Rijksoverheid

Pilot automatische e-mailclassificatie

Rapport

24 april 2026

Versie: 1.0

Rijksprogramma
Duurzaam
Digitale
Informatiehuishouding

Versiebeheer

Datum	Versie
17 december 2025	Versie 0.1, eerste opzet van het rapport.
19 januari 2026	Versie 0.2, verwerking feedback van projectteam.
29 januari 2026	Versie 0.3, verwerking feedback van projectteam.
10 februari 2026	Versie 0.4, verwerking feedback van projectteam.
16 maart 2026	Versie 0.5, verwerking feedback Nationaal Archief.
1 april 2026	Versie 0.6, aangeboden aan het MT RDDI.
7 april 2026	Versie 0.7, verwerking feedback MT RDDI en opdrachtgever.

Voorwoord

De Nederlandse overheid is aan strikte regels gebonden. Op grond van de Archiefwet en de Wet open overheid (Woo) zijn overheidsorganisaties verplicht om relevante informatie, waaronder e-mail, die van belang is voor besluitvorming en taakuitvoering duurzaam toegankelijk en vindbaar te houden.

In de praktijk blijkt het lastig om structureel aan deze verplichtingen te voldoen. E-mail vormt een essentieel onderdeel van de informatiehuishouding, tegelijk lopen bestaande achterstanden in de afgelopen jaren op door de toenemende volumes, uitstel, beperkte investeringen en het ontbreken van samenhangend e-mailbeheer. Hoewel de verantwoordelijkheid formeel bij individuele medewerkers en bestuurders ligt, maken de omvang van het e-mailverkeer, de verwevenheid van zakelijke en niet-zakelijke communicatie en het gebrek aan ondersteunende voorzieningen deze verantwoordelijkheid moeilijk uitvoerbaar.

Tegen deze achtergrond start de pilot automatische e-mailclassificatie. In deze pilot is onderzocht of geautomatiseerde ondersteuning kan bijdragen aan het verbeteren van e-mailbeheer. Daarbij is gewerkt met vijf categorieën, waarbij het onderscheid tussen zakelijke en niet-zakelijke e-mail centraal staat.

De pilot bouwt voort op een eerdere verkenning van het Nationaal Archief uit 2018, waarin wordt vastgesteld dat automatische e-mailclassificatie technisch mogelijk is. In dit vervolg is zichtbaar dat een volgende stap wordt gezet. Niet alleen ontwikkelt de techniek zich verder, ook de juridische duiding scherpt aan. Daarmee ontstaat een steviger fundament om deze aanpak niet langer uitsluitend als experiment te beschouwen, maar als een mogelijke bouwsteen voor toekomstig e-mailbeheer binnen de rijksoverheid.

"Deze pilot maakt duidelijk dat automatische e-mailclassificatie kan bijdragen aan privacy- en gegevensbescherming, mits zorgvuldig ingericht."
- Chief Privacy Officer, Ministerie van Algemene Zaken

Een centrale vraag in deze pilot is niet alleen of automatische classificatie technisch mogelijk is, maar ook of deze juridisch toelaatbaar is. De wijze waarop e-mails worden gecategoriseerd raakt direct aan privacy en gegevens-bescherming. De resultaten laten zien dat automatische classificatie binnen de geldende juridische en privacykaders mogelijk is, mits deze zorgvuldig wordt ingericht en transparant wordt toegepast.

De pilot is uitgevoerd door een multidisciplinair projectteam en in nauwe samenwerking tussen verschillende disciplines. Waaronder het ministerie van SZW, het Nationaal Archief en andere betrokken partijen. In deze samenwerking wordt expertise op het gebied van archiefvorming, recht, privacy, beleid en techniek samengebracht. Deze werkwijze is essentieel om een integrale benadering van het vraagstuk mogelijk te maken.

"Deze pilot biedt concrete aanknopingspunten en vervolgstappen binnen de rijksbrede verbetering van de informatiehuishouding."
- Directeur Open Overheid, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

De opbrengst van deze pilot is dat nieuwe kennis wordt verworven. Tegelijkertijd bieden de resultaten concrete aanknopingspunten voor verdere richting en besluitvorming. De pilot maakt zichtbaar wat op dit moment technisch, juridisch en organisatorisch mogelijk is, waar beperkingen liggen en welke randvoorwaarden nodig zijn voor verantwoorde toepassing. Daarmee levert zij niet alleen inzicht, maar ook een kader voor vervolgstappen binnen de rijksbrede verbetering van de informatiehuishouding.

De blik is daarbij nadrukkelijk gericht op de toekomst. Automatische e-mailclassificatie kan bijdragen aan het beheersbaar maken van bestaande achterstanden in veiliggestelde mailboxen en de grootste meerwaarde ligt aan de voorkant van het proces. Door ondersteuning te bieden op het moment dat e-mails worden ontvangen en verstuurd, voor alle medewerkers, wordt structureel beter informatiebeheer mogelijk. Dit vraagt om bewuste keuzes in inrichting, governance en verantwoordelijkheid, en vormt een belangrijke bouwsteen voor een duurzame, uniforme en transparante aanpak van e-mailbeheer binnen de rijksoverheid.

Met deze pilot wordt een volgende stap gezet: van pilot naar verdere verkenning van opschaling en van correctie achteraf naar ondersteuning vooraf, als basis voor een zorgvuldige en samenhangende aanpak van e-mailbeheer binnen de rijksoverheid.

Met vriendelijke groet,

Het projectteam Automatische e-mailclassificatie
Chief Privacy Officer, Ministerie van Algemene Zaken
Directeur Open Overheid, Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Inhoudsopgave

Begrippenkader

Begrip	Uitleg
Accuracy	Accuracy geeft het percentage correcte voorspellingen weer ten opzichte van het totaal aantal voorspellingen.
Artificial Intelligence (AI)	Artificial Intelligence is een verzamelnaam voor systemen of machines die taken kunnen uitvoeren waarvoor normaal gesproken menselijke intelligentie nodig is.
AUC (Area Under the Curve)	De AUC geeft aan hoe goed een model onderscheid maakt tussen categorieën over alle mogelijke drempelwaarden. Een waarde van 1 betekent perfecte classificatie, 0,5 staat gelijk aan willekeurig gokken.
Average Precision (AP)	Average Precision is de oppervlakte onder de precision-recall curve en geeft een realistischer beeld van prestaties bij ongebalanceerde datasets.
Classificering	Classificering is het labelen van e-mails op basis van de inhoud van een e-mail. De inhoud van de e-mail bepaalt welke label de e-mail krijgt.
Confidence score	De confidence score is de waarschijnlijkheid die een model toekent aan een voorspelling en geeft aan hoe zeker het model is van een classificatie.
Confidence thresholds	Betrouwbaarheidsdrempels zijn vooraf gedefinieerde minimumscores, vaak variërend van 0 tot 1 of van 0% tot 100% die in machine learning worden gebruikt om te bepalen of de voorspelling van een model betrouwbaar genoeg is om op te reageren.
Confusion matrix	Een confusion matrix is een overzicht dat laat zien hoeveel voorspellingen correct en incorrect zijn, uitgesplitst naar type fout.
F1-score	De F1-score combineert precision en recall in één maat en geeft een gebalanceerd beeld van de modelprestaties.
Human-in-the-loop	Een werkwijze waarbij menselijke controle wordt ingezet bij twijfelgevallen om de kwaliteit en betrouwbaarheid van classificaties te waarborgen.
Machine Learning (ML)	Machine learning is de techniek binnen AI die computers leert om zelfstandig patronen te ontdekken in data.
Multinominal Logistic Regression	Multinominale logistische regressie is een classificatiealgoritme dat wordt gebruikt om een nominale afhankelijke variabele met drie of meer ongeordende categorieën te voorspellen op basis van een of meer onafhankelijke variabelen.
Multinomial Naive Bayes	Multinomial Naive Bayes is een probabilistisch, supervised learning-algoritme dat vaak wordt gebruikt voor tekstclassificatie en categorieën voorspelt op basis van woordfrequenties.
Ongebalanceerde dataset	Een dataset waarin sommige categorieën veel vaker voorkomen dan andere, wat invloed heeft op de prestaties van het model.
Parameter Efficient Fine-Tuning (PEFT)	Bij Parameter-Efficient Fine-Tuning (PEFT) worden grote, vooraf getrainde modellen aangepast voor specifieke taken door slechts een kleine subset van parameters te trainen, wat de rekenkosten verlaagt.
Precision	Precision geeft aan in hoeverre positieve voorspellingen daadwerkelijk correct zijn.
PR-curve	De precision-recall curve toont de verhouding tussen precision en recall en geeft een betrouwbaarder beeld bij ongebalanceerde datasets.
PST-bestand	Een PST-bestand (Personal Storage Table) is een Outlook-gegevensbestand (.pst) dat wordt gebruikt voor het opslaan van e-mails, agenda-items, contactpersonen en andere gegevens, vaak voor archivering of back-ups.
Recall	Recall geeft aan in welke mate het model erin slaagt om alle relevante gevallen te identificeren.

Begrip	Uitleg
ROC-curve	De ROC-curve toont de verhouding tussen het percentage correct herkende positieve gevallen en het percentage fout-positieve voorspellingen bij verschillende drempelwaarden.
Regel gebaseerd	Een regelgebaseerd model neemt beslissingen op basis van vooraf gedefinieerde “als-dan”-regels die door mensen zijn opgesteld.
Threshold (drempelwaarde)	De threshold is de grenswaarde die bepaalt vanaf welke confidence score een classificatie wordt geaccepteerd of automatisch wordt verwerkt.
Transformer model	Een Transformer model is een neurale netwerkarchitectuur die sequentiële data, zoals tekst, gelijktijdig analyseert en daardoor context beter begrijpt.

Leeswijzer

Dit rapport doet verslag van de resultaten van een pilot waarbij artificiële intelligentie (AI) is ingezet voor de automatische classificatie van e-mails. Naast de technische en organisatorische uitvoering van de pilot komt ook de ontwikkeling van de techniek aan bod, inclusief de modellen die daarvoor zijn toegepast en de resultaten die zijn behaald. De opzet van het rapport is bewust breed gehouden om tegemoet te komen aan verschillende lezersbehoeften, van beleidsmatig geïnteresseerden tot technisch betrokkenen.

Voor lezers die primair geïnteresseerd zijn in de hoofdbevindingen en aanbevelingen volstaat de managementsamenvatting in combinatie met hoofdstuk 7 en 8. Lezers die meer achtergrond zoeken bij de technische en methodologische keuzes worden verwezen naar de hoofdstukken 3 tot en met 6 en de bijlagen.

De opbouw van het rapport is als volgt:

- **Hoofdstuk 1** introduceert de pilot: de aanleiding, de achtergrond en het doel waarmee de pilot is opgezet.
- **Hoofdstuk 2** beschrijft de organisatie en uitvoering van de pilot, waaronder de gevolgde processtappen, de betrokken expertises en rollen, en de waarborgen die zijn genomen voor een zorgvuldig en betrouwbaar resultaat.
- **Hoofdstuk 3** biedt een theoretisch kader ter verduidelijking van de begrippen Artificial Intelligence en Machine Learning, en licht toe hoe deze technieken zijn toegepast binnen de pilot.
- **Hoofdstuk 4** werkt het classificatiekader uit en definieert de categorieën die in de pilot zijn gehanteerd, van functionele en niet-functionele e-mails tot privé-, personeelsvertrouwelijke en partijpolitieke berichten.
- **Hoofdstuk 5** gaat in op de gebruikte data: welke e-mailboxen zijn betrokken, hoe de handmatige classificatie heeft plaatsgevonden en welke voorbereidingsstappen de data heeft doorlopen voordat deze in de modellen is ingevoerd.
- **Hoofdstuk 6** beschrijft de toegepaste modellen en legt uit welke factoren bepalend zijn voor hun uitlegbaarheid en betrouwbaarheid.
- **Hoofdstuk 7** presenteert de resultaten van het experiment, zowel de kwantitatieve trainings- en validatie-resultaten als de bredere bevindingen over de toepasbaarheid en betekenis van de uitkomsten.
- **Hoofdstuk 8** vertaalt de opgedane inzichten naar aanbevelingen en mogelijke vervolgroutes, gericht op opschaling, productontwikkeling en een rijksbrede toepassing.

De bijlagen bevatten aanvullende technische en procedurele documentatie, waaronder het stappenplan met waarborgen, een toelichting op de uitlegbaarheid van de modellen, filterlogica, kwaliteitsborging en een vergelijking tussen de pilotsituatie en een toekomstige productieomgeving.

Managementsamenvatting

De pilot automatische e-mailclassificatie laat zien dat geautomatiseerde ondersteuning bij e-mailbeheer technisch haalbaar is en binnen de geldende juridische en privacy kaders kan worden ingezet. Daarmee zet de pilot een concrete stap van experiment naar fundament voor rijksbreed e-mailbeheer.

Overheidsorganisaties zijn verplicht op grond van de Archiefwet en de Wet open overheid (Woo) om relevante informatie, waaronder e-mail, die van belang is voor besluitvorming en taakuitvoering duurzaam toegankelijk en vindbaar te houden. In de praktijk is dit moeilijk uitvoerbaar door de omvang van het e-mailverkeer, bestaande achterstanden en het ontbreken van passende ondersteuning. De pilot onderzoekt of automatische classificatie van e-mails in vijf categorieën; functioneel, niet-functioneel, privé, personeelsvertrouwelijk en partijpolitiek een bijdrage kan leveren aan het verbeteren van het e-mailbeheer.

De pilot is uitgevoerd met mailboxen van voormalige bewindspersonen en bestuursraadleden van het ministerie van SZW. Een multidisciplinair team classificeert 13.709 e-mails handmatig als trainingsdata. Twee modellen zijn ontwikkeld en vergeleken: een Logistic Regression-model (hoge transparantie) en een PEFT-taalmodel op basis van BERTje (sterkere contextuele classificatie).

Beide modellen behalen een accuracy van 85% en presteren sterk op de grote categorieën functioneel en niet-functioneel (F1-scores van respectievelijk 0,85 en 0,88). Voor de kleinere categorieën privé, personeelsvertrouwelijk en partijpolitiek zijn de prestaties lager, voornamelijk door beperkte trainingsdata. Met slechts 37 trainingsvoorbeelden voor partijpolitiek is betrouwbare herkenning van deze categorie momenteel niet mogelijk. Eenvoudige filterregels (business rules) blijken daarnaast al in staat om gemiddeld een derde van de niet-functionele e-mails automatisch te identificeren, zonder tussenkomst van een model.

De pilot bevestigt dat datakwaliteit en labeleenduidigheid bepalender zijn voor modelprestaties dan model-complexiteit. Filtering vooraf levert meer winst op dan verwacht. De beveiligingsmaatregelen voor de verwerking van gevoelige data blijken in de praktijk werkbaar, mits expliciet vastgelegd en nageleefd. De inzet van Outlook als classificatieomgeving is kwalitatief sterk maar technisch complex, een afweging die in vervolgtrajecten bewust moet worden gemaakt.

De pilot legt een fundament. De volgende stap is het versterken en toepassen van dat fundament langs twee sporen. Spoor 1 richt zich op het professionaliseren van de techniek, het vastleggen van de doelarchitectuur en het beleggen van governance. Spoor 2 richt zich op verbreding naar andere departementen en doorontwikkeling van functionaliteit. Randvoorwaarde voor beide sporen is uitbreiding van de trainingsdata voor de drie kleinste categorieën (de “3P’s”: privé, personeelsvertrouwelijk en partijpolitiek) en herziening van de handreiking “welke e-mail kan weg”.

Geleerde lessen

De pilot levert nieuwe kennis op. Naast de technische resultaten levert de uitvoering inzichten op die relevant zijn voor toekomstige trajecten. De belangrijkste lessen worden hieronder samengevat.

Data is bepalender dan modelkeuze

De modelprestaties worden niet primair begrensd door de keuze voor een eenvoudig of complex model, maar door de kwaliteit en omvang van de trainingsdata. Voor categorieën met weinig voorbeelden, met name partijpolitiek (37 voorbeelden), is betrouwbare classificatie onmogelijk, ongeacht het gebruikte model. Investeren in meer en beter gelabelde data heeft een grotere impact dan het inzetten van een geavanceerder model.

Eenvoudige filterregels zijn krachtig en onderschat

Het toepassen van business rules gebaseerd op afzenderpatronen, onderwerpregels en berichttypen blijkt gemiddeld een derde van de niet-functionele e-mails te kunnen identificeren zonder tussenkomst van een ML-model. Dit resultaat is snel realiseerbaar, volledig transparant en goed uitlegbaar richting toezicht en verantwoording. In vervolgtrajecten verdient deze aanpak een expliciete en vroege rol in de verwerkingsketen.

Labeleenduidigheid vereist continue aandacht

In de praktijk blijken grensgevallen tussen categorieën, zoals het onderscheid tussen privé en niet-functioneel of tussen functioneel en partijpolitiek, regelmatig tot discussie te leiden onder classificerende medewerkers. De 5% overlap in de dataset maakt dit zichtbaar en helpt bij het aanscherpen van definities. Dit onderstreept dat classificatierichtlijnen geen statisch document zijn, maar continu onderhoud vragen naarmate er meer e-mailtypen worden gezien.

Beveiligingsmaatregelen zijn werkbaar, niet belemmerend

Voorafgaand aan de pilot bestaat de verwachting dat strenge beveiligingseisen, zoals de volledige isolatie van de ontwikkelomgeving, de uitvoering sterk bemoeilijken. In de praktijk blijken deze maatregelen goed uitvoerbaar, mits tijdig aangevraagd en vooraf expliciet vastgelegd. De veiligheidsmaatregelen zijn daarmee geen belemmering, maar een randvoorwaarde die vroegtijdig in de planning moet worden opgenomen.

De keuze voor Outlook als classificatieomgeving heeft voor- en nadelen

Het gebruik van Outlook voor de handmatige classificatie biedt grote voordelen voor de kwaliteit: classificerende medewerkers hebben direct toegang tot metadata, bijlagen en de volledige e-mailcontext. Dit verhoogt de snelheid en nauwkeurigheid van het labelproces. Tegelijkertijd maakt de afhankelijkheid van de Outlook-omgeving automatisering van de dataverwerkingsketen complexer. In een vervolgtraject is het verstandig om deze afweging bewust te maken en te onderzoeken of een andere classificatieomgeving hetzelfde kwaliteitsniveau kan bieden met minder technische complexiteit.

Multidisciplinaire samenwerking is essentieel

De betrokkenheid van expertises op het gebied van archief, recht, privacy, beleid en techniek blijkt onmisbaar. Vraagstukken rondom de classificatiecategorieën kunnen niet puur technisch worden opgelost; ze vragen om juridische duiding en organisatorisch inzicht. In toekomstige trajecten moet deze multidisciplinaire samenwerking vanaf het begin worden ingericht, niet als aanvulling maar als structurele werkwijze.

De voorbereidingsfase

De voorbereidingsfase vraagt in de praktijk het grootste deel van de totale doorlooptijd van de pilot. Een aanzienlijk deel van de tijd, circa twee derde van het traject, wordt besteed aan afstemming met stakeholders, het organiseren van randvoorwaarden en het inrichten van de technische en organisatorische basis.

Deze fase omvat onder andere het verkrijgen van formele goedkeuringen, het voldoen aan eisen op het gebied van privacy en informatiebeveiliging, het doorlopen van inkooptrajecten en het voorbereiden van de uitvoering. Dit inzicht onderstreept dat de voorbereiding geen randvoorwaarde is, maar een bepalende fase binnen het traject. Voor toekomstige implementaties en opschaling is het van belang om deze fase expliciet te plannen en hier voldoende tijd en capaciteit voor te reserveren.

1. Inleiding

1.1 Aanleiding en achtergrond van het project

Binnen de Nederlandse overheid is informatiebeheer gebonden aan strikte regels. Op grond van de Archiefwet en de Wet open overheid (Woo) zijn overheidsorganisaties verplicht om relevante e-mails, informatie die van belang is voor besluitvorming of de uitvoering van taken, duurzaam toegankelijk en vindbaar te houden.

In een digitale werkomgeving waarin e-mail een essentieel onderdeel vormt van de informatiehuishouding, vraagt het voldoen aan deze verplichtingen om duidelijke kaders en passende ondersteuning. Het programma Open Overheid werkt aan het verbeteren van de informatiehuishouding van de Rijksoverheid, waaronder het e-mailbeheer. Hierbij wordt ingezet op twee sporen: het hoofdspoor waarbij medewerkers worden gevraagd e-mails toe te voegen aan dossiers en het tijdelijk Beleidskader waarbij e-mails in bulk worden veiliggesteld. De resultaten van het onderzoek richten zich op het geautomatiseerd ondersteunen van classificatie van e-mails bij toepassing van het hoofdspoor en van het tijdelijk Beleidskader.

De wijze waarop e-mails worden beheerd raakt zowel de dagelijkse uitvoering van werkzaamheden als de bredere opgave van transparantie en verantwoording. Tegen deze achtergrond start de pilot automatische e-mailclassificatie.

In deze pilot is onderzocht of geautomatiseerde ondersteuning van e-mailclassificatie kan bijdragen aan het verbeteren van e-mailbeheer. Daarbij is gewerkt met het categoriseren van e-mails in vijf categorieën; functioneel, niet-functioneel, privé, personeelsvertrouwelijk en partijpolitiek e-mails[1]. Waarbij het onderscheid tussen zakelijke en niet-zakelijke e-mail centraal staat.

Om dat te realiseren is er door verschillende organisaties onderzoek gedaan naar het inzetten van automatisering om de ambtenaar te ondersteunen in het selecteren van de e-mails.[2] [3] De eerder uitgevoerde onderzoeken hebben de volgende conclusies opgeleverd:

- Zelflerende systemen kunnen bijdragen aan een betere informatiehuishouding (Mette, 2018).
- Automatisering kan in hoge mate de ambtenaar ondersteunen in het classificeren van e-mails (Mette, 2018).
- Een basismodel kan tot 20% datareductie op de te selecteren functionele e-mails realiseren (One-Fox, 2025).
- Een basismodel kan tot 80% van de privé e-mails uitfilteren zonder dataverlies (One-Fox, 2024).
- Het is mogelijk e-mails te koppelen aan een bewaartermijn met zelflerende systemen (Deeploy, 2025).

Hoewel de resultaten van de voorgaande onderzoeken veelbelovend zijn is er nog voldoende te verkennen om een medewerker volledig (geautomatiseerd) te ondersteunen in het selecteren van functionele e-mails en uitfilteren van niet-functioneel, privé, personeelsvertrouwelijk en partijpolitiek e-mails. Door stap voor stap de inzet van de automatisering te toetsen ontstaat er voortschrijdend inzicht dat kan leiden tot een overheidsbrede geautomatiseerde oplossing om medewerkers te ondersteunen.

Zodoende is naar aanleiding van de vorige onderzoeken het verzoek[4] gekomen een pilot op te starten dat inzichten geeft:

- In welke mate geautomatiseerde e-mailclassificering mogelijk is op basis van overheidsdata;
- Welke waarde het toevoegt voor individuele medewerkers (privacybescherming) en voor de organisatie;
- Welke waarde het heeft op de werkdruk van medewerkers.

---

1 Handleiding "Welke e-mail kan weg?"

2 Rapport "Machine Learning en Automatische Classificatie"

3 Onderzoek Ministerie van Algemene Zaken "Automatische e-mailarchivering: Doorontwikkeling van het Test Purposes Model".

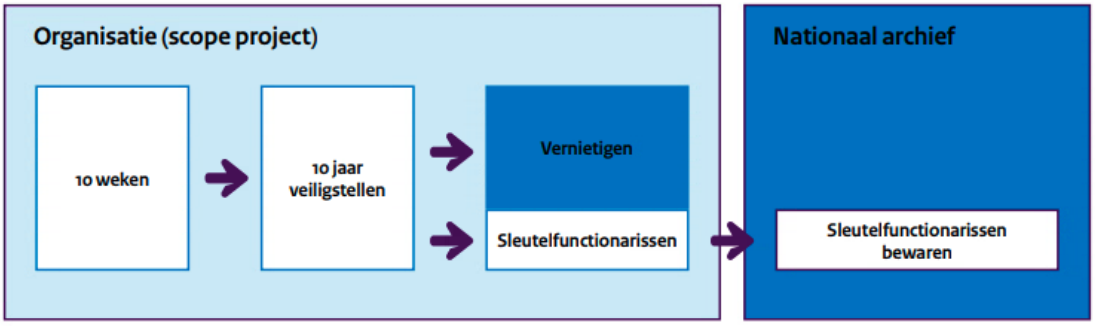
4 Vanuit CIO-Rijk zijn aanvullende inzichten gevraagd over het operationaliseren van geautomatiseerde e-mailclassificering binnen de Rijksoverheid, om daarmee de privacy van individuele medewerkers te verbeteren en voor de organisatie het aantal e-mails om veilig te stellen te beperken. Onderliggend is het PAR-advies "DPIA Beleidskader voor geautomatiseerde archivering van e-mails op basis van de sleutelfunctiemethodiek".

In het “Tijdelijk Beleidskader voor geautomatiseerde archivering van e-mails op basis van de sleutelfunctie-methodiek” (hierna: tijdelijk Beleidskader) en de handleiding “Welke e-mail kan weg?” (hierna: Handleiding)[5] wordt de werkwijze voor het bewaren en selecteren van e-mail beschreven. De afgelopen jaren is de hoeveelheid digitale overheidsinformatie sterk gegroeid, waarbij e-mail een van de belangrijkste communicatiemiddelen binnen de overheid is geworden.

Archiefwettelijk geldt dat e-mails moeten worden geselecteerd op basis van een vastgestelde selectielijst. Deze selectie kan plaatsvinden op verschillende manieren, bijvoorbeeld op basis van processen, zaken, hotspots of via de sleutelfunctiemethodiek. Zolang e-mails niet zijn beoordeeld volgens een geldende selectiemethode, mogen zij in principe niet worden vernietigd.

De in het tijdelijk Beleidskader beschreven werkwijze gaat uit van het veiligstellen van e-mails voor een periode van tien jaar, zie figuur 1. Na deze periode worden de e-mails, op basis van de uitgevoerde selectie, óf vernietigd óf overgebracht naar het Nationaal Archief. Deze werkwijze ondersteunt organisaties bij het voldoen aan de Archiefwet, de Wet open overheid (Woo) en de AVG.

Figuur 1: Werkwijze tijdelijk Beleidskader



Hoewel de pilot voortkomt uit het traject e-mailarchivering, is de werkwijze toepasbaar binnen meerdere selectiesporen. De pilottechniek is daarbij generiek bruikbaar en ondersteunt organisaties ongeacht de gekozen selectiemethode.

Met de pilot wordt onderzocht of een algoritme toegepast zou kunnen worden om e-mails te classificeren. Zo zouden veiliggestelde e-mails, die op grond van het tijdelijk Beleidskader tien jaar worden bewaard, geautomatiseerd kunnen worden gecategoriseerd en daarmee doelgericht kunnen worden opgeschoond van niet-functionele e-mails.

1.2 Casus

Deze onderzoeksvraag krijgt concrete invulling in een praktijkcasus bij het ministerie van Sociale Zaken en Werkgelegenheid (SZW). Vanuit de directie Bestuursondersteuning (BO) bestaat de wens om, ten opzichte van de huidige situatie, efficiënter en sneller te kunnen beschikken over (delen van) de mailboxen van voormalige bewindspersonen (BWP) en bestuursraadleden (BR-leden). Deze mailboxen worden regelmatig opgevraagd in het kader van informatieverzoeken, waaronder Woo-verzoeken.

Voorgesteld is om deze mailboxen te classificeren en op een verantwoorde manier structureel en duurzaam toegankelijk op te nemen in het Document Management Systeem (DMS). Daarmee wordt voorkomen dat voor elk verzoek opnieuw een arbeidsintensieve en repetitieve opvraging (van een deel) van een mailbox nodig is. Voorwaarde voor opname in het DMS is dat de mailboxen eerst worden opgeschoond van niet-functionele, privé-, personeelsvertrouwelijke en partijpolitieke e-mails. Ook is een voorwaarde dat de dossiers waarin de e-mails worden geplaatst zijn afgeschermd en alleen worden vrijgegeven voor een onderzoek op basis van een formeel informatieverzoek. Dit draagt bij aan een betere borging van de AVG en zorgt ervoor dat alleen relevante informatie wordt bewaard en doorzocht.

5 De handleiding is eind 2025 geactualiseerd en kan gebruikt worden door organisaties die volgens het hoofdspoor werken en aan het eind 2025 op te stellen toetsingskader voldoen.

1.3 Doel van de pilot

Het doel van dit onderzoek is om, aan de hand van een concrete use case, te verkennen of geautomatiseerde e-mailclassificatie kan bijdragen aan:

1. Het automatisch herkennen en toewijzen van categorieën (zoals functioneel, niet-functioneel, privé, personeelsvertrouwelijk en partijpolitiek) aan e-mails;
2. Een basis leggen voor toekomstig geïntegreerde oplossingen die rijks medewerkers ondersteunen bij het selecteren van e-mails;
3. Inzicht in de afweging tussen transparantie en nauwkeurigheid van de gebruikte classificatiemethodes;
4. Eerste inzichten verschaffen in wat ervoor nodig is om een oplossing, ondersteund door AI, in de mailbox van medewerkers te kunnen laten functioneren;
5. Het opleveren van twee batches met e-mails/mailboxen. Eén met zakelijke informatie die door de stakeholders gevalideerd kan worden en in het DMS opgenomen kan worden en één set met niet functionele, privé, partijpolitieke en personeelsvertrouwelijke mails die niet in het DMS opgenomen gaan worden.

1.4 Doelgroep

Dit onderzoek is uitgevoerd in opdracht van de CIO-Rijk en richt zich op meerdere doelgroepen binnen de overheid. Op strategisch niveau biedt het inzicht aan CIO-Rijk en bestuurders in de haalbaarheid, risico's en governance-aspecten van geautomatiseerde e-mailclassificatie.

De Privacy Adviseur Rijk (PAR) adviseert "onderzoek te doen naar een geautomatiseerde manier van archiveren en uitzonderen, zodat de verantwoordelijkheid niet langer primair bij de betreffende medewerker ligt" in relatie tot e-mail en het tijdelijk Beleidskader. De uitkomsten van de pilot worden mede gebruikt voor terugkoppeling aan de PAR, met het oog op een zorgvuldige beoordeling van rechtmatigheid, proportionaliteit en privacyaspecten.

Op organisatieniveau is het ministerie van SZW (in het bijzonder de directie Bestuursondersteuning) een belangrijke doelgroep. Zij zijn het eerste ministerie dat de mailboxen van voormalige bewindspersonen en bestuursraadleden doormiddel van AI opschonen en ontsluiten in het DMS.

Daarnaast richt het onderzoek zich op informatieprofessionals en medewerkers, die in hun dagelijkse werkzaamheden elke dag werken met e-mails en zelf verantwoordelijk zijn voor de archivering van hun e-mails.

2. De organisatie en uitvoering van de pilot

In dit hoofdstuk wordt beschreven hoe de pilot is georganiseerd, welke stappen in het proces centraal staan, welke expertises erbij betrokken zijn en welke waarborgen zijn genomen om tot een zorgvuldig en betrouwbaar resultaat te komen. De pilot is ingericht als een gecontroleerd traject waarin zowel de functionaliteit als de werkbaarheid van de nieuwe aanpak konden worden getest. De pilot is opgebouwd uit een voorbereidende fase, een uitvoerende fase en een afronding.

2.1 Voorbereidingsfase

Voorafgaand aan de uitvoering van de pilot wordt een omvangrijke voorbereidingsfase doorlopen. In deze fase vindt afstemming plaats met verschillende stakeholders, waaronder betrokken ministeries, uitvoeringsorganisaties en specialisten op het gebied van privacy, informatiebeveiliging en archivering.

Daarnaast wordt tijd geïnvesteerd in het organiseren van de benodigde randvoorwaarden. Dit omvat onder andere het doorlopen van inkooptrajecten, het verkrijgen van formele goedkeuringen en het voldoen aan geldende eisen op het gebied van informatiebeveiliging en privacy.

Deze voorbereidende werkzaamheden vragen een aanzienlijke inspanning en doorlooptijd, maar vormen een noodzakelijke basis om de pilot zorgvuldig, rechtmatig en uitvoerbaar te kunnen starten. Deze fase blijkt in de praktijk bepalend voor de haalbaarheid en doorlooptijd van de pilot. De belangrijkste activiteiten in deze fase zijn:

- Opstellen van het plan van aanpak.
- Afstemming met stakeholders en ophalen van eisen en randvoorwaarden.
- Inrichten van governance en verkrijgen van formele goedkeuringen.
- Doorlopen van inkoop- en aanbestedingstrajecten (indien van toepassing).
- Inrichten van de technische omgeving (initieel).
- Opstellen van handleidingen en instructies voor deelnemers.
- Selecteren en voorbereiden van deelnemers en datasets.

2.2 Uitvoeringsfase

Tijdens de uitvoering staat de begeleiding van labelaars centraal. Tijdens de uitvoering staat de begeleiding van deelnemers centraal. Zij krijgen instructies over wat er van hen wordt verwacht en kunnen gedurende het traject vragen stellen of knelpunten melden. De voortgang wordt continu gemonitord, zowel inhoudelijk als technisch, via regulier overleg en een logboek. Eventuele verstoringen worden geregistreerd, geanalyseerd en waar nodig direct opgelost. Daarnaast wordt tussentijds informatie verzameld om inzicht te krijgen in het functioneren van de automatische oplossing in de praktijk. De belangrijkste activiteiten in deze fase zijn:

- Begeleiden en ondersteunen van labelaars.
- Uitvoeren van handmatige classificatie van e-mails.
- Doorontwikkelen en finetunen van modellen.
- Monitoren van voortgang en kwaliteit classificatie.
- Registreren en oplossen van technische en procesmatige issues.
- Verzamelen van data voor analyse en evaluatie.

2.3 Analysefase

De pilot wordt afgesloten met een analyse van alle bevindingen. De resultaten worden gevalideerd met de betrokken stakeholders, waarna conclusies worden geformuleerd en de belangrijkste lessen worden vastgelegd. Deze eindfase resulteert in een evaluatie die niet alleen terugkijkt, maar ook vooruitwijst naar mogelijke vervolgstappen. De belangrijkste activiteiten in deze fase zijn:

- Analyseren van modelprestaties en classificatieresultaten.
- Valideren van resultaten met stakeholders.
- Evalueren van proces, techniek en organisatie.
- Formuleren van conclusies en geleerde lessen.
- Opstellen van rapportage en aanbevelingen.

2.4 Betrokken Expertises en Rollen

Bij de voorbereidende fase van de pilot zijn verschillende expertises betrokken, ieder met een duidelijk omschreven rol binnen het proces. Een belangrijk formeel startpunt wordt gevormd door de goedkeuring van de Secretaris-Generaal (SG) en plaatsvervangend Secretaris-Generaal (pSG). Met deze instemming wordt niet alleen het mandaat voor de pilot bevestigd, maar wordt ook officieel toestemming verkregen om de benodigde data op te vragen en te verwerken. Hiermee ontstaat de noodzakelijke juridische en bestuurlijke basis om het traject te kunnen starten.

De Chief Information Security Officer (CISO) speelt vervolgens een centrale rol bij het vaststellen van de informatie-beveiligingseisen en de voorwaarden waaronder de pilot mag worden uitgevoerd. Op basis van deze kaders adviseert het informatiebeveiligingsteam (IB) verder over de praktische mogelijkheden, de toepasbare maatregelen en de ruimte die er binnen de bestaande veiligheidsrichtlijnen bestond om de pilot verantwoord vorm te geven.

Daarnaast is de beveiligingsautoriteit (BVA) betrokken, omdat deze verantwoordelijk is voor het aanvragen van de benodigde data bij SSC-ICT.

De uiteindelijke uitvoering van de pilot zelf is mede gedragen door een groep inhoudelijke experts die handmatig e-mails classificeert. Ten tweede is er een projectleider betrokken vanuit het Ministerie van Sociale Zaken en Werkgelegenheid voor de nodige afstemming binnen de organisatie. Ten derde zijn er twee adviseurs vanuit de directie Bestuursondersteuning aangehaakt om te adviseren en toetsen of de oplossing aansluit bij de geformuleerde behoefte. Waarbij er als laatste ook een AI-adviseur betrokken is voor de ontwikkeling van het model en de technische inrichting en een coördinator voor het geheel van de pilot.

Gezamenlijk vormden deze expertises een zorgvuldig opgebouwd geheel van bestuurlijke, juridische, beveiligings- en inhoudelijke rollen, die samen de basis creëerden voor een pilot die zowel praktisch uitvoerbaar als veilig en verantwoord was.

3. De basis van Artificial Intelligence en Machine Learning

Binnen de pilot is de automatisering waar we over spreken Kunstmatige intelligentie (Artificial Intelligence, of kortweg AI). AI is de verzamelnaam voor technologieën die proberen menselijke denkprocessen na te bootsen. AI helpt computers om niet alleen data te verwerken, maar ook om te leren, redeneren en beslissingen te nemen. Een belangrijk onderdeel van AI is machine learning (ML), systemen die zichzelf verbeteren door te leren van voorbeelden. In dit hoofdstuk wordt uitgelegd wat AI en ML zijn, hoe ze werken, en hoe ze worden toegepast bij het automatisch herkennen en categoriseren van e-mails.

3.1 Wat is Artificial Intelligence (AI)?

AI verwijst naar computersystemen die taken kunnen uitvoeren die normaal gesproken menselijke intelligentie vereisen, zoals het begrijpen van taal, het herkennen van beelden of het nemen van beslissingen. AI is niet één technologie, maar een verzamelterm voor verschillende technieken die samenwerken om ‘slim gedrag’ mogelijk te maken.

In het dagelijks leven komen we AI tegen in allerlei vormen: denk aan digitale assistenten zoals Siri of Alexa en aanbevelingen op Netflix. Al deze toepassingen gebruiken AI om patronen te herkennen en op basis daarvan te reageren.

3.2 Wat is Machine Learning (ML)?

Machine learning is de techniek binnen AI die computers in staat stelt om zelfstandig patronen te ontdekken in data. In plaats van een reeks vaste regels te volgen, leert een ML-model van voorbeelden. Bij e-mailclassificatie wordt het model bijvoorbeeld getraind met e-mails die al een label hebben, zoals “functioneel”, “privé” of “partijpolitiek”. Door deze voorbeelden leert het model kenmerken herkennen, zoals specifieke woorden, afzenders of schrijfstijlen en kan het nieuwe e-mails automatisch in de juiste categorie plaatsen. Er bestaan verschillende vormen van machine learning, waaronder:

- Supervised learning – het model leert van gelabelde voorbeelden (zoals spam vs. niet-spam).
- Unsupervised learning – het model ontdekt zelf patronen, bijvoorbeeld door e-mails te groeperen op onderwerp of toon.
- Reinforcement learning – het model leert door te experimenteren en feedback te krijgen, vergelijkbaar met hoe mensen leren door trial-and-error.

3.3 Toepassen binnen de pilot

Binnen de pilot worden twee benaderingen gebruikt om te kunnen vergelijken op uitkomsten én op uitlegbaarheid van de classificatie. Beide benaderingen zijn gebaseerd op supervised learning. Ze leren in beide gevallen patronen te herkennen op basis van trainingsdata, maar verschillen sterk in complexiteit en interpreteerbaarheid.

- Lichtgewicht classificatiemodellen: De eerste benadering maakt gebruik van relatief eenvoudige, statistische machine learning-modellen: **Multinomial Naive Bayes en Multinomial Logistic Regression**. Deze modellen leren om patronen te herkennen tussen de aanwezige woorden en de mogelijke categorieën. Ze zijn goed interpreteerbaar: het is inzichtelijk welke woorden of kenmerken het meest bijdragen aan een classificatie, en de modellen zijn eenvoudig te analyseren. Dezelfde invoer leidt hierbij altijd tot dezelfde uitkomst.
- Taalmodel: In deze complexere benadering wordt een **Transformer-model ge-finetuned met Parameter Efficient Fine-Tuning (PEFT)**. Een transformer-model bevat de capaciteit om e-mails te classificeren op basis van zowel inhoud als context. Dit maakt het mogelijk om complexere patronen en nuances te identificeren. Echter zijn transformer-modellen moeilijker te interpreteren: de factoren die leiden tot een keuze zijn complex om te analyseren, en bovendien leidt dezelfde invoer bij veel modellen niet altijd tot dezelfde uitvoer.

Door deze twee aanpakken naast elkaar te gebruiken, kan worden vastgesteld wat het effect is op kwaliteit en consistentie van de classificatie én op de mate waarin de uitkomst goed te verklaren is. In Bijlage I: Technische uitleg is de beschrijving te vinden van de techniek.

4. Classificering

In dit hoofdstuk is beschreven wat wordt bedoeld met classificering en wordt uitgelegd welke classificaties worden gebruikt. Met classificering wordt in dit onderzoek bedoeld: het labelen van e-mails op basis van de inhoud van een e-mail. De inhoud van de e-mail (en bijlagen) bepaald welke classificatie de e-mail krijgt. E-mails worden binnen de pilot geclassificeerd als functioneel, niet-functioneel, privé, personeelsvertrouwelijk en partijpolitiek in lijn met het tijdelijk Beleidskader en de Handleiding.

4.1 Functioneel

Een functionele e-mail is een e-mail die onderdeel uitmaakt van het uitvoeren van een publieke taak of van het bestuurlijk handelen van een overheidsorganisatie.

De Archiefwet verplicht overheidsorganisaties om alle informatie die deel uitmaakt van hun bestuurlijke en uitvoerende werkzaamheden duurzaam te bewaren toegankelijk te maken en te houden. De Woo (Wet open overheid) bouwt hierop voort: burgers mogen inzage vragen in overheidsinformatie, inclusief functionele e-mails die onder de Archiefwet vallen.

Voorbeelden van e-mails die in deze categorie vallen:
- Een advies of afstemming tussen ambtenaren over beleid of uitvoering.
- Een goedkeuring of opdracht (“Ga akkoord met voorstel X”).
- Correspondentie die leidt tot, of inzicht geeft in, een besluit of actie.

4.2 Niet-functioneel

Een niet-functionele e-mail is een e-mail die geen onderdeel uitmaakt van het uitvoeren van een publieke taak of van het bestuurlijk handelen van een overheidsorganisatie. Een niet-functionele e-mail is dus alles wat niet direct bijdraagt aan die taak, maar wél via het werkaccount kan lopen.

De Archiefwet verplicht overheidsorganisaties om alle informatie die deel uitmaakt van hun bestuurlijke en uitvoerende werkzaamheden duurzaam te bewaren toegankelijk te maken en te houden. Informatie die geen deel uitmaakt van de bestuurlijke en uitvoerende werkzaamheden van overheidsorganisaties zoals niet-functionele e-mailberichten, valt buiten de reikwijdte van de Archiefwet.

Niet-functionele e-mails zijn berichten die wel binnen de werkomgeving ontstaan, maar geen bijdrage leveren aan de uitvoering van taken, besluitvorming of verantwoording. Dergelijke informatie wordt dan ook niet gewaardeerd aan de hand van de selectielijst. Niet-functionele e-mail valt ook buiten het bereik van de Woo.

Enkele voorbeelden van e-mails die in deze categorie vallen:
- Nieuwsbieven.
- Out-of-Office e-mails.
- E-mails van informatiesystemen met notificaties (“U wachtwoord verloopt”, “Er is een taak aan u toegekend vanuit systeem X”).
- Correspondentie over evenementen, borrels etc.

4.3 Privé

Een privé-e-mail is een bericht dat geen verband houdt met de uitvoering van de publieke taak of het bestuurlijk handelen van een overheidsorganisatie.

De Archiefwet verplicht overheidsorganisaties om alle informatie die deel uitmaakt van hun bestuurlijke en uitvoerende werkzaamheden duurzaam te bewaren. Privé e-mails vallen hier niet onder. Ook in deze categorie geldt: informatie die geen deel uitmaakt van de bestuurlijke en uitvoerende werkzaamheden van overheidsorganisaties zoals privé e-mailberichten, vallen buiten de reikwijdte van de Archiefwet. Privé e-mails vallen ook buiten het bereik van de Woo.

Privé e-mails en niet-functionele e-mails worden in het onderzoek onderscheiden als afzonderlijke categorieën. Waarbij de nuance zit in het volgende: Privé e-mails betreffen berichten met een persoonlijk karakter die geen relatie hebben met het werk, zoals correspondentie met familie of vrienden. Niet-functionele e-mails zijn berichten die wel binnen de werkomgeving ontstaan, maar geen bijdrage leveren aan de uitvoering van taken, besluitvorming of verantwoording. Het onderscheid is relevant, omdat beide categorieën geen directe archiefwaarde hebben, maar een andere aard en context kennen.

Voorbeelden van e-mails die in deze categorie vallen:
- E-mails naar en van eigen privé e-mailaccount.
- Foto's van uitjes.
- E-mails van privé afgesloten abonnementen.
- Uitnodigingen voor events die niet werk gerelateerd zijn.
- E-mails naar/van (leden van) organisaties waarvan de overheidsmedewerker niet uit hoofde van zijn functie lid is.

Uitzondering: Als een e-mail deels functioneel en deels privé is (bijv. een werkmail met de planning waarin je ook privéinformatie deelt), dan moet de informatie als geheel worden gearchiveerd (met het label functioneel). Bij openbaarmaking zal het privé deel in de mail moeten worden gelakt.

4.4 Personeelsvertrouwelijk

Een personeelsvertrouwelijke e-mail is een bericht dat (bijzondere) persoonsgegevens bevat over de arbeidsrelatie tussen werkgever en medewerker, en dat primair thuishoort in een afgeschermd personeelsdossier.

De AVG maakt onderscheid tussen gewone en bijzondere persoonsgegevens. Gewone persoonsgegevens zijn alle gegevens die direct of indirect naar een persoon te herleiden zijn, zoals naam, adres of telefoonnummer. Bijzondere persoonsgegevens zijn privacygevoeliger van aard, denk aan gegevens over gezondheid, religie of etnische afkomst, en vallen onder een verzwaard beschermingsregime. Een volledig overzicht van beide categorieën is opgenomen in Bijlage II: Gewone en bijzondere persoonsgegevens.

De AVG verplicht organisaties tot een zorgvuldige omgang met (bijzondere) persoonsgegevens. De Archiefwet verplicht overheidsorganisaties om alle informatie die deel uitmaakt van hun bestuurlijke en uitvoerende werkzaamheden duurzaam te bewaren toegankelijk te maken en te houden. Een personeelsvertrouwelijke e-mail kan functioneel of niet-functioneel zijn, afhankelijk van de context waarin deze is verstuurd of ontvangen.

Voorbeelden van e-mails die in deze categorie vallen:
- E-mails met CV's.
- E-mails met ziektemeldingen.
- E-mails over disciplinaire maatregelen of beoordelingen.

Uitzondering: Een functionele e-mail met (bijzondere) persoonsgegevens die al elders binnen de organisatie is vastgelegd, bijvoorbeeld in een personeelsdossier of HRM-systeem, dient te worden uitgezonderd van veiligstelling in het e-mailarchief. Dubbele opslag is niet noodzakelijk en vergroot onnodig de privacyrisico's.

4.5 Partijpolitiek

Partijpolitieke e-mails zijn communicatie met partijgenoten over onderwerpen die de partij aangaan (zoals over interne partij-aangelegenheden en inhoudelijke partijpolitieke standpunten).

De Archiefwet verplicht overheidsorganisaties om alle informatie die deel uitmaakt van hun bestuurlijke en uitvoerende werkzaamheden duurzaam te bewaren toegankelijk te maken en te houden. Partijpolitieke e-mails vallen hier niet onder.

Partijpolitieke informatie valt daar niet onder, want: het is geen overheidsinformatie, het valt niet onder de verantwoordelijkheid van het bestuursorgaan, maar van de politieke partij. Partijpolitieke e-mails vallen ook buiten het bereik van de Woo.

Voorbeelden van e-mails die in deze categorie vallen:
- E-mails ter voorbereiding van de ministerraad vanuit het perspectief van de politieke partij.
- Partijgebonden activiteiten (flyeren, campagne etc.).
- Benoeming van rollen van politieke partijen.

Uitzondering: Als een partijpolitieke mail óók elementen bevat van bestuurlijk handelen (bijvoorbeeld een fractievoorzitter die beleidsinhoudelijk afstemt met een wethouder over een raadsvoorstel), dan is de mail wél functioneel en moet deze worden gearchiveerd.

Ter ondersteuning van de pilot is een handleiding opgesteld waarin de e-mailcategorieën worden toegelicht, inclusief uitleg en voorbeelden. Deze handleiding is op aanvraag beschikbaar.

5. De data

In dit hoofdstuk wordt toegelicht welke soorten data wordt gebruikt binnen de pilot en op welke wijze deze data wordt verwerkt. Daarnaast wordt beschreven welke stappen zijn doorlopen om de data te classificeren, van de voorbereiding en selectie van de gegevens tot en met de kwaliteitsborging. Hiermee wordt inzicht gegeven in het classificatieproces en de keuzes die daarbij worden gemaakt.

5.1 E-mailboxen en classificatie

De data die binnen de pilot wordt gebruikt, bestaat uit e-mails uit e-mailboxen van voormalige bewindspersonen en bestuursraadleden. De mailboxen zijn aangeleverd als PST-bestanden (Outlook-archieven). Er zijn 20 e-mailboxen aangeleverd.

5.2 Handmatig classificeren

Als onderdeel van de pilot is een deel van deze totale e-mailcollectie handmatig geclassificeerd. In de praktijk is een goed samengestelde ground truth-set[6] van enkele duizenden tot tienduizenden e-mails nodig voor het trainen van een model. Die set vormt de referentie ('waarheid') waarmee het model wordt getraind en gevalideerd. De precieze omvang van de handmatige set hangt af van het aantal classificaties, de gewenste nauwkeurigheid en hoe efficiënt het classificeren wordt aangepakt.

Eerder onderzoek laat zien dat het handmatig classificeren van e-mails relatief kennis- en tijdsintensief kan zijn: gemiddeld worden ongeveer 20 e-mails per uur geclassificeerd[7]. Dit betekent dat de inzet van medewerkers die zo'n beoordeling kunnen uitvoeren expliciet moet worden ingepland. In de pilot is er daarom voor gekozen om stapsgewijs handmatig e-mails te classificeren. Voor de eerste kalibratie van het model was minimaal een set van 5000 e-mails nodig, waarna wordt toegewerkt naar 13.709 e-mails zodat er een robuuste basis ligt voor training en evaluatie.

5.3 Voorbereiding data

Om een model te ontwikkelen is het van belang dat de data van goede kwaliteit is. Daarom worden er meerdere stappen doorlopen om de dataset(s) te prepareren en geschikt te maken voor classificatie en modelontwikkeling. Daarbij is niet alleen inhoudelijk (relevantie) gekeken, maar ook technisch (inlaadbaarheid en integriteit van PST's).

5.3.1 Stap 1: Ontvangst, inladen en technische validatie (PST)

Na ontvangst van de PST-bestanden wordt eerst gecontroleerd of de bestanden correct konden worden ingeladen en of de data niet corrupt[8] is. Hierbij is vastgesteld dat 2 PST-bestanden van de 20 corrupt waren en dat 3 PST-bestanden niet via de gebruikte code ingeladen konden worden. Deze bestanden zijn buiten de geautomatiseerde verwerkingsstappen gehouden voor de eerste datasets die uitgegeven zijn en later meegenomen in de vervolg datasets. In Bijlage III: Verdeling dataset is meer informatie te vinden over de dataset.

6 Verwijst naar echte, direct waargenomen gegevens die worden gebruikt om AI-modellen te trainen en valideren, en die dienen als maatstaf voor de nauwkeurigheid bij taken als voorspellende analyses.

7 Onderzoek Ministerie van Algemene Zaken "Automatische e-mailarchivering: Doorontwikkeling van het Test Purposes Model". Hierin worden e-mails geclassificeerd op basis van categorieën en gekoppeld aan een categorie in de selectielijst.

8 Corrupt betekent hier dat mailboxen moesten worden teruggezet naar de oorspronkelijke mappenstructuur; PST-bestanden kunnen beschadigd raken, vooral bij grote bestandsgroottes.

5.3.2 Stap 2: Onderzoeken hoeveelheden e-mails per mailbox en samenvoegen

Vervolgens wordt geanalyseerd wat de verdeling is van de hoeveelheid e-mails over de mailboxen, waarbij als grootste mailbox een mailbox aantroffen is met 123.325 e-mails, maar ook een mailbox met 59 mails. Het gemiddelde over alle mailboxen is 41.727 e-mails. Alle e-mails worden geconsolideerd tot één totale e-mailcollectie. Hierdoor kunnen vervolgstappen zoals set-opbouw en filtering consistent worden uitgevoerd over alle mailboxen heen.

5.3.3 Stap 3: Inventarisatie itemtypen en uitsluiten niet-email items

Omdat PST-bestanden naast e-mails ook andere Outlook-items kunnen bevatten (zoals agenda-items, taken, contactgegevens of notities), wordt geïnventariseerd welke itemtypen aanwezig waren. Niet-email items worden uitgesloten, zodat de verzameling uitsluitend e-mailcommunicatie bevatte. In deze stap zijn 493.499 niet-email/ niet-communicatie items aangetroffen (circa 37%) waarna een netto e-mailcollectie resteert van 844.433 e-mails.

Het doel van deze stap is om de hoeveelheid te classificeren items te beperken door irrelevante of ruis-gevende berichten te verwijderen. Door filtering blijft alleen relevante inhoud over voor de automatische oplossing, wat ook de efficiëntie verhoogt en de kwaliteit van de classificatie verbetert. Aan het eind van deze stap is er een opgeschoonde dataset van relevante item-types, inclusief een logbestand van de toegepaste filters.

5.3.4 Stap 4: Evenredige sets maken

Op basis van de netto e-mailcollectie zijn evenredige en representatieve sets gemaakt, zodat de handmatige classificatie goed gepland kan worden en de werkdruk gespreid. Aan het eind van deze stap wordt de e-mailverzameling verdeeld in genummerde en geregistreerde sets met elk ongeveer 2000 e-mails. In de pilot wordt ervoor gekozen om eerst een dataset te creëren en daarna een gedetailleerde filtering toe te passen. De reden hiervoor is dat het binnen de beschikbare tijd niet haalbaar was om alle filtering al volledig in de code te automatiseren en daarna opnieuw de sets op te bouwen.

5.3.5 Stap 5: Filtering binnen Outlook

Na het maken van de datasets is de filtering grotendeels handmatig uitgevoerd nadat de e-mails in Outlook zijn ingeladen voor handmatige classificatie. Het doel van deze stap was om de hoeveelheid te classificeren e-mails te beperken door irrelevante of ruisgevende berichten te verwijderen. In de pilot betrof dit voornamelijk e-mails die als niet-functioneel zijn aangemerkt (dit is één van de classificatiecategorieën).

Een bijkomstigheid van het filteren nadat de datasets al waren gemaakt, is dat dit meer controle en inzicht geeft in welke filters effectief zijn en welke niet. De toegepaste filters worden vastgelegd in Bijlage IV: Filteren van e-mails. Wanneer binnen een vervolgtraject meer tijd beschikbaar is, kan deze filtering (deels) worden geautomatiseerd op basis van de inzichten die er nu zijn. Hierdoor kan de doorlooptijd afnemen en de handmatige inspanning gerichter worden ingezet.

5.3.6 Stap 6: Handmatig classificeren

Vervolgens zijn de overgebleven e-mails handmatig geclassificeerd in Outlook, door elke mail te verplaatsen naar het mapje voor de betreffende categorie; functioneel, niet-functioneel, privé, personeelsvertrouwelijk en partijpolitiek. Zodat per e-mail (en waar relevant inclusief bijlagen) een classificatielabel wordt toegekend. Aan het eind van deze stap zijn 13.709 e-mails geclassificeerd.

Ter voorbereiding op de handmatige classificatie zijn daarnaast verschillende scenario's onderzocht om het classificatieproces te versnellen en te vergemakkelijken, rekening houdend met informatiebeveiliging- (IB) en securitykaders, beschikbare tijd van experts en de doorlooptijd van de pilot. Deze scenario's zijn in hoofdlijnen beschreven in Bijlage V: Stappenplan en waarborgen kunnen de genomen stappen teruggevonden worden.

5.3.7 Stap 7: Export van geclassificeerde e-mails

Na afronding van de classificatie worden de geclassificeerde e-mails geëxporteerd, waarbij zowel de e-mailinhoud als de toegekende classificatie wordt vastgelegd in een verwerkbaar formaat (PST) voor analyse en modelontwikkeling. De export bevat per e-mail onder andere de datum, afzender, ontvanger(s), onderwerp, body, label, etc.

5.3.8 Stap 8: Expliciete scheiding in train-, validatie- en testset

Tot slot wordt de gelabelde dataset opgesplitst in afzonderlijke deelsets: een trainingsset een validatieset.[9]

- *Trainingsset:* De trainingsset wordt gebruikt om het model te leren patronen te herkennen in de data. Tijdens deze fase worden de modelparameters geoptimaliseerd om de fout tussen de voorspellingen en de werkelijke waarden te minimaliseren.
- *Validatieset:* De validatieset simuleert nieuwe, ongeziene data en biedt een onafhankelijke maatstaf voor de uiteindelijke prestaties van het model. Zo kan worden vastgesteld hoe goed het model generaliseert naar nieuwe situaties, die nog niet eerder gezien zijn tijdens het trainen.

Deze scheiding is nodig om betrouwbare en generaliseerbare machine learning-modellen te ontwikkelen, overfitting[10] te voorkomen en prestaties eerlijk te evalueren. Tijdens de pilot is een verdeling toegepast van 80% trainingsdata en 20% testdata.

Om tot betrouwbare resultaten te komen is verder gebruik gemaakt van cross-validation. Hierbij wordt de data in vijf gelijke delen gesplitst, waarbij het model vijf keer wordt getraind en gevalideerd, elke keer met een ander deel als testset en de overige vier als trainingsset. De gerapporteerde prestaties zijn het gemiddelde over deze vijf iteraties.

5.3.9 Stap 9: Kwaliteitsborging van classificatie en steekproef

Tijdens het classificatieproces wordt onderzocht op welke wijze de kwaliteit en consistentie van de handmatige classificaties het beste kan worden geborgd. Twee opties zijn daarbij verkend. Zie Bijlage VI: Kwaliteitsborging voor de andere onderzochte opties. Op basis hiervan is gekozen voor het opnemen van 5% overlap in de te classificeren data, omdat deze aanpak het beste aansluit bij het doel om de kwaliteit en betrouwbaarheid van de uiteindelijke dataset zo hoog mogelijk te maken. Concreet betekent dit dat er 13.709 e-mails zijn geclassificeerd. Van deze e-mails zijn er 12.947 e-mails één keer gelabeld. 753 e-mails zijn twee keer gelabeld en 9 e-mails zijn drie keer gelabeld. Daarnaast hebben er nog correcties plaatsgevonden op de e-mail sets, in totaal zijn 670 e-mails gecorrigeerd.

9 Bij het trainen van machine learning-modellen wordt de validatieset soms ook nog gesplitst in een validatieset en een testset, om na alle optimalisaties een volledig onafhankelijke evaluatie van het model te kunnen doen. Gezien de beperkte hoeveelheid data is er in dit geval voor gekozen om alleen een validatieset te gebruiken.

10 Overfitting (over aanpassing) is een veelvoorkomend probleem in machine learning waarbij een model de trainingsdata te goed leert, inclusief ruis en toevallige patronen, waardoor het slecht presteert op nieuwe, onbekende data.

6. De modellen en uitlegbaarheid

In dit hoofdstuk worden de toegepaste modellen toegelicht en wordt uiteengezet welke elementen bepalend zijn voor hun uitlegbaarheid. We beginnen met het beoordelingskader, de criteria waarop de modellen worden geëvalueerd, waarna de opbouw, aannames en werking van de modellen zelf worden beschreven. De daadwerkelijke modelresultaten worden pas in hoofdstuk 7 gepresenteerd; in dit hoofdstuk richten we ons op het creëren van het benodigde begrip om die resultaten goed te kunnen duiden.

6.1 Het beoordelingskader

Voor een zorgvuldige inzet van de modellen is inzicht in een aantal kernaspecten essentieel. Deze aspecten vormen samen het kader waarbinnen de classificatie-oplossing moet functioneren. De definities zijn zowel technisch als organisatorisch van belang, omdat zij richting geven aan de beoordeling van modelprestaties en de randvoorwaarden voor implementatie.

6.1.1 Functionele juistheid

De classificatie moet met een hoge mate van zekerheid onderscheid kunnen maken tussen zakelijke en niet-zakelijke e-mails. Omdat e-mails zelden volledig zakelijk of volledig niet-zakelijk zijn, wordt gestreefd naar een balans tussen classificatiezekerheid en informatiebehoud.

Betrouwbaarheid kan hierbij worden gedefinieerd via de confidence score of betrouwbaarheidsdrempel. Dit zijn vooraf ingestelde minimumscores vaak variërend van 0 tot 1 die bepalen of een voorspelling betrouwbaar genoeg is om automatisch te verwerken, of dat menselijke beoordeling noodzakelijk is.

Het kiezen van de juiste drempelwaarde is een afweging:
- Een te hoge waarde kan leiden tot gemiste detecties (false negatives).
- Een te lage waarde verhoogt het aantal onjuiste classificaties (false positives).

Tabel 1: Drempelwaardes

Soort drempelwaarde	Waarde
Hoge drempelwaarden	>0,90
Lage drempelwaarden	>0,50

Meetaanpak binnen de pilot:
- Precision en recall op zakelijke e-mails.
- Confusion matrix (TP/FP/FN/TN).
- Percentage e-mails boven threshold.
- Aantal handmatige correcties tijdens pilot.

6.1.2 Consistentie

Het systeem moet vergelijkbare e-mails uniform classificeren, ongeacht afzender, formulering of context. Dit bevordert reproduceerbaarheid en voorkomt willekeur in de classificatie-uitkomsten. Consistentie is belangrijk omdat inconsistent gedrag het vertrouwen van gebruikers kan ondermijnen en leidt tot onvoorspelbare werkprocessen.

Meetaanpak binnen de pilot:
- Test op herhaalde classificatie van dezelfde e-mail (stabiliteit).
- Variatieanalyse tussen vergelijkbare berichten.
- Consistentie tussen menselijke en automatische classificatie.
- Percentage inconsistent gelabelde e-mails binnen clusters.

6.1.3 Transparantie

De onderliggende techniek mag deels functioneren als een 'black box', mits de gekozen strategie voor labeltoekenning helder wordt toegelicht. Dit omvat onder andere inzicht in de gehanteerde confidence thresholds, de rol van beslisregels en eventuele regel gebaseerde componenten, en de manier waarop deze elementen de betrouwbaarheid beïnvloeden. Het doel is om zo transparant mogelijk te zijn binnen de grenzen van technische haalbaarheid.

Meetaanpak binnen de pilot:
- Uitlegbaarheidsinformatie per classificatie beschikbaar gesteld.
- Thresholds en beslisregels gedocumenteerd.

6.2 Model 1: Logistic Regression

Logistic Regression is een relatief eenvoudig en veelgebruikt machine-learningmodel voor classificatie. In het geval van e-mailclassificatie leert het model welke woorden, patronen of tekstkenmerken vaak voorkomen in bepaalde categorieën, zoals functioneel, niet-functioneel, privé, partijpolitiek of personeelsvertrouwelijk. Op basis van deze kenmerken berekent het model de kans dat een e-mail tot een bepaalde categorie behoort en kiest vervolgens de categorie met de hoogste kans.

Het model werkt met een lineaire combinatie van features (bijvoorbeeld woorden of tekstrepresentaties zoals TF-IDF) en zet deze om naar een waarschijnlijkheidsscore tussen 0 en 1 met behulp van een logistieke functie. Hierdoor kan het model relatief efficiënt bepalen bij welke categorie een e-mail het beste past.

Een belangrijk kenmerk van Logistic Regression is dat het model goed uitlegbaar en transparant is. Elk kenmerk in de tekst krijgt een gewicht dat aangeeft hoe sterk het bijdraagt aan een bepaalde classificatie. Daardoor is het mogelijk om redelijk precies te analyseren waarom een e-mail bijvoorbeeld als partijpolitiek of privé wordt geclassificeerd. Dit maakt het model geschikt voor situaties waarin verantwoording of controle van beslissingen belangrijk is.

Tabel 2: Voordelen en nadelen Logistic Regression

Voordelen	Nadelen
Hoge mate van uitlegbaarheid en transparantie.	Beperkt vermogen om complexe taalstructuren of context te begrijpen.
De invloed van individuele woorden of features is zichtbaar via modelcoëfficiënten.	Relaties tussen woorden worden vaak niet volledig meegenomen.
Relatief eenvoudig model dat snel te trainen is.	Kan minder goed omgaan met subtiele verschillen tussen categorieën.
Werkt goed met kleinere datasets.	Prestaties kunnen lager zijn bij complexere NLP-taken[11].
Lage rekenkosten en makkelijk te implementeren.	

6.3 Model 2: Parameter-Efficient Fine-Tuning (PEFT)

PEFT is een methode om grote taalmodellen, zoals transformer-gebaseerde modellen, aan te passen aan een specifieke taak zoals e-mailclassificatie. Binnen de pilot hebben we gebruik gemaakt van het taalmodel *bert-base-dutch-cased* ook wel genaamd *BERTje*. In plaats van alle parameters van een groot model opnieuw te trainen, worden de meeste parameters vastgezet en wordt slechts een klein deel van het model aangepast. Hierdoor kan het model efficiënt worden afgestemd op de classificatie van e-mails zonder de volledige rekenkosten van een complete hertraining.

11 Natural Language Processing (NLP) is het vakgebied binnen Kunstmatige Intelligentie (AI) dat zich bezighoudt met het analyseren van menselijke taal en de communicatie tussen mens en computer.

Omdat deze modellen gebaseerd zijn op grote taalmodellen die getraind zijn op enorme hoeveelheden tekst, zijn ze vaak beter in staat om context, zinsstructuur en semantiek te begrijpen. Daardoor kunnen ze subtielere verschillen herkennen tussen categorieën zoals functioneel en niet-functioneel, of tussen privé en personeelsvertrouwelijk.

Tegelijkertijd zijn deze modellen aanzienlijk complexer. De interne beslissingen worden gemaakt via duizenden tot miljoenen parameters, waardoor het vaak lastig is om precies te verklaren waarom een bepaalde classificatie wordt gemaakt. De transparantie is daardoor lager dan bij eenvoudige modellen zoals Logistic Regression. Uitlegbaarheid kan wel deels worden verkregen via aanvullende technieken zoals attention-visualisatie of feature-attribution methoden, maar deze geven meestal slechts een indicatie van het besluitvormingsproces.

Tabel 3: Voordelen en nadelen BERTje

Voordelen	Nadelen
Sterker vermogen om context en betekenis van tekst te begrijpen	Minder transparant en moeilijker uitlegbaar
Vaak hogere classificatieprestaties bij complexe tekst	Complexere architectuur en moeilijker te analyseren
Kan subtiele verschillen tussen categorieën beter herkennen	Meer rekenkracht nodig dan eenvoudige modellen
Efficiënter dan volledige fine-tuning van grote taalmodellen	Beslissingen zijn moeilijker te controleren of te interpreteren

6.4 Modelvergelijking

De twee modellen vullen elkaar aan en vertegenwoordigen een bewuste afweging tussen uitlegbaarheid en classificatievermogen. Logistic Regression biedt maximale transparantie en is eenvoudig te controleren, maar mist de contextuele diepgang die nodig is voor subtiele categorieën. PEFT presteert doorgaans beter op functionele juistheid, maar gaat gepaard met een lagere transparantie en hogere rekenkosten.

Tabel 4: Vergelijking modellen

Criterium	Logistic Regression	PEFT (BERTje)
Functionele juistheid	Matig bij complexe tekst	Sterk
Consistentie	Stabiel en voorspelbaar	Over het algemeen stabiel
Transparantie	Hoog	Beperkt
Rekenkosten	Laag	Gemiddeld
Uitlegbaarheid	Hoog	Beperkt, via aanvullende technieken

De criteria die buiten de scope van de pilot vallen maar wel in het plan van aanpak zijn genoemd, zijn opgenomen in Bijlage VII als aandachtspunten voor een toekomstige productieomgeving. In hoofdstuk 7 worden de daadwerkelijke resultaten van beide modellen gepresenteerd en geïnterpreteerd aan de hand van bovenstaand kader.

7. Resultaten

Dit hoofdstuk presenteert de eindresultaten van de pilot. De bevindingen worden uitgewerkt voor de twee gebruikte modellen en bieden inzicht in de classificatie-prestaties per categorie. Hiermee vormt dit hoofdstuk de inhoudelijke basis voor de conclusies en aanbevelingen in hoofdstuk 8.

7.1 Eerste bevindingen

Het eerste trainings- en iteratieproces wordt uitgevoerd met drie modellen: twee varianten van Logistic Regression en één PEFT-taalmodel gebaseerd op BERTje. De twee Logistic Regression-varianten verschilden in de manier waarop de tekstrepresentatie was opgezet. Omdat de prestaties van de drie modellen dicht bij elkaar liggen, wordt gekozen om verder te gaan met één Logistic Regression-variant en het PEFT-taalmodel. De best presterende Logistic Regression-variant wordt geselecteerd op basis van de hoogste gewogen F1-score op de validatieset, waarbij de prestaties op de categorieën functioneel en niet-functioneel het zwaarst wegen vanwege hun aandeel in de dataset.

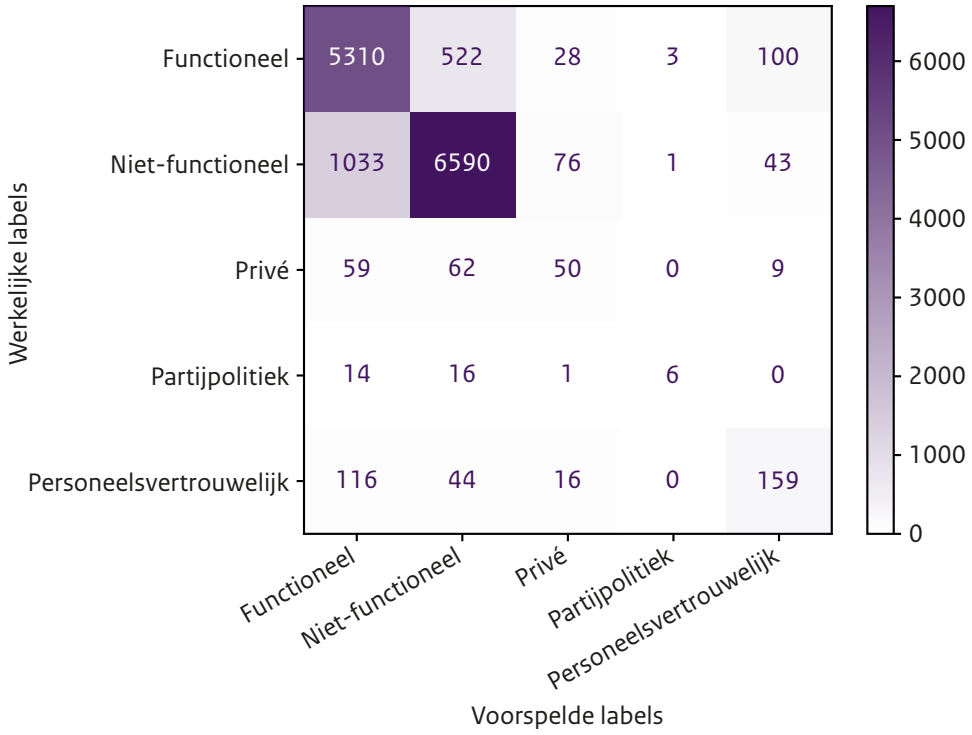
Voor model 1 (Logistic Regression) laten de resultaten zien dat een duidelijke scheiding tussen functionele en niet-functionele e-mails goed mogelijk is. Deze categorieën vormen het grootste deel van de dataset en worden door het model relatief betrouwbaar herkend. De prestaties worden echter begrensd door het feit dat kleinere categorieën zoals privé, partijpolitiek en personeelsvertrouwelijk aanzienlijk moeilijker te classificeren zijn. Dit hangt vooral samen met het beperkte aantal trainingsvoorbeelden voor deze categorieën en met inhoudelijke overlap tussen sommige categorieën.

Voor model 2 (PEFT-taalmodel gebaseerd op BERTje) blijkt dat dit model iets beter kan omgaan met variatie in taalgebruik en formuleringen. Dit is bijvoorbeeld zichtbaar in een hogere recall voor bepaalde kleinere categorieën, zoals privé-e-mails. Tegelijkertijd blijft ook bij dit model zichtbaar dat categorieën met weinig voorbeelden of met inhoudelijke overlap moeilijker te herkennen zijn. Over beide modellen geldt dat de kwaliteit sterk samenhangt met de eenduidigheid van definities, de beschikbaarheid van representatieve voorbeelden en de mate waarin categorieën elkaar inhoudelijk overlappen.

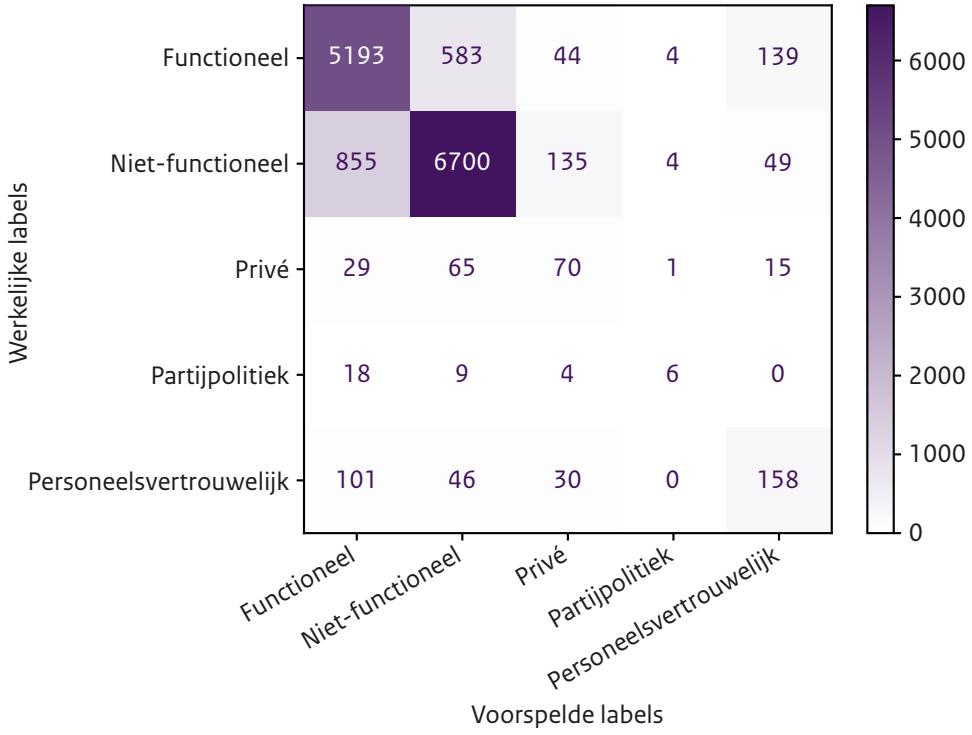
7.2 Confusion matrices

Een confusion matrix is een overzicht dat inzicht geeft in hoe een model zijn voorspellingen heeft gedaan ten opzichte van de werkelijkheid. Het laat zien hoeveel voorspellingen correct waren en waar het model fouten heeft gemaakt. Daarbij wordt onderscheid gemaakt tussen correct positieve voorspellingen, correct negatieve voorspellingen, fout-positieve voorspellingen (ten onrechte als positief aangemerkt) en fout-negatieve voorspellingen (ten onrechte gemist). Onderstaande confusion matrices laten zien hoe model 1 en model 2 presteren en maken inzichtelijk waar de verschillen in prestaties tussen beide modellen zich bevinden.

Figuur 2: Model 1 – Logistic Regression



Figuur 3: Model 2 – PEFT-taalmodel (BERTje)



Beide modellen presteren het beste bij de grootste categorieën in de dataset: functioneel en niet-functioneel. Deze categorieën bevatten veel trainingsvoorbeelden en hebben relatief herkenbare patronen in taalgebruik, wat leidt tot hoge precision- en recallwaarden. De matrices laten tegelijkertijd zien dat een deel van de e-mails tussen deze twee categorieën wordt verwisseld, wat erop wijst dat het onderscheid in de praktijk niet altijd scherp is.

Voor de kleinere categorieën: privé, partijpolitiek en personeelsvertrouwelijk, zijn de prestaties duidelijk lager. Dit komt vooral doordat deze categorieën relatief weinig voorbeelden bevatten, waardoor de modellen minder goed kunnen leren welke kenmerken specifiek bij deze categorieën horen. De categorie partijpolitiek vormt hierbij een bijzonder geval: met slechts 37 voorbeelden in de dataset hebben beide modellen grote moeite om deze categorie betrouwbaar te herkennen[12].

Het PEFT-model laat op enkele punten een kleine verbetering zien ten opzichte van Logistic Regression, met name bij categorieën waarbij de formulering sterk kan variëren. De verschillen blijven echter beperkt, wat erop wijst dat de prestaties momenteel vooral worden begrensd door de dataset en de labeldefinities.

7.2.1 Wat zeggen de cijfers?

De confusion matrix vormt de basis voor kerngetallen zoals accuracy, precision, recall en de F1-score. Deze cijfers maken gericht inzichtelijk waar het model goed presteert en waar verbetering mogelijk is.

Bij de beoordeling van het AI-model is gebruikgemaakt van vier gangbare kwaliteitsmaten: accuracy, precision, recall en de F1-score.

- **Accuracy** geeft het totale percentage correcte voorspellingen weer en laat zien hoe vaak het model over het geheel genomen de juiste uitkomst geeft. Deze maat is vooral informatief wanneer de verschillende categorieën in de data ongeveer gelijk verdeeld zijn.
- **Precision** geeft aan in hoeverre positieve voorspellingen daadwerkelijk correct zijn. Met andere woorden: als het model iets als “positief” aanmerkt, hoe betrouwbaar is die beslissing?
- **Recall** laat zien in welke mate het model erin slaagt om alle daadwerkelijk positieve gevallen te identificeren. Dit is met name van belang wanneer het missen van relevante gevallen ongewenste of risicovolle gevolgen heeft.
- De **F1-score** combineert precision en recall in één maat en geeft daarmee een gebalanceerd beeld van de prestaties, vooral wanneer beide typen fouten (onterecht positief en onterecht negatief) impact hebben.

Tabel 5: Model 1 – Logistic Regression, verschillende kwaliteitsmaten

Categorie	Precision	Recall	F1-score	Support
Functioneel	0.81	0.89	0.85	5963
Niet-functioneel	0.91	0.85	0.88	7743
Privé	0.29	0.28	0.28	180
Personeelsvertrouwelijk	0.51	0.47	0.49	335
Partijpolitiek	0.60	0.16	0.26	37

Tabel 6: Model 1 – Logistic Regression, de scores

Metric	Score
Accuracy	0.85
Macro average F1	0.55
Weighted average F1	0.85

12 Het is goed om te vermelden dat de mailboxen van Politieke Assistentes geen onderdeel van de geleverde mails waren, e-mails van deze functionaris konden wel in de correspondentie zitten.

Tabel 7: Model 2 – PEFT-taalmodel (BERTje), de verschillende kwaliteitsmaten

Categorie	Precision	Recall	F1-score	Support
Functioneel	0.84	0.87	0.85	5963
Niet-functioneel	0.91	0.87	0.88	7743
Privé	0.25	0.39	0.30	180
Personeelsvertrouwelijk	0.44	0.47	0.45	335
Partijpolitiek	0.40	0.16	0.23	37

Tabel 8: Model 2 – PEFT-taalmodel (BERTje), de scores

Metric	Score
Accuracy	0.85
Macro average F1	0.55
Weighted average F1	0.85

7.2.2 Wat goed werkt

Beide modellen behalen een accuracy van 0,85, wat betekent dat ongeveer 85% van de e-mails correct wordt geclassificeerd. Voor de twee grootste categorieën zijn de resultaten sterk: functioneel scoort een F1 van 0,85, niet-functioneel een F1 van 0,88. Dit betekent dat de kern van het classificatievraagstuk: het onderscheid tussen zakelijke en niet-zakelijke e-mail betrouwbaar werkt. Dat is een directe bevestiging van de haalbaarheid van geautomatiseerde e-mailclassificatie voor het meest voorkomende gebruik.

7.2.3 Waar verbetering nodig is

De accuracy van 0,85 moet voorzichtig worden geïnterpreteerd, omdat de dataset sterk wordt gedomineerd door de categorieën functioneel en niet-functioneel. Wanneer een model deze grote categorieën goed herkent, kan de totale accuracy relatief hoog uitvallen, ook als kleinere categorieën minder goed worden geclassificeerd. Dit wordt zichtbaar in de macro average F1-score van 0,55, die aangeeft dat de prestaties gemiddeld over alle categorieën aanzienlijk lager liggen.

Voor de kleinere categorieën liggen de scores aanmerkelijk lager. Privé haalt een F1-score van 0,28 bij model 1 en 0,30 bij model 2. Personeelsvertrouwelijk scoort iets beter met F1-scores rond 0,45 tot 0,49, maar ook hier bestaat nog aanzienlijke verwarring met andere categorieën. Partijpolitiek vormt de grootste uitdaging: de recall ligt bij beide modellen slechts op 0,16, wat betekent dat het merendeel van deze e-mails niet correct wordt herkend. De voornaamste verklaring is het zeer kleine aantal trainingsvoorbeelden.

Bij de interpretatie van de ROC- en PR-curves worden twee aanvullende kwaliteitsmaten gebruikt: de AUC en de Average Precision (AP).

- De **Area Under the Curve** (AUC) geeft aan hoe goed het model over alle mogelijke drempelwaarden heen onderscheid maakt tussen een categorie en de rest. De waarde loopt van 0 tot 1: een AUC van 1,0 duidt op perfecte classificatie, een waarde van 0,5 staat gelijk aan willekeurig gokken. In de praktijk geldt een AUC boven 0,80 als goed en boven 0,90 als sterk.
- De **Average Precision** (AP) is het equivalent van de AUC, maar berekend onder de PR-curve. Bij sterk ongebalanceerde datasets waarbij sommige categorieën veel minder voorbeelden bevatten dan andere geeft de AP een eerlijker beeld van de werkelijke bruikbaarheid van het model. Een hoge AUC in combinatie met een lage AP wijst erop dat het model categorieën wel kan onderscheiden, maar dat dit in de praktijk minder betrouwbaar is dan de AUC suggereert.

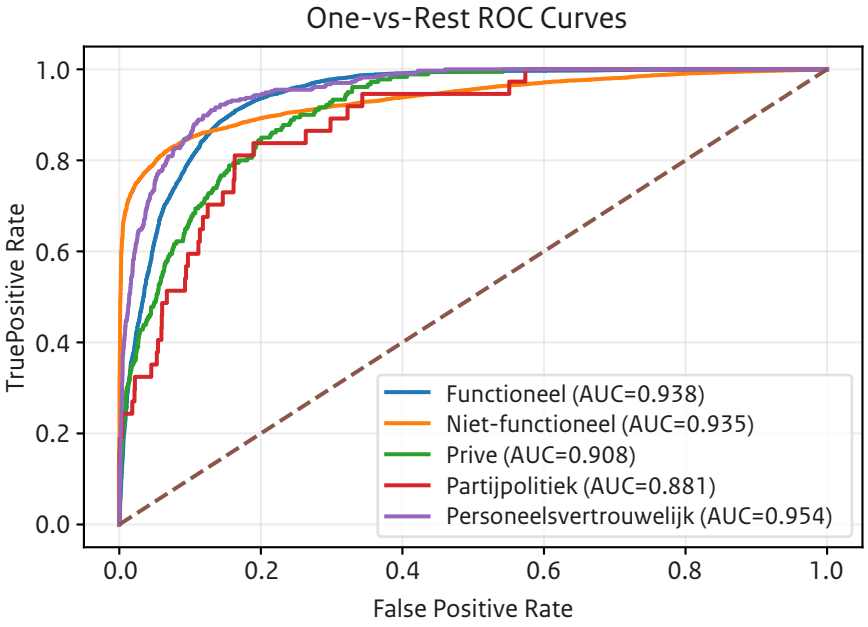
7.3 ROC-curves en PR-curves

Naast de confusion matrix en de kerngetallen bieden de ROC-curve en de PR-curve aanvullend inzicht in hoe de modellen presteren over verschillende drempelwaarden heen. Waar de confusion matrix een momentopname geeft bij één vaste drempelwaarde, laten deze curves zien hoe de prestaties zich ontwikkelen wanneer die drempelwaarde wordt gevarieerd.

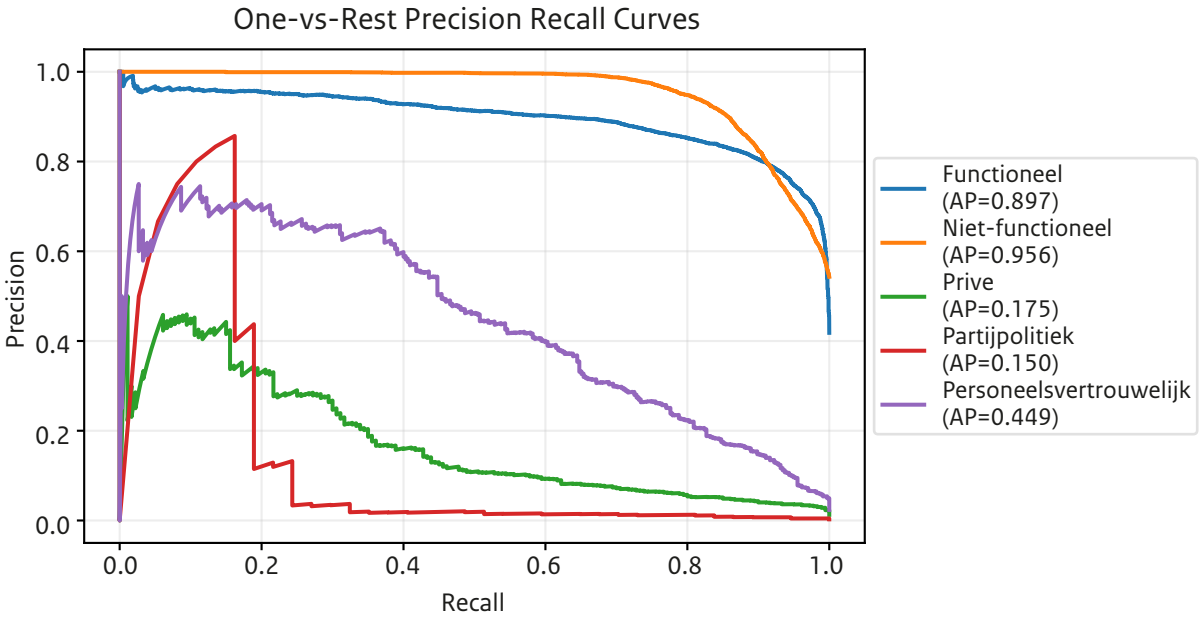
De ROC-curve (Receiver Operating Characteristic) zet de verhouding af tussen het percentage correct herkende positieve gevallen (true positive rate) en het percentage ten onrechte als positief aangemerkte gevallen (false positive rate). Een model dat perfect onderscheid maakt, bereikt een AUC-waarde (Area Under the Curve) van 1,0. Een waarde van 0,5 staat gelijk aan willekeurig gokken.

De PR-curve (Precision-Recall curve) zet precision af tegen recall en geeft bij ongebalanceerde datasets een betrouwbaarder beeld dan de ROC-curve. Bij datasets waarbij bepaalde categorieën sterk ondervertegenwoordigd zijn, zoals in deze pilot het geval is voor privé, partijpolitiek en personeelsvertrouwelijk, kan een hoge AUC-score een rooskleuriger beeld geven dan gerechtvaardigd is. De bijbehorende maatstaf, de Average Precision (AP), is daarom in deze context een eerlijkere graadmeter voor modelprestaties.

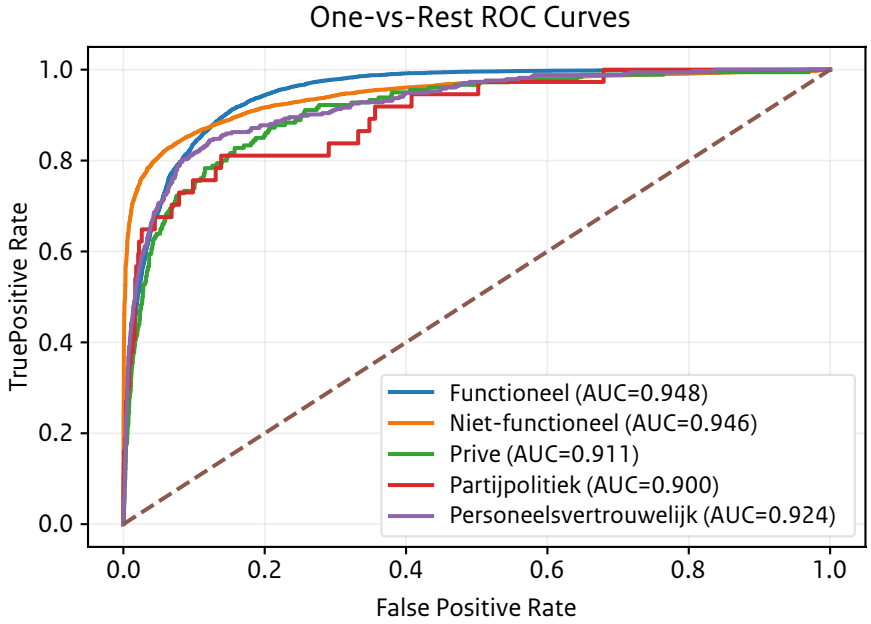
Figuur 4: Model 1 – Logistic Regression, ROC-curve



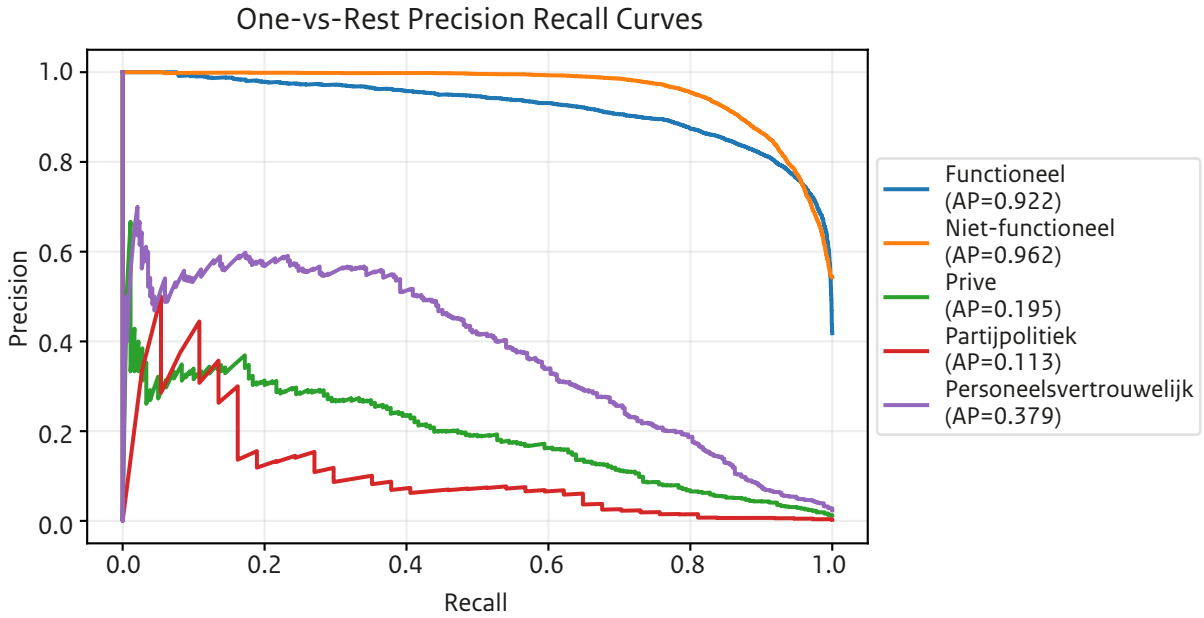
Figuur 5: Model 1 – Logistic Regression, PR-curve



Figuur 6: Model 2 – PEFT-taalmodel (BERTje), ROC-curve



Figuur 7: Model 2 – PEFT-taalmodel (BERTje), PR-curve



7.3.1 ROC-curves: beide modellen onderscheiden categorieën beter dan de F1-scores suggereren

De ROC-curves laten voor beide modellen opvallend hoge AUC-waarden zien, ook voor de kleinere categorieën. Bij Logistic Regression liggen de AUC-scores tussen 0,881 (partijpolitiek) en 0,954 (personeelsvertrouwelijk). Bij het PEFT-model liggen deze tussen 0,900 (partijpolitiek) en 0,948 (functioneel). Dit betekent dat beide modellen in principe in staat zijn om elke categorie te onderscheiden van de overige categorieën, ook wanneer de drempelwaarde wordt gevarieerd. Het PEFT-model behaalt op vrijwel alle categorieën een iets hogere AUC dan Logistic Regression, met uitzondering van personeelsvertrouwelijk, waar Logistic Regression (0,954) hoger scoort dan PEFT (0,924).

7.3.2 PR-curves: het eerlijkere beeld bij een ongebalanceerde dataset

De PR-curves vertellen een genuanceerder verhaal. Voor functioneel en niet-functioneel presteren beide modellen sterk, met AP-scores van respectievelijk 0,897 en 0,956 bij Logistic Regression, en 0,922 en 0,962 bij PEFT. Voor de kleinere categorieën dalen de scores aanzienlijk. Bij Logistic Regression haalt privé een AP van 0,175, partijpolitiek slechts 0,150 en personeelsvertrouwelijk 0,449. Bij PEFT liggen deze waarden op respectievelijk 0,195, 0,113 en 0,379.

Opvallend is dat Logistic Regression ondanks zijn eenvoudigere architectuur beter presteert op de twee kleinste categorieën: partijpolitiek (AP 0,150 versus 0,113) en personeelsvertrouwelijk (AP 0,449 versus 0,379). Dit wijst erop dat Logistic Regression bij schaarse trainingsdata een gunstiger balans weet te vinden tussen precision en recall.

7.3.3 Het contrast tussen ROC en PR als leidraad voor modelkeuze

Het verschil tussen de hoge AUC-scores en de lage AP-scores is kenmerkend voor situaties met sterk ongebalanceerde datasets. De ROC-curve suggereert dat beide modellen de categorieën in beginsel goed kunnen onderscheiden, maar de PR-curve laat zien dat dit onderscheid in de praktijk, bij het daadwerkelijk toewijzen van een label, aanzienlijk moeilijker is voor de ondervertegenwoordigde categorieën. Voor de beoordeling van modelprestaties in deze pilot is de PR-curve daarom de meest relevante maatstaf. De AP-scores bevestigen de conclusie die ook uit de F1-scores naar voren komt: de grootste winst is te behalen door uitbreiding en verbetering van de trainingsdata voor de kleinere categorieën, niet door verdere optimalisatie van het model zelf.

7.4 Besliszones en handelingsperspectief

De technische resultaten uit de ROC- en PR-curves krijgen pas bestuurlijke betekenis wanneer ze worden vertaald naar concrete keuzes over hoe het systeem wordt ingezet. De centrale vraag is niet alleen hoe goed het model presteert, maar ook: waar kan op gestuurd worden, en welke afwegingen horen bij welke categorie? Op basis van de confidence-scores en de drempelanalyse uit paragraaf 7.3 worden drie besliszones onderscheiden die als handelingskader dienen.

Zone A – Hoge zekerheid (confidence ≥ 0,90)
In deze zone classificeert het model met hoge mate van zekerheid. Uit de confidence-distributietabel blijkt dat bij een drempel van 90% recall op functionele e-mails nog slechts 14,5% van de niet-functionele e-mails wordt meegenomen, en 27,8% van de privé-e-mails. Voor functioneel en niet-functioneel, de twee grootste categorieën met de sterkste AP-scores, is deze zone geschikt voor geautomatiseerde verwerking. Het risico op foutieve classificatie is hier beheersbaar, en automatisering levert efficiëntiewinst op zonder onevenredig hoge foutmarge.

Voor personeelsvertrouwelijk geldt een uitzondering: ook al valt een e-mail in Zone A op basis van de functionele confidence-score, dan nog is aanvullende controle aan te raden vanwege het hogere risicoprofiel van deze categorie. Een gemiste personeelsvertrouwelijke e-mail heeft immers grotere consequenties dan een gemiste niet-functionele e-mail.

Zone B – Twijfelzone (0,70 ≤ confidence < 0,90)
In deze zone is het model niet eenduidig. De confidence-score is onvoldoende hoog om automatisch te verwerken, maar ook niet laag genoeg om zonder meer uit te sluiten. Uit de distributietabel blijkt dat in dit gebied de meeste variatie zit tussen categorieën: privé en personeelsvertrouwelijk schalen hier relatief snel mee met functioneel, wat betekent dat de kans op ongewenste meename het grootst is in dit gebied.

Deze zone fungeert daarom als het primaire kwaliteits- en beheersmechanisme. Hier zijn drie opties bestuurlijk verdedigbaar: menselijke controle via een human-in-the-loop werkwijze, aanvullende validatieregels op basis van inhoudelijke kenmerken, of tijdelijke classificatie met een expliciete mogelijkheid tot correctie. De twijfelzone voorkomt dat onzekerheid automatisch wordt weggedrukt naar bewaren of verwijderen, en maakt de besluitvorming controleerbaar en toetsbaar.

Zone C – Lage zekerheid (confidence < 0,70)
In deze zone acht het model de kans klein dat een e-mail functioneel bewaard moet worden. Uit de distributietabel blijkt dat bij 70% recall op functioneel slechts 6,1% van de niet-functionele e-mails wordt meegenomen. Deze zone is in beginsel geschikt voor automatische uitsluiting van bewaring.

Voor partijpolitiek geldt hier een specifieke kanttekening. De zeer kleine dataset met slechts 37 voorbeelden maakt dat het model deze categorie onbetrouwbaar herkent, ook in Zone C. De grillige PR-curve voor partijpolitiek bevestigt dit. Zolang de trainingsdata voor deze categorie niet substantieel is uitgebreid, is het verstandig om partijpolitieke e-mails niet volledig over te laten aan automatische uitsluiting, maar een afzonderlijke controleregel te hanteren.

Samenvatting: gedifferentieerd sturen per categorie
De drie zones bieden een werkbaar kader, maar vragen om gedifferentieerde toepassing per categorie. Onderstaande tabel geeft een overzicht van het aanbevolen handelingsperspectief:

Tabel 9: Drempelwaardes

Categorie	Zone A (≥ 0,90)	Zone B (0,70–0,90)	Zone C (< 0,70)
Functioneel	Automatisch bewaren	Menselijke controle	Automatisch uitsluiten
Niet-functioneel	Automatisch uitsluiten	Menselijke controle	Automatisch uitsluiten
Privé	Altijd controleren	Altijd controleren	Automatisch uitsluiten
Personeels-vertrouwelijk	Altijd controleren	Altijd controleren	Automatisch uitsluiten
Partijpolitiek	Altijd controleren	Altijd controleren	Aparte controleregel

De drempelwaarden in dit kader zijn indicatief en gebaseerd op de empirische resultaten uit de pilot. In een productieomgeving verdient het aanbeveling om deze drempelwaarden te valideren op basis van een bredere en evenwichtiger samengestelde dataset, en om de zones periodiek te herijken naarmate het model meer data verwerkt en de trainingsdata wordt uitgebreid. De bestuurlijke vaststelling van deze drempelwaarden en de governance daaromheen wordt uitgewerkt in hoofdstuk 8.

7.5 Confidence-distributie

Aanvullend op de ROC- en PR-curves is gekeken naar de verdeling van de confidence-scores per categorie. De onderstaande tabel laat zien wat er gebeurt wanneer de drempelwaarde voor functionele e-mails wordt gevarieerd: welk percentage van de functionele e-mails wordt bewaard bij een bepaalde drempel, en hoeveel e-mails uit de overige categorieën worden dan automatisch meegenomen. Omdat de resultaten van beide modellen dicht bij elkaar liggen, is één gecombineerde tabel opgesteld die voor beide modellen representatief is.

Tabel 10: Confidence distributie

Bewaard van functioneel	Functioneel	Niet-functioneel	Privé	Partijpolitiek	Personeels-vertrouwelijk
50%	209.146	17.955 (3,3%)	351 (2,8%)	281 (10,8%)	1.262 (5,4%)
70%	292.748	33.034 (6,1%)	842 (6,7%)	561 (21,6%)	2.665 (11,3%)
80%	334.549	50.217 (9,2%)	1.753 (13,9%)	842 (32,4%)	4.839 (20,6%)
90%	376.420	78.833 (14,5%)	3.507 (27,8%)	1.052 (40,5%)	7.715 (32,8%)
95%	397.321	114.743 (21,1%)	6.032 (47,8%)	1.262 (48,6%)	12.414 (52,8%)
99%	414.013	191.191 (35,2%)	9.889 (78,3%)	1.964 (75,7%)	18.235 (77,6%)
100%	418.221	543.064 (100%)	12.624 (100%)	2.595 (100%)	23.496 (100%)

De tabel maakt inzichtelijk dat de keuze voor een drempelwaarde directe gevolgen heeft voor de volledigheid van de classificatie over alle categorieën. Niet-functioneel daalt relatief langzaam: zelfs bij 50% recall op functioneel wordt slechts 3,3% van de niet-functionele e-mails meegenomen. Privé en personeelsvertrouwelijk schalen iets sneller mee, wat erop wijst dat deze categorieën vaker een matige confidence-score krijgen die dicht bij de functionele grens ligt. Partijpolitiek laat een grilliger patroon zien, wat samenhangt met het zeer kleine aantal voorbeelden in de dataset.

Voor de praktische inzet betekent dit dat de drempelwaarde een bewuste beleidsmatige keuze vereist. Een hogere drempel leidt tot een schonere selectie maar verhoogt het risico dat relevante e-mails worden gemist. Een lagere drempel borgt een hogere volledigheid, maar brengt meer e-mails uit andere categorieën mee die alsnog handmatig beoordeeld moeten worden.

Een belangrijk inzicht is dat de gekozen drempelwaarde niet generiek hoeft te zijn voor alle mailboxen of doelgroepen. De optimale drempelwaarde kan verschillen afhankelijk van het risicoprofiel, de functie en de context waarin de e-mails worden gebruikt.

Zo kan voor mailboxen van bestuurders een hogere drempelwaarde worden gehanteerd, om de kans op onjuiste classificatie van gevoelige of bestuurlijk relevante informatie te minimaliseren. Voor andere doelgroepen, zoals sleutelfunctionarissen en medewerkers, kan juist een lagere drempelwaarde passend zijn, waarbij meer nadruk ligt op volledigheid en ondersteuning van het werkproces. Dit benadrukt dat de keuze voor drempelwaarden niet alleen een technische, maar vooral een beleidsmatige en contextafhankelijke afweging is.

7.6 Vergelijking en conclusie modellen

Wanneer beide modellen naast elkaar worden gezet, vallen de verschillen mee. De algemene prestaties zijn vergelijkbaar: beide modellen behalen een accuracy van 0,85 en een macro average F1 van 0,55. Toch zijn de modellen op verschillende punten sterk. Logistic Regression behaalt zijn score vooral via een hoge precision op de categorie niet-functioneel en een gunstiger balans tussen precision en recall op de kleinste categorieën. Het PEFT-taalmodel laat een iets hogere recall zien op privé-e-mails en scoort beter op de ROC-curve voor de meeste categorieën.

Een belangrijk inzicht dat uit de cijfers volgt, is dat niet alle misclassificaties even ernstig zijn. Een personeelsvertrouwelijke e-mail die ten onrechte als functioneel wordt geclassificeerd brengt een groter risico met zich mee dan omgekeerd. Voor categorieën waarbij een gemiste classificatie grote gevolgen kan hebben, verdient recall daarom meer gewicht dan precision. Dit heeft directe consequenties voor de instelling van de confidence threshold: voor risicovolle categorieën is het verstandig om een lagere drempelwaarde te hanteren, zodat twijfelgevallen vaker ter menselijke beoordeling worden aangeboden.

7.6.1 Welk model voor welke situatie?

Op basis van de resultaten zijn beide modellen inzetbaar, maar met een verschillende voorkeur afhankelijk van de context. Logistic Regression verdient de voorkeur wanneer uitlegbaarheid en transparantie zwaar wegen bijvoorbeeld bij verantwoording richting toezichthouders of bij categorieën met een hoog risicoprofiel zoals personeelsvertrouwelijk en partijpolitiek, waar het model ondanks zijn eenvoud beter presteert dan PEFT. Het PEFT-taalmodel verdient de voorkeur wanneer classificatieprestaties op de grote categorieën (functioneel en niet-functioneel) het zwaarst wegen en wanneer variatie in taalgebruik een grotere rol speelt.

In de praktijk hangt de modelkeuze af van meer dan alleen classificatieprestaties. Onderstaande tabel illustreert hoe de twee modellen zich verhouden op de criteria uit het beoordelingskader in paragraaf 6.1, aangevuld met een aantal factoren die specifiek relevant zijn in een overheidscontext.

Tabel 11: Vergelijking modellen

Factor	Logistic Regression	PEFT (BERTje)
Functionele juistheid	~ Goed op functioneel en niet-functioneel; betere AP op schaarse categorieën	√ Iets sterker op functioneel en niet-functioneel; lagere AP op schaarse categorieën
Consistentie	√ Zelfde invoer geeft altijd zelfde uitkomst	√ Over het algemeen stabiel
Transparantie	√ Classificatie herleidbaar tot concrete tekstkenmerken	~ Beperkt, aanvullende technieken nodig
Wetgeving en compliance	√ Transparantie eenvoudig aantoonbaar, passend bij AVG, Archiefwet en Woo	~ Vraagt aanvullende documentatie voor de EU AI Act
Schaalbaarheid en rekenkracht	√ Lage rekenkosten, snel en licht	~ Hogere rekenkracht vereist
Duurzaamheid	√ Eenvoudig te onderhouden en aan te passen	~ Vraagt meer expertise en infrastructuur
Adoptie en vertrouwen	√ Beslissing uitlegbaar aan medewerkers en bestuurders	~ Minder inzichtelijk voor niet-technisch publiek

De dataset zelf vormt momenteel de belangrijkste bottleneck voor betere modelprestaties, ongeacht de modelkeuze. Met slechts 37 voorbeelden voor partijpolitiek is het voor een model vrijwel onmogelijk om consistente patronen te leren. Uitbreiding van de trainingsdata voor de drie kleinste categorieën zal naar verwachting meer effect hebben op de prestaties dan de keuze voor een complexer model. De aanbevelingen hiervoor zijn uitgewerkt in hoofdstuk 8.

7.7 Organisatorische inzichten

De pilot laat zien dat een zorgvuldige inzet van (zelflerende) systemen begint bij de juiste goedkeuring, borging en risicoafweging. Het moet expliciet zijn onder welke voorwaarden de toepassing wordt ingezet, wie hierover besluit en hoe risico's worden geïdentificeerd, gewogen en beheerst. Deze afwegingen moeten bovendien aantoonbaar zijn, zodat achteraf inzichtelijk is waarom bepaalde keuzes zijn gemaakt en hoe verantwoord gebruik wordt geborgd.

Rol en taakafbakening
Daarnaast vraagt het werken met het systeem om heldere rol- en taakafbakening. Het moet duidelijk zijn wie eigenaar is van de categorieën, wie verantwoordelijk is voor het beheer van regels, wie uitzonderingen beoordeelt en wie toeziet op de kwaliteit van de uitkomsten. Zonder deze expliciete inrichting ontstaat onduidelijkheid, wat direct invloed heeft op de consistentie en betrouwbaarheid van het proces.

Werken vanuit een gedeeld kader
Adoptie speelt hierbij een cruciale rol. Medewerkers moeten werken vanuit een gedeeld begrip van de categorieën en de bijbehorende beoordelingscriteria. Heldere instructies, training en consistente toepassing zijn nodig om te voorkomen dat verschillende interpretaties ontstaan, wat de kwaliteit en reproduceerbaarheid van de classificaties ondermijnt. Inhoudelijke deskundigheid blijft daarbij onmisbaar: in bepaalde gevallen vraagt het duiden van e-mails om context en expertise die niet volledig te vangen zijn in regels of modellen. De pilot biedt, door het gebruik van mailboxen van voormalige bewindspersonen en bestuursraadleden, een geschikte en afgebakende casus om deze werkwijze te ontwikkelen en te toetsen.

AVG en de AW
De pilot heeft een bredere organisatorische vraag blootgelegd die nog niet is beantwoord: welke drempelwaarde voor privacygevoelige e-mails in een te archiveren dataset is acceptabel, en hoe weegt de organisatie de verplichtingen uit de Archiefwet af tegen de vereisten van de AVG? Dit zijn geen technische, maar bestuurlijke afwegingen die op het juiste niveau belegd moeten worden.

Dit vraagstuk werd concreet zichtbaar in de omgang met de dataset vanuit SZW. Het doel was om een drempelwaarde te kiezen waarboven de functionele e-mails met voldoende zekerheid geïdentificeerd zijn en de set opgenomen kon worden in het DMS. De onbalans in de dataset waarbij functionele e-mails sterk ondervertegenwoordigd zijn ten opzichte van privé, partijpolitieke en personeelsvertrouwelijke e-mails maakt echter dat bij elke gekozen drempelwaarde nog een te groot aandeel privacygevoelige berichten in de set overblijft. SZW heeft dit beoordeeld als een onacceptabel privacyrisico, waardoor opname in het DMS in deze fase niet mogelijk is. Zolang hierover geen heldere bestuurlijke kaders bestaan, kunnen de resultaten van de pilot niet worden omgezet in concrete vervolgstappen. Het is dan ook van belang dat deze vraagstukken breed worden opgepakt binnen de eigen organisatie én mogelijk in samenwerking met andere organisaties die voor vergelijkbare dilemma's staan.

7.8 Technische inzichten

De pilot is uitgevoerd in een gecontroleerde setting, waarbij is gewerkt met een afgeschermde laptopomgeving. De risico's rondom het gebruik van de data zijn hierbij ondervangen door duidelijke procedures en maatregelen voor datagebruik en toegang. Dit laat zien dat het mogelijk is om ook met gevoelige data een werkbare en verantwoorde ontwikkel- en testomgeving in te richten, mits de randvoorwaarden expliciet zijn vastgelegd en nageleefd.

Voor de hoeveelheid data binnen de pilot zijn geen performanceproblemen geconstateerd. Dit geeft vertrouwen in de technische haalbaarheid op kleine tot middelgrote schaal, maar zegt nog beperkt iets over gedrag en prestaties bij verdere opschaling.

Een aandachtspunt dat naar voren komt, is de gekozen werkwijze binnen de Microsoft-omgeving. Het gebruik van Outlook voor classificatie blijkt zeer waardevol voor de kwaliteit en snelheid: classificerende medewerkers hebben direct toegang tot metadata, bijlagen en de volledige e-mailcontext. Tegelijkertijd brengt deze keuze complexiteit met zich mee voor de verwerking van data en de automatisering van de verwerkingsketen. Dit is een bewuste afweging die in vervolgtrajecten opnieuw gemaakt moet worden.

Binnen de pilot blijkt daarnaast dat door het toepassen van eenvoudige filters en business rules op de handmatig te classificeren datasets een groot deel van de niet-functionele e-mails snel kon worden geïdentificeerd. Gemiddeld kon ongeveer een derde van de dataset op deze manier worden opgeschoond. Dit benadrukt dat al met relatief eenvoudige en transparante maatregelen aanzienlijke winst kan worden behaald in het reduceren van e-mailvolume.

7.9 Betekenis van de resultaten voor de beleidssporen

De resultaten van de pilot hebben directe betekenis voor zowel het hoofdspoor als het tijdelijk Beleidskader. De classificatietechniek is inzetbaar langs drie toepassingsrichtingen: categorisering aan de voorkant in de mailbox, categorisering tijdens het uitlezen of overbrengen voor bulkarchivering, en het opschonen van bestaande e-mailvoorraden. De inhoudelijke uitwerking van deze toepassingsrichtingen, inclusief de strategische keuzes die daarbij horen, is opgenomen in hoofdstuk 8.

Wat de resultaten in elk geval duidelijk maken, is dat het classificatiesysteem zijn meerwaarde het duidelijkst laat zien bij grote, goed afgebakende e-mailvoorraden. De combinatie van business rules en ML-classificatie biedt een werkbare en verantwoorde aanpak voor het reduceren van e-mailvolume en het identificeren van archiefwaardige informatie. De randvoorwaarden voor verdere toepassing, waaronder uitbreiding van trainingsdata, technische professionalisering en governance, worden uitgewerkt in hoofdstuk 8.

8. Aanbevelingen en vervolg

De pilot bevestigt dat geautomatiseerde e-mailclassificatie technisch haalbaar is en dat een solide fundament voor verdere ontwikkeling wordt gelegd. Tegelijkertijd maakt de pilot duidelijk dat er geen enkelvoudig vervolgpad bestaat. Organisaties verschillen in e-mailcultuur, informatiebeheer en technische uitgangssituatie. Het vervolg vraagt daarom om een combinatie van ontwikkelroutes die elkaar versterken en samen toewerken naar een duurzame, rijksbrede toepassing.

De aanbevelingen zijn geordend langs twee sporen en een randvoorwaardelijk derde onderdeel. Spoor 1 richt zich op het versterken van het fundament: de technische, financiële en bestuurlijke basis die nodig is voordat opschaling verantwoord kan plaatsvinden. Spoor 2 richt zich op toepassing en verbreding: het inzetten van de opgebouwde kennis in nieuwe contexten en het doorontwikkelen van functionaliteit. Als randvoorwaarde voor beide sporen geldt de aanscherping van de categorisering en de herziening van de handreiking “welke e-mail kan weg”. De onderstaande tabel geeft een overzicht van de prioritering en de mogelijke actoren:

Tabel 12: Prioriteiten

Prioriteit	Aanbeveling	Spoor	Actoren
1	Uitbreiding trainingsdata voor privé, personeelsvertrouwelijk en partijpolitiek	Spoor 1	PrOO/CIO-Rijk (opdrachtgever) RDDI (faciliterend) Ministerie (deelnemer en data-leverancier)
2	Vastleggen doelarchitectuur en technische professionalisering	Spoor 1	PrOO/CIO-Rijk (opdrachtgever) RDDI (faciliterend) Beheerpartij (betrokken) Ministerie (deelnemer met testomgeving)
3	Beleggen van governance en eigenaarschap	Spoor 1	PrOO/CIO-Rijk (opdrachtgever) RDDI (faciliterend) Ministeries (betrokken) Beheerpartij (betrokken)
4	Validatie van drempelwaarden in een productieomgeving	Spoor 1	RDDI (faciliteren) Ministerie (deelnemer met testomgeving) Beheerpartij (betrokken)
5	Toepassing bij andere departementen	Spoor 2	RDDI (faciliteren) Ministerie (deelnemer met testomgeving) Beheerpartij (betrokken)
6	Doorontwikkeling: dossierkoppeling en meertaligheid	Spoor 2	RDDI (faciliteren) Ministeries (deelnemer met testomgeving) Beheerpartij (betrokken)
7	Herziening handreiking en aanscherping categorisering	Rand-voorwaarden	PrOO/CIO-Rijk (opdrachtgever) RDDI (faciliteren) NA (inhoudelijk) Ministeries (betrokken)

8.1 Toepassingsgebieden

De pilot heeft niet alleen een werkende toepassing opgeleverd, maar vooral een technisch en inhoudelijk fundament voor automatische e-mailclassificatie. Op basis van dit fundament kunnen meerdere toepassingen worden ontwikkeld. Deze toepassingen maken gebruik van dezelfde onderliggende techniek, classificatieregels en inzichten, maar verschillen in het moment en de context waarin zij worden ingezet.

Het is daarom behulpzaam om deze ontwikkeling te benaderen als één basisproduct met meerdere toepassingsrichtingen. Eerst wordt het fundament verder ontwikkeld en gestabiliseerd, dit is de kern van Spoor 1 in paragraaf 8.2. Vervolgens kan gericht worden gekozen welke toepassingsrichting als eerste wordt uitgewerkt en geïmplementeerd. Andere toepassingen kunnen daarna op hetzelfde fundament worden toegevoegd. In de praktijk zijn drie toepassingsrichtingen te onderscheiden:

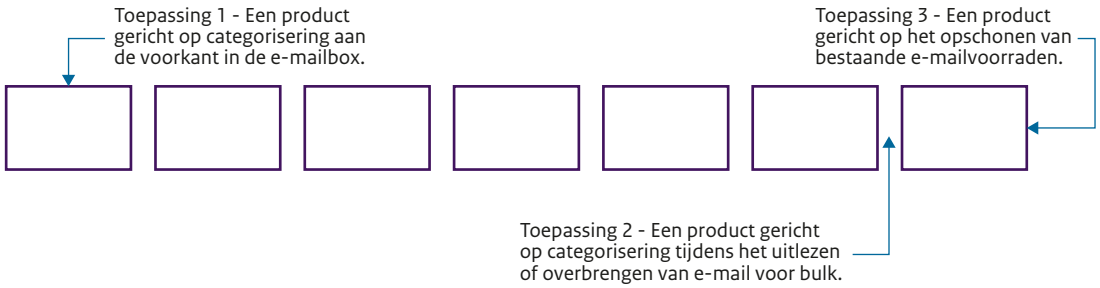
Categorisering aan de voorkant in de mailbox
Deze toepassingsrichting richt zich op het structureel categoriseren van e-mails bij binnenkomst of verzending, voor alle medewerkers. Het doel is om e-mails direct bij ontvangst of gebruik te classificeren, zodat archiefwaardige informatie tijdig kan worden herkend en beheerd. Dit speelt in op het groeiende onvermogen om met de huidige middelen het volume en de complexiteit van e-mail structureel als archiefbescheiden te verwerken. Deze toepassing heeft de grootste potentiële impact, maar vraagt ook de zwaarste organisatorische inbedding.

Categorisering tijdens het uitlezen of overbrengen voor bulkarchivering
Deze toepassingsrichting richt zich op het classificeren van e-mails op het moment dat deze worden uitgelezen, overgebracht of geselecteerd voor bulkarchivering. Door voorafgaand aan opname selectie en opschoning toe te passen, wordt voorkomen dat niet-relevante of dubbel aanwezige informatie structureel wordt opgeslagen. Dit draagt bij aan een beheersbaar en kwalitatief e-mailarchief.

Opschonen van bestaande e-mailvoorraden
Deze toepassingsrichting richt zich op het analyseren, classificeren en opschonen van bestaande e-mailvoorraden waarvoor nog geen waardering en selectie heeft plaatsgevonden. De pilot laat zien dat deze aanpak werkt en dat het mogelijk is om grote voorraden op een verantwoorde manier te verwerken. Dit is de meest directe vervolgstap op de uitgevoerde casus bij het ministerie van SZW.

De drie toepassingsrichtingen delen een groot deel van dezelfde bouwstenen; classificatiemodellen, analysetechniek en verwerkingslogica, maar verschillen in positionering binnen het e-mailproces, organisatorische inbedding en technische integratie. De keuze voor een eerste toepassingsrichting is daarom primair een strategische en organisatorische keuze, niet zozeer een technische. Gezamenlijk vormen de drie richtingen een ketenbenadering: nieuwe achterstanden worden voorkomen, archiveringsprocessen worden beheersbaar ingericht en bestaande voorraden worden verantwoord opgeschoond.

Figuur 8: De e-mailketen en toepassingen



8.2 Spoor 1: Verstevigen van het fundament

Voordat verdere toepassing plaatsvindt, is het noodzakelijk om het fundament van de oplossing te versterken en te professionaliseren. Ongeacht de gekozen toepassingsrichting vraagt dit om gerichte investeringen in vier samenhangende onderdelen.

8.2.1 Uitbreiding van de trainingsdata voor de 3P-categorieën

De pilot heeft een bredere organisatorische vraag blootgelegd die nog niet is beantwoord: welke drempelwaarde voor privacygevoelige e-mails in een te archiveren dataset is acceptabel, en hoe weegt de organisatie de verplichtingen uit de Archiefwet af tegen de vereisten van de AVG? De pilot biedt op dit bestuurlijke vraagstuk geen definitief antwoord, maar wijst wel een richting. Uit de resultaten blijkt dat een uitbreiding van de dataset met e-mails vanuit de categorieën; privé, personeelsvertrouwelijk en partijpolitiek naar verwachting de balans tussen functionele en privacygevoelige e-mails aanzienlijk verbetert. Dit biedt handelingsperspectief: een evenwichtiger dataset vergroot de kans dat bij een aanvaardbare drempelwaarde het aandeel privacygevoelige e-mails in de te archiveren set daalt tot een niveau dat bestuurlijk en juridisch verdedigbaar is. Een concrete aanbeveling is dan ook om bij eventuele doorontwikkeling van de tool als eerste stap de dataset uit te breiden, en parallel het gesprek te voeren over de bestuurlijke kaders rondom de AVG-Archiefwet-afweging.

De resultaten laten zien dat de modellen goed presteren op functioneel en niet-functioneel, maar dat de kleinere categorieën; privé, personeelsvertrouwelijk en partijpolitiek te weinig trainingsvoorbeelden bevatten voor betrouwbare classificatie. Voor partijpolitiek geldt dit in het bijzonder: met slechts 37 voorbeelden is consistente herkenning vrijwel onmogelijk. Een algemeen gehanteerde vuistregel is dat een categorie minimaal 100 tot 200 representatieve voorbeelden nodig heeft. Uitbreiding van de dataset voor deze drie categorieën heeft dan ook de hoogste prioriteit en zal naar verwachting een grotere positieve impact hebben op de modelprestaties dan de inzet van een complexer model.

Resultaat: uitgebreide en evenwichtig samengestelde trainingsset wat leidt tot betere model resultaten en kan helpen in het beantwoorden van het organisatievraagstuk.

8.2.2 Doelarchitectuur en technische professionalisering

De pilotcode en -omgeving zijn ontwikkeld als proefsetting. Voor duurzame inzet is refactoring en professionalisering noodzakelijk. Dit omvat het opschonen en standaardiseren van de codebase, het volledig en actueel maken van documentatie, en het vastleggen van ontwikkelkeuzes op een navolgbare manier. Daarnaast moet een doelarchitectuur worden uitgewerkt die beschrijft hoe de oplossing past binnen bestaande infrastructuren en standaarden, inclusief keuzes rond hosting, beveiliging en schaalbaarheid.

Resultaat: vastgesteld architectuurdocument en gestandaardiseerd technisch fundament met eerste werkbare toepassing.

8.2.3 Governance en eigenaarschap

Een werkende techniek is onvoldoende als niet duidelijk is wie eigenaar is van de classificatiecategorieën, wie verantwoordelijk is voor het beheer van regels en drempelwaarden, wie uitzonderingen beoordeelt en wie toeziet op de kwaliteit van de uitkomsten. Dit geldt zowel op het niveau van de afzonderlijke organisatie als rijksbreed. Het beleggen van eigenaarschap en het inrichten van wijzigingsbeheer en kwaliteitscontrole zijn dan ook een randvoorwaarde voor verantwoorde opschaling.

Daarbij is de keuze voor de juiste drempelwaarden per categorie geen puur technische beslissing, maar een bestuurlijke. De zones die in paragraaf 7.5 zijn beschreven; hoge zekerheid, twijfelzone en lage zekerheid, bieden een werkbaar kader, maar vereisen expliciete bestuurlijke vaststelling. Wie beslist dat een e-mail in zone B automatisch wordt verwerkt? Wie is aanspreekbaar als een classificatie onjuist blijkt? Deze vragen moeten worden beantwoord voordat het systeem productief wordt ingezet.

Resultaat: governance- en beheermodel, inclusief vastgestelde drempelwaarden per categorie en escalatieproces voor twijfelgevallen.

8.2.4 Financieel kader

Voor structurele inzet is inzicht nodig in de ontwikkel- en beheerkosten en de mogelijke financierings- en exploitatiemodellen. Een businesscase maakt zichtbaar wat de investering oplevert in termen van tijdwinst, reductie van e-mailvolume en verbetering van archiefkwaliteit.

Resultaat: businesscase en exploitatiemodel.

8.3 Spoor 2: Toepassen, verbreden en doorontwikkelen

Naast het versterken van het fundament is er ruimte voor toepassing en verdere ontwikkeling. Deze activiteiten dragen bij aan draagvlak, verfijning en zichtbaarheid van de oplossing, mits zij aansluiten op het fundament uit Spoor 1.

8.3.1 Toepassing bij andere departementen

Een eerste lijn van verbreding ligt in het toepassen van de aanpak bij andere ministeries, bij voorkeur gericht op mailboxen van vertrokken medewerkers of sleutelfunctionarissen. Door de werkwijze in verschillende organisatorische contexten toe te passen, ontstaat beter inzicht in de herhaalbaarheid van de aanpak en de mate waarin de oplossing generiek inzetbaar is. Daarbij is het van belang te erkennen dat de inzet bij mailboxen van zittende medewerkers of sleutelfunctionarissen complexer is: in die situaties spelen aanvullende vraagstukken rondom zeggenschap, zorgvuldigheid en impact op de werkpraktijk, die een zwaardere organisatorische inbedding vereisen.

8.3.2 Doorontwikkeling: dossierkoppeling en meertaligheid

Een belangrijke vervolgstap is het automatisch koppelen van geclassificeerde functionele e-mails aan dossiers in het DMS op basis van metadata. Dit maakt verdere automatisering van dossiervorming mogelijk en versterkt de aansluiting op het informatiebeheerproces. Daarnaast is uitbreiding richting meertalige e-mails van belang, gezien de internationale correspondentie die in overheidsomgevingen voorkomt.

8.3.3 Beleids- en inrichtingsvraagstukken

Tot slot spelen bredere vraagstukken rondom beleid en inrichting, zoals de omgang met zakelijke e-mails na classificatie, de inrichting van duurzame opslag en de juridische en archiefrechtelijke implicaties van de werkwijze. Deze vragen zijn essentieel voor structurele borging en vragen om nadere uitwerking in samenhang met de technische ontwikkeling.

8.4 Categorisering en handreiking “welke e-mail kan weg”

Een randvoorwaarde voor zowel het toepassen van filterregels als AI-classificatie is dat de categorisering en het bijbehorende handelingsperspectief eenduidig zijn. De pilot heeft laten zien dat grensgevallen zoals e-mails die deels functioneel en deels privé zijn, of partijpolitieke e-mails met beleidsinhoudelijke elementen in de praktijk tot interpretatieverschillen leiden. Eenduidigheid is niet alleen van belang voor de kwaliteit van de classificatie, maar ook voor de uitlegbaarheid richting toezicht, medewerkers en externe partijen.

Aanbevolen wordt om de bestaande categorisering aan te scherpen met expliciete definities, voorbeelden en uitgewerkte grensgevallen per categorie. Dit vormt de basis voor zowel de handleiding als het trainingsmateriaal voor toekomstige modellen.

De handreiking “welke e-mail kan weg” dient daarnaast te worden herzien. Het document is cruciaal voor brede adoptie, maar beschrijft momenteel vooral wat weg kan. Voor de praktijk is minstens zo belangrijk wat te doen bij twijfel: wanneer labelen, wanneer archiveren, wanneer doorzetten naar een dossier en wanneer escaleren. De herziene handreiking moet aansluiten op de vijf classificatiecategorieën zoals gehanteerd in de pilot en moet ook het handelingsperspectief beschrijven bij de drie besliszones (hoge zekerheid, twijfelzone, lage zekerheid) uit paragraaf 7.5.

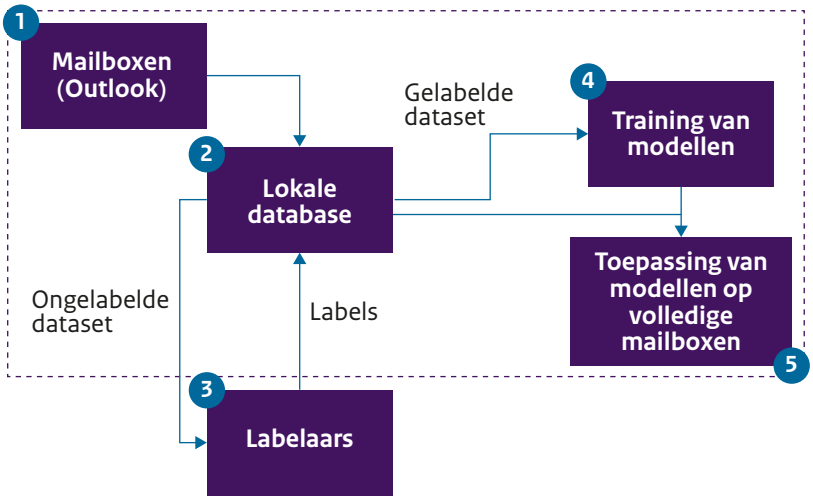
Resultaat: herziene handreiking “welke e-mail kan weg” en aangescherpte categoriseringsdocumentatie, bruikbaar als basis voor zowel medewerkers instructie als modeltraining.

Bijlagen

Bijlage I: Technische uitleg

De technische werkzaamheden die uitgevoerd zijn binnen deze pilot zijn, met uitzondering van het handmatig labelen, volledig uitgevoerd op één laptop. De verschillende technische uitvoerwerkzaamheden zijn gecoördineerd via de programmeertaal Python.

Figuur 9: Componenten van pilot techniek



Stap 1: De Outlook-gegevensbestanden zijn uitgelezen en verwerkt via Outlook. Dit is op een geautomatiseerde manier gegaan via de Component Object Model (COM)-interface, een interface die via de Python-package pywin32 automatisering van Outlook-handelingen mogelijk maakt binnen Python.

Stap 2: Voor snelle toegang tot alle informatie rondom de mailboxen en classificaties, is een SQLite-database opgezet. In deze database werd (tijdelijk) opgeslagen:

- Gegevensbestanden en eigenaar mailbox
- Folders binnen gegevensbestanden
- (Mail-)items binnen folders
- Relevante (meta)data van items
- Labelset, oftewel: welke items wanneer door welke mensen zijn gelabeld, inclusief de labels zelf
- Labels die door het uiteindelijk gekozen model zijn gegeven aan de volledige mailboxen

Stap 3: Via aansturing vanuit de database en Python zijn willekeurige mails geselecteerd en vanuit mailboxen gekopieerd naar aparte Outlook-gegevensbestanden, die in de omgeving van labelaars zijn ingeladen om te kunnen labelen. Na het labelen, waarbij labelaars mails naar specifieke mapjes hebben gesleept, zijn de gelabelde gegevensbestanden ingelezen naar de database.

Stap 4 en 5: Na het creëren van een dataset met labels zijn de modellen getraind. Dit is gedaan in de volgende stappen:

- De ongebalanceerde dataset is tijdens de training gecompenseerd door gewichten toe te kennen aan individuele voorbeelden. Deze gewichten bepalen in welke mate het model zichzelf aanpast aan de hand van een trainingsvoorbeeld. Normaal zijn deze gewichten omgekeerd evenredig aan het aantal voorbeelden van elk label. Om de prestaties verder te optimaliseren, is er in dit geval gekozen voor een instelbare parameter die bepaalt in welke mate de labelgewichten omgekeerd evenredig worden toegepast.
- Als features voor de modellen zijn de body, onderwerp en de namen van de bijlages gebruikt. Hierbij zijn HTML/RTF-opmaak uit de body van de email opgeschoond.
- Voor het Logistic Regression-model is gebruik gemaakt van de implementatie uit de Python-package scikit-learn. De features zijn via TF-IDF verwerkt.
- Bij het Transformer-model is als basismodel bert-dutch-base-cased (BERTje) gebruikt. Dit model is met behulp van Low-Rank Adaptation (LORA) gefinetuned.

Bijlage II: Gewone en bijzonder persoonsgegeven

De AVG maakt onderscheid tussen ‘gewone’ en ‘bijzondere’ persoonsgegevens. Bijzondere persoonsgegevens zijn gegevens die zó privacygevoelig zijn dat het grote(re) impact op iemand kan hebben als deze gegevens worden verwerkt. Gewone persoonsgegevens zijn alle informatie die direct over een persoon gaat of die indirect naar die persoon te herleiden is.

Tabel 13: Voorbeelden van gegevens

Voorbeelden bijzondere persoonsgegevens	Voorbeelden gewone persoonsgegevens
Persoonsgegevens waaruit iemands ras of etnische afkomst blijkt.	Iemands naam.
Persoonsgegevens waaruit iemands politieke opvattingen blijken.	Het adres van een persoon.
Persoonsgegevens waaruit iemands religieuze of levensbeschouwelijke overtuigingen blijken.	Het telefoonnummer van iemand.
Persoonsgegevens waaruit het lidmaatschap van een vakbond blijkt.	Een pasfoto van een persoon.
Gegevens over iemands gezondheid.	
Gegevens over iemands seksueel gedrag of seksuele gerichtheid.	
Genetische gegevens.	
Biometrische gegevens (bedoeld voor de unieke identificatie van een persoon).	

Strafrechtelijke gegevens zijn ook heel gevoelige persoonsgegevens, maar die vallen niet onder het begrip ‘bijzondere persoonsgegevens’ volgens de AVG. Er gelden wel speciale regels voor het verwerken van strafrechtelijke gegevens. In dit onderzoek nemen we deze type gegevens niet mee als aparte categorisering, maar bevelen we wel aan bij het ontwikkelen van een duurzame oplossing die Rijksbreed kan worden toegepast deze categorie te onderzoeken. Het Ministerie van Justitie en Veiligheid heeft bijvoorbeeld veel te maken met deze gegevens.

Bijlage III: Verdeling dataset

Tabel 14: Cijfers over de e-mails

Getal	Duiding
20	E-mailboxen
834.531	E-mails in totaal
752.074	Unieke e-mails in totaal
41.727	Gemiddelde hoeveelheid e-mails per mailbox
13.709	E-mails handmatig geclassificeerd
753	E-mails zijn twee keer gelabeld
9	E-mails zijn drie keer gelabeld
670	E-mails gecorrigeerd (4,9% van het geheel)[13]

Tabel 15: Verdeling gelabelde e-mails

Classificatie	Aantal keer gelabeld
Functioneel	5958 (41.8%)
Niet-functioneel	7739 (54.3%)
Privé	180 (2.4%)
Personeelsvertrouwelijk	335 (1.3%)
Partijpolitiek	37 (0.3%)

13 Het is goed om te vermelden dat de mensen die onderdeel zijn geweest van de handmatige classificatie, inhoudelijk bekend zijn met de e-mails of eerdere ervaring hebben met classificeren.

Bijlage IV: Filteren van e-mails

Om het handmatige classificatieproces te optimaliseren, doorlopen de mailboxen eerst een voorbereidende fase. Deze preparatie omvat twee hoofdstappen 1) voorfiltering, items die niet aan de definitie van een e-mail voldoen worden automatisch uitgefilterd uit de Outlook omgeving. 2) automatische classificatie via filterregels, e-mails die met 100% zekerheid geclassificeerd kunnen worden, worden automatisch verwerkt volgens vooraf gedefinieerde filterregels. De onderstaande filterregels zijn toegepast op de datasets om niet-functionele e-mails automatisch te categoriseren.

Tabel 16: Toegepaste filters

Type	Filtering
Agendaverzoek	“Geaccepteerd”/ “Accepted”
	“Geweigerd”/ “Declined”
	“Voorlopig”/ “Tentative”
	“Verplaatst”/ “Changed”
	“Nieuw tijdstip”/ “New time”
	“Geannuleerd”/ “Cancelled”
Automatische antwoorden	“Automatisch antwoord”/ “Automatic reply”
Niet te leveren	“Onbestelbaar”
Kalenderbestanden	(.ics)
No-reply-adressen	No-reply@
	Noreply@
	“Do not reply”
Systeem notificaties	“Notifications@”
Titel start met onderwerp	“Newsletter”
	“Nieuwsbrief”
	“Weekbericht”
	“Promo”
	“Aanbieding”
	“ANP”
	“FD Ochtendnieuws”
	“CAR Redactie”
Systeem notificaties	“DigiDoc Bericht”
Out-of-office	“Automatisch antwoord”
	“Automatic answer”
	“Automatic reply”
Wachtwoord	“Wachtwoord wijziging”/ “Password reset”
Test	“Test”
	“Proef”
Headers	Mailer
	List-Unsubscribe
	Precedence: bulk of Precedence
	Auto-Submitted: auto-generated
Ontvanger	team@
	info@

Bijlage V: Stappenplan en waarborgen

Dit stappenplan beschrijft een algemeen toepasbare stappen voor organisaties die een pilot willen uitvoeren waarbij AI-modellen lokaal en volledig afgeschermd worden ontwikkeld en toegepast op e-mails. Het biedt houvast voor zowel technische als organisatorische stappen en legt de nadruk op veiligheid, reproduceerbaarheid en controle.

Stap 1 – Aanvragen en inrichten van een geschikte ontwikkelomgeving
Voor de pilot is een lokale ontwikkelomgeving nodig in de vorm van een laptop of een afgeschermde server met voldoende rekenkracht. Minimaal vereist is een moderne processor, voldoende RAM en een grafische kaart met voldoende VRAM en rekenkracht. Daarnaast moet de omgeving beschikken over ontwikkelrechten om Python, machine learning-libraries en andere tools te kunnen installeren.

Deze ontwikkelomgeving wordt aangevraagd via de reguliere ICT-leverancier van de organisatie. Belangrijk is dat vooraf wordt bepaald of standaard beveiligingsmaatregelen (zoals hardening of restrictief patchbeheer) kunnen worden toegepast. Als dat niet mogelijk is vanwege ontwikkelbehoeften, worden in latere stappen aanvullende beveiligingsmaatregelen ingezet om de risico's te mitigeren.

Aanpak pilot
Ten behoeve van de pilot is een laptop met verhoogde ontwikkelrechten beschikbaar gesteld, met expliciete goedkeuring van de afdeling informatiebeveiliging.

Stap 2 – Formeel verzoek tot datalevering via de daarvoor aangewezen afdeling
De benodigde data mogen pas worden gebruikt nadat hiervoor een formeel verzoek is ingediend bij de afdeling die binnen de organisatie verantwoordelijk is voor het afhandelen van informatieverzoeken. In sommige organisaties is dit de BVA; elders kan dit een informatiebeheer- of privacyteam zijn dat alle dataverzoeken toetst en doorgeleidt naar de ICT-leverancier. Het verzoek bevat altijd een onderbouwing van het doel, de reikwijdte van de data en de maatregelen die worden genomen voor veilige verwerking. Pas na formele goedkeuring wordt de dataset aangeleverd.

Aanpak pilot
Binnen de pilot was de BVA verantwoordelijk voor het aanvragen van de e-mailboxen bij SSC-ICT via een vaste procedure. Hierbij is ervoor gekozen om de data aangeleverd te krijgen via het netwerk van SZW. Hierbij is de data in een afgesloten map geplaats en via een versleutelde harde schijf op de laptop geplaats.

Stap 3 – Installatie van de technische omgeving
De ontwikkelomgeving wordt voorzien van alle benodigde software, waaronder Python, machine learning-libraries en tooling voor data-analyse. Hier is een data scientist of AI-adviseur voor nodig. De internetverbinding wordt alleen gebruikt voor installatie van vertrouwde pakketten. Na installatie wordt gecontroleerd of de omgeving functioneert zoals bedoeld, inclusief logging, encryptie en beveiligingsinstellingen. Voor deze stap is er specifieke kennis nodig voor de installatie. Ook is er een mogelijkheid om deze stap door een leverancier te laten uitvoeren als er geen expertise aanwezig is binnen de organisatie.

Aanpak pilot
De installatie van de ontwikkelomgeving is tijdens de pilot verzorgd door interne expertise vanuit de organisatie.

Stap 4 – Fysieke en digitale isolatie van het apparaat
Na installatie wordt de ontwikkelomgeving, bij een laptop, volledig geïsoleerd. Alle netwerkmogelijkheden worden uitgeschakeld via BIOS/UEFI of vergelijkbare instellingen, inclusief wifi, ethernet, bluetooth en andere interfaces. Hierdoor ontstaat een afgezonderde omgeving waar data noch uit kan stromen, noch kan worden benaderd van buitenaf. Voor deze stap is er specifieke kennis nodig over de benodigde hardware en de mogelijkheden om de hardware te voorzien van de waarborgen. De ICT-leverancier kan hier normaliter mee helpen.

Aanpak pilot
SSC-ICT heeft tijdens de pilot geadviseerd over de isolatiemogelijkheden. De feitelijke configuratie van de laptop is verzorgd door het projectteam.

Stap 5 – Plaatsing in een beveiligde werkruimte
De ontwikkelomgeving wordt fysiek geplaatst in een beveiligde, afgesloten ruimte. Alleen vooraf aangewezen medewerkers hebben toegang. Toezicht en logging van toegang kunnen hier onderdeel zijn van de procedure. Dit is bijvoorbeeld een afgesloten kantoorruimte waar alleen gescreende medewerkers binnenkomen, en waar het apparaat niet onbeheerd toegankelijk is.

Aanpak pilot
Gedurende de pilot is de laptop opgeslagen in een afgesloten werkruimte, toegankelijk uitsluitend voor gescreende medewerkers.

Stap 6 – Inlezen van de data
De dataset wordt éénmalig aangeleverd via een beveiligde, versleutelde drager. Het inlezen gebeurt volgens een vastgestelde procedure die voorkomt dat de data wordt gekopieerd, verplaatst of buiten de omgeving terechtkomt. Na overdracht wordt gecontroleerd of de data volledig is en correct is geïmporteerd.

Aanpak pilot
Na formele goedkeuring van het dataverzoek via de daartoe geldende organisatieprocedure, is de dataset overgedragen via een versleutelde externe harde schijf en ingelezen op de geïsoleerde hardware.

Stap 7 – Ontwikkeling, training en uitvoering van het model
Alle modelontwikkeling, training en classificatie vindt lokaal plaats. Het model wordt getest, gevalideerd en alleen binnen de afgeschermde omgeving uitgevoerd. De volledige verwerkingsketen blijft lokaal en is herleidbaar vastgelegd.

Aanpak pilot
Binnen de pilot zijn alle stappen van modelontwikkeling, training en classificatie uitgevoerd op de geïsoleerde laptop. Er is geen gebruik gemaakt van externe systemen of cloudoplossingen. De resultaten zijn gedocumenteerd en intern gearchiveerd.

Stap 8 – Analyse en verwerking van resultaten
De resultaten van het model worden binnen de geïsoleerde omgeving geanalyseerd. Hierbij wordt gecontroleerd of classificaties logisch, consistent en reproduceerbaar zijn. Alleen geaggregeerde of niet-herleidbare resultaten mogen (na controle) buiten de omgeving worden overgenomen.

Aanpak pilot
De resultaten van de modelclassificatie zijn binnen de afgeschermde omgeving geanalyseerd en geverifieerd op consistentie en reproduceerbaarheid. Enkel geaggregeerde uitkomsten zijn als rapportage buiten de omgeving gebracht, volgens de vastgestelde procedure.

Stap 9 – Oplevering van het eindresultaat
De geclassificeerde gegevens worden voorbereid voor overdracht naar reguliere systemen van de organisatie. De overdracht vindt plaats volgens een vooraf vastgelegde procedure.

Aanpak pilot
De eindresultaten zijn in overeenstemming met het vastgestelde overdrachtsprotocol overgedragen aan de verantwoordelijke afdeling binnen de organisatie, voorzien van de bijbehorende documentatie en kwaliteitscontroles.

Stap 10 – Afronding, vernietiging en inlevering
Na afronding wordt de laptop volledig gewist, inclusief alle data, modellen, logs en tijdelijke bestanden. De opschoning gebeurt volgens een onherstelbare verwijdermethode. Daarna wordt de laptop ingeleverd en wordt de pilotomgeving formeel afgesloten.

Aanpak pilot
Na afronding van de pilot is de laptop volledig gewist volgens de organisatiestandaard voor veilige gegevensverwijdering. De hardware is teruggekeerd aan het AI en datateam van SZW en de pilotomgeving is formeel afgesloten.

Bijlage VI: Kwaliteitsborging

In deze bijlage worden de verschillende opties beschreven die in het kader van de pilot zijn verkend. Daarnaast wordt toegelicht welke keuze uiteindelijk is gemaakt en op basis van welke overwegingen.

Optie 1: 5% overlap in te classificeren data
Bij deze aanpak krijgt iedere labelaar een deel van de dataset toegewezen, waarbij voor ongeveer 5% van de items bewust overlap wordt gecreëerd. Dit betekent dat een klein deel van de mails door twee verschillende mensen wordt geclassificeerd. Door deze overlap ontstaan controlepunten waarmee afwijkende classificaties kunnen worden geïdentificeerd en direct tijdens het proces kunnen worden besproken, waarna een definitieve classificatie wordt vastgesteld.

Een belangrijk voordeel van deze werkwijze is dat inconsistenties vroegtijdig aan het licht komen, waardoor de kwaliteit al tijdens het classificeren kan worden verbeterd. Een nadeel is dat het classificatieproces hierdoor meer tijd kost, omdat een deel van de data bewust dubbel moet worden geclassificeerd.

Optie 2: 10% van de geclassificeerde data achteraf apart nemen
Bij deze optie wordt pas aan het einde van het classificatieproces een steekproef van 10% uit de al geclassificeerde data geselecteerd. Deze subset wordt opnieuw beoordeeld om te controleren of er afwijkingen bestaan in de classificaties, waarna eventuele inconsistenties alsnog kunnen worden gecorrigeerd en een definitieve classificatie wordt vastgesteld.

Het voordeel van deze aanpak is dat labelaars geen extra overlappende data hoeven te verwerken, waardoor de werkdruk tijdens het proces lager is. Het nadeel is dat eventuele structurele verschillen in interpretatie pas achteraf zichtbaar worden, waardoor er minder gelegenheid is om tijdens het proces bij te sturen.

Gemaakte keuze
Op basis van de verkenning van beide opties is gekozen voor optie 1: het opnemen van 5% overlap in de te classificeren data. Deze keuze is gemaakt omdat deze aanpak beter aansluit bij het doel om de kwaliteit en betrouwbaarheid van de uiteindelijke dataset zo hoog mogelijk te maken.

Een belangrijk voordeel van optie 1 is dat de kwaliteitscontrole integraal onderdeel wordt van het classificatieproces. Doordat overlappende items direct door een labelaar worden beoordeeld, worden verschillen in interpretatie vroegtijdig zichtbaar. Dit maakt het mogelijk om tijdens het proces gezamenlijke beslissingen te nemen en classificatieregels waar nodig aan te scherpen. Deze iteratieve terugkoppeling voorkomt dat inconsistenties zich door het hele proces heen ophopen en pas aan het einde aan het licht komen.

Daarnaast zorgt optie 1 ervoor dat we meer inzicht krijgen in de mate van subjectiviteit binnen de classificaties. Door structurele verschillen of twijfelgevallen al tijdens de uitvoering te signaleren, kan het team beter bepalen welke typen berichten gevoelig zijn voor variatie in interpretatie. Dit bevordert niet alleen de kwaliteit van de dataset, maar levert ook waardevolle kennis op voor toekomstige classificatie- of AI-trainingsprocessen.

Bijlage VII: Aandachtspunten voor een productieomgeving

Onderstaande kwaliteitscriteria zijn geïdentificeerd in het Plan van Aanpak, maar konden binnen de scope van de pilot niet worden onderzocht. In een vervolgtraject of bij de overgang naar een productieomgeving dienen deze criteria alsnog te worden geadresseerd.

Gebruiksvriendelijkheid
De oplossing dient eenvoudig inpasbaar te zijn in het dagelijkse werkproces, met minimale extra handelingen voor medewerkers. De classificatie moet ondersteunend zijn en geen onnodige administratieve last introduceren.

Aanbevolen meetaanpak voor vervolg:
- Tijdwinst per medewerker meten via een voor- en nameting.
- Aantal benodigde handmatige acties per classificatiecyclus registreren.
- Gebruikerstevredenheid meten via een gestandaardiseerde vragenlijst (bijvoorbeeld SUS-score).
- Adoptiegraad binnen de pilotgroep monitoren over tijd.

Beveiliging en privacy
Omdat e-mails persoonsgegevens en vertrouwelijke informatie kunnen bevatten, dient de verwerking te voldoen aan de AVG en de relevante departementale richtlijnen. Dit omvat dataminimalisatie, zorgvuldig toegangsbeheer en het beperken van risico's op onbedoelde verwerking of datalekken.

Aanbevolen meetaanpak voor vervolg:
- AVG-compliancechecklist opstellen en periodiek doorlopen.
- Toegangsrechten en autorisatiestructuur formeel vastleggen en reviewen.
- Privacy Impact Assessment (PIA) uitvoeren voorafgaand aan productie.
- Aantal geconstateerde afwijkingen of incidenten registreren en opvolgen.

Bijlage VIII: Toelichting publicatie broncode pilotsoftware

In het kader van de pilot is software ontwikkeld voor het automatisch classificeren van e-mails op basis van AI-analyse van de inhoud. Deze software is ontwikkeld met publieke middelen en heeft als doel om processen binnen de overheid efficiënter, transparanter en beter beheersbaar te maken.

Bij de afronding van de pilot komt de vraag aan de orde op welke wijze de ontwikkelde broncode beschikbaar wordt gesteld. Gezien de publieke aard van de ontwikkeling en de wens tot transparantie is onderzocht op welke manier publicatie van de software kan plaatsvinden. In deze bijlage worden drie mogelijke publicatievormen toegelicht:

- Publicatie onder een permissieve open source licentie (MIT);
- Publicatie onder een copyleft open source licentie (GNU GPL v3);
- Publicatie zonder open source licentie, uitsluitend ten behoeve van transparantie.

Wettelijk kader en uitgangspunten

De keuze om broncode openbaar te maken sluit aan bij de Wet open overheid (Woo). De Woo bevat een inspanningsverplichting voor overheidsorganisaties om informatie die met publieke middelen is ontwikkeld zoveel mogelijk proactief openbaar te maken. Broncode van software kan hieronder vallen.

Daarnaast bepaalt de Wet hergebruik van overheidsinformatie (Who) dat openbaar gemaakte informatie zoveel mogelijk geschikt moet worden gepubliceerd voor hergebruik. Wanneer software daadwerkelijk voor hergebruik beschikbaar moet worden gesteld, gebeurt dit doorgaans via een open source licentie.

Bij publicatie van software moet ook rekening worden gehouden met de Wet Markt & Overheid (Wet M&O). Deze wet voorkomt dat overheidsorganisaties commerciële activiteiten ontplooien die de markt kunnen verstoren. In dit geval wordt de software niet commercieel geëxploiteerd en wordt geen exclusief economisch voordeel nagestreefd.

Een verkenning van de markt laat zien dat er verschillende commerciële oplossingen bestaan voor AI-gestuurde e-mailclassificatie. Deze oplossingen richten zich echter op bredere inbox-automatisering, documentverwerking of juridische eDiscovery. Er is geen partij aangetroffen die exact dezelfde positionering en doelstelling heeft als de in deze pilot ontwikkelde oplossing. Openbaarmaking van de broncode leidt daarom niet tot het verdringen van een specifieke marktpartij.

Overwegingen bij de wijze van publicatie

Bij de afronding van de pilot worden drie mogelijke vormen van publicatie onderzocht.

1. Publicatie onder een permissieve open source licentie (MIT)

De MIT-licentie is een internationaal veelgebruikte open source licentie die maximale vrijheid biedt voor gebruik, aanpassing en verdere verspreiding van software. Organisaties mogen de software vrij gebruiken, aanpassen en integreren in andere software, zowel in open source als in commerciële toepassingen.

De licentie stelt slechts beperkte voorwaarden, zoals het behouden van de oorspronkelijke copyrightvermelding en het opnemen van de licentietekst bij verdere verspreiding. Voordelen van een MIT-licentie zijn onder andere:

- Lage juridische drempel voor hergebruik;
- Brede toepasbaarheid door andere overheden, kennisinstellingen en marktpartijen;
- Eenvoudige integratie in bestaande systemen.

Een mogelijk nadeel is dat afgeleide werken niet verplicht open source hoeven te blijven. Marktpartijen zouden de software verder kunnen ontwikkelen en als gesloten product aanbieden.

2. Publicatie onder een copyleft licentie (GNU GPL v3)

De GNU General Public License v3 is een zogenoemde copyleft-licentie. Deze licentie vereist dat afgeleide werken eveneens onder dezelfde open source licentie beschikbaar worden gesteld. Dit betekent dat aanpassingen, uitbreidingen en verdere ontwikkelingen van de software ook openbaar moeten blijven wanneer deze worden verspreid. Voordelen van deze licentie zijn:

- Waarborg dat doorontwikkelingen open blijven;
- Blijvende beschikbaarheid van met publieke middelen ontwikkelde functionaliteit;
- Stimulering van samenwerking binnen de open source gemeenschap.

De GPL v3 stelt echter strengere voorwaarden aan gebruik en distributie dan permissieve licenties. Dit kan integratie in bepaalde software-omgevingen of commerciële producten complexer maken.

3. Publicatie zonder open source licentie (transparantiepublicatie)
Een derde mogelijkheid is om de broncode wel openbaar te maken, maar zonder open source licentie. In dat geval blijft het volledige auteursrecht op de software van kracht. De code wordt zichtbaar gepubliceerd, bijvoorbeeld via een openbare repository, maar derden mogen deze niet gebruiken, aanpassen of verspreiden zonder toestemming van de rechthebbende. Publicatie zonder licentie kan worden gebruikt wanneer:

- De software zich nog in een experimentele fase bevindt;
- Nog geen besluit is genomen over structurele doorontwikkeling;
- De organisatie eerst transparantie wil bieden over de ontwikkelde oplossing;
- Juridische of organisatorische afwegingen rondom open source nog lopen.

Deze vorm van publicatie draagt bij aan transparantie over het bestaan en de werking van de software, zonder dat direct rechten voor hergebruik worden.

Verantwoordelijkheid voor publicatie
In overleg tussen de opdrachtgever en de uitvoerende partij is besloten dat het Ministerie van Sociale Zaken en Werkgelegenheid (SZW) verantwoordelijk is voor de publicatie van de broncode. Dit is passend omdat de software is ontwikkeld en toegepast binnen de context van SZW.

Beheer en onderhoud na afronding pilot
De in het kader van deze pilot ontwikkelde software betreft experimentele software die is ontwikkeld voor onderzoeks- en verkenningsdoeleinden. De software is niet aangemerkt als productievoorziening en maakt geen onderdeel uit van een structurele beheervoorziening binnen de organisatie. Na afronding van de pilot vindt vanuit de organisatie geen actief functioneel of technisch beheer meer plaats op de codebasis.

Er worden geen garanties gegeven ten aanzien van:

- Doorontwikkeling;
- Beveiligingsupdates;
- Compatibiliteit met toekomstige systemen;
- Ondersteuning of servicedeskfunctionaliteit.

Conclusie
De opdrachtgever en de uitvoerende partij zijn gezamenlijk tot de conclusie gekomen dat de broncode van de pilotsoftware openbaar wordt gemaakt onder een open source licentie. Publicatie onder een open source licentie sluit aan bij de uitgangspunten van de Wet open overheid en de Wet hergebruik van overheidsinformatie, en draagt bij aan transparantie over de inzet van publiek gefinancierde technologie.

De keuze voor de specifieke licentievorm, permissief (MIT) of copyleft (GNU GPL v3), wordt in een later stadium definitief vastgesteld, mede op basis van de gewenste mate van openheid voor afgeleide werken.

De verantwoordelijkheid voor publicatie ligt bij SZW, als de organisatie waar de software is ontwikkeld en toegepast. SZW beschikt op dit moment nog niet over de benodigde technische en organisatorische randvoorwaarden om publicatie op verantwoorde wijze te realiseren. Deze randvoorwaarden worden op korte termijn ingevuld, waarna publicatie via een openbaar toegankelijke repository zal plaatsvinden.

Bijlage IX: Registratie AI-toepassing

Binnen de pilot automatische e-mailclassificatie is kunstmatige intelligentie (AI) ingezet voor het automatisch classificeren van e-mails. De toepassing analyseert de inhoud van e-mailberichten en doet een voorstel voor classificatie, met als doel medewerkers te ondersteunen bij het ordenen en verwerken van e-mailverkeer. In het kader van transparantie over algoritmisch gebruik binnen de overheid is de vraag aan de orde gekomen of en hoe deze toepassing dient te worden opgenomen in het Algoritmeregister, en of de toepassing daarnaast publiek toegankelijk moet worden gemaakt. Hieronder wordt de gemaakte afweging toegelicht.

Risicoprofiel en registratieplicht

Het Algoritmeregister onderscheidt algoritmes op basis van risicoprofiel. Voor algoritmes met een hoog risico toepassingen die direct van invloed zijn op rechten, plichten of voorzieningen van burgers geldt een zwaardere verantwoordingsplicht en een sterke aansporing tot publieke transparantie. De AI-toepassing voor e-mailclassificatie kent een laag risicoprofiel. Daarvoor zijn de volgende redenen:

- De toepassing ondersteunt medewerkers bij het ordenen van e-mail; er is geen sprake van geautomatiseerde besluitvorming met rechtsgevolgen.
- Medewerkers behouden volledige controle over de uiteindelijke classificatie; voorstellen kunnen eenvoudig worden aangepast of overschreven.
- De toepassing heeft geen directe gevolgen voor rechten, plichten of voorzieningen van burgers of medewerkers.
- De inzet vindt plaats binnen bestaande kaders voor privacy, informatiebeveiliging en toezicht.

Gekozen richting: registratie in het Algoritmeregister, geen publieke publicatie

Op basis van het lage risicoprofiel is besloten de toepassing niet te registreren in het Algoritmeregister. Publieke publicatie in de zin van het breed toegankelijk maken van de toepassing zelf is voorbehouden aan toepassingen met een hoog risicoprofiel waarbij actieve publieke verantwoording is aangewezen. Dat is hier niet het geval.

Colofon

Programma	RDDI
Projectnaam	Pilot Automatische E-mailclassificatie
Versienummer	1.0
Projectleider	Suzanne van Poelgeest
Projectadviseur	Luuk Visser
Projectsecretaris	Salwa Hammoudi
	T +31 6 25723142
	Salwa.hammoudi@minocw.nl
	Rijnstraat 50 | Den Haag
	Postbus 16375 | 2500 BJ Den Haag
Auteur	Suzanne van Poelgeest