met gevalideerde trefwoorden is het mogelijk gebleken om een lijst van 1.590 bedrijven samen te stellen met sterke indicatoren voor AI-ontwikkelingen (tweede kwartaal 2021). De verwachting is dat dit aantal bedrijven snel groeit. Deze populatie bestaat niet enkel uit ontwikkelaars van AI (aanbod) maar ook uit bedrijven die op andere manieren betrokken zijn bij AI. Deelname aan een AI-netwerk alleen blijkt niet voldoende basis om een bedrijf als AI-ontwikkelaar te kenmerken. Er is aanvullende analyse (onder andere handmatige controle) voor nodig om vast te stellen welke bedrijven daadwerkelijk AI-ontwikkelaar zijn. Bedrijven die wel AI ontwikkelen maar hiermee niet in de openbaarheid treden blijven met deze methode buiten beeld.

3.3 VERGELIJKING VAN DE RESULTATEN MET DE WBSO

RVO heeft in een parallel pilot-traject in kaart proberen te brengen wat de omvang van AI is binnen de WBSO. Via de WBSO stimuleert het ministerie van Economische Zaken en Klimaat ondernemers om meer te investeren in speur- en ontwikkelingswerk (S&O), ook wel bekend als Research en Development (R&D). Op basis van alle projecten in 2019 en een gewogen-trefwoordenmodel, is een totaal aantal van 1.515 bedrijven gevonden met één of meerdere AI-projecten. Het betreft zo’n 8,4% van alle bedrijven (exclusief zelfstandigen) in de WBSO en 3,3% van alle projecten in 2019. Van de gevonden bedrijven was 96% MKB. Het volledige rapport, met meer informatie over de aanpak, de resultaten en een uitgebreide validatie van de resultaten is op moment van schrijven nog niet gepubliceerd. RVO is voornemens om het eindrapport op korte termijn te publiceren op de website www.bedrijvenbeleidinbeeld.nl, namens het ministerie van EZK.

Verschillen in onderzoeken

Het in de WBSO gevonden aantal van 1.515 bedrijven komt aardig overeen met het uiteindelijke aantal dat is gevonden in dit project (1.590). Ogenscheinlijk lijken dit vergelijkbare resultaten. Er zijn echter belangrijke verschillen in de uitgangspunten van beide onderzoeken. In de tabel hieronder zijn deze opgesomd. Een goed begrip van deze verschillen is noodzakelijk bij het vergelijken van de resultaten uit beide onderzoeken.

Tabel 3 Verschillen in onderzoeken Innovatiespotter/CBS en WBSO

	Innovatiespotter/CBS	WBSO	Invloed op uitkomsten
Gebruikte bronnen	Algemene publiek toegankelijke teksten van websitepagina's (waaronder 150 KvK-inschrijvingen die zijn opgericht in 2020-2021). <i>Doel van teksten: Aanprijzen diensten aan een breed publiek.</i>	Teksten uit 135.000 projectomschrijvingen (niet-uniek) in 2019 (technisch van aard) van bedrijven, excl. zelfstandigen. <i>Doel van teksten: Aantonen technische nieuwheid ter verkrijging S&O-verklaring.</i>	Door de algemene aard van de websites kunnen er bij de CBS/Innovatiespotter-methode AI-bedrijven 'verborgen' blijven. Deze bedrijven kunnen wel gevonden worden als zij een AI-project in de WBSO uitvoeren. Illustratie: een AI-bedrijf dat zich volledig op consumenten richt en daardoor geen technische termen opneemt op de website. Of een bedrijf dat een algoritme ontwikkelt voor de kwaliteitscontrole van een fabriekslijn, en waar AI dus onderdeel is van een proces.
Classificatie	Bedrijven die aan AI-ontwikkeling doen.	Projecten die AI-ontwikkeling bevatten. Hiermee kunnen bedrijven geteld worden die tenminste één AI-project toegekend kregen binnen de WBSO alleen in 2019.	De verwachting is dat de groep bedrijven die één of meer AI-projecten doet, veel groter is dan de groep bedrijven die puur AI-ontwikkelaar is. Illustratie; een groot supermarktconcern dat een AI-project met een ontwikkelingscomponent in de WBSO heeft, zal dit waarschijnlijk niet op hun website aangeven.
Methode	Het betreft de trainingsset die is gebruikt om een algoritme te trainen. De bedrijven zijn geselecteerd omdat ze in het netwerk zitten van Nederlandse AI-netwerken. Ook is met trefwoorden gezocht naar bedrijven en handmatig gevalideerd. <i>Validatie door inhoudelijke experts de websites te laten beoordelen of er sprake is van een bedrijf dat AI ontwikkelt.</i>	Strikt zoeken naar vooraf opgestelde zoektermen met weging. Door een script worden de projecten gevonden en voorzien van een projectscore. <i>Validatie door inhoudelijk expert te laten beoordelen en een grenswaarde te bepalen waarboven de projecten als AI- worden geclassificeerd.</i>	De methode heeft op zichzelf weinig invloed; wel van invloed zijn de gebruikte bronnen, zoals hierboven reeds genoemd.
Knelpunten	Het trainen van het model om correcte resultaten te krijgen voor AI-bedrijven bleek lastiger dan voor andere thema's. - De populatie was te klein. - De scheidslijn tussen toepassers en	Na zes iteraties (aanpassingen in de zoektermenlijst en wegingen) bleek het mogelijk om resultaten te vinden in de zoektocht naar AI-projecten. We verwachten dat deze aanpak de totale omvang van de populatie goed schat, maar we verwachten dat ongeveer 1/3 van de gevonden	Niet van toepassing.

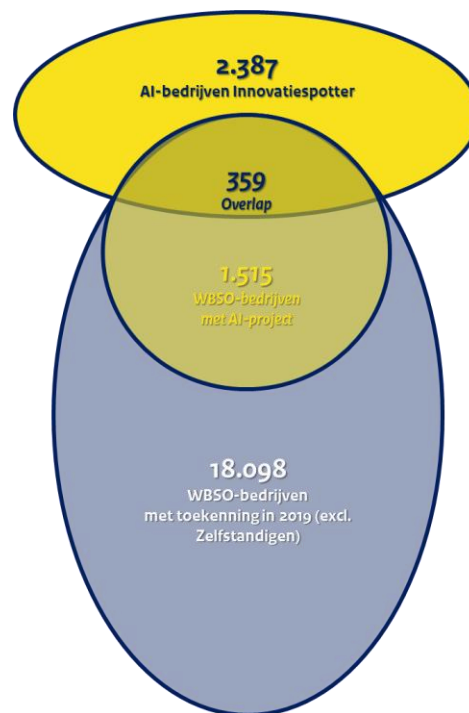
	ontwikkelaars was lastig te bepalen. <i>Het vinden van AI-ontwikkelaars is niet zondermeer met één druk op de knop uit te voeren, validatie door de mens blijft noodzakelijk.</i>	set niet juist zal zijn. Dit wordt gecompenseerd door het deel dat het model heeft gemist. <i>Het vinden van AI-projecten is niet zondermeer met één druk op de knop uit te voeren, validatie door de mens blijft noodzakelijk.</i>	
--	--	--	--

Overlapanalyse

Ondanks deze verschillen, is het toch interessant om te kijken in hoeverre deze populaties overeenkomen. Dat is onderzocht; van de lijst met 2.387 unieke KvK-nummers die horen bij 1.590 unieke bedrijven, is gezocht of deze in 2019 een AI-project bij de WBSO hadden. De overlap blijkt beperkt: 359 KvK-nummers komen overeen. Dit is 15% van de totale lijst van Innovatiespotter. Andersom beredeneerd, vinden we 24% van alle gevonden AI-bedrijven in de WBSO in 2019 terug in de lijst van Innovatiespotter (359 van 1.515 bedrijven).

De eerdergenoemde verschillen in opzet van de onderzoeken verklaren voor een deel waardoor de resultaten verschillen. Daarnaast is er nog een aantal andere, kleinere zaken die de overlap (kunnen) beperken¹¹.

Dan blijft nog over de bedrijven die wél AI-ontwikkelaars zijn, maar geen WBSO-subsidie ontvangen. Ofwel omdat ze geen WBSO-subsidie hebben aangevraagd, ofwel omdat hun aanvraag is afgewezen.



3.4 INTERNATIONALE CONTEXT

Dit door CBS, RVO en Innovatiespotter uitgevoerde onderzoek staat niet op zichzelf. Er zijn diverse internationale onderzoeken uitgevoerd om met machine learning (de adoptie van) AI in kaart te brengen. Deze onderzoeken verschillen wel op een wezenlijk punt. Bij deze onderzoeken worden populaties van bedrijven middels unsupervised learning ingedeeld in topics (clusters). Hoe deze populaties zijn vastgesteld, en of ze volledig zijn, wordt niet beschreven. In dit onderzoek wordt met een trainingset naar nieuwe bedrijven gezocht. Een vergelijkbaar experiment om binnen de trainingset te kijken naar topics is te lezen in de bijlage (Clusters binnen de trainingsset).

¹¹ Bij de WBSO-analyse is alleen gekeken naar projecten uit 2019. Er kunnen bedrijven zijn die eerder wel WBSO gebruikten en nu niet meer. Daarnaast had ongeveer 10% van de bedrijven in de dataset van Innovatiespotter een oprichtingsdatum in 2020 of 2021. De WBSO-analyse maakt gebruik van projecten uit 2019, waardoor deze nieuwe bedrijven niet meegenomen worden. Als we deze bedrijven zouden uitsluiten, zou de procentuele overlap groter zijn. Als laatste zou de overlap ook groter kunnen zijn, als er was gekeken naar bedrijfseenheden in plaats van KvK-nummers. Een bedrijf heeft doorgaans een structuur met van verschillende KvK-inschrijvingen. De WBSO-aanvraag wordt gedaan voor de medewerkers die op de loonlijst staan van de entiteit/KvK-nummer waar daadwerkelijk de S&O-werkzaamheden. In zo'n geval zal de website een algemeen totaalbeeld schetsen van de activiteiten die worden uitgevoerd door de verschillende onderdelen. Afhankelijk hoe de verschillende werkmaatschappijen gepositioneerd zijn en beschreven zijn zou dit uit een website kunnen blijken, maar dat hoeft natuurlijk niet.

In Tabel 4 zijn de belangrijkste eigenschappen van deze onderzoeken weergegeven. Wat opvalt is dat de verschillende ‘subtopics’ sterk uiteen lopen. Dit geeft nog maar eens aan dat AI klaarblijkelijk lastig te meten is. De volledige analyse is te vinden in BIJLAGE 9 Internationale context.

Tabel 4 (Internationale) onderzoeken naar de adoptie van AI

Organisatie	Ipsos	OECD	Stanford University
Onderzoekstitel	European enterprise survey on the use of technologies based on artificial intelligence	Measuring the Digital Transformation: A Roadmap for the Future,	The AI Index 2019 Annual Report
Jaar	2020	2019	2019
Geografisch onderzoeksgebied	EU & Nederland	UK	Mondiaal & EU
Gegevensbron	Enquêtevragen (n=9.640)	Websites	Onbekend
Aantal bedrijven in steekproef/selectie	9.640 EU (w.v. 42% ≥1 AI-technologie toepast) 500 NL (w.v. 40% ≥1 AI-technologie toepast).	2.725 UK	4.403 mondiaal 993 EU 55 NL
Selectie van AI-bedrijven	Bedrijven die zeggen AI toe te passen (niet alleen ontwikkelaars)	Vermelding van AI-gerelateerde woorden op de website	Vermelding van AI-gerelateerde termen, niet bekend of dit om websites gaat.
AI-methode	Unsupervised clustering	Unsupervised clustering	Unsupervised clustering
Naamgeving dimensies	Low AI use <----> High AI use Enterprise app <----> customer-facing app	Natural language processing <----> analytics Recognition <----> automation	Onbekend
Aantal subtopics	10	9	36
Subtopics	NLP Computer vision Anomaly detection Sentiment analysis Forecasting Process optimisation Recommendation engines Process automation Autonomous machines Creative activities	Business solution E-commerce Language processing Deep learning Data mining Autonomous vehicles Image recognition Robotics, IoT & VR Intrusion detection	Data tools Fashion, Retail Tech Automation, Oil & Gas Text analytics Fintech Medtech Autonomous vehicles Chatbots Marketing, adtech Energy management Medical Imaging EdTech Facial recognition Robotic process automation
Hoe zijn subtopics tot stand gekomen?	Enquêteantwoorden	Topic modeling	Topic modeling

4 CONCLUSIE

Door de OECD¹² wordt in een rapport vastgesteld dat het meten van AI lastig is omdat de scheiding tussen AI en andere innovaties onscherp is en voortdurend wijzigt. Het AI Index Report van de Stanford University stelt dat AI zich lijkt te ontwikkelen tot een *general purpose technology* (GPT).¹³ Het resultaat van ons onderzoek lijkt deze constatering te onderschrijven.

De doelstelling van dit project was het identificeren van de populatie bedrijven (in omvang, uitgesplitst naar sector, grootteklasse en eventueel leeftijd en regio) die AI-producten en -diensten aanbieden. Dat doel hebben we niet bereikt op de manier zoals we die voor ogen hadden, namelijk met een getraind model dat in staat is om met voldoende nauwkeurigheid op basis van de websitetekst vast te stellen of een bedrijf ontwikkelaar is van AI-producten en -diensten. Het model dat we hiervoor wilden inzetten presteerde hiervoor ook na verschillende iteraties nog onvoldoende. Dit bleek uit handmatige validatie. Het bleek moeilijk te zijn om op basis van de website tekst onderscheid te maken tussen ICT-bedrijven en AI-bedrijven. Het onderscheiden van AI-toepassers en AI-ontwikkelaars bleek vervolgens nog een extra uitdaging te zijn. De conclusie was daarom dat op basis van deze resultaten geen betrouwbare schatting af te geven is van het aantal bedrijven dat AI-producten en -diensten ontwikkelt.

Via alternatieve methoden zijn we alsnog tot een lijst van 1.590 bedrijven gekomen. Deze lijst kan als basis dienen voor een eerste inventarisatie van de AI-populatie in Nederland. Belangrijke kanttekening daarbij is dat deze populatie van 1.590 bedrijven niet enkel bestaat uit ontwikkelaars van AI (aanbod) maar ook uit bedrijven die op andere manieren betrokken zijn bij AI. Bedrijven die AI ontwikkelen, maar hiermee niet via hun website in de openbaarheid treden, blijven met deze methode buiten beeld.

¹² Paragraaf: *How can AI-related developments in science and technology be tracked?* OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264311992-en>

¹³ Stanford University (2019), "The AI Index 2019 Annual Report", p.91 <https://hai.stanford.edu/research/ai-index-2019>

5 SUGGESTIES VOOR VERVOLG

5.1 VERVOLGSTAPPEN

We zien twee mogelijke opties voor vervolgstappen om het AI-aanbod in Nederland beter in kaart te brengen:

1. De nu (deels handmatig) vastgestelde lijst van 1.590 bedrijven kan handmatig worden onderhouden. Op die manier ontstaat een groeiende populatie bedrijven die Innovatiespotter beschikbaar kan blijven stellen. CBS kan deze gevalideerde lijst gebruiken om statistische informatie over deze populatie bedrijven te publiceren.
2. Om beter onderscheid te maken tussen aanbieders en ontwikkelaars, kunnen we in een volgend project experimenteren met een aanpak die uit 2 stappen bestaat:
 - a. Een AI-model ontwikkelen die zowel ontwikkelaars als toepassers classificeert
 - b. Daarna met alternatieve technieken (zoals unsupervised clustering) binnen deze selecte onderscheid maken tussen ontwikkelaar en toepasser. Als onderdeel van dit onderzoek hebben we hiervoor al een korte verkenning uitgevoerd.

5.2 ONDERSCHIED TUSSEN ONTWIKKELAAR EN TOEPASSER

Om in de toekomst beter een onderscheid te kunnen maken tussen AI-ontwikkelaars en AI-toepassers, is een set van 16 bedrijven uitgewerkt door op gestructureerde manier naar een bedrijf te kijken en dit in een tabel samen te vatten. Deze set van 16 bedrijven bestond uit bedrijven waarover de beoordelaar tijdens het valideren de optie “second opinion” had geselecteerd, aangevuld met bedrijven uit de tweede validatieronde waarover de beoordelaars niet unaniem waren.

Bij elk bedrijf is gekeken naar de website en is onderzocht of en zo ja welke relevante trefwoorden voorkomen. Verder is naar de medewerkers van het bedrijf op LinkedIn gekeken (hebben zij een data science- of nog beter een AI-achtergrond) en is soms naar vermeldingen van het bedrijf gezocht op internet.

De 16 bedrijven zijn vervolgens op de volgende aspecten beoordeeld:

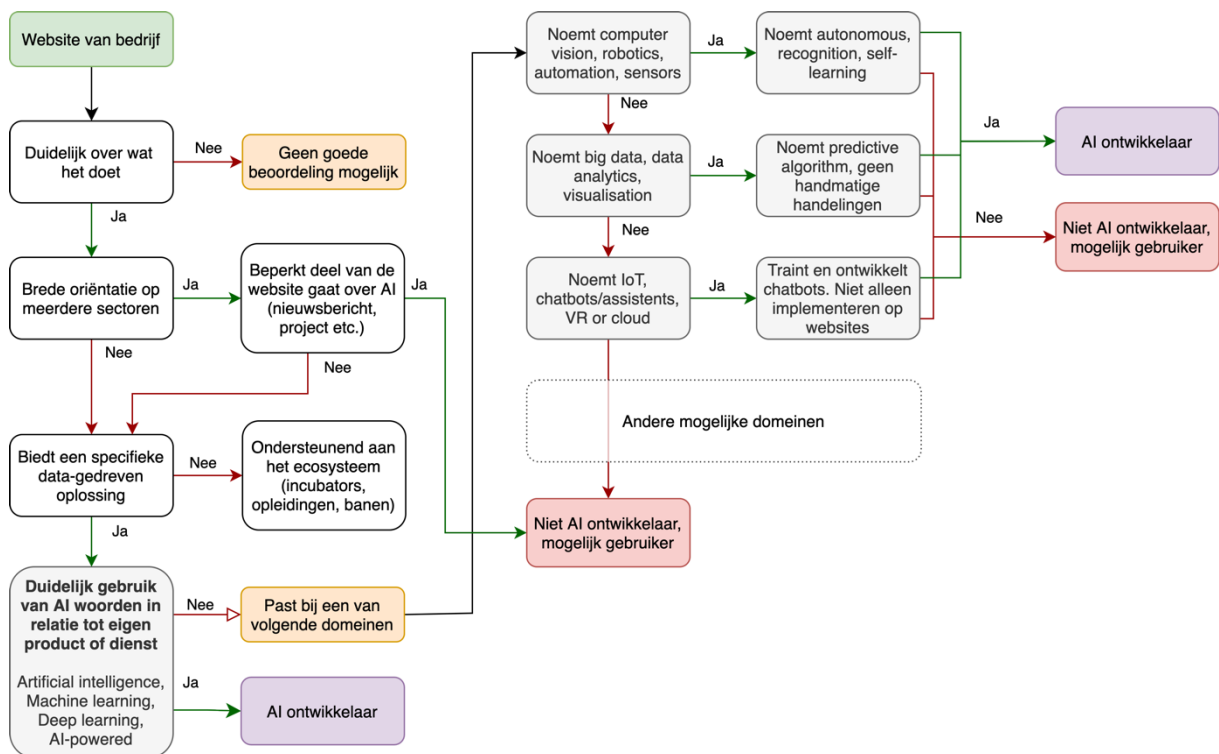
- Ontwikkelaar ja/nee
- Toepasser ja/nee
- Behoort tot het ecosysteem ja/nee
- Past in een trainingsset ja/nee

Het document is opgesteld door Innovatiespotter, daarna beoordeeld door en besproken met 2 stagiaires van het CBS en vervolgens door een AI-expert van RVO beoordeeld. Deze expert beoordeelt AI-gerelateerde WBSO-aanvragen. De conclusie was dat de verschillende beoordelingen vooral voortkwamen uit onduidelijkheid rond informatie die al dan niet op de websites te vinden is, en minder door verschil van inzicht over wat AI en AI-ontwikkeling is. Deze bevinding verklaart mede de matig presterende ML-modellen, omdat de website-inhoud vaak geen of onvoldoende AI-gerelateerde indicatoren bevat.

Op basis van de bevindingen is een stroomschema gemaakt om het beoordelingsproces te vergemakkelijken, zie Figuur 4. Ook draagt een stroomschema bij aan een reproduceerbare aanpak.

Dit experiment heeft eveneens geleid tot het inzicht om voor een volgende stap te experimenteren met een aanpak die uit 2 stappen bestaat:

- Een AI-model die zowel ontwikkelaars als toepassers classificeert
- Daarna met alternatieve technieken (zoals unsupervised clustering) binnen deze selectie onderscheid te maken tussen ontwikkelaar en toepasser.



Figuur 4. Stroomschema om het beoordelingsproces te vereenvoudigen.

BIJLAGE 1 DETAILRESULTATEN PER UITGEVOERDE STAP

STAP 1: DEFINITIE

Er is gezocht naar definities van AI die worden gebruikt door overheid, kennisinstellingen en betrokken partijen uit het bedrijfsleven. Er is geen gangbare definitie die door iedereen wordt gedeeld. Het Strategisch Actieplan voor Artificiële Intelligentie (SAPAI)¹⁴ en ook de Nederlandse Digitaliserings-strategie¹⁵ baseren zich op de definitie van de Europese Commissie¹⁶, die is opgesteld door de High-Level Expert Group on AI (AI HLEG)¹⁷.

“Er is geen algemeen geldige definitie van AI die consistent wordt gebruikt door alle belanghebbenden. Wij gebruiken de omschrijving van AI door de Europese Commissie: “AI verwijst naar systemen die intelligent gedrag vertonen door hun omgeving te analyseren en - met een zekere mate van zelfstandigheid - actie ondernemen om specifieke doelen te bereiken.” (SAPAI, 2019)

De Engelse definitie van de Europese Commissie is:

“Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).” (Europese Commissie, 2018)

De AI HLEG heeft deze definitie uitgebreid om een verduidelijking te brengen in aspecten van AI als wetenschappelijke discipline en als technologie. Deze definitie is ook geschreven voor non-AI experts en voor discussies rondom AI-richtlijnen en beleidsaanbevelingen.

*“Artificial intelligence (AI) systems are **software** (and possibly also **hardware**) systems that, given a complex goal, act in the **physical** or **digital** dimension by **perceiving** their environment through data acquisition, **interpreting** the collected structured or unstructured data, **reasoning** on the knowledge derived from this data and **deciding** the best action(s) to take to achieve the given goal. AI systems can also be designed to learn to **adapt** their behaviour by analysing how the environment is affected by their previous actions.*

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes

¹⁴ Strategisch Actieplan voor Artificiële Intelligentie, 2019, p11

<https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie/Rapport+SAPAI.pdf>

¹⁵ Nederlandse Digitaliseringsstrategie, 2020, p11 [https://www.nederlanddigitaal.nl/nederlandse-](https://www.nederlanddigitaal.nl/nederlandse-digitaliseringsstrategie/documenten/rapporten/2020/06/25/nederlandse-digitaliseringsstrategie-2020)

[digitaliseringsstrategie/documenten/rapporten/2020/06/25/nederlandse-digitaliseringsstrategie-2020](https://www.nederlanddigitaal.nl/nederlandse-digitaliseringsstrategie/documenten/rapporten/2020/06/25/nederlandse-digitaliseringsstrategie-2020)

¹⁶ Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM/2018/237 final. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN>

¹⁷ A definition of AI: main capabilities and scientific disciplines, AI HLEG, April 2019.

http://www.pcci.gr/evapiimages/0101_F483.pdf

planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

(AI HLEG, 2019)

AI-systemen zijn voornamelijk software, en in sommige gevallen ook hardware. Er vinden bepaalde processen plaats (vergaring, verwerking, redenering, beslissing en adaptatie). Volgens de AI HLEG kunnen de meeste AI-technieken en -sub-disciplines hiermee worden ingedeeld.

Daarnaast stelt de AI HLEG dat de huidige ontwikkelingen vooral nog *narrow AI*-systemen betreft. Er is nog geen sprake van *general AI*; systemen die zelfstandig meerdere taken kunnen doen die mensen ook kunnen. AI is daarom in veel gevallen *goal-directed*, het is getraind op een specifieke taak binnen vastgestelde kaders. Binnen die kaders is een mate van autonomie of zelflerend vermogen mogelijk.

STAP 2: OPERATIONALISATIE

De operationalisatie is bedoeld om de definitie in de praktijk te kunnen toepassen. Daarvoor moet worden gekeken naar praktische activiteiten, producten en diensten. De NL AI Coalitie gebruikt bijvoorbeeld geen strikte definitie van AI, maar richt zich op specifieke, voor Nederland relevante sectoren. "*Artificial Intelligence (AI) is een generieke technologie die uiteindelijk haar toepassing vindt in alle bedrijfssectoren en maatschappelijke uitdagingen.*"¹⁸ Voor elk aangewezen toepassingsgebied is een werkgroep opgericht.

- Cultuur
- Defensie
- Energie en Duurzaamheid
- Financiële Dienstverlening
- Gezondheid en Zorg
- Haven en Maritiem
- Landbouw en Voeding
- Mobiliteit, Transport en Logistiek
- Onderwijs
- Publieke Diensten
- Technische Industrie
- Veiligheid, Vrede en Recht
- Gebouwde omgeving

Ook de AI Index Report van de Stanford University stelt dat AI een technologie is die breed wordt toegepast door bedrijven, "*AI appears to be transforming into a general purpose technology (GPT)*"¹⁹. Er worden voorbeelden genoemd in sectoren als vastgoed, game-industrie, financiële sector, gezondheidszorg of beveiliging. De OECD stelt ook dat het monitoren van ontwikkelingen in AI lastig is: "*Measuring the development of AI is challenging as the boundaries between AI and other innovations are blurred and change over time.*"²⁰

De Joint Research Center van de Europese Commissie heeft op basis van de AI HLEG en ruim 50 andere definities een taxonomie gemaakt in AI-domeinen²¹. Naast politieke en wetenschappelijke aspecten is ook gezocht naar definities vanuit het perspectief van het bedrijfsleven. De domeinen moeten niet worden gezien als losstaande categorieën. Binnen AI is veel overlap en kruisbestuiving.

¹⁸ De Nederlandse AI Coalitie, toepassingsgebieden <https://nlaic.com/toepassingsgebied/>

¹⁹ Stanford University (2019), "The AI Index 2019 Annual Report", p.91 <https://hai.stanford.edu/research/ai-index-2019>

²⁰ Paragraaf: *How can AI-related developments in science and technology be tracked?* OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264311992-en>

²¹ Samoil, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., and Delipetrev, B., *AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence*, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163.

De AI-domeinen worden opgedeeld in Core en Transveral, wat kan worden opgevat als horizontale en verticale gebieden. De kern van AI bestaat uit reasoning, planning, learning, communication en perception. Waaronder bekende technologieën vallen als natural language processing, computer vision, spraakherkenning.

In de horizontale domeinen komen veel van deze core-domeinen samen. Een zelfrijdende auto heeft perception, reasoning, planning en learning allemaal nodig om zelfstandig in het verkeer te navigeren.

AI services is een ruim begrip van cognitive computing, machine learning, bots en virtual assistants. Het wordt vaak aangeboden als system-as-a-service en maakt gebruik van de cloud.

AI taxonomy		
	AI domain	AI subdomain
Core	Reasoning	Knowledge representation
		Automated reasoning
		Common sense reasoning
	Planning	Planning and scheduling
		Searching
		Optimisation
	Learning	Machine learning
	Communication	Natural language processing
	Perception	Computer vision
		Audio processing
Transversal	Integration and interaction	Multi-agent systems
		Robotics and Automation
		Connected and Automated vehicles
	Services	AI Services
	Ethics and Philosophy	AI Ethics
		Philosophy of AI

Tabel 5. AI-domeinen en sub-domeinen. Overgenomen van Joint Research Center, AI Watch (2020, p11).

Omdat de Joint Research Center zoveel verschillende onderzoeken heeft samengebracht is dit een erg bruikbare taxonomie van sub-domeinen. Twee veelgebruikte indelingen binnen het Nederlandse AI-landschap sluiten hier goed op aan. De Special Interest Group on AI (SIGAI)²² is een Nederlandse groep van AI-onderzoekers. De indeling die zij maken is vaak terug te zien in analyses door kennisinstellingen of regionale hubs.

SIGAI onderscheidt 7 onderzoeksgebieden die in Nederland sterk vertegenwoordigd zijn.

- Autonomous agents & robotics
- Computer vision
- Decision making
- Information retrieval
- Knowledge representation & reasoning
- Machine learning
- Natural language processing

In de *Community survey on ICT usage and e-commerce in enterprises* (2021) wordt onderscheid gemaakt tussen *software-based* en *embedded* AI-systemen. AI wordt verdeeld in 7 technologieën.

- Text mining
- Speech recognition
- Natural language generation
- Image recognition, image processing
- Machine learning, deep learning
- AI-based software robotic process automation
- Autonomous robots, self-driving vehicles, autonomous drones

²² SAGAI, *Dutch Artificial Intelligence Manifesto*, 2019, <http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf>

Ook wordt in de Community survey onderscheidt gemaakt tussen 7 toepassingsgebieden:

- Marketing or sales
- Production processes
- Organisation of business administration processes
- Management of enterprises
- Logistics
- ICT security
- Human resources management or recruiting

STAP 3: PERSONA'S

Aan de hand van de eigenschappen in Tabel 1 zijn 12 combinaties mogelijk (software/hardware, fysiek/virtueel, consument / zakelijk / industrie). Dit hoeft niet te betekenen dat bij alle 12 combinaties passende bedrijven bestaan. Zie BIJLAGE 3 PERSONAS voor uitgewerkte voorbeelden. Ook zijn persona's opgesteld van archetypen die mogelijk ongewenst zijn, of lastig kunnen worden geïdentificeerd.

Software voor de zakelijke markt

Slimmer.ai en Blitz beschouwen we als goede voorbeelden voor partijen die software ontwikkelen voor een virtuele omgeving en met een zakelijke toepassing. BodyGraph maakt software voor de zakelijke markt, maar werkt in de fysieke omgeving.

Industriële toepassingen

Accerion en Cygnify worden beschouwd als goede voorbeelden voor toepassingen voor de industrie, waarbij Accerion zich richt op hardware (autonome robots) en Cygnify software ontwikkelt voor o.a. gezichtsherkenning.

Moeilijk te bepalen

Er zijn ook voorbeelden waarbij het moeilijk te identificeren is of er AI-producten en diensten worden ontwikkeld. Voorbeelden daarvan zijn HOOOOP en Sleep.ai.

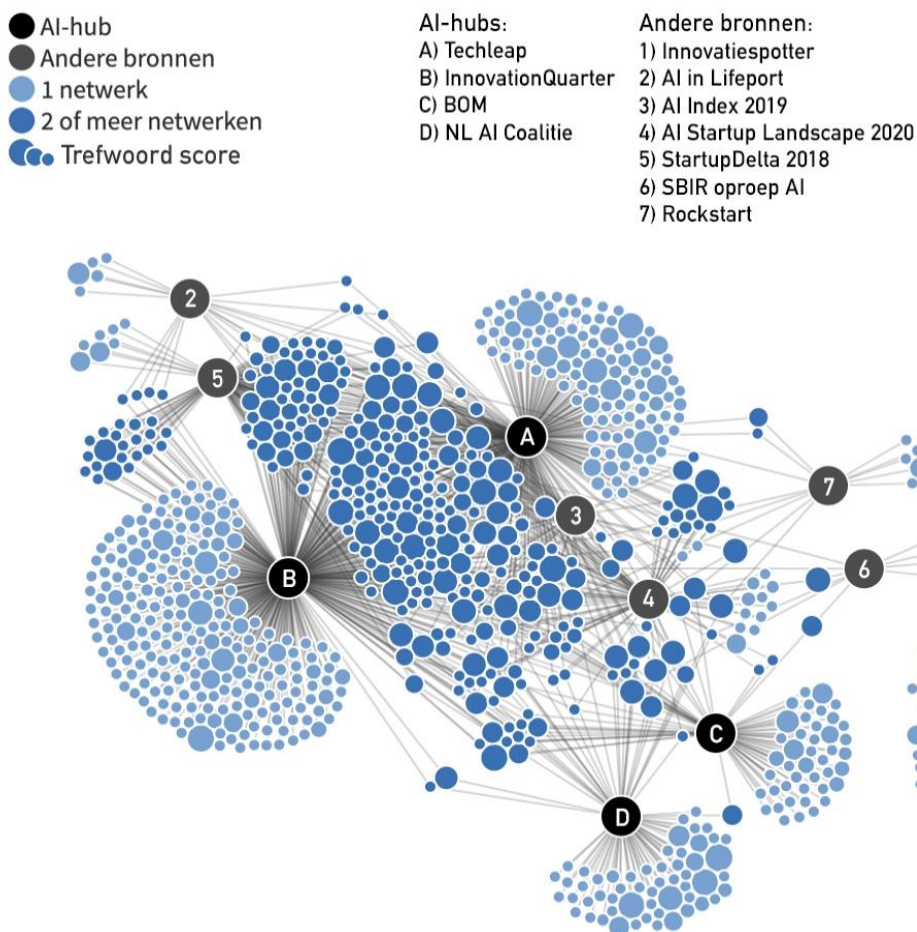
Bij nader inzien toch geen AI

Er zijn ook voorbeelden waarbij de terminologie overeen lijkt te komen met die van AI, maar waarbij een nadere beschouwing leert dat deze bedrijven niet mee gerekend zouden moeten worden. InnerQI bijvoorbeeld gebruikt NLP op de website in de context van neurolinguïstisch programmeren, en niet in de context van natural language processing. AI komt voor ook weleens op de website voor in de context van artificial insemination (of kunstmatige inseminatie) zoals bij ai-total en niet in de context van artificial intelligence of kunstmatige intelligentie. Er zijn bedrijven die publiceren over AI, maar het niet zelf ontwikkelen, zoals Analytics in HR en Talent Factor Consultancy. En er zijn voorbeelden van bedrijven die wel AI toepassen, maar het niet zelf ontwikkelen zoals PicNic en KOAT.

STAP 4. SAMENSTELLEN TRAININGSSET

Omdat vooraf aan dit pilotproject geen overzicht van Nederlandse AI-startups en bedrijven beschikbaar is moeten bedrijven worden geselecteerd. Voor de initiële trainingsset zijn bedrijven samengesteld op basis van hun deelname aan de verschillende AI-netwerken in Nederland. De NL AI Coalitie en alle AI-hubs zijn benaderd voor ledenlijsten of lijsten andere bedrijven die zij in beeld hebben. Ook is gekeken hoe de websites van deze bedrijven scoren op beschikbare trefwoordlijsten (zie BIJLAGE 5 Trefwoordenlijst en BIJLAGE 6 Trefwoordenlijst JRC). Dit leverde een lijst op van 951 bedrijven (waarvan 782 afkomstig zijn uit de netwerken, en 169 zijn geselecteerd op basis van trefwoorden).

In Figuur 5 en Tabel 6 zijn de relaties te zien tussen de netwerken en bedrijven. Elk bedrijf is een blauwe stip, en de lijnen laten zien uit welke netwerken het afkomstig is. Een donkere kleur betekent dat het bedrijf in meerdere netwerken voorkomt. Een grotere punt betekent dat de website van het bedrijf beter scoort op een lijst met trefwoorden (zie BIJLAGE 5 Trefwoordenlijst).



Figuur 5 Relaties tussen de netwerken en bedrijven uit de trainingsset.

Tabel 6 Aantal bedrijven per netwerk.

Netwerk	Aantal bedrijven
InnovationQuarter	485
Techleap	361
Innovatiespotter*	169
StartupDelta	116
AI Startup Landscape	87
NL AI Coalitie	84
BOM	79
AI Index 2019	38
AI in Lifeport	20
Rockstart AI	13
SBIR AI	10
Totaal	951

*Gevonden op basis van trefwoorden en handmatig beoordeeld.

Het grootste aantal bedrijven is afkomstig uit de netwerken van InnovationQuarter en Techleap, of beoordeeld op basis van trefwoorden door Innovatiespotter. Een derde van de bedrijven komt in 2 of meer netwerken voor (40% als de groep van Innovatiespotter niet wordt meegerekend). Veel van deze bedrijven hebben ook een hoge trefwoordscore.

Relevantie op trefwoorden

Er is gekeken hoe goed de websites scoren op relevante trefwoorden. Hiervoor is een lijst gebruikt van RVO (Inzet op sleuteltechnologieën vanuit de WBSO, sept 2019), zie BIJLAGE 5 Trefwoordenlijst. Te specifieke of algemene trefwoorden zijn weggelaten, en de lijst is aangevuld met enkele trefwoorden van Innovatiespotter. Er is ook gekeken met een trefwoordenlijst van het Joint Research Center (AI Watch: Defining Artificial Intelligence, 2020), zie BIJLAGE 6 Trefwoordenlijst JRC.

De gebruikte trefwoordenlijsten zijn niet opgesteld met het doel om websiteteksten van bedrijven te doorzoeken. De trefwoordenlijst van RVO wordt bijvoorbeeld ingezet voor projectbeschrijvingen van WBSO-aanvragen. In tegenstelling tot een websitetekst bevat dit veelal technisch taalgebruik. De doelgroepen van een website zijn consumenten of de zakelijke markt. Daarom vermelden zij vaker alleen de oplossingen of toepassingen van AI, en niet de achterliggende techniek.

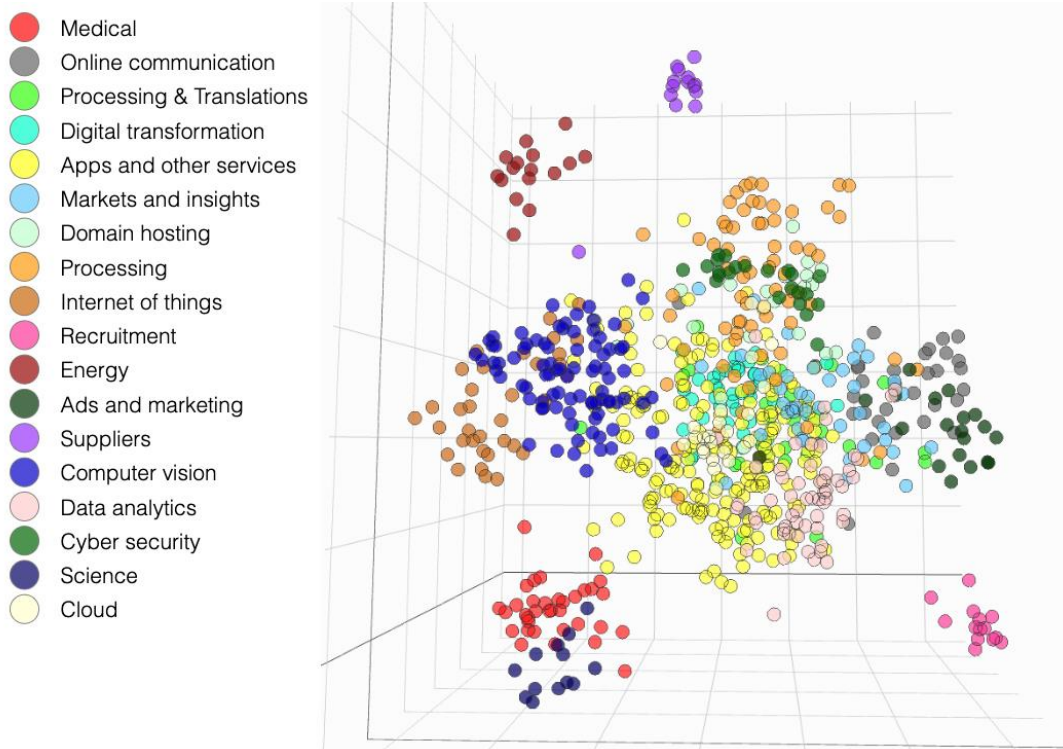
De trefwoordscore is een getal tussen de 0 en 1 (2 decimalen) en wordt berekend op basis van het aantal verschillende trefwoorden die op een website worden vermeld en de frequentie hiervan. Een score van 1 komt maar bij enkele bedrijven voor, en de score loopt daarna relatief snel af. Een gemiddelde score van 0.2-0.3 hoeft daarom niet slecht te zijn. Ook als rekening wordt gehouden met het oorspronkelijke doel van deze trefwoordenlijsten. Veel van de bedrijven hebben een hoge trefwoordscore. Echter zijn er ook bedrijven met een lagere score, of niet op trefwoorden scoren.

Tabel 7. Gemiddelde trefwoordscore per netwerk.

Netwerk	Aantal bedrijven	Gem. trefwoordscore RVO	Gem. trefwoordscore JRC
InnovationQuarter	485	0.19	0.15
Techleap	361	0.25	0.18
Innovatiespotter*	169	0.40	0.34
StartupDelta	116	0.26	0.19
AI Startup Landscape	87	0.33	0.21
NL AI Coalitie	84	0.27	0.18
BOM	79	0.23	0.21
AI Index 2019	38	0.37	0.19
AI in Lifeport	20	0.25	0.21
Rockstart AI	13	0.28	0.16
SBIR AI	10	0.35	0.31
Totaal	951	0.22	0.18

Clusters binnen de trainingsset

Uit de operationalisatie is gebleken dat er veel AI-domeinen zijn, en dat voor veel verschillende sectoren AI-oplossingen worden ontwikkeld. Met unsupervised machine learning is gekeken wat voor groepen AI-bedrijven in de trainingsset aanwezig zijn. “Unsupervised” betekent dat een algoritme verbanden legt in de data, zonder dat daar menselijke input voor nodig is. Er ontstaan clusters van bedrijven die gemeenschappelijke woorden gebruiken. Aan de hand van deze woorden is gekeken naar een geschikte naam voor deze clusters.



Figuur 6. Clusters binnen AI trainingsset. Met unsupervised machine learning worden clusters gemaakt zonder dat daar menselijke input voor nodig is. Elk bedrijf is een punt in de driedimensionale wolk.

Figuur 6 toont het resultaat van dit experiment. Dit laat zien dat er veel verschillende onderwerpen zijn in de populatie van AI-bedrijven. Dit is vergelijkbaar met internationale onderzoeken die zijn beschreven hoofdstuk 3.4. Een samenvatting van deze onderzoeken is te vinden in BIJLAGE 9 Internationale context.

De clusters Medical, Science, Energy, Suppliers en Recruitment staan verder af van de “wolk” in het midden. Ook is er een onderscheid te zien tussen de Computer vision en Internet of things toepassingsgebieden (links) en de meer online/software gerichte gebieden rechts.

Unsupervised machine learning is eerst gebruikt om binnen de trainingsset verbanden te vinden. Daarna is geëxperimenteerd met grotere populaties, zoals ICT en High Tech. Binnen deze 2 gebieden zijn de meeste AI-bedrijven te verwachten. Dit leverde beter inzicht in het landschap op. AI werd in deze experimenten op grotere populaties door het algoritme niet per se als losstaand cluster gezien. Wel leverde het doorzoeken van “ICT” of “Data” gerelateerde clusters steeds kleine aantallen nieuwe AI-gerelateerde bedrijven op.

STAP 5. TRAINEN EN TESTEN VAN MODEL

Aanpak

Opschonen van de data

Voor een model getraind kan worden dient eerst de data te worden opgeschoond. Bij het opschonen worden de volgende acties uitgevoerd:

- Verwijderen van datums, weekdays, maanden, tijden etc. Aangezien deze waarden vaak in verband stonden met het moment van scrapen en zijn daardoor ook niet wenselijk.

- Verwijdering van interpunctie: Voor het doel van dit project heeft interpunctie geen invloed op de voorspellingen, vandaar dat dit ook wordt verwijderd.
- Stop word removal (NL en EN): Veel voorkomende woorden worden verwijderd (e.g. ik, jij, van, de, een, etc.). Aangezien deze woorden over het algemeen genomen weinig informatie dragen, immers als ze overal voorkomen, zullen ze ook niet helpen met een voorspelling.
- Tekst naar lowercase: Een 'a' en een 'A' zijn voor een computer niet hetzelfde, daarom trekken we dit op een simpele manier gelijk.
- Verwijder getallen uit de tekst: Over het algemeen zit er weinig voorspellende waarde in getallen als het gaat om voorspellingen op basis van tekst en wordt dit meer als ruis gezien.
- Stemming/Lemmatization: Vervoegingen van woorden, bijvoorbeeld: lopen, loop en loopt, duiden allemaal dezelfde actie aan, maar worden anders geschreven, waardoor NLP-technieken deze mogelijk niet als hetzelfde zien. Vandaar dat vervoegingen weggehaald worden. In dit geval is stemming gebruikt als de techniek om dit te doen.

Training

Tijdens de trainingsfase zijn verschillende modellen getest. Hiervoor is Python als software gebruikt en met name de pakketten "scikit-learn" en "fasttext".

Fasttext

Fasttext-modellen hebben als bijzonderheid dat ze een mix tussen combinaties van woorden en combinaties van letters (bijvoorbeeld: lopen kan je opdelen in "lop", "ope", "pen") gebruiken. Het nadeel van dit soort model is dat het lastig(er) is om uit te leggen waar het algoritme precies de beslissing op baseert.

Scikit

In het Scikit-pakket zijn verschillende modellen beschikbaar zoals Logistische regressie, Naive Bayes, Support Vector Machines, Decision Trees, Random Forest en AdaBoost. Zie de begrippenlijst in de bijlage voor meer uitleg. Uiteindelijk is gekozen voor Logistische regressie (LR) vanwege de relatief goede performance en de simpele verklaarbaarheid van de uitkomsten. Bij een LR model krijgen woorden een gewicht toegekend. Op basis van een combinatie van de woorden die voorkomen in een tekst en het gewicht van deze woorden wordt vervolgens besloten of een tekst bij een AI-bedrijf behoort of niet.

Resultaten

1. Input

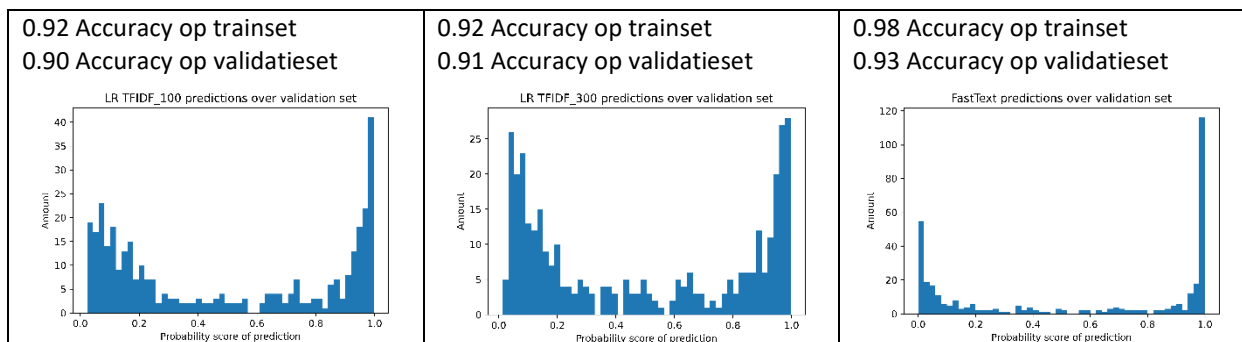
Innovatiespotter heeft 951 bedrijven met een positieve classificatie geïdentificeerd op basis van input uit de netwerken, aangevuld met eigen classificaties. Deze lijst is aangevuld met (random) negatieve bedrijven door het CBS. In totaal zijn er 1.779 unieke URL's gebruikt die gescraped konden worden: 80% voor trainingsdoeleinden, 20% voor validatie.

2. Trainen

Er zijn 3 verschillende modellen getraind en getest. Deze modellen zijn alleen getraind op teksten van de homepage van de bedrijven. De resultaten staan in Tabel 8. De figuren laten zien dat er pieken optreden bij bedrijven die duidelijk als AI worden herkend (een hoge score) en bij bedrijven die duidelijk als niet-AI worden aangemerkt (lage score).

Tabel 8 Verdeling van bedrijven uit het de training-/testset per probability score per model

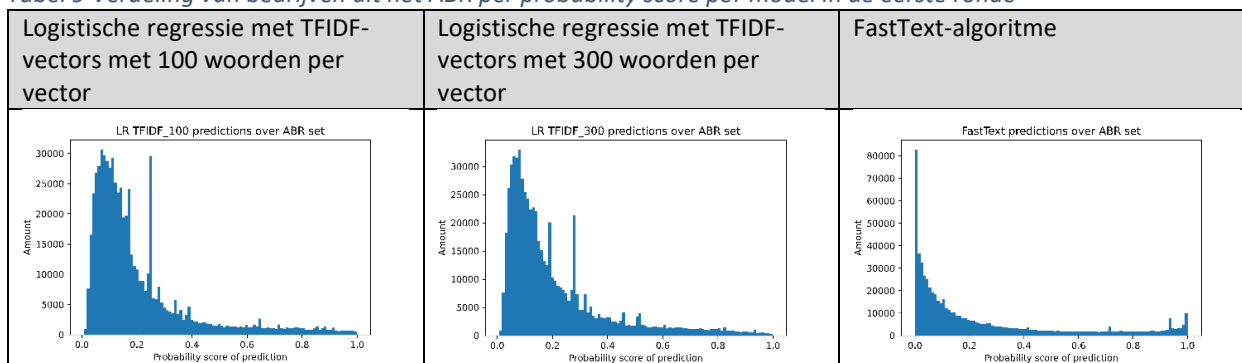
Logistische regressie met TFIDF-vectors met 100 woorden per vector	Logistische regressie met TFIDF-vectors met 300 woorden per vector	FastText-algoritme
Performance	Performance	Performance



3. Output

Het getrainde model is vervolgens toegepast op de 600.000+ bedrijven uit het ABR. De uitkomsten hiervan staan in Tabel 9 en Tabel 10.

Tabel 9 Verdeling van bedrijven uit het ABR per probability score per model in de eerste ronde



Tabel 10 Aantal AI bedrijven per confidence interval

Confidence interval	Logistische regressie met TFIDF-vectors met 100 woorden per vector	Logistische regressie met TFIDF-vectors met 300 woorden per vector	FastText-algoritme
0.5-1.0	57.322	61.194	120.137
0.55-1.0	50.138	48.525	109.757
0.6-1.0	43.217	40.986	100.295
0.65-1.0	34.858	33.415	91.016
0.7-1.0	29.063	26.439	82.136
0.75-1.0	22.974	20.476	70.995
0.8-1.0	17.165	14.935	61.858
0.85-1.0	12.455	9.414	53.068
0.9-1.0	7.081	5.652	42.799
0.95-1.0	3.128	2.183	24.367

4. Controle en evaluatie

De uitkomsten in Tabel 9 en Tabel 10 vinden we niet realistisch. Normaliter zou er bij 0.5 een standaard ‘afkap’ worden gehanteerd: alles met een score > 0.5 behoort dan tot AI, alles eronder niet tot AI. Met deze uitkomsten zou dat betekenen dat er ergens tussen de 57.322 en 120.137 bedrijven zijn die AI-producten en -diensten ontwikkelen. Dat is wel erg onwaarschijnlijk. Er lijkt er een enorme overschatting plaats te vinden. Die overschatting lijkt te maken te hebben met de scheve samenstelling van de populatie Nederlandse bedrijven; de populatie AI-bedrijven is erg klein vergeleken met het totaal aantal bedrijven. Er zijn manieren om hiervoor te kunnen corrigeren als dit probleem optreedt. In de volgende fase (validatie) is de uitdaging de verschillende

uitkomsten van de modellen nader te verklaren, te corrigeren voor de uitkomsten en te bepalen wat een goed ‘afkap’-punt zou moeten zijn. Daarvoor zullen de door het model gevonden bedrijven ook handmatig worden gevalideerd.

STAP 6 VALIDATIE

Eerste validatieronde

Aanpak

Uit de resultaten van fase 2 is een willekeurige steekproef van bedrijven geselecteerd om handmatig te valideren. Deze bedrijven zijn met openbare data geselecteerd. Voor de validatie zijn de uitkomsten van het tweede model gebruikt: logistische regressie met TFIDF-vectors met 300 woorden per vector.

Voor de validatie is gebruik gemaakt van Q-Nection, een product van Innovatiespotter (ISP)²³. Hiermee wordt alle relevante informatie van een bedrijf getoond om tot een goede afweging te komen en kunnen labels en classificaties worden toegevoegd.

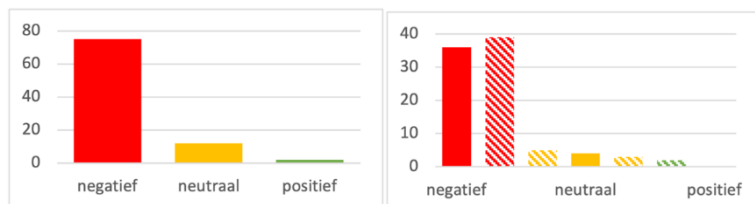
Positieve en negatieve classificaties spreken voor zich. Handmatig positief beoordeelde bedrijven kunnen worden toegevoegd aan de positieve trainingsset en negatief beoordeelde bedrijven aan de negatieve trainingsset. Bedrijven kunnen ook als ‘neutraal’ worden beoordeeld als ze wel iets met AI doen, maar ze niet als ontwikkelaar worden gezien. Deze bedrijven nemen we vervolgens ook mee in de negatieve trainingsset.

Tevens is het mogelijk de optie “second opinion” aan te vinken. Dit geeft de beoordelaar de mogelijkheid om bij twijfelgevallen deze ook door anderen te laten beoordelen.

De validatie is uitgevoerd door CBS, ISP en RVO. Sommige bedrijven zijn individueel beoordeeld en een groep van 100 bedrijven is door iedereen beoordeeld om onderlinge vergelijkbaarheid mogelijk te maken.

Resultaten

- De eerste validatieronde werd voortijdig beëindigd, want vanuit de meeste bedrijven werden overwegend negatief (niet-AI) beoordeeld.
- Er zijn slechts 2 bedrijven in de steekproef die door de meerderheid als AI werden gezien.
- Er zijn 12 bedrijven die overwegend neutraal werden beoordeeld, of waar geen doorslaggevende beoordeling is.



Figuur 7 Resultaten eerste validatie: 75 negatief, 12 neutraal, 2 positief (effen = 100% overeenstemming, gearceerd is merendeel overeenstemming)

²³ Qnection.innovatiespotter.nl

Inhoudelijk viel op dat er overwegend ICT-bedrijven positief geïclassificeerd werden door het model, terwijl de SBI geen feature in het model is. Binnen die groep bedrijven blijkt het bij een handmatige controle lastig vast te stellen: betreft het alleen ICT, Business Intelligence of zijn de producten en diensten Artificial Intelligence? Er zaten ook een paar vreemde gevallen tussen zoals groothandels en massagesalons. Al met al liet de feedback uit de eerste ronde zien dat deze versie van het model onvoldoende presteert.

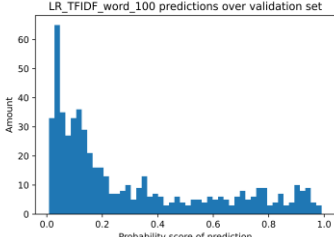
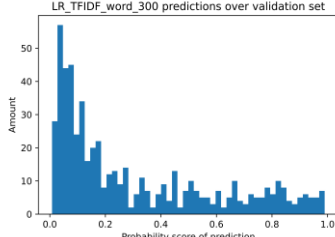
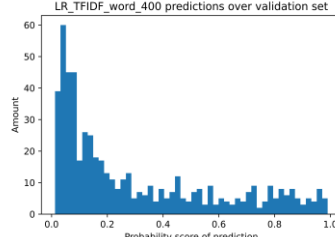
Tweede validatieronde

Aanpak

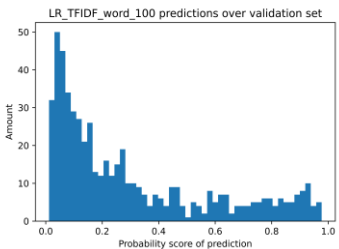
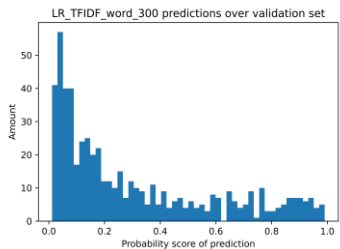
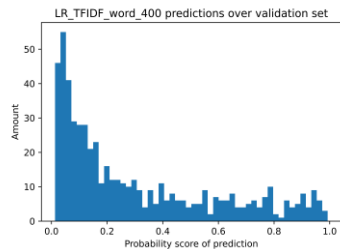
Voor de tweede validatieronde is de trainingsset allereerst uitgebreid met extra cases (positief en negatief), deels op basis van de geïnoteerde websites uit de vorige ronde. De bedrijven die daarbij als ‘tweifel’ werd aangemerkt zijn toegevoegd als niet-AI. De verdeling positief-negatief is aangepast van 50-50 naar 33-66. In totaal bestond de trainingsset daarmee uit 2.794 websites, waarvan 898 AI.

In de tweede ronde is FastText achterwege gelaten en zijn er drie modellen verder getest. De modellen zijn eerst alleen getraind op de homepage van de websites en daarna op de gehele website. Na hertrainen van het model daalde het aantal positief geïclassificeerde bedrijven (met probability >0,5) van 57.322 naar 8.442 bedrijven voor het model dat op de voorpagina is getraind. Het aantal positief geïclassificeerde bedrijven daalde naar 8.802 bedrijven voor het model dat op de gehele website is getraind. Het model gebaseerd op 400 woorden hield maar 5.246 bedrijven over die als positief werden geïclassificeerd. Bij het verhogen van het aantal woorden in de vectoren van 100 naar 300 en 400 woorden was ook een duidelijke stijging te zien in de nauwkeurigheid van de getrainde modellen. Dit is onder andere te zien in Tabel 11 en Tabel 12, waar de nauwkeurigheid op de validatieset steeg met minimaal 4% punten tot maximaal 10% punten. Daarom zijn in de volgende fases 400 woorden gebruikt in plaats van 100 woorden zoals in de eerste validatie ronde.

Tabel 11 Verdeling van bedrijven uit het ABR per probability score per model in de tweede ronde, gebaseerd op de homepage.

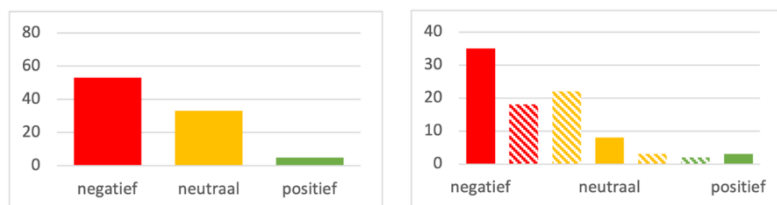
Logistische regressie met TFIDF-vectors met 100 woorden per vector	Logistische regressie met TFIDF-vectors met 300 woorden per vector	Logistische regressie met TFIDF-vectors met 400 woorden per vector
Performance 0 85 Accuracy op trainset 0 81 Accuracy op validatieset	Performance 0 86 Accuracy op trainset 0 84 Accuracy op validatieset	Performance 0 87 Accuracy op trainset 0 84 Accuracy op validatieset
		

Tabel 12 Verdeling van bedrijven uit het ABR per probability score per model in de tweede ronde, gebaseerd op de volledige website.

Logistische regressie met TFIDF-vectors met 100 woorden per vector	Logistische regressie met TFIDF-vectors met 300 woorden per vector	Logistische regressie met TFIDF-vectors met 400 woorden per vector
Performance 0 83 Accuracy op trainset 0 74 Accuracy op validatieset	Performance 0 86 Accuracy op trainset 0 84 Accuracy op validatieset	Performance 0 87 Accuracy op trainset 0 84 Accuracy op validatieset
		

Resultaten

- Op basis van de resultaten van het 400-woorden-model op de gehele websites is er een groep van 100 bedrijven door meerdere mensen beoordeeld. Er zijn een paar afgevallen omdat de website offline is, of niet bij het bedrijf hoort. Er zijn er 91 overgebleven.
- De meeste bedrijven worden door alle 5 beoordelaars als negatief beoordeeld. In tegenstelling tot de eerste steekproef worden er meer neutrale beoordelingen gegeven.



Figuur 8 Resultaten tweede validatie: 53 negatief, 33 neutraal, 5 positief

Inhoudelijk viel op dat er in de tweede validatieronden wederom minder overduidelijk foute resultaten waren en veel meer twijfelgevallen. Al met al werden er 2,5x zoveel 'goede' bedrijven gevonden als in de eerste ronde. Ook hier zien we wel dat het lastig blijft om te beoordelen of een bedrijf een echt AI-bedrijf is volgens de eerder vastgestelde operationalisatie. De twijfel zit vooral bij bedrijven waarbij de AI-producten of -diensten geen core business lijken te zijn of waarbij er wordt geïmpliceerd dat er AI wordt gebruikt, maar de manier waarop niet duidelijk is. Het meest subtiele verschil is zichtbaar bij websites met BI-achtige applicaties, die wel data verwerken en inzicht bieden, maar het probleem op laten lossen door mensen in plaats van algoritmes.

Het ontwikkelde model blijkt na twee iteraties een stuk beter te presteren, in elk geval als eerste selectie. Bij nadere bestudering blijkt een klein deel (~6%) van deze selectie als 100% AI te worden aangemerkt en een iets groter deel (~36%) als twijfelgeval. Dit is niet wat we hadden gehoopt, maar ook niet heel slecht gegeven de context dat de overgrote meerderheid (>99%) van de bijna 600.000 bekeken websites geen AI betreft.

HERHALING STAPPEN 5 EN 6

Op basis van de resultaten van fase 2 en 3 is nog een laatste iteratie uitgevoerd om het model verder te proberen te optimaliseren. Hierbij is verder gewerkt met het model dat gebruik maakt van 400 woorden.

Aanpak

Extra niet-AI voorbeelden

Er zijn in deze ronde nog meer niet-AI bedrijven toegevoegd aan de trainingsset om een grotere set (met grotere spreiding) aan negatieve voorbeelden te hebben; in deze set zaten ook iets meer IT-bedrijven. De uiteindelijke trainingsset bevat nu 3.895 niet-AI bedrijven en 898 AI-bedrijven.

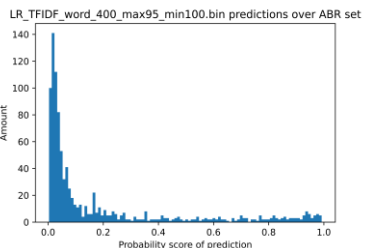
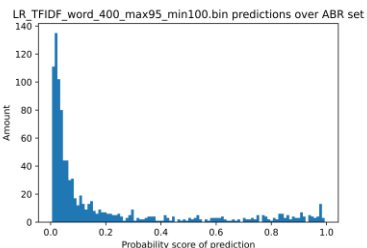
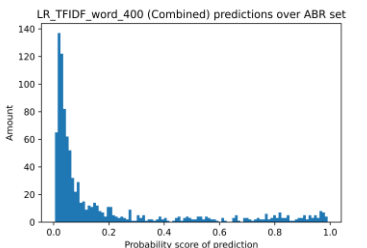
Engels of Nederlands

Er is bepaald of het zinvol is om twee modellen te hanteren, afhankelijk van de taal. Hieruit bleek dat een Engels model beter werkt voor Engelstalige websites, maar dat een Nederlands model daarentegen weer slechter presteert voor Nederlandstalige websites. In het merendeel van de gevallen is de website Nederlands en levert het initieel gebruikte gemengde model voor Engelstalige en Nederlandstalige websites over het geheel genomen het beste resultaat op.

Volledige website of alleen de homepage

Er is bepaald of het zinvol is het model alleen te trainen op basis van de hoofdpagina, of om de volledige website mee te nemen. Daaruit bleek dat het model dat de volledige website meeneemt vooral goed lijkt te presteren op grote bedrijven en minder goed op kleine bedrijven. Het model dat alleen de hoofdpagina meeneemt presteert daarentegen beter op de kleinere start-ups. Mogelijk dat start-ups dit ook groter op hun hoofdpagina aankaarten en er lijkt ook wel een trend dat nieuwe bedrijfjes enkel één grote hoofdpagina hanteren. De uitkomst voor ons onderzoek was dat het gemiddelde van de score van beide soorten modellen (homepage en volledige website) de beste resultaten opleverde. Deze combinatie wordt in de rest van het rapport het gecombineerde model genoemd. Dit gecombineerde model leverde 3.019 positief geïdentificeerde bedrijven (met probability >0.5) op.

Tabel 13 Verdeling van bedrijven uit het ABR per probability score per model in de derde ronde, gebaseerd op de homepage, de volledige website en het gemiddelde van beide scores.

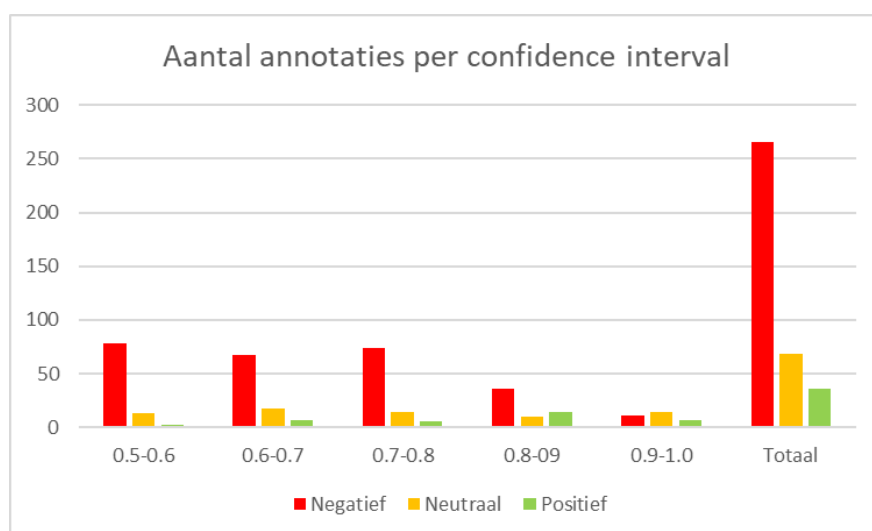
Logistische regressie met TFIDF, 400 woorden (gehele website)	Logistische regressie met TFIDF, 400 woorden (alleen voorpagina)	Logistische regressie met TFIDF, 400 woorden (gecombineerd)
Performance 0 91 Accuracy op trainset 0 91 Accuracy op validatieset 	Performance 0 91 Accuracy op trainset 0 91 Accuracy op validatieset 	Performance 0 92 Accuracy op trainset 0 92 Accuracy op validatieset 

Resultaten

Er is nog een laatste evaluatie uitgevoerd met een gestratificeerde steekproef op basis van de confidence intervallen van de classifier, vanaf 0.5 in stappen van 10% (bijv 0.5-0.6, 0.6-0.7, enz.). In principe wilde we 100 bedrijven per interval selecteren, maar niet ieder interval had genoeg bedrijven.

Tabel 14 Aantal gevalideerde bedrijven per confidence interval

Confidence interval	Niet-AI		Twijfel		AI		Totaal
	#	%	#	%	#	%	
0.5-0.6	78	84%	13	14%	2	2%	93
0.6-0.7	67	74%	17	19%	7	8%	91
0.7-0.8	74	79%	14	15%	6	6%	94
0.8-0.9	36	60%	10	17%	14	23%	60
0.9-1.0	11	34%	14	44%	7	22%	32
Totaal	266	72%	68	18%	36	10%	370



Figuur 9 Aantal annotaties (niet-AI, neutraal, AI) per confidence interval en in totaal

Uit de validatie wordt duidelijk dat er geen goed cut-off point is vast te stellen. Zowel in het 0.5-0.6-interval als in het 0.9-1.0-interval worden niet-AI bedrijven gevonden. De verhoudingen tussen niet-AI en wel-AI worden wel wat beter in de hogere intervallen. Het interval 0.8-0.9 valt het meeste op, omdat hier meer AI- dan neutrale beoordelingen tussen zitten. Het gaat maar om kleine aantallen, dus het is lastig om hier conclusies aan te verbinden. Wat ook opvalt (niet zichtbaar in deze grafiek) is dat de verhouding tussen positieve en negatieve validaties verschilt per persoon die de validatie heeft uitgevoerd.

In het eerder aangehaalde rapport van de OECD²⁴ wordt vastgesteld dat het meten van AI lastig is omdat de scheiding tussen AI en andere innovaties onscherp is en voortdurend wijzigt. Het resultaat van ons onderzoek lijkt dit te onderschrijven. Op basis van deze resultaten is geen betrouwbare schatting af te geven van het aantal bedrijven dat AI-producten en -diensten ontwikkelt.

²⁴ Paragraaf: How can AI-related developments in science and technology be tracked? OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264311992-en>

BIJLAGE 2 VERGELIJKING MET AI-PROJECTEN IN DE WBSO NADER BEKEKEN

In de tabellen hieronder wordt de overlap van de twee populaties nader bekeken. Er wordt gekeken aan de hand van sectorindeling (SBI-code) en werknemerscategorieën.

Wat betreft SBI-codes en leeftijdsopbouw komen de populaties aardig overeen. Specifiek valt het volgende op:

1. De meeste bedrijven bevinden zich in sector 62-Dienstverlenend informatietechnologie, gevolgd door sector 64-Financiële instellingen. Uit de lijst van Innovatiespotter omvatten deze sectoren respectievelijk 39% en 15% van de bedrijven. Bij de WBSO is dit respectievelijk 55% en 9%. Deze sectoren zijn dus kenmerkend voor zowel AI-bedrijven van Innovatiespotter als WBSO-bedrijven met een AI-project. 58% van de totale populatie die overlap is gevonden bij bedrijven actief in sector 62-Dienstverlenend informatietechnologie.
2. De grootste overlap (in %) zit in sector 72-Speur- en ontwikkelingswerk; 24% van alle bedrijven in deze sector uit de lijst van Innovatiespotter komt ook in de WBSO-set voor. Vanuit de WBSO-populatie gezien, komt 41% in deze sector ook voor in de lijst van Innovatiespotter.
3. Er is relatief veel overlap gevonden bij bedrijven met in de werknemerscategorie 10-19 personen; 38% van de bedrijven in de innovatiespotter-set komt ook bij de WBSO voor. Bij de categorie 'onbekend' was de overlap slechts 7 procent.
4. In de WBSO-set vinden we relatief iets minder kleinere bedrijven; bijna 50% van de bedrijven telt minder dan 9 werkzame personen. Van de AI-bedrijven Innovatiespotter heeft 63 procent minder dan 9 werkzame personen. Een kanttekening hierbij is dat voor meer dan 40% de werknemerscategorie onbekend is. In de tabel hieronder kan dit overigens niet direct worden afgeleid; voor de berekening van de genoemde percentages is de categorie 'onbekend' hier buiten beschouwing gelaten.

In de tabellen hieronder zijn specifieke cellen groen gearceerd waar naar wordt verwezen bij de punten 1 tot en met 4 hierboven.

Tabel 15 Overlapanalyse op SBI-codes

SBI-code 2-digits	Hoofdactiviteit	Innovatiespotter AI-bedrijven	WBSO-bedrijven met AI-project 2019	Overlap	Overlap als % van Innovatiespotter	Overlap als % van WBSO
62	Dienstverlenende activiteiten op het gebied van informatietechnologie	923 (39%)	839 (55%)	208 (58%)	23%	25%
63	Dienstverlenende activiteiten op het gebied van informatie	62 (3%)	29 (2%)	10 (3%)	16%	34%
64	Financiële instellingen (geen verzekeringen en pensioenfondsen)	362 (15%)	141 (9%)	11 (3%)	3%	8%
70	Holdings (geen financiële), conserndiensten binnen eigen concern en managementadvisering	221 (9%)	95 (6%)	23 (6%)	10%	24%
71	Architecten, ingenieurs en technisch ontwerp en advies; keuring en controle	116 (5%)	77 (5%)	23 (6%)	20%	30%
72	Speur- en ontwikkelingswerk	109 (5%)	63 (4%)	26 (7%)	24%	41%
--	Overige sectoren	594 (25%)	271 (18%)	58 (16%)	10%	21%
Totaal		2.387 (100%)	1.515 (100%)	359 (100%)	15%	24%

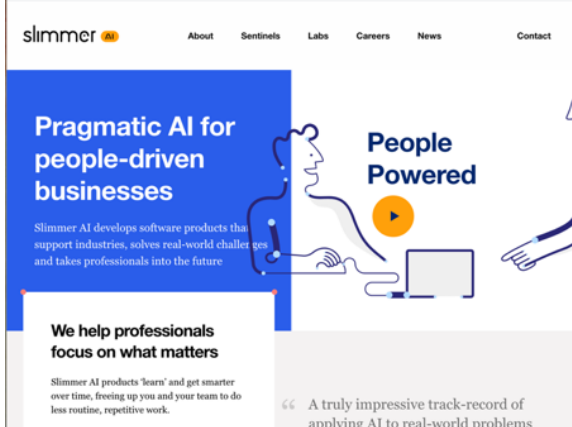
Tabel 16 Overlapanalyse op werknemerscategorieën

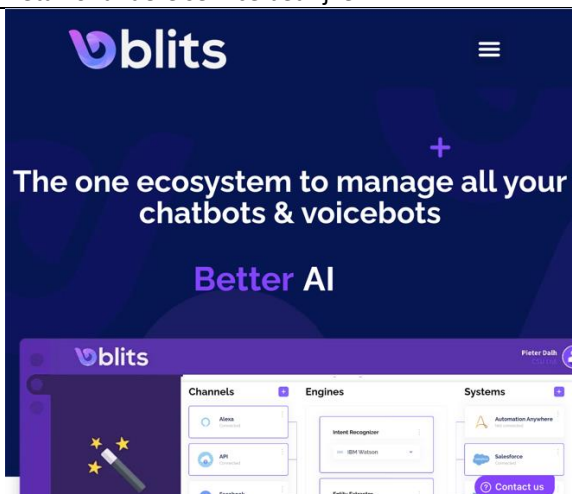
Werknemerscategorie (a.d.h.v. Innovatiespotter)	Innovatiespotter	WBSO-bedrijven met AI-project 2019	Overlap	Overlap als % van Innovatiespotter	Overlap als % van WBSO
Cat. 01: 1 - 1	326 (14%)	249 (16%)	38 (11%)	12%	15%
Cat. 02: 2 - 4	322 (13%)	206 (14%)	56 (16%)	17%	27%
Cat. 03: 5 - 9	209 (9%)	230 (15%)	60 (17%)	29%	26%
Cat. 04: 10 - 19	170 (7%)	241 (16%)	64 (18%)	38%	27%
Cat. 05: 20 - 49	147 (6%)	268 (18%)	30 (8%)	20%	11%
Cat. 06: 50 - 99	63 (3%)	106 (7%)	14 (4%)	22%	13%
Cat. 13: 100 en meer	121 (5%)	152 (10%)	27 (8%)	22%	18%
Cat. 00: Onbekend	1.029 (43%)	63 (4%)	70 (19%)	7%	111% ²⁵
Totaal	2.387 (100%)	1.515 (100%)	359 (100%)	15%	24%

²⁵ In de set van Innovatiespotter bevinden zich dubbele KvK-nummers. Van de 1 029 bedrijven met werknemerscategorie 'onbekend' uit de set van Innovatiespotter, kwamen er 70 (niet-unieke) KvK-nummers ook voor in de WBSO-analyse. Deze 70 KvK-nummers horen bij totaal 63 unieke KvK-nummers.

BIJLAGE 3 PERSONAS

Goede voorbeelden

Website: Slimmer ai	
Software	Software voor detectie van onregelmatigheden
Virtueel	Gebruikt digitale informatie en geen interactie met fysieke omgeving
Zakelijk	Toegepast door overheden of grote bedrijven
Ontwikkelt algoritmes voor risk-management en forensic profiling. Levert diensten aan de zakelijke markt	
	

Website: Blits	
Software	Low-code ecosysteem voor chatbots en voicebots
Virtueel	Online platform
Zakelijk	Kan worden toegepast door retail of andere service bedrijven
low-code ecosystem voor bouwen, trainen en toepassen van chat- & voicebots Omgeving waarin verschillende AI-engines (Google, IBM, Microsoft, Nuance, ...) kunnen worden gebruikt.	
	

Website: BodyGraph	
Software	Online platform en app voor wearables
Fysiek	Interpreteert informatie van wearables en geeft advies, ontwikkelt ook eigen shirt
Zakelijk	Toegepast op sport en zorg

Health is hot. More and more often, people want insights into their health. The current generation of smart wearables makes this possible, and provides the user with a lot of data. This data still needs to be interpreted, however. Because: what these smart wearables cannot do yet, is give you real-time feedback that is precisely tailored to your body and your situation

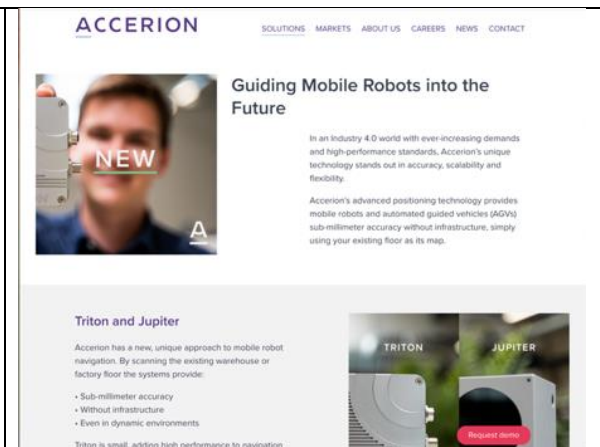


Website: Accerion

Hardware	Embedded in hardware
Fysiek	Stelt in staat om de fysieke omgeving waar te nemen
Industrie	Toegepast om robots autonoom te laten navigeren in fabrieken

Ontwikkelt robot modules die autonoom kunnen navigeren in o.a. magazijnen, fabrieken of kassen

Positioning technologie voor automated guided vehicles zonder hulp van signalering



Website: Cygnify

Software	Ontwikkelt software voor o.a. gezichtsherkenning
Fysiek	AI herkent veranderingen in het verkeer of bestuurder
Industrie	Software wordt toegepast door bedrijven in de automotive industrie

Ontwikkelt algoritmes voor autonomous vehicles en monitoren en assistentie van bestuurders.



Moeilijk te bepalen

Website: HOOOOP Een spraakgestuurde vacaturebank voor analfabeten, immigranten en andere groepen die nu geen of lastig vacatures kunnen lezen Spraakherkenning in vele talen. Vacatures worden ook in meerdere talen vertaald. Maar de AI-technologie wordt niet genoemd (Software – fysiek – consument)	
Website: Sleep ai Een app die geluidsopnames analyseert om te bepalen of je 's nachts snurkt of tandenknarst De website gaat in op de diagnose en hoe het is te voorkomen. Maar de AI-technologie wordt niet genoemd (Software – fysiek – consument)	

Foute context

Website: InnerQi

NLP komt voor op de website in de context van neurolinguïstisch programmeren, en niet in de context van natural language processing



Website: Ai Total

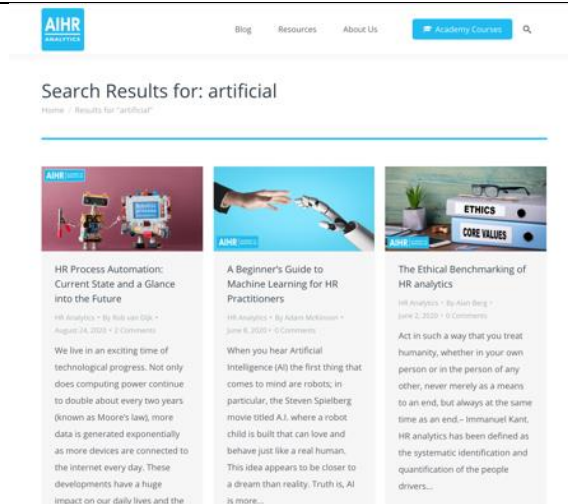
AI komt voor op de website in de context van artificial insemination (of kunstmatige inseminatie), en niet in de context van artificial intelligence of kunstmatige intelligentie



Blogs & nieuws

Website: Analytics in HR

Geeft cursussen voor human resource studenten en publiceert over ontwikkelingen binnen het vakgebied. Waaronder veel over AI en analytics



Website: Talent Factor Consultancy

De website bevat een blog over een seminar over toegepaste AI in de psychologie.



Gebruiker

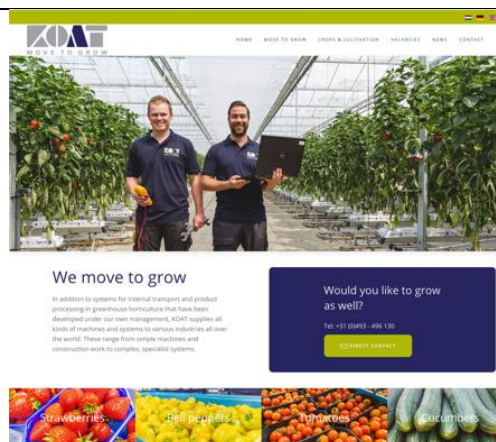
Website: Picnic

Past AI toe in het doen van aanbevelingen aan de klant, en de logistiek bij het bezorgen van boodschappen



Website: KOAT

Leverancier van mechatronica en automatisering (computer vision) in de landbouw en horticultuur



BIJLAGE 4 AI NETWERKEN

Netwerk	Bron
InnovationQuarter	Vertrouwelijke lijst
Techleap	Vertrouwelijke lijst
BOM	Vertrouwelijke lijst
Brainport	Vertrouwelijke lijst
OostNL	Vertrouwelijke lijst
NL AI Coalitie	http://nlaic.com/coalitie-deelnemers/
StartupDelta 2018	http://www.startupdelta.org/startupdelta-x-global-startup-ecosystem-report-2018-lets-zoom-in-on-breakthrough-technology/dutch-ai-startup-infographic-v0-2/
AI Startup Landscape 2020	https://ai.nl/ai-startups/
AI Index 2019	https://drive.google.com/drive/folders/15ilwouvGoQoOwpbEKw45gRqGg0oGgV13
AI Index 2021	https://drive.google.com/drive/folders/11dlp56lhuka6bbe52nbkVCywJ0orZG54
AI in Lifeport	http://www.aiforlife.nl/
SBIR Artificiele Intelligentie	https://www.rvo.nl/subsidie-en-financieringswijzer/sbir/sbir-oproep-artifici%C3%ABle-intelligentie-voor-de-publieke-diensten
Rockstart (AI)	https://www.rockstart.com/startups/

BIJLAGE 5 TREFWOORDENLIJST ELSEVIER, RVO, INNOVATIESPOTTER

Keyword source	Keywords	Relevant
Zoektermen van Elsevier	"Artificial intelligence"	Ja
Zoektermen van Elsevier	"Machine learning"	Ja
Zoektermen van Elsevier	"Deep Learning"	Ja
Zoektermen van Elsevier	"Neural networks"	Ja
Zoektermen van Elsevier	"Autonomous systems"	Ja
Zoektermen van Elsevier	"Reinforcement Learning"	Ja
Zoektermen van Elsevier	"Supervised Learning"	Ja
Zoektermen van Elsevier	"Supervised Machine Learning"	Ja
Zoektermen van Elsevier	"Semi-supervised Learning"	Ja
Zoektermen van Elsevier	"Unsupervised Machine Learning"	Ja
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Predictive"	Ja
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Zelflerend"	Ja
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Neurale netwerken"	Ja
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"TensorFlow"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"algorithm"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"algoritme"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"kunstmatige intelligentie"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"spraakherkenning"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"natural language processing"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"speech recognition"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"fuzzy logic"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"machine learning platform"	Ja
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"unsupervised learning"	Ja
Innovatiespotter	"ai data"~3	Ja
Innovatiespotter	"ai technology"~3	Ja
Innovatiespotter	"ai solution"~3	Ja
Innovatiespotter	"ai powered"~3	Ja
Innovatiespotter	"ai application"~3	Ja
Innovatiespotter	"ai driven"~3	Ja
Innovatiespotter	"ai digital"~3	Ja

Innovatiespotter	"ai technologie"~3	Ja
Innovatiespotter	"sentiment analysis"	Ja
Innovatiespotter	"anomaly detection"	Ja
Innovatiespotter	"ai enabled"~3	Ja
Innovatiespotter	"conversational ai"	Ja
Innovatiespotter	"cognitive ai"	Ja
Zoektermen van Elsevier	"Co-training"	te algemeen
Zoektermen van Elsevier	"Ant Colony Optimization"	te specifiek
Zoektermen van Elsevier	"Autonomous decision making"	te specifiek
Zoektermen van Elsevier	"Back-propagation Neural Network"	te specifiek
Zoektermen van Elsevier	"Bam Neural Networks"	te specifiek
Zoektermen van Elsevier	"Boltzmann Machine"	te specifiek
Zoektermen van Elsevier	"Bp Neural Network"	te specifiek
Zoektermen van Elsevier	"Causal interference"	te specifiek
Zoektermen van Elsevier	"Cohen-grossberg Neural Networks"	te specifiek
Zoektermen van Elsevier	"Computational Creativity"	te specifiek
Zoektermen van Elsevier	"Delayed Neural Networks"	te specifiek
Zoektermen van Elsevier	"Echo State Network"	te specifiek
Zoektermen van Elsevier	"Elman Neural Network"	te specifiek
Zoektermen van Elsevier	"Extreme Learning Machine"	te specifiek
Zoektermen van Elsevier	"Genetics-based Machine Learning"	te specifiek
Zoektermen van Elsevier	"Neuroevolution"	te specifiek
Zoektermen van Elsevier	"Probabilistic Neural Network"	te specifiek
Zoektermen van Elsevier	"Pulse Coupled Neural Network"	te specifiek
Zoektermen van Elsevier	"Radial Basis Function Neural Network"	te specifiek
Zoektermen van Elsevier	"Recurrent Neural Networks"	te specifiek
Zoektermen van Elsevier	"Sentiment Classification"	te specifiek
Zoektermen van Elsevier	"Spiking Neural Networks"	te specifiek
Zoektermen van Elsevier	"Stochastic Neural Networks"	te specifiek
Zoektermen van Elsevier	"Swarm Intelligence"	te specifiek
Zoektermen van Elsevier	"Turing Test"	te specifiek
Zoektermen van Elsevier	"Wavelet Neural Network"	te specifiek
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Voorspellen"	te algemeen
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Learning"	te algemeen
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"CNN"	te algemeen
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"R (als programmeertaal)"	te algemeen

Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Python"	te algemeen
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Tensor"	te algemeen
Termen en concepten uit de feedback van de betreffende WBSO-adviseur uit de validatieronde	"Genetische algoritmen"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"neural networks"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"neural network"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"neuraal netwerk"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"neurale netwerken"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"deep learning"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"machine learning"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"unsupervised machine learning"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"machine learning platforms"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"text normalization"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"artificial intelligence"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"algoritmen"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"algorithms"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"voorspellingen"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"predictions"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"bots"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"computationele fysische simulaties"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"complex models"	Dubbel
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"foto"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"API"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"cloud"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"NLP"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"natural language"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"natuurlijke taal"	te algemeen

Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"social media"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"AI"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"performance verbetering"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"service platform"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"input text"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"information processing"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"informatieverwerking"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"taal"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"language"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"database"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"voorspelling"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"prediction"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"bot"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"complex model"	te algemeen
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"text normalisation"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"tekstnormalisatie"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"PoS tags"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"text modules"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"tekstmodules"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"computationele fysische simulatie"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"decision tree"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"point cloud"	te specifiek
Termen en concepten die voorkomen in de door de experts gevalideerde projecten	"patronen herkennen"	te specifiek
Innovatiespotter	"data decision"~3	Te algemeen
Innovatiespotter	"deep vision"	te specifiek
Innovatiespotter	"predictive solution"	te specifiek

BIJLAGE 6 TREFWOORDENLIJST JRC

AI domain	AI Subdomain	Keyword
Reasoning	Reasoning	"case-based reasoning"
Reasoning	Reasoning	"causal inference"
Reasoning	Reasoning	"causal models"
Reasoning	Reasoning	"common-sense reasoning"
Reasoning	Reasoning	"expert system"
Reasoning	Reasoning	"fuzzy logic"
Reasoning	Reasoning	"graphical models"
Reasoning	Reasoning	"inductive programming"
Reasoning	Reasoning	"information theory"
Reasoning	Reasoning	"knowledge representation & reasoning"
Reasoning	Reasoning	"latent variable models"
Reasoning	Reasoning	"semantic web"
Reasoning	Reasoning	"uncertainty in artificial intelligence"
Planning	Planning	"bayesian optimisation"
Planning	Planning	"constraint satisfaction"
Planning	Planning	"evolutionary algorithm"
Planning	Planning	"genetic algorithm"
Planning	Planning	"gradient descent"
Planning	Planning	"hierarchical task network"
Planning	Planning	"metaheuristic optimisation"
Planning	Planning	"planning graph"
Planning	Planning	"stochastic optimisation"
Learning	Machine Learning	"active learning "
Learning	Machine Learning	"adaptive learning "
Learning	Machine Learning	"adversarial machine learning "
Learning	Machine Learning	"adversarial network "
Learning	Machine Learning	"anomaly detection "
Learning	Machine Learning	"artificial neural network "
Learning	Machine Learning	"automated machine learning "
Learning	Machine Learning	"automatic classification "
Learning	Machine Learning	"automatic recognition "
Learning	Machine Learning	"bagging "
Learning	Machine Learning	"bayesian modelling "
Learning	Machine Learning	"boosting "
Learning	Machine Learning	"classification "
Learning	Machine Learning	"clustering "
Learning	Machine Learning	"collaborative filtering "
Learning	Machine Learning	"content-based filtering "

Learning	Machine Learning	"convolutional neural network "
Learning	Machine Learning	"data mining "
Learning	Machine Learning	"deep learning "
Learning	Machine Learning	"deep neural network "
Learning	Machine Learning	"ensemble method "
Learning	Machine Learning	"feature extraction"
Learning	Machine Learning	"generative adversarial network"
Learning	Machine Learning	"generative model"
Learning	Machine Learning	"multi-task learning"
Learning	Machine Learning	"neural network"
Learning	Machine Learning	"pattern recognition"
Learning	Machine Learning	"probabilistic learning"
Learning	Machine Learning	"probabilistic model"
Learning	Machine Learning	"recommender system"
Learning	Machine Learning	"recurrent neural network"
Learning	Machine Learning	"recursive neural network"
Learning	Machine Learning	"reinforcement learning"
Learning	Machine Learning	"semi-supervised learning"
Learning	Machine Learning	"statistical learning"
Learning	Machine Learning	"statistical relational learning"
Learning	Machine Learning	"supervised learning"
Learning	Machine Learning	"support vector machine"
Learning	Machine Learning	"transfer learning"
Learning	Machine Learning	"unstructured data"
Learning	Machine Learning	"unsupervised learning"
Communication	Natural Language Processing	"chatbot"
Communication	Natural Language Processing	"computational linguistics"
Communication	Natural Language Processing	"conversation model"
Communication	Natural Language Processing	"coreference resolution"
Communication	Natural Language Processing	"information extraction"
Communication	Natural Language Processing	"information retrieval"
Communication	Natural Language Processing	"natural language understanding"
Communication	Natural Language Processing	"natural language generation"
Communication	Natural Language Processing	"machine translation"
Communication	Natural Language Processing	"question answering"
Communication	Natural Language Processing	"sentiment analysis"
Communication	Natural Language Processing	"text classification"
Communication	Natural Language Processing	"text mining"
Perception	Computer Vision	"action recognition"
Perception	Computer Vision	"face recognition"
Perception	Computer Vision	"gesture recognition"
Perception	Computer Vision	"image processing"

Perception	Computer Vision	"image retrieval"
Perception	Computer Vision	"object recognition"
Perception	Computer Vision	"recognition technology"
Perception	Computer Vision	"sensor network"
Perception	Computer Vision	"visual search"
Perception	Audio Processing	"computational auditory scene"
Perception	Audio Processing	"music information retrieval"
Perception	Audio Processing	"sound description"
Perception	Audio Processing	"sound event recognition"
Perception	Audio Processing	"sound source separation"
Perception	Audio Processing	"sound synthesis"
Perception	Audio Processing	"speaker identification"
Perception	Audio Processing	"speech processing"
Perception	Audio Processing	"speech recognition"
Perception	Audio Processing	"speech synthesis"
Integration and interaction	Multi-agent systems	"agent-based modelling"
Integration and interaction	Multi-agent systems	"agreement technologies"
Integration and interaction	Multi-agent systems	"computational economics"
Integration and interaction	Multi-agent systems	"game theory"
Integration and interaction	Multi-agent systems	"intelligent agent"
Integration and interaction	Multi-agent systems	"negotiation algorithm"
Integration and interaction	Multi-agent systems	"network intelligence"
Integration and interaction	Multi-agent systems	"q-learning"
Integration and interaction	Multi-agent systems	"swarm intelligence"
Integration and interaction	Robotics and Automation	"cognitive system"
Integration and interaction	Robotics and Automation	"control theory"
Integration and interaction	Robotics and Automation	"human-ai interaction"
Integration and interaction	Robotics and Automation	"industrial robot"
Integration and interaction	Robotics and Automation	"robot system"
Integration and interaction	Robotics and Automation	"service robot"
Integration and interaction	Robotics and Automation	"social robot"
Integration and interaction	Connected and Automated vehicles	"autonomous driving"
Integration and interaction	Connected and Automated vehicles	"autonomous system"
Integration and interaction	Connected and Automated vehicles	"autonomous vehicle"
Integration and interaction	Connected and Automated vehicles	"self-driving car"
Integration and interaction	Connected and Automated vehicles	"unmanned vehicle"
Services	AI Services	"ai application"
Services	AI Services	"ai benchmark"
Services	AI Services	"ai competition"
Services	AI Services	"ai software toolkit"
Services	AI Services	"analytics platform"
Services	AI Services	"big data"

Services	AI Services	"business intelligence"
Services	AI Services	"central processing unit"
Services	AI Services	"computational creativity"
Services	AI Services	"computational neuroscience"
Services	AI Services	"data analytics"
Services	AI Services	"decision analytics"
Services	AI Services	"decision support"
Services	AI Services	"distributed computing"
Services	AI Services	"graphics processing unit"
Services	AI Services	"intelligence software"
Services	AI Services	"intelligent control"
Services	AI Services	"intelligent control system"
Services	AI Services	"intelligent hardware development"
Services	AI Services	"intelligent software development"
Services	AI Services	"intelligent user interface"
Services	AI Services	"internet of things"
Services	AI Services	"machine learning framework"
Services	AI Services	"machine learning library"
Services	AI Services	"machine learning platform"
Services	AI Services	"personal assistant"
Services	AI Services	"platform as a service"
Services	AI Services	"tensor processing unit"
Services	AI Services	"virtual environment"
Services	AI Services	"virtual reality"

BIJLAGE 7 FEATURES (WOORDEN) PER MODEL

LR_TFIDF_word_100

feature	value	new	-0,38	busi	0,32	ontwikkel	0,94
mail	-1,97	media	-0,37	process	0,35	custom	0,95
hom	-1,89	innov	-0,30	werk	0,37	experi	0,97
info	-1,64	contact	-0,30	softwar	0,43	user	0,98
dat	-1,49	offer	-0,28	person	0,44	creat	1,00
nl	-1,49	lat	-0,27	zorg	0,46	news	1,15
goed	-1,20	work	-0,19	manag	0,51	model	1,23
grag	-1,18	bekijk	-0,18	inform	0,55	blog	1,26
grot	-1,13	onlin	-0,15	social	0,55	cloud	1,27
world	-1,13	industri	-0,07	amp	0,58	help	1,42
kom	-1,07	provid	-0,06	open	0,63	start	1,59
design	-0,94	nem	-0,04	email	0,64	digit	1,60
snel	-0,86	product	-0,03	com	0,65	research	1,64
websit	-0,84	vind	0,02	content	0,66	team	1,69
bedrijf	-0,78	informatie	0,02	amsterdam	0,68	klant	1,76
zoek	-0,75	https	0,07	develop	0,70	oplos	1,77
stat	-0,59	menu	0,10	compani	0,71	learn	1,88
algemen	-0,58	best	0,10	right	0,74	app	1,92
nederland	-0,58	support	0,15	privacy	0,77	solut	1,99
project	-0,57	wet	0,15	applic	0,79	technolog	2,16
servic	-0,51	home	0,19	system	0,84	platform	2,92
dienst	-0,47	read	0,22	gebruik	0,85	data	4,17
ervar	-0,47	vrag	0,23	need	0,85	ai	4,51
nieuw	-0,43	high	0,29	privaci	0,87		
gan	-0,42	event	0,30	engin	0,90		
market	-0,39	partner	0,32	direct	0,91		

LR_TFIDF_word_300

feature	value	nl	-1,14	cursus	-0,90	design	-0,77
welkom	-1,92	kvk	-1,11	goed	-0,89	zoek	-0,75
mail	-1,73	foto	-1,10	natur	-0,88	global	-0,74
hom	-1,70	lev	-1,09	complet	-0,85	advies	-0,73
consult	-1,41	mooi	-1,06	gebied	-0,84	nieuw	-0,73
info	-1,36	view	-1,04	behandel	-0,82	snel	-0,72
dat	-1,25	begeleid	-1,03	mat	-0,82	wilt	-0,71
afsprak	-1,24	websit	-1,02	internet	-0,80	bedrijf	-0,70
diver	-1,20	bestell	-0,92	praktijk	-0,78	grag	-0,68

project	-0,68	share	-0,43	auto	-0,15	laatst	0,07
world	-0,68	transport	-0,42	dutch	-0,14	netherlands	0,07
dienst	-0,67	ervar	-0,41	biedt	-0,14	impact	0,08
ontwerp	-0,67	web	-0,41	requir	-0,13	https	0,09
kom	-0,67	winkel	-0,40	organ	-0,12	softwar	0,11
nieuwsbrief	-0,67	contact	-0,40	test	-0,11	public	0,11
shop	-0,66	beschik	-0,39	vind	-0,11	review	0,12
btw	-0,66	kwaliteit	-0,38	lead	-0,11	opleid	0,12
problem	-0,66	success	-0,35	energi	-0,10	bied	0,13
bel	-0,63	bekijk	-0,35	like	-0,10	connect	0,14
grot	-0,63	del	-0,35	volg	-0,08	busines	0,14
special	-0,61	train	-0,34	client	-0,08	googl	0,15
work	-0,60	chang	-0,33	function	-0,07	woning	0,16
group	-0,59	workshop	-0,32	menu	-0,06	tool	0,17
management	-0,57	member	-0,32	wet	-0,06	video	0,19
market	-0,57	onderhoud	-0,31	verschill	-0,06	vrag	0,20
zowel	-0,56	bv	-0,31	informatie	-0,06	meet	0,21
duurzaam	-0,56	look	-0,30	network	-0,05	secur	0,22
training	-0,56	intern	-0,29	document	-0,05	term	0,22
brand	-0,55	process	-0,28	manier	-0,04	event	0,23
stat	-0,55	financ	-0,27	support	-0,03	effect	0,23
kijk	-0,54	hand	-0,27	samenwerk	-0,03	provid	0,23
algemen	-0,54	rotterdam	-0,26	continu	-0,02	peopl	0,25
specialist	-0,53	kennis	-0,26	basis	-0,02	valu	0,26
nederland	-0,52	merk	-0,26	echt	-0,02	find	0,26
le	-0,52	huis	-0,24	offic	-0,01	gratis	0,26
facebok	-0,52	houd	-0,24	report	-0,01	read	0,27
water	-0,51	ler	-0,24	resourc	-0,01	latest	0,28
plat	-0,51	artikel	-0,23	nem	0,00	partner	0,30
webshop	-0,50	way	-0,22	industri	0,01	ontdek	0,30
product	-0,50	qualiti	-0,21	activ	0,02	social	0,32
klik	-0,50	gemaakt	-0,21	onlin	0,02	zorg	0,32
gat	-0,50	focus	-0,20	stan	0,03	bouw	0,32
innov	-0,49	gan	-0,20	manag	0,03	high	0,33
klein	-0,48	domain	-0,19	home	0,04	detail	0,33
januari	-0,48	new	-0,18	offer	0,04	medewerker	0,34
belangrijk	-0,48	perform	-0,18	set	0,04	ict	0,34
media	-0,47	date	-0,17	english	0,04	best	0,36
sit	-0,46	februari	-0,17	deliv	0,05	free	0,37
post	-0,46	zie	-0,17	fal	0,05	bedrijv	0,37
servic	-0,45	co	-0,16	lever	0,06	werk	0,38
sustain	-0,44	lat	-0,16	includ	0,06	inform	0,39
mee	-0,43	account	-0,16	zien	0,07	ondersteun	0,40

custom	0,41	experi	0,66	download	0,98	start	1,52
busi	0,42	right	0,66	vacatures	0,98	real	1,52
polici	0,44	develop	0,66	cloud	0,99	learn	1,54
proces	0,45	know	0,66	news	0,99	team	1,62
ga	0,46	privacy	0,68	digital	1,01	scienc	1,62
base	0,46	build	0,68	creat	1,07	oplos	1,65
voorwaard	0,46	plan	0,70	result	1,11	app	1,69
com	0,47	privaci	0,73	blog	1,13	technolog	1,77
engin	0,47	job	0,74	search	1,13	analyt	1,80
futur	0,47	career	0,76	cost	1,13	insight	1,95
applic	0,48	amp	0,77	model	1,14	improv	1,98
control	0,48	user	0,78	twitter	1,16	autom	2,11
collect	0,53	gebruik	0,78	case	1,24	machin	2,23
gebruikt	0,54	content	0,78	werkt	1,25	monitor	2,39
men	0,56	integr	0,80	expert	1,28	platform	2,54
maakt	0,56	end	0,80	verbeter	1,30	inzicht	2,59
veilig	0,57	email	0,82	mobil	1,30	smart	2,90
open	0,57	compani	0,82	digit	1,30	intellig	3,30
amsterdam	0,58	risk	0,85	enabl	1,33	demo	3,34
system	0,60	need	0,88	help	1,38	data	3,82
person	0,60	oper	0,89	effici	1,39	ai	4,45
organisatie	0,64	linkedin	0,94	solut	1,40		
direct	0,64	power	0,94	research	1,45		
nodig	0,65	ontwikkel	0,94	klant	1,51		

LR_TFIDF_word_400

feature	value	begeleid	-0,95	juni	-0,77	program	-0,66
welkom	-1,82	onderdel	-0,89	kinder	-0,76	recht	-0,66
hom	-1,68	behandel	-0,89	praktijk	-0,75	internet	-0,66
mail	-1,62	bestell	-0,88	duurzam	-0,75	locatie	-0,65
info	-1,35	cursus	-0,88	nieuw	-0,74	tel	-0,65
consult	-1,30	zoek	-0,87	mat	-0,74	groep	-0,65
nl	-1,29	server	-0,87	advies	-0,74	project	-0,64
afsprak	-1,23	gebied	-0,87	complet	-0,73	wilt	-0,64
dat	-1,17	komt	-0,87	april	-0,73	management	-0,63
foto	-1,12	natur	-0,85	shop	-0,71	ontwerp	-0,63
diver	-1,11	maart	-0,84	grag	-0,71	knowledg	-0,63
view	-1,05	design	-0,84	sustain	-0,68	professionel	-0,62
kvk	-1,05	wen	-0,79	product	-0,67	kom	-0,62
lev	-1,02	goed	-0,79	concept	-0,67	grot	-0,61
websit	-1,00	bedrijf	-0,78	snel	-0,66	dienst	-0,61
mooi	-0,98	global	-0,77	btw	-0,66	world	-0,61

work	-0,60	belangrijk	-0,38	look	-0,22	stan	-0,05
bel	-0,60	winkel	-0,38	qualiti	-0,20	menu	-0,05
webshop	-0,60	corpor	-0,38	domain	-0,20	basis	-0,05
problem	-0,59	beschik	-0,37	resultat	-0,20	differ	-0,05
le	-0,59	del	-0,37	co	-0,20	offer	-0,05
prijs	-0,58	mee	-0,37	date	-0,20	softwar	-0,04
training	-0,57	bekijk	-0,36	artikel	-0,19	onlin	-0,04
kind	-0,57	kwaliteit	-0,36	techniek	-0,19	wereld	-0,03
activiteit	-0,56	process	-0,35	financ	-0,19	verschill	-0,03
ruim	-0,56	train	-0,35	like	-0,18	informatie	-0,02
nieuwsbrief	-0,55	cour	-0,34	test	-0,18	lead	0,00
nederland	-0,55	januari	-0,34	services	-0,18	network	0,00
market	-0,54	ervar	-0,34	challeng	-0,18	nem	0,00
kijk	-0,53	kennis	-0,34	gemaakt	-0,18	includ	0,01
group	-0,53	elkar	-0,33	huis	-0,18	field	0,01
water	-0,53	pagina	-0,33	houd	-0,18	document	0,02
servic	-0,52	onderhoud	-0,33	krijg	-0,17	laatst	0,03
algemen	-0,52	share	-0,33	lat	-0,16	manier	0,03
kop	-0,51	web	-0,33	technisch	-0,16	samenwerk	0,03
post	-0,51	bv	-0,32	account	-0,15	home	0,04
media	-0,50	strategi	-0,31	support	-0,15	zodat	0,04
zowel	-0,50	covid	-0,31	organ	-0,15	aantal	0,04
plat	-0,49	way	-0,31	ler	-0,14	deliv	0,04
sit	-0,49	requir	-0,31	gan	-0,14	zien	0,04
special	-0,49	member	-0,30	success	-0,14	echt	0,05
gat	-0,49	beher	-0,29	zie	-0,13	activ	0,05
possibl	-0,48	omgev	-0,29	energi	-0,13	resourc	0,06
specialist	-0,48	profess	-0,29	report	-0,13	continu	0,06
locat	-0,48	extra	-0,29	stori	-0,13	netherland	0,06
klik	-0,48	eenvoud	-0,28	center	-0,12	bereik	0,07
juli	-0,47	hand	-0,28	industri	-0,10	set	0,07
gezond	-0,47	site	-0,28	vind	-0,10	lever	0,08
stat	-0,46	function	-0,27	focus	-0,10	review	0,09
perform	-0,46	intern	-0,27	client	-0,09	peopl	0,09
facebok	-0,45	vast	-0,26	bericht	-0,09	order	0,09
brand	-0,45	merk	-0,26	kun	-0,09	ieder	0,09
contact	-0,44	rotterdam	-0,26	februari	-0,09	offic	0,09
klein	-0,43	central	-0,25	wet	-0,08	public	0,10
boek	-0,41	big	-0,25	dutch	-0,07	impact	0,10
chang	-0,40	workshop	-0,24	biedt	-0,07	vrag	0,11
innov	-0,40	auto	-0,23	aanvrag	-0,07	secur	0,12
hog	-0,40	leuk	-0,23	volg	-0,06	effect	0,12
transport	-0,39	new	-0,23	featur	-0,06	opleid	0,12

latest	0,13	inform	0,34	experi	0,63	werkt	1,15
english	0,13	avail	0,34	privaci	0,65	model	1,15
manag	0,15	ontdek	0,35	build	0,66	cloud	1,16
gratis	0,16	collect	0,35	user	0,68	expert	1,17
free	0,17	medewerker	0,36	email	0,69	research	1,18
provid	0,17	engin	0,36	organisatie	0,69	case	1,19
join	0,17	werk	0,37	gebruik	0,72	effici	1,20
connect	0,18	explor	0,37	job	0,73	blog	1,21
googl	0,18	custom	0,37	onderzoek	0,73	verbeter	1,24
https	0,18	polici	0,39	need	0,76	help	1,26
woning	0,18	ondersteun	0,40	necessari	0,79	solut	1,29
term	0,19	best	0,41	news	0,80	human	1,32
tool	0,19	futur	0,42	content	0,81	digit	1,39
valu	0,19	gebruikt	0,45	oper	0,82	real	1,45
partner	0,19	busi	0,46	risk	0,82	klant	1,49
bied	0,19	com	0,46	integr	0,82	scienc	1,52
terug	0,20	men	0,47	studi	0,83	oplos	1,56
level	0,20	voorwaard	0,47	access	0,84	increa	1,57
busines	0,21	updat	0,48	power	0,85	technologie	1,59
video	0,21	know	0,49	toepas	0,85	technolog	1,60
moment	0,22	betal	0,49	linkedin	0,87	start	1,61
meet	0,23	health	0,50	ontwikkel	0,88	learn	1,69
social	0,24	control	0,50	price	0,88	improv	1,72
event	0,24	veilig	0,50	optim	0,88	team	1,75
transform	0,25	maakt	0,52	inlogg	0,91	app	1,77
read	0,25	environ	0,54	mobil	0,91	analyt	1,78
let	0,26	plan	0,54	cost	0,91	autom	1,84
find	0,26	nodig	0,55	digital	0,92	insight	1,91
statement	0,28	develop	0,56	search	0,92	machin	2,07
detail	0,28	person	0,56	result	0,92	imag	2,14
applic	0,29	understand	0,56	creat	0,93	platform	2,40
bouw	0,29	right	0,57	request	0,94	monitor	2,45
sector	0,30	volled	0,58	download	0,96	inzicht	2,48
system	0,31	direct	0,58	vacatures	1,00	smart	2,72
follow	0,31	base	0,58	better	1,02	robot	2,78
high	0,32	zet	0,58	complex	1,03	demo	3,13
bedrijv	0,32	amsterdam	0,60	twitter	1,04	intellig	3,15
ga	0,32	career	0,61	compani	1,04	data	4,08
proces	0,33	live	0,62	asset	1,05	ai	4,41
check	0,33	open	0,62	devic	1,08		
zorg	0,33	end	0,63	enabl	1,11		
ict	0,34	privacy	0,63	kost	1,15		

BIJLAGE 8 BEGRIPPENLIJST

Actionable insights	Betekenisvolle inzichten die voortkomen uit het analyseren van data.
Actionable Intelligence	Beschikken over de benodigde informatie om met een situatie om te gaan. Daarbij gaat het om real-time verzamelen, organiseren en analyseren van data.
AdaBoost	Een afkorting van Adaptive Boosting, een meta-algoritme voor statistische classificatie. Een meta-algoritme is een algoritme dat een set andere algoritmes gebruikt om weer een nieuw algoritme te creëren.
Algoritme	Een reeks stappen om een beoogd doel te bereiken. In de context van AI betreft dit een reeks instructies in een bepaalde programmeertaal om de computer een probleem op te laten lossen. Algoritmes worden vaak gebruikt om bepaalde beslissingen (deels) te automatiseren.
Automated Guided Vehicles	Een voertuig zonder bestuurder dat geleid wordt door software en sensoren.
Big data	Zeer grote databronnen met als eigenschap dat de data ongestructureerd is, met een wisselende kwaliteit en betrouwbaarheid.
Business Intelligence	Het verzamelen en analyseren van bedrijfsgegevens en –informatie. Dit gebeurt binnen de eigen organisatie en zou moeten leiden tot kennis en het aanzetten tot (een) actie(s).
Chatbot	Een geautomatiseerde (virtuele) gesprekspartner.
Cognitive computing	Geautomatiseerde modellen die het menselijke denkproces nabootsen in complexe situaties.
Computer vision	Het verwerken van beelden door computers. Het gaat hierbij om het detecteren, analyseren en interpreteren van afbeeldingen.
Data Science	Onderzoeksveld met betrekking tot wetenschappelijke methoden, processen en systemen om kennis en inzichten te onttrekken uit data.
Decision trees	Een beslisboom, voor de weergave van een besluitvormingsproces. Wordt onder andere gebruikt bij het random-forest-classificatie-algoritme (zie onder).
Deep learning	Nabootsen van het menselijke brein bij het verwerken van data en het creëren van patronen voor besluitvorming.
Internet of Things	Het geheel aan apparaten dat via internetverbinding met andere apparaten of systemen in contact staat en gegevens uitwisselen.
Logistische regressie	Een voorspellend model met categorische, vaak dichotome, variabelen. Voorbeelden van dichotome variabelen zijn: waar of niet waar, man of vrouw, etc.
Machine learning	De ontwikkeling van algoritmes waarmee systemen automatisch kunnen leren van data, patronen kunnen identificeren en beslissingen kunnen maken. 1. Supervised machine learning: hierbij wordt de input en de gewenste output aan de computer meegegeven om patronen in de data te vinden waarmee de input om kan worden gezet in de gewenste output 2. Unsupervised machine learning: hierbij wordt geen gewenste output meegegeven en wordt het aan de computer zelf overgelaten om patronen in de data te ontdekken
Mechatronica	Technische discipline die bestaat uit de combinatie van elektrotechniek, werktuigbouwkunde, en meet- en regeltechniek.
Multi-agent systems	Meerdere (virtuele) actoren die communiceren in een gedeelde omgeving om gemeenschappelijke of tegenstrijdige doelen te bereiken.
Naive Bayes	Een verzameling classificatie-algoritmen die allen een gemeenschappelijk principe delen, gebaseerd op de Bayesiaanse statistiek. Het is een methode om de kans dat iets gebeurt te berekenen, op basis van kennis over voorgaande keren dat hetzelfde gebeurde.

Natural Language Processing	Interactie tussen computers en menselijke taal, in het bijzonder het verwerken en analyseren van grote hoeveelheden natuurlijke taalgegevens.
Overfitten	De productie van een model dat te nauw overeenkomt met de input gegevens waardoor het in staat is om aanvullende gegevens correct te verwerken of te voorspellen. Een model is daardoor niet goed toe te passen op nieuwe gegevens.
Precision en Recall	Precision (precisie of nauwkeurigheid): als een model positief <i>voorspelt</i> (bv.: dit is een AI-bedrijf), hoe vaak is dit dan correct? Een hoge precision betekent dat er weinig false positives zijn. De formule voor precision is: $\frac{\text{True positive}}{\text{True positive} + \text{False positive}}$ Recall: hoe veel van de <i>werkelijke</i> positieve gevallen (bv. werkelijke AI-bedrijven) worden als zodanig door het model herkend. Een hoge recall betekent dat er weinig false negatives zijn. De formule voor recall is: $\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$
Random forest	Een classificatie-algoritme bestaande uit vele beslisbomen.
Scrapen	Ook wel <i>web scraping</i> . Het gebruik van software om informatie van webpagina's te extraheren om deze vervolgens te analyseren.
Support Vector Machines	Een binaire classificeerder. Aan de hand van een aantal kenmerken wijst het algoritme objecten toe aan een van twee klassen.
Taxonomie	Het indelen van individuen of objecten in groepen.
Tekstmining	Het proces om met allerlei ICT-technieken waardevolle informatie te halen uit grote hoeveelheden tekstmateriaal.
Testset	De testset zorgt voor de laatste evaluatie van het model na de training fase. De test set is nodig om de nauwkeurigheid (zie precision en recall) van het model te bepalen.
Trainingsset	Een deel van de dataset dat gebruikt wordt om het model te trainen. Het model leert hierbij van een fractie van de werkelijke data.
Validatieset	Om ervoor te zorgen dat het model getraind kan worden met behulp van de trainingsset, moet het regelmatig geëvalueerd worden. De validatie set zorgt voor deze evaluatie, door het berekenen van de regelmaat waarop fouten worden gemaakt door het model. Uiteindelijk zal het model haar variabelen op basis van de validatie afstemmen.
Virtual assistant	Een softwareprogramma die door middel van spraakherkenning taken of diensten uit kan voeren voor de gebruiker.
Voicebot	Een computergestuurd programma dat spraak van mensen kan begrijpen en aan de hand van de input kan reageren.
Word-embedding	In word-embedding modellen wordt de betekenis van woorden vertaald naar een reeks van getallen. Woorden met een soortgelijke betekenis krijgen op die manier een soortgelijke reeks getallen. Dit maakt het voor de computer mogelijk om woorden op basis van hun betekenis te kunnen vergelijken.

BIJLAGE 9 INTERNATIONALE CONTEXT

Dit door CBS, RVO en Innovatiespotter uitgevoerde onderzoek staat niet op zichzelf. Er zijn diverse internationale onderzoeken uitgevoerd om (de adoptie van) AI in kaart te brengen. Hieronder staan deze onderzoeken beschreven.

Het is belangrijk om te vermelden dat de criteria waarop bedrijven zijn geselecteerd niet duidelijk zijn beschreven. Er wordt enkel vermeld dat AI-gerelateerde woorden voorkomen. Ook is niet duidelijk hoe volledig de onderzochte populatie is. In geen van deze onderzoeken is men transparant over aard, omvang, bron en volledigheid van de gebruikte dataset.

European enterprise survey on the use of technologies based on artificial intelligence (2020).

Uit onderzoek van Ipsos²⁶ blijkt dat 40% van de Nederlandse bedrijven ten minste één AI-technologie toepassen. 27% past AI nog niet toe, maar is dat wel van plan in de komende twee jaar. 32% past AI nog niet toe, en is dat ook niet van plan. De enquête is ingevuld door 9640 ondernemingen in 30 Europese landen, waarvan 500 in Nederland. Dit is een onderzoek naar de adoptie van AI, er wordt dus niet uitsluitend gekeken naar ontwikkelaars. Het onderzoek gebruikt een taxonomie van 10 technologieën.

Tabel 17. Adoption per AI technology (Ipsos, 2020, p.114)

AI Technology	NL	EU27
Process automation	14%	12%
Process optimisation	13%	13%
Forecasting	13%	10%
Recommendation engines	11%	9%
NLP	10%	10%
Anomaly detection	9%	13%
Computer vision	8%	8%
Autonomous machines	7%	9%
Creative activities	6%	7%
Sentiment analysis	2%	3%

Er is ook een analyse gemaakt van AI-adoptie per sector, met gebruik van unsupervised machine learning (K-Means). Voor het clusteren zijn de antwoorden van de respondenten gebruikt, elk van de 9 vragen is een dimensie. Dit is teruggebracht tot twee dimensies om het te kunnen visualiseren. Deze dimensies kunnen als volgt worden geïnterpreteerd:

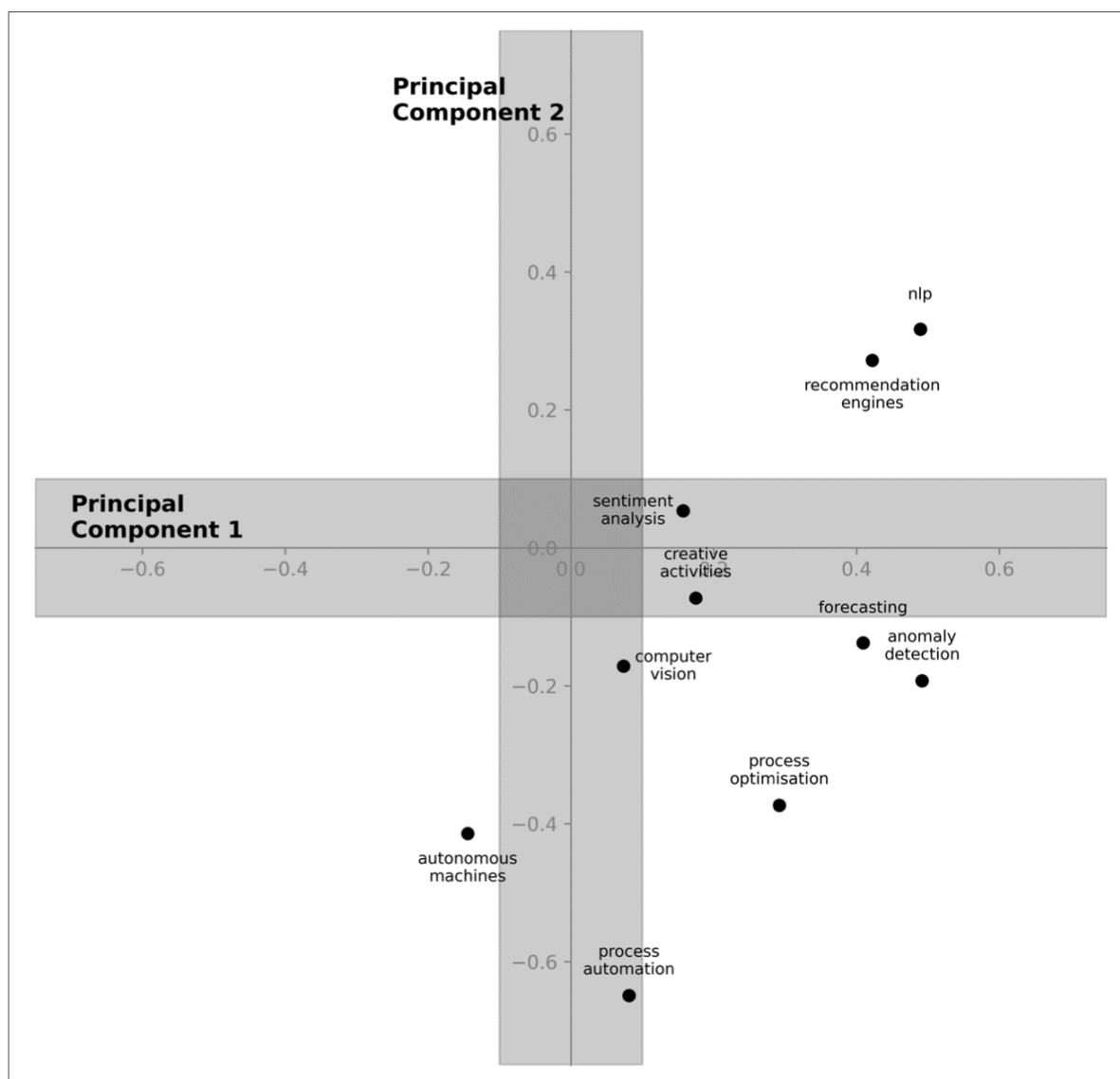
X of Principal component 1: Lower use of AI Technologies (<0) – Higher use of AI Technologies (>0)

Y of Principal component 2: Enterprise Applications (<0) – Customer-facing Applications (>0)

²⁶ Ipsos (2020), *European enterprise survey on the use of technologies based on artificial intelligence*, Publications Office of the European Union, 2020. <http://doi.org/10.2759/759368> (p.31).

Tabel 18. Principal components (Ipsos, 2020, p.41)

AI Technology	X	Y
NLP	0.489	0.317
Computer vision	0.073	-0.171
Anomaly detection	0.492	-0.193
Sentiment analysis	0.157	0.054
Forecasting	0.409	-0.138
Process optimisation	0.292	-0.373
Recommendation engines	0.422	0.272
Process automation	0.081	-0.649
Autonomous machines	-0.145	-0.414
Creative activities	0.175	-0.073



Figuur 10. Scatter plot (Ipsos, 2020, p.43).

Vervolgens is gekeken welke technologieën vaak samen voorkomen (bundles), en in welke SBI-sectoren dit het meest wordt toegepast.

Tabel 19. Meest voorkomende AI-technologie bundles per SBI-sector (Ipsos, 2020, p.47-48)

SBI-sector	Bundle	Support
[A] Landbouw, bosbouw en visserij	Autonomous Machines Computere vision	22.57%
[C] Industrie	Process Automation Process Optimisation	26.96%
[D] Productie en distributie van en handel in elektriciteit, aardgas, stoom en gekoelde lucht	Process Automation Process Optimisation	28.64%
	Forecasting Processs Optimisation	17.37%
	Forecasting Process automation	16.43%
[E] Winning en distributie van water; afval- en afvalwaterbeheer en sanering	Process Automation Process Optimisation	21.79%
[H] Vervoer en opslag	Forecasting Processs Optimisation	23.11%
[J] Informatie en communicatie	Anomaly Detection NLP	34.56%
	Forecasting Process Automation	26.65%
	Forecasting Processs Optimisation	19.53%
	Anomaly Detection Process Optimisation	18.21%
	Anomaly Detection Recommendation Engines	17.68%
	Forecasting Recommendation Engines	17.41%
	NLP Process Optimisation	17.41%
	Process Optimisation Recommendation Engines	17.15%
	Anomaly Detection Forecasting	16.89%
	NLP Recommendatiooon Engines	15.57%
	Forecasting NLP	15.30%
[K] Financiële instellingen	Anomaly Detection Forecasting	15.30%
	Anomaly Detection Process Automation	24.38%
[P] Onderwijs	Creative Activities NLP	28.15%
	Creative Activities Process Optimisation	24.07%

OECD Measuring the digital transformation (2019)

In een grootschalig onderzoek van het OECD²⁷ is ook een analyse uitgevoerd met unsupervised machine learning. Deze analyse is uitgevoerd door het Engelse bedrijf Glass.ai. Waar Ipsos (2020) gebruikt maakt van enquêteresultaten kijkt Glass.ai naar websites van bedrijven. Omdat de methode enigszins vergelijkbaar is kunnen de uitkomsten een indicatie zijn van het soort activiteiten die Innovatiespotter & CBS zullen identificeren.

In 2018 waren er in het Verenigd Koninkrijk 6.000 AI-gerelateerde bedrijven. Glass.ai is niet duidelijk op basis van welke criteria deze bedrijven zijn geselecteerd, maar het is aannemelijk dat dit met een vorm van text-retrieval of financieringsinformatie is gedaan. Van deze 6.000 bedrijven gebruikten 2.725 expliciet AI-activiteiten op de website²⁸.

Deze ±2.800 bedrijven zijn gebruikt in een unsupervised machine learning analyse. Voor de topic modeling is een Dirichlet Allocation algorithm en Gibbs sampling gebruikt, waarbij is gekozen voor 9 topics. De frequentie en combinatie van woorden bepaald aan welk topic een bedrijf wordt toegewezen. De X- en Y-as zijn bepaald met een principal component analysis. In

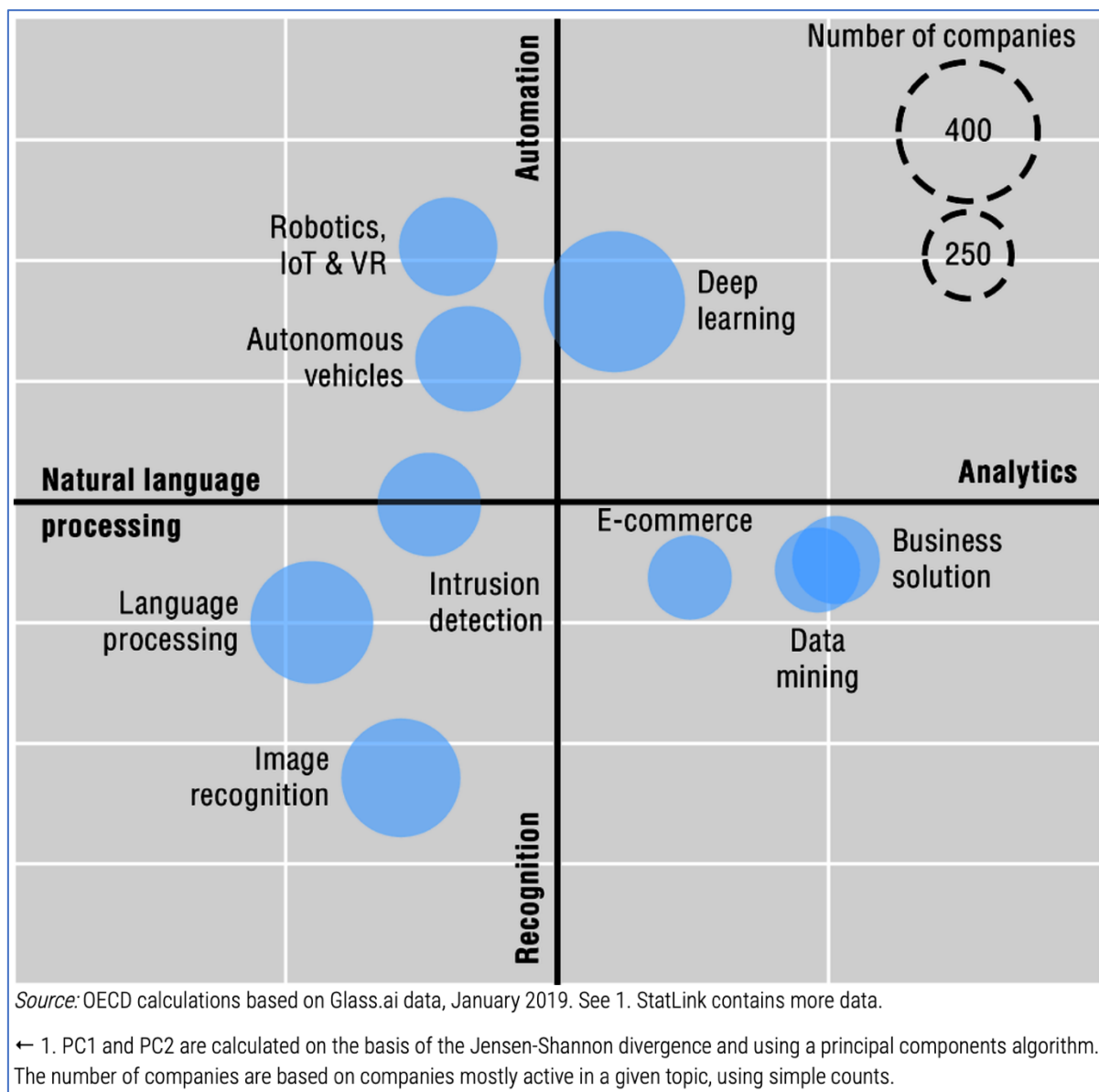
Tabel 21 zijn de belangrijkste woorden weergegeven.

Tabel 20. AI-topics en coördinaten (Glass.ai, 2019)

Topic	X (PC1)	Y (PC2)	Aantal bedrijven
Business solution	0.20	-0.05	253
E-commerce	0.10	-0.06	243
Language processing	-0.18	-0.10	353
Deep learning	0.04	0.16	406
Data mining	0.19	-0.05	245
Autonomous vehicles	-0.07	0.12	303
Image recognition	-0.11	-0.22	342
Robotics, IoT & VR	-0.08	0.21	283
Intrusion detection	-0.09	0.00	297

²⁷ OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264311992-en>

²⁸ Glass.ai (2019), *AI-related companies in the United Kingdom, by focus of activity, 2018*, <https://doi.org/10.1787/888933928388>



Figuur 11. OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264311992-en>. (p. 30-43)

Tabel 21. Meest belangrijke woorden per principal component (Glass.ai, 2019)

Top 25 words by principal components					
Term	PC1	Share	Term	PC2	Share
solution	0.06	8.7	company	0.04	6.5
business	0.05	7.8	world	0.04	4.6
datum mine	0.05	8.7	industry	0.03	5.1
predictive analytics	0.04	5.1	develop	0.03	6.7
help	0.04	6.2	digital	0.03	5.0
process	0.04	6.3	robotic	0.03	4.0
analytics	0.04	5.5	research	0.03	5.4
analysis	0.04	5.3	automation	0.03	3.7

management	0.04	5.5	deep learn	0.03	4.7
platform	0.04	5.8	team	0.03	4.8
model	0.03	4.1	focus	0.03	4.3
deliver	0.03	3.9	experience	0.03	6.1
tool	0.03	4.4	iot	0.03	3.3
allow	0.03	3.5	technique	0.02	4.0
information	0.03	4.3	drive	0.02	5.2
text	0.02	3.0	vehicle	0.02	4.0
client	0.02	3.1	sector	0.02	2.5
customer	0.02	6.1	automate	0.02	3.6
combine	0.02	3.4	change	0.02	2.3
capability	0.02	2.9	car	0.02	3.9
expertise	0.02	2.8	autonomous	0.02	3.9
enable	0.02	3.6	specialise	0.02	2.0
predictive	0.02	2.6	autonomous vehicle	0.02	4.0
market	0.02	6.1	lead	0.02	4.9
decision	0.02	2.2	emerge	0.02	1.9

Bottom 25 words by principal components

Term	PC1	Share	Term	PC2	Share
software	-0.06	10.1	range	-0.03	4.4
service	-0.05	8.3	recognition	-0.03	3.9
translation	-0.03	3.3	software	-0.03	10.1
application	-0.03	6.3	image	-0.03	3.4
machine	-0.03	4.8	anpr	-0.03	3.0
development	-0.03	6.2	offer	-0.03	4.7
offer	-0.02	4.7	computer vision	-0.03	2.8
nlp	-0.02	2.8	facial recognition	-0.03	3.6
cloud	-0.02	3.6	control	-0.03	3.2
compute	-0.02	2.6	service	-0.02	8.3
speech recognition	-0.02	2.5	video	-0.02	2.3
range	-0.02	4.4	time	-0.02	3.0
recognition	-0.02	3.9	product	-0.02	4.8
language	-0.02	1.9	application	-0.02	6.3
web	-0.02	2.3	integrate	-0.02	2.4
digital	-0.02	5.0	mobile	-0.02	2.6
understand	-0.02	2.7	camera	-0.02	2.1
anpr	-0.02	3.0	ocr	-0.02	1.8
network	-0.02	3.7	device	-0.02	2.2
computer vision	-0.01	2.8	translation	-0.02	3.3
facial recognition	-0.01	3.6	power	-0.02	3.9
security	-0.01	3.9	machine	-0.02	4.8
industry	-0.01	5.1	access	-0.02	2.2
human	-0.01	3.7	datum mine	-0.01	8.7
intrusion detection	-0.01	3.5	process	-0.01	6.3

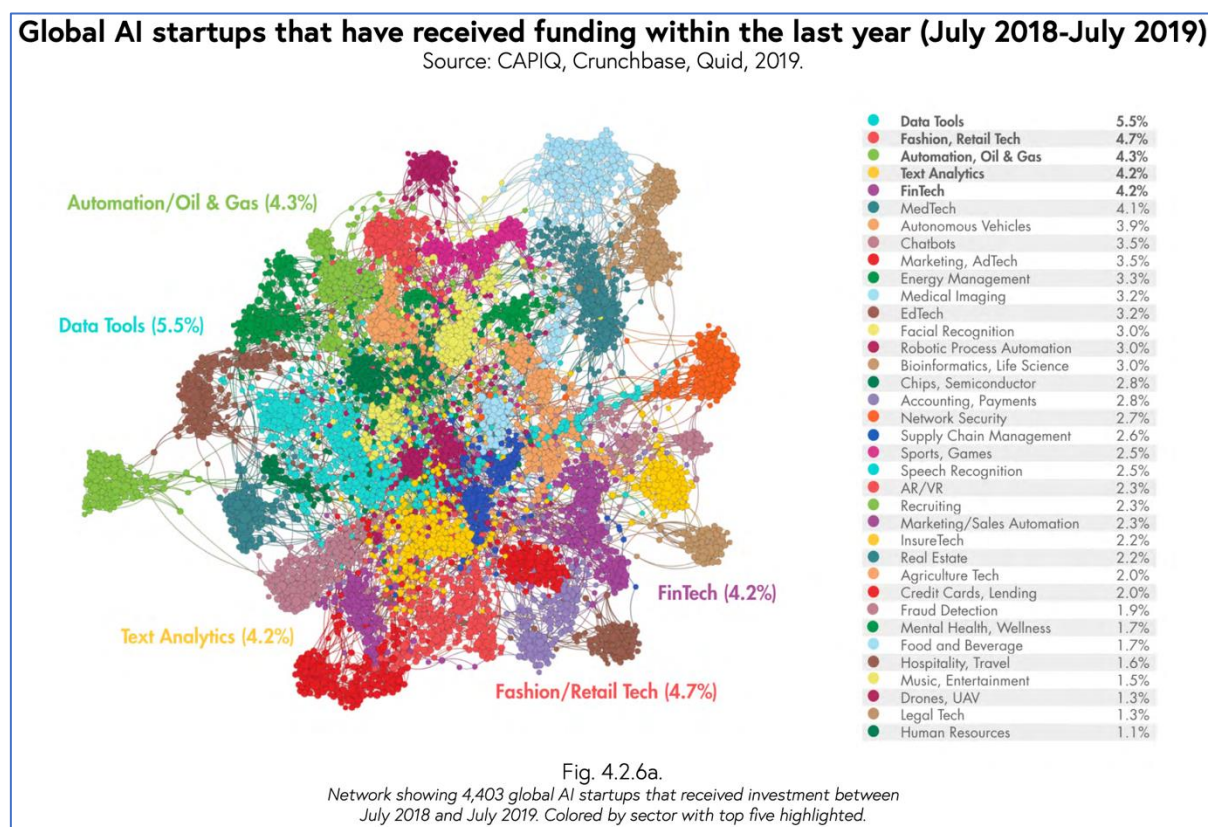
PC1 and PC2 are calculated on the basis of Jensen-Shannon divergence and principal components algorithm. The number of companies are based on companies mostly active in a given topic, using simple counts.

Stanford University, AI Index 2019 annual report.

Een ander onderzoek is van de Human-Centered Artificial Intelligence Institute (HAI) van Stanford University²⁹. In dit onderzoek is onder andere een unsupervised clustering uitgevoerd. Met bedrijfsdata van CAPIQ en Crunchbase zijn wereldwijd 4403 AI-startups geïdentificeerd uit een database van 1.8 miljoen profielen. Startups zijn geselecteerd op de termen “artificial intelligence”, “AI”, “machine learning” of “deep learning”. Als databron is een “beschrijving” gebruikt, het is niet bekend of dit om bedrijfswebsites gaat of om investeringsinformatie (Stanford University, 2019, p.246).

Van de 4.403 AI-startups komen 993 uit europa, en 55 uit Nederland.³⁰ De startups zijn geclusterd met software van NetBase Quid. De grootste focusgebieden zijn: data tools, fashion & retail tech, automation, oil & gas, tekst analytics, en fintech, zie Figuur 12.

In Europa zijn de grootste focusgebieden: fashion & retail tech, medtech, tekst analytics, marketing & adtech, autonomous vehicles, zie Figuur 13.



Figuur 12. Stanford University, “The AI Index 2019 Annual Report” (p.91).

²⁹ Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Nieves, “The AI Index 2019 Annual Report”, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, December 2019.

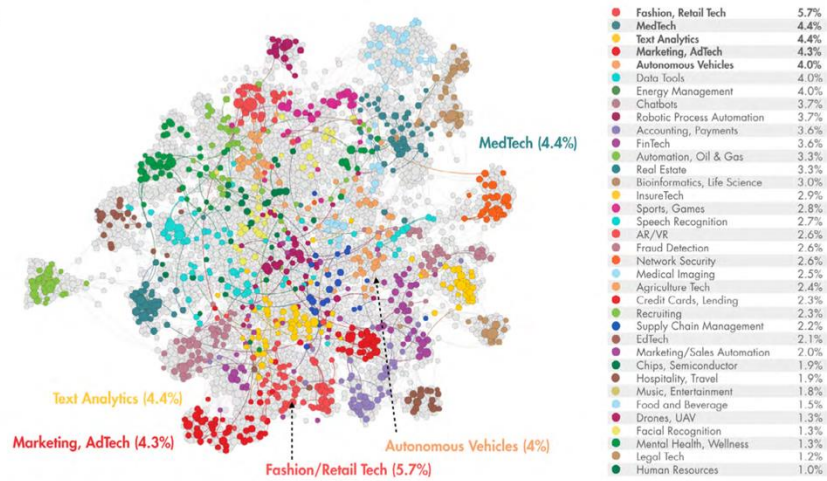
³⁰ 2019 AI Index Data, Chapter 4.2. CAPIQ, CB, Quid - Investment Activity.xlsx.

AI startups in the European Union: Companies that have received any funding within the last year, by focus area (July 2018-July 2019)

Source: CAPIQ, Crunchbase, Quid, 2019.

Fig. 4.2.6b.

Notes: Network highlighting 993 AI startups in Europe that received investment between July 2018 and July 2019. Colored by focus area with top five labeled.



Figuur 13. Stanford University, "The AI Index 2019 Annual Report" (p.92).