Textmining in toezicht

Een exploratieve studie
naar de toegevoegde waarde van
textmining voor de
Inspectie voor de Gezondheidszorg en Jeugd

Januari 2019

Radboud Universiteit

Radboudumc

Dit onderzoek is uitgevoerd in de Academische Werkplaats Toezicht. In deze werkplaats werken samen ZonMw, de IGZ en vier kennisinstituten: IQ healthcare (Radboudumc, Nijmegen), instituut Beleid & Management Gezondheidszorg (Erasmus Universiteit Rotterdam), NIVEL (Utrecht) en EMGO+ (VUmc, Amsterdam). In de Academische Werkplaats Toezicht wordt een door ZonMw gefinancierd onderzoeksprogramma uitgevoerd naar de effectiviteit van toezicht en de determinanten daarvan. Doel van het onderzoek is een bijdrage te leveren aan de verbetering en vernieuwing van het toezicht en de effectiviteit van het toezicht te verhogen.

Leden projectgroep
Iris Hendrickx, Centre for Language and Speech Technology, Radboud University
Tim Voets, Centre for Language and Speech Technology, Radboud University
Sander Ranke MSc, onderzoeker, IQ healthcare, Radboudumc
Tijn Kool MD PhD, senior onderzoeker, IQ healthcare, Radboudumc

Dit is een publicatie van het Scientific Center for Quality of Healthcare (IQ healthcare), Radboudumc en het Centre for Language and Speech Technology, Radboud University, De studie is uitgevoerd in opdracht van ZonMw.

Radboudumc ZonMw Radboud Universiteit IN DEI NOMINE FELICITER

Inhoudsopgave

Bijbehorend supplement

Hendrickx I, Voets T, Ranke S, Kool T. Supplement bij het rapport Textmining in toezicht. Een exploratieve studie naar de toegevoegde waarde van textmining voor de Inspectie voor de Gezondheidszorg en Jeugd. Nijmegen: IQ healthcare, Radboudumc/ Centre for Language and Speech Technology, Radboud University, 2018.

Samenvatting

Inleiding

Het doel van deze studie is om automatische tekstanalysetechnieken (textmining) toe te passen op de klachten die bij het LMZ binnenkomen om zo eventuele trends en patronen te vinden die meer inzicht geven in de aard en de relevantie van de klachten. We richten ons op de volgende onderzoeksvragen:

- Wat is er bekend in de literatuur over textmining en toezicht in de zorg? In welke mate gebruiken andere inspecties of inspecties in het buitenland textmining?
- Welke groepen van klachten kunnen we automatisch onderscheiden en wat zijn de meest voorkomende onderwerpen in de klachten?
- Kunnen we de ernst van de klachten bepalen door middel van automatische classificatie en met behulp van onder meer sentimentanalyse?
- Welke patronen in de set van klachten kunnen wijzen op een verhoogd risico dat de IGJ nader moet bekijken?

Zo hopen we automatische tools te ontwikkelen om trends en patronen te vinden die meer inzicht zullen geven in de aard en de relevantie van de klachten. Het ontwikkelen van deze textmining tools die automatisch ernst kunnen bepalen of relevante patronen kunnen ontdekken, kost eenmalig tijd en inspanning, maar het voordeel is dat de tools daarna volledig automatisch kunnen worden ingezet in de praktijk.

Methode

Voor het beantwoorden van de onderzoekvraag zijn de volgende deelonderzoeken uitgevoerd.

1. Literatuuronderzoek

Als eerste stap van dit onderzoek is gestart met een verkennende analyse van de wetenschappelijke en grijze literatuur van het gebruik van textmining in de zorg in het algemeen en voor het toezicht in het bijzonder.

2. Ervaringen van andere inspecties

Om een beeld te krijgen van de activiteiten van de andere inspecties met textmining in het toezicht heeft op 13 juni 2018 een bijeenkomst plaatsgevonden met vertegenwoordigers van de IGJ, de Nederlandse Voedsel- en Warenautoriteit (NVWA), de Inspectie Sociale Zaken en Werkgelegenheid (ISZW) en de Inspectie Leefomgeving en Transport (ILT).

3. Voorbereiding

In dit deelonderzoek is ten eerste een veilige digitale omgeving gecreëerd om de analyses te doen met de gewenste textminingsoftware. Ten tweede zijn de klachten die gearchiveerd zijn door het LMZ geëxtraheerd en verwerkt tot een formaat dat geschikt is voor textmining.

4. Groepering klachten

In dit deelonderzoek hebben we bestudeerd in hoeverre we de klachten op automatische wijze kunnen groeperen en hoe we de meest voorkomende onderwerpen in de klachten kunnen vaststellen. We hebben drie verschillende technieken toegepast.

5. Bepalen van de ernst

Dit deelonderzoek richtte zich op de vraag in hoeverre we de ernst van de klachten automatisch kunnen bepalen. We hebben hiervoor automatische classificatiemodellen gebruikt die we trainden op de door het LMZ handmatig toegekende labels die ernst indiceren. Door te trainen op de handmatig gelabelde klachten, leert het model welke woorden in de klachttekst geassocieerd worden met ernstige of niet-ernstige gevallen.

6. Identificeren van patronen die wijzen op risico's

Ten slotte hebben we een analyse gedaan per organisatie waarover meerdere klachten zijn binnengekomen. Hierbij hebben we zowel de informatie uit de voorgaande deelonderzoeken gebruikt zoals de voorspelde ernst en de clusterlabels per klacht, als een analyse om de patronen per organisatie op te sommen. We hebben de klachten per organisatie geordend en een lijst

gemaakt van de organisaties waar minstens tien meldingen voor zijn gemaakt. Vervolgens hebben we voor iedere organisatie een gemiddelde berekend van de voorspelde ernstlabels en de lijst gesorteerd op voorspelde ernst.

Kernbevindingen

Literatuuronderzoek en ervaringen van andere inspecties

Er is weinig literatuur beschikbaar over de relatie tussen textmining en gebruik ervan in toezicht. Er zijn een paar studies die positieve resultaten laten zien, wanneer aan een aantal randvoorwaarden is voldaan zoals eenduidig (klinisch) taalgebruik, het overeenkomen van gegevens uit verschillende bronnen en dat databestanden volgens een standaard format ingevuld worden.

De vier inspecties concludeerden dat zij met beperkte analyse en inspanningen interessante inzichten kunnen verkrijgen. Een essentiële randvoorwaarde hiervoor is investeren in datasciencevaardigheden en faciliteiten zoals een datasciencelab, mede omdat de datapreparatie veel tijd kost.

Voorbereiding

Het gereedmaken van een veilige digitale omgeving waar de software voor textmining kon draaien heeft vele maanden geduurd. De uiteindelijke oplossing was omslachtig waardoor veel tijd verloren is gegaan tijdens het onderzoek aan connectieproblemen. We merken hierbij op dat de voorbereidende fase bij een praktijktoepassing van textmining slechts eenmalig inspanning en tijd kost, en dat daarna de ontwikkelde software volledig automatisch en snel kan worden toegepast.

Groepering klachten

We hebben drie technieken gebruikt die de klachten groeperen. Als eerste hebben we gekeken naar de handmatig toegekende categorieën en een *classifie*r op deze 45 categorieën getraind. De prestatie hiervan was matig tot redelijk. De meest frequente categorie bleek een onduidelijke categorie te zijn waardoor de classifier deze vaak verwarde met andere categorieën waardoor de gemiddelde score veel lager uitkwam. Als tweede hebben we de clusteringtechniek toegepast die gelijkende klachtbeschrijvingen groepeert in clusters. De resulterende clusters zijn handmatig geëtiketteerd door het LMZ en beoordeeld op de samenhang binnen iedere groep op basis van de top 20 woorden van ieder cluster. De indeling in 50 clusters werd het meest geschikt geacht en 75% van deze clusters kan worden gezien als voldoende samenhangend. De derde techniek die we hebben gebruikt waarbij we automatisch onderwerpen binnen klachtbeschrijvingen zochten, leverde minder goede resultaten.

Bepalen van de ernst

Hoewel we hebben geëxperimenteerd met verschillende manieren om de informatie in de klachtteksten te representeren (ook *featurerepresentaties* genoemd), bleek de beste oplossing voor de ernstclassificatie een zogenoemde eenvoudige bag-of-words representatie. Hierbij wordt een document gerepresenteerd als een ongeordende lijst van de meest voorkomende woorden en woordparen in het document. De woorden in de klachtenbeschrijvingen bevatten voldoende informatie om ernst te kunnen herkennen met een redelijk resultaat. Alle andere featurerepresentaties die we hebben gebruikt leverden geen verbetering op, noch een combinatie ervan.

Identificeren van patronen die wijzen op risico's

We hebben een overzicht gemaakt van alle organisaties waar tenminste tien klachten voor waren. Voor deze organisatie hebben we een zogenoemde n-gram analyse gedaan. N-grammen zijn woorden die (vaak) naast elkaar in de tekst voorkomen. Het analyseprogramma telt hoe vaak woorden los van elkaar en naast elkaar voorkomen om zo veel voorkomende woordparen te ontdekken. Hiermee hebben we per organisatie de termen inzichtelijk gemaakt die specifiek zijn voor deze organisatie in de aan deze organisatie gerelateerde klachten. De resultaten laten zien dat de opschoning die in Deelonderzoek 3 heeft plaatsgevonden niet alle delen van de teksten heeft verwijderd waar adresgegevens staan of de termen die bij de invulvelden van een e-mail of contactformulier horen. Hier is duidelijk nog ruimte voor verbetering.

Naar aanleiding van onze bevindingen doen we de volgende aanbevelingen.

1. Gebruik de uitkomsten van dit onderzoek bij het aanpassen van de categorie-indeling van klachten die het LMZ gaat doen.

Zowel de gevonden wetenschappelijke literatuur over klachten categorisatie als de analyses waarin klachten op automatische manieren gegroepeerd werden in verschillende indelingen, kan het LMZ gebruiken als leidraad bij de op handen zijnde categorie-herindeling.

2. Ontwerp en implementeer een toepassing die klachten, binnengekomen via de e-mail of het contactformulier, automatisch voorziet van een ernstlabel

De klachtencollectie zal gevallen bevatten die wel redelijk ernstig waren maar niet ernstig genoeg voor triage, of die niet getriageerd zijn maar waarvan de inspecteurs wel op de hoogte gebracht zouden moeten worden. Met deze toepassing kan de IGJ een deel van de redelijk ernstige klachten selecteren. Voor deze toepassing is een groot deel van het werk al in dit onderzoek verricht.

3. Start een vervolgonderzoek om de resultaten van het aanwijzen van risico's te evalueren

Dit onderzoek heeft voor een flink aantal organisaties lijsten opgeleverd met aanwijzingen over mogelijke risico's voor deze specifieke zorgorganisaties. Om de gevonden resultaten te plaatsen in een brede context en te beoordelen in hoeverre de automatische analyse inderdaad nuttig is, is interpretatie door inspecteurs nodig.

4. Creëer een aparte veilige digitale omgeving waar de specifieke textminingsoftware kan functioneren

Dit onderzoek heeft veel vertraging ondervonden door het gebrek aan een goed werkende digitale omgeving voor textmining. Bij vervolgonderzoek is een dergelijk omgeving een absolute randvoorwaarde om effectief en efficiënt praktische toepassingsmogelijkheden van textmining voor de IGJ te onderzoeken.

INLEIDING
&
METHODE

Achtergrond

Inleiding en doel onderzoek

Het Landelijk Meldpunt Zorg (LMZ) ontving in 2016 6455 klachten. Deze klachten worden handmatig gecategoriseerd en vervolgens beoordeeld op ernst. Een klein deel van de klachten, in 2016 waren dat er 1106, wordt nader onderzocht door de Inspectie voor de Gezondheidszorg en Jeugd (IGJ). Zij beoordeelde in 2016 voor ongeveer de helft van deze geselecteerde klachten dat nader onderzoek nodig was. Een groot deel van de klachten wordt dus niet nader geanalyseerd omdat daar in eerste instantie geen aanleiding toe is. De IGJ gaat ervan uit dat deze gegevens potentieel wel van belang zijn voor het toezicht. Mogelijk bevatten deze klachten trends en patronen die door individuele analyse niet zichtbaar worden maar wel zinvol zijn voor het risicotoezicht. De IGJ wil derhalve actief op zoek naar informatie die zij mogelijk heeft gemist door analyse van de individuele klachten. De ervaring heeft de IGJ geleerd dat er mogelijkheden zijn om met tools data te extraheren uit ongestructureerde tekst. Om echter optimaal gebruik te kunnen maken van deze bronnen, is het belangrijk de ongestructureerde data te structureren maar de juiste tools ontbreken momenteel.

Het doel van deze studie is om automatische tekstanalysetechnieken (textmining) toe te passen op de klachten die bij het LMZ binnenkomen om zo eventuele trends en patronen te vinden die meer inzicht geven in de aard en de relevantie van de klachten.

We richten ons op de volgende onderzoeksvragen:

- Wat is er bekend in de literatuur over textmining en toezicht in de zorg? In welke mate gebruiken andere inspecties of inspecties in het buitenland textmining?
- Welke groepen van klachten kunnen we automatisch onderscheiden en wat zijn de meest voorkomende onderwerpen in de klachten?
- Kunnen we de ernst van de klachten bepalen door middel van automatische classificatie en met behulp van onder meer sentimentanalyse?
- Welke patronen in de set van klachten kunnen wijzen op een verhoogd risico dat de IGJ nader moet bekijken?

Zo hopen we automatische tools te ontwikkelen om trends en patronen te vinden die meer inzicht zullen geven in de aard en de relevantie van de klachten. De verwachting is om zo potentiële risicogevallen te ontdekken die op basis van de behandeling van individuele klachten nog niet waren gesignaleerd. Hiermee kan de IGJ meer inzicht krijgen in de toegevoegde waarde van textmining voor het toezicht op de zorg

Zes deelonderzoeken

Om deze onderzoeksvragen te beantwoorden hebben we zes deelonderzoeken uitgevoerd.

Deelonderzoek 1: Literatuuronderzoek

Als eerste stap van dit onderzoek is gestart met een verkennende analyse van de wetenschappelijke en grijze literatuur van het gebruik van textmining in de zorg in het algemeen en voor het toezicht in het bijzonder. In Pubmed zijn twee searches gedaan om relevante artikelen op te sporen.

Deelonderzoek 2: Ervaringen van andere inspecties

Om een beeld te krijgen van de activiteiten van de andere inspecties met textmining in het toezicht heeft op 13 juni 2018 een bijeenkomst plaatsgevonden met vertegenwoordigers van de IGJ, de

Nederlandse Voedsel- en Warenautoriteit (NVWA), de Inspectie Sociale Zaken en Werkgelegenheid (ISZW) en de Inspectie Leefomgeving en Transport (ILT). De inzichten uit deze bijeenkomst, uit contact met de General Medical Council in het Verenigd Koninkrijk, gecombineerd met inzichten uit de literatuur, kunnen worden gebruikt om de randvoorwaarden te formuleren voor het toepassen van textmining als tool voor risicotoezicht.

Deelonderzoek 3: Voorbereiding

In dit deelonderzoek is ten eerste een veilige digitale omgeving gecreëerd om de analyses te doen met de gewenste textminingsoftware. Ten tweede zijn de klachten die gearchiveerd zijn door het LMZ geëxtraheerd en verwerkt tot een formaat dat geschikt is voor textmining. Eerst is op basis van vooraf geformuleerde criteria een subset van klachten gekozen om te exporteren. Vervolgens zijn de invulvelden in het Topdesksysteem, dat het LMZ gebruikt om de klachten te registreren, geselecteerd die mogelijk relevant waren voor textmining. Deze informatie is uit het Topdesksysteem geëxporteerd in het juiste format en is verder geanalyseerd en bewerkt. Ten slotte zijn de invulvelden van de klachten gekozen die we daadwerkelijk gebruiken in de textminingexperimenten.

Deelonderzoek 4: Groepering klachten

In dit deelonderzoek hebben we bestudeerd in hoeverre we de klachten op automatische wijze kunnen groeperen en hoe we de meest voorkomende onderwerpen in de klachten kunnen vaststellen. We hebben drie verschillende technieken toegepast:

1. automatische documentclassificatie van klachten om categorieën te voorspellen;
2. Clustering, een niet-gesuperviseerde techniek[1] om de klachten die op elkaar lijken te groeperen.
3. Latent Diriclet Allocation (LDA), ook een niet-gesuperviseerde techniek om onderwerpen in klachten te vinden.

We maakten gebruik van bekende algoritmen voor Machine Learning, een techniek om computers nieuwe stappen te leren die niet expliciet zijn geprogrammeerd. We hebben gekozen voor algoritmes die geschikt zijn voor tekstclassificatie en die algemeen toegankelijk zijn doordat ze in openbare tool kits geïmplementeerd zijn. Voor de niet-gesuperviseerde technieken (2. en 3.) hebben we een beoordeling laten doen door een een LMZ-medewerker om te bepalen in hoeverre de gevonden LDA-onderwerpen en clusters samenhangend waren.

Deelonderzoek 5: Bepalen van de ernst

Dit deelonderzoek richtte zich op de vraag in hoeverre we de ernst van de klachten automatisch kunnen bepalen. We hebben hiervoor automatische classificatiemodellen gebruikt die we trainden op de door het LMZ handmatig toegekende labels die ernst indiceren. Door te trainen op de handmatig gelabelde klachten, leert het model welke woorden in de klachttekst geassocieerd worden met ernstige of niet-ernstige gevallen. In overleg met het LMZ hebben we de drie volgende velden uit Topdesk gebruikt als labels in onze experimenten om het model op te trainen:

- Start triageproces
- Triagebeslissing
- Mogelijk IT

We hebben de dataset van ongeveer 22000 klachten gesplitst in een training- en een testset. We hebben de eerste helft van 2017 als testset genomen en alle andere klachten als trainingsmateriaal. Voor optimalisatie-experimenten is steeds de trainingset gebruikt waarna het beste systeem is gebruikt om de klachten in de testset te labelen. We hebben geëxperimenteerd met verschillende featurerepresentaties.

---

1. [1] Niet-gesuperviseerd betekent dat er geen labels van tevoren als uitgangspunt is meegegeven bij de uitvoering.

Deelonderzoek 6: Identificeren van patronen die wijzen op risico's

In de laatste stap van deze exploratieve studie hebben we een analyse gedaan per organisatie waarover meerdere klachten zijn binnengekomen. Hierbij hebben we zowel de informatie uit de voorgaande deelonderzoeken gebruikt zoals de voorspelde ernst en de clusterlabels per klacht. Uit Deelonderzoek 4 bleek bij handmatige evaluatie dat de LDA topics niet informatief waren voor onze doeleinden en zijn daarom ook niet gebruikt in Deelonderzoek 6.

Daarnaast hebben we ook een n-gram analyse uitgevoerd om de specifieke termen per organisatie op te sommen. We hebben de klachten per organisatie geordend en een lijst gemaakt van de organisaties waar minstens tien meldingen voor zijn gemaakt. Vervolgens hebben we voor iedere organisatie een gemiddelde berekend van de voorspelde ernstlabels en de lijst gesorteerd op voorspelde ernst. Hierbij zijn ook de frequenties en gemiddelden van de handmatig gelabelde indicaties van ernst betrokken. Dit overzicht van organisaties met ernst-indicatie, clusters en specifieke termen kunnen IGJ-inspecteurs gebruiken om snel inzicht te krijgen in de klachtencollectie van het LMZ.

ONDERZOEKS-RESULTATEN

Deelonderzoek 1: Literatuuronderzoek

Er is weinig literatuur beschikbaar over de relatie tussen textmining en gebruik ervan in toezicht. Binnen de gezondheidzorg zijn er wel verschillende studies gedaan over het gebruik van textmining. Dit is met name gedaan bij het categoriseren van klachten en het ontdekken van patronen in patiëntendossiers. Hoewel deze studies positieve resultaten laten zien, zijn er wel een aantal randvoorwaarden waaraan voldaan moet zijn:

- Eenduidig (klinisch) taalgebruik;
- Gegevens uit verschillende bronnen moeten met elkaar overeenkomen;
- Databestanden die gebruikt worden moeten volgens een standaard format ingevuld worden en ook consequent. Op deze manier kan veel missende, incomplete, niet-gestandaardiseerde data voorkomen worden.

Wanneer aan deze voorwaarden wordt voldaan lijkt textmining een veelbelovende techniek te zijn om patronen in ongestructureerde data op te sporen.

Deelonderzoek 2: Ervaringen van andere inspecties

De vier gevraagde Nederlandse inspecties zijn alle bezig de mogelijkheden te onderzoeken om patronen, trends en informatie te halen uit allerlei bronnen zoals meldingen die binnenkomen of internetfora. Meestal gebeurt dat voor specifieke thema's. Hiermee hopen de inspecties snel inzicht te krijgen in onderwerpen die spelen op een specifiek thema.

Een belangrijke conclusie was voor alle inspecties dat de datapreparatie veel tijd kost. Het is vaak nodig om de verschillende gegevens met elkaar te koppelen en deze gegevens op te schonen wat in de praktijk een uitdaging blijkt te zijn. De kwaliteit van de meldingen is vaak beperkt met veel spellingsfouten en vaak te korte teksten. Ook werkt informatie zoals namen en adressen storend op het proces en is er soms een beperkte capaciteit van de hardware. De thematische benadering vergt een intensieve samenwerking tussen materiedeskundigen en data-analisten.

De vier inspecties concludeerden dat zij met beperkte analyse en inspanningen interessante inzichten kunnen verkrijgen. Alle vier de inspecties zullen de komende jaren het onderzoek naar de toepassing van textmining uitbreiden en textmining verder toepassen om risicogerichter te werken en de signalering te verbeteren. Een essentiële randvoorwaarde hiervoor is investeren in data-sciencevaardigheden en faciliteiten zoals een data-sciencelab.

Deelonderzoek 3: Voorbereiding

Het gereedmaken van een veilige digitale omgeving waar de software voor textmining kon draaien heeft vele maanden geduurd. De formele belemmering om als onderzoeker direct contact te hebben met de operationeel verantwoordelijke van de externe organisatie die verantwoordelijk was voor het gereedmaken van de omgeving, was hierin een belangrijke vertragende factor. Er waren technische uitdagingen maar die waren bij direct contact vaak oplosbaar. De uiteindelijke oplossing was omslachtig waardoor veel tijd verloren is gegaan tijdens het onderzoek aan connectieproblemen.

De klachten moesten worden geëxporteerd uit het Topdesksysteem dat het LMZ gebruikt om de klachten te registreren. Voor het prepareren van de data hebben we de volgende stappen doorlopen:

- Het selecteren van de klachten;
- Het selecteren van de invulvelden uit het Topdesksysteem;
- Het exporteren van de klachten naar een formaat geschikt voor de textmininganalyse;

- Het bewerken van de data door anonimisatie en door metadata te normaliseren door te zorgen voor consistente labels en door opmaak uit de formulieren, brieven of e-mails te verwijderen;
- Het taalkundig verrijken van de tekst.

Om textmining in de praktijk toe te passen, zijn deze voorbereidende stappen slechts eenmalig heel tijdrovend omdat er software moeten worden geschreven om de stappen uit te voeren. Wanneer deze software klaar is, kan deze in het vervolg in de praktijk gewoon automatisch worden toegepast. Ook de taalkundige verrijking is een automatisch proces dat geen verdere handmatige voorbereiding nodig heeft. Wanneer een andere database wordt onderzocht, zullen wel weer de basisstappen van textmining doorlopen moeten worden.

Deelonderzoek 4: Groepering klachten

We hebben drie technieken gebruikt die de klachten groeperen. Als eerste hebben we gekeken naar de handmatig toegekende categorieën en een *classifie*r op deze 45 categorieën getraind. De gemiddelde F1-score[2] van deze classifier op de testset, is matig tot redelijk (met een waarde van 0,49) en onvoldoende om automatisch toe te passen in de praktijk. Voor slechts sommige categorieën zoals 'medicatie' en 'medisch hulpmiddel' was de classifier in staat om een hogere F-score te halen die wel betrouwbaar genoeg is om toe te passen. De meest frequente categorie bleek een onduidelijke categorie te zijn waardoor de classifier deze vaak verwarde met andere categorieën waardoor de gemiddelde score veel lager uitkwam.
Als tweede hebben we de clusteringtechniek toegepast waarbij de resultaten handmatig zijn geëtiketteerd door het LMZ en beoordeeld op samenhang op basis van de top 20 woorden van ieder cluster. De indeling in 50 clusters werd het meest geschikt geacht en 75% van deze clusters kan worden gezien als voldoende samenhangend en bruikbaar als groepsindeling.

Als derde hebben we de LDA-topics techniek gebruikt. De resultaten hiervan bleken veel minder goed te interpreteren te zijn. Er was zelfs een extra opschoningsronde nodig omdat er te veel getallen, afkortingen en betekenisloze woorden in de topics stonden. Na opschoning bleek 45% van de topics een coherent en herkenbaar onderwerp te bevatten.

Deelonderzoek 5: Bepalen van de ernst

De classifier op basis van de zogenoemde Bag-of-Words-techniek, bleek de beste oplossing voor de ernstclassificatie. De woorden in de klachtenbeschrijvingen bevatten voldoende informatie om ernst te kunnen herkennen met een redelijk resultaat. Alle andere featurerepresentaties die we hebben gebruikt (word embeddings, Linguistic Inquiry and Word Count, documentkenmerken, sentiment)[3] leverden geen verbetering op, noch een combinatie van alle features. We gebruikten daarbij drie soorten labels ('start triageproces', Triagebeslissing' en 'Mogelijk IT') om te leren welke klachten ernstig waren.

Het valt op dat de 'Mogelijk IT' labels lastiger te voorspellen zijn dan de andere twee ernstlabels. Bij navraag bij het LMZ bleek dat dit label alleen is gebruikt bij digitaal binnengekomen klachten en niet bij telefonische klachten. Ze betreffen dus maar een deel van de ernstige gevallen. De classifier heeft alleen kennis van de omschrijving in het verzoekveld en heeft geen kennis van de manier van indienen. Het toevoegen van informatie uit het invulveld 'Soort binnenkomst' of vooraf filteren op dit veld, zijn opties om de score van de classifier verbeteren voor deze gevallen. Dit

[2] De F1-score is het gemiddelde van *volledigheid* (hoeveel van de correcte labels zijn geselecteerd?) en de *precisie* (hoeveel geselecteerde labels zijn ook echt correct?)
[3] Zie het Supplement voor een gedetailleerde uitleg van deze verschillende featurerepresentaties.

levert echter een classifier op die slechts toepasbaar is op een deel van de klachten. Het ‘Mogelijk IT’ label bleek daarom minder geschikt als ernstindicator.

Onze verwachting dat de verschillende featurerepresentaties zouden helpen bij de ernstclassificatietaak is niet bevestigd. We hadden ook verwacht dat we een samenhang zouden vinden tussen negatieve sentimentwaarden en ernstige klachten, en dat de sentiment features zouden bijdragen aan een verbeterde prestatie van de ernstclassifier. Dat bleek niet het geval. Bij een aanvullende analyse waarbij we hebben gekeken naar correlaties tussen verschillende categorieën en sentiment, bleek dat ook op dit vlak de sentiment features niet informatief waren om categorieën van elkaar te onderscheiden.

Deelonderzoek 6: Identificeren van patronen die wijzen op risico’s

We hebben een overzicht gemaakt van alle organisaties waar tenminste tien klachten voor waren en ook een n-gram analyse gedaan voor iedere organisatie om de specifieke termen inzichtelijk te maken in de aan deze organisatie gerelateerde klachten. Zo konden we per organisatie de termen en frasen die specifiek zijn voor deze organisatie, in beeld brengen. De resultaten laten zien dat de opschoning die in Deelonderzoek 3 heeft plaatsgevonden niet alle delen van de teksten heeft verwijderd waar adresgegevens staan of de termen die bij de invulvelden van een e-mail of contactformulier horen. Hier is duidelijk nog ruimte voor verbetering.

DISCUSSIE

Implicaties voor het toezicht van de IGJ

Bovengenoemde resultaten bieden voor de IGJ op verschillende manieren mogelijkheden om hun werkprocessen te verbeteren en te ondersteunen.

Ten eerste kan deze analyse inzicht bieden aan het LMZ dat op korte termijn zijn categorie-indeling wil evalueren. De categorie-indeling die het LMZ nu gebruikt, bleek niet optimaal leerbaar te zijn voor de classifier, grotendeels door een zeer dominante onduidelijke categorie ('oneens medisch handelen') en omdat sommige categorieën vaak met elkaar werden verward. De door ons gehanteerde textminingtechnieken zijn bruikbaar om klachten op een automatische manier te groeperen in verschillende indelingen. De clustering en LDA-technieken bieden een indeling die geheel losstaat van enige handmatige indelingen en is puur gebaseerd op de klachtenteksten zelf.

Ten tweede kan de IGJ dit onderzoek gebruiken om klachten die via mail of een contactformulier zijn verstuurd naar het LMZ, direct bij binnenkomst te sorteren op hun prioriteit aan de hand van een automatisch voorspeld label wat betreft risico. Ook inspecteurs van de IGJ die een overzicht willen hebben over de gehele klachtencollectie voor een organisatie, een zorgverlener of een sector, kunnen baat hebben bij een automatisch voorspeld label. Deze klachtencollectie zal gevallen bevatten die wel redelijk ernstig waren maar niet ernstig genoeg voor triage, of die niet getriageerd zijn om een andere reden zoals de wens van de burger om niet te worden voorgelegd. Het ontwikkelen van textmining tools die toepasbaar zijn in de praktijk voor de inspecteurs, kost een eenmalige inspanning maar kunnen daarna volledig automatisch worden toegepast.

Mogelijkheden voor vervolgonderzoek

We hebben een aantal verschillende mogelijkheden geëvalueerd om de klachten te representeren. We noemen hier nog even kort een aantal mogelijke opties die niet binnen de huidige exploratieve studie zijn onderzocht maar goede vervolgstappen zouden kunnen zijn. Zo zou het vervangen van de algemene LiWC woordenlijst door een specifieke woordenlijst met termen die door inspecteurs en LMZ-medewerkers is samengesteld, misschien wel een verbetering kunnen opleveren.
In deze studie hebben we ons gericht op een zo hoog mogelijke F1-score, het gemiddelde van *recall* (hoeveel relevante klachten zijn geselecteerd?) en *precisie* (hoeveel geselecteerde klachten zijn relevant?) maar bij een praktische toepassing waar een LMZ-medewerker of inspecteur een gesorteerde lijst klachten doorloopt, zou het nuttiger kunnen zijn om de ernstclassifier te optimaliseren op volledigheid waardoor zoveel mogelijk ernstige gevallen worden gedetecteerd ten koste van de precisie.

De resultaten uit Deelonderzoek 6 zijn niet expliciet geëvalueerd omdat hiervoor de inzet van een inspecteur nodig is die de gevonden resultaten kan plaatsen in een brede context en kan beoordelen in hoeverre de automatische analyse inderdaad nuttig is, en in hoeverre deze analyse helpt om de klachtencollectie inzichtelijk en overzichtelijk te maken. In deze studie hebben we ons alleen gericht op een evaluatie van de textminingtechnieken zelf. In een vervolgonderzoek zou een expliciete evaluatie door de inspecteurs die de betreffende toezichtsobjecten kennen, een logische vervolgstap zijn.

Een fundamenteel onderdeel bij een dergelijke evaluatie is om niet alleen naar de inhoud van de resultaten te kijken, maar ook om de resultaten op een goede en bruikbare manier zodanig te visualiseren dat een inspecteur gemakkelijk door de resultaten kan bladeren waarbij de resultaten ook gekoppeld zijn aan de oorspronkelijke bron, de klachten zelf ("provenance").

Aanbevelingen

Naar aanleiding van onze bevindingen doen we de volgende aanbevelingen.

1. Gebruik de uitkomsten van dit onderzoek bij het aanpassen van de categorie-indeling van klachten die het LMZ gaat doen.
Zowel de gevonden wetenschappelijke literatuur over klachten categorisatie als de analyses waarin klachten op automatische manieren gegroepeerd werden in verschillende indelingen, kan het LMZ gebruiken als leidraad bij de op handen zijnde categorie-herindeling.

2. Ontwerp en implementeer een toepassing die klachten, binnengekomen via de e-mail of het contactformulier, automatisch voorziet van een ernstlabel
De klachtencollectie zal gevallen bevatten die wel redelijk ernstig waren maar niet ernstig genoeg voor triage, of die niet getriageerd zijn maar waarvan de inspecteurs wel op de hoogte gebracht zouden moeten worden. Met deze toepassing kan de IGJ een deel van de redelijk ernstige klachten selecteren. Voor deze toepassing is een groot deel van het werk al in dit onderzoek verricht.

3. Start een vervolgonderzoek om de resultaten van het aanwijzen van risico's te evalueren
Dit onderzoek heeft voor een flink aantal organisaties lijsten opgeleverd met aanwijzingen over mogelijke risico's voor deze specifieke zorgorganisaties. Om de gevonden resultaten te plaatsen in een brede context en te beoordelen in hoeverre de automatische analyse inderdaad nuttig is, is interpretatie door inspecteurs nodig.

4. Creëer een aparte veilige digitale omgeving waar de specifieke textminingsoftware kan functioneren
Dit onderzoek heeft veel vertraging ondervonden door het gebrek aan een goed werkende digitale omgeving voor textmining. Bij vervolgonderzoek is een dergelijk omgeving een absolute randvoorwaarde om effectief en efficiënt praktische toepassingsmogelijkheden van textmining voor de IGJ te onderzoeken.