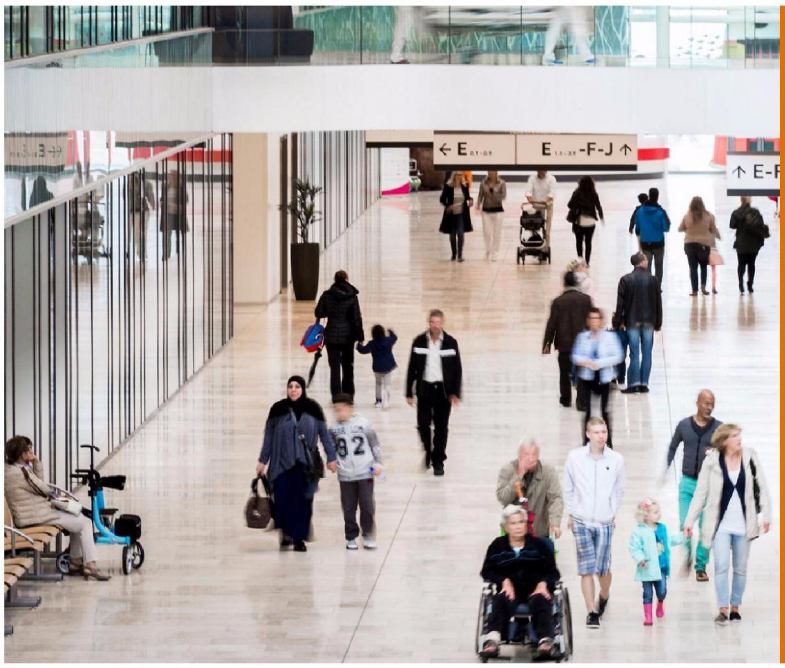
10034673



Voorspelmodel

Wat weten we nu

- Voorspelmodel heeft accuraatheid van 86% voor het voorspellen van VOV en VNOV
- Op 100 documenten worden 86 documenten hetzelfde beoordeeld als een jurist heeft gedaan
- De vraag is heeft een jurist het zelf goed beoordeeld?

Geconstateerd:

- In dataset zijn er identieke documenten op verschillende matters die inconsistent beoordeeld zijn.

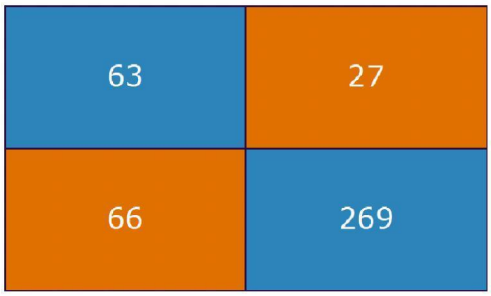
Om te kunnen valideren hoe goed het model presteert, is het noodzakelijk te weten hoe consistent een jurist een document beoordeeld.

Wat hebben we gedaan

- volkskrant special (425 documenten) voorspelt met het Voorspelmodel
- 5.1.2e heeft standaard quickscan gedaan zonder naar predicties te kijken
- 5.1.2e kon alleen predicties inzien en bekeken of hij het eens was met de predictie van het model

Voorspelmodel vs. 5.1.2e

- 425 documenten
- 332 goed voorspeld
- 93 fout voorspeld



(332/425) * 100 = 78% accuracy

5.1.2e VS. 5.1.2e

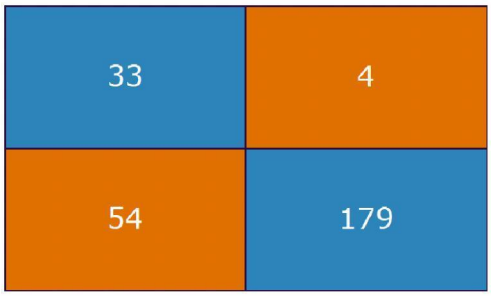
- 303 documenten
- 33 documenten anders beoordeeld

"De voorspelling "valt onder verzoek" bij omgevingsanalyses / overzichten van sociale media / nieuwsberichten. De begeleidende- / vervolgmails van omgevingsbeelden tag ik VNOV, al vermoed dat ik collega's hier ook 'administratief' taggen."

- Zonder administratief 21

Voorspelmodel vs. (5.1.2e & 5.1.2e)

- Alleen consistent beoordeelde documenten (270)

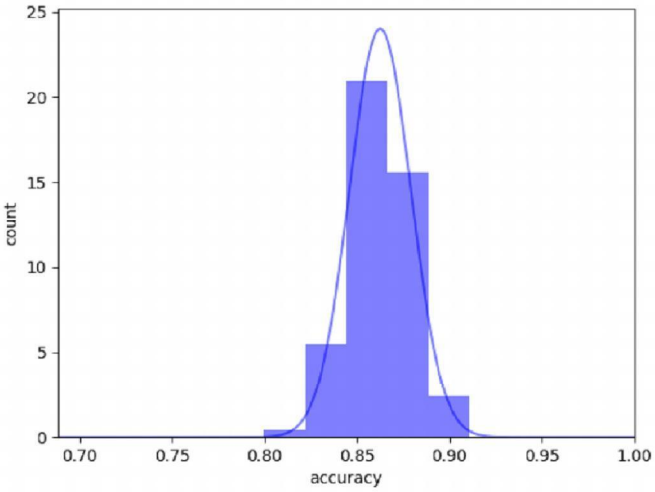


(212/270) * 100 = 78% accuracy

Konden we dit verwachten?

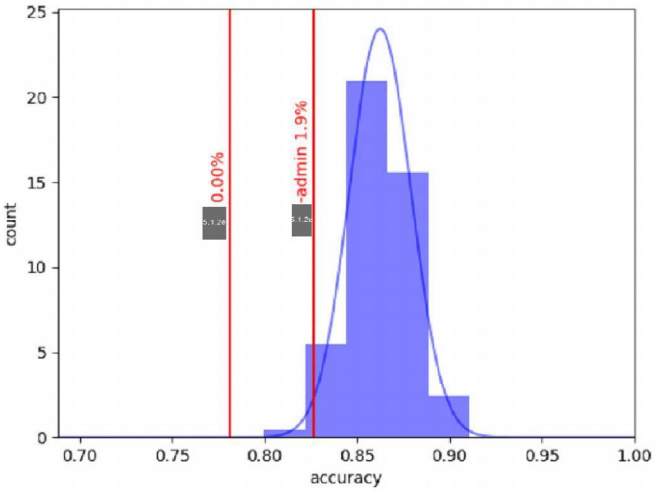
- Over de hele test set hadden we 86% goed

- 20.000 samples
- 425 documenten per sample



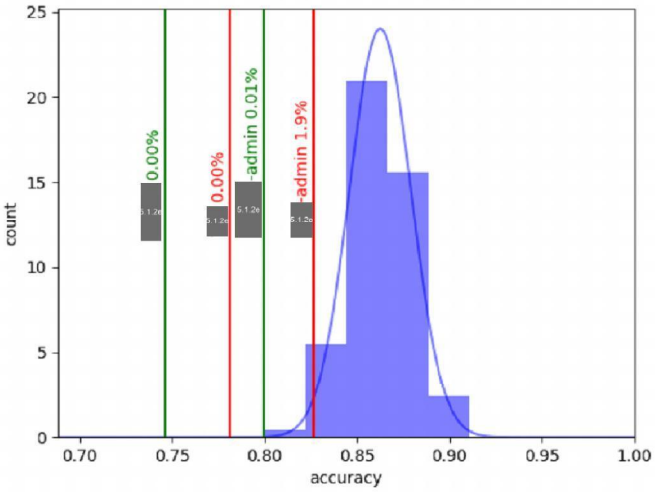
Konden we dit verwachten?

- Over de hele test set hebben we 86% goed
- Volgens 5.1.2e hebben we 78% goed
- Dit kwam in geen van de samples voor
- Zonder administratief scoren we 82%, dit komt in 1.9% van de samples voor



Konden we dit verwachten?

- Over de hele test set hebben we 86% goed
- Volgens [redacted] hebben we 75% goed
- Dit kwam in geen van de samples voor
- Zonder administratief scoren we 79%, dit komt in 0,1% van de samples voor



Dit is niet wat we hadden verwacht

- De Volkskrant special lijkt niet te lijken op waar wij op getrained en getest hebben

- Er zijn dubbele documenten, model maakt meerdere keer dezelfde fout
- We houden geen rekening met de *administratief* tag

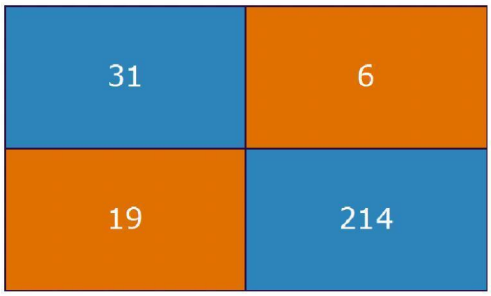
Wat doen we met *adminstratief?*

Opties:

1. Voorspellen als VNOV
2. Voorspellen als VOV
3. Voorspellen als VNOV. Als doc=VNOV en in family VOV -> Administratief

Voorspelmodel2 vs. ([redacted] & [redacted])

- Alleen consistent beoordeelde documenten (270)

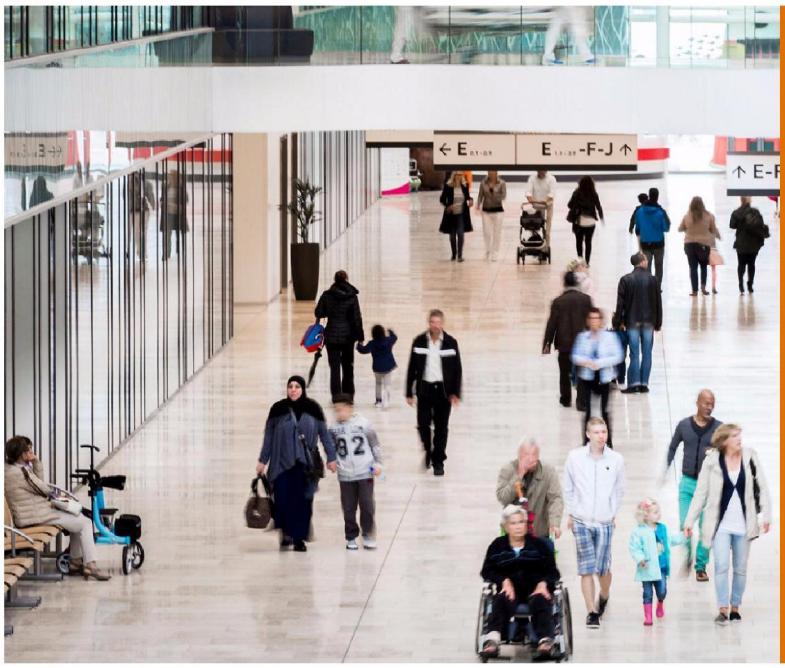


(1 - 25/270) * 100 = 91% accuracy
Waaronder herhaalde fouten

Toekomst

Om te kunnen valideren hoe goed het model presteert, is het noodzakelijk te weten hoe consistent een jurist een document beoordeeld.

Zo kunnen we beter inschatten hoe goed het model presteert en in hoeverre het te verbeteren is.



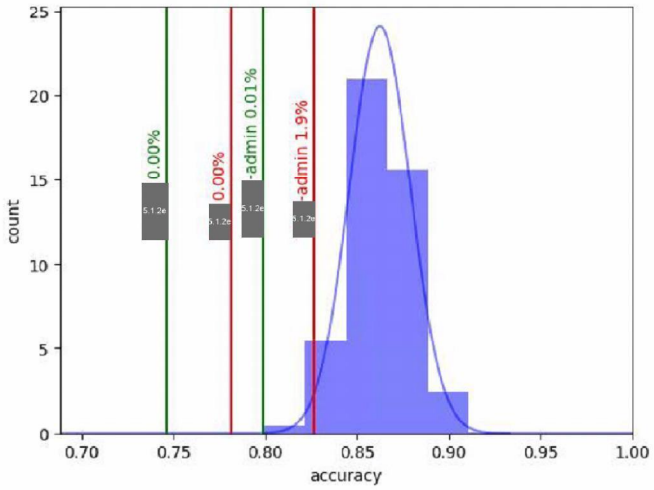
Experiment



1 sample naar meerdere juristen

1. Hoe groot moet het sample zijn?
2. Hoeveel juristen hebben we nodig?
3. Hoe maken we het sample?

Met een Monte Carlo simulatie van 426 documenten



Hoe groot moet het sample zijn?

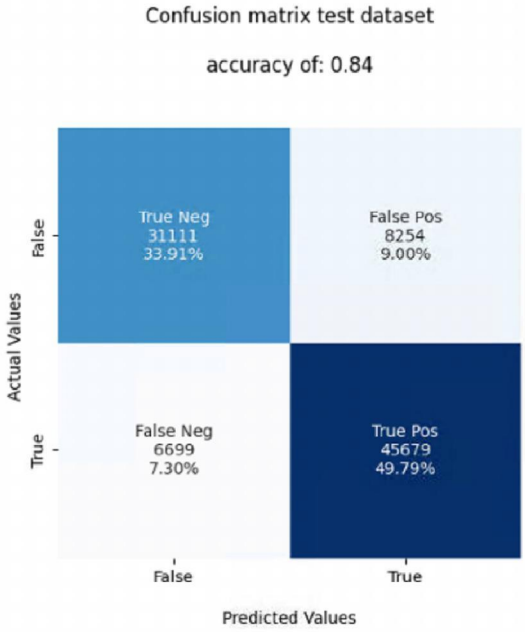
Hoeveel juristen hebben we nodig

Distinctie validatiesets

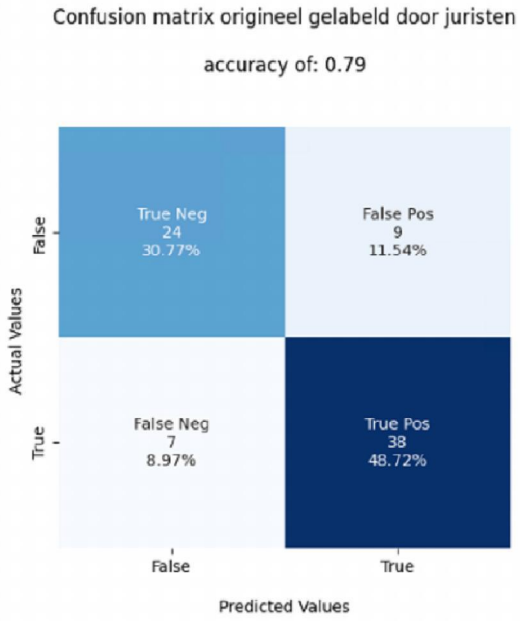
- Test-set:
  - Model is getraind op 80% van de gelabelde documenten en getest op 20% van de gelabelde documenten
  - Ieder document is door één jurist gelabeld.
  - ~80k documenten

- Sample-set:
  - 35 juristen beoordelen 78 dezelfde documenten
  - Deze set is representatief voor de test-dataset
    - Zelfde verhouding labels
    - Zelfde verhouding matters

Prestatie van Model op huidige test-set

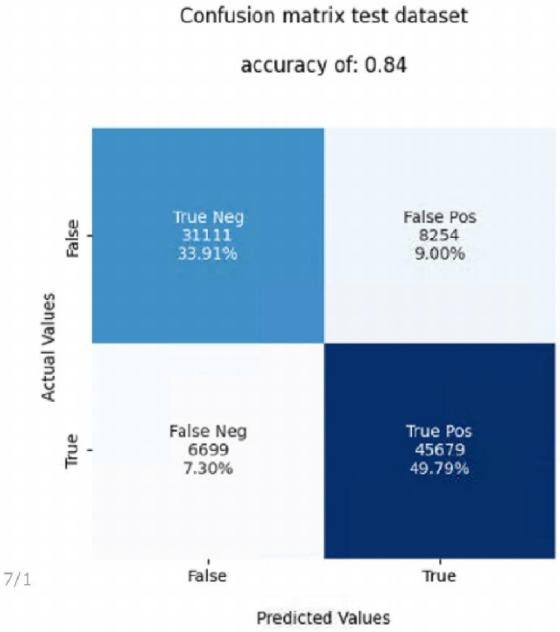


Prestatie van Model op Sample set

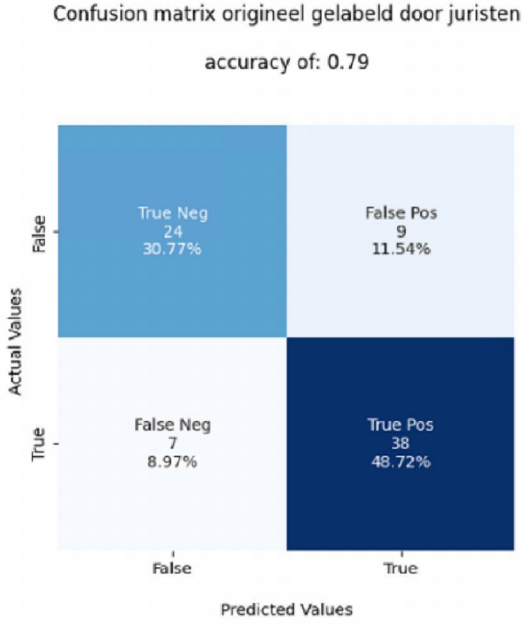


- Dit zijn de labels die ooit zijn gegeven door een jurist. Ieder document is dus één keer gelabeld
- Accuraatheid van 79 procent
- 7 documenten valt niet onder verzoek gelabeld die eigenlijk onder de wob vallen (valse negatieven)

Test-set



Sample-set

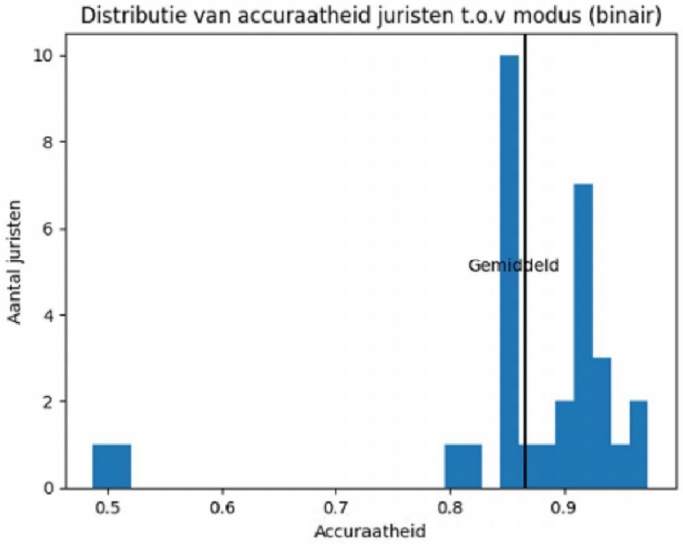


Consistentie Juristen onderling

- Ieder document is 31 keer beoordeeld
- Voor ieder document is gekeken hoe vaak ieder label is gegeven.
- Met deze informatie heeft 5.1.2e een "ground truth" opgesteld. Dit is dus het definitieve antwoord. Aan de hand van de "ground truth" kunnen we het model opnieuw vergelijken en individuele juristen testen.

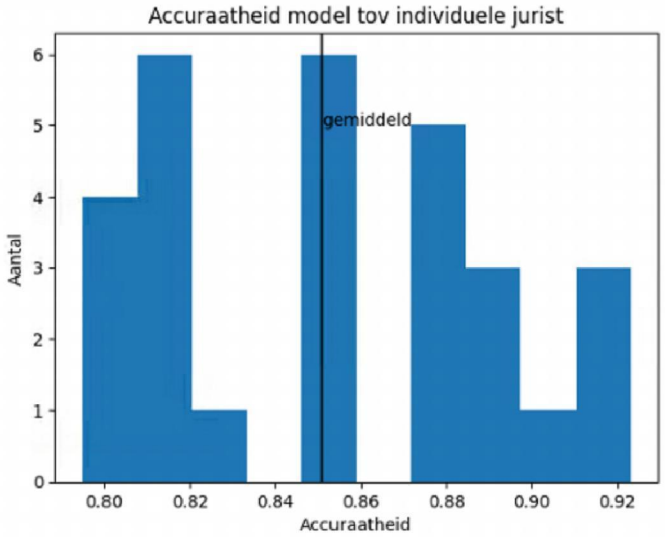
	id	0	1	2	3	4	5	6	7	8	...	26	27	28	29	30	VNOV	VOV	ADMIN	modus	ground_with_admin
0	31923613	2.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	26.0	3.0	2.0	26.0	0
1	11163934	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	27.0	4.0	0.0	27.0	0
2	6119438	1.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	...	0.0	0.0	0.0	0.0	0.0	27.0	2.0	2.0	27.0	0
3	16154038	1.0	0.0	1.0	2.0	1.0	2.0	0.0	2.0	2.0	...	2.0	1.0	1.0	1.0	2.0	8.0	15.0	8.0	15.0	1
4	6922296	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	1.0	0.0	0.0	1.0	0.0	27.0	3.0	1.0	27.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
73	12809033	2.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	2.0	...	0.0	0.0	0.0	0.0	0.0	25.0	0.0	6.0	25.0	0
74	12509339	1.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	2.0	...	1.0	0.0	0.0	0.0	0.0	25.0	3.0	3.0	25.0	0
75	23294441	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	...	2.0	2.0	0.0	2.0	2.0	1.0	2.0	28.0	28.0	2
76	4683347	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	1.0	2.0	1.0	1.0	1.0	2.0	26.0	3.0	26.0	1
77	1620276	2.0	0.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	...	2.0	2.0	2.0	0.0	2.0	2.0	1.0	28.0	28.0	2

Outliers en imputatie



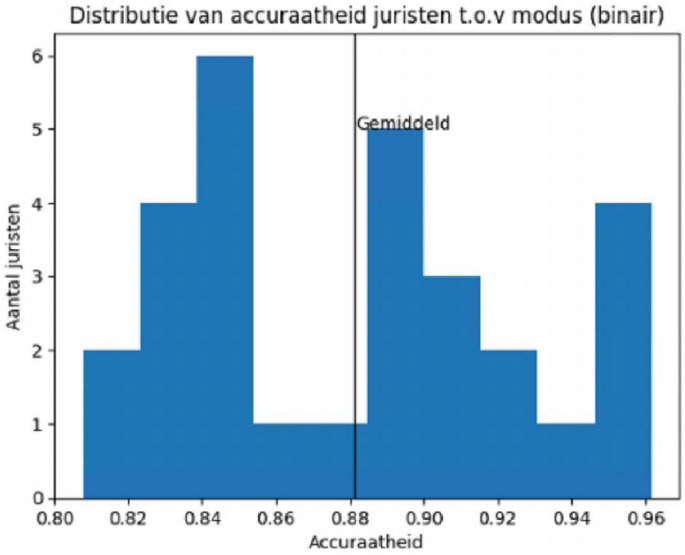
- Twee juristen scoren opvallend slecht, deze zijn in de rest van de grafieken niet meegenomen.
- Één jurist heeft één document niet geïnventariseerd. Hierbij zijn we ervanuit gegaan dat de jurist het goede antwoord had gegeven.

Model t.o.v individuele jurist



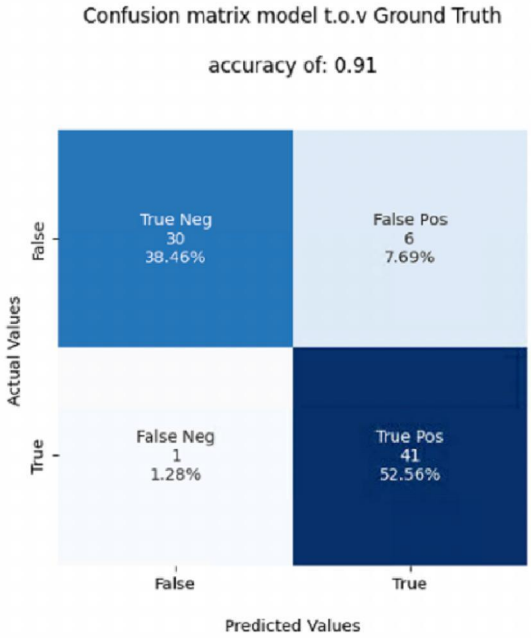
- Als je het model test op iedere individuele jurist zou je verwachten dat je eenzelfde accuraatheid krijgt als de test dataset.
- Gemiddelde accuraatheid is 85% dus zit inderdaad heel dichtbij

Individuele juristen t.o.v “ground truth”



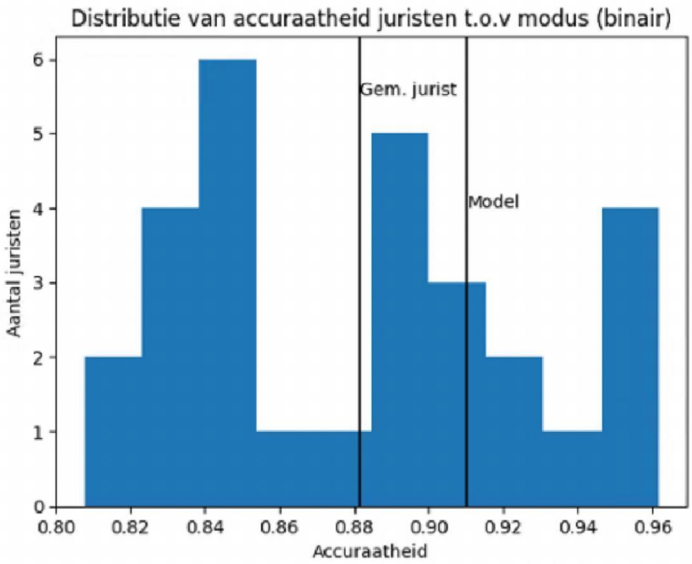
- Voor iedere jurist is de accuraatheid ten opzichte van de ground truth bepaald
- Zwarte streep is de gemiddelde accuraatheid van de juristen
- Gemiddeld labelt een jurist 88% procent van de documenten correct.

Model t.o.v Ground Truth



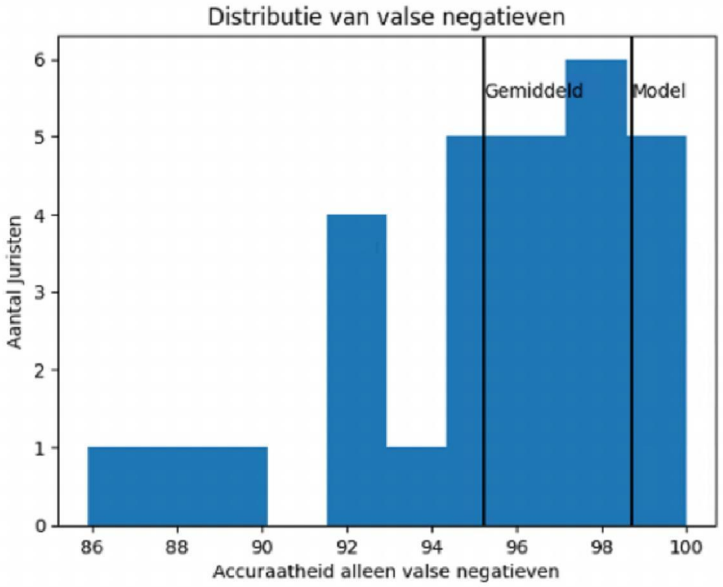
- Het model wordt in deze confusion matrix vergeleken met deze ground truth.
- Model heeft één valse negatieve, dit document had administratief moeten zijn.

Model t.o.v Ground Truth



- Model presteert in dit geval iets beter dan de gemiddelde jurist

Vergelijking Valse Negatieven

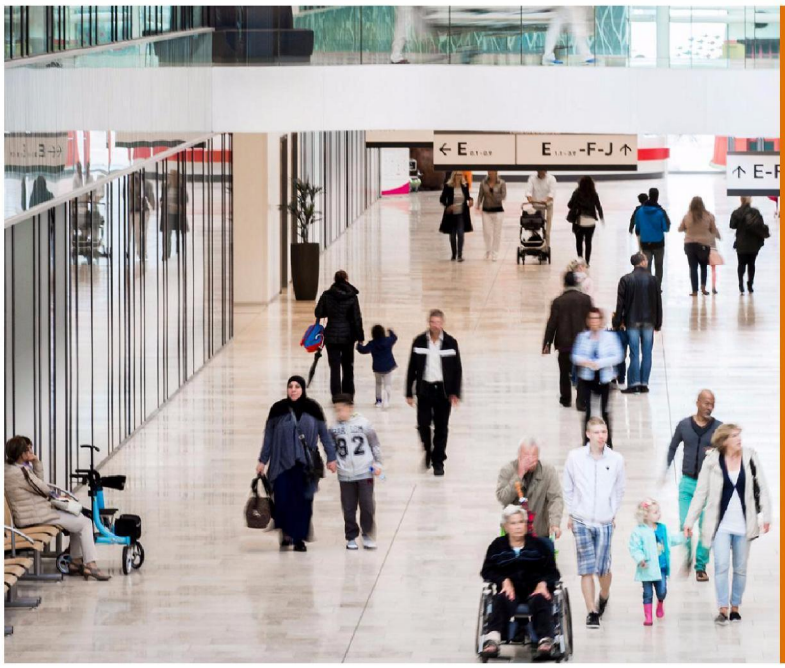


- Doel is het reduceren van valse negatieven, dit zijn documenten gelabeld als Valt niet onder verzoek terwijl deze Valt onder verzoek had moeten zijn.
- Distributie zijn juristen ten opzichte van Ground Truth
- Verticale lijn voor Model ten opzichte van Ground Truth

Conclusie

- Model presteert vergelijkbaar met een gemiddelde jurist

10034673

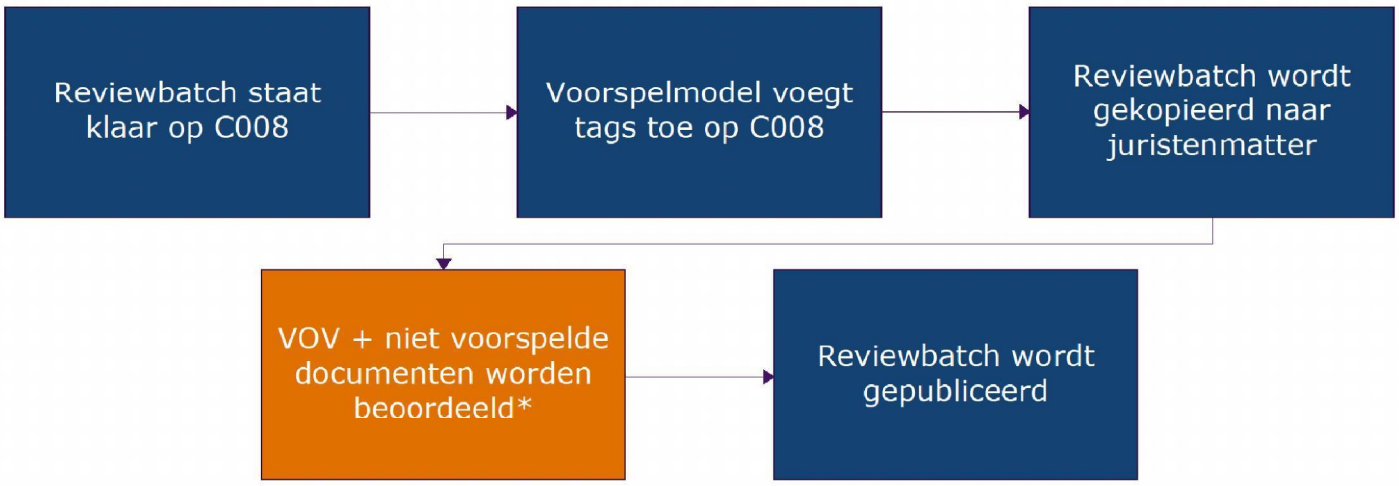


Monitoring

Monitoring

- Er is geconstateerd dat het model vergelijkbaar presteert met een gemiddelde jurist op de huidige dataset
- Het is belangrijk te blijven monitoren of het model nog steeds presteert zoals verwacht wordt.

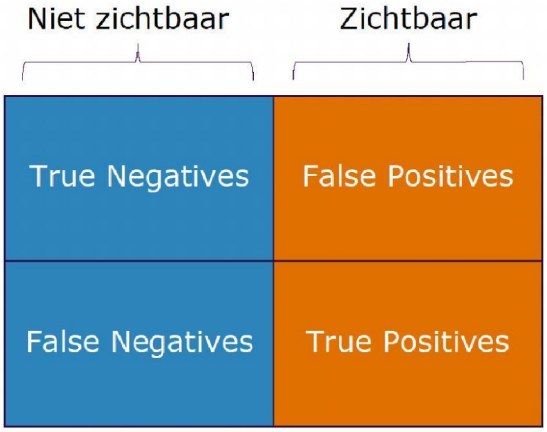
Voorspelmodel Flow



*Na de beoordeling zal validatie plaatsvinden

Welke documenten krijgen een jurist te zien bij de beoordeling?

- Te zien door de jurist zijn documenten die het model 'valt onder verzoek heeft voorspeld'
- Hier zitten dus documenten in die correct zijn voorspeld (True Positives) en documenten die onterecht 'valt onder verzoek' zijn voorspeld (False Positives)
- Na een boordeling zullen twee verschillende validaties plaatsvinden:
  - **Validatie van de 'False Positives'.** Meteen na de beoordeling is te zien hoeveel documenten na het bekijken door een jurist toch een 'valt niet onder verzoek' label hebben gekregen
  - **Validatie van de 'False Negatives'.** Deze documenten hadden eigenlijk het label 'valt onder verzoek' moeten hebben en zijn niet bekeken. Door middel van een steekproef wordt bekeken of het aantal documenten die worden gemist naar verwachting is.



Validatie False Postives

- De beoordeling wordt nog steeds gedaan door één jurist.
- De labels van de testdataset waarop het model is getrained zijn ook gegeven door één jurist
- De verwachting met een threshold van 0.6 is als volgt
- Reviewbatches zijn over het algemeen een stuk kleiner, hier moet rekening mee worden gehouden

