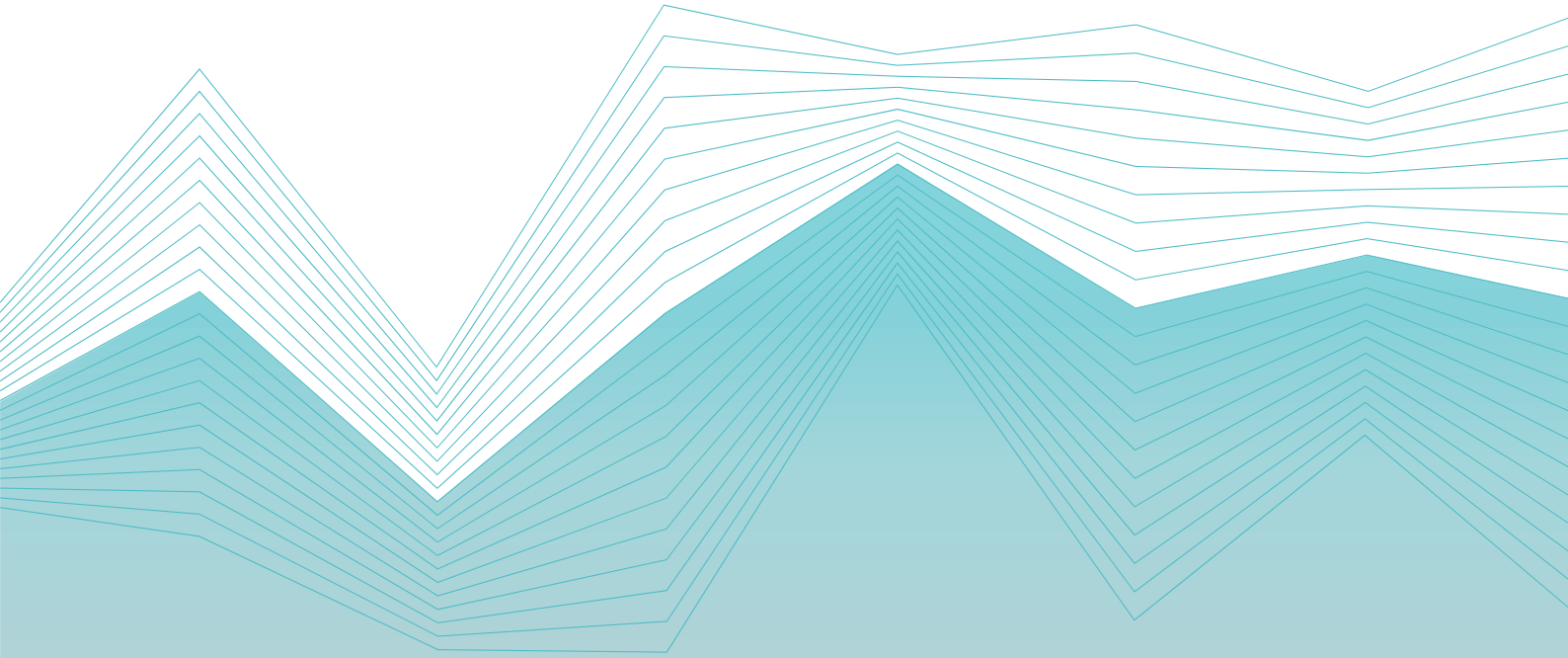
Analyse van uitstroomredenen op basis van topic modellering

Deelrapport 1: Machine learning als methode om verborgen thema's te ontdekken in grote hoeveelheden open antwoorden



Art van Schaaijk
Daan van Kooten

NIVEL
Kennis voor betere zorg

Het Nivel levert kennis om de gezondheidszorg in Nederland beter te maken. Dat doen we met hoogwaardig, betrouwbaar en onafhankelijk wetenschappelijk onderzoek naar thema's met een groot maatschappelijk belang. 'Kennis voor betere zorg' is onze missie. Met onze kennis dragen we bij aan het continu verbeteren en vernieuwen van de gezondheidszorg. We vinden het belangrijk dat mensen in staat zijn om deel te nemen aan de samenleving. Ons onderzoek draait uiteindelijk om de vraag hoe we de zorg voor de patiënt kunnen verbeteren. Alle onderzoeken publiceert het Nivel openbaar, dat is statutair vastgelegd.

Maart 2024

ISBN 978-94-6122-844-4

030 272 97 00
nivel@nivel.nl
www.nivel.nl

Voorwoord

Dit is een deelrapport over de eerste fase van een verdieping op de data van het landelijk uitstroomonderzoek van RegioPlus en Presearch. In deze fase staat de ontwikkeling van een innovatieve methode voor het analyseren van open antwoorden met betrekking tot de uitstroomredenen van medewerkers in de sector zorg en welzijn centraal. In dit rapport wordt een verdiepingsslag aangebracht op de analyse van data van het uitstroomonderzoek van RegioPlus en Presearch. Het begrijpen van de drijfveren achter personeelsverloop in deze cruciale sector is van onschatbare waarde voor het verbeteren van arbeidsomstandigheden en het waarborgen van een stabiele en kwalitatieve dienstverlening.

Het onderzoek is gebaseerd op een zorgvuldige analyse, innovatieve methodologieën en een nauwkeurige interpretatie van de verzamelde gegevens. Ons uiteindelijke doel is niet alleen het blootleggen van de redenen achter uitstroom, maar ook het bieden van inzichten die organisaties in staat stellen proactief te handelen en strategieën te ontwikkelen die het behoud van personeel bevorderen.

Uit dit rapport komen nog geen concrete resultaten. Desondanks biedt dit rapport inzichten in wat de open antwoordanalyse kan bijdragen aan de analyse van uitstroom van medewerkers in de sector zorg en welzijn. Dit rapport is slechts een eerste stap op weg naar een meer gefundeerd begrip van de uitdagingen waarmee deze sector wordt geconfronteerd. Desondanks is dit document een waardevolle bron voor verdere discussie en actie.

De auteurs.

Inhoud

Samenvatting

In het landelijk uitstroomonderzoek van RegioPlus en Presearch wordt vertrekkende werknemers in een open antwoordveld gevraagd naar toelichting bij hun vertrek. Om deze open antwoorden systematisch te analyseren wordt een methode ontwikkeld die in dit rapport wordt toegelicht.
Het onderzoek omvat twee fasen: de ontwikkeling van de analysemethode in fase 1 en de analyse van profielen van vertrekkende medewerkers in relatie tot uitstroom- en preventietopics in fase 2.
In de eerste fase van dit onderzoek ligt de nadruk op de ontwikkeling van de methode en een vergelijking met de bestaande uitstroomthema's. Onderzoeksvragen richten zich op het identificeren van topics en het vergelijken van open antwoorden met beantwoorde multiple-choice-vragen over uitstroomredenen.

De onderzoeksmethode start met data-voorbereiding, een woordfrequentieanalyse en het creëren van een document-term-matrix. Door vervolgens Latent Dirichlet Allocation (LDA) (een vorm van topic modellering), toe te passen kan meer inzicht verkregen worden in de uitstroom van medewerkers. Deze techniek maakt het mogelijk maken om in grote hoeveelheden tekst verborgen verbanden te ontdekken. Topics werden door LDA ingedeeld op een basis waarbij woorden die vaak samen worden genoemd gezien worden als bij elkaar horend. Het proces maakt gebruik van LDAvis voor een interactieve visualisatie, waarbij de relevantie van termen binnen topics kan worden aangepast met behulp van een slider. In het onderzoek waren de antwoorden van 1671 respondenten bruikbaar voor de topicmodellering van uitstroomredenen en 1226 voor preventiemogelijkheden. De analyse omvatte het identificeren van topics en de meest voorkomende woorden, en richt zich op zowel uitstroomredenen als preventiemogelijkheden, met het identificeren van respectievelijk 11 en 6 topics. Ondanks dat uit deze fase nog geen concrete uitkomsten naar voren komen, geeft de ontwikkeling van de methode een beeld van wat de open antwoordanalyse kan opleveren.

De congruentieanalyse analyseerde in hoeverre de uitstroomtopics overeenkwamen met thema's uit de multiple-choice-vragen. De vergelijking tussen LDA-geïdentificeerde topics en de door respondenten geselecteerde uitstroomthema's toonde enige overeenkomst, maar ook verschillen. Omdat respondenten vaak meerdere thema's benoemden in hun open antwoord was het scharen van dit antwoord onder 1 topic ingewikkeld. Hierdoor ontstond er overlap tussen de verschillende topics. De inhoudsanalyse onthulde dat bepaalde subredenen en nuances werden gemist in de uitstroomredenen in de vragenlijst, zoals het niet nakomen van afspraken over doorgroeien. Dit alles heeft geresulteerd in zowel meerdere overwegingen ter aanpassing van de vragenlijst, als een verbetering van de methode in fase 2.
Zo komen verschillende vertrekredenen naar voren die mogelijk niet terug te vinden zijn in de multiple-choice vragen in de vragenlijst van het uitstroomonderzoek. Daarnaast is een vaakgenoemde reden als verhuizing niet terug te vinden als hoofdthema waardoor deze moeilijker vindbaar is voor respondenten. Voor veel voorkomende thema's is het verstandig om deze op te nemen als hoofdthema. Enkele suggesties hiervan betreffen: Reistijd, Werksfeer, Werktijden en Waardering.

In Fase 2 is meer data beschikbaar waardoor meer topics kunnen worden gevormd. Zo komen thema's die vaak samen worden genoemd in een onderscheidend topic terecht. Om de methode te verbeteren wordt de suggestie gedaan om de document-term matrix (DTM) te integreren in deze analyse. Dit biedt de mogelijkheid om op basis van vaak voorkomende woorden en

woordcombinaties meerdere thema's toe te kennen wanneer respondenten ook meerdere redenen opgeven in hun antwoord. Op deze wijze kunnen ook bepaalde woorden die synoniemen van elkaar zijn samengevoegd worden zoals fysiek en lichamelijk, en loon en salaris zodat deze binnen hetzelfde thema vallen.

Ten slotte toonde de relatieanalyse tussen uitstroom- en preventietopics behoorlijke overlap. Ook hier werd dit bemoeilijk door de aard van de open antwoorden. Verbeteringen in de methode, waardoor er minder overlap tussen topics ontstaat, dragen bij aan het verbeteren van de zichtbaarheid van de relatie tussen uitstroomredenen en preventiemogelijkheden. Dit zal resulteren in meer inzicht in welke preventiemaatregelen effectief zijn bij de verschillende uitstroomredenen.

De LDA-methode in het onderzoek biedt diepgaand inzicht in uitstroomredenen, met toegevoegde waarde van de multiple-choice antwoorden. De ontwikkelde methode is bruikbaar, maar behoeft nog wel verbetering voor de volgende fase. Ondanks dat aanpassingen nodig zijn, heeft de huidige analyse al belangrijke inzichten opgeleverd over de uitstroom in de zorg- en welzijnssector. Een voorbeeld hiervan is dat uitstroom vaak niet het gevolg is van een enkele oorzaak, maar vaak meerdere redenen kent. De volgende fase richt zich op het koppelen van uitstroom- en preventietopics aan persoonskenmerken en het ontdekken van associaties. Ondanks de complexiteit van individuele antwoorden, biedt de LDA-methode een beter begrip van de uitstroom, waardoor preventieve acties gerichter kunnen worden ondernomen. In fase 2 wordt dit duidelijker vormgegeven waarbij concrete aanbevelingen voor de preventie van uitstroom worden gedaan.

1 Inleiding

De arbeidsmarkt Zorg en Welzijn kent grote uitdagingen. Op dit moment is er een tekort ontstaan aan zorgprofessionals. Daarbij stijgt de zorgvraag door de toenemende vergrijzing en stijgende complexiteit harder dan het aanbod. Het is hierdoor de verwachting dat de tekorten in de nabije toekomst alleen maar zullen toenemen van 3,9% in 2023 tot 9,5% in 2032. [1] Dit leidt tot hogere werkdruk en onvoldoende capaciteit om aan de zorgvraag te voldoen.

Omdat de instroom ontoereikend is om deze ontwikkeling op te vangen, is het van belang om de uitstroom tegen te gaan. Een significant deel van de uitdaging in de zorgsector heeft betrekking op de hoge uitstroom van medewerkers bij hun werkgever, die momenteel meer dan 20% bedraagt. [2] De netto uitstroom uit de sector bedraagt ruim 10%. Dat wil zeggen dat bij werkgevers 20% van de werknemers per jaar vertrekt. Deze stoppen of gaan dan elders aan de slag. De helft van deze uitstromers verlaat de sector zorg en welzijn. Deze voortdurende uitstroom brengt niet alleen operationele verstoringen met zich mee, maar ook aanzienlijke financiële lasten. Het vervangen van een werknemer wordt geschat op 16 tot 20% van hun jaarsalaris Boushey (2012). Bij een organisatie van 100 medewerkers komt dit per jaar neer op 3 tot 4 volledige jaarsalarissen. Geld dat anders besteed kan worden om de werkdruk te verlichten door meer capaciteit te creëren. In deze context biedt het verminderen van de uitstroom aanzienlijke voordelen: het helpt niet alleen de tekorten te beperken, maar verlaagt ook de kosten voor werkgevers. Het behoud van zorgprofessionals is een essentiële factor voor het garanderen van een stabiele en effectieve gezondheidszorg.

1.1 Uitstroomonderzoek

In het landelijk uitstroomonderzoek van RegioPlus en Presearch wordt bij werkgevers vertrekkende werknemers in de sector zorg en welzijn gevraagd naar toelichting bij hun vertrek. Dit biedt mogelijk inzichten en aanknopingspunten om preventief beleid te voeren om de uitstroom in de sector zorg en welzijn terug te dringen. Allereerst dient hiervoor meer inzicht te worden verkregen in de uitstroomredenen van medewerkers. In een later stadium kan dit gekoppeld worden aan preventieve handelingen die de werkgever had kunnen treffen om dit vertrek te voorkomen.

1.2 Vernieuwende methoden

In de laatste jaren heeft de opkomst van kunstmatige intelligentie een breed scala aan mogelijkheden geboden in verschillende domeinen. AI-systemen hebben indrukwekkende prestaties geleverd in taken zoals beeldherkenning en natuurlijke taalverwerking. Binnen de context van natuurlijke taalverwerking heeft de vooruitgang van Latent Dirichlet Allocation (LDA) de capaciteit vergroot om verborgen thema's in grote tekstcorpora te ontdekken. Hierdoor is LDA, een methode voor topic modellering, een bruikbaar instrument geworden om informatie te extraheren uit ongestructureerde data zoals grote hoeveelheden open tekstantwoorden. Topic modellering, en

[1] Bron: Prognosemodel Zorg en Welzijn, Nieuwbeleidscenario, cijfers 2023

[2] CBS Staline: Werknemers met een baan in de zorg en welzijn & Mobiliteit van werknemers

meer specifiek LDA wordt in toenemende mate gebruikt bij het analyseren van grote hoeveelheden tekst (Goudzarvand, 2019; Li, 2019; Westrupp, 2022; Inoue, 2023).

Latent Dirichlet Allocation

Het onderzoek maakt gebruik van Latent Dirichlet Allocation (LDA), een methode voor topic modellering op basis van machine learning, om de antwoorden op open vragen over uitstroomredenen te analyseren.

De groeiende kennis van LDA biedt kansen om verdere analyse uit te voeren op tekstantwoorden van het uitstroomonderzoek van RegioPlus en Presearch. In dit uitstroomonderzoek wordt aan vertrekkende medewerkers gevraagd wat de reden is voor hun vertrek. Door de antwoorden te analyseren kan meer inzicht gecreëerd worden in de verschillende redenen achter de uitstroom. Wanneer deze redenen gecategoriseerd worden, biedt dit mogelijkheden tot gerichte preventieve maatregelen. In een eerder onderzoek met verdieping naar uitstroomredenen in de RegioPlus regio Utrecht kwam bijvoorbeeld naar voren dat jongere medewerkers over het algemeen andere uitstroomredenen opgeven dan oudere werknemers (Lemmelijn, 2023).

1.3 Onderzoeksvragen

Dit leidt tot de volgende onderzoeksvragen:

1) Welke thema's zijn er te onderscheiden in de antwoorden op de twee open vragen;
    - Een open vraag naar de reden van uitstroom
    - Een open vraag naar hoe de werkgever dit vertrek had kunnen voorkomen
2) In welke mate komen de topics die uit de analyse voortkomen overeen met de huidige thema's uit de multiple-choice-vragen?
3) In welke mate komen de redenen binnen de uitstroomtopics die uit de analyse voortkomen overeen met de huidige redenen binnen de uitstroomtopics uit de multiple-choice-vragen?
4) Wat is op basis van de antwoorden op bovenstaande vragen de toegevoegde waarde om deze methode toe te passen om landelijke data te analyseren?

2 Methoden

Hoewel er reeds analyses zijn uitgevoerd met betrekking tot de uitstroomredenen, blijft een groot deel van de beschikbare data onderbenut. Eerdere analyses waren gebaseerd op meerkeuzevragen uit het Landelijk uitstroomonderzoek van RegioPlus en Presearch in de RegioPlus regio Utrecht (Lemmelijn, 2023). Het benutten van open antwoorden in combinatie met technieken zoals LDA, biedt nieuwe mogelijkheden.

Door de toepassing van LDA en de analyse van open antwoorden is het mogelijk om actuele uitstroomtopics en nuances in de uitstroomredenen beter te begrijpen. Dit levert waardevolle inzichten op die niet worden gestuurd door vooraf gedefinieerde categorieën. Het biedt ruimte voor respondenten om extra informatie en motivering te geven, wat tot een dieper begrip van hun uitstroomredenen leidt.

2.1 Opzet van het onderzoek

In het onderzoek worden open antwoorden geanalyseerd door middel van Latent Dirichlet Allocation (LDA). Deze methode staat bekend als "topic modellering." LDA is een statistische techniek die wordt gebruikt om de latente (verborgen) topics in een verzameling tekstuele gegevens te ontdekken. Het identificeert clusters van woorden die vaak samen voorkomen in documenten en wijst toe welke woorden tot welk topic behoren. Hierdoor kunnen onderzoekers de belangrijkste thema's of onderwerpen in een grote hoeveelheid tekst identificeren en begrijpen. Topic modellering met LDA wordt veel gebruikt in tekstanalyse, informatieherwinning en natuurlijke taalverwerking om inzicht te krijgen in de inhoud van documenten en om patronen en trends te ontdekken.[3,4,5]

Het onderzoek bestaat uit twee fasen, welke een vergelijkbare start kennen: In de eerste fase wordt de data gebruikt uit het uitstroomonderzoek van RegioPlus en Presearch die is verzameld in de RegioPlus-regio Utrecht. In deze fase wordt de onderzoeksmethode ontwikkeld en toegepast op een kleinere groep antwoorden. In eerste instantie bestond de dataset uit 3399 respondenten waarvan een selectie is gemaakt. Hierbij wordt bepaald in welke mate de ontwikkelde methode geschikt is, en of de huidige thema's in de multiple-choice vragen passend zijn bij de open antwoorden die gegeven worden door de respondent. Na de eerste fase vindt een evaluatie plaats om te bepalen of deze methode ook toegepast wordt op de volledige Nederlandse dataset.

In de tweede fase wordt de data gebruikt uit het uitstroomonderzoek van RegioPlus en Presearch die landelijk is verzameld. Deze data wordt door RegioPlus beschikbaar gesteld aan het Nivel. Er wordt opnieuw een afweging gemaakt van de gebruikte methoden en scripts. Deze worden vervolgens

[3] Inoue, M., Fukahori, H., Matsubara, M., Yoshinaga, N., & Tohira, H. (2023). Latent Dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing. Japan Journal of Nursing Science, 20(2), e12520.

[4] Shah, A. M., Yan, X., Qayyum, A., Naqvi, R. A., & Shah, S. J. (2021). Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach. International Journal of Medical Informatics, 149, 104434.

[5] Li, Y., Rapkin, B., Atkinson, T. M., Schofield, E., & Bochner, B. H. (2019). Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. Quality of Life Research, 28, 1441-1455.

aangepast om ze passend te maken bij de grotere dataset. Daarna wordt een verdiepende analyse uitgevoerd op het uitstroomtopic door profielen op te stellen op basis van enkele persoonskenmerken die in overleg met de begeleidingscommissie worden geselecteerd. Hoe dit onderzoek wordt uitgevoerd, en wat de verwachte resultaten zijn, wordt hieronder weergegeven.

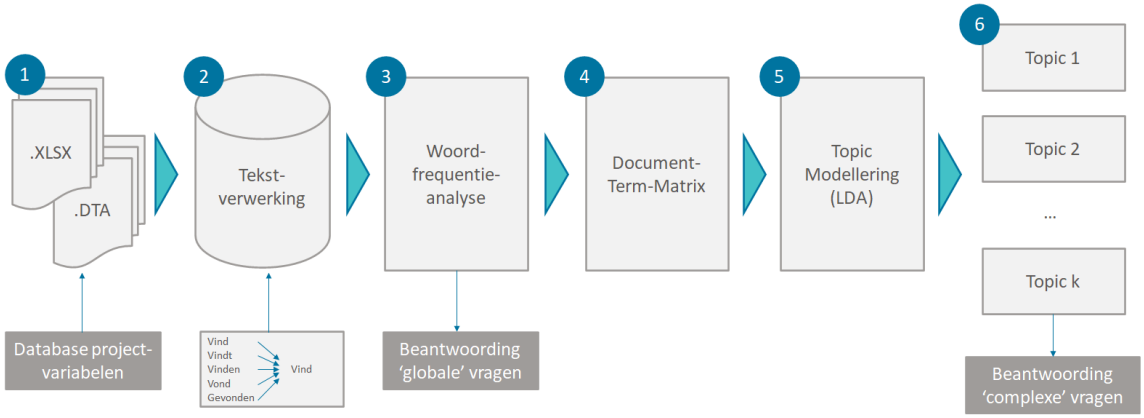
2.2 Onderzoeksmethode

Voor dit onderzoek worden de volgende variabelen uitgewerkt in de vrije tekstanalyse;
Open antwoorden op de vragen:

1) Wat is de reden dat je weggaat bij [organisatie]?
2) Wat had [organisatie] anders kunnen doen om jouw vertrek te voorkomen?

Om tot meer inzicht te komen in de antwoorden worden verschillende technieken gebruikt om het groot aantal open antwoorden in te delen en te categoriseren. In het proces van het analyseren van tekstdata zijn er verschillende stappen die gevolgd moeten worden om zinvolle inzichten te verkrijgen. Voor onderstaande analyse, wordt gebruikgemaakt van R en R-Studio waarin verschillende scripts worden geschreven in de bijbehorende programmeertaal. Deze tools worden veel gebruikt voor tekstverwerking en bieden een solide basis voor de uitvoering van onderstaande stappen. Allereerst wordt gestart met het voorbereiden van de dataset.

Figuur 1 Stappenplan onderzoeksmethode



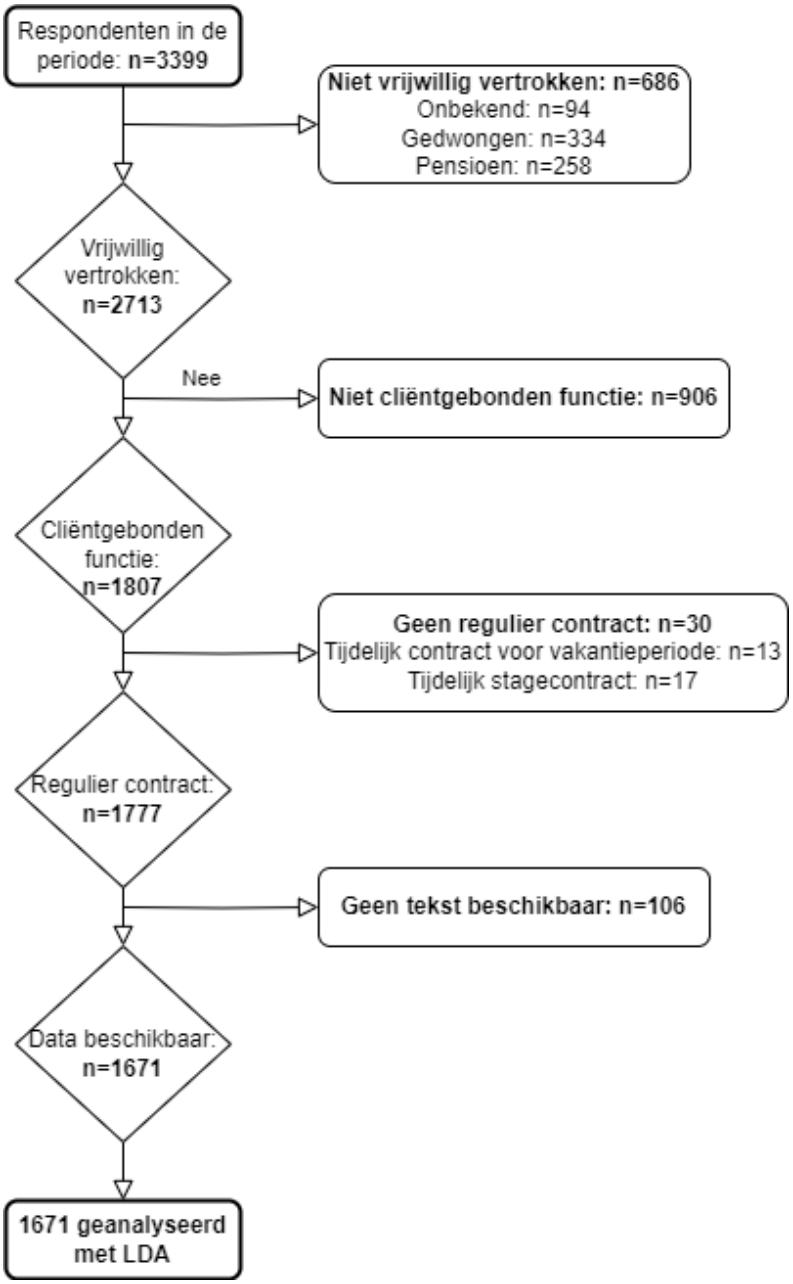
2.3 Stap 1: Gegevensverzameling en -voorbereiding

In de eerste stap van dit proces is het essentieel om de benodigde gegevens uit de dataset te selecteren. Het betreft hier een grote dataset die wordt opgesplitst in twee verschillende datasets: Een dataset met daarin het open antwoord op de vraag: 'Wat is de reden dat je weggaat bij [organisatienaam]?'. En een tweede dataset met daarin het open antwoord op de vraag: 'Wat had [organisatienaam] anders kunnen doen om jouw vertrek te voorkomen?'. Daarbij zit een ID-variabele om de resultaten met betrekking tot het uiteindelijk geselecteerde topic weer aan de dataset te koppelen voor verdere analyse. Vervolgens wordt de data geïmporteerd in een geschikte tekstverwerkings- en analysetool, in dit geval R en R-studio.

Data-selectie

In samenspraak met arbeidsmarktonderzoekers van RegioPlus organisaties is besloten tot het analyseren van de data van een specifieke groep werknemers uit de sector zorg en welzijn die zijn uitgestroomd. Respondenten die worden meegenomen hebben een vragenlijst ingevuld tussen 14-2-2022 en 31-1-2024 [6]. Respondenten worden geïncludeerd wanneer het vrijwillig vertrek betrof in cliëntgebonden functies. Daarnaast zijn tijdelijk vakantie- en stagecontracten geëxcludeerd uit de analyse. In figuur 2 is een flowchart weergegeven van de inclusie.

Figuur 2 Flowchart van ge-in- en excludeerde respondenten op basis van uitstroomredenen:



[6] De datum 14-2-2022 is gekozen omdat vanaf deze periode een vernieuwde vragenlijst is verstuurd naar de deelnemers.

Van de personen die binnen de inclusiecriteria vielen, zijn de antwoorden gebruikt voor de analyse. Daarvan zijn de teksten bewerkt zoals aangegeven in de volgende stappen. Na bewerking bleven vervolgens 1671 personen over waarvan de data geschikt was voor analyse. Van de antwoorden van 106 respondenten bleef geen data over na de bewerking van de tekst in de volgende stappen van de voorbereiding van de data.

Met betrekking tot het preventietopic zijn 1226 antwoorden van respondenten bruikbaar. Antwoorden die starten met het woord ‘niks’ of ‘niets’ zijn verwijderd uit de analyse omdat de werkgever in dit geval geen mogelijkheid had om het vertrek te voorkomen. Dit blijkt een valide methode om te voorkomen dat de positieve aspecten die werknemers noemen worden gecategoriseerd als preventiemogelijkheid. Van de 242 antwoorden die begonnen met ‘niks’ of ‘niets’ kwam uit slechts twee antwoorden nog een mogelijkheid tot preventie. In alle andere gevallen wanneer het antwoord startte met het woord ‘niks’ of ‘niets’, volgde daarna redenen waar de werkgever inderdaad niets aan kon doen zoals: verhuizing, reistijd, privéomstandigheden, van 2 naar 1 werkgever, bewust gekozen voor ander werk, tijd voor verandering of het gaan volgen van een vervolgopleiding. Een deel van de respondenten geeft zelfs aan na een periode weer terug in dienst te willen bij de werkgever.

2.4 Stap 2: Tekstverwerking

Na het importeren van de dataset begint de tekstverwerking. Hierbij worden verschillende stappen doorlopen om de tekstdata te optimaliseren voor verdere analyse. Allereerst worden speciale tekens, overbodige spaties en andere niet-relevante karakters uit de tekstdata verwijderd. De tekstdata wordt vervolgens omgezet in kleine letters om variaties in hoofdletters te normaliseren. Hierna wordt tokenisatie toegepast om de tekst op te splitsen in individuele woorden en/of woordgroepen. Daarnaast worden stopwoorden zoals "de, het, een, op, bij, en” etc. verwijderd om veelvoorkomende, maar weinig informatieve woorden te elimineren. Tot slot wordt stemming of lemmatisering uitgevoerd om woorden terug te brengen tot hun basisvorm, bijvoorbeeld van "lopende" naar "lopen". Deze stemming vindt plaats door het schrijven en uitvoeren van scripts.

Tokens

In topic modellering verwijst een token naar een elementair stukje van een tekst, dat kan worden geïdentificeerd als een eenheid.

Uit de tekst zijn nog specifieke woorden en woordcombinaties verwijderd. Deze zijn verwijderd omdat ze niets toevoegen of het beeld vertroebelen. Daarin wordt onderscheid gemaakt tussen tokens en zinsdelen (phrases).
De verwijderde tokens als het gaat om de uitstroomredenen zijn: 'weinig', 'heel', 'nieuwe', 'gevonden', 'beter', 'teveel', 'minder', 'betere', 'gaan', 'geven', 'goed', 'moeten', 'maken', 'nemen', 'verder', 'binnen', 'alleen', 'wel', 'vanuit'. Alleen het losse token is verwijderd. Combinaties met bovenstaande woorden niet. Het losse token ‘weinig’ is dus verwijderd, maar ‘weinig uitdaging’ blijft bestaan.

Daarnaast zijn bepaalde zinsdelen verwijderd, wat betekent dat deze ook niet meer terugkomen in de te analyseren teksten. Hieronder vallen ook de persoonsnamen en namen van organisaties. Qua zinsdelen is bijvoorbeeld ‘andere uitdaging’ eruit gefilterd. Dit betekent dat de combinatie ‘andere

uitdaging' of 'andere uitdaging gevonden' niet meer voorkomt. De losse tokens 'uitdaging' en 'andere' blijven wel in de dataset. De combinatie 'andere uitdaging' is verwijderd omdat dit gezien wordt als synoniem van 'nieuwe baan', wat verder geen informatie over de uitstroomreden bevat. Het woord 'uitdaging' is behouden omdat het mogelijk is dat een respondent vertrekt omdat 'uitdaging mist' in de huidige baan. Een overzicht van de verwijderde zinsdelen is te vinden in Bijlage A.

Er is een beoordeling gemaakt van de maximale tokenlengte. Tokens met hogere lengte bevatten meer informatie. Het token 'dichter_huis', met lengte 2) is dus specifieker en meer inhoudelijk dan het token 'huis'. Door de woordverwijderingen komen er minder vaak langere tokens voor. Daarnaast wordt de laagste 1% van alle tokens verwijderd omdat deze te specifiek zijn, en slechts iets zegt over een erg klein deel van de inhoud van een topic. Dit heeft ook tot gevolg dat de minst vaak voorkomende token-3 woorden worden verwijderd. Er zijn nog enkele tokens met lengte 3 die informatie bevatten en vaak genoeg voorkomen om in de analyse te blijven. Er is gekozen om deze tokens te behouden en de maximale tokenlengte op 3 te zetten. Tokens met een lengte van 4 woorden kwamen niet vaak genoeg voor om relevant te blijven.

De verwijderde tokens voor preventiemogelijkheden zijn dezelfde als voor de uitstroomredenen. Voor de analyse van de preventiemogelijkheden worden andere zinsdelen verwijderd. De verwijderde zinsdelen voor de preventiemogelijkheden zijn te vinden in Bijlage A. Ook hier zijn de namen van organisaties en persoonsnamen verwijderd die vaker voorkwamen en terug te zien waren. Daarnaast zijn de antwoorden die begonnen met de woorden: 'niks' of 'niets' verwijderd omdat hierna vaak werd verteld dat: de werknemer of het team juist niet meespeelden in het vertrek, of er niet een mogelijkheid tot het voorkomen van het vertrek was. Het sentiment van deze antwoorden is over het algemeen positief. Dat wil zeggen dat de werknemer na de woorden 'niks' of 'niets' de positieve punten benoemde van het werken voor de voormalig werkgever.

2.5 Stap 3: Woordfrequentieanalyse + Woordwolk

Na de tekstverwerking volgt de woordfrequentieanalyse. Hierbij wordt de frequentie van elk woord in de dataset berekend om inzicht te krijgen in de meest voorkomende termen. Zeldzame woorden of woorden met lage frequentie worden geïdentificeerd en vervangen of verwijderd om ruis in de gegevens te verminderen. Deze frequentiegegevens worden vervolgens gebruikt om een woordwolk en/of staafgrafiek te genereren, waarbij de meest voorkomende woorden visueel worden weergegeven.

2.6 Stap 4: Document-Term-Matrix/Term-Document-Matrix

In stap 4 wordt een term-documentmatrix opgebouwd. Op basis van deze matrix kan worden bepaald welke woorden in een antwoord voorkomen, en welke combinaties er bestaan. In deze matrix vertegenwoordigt elke rij een document, of in dit geval een antwoord van een respondent, en elke kolom geeft de frequentie van een woord in dat 'document' weer. Dit is een cruciale stap voor verdere analyse, omdat het de basis vormt voor technieken zoals topic modellering. Een DTM is kort gezegd een document dat telt hoe vaak een woord in een document voorkomt.

Een kort voorbeeld van hoe een document-term-matrix eruit ziet op basis van 3 respondenten, nog zonder woorden te verwijderen:

Document 1: De kat is zwart
Document 2: De kat is wit
Document 3: De kat is zwart en de hond is wit

Tabel 1 Fictieve document matrix ter illustratie met 3 respondenten

	de	kat	cavia	hond	en	is	zwart	wit
Document 1	1	1	0	0	0	1	1	0
Document 2	1	1	0	0	0	1	0	1
Document 3	2	1	0	1	1	2	1	1

In dit fictieve voorbeeld worden slechts enkele respondenten getoond, met korte antwoorden die veel op elkaar lijken. In de dataset in deze fase betreft het vele respondenten, met langere uiteenlopende antwoorden.

2.7 Stap 5: Topic Modellering

De laatste stap omvat topic modellering, waarbij het Latent Dirichlet Allocation (LDA) algoritme wordt geïmplementeerd. Dit algoritme identificeert topics en bijbehorende beschrijvende woorden in de feedbackteksten. Hierdoor kan men patronen en thema's ontdekken in grote hoeveelheden ongestructureerde tekst door deze automatisch te laten organiseren, wat nuttig kan zijn voor het begrijpen van de inhoud en het verkrijgen van waardevolle inzichten. In deze stap wordt gebruik gemaakt van LDAvis: Een methode voor het visualiseren en interpreteren van onderwerpsmodellen.

Corpus

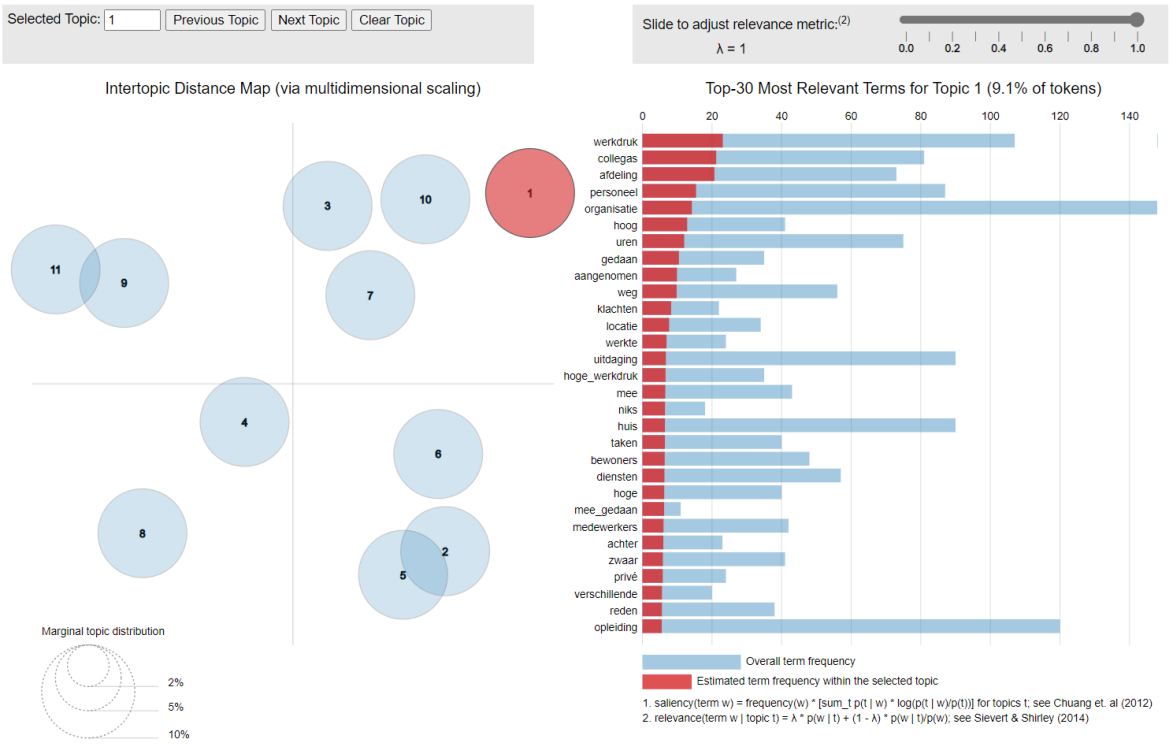
Voor LDAvis wordt een corpus gebruikt als input voor het trainen van het LDA-model. Dit corpus bevat de tekstuele gegevens waaruit het model de verborgen onderwerpen zal ontdekken. In dit geval gaat het om alle ingevoerde tekst.

Het proces van visualisatie van de LDA topics omvat verschillende stappen. Eerst wordt een LDA-model getraind op een corpus van tekstuele gegevens, waarbij de onderliggende verborgen onderwerpen in de dataset worden geïdentificeerd. In de visuele weergave van de topics zijn verschillende onderdelen te onderscheiden. LDAvis genereert een interactieve visualisatie die bestaat uit een plot waarbij elk cirkelsegment een topic vertegenwoordigt. De afmeting van elke cirkel staat voor de prevalentie van de inhoud in de corpus, terwijl de relatieve afstand tussen de cirkels de gelijkenis tussen de clusters weergeeft op basis van de frequentie van overlappende woorden. Cirkels die dichterbij elkaar op de kaart staan, duiden erop dat de bijbehorende onderwerpen gemeenschappelijke termen delen, wat wijst op semantische gelijkenis. Overlappende cirkels geven aan dat documenten in de corpus termen bevatten uit meerdere onderwerpen die gelijkenis met elkaar vertonen. De assen waarop de cirkels staan zijn hulplijnen om de onderlinge afstand in te schatten. Daarnaast biedt LDAvis een zogenaamde 'top terms'-weergave die de belangrijkste woorden voor elk onderwerp toont en een 'document-term matrix' die de verdeling van onderwerpen in individuele documenten weergeeft. Dit alles stelt gebruikers in staat om intuïtief en interactief door de onderwerpen te navigeren, hun relaties te begrijpen en diepgaande inzichten te verkrijgen in de structuur van de gegevens die door het LDA-model zijn onthuld. Het biedt

interactieve visualisaties om de onderwerpen en hun relaties in de modellen te verkennen. Het helpt bij het analyseren en communiceren van complexe resultaten.

LDAvis

LDAvis is een instrument dat is ontworpen voor het visualiseren en interpreteren van onderwerpen in een topic model, en biedt interactieve visualisaties om de onderwerpen en hun relaties in de modellen te verkennen. Het helpt bij het analyseren en communiceren van complexe resultaten.



In deze topic modellering worden woorden die vaak bij elkaar genoemd worden in een corpus gezien als een cluster. Als input is gekozen dat de modellering wordt vormgegeven in 11 uitstroomtopics, gelijk aan het aantal thema's in de multiple-choice-vraag die wordt gesteld in het uitstroomonderzoek van RegioPlus en Presearch. De reden hierachter is dat op deze manier bepaald kan worden in welke mate de topics die naar voren komen uit de topic modellering overeenkomen met de reeds bestaande thema's. Voor de preventietopics is gekozen voor 6 topics. Deze keuze is gebaseerd op de hoeveelheid overlap die er ontstond tussen de topics wanneer voor meer topics was gekozen.

Aan de rechterkant van de visual staat een staafdiagram met "most relevant terms topics". Het staafdiagram toont de prevalentie van elk woord in de gehele corpus (blauw) en binnen het geselecteerde onderwerp (rood). In de LDA analyse kan op twee manieren worden gekeken naar de relevantie:

1. De meest voorkomende woorden binnen een onderwerp zijn relevant voor het onderwerp (λ =1). Dus de woorden met het grootste rode balkje.

2. De woorden die het meest voorkomen in het onderwerp in vergelijking met het corpus zijn relevant voor het onderwerp. (λ=0) Dus waar het rode balkje een groot deel onderdeel uitmaakt van het blauwe balkje.

De interactieve slider 'adjust relevance matrix' geeft de mogelijkheid om de λ waarde te wijzigen. Hiermee kan het gewenste detailniveau in de weergave van de topics aangepast worden wat een dieper inzicht kan geven in de subonderwerpen of de fijnere nuances van dat topic. De slider met de lambda in LDAvis is een interactief onderdeel van de visuele interface. Door deze slider aan te passen verandert de weergave van de meest prominente termen die aan een bepaald topic zijn gekoppeld. Met een hogere lambda-waarde worden de meest voorkomende termen in een topic getoond, terwijl met een lagere lambda-waarde de meer specifieke, minder voorkomende termen worden weergegeven.

Bij de keuze voor optie 1 wordt gekeken naar welk woord het vaakst voorkomt binnen een topic. Dit woord staat bovenaan weergegeven. Bij optie 2 wordt gekeken naar hoe vaak het token voorkomt in het geselecteerde topic ten opzichte van het totaal aantal keren dat het token voorkomt. Dit is te zien als een percentage. Hoe hoger dit percentage, hoe relevanter het woord is voor het desbetreffende topic. Het is dus een woord dat erg typerend is voor dat topic. Woorden die erg relevant lijken in optie 2 hoeven niet relevant te zijn volgens optie 1. Bijvoorbeeld een woord dat maar één keer voorkomt in alle antwoorden, zal dus 100% scoren als wordt uitgegaan van relevantie onder optie 2. Omdat het woord slechts één keer voorkomt, zal het niet relevant zijn volgens optie 1. Om te komen tot de inhoud van de topics dient er een middenweg te worden gevonden tussen beide opties. Door een tussenwaarde te kiezen kan een onderzoeker zelf het gewenste detailniveau bepalen en bestuderen.

2.8 Stap 6: Analyse

De uitkomsten van LDA topic modellering worden gebruikt voor het categoriseren van de open antwoorden van de respondenten op de vragenlijst. Op basis van de antwoorden van iedere respondent wordt één van de topics die in de vorige stap zijn ontwikkeld toegekend de respondent op basis van het ID-nummer. Door middel van een script wordt dit teruggekoppeld aan de oorspronkelijke dataset. Dit maakt het mogelijk dat wordt bepaald in welke mate de inhoud van de topics overeenkomt met de geselecteerde uitstroomtopics van de respondent in de multiple-choice vraag in het uitstroomonderzoek

De gebruikte methode wordt geëvalueerd waarbij wordt bepaald welke mogelijke aanpassingen kunnen worden doorgevoerd in fase 2 van het project, en de vragenlijst die gebruikt wordt in het landelijke uitstroomonderzoek. Daarnaast wordt de meerwaarde van het gebruik van deze methode voor de specifieke data in het uitstroomonderzoek bepaald.

Het onderzoek bestaat uit 2 fasen. De eerste fase wordt gezien als ontwikkelfase voor de methode van analyse. Hierbij wordt bepaald of deze methode aanvullende inzichten teweeg brengt. Met de resultaten wordt in fase 1 nog geen verdiepende analyse uitgevoerd op basis van de persoons- en baan-kenmerken. In de tweede fase wordt bepaald welke profielen er te ontdekken zijn in vertrekkende medewerkers en hun uitstroomtopics, en wordt bepaald welke preventietopics aansluiten bij deze uitstroomtopics.

3 Resultaten

In dit hoofdstuk worden de antwoorden op de onderzoeksvragen gegeven. In het deelrapport van Fase 1 ligt de nadruk op de ontwikkeling van de methode en een vergelijking met de bestaande uitstroomthema's. Zoals hierboven beschreven zijn er in de ontwikkeling van de methoden bepaalde beslissingen genomen. Deze zijn op dit moment passend voor deze dataset, en zouden voor een andere dataset mogelijk anders genomen zijn. Een voorbeeld hiervan is het aantal topics waarvoor gekozen is. In dit hoofdstuk worden eerst de resultaten van de analyse besproken. Aan de opdrachtgever van dit onderzoek is daarnaast de interactieve LDA-visual opgeleverd waardoor arbeidsmarktonderzoekers zich kunnen verdiepen in de gebruikte methoden. In deze LDA-visual kunnen gebruikers van globaal tot in detail door de topics bladeren.

Op basis van de open antwoorden worden verschillende uitstroomtopics onderscheiden in de uitstroomredenen van medewerkers. Daarbij moet worden aangetekend dat veel medewerkers een combinatie van uitstroomredenen opgeven.

3.1 Woordfrequentieanalyse

In totaal waren de antwoorden van 1671 respondenten bruikbaar voor topic modellering van de uitstroomredenen en 1226 voor de preventiemogelijkheden. In alle antwoorden is gekeken naar welke woorden het vaakst voorkwamen. Wanneer wordt gekeken naar welke zelfstandig naamwoorden het vaakst voorkomen in de dataset, zijn veel woorden op zichzelf niet informatief. Als deze worden gecombineerd met de meest voorkomende woorden binnen een specifiek topic, komt meer informatie naar voren. Binnen de topics komen niet alleen de zelfstandig naamwoorden naar voren maar ook de bijvoeglijk naamwoorden en werkwoorden. Al deze woorden in relatie tot elkaar bieden context. In Tabel 1 staan de meest voorkomende zelfstandig naamwoorden. Dit geeft een beeld van welke onderwerpen worden genoemd door vertrekkende werknemers.

Tabel 2 Meest voorkomende zelfstandig naamwoorden voor de uitstroomredenen en preventiemogelijkheden

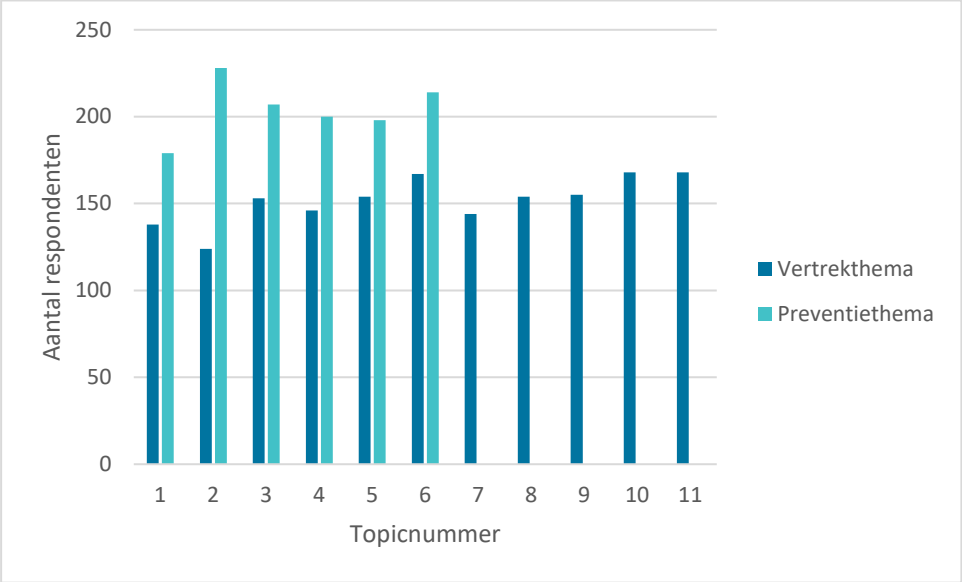
Uitstroomredenen (n=1671)		**Preventiemogelijkheden** (n=1226)	
Zelfstandig naamwoord	*Aantal keer in de dataset*	*Zelfstandig naamwoord*	*Aantal keer in de dataset*
organisatie	148	personeel	131
team	127	team	116
opleiding	120	collega's	93
werkdruk	107	zorg	87
tijd	105	tijd	86
huis	90	functie	84
uitdaging	90	waardering	81
personeel	87	organisatie	74
collega's	81	medewerkers	68
uren	75	salaris	64
afdeling	73	gesprek	53
reistijd	62	mogelijkheden	53

diensten	57	afdeling	52
contract	53	werkdruk	52
ziekenhuis	50	werkvloer	52
bewoners	48	communicatie	50
gevoel	45	baan	48
manager	45	opleiding	48
cliënten	44	aandacht	45
uur	43	mensen	45
medewerkers	42	cliënten	40
salaris	41	diensten	39
dagen	40	situatie	38
sfeer	40	uren	38
communicatie	39	uitdaging	33
reden	38	contract	31
hoge_werkdruk	35	rooster	30
management	35	bewoners	29
ontslag	35	gevoel	28
locatie	34	dingen	27
veranderingen	33	manager	27
verhuizing	33		
week	32		
werkvloer	32		
mogelijkheden	31		
afstand	30		
rooster	30		
doelgroep	29		
waardering	29		
beleid	28		
teamleider	28		
jaren	27		
mensen	27		
studie	27		
werkzaamheden	27		

3.2 Topic modellering

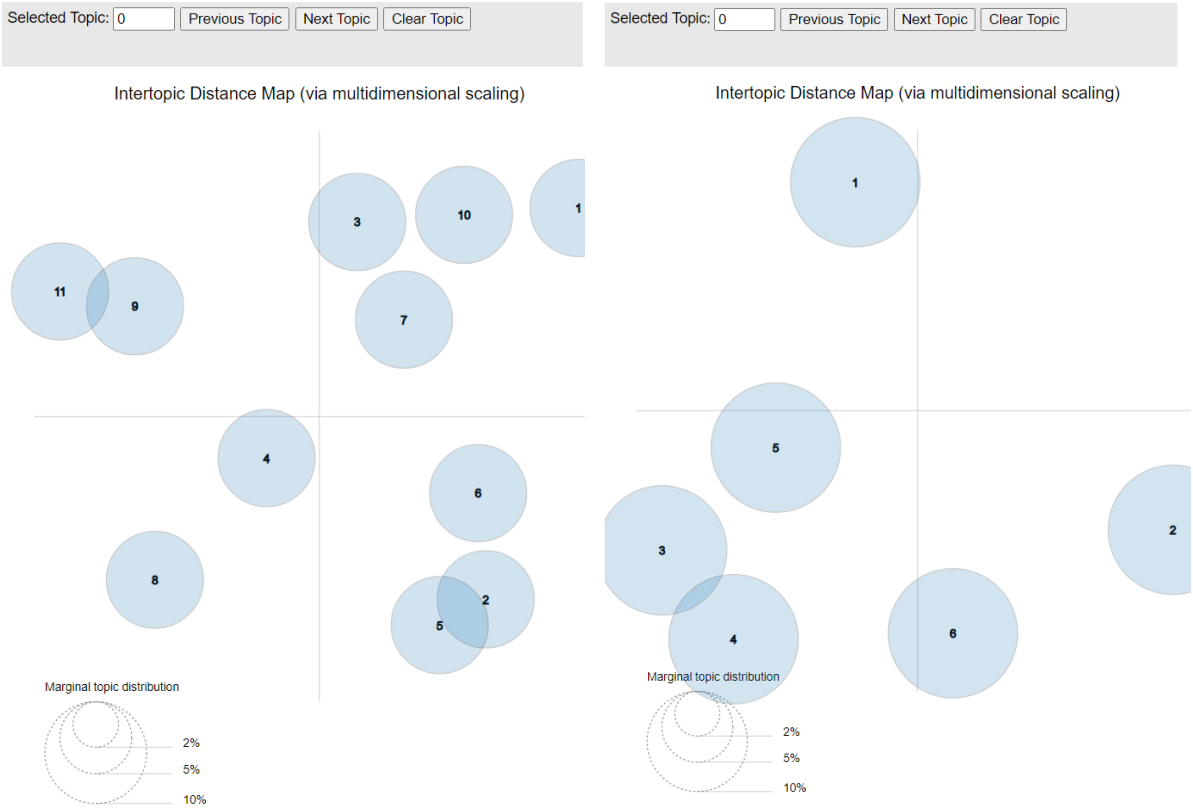
Op basis van de keuzes in de methoden zijn 11 uitstroomtopics en 6 preventietopics ontwikkeld. Wanneer gekeken wordt naar hoe vaak verschillende topics naar voren komen is te zien dat de respondenten relatief gelijk zijn verdeeld over de topics (Figuur 3). Door in de LDA-visual een topic te selecteren is het mogelijk om door de woorden heen te bladeren. Afhankelijk van de stand van de slider geeft dit de meest voorkomende woorden in het topic, de meest typerende woorden voor een topic, of een tussenvariant weer.

Figuur 3 Verdeling van de respondenten over de topicnummers voor uitstroomredenen (n=1671) en preventiemogelijkheden (n=1226).



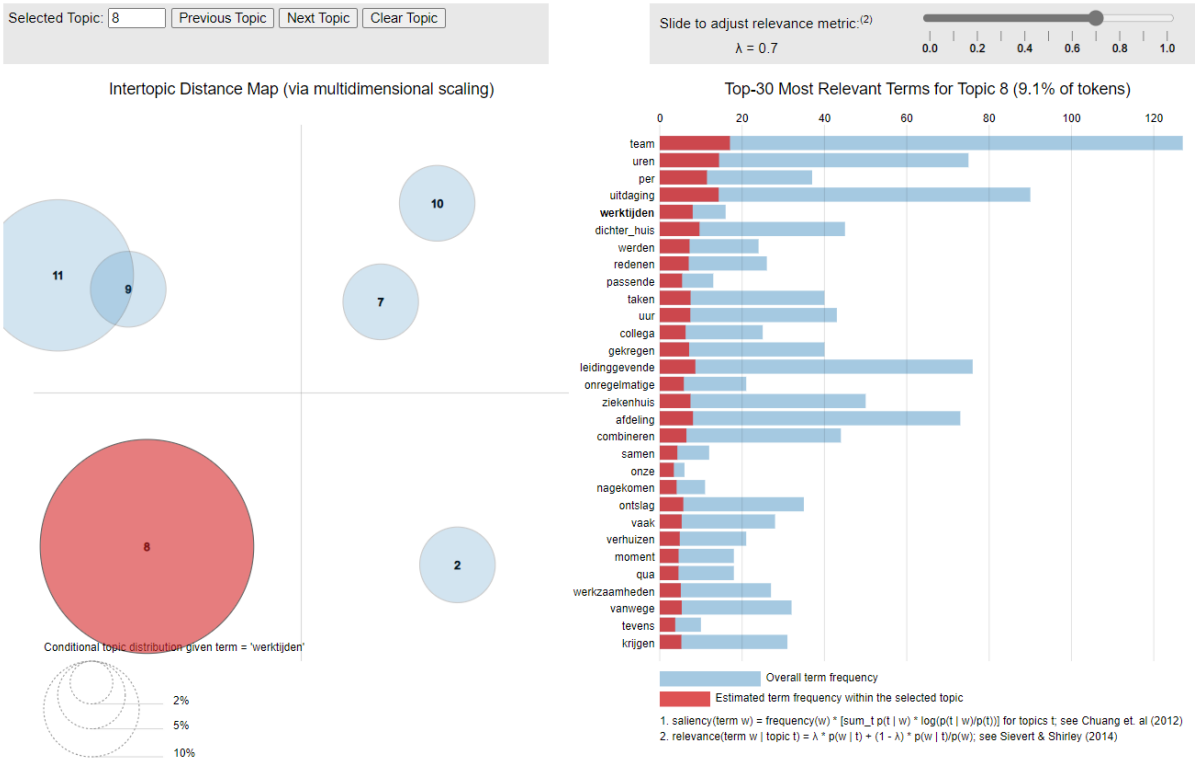
Aan de linkerkant van de LDA-visual zijn de 11 topics weergegeven als cirkels. Daarin geeft de relatieve afstand tussen de topics weer in welke mate de topics op elkaar lijken en dezelfde woorden bevatten. In figuur 4 is de linkerkant van de LDA-visual weergegeven voor zowel de uitstroomtopics als de preventietopics.

Figuur 4 Schermafbeelding van de linkerkant van de LDA-visual voor uitstroomtopics (links) en preventietopics (rechts) met daarin de relatieve afstand tussen de topics.



Door met de muis aan de rechterkant van de LDA-visual op een woord te staan is te zien in welke topics dit woord voorkomt, en in welke mate. Zo is te zien dat het woord ‘werktijden’ het vaakst voorkomt in topic 8.

Figuur 5 Schermafbeelding van de LDA-visual van uitstroomredenen waarbij het woord ‘werktijden’ is geselecteerd bij λ=0.7



Op basis van de woorden binnen een topic krijgen de topics een overkoepelende naam die overeenkomt met de inhoud van het topic. (Tabel 3) Omdat respondenten meerdere uitstroomredenen opgeven, ontstaat er overlap tussen de topics. Deze benamingen zijn bepaald door te kijken naar een combinatie van de meest voorkomende woorden binnen een topic, en de woorden die het meest typerend zijn voor een topic. Daarbij is λ=0.7 aangehouden als middenweg om de inhoud van zowel de uitstroontopic als de preventietopics te bepalen. Hierdoor komen de genoemde tokens nog in voldoende mate voor om iets te zeggen over de gehele groep. Door de verkregen informatie en benaming kan de eerste onderzoeksvraag beantwoord worden:

3) *Welke topics zijn er te onderscheiden in de antwoorden op de twee open vragen;*
   - *Wat is de reden dat je weggaat bij [organisatie]?*
   - *Wat had [organisatie] anders kunnen doen om jouw vertrek te voorkomen?*

Tabel 3 Topicnamen voor de 11 uitstroomtopics

Topicnummer	Topicnaam
1	Werkdruk, werksfeer, teamgevoel en collega's
2	Ontwikkeling, werkuren, contractinhoud en diensten, een opleiding gaan volgen
3	Reorganisatie, veranderingen, leidinggevende, fysieke beperkingen
4	Niet gehoord voelen, organisatie binnen de afdeling, gebrek aan begeleiding
5	Werkinhoud, fysieke klachten, elders betere baan/doorgroeimogelijkheden
6	Begeleiding/communicatie vanuit het management/leiding, verhuizing, van 2 naar 1 werkgever, te veel onregelmatigheid
7	Reistijd en baan dichter bij huis genomen, combinatie met privéleven
8	Onregelmatige werktijden, baan dichter bij huis, takenpakket
9	Leidinggevende, begeleiding en afspraken nakomen
10	Doorontwikkelen, doorgroeien, diensten, werk-privé-balans
11	Niet serieus genomen worden, geen waardering, nakomen afspraken, reorganisatie

De woorden binnen topics in relatie tot elkaar geven meer informatie. Per topic zijn de 5 meest voorkomende woorden bepaald. Deze woorden komen uit de LDA- visual naar voren bij een lambda van λ=1.0. Omdat er veel overlap is in de gegeven antwoorden met daarin meerdere thema's, komen woorden 'collega's', 'organisatie', 'werkdruk' en anderen voor in meerdere topics.

Tabel 4 Vijf meest voorkomende woorden per uitstroomtopic

Topicnummer	5 meest voorkomende woorden
1	Werkdruk, collega's, afdeling, personeel, organisatie
2	Opleiding, collega's, uur, personeel, week
3	Organisatie, leidinggevende, uren, gewerkt, taken
4	Organisatie, team, afdeling, werkdruk, tijd
5	Tijd, opleiding, team, ziekenhuis, gewerkt
6	Werkdruk, personeel, ziekenhuis, team, week
7	Huis, management, uitdaging, dichter_huis, opleiding
8	Team, uren, uitdaging, per, dichter_huis
9	Organisatie, huis, team, leidinggevende, tijd
10	Opleiding, organisatie, collega's, tijd, personeel
11	Organisatie, team, reistijd, dichter, genomen

De preventiemogelijkheden zijn onderverdeeld in 6 topics. In Tabel 5 zijn de topicnamen voor de preventietopics weergegeven met daaronder in Tabel 6 de 5 meest voorkomende woorden per preventietopic bij een lambda van λ=1.0.

Tabel 5 Topicnamen voor de 6 preventietopics

Topicnummer	Topicnaam
1	Uren/diensten aanpassen, werkdruk verlagen, afspraken nakomen
2	In gesprek gaan met medewerkers, klachten serieus nemen, samenwerking in het team verbeteren
3	Begrip voor medewerkers, luisteren naar medewerkers
4	Waardering geven aan medewerkers, doorgroeimogelijkheden bieden, beter salaris bieden
5	Inspraak in rooster, diensten, zorgkwaliteit behouden
6	Serieus nemen van medewerkers, meedenken met en van medewerkers

Tabel 6 Vijf meest voorkomende woorden per preventietopic

Topicnummer	5 meest voorkomende woorden
1	Team, collega's, functie, per, gesprek
2	Mee, gesprek, collega's, leidinggevende, baan
3	Personeel ,luisteren, team, collega's, laten
4	Team, waardering, leidinggevende, salaris, personeel
5	Zorg, luisteren, werkvloer, mogelijkheden, organisatie
6	Echt, leidinggevende, serieus, personeel, zorg

3.3 Congruentieanalyse

Op basis van ID-nummers is het topicnummer dat door LDA aan een respondent wordt toegekend, gekoppeld aan de oorspronkelijke dataset met daarin de door de respondent aangevinkte thema's. Dit biedt de mogelijkheid om de tweede onderzoeksvraag te beantwoorden:

4) *In welke mate komen de topics die uit de analyse voortkomen overeen met de huidige thema's uit de multiple-choice-vragen?*

Hiervoor is een vergelijking gemaakt tussen de door LDA geïdentificeerde topics en de uitstroomthema's die de respondent geselecteerd heeft in de vragenlijst. Om te bepalen in welke mate de topics ontwikkeld door de LDA-methode overeenkomen met thema's uit de vragenlijst is per topic dat door de LDA aan een respondent gekoppeld is, bepaald hoe vaak een thema uit de vragenlijst geselecteerd is door de respondent. De thema's in de multiple choice vragenlijst komen tot op zekere hoogte overeen met de topics die naar voren komen uit de LDA analyse. In Tabel 7 staat weergegeven hoe vaak een respondent met een bepaald LDA-topic een thema uit de vragenlijst heeft aangevinkt. Zo is bijvoorbeeld te zien dat van alle respondenten die volgens de LDA-analyse binnen topic 1 vallen, 31% het thema werkdruk heeft geselecteerd op de vragenlijst. Dit komt overeen met de inhoud van topic 1: Werkdruk, werksfeer, teamgevoel en collega's.
Toch zijn er zeker ook verschillen te ontdekken. Respondenten geven vaak meerdere redenen voor hun vertrek aan in hun open antwoorden. Dit zorgt voor relatief veel overlap tussen de topics. Uit de LDA analyse komen bepaalde nuances naar voren, die niet uit de multiple choice antwoorden te halen zijn. Een voorbeeld hiervan is het verlaten van een werkgever omdat de behoefte bestaat niet meer voor 2, maar voor 1 werkgever te werken. Daarnaast is het onderwerp reistijd, of dichter bij huis willen werken een belangrijk onderwerp, en mogelijk nieuwe hoofdreden voor de vragenlijst, waar dit nu een subthema is onder privésituatie.

Tabel 7 Overeenkomst tussen de uitstroomtopics en de geselecteerde thema's in de vragenlijst

	Leidinggevende(n)	Cliënten	Planning & tijd	Samenwerking collega's	Ontwikkelmogelijkheden	Inspraak & invloed	Werkdruk	Arbeidsvoorwaarden	Privésituatie	Ongewenst of vervelend gedrag	Werkinhoud	Anders, namelijk
1	23%	5%	12%	11%	19%	8%	31%	6%	16%	8%	18%	20%
2	18%	4%	18%	5%	33%	9%	8%	20%	15%	6%	18%	12%
3	17%	6%	13%	11%	12%	14%	20%	6%	16%	6%	26%	19%
4	10%	2%	19%	16%	15%	8%	19%	15%	19%	5%	16%	23%
5	12%	5%	28%	10%	15%	5%	20%	4%	20%	5%	25%	20%
6	12%	2%	14%	5%	27%	9%	12%	4%	31%	4%	18%	22%
7	12%	2%	16%	10%	15%	10%	13%	10%	29%	5%	21%	25%
8	10%	2%	19%	15%	22%	8%	15%	7%	24%	5%	14%	22%
9	26%	4%	15%	17%	14%	7%	15%	15%	14%	8%	13%	19%
10	7%	3%	20%	10%	22%	14%	23%	11%	24%	7%	20%	12%
11	18%	5%	14%	13%	27%	13%	17%	8%	21%	7%	19%	18%

3.4 Inhoudsanalyse

Door middel van de interactieve slider is het mogelijk om tot in detail te kijken welke woorden vaak, en minder vaak binnen een topic worden genoemd. Hierdoor kan de nuance die door de respondent wordt gegeven in het open antwoord naar voren komen. Bepaalde woorden komen dan naar voren die minder vaak worden ingevuld door de respondenten.
Om te bepalen welke redenen binnen een uitstroomtopic naar voren komen, is het belangrijk om ook te kijken naar de minder vaak voorkomende combinaties. Dit geeft de mogelijkheid tot het beantwoorden van de derde onderzoeksvraag:

5) *In welke mate komen de redenen binnen de uitstroomtopics die uit de analyse voortkomen overeen met de huidige redenen binnen de uitstroomthema's uit de multiple-choice-vragen?*

Omdat bepaalde uitstroomredenen vaker voorkomen in de dataset, komen met name de hoofdonderwerpen duidelijk naar voren uit de LDA-analyse. Om te bepalen in welke mate de subredenen binnen de uitstroomtopics overeen komen is het ook van belang te kijken naar uitstroomredenen die minder vaak voorkomen. Hier kan naar gekeken worden door de lambda slider (λ) te schuiven naar een kleiner getal. Wanneer een uitstroomreden (en dus ook een subthema) erg weinig voorkomt zal deze minder snel opgemerkt worden omdat de laagste 1% van alle tokens uit de LDA analyse gefilterd wordt. Hierdoor is minder zich op de redenen die niet vaak voorkomen. De methode van LDA-analyse is met name geschikt om in een grote hoeveelheid tekst verborgen verbanden te ontdekken.

Uit de LDA analyse blijkt dat nuance nog steeds belangrijk is. Dit is een groot voordeel van het analyseren van de open antwoorden. Zo komt het voor dat afspraken over doorgroeien en ontwikkelen niet worden nagekomen. Dit wordt in de huidige vragenlijst geschaard onder

ontwikkelmogelijkheden, terwijl er meer aan ten grondslag ligt. Het gegeven dat afspraken over dit onderwerp niet worden nagekomen is niet terug te vinden als subreden.
Wat opvalt is dat sommige redenen vaker worden genoemd, terwijl dit in het uitstroomonderzoek een subreden is binnen een hoofdthema. Daarnaast is een belangrijke bevinding dat er bepaalde woorden voorkomen in de open antwoorden, die niet terugkomen als uitstroomreden binnen een thema in de vragenlijst. In de LDA-analyse zijn verschillende van die woorden te ontdekken. Dat kan er op duiden dat er mogelijk subthema's missen in de vragenlijst van het uitstroomonderzoek. Een voorbeeld hiervan is dat het niet nakomen van afspraken (over doorgroeien) niet terugkomt als antwoordmogelijkheid in de vragenlijst of niet makkelijk te vinden zijn door de respondent. Andere voorbeelden betreffen de sfeer binnen het team, het twijfelen over de kwaliteit en kundigheid van de leidinggevende, of zorgen over de prioriteiten van een organisatie (kwaliteit van zorg of economisch belang). Dit zijn op dit moment geen directe opties in de vragenlijst.
Een voorbeeld van een thema dat tot nu toe onderbelicht bleef is het krijgen van onvoldoende waardering, of het idee hebben dat er niet geluisterd wordt wanneer er problemen aangekaart worden door de medewerkers. Het lijkt er op dat medewerkers vaak het probleem waarvoor ze uitstromen eerder aankaartten bij de leidinggevende. De uitstroomreden ligt vervolgens in het feit dat de medewerker het idee krijgt dat er niets mee wordt gedaan, en dat er niet geluisterd wordt. Dit geeft de medewerker het gevoel dat hij of zij niet belangrijk gevonden wordt. Dit geeft aan dat in veel gevallen er minimaal twee redenen zijn: het oorspronkelijke probleem, en het gevoel dat er niet geluisterd wordt of weinig waardering is.

3.5 Relatie tussen uitstroomtopics en preventietopics

Wanneer de uitstroomtopics en de preventietopics aan elkaar gekoppeld worden geeft dit nog weinig concrete informatie (tabel 8). De verschillende preventietopics komen nu verspreid voor onder de uitstroomtopics. Dit is een direct gevolg van de aard van de data waarin respondenten in hun open antwoord meerdere uitstroomredenen en preventiemogelijkheden aangeven. Het duidelijkst komt de relatie tussen preventietopic 5 (Inspraak in rooster, diensten, zorgkwaliteit behouden) en uitstroomtopic 5 (Werkinhoud, fysieke klachten, elders betere baan/doorgroeimogelijkheden) naar voren. Van alle respondenten met uitstroomtopic 5, wordt 27% gekoppeld aan preventietopic 5. Dat wil dus zeggen dat van de personen die vertrekken omdat de werkinhoud niet passend was, er fysieke klachten in het spel waren, of het gebrek aan doorgroeimogelijkheden of dat elders een betere baan gevonden was, in 27% van de gevallen mogelijk voorkomen had kunnen worden door een beter rooster, diensten aan te passen, of door zorg voor bewoners op niveau te houden. Het serieus nemen van medewerkers, en meedenken met en van medewerkers als preventietopic 6, wordt juist minder vaak gekoppeld aan uitstroomtopic 5 (7%) maar wel vaker aan topic 9 en 7.

Tabel 8 Mate waarin het preventietopic bij respondenten voorkomt per uitstroomtopic

Uitstroomtopic	Preventietopic 1	2	3	4	5	6
1	15%	16%	23%	13%	14%	19%
2	17%	18%	18%	19%	13%	14%
3	11%	26%	18%	17%	12%	16%
4	15%	20%	19%	16%	16%	15%
5	21%	13%	15%	17%	27%	7%
6	13%	20%	22%	18%	12%	15%
7	14%	20%	17%	15%	10%	24%
8	14%	17%	12%	17%	19%	21%
9	17%	16%	12%	11%	18%	25%
10	14%	20%	12%	21%	15%	19%
11	10%	19%	18%	16%	18%	18%

Een verbetering van de methode waardoor er minder overlap is tussen topics, en minder onderwerpen in een topic staan kan bijdragen aan een nauwkeurigere bepaling van welke preventiemogelijkheden een vertrek om een bepaalde reden kan voorkomen. De relatie tussen de uitstroomredenen en preventiemogelijkheden wordt in fase 2 dieper bestudeerd.

4 Evaluatie en aanbevelingen

In dit hoofdstuk wordt de eerste fase geëvalueerd en worden aanbevelingen gedaan over hoe de analyse in fase 2 effectiever ingezet kan worden. Daarnaast worden suggesties gedaan hoe de vragenlijst van het landelijk uitstroomonderzoek van RegioPlus en Presearch mogelijk verbeterd kan worden.

4.1 Toegevoegde waarde van het gebruik van de LDA methode ten opzichte van de multiple-choice-vraag

De huidige analyse van de uitstroomgegevens is voornamelijk gebaseerd op het rapporteren van percentages aangevinkte hoofdredenen uit een multiple choice vraag. De LDA methode is gebaseerd op de open antwoorden van respondenten. Open antwoorden bieden de respondenten de mogelijkheid om naar eigen keuze vrije tekst in te voeren wat een groot voordeel heeft ten opzichte van voorgeformuleerde thema's. Respondenten kunnen hun uitstroomreden exact weergeven, zonder dit te moeten categoriseren onder voorgeformuleerde thema's. De informatie die uit de open antwoorden komt is dus direct afkomstig van de respondent. In de voorgeformuleerde thema's is het mogelijk dat de respondent de uitstroomreden die het beste bij hem of haar past niet terug kan vinden in de themalijst, of de reden niet direct kwijt kan onder de noemer van het thema.
Het nadeel van het gebruiken van open antwoorden is dat het de analyse en het onderzoek een stuk ingewikkelder, tijdrovender en arbeidsintensiever maakt. De LDA methode maakt het mogelijk om in grote hoeveelheden tekst verborgen structuren te ontdekken, met een minder arbeidsintensief proces. De LDA methode draagt bij aan het nauwkeuriger in kaart brengen, en later beter analyseren van de uitstroomredenen.

De huidige analyse in fase 1 geeft al een beter beeld van welke redenen worden opgegeven als uitstroomreden, en in welke mate dit voorkomt dan de bestaande methode waarbij slechts wordt ingegaan op hoe vaak een vooraf vastgesteld thema voorkomt. Daarnaast is het op dit moment niet mogelijk om de preventiemogelijkheden te analyseren omdat een multiple-choice vraag hierover ontbreekt in de vragenlijst. Hier zijn alleen gegevens over bekend uit een open antwoordveld. Analyse door middel van LDA brengt verschillende inzichten met zich mee die gebruikt worden in de analyse, en mogelijk later kan bijdragen aan de preventie van uitstroom. De LDA methode maakt de antwoorden van respondenten inzichtelijk zodat er een beeld ontstaat van wat er daadwerkelijk nodig is om een persoon te behoeden voor het vertrek.

4.2 Toegevoegde waarde van het gebruik van de landelijke data voor de analyse

In de eerste fase is gekeken naar data van respondenten uit de RegioPlus-regio Utrecht. Omdat in deze huidige fase zowel gekeken wordt naar de uitstroomtopics als de preventietopics, kan een inschatting worden gemaakt van de bruikbaarheid van de methode voor de volledige (landelijke) dataset. Met deze informatie kan de vierde onderzoeksvraag worden beantwoord:

6) *Wat is op basis van de antwoorden op bovenstaande vragen de toegevoegde waarde om deze methode toe te passen om landelijke data te analyseren?*

Er zijn verschillende redenen waarom het nuttig is om de analyse op de landelijke dataset uit te voeren:

1. Een analyse op een veelvoud van input geeft robuustere resultaten: Analyse op de volledige dataset zorgt voor robuustere resultaten door een verhoogde statistische kracht, waardoor kleinere effecten detecteerbaar worden en conclusies met grotere zekerheid getrokken kunnen worden.

2. Diepgaandere inzichten: Met een grotere dataset kunnen meer gedetailleerde en specifieke patronen en trends worden geïdentificeerd, wat leidt tot diepgaandere inzichten in de beweegredenen achter uitstroom en preventie van uitstroom.

3. Verhoogde generaliseerbaarheid: Resultaten gebaseerd op de volledige dataset zijn representatiever voor de gehele populatie, wat de generaliseerbaarheid van de bevindingen vergroot.

4. Nuance en diversiteit in redenen: Een grotere dataset onthult een breder spectrum aan redenen achter uitstroom, inclusief minder voorkomende maar even belangrijke factoren.

5. Ontdekking van subgroepen: Een uitgebreide dataset biedt de mogelijkheid om analyses uit te voeren op subgroepen, waardoor het mogelijk wordt om te identificeren hoe verschillende groepen werknemers uniek worden beïnvloed. Er kunnen meer topics worden gecreëerd waardoor subgroepen zichtbaar worden. Zo kan bijvoorbeeld beoordeeld worden welke topics vaak bij elkaar worden genoemd omdat er voldoende input is om extra topics te ontwikkelen.

6. In fase 2 zal gekeken worden naar een combinatie van uitstroomredenen en preventiemogelijkheden. Dit geeft informatie met betrekking tot hoe een vertrek voorkomen had kan worden, en geeft concrete handvatten waar werkgevers mee aan de slag kunnen.

7. In fase 2 wordt daarnaast gekeken naar of bepaalde profielen te ontdekken zijn in de uitstroom van medewerkers. Hierdoor kan niet alleen beoordeeld worden welke preventiemogelijkheid bij welke uitstroomreden past, maar daarnaast ook of er verschillen te ontdekken zijn in de uitstroomredenen tussen personen met bepaalde persoons- of baankenmerken. (Hier zijn al aanwijzingen voor, zie: Lemmelijn (2023).)

Met enkele aanpassingen aan het script en de methode, is het uitvoeren van deze analyse op de landelijke data van toegevoegde waarde.

4.3 In welke mate is de ontwikkelde methode bruikbaar voor fase 2?

De LDA-methode is uitermate geschikt om verborgen verbanden en thema’s te ontdekken in grote hoeveelheden tekst. In fase 2 gaat het om nog veel grotere hoeveelheden tekst dan in de eerste fase. De methode die in fase 1 is ontwikkeld, heeft enkele nieuwe inzichten opgeleverd. Het gebruik ervan heeft geleid tot aanbevelingen voor aanpassingen in fase 2. Een van de belangrijkste wensen is om respondenten meerdere thema's toe te kennen op basis van hun open antwoorden. Verschillende opties hiervoor worden overwogen.

De LDA analyse kan bruikbaar zijn waarbij op basis van meer data, meer topics kunnen worden gevormd. Zo komen thema's die vaak samen worden genoemd in een onderscheidend topic terecht. De toepassing van Latent Dirichlet Allocation (LDA) analyse op een de grotere dataset van fase 2 biedt het voordeel van het mogelijk maken van meer clusters, wat resulteert in een gedetailleerder inzicht in de data. Dit komt door het feit dat grotere datasets een breder scala aan informatie bevatten, wat de identificatie van een uitgebreider aantal thema's of onderwerpen mogelijk maakt. Echter, uit fase 1 blijkt dat individuele antwoorden vaak meerdere thema's omvatten. Dit leidt tot clusters die naar verwachting diverse onderwerpen bevatten. Deze karakteristiek van de data is onveranderlijk en is typisch voor de data.

Een onderdeel van de huidige methode is de frequentie-analyse. Het uitvoeren van een analyse op de meest frequente woorden binnen de dataset is uitermate waardevol, aangezien het een algemeen overzicht biedt van de thema's of onderwerpen die prevalent zijn binnen de verzamelde gegevens. Een praktische benadering kan zijn om handmatig een clustering van de top x meest voorkomende woorden uit te voeren om zo patronen of thematische groeperingen te identificeren. Een alternatieve strategie zou zijn om de document-term matrix (DTM) een prominentere plek te geven in deze analyse. Door dit te doen, kan men specifiek zoeken naar antwoorden die bepaalde woorden bevatten binnen hulpprogramma's zoals Excel of Stata. Dit biedt een methodische benadering om diepgaand inzicht te krijgen in hoe specifieke termen zich verhouden tot de bredere thema's binnen de dataset, en kan helpen bij het verder verfijnen van de analyse en interpretatie van de gegevens. Op deze wijze kunnen ook bepaalde woorden die synoniemen van elkaar zijn samengevoegd worden zoals fysiek en lichamelijk, en loon en salaris.

Het is belangrijk om de typen respondenten in gedachten te houden. In deze analyse is ervoor gekozen om personen die vrijwillig vertrekken uit cliëntgebonden functies te analyseren. Daarom zijn in de huidige analyse respondenten met tijdelijke stage- of vakantiecontracten buiten de analyse gelaten. Het is mogelijk wenselijk om in fase 2 personen uit de analyse te laten die het woord bijbaan gebruiken om zo de uitstroomredenen van werknemers die dit slechts als bijbaan deden (naast hun studie) eruit te filteren. Daarnaast kan door middel van een vooraf gedefinieerde lijst in een eerder stadium de namen van werkgevers verwijderd worden uit de antwoorden.

4.4 Suggesties vragenlijst uitstroomonderzoek RegioPlus en Presearch

De analyse en methode hebben waardevolle inzichten opgeleverd met betrekking tot de data. Dit draagt bij aan de doorontwikkeling van de bestaande vragenlijst. Op basis van de verkregen informatie worden enkele suggesties gedaan voor de vragenlijst die wordt gebruikt in het uitstroomonderzoek van RegioPlus en Presearch.

- Omdat veel medewerkers een combinatie van uitstroomredenen opgeven is het belangrijk in de vragenlijst het mogelijk te maken om meerdere uitstroomtopics te selecteren. Op dit moment is dat beperkt tot 2 thema's.
- Waardering of niet gehoord voelen valt onder 3 verschillende thema's door de nuance die eraan gegeven is. Niet gehoord worden door: de organisatie (Inspraak en invloed), de leidinggevende (Leidinggevende) of collega's (Samenwerking collega's), en waardering door: de leidinggevende (Leidinggevende), cliënten (Cliënten) of collega's (Samenwerking collega's). Door 'niet gehoord worden' of 'gebrek aan waardering' op te nemen als hoofdreden, kan mogelijk daarna de nuance gemaakt worden naar voor wie dit geldt.

- In het thema Leidinggevende staan veel verschillende zaken genoemd. Een onderwerp dat voortkwam uit de LDA analyse was dat de leidinggevende niet geschikt was, of geen gevoel had voor de verhoudingen in het team. Een ander onderwerp betrof het niet nakomen van gemaakte afspraken door de leidinggevende of het management.
- Daarnaast wordt regelmatig aangegeven dat medewerkers de organisatie te groot of onpersoonlijk vinden. De grootte van de organisatie wordt nu niet gevraagd in de vragenlijst. Tevens kan dit niet als reden worden aangevinkt voor vertrek.
- Sommige uitstroom kan worden verklaard omdat personen stoppen met hun bijbaan. Dit kan nu niet direct bepaald worden uit de vragenlijst.
- Enkele respondenten verlieten hun werkgever omdat ze de wens hadden niet meer voor meerdere werkgevers te werken. Dit kan komen vanwege privéredenen, maar ook door inflexibiliteit van de werkgever, of het wijzigen van bestaande afspraken (bijvoorbeeld rondom werktijden/dagen). Nu is onbekend of respondenten meerdere werkgevers hebben.
- Er zijn enkele thema's die nu als uitstroomreden pas kunnen worden geselecteerd na het selecteren van een hoofdreden, maar niet altijd worden gevonden door de respondenten. Om deze uitstroomreden te zien te krijgen in de vragenlijst moet eerst het juiste hoofdthema geselecteerd worden. Zo valt bijvoorbeeld de reistijd of een verhuizing onder het thema 'Privésituatie'. Voor veel voorkomende thema's is het verstandig om deze op te nemen als hoofdthema.
- Een aanvullende analyse van de open antwoorden van respondenten die in de multiple-choice-vraag 'Anders, namelijk' hebben aangevinkt heeft aangetoond dat de antwoorden vaak in te delen zijn onder (bestaande) thema's.
- Het is belangrijk dat redenen die vaak worden genoemd, niet verstopt zitten onder andere hoofdthema's. Momenteel wordt reistijd vaak ingevuld bij 'Anders namelijk'. Dit geeft aan dat deze uitstroomreden niet direct in te schatten is onder de huidige thema's in de vragenlijst. Een andere optie is meer toelichting over de inhoud te geven per hoofdthema om zo de respondent te helpen het juiste thema te selecteren.
- Mogelijk kan er ook een extra multiple-choice vraag over preventiethema's worden opgenomen in de vragenlijst. Hier kunnen dan thema's ontwikkeld worden op basis van de meest voorkomende woorden. Dit maakt in de toekomst analyse eenvoudiger omdat zo relaties beter kunnen worden blootgelegd.
- Enkele mogelijke thema's voor een multiple choice vraag over preventiemogelijkheden zijn:
    - In gesprek gaan met de medewerkers
    - Luisteren naar problemen als deze worden aangekaart
    - Waardering tonen
    - Betere communicatie
    - Leidinggevende die beter gevoel heeft met het team
    - Werksfeer verbeteren
    - Werkdruk verlagen
    - Rooster aanpassen
        1. Flexibeler omgaan met tijden
        2. Meer vaste werktijden bieden
    - In gesprek gaan over doorgroeimogelijkheden
    - Beter salaris bieden
    - Meer aandacht voor de kwaliteit van zorg
    - De werkgever had niets kunnen doen
        1. Ik was heel tevreden
        2. Mijn vertrekreden lag binnen de privésfeer
- Enkele uitstroomredenen die vaak genoemd worden, en waarvan overwogen kan worden dit als hoofdthema op te nemen in de vragenlijst:

    - Reistijd
    - Werksfeer
    - Werktijden
    - Waardering
- Een mogelijk aanvullend subthema is: Er is niet geluisterd naar mijn problemen

Er worden door respondenten in de open antwoorden bepaalde redenen genoemd die niet voorkomen in de vragenlijst. Enkele zijn hierboven genoemd. Het is een afweging hoe belangrijk het is om compleet te zijn in de vragenlijst, of dat deze informatie uit de open antwoorden kan worden gehaald. Het is bekend dat het bieden van een mogelijkheid om zelf tekst in te voeren zorgt voor een betere respons. Daarnaast biedt dit de mogelijkheid tot nuance, zoals nu naar voren komt uit de LDA-analyse.

5 Conclusie en discussie

In dit hoofdstuk wordt kort benoemd wat de geleerde lessen zijn uit Fase 1 van dit onderzoek. Hierin staan de toegevoegde waarde van het gebruik van de open antwoorden, en de methode centraal.

5.1 Conclusie

De ontworpen LDA-methode is bruikbaar, en geeft relevante informatie die niet direct uit andere variabelen te extraheren is. Desondanks dient voor andere data beoordeeld te worden hoe de geschreven codes aangepast dienen te worden zodat dit aansluit bij de dataset. Ondanks dat fase 1 vooral gericht was op het ontwikkelen en testen van de methode biedt de minimale analyse van de resultaten een beter inzicht in de uitstroomredenen van respondenten dan dat de multiple-choice antwoorden bieden. De analyse in huidige vorm geeft als gevolg van de aard van de data veel overlap tussen de ontwikkelde topics waardoor het moeilijk is om deze te vergelijken met de bestaande hoofdthema's

Doordat er van globaal tot in detail door de topics kan worden gebladerd biedt dit andere inzichten dan classificatie op hoofdthema's. Dit brengt extra informatie met zich mee waardoor de vragenlijst van het uitstroomonderzoek van RegioPlus en Presearch mogelijk verbeterd kan worden. Deze aanpassingen sluiten beter aan bij de werkelijkheid en de verdeling van de uitstroomredenen door respondenten. Dat neemt niet weg dat er nog een verdiepingsslag te maken is met betrekking tot de gebruikte LDA methode. De huidige resultaten bieden onvoldoende aanknopingspunten omdat respondenten niet in onderscheidende categorieën ondergebracht worden. De reden hiervoor ligt in het feit dat er regelmatig meerdere verschillende redenen worden aangedragen voor het vertrek van een medewerker. Om de methode te verbeteren worden in voorgaand hoofdstuk enkele aanbevelingen gedaan. Een aanpassing van de methode is gewenst om zo recht te doen aan het antwoord van de respondent. Dit zal een verbetering opleveren van inzicht in verbanden die gelegd kunnen worden tussen de uitstroomtopics en de preventietopics.

Desondanks heeft analyse met de huidige methode al een toegevoegde waarde voor de inzichten met betrekking tot het vertrek, en de preventie van het vertrek van werknemers in de sector zorg en welzijn. De huidige analyse biedt een verdiepend inzicht in de nuance van de uitstroomredenen. Wanneer personen zelf de mogelijkheid hebben om exact hun situatie en reden in te vullen geeft dit een veel nauwkeuriger beeld van wat er speelt binnen de sector zorg en welzijn. Enkele thema's die tot nog toe onderbelicht bleven, zijn komen bovendrijven als vaak voorkomende reden voor vertrek.

5.2 Discussie

Ondanks dat de huidige methode voor deze dataset nog niet perfect is, heeft deze veel extra informatie opgeleverd. Het indelen van respondenten naar 1 topic lijkt niet passend te zijn voor de werkelijkheid. De reden die hieronder ligt, is dat de ingevoerde data niet slechts één reden van uitstroom bevat, maar dat het vertrek vaak een combinatie is van meerdere oorzaken. Het koppelen van meerdere thema's aan een respondent past beter bij de uitstroomreden.

Wanneer we de gehanteerde methode vergelijken met alternatieve methoden valt op dat de meeste alternatieve methoden een handmatige analysemethode hanteren. Een veelgebruikte methode om teksten te analyseren is MAXQDA. Op basis van handmatige toekenning van thema's aan stukken tekst worden later de thema's geanalyseerd. Deze methode wordt vooral gebruikt voor interviews in kleine aantallen. Voor een dataset met de huidige omvang is deze methode niet haalbaar. Een andere methode is het analyseren van sentimenten. Deze methode is voornamelijk gebaseerd op het analyseren van bijvoeglijke naamwoorden. Verschillende bijvoeglijk naamwoorden krijgen dan een bepaalde lading mee, positief of negatief. Het gaat hierbij vaak om de houding van de respondent ten aanzien van een product of dienst. Omdat het hier gaat om werknemers en de redenen achter hun uitstroom is deze methode niet geschikt. Er is namelijk geen schaal tussen 2 punten zoals bij positief en negatief. Het weergeven van een woordwolk kan een methode zijn om zo via de veelvoorkomende woorden een overzicht te creëren van vaak genoemde vertrekredenen. Deze methode geeft dan alleen de meest genoemde woorden, en niet de context waarin deze worden genoemd. In de huidige analyse wordt al een woordfrequentieanalyse getoond waardoor het weergeven van een woordwolk slechts een visueel andere interpretatie geeft. In de toekomst is het wellicht mogelijk om de analyse uit te laten voeren door niet alleen te kijken naar een toepassing die de tekst verwerkt en kijkt naar welke woorden vaak samen worden genoemd, maar ook naar de betekenis van die woorden in de context waarin ze genoemd worden.

Desondanks bieden de ontworpen LDA-methode en resultatenanalyse diepgaander inzicht in uitstroomredenen dan multiple-choice antwoorden, waardoor de uitstroomvragenlijst van RegioPlus en Presearch mogelijk verbeterd kan worden. Enkele suggesties bieden de werkgroep die over de vragenlijst beslist mogelijkheden om de vragenlijst aan te passen. Hoewel er nog mogelijkheden zijn voor methodologische verbeteringen, levert de huidige analyse reeds waardevolle inzichten op met betrekking tot het vertrek van werknemers in de zorg- en welzijnssector en de preventie daarvan. Daarnaast zijn enkele onontdekte thema's die tot nu toe onderbelicht waren aan het licht gekomen.

De eerste fase van dit onderzoek betrof de verdieping en ontwikkeling van de LDA-methode. De tweede fase van dit onderzoek richt zich op het extraheren van de gewenste data en inzichten die werkgevers direct kunnen gebruiken bij de preventie van de uitstroom van medewerkers in zijn of haar organisatie. Door in de tweede fase de uitstroomtopics te koppelen aan persoonskenmerken en de preventietopics kan beoordeeld worden op welke wijze de werkgever een bepaald soort vertrek mogelijk kan voorkomen.

Om daadwerkelijk de gewenste informatie met betrekking tot welke preventieve activiteiten kunnen worden ondernomen bij welke uitstroomreden en type medewerker gewenst is te extraheren, is het van belang de analyse aan te passen, en de tweede fase van het onderzoek te starten. Met die resultaten uit fase 2 worden waardevolle inzichten verkregen waar de werkgever mee aan de slag kan gaan. In de tweede fase worden de volgende onderzoeksvragen beantwoord:

7) *Hoe moeten de scripts en methoden uit fase 1 worden aangepast om deze te kunnen gebruiken voor de complete dataset van heel Nederland?*
8) *Welke profielen zijn er te ontdekken in de uitstroomthema's? (Nadat een uitstroomtopic uit de analyse is toegekend aan een respondent) Deze profielen worden ontwikkeld op basis van enkele persoonskenmerken die in overleg met de begeleidingscommissie worden geselecteerd. Enkele mogelijke selecties:*
    - *Regionale verschillen*
    - *Verschillen tussen leeftijdscategorieën*
    - *Verschillen tussen de beroep(sgroep)en*
    - *Verschillen tussen cliëntgebonden en niet-cliëntgebonden functies*

- *Verschillen in werkervaring*
- *Verschillen in functieniveau*
- *Verschillen in contractsoort*
- *Verschillen in type uitstroom (uitstroom bij de werkgever, uit de branche en uit de sector)*

9) *Welke associaties zijn er te ontdekken tussen uitstroomtopic en preventietopic?*

Het onderzoek

Dit onderzoek is uitgevoerd met data van het landelijk uitstroomonderzoek van RegioPlus en Presearch. Fase 1 betrof data verzameld in de RegioPlus-regio Utrecht in de periode tussen 14-2-2022 en 31-1-2024. Het landelijk uitstroomonderzoek biedt inzicht in de uitstroom van werknemers uit de sector zorg en welzijn. Hierin wordt aan vertrekkende werknemers onder andere gevraagd wat de reden is van hun vertrek, en hoe de werkgever hun vertrek had kunnen voorkomen. In de bovengenoemde periode betrof het 3399 personen die zijn uitgestroomd bij de aangesloten werkgevers.

Titelgegevens van deze publicatie

De gegevens uit deze publicatie mogen met de volgende bronvermelding worden gebruikt: van Schaaijk, A. & van Kooten, D. Analyse van uitstroomredenen op basis van topic modellering; Deelrapport 1: Machine learning als methode om verborgen thema's te ontdekken in grote hoeveelheden open antwoorden. Utrecht: Nivel, 2024.

Literatuur

Lemmelijn, M., Schaaijk, A. van. Van uitstroom naar behoud in de sector zorg en welzijn: een verdieping naar de uitstroomredenen in verschillende functiegroepen in de regio Utrecht. Utrecht: Nivel, 2023. 29 p.

Boushey, H., & Glynn, S. J. (2012). There are significant business costs to replacing employees. Center for American Progress, 16, 1-9.

Goudarzvand, S., St. Sauver, J., Mielke, M. M., Takahashi, P. Y., Lee, Y., & Sohn, S. (2019). Early temporal characteristics of elderly patient cognitive impairment in electronic health records. BMC medical informatics and decision making, 19, 1-14.

Li, Y., Rapkin, B., Atkinson, T. M., Schofield, E., & Bochner, B. H. (2019). Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. Quality of Life Research, 28, 1441-1455.

Westrupp, E. M., Greenwood, C. J., Fuller-Tyszkiewicz, M., Berkowitz, T. S., Hagg, L., & Youssef, G. (2022). Text mining of Reddit posts: Using latent Dirichlet allocation to identify common parenting issues. PLoS One, 17(2), e0262529.

Inoue, M., Fukahori, H., Matsubara, M., Yoshinaga, N., & Tohira, H. (2023). Latent Dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing. Japan Journal of Nursing Science, 20(2), e12520.

Bijlage A

De verwijderde zinsdelen bij de uitstroomredenen: 'jaar', 'ivm', 'andere organisatie', 'nieuwe organisatie', 'nieuwe baan', 'andere baan', 'ander werk', 'nieuw werk', 'werken', 'waardoor', 'verder', 'waar', 'toe', 'wel', 'nieuwe uitdaging', 'andere functie', 'andere uitdaging', 'nieuwe functie', 'werk', 'baan', 'baan gevonden', 'binnen', 'zorg', 'gaan', 'goed', 'wilde', 'graag', 'steeds', 'erg', 'vind', 'heel', 'vanuit', 'kwam', 'helaas', 'willen', 'baan gevonden', 'past', 'daarnaast', 'anders', 'kreeg', 'functie', 'andere organisatie', 'ander', 'nieuwe organisatie'. Daarnaast zijn alle persoonsnamen en namen van organisaties verwijderd.

De verwijderde zinsdelen voor de preventiemogelijkheden: 'alle', 'waar', 'toe', 'ivm', 'werken', 'werk', 'ipv','jaar', 'etc', 'ten', 'voorkomen', 'wij', 'niks', 'niets', 'nee', 'waardoor', 'dmv', 'zie vorige', 'denk', 'nvt', 'ontslag', 'ontslag genomen' , 'hadden', 'ander', 'daarnaast', 'vpk', 'tevens', 'qua', 'zoals eerder'. Ook hier zijn alle persoonsnamen en namen van organisaties verwijderd.