Auteurs: Sandra Kagie is freelance tekstschrijver (Sanscript Tekstproducties). Sandra is bereikbaar via info@sanscriptproducties.nl. Chris de Vries is redactielid van iB-Magazine. Hij werkt als zelfstandig professional onder de naam: De Vries Impuls Management. Chris is bereikbaar via impuls@euronet.nl.

INTERVIEW

Ongestructureerde data beveiligen: onderzoek naar dit hoofdpijndossier voor menig CISO

Maaike de Boer

Rick van der Kleij

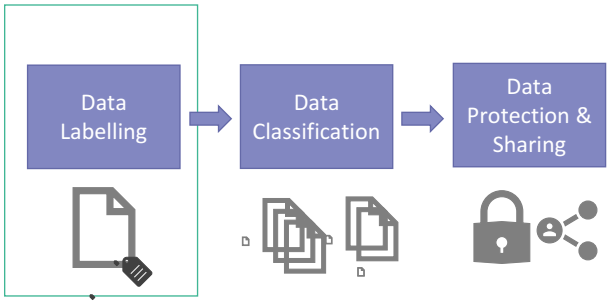
Binnen grote (financiële) organisaties circuleert een enorme hoeveelheid data, waarvan een groot deel ongestructureerd. Denk aan e-mails, audio- en videobestanden, maar ook allerhande tekstdocumenten. Deze bestanden en documenten zijn vaak niet geclassificeerd, maar kunnen vertrouwelijke informatie bevatten. Medische gegevens, fraude gerelateerde data of personal identifiable information (PII) bijvoorbeeld. Een hoofdpijndossier voor menig CISO. "Hier ligt een CISO wakker van", concludeerde het Partnership for Cyber Security Innovation (PCSI) (1) waarbinnen TNO en een aantal grote financiële instellingen als Achmea, ABN AMRO, ING en de Volksbank samenwerken om cybersecurity op een hoger plan te brengen.

Uit een ideation-sessie kwam het idee naar voren om machine learning (ML) toe te passen met betrekking tot de dreigingen van ongestructureerde data. De uitkomsten waren voor PCSI reden om eind 2020 een project te starten met als doel het zoeken naar de mogelijkheid om middels geautomatiseerde labelling als organisatie grip te krijgen op ongestructureerde data. Daarbij kwamen PII-, medische- en fraude gerelateerde data als eerste in beeld.

Waarom geautomatiseerde labelling?

Geautomatiseerde labelling omdat het handmatig etiketteren van ongestructureerde data zeer complex en tijdrovend is, waardoor het bijna onmogelijk is om de grote hoeveelheden ongestructureerde gegevens goed te labelen, vervolgens te classificeren en uiteindelijk te beschermen. Het proces dat binnen het project doorlopen wordt, kent vier stadia. Allereerst de explore-fase waarin geconstateerd werd dat de berg aan ongestructureerde data veelal niet op het gewenste niveau beveiligd was. Vervolgens de Proof of Concept-fase, gerelateerd aan Open Source data. Als derde de pilot-fase die al een werkend prototype

opleverde, kort cyclisch qua opzet en gebaseerd op een agile werkwijze. En tot slot de exploit-fase waarin het project zich nu bevindt.
De volgende afbeelding geeft een impressie van de gewenste werking van het prototype.



Figuur 1 - Van data-etikettering tot databescherming en -deling (2).

Begin dit jaar presenteerde PCSI zijn eerste conclusies in een whitepaper. Reden voor een gesprek met dr. Rick van der Kleij, senior research psychologist cybersecurity en projectleider namens TNO, en dr. Maaike de Boer, data-scientist bij TNO.

De centrale vraag

In het nog lopende project staat volgens Rick van der Kleij de vraag centraal of met kunstmatige intelligentie (KI/AI) op basis van machine learning er een hoge mate van betrouwbaarheid kan worden gerealiseerd als het gaat om het labelen van ongestructureerde data. "Komen er op deze manier betrouwbare labels tot stand op basis waarvan data geclassificeerd en uiteindelijk ook beter beschermd kunnen worden, zodat we uiteindelijk met z'n allen beter zicht kunnen houden op deze data?", specificeert Van der Kleij de onderzoeksvraag. Hij noemt de resultaten tot nu toe 'veelbelovend' met een 'nauwkeurigheid van meer dan 80%' ook als het gaat om meer complexe of gedetailleerde labels. Labels dus die verder gaan dan het etiket 'CV' of 'contract'. "Basis van het project zijn nu vooral tekst(document)en", legt Maaike de Boer uit. In dit soort documenten is de woordvolgorde belangrijk om tot een juiste herkenning en daarmee classificatie te komen."
Ze benadrukt verder dat er binnen het project gebruik wordt gemaakt van de bredere definitie van Artificial Intelligence. "In de nauwe definitie (strong AI) leert de 'robot' zoals de mens en neemt deze het proces geheel over. Terwijl de bredere definitie (weak AI) uitgaat van bijvoorbeeld het goed uitvoeren van één taak, waarbij in dit geval een systeem wordt gevoed met documenten op basis waarvan een kansberekening, geavanceerder dan statistiek, plaatsvindt om vast te stellen of een document bijvoorbeeld PII-gegevens bevat, een CV of een medisch document is. Om hier vervolgens een bijbehorende classificatie en beveiliging aan te koppelen."

Kansen nieuwe methodiek

Het doel van het project is volgens Van der Kleij ook om met een flexibele en schaalbare aanpak te komen, zodat er in de loop der tijd labels kunnen worden toegevoegd. Daarnaast biedt de nieuwe methodiek organisaties volgens hem betere mogelijkheden om transparanter te kunnen communiceren en (vertrouwelijke) informatie te delen.
De onderzoekers stellen in hun whitepaper dat er met dit doel weliswaar diverse tools en pakketten op de markt zijn, maar dat er nog vele (ontwikkel)uitdagingen bestaan die grootschalige toepassing van geautomatiseerde data labelling binnen (financiële) organisaties in de weg staan. Er zijn ten aanzien van het PCSI-prototype vijf waardeproposities te onderscheiden, te weten:

- Nauwkeurigheid
- Flexibiliteit
- Complexiteit
- Granulariteit
- Uitlegbaarheid

Voor een uitgebreide toelichting op de vijf punten zie de whitepaper (3).
Juist financials hebben in de woorden van de TNO-onderzoekers vaak 'net wat meer nodig' op de genoemde vijf waarden. Dit omdat de financiële sector rekening heeft te houden met de eisen van toezichthouders zoals De Nederlandsche Bank en de Autoriteit Financiële Markten. "We vertrouwen erop met ons onderzoek marktpartijen te stimuleren dat stapje extra te zetten", stelt Van der Kleij. Organisaties zitten vaak niet te wachten op nóg een tool zo blijkt volgens hem uit diverse gesprekken met zowel organisaties binnen PCSI als daarbuiten. "Ze zijn veel meer geïnteresseerd in een verbetering op de voor hen cruciale punten van de tools van bijvoorbeeld Microsoft en Proofpoint die ze nu gebruiken."

De Boer geeft aan dat het probleem dat automatische labelling oplost, ook nadrukkelijk een menselijk aspect omvat. Namelijk de belasting van de mens. Handmatig labelen is namelijk een tijdrovende en intensieve taak. De Boer: "Geautomatiseerde labelling voorkomt fouten die ondanks goede bedoelingen van medewerkers ontstaan.

Het ontlast medewerkers en het voorkomt de inwerkingtreding van het bekende adagium: 'Garbage in, garbage out!'"

"Onze methodiek van geautomatiseerde labelling is een semi-supervised methode", gaat ze verder. "Door menselijke terugkoppeling leert het systeem en zorgt het voor vastlegging en vorming van noodzakelijke trainingen binnen een organisatie. Aan de werking hiervan gaat een discussie met materiedeskundigen vooraf. Waarin ze kunnen aangeven wat voor hen belangrijk is en wat zij als essentiële definities zien. Dat betekent dat per sector verschillen kunnen en mogen bestaan, ook in geval van gelijknamige begrippen."

Uitnodiging aan marktpartijen

De TNO-onderzoekers zien hun eerste conclusies nadrukkelijk als een uitnodiging naar marktpartijen, sectorpartijen en vendors, om met elkaar in gesprek te gaan. "We nodigen leveranciers en andere geïnteresseerden uit om samen te bekijken hoe we onze bevindingen in de (inter)nationale praktijk kunnen brengen. Dit niet alleen binnen de financiële sector, maar juist ook in de bredere security community. Zodat we samen antwoorden kunnen vinden op dit vraagstuk waarvan een CISO wakker ligt", besluit Van der Kleij.

Referenties

(1) https://pcsi.nl/

(2) Whitepaper PCSI, 'Protecting unstructured data – challenges and opportunities of automated labelling

(3) https://pcsi.nl/news/protecting-unstructured-data-challenges-and-opportunities-of-automated-labelling/

In onderstaande interviews geven ABN AMRO en Achmea weer hoe zij omgaan met de classificering van ongestructureerde data binnen hun organisaties.

ABN AMRO verwacht PCSI-model binnen een jaar toe te passen

"Het probleem van het classificeren van grote hoeveelheden ongestructureerde data binnen organisaties bestaat al heel lang. Er zijn een aantal commerciële oplossingen op de markt, maar deze lossen het probleem niet écht op. We hebben gegevens waarvan we geacht worden dat we er zorgvuldig mee omgaan of waarvan we zelf willen weten waar ze blijven, maar in het geval van ongestructureerde data is er geen gemakkelijke manier voor ons om die gegevens op te sporen. We zoeken een oplossing aan de voorkant om de data te kunnen vinden om deze aan de achterkant te kunnen beveiligen."

Noor Spanjaard

Olaf Streutker

Deze aftrap doet Olaf Streutker, Strategisch Adviseur bij het Corporate Information Security Office van ABN AMRO Bank. We spreken hem en zijn collega Noor Spanjaard, Enterprise Data Governance Adviseur bij ABN AMRO Bank, over het Automated Data Labelling project. De bank is als partner binnen PCSI een van de deelnemers aan het project.

"De documenten zoals word-bestanden, excel-sheets en pdf's, waar het om gaat, kunnen echt door iedereen binnen de organisatie worden gemaakt met allerlei verschillende doeleinden", verduidelijkt Spanjaard. "Terwijl gestructureerde informatie vaak via een applicatie, via een front-end, op een bepaalde manier en met een bepaald doel wordt ingevuld. Veel makkelijker van tevoren te classificeren dus, omdat er sprake is van een duidelijke doelkoppeling. Een koppeling die in het geval van ongestructureerde data ontbreekt."

Onderscheid labelen en classificeren

Heel veel oplossingen die nu op de markt zijn, richten zich volgens Spanjaard op het classificatieprobleem, maar dan mis je in haar woorden 'de granulariteit van het labelen'. Precies de reden waarom in het PCSI-onderzoek volgens Streutker zo nadrukkelijk het onderscheid wordt gemaakt tussen labelen, 'het objectief vaststellen of een document bijvoorbeeld persoonsgegevens bevat', en classificeren, 'in welk bakje stop ik het en hoe moet ik het beschermen, een subjectieve beoordeling'.

"Voor bedrijven die niet zo streng gereguleerd zijn als een financiële instelling kan zo'n classificatieoplossing prima werken", vervolgt zijn collega. "Maar juist wanneer je heel specifieke regelgeving hebt waaraan je moet voldoen dan blijkt het in de praktijk onvoldoende te werken." Een generiek probleem, zo heeft de bank in zijn zoektocht naar de oplossing geconstateerd.

Lesson learned

Het meest in het oog springende dat ABN AMRO tot nu toe heeft geleerd van dit PCSI-project is volgens Spanjaard dat het begrip van de voorwaarden waaronder Machine Learning (ML-)technieken goed kunnen werken, binnen de organisatie niet wijd verspreid is. "Dit maakt het lastig omdat je voor een succesvolle toepassing van deze techniek vooraf een klein beetje moet investeren voordat je de vruchten kunt plukken", legt ze uit. De investering behelst het annoteren van gegevens in de systeemdataset. Je moet een aantal voorbeelden geven van hoe jij informatie wilt labelen en vervolgens kan het systeem die labelling veel breder gaan toepassen op alle documenten. Je moet dus een klein voorzetje geven, voordat het algoritme zelf verder kan leren."

"Probleem is nu dat mensen binnen de organisatie de urgentie van het probleem weliswaar zien, maar dat ze door een gebrek aan begrip van de techniek geen of te weinig tijd vrij maken voor die kleine investering wat betreft het annoteren van data."

Het grote voordeel van het PCSI-model is volgens beiden juist dat je niet langer afhankelijk bent van de individuele inschatting van medewerkers die hun documenten moeten classificeren, maar dat het model zelf steeds beter wordt in het betrouwbaar labelen van documenten. Het aantal labels dat je potentieel kunt toekennen aan documenten is onbegrensd."

Potentie PCSI-model

Wanneer je de basis van labelling goed op orde hebt, kan het PCSI-model volgens Streutker veel breder worden ingezet dan voor security-doeleinden alleen. Een goed gelabeld document maakt volgens hem namelijk meerdere classificaties naast elkaar mogelijk. "Een security classificatie is wat anders dan een business classificatie of een krediet classificatie."

Voor een concrete implementatie binnen de bank van het PCSI-model is het volgens Streutker nog net wat te vroeg. Die opschaling moet volgens hem nog komen. "We zitten nog in de fase waarin we aantonen dat het werkt om het probleem van ongestructureerde data op te lossen met behulp van dit model", geeft hij aan. "Gelukkig zijn we een stuk verder dan in 2013, toen het alleen nog maar in een research omgeving mogelijk was om experimenten uit te voeren", vult Spanjaard aan. Toen bleek de implementatie in de organisatie van zo'n ML-oplossing volgens haar te moeilijk.

"Daar waren destijds te veel expertise en te grote investe-

ringen voor nodig. Wat we nu samen met TNO hebben ontdekt is dat met een kleine effort dit model al getraind kan worden. Het loopt nu alleen nog vast op het op weg helpen van het model met de juiste trainingsdata. Wat we wel hebben gecheckt is dat onze IT-organisatie dit model lokaal kan draaien. Qua volwassenheid van de techniek binnen onze financiële organisatie zijn we de afgelopen jaren enorm gegroeid."

Toepassing binnen een jaar

Streutker verwacht dan ook dat het model binnen ABN AMRO binnen een jaar daadwerkelijk gebruikt gaat worden. "De voorwaarden daarvoor zijn aanwezig", weet hij. Voor organisaties met een vergelijkbaar volwassenheidsniveau zien hij en Spanjaard zeker ook mogelijkheden om concrete data beveiligingsproblemen aan te pakken. Waarbij ze nadrukkelijk wijzen op overheidsorganisaties en andere financiële instellingen.

Omdat zij net als ABN AMRO beschikken over een goed toegerust data science team. "Een voorwaarde omdat het toepassen van kunstmatige intelligentie (KI) in informatiebeveiliging kansen biedt, maar ook risico's met zich meebrengt", waarschuwt Spanjaard. "Risico's van misbruik. Om dat te voorkomen heb je wel een bepaald volwassenheidsniveau op het gebied van KI binnen je organisatie nodig." Organisaties die hierover niet beschikken adviseert ze daarom te kiezen voor een oplossing van een van de partijen die al beschikbaar is, 'off the shelf' dus.
"Dit model maakt het mogelijk dingen die je kwijt bent te vinden én te beveiligen", concludeert Streutker. Daar moet iedere informatiebeveiliger volgens hem warm van worden. "Zeker als je kijkt dat de labels uit ons model ook uitleesbaar moeten zijn voor andere security protection tooling", haakt Spanjaard in. Om zo een vendor lock-in te voorkomen. De twee benadrukken dat ze hopen dat de bevindingen uit het huidige PCSI-onderzoek ontwikkelaars van off the shelf-oplossingen inspireren hun pakketten te verbeteren. Vooral op het vlak van flexibiliteit en granulariteit. "Ze beloven de oplossing voor alle problemen te zijn als het gaat om ongestructureerde data. Maar wanneer je doorvraagt en verder kijkt, zijn ze dat voor een organisatie als de onze nét niet helemaal."

Verdere toekomst

In een utopische wereld tot slot ziet Streutker ook kansen voor het PCSI-model als het gaat om het attribute based access control waarin je op basis van attributen toegang verleent. "Labels zijn attributen. Je zou dus kunnen zeggen wanneer ik meer gegranuleerde labels kan toekennen, kan ik ook betere beslissingen nemen over wie er wel of niet bij bepaalde informatie mag."

'We hebben als Achmea al gigantische vooruitgang geboekt'

Michaël Stekkinger

Achmea is als partner binnen PCSI een van de deelnemers aan het project Automated Data Labelling. "We beschikken bij Achmea, zoals vrijwel alle moderne organisaties, over een enorme hoeveelheid data. Lang niet allemaal geordend of gestructureerd binnen een bepaalde applicatie of rond een naam of andere persoonsgegevens die nodig zijn voor het afsluiten van een verzekering of het organiseren van een afhandeling bij schade. Het is nogal een taak om die stroom aan data goed en veilig te organiseren", legt Michaël Stekkinger, Information Security en Compliance Officer bij Achmea, de urgentie van het project uit.

"Bij Achmea staan we voor duurzaam samen leven."

Waarbij het volgens hem een belangrijk punt is dat je als organisatie wilt weten welke gevoeligheid een bepaald document heeft binnen je organisatie. Zodat je zaken die extra gevoelig liggen, ook extra kunt beveiligen. Dit in samenhang met eisen van het Information Rights- en Data Loss Management. Zodat je in zijn woorden 'tegenstanders het voortouw kunt ontnemen'. Juist dit inzicht verkrijgen in die enorme hoeveelheid aan ongestructureerde data noemt Stekkinger 'een noodzakelijk kwaad en niet zo gemakkelijk'.

"Niet zo gemakkelijk omdat we hierin, tot op heden, grotendeels afhankelijk zijn van medewerkers", gaat hij verder. "Zij moeten documenten classificeren en op de juiste plek opslaan: stelselmatig en met het juiste label", legt hij uit. "Een kennisintensief proces dat ook tijd en aandacht vergt. En hoe welwillend en getraind medewerkers ook zijn. Het gaat wel eens mis."

Behoefte aan een zelflerend algoritme

Als problemen waar je tegen aanloopt, noemt hij bijvoorbeeld data die in het verleden al opgeslagen zijn en zelden tot nooit geraadpleegd worden, data die niet gelabeld zijn en data met vergelijkbare labels, maar in verschillende talen. Dit alles vereist volgens Stekkinger een algoritme dat multi-lingual is, dat de mogelijkheid biedt eigen labels toe te voegen en dat het toestaat dat een label binnen de ene organisatie een andere betekenis heeft dan binnen een andere organisatie. "Dit zijn gezamenlijke eisen vanuit de deelnemende financiële dienstverleners. Complexe materie die vraagt om de inzetbaarheid van een zelflerend algoritme", concludeert hij.

Randvoorwaarden die hierbij voor Achmea een belangrijke rol spelen zijn onder meer:

- compliance vereisten (waaronder uitlegbaarheid van de gekozen labels door een algoritme)
- bruikbaarheid in een complexe IT-omgeving en niet gelimiteerd aan één vendor/product.

Kan het niet efficiënter?

Kernvraag voor Achmea binnen het project is nu uit te zoeken of het labelen van ongestructureerde data efficiënter en georganiseerder kan door de inzet van een zelflerend algoritme. "Wetende dat dit een complexe zaak is, dat het accuraat moet gebeuren en dat resultaten ook uitlegbaar moeten zijn", benadrukt Stekkinger.

Hij ziet voor Achmea nu al, terwijl het project nog loopt, een 'gigantische vooruitgang'. "We weten meer en we kunnen meer", geeft hij aan. Zo kunnen we in de toekomst dankzij automatische labelling middels een zelflerend algoritme meer duidelijkheid creëren in ongestructureerde data. Daarbij kunnen we medewerkers ontlasten in het complexe proces zodat ze meer tijd beschikbaar hebben voor andere taken. En we zien ook nog eens mogelijkheden om eenzelfde soort algoritme te gebruiken om andere uitdagingen te tackelen, zoals data retentie."

Wat hij ook vooral waardeert, is het innovatieve aspect van het onderzoek binnen PCSI met als facilitator TNO. "Je ziet hierdoor dat bijvoorbeeld ook bepaalde vendors interesse hebben in onze onderzoeksresultaten. Waardoor je beweging ziet in het hele veld. Een ontwikkeling waar uiteindelijk niet alleen wij en andere financials binnen PCSI profijt van hebben, maar iedereen."

Hoog op agenda

"Bij Achmea staan we voor duurzaam samen leven", stelt hij tot slot. Hierbinnen lossen we samen met onze klanten, strategische partners en relaties grote maatschappelijke vraagstukken op rond gezondheid, wonen & werken, mobiliteit en inkomen. Dan moet je zorgen dat je meegaat in digitale ontwikkelingen en dat je datahuishouding op orde is. Je wordt geen digitale verzekeraar zónder dat dit hoog op de agenda staat. En dan doel ik op het allerhoogste niveau: bij de Raad van Bestuur. Een probleem als ongestructureerde data tackelen, staat daarom als vanzelfsprekend hoog op onze IT-agenda. Heel mooi dus dat we deze uitdaging binnen PCSI samen met partners/peers met een gelijke focus kunnen oppakken."