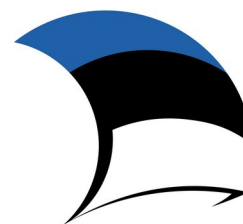


TEXTA



Euroopa Liit  
Euroopa  
Regionaalarengu Fond



Eesti  
tuleviku heaks

# Kodaniku pöördumiste klassifitseerija analüüs ja pilootprojekt: detailanalüüs

Majandus- ja Kommunikatsiooniministeeriumi riigihange  
„Kodaniku pöördumiste klassifitseerija analüüs ja pilootprojekt”

Riigihanke objekt: Tarkvara arendusteenused  
(kood: 72262000-9)

Autorid:

Linda Freienthal  
Raul Sirel  
Mart Laur  
Hele-Andra Kuulmets  
Silver Traat

# Sisukord

<b>Sissejuhatus</b>	<b>4</b>
<b>Terminoloogia</b>	<b>5</b>
<b>Asutuste vajaduste kaardistamine</b>	<b>8</b>
Pöördumiste kanalid ja mahud	8
Huvi juturobotite vastu	10
Pöördumiste keeleline osakaal	11
Valele asutusele saadetud kirjade arv	13
Pöördumiste menetlemine	14
Klienditeenindajate arv	14
Pöördumise menetlemise ajakulu	15
Pöördumiste hoiustamise keskkonnad	15
Ühe pöördumise maksumus	16
Kirjade märgendamise hetkeolukord	16
Nõuded pöördumiste statistikale ja klassifitseerija(te)le	19
<b>Andmed</b>	<b>23</b>
Ülevaade andmetest	23
Andmete puhastamine ja eeltöötuse juhend	29
<b>Asutusteülene klassifitseerimine</b>	<b>32</b>
Lineaarsed klassifitseerijad	32
Eraldatud fraaside lisamise mõju meili pöördumiste ennustamisele	34
Stoppsõnad	38
Klassi tasakaalutus	40
Sõnavektorite lisamine	42
Lineaarsed klassifitseerijad juturobotite andmestikul	43
Tehisnärvivõrkudel põhinevad klassifitseerijad	45
Teksiklassifitseerija <code>fastText</code>	45
Transformeritel põhinevad klassifitseerijad	47
<b>Asutusesisene klassifitseerimine</b>	<b>48</b>
Soovitused klassifitseerija treeningandmete koostamiseks	48
Märgendatud andmetel klassifitseerimine	49
Solvangute tuvastamine veebikommentaariidel treenitud mudeliga	52
Märgendamata andmetel klassifitseerimine ehk klasterdamine PPA-s	54
Andmestik	54
Klastrite loome	54

Mudelite treenimine ja tulemused	56
<b>PoC ehk Siimuke</b>	<b>59</b>
siimuke.mkm.ee kirjeldus	59
PoC-is kasutatavad mudelid	61
Vastused mõnele hanke tehnilise kirjelduse küsimusele	62
<b>Prototüübi arhitektuur</b>	<b>64</b>
Asutusesisene klassifitseerimine	64
Eesmärk	64
Lahendus	65
Peamine kasutuslugu	66
Kokkuvõte	67
Asutusteülene klassifitseerimine	67
Eesmärk	67
Valikud	67
Föderatiivõpe	68
Teenusmudelite foorum	70
Asutuseülese mudeli arhitektuuri valikud	73
Peamine kasutuslugu	74
Kokkuvõte	74
Üldine turvalisus	74
Kasutajate rollid	75
Logimine	75
<b>Tasuvusanalüüs</b>	<b>78</b>
Klassifitseerija kasutegur	78
Mudelite jätkuarendused	79
Nõuded arendusmeeskonnale	80
<b>Kodaniku pöördumiste klassifitseerijaga isikuandmete töötlemisele kohalduvad andmekaitseenõuded</b>	<b>83</b>

# Sissejuhatus

Antud projekti tehniline kirjeldus selgitab projekti ajendit nii: „Vestlustest erinevate avaliku ja erasektori osapooltega on tekkinud vajadus parendada klientide (kodanikud ja ettevõtjad) koormust riigiga suhtlemisel ning ka vähendada riigi (ministeeriumid ning erinevad asutused) koormust klientidega suhtlemisel. Vajadust vähendada suhtluskoormust ja lihtsustada riigiga suhtlust on välja toodud ka mujal, näiteks 2018. aastal Bürokratia vähendamise rakkerühma poolt ning ka Infoühiskonna arengukavas.

/.../ Samas puudub täna teadmine, kas oleks võimalik klientide päringute suunamise toetamiseks kasutada päringute asutuse põhist automaatset klassifitseerimist, millistele nõuetele peaksid asutuste alusandmed vastama, millises ulatuses ja kas on erinevat liiki pöördumisi (e-mail, vestlusaken (sh. juturobot), kliendikõned jm) võimalik ühtses mudelis kasutada, millised on rakendusele kohalduvad tehnilised nõuded, millised riskid ja piirangud võivad kaasneda – nende probleemide analüüsimisele ja lahendamisele projekt keskendubki.”

Selle projekti käigus analüüsimegi, kas kliendipöördumisi on võimalik ning mõistlik üle riigi, ministeeriumi ja asutuse tasemel klassifitseerida ning mida selle efektiivseimaks tegemiseks tuleks teha. Samuti toome välja, millega tasuks edaspidi arvestada, millised on riskid, kas projekt tasuks ära ning millised on vajalikud tegevused jätkuarendusteks. Lisaks lõime kliendipöördumiste klassifitseerimise PoC-i kontseptsiooni tehnoloogiliseks katsetamiseks, mille kirjelduse (leiate siit) ja asutustele juhendi kliendipöördumiste märgendamiseks (leiate siit).

Detailanalüüs teeb kokkuvõtte intervjuudest riigiasutustega, annab ülevaate projektis kasutatud andmetest ja asutusteülese ja -sisese klassifitseerimise analüüsi tulemustest, kirjeldab Siimukest ehk PoC-i ning prototüübi arhitektuuri, tehnilisi nõudeid, andmekaitseenõudeid, projekti mõju ja vajalikke jätkutegevusi ning nende mahte.

Tehnilises kirjelduses toodud küsimuste, millele detailanalüüs vastama pidi, vastused on eristatud tekstist hallide lilla äärisega kastikeste abil.

# Terminoloogia

## Asutused

Lühend	Asutus
EMTA	Maksu- ja Tolliamet
TTJA	Tarbijakaitse ja Tehnilise Järelevalve Amet
RIA	Riigi Infosüsteemide Amet
PPA	Politsei- ja Piirivalveamet
PRIA	Põllumajanduse Registrate ja Informatsiooni Amet
KRA	Kaitseressursside Amet
KKA	Keskkonnaagentuur
SKA	Sotsiaalkindlustusamet
STAT	Statistikaamet
TA	Transpordiamet
HARNO	Haridus- ja Noorteamet
HK	Haigekassa
RK	Riigikantselei
RR	Rahvusraamatukogu

## Töövahendid

TTK	TEXTA Toolkit	Vabavaraline riigikratt. Dokumentatsioon kätte saadav aadressil docs.texta.ee või lisast 4
-----	---------------	--

## Tõesusklassid

Sümbol	Termin	Definitsioon	Näide
TP	Tõsiposiitiv ( <i>true positive</i> )	Õigesti klassifitseeritud positiivne märgend.	Pöördumisega on seotud märgend "PPA" ning märgendaja lisab vastava märgendi.
FP	Väärpositiiv või valepositiiv ( <i>false positive</i> )	Valesti klassifitseeritud positiivne märgend.	Pöördumisega ei ole seotud märgend "PPA", ent märgendaja lisab selle.
TN	Tõsinegatiiv ( <i>true negative</i> )	Õigesti klassifitseeritud negatiivne märgend.	Pöördumisega ei ole seotud märgend "PPA" ning märgendaja ei lisa seda.
FN	Väärnegatiiv või	Valesti klassifitseeritud	Pöördumisega on seotud märgend

	valenegatiiv ( <i>false negative</i> )	negatiivne märgend.	"PPA", ent märgendaja ei lisa seda.
--	---	---------------------	-------------------------------------

## Masinõppe mõõdikud

Sümbol	Nimetus	Valem	Kirjeldus
TPR	Tõsiposiitivide määr ( <i>true positive rate</i> )	$TP/(TP+FN)$	Tõsiposiitivide osakaal kõikide positiivsete elementide hulgas.
FPR	Väärpositiivide määr ( <i>false positive rate</i> )	$FP/(TN+FP)$	Väärpositiivide osakaal kõikide negatiivsete elementide hulgas.
	Täpsus ( <i>precision</i> )	$TP/(TP+FP)$	Mõõdik mudelite kvaliteedi hindamiseks, mis väljendab ennustatud tõeste märgendite suhet kõikide ennustatud märgenditega (väärtused jäävad vahemikku 0.0-1.0 või 0%-100%). Teisisõnu: palju ennustatult õigete seas tegelikult õigeid oli.
	Saagis <sup>1</sup> ( <i>recall</i> )	$TP/(TP+FN)$	Mõõdik mudelite kvaliteedi hindamiseks, mis väljendab ennustatud tõeste märgendite suhet kõikide relevantsete märgenditega (väärtused jäävad vahemikku 0.0-1.0 või 0%-100%). Teisisõnu: kui suur osa õigetest õige märgendi sai.
	F1-skoor	$2*(\text{Täpsus}*\text{Saagis})/(\text{Täpsus}+\text{Saagis})$	Mõõdik mudelite kvaliteedi hindamiseks, mis kombineerib täpsuse ja saagise (väärtused jäävad vahemikku 0.0-1.0 või 0%-100%).d

## Üldised definitsioonid

- **Asutusesisene mudel** – Mudel, mis ennustab asutusesiseseid märgendeid ehk teemasid, klasse, sarju.
- **Asutuseülene mudel** - Mudel, mis ennustab, millise asutuse pöördumisega on tegu.
- **Juhendamata masinõpe** (*unsupervised machine learning*) - Mudeldamine märgendamata andmestikul.
- **Juhendatud masinõpe** (*supervised machine learning*) - Mudeldamine märgendatud andmestikul.
- **Lemmatiseerimine** on sõna sõnastikuvormi kujule viimine (sõnad *sõi*, *sööksin* ja

1 Saagis on tegelikult sama, mis tõsiposiitivide määr, aga kuna need terminid on antud dokumendis kasutusel pisut erinevates kontekstides, on need tabelis lahku löödud.

söönud saavad kuju *sööma*). See vähendab sõna varieeruvust: arvuti jaoks erinevad sõnad saavad ühe kindla kujuga sõnaks.

- **Märgend** - on silt, mille olemasolu mudelid ennustavad. Antud dokumendi kontekstis kasutatakse seda sõnade "**klass**", "**teema**", "**valdkond**" sünonüümina. Sõltuvalt asutusest võib tähendada pöördumise teemat, alateemat, valdkonda või sarja.
- **Märgendamine** - tekstile märgendi andmine.
- **Märgendipõhine hindamine** (*label-based evaluation*) - Mudelite hindamine kasutades hindamise aluseks dokumentidega seotud märgendeid ehk hinnatakse, kui hästi on mudel keskmiselt võimeline ennustama iga märgendi sobivust juhuslikule testandmestiku dokumendile.
- **Näitepõhine hindamine** (*example-based evaluation*) - Mudelite hindamine kasutades hindamise aluseks dokumente ehk hinnatakse, kui hästi on mudel keskmiselt võimeline ennustama iga dokumendiga seotud märgendite hulka.
- **Stemmimine** võtab ära sõnade käände- ja pöördelõpud, vähendades samuti sõna varieeruvust.
- **Stopp-sõna** on sõna, mida masinõppes treenimise ja ennustamise ajal ignoreeritakse (ehk neid ei vaadata ega kaaluta klassitunnustena). Sinna alla kuuluvad üldjuhul asesõnad, mille tähendus sõltub kontekstist (*see, üleval, nemad*), isikunimed (nt kirjalõppudes korduvad vastajanimed *Mari, Jaan Tamm*), asukohad (*Tartu, Jõhvi*), numbrid (nt telefoninumbrid) ning andmestikuspetsiifilised korduvad sõnad (e-kirjades tervitused, jalustes *lugupidamisega* või asutuse nimed, puhastamata kirja korral *To:, From:* read).
- **Tokeniseerimine** on teksti tokeniteks ehk väiksemateks tekstiühikuteks tükeldamine. Tegu ei ole lihtsa ülesandega, kuna näiteks ainult tühikutega lõikumine võib anda arvuti jaoks erinevad tähejärjendid nagu „tere!”, „tere,” ja „tere”.
- **Ülesobitamine** (*overfitting*) - Mudelite liigne sõltuvus treeningmaterjalist, mis tähendab nõrka üldistusvõimet teistele andmetele.

# Asutuste vajaduste kaardistamine

Antud projekti raames kaardistasime asutuste PoC arhitektuuri välja töötamiseks asutuste tegelikke vajadusi ning pöördumistega tegelemise hetkeolukorda. Antud peatükis toome välja intervjuudel selgunud olulised punktid, millele tuleb antud hankes tähelepanu pöörata. Iga peatükk lõppeb kokkuvõtva arutlusega.

Täname tabelis 1 toodud asutusi, kelle esindajad lahkelt meie küsimustele vastasid.

Asutus	Intervjuu kuupäev
Maksu- ja Tolliamet	03.05.2021
Haigekassa	20.04.2021
Haridus- ja Noorteamet	27.05.2021
Kaitseressursside amet	04.05.2021
Keskonnaagentuur	03.05.2021
Politsei- ja Piirivalveamet	06.05.2021
Põllumajanduse Registrate ja Informatsiooni Amet	13.05.2021
Rahvusraamatukogu	15.04.2021
Riigi Infosüsteemi Amet	06.05.2021
Riigikantselei	13.05.2021
Sotsiaalkindlustusamet	27.04.2021
Statistikaamet	28.04.2021
Tarbijakaitse ja Tehnilise Järelevalve Amet	16.04.2021
Transpordiamet	23.04.2021

Tabel 1. Intervjueeritud asutused intervjuu kuupäevaga.

## Pöördumiste kanalid ja mahud

Detailanalüüsi lisa 1 on toodud iga asutuse pöördumise kanalid, üldised teemad, mahud ja klienditeenindajate arv. Seal on näha, et laias laastus jagunevad kanalid seitsmeks:

- e-postkastid (ühel asutusel võib neid olla 1-7, keskmiselt 2),
- tavapost,
- telefoniliin(id),
- vestlusaknad (inimene vastab) või juturobotid (vastab mudel ja ebaõnnestumise

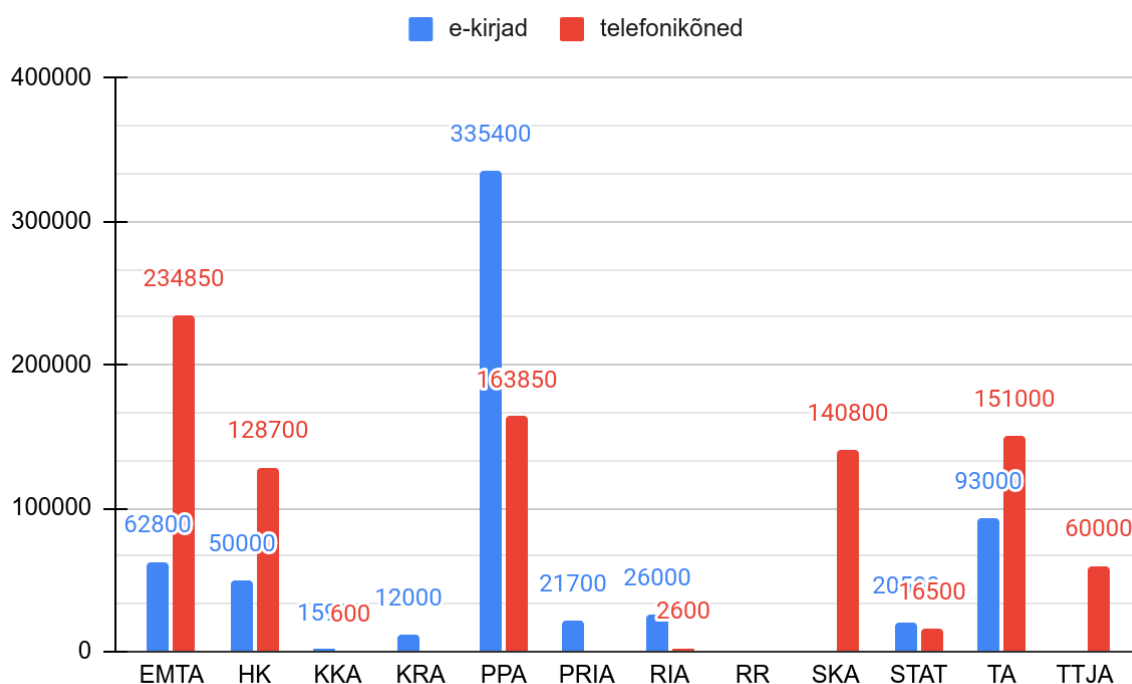


korral suunatakse inimesele),

- kohapealne teenindus büroodes,
- iseteeninduskanalid (kodulehe vormid, iseteeninduskeskkonnad),
- sotsiaalmeedia (Facebook).

Nende seitsme kombinatsioonid või lisandused neile varieeruvad asutuse ti. Pöördumise mahud erinevad väga: kui näiteks PPA saab keskmiselt üle 1350 e-kirja päevas, siis Keskkonnaagentuur saab 1600 e-kirja aastas. Kui RIA vastab Facebookis ühele pöördumisele nädalas, siis Kaitseressursside amet vastab ca 5-10 Facebooki pöördumisele päevas.

Joonisel 1 on toodud hinnangulised e-kirjade ja teenindatud telefonikõnede mahud asutuse ti. Mahud on saadud intervjuudes välja toodud üldjuhul (üle)elmise aasta statistika või hinnangulise päevamahu põhjal.



*Joonis 1. Hinnanguline sissetulevate e-kirjade (ja asünkroonne iseteeninduse vastamise) ning teenindatud telefonikõnede maht aastas. Info puudumisel tulp puudub. KKA saab 1597 e-kirja aastas ning STAT 20500.*

Möödunud aasta oktoobrikuus tehtud arvestuse kohaselt tuleb RR-ile aastas ca 16 000 pöördumist (arv sisaldab nii ekirju kui ka terveid, mitme küsimusega veebivestluseid). SKA-l on erinevate kanalite peale kokku 565 000 pöördumist, sinna lisanduvad meilid ja lehekülgede külastused. Statistikaamet märkis, et just juturobotil tõuseb kontaktide arv - seda kasutavad nii tarbijad kui ka andmesisestajad.

## Huvi juturobotite vastu

**Huvi juturobotite vastu** ilmutas mitu asutust (vaata kokkuvõtvat tabelit 2). EMTA, RIA, PPA kavatsevad liituda Bürokrati juturoboti jätkuprojektiga. Viimased kaks osalesid ka riikliku juturoboti PoC projektis ning on selle edenemisest väga huvitatud. EMTA loodab ilma robotita saada vestlusakna käima juba selle aasta septembrikuus.

HK-I, HARNO-I, TTJA-I ja RR-il on robotita vestlusaken. TTJA loodab tulevikus vestlusaknas genereerida automaatvastuseid. Hetkel testib TTJA vestlusakent, alustades väiksema mahuga lehtedelt ja robotita. Riigikantselei pakkus neile tehnilise platvormi. Rahvusraamatukogul on olnud ka Krati projekt veebivestluses tüüpvastuste automaatseks vastamiseks. Selle projektiga liiguvad nad varsti edasi. Eelmises Krati projektis leiti, et tüüpvastuste loomine on keeruline, kuna sisse tuleb nii üldküsimusi, konkreetseid küsimusi kui ka mitmest küsimusest koosnevat kirju (näiteks küsitakse ühes kirjas lahtiolekuaegu, kas on võimalik kohapeal ennast lugejaks registreerida ning kas teatud kindlat raamatut on). Ka Haigekassa on katsetanud vestlusakna automatiseerimist, kuid pole lõppversioonini jõudnud. Feelingstreamingu<sup>2</sup> pilootprojekti abil tuvastati, et telefonikõned on nii personaalsed, et automatiseerimist on võimalik vähestel juhtudel teha.

Statistikaametil on juba oma juturobot olemas ning seda kasutatakse järjest enam. STAT-i ITi suudab 90 protsendile küsimustest vastata automaatselt ning orienteerub 20 tuhande nähtuse seas. Seda, erinevate parameetritega tabelist õige arvu eraldamise võimekust või päringutele vastamist (eeldab ligipääsu andmebaasile) soovitakse juurde. Neile küsimustele, millele juturobot vastata ei suuda, vastab teenindaja või saadetakse töövälisel ajal vastamata pöördumine meilile.

Transpordiamet tegeleb hetkel juturoboti testimisega väiksematel kodulehtedel ning KRA-s ei ole juturobot või vestlusaken arutlusel olnud.

Eraldi toome välja Sotsiaalkindlustusameti, kes tundis antud projekti suhtes väga suurt huvi. Nende lasteabi osakond on huvitatud sellest, et ühtne riigi juturobot tunneks ära ka teistes asutustes tehtud lastega seotud küsimusi nii, et abi vajav laps jõuaks lasteabisse. Lisaks leiab SKA, et ideaalis võiks juturobot suuta inimesega ka niisama vestelda. Näiteks köidaks lapsi, mille abil saab lastega tekitada usaldusväärset suhtlust. Ohvriabis võiks aidata ohvri ennast diagnoosida ning leida õige kontakt koos info ja julgustusega selle poole pöörduda. On oluline, et sõltumata asutusest, tuvastataks taolise kahtlustusega diskussioonid ning üritatakse seejärel inimest toetada ja suunata edasi. Kindlasti ei tähenda taoline tuvastus ja

2 <https://www.feelingstream.com/>

Julgustus psühholoogi tasemel nõustamise täisteenust, kuid abi on juba märkamisest ja õigele kontaktile suunamisest. Mõned reaalsed pöördumised kriisi juturobotilt SUVI: *ma ei taha enam elada, ma olen nii kurb, ma hakkan lõikuma ennast.*

Asutus	Eksisteerib vestlusaken	Eksisteerib juturobotiga vestlusaken	Huvi juturoboti (edasi)arenduse suhtes olemas	On tegelenud juturoboti PoC-iga
EMTA	jah (alates 09.2021)	ei	jah	Ei, aga loodab sügisel Bürokrati projektiga liituda
TTJA	jah	ei	jah	ei
RIA	ei (Facebookis teenindatakse pöördumisi)	ei	jah	jah
PPA	ei	ei	jah	jah
PRIA	ei	ei	seisukoht puudub (vaja asutusesiseselt arutada)	ei
KRA	ei (Facebookis teenindatakse pöördumisi)	ei	pole asutusesiseselt arutluseks olnud	ei
KKA	ei	ei	N/A	ei
SKA	jah	ei	jah	ei
STAT	jah	jah	jah	jah
TA	jah	ei	jah	ei
HARNO	jah	On algeline juturobot, mis saadab mure edasi	N/A	ei
HK	jah	jah (kodulehel liikumiseks)	Pilootprojektis tuvastati, et automatiseerimine on keeruline	jah
RR	jah	jah	jah	jah

Tabel 2: Asutuste huvi juturobotite vastu

## Pöördumiste keeleline osakaal

Tabelis 3 on toodud pöördumiste keeleline osakaal asutuste lõikes. 15% Statistikaameti pöördumistest on andmeesitajatelt, need on eesti või vene keeles. Keskkonnaametil on aasta jooksul üldse vaid paar kirja võõrkeeles. Tabelis on näha, et kõikide asutuste valdav pöördumiste suhtluskeel on eesti keel. Mitmel asutusel nagu Keskkonnakaitse Agentuuril, Kaitseressursside keskusel ja PRIA-l on teised keeled marginaalsed. Haigekassas, PPA-s, RIA-s, SKA-s ja Transpordiametis teenindatakse ka suurel hulgal (vähemalt 30%) pöördumisi vene keeles. Haigekassa ning Transpordiamet märkavad ingliskeelsete

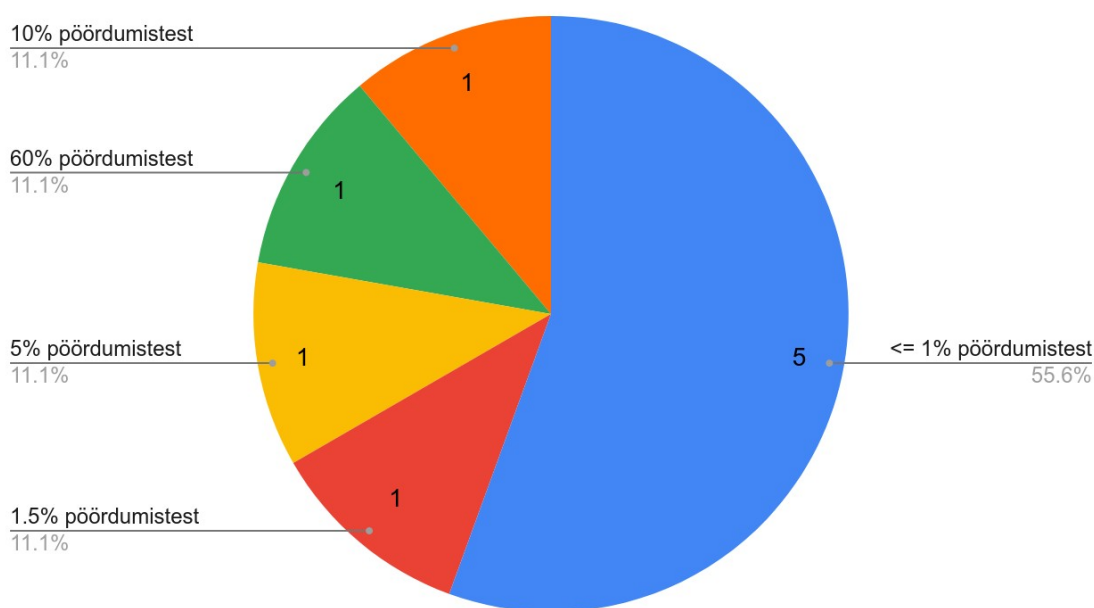
pöördumiste tõusu. Mitmekeelsust tuleks arvestada ka Bürokratis. Näiteks on võimalik treenida mitu ükskeelset klassifitseerijat ja rakendada keeletuvastust. Samuti on suuremate (BERT) närvivõrkude puhul luua mitmekeelset klassifitseerijat.

ASUTUS	eesti keel	vene keel	inglise keel
HARNO	Põhilised keeled. Vene keelt tuleb näiteks Narva ja Ida-Virumaa õpiraskustega klientidelt		Hakkab tulema just sisserändajate arvelt
HK	enamus	40% kõnedest, 30% kirjadest.	pöördumiste arv kasvab kõnes ja kirjas. Üldjuhul pöörduvad inglise keeles mitteinglased, st. kirjepilt ja kõne pole n-õ standard inglise keelega kooskõlas.
KRA	Valdav enamus pöördumistest on eestikeelsed. Ca 5% on vene- või inglisekeelsed.		
KKA	enamus	Üldjuhul kirjutavad muukeelsed ikkagi eesti keeles, mõnikord võib telefonipöördumine olla vene keeles	Rohkem, kui vene keeles.
EMTA	N/A		
PPA	Üle poole pöördumistest. 56% kõnedest	41% kõnedest.	3% kõnedest (üldjuhul keerulisemad ja pikemad kõned).
PRIA	Valdav enamus	marginaalne	marginaalne
RR	Pöördatakse eesti, vene, inglise, saksa ja prantsuse keeles.		
RIA	ca 60%	30%	10% (enamus pöördumistest Single Digital Gate raames)
RK	N/A		
SKA	90% on eesti- ja venekeelsed pöördumised. Enamus eesti keeles.		pöördatakse inglise keeles, osakaalu ei tea.
STAT	N/A		
TTJA	N/A		
TA	Enamus	Vanasti olid pooled kõnedest venekeelsed, nüüd on see vähenenud ca 30 protsendini. Kirju vähe.	Kirjade arv tõuseb (tõus tuleb lennunduse ja merenduse valdkonnast). Ca 3-4% kirjadest.

Tabel 3. Pöördumiste keeleline osakaal asutuste lõikes

## Valele asutusele saadetud kirjade arv

Joonisel 2 on toodud valele asutusele saadetud pöördumiste osakaal. Viis asutust andsid täpse statistika puudumise tõttu mahuhinnanguks tunnetuslikke arve nagu *kuni kümme korda kvartalis või maksimaalselt 10 kirja kuus*, mis nende pöördumiste üldmahte arvestades jäid kõik alla ühe protsendi pöördumistest. Kolm pidasid valele asutusele suunatud pöördumiste arvu nii marginaalseks, et arvulist hinnangut ei andnudki (TA, TTJA, PRIA). RR leidis, et hinnanguliselt ehk 5% sissetulevatest kirjadest ei kuulu nende pädevusse. KKA saab ehk paar korda kuus taolisi kirju, mida tuleb edasi saata. Nende puhul on see ca 1,5% sissetulevatest kirjadest. SKA-s leiti, et tunnetuslikult 3% e-kirjadest suunatakse edasi õigesse asutusse ning 10% kõnedest pole nende haldusalas. Vaid RIA (lausa 60% pöördumistest) ja SKA nimetasid sissetulevate valele asutusele suunatud pöördumiste arvu probleemiks.



Joonis 2. Valele asutusele saadetud pöördumiste osakaal erinevates asutustes

Mitu asutust tõi välja, et tihti on pöördumised, mida nad edasi saadavad, osati ka nende enda poolt vastata. Ehk kliendil on kompleksne probleem, mille lahendamise või millele vastamisega **peabki tegelema mitu asutust**. Asutus vastab enda osa ära ning suunab edasi (klient koopias) või palub kliendi ise teise asutuse poole pöörduda. Seega võib neid lugeda osaliselt vale asutuse pöördumise alla. Taolisi piiripealsed juhtumeid on võimalik treeningandmete olemasolu korral eraldi märgendada. Nende ning täielikult valele asutusele saadetud kirja eristamine võib osutuda keeruliseks väljakutseks, kus 100% täpsust oodata ei saa. Seetõttu tuleb olla ettevaatlik näiteks töövoogude automatiseerimisel: valele asutusele

suunatud kirja klassifitseerija töös peab olema kindel (see peab olema korralikult testitud). Teine variant on jätta osaliselt ja täielikult valele asutusele suunatud kirjad ühte klassi. Klassifitseerija poolt sinna klassi liigitatud e-kirjad vaatab üle klienditeenindaja, kes siis otsustab, kas kiri saadetakse otse edasi või vastatakse osaliselt ja saadetakse edasi. Masinõppe seisukohast on viimane variant mõistlikum ja töökindlam. Eriti arvestades võimalusega luua asutuseülene klassifitseerija. Kolmas variant on kasutada mitme märgendiga klassifitseerijat (ingl *multi-label classification*): kompleksed e-kirjad, mis nõuavad vastuseid erinevatelt asutustelt, saavadki mitu märgendit külge. Mitme märgendiga e-kirjad lähevad klienditeenindajale (võib ju eeldada, et vähemalt üks märgenditest on ikka kirja vastu võtnud asutus), ühe märgendiga suunatakse automaatselt vastavale asutusele edasi. See variant sobitub hästi ka PoC-i üleüldisesse arhitektuuri.

Üldjuhul oskavad asutused välja tuua, milliste asutustega neid segi aetakse või millistele asutustele nad tüüpiliselt kirju edasi suunavad. Asutused oskavad välja tuua ka kindlaid olukordi, millal nende poole valesti pöörduakse. See vihjab potentsiaalile lahendada tüüpilised valele asutusele saadetud juhtumid automaatselt asutuse sees. Näiteks tegeles vanasti SKA matuseteenuse infoga, nüüd tegelevad sellega kohalikud omavalitsused, kuid aeg-ajalt pöörduvad inimesed selle teemaga ikka SKA poole. KRA tõi välja, et reservväelaste õppuse ajal saavad nad 10-15 kirja päevas, millega nemad tegelikult ei tegele. Muul ajal saavad nad samuti ca 10 kirja kuus. HARNØ saab siiani regulaarselt karjääriteemalisi kirju, mis varasemalt oli nende teema, aga nüüd kuulub Töötukassa haldusalasse. Leiame, et seda potentsiaali tasub kindlasti ära kasutada, sest **asutusesisene valele asutusele saadetud kirjade klassifitseerimine tüüpjuhtude abil võib väga häid tulemusi anda**. Lihtsam ja odavam on klassifitseerida üksikuid tüüpjuhte asutusesiselt, kui sealsed klienditeenindajad on need juba niivõinää tuvastanud ja on neile koostanud tüüpvastused. Asutusesisese mudeli ennustatavad klassid on asutuse enda määrata ning soovi korral võivad nad ka sagedaseimad valepöördumised oma mudeliga ära lahendada, selmet asutusteülese mudelitesüsteemiga liituda.

## Pöördumiste menetlemine

### Klienditeenindajate arv

Viiel asutusel on töötajad (ca 5-12), kes tegelevad ainult klienditeenindamisega. Seitsmes asutuses tegeleb 1-2 inimest kas täiskohaga või muu töö kõrvalt kirjadele vastamise või nende asutusesisese edasi saatmisega, kus vastab spetsialist kirjaga edasi tegeleb. Kahe puhul teenindajate arvu ei tuvastatud. Täpsemat infot asutuse kaupa saab lugeda lisast 1.

## Pöördumise menetlemise ajakulu

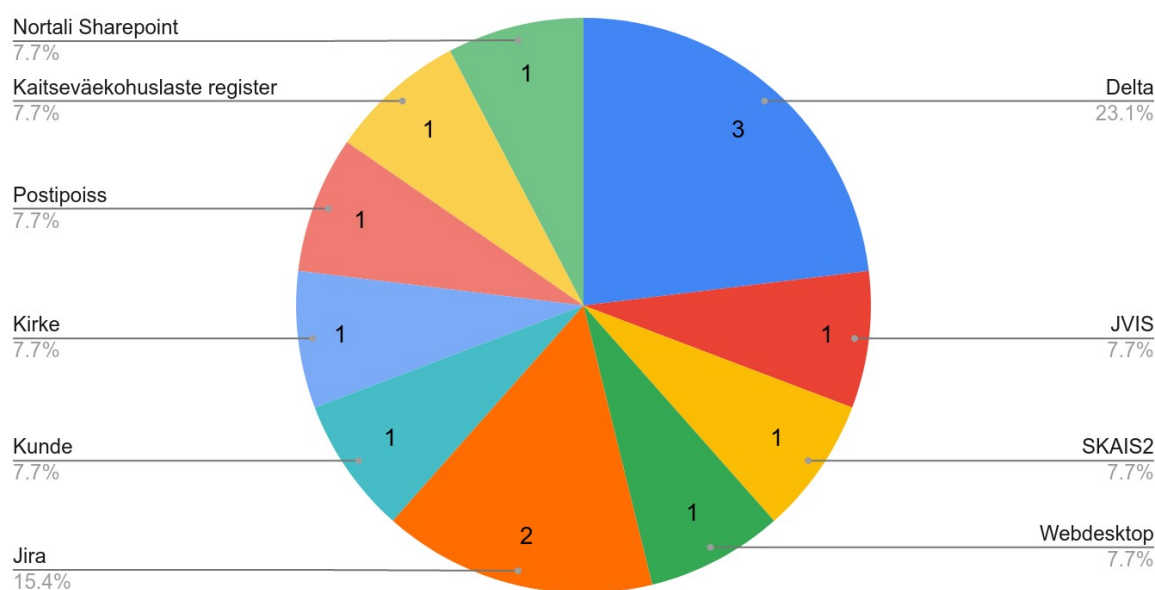
See, kui kaua aega võtab pöördumisele vastamine, sõltub pöördumise iseloomust. Näiteks infopäringud ja kergemad isikustatud küsimused võivad võtta aega vastavalt asutusele 1-30 minutit (alates kirja avamise hetkest). Teiste puhul võib asutusesiseselt sisendi küsimine võtta mitu päeva. Kirjad, millega peab mõni ametnik tegelema ja mille puhul peab suhtlema erinevate asutuste ja pooltega, võivad venida mitme kuu pikkuseks kirjavahetuseks.

Tegelik vastuse saamise aeg kirja saatmise hetkest sõltub aga järjekorrast. Üldjuhul hoitakse kinni asutusesisesest kokkulepitud standardist (nt 3 või 5 päeva) ja saadetakse kliendile selle kohta automaatteade. Täpsemalt saab lugeda ajakulu asutuste kaupa lisast 1.

## Pöördumiste hoiustamise keskkonnad

Meiliserverina kasutavad Exchange'i kõik intervjueeritud kliente haldavad asutused. Üks nendest kasutab seda ka kirja salvestuskohana.

Üks asutus hoiustab kirju JVIS-is ja Deltas. Teine kasutab nii SKAIS2-te kui ka Deltat. Kolmas ainult Deltat. Webdesktopi, mis annab kasutajatele ka pöördumiste statistikast aru, kasutab üks asutus. Jiras haldab pöördumisi 2 asutus. Üks asutus kasutab Saksamaa kliendihaldusprogrammi Kunde, läbi mille liiguvad pöördumised dokumendihaldussüsteemi. Teisel salvestuvad kirjad Kirkesse, kolmandal Postipoissi ja kaitsevääkoruslaste registrisse, neljandal on Nortali loodud Sharepointil töötav dokumendihaldussüsteem. Sama info on toodud joonisel Viga: viidatud allikat ei leitud.



Joonis 3: Pöördumiste hoiustamise keskkonnad ning seda kasutatavate asutuste arv

## Ühe pöördumise maksumus

7 asutust ei ole ühe kliendipöördumise teenindamise maksumust hinnanud, üks ei avaldanud. Need tulemused ei ole üks-ühele võrreldavad, kuna on arvatud erinevalt ning sõltuvad pöördumiste hulgast ning teenindamistavadeast. Erinevad asutused on hinnanud kliendipöördumiste maksumuseks 1,9 eurot, 3,71 eurot ja 8 eurot. Üks asutus hindab oma päevakuluks viie infoliiniteenindaja peale 6 eurot. Teisel maksab üks kõne ca 2,5 eurot ning kiri 10-11 eurot. Kolmandal on kõneminuti maksumus koos administratiivsete tegevustega 1,18 eurot ning dokumendihaldussüsteemi suunatav kiri maksab ca 40 senti ning vastust nõudev keskmine kiri maksab tööjõukulu alusel 2 eurot.

Toodi ka välja, et kuna kõne hind on odavam ja kiirem, **eelistatakse helistamist** näiteks pika ja segase kirjalike kirjaliku vastamise asemel.

## Kirjade märgendamise hetkeolukord

Intervjuude käigus proovisime tuvastada, kas praegu juba mingil kujul pöördumisi märgendatakse. Selgus, et pöördumisi kategoriseeritakse erinevatel viisidel. See võib raskendada ka asutuseülest rakenduse kasutuselevõttu. Seda kinnitas ka läbiviidud katseprojekt, kus asutused olid märgendanud pöördumisi erinevalt ning kohati ka viisil, mis muutis andmete töötlemise sisuliselt võimatuks. Näiteks võib märgendamiseks lugeda pöördumiste suunamist Outlooki erinevatesse alamkaustadesse, dokumendihaldussüsteemis (nt Delta või Web Desktop) pöördumistele funktsiooni, sarja või teema (koos alamteemadega) lisamist etteantud valikust või vabakirjalise kokkuvõtte/teema lisamist metaandmetesse. Mitu asutust tegeleb hetkel aktiivselt uue pöördumiste menetlemise süsteemide loomisega ning loodab kodulehel olevate vormide või iseteeninduskeskkondade abil teenindajate töökoormust vähendada. See näitab klienditeenindusjuhtide motiveeritust ja huvi kaasa lüüa projektides, mis tulemusena aitavad nende klienditeenindajaid ning vähendada pöördumistele vastamisele kuluvat aega.

Mõni asutus lisab metaandmetesse **mitu erinevat märgendit ühele kirjale**: nt teenindaja, sari ja valdkond või teema ja kirja liik nagu *infopäring*, *selgitustaotlus*, *vaie*. Märgendite arv sõltub asutuse teenindusvaldkonnast, -praktikast ja pöördumiste mahust. Näiteks leiab üks asutus, et nende mahtude juures tundub 16 funktsiooni ja nende all olevad mitmed sarjad täiesti piisavad. Teine Outlookis ei tee midagi, kuid dokumendihaldussüsteemis jaotab dokumendid üle 300 sarja. Kaks asutust tõid välja, et **tegelikult liiga palju liigitusvalikuid toob endale kaasa vigu ning segadust**. Ka projekti katsetest ilmses, et liiga detailseid ja



omavahel sarnaseid klasse on mudelil keeruline eristada. Seetõttu **soovitame klasside valikul võimalusel eelistada üldistamist**. Praeguste märgendite põhifunktsiooniks on üldjuhul:

- a) pöördumisest ülevaadete saamine statistika tarbeks,
- b) hiljem pöördumise otsimise lihtsustamine,
- c) teavitamine, mida kirjaga tehti,
- d) teavitamine, mida kirjaga teha tuleb.

Kindlasti tuleb arvestada ka sellega, et praegune olukord ei pruugi vastata sellele, mida tegelikult vaja on. Paljud asutused on vaeva näinud ning vähendanud ja korrastanud märgendite arvu ning täiustanud vastavalt pöördumiste analüüsile iseteeninduskeskkondi, et vähendada klienditeeninduskoormust. Intervjuude käigus saime ülevaate praegusest märgendamisskeemist. Selleks, et teada saada, kas tegelikult leidub veel võimalike märgendeid, mis teenindajate töövoogu hõlbustaks, **tuleb analüüsida iga asutuse pöördumisi ja nende menetlemist süvitsi koostöös klienditeenindusspetsialistidega**. Nagu üks intervjuueeritav ütles: “See, mis näib, ei ole see, mida vajame. Ja mida me tahame, ei ole see, mida me vajame”.

Märgendite kirjeldamisel toodi välja ka asjaolu, et **märgendamine (eriti DHS-ides) on aeganõudev** ning 9 asutust nõustus, et automaatne märgendamine kasvõi teenindajale soovitusena pakkumisena kiirendaks oluliselt pöördumise teenindusaega.

Ühel asutusel on Outlookis **defineeritud reeglid**, mille abil suunatakse lihtsamad tavalised ja korduvad kirjad automaatselt kindlasse kausta edasi (näiteks kindlalt aadressilt kirja kättesaamise kinnitused). Moodsas masinõppevaimustuses ei tasu reeglipõhiseid pöördumiste tuvastamisi unustada, kui nende kasutamine võimalik on. Reeglipõhiste lähenemiste eelis on robustsus ja sellest tulenev kiirus. Samas päris kõike nende abil teha ei saa – klassifitseerimiseks ei saa alati tuvastada ilmselgeid teisi klasse ja olukordi välistavaid reegleid. Pigem tasubki neid eelistada sellistel juhtudel, mida eeltoodud kindlalt aadressilt saadud kirja näide kirjeldas.

Laias laastus võib asutuseülesed pöördumised jagada kaheks:

- a) infopäringud, millest osadele on teenindajatel olemas ka enda loodud või ühises kasutuses (nt Drive kaustas) olevad **tüüpvastused**,
- b) isikustatud pöördumised, mille **vastus sõltub kliendi andmetest** erinevates andmebaasides või kirja sisus.

Esimesele variandile vastamise koormuse leevendamine ehk automaatne vastamine või tüüpvastuse soovitamine on lihtsam, kuna ei eelda tehniliselt ja andmekaitseliselt keerukat ühendust andmebaasidega. Teise osakaal kogupöördumistest sõltub natukene asutusest. Näiteks tegelevad Haigekassa ja Sotsiaalkindlustusamet küsimustega, millele korralikuks vastamiseks ja võib-olla küsimusevälise vajaliku lisainfo andmiseks on vaja teada inimese tausta ja olukorda. Haigekassa on läbi teinud Feelingstreamingu pilootprojekti, milles leiti, et telefonikõned on nii personaalsed, et automatiseerimist on võimalik vähestel juhtudel teha. Seetõttu võib nendele asutustele klassifitseerijaid tehes meeles pidada, et lihtsamaid, tüüpvastustega lahendatavaid juhtumeid on teiste asutustega võrreldes märgatavalt vähem ning teenindajate koormus väheneb alles siis, kui süsteem suudab vastata isikustatult. Ka Transpordiamet liigub selles suunas, et teenindajad kasutaksid võimalikult palju küsija tausta (uurides kättesaadavaid andmebaase), et vältida infomüra ja üleliigse info külvamist kirja. Samuti tõid nad välja, et tihti pöörduakse nende poole küsimusega, mis ei sisalda vastamiseks piisavalt lisainfot. Kui küsitakse *kuidas autot ümber registreerida?* sõltub vastus sellest, kas tegu on kliendi või liisingu autoga. Seega on vajadus liikuda asünkroonse suhtluse poole, kus klientidelt saaks täpsustavaid küsimusi küsida vastavat vajadusele pea reaalajas

Sellest johtub ka **vajadus sisukontrolli järele** (see säästaks kindlasti teenindajate aega). Näiteks tunti huvi, kas on võimalik kontrollida pöördumistel teatud nõuete täitmist (kasvõi see sama isiklik vs liisinguauto info olemasolu) ning vajadusel lisainfot automaatselt juurde küsida või teenindajale sellest märku anda. Kontrollida võib ka näiteks manuse olemasolu (kui kirjas on manus mainitud) ning taotlusel digiallkirja olemasolu. Üks asutus kontrollib hetkel kahtlaseid linke (kui näiteks saadetakse dokument Drive'i lingiga) SMIT-i abivahendiga, mille jaoks peab lingi kontrollijasse kopeerima. Neil on huvi see kontroll automatiseerida. Lisaks tunnevad nad puudust **isikutuvastusest**<sup>3</sup> - hetkel võib väga juhuslike nimedega e-maili aadressidelt laekuda isikustatud pöördumisi ning teenindajad ei saa hästi kontrollida, kas tegu on selle isikuga, kelleks kirja sisu väidab end olevat. Ühes teises asutuses on jälle huvi spämmi, reklaamide ja koolituspakkumise välja filtreerimise kohta. Kolmas töö välja, et tänukirjade (“aitäh veelkord”) lugemine võtab ka aega ja need võiks automaatselt eraldi kausta või prügikasti saata. PPA puhul on oluline kiirreageerimine. Mõnikord saadetakse nende üldmeilile kirju, millele tuleb vastata kas kohe või millele peab reageerima jõudude välja saatmisega. Hetkel pannakse nendele eraldi käsitsi “Kõrge prioriteet”-lipuke külge. Ka seda võib võtta omamoodi märgendina, mis aitaks hõlpsamini üles leida päevas üle tuhande saabuva kirja seest kohest reageerimist vajavad sündmused.

3 v.a juhtudel, kui pöördumine on soovitud anonüümne vihje vihjeliinile.

Kui eelmises peatükis tõime välja ühe pöördumise mitme asutuse vahel jagunemise keerukuse, siis tegelikult **võib ka asutusesiseselt olla pöördumine mitmeteemaline**. Näiteks küsitakse ühes kirjas Rahvaraamatukogule tolle lahtioleku aegu, kas on võimalik kohapeal ennast lugejaks registreerida ning kas teatud kindlat raamatut on. Märgendajana on loomulik näha selles pöördumises kolme erinevat märgendit ja tüüpvastust, mis tegelikult tuleb siinkohal ühele kirjale kohe külge panna. Seetõttu leiti ka nende juturoboti pilootprojektis, et tüüpvastuste loomine on keeruline, kuna sisse tuleb nii üldküsimusi, konkreetseid küsimusi kui ka mitmest küsimusest koosnevaid kirju. Mitmeteemalisuse probleem on ammu tuntud klassifitseerimise probleem (näiteks juturobotites tegeletakse nendega eraldi<sup>4</sup>). E-kirjade kontekstis soovitame taolised kirjad jaotada teemade kaupa osadeks ning märgendada lisaks kogu kiri sobivate teemadega. Sel viisil saab treenida klasse ning testida klassifitseerija tulemust kogu kirjal. **Mitmeteemalisuse puhul soovitame lubada klassifitseerijal määrata mitu märgendit**. Sel viisil näeb klienditeenindaja ära, et mudel on kas segaduses või pöördumine sisaldab mitut teemat. Seejärel saab teenindaja ise otsustada, kumba klassi pöördumine tegelikult kuulub, või jättagi pöördumine kahe- või enama klassiliseks. Mitmeteemalisuse miinuseks on võimalik suurem valepositiivide hulk ning väiksem võimalus töövooge automatiseerida (kui kindla märgendi ehk klassi saanud pöördumine muidu automaatselt kuhugi edasi läheks, siis ebamäärasuse puhul seda teha ei saa).

Vaid ühel asutusel on toimiv reegli- ja statistikapõhine klassifitseerimissüsteem, mille puhul neil puudub vajadus muudatuste, sh tehisintellekti ja lisaklassifitseerija järele.

## Nõuded pöördumiste statistikale ja klassifitseerija(te)le

Vähemalt neli asutust kasutab märgendeid ning klientidelt küsitud rahuloluhinnangut hiljem teenuse arendamise ja statistika sisendina või saab juba oma süsteemides mingil määral statistilist infot kätte erinevate pöördumislükkide (kõned, kohapealne pöördumine, ekiri) lõikes. Osadel asutustel hallatakse erinevat liiki pöördumiseid erinevates süsteemides, mistõttu ei saa neid korraga analüüsida.

Huvi analüüsi vastu on suur, kuna nähakse potentsiaali teenuste parendamiseks. Mõni asutus ütles, et hetkel ei jätku käsitsi analüüsimise jaoks ressursi, kuid hea meelega kasutaks automaatset kokkuvõtet.

4 Nt Rasas: <https://blog.rasa.com/how-to-handle-multiple-intents-per-input-using-rasa-nlu-tensorflow-pipeline/>

Kliendipöördumiste ülevaade võiks olla nii kuupõhine kui ka jooksev info (nt häirete puhul on vaja kohe sisendit). Oluline on, et seda oleks võimalik lihtsalt, selgelt ja minimaalsete sammudega kasutada.

Asutused töid välja erinevaid aspekte, mida võiks nende arvates kuvada. Aega säästaksid neist hinnanguliselt järgmiste asjade kuvamine:

- märgendite<sup>5</sup> mahtu kanalite kaupa ja üheskoos (kõned transkribeeritult, uute esile kerivate teemade info teenindajatele valmistumiseks),
- ülevaade välja saadetud infost (ühe kontakti põhimõtte järgimine, keeled, teemad, sisu),
- kas pöördumine sai asutuselt vastuse,
- kas klient sai esimese vastusega kõik vajalikud vastused kätte või oli sunnitud uuesti pöörduma,
- kas klient käitus saadud soovitudele ja juhistele vastavalt (vastuste kvaliteedi parandamine, et ei peaks korduvalt kirjutama),
- kliendi pöördumise põhjus (kas mingile teenusele on vaja lisada paremaid juhiseid vms),
- millised on ühetaoliste pöördumiste tüüpvastused (andmebaas teenindajatele),
- kes oli saatja ja selle kordumine (top saatjad, kirjutab mitu korda erinevatel aegadel – seose märkamine).

Klienditeeninduse haldamisel ja kvaliteedi kontrolliks on abi järgmiste asjade kuvamisest:

- töötaja vastuste kvaliteeti (ei kasuta keelatud sõnu, ei halvusta ega kopeeri seadusakte),
- pöördumise ja/või kanali teenindamise kulu,
- kliendirahulolu (osad asutused küsivad pärast teenindamist näiteks automaatse e-kirja abil),
- dialoogi pikkus (ajaliselt ja pöördumiste arvu järgi),
- pöördumise teenindamise aeg (anomaaliate tuvastus, kas vastati ette määratud aja jooksul (nt 3 või 5 päeva või 3 kuud)),
- kui kaua pöördumine ootas enne, kui sellega tegelema hakati,
- ajaanalüüs ehk kuidas muutub teatud teemadel pöördumiste maht ajas.

Kui eelnevalt kirjeldasime **vajadust tuvastada ohvreid** ja neid suunata õigetele kontaktidele, siis Sotsiaalkindlustusamet lisas veel ühe soovi. Nimelt on nende poole riik

<sup>5</sup> Märgend on siinkohal ebamäärane mõiste, mis sõltuvalt asutusest tähendab teemat, alateemat, valdkonda või sarja.

mitmel korral pöördunud küsimusega, kuidas oleks hea midagi teha, kuid neil pole teistelt asutustelt andmeid selle kohta, mille põhjal otsustada. Näiteks ei teata, milline on seis Transpordiametile tulnud puuetega seotud küsimustega (kas ja kui palju ja mis teemadel täpselt neid on?). Seetõttu on huvi kätte saada **statistikat enda asutuse teemade kohta teiste asutuste pöördumistes**.

Klassifitseerija loomisel võib probleemiks kujuneda ka valdkondade paljusus ja asjaolu, et kogu aeg tuleb uusi teemasid ja teenuseid juurde. Seetõttu peab klassifitseerija olema **ületreenitav** ning peab jääma **võimalus lisada uusi märgendeid**. Samuti tuleb silmas pidada pöördumiste **mitmekeelsust**. Kaheldakse ka klassifitseerija võimes: kuidas suudab klassifitseerija klassifitseerida, kui teenindajagi mõtleb tükk aega, kuhu alla see kiri läheb ning üks kiri võib olla **mitmeteemaline** või võib selle alt välja kooruda mitu teemat. Seetõttu on mõni asutus äraootaval seisukohal ning kaks asutust tunneb huvi pigem asutusesisese klassifitseerijate vastu. Neli asutust leiavad, et hoolimata näilisest valele asutusele pöördumiste väiksest mahust teeks asutuseülene klassifitseerimine siiski elu lihtsamaks.

Klassifitseerijad peavad arvestama järgmiste asjaoludega:

- personaalne info ei tohi majast välja minna ning e-kirjad võivad neid sisaldada,
- sissetulevad kirjad võivad olla krüpteeritud,
- pilveteenuseid ei pruugi olla võimalik kasutada (piirangud seonduvalt infoturbest ja IKÜMist),
- e-mailide ja juturobotite treenimine võiks luua ühtse terviku ning olema koos analüüsitav,
- vastused võivad sõltuda andmebaasidest (nt statistikapäringud STAT ametile või isikuandmetest sõltuvad vastused Sotsiaalkindlustusametis),
- veaprotsent peab olema võimalikult väike<sup>6</sup>.

Automaatset märgendamist saab kasutada üldjoontes kahel moel:

- a) märgendit soovitatakse teenindajale, st teenindaja kinnitab või muudab,
- b) ennustatud märgendi abil tehakse juba mingi tegevus, nt kiri saadetakse automaatselt edasi või edastatakse dokumendihaldussüsteemi.

Kui variant b on ideaal, siis alustada soovitakse variandiga a: automaatseks võib asja teha siis, kui on kindel, et klassifitseerija on piisavalt hea. Märgend võib olla nii reeglipõhine (märksõnade leidmise järgi sisus või saatja aadress põhjal) kui ka masinõppepõhine. Asutused leiavad, et automaatne märgendamine vähendaks tööaja raiskamist (nt ajakulu

6 Üldjuhul leiti, et 10% oleks maksimaalne, kuna ka teenindajad teevad vigu. Üks asutus tõi välja, et neil on juba praegu problemaatiline, kui asutusesiselt kiri valele vastajale edasi saadetakse, seega nad ootavad väga kõrget täpsust.

kirjadele registreerimiseks ja edasi saatmiseks) ning lihtsustaks tööprotsesse ning aitaks kliente teenindada ka töövälisel ajal. Samuti aitavad automaatsed märgendid pöördumiste sisu analüüsida, mille abil planeerida ressursse ja töökorraldust paremini.

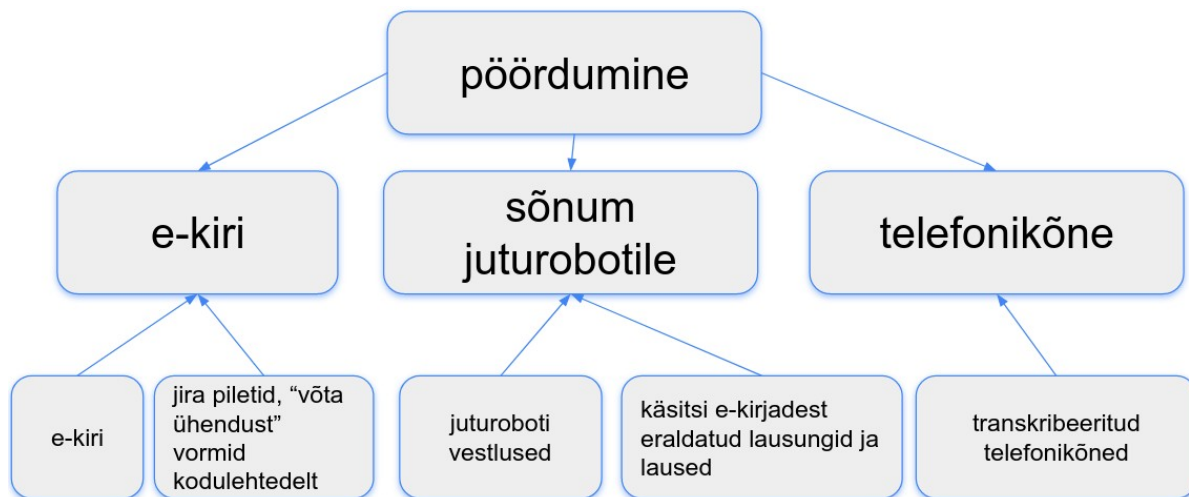
# Andmed

## Ülevaade andmetest

See peatükk annab ülevaate analüüsi aluseks olevatest märgendatud algandmetest. Tabelis 4 on toodud erinevatelt asutustelt saadud andmete arv ja liik. Joonis 4 illustreerib pöördumiste jaotust kolme žanrisse – e-kiri, juturobot ja telefonikõne (välja on jäänud kohapealne teenindus) – ning saadud andmete liigitamsit žanritesse.

Asutus	Pöördumiste arv	Pöördumiste liik
PPA	995 (647 + 348)	e-kirjad
Keskonnaagentuur	373	e-kirjad
Keskonnaagentuur	213	juturobotid
Transpordiamet	582 (78 + 504)	e-kirjad
Transpordiamet	28K	automaatselt anonümiseeritud e-kirjad
TTJA	73	juturobotid
TTJA	5996 (spämmita)	e-kirjad
RIA	124	Jira piletid, telefonikõnede kokkuvõtted eemaldatud.
Rahvusraamatukogu	1218	juturobotid
Rahvusraamatukogu	5813	juturobotid
Haigekassa	80	juturobotid
Sotsiaalkindlustusamet	254 + 194	“võta ühendust” vorm
Sotsiaalkindlustusamet	896	kodulehe tagasisidevorm
Statistikaamet	205	e-kirjad
Statistikaamet	17505	juturobotid
Kriisi juturoboti vestluslogid	23K	juturobotid

Tabel 4. Ülevaade projektist saadud märgendatud andmetest



Joonis 4. Pöördumiste ja andmete liigitus

**Politsei- ja Piirivalveamet** katalogiseeris 647 e-kirja 13 teemasse, igas teemas 50 näidet (v.a *väljastuskoha muutmine*, kus on 47 näidet). Teemade valikul lähtuti enda loogikast ning toodi välja teemad, mida kõige rohkem küsiti. Osad teemad on valitud sarnasest valdkonnast (nt teemad *Elamisloa lisadokumendid*, *Elamisloa pikendamine* ja *Infopäring elamisloa kohta* tegelevad kõik elamisloaga seotud küsimustega). Seega võib eeldada nende teemade omavahelise eristamise keerukust. Sellegipoolest on teemad hästi piiritletud ning nende automaatse klassifitseerimise korral on võimalus ette anda tüüpvastuseid. E-kirjad eraldati käsitsi info@politsei.ee postkastist (msg failid). Need on eelnevalt anonümiseeritud - nimed ja asutused on asendatud tühikuga. Muidu muudatusi kirjades ei tehtud. Tervitused ja jalused (nt “Lugupidamisega”) on alles. Taoline korrektne andmestik sobib väga hästi asutusesiseseks ja asutuseüleseseks klassifitseerimiseks.

	(995) TEEMA
	(75) Dokumendifoto saatmine
	(75) Elamisloa lisadokumendid
	(75) Elamisloa pikendamine
	(75) ID-kaardi ja passi valmimine
	(75) Infopäring elamisloa kohta
	(75) Kuidas riigist lahkuda liikumispiirangu ajal
	(75) Kuidas riiki siseneda liikumispiirangu ajal
	(75) Kuidas tohin liikuda liikumispiirangu ajal
	(75) PIN koodid
	(75) Teade vara- või isikukahju kohta
	(75) menetluse päring
	(74) Maksekorraldus
	(71) väljastuskoha muutmine
	(25) Dokumentide uuendamine

Joonis 5. PPA anonümiseeritud ja märgendatud kirjade teemad ning arv

16. aprillil lisandus igale teemale meie soovil veel 24-25 näidet, kokku 348. Joonisel 5 on toodud PPA märgendatud kirjade teemad ja näidete arv selles.



**Keskkonnaagentuur** jagas 373 käsitsi märgendatud e-kirja aastatest 2016–2020. Need on kõik sellel ajaperioodil laekunud e-kirjad. Nendest 27 on saadetud teisele asutusele (näiteks Eesti Keele Instituut või Keskkonnaamet) edasi. E-kirjad on enamasti tervituste ja jalusteta. Kokku leidub andmestikus 20 unikaalset teemat ning igal pöördumisel võib olla mitu erinevat teemamärgendit. Üks ja sama teemamärgend on mõnikord kirjutatud erinevalt ja seetõttu jagunevad need erinevateks teemadeks. Näiteks on teemad *tuule suund*, *tuulte suund* ja *tuulesuund* samaväärsed, kuid kirjutatud erinevalt. Kõige sagedasemad teemad on *Vaatlusandmed* (180), *Ekskursioon* (25), *Ilmaprognosis* (25), *sademed* (22) ja *õhutemperatuur* (20). Teemade korduvuse (üks pöördumine võib olla kategoorias *Vaatlusandmed* aga ka näiteks *päikesepaiste* või *vooluhulgad*) ja andmete vähesuse tõttu ei saa teha asutusesisest klassifitseerimist. Samuti ei pruugi sellele väikese kliendikirja hulga tõttu olla ka tarvidust. Neid kirju saab kasutada asutuseüleseks klassifitseerimiseks. Väga spetsiifilise temaatika tõttu ennustame, et sellele asutusele kuuluvate kirjade tuvastamine võib kujuneda edukaks.

Keskkonnaagentuuri jagatud teine andmestik koosnes 213 pöördumisest. Pöördumised koosnevad enamasti ühelauselise küsilausest, mis on otse välja võetud keskkonnaametile suunatud ekirjadest või siis mõnel juhul oma sõnadega kokku võetud. Osa sellest andmestikust kattub keskkonnaagentuuri poolt antud esimese andmestikuga. Kuna tegu on pigem ühelauseliste pöördumistest eraldatud küsimustega, sobib see andmestik rohkem juturoboti žanrisse, kui emaili klassifitseerimisse. Eelnevad kommentaarid teemade ja klassifitseerimisvõimekuse kohta kehtivad ka sellel andmestikul.

Huvitav on märkida, et mõlemad andmestikud sisaldavad infot selle kohta, kuhu antud pöördumine edasi suunati (vastati Keskkonnaagentuurisiseselt või saadeti edasi Maanteeametile, Eesti Geoloogia Keskusele).

**Transpordiamet** andis 78 e-kirja, mida hiljem täiendas 504 kirjaga. Need on jaotatud kaheksasse kategooriasse (vt joonist 6). Kirjadest on eraldatud tervitused ja jalused ning telefoninumbrid anonümiseerisime automaatselt. Nendest kolmel on piisavalt andmeid asutusesiseseks klassifitseerimiseks. Lisatud on ka pöördumiste kokkuvõtted, millest osa kordub (nt *Juhiload ja juhtimisõigus* (72 korda) või *Sõiduk* (34 korda)). Sobib asutuseüleseks klassifitseerimiseks ning ekirja žanrisse. Asutusesiseseks klassifitseerimiseks on näiteid liiga vähe.

Lisaks panustas Transpordiamet 27580 osaliselt automaatselt ja osaliselt käsitsi anonümiseeritud ekirjaga. See on skriptiga ja manuaalselt anonümiseeritud ning me ei tea detailselt, kuidas see tehti. Just see manuaalsus tähendab, et meil pole võimalik automaatselt kõik ära puhastada. Lisaks anonümiseeritusele peame eraldama ka teksti sisu, millest peale vaadates tavainimene hetkel aru ei saa. Kõik see eeltöötlus on väga mahukas ega taga meile originaalandmestikulähedast tulemit. Kui tekst pole puhas (me ei saa garanteerida, et on võimalik kätte saada tekst kujul, millisena tulevased tekstid sisse hakkavad tulema), ajab see vaid mudelit segadusse. Seetõttu jätame selle andmestiku välja. Edastatud kujul andmete klasterdamine teemade tuvastamiseks ei ole mõistlik, sest need on liiga raskesti loetavad klatri sisu kiireks hindamiseks.

▼	(582) TEEMA	
	(253) Juht	
	(205) Sõiduk	
	(64) Tehno	
	(30) Eksam	
	(16) E-teenindus	
	(9) Liiklusohutus	
	(4) Üldine	
	(1) Sõiduk, Tehno	

**Tarbijakaitse ja Tehnilise Järevalve Amet** eraldas 73 *Joonis 6. Transpordiameti pöördumist, nendest kõik kuuluvad Tarbimiskeskonna e-kirjade jaotus nõustamise* alamteema *tarbjaõigused* kategooriasse. Pöördumised on omakorda jagatud 18 kategooriasse nagu *puudustega\_asi* (21), *e-poe\_tellimus* (10) ja *Teenuse\_mittevastavus* (9). Kategooriad on täpsed (nt *Sideteenuse lepingu ühepoolne muutmine sideettevõtja poolt* ja *Sideteenuse arvega mittenõustumine, tellimata teenused*) ning on tõenäoliselt tehtud pidades silmas juturoboti andmestiku loomisjuhendit. Pöördumised ise on muutmata küsimused kopeeritud kliendi pikemast kirjast ning sobivad taolise ekirjast eraldamise tõttu pigem juturoboti žanrisse. Kategooriate maht ei ole laiapõhjaliste klassifitseerimiskatsete jaoks piisav.

Lisaks on Textal eelnevast koostööprojektist TTJA-ga kasutada 11 978 märgendatud kirja, mis on kas saadetud ose TTJA meilile või on sinna jõudnud asutuse veebivormide kaudu. Nendest 5996 on märgendi poolest spämmi ja/või reklaamivabad. Kirjad on märgendatud poolenisti automaatselt, poolenisti käsitsi Texta töötaja poolt TTJA edastatud teemade nimekirja alusel, ning sisaldavad tervitusi ja jaluseid. Kirjad jagunevad 13-sse teemasse, millest igaühe kohta on vähemalt 100 näidet. Kuna iga teema kohta olid teada mõningad alamteemad või kirjatüübid, mis sinna alla kuuluvad, on ka nende kohta osaline märgendus olemas. Kokku 11 alamteemat, millest rohkem kui 50 näidet 9 alamteema puhul.

**RIA** lisas omalt poolt treeningandmestikku 762 näidet. Tegemist on väljavõttega Jira piletitest, mis sisaldavad ekirju. Kui ekirjade tervitustes või jalutustes on isikuandmeid, on need eemaldatud. Piletid on märksõnastatud 92 erineva teemamärksõnaga, kusjuures üks

pilet võib olla ka rohkem kui ühe märksõnaga seotud. Märksõnad omakorda on jagatud 16-sse ülemteemasse, millest umbes nelja puhul on piisavalt andmeid teemapõhise klassifitseerija treenimiseks. Teemade kategoorias on kaheksal piisavalt näiteid teemapõhise klassifitseerija treenimiseks. Andmestik sobib ekirjade žanrisse.

Hiljem selgus, et andmestik sisaldab ka neid kirju, mis RIA teistele asutustele edasi saadab. Asutuseüleseks klassifitseerimiseks andsid nad meile 124 eelnevaga sarnast, kuid teemade suhtes märgendamata e-kirja, millega nad tegelevad RIA-siseselt.

**Rahvusraamatukogu (RR)** panustas andmestikku 1 210 küsimuse või fraasiga. Algandmed koguti kahel erineval viisil: veebivestluste logidest võeti iga vestluse kaks esimest küsimust. Lisaks koostasid RR lugemissaalide töötajad näidisküsimused, mida võiksid kasutajad küsida. Küsimuste variatiivsuse tagas koostajate suur arv (ca 10). Küsimuste koostamiseks eraldi mingist kindlast juhendist ei juhitud. Teemade valik (<https://tinyurl.com/margendamine>) põhines e-kirjade teemade analüüsil ja selle valiku koostas prototüüpiprojekti (mitte antud hanke) meeskond. Eelistati RR teenuste teemasid ja seejärel üldinfot ning suuremaid valdkondi (*kirjastajaportaal, haridusprogrammid*). Kirjade seest küsimuse/fraasi eraldamine osutus liiga ajamahukaks ja seda ei tehtud. Koostajad lisasid ka selliseid vorminguid, mida on küsitud vahetu suhtluse või telefoni teel.

Andmestik sobib juturoboti žanrisse. Pöördumised on jagatud 48 teemasse (135 näite puhul teema puudub). Teemades esineb üksikuid semantilisi duplikaate (nt *Ajalehed/Ajakirjad* ja *ajalehed*). Teemasid, mille kohta on piisavalt palju näiteid asutusesisese mudeli treenimiseks (ca 50 näidet), on kümne ringis.

Lisaks andis Rahvusraamatukogu Textale ka juturoboti Vesta 5813 andmed. Need sisaldavad nii treenimiseks kasutatud fraase ja küsimusi, kui ka juturobotile väliskasutajate poolt kirjutatud lausungeid. Teemasid, mille kohta on asutusesisese mudeli treenimiseks piisavalt näiteid, on 38. Kokku on teemasid 69. Teemad sisaldavad ka tüüpilisi juturoboti žanrile omapäraseid teemasid nagu *Kuidaslaheb, Head aega* ja *Misteed*. Teemad on hästi jaotatud ega kordu.

**Haigekassa** edastas Textale 79 pöördumist (küsimust), mis on jaotatud 21 teemasse. Pöördumised on kogutud käsitsi e-kirjadest, eraldades kirjadest kliendi muutmata küsimused. Andmeid saab kasutada asutuseüleisel pöördumiste klassifitseerimisel.

**Sotsiaalkindlustusamet** panustas treeningandmestikku 254 pöördumisega, mis on neile saadetud kodulehe "võta ühendust" vormi kaudu. Andmestik sobib pigem juturoboti treenimiseks, koosnedes küsimustest ja fraasidest, millele on määratud pöördumise

eesmärki detailselt kirjeldav intent ehk teema. Pöördumisi pole muul viisil muudetud kui et eemaldatud on isikuandmed. Kuna iga teema kohta on näiteid väga vähe (alla kümne), siis asutusesiseseks klassifitseerimiseks seda märgendust kasutada ei saa. Lisaks on andmed veel jaotatud 11-sse ülemkategoriasse (20 näitel puudul sellekohane märgendus) ja 82-sse kategooriasse (22 märgend puudu). Nagu ka Keskkonnaagentuuri puhul, on ühte teemat kirjutatud erineval viisil (nt *staaaz* vs *Staaaz*), mis muudab teemade arvu kunstlikult kõrgeks.

Asutuseülese klassifitseerija paremaks treenimiseks küsisime juurde märgendamata pöördumisi. Neid anti meile kahes osas. Esimeses on 194 eelneva andmestikuga samast kanalist saadud märgendamata pöördumist. Teise osana anti meile SKA kodulehekülje tagasisidevormi automaatne väljavõte. Parsisime sealt *Kommentaari* väljalt eestikeelsed (eraldatud *langdetect*-iga<sup>7</sup>) kommentaari. Valisime pikemad kui 15 tähemärki kommentaari ning eraldasime mõningaid märksõnu sisaldavaid kommentaare, mis sisuliselt ütlesid, et tagasisidet ei ole ja koduleht andis piisavalt infot. Niiviisi saime kätte need kommentaarid, millega pöörduti otsese küsimusega SKA poole (mõni neist sisaldas ka nime ja telefoninumbrit!). Andmestik võib veel sisaldada vähesel määral "koduleht on OK" stiilis kommentaare. Selles andmestikus on 896 küsimust/fraasi.

**Vestluslogid** on kõige suurem meile edastatud märgendatud andmestik, koosnedes 23 tuhandest pöördumisest. Kuna tegemist on juturoboti poole pöördumistega, siis on need pöördumised lühikesed, koosnedes mõnikord isegi ühest sõnast ja tihti omades pöördumise teema tuvastamise kontekstis mitte mingisugust tähendust. Pöördumistest 10 910 näidet on märgendamata ja 11 949 näidet ehk 52% andmetest on märgendatud. Sagedaseimad teemad on *Üldised küsimused*, *Koroonaviirus*, *Hetkeseis riigis*, *Reisimine* ja *Arv*. Andmetesse sisse vaadates ilmneb, et märgendus pole kuigi adekvaatne, näiteks on teemaga *Arv* märgendatud pöördumised *Millega sa aidata oskad?* ja *kas sa tunnend onu roobertit* ja teemaga *Reisimine* märgendatud *mul on halb tuju ja asoo, mis sa plärad*. Üleüldiselt jätab märgenduse kvaliteet halva mulje ja masinõppes teemade ennustamiseks see sellisel kujul ei sobi.

**Statistikaamet** andis 205 ekirja 2021. aasta I kvartalist. Sisse jäid vaid need pöördumised, millele STAT vastas. Vähendati ülekaalukalt koroonaga seotud teemasid ja korduvaid surmade küsimusi. E-kirjadest eemaldati viited isikule (kirja lõpud või tutvustused alguses). Huvitav on märkida, et nende poole pöörduvad nii eraisikud kui ka riigiasutused. Kirjad koosnevad vahemikus 01.2021-03.2021 aadressile [stat@stat.ee](mailto:stat@stat.ee) saabunud kirjadest. Exceli tabelisse tõsteti need käsitsi Inboxist. Eemaldati kohad, milles on nimi: reeglina kirjade jalused, aga ka algused (tervitused ja tutvustused). Kõik pöördumised on

<sup>7</sup> <https://pypi.org/project/langdetect/>

originaalpöördumised ning müra, st mittesisulised pöördumised jäeti välja.

Statistikaamet panustas veel andmestikku 17505 juturoboti klassifitseeritud pöördumisega. Andmestik sisaldab ainult märgendatud kliendipöördumisi ehk lausungeid ega sisalda juturoboti enda vastuseid. Kaheksakümnel teemal on vähemalt 50 näidet. Enamus teemasid on statistikaametispetsiifilised, kuid leidub ka standardseid juturobotiteemasid nagu *tere*, *roppus*, *KasRobot*, *tänu*, *Armas* (*ma armastan sind ja oled ilus pöördumiste klass*).

Andmebaasi ei kuulu ühtegi transkribeeritud kliendikõnede kogu.

## Andmete puhastamine ja eeltötluse juhend

Andmete puhastamise ja eeltötlumisega seoses suurim väljakutse on asjaolu, et kõik asutused on enne oma andmete Textale edastamist juba mingit eeltötlust ja puhastamist rakendanud. Põhimotivatsioon selleks on olnud andmekaitseõuetest tulenev vajadus viia andmed isikustamata kujule. Enamik märgendatud e-kirjadest või nendest eraldatud pöördumistest (PPA, KKA, TA, RIA, HK, SKA) on anonümiseeritud käsitsi, kustutades lihtsalt kirja seest nõutud informatsiooni ning eemaldades kirja päised ja jalused. Vestlusrobotite (Vesta, Kriisi juturobot) logisid ei ole anonümiseeritud, kuid seda pole ka vaja, sest pöördumised ei tohiks sisalda isikuandmeid.

Transpordiameti märgendamata andmekogus on käsitsi järeltoimetusega anonümiseerimise käigus asendatud aja-, koha- ja nimejärjendid sümbolitesüsteemiga, kuid täpne loogika ja algoritm pole siinkohal teada.

TTJA andmed edastati Textale toorel kujul ning seega töödeldud praegusele kujule täiesti nullist - alustades postkastide väljavõtete parsimisest ning lõpetades HTML-ist ja CSS-ist puhastamisega.

Viimase sammuna rakendati andmetele MLP-d<sup>8</sup> (*Multilingual Preprocessor* ehk mitmekeelset eeltötlejat), mis tokeniseerib ja lemmatiseerib tekstid.

Millistele nõuetele peaksid alusandmed vastama, et neid saaks mudeli(te)s kasutada? Kas ja millises ulatuses erikanalitest tulnud pöördumised vajavad eeltötlust?

Kanalist sõltumata tuleb **UTF-8 kodeeringus** andmed enne treenimist **puhastada** HTML-i, CSS jms formaadist tavatekstile, **tokeniseerida ja lemmatiseerida või stemmida**<sup>9</sup>.

<sup>8</sup> Vt <https://docs.texta.ee/et/mlp.html> või lisast 4.

<sup>9</sup> BERT mudelite puhul pole seda tarvis. BERT mudeli algoritmi arendas 2018. aastal Google ning see vajab suuri andmemassiive. Seda saab kasutada ka klassifitseerimiseks.

Tokeniseerimine on teksti tokeniteks ehk väiksemateks tekstiühikuteks tükeldamine. Tegu ei ole lihtsa ülesandega, kuna näiteks ainult tühikutega lõikumine võib anda arvuti jaoks erinevad tähejärjendid nagu „tere!”, „tere,” ja „tere”. Lemmatiseerimine on sõna sõnastikuvormi kujule viimine (sõnad *sõi*, *sööksin* ja *söönud* saavad kuju *sööma*), stemmimine võtab ära sõnade käände- ja pöördelõpud, vähendades samuti sõnade varieeruvust: arvuti jaoks erinevad sõnad saavad ühe kindla kujuga sõnaks. Kuigi stemmimise eelis lemmatiseerimise ees on selle märgatav kiirus suurte andmemassiivide töötlemisel, on lemmatiseeritud teksti kasutajal lihtsam analüüsida. Kõike kolme saab teha näiteks Texta Toolkitis (TTK-s) ning `texta-mlp`<sup>10</sup> Pythoni paki abil saab tokeniseerida ja lemmatiseerida.

Väga oluline on tähele panna, et üleriigilise mudelitesüsteemi puhul, kus erineva asutuse mudelid ennustavad pöördumisele oma märgendeid ja võidab enesekindlaim (vt ptk Prototüübi arhitektuur), peavad mudelid olema treenitud sama tokeniseerijaga tokeniseeritud ning lemmatiseerijaga lemmatiseeritud andmetel. Selleks soovitame arhitektuuri arvestada ka **ühtset eeltöötlust** (näiteks `texta-mlp` pakki), mis sõltuvalt asutustest võib arvestada ka erivajadustega (asutusesiseselt võib tekkida vajadus lisada eeltöötlemisse ühest formaadist teise konverteerimine vms, st ühtsesse tokeniseerimisse ja lemmatiseerimisse läheb juba puhastatud tavatekst). Ühtne eeltöötlus on oluline, kuna juhul, kui erinevad mudelid on treenitud erinevalt eeltöödeldud andmete peal, võivad mudelid hakata ennustama mitte sisu vaid töötlemise eripärade põhjal.

Esialgset treeningandmestikku koostades tuleb silmas pidada üht masinõppe põhireeglitest: *treeningandmestik peab olema täpselt selline nagu tulevikus sisend olema hakkab*. See tähendab seda, et treenides lemmatiseeritud teksti peal peab ka sisend olema lemmatiseeritud. **Uus, ennustusele minev pöördumine peab olema läbinud sama eeltöötlemise, mis treeningandmedki**. Seda, kas ja kuidas erinevaid pöördumise žanreid saab ühes mudelis korrata treenida, uurib peatükk Asutusteülene klassifitseerimine.

Mudeli treenimisele eelneb ka stopp-sõnade loetelu koostamine. Sinna kuuluvad sõnad, mida algoritm treenimise ning ennustamise ajal ignoreerib (vt Terminoloogiat). Selleks, et vältida olukorda, kus mõni mudel võtab tunnuseks tühisõnu, mida teised mudelid ignoreerivad, ning tõstab seetõttu valepositiivide hulka, soovitame kasutada stopp-sõnade loetelu põhja (vt nt lisa 2), mis sisaldab klassikalisi stopp-sõnu. Treenimise ajal saab tunnuseid uurida ning mudelispetsiifiliselt stopp-sõnade loetelu täiendada.

Kindlasti ei tasu unustada, et kirjadega kaasnevad lisaks sisule ka manused ja metaandmed,

<sup>10</sup> <https://pypi.org/project/texta-mlp/>

mis võivad klassifitseerimisel olulisi tunnuseid sisaldada. Näiteks võib kirja sisu olla „*Tere! Edastan manuses oleva dokumendi. Terv. Mari*”, mille puhul saab alles manuses oleva dokumendi järgi otsustada, kas tegu on näiteks avalduse või taotlusega. Samuti võib infot anda saatja aadress (nt PPA-s lähevad @kohus.ee lõpuga saatja aadressilt tulnud kirjad kohe kindlasse kausta) või kirja teema (näiteks kiri pealkirjaga *Taotlus kalapüügi töötleva praktikatoetuseks* PRIA-s). Antud PoC-i raames me lisaandmetega ei tegele, kuid lõplikus töövoos tuleb arvestada sellega, et ka **lisaandmed tasub võimalusel lisada ning siis tuleb neid samamoodi eeltöödelda** (kergeima viisina lisades manuse sisu või pikema manuse korral selle alguse / lõike sellest lihtsalt kirja lõppu).

TTK toetab järgnevaid formaate: JSON Lines, CSV, Excel. Nendes formaatides andmeid saab kõige hõlpsamini andmed kanda TTK-sse sisu uurimiseks, analüüsimiseks, treening- ja testandmete koostamiseks ning klassifitseerijate treenimiseks ja viimase tulemuste analüüsimiseks. E-mailide puhul sobivad formaadid .mbox, .pst ja .eml. Lisaks suudab Texta Filescanner TTK-sse parsida järgnevad formaadid: .ddoc, .bdoc, .asice, .asics, .doc, .docx, .odt, .pdf, .rtf, .htm, .html, .epub, .txt, .jpg, .jpeg, .gif, .png, .bmp, .tiff, .csv, .xls, .xlsx.

# Asutusteülene klassifitseerimine

Eelnimetatud andmete põhjal lõime asutusteüleseid klassifitseerijaid. Masinõppes on õige meetodi leidmiseks alati vaja teha erinevaid katseid. Käesolev peatükk annab ülevaate tehtud katsete parimatest tulemustest ja kirjeldab asutuse- ja žanriülese klassifitseerimise võimalikkust ning tingimusi.

## Lineaarsed klassifitseerijad

OVR (OneVsRestClassifier<sup>11</sup>) on märgendaja, mis õpib mitme märgendi vahel (antud olukorras asutuste nimed) valima õiget märgendit, luues binaarseid mudeleid, kus negatiivideks võetakse kõik ülejäänud asutused.

Vaatame esialgu ainult emaille. Katsete alustamise ajal olemasolevad andmed jagati treenimis- ja testandmestikuks. Treenimisandmestikul treeniti mudeleid ja osa sellest läks mudeli treeningaegseks valideerimiseks. Mudel ei näinud treenimise ajal kordagi testandmestikku ja sellel hinnatakse tegelikku mõju, vältimaks ületreenimisest saadud optimistlikke näitajaid. Treening- ja testandmete jaotus esimestes katsetes, mil koguandmestik polnud veel kättesaadav, on toodud tabelis 5.

Selleks, et TTJA suur ülekaal andmestikus mõju avaldama ei hakkaks (ilma spämmita on andmestikus 5996 kirja), võtsime TTK Reindexeri abil välja juhuslikult valitud 800 ekirja. Testandmestikus klasside osakaalu valides peeti silmas klasside reaalselt osakaalu andmestikus, aga ka seda, kui palju jääb sel juhul treeningandmetes näiteid alles. KKA andmestikust on välja jäetud need kirjad, millele oli määratud rohkem kui üks asutus (nt saatis KKA kirja edasi Eesti Keele Instituudile või Päästeametile).

Treenisime antud andmetel erinevaid mudeleid, uurides, milline vektoriseerija ning algoritmi kombinatsioon TTK Taggeris annab lemmatiseeritud testandmetel kõige paremaid tulemusi ja valisime selle põhjal oma **baasmudeliks TF\_SVC\_1**, mis kasutab lineaarset SVC algoritmi (ingl *linearSVC*) ja Tfidf vektoriseerijat (ingl *Tfidf vectorizer*), õpib ennustama lemmatiseeritud teksti peal ning treeningandmestikus on negatiivsete ja positiivsete näidete suhe 1 : 1. Selle tulemused testandmestikul on toodud tabelis 6 ja segadusmaatriksis joonisel 7.

11 <https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html>



Asutus	Treeningandmed	Testandmed
TTJA	700	100
RIA	662	100
PPA	547	100
KKA	296	50
SKA	200	43
TA	60	18

Tabel 5: Klasside ehk asutuste jaotus treening- ja testandmetes

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.927	0.776	0	0.845
PPA	0.886	0.939	0	0.912
RIA	0.689	0.919	1	0.788
SKA	0.952	0.465	0	0.625
TTJA	0.927	0.918	3	0.922
TA	1	0.333	0	0.5

Tabel 6: Mudeli TF\_SVC\_1 tulemused testandmestikul

Tegelik märgend	KKA	PPA	RIA	SKA	TTJA	TA	no_label
KKA	38	2	7	0	2	0	0
PPA	0	93	5	0	1	0	0
RIA	0	5	91	1	1	0	1
SKA	1	1	19	20	2	0	0
TTJA	1	0	4	0	89	0	3
TA	1	4	6	0	1	6	0
no_label	0	0	0	0	0	0	0
Ennustatud märgend	KKA	PPA	RIA	SKA	TTJA	TA	no_label

Joonis 7: Mudeli TF\_SVC\_1 segadusmaatriks testandmestikul

Segadusmaatriksist joonisel 7 on näha, et mudelil on **raskusi RIA ja SKA pöördumiste eristamisel**. Kuigi TA märgend on pandud ainult TA pöördumistele, on siiski üle poole TA pöördumistest saanud mõne teise sildi. F1 on suhteliselt madal TA, SKA, RIA ja KKA märgendil. RIA puhul on see üllatav, kuna võiks eeldada, et rohkemad andmed tõstavad

täpsust. Selgust toob RIA ja SKA jaoks eraldi treenitud binaarsed mudelid - nende tunnused kattuvad. Näiteks on mõlema mudeli tunnuseks sõnaüksused *laps, teadma, tud, uus ja küsimus*. Sisuanalüüs näitab, et kui RIA-s küsitakse lapsega seoses dokumentide tegemise võimalust, siis SKA-s pensioni sõltuvust laste arvuga. Taolist tunnuste kattumise probleemi sõnadega, mis tegelikult ei ole semantiliselt märkimisväärsed klassitunnused, aitab lahendada korralik stoppsõnade sõnastik. Stoppsõnad on sõnad, millega mudel tunnuste valimisel ei arvesta.

## Eraldatud fraaside lisamise mõju meili pöördumiste ennustamisele

Enne tunnuste uurimist ja stoppsõnastiku loomist loome uue mudeli **TF\_SVC\_1\_fraasidega**, kus lisame samale baasmudelile treeningandmestikku TTJA 73 käsitsi eraldatud fraasi ning 213 KKA käsitsi eraldatud fraasi. Need fraasid on käsitsi eraldatud küsimused või märksõnad kliendipöördumistest ning on mõeldud juturoboti treenimiseks. 10 KKA fraasi korduvad juba emailides. Nendest kaks on testandmestikus olemas. Sellega uurime, kas ja kuidas on võimalik eri liiki pöördumisi ühes mudelis kasutada. **Vaata järeldusi lk 37.**

Tabelist 7 ja segadusmaatriksist joonisel 8 on näha, et fraaside lisamine treeningandmetele suurendas või jättis samaks KKA, PPA ja TTJA saagise ning PPA, RIA, TTJA ja TA täpsuse (hinnantud samal testandmestikul, nagu eelmine mudelgi). Samas süvenes RIA ja SKA eristamise probleem, kuid selle parendamiseks tuleb lisaandmete lisamise asemel uurida tunnuseid. Nendest tulemustest võime järeldada, et emailide klassifitseerimiseks võib treeningandmetes vähesel määral kasutada ka emailidest eraldatud fraase.

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.849	0.918	0	0.882
PPA	0.889	0.97	0	<b>0.927</b>
RIA	0.725	0.879	1	0.794
SKA	0.947	0.419	0	0.58
TTJA	0.927	0.918	3	<b>0.922</b>
TA	1	0.278	0	0.435

Tabel 7: Mudeli TF\_SVC\_1\_fraasidega tulemused testandmestikul

Uurimise hetkel (märtsi lõpp) käesolevatest andmetest on meie käes ka kriisi juturoboti vestluslogid. Need on päris juturobotile suunatud lausungid (ingl *utterance*), peamiseks teemaks käitumine eriolukorra ajal. Eraldasime sealt 582 PPA-le suunatud lausungit ning 7 TTJA-le suunatud lausungit (andmestikus oli juba osad lausungid (tõenäoliselt automaatselt) PPA ja TTJA asutuse märgendiga märgendatud). PPA eristamine teistest asutustest on juba praegu üsna hea, kuid vaatame, mis juhtub, kui teemasid ning pöördumisi teises žanris juurde tuleb. Oluline on tähele panna, et nüüd on PPA mudelil

TF\_SVC\_1\_fraasidega\_ja\_kriisibotiga tugevalt ülekaalus. Mudeli TF\_SVC\_1\_fraasidega\_ja\_kriisibotiga tulemused testandmestikul on toodud tabelis 8 ja joonisel 9.

Tegelik märgend	KKA	45	0	2	0	2	0	0
	PPA	0	96	3	0	0	0	0
	RIA	2	5	87	1	3	0	1
	SKA	3	2	18	18	2	0	0
	TTJA	1	1	3	0	89	0	3
	TA	2	4	7	0	0	5	0
	no_label	0	0	0	0	0	0	0
no_label	KKA	PPA	RIA	SKA	TTJA	TA	no_label	
		Ennustatud märgend						

Joonis 8: Mudeli TF\_SVC\_1\_fraasidega segadusmaatriksi testandmestikul

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.886	0.796	0	0.838
PPA	0.72	0.96	0	0.822
RIA	0.779	0.818	1	0.798
SKA	0.947	0.419	0	0.58
TTJA	0.918	0.928	3	0.922
TA	1	0.222	0	0.363

Tabel 8: TF\_SVC\_1\_fraasidega\_ja\_kriisibotiga tulemused testandmete

Tegelik märgend	KKA	39	4	5	0	1	0	0
	PPA	0	95	4	0	0	0	0
	RIA	1	13	81	1	2	0	1
	SKA	1	12	9	18	3	0	0
	TTJA	1	2	1	0	90	0	3
	TA	2	6	4	0	2	4	0
	no_label	0	0	0	0	0	0	0
no_label	KKA	PPA	RIA	SKA	TTJA	TA	no_label	
		Ennustatud märgend						

Joonis 9: Mudeli TF\_SVC\_1\_fraasidega\_ja\_kriisibotiga segadusmaatriksi tulemused testandmetel

Näeme, et PPA-d aetakse rohkem sassi RIA ja SKA-ga. Siin võib rolli mängida asjaolu, et tunnused ehk märksõnad, mille põhjal klassifitseerija klassi kuulumist otsustab, kattuvad. Samuti selgus selle mudeli analüüsi käigus, et RIA andmed sisaldavad ka neid pöördumisi, mille puhul nad kliendil soovitasid pöörduda õige asutuse poole (ei saatnud edasi andmekaitselistel põhjustel). Seetõttu jätame RIA andmestiku järgnevatest katsetest kuni korrastamiseni välja.

Katsed eelmiste mudelitega näitavad, et **väikse arvu fraaside lisamine parandas klassifitseerija tulemusi ning juturoboti logide lisamine halvendas**. Siin tuleb arvestada ka mahtude küsimust - klassid on tasakaalust väljas ning see võib mõjutada tulemusi. Klasside näidete arvu mõju uurime ülejäärgmises peatükis, kui meile on kättesaadavad kõik projekti andmed.

Treenime TF\_SVC\_1\_fraasidega uuesti, aga RIA andmeteta (TF\_SVC\_1\_fraasidega\_RIAta). Meeletuletusena on antud mudeli test- ja treeningandmed erinevad baasmudeli andmetest. Jätkuvalt on treeningandmetes lisaks ekirjadele ka fraasid. Testandmed on samad, v.a asjaolu, et RIA näited võtsime mõlemast andmestikust välja.

Asutus	Treeningandmed	Testandmed
TTJA	700 ekirja + 73 fraasi = 773	100
PPA	547	100
KKA	296 ekirja + 182 fraasi = 478	50
SKA	200	43
TA	60	18

Tabel 9: Treening- ja testandmestik

Tabelis 10 ja joonisel 10 on toodud uue mudeli tulemused. Tabel 10 ei ole tabeliga 7 otseselt võrreldav, kuna absoluutarvud on RIA eemaldamise tõttu erinevad. Suurima tähelepanu tegi TTJA saagis. Samas tekkis TTJA märgendiga valepositiivseid juurde. Kadunud on SKA pea pooltelt juhtudel RIA klassi ennustamise probleem, kuid seda aetakse siiani teiste klassidega (KKA, PPA ja TTJA-ga) segi. Ka TA-l on jätkuvalt nõrgad tulemused, mille tõenäoliselt põhjustab TA andmete vähesus.

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.759	0.936	0	0.838
PPA	0.897	0.98	0	0.936
SKA	1	0.605	0	0.753
TTJA	0.905	0.96	3	0.931

TA	1	0.333	0	0.499
----	---	-------	---	-------

Tabel 10: Mudeli TF\_SVC\_1\_fraasidega\_RIAta tulemused testandmestikul

Tegelik märgend	KKA	PPA	SKA	TTJA	TA	no_label
KKA	44	1	0	2	0	0
PPA	1	96	0	1	0	0
SKA	7	5	26	5	0	0
TTJA	1	0	0	95	0	3
TA	5	5	0	2	6	0
no_label	0	0	0	0	0	0
	Ennustatud märgend					
	KKA	PPA	SKA	TTJA	TA	no_label

Joonis 10: Mudeli TF\_SVC\_1\_fraasidega\_RIAta segadusmaatriks testandmestikul

Millisel määral saab mudeli treenimisel kasutada ühiselt vähemalt e-maile ja vestlusakende (sh juturoboti) pöördumisi?

Eelnevate katsetega uurisime, kas ja kuidas saab asutuseülese mudeli treenimisel kasutada ühiselt erinevat liiki pöördumisi. Transkribeeritud kõnede puudumise tõttu ei saanud me selle liigi eripärasid ja mõju siinsete katsetega hinnata (vt peatüki lõpust selle hinnangulist mõju). Uurimiseks on meil kahte liiki andmeid: käsitsi e-mailidest eraldatud TTJA ja KKA fraasid (vastavalt 73 ja 128 fraasi) ning kriisiboti lausungid, kus on (tõenäoliselt automaatselt) märgendatud PPA ja TTJA asutusele kuuluvad lausungid (vastavalt 582 ja 7 tk). Vaid nende asutuste puhul oli meil olemas nii näidis e-kirju kui ka juturobotile sobivaid fraase või juturoboti lausungeid. Seda on põhjalike katsete jaoks väga vähe, kuid juba nende põhjal oli näha, et väikse arvu fraaside lisamine parandas klassifitseerija tulemusi ning juturoboti logide lisamine halvendas.

Seega võiks justkui antud katsete põhjal öelda, et juturoboti logisid ei saa kasutada e-kirjadega ühises mudelis. Tegelikult päris nii väita ei saa. Lähemal vaatlusel selgub, et kriisiboti logist eraldatud lausungid sisaldavad palju tühilausungeid nagu *Kes tegeleb noortegruppidega?*, *saared kinni*, *koosoleku korraldamine?*, *Kas augustikuus on mõistlik*

*sõita Pärnusse?*, mis tegelikult otseselt asutuste antud e-kirja teemadele või üleüldse kindlale asutusele ei viita. Seega on loomulik, et segane treenimismaterjal tõmbas mudeli tulemused alla: lühikesed lausungid ei anna (ka inimotsustajale) otsustamiseks piisavalt konteksti ning selgelt kindla asutuse lausungid ei kattu nende e-kirjade teemadega, millel mudelit testiti. Kuna hoolikama kontrolli läbinud käsitsi eraldatud fraasid tegelikult tõstsid tulemusi e-kirjade klassifitseerimise, võib väita, et **vestlusakende pöördumisi saab kasutada e-kirjade asutuseülese klassifitseerijate treeningmaterjalis tingimusel, et need pöördumised on puhtad**. Puhas tähendab seda, et antud pöördumine tegelikult ka sisaldab pöördumise põhjust. Juturoboti puhul võib juhtuda, et juturoboti esimesed versioonid alles õpivad ning segadusse ajavad lausungid saavad enda külge vale märgendi, mistõttu ei pruugi automaatsed märgendid olla piisavalt puhtad.

Selle, kuidas mõjutab e-kirjade lisamine juturoboti klassifitseerijate treeningmaterjali, jagus uurimiseks vähe andmestikku. Siiski võib sellel andmestikul tehtud katsete põhjal öelda, et **e-kirjade pöördumisi saab kasutada vestlusakende asutuseülese klassifitseerijate treeningmaterjalis**: e-kirjadel ja fraasidel / juturoboti lausungitel treenitud mudel klassifitseeris väga hästi, ennustades õigesti 85% KKA, 94% PPA ja 96% TTJA treenimisel nägemata fraasidest või lausungitest.

Tähele tuleb panna, et **juturoboti ja e-kirjade klassifitseerijad töötavad erinevalt**: kui juturobot võib kasutada oma tõlgenduses vestluse mälu, siis e-kirjade klassifitseerija saab kogu info otsustamiseks korruga kätte. Seega võib juturoboti ja e-kirjade klassifitseerija arhitektuurist lähtudes juhtuda, et mõistlikum on nende mudelite treenimised hoida lahus.

## Stoppsõnad

Käesolevas alapeatükis uurime stoppsõnade mõju tulemustele. Eelneva peatüki viimasele mudelile rakendatakse stoppsõnad, mis on käsitsi kaevandatud TEXTA Toolkiti Lexicon Mineriga<sup>12</sup>. Seemnesõnad valiti binaarsete mudelite tunnuste analüüsi ning eestikeelsete stoppsõnade loetelu<sup>13</sup> abil. Saadud stopp-sõnad on toodud lisas 2. Kuna mudel kipub olulisteks tunnusteks panema ka numbreid (nt aastaarve ja kuupäevi), keelame tal selle ära: aastaarv ei ole piisav vihje pöördumise mingile asutusele edasisaatmiseks.

Kuna andmed on anonümiseeritud, ei pea muretsema, et mudel isikuandmeid tunnusteks võtab või neid hiljem, pärisandmetel klassifitseerides, oluliseks peab.

Tulemused on toodud tabelis 11 ja segadusmaatriksis joonisel 11. Tulemused halvenesid silmnähtavalt. Üheks põhjuseks on andmete erinev käsitsi eraldamine: näiteks PPA ja TTJA

12 [Vt https://docs.texta.ee/lexicon\\_miner.html](https://docs.texta.ee/lexicon_miner.html) või lisast 4.

13 <https://datadoi.ee/handle/33/78>

kirjades olid jalused alles, mistõttu sai sõna *lugupidamine* oluliseks tunnuseks. Ilmselgelt hakkavad tulevikus kõik kirjad sisse tulema täisvormis (st tervituste ja jalustega), mistõttu panime nende võimalikud vormid stoppsõnadesse. Tulemuste halvenemine näitab, et eelnevad mudelid ülesobitasid end valedete sõnadele ehk algoritm võttis tunnusteks sõnu, mis tegelikult klassi näitama ei pea. Näiteks osad e-kirjad andmestikus on koos jalustega (*heade soovidega, lugupidamisega*), osad mitte. Mudel õpib selgeks, et jalustega e-kirjad kuuluvad PPA-le. See aga on puhastamise ja parsimise küsimus, mida saab hõlpsasti lahendada stoppsõnadega. Seetõttu on praegune mudel parem.

Tulemustest on veel näha, et kõige raskem on mudelil märgendada SKA ja TA pöördumisi õigesti. Mõlema puhul mängib rolli väike treening- ja testandmestik. Samuti on mudeli tunnuseid vaadates märgata, et enamus tunnustest on ilmaste (Keskkonnaamet) kohta. Klasside tasakaalutusega tegeleb järgmine alapeatükk.

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.726	0.9	0	<b>0.803</b>
PPA	0.897	0.96	0	<b>0.927</b>
SKA	0.958	0.535	0	0.686
TTJA	0.855	0.94	3	<b>0.895</b>
TA	1	0.278	0	0.435

Tabel 11: Stoppsõnadega mudeli tulemused testandmestikul

KKA	45	2	1	2	0	0
PPA	1	96	0	3	0	0
SKA	8	4	23	8	0	0
TTJA	3	0	0	94	0	3
TA	5	5	0	3	5	0
no_label	0	0	0	0	0	0
	KKA	PPA	SKA	TTJA	TA	no_label

Ennustatud märgend

Joonis 11: Stoppsõnadega mudeli segadusmaatriks testandmestikul

## Klassi tasakaalutus

Aprillikuu jooksul lisandus andmeid juurde. Uutest andmetest koostasime järgneva uue andmestiku (tabelis 12). Kuna TTJA spämmita märgendatud kirjade arv (pea 6000) on

teistest mitmekordselt üle, eraldasime TTK Reindexeriga juhuslikkuse alusel 1500 kirja, millele lisasime ka 73 käsitsi eraldatud fraasi. Sotsiaalkindlustusamet (SKA) suutis oluliselt suurendada oma pöördumiste näidete arvu: 448 kodulehe kliendivormi pöördumist (sh 254 märgendatud teemade suhtes) ning 896 kodulehe tagasiside kommentaari. Viimasest parsiti välja eestikeelsed pikemad kommentaarid, mis ei sisalda stopp-sõnu, mis viitavad tühipöördumistele (tänamised, *koduleht arusaadav* jms). Andmestik ei ole täielikult tühipöördumistest puhas, kuid enamus pöördumistest sisaldavad siiski küsimusi, millele kliendid kodulehel vastust ei leidnud. Täpsemalt saab andmete kohta lugeda andmete peatükist.

Asutus	Treeningandmestik (85%)	Testandmestik (15%)
TTJA	1397 (sh 73 fraasi)	150
SKA	940	150
PPA	807	150
TA	473	100
KKA	389	100
RIA	100	24

Tabel 12: Test- ja treeningandmestik

Selle treeningandmestiku peal treeniti ning testandmestiku peal testiti uuesti TfIdf vektoriseerijaga LinearSVC algoritmi ning stoppsõnu kasutades lemmadel treenitud mudelit. Parameetrid on identsed eelmise alapeatüki mudeliga. Vaid treening- ning testandmed on muutunud. Testandmete muutuse tõttu ei saa tulemusi täpselt võrrelda, kuid üldist tendentsi võib küll kontrollida. Antud treeningandmestikul treenitud mudeli tulemused testandmestikul on toodud segadusmaatriksis joonisel 12 ja tabelis 13.

Tulemustest on näha, et saagised on kõrged PPA-l ja SKA-l. Kuigi SKA saagis on kõrge (vaid 13 SKA asutusele kuulunud pöördumist märgendati valesti), on mudel SKA määramisel liigentusiastlik: igast asutusest sai SKA märgendi neli või enam pöördumist. Samuti tekitab segadust Transpordiametile (TA) kuuluvate pöördumiste korrektne märgendamine: 9% SKA pöördumistest said TA märgendi. SKA on siin mudelis ainukene segadusetekitaja: muid asutusi teiste asutustega eriti sassi ei aeta. Siin võib rolli mängida SKA andmete kvaliteet, kuna osa andmestikust puhastati poolautomaatselt.

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.965	0.83	0	0.892
PPA	0.942	0.973	0	0.957



RIA	1	0.542	1	0.702
SKA	0.678	0.913	0	0.778
TTJA	0.899	0.773	0	0.831
TA	0.943	0.83	0	0.882

Tabel 13: Uuenenud andmetel treenitud mudeli tulemused testandmestikul

Tegelik märgend	KKA	83	2	0	10	4	1	0
	PPA	0	146	0	4	0	0	0
	RIA	0	0	13	6	3	1	1
	SKA	1	6	0	137	5	1	0
	TTJA	1	0	0	31	116	2	0
	TA	1	1	0	14	1	83	0
	no_label	0	0	0	0	0	0	0
	Ennustatud märgend	KKA	PPA	RIA	SKA	TTJA	TA	no_label

Joonis 12: Uutel andmetel treenitud mudeli segadusmaatriks testandmestikul

TTK pakub klasside tasakaalustamise võimalust näidete treeningandmestikus korrutamise näol. See on klassikaline võtte vähema näidetega klassidel tulemuse parandamiseks. Seda võimalust kasutades eelnevale samade parameetritega mudeli tulemused on toodud tabelis 14 ja segadusmaatriksis 13. Tasakaalustatud treeningandmestik tähendab, et klassi juhuslikke elemente kahekordistatakse nii kaua, kuni klassielementide arv on võrdne suurima klassi elementide arvuga (antud juhul TTJA 1397).

Tulemused paranevad kõigil näitajatel. Eriti suure hüppe on teinud eelnevalt kõige väiksema andmestikuga olnud TA. Ainukesena halvenesid tulemused TTJA-l, mida mudel ajas SKA-ga segi: igale viiendale TTJA pöördumisele määrati SKA märgend. Mõned näited taolistest pöördumistest (loetavuse nimel lemmatiseerimata): *Minu ostetud asi ei tööta. Mida teha?*, *Mis ma pean tegema selleks, et raha tagasi küsida?*, *Kust ma saan infot kuhu pean pöörduma kui mul on vaidlus ostu või teenuse osas?*

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.776	0.9	0	0.833
PPA	0.942	0.973	0	<b>0.957</b>
RIA	0.8	0.667	1	0.727
SKA	0.761	0.913	0	0.83
TTJA	0.943	0.66	0	0.776
TA	0.887	0.86	0	0.873

Tabel 14: Balansseeritud andmetel treenitud mudeli tulemused testandmestikul

Tegelik märgend	KKA	PPA	RIA	SKA	TTJA	TA	no_label
KKA	90	1	0	5	2	2	0
PPA	1	146	0	2	0	1	0
RIA	1	0	16	3	2	1	1
SKA	6	4	0	137	2	1	0
TTJA	9	2	4	30	99	6	0
TA	9	2	0	3	0	86	0
no_label	0	0	0	0	0	0	0

Ennustatud märgend

Joonis 13: Balansseeritud andmetel treenitud mudeli segadusmaatriks testandmetel

## Sõnavektorite lisamine

Sõnavektor on sõna numbriline kodeering, mille abil saab arvutile (antud juhul mudelile treeningandmestiku lisatunnustena) edasi anda sõnade tähendusi. Lisasime eelneva alapeatüki parimale mudelile (andmete balanseerimist kasutav mudel, vt tabel X) erinevaid, kogu andmestikul treenitud sõnavektoreid lisatunnusteks. Word2Vec mudeliga koostatud sõnavektorite lisamine parandas silmnähtavalt tulemusi (tabel X ja joonis Y).

TTJA pöördumiste eristamine SKA pöördumistest on siiani mudelile kõige keerulisem: muuga saab hästi hakkama. Neile valesti märgendatud pöördumistele peale vaadates ei jää silma mingit kindlat korduvat joont, mida võiks SKA-ga sassi ajada. Samuti ei esine paljudes nendes pöördumistes sõnu, mis on OneVsRest mudeli tunnuste seas. Kõige paremini saab mudel hakkama PPA märgendamisega (F1 on lausa 0,96). Ka SKA, TA ning KKA tulemused

on väga head F1 skooridega, mis jäävad üle 0,8.

Märgend	Täpsus	Saagis	Märgendamata	F1
KKA	0.812	0.91	0	0.858
PPA	0.954	0.973	0	0.963
RIA	0.842	0.667	1	0.744
SKA	0.735	0.907	0	0.811
TTJA	0.954	0.687	0	0.798
TA	0.917	0.88	0	0.898

Tabel 15: Word2Veciga mudeli tulemused testandmetel

Tegelik märgend	Ennustatud märgend						
	KKA	PPA	RIA	SKA	TTJA	TA	no_label
KKA	91	1	0	6	1	1	0
PPA	0	146	0	2	0	2	0
RIA	1	0	16	3	2	1	1
SKA	8	4	0	136	1	1	0
TTJA	5	2	3	34	103	3	0
TA	7	0	0	4	1	88	0
no_label	0	0	0	0	0	0	0

Joonis 14: Word2Veciga mudeli segadusmaatriksi testandmestikul

## Lineaarsed klassifitseerijad juturobotite andmestikul

Projekti käigus anti analüüsimiseks kolm suurt juturoboti andmestikku:

1. Rahvusraamatukogult 1 218+5 813 pöördumist Vestale,
2. Statistikaametilt 17 505 pöördumist nende juturobotile Itile,
3. Kriisi vestluslogi SUVI 22 859 lausungit.

Lisaks andis Haigekassa 80 ning TTJA 73 käsitsi koostatud lausungit juturoboti treenimise tarbeks. Andmete vähesuse tõttu jäävad need siit uuringust välja. Treenisime puhastatud andmetel binaarseid stoppsõnadega mudeleid lemmatiseeritud andmestikul, kus positiivseteks näideteks oli Statistikaameti pöördumised ning negatiivseteks Rahvusraamatukogu omad. Puhastamine tähendab siin seda, et jätsime välja teemad (ingl

*intents*), mis ei olnud asutusespetsiifilised (nagu *tervitus*, *Armas* (see viitas armastusavaldusele juturobotile), *headAega*). Samuti jäid välja teemad, milles oli alla kümne näite. Tulemused on näha allolevas tabelis.

Mudel	Täpsus	Saagis	F1
TF-IDF + LinearSVC	0.88	0.97	0.92
TF-IDF + LR	0.91	0.93	0.92

*Tabel 16: Juturoboti lausungitel binaarse asutuseülese mudelite tulemused testandmestikul*

Tulemustest on näha, et asutusteülene klassifitseerimine töötab ka juturoboti lausungite puhul ning seda saab teha (masinõppe praktikas eelistatum variant) ka žanrispetsiifiliselt. See tähendab, et nii juturobotite kui ka e-mailide klassifitseerimiseks on eraldi mudelid.

## Tehisnärvivõrkudel põhinevad klassifitseerijad

Piisava koguse treeningandmete olemasolul on tehisnärvivõrkudel põhinevatel klassifitseerijatel potentsiaali ennustustäpsuse poolest lineaarseid klassifitseerijaid märkimisväärselt ületada. Antud peatükis uuritakse, mis tulemusi on võimalik tehisnärvivõrkudega saavutada, arvestades projekti skoobis olevate andmete mahtusid.

### Teksiklassifitseerija `fastText`

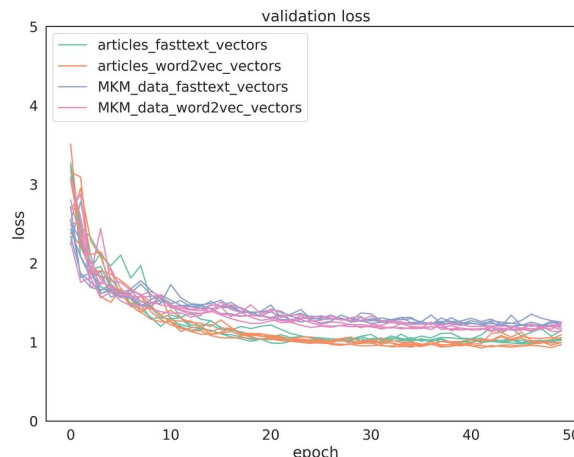
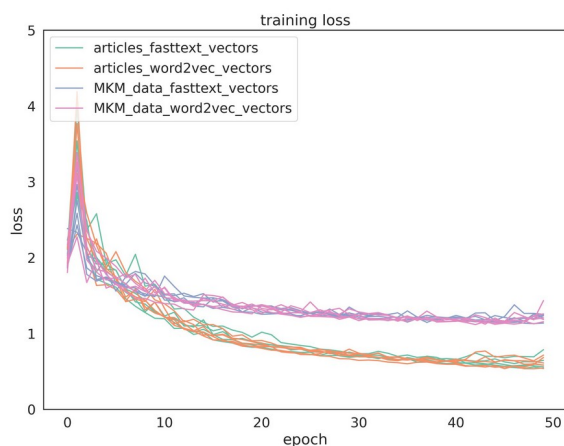
Teksiklassifitseerimisalgoritm `fastText` on Facebook AI Researchi loodud algoritm, mis, nagu nimigi viitab, on arvutuslikult üks kiiremaid meetodeid tekstide klassifitseerimiseks. On näidatud, et see töötab tihtipeale sama hästi kui süvanärvivõrkudel põhinevad klassifitseerijad, mis oma treenimisaja ja ressursivajaduse poolest on aga oluliselt nõudlikumad. Lihtsa arhitektuuri ja töökindluse tõttu kasutatakse ka `fastText`'i tihti baasmudelina võrdluses keerulisemate algoritmidega.

Mudel kasutab ennustamiseks eeltreenitud sõnavektoreid. Kuna nii sõnavektorite kvaliteet kui ka üldine sobivus antud ülesande lahendamiseks mõjutab ka lõpptulemust, eksperimenteerime siinkohal nelja erineva sõnavektorite süsteemiga. Vektorsüsteemid on saadud, kombineerides omavahel erinevaid treeningalgoritme ja alusandmestikke. Algoritmidest katsetame nii `fastText`<sup>14</sup> kui ka `word2vec` ning alusandmeteks on kas suur uudisartiklite korpus või käesoleva projekti tarbeks kokku pandud andmestik. Viimane neist on mahu poolest märgatavalt väiksem kui artiklite korpus. Artiklite korpus ei pruugi aga semantiliselt projekti andmetega nii hästi kattuda.

Kõigi vektorsüsteemidega treeniti klassifitseerijat 50 epohhi ja protsessi korrati 5 korda, et välistada juhuslikkusest tingitud kõrvalekaldeid. Ilmnes, et artiklikorpuse peal treenitud sõnavektoritega õpib mudel paremini ennustama. Seda näitavad tulemused nii treening- kui ka testandmestiku peal (vt graafikuid joonisel 15).

Treenimise käigus salvestati mudelid pärast 10, 25 ja 50 epohhi. Osutus, et valideerimisandmestikul on neist kõige madalama kahjuga (ingl *loss*) pärast 25 epohhi salvestatud mudel, seda nii `fastText` vektorite kui `word2vec` vektorite puhul. Lisaks, tabelis 17 toodud tulemustest on näha, et konkreetse vektorsüsteemi valik ei mõjuta oluliselt lõpptulemusi: `fastText` vektoreid kasutades on mudelite keskmine f1 skoor 0.612 ja `word2vec` vektorite korral 0.605 ehk erinevus on ainult 1.1%.

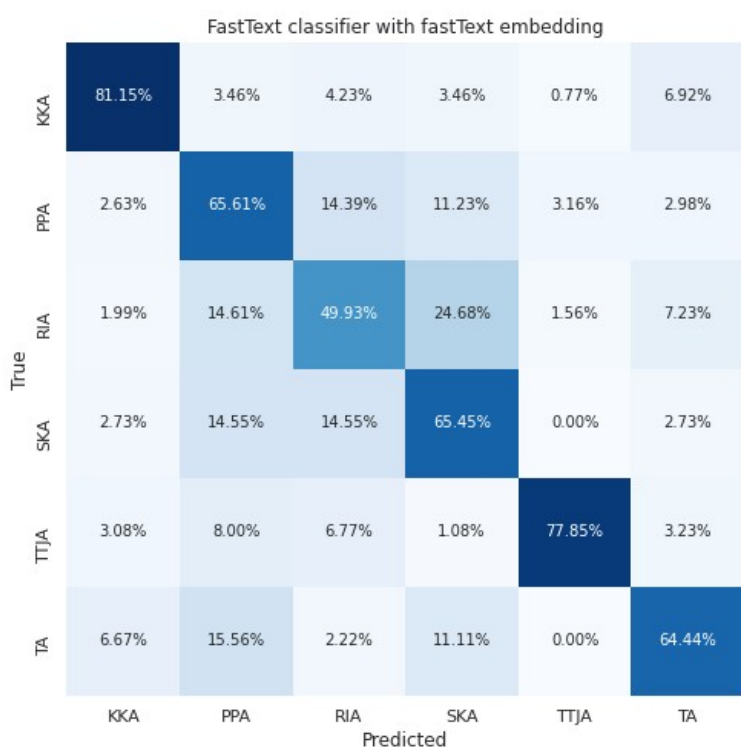
14 Mitte ajada `fastText` vektoriseerijat segamini klassifitseerimisalgoritmiga `fastText`.



Joonis 15: Erinevate vektorsüsteemidega klassifitseerijate kahju

Fasttext-il põhinevad klassifitseerijad						
	täpsus		saagis		F1	
märgend	fasttext	word2vec	fasttext	word2vec	fasttext	word2vec
KKA	0.793	<b>0.835</b>	<b>0.812</b>	0.8	0.802	<b>0.817</b>
PPA	0.655	<b>0.772</b>	<b>0.656</b>	0.596	0.656	<b>0.673</b>
RIA	<b>0.681</b>	0.635	<b>0.499</b>	0.489	<b>0.576</b>	0.553
SKA	<b>0.379</b>	0.352	0.655	<b>0.686</b>	<b>0.48</b>	0.465
TTJA	0.944	<b>0.964</b>	<b>0.778</b>	0.772	0.853	<b>0.858</b>
TA	<b>0.219</b>	0.168	<b>0.644</b>	0.578	<b>0.326</b>	0.261

Tabel 17: Fasttext-il põhinevate klassifitseerijate tulemused



Kõrvaloleval segadusmaatriksil (joonis 16) on toodud viie korduse keskmised tulemused pärast 25. epohhi fastText vektoritega klassifitseerijat kasutades. Tulemused on asutuse kaupa normaliseeritud vahemikku [0,1] ja viidud protsentkujule. On näha, et enim segadust on mudelil RIA, PPA ja SKA eristamisega. RIAGA seotud probleemid tulid juba eelmistes eksperimentides välja, seega pole need eriti üllatavad. Küll aga saab tõdeda, et võrreldes eelnevate eksperimentidega on tulemused

Joonis 16: fastText vektoritega klassifitseerija tulemused pärast 25. epohhi

testandmestikul madalamad.

## Transformeritel põhinevad klassifitseerijad

Transformer on tehismärgivõrgu arhitektuur, millele ehitatud ennustusmudelid on empiirilisel osutunud parimaks paljude erinevate keeletöötlusülesannete lahendamisel. Nende edu aluseks on kontekstuaalsed keelemudelid, mis on eelnevalt treenitud erakordselt suuremahulise andmekogu peal viisil, mis ei vaja märgendatud andmete olemasolu. Selliseid keelemudeleid saab kohandada spetsiifilisi ülesandeid lahendama mudeli peenhäälestuse abil ning on näidatud, et sel juhul heade ennustustulemuste saavutamiseks piisab juba väikesest arvust märgendatud andmetest.

Ka eesti keele jaoks on sellised keelemudelid juba olemas, neist üks näiteks EMBEDDIA projekti raames loodud FinEst BERT, mida ka antud projektis klassifitseerimisülesande jaoks peenhäälestati.<sup>1516</sup>

FinEst BERT fine-tuned for classification

True \ Predicted	KKA	PPA	RIA	SKA	TTJA	TA
KKA	86	0	0	0	14	0
PPA	0	134	1	0	13	2
RIA	0	0	7	0	15	2
SKA	5	1	0	126	17	1
TTJA	0	0	0	0	149	1
TA	4	0	0	0	11	85

Joonis 17: FinEst BERT-il 5 epohhiga treenitud klassifitseerija segadusmaatriks testandmetel mis tegelikult on müra.

Peenhäälestamise käigus treeniti keelemudeli peale klassifitseerija, kokku 5 epohhi. Treenimise käigus valideeriti mudelit iga 200 sammu järel ning testandmete tulemused on raporteeritud sellisel viisil parima leitud mudeli peal ja näidatud segadusmaatriksiga joonisel 17. On näha, et kui enamike asutuste ennustamisel teeb mudel üksikuid eksimusi, siis TTJA suunas saadaks see mudel väga palju kirju, mis tegelikult sinna ei peaks minema. Konkreetset põhjust on siinkohal raske öelda, kuid kaaluda tasub asjaolu, et TTJA andmed on palju vähem puhastatud ja mudel võib olla neist õppimisel leidnud tunnuseid,

15 <https://huggingface.co/EMBEDDIA/finest-bert>

16 Need katsed tehti FinEst BERT-ga, kuna need olid sel hetkel saadaval. Nüüd on avalikud ka TartuNLP EstBERT ja EMBEDDIA est-roberta, mis annavad häid tulemusi ja mida tasub edasistes katsetes kindlasti proovida.

## Asutusesisene klassifitseerimine

Kui eelmine peatükk testis pöördumiste klassifitseerimist asutustesse, siis käesolev peatükk keskendub asutusesisesele klassifitseerimisele. Peatükk annab ülevaate tehtud töödest ning nendega seonduvatest järeldustest.

Asutuste vajaduste kaardistamise peatükist tuli välja, et asutused on rohkem huvitatud sisemiste töövoogude hõlbustamiseks pöördumiste teemade / sarjade / valdkondade kaupa klassifitseerimisest. Lisas 1 on toodud iga asutuse pöördumise üldised teemad, mille alusel juba praegu pöördumisi jaotatakse. Intervjuudest ning andmete analüüsist ilmses, et kuus asutust tegeleb juba mingil tasemel pöördumiste märgendamisega (näiteks postkastis kirjade alamkaustadesse suunamine või DELTA-s sarjade pidamine), kuid tunneb vajadust märgendada detailsemalt. Kolm asutust tõid välja, et nad on teinud tööd teemade arvu vähendamise nimel või tunnevad, et liiga palju märgendeid teeb liigitamise segaseks ning suurendab valede valikute ohtu. Kaks asutust ei märgenda oma pöördumisi.

## Soovitused klassifitseerija treeningandmete koostamiseks

Mida tuleb asutusesisese kliendipöördumiste klassifitseerimise puhul tähele panna?

Klassifitseerijate klassid tuleb disainida asutusepõhiselt, **lähtudes vastava asutuse vajadustest**. Oluline on mõista, et see, mida asutused soovivad, ei pruugi olla see, mida nad tegelikud vajavad. Samuti on asutuste pöördumiste mahud ja teemade varieerumised nii erinevad, et iga klassifitseerijat tuleb kohandada vastavalt asutuste vajadustele. Siin all mõtleme pigem klassifitseeritavate teemade valikut. Asutusevajadustest lähtumine võib tuua olukorra, kus erinevad asutused on valinud erineva täpsusastmega teemad. Näitlikult tähendab see, et osadele asutustele on piisav, kui pöördumine pannakse klassi *koroona*, teistele on vaja aga ühe teema täpsemaid alamjaotusi nagu *koroona abirahastus ettevõtjatele ja hetkel kehtivad koroonapiirangud*. See ei ole iseenesest probleem, kuid arvestada tuleb, et nii nagu inimesel on keeruline vahel otsustada, millisesse klassi pöördumine kuulub, **võib ka automaatne klassifitseerija sarnaste klasside vahel valimisel hätta jääda ning soovitame võimalusel klasse üldistada**.

Kuigi koroonateemalised küsimused võivad esineda erinevatel asutustel, ei jäänud antud andmete analüüsimisel silma asutusesisesele teemasid, mis oleks asutuseülele (st vähemalt kahes erinevas asutuses) sarnased. Kui taolised teemad või klassid siiski mingil hetkel ilmnevad ehk kahe erineva asutuse klassifitseerijad peavad pöördumist enesekindlalt enda teemaks, soovitame eelistada seda asutust, mille poole pöörduti. Samuti võib aidata



treenimise ajal tehtav klassifitseerijate lisaanalüüs, mida alguses tuleb niivõinää tõhususe kindlustamiseks teha ja mille abil saab tuvastada taolise olukorra põhjuseid ning lahendusi.

Mitmeteemalisuse probleem on ammu tuntud klassifitseerimise probleem. **E-kirjade kontekstis soovitame taolised kirjad jaotada teemade kaupa osadeks ning märgendada lisaks kogu kiri sobivate teemadega.** Sel viisil saab treenida klasse ning testida klassifitseerija tulemust kogu kirjal. Mitmeteemalisuse puhul soovitame lubada klassifitseerijal määrata mitu märgendit. Sel viisil näeb klienditeenindaja ära, et mudel on kas segaduses või pöördumine sisaldab mitut teemat. Seejärel saab teenindaja ise otsustada, kumba klassi pöördumine tegelikult kuulub, või jättagi pöördumine kahe- või enama klassiliseks.

Asutusesisese klassifitseerijate andmestiku loomisel tasub silmas pidada ka arhitektuuriülesust. **Soovitame luua lisaklassifitseerija, mis tegeleb asutusteväliste teemadega nagu tervitused ja roppused** (vt peatükki Solvangute tuvastamine veebikommentaari del treenitud mudeliga) **või käsitleda neid juhte muul moel eraldi.** See välistab võimalikku topelttööd asutustes (asutused koostavad vaid endale vajalikke klasse).

**Andmestiku asutusesiseseks klassifitseerimiseks koostamine** Kodaniku pöördumiste klassifitseerija jätkuprojektis ning arhitektuuri seadistamisel<sup>17</sup> **on võtmetähtsusega. Sellele peab eelnema asutusepõhine analüüs,** mis tuvastab asutuste tegelikud vajadused ning võimalikud teemad. Soovitame analüüsi läbi viia tihedas koostöös asutuste klienditeenindajatega, kellel on ülevaade ja sisetunne sobivate jaotuste kohta, ning andmeanalüütikuga või keeletehnoloogiga, kes saavad hinnata märgendite / klasside sobivust ja abistada andmestiku koostamisel.

Asutustel, kellel puudub kontrollitud ja välja töötatud klassifitseerimissüsteem (olgu selleks siis lipukesed e-kirjal, postkastis alamkaustadesse jaotamine või DELTA-s sarjadesse jaotamine), on võimalik kasutada Texta Toolkitti<sup>18</sup> sobivate klasside tuvastamiseks ning andmestiku koostamiseks. Antud PoC-i raames viisime selle ka Politsei- ja Piirivalveameti näitel läbi (vt peatükki Märgendamata andmetel klassifitseerimine ehk klasterdamine PPA-s).

Eeltööluse juhised leheküljel 29 kehtivad ka asutusesisese klassifitseerimise puhul.

## Märgendatud andmetel klassifitseerimine

Politsei- ja Piirivalveamet andis meile lahkelt käsitsi 13 teemasse lahterdatud e-kirja, igas teemas 75 näidet. Jagasime andmed treening- (50 kirja) ja testandmeteks (25 kirja) ja treenisime erinevaid mudeleid. Joonis 18 illustreerib katsete põhjal seni parima

<sup>17</sup> Vt prototüübi arhitektuuri peatükist, kuidas seostub asutusesisene klassifitseerija koguarhitektuuriga.

<sup>18</sup> Vt docs.texta.ee või lisa 4.

klassifitseerija tulemust testandmestikul. Edukaimaks osutus 5 epohhiga treenitud 300dimensiooniline EstBERT-il treenitud mudel täpsuse ja saagisega 0.6. See ei ole kasutuskõlbulik tulemus. Segadusmaatriksilt joonisel 18 on näha, milliseid teemasid klassifitseerija kõige enam segamini ajab.

Confusion matrix

Dokumentifoto saatmine	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	
Dokumentide uuendamine	0	0	1	0	0	2	0	0	0	0	0	0	0	0	4	0	0	1	
Elamisloa pikendamine	0	0	19	0	4	0	2	0	0	0	0	0	0	0	0	0	0	0	
Elamisloa lisadokumendid	2	0	0	20	2	0	0	0	0	0	0	0	0	0	0	0	1	0	
Infopäring elamisloa kohta	0	0	6	0	18	0	0	0	0	0	1	0	0	0	0	0	0	0	
ID-kaardi ja passi valmimine	1	0	0	0	0	12	0	0	0	0	11	0	0	0	0	0	0	1	
Kuidas riiki siseneda liikumispiirangu ajal	0	0	3	0	0	0	7	4	10	0	0	0	1	0	0	0	0	0	
Kuidas tohin liikuda liikumispiirangu ajal	0	0	0	0	0	0	4	10	11	0	0	0	0	0	0	0	0	0	
Kuidas riigist lahkuda liikumispiirangu ajal	0	0	2	0	0	0	2	10	9	0	0	0	0	0	0	2	0	0	
menetluse päring	0	0	2	0	1	5	0	0	0	14	1	1	1	0	0	0	0	0	
Maksekorraldus	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	1	
PIN koodid	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	3	
Teade vara- või isikukahju kohta	0	0	1	0	2	1	0	1	2	0	0	0	0	18	0	0	0	0	
väljastuskoha muutmine	0	0	0	0	0	5	0	0	0	2	1	0	0	0	0	0	0	15	
	Dokumentifoto saatmine	Dokumentide uuendamine	Elamisloa pikendamine	Elamisloa lisadokumendid	Infopäring elamisloa kohta	ID-kaardi ja passi valmimine	Kuidas riiki siseneda liikumispiirangu ajal	Kuidas tohin liikuda liikumispiirangu ajal	Kuidas riigist lahkuda liikumispiirangu ajal	menetluse päring	Maksekorraldus	PIN koodid	Teade vara- või isikukahju kohta	väljastuskoha muutmine					

Joonis 18: Asutusesisene e-kirja klassifitseerija PPA anonümiseeritud märgendatud e-kirjadel

Mudelil on raskusi üsna loogilistes kohtades – klassidega, mis on omavahel sarnased ja võivad olla eristamisel inimeselegi keerulised. Näiteks on klassid *Kuidas riiki siseneda liikumispiirangu ajal*, *Kuidas tohin liikuda liikumispiirangu ajal* ja *Kuidas riigist lahkuda liikumispiirangu ajal* temaatiliselt väga sarnased, sisaldades samu märksõnu. Segadusmaatriksist näha, et kõiki nende kolme klassi e-kirju ennustatakse pigem riigist

lahkumise või riigis ringi liikumise klassi. Samamoodi on näha näiteks, et 6 (25-st) *Infopäring elamisloa kohta* e-kirja liigitas mudel klassi *Elamisloa pikendamise*, mis mõlemad tegelevad elamislubade ja ajaga (*kaua veel elamisloa kättesaamiseni läheb?*). Nii klassi *menetluse päring* kui ka *ID-kaardi ja passi valmistamine* e-kirjad sisaldavad küsimusi selle kohta, kas taotlus jõudis kohale ja miks mingeid teateid dokumendi valmimise kohta pole tulnud. Inimmärgendajaga võib segadusse sattuda nende kahe klassi eristamisel ja otsustamisel, milline e-kiri kuhu läheb.

See näide illustreerib hästi peatüki alguses rõhutatud vajadust luua treenimisandmestik andmeanalüütiku või keeletehnoloogi ning klienditeenindajate koostöös. Juba teemade nimetustele peale vaadates võib ära näha, millised teemad on omavahel nii sarnased, et nende eristamisega võib tekkida raskusi. Näiteks siin on näide klassist *ID-kaardi ja passi pikendamine*:

*Käisin augusti alguses passi taotlust esitamas, kuid seni ei ole pass veel valmis saanud või ei ole vastav info minuni jõudnud. Uurin, kas sellega väljastamisega on ikka kõik korras ning mis ajaks võiks pass valmis saada?*

See on üsna sarnane näitega teisest klassist (*menetluse päring*):

*Tere hommikust, Soovin teada, kas minu Eesti pass ja id kaart on valmis. Minu isikukood on Xx Lugupidamisega, Xx*

Muidugi me ei väida siinkohal, et treeningandmestiku koostamisel tuleb luua võimalikult erinevaid klasse ja sarnased ei sobi mitte mingil juhul. Rõhutame hoopis seda, et klasside valikult tuleb lähtuda asutuste tegelikest vajadustest (millisest märgendist on hiljem ka kasu?) ning sarnaste klasside loomisel ei saa oodata klassifitseerijalt 100protsendilist täpsust. Vajadusel saab kindlasti tulemusi parandada analüüsides mudeli tunnuseid, treenides uusi mudeleid, suurendades treeningandmestiku hulka ja valides hoolikamalt treenimisel negatiivseid näiteid klassile.

Asutusesisese klassifitseerimise katsed Rahvusraamatukogu ning Statistikaameti juturoboti märgendatud andmestkel tõid välja samad probleemid, mis tuvastati e-mailidel. Juturobotite pöördumiste arvukuse tõttu andis Texta Hybrid Tagger<sup>19</sup> 10% kõrgemaid F1 skooore, jäädes valideerimisandmestikul Statistikaametil 0.88 ja Rahvusraamatul 0.8 kanti. Päril juturobotis arvestab robot ka konteksti (st eelnevalt saadetud sõnumeid), mis võib mõjutada tulemust võrreldes meie praeguste kontekstivabade katsetega. Kuigi nii tavalised e-mailid kui ka

<sup>19</sup> Hybrid Tagger on Toolkitis Tagger Group nime all ning see optimeerib suure arvu klassidega mudeli klassifitseerimist, kasutades binaarseid mudeleid ja Elasticsearchi. Vaata lähemalt: <https://zenodo.org/record/4306169>

juturobotid võivad (eriti infopäringutes) olla üsna lühikesed, on juturobotid oma iseloomult e-mailidest lühemad ja kannavad endas klassifitseerimiseks vähem infot. Suuremate andmete puhul tasub kindlasti uuesti proovida BERT taggerite üle treenimist (see vähendaks ka sõnumite eeltöötuse vajadust).

## Solvangute tuvastamine veebikommentaariidel treenitud mudeliga

Juturobotitele saadetud sõnumeid uurides oli näha, et arvestatav hulk neist on lihtsalt roppused, solvangud või sõim. Et juturobot oskaks nendega toime tulla, **on vaja luua eraldi roppuste teema** (ingl *intent*) ning see näidetega illustreerida. Rahvusraamatukogu ja Statistikaamet ongi seda teed läinud. Kriisi juturobotil vastav kavatsus puudub, mille tulemusena märgendatakse ebasündsad sõnumid vale kavatsusega näiteks *Üldised küsimused*, mille tõttu omakorda saadetakse kasutajale tagasi vastus, millest ta ilmselt huvitatudki ei ole.

Et hinnata, kui palju ebasündsaid sõnumeid vastavad kavatsused kinni suudavad püüda, söötis Texta kõik juturobotitele saadetud sõnumid sisse Texta arendatud veebikommentaare modereerivale mudelile. Teisalt oli Textal soov selle eksperimendiga ka uurida, kui hästi võiks kommentaaridel treenitud mudel üldistuda vestlusrobotitele saadetud sõnumitest pahatahtlike kavatsuste tuvastamiseks.

Järgnevalt toodud tabelist 18 on näha, veebikommentaare modereeriv mudel saavutab juturobotite puhul üsna hea saagise, kuid täpsus on väga madal. Kuldmärgendusena on siinkohal kasutatud kavatsusega *Roppus* märgitud sõnumeid.

Asutus	Täpsus	Saagis
Statistikaamet	0.19	0.83
Rahvusraamatukogu	0.26	0.92

Tabel 18: Sündsusetute lausungite tuvastamine

Kvalitatiivne analüüs heidab valgust selle põhjustele. Näiteks statistikaameti puhul peab kommentaarimudel ekslikult ebasobivaks nii suurt kogust tervitusi, kui ka boti nime lti. Põhjuseks võib olla tendents kommentaarimudeli treeningandmetes, kus just väga lühikesed sõnumid on tihti ebasündsast laadi. Teisalt on näha, et mudel on tuvastanud ka selliseid ebaviisakaid sõnumeid, mida kavatsuse *Roppus* alla pole pandud. Näiteks *sure maha ja türa kao minema siit* kavatsuse *liti segab* all.

Rahvusraamatukogu puhul on probleemid üldjuhul samad. Vesta puhul kerkib esile ka domeeni erinevustest põhjustatud märgendamisvead. Näiteks sõnu *autism* või *kommunism* vestale öelda pole ilmselt midagi imelikku, kui kasutaja just nende teemade kohta materjale otsibki. Samas veebikommentaariumites esinevad need sõnad peaausjalikult just solvangute, sõimu ja vihakõne kontekstis.

Ülaltoodud probleemid saaks ilmselt leevendus mõningase kommentaarimudeli peenhäälestamisega sõnumite žanrile. Ebaviisakate, solvavate või sõimavate sõnumite tõhus tuvastamine aitaks muuta juturobotit inimkäitumisele sarnasemaks. Näiteks nii, et vestlusrobot ei üritaks pakkuda järjekindlalt abi inimesele, kes sellest ilmselgelt huvitatud ei ole.

## Märgendamata andmetel klassifitseerimine ehk klasterdamine PPA-s

Peatüki alguses soovitasime kasutada asutustel, kellel pole pöördumiste klassid ehk teemad / jaotus välja töötatud, kasutada Texta Toolkitti sobivate klasside tuvastamiseks ning andmestiku koostamiseks. See peatükk kirjeldab projekti raames läbiviidud väiksemahulist uurimust, kus proovisime läbi klasside loomise töövoos ning tuvastasime, mida tuleb selles töövoos ning PPA pöördumiste klasterdamise puhul tähele panna.

### Andmestik

Politsei ja Piirivalve Amet eraldas oma postkasti kaustades arhiveeritud e-mailid. Me tõstsime need lokaalses arvutis Stanza mudelitega lemmatiseeritud kujul (TTK-s ei saanud lemmatiseerida, kuna võrguliikluspiirangute tõttu polnud Stanza mudeleid TTK-sse sel hetkel laetud) PPA Texta Toolkitti. Ca 31% e-mailidest sisaldab PPA e-mailide eraldamisel tuvastamata kodeeringuhäirete tõttu suures osas vaid küsimärke. Tõenäoliselt on tegu endiste kirillitsa karakteritega. Need e-kirjad jäid uuringust välja. Kokku on andmestikus 58561 e-kirja. Nendest 25012 kuuluvad postkasti alamkausta *Kohus*, *Vangla*, *Prokuratuur*, 24967 kausta *kustutatud*, 5317 kausta *koroon*, 2200 kausta *sisendkaust*, 1065 kausta *identiteet*. Enamus kirju on pärit selle aasta I kvartalist.

### Klastrite loome

PPA kasutas juhendi<sup>20</sup> abil Texta Toolkiti Teemaanalüsaatorit<sup>21</sup> ning Regex Taggerit<sup>22</sup> erinevate nende jaoks kasulike klastrite moodustamiseks. Kasulikkust hindasid need, kes ise ka klienditeenindusega tegelevad. Silmas peeti seda, et väga heade tulemuste puhul saaks töövooge automatiseerida ning heade tulemuste puhul klienditeenindaja ajakulu vähendada.

Kokku loodi 22 klassist, millest kuus valiti PoC-i raames asutusesisese mudeli treenimiseks. Valiku aluseks oli maht: need olid suurima näidete arvuga klassist. Kuigi Teemaanalüsaator tõestas ennast TEXTA hinnangul üsna hea teemadetuvastajana, grupeeris see PPA

20 Juhendi leiab aadressilt [docs.texta.ee](https://docs.texta.ee) või lisast 4. Lisaks tehti neile PPA-sisene videojuhend eraldi nende andmete näidisel.

21 Teemaanalüsaator ehk Topic Analyzer on tööriist, mis tuvastab sarnaste dokumentide gruppe andmestikust. Seda saab kasutada näiteks märgendamata andmestiku struktuuri uurimiseks, et saada aimu, mida see endas sisaldab. Tööriista peamine eesmärk on siiski astuda samm edasi ja muuta saadud avastused märgenditeks. Seda saab seejärel kasutada juhendatud masinõppe mudelite treenimiseks.

22 Regex Tagger on tööriist tekstide mustripõhiseks märgendamiseks regulaaravaldiste abil. Kasutaja saab defineerida mustreid (sõnu/fraase/regexeid), mida otsida tekstist, ning kindla informatsiooni eraldamiseks ja saadud informatsiooni põhjal tekste märgendamiseks peenhäälestada mustrite ja otsingute parameetreid.

hinnangul kokku e-maile, mille ühe märgendamise alla koondamine annab küll lisainfot, kuid ei tõhusta teenindusvoogu. PPA pidas silmas juba kindlaid klastreid ega vajanud seetõttu teemade tuvastamist nii väga ning läks Regex Taggeri kasutamise teed. Hiljem TEXTA täiendas Regex Taggeri reegleid ja parandas saadud klastreid vastavalt kliendi soovile. Tulemusena saadi asutusesisese klassifitseerimise katseteks:

- 955 **KoroonaRiigipiir** klustrisse kuuluvat e-kirja, mis küsivad eriluba riigipiiri ületuseks või täpsustavad riigipiiri ületamise tingimusi koroona ajal;
- 1439 **Karantiin** klustrisse kuuluvat e-kirja, mis sisaldavad märksõnu *eneseisolatsioon, isolatsioon, karantiin, koroonakarantiin*;
- 1014 **Dokumendifoto** klustrisse kuuluvat e-kirja, mis edastavad dokumendifoto või küsivad täpsustusi selle tegemise ja tingimuste kohta;
- 7889 **Täitemenetlus** klustrisse kuuluvat e-kirja, mis sisaldab sõnu nagu *sundtäitmine, kohtutäitur ja täitedokument*;
- 16493 **Automaatvastus** klustrisse kuuluvat e-kirja, mis sisaldab automaatvastuseid (nt oli kaustas nimega Kohus, Vangla, Prokuratuur 32% kirjadest (16317 kirja) automaatvastused stiilis *Käesolevaga teatame Teile, et Harju Maakohtu Tallinna kohtumaja kantselei on vastu võtnud Teie poolt aadressile hmk tallinn.menetlus@kohus.ee saadetud elektronkirja koos võimalike manustega*);
- 20336 **Kohtudokumendid** klustrisse kuuluvat e-kirja, mis sisaldab *Kohus, Vangla, Prokuratuur* kausta mitte Automaatvastuse või Täitemenetluse kuuluvad kirju, mis sisaldavad sõnu nagu *dokument, kohtumäärus, kohtuotsus*.

Mitmed klasterid nendest (Automaatvastus ja Täitemenetlus) on lahendatavad reeglipõhiselt Regex Taggeriga. Dokumendifotoga on juba keerulisem: dokumendifototeemalises e-kirjas võib olla mainitud koroona või karantiin, mistõttu fotot kohapeal tegemas käia ei saa. Seetõttu on oluline ka reeglite järjekord: kui dokumendifoto märgendit antud e-kiri ei saa, alles siis võib pidada märksõnu *koroona* ja *karantiin* koroonateemalisele e-kirjale viitavaks. Kohtudokumentide klaster kasutas ära eelteadmused ehk otsis dokumentide mainimisi juba *kohtu* kausta käsitsi lohistatud kirjadest.

TEXTA soovib PPA e-kirjade puhul kasutada reeglipõhise ja masinõppepõhise lähenemise hübriidvarianti, kus keerulisemad juhud lahendatakse masinõppega ning selgete reeglitega tuvastatavad juhud reeglipõhiselt. Juba praegu kasutatakse PPA postkastis kindlaid reegleid (nt saatja aadressi alusel), mille abil kirjad automaatselt tegevuskaustadesse suunatakse. Suurema osa sissetulevatest kirjadest sorteerib infohaldur siiski veel käsitsi.

Tegevuskaustad on loodud e-kirjade sisu ja tehtavate tegevuste järgi, et lihtsustada sisendkausta haldamist, märgendamist, statistika kogumist ning jagada infohaldurite vahel töökoormust. Suve alguses toimus ka tegevuskaustade ümberstruktureerimine kümneks kaustaks ning on märgata teemasid, mille automatiseerimisest või automaatselt märgendamisest oleks palju abi. Näiteks kausta **Registreerimisele** juba praegu saatja aadressi järgi automaatselt suunatud kirjad lohistatakse käsitsi edasi Deltasse. Infopäringute märgendamist tüüpvastuste tasemel veel ei toimu, selle jaoks on eraldi üks suur kaust ning PPA-l puudub huvi liiga paljude märgendite vastu. Siiski näiteks Deltasse suunduvate kirjade DELTA funktsioonidesse ja sarjadesse automaatne klassifitseerimine DELTA-s või juba postkastis on PPAs arutlusel olnud. Samuti soovitame klassifitseerida kõrge prioriteediga e-kirju, millele tuleb koheselt reageerida (praegu tuvastatakse neid käsitsi). See aitaks kiirendada taoliste e-kirjade tuvastamist ja seega ka nende reageerimist.

## Mudelite treenimine ja tulemused

Antud klastrid jagasime treening- ja testandmeteks (vt tabelit 19). Kohtudokumentide ja automatvastuste seast valisime andmestikku juhuslikult vaid osa meilidest, et vähendada nende suurt osakaalu koguandmestikus. Tähele tuleb panna, et andmestik sisaldab palju vastuskirju ning edasisaatmiseid, mistõttu võib ühes e-kirjas olla pöördumine, teises see sama pöördumine koos politsei vastusega ning reageering vastusele. Seetõttu ei saa me garanteerida, et testandmestik ei sisalda e-maile, mis üheski osas ei kattu treeningandmestikus olevatega, ning tulemuste tõlgendamisel ei tasu olla liiga optimistlik.

Klaster	Treeningandmed	Testandmed
Automaatvastus	3000 (juhuvalim)	1000
Kohtudokumentid	3000 (juhuvalim)	1000
Täitemenetlus	2970	500
Karantiin	1224	215
Dokumendifoto	830	145
KoroonaRiigipiir	810	145

*Tabel 19: Texta Toolkitiga koostatud treening- ja testandmed PPA näitel*

Joonis 19 näitab mudelite treenimistulemusi valideerimisandmestikul. Joonis 20 näitab samade mudelite tulemusi testandmestikul. Nimes olevad lühendid kirjeldavad mudeli arhitektuuri: CV on Count Vectorizer, TF on TfIdf Vectorizer, LSVC on LinearSVC, LR on Logistic Regression, w2vp on fraasijaga word2vec sõnamudelite lisandus, ft on FastTexti



sõnamudelite lisandus. Kõik mudelid on treenitud lemmatiseeritud e-kirjade sisul ja pealkirjal.

Joonistelt on näha, et mudel sobitab üle: treeningtulemused on kõik saja protsendi lähedase F1 skooriga, testtulemused 75 kuni 80 protsenti. Segadusmaatrikseid uurides on näha, et TF mudelite testandmestikul saadud skoorid toob alla Karantiini ennustamine, CV mudelitel Karantiin ja Dokumendifoto. Segadus Karantiini ja Koroonariigipiiri klastrite vahel oli oodatav, kuna need sisaldavad üsna sarnaseid märksõnu (riigipiiri ületamise lubatuse uurimisel seletatakse, et plaanitakse olla ilusti karantiinis jne) või olla kaheosalised: küsides nii riigipiiri võimalikkuse kui ka karantiini kohustuse kohta. Samuti ei saa garanteerida andmestiku puhtust, kuna klastrid loodi pigem reeglitega kui ükshaaval käsitsi e-kirju lahterdades.

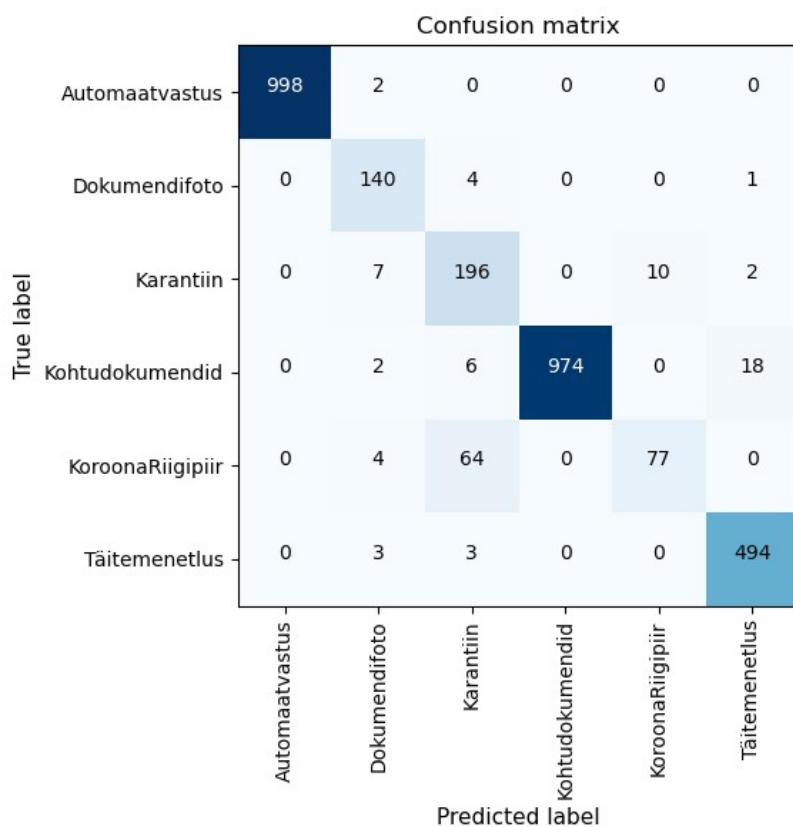
Description	F1 score ↓	Precision	Recall
ML_TF_LR_w2vp_Jemmas	0.80	0.71	0.97
ML_TF_LR_ft_Jemmas	0.78	0.69	0.97
ML_TF_LR_Jemmas	0.77	0.68	0.97
ML_CV_LSVC_ft_Jemmas	0.77	0.74	0.96
ML_TF_LSVC_w2vp_Jemmas	0.76	0.68	0.96
ML_CV_LSVC_w2vp_Jemmas	0.76	0.72	0.96
ML_TF_LSVC_Jemmas	0.76	0.67	0.97
ML_TF_LSVC_ft_Jemmas	0.75	0.66	0.97
ML_CV_LSVC_Jemmas	0.75	0.72	0.95

Joonis 19: Mudeli treenimistulemused valideerimisandmestikul

Parandasime seni parimat mudelit stoppsõnadega, kuna mudeli õpitud tunnusteks sattus ka nimesid ja asukohti, mis tegelikult e-kirja teemat näidata ei tohiks. Selle mudeli F1 testandmestikul on samuti 0.8, kuid täpsus on 0.74 ja saagis 0.95. Segadusmaatriks on toodud joonisel 21. Ka sellel mudelil on kõige keerulisem eristada *Karantiini* ja *Koroonariigipiiri* ning vähesel määral ka *Täitemenetlust Kohtudokumentidest*.

Description	F1 score ↑	Precision	Recall
ML_CV_LSVC_w2vp_lemmas	0.97	0.97	0.97
ML_TF_LSVC_w2vp_lemmas	0.97	0.97	0.97
ML_HV_LSVC_w2vp_lemmas	0.98	0.98	0.98
ML_TF_LR_ft_lemmas	0.98	0.98	0.98
ML_TF_LR_lemmas	0.98	0.98	0.98
ML_TF_LR_w2vp_lemmas	0.98	0.98	0.98
ML_TF_LSVC_ft_lemmas	0.98	0.98	0.98
ML_TF_LSVC_lemmas	0.98	0.98	0.98
ML_CV_LSVC_ft_lemmas	0.98	0.98	0.98
ML_CV_LSVC_lemmas	0.99	0.99	0.99

Joonis 20: Mudeli treenimistulemused testandmestikul



Joonis 21: Mudeli segadusmaatriks testandmestikul

# PoC ehk Siimuke

## siimuke.mkm.ee kirjeldus

Aadressil siimuke.mkm.ee on üleval PoC-i näidis. See sisaldab endas TTK versiooni 2.39.x, koos projektiga MKM. Projektile saavad uutele kasutajatele ligipääsu anda adminkasutajad. Projektis on kättesaadavad tabelis 20 toodud anonümiseeritud andmestikud (täpsemalt loe nende kohta andmete peatükist) ja tabelis 21 toodud prototüübi raames treenitud mudelid (loe täpsemalt asutuseülese ja -sisese klassifitseerimise peatükki).

Andmeid saab uurida ning nendes otsinguid teha Searcheris aktiveerides üleval sinisel ribal MKM-i projekti ning valides huvipakkuv(ad) indeks(id) (vt joonist 22). Searcheri kasutust saab täpsemalt uurida TTK dokumentatsioonist<sup>23</sup>.

Mudelite pöördumiste klassifitseerimist saab:

- a) testida minnes huvipakkuva mudeli juurde (vt tabelist 21 mudeli asukohti) ning Actions alt kasutades variante „Tag text” (enda trükitud/kopeeritud teksti hindamine), „Tag doc” või „Tag random doc” (juhuslikult valitud pöördumise hindamine),
- b) uurida vaadates Searcheris mudelite tulemusi testandmestikul,
- c) uurida vaadates Evaluatoris ehk Hindajas kokkuvõtet testandmestiku tulemustes<sup>24</sup>.

Projekti ajal uuenes TTK versioonilt 2.24 versioonile 2.39. Uuendused sisaldasid ka muudatusi, mis aitasid masinõppe treenimisele ja tulemuste analüüsimisele kaasa. TTK on pidevas täienemises, kuid antud projekti vaatest jäid põhiasjad samaks ja midagi olulist ei muutunud.

Lisaks Siimukese demoversioonile viisime 21. juunil 2021 läbi ka TTK töötoa, mille kasutajatelt küsisime hiljem TTK kasutavuse kohta tagasisidet. 21-st osalejast 11 täitis tagasisideküsitluse. Ühe kasutaja SUS skooriks<sup>25</sup> kujunes 82.5 ehk A, üldkeskmiseks kujunes 48 ehk F. Vabatekstilises tagasisides toodi välja, et Texta Toolkiti paljud võimalused võtavad alguses silme eest kirjuks ning nendega tutvumiseks kulub rohkem kui üks pooleteisetunnine töötuba. Samas leiti, et tegu on võimsa tööriistaga, mis kindlasti tuleb tekstiliste andmete analüüsimisel kasuks<sup>26</sup> ja mida soovitakse edaspidigi kasutada.

23 Aadressilt <https://docs.texta.ee/searcher.html> või lisast 4.

24 Pane tähele, et Evaluator ei tekita segadusmaatrikseid juhul, kui andmestikus on üle 30 klassi.

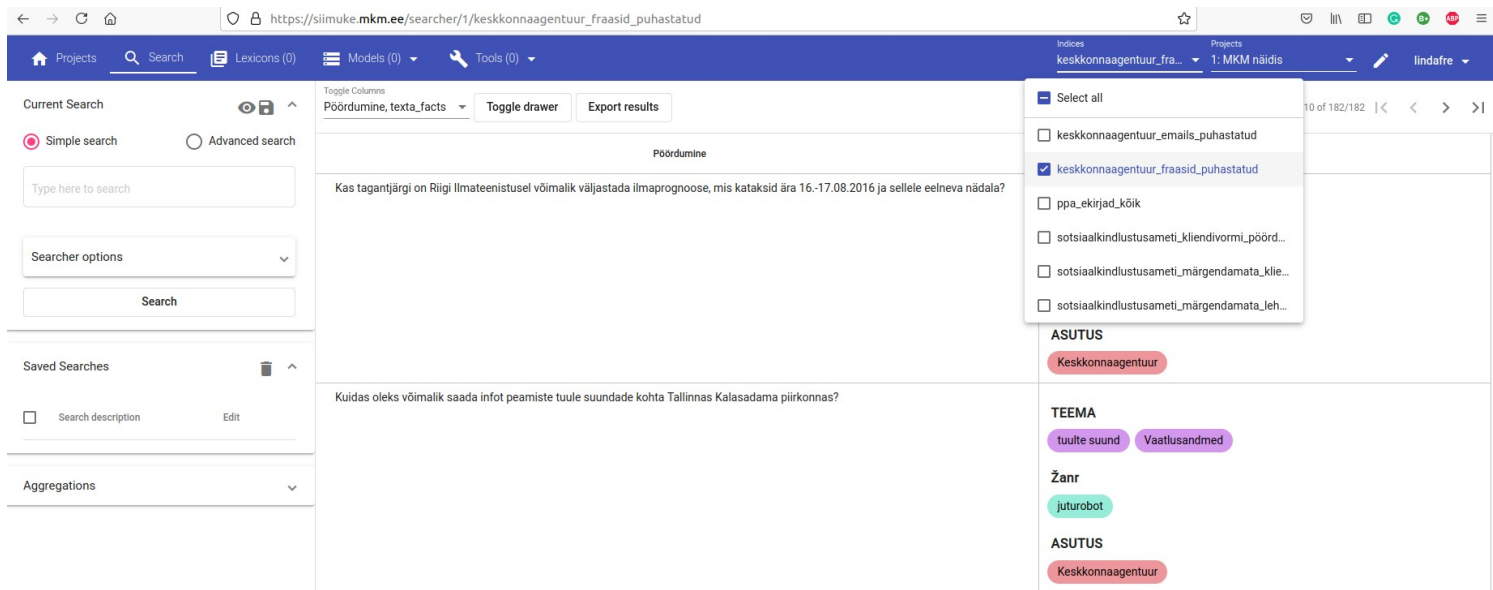
25 System Usability Scale konverteerib 10 lauset viiepallisel skaalal „Nõustun igati” kuni „Kindlasti ei nõustu” numbriliseks väärtuseks, millega saab hinnaata süsteemi keerukust esmakasutaja jaoks. Vt täpsemalt siit: <https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html>.

26 Suurem osa osalejatest oli korpuslingvistika või keeletehnoloogia taustaga ning mudelite treenimisele keskenduda ei jõutud.

Texta Toolkiti lahendus on kasutatav juba praegu – kõik siinsete katsete töövood viidi läbi TTK-s – ning töötoa tagasiside põhjal alustati TTK veelgi intuitiivsemaks disainimist.

indeksi nimi	kirjeldus
keskkonnaagentuur_emails_puhastatud	373 KKA e-kirja
keskkonnaagentuur_fraasid_puhastatud	182 KKA e-kirjadest eraldatud fraasi (filtreeritud 213 fraasist)
sotsiaalkindlustusameti_märgendamata_kliendivormi_pöördumised	254 SKA „võta ühendust” vormi pöördumised
sotsiaalkindlustusameti_märgendamata_lehe_tagasiside_pöördumised	194 SKA klassifitseerimata „võta ühendust” vormi pöördumised
sotsiaalkindlustusameti_kliendivormi_pöördumised_v2_puhastatud	896 klassifitseerimata kodulehe tagasiside filtreeritud väljavõte
ppa_ekirjad_kõik	995 PPA e-kirjad
transpordiamet_emailid	573 TA e-kirja
ria_märgendamata_piletid	124 RIA e-kirja
rahvusraamatukogu_juturobot_treeningandmestik	1210 RR-i käsitsi ette valmistatud juturoboti lausungit
haigekassa_pöördumised	79 HK e-kirjadest eraldatud fraasi
aü_big_test2_	Asutuseülese e-kirjade ja fraasidega mudeli testandmestik
aü_big_train2_	Asutuseülese e-kirjade ja fraasidega mudeli treeningandmestik
juturobotid_rr_stat_test	RR ja STAT-i juturoboti mudeli testimiseks mõeldud puhastatud testandmestik
juturobotid_rr_stat_train	RR ja STAT-i juturoboti mudeli treenimiseks mõeldud puhastatud treeningandmestik
ppa_ekirjad_test	Indeksist ppa_ekirjad_kõik eraldatud testandmestiku osa
ppa_ekirjad_train	Indeksist ppa_ekirjad_kõik eraldatud treeningandmestiku osa

Tabel 20: Siimukese PoC-is olevad projektandmete kirjeldus



Joonis 22: Siimukeses on akitveeritud indeks keskkonnaagentuur\_fraasid\_puhastatud. Näha on fraase ja nendega seotud teksti fakte metainfoga: pöördumise teema(d), žanr ja asutuse nimi.

## PoC-is kasutatavad mudelid

Millised meetodid/algoritmid annavad kõige täpsema klassifitseerimise tulemuse?

Katsed peatükis Asutuseülene klassifitseerimine näitasid, et kõige paremaid tulemusi asutuseülel e-kirjade klassifitseerimisel andis mudel, mis treeniti lemmatiseeritud e-kirjadel või nendest käsitsi eemaldatud fraasidel (vt lk 42) ning kasutas stoppsõnu, Word2Vec sõnavektoreid ja treeningandmete balansseerimist (väiksema näidete arvuga klasside mahu suurendamist näidete kordamise näol). Mudel kasutas linearSVC algoritmi ning TfIdf vektoriseerijat.

Tehisnärvivõrgud ei andnud häid tulemusi. Neid tasub uuesti kaaluda siis, kui tegeletakse kordades suuremate andmetega. Samuti võib proovida BERT mudelite peenhäälestamist (ingl *fine-tuning*). See tähendab olemasolevate eestikeelsete BERT mudelite ületreenimist pöördumistel.

Asutusesisese klassifitseerimine tõstas uuesti klasside hoolika valimise olulisuse ning tõestas asutusesisese klassifitseerimise võimalikkust.

Siimukesse testimiseks üle kantud mudelid on toodud tabelis 21.

Mudeli nimi	Asukoht	kirjeldus
Asutusteülene klassifitseerija	Models > Taggers	Asutusteülene e-kirjade mudel koos juturoboti fraasidega (parim versioon, mis asutuseüleses klassifitseerimises saadi)
Asutusteülene juturoboti binaarne klassifitseerija	Models > Taggers	Asutusteülene binaarne juturoboti mudel, kus True on STAT, ja False on RR
PPA asutusesisene klassifitseerija	Models > Bert Taggers	Asutusesisene PPA e-kirjade klassifitseerija (erinevalt näitest treenitud EMBEDDIA/fineest-bert mudelil, kuna EstBERT polnud Siimukeses kättesaadav)

Tabel 21: Siimukese PoC-is olevad projekti raames treenitud mudelid

## Vastused mõnele hanke tehnilise kirjelduse küsimusele

Siin alapeatükis vastame kokkuvõtvalt hanke tehnilises kirjelduses toodud küsimustele, mida mujal selgelt eristatud ei ole.

Millistel tingimustel on ministeeriumi ja asutuse tasemel kliendipöördumiste klassifitseerimine võimalik ning kuidas on seda kõige mõistlikum teha?

Leheküljel 29 kirjeldatakse pöördumiste eeltöötlemise tingimusi. Leheküljel 34 kirjeldatakse, millisel määral saab mudeli treenimisel kasutada ühiselt vähemalt e-maile ja vestlusakende (sh juturoboti) pöördumisi. Leheküljel 48 tuuakse välja, mida tuleb asutusesisese kliendipöördumiste klassifitseerimise puhul tähele panna. Kõik sealtoodu kehtib ka asutuseülese klassifitseerimise puhul. Prototüübi arhitektuuri peatükk kirjeldab võimalikke mõistlikke lahendusi arhitektuuriliselt.

Kuidas lahendada vastuste loomine asutuste tüüpsemadele? Kas seda on mõistlik teha klassifitseerijas?

Seni oleme kasutanud terminit klassifitseerija viidates treenitud masinõppe mudelile, mitte mudelit kasutavale kogusüsteemile. Üldjuhul treenitakse juhendatud, st märgendatud andmetel, mudeleid märgendite abil. See tähendab, et algoritm saab sisse erinevate pöördumiste klasside näidised ja õpib selle peal nähtud klasse ennustama. Ka meie prototüübi arhitektuuri peatükis on arvestatud, et klassifitseerija väljastab treenitud klassi nime, mitte terve tüüpvastuse. Seega **ei ole mõistlik luua tüüpvastuseid klassifitseerijas**

ning klasside ja tüüpvastuste sidumiseks peab olema eraldi teenus/loogika.

Asutuste vajaduste kaardistamise peatükk tõi välja, et asutused kasutavad erinevaid meiliservereid ja dokumendihaldussüsteeme, mistõttu peab iga asutus ise leidma loogika, kuidas saadud märgendiga edasi tegutseda. Sama käib ka tüüpvastuste kohta.

Nagu asustusesisese klassifitseerimise peatükk välja tõi, võib tekkida raskusi selliste klasside treenimisel, millele saab vastata tüüpvastusega. Ka intervjuudel asutustega (vt asutuste vajaduste kaardistamise peatükki) tõi need, kes on juba klassifitseerimise pilootprojekte teinud, välja, et raske oli leida tüüpvastustega vastataavaid pöördumisi. Tihti sisaldavad pikemad kirjad mitmeteemalisi nii asustusesiselt kui ka -üleselt pöördumisi, mida on inimeselgi keeruline kategoriseerida. Samuti võivad olla pöördumised nii spetsiifilised, et neile ei ole mõtet klassi treenida või neid peab hästi üldistama, või sõltub vastus inimesest (ja temaga seonduvast infost kuskil andmebaasides).

Sellegipoolest on mitme asutuste klienditeenindajatel oma isiklik või jagatav tüüpvastuste andmebaas (näiteks Excel) ja me ennustame, et paljusid klasse ehk teemasid saab hõlpsasti automaatselt lahendada (st. neile automaatselt vastata nagu on tehtud Bürokrati juturoboti projektis). **Tüüpvastuste loomine käib käsikäes treeningandmete loomisega.** St, klassi luues peab juba mõtlema sellele, mida selle klassimärgendiga edasi teha soovitakse ning kas selle klassi pöördumistele saab vastata tüüpvastusega.

Millisel määral saab loodud mudel(eid) kasutada kliendikõnede kaudu tulnud kliendipöördumiste klassifitseerimiseks?

Projekti raames transkribeeritud kõnesid, mille teemad kattuvad e-kirjadega, ei õnnestunud saada. Transkribeeritud kõnedele ja e-kirjadele ühise mudeli treenimine pole tavapärase praktika, mida oleks palju uuritud. Transkribeeritud tekst erineb tavatekstist oluliselt. Isegi arvestades, et kliendid võivad pöörduda asutuse poole trüki- ja kirjavigadega mitte puhtas kirjakeeles, on transkribeeritud tekst veaohlikum sisaldades endas kõnekeelt, vahehäälsusi (*aamm, ah, jajah*) ja valesti transkribeeritud sõnu (näiteks *alkoovist* saab *alkohol*). Kõnekeel erineb kirjakeelest väga – mõnikord ei pruugigi kirja pandud kõnet lugedes aru saada, millest räägiti, kuna loogiline tundub see alles kuulates. Samuti sisaldab kõne dialoogi, kiri aga ainult küsimust.

Neid erinevusi ja vähest teaduslikku uurimustööd sellel teemal arvestades leiame, et **loodud mudeleid ei saa kliendikõnede kaudud tulnud kliendipöördumiste klassifitseerimisel edukalt kasutada.**

# Prototüübi arhitektuur

Klassifitseerijate arhitektuuri eesmärk on toetada mitmete erinevate asutuste suurandmete suunamise ning märgendamise vajadusi. Lisaks peab arhitektuur ka toetama asutusteülest suunajat ning mudelite pidevat täiendamist (märgendamist). Lisaks peab lahendus olema lihtsalt skaleeritav ning andmed peavad resideerima turvaliselt neid omava asutuse juures.

Eelnevatest eesmärkidest lähtuvalt kirjeldatakse võimalikud arhitektuurilised lahendused läbi kahe peamise kasutusjuhu:

- **asutusesisene** klassifitseerimine, mis võimaldab asutusel treenida ja rakendada klassifitseerimismudeleid enda taristul eesmärgiga märgendada sissetulevaid infopäringuid;
- **asutusteülene** klassifitseerimine, mis võimaldab treenida ja rakendada klassifitseerimismudeleid, mis on mõeldud kliendipöördumisele vastava asutuse tuvastamiseks.

Arhitektuurne lahendus on loodud abstraktsena: kuigi kirjelduses on kasutatud klassifitseerimismudelite treeningkeskkonnana vabavaralist TEXTA Toolkit<sup>27</sup> ja tekstilise andmebaasina Elasticsearch<sup>28</sup>, on kõik komponendid analoogsete toodetega asendatavad.

## Asutusesisene klassifitseerimine

### Eesmärk

Selleks, et efektiivselt klassifitseerida näiteks ühe ministeeriumi valdkonna küsimus õigesse allastusse ning selle asutuse õigesse osakonda on vaja lahendust, mis baseeruks masinõppe süsteemil, mis on treenitud suure hulga antud valdkonna andmete pealt. Lisaks esialgsele treenimisele peab mudelit ka kogu aeg edasi treenima kuna ajaga teenuste tähtsused muutuvad ning tekivad uued vastutusvaldkonnad ja teenused. See tähendab, et mudel peaks lõpuks omama terviklikku ülevaadet ühe asutuse või ministeeriumi teenustest ning küsimustest, mis kodanikel sellega seoses tekivad. Lisaks peab lahendus olema turvaline, lihtsalt laiendatav ning andmed jääma vastavad asutuse kätte.

27 [https://et.wikipedia.org/wiki/Texta\\_toolkit](https://et.wikipedia.org/wiki/Texta_toolkit)

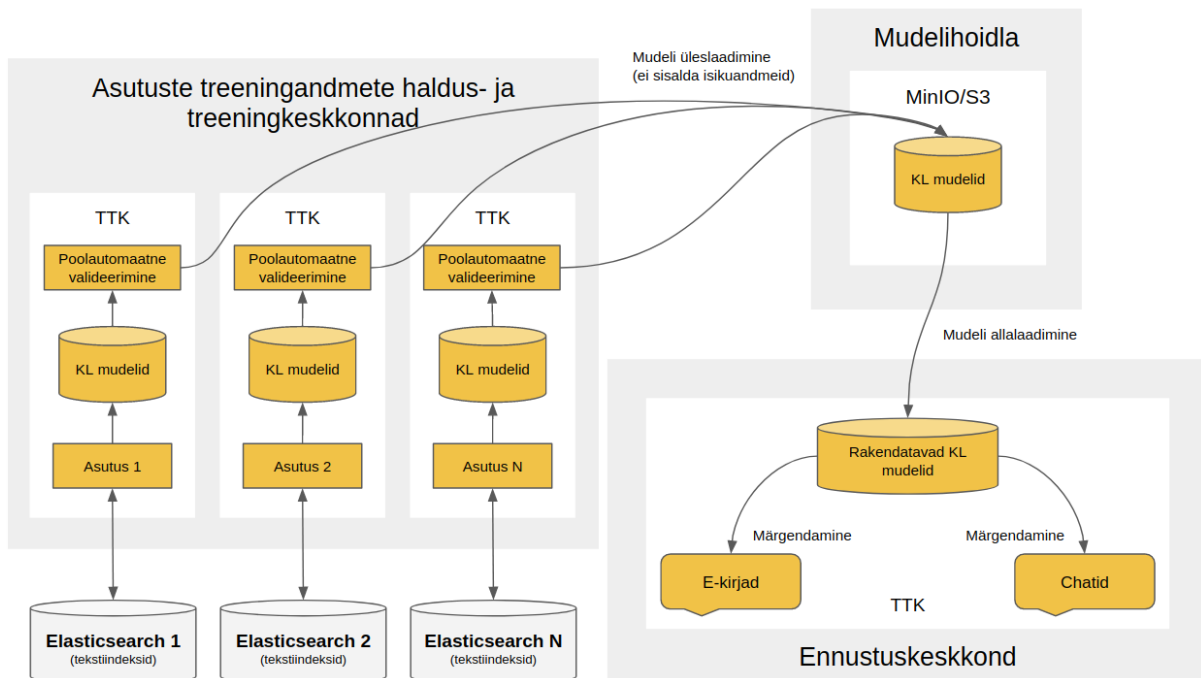
28 <https://en.wikipedia.org/wiki/Elasticsearch>



## Kuidas peaks olema lahendatud pöördumiste teemapõhine klassifitseerimine?

### Lahendus

Väljapakutav lahendus (vt joonist 23) põhineb horisontaalselt skaleerival arhitektuuril, kus andmed salvestatakse vaid andmeid omava asutuse juures, mudelid on taastreenitavad, valideeritud ning versioneeritud. **Kõik klassifitseerimismudelid treenitakse asutuse infrastruktuuril, et välistada andmete liikumist asutusest väljapoole.** Kõik treenitud mudelid on üleslaetavad mudelihoidlasse, mis töötab repositooriumi põhimõttel ning on paigaldatav riigiüleselt, mis tagab mudelite riskasutuse. Samas on mudelihoidla paigaldatav ka asutusesiseselt, kui selleks peaks asutusel vajadus tekkima.



Joonis 23: Asutusesisene klassifitseerija

### Lahenduse komponendid:

#### 1. Asutuse treeningandmete haldus- ja treeningkeskkond

- Treeningandmete halduskeskkonna eesmärk on märgendada sissetulevaid emaille, vestluseid jms. Kui tekst on märgendatud, st talle on külge määratud näiteks valdkond või teema, siis on võimalik andmeid kasutada mudeli treenimiseks.
- Antud joonisel on igal asutusel enda TTK installatsioon ning Elasticu andmebaas.
- Igal TTK projektil on enda andmetabelid, mis sisaldavad treeningandmeid.

Kasutajad ei näe üksteise andmeid ehk teisi projekte. Kõikide projektide andmed on indekseeritud Elasticsearchis. Ligipääs teiste asutuste andmetele puudub, kuna igal asutusel on enda versioon oma sisevõrgus.

- d. Klassifitseerimismudelite treenimisel toimub andmete ligipääsude haldamine TTK projektidena<sup>29</sup>, mis tagavad asutuse kasutajatele ligipääsu nende andmetikele. Alternatiivselt on võimalik õiguseid hallata Elasticsearchis<sup>30</sup>, kui kasutada ei soovita juba olemasolevaid õiguste lahendusi (nn TTK).

## 2. Mudelihoidla

- a. Treenitud mudelid laetakse üles klassifitseerimismudelite hoidlasse, kus toimub nende versioonihaldus. Hoidlasse laetud klassifitseerimismudelid ei sisalda isikuandmeid.
- b. Klassifitseerimismudelite hoidla võimaldab hoiustada ja versioneerida nii avalikke (asutustevahelisi) kui privaatseid (asutusesiseseid) mudeleid. Antude *usecase*-i puhul laeb iga asutus Mudelihoidlasse enda asutuse spetsiifilised mudelid.
- c. Mudelihoidla on S3 protokollil töötav vabavaraline objektihoidla (*Object Storage*), kuhu laetakse üles mudelite failid (kokkupakitud kaust faile). S3 süsteemid võimaldavad ühest objektist hoida mitu versiooni ühes *bucketis*<sup>31</sup>.
- d. Mudelihoidla võib olla asutusepõhine, kuid võib olla ka riigiülene, nt nagu Koodivaramu. Riigiülese lahenduse puhul on soovi korral võimalik piirata kasutuse skoopt asutuseti.

## 3. Ennustuskeskkond

- a. Ennustuskeskkond on vastava asutuse server, kus jookseb eelnevalt loodud ja treenitud masinõppe mudel.
- b. Treenitud mudelid on kasutatavad ennustamiskekkonnas, kuhu need laetakse klassifitseerimismudelite hoidlast.
- c. Ennustuskeskkond on suuteline automaatselt mudeleid uuendama, kui hoidlasse ilmub sama mudeli uus versioon.
- d. Ennustuskeskkonda on võimalik kasutada nii e-kirjade märgendamiseks kui vestlusrobotite kasutatavate suhtluseesmärkide (*intent*) määramiseks.

## Peamine kasutuslugu

Järgnevalt kirjeldatakse ühe asutuse mudeli loomise ja treenimise protsessi:

29 [Vt https://docs.texta.ee/et/terminology.html#project](https://docs.texta.ee/et/terminology.html#project) või lisast 4.

30 <https://www.elastic.co/guide/en/elasticsearch/reference/6.8/field-and-document-access-control.html>

31 <https://docs.min.io/docs/minio-bucket-versioning-guide.html>

- Kasutaja laeb üles või talle imporditakse automaatselt andmed, mida on vaja märgendada (nt. emailid või sõnumivahetused).
- Kasutaja avab TTK-s enda asutuse projekti, märgendab andmed.
- Pärast märgendamist paneb kasutaja mudeli treenima. Mudel treenitakse uute lisatud andmete pealt.
- Kasutaja valideerib mudeli
- Mudel laetakse üles Mudelihoidlasse.
- Ennustuskeskkond laeb alla uuendatud targema mudeli

## Kokkuvõte

Lahendus põhineb mikroteenuste arhitektuuril ning on paindlik nii laiendamises kui ka mudelite treenimises. Iga asutuse saab endale paigaldada oma versiooni TTK-st ja Elasticust, seeläbi hoides andmeid enda juures ning üles laadides vaid masinõppe mudeli. Lahendus võimaldab jagada ja täiendada anonüümseid masinõppe mudeleid, et luua nende pealt uusi ja targemaid mudeleid. Tänu versioneerimisele on võimalik alati minna tagasi eelnevatesse mudelitesse ning on võimalik kasutada ka teiste asutuste mudeleid.

## Asutusteülene klassifitseerimine

### Eesmärk

Asutusteülse klassifitseerimise vajadus tuleneb asutuste ja valdkondade paljususest - selleks, et leida kodaniku vajadusele vastus on kõigepealt vaja aru saada, mis valdkonna või valdkondade kohta tema küsimus on. Selleks, et efektiivselt klassifitseerida ükskõik millise valdkonna kliendi pöördumine õigesse asutusse või teenusesse on vaja infosüsteemi, mis omaks masinõpet, mis on treenitud kõikide ametkondade andmete pealt. See tähendab, et masin peab peab "teadma" kõikide asutuste spetsiifikat, selleks, et efektiivselt suunata kodanik õigesse valdkonda või asutusse.

Kuidas on kõige mõistlikum ministeeriumi ja asutuse tasemel kliendipöördumisi klassifitseerida?

### Valikud

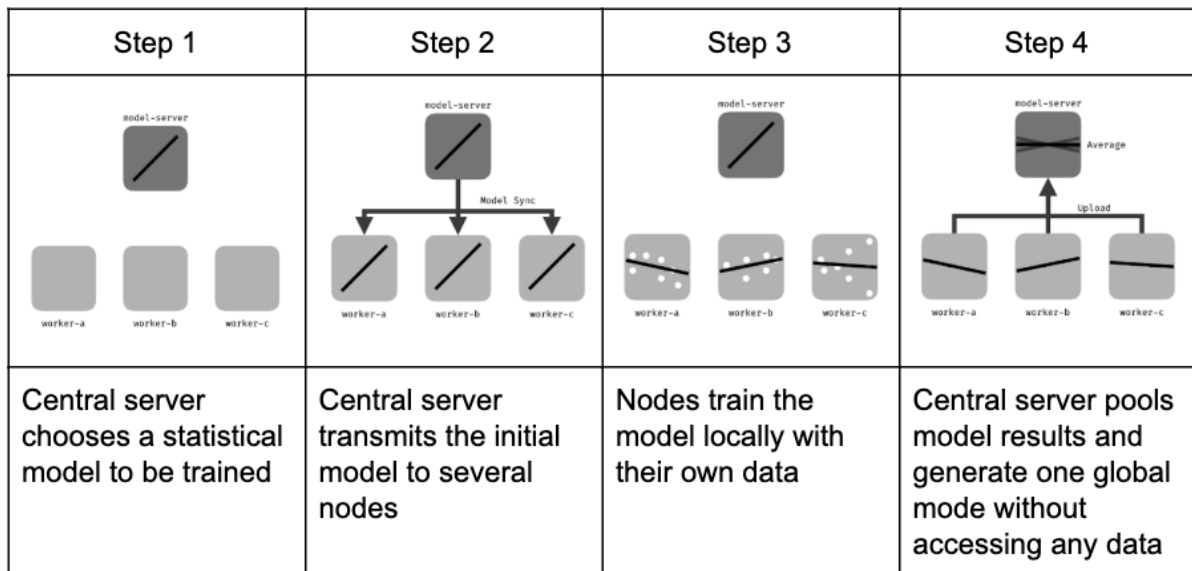
Alljärgnevalt kirjeldatakse kahte arhitektuurilist valikut asutusteülse klassifitseerimise implementeerimiseks - keskne födereeritud õpe (*Centralized Federated Learning*<sup>32</sup>) ning teenusmudelite foorum (*Voting Based*). Mõlema lahenduse puhul tuuakse välja tema plussid

32 [https://en.wikipedia.org/wiki/Federated\\_learning#Centralized\\_federated\\_learning](https://en.wikipedia.org/wiki/Federated_learning#Centralized_federated_learning)

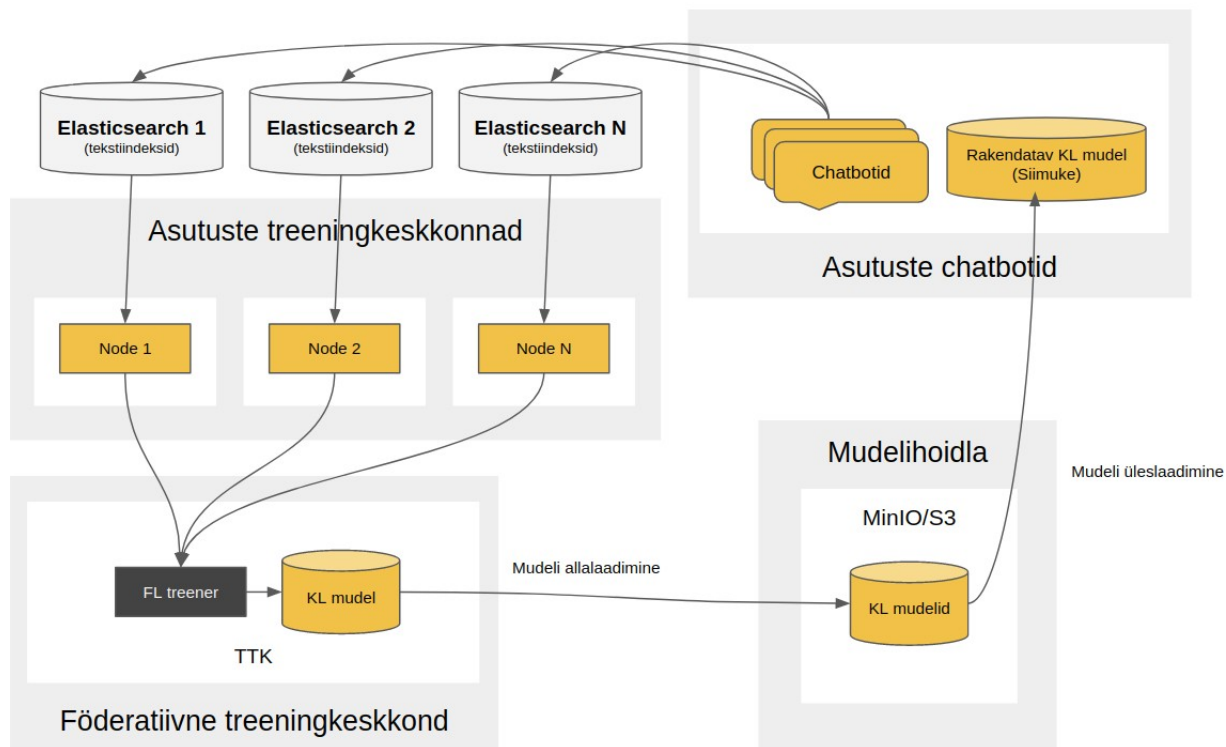
ja miinused ning lõpus antakse soovitusi, milliste eesmärkide täitmise jaoks on üks või teine arhitektuur mõistlikum.

### Föderatiivõpe

Födereeritud õpe põhineb detsentraliseeritud treenimise põhimõttel. See tähendab, et iga andmeid valdav asutus kasutab oma andmeid asutuse sees mudeli treenimiseks ning keskse mudeli uuendamiseks saadab iga asutus kesksesse rakendusse metainfot keskse mudeli uuendamiseks (joonis 24). Lahendus tagab arhitektuuriliselt selle, et treenitakse iga asutuse juures ning toorandmed ei liigu asutuse piiridest kunagi välja, kuid iga astus peab andmeid treenima sama tarkvara ja versiooniga. Väljapakutav lahendus põhineb horisontaalselt skaleeruval arhitektuuril (joonis 25), kus andmed salvestatakse vaid andmeid omava asutuse juures, mudelid on taastreenitavad, riskasutatavad ning versioneeritud.



Joonis 24: Tsentraliseeritud föderatiivõpe



Joonis 25: Asutusteülene suunaja föderatiivõpet kasutades

## Lahenduse komponendid:

### 1. Föderatiivne treeningkeskkond (Federated Learning Engine)

- Federate Learning Engine, kuhu *node*'id (lokaalne treener + Elastic) postitavad oma vektormudeli uuendused. Iga uue vektori uuenduse peale pannakse keskne mudel uuesti treenima. Uuendatud mudel laetakse üles S3-e (Mudelihoidlasse). Tarbijad tõmbavad uuendatud mudeli alla ning kasutavad uuendatud versiooni klassifitseerimisel.

### 2. Asutuste treeningandmete halduskeskkonnad

- Treeningandmete halduskeskkonna eesmärk on märgendada sissetulevaid emailde, vestluseid jms. Kui tekst on märgendatud, st talle on külge määratud näiteks valdkond, siis on võimalik andmeid kasutada mudeli treenimiseks.
- Iga ministeerium või asutus võib endale paigaldada enda oma TTK koos andmebaasiga, kuid samuti on võimalik kasutada näiteks ühe ministeeriumi haldusala sees üht TTK installatsiooni, kus igal asutusel on enda privaatsed projektid. Ühe TTK ja Elasticu kasutajate hulk pole arhitektuuriliselt piiratud, kuid skaleeruvuse huvides tundub kõige loogilisem tükeldus ministeeriumite järgi.
- Igal TTK projektil on enda andmetabelid, mis sisaldavad treeningandmeid. Asutused ei näe üksteise andmeid ehk teisi projekte. Kõikide asutuste andmed on indekseeritud Elasticsearchis. Ligipääs teiste asutuste andmetele

puudub nii eraldi installatsioonides kui ka ühe rakenduse sees.

- Klassifitseerimismudelite treenimisel toimub andmete ligipääsude haldamine TTK projektidena<sup>33</sup>, mis tagavad asutuse kasutajatele ligipääsu nende andmestikele.
- Federated Learningus toimub mudeli treenimine iga eraldi asutuse juures, pärast mudeli treenimist saadab iga *node* mudeliuudendused kesksesse Federated Learning Engine-isse.

### 3. Mudelihoidla

- Treenitud mudelid laetakse üles klassifitseerimismudelite hoidlasse, kus toimub nende versioonihaldus. Hoidlasse laetud klassifitseerimismudelid ei sisalda isikuandmeid.
- Klassifitseerimismudelite hoidla võimaldab hoiustada ja versioneerida nii avalikke (asutustevahelisi) kui privaatseid (asutusesiseseid) mudeleid.
- Mudelihoidla võib olla asutusepõhine, kuid võib olla ka riigiülene, nt nagu Koodivaramu. Riigiülese lahenduse puhul on soovi korral võimalik piirata kasutuse skooopi asutuseti.
- Mudelihoidla on S3 protokollil töötav vabavaraline objektihoidla (*Object Storage*), kuhu laetakse üles mudelite failid (kokkupakitud kaust faile). S3 süsteemid võimaldavad ühest objektist hoida mitu versiooni ühes *bucketis*<sup>34</sup>.

### 4. Ennustuskeskkond (Siimuke)

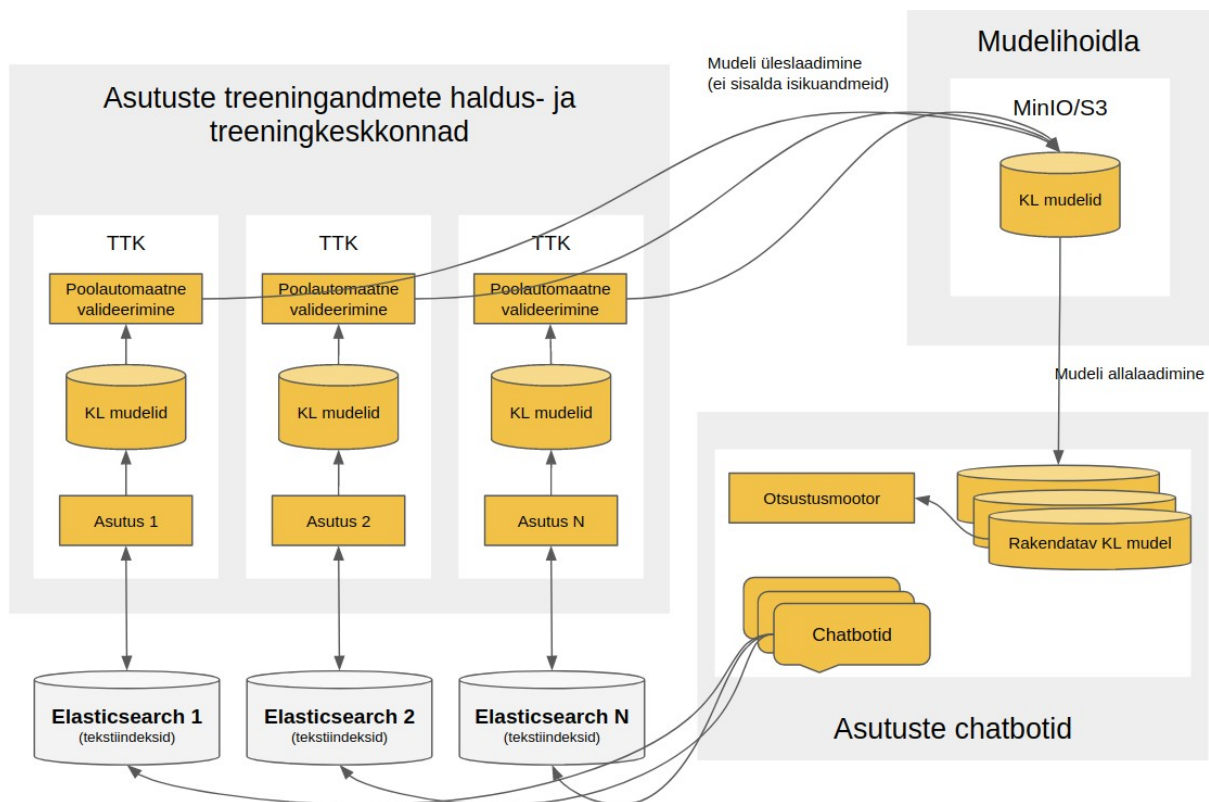
- Treenitud mudel(id) on kasutatavad ennustamiskekkonnas, kuhu need laetakse klassifitseerimismudelite hoidlast.
- Ennustuskeskkond on suuteline automaatselt mudeleid uuendama, kui hoidlasse ilmub sama mudeli uus versioon.
- Ennustuskeskkonda on võimalik konfigurereida nii kasutamiseks nii asutusesiseselt kui nende vaheliselt.
- Ennustuskeskkonda on võimalik kasutada nii e-kirjade märgendamiseks kui vestlusrobotite kasutatavate suhtluseesmärkide (*intent*) määramiseks.

## Teenusmodelite foorum

Alternatiivselt tsentraliseeritud föderatiivõppele on sama probleemi võimalik lahendada ka taaskasutades asutuste juures treenitud teenusmudeid (vt asutusesisese klassifitseerimise peatükki) läbi nende ennustustulemuste agregeerimise. Teenusmodelite foorumi eesmärk on taaskasutada asutusesiseseks klassifitseerimiseks loodud infrastruktuuri ja klassifitseerimismudeid, mis on allalaetavad S3 mudelihoidlast.

33 Vt <https://docs.texta.ee/et/terminology.html#project> või lisast 4.

34 <https://docs.min.io/docs/minio-bucket-versioning-guide.html>



Joonis 26: Asutusteülene klassifitseerimine teenusmodelite foorumit kasutades

## Lahenduse komponendid:

### 1. Asutuste treeningandmete halduskeskkonnad

- Treeningandmete halduskeskkonna eesmärk on märgendada sissetulevaid emaille, vestluseid jms. Kui tekst on märgendatud, st talle on külge määratud näiteks valdkond, siis on võimalik andmeid kasutada mudeli treenimiseks.
- Iga ministerium või asutus võib endale paigaldada enda oma TTK koos andmebaasiga, kuid samuti on võimalik kasutada näiteks ühe ministeriumi haldusala sees üht TTK installatsiooni, kus igal asutusel on enda privaatsed projektid. Ühe TTK ja Elasticu kasutajate hulk pole arhitektuuriliselt piiratud, kuid skaleeruvuse huvides tundub kõige loogilisem tükeldus ministeriumite järgi.
- Igal TTK projektil on enda andmetabelid, mis sisaldavad treeningandmeid. Asutused ei näe üksteise andmeid ehk teisi projekte. Kõikide asutuste andmed on indekseeritud Elasticsearchis. Ligipääs teiste asutuste andmetele puudub nii eraldi installatsioonides (segmenteeritus tagatud OS ja Elasticsearchi tasandil) kui tsentraalses installatsioonis (segmenteeritus tagatud TTK tasandil).
- Klassifitseerimismudelite treenimisel toimub andmete ligipääsude haldamine

TTK projektidena<sup>35</sup>, mis tagavad asutuse kasutajatele ligipääsu nende andmestikele.

## 2. Mudelihoidla

- Treenitud mudelid laetakse üles klassifitseerimismudelite hoidlasse, kus toimub nende versioonihaldus. Hoidlasse laetud klassifitseerimismudelid ei sisalda isikuandmeid.
- Klassifitseerimismudelite hoidla võimaldab hoiustada ja versioneerida nii avalikke (asutustevahelisi) kui privaatseid (asutusesiseseid) mudeleid.
- Mudelihoidla võib olla asutusepõhine, kuid võib olla ka riigiülene, nt nagu Koodivaramu. Riigiülese lahenduse puhul on soovi korral võimalik piirata kasutuse skoopt asutusesti.
- Mudelihoidla on S3 protokollil töötav vabavaraline objektihooldla (*Object Storage*), kuhu laetakse üles mudelite failid (kokkupakitud kaust faile). S3 süsteemid võimaldavad ühest objektist hoida mitu versiooni ühes *bucketis*<sup>36</sup>.

## 3. Otsustusmootor

- Otsustusmootor laeb alla ennustusmudelid mudelihoidlast.
- Otsustusmootor pöörduv reaalajas kõikide kasutatavate ennustusmudelite poole ning kogub kokku tulemused. Mudeleid, mis andsid kõrgeima ennustuskooriga vastuse, kasutatakse kombineeritult õigeima vastuse leidmiseks. Otsustusmootoris on võimalik anda erinevatele ennustusmudelitele kaalusid ning määrata prioriteete, mis võimaldab kasutustäpsust täppisestadistada. Otsustusmootor ei salvesta sissetulnud päringute sisu ega nendele antud vastuseid.
- Ennustuskeskkond on suuteline automaatselt mudeleid uuendada, kui hoidlasse ilmub sama mudeli uus versioon.
- Ennustuskeskkonda on võimalik kasutada vestlusrobotite kasutatavate suhtluseesmärkide (*intent*) määramiseks.
- Ennustusmudeleid ja vestlusroboteid kasutatakse koos Bürokratis, et pidada vestluseid. Vestlused, mis tekivad vestlusrobotis või üleüldiselt Bürokratis saadetakse tagasi asutuse treeningkeskkonda.

35 [Vt https://docs.texta.ee/et/terminology.html#project](https://docs.texta.ee/et/terminology.html#project) või lisast 4.

36 <https://docs.min.io/docs/minio-bucket-versioning-guide.html>



## Asutuseülese mudeli arhitektuuri valikud

	<b>Federated Learning</b>	<b>Teenusmudelite foorum</b>
<b>Mudelite lähteandmed</b>	Treenimine toimub asutuse juures. Lähteandmed asutuse Elasticust ei lahku. Keskmesse mudelisse saadetakse uuendamiseks mudeli metainfot.	Mudelid töötavad sõltumatult nagu chatboti arhitektuuris. Lähteandmed asutuse Elasticust ei lahku.
<b>Mudelite treenimine</b>	Treenimine asutuse TTK-s, aluseks sisendtekstid.	Treenimine asutuse TTK-s, aluseks sisendtekstid. Taaskasutatavad on ka asutusesisese klassifitseerimise mudelid.
<b>Lahenduse tehniline keerukus</b>	Tuleb implementeerida kogu födereeritud õppe tarkvara. <ul style="list-style-type: none"> <li>keskne mudelite haldur (FL treener), mis võtab vastu asutuse treeninginfot ja paneb tulemused kokku</li> <li>Asutuse treeningkeskkond, mis suhtleb keske halduriga (FL treener)</li> </ul>	Tehniliselt suures osas teostatud asutusesisese lahendusega. Lahendus põhineb mikroteenustel ja on horisontaalselt skaleeritav. Tuleb implementeerida otsustusmootor mudelite ennustustulemuste agregeerimiseks.
<b>Mudelite interpreteerimise ja täiendamise keerukus</b>	Treeneril puudub ligipääs kogu mudeli aluseks olevatele treeningandmetele (näeb ainult enda asutuse andmeid). Sellest tulenevalt on interpreteerimine/debugimine keerulisem.	Kõik mudelid treenitakse asutuses juba olemasoleva klassifitseerimistarkvaraga (tekib asutusesisese klassifitseerija implementeerimisega), mistõttu saavad asutuse spetsialistid algandmetega tutvuda ja neid vajadusel täiendada.
<b>Turvalisus</b>	Lähteandmed ja ennustuskeskkond asutuse sees, üle võrgu liiguvad <ul style="list-style-type: none"> <li>Küsimused mudelisse ja skoor tagasi</li> <li>Metainfo keskmesse FL treenerisse mudeli uuendamiseks</li> </ul>	Lähteandmed ja ennustuskeskkond asutuse sees, üle võrgu liiguvad ainult küsimused mudelisse ja skoor tagasi.
<b>Ohud</b>	<ul style="list-style-type: none"> <li>Olukorras, kus üks asutus täiendab enda alammudelit vigaste algandmetega, võidakse selle tagajärjel rikkuda födereeritud keskne mudel.</li> <li>Arhitektuuriliselt on keeruline lahendus, sest ei saa taaskasutada olemasolevaid mudeleid, mida kasutatakse asutusesiseseks klassifitseerimiseks.</li> <li>Federated Learningu arhitektuuri aluseks oleva algoritmi väljavahetamine on suurema keerukusega kui teemamudelite foorumi puhul. Selleks et võtta kasutusele uus klassifitseerimismudel (nt luuakse uus algoritm), tuleb kõikide asutuste treeningkeskkonnad uuendada ning nendega treenida uued mudelid (nii alam- kui keskne mudel). Sisuliselt on tegemist monoliitse arhitektuuriga (<i>Distributed monolith</i>), kuna mudelid ega treeningkeskkonnad ei ole eraldiseisvalt väljavahetatavad</li> </ul>	<ul style="list-style-type: none"> <li>Üks asutus võib üldise mudeli katki teha (versioneerimine võimaldab alati tagasi minna), kui nt ühe asutuse mudel on vigane.</li> <li>Mudelite väljund võib olla ebaühtlane, kui mudelite treenituse tase ei ole standardne (nt mõnel klassil liiga palju või valed andmed).</li> <li>Tasakaalustamine võib olla keeruline ja/või ajamahukas, kuna mudeleid on palju ja mudelite uuenemisel tuleb kaalusid potentsiaalselt muuta.</li> <li>Ennustamine on aeglasem kui FL puhul ning rikkealtim, kuna tuleb suhelda paljude erinevate teenustega kogumaks kokku kõikide mudelite ennustused.</li> </ul>

## Peamine kasutuslugu

Järgnevalt kirjeldatakse asutuste ülese klassifitseerija töötamise ja treenimise protsessi:

- Kasutaja laeb üles või talle imporditakse automaatselt andmed, mida on vaja märgendada (nt. vestlused või sõnumivahetused).
- Kasutaja avab TTK-s asutusteülese suunaja projekti, märgendab andmed.
- Pärast märgendamist paneb kasutaja mudeli treenima. Mudel treenitakse uute lisatud andmete pealt.
  - FL-i puhul peab treenima temale spetsiifilist mudelit. Teenusmudelite foorumi puhul saab kasutada olemasolevat asutusesisese klassifitseerija mudelit.
- Kasutaja valideerib mudeli ja lõpetab treenimise.
  - FL-i puhul genereeritaks metainfo, millega uuendada kesksel mudelit.
  - FL-i mudeliuudendused saadetakse kesksesse mudelisse ning lisatakse mudelisse.
- Mudel laetakse üles Mudelihoidlasse.
- Ennustuskeskkond laeb alla uuendatud targema mudeli.

## Kokkuvõte

Asutusteülese klassifitseerija arhitektuuris pakuti välja kaks erinevat lahendust, kirjeldati nende omadusi ning toodi välja plussid ja miinused. Empiirilisel on keeruline öelda, kumb lahendus on parem või sobivam Bürokrati arhitektuuri, kuna väga palju sõltub ka konkreetsest arhitektuuri implementatsioonist, kuid kindlalt võib väita, et födereeritud õpe lahendus on töö- ja ajamahukam. Kuid tulenevalt klientidega peetud aruteludest tundub Bürokrati visooni hetkel rohkem sobivat födereeritud õpe.

## Üldine turvalisus

- Kõik väline võrguliiklus üle HTTPS.
- Elasticsearch ega teised andmebaasid ei ole välismaailmale nähtavad. Igal asutusel on enda Elasticsearchi installatsioon. Seega turvalisust on võimalik tagada TTK või OS'i ja Elasticsearch'i tasandil.
- Kõikide toimingute logimine TTK-s.
- Tsentraalse autentimise kasutamine läbi TARA teenuse tulevikus. Kasutaja logib TARASSE sisse, mille alusel tehakse lokaalne TTK konto. Kuna TARAST rolle ei saa, siis TTK rakenduse moderaator määrab kasutajale projektid, millele tal on ligipääs.

## Kasutajate rollid

**Andmebaasi/rakenduse admin** - Haldab andmebaasi ning rakendust, äriliselt mingit rolli pole. Vajadusel pääseb ligi andmebaasile.

**Rakenduse moderaator** - loob uusi projekte, lisab inimesi projektidesse (saab ka luua ning lisada vajadusel läbi API). Ise projektide andmetele ligi ei pääse (ainult kasutajate õigused ning projektide loomine).

**Kasutaja** - Näeb ainult enda asutuse/projekti andmeid (mudelite treenimisel liidetakse teiste asutuste andmed, mida kasutaja ei näe).

## Logimine

Hetkel logitakse kindlalt ainult sissetulevate päringute otspunktide (ingl endpoints) nimed ja mõnel puhul ka suuremate taskide (TTK-s algatatud tööde) algused, perioodiline progress ja nende käigus tekkivad vead. Tehniliselt saab ka kõiki andmebaasi päringuid välja logida, sh Elasticsearchi päringu otspunkte. TTK tuleb täiendada, et viia see MKM-i IT-süsteemide logimise korrale (vt lisa 5) vastavaks.

Rakenduse logi:

```
{ "levelname": "INFO", "module": "tasks", "funcName": "start_mlp_worker", "lineno": 52, "name": "info_logger", "process": 147, "thread": 140248323344192, "message": "Starting applying mlp on the index for model ID: 34", "asctime": "2021-04-13 13:59:53,222" }
{ "levelname": "INFO", "module": "tasks", "funcName": "reindex_task", "lineno": 129, "name": "info_logger", "process": 148, "thread": 139733773530944, "message": "Reindexing succesfully completed.", "asctime": "2021-04-13 14:00:13,419" }
{ "levelname": "INFO", "module": "helpers", "funcName": "process_actions", "lineno": 18, "name": "info_logger", "process": 150, "thread": 140248323344192, "message": "Starting the processing of indices for MLP worker with ID of 34!", "asctime": "2021-04-13 14:00:29,875" }
{ "levelname": "INFO", "module": "tasks", "funcName": "reindex_task", "lineno": 129, "name": "info_logger", "process": 152, "thread": 139733773530944, "message": "Reindexing succesfully completed.", "asctime": "2021-04-13 14:01:54,788" }
{ "levelname": "INFO", "module": "tasks", "funcName": "reindex_task", "lineno": 89, "name": "info_logger", "process": 157, "thread": 139733773530944, "message": "Starting task 'reindex'.", "asctime": "2021-04-13 14:04:27,620" }
{ "levelname": "INFO", "module": "tasks", "funcName": "reindex_task", "lineno": 107, "name": "info_logger", "process": 148, "thread": 139733773530944, "message": "Updating index schema.", "asctime": "2021-04-13 14:04:27,620" }
```

```

"2021-04-13 14:04:53,869"}
{"levelname": "INFO", "module": "tasks", "funcName": "reindex_task",
"lineno": 112, "name": "info_logger", "process": 148, "thread":
139733773530944, "message": "Creating new index.", "asctime": "2021-
04-13 14:04:53,916"}
{"levelname": "INFO", "module": "tasks", "funcName": "reindex_task",
"lineno": 116, "name": "info_logger", "process": 148, "thread":
139733773530944, "message": "Indexing documents.", "asctime": "2021-
04-13 14:04:55,015"}
{"levelname": "INFO", "module": "tasks", "funcName":
"start_mlp_worker", "lineno": 52, "name": "info_logger", "process":
147, "thread": 140248323344192, "message": "Starting applying mlp on
the index for model ID: 35", "asctime": "2021-04-13 14:05:44,937"}
{"levelname": "INFO", "module": "helpers", "funcName":
"process_actions", "lineno": 18, "name": "info_logger", "process":
147, "thread": 140248323344192, "message": "Starting the processing
of indices for MLP worker with ID of 35!", "asctime": "2021-04-13
14:06:17,519"}

```

Docker logi:

```

[pid: 89|app: 0|req: 2489/8437] 90.190.180.90 () {46 vars in 768
bytes} [Wed Apr 14 09:14:26 2021] GET /api/v1/projects/8/ =>
generated 42426 bytes in 244 msec (HTTP/2.0 200) 5 headers in 188
bytes (1 switches on core 0)
[pid: 89|app: 0|req: 2490/8438] 90.190.180.90 () {46 vars in 773
bytes} [Wed Apr 14 09:14:27 2021] GET
/static/rest_framework/css/prettify.css => generated 817 bytes in 12
msec via sendfile() (HTTP/2.0 200) 5 headers in 153 bytes (0
switches on core 0)
[pid: 88|app: 0|req: 5948/8439] 90.190.180.90 () {46 vars in 789
bytes} [Wed Apr 14 09:14:27 2021] GET
/static/rest_framework/css/bootstrap-tweaks.css => generated 3385
bytes in 12 msec via sendfile() (HTTP/2.0 200) 5 headers in 154
bytes (0 switches on core 0)
[pid: 89|app: 0|req: 2491/8440] 90.190.180.90 () {46 vars in 771
bytes} [Wed Apr 14 09:14:27 2021] GET
/static/rest_framework/css/default.css => generated 1131 bytes in 3
msec via sendfile() (HTTP/2.0 200) 5 headers in 154 bytes (0
switches on core 0)
[pid: 88|app: 0|req: 5949/8441] 90.190.180.90 () {46 vars in 783
bytes} [Wed Apr 14 09:14:27 2021] GET
/static/rest_framework/css/bootstrap.min.css => generated 121457
bytes in 6 msec via sendfile() (HTTP/2.0 200) 5 headers in 156
bytes (0 switches on core 0)
[pid: 89|app: 0|req: 2492/8442] 90.190.180.90 () {48 vars in 844
bytes} [Wed Apr 14 09:14:35 2021] GET /api/v1/projects/8/reindexer/
=> generated 49588 bytes in 190 msec (HTTP/2.0 200) 5 headers in

```

```
174 bytes (1 switches on core 0)
[pid: 88|app: 0|req: 5950/8443] 90.190.180.90 () {46 vars in 885
bytes} [Wed Apr 14 09:14:36 2021] GET
/static/rest_framework/fonts/glyphicons-halflings-regular.woff2 =>
generated 18028 bytes in 11 msec via sendfile() (HTTP/2.0 200) 5
headers in 157 bytes (0 switches on core 0)
[pid: 89|app: 0|req: 2493/8444] 90.190.180.90 () {48 vars in 806
bytes} [Wed Apr 14 09:14:56 2021] GET /api/v1/rest-auth/user/ =>
generated 203 bytes in 13 msec (HTTP/2.0 200) 5 headers in 170
bytes (1 switches on core 0)
[pid: 89|app: 0|req: 2494/8445] 90.190.180.90 () {48 vars in 813
bytes} [Wed Apr 14 09:14:56 2021] GET /api/v1/projects/?undefined =>
generated 34407 bytes in 159 msec (HTTP/2.0 200) 5 headers in 166
bytes (1 switches on core 0)
[pid: 89|app: 0|req: 2495/8446] 90.190.180.90 () {48 vars in 788
bytes} [Wed Apr 14 09:14:57 2021] GET /api/v1/users/ => generated
33389 bytes in 56 msec (HTTP/2.0 200) 5 headers in 160 bytes (1
switches on core 0)
```

# Tasuvusanalüüs

„Riigi keskse juturoboti Bürokratt järelanalüüsis” (Riigihange 223660) on juba mitmed siin kontekstis olulised tegurid analüüsitud, mistõttu soovitame seda lugeda.

## Klassifitseerija kasutegur

Bürokraati mõju hindamiseks ameti ja riigi kontekstis tuleb vaadelda kahte eraldiseisvat juhtumit/ lähteülesannet. Esimeseks lähteülesandeks on valele asutusele tehtud pöördumise automaatne suunamine õigele asutusele ning seeläbi klienditoe koormuse vähendamine. Seda saab lahendada nii asutuseüleselt kui ka asutusesiseselt. Juhul, kui selliseid valepöördumistega asutusi on vähe (läbiviidud intervjuudest oli ainult RIA’l üle 10% selliseid pöördumisi), on soovitatav lahendada see läbi asutusesisesel klassifitseerimisel. Samas on sellise ülesande mõju asutusele klienditeenindusele kõige suurem ja lihtsamini mõõdetav (vt tabelit 22).

Vestluse maksumus	Pöördumiste maht kuus	Valele asutusele pöördumiste maht	Sääst kuus
2 eurot	2000	1200 (60%)	2400 eurot
4 eurot	2000	1200 (60%)	4800 eurot
8 eurot	2000	1200 (60%)	9600 eurot

Tabel 22: Vale asutuse mõjutegur RIA valepöördumiste mahu näitel

Kuigi valele asutusele saadetud pöördumiste hulk oli enamikes asutustes madal, siis selgus vajaduste kaardistamisel, et enamik saaksid kasu asutusesiseselt klassifitseerimiselt. Siin saab jällegi tuua kaks konkreetset ja mõõdetavat kasutusvaldkonda. Esiteks on paljude asutuste klienditeenindajatel olemas mallid, millega vastatakse tüüp küsimustele. Läbi automaatse klassifitseerimise on võimalik selliseid küsimusi tuvastada ning neile ka koheselt vastata (vt tabel 23).

Teisalt töid osad asutused ka välja, et nende asutuses on konkreetsed inimesed, kelle tööülesandeks on kirjade suunamine õigetele inimestele. Automaatne teemade tuvastamine aitaks nende inimeste tööaega kokku hoida, sest süsteem saaks neile kohe soovitada vastutava inimese, kes antud teemaga tegeleb. Võimalik on ka sellise töö täisautomatiseerimine (vt tabel 24).

Pöördumise hind	Tüüpküsimuste maht	Keskmine pöördumiste hulk asutustes kuus	Kokkuhoid eurodes
€4.00	10.00%	5000	€2,000.00
€4.00	20.00%	5000	€4,000.00
€4.00	30.00%	5000	€6,000.00
€4.00	40.00%	5000	€8,000.00
€4.00	50.00%	5000	€10,000.00
€8.00	10.00%	5000	€4,000.00
€8.00	20.00%	5000	€8,000.00
€8.00	30.00%	5000	€12,000.00
€8.00	40.00%	5000	€16,000.00
€8.00	50.00%	5000	€20,000.00

Tabel 23: Tüüpvastuste automatiseerimise kasutegur

Inimeste arv	Igakuine kulu	Kokkuhoid aastas
1	€1,970.00	€23,640.00
2	€3,940.00	€47,280.00
3	€5,910.00	€70,920.00
4	€7,880.00	€94,560.00

Tabel 24: Pöördumiste edasisuunaja abivahendi kasutegur

## Mudelite jätkuarendused

Töötava klassifitseerija saamiseks tuleb läbida tabelis 25 toodud sammud. Rahaline kulu on arvestatud eeldusega, et tekstianalüütik-andmeteadlase tunnihind on 70 eurot.

Samm	Seotud isikud	Ajakulu nädalates	Rahaline kulu
Treeningandmete sobival kujul kätte saamine (vt eeltötluse juhendit lk 29)	Asutus	4	0
Andmete eeltöötlemine juhendile vastavalt	Asutus või tekstianalüütik-andmeteadlane	3	8400
Märgendatud andmete puhul klasside tuvastamine intervjuude ja seniste märgendite/teemade/klasside analüüsi käigus	Asutuse klienditeenindajad ja tekstianalüütik-andmeteadlane	5	14000
Märgendamata andmete puhul klasside tuvastamine klasterdamisega <sup>37</sup>	Asutuse klienditeenindajad ja tekstianalüütik-andmeteadlane	5	14000
Märgendatud andmete koostamine või korrigeerimine tuvastatud klasside põhjal	Asutus ja tekstianalüütik-andmeteadlane	4	11200
Sobiva mudeli välja töötamine	tekstianalüütik-andmeteadlane	4 <sup>38</sup>	11200
Mudeli implementeerimine Siimukesse	tekstianalüütik-andmeteadlane	0.5	1400
Koolitus Siimukese kasutamiseks pöördumiste analüüsimiseks	Tekstianalüütik-andmeteadlane või projektijuht	1	2800
Koolitus Siimukeses mudelite treenimiseks, sh klassihaldus	Tekstianalüütik-andmeteadlane või projektijuht	1	2800

Tabel 25: Hinnanguline mudelite treenimise töövoog ja maksumus

## Nõuded arendusmeeskonnale

Täitjapoolne hinnang tarvilikust meeskonna koosseisust viperusteta arendamise jaoks.

**1 projektijuht** - viimase 5 (viie) aasta jooksul vähemalt 3 (kolme) aastane projektijuhtimise kogemus, viimase kolme aasta jooksul juhtinud vähemalt 3 (kolme) projekti, mille töömaht on vähemalt 600 (kuussada) inimtöötundi projekti kohta. Juhtimiseks loetakse, kui vastavas rollis on teostatud töid vähemalt 100 (ühesaja) inimtöötunni ulatuses.

<sup>37</sup> Selle sammu saab ära jätta, kui eelmises sammus said klassid tuvastatud.

<sup>38</sup> Ühe klassi treenimine võtab aega pool päeva.



**Vähemalt 1 süsteemianalüütik** - viimase 5 (viie) aasta jooksul vähemalt 3 (kolme) aastane töökogemus infosüsteemide analüütikuna.

**1 arhitekt** - viimase 5 (viie) aasta jooksul vähemalt 2 (kahe) aastane töökogemus infosüsteemide arhitektina. Vähemalt 1 infosüsteemi loomise või modifitseerimise ja juurutamise projektis on omandanud kogemuse, mis vastab järgnevale:

- andmete sisestamine, muutmine, kinnitamine ja tühistamine/kustutamine;
- kogemus Apache Kafka lahenduse implementeerimisel ja disainimisel;
- kogemus asünkroonsete hajussüsteemide disainimisel;
- kogemus REST ja WebSocket tehnoloogiatel põhinevate liideste disainimisel;
- kogemus vabavaraliste NLP tehnoloogiatega.

**1 täispinu tarkvaraarendaja** - viimase 5 (viie) aasta jooksul vähemalt 1 (ühe) aastane programmeerimise töökogemus. Vähemalt järgmiste tehnoloogiate/platvormidega: Java, React, PostgreSQL, Python. Vähemalt 1 infosüsteemi loomise või modifitseerimise ja juurutamise projektis on omandanud kogemuse, mis vastab järgnevale:

- kogemus REST ja WebSocket tehnoloogiatel põhinevate liideste arendamisel.
- kogemus vabavaraliste NLP lahendustega liidestumisel.
- kogemus Apache Kafka lahendusega liidestumisel.

**1 täispinu (noorem) tarkvaraarendaja** - viimase 5 (viie) aasta jooksul vähemalt 1 (ühe) aastane programmeerimise töökogemus. Vähemalt järgmiste tehnoloogiate/platvormidega: Java, React, PostgreSQL, Python. Vähemalt 1 infosüsteemi loomise või modifitseerimise ja juurutamise projektis on omandanud kogemuse, mis vastab järgnevale:

- kogemus REST ja WebSocket tehnoloogiatel põhinevate liideste arendamisel.
- kogemus vabavaraliste NLP lahendustega liidestumisel.
- kogemus Apache Kafka lahendusega liidestumisel.

**1 tekstianalüütik-andmeteadlane** - Omama kõrgharidust keelteaduse, loodusteaduste/matemaatika, rakendusliku füüsika või informaatika alal. Viimase 5 (viie) aasta jooksul vähemalt 3 (kolme) aastane andmeteadlase töökogemus, mille jooksul teinud töid vähemalt 3 (kolmes) projektis, mille sisuks on tekstiliste andmete analüüs, masinõppe mudelite arendus ja treenimine ning mille töömaht on vähemalt 600 (kuussada) inimtöötundi projekti kohta. Tööde tegemiseks loetakse, kui vastavas rollis on teostatud töid vähemalt 100

(ühesaja) inimtöötunni ulatuses. Vähemalt ühe projekti sisuks peab olema vabavaraliste juturobitite abil vestluste loomine. Vestluste loomise all peetakse silmas treeningandmete loomist, keeletötlusvahenditega töötlemist, töövoos ("pipeline") seadistamist vastavalt vestluskeelele ning juturoboti treenimist.

# Kodaniku pöördumiste klassifitseerijaga isikuandmete töötlemisele kohalduvad andmekaitseõuded

Käesoleva projekti käigus kaardistasime kodaniku pöördumiste klassifitseerijaga isikuandmete töötlemisele kohalduvad peamised andmekaitseõuded, mille järgimine aitab maandada ohtusid ja riske, mis võivad klassifitseerijaga isikuandmete töötlemisega seoses füüsiliste isikute õigustele ja vabadustele kaasneda. Klassifitseerijaga isikuandmete töötlemisel võib olulisemate ohtude ja riskidena välja tuua järgmised aspektid:

- isikuandmete väärkasutuse oht (sh pöördumiste ja nendes sisalduvate isikuandmete või treenitud klassifitseerimismudelite töötlemine eesmärkidel, mis ei seonu asutustele edastatud pöördumiste lahendamiseks ja nendele vastamiseks pöördumiste märgendamise);
- võimalus teha järeldusi füüsiliste isikute (sh eriliselt haavatavate andmesubjektide kategooriatesse kuuluvate andmesubjektide) eraeluliste asjaolude kohta;
- isikuandmete ülemäärane töötlemine (sh klassifitseerimismudelite treenimiseks ja pöördumiste märgendamiseks üleliigsete isikuandmete töötlemine);
- valede, ebasobivate või aegunud isikuandmete kasutamine klassifitseerimismudelite treenimiseks;
- klassifitseerija poolt isikuandmete töötlemine (sh nii klassifitseerimismudelite treenimise kui ka pöördumiste märgendamise kontekstis) ei ole andmesubjektidele läbipaistev ja arusaadav;
- pöördumiste märgendamise võimalik negatiivne mõju pöördumiste edasisele töötlemisele (sh diskrimineerimise soodustamine);
- oht isikuandmete konfidentsiaalsusele, terviklusele ja kättesaadavusele.

Andmekaitseõuete kaardistamise käigus koostatud andmekaitseõudeid kirjeldav tabel lähtub detailanalüüsis kirjeldatud arhitektuurilistest lahendustest ning hõlmab nii asutusesisese klassifitseerija lahendust kui ka asutusteülese klassifitseerija lahendusi. Kaardistatud andmekaitseõuete puhul võib eristada rakenduse arendust mittemõjutavaid andmekaitseõudeid ja arendust mõjutavaid andmekaitseõudeid.

Arendust mittemõjutavad andmekaitseõuded ei mõjuta otseselt klassifitseerija kui rakenduse arhitektuuri, funktsionaalsusi ja arendust. Nende nõuete täitmine on eelkõige klassifitseerijaga isikuandmeid töötleva iga vastutava töötleja ehk asjakohase asutuse pädevuses ja vastutusallas ning neid on suuresti võimalik täita sõltumata rakenduse arhitektuurist ja funktsionaalsustest. Lisaks eeldab kõnealuste andmekaitseõuete sisuline

täitmine muuhulgas arvestamist asjaoludega, mis on iga asutuse poolt isikuandmete töötlemise kontekstis iseloomulikud just vastavale asutusele. Arendust mittemõjutavate andmekaitse nõuete kontekstis tuleks asutustel klassifitseerijaga isikuandmete töötlemisel tähelepanu pöörata eelkõige järgnevale:

- andmesubjektide kategooriate määramine ja sellest johtuv töötlemise riskitaseme määramine;
- isikuandmete liikide määramine ja sellest johtuv töötlemise riskitaseme määramine;
- asutuse kui vastutava töötleja määramine ning vajadusel kaasvastutavate töötlejate vastutusvaldkondade kokkuleppimine;
- volitatud töötlejate valimine ja töötlemiskokkulepete sõlmimine;
- töötlemise eesmärkide määramine, sh hindamine, kas varem kogutud andmete töötlemine klassifitseerimismudelite treenimiseks on kooskõlas algse andmete kogumise eesmärgiga;
- töötlemise läbipaistvuse tagamine ja andmesubjektide kohane teavitamine;
- korralduslike (organisatsiooniliste ja füüsiliste) turvameetmete rakendamine vastavalt asutuse organisatsiooni eripäradele.

Arendust mõjutavate andmekaitse nõuete täitmine on samuti iga isikuandmeid töötleva vastutava töötleja ehk asjakohase asutuse pädevuses ja vastutusalas, kuid nende nõuete täitmise võimalikkus sõltub suuresti rakenduse funktsionaalsustest ja omadustest. Seetõttu tuleb vastavad nõuded rakenduse arendamisel otseselt aluseks võtta ja rakenduse funktsionaalsustega tagada. Olulisemate arendust mõjutavate andmekaitse nõuetega seoses peab klassifitseerija arendamisel olema tagatud, et:

- treenitud klassifitseerimismudelid ei sisalda üheski etapis isikuandmeid ning treenimisel ei tehta toiminguid andmesubjekti isiku tuvastamiseks ega isiku identiteedi või muude eraeluliste asjaolude kohta järelduste tegemiseks;
- ennustuskeskkonnas töödeldakse iga pöördumist eraldi ning andmeid erinevate pöördumiste vahel ei koondata, võrrelda, seostata ega tehta selle pinnalt järeldusi füüsilise isikute kohta;
- on välistatud võimalus kasutada treenitud klassifitseerimismudeleid füüsiliste isikute kohta järelduste tegemiseks või mistahes muudel eesmärkidel, mis ei seondu otseselt asutustele edastatud pöördumiste lahendamiseks ja nendele vastamiseks pöördumiste märgendamiseks;
- klassifitseerimismudelite kvaliteetne treenimine ning pöördumiste märgendamine saavutatakse töödeldes selleks võimalikult vähe isikuandmeid;

- klassifitseerimismudelite treenimiseks kasutatavad pöördumised oleks ajakohased, õiged ja sobivad;
- pöördumiste märgendamisel rakenduse siseselt töödeldud andmeid edasiseks töötlemiseks ei salvestata ega kasutata hiljem muul eesmärgil töötlemiseks;
- mudelite treenimisel isikuandmete väärkasutuse võimalus on välistatud (rakendatud meetmeid selle riski maandamiseks);
- pöördumiste märgendamisel ei tehta järeldusi konkreetse andmesubjekti kohta ning ka pöördumise sisu osas tehakse ennustus üksnes pöördumise teema / pädeva adressaadi kohta;
- märgendamisel ei oleks sisulist mõju märgendamisele järgnevale pöördumise sisulisele lahendamisele ja vastamisele (ei soodustaks diskrimineerimist);
- rakenduse süsteemis on treenimisel kasutatud pöördumiste ja nendes sisalduvate isikuandmete pöördumatu kustutamise võimalus (võimalusel automaatne kustutamine) pärast asutuse poolt säilitamiseks määratud tähtaja möödumist;
- pöördumiste märgendamisel salvestatakse andmed ennustuskeskkonnas üksnes töötlemiseks vajaliku aja vältel;
- on võimalik tuvastada, kas ja milliseid andmesubjekti isikuandmeid on töödeldud klassifitseerimismudelite treenimiseks ja pöördumis(t)e märgendamiseks;
- treeningandmete haldus- ja treeningkeskkonnas salvestatud isikuandmeid on võimalik põhjendamatu viivitusega pöördumatult kustutada;
- treeningandmete haldus- ja treeningkeskkonnas salvestatud isikuandmete töötlemist on võimalik piirata;
- isikuandmete ja süsteemi konfidentsiaalsuse, tervikluse, kättesaadavuse ja vastupidavuse/jätkusuutlikkuse tagamiseks rakendatakse asjakohaseid tehnilisi turvameetmeid;
- isikuandmete töötlemine rakenduses on logitud.

Kaardistatud andmekaitseõuete järgimisel on klassifitseerijaga isikuandmete töötlemisel tekkivat võimalikku ohtu ja riske füüsiliste isikute õigustele vabadustele võimalik olulises osas vähendada. Olenemata eeltoodust tuleks siiski enne klassifitseerija tegelikku asutuste igapäevatoos kasutuselevõtmist täiendavalt analüüsida klassifitseerijaga isikuandmete töötlemise asjaolusid lõpliku lahenduse ja iga vastava asutuse kontekstis eraldi.

Isikuandmete töötlemist reguleerib eelkõige Euroopa Parlamendi ja nõukogu määrus (EL) 2016/679, 27. aprill 2016, füüsiliste isikute kaitse kohta isikuandmete töötlemisel ja selliste andmete vaba liikumise ning direktiivi 95/46/EÜ kehtetuks tunnistamise kohta (isikuandmete kaitse üldmäärus, edaspidi **IKÜM**) ja isikuandmete kaitse seadus (**IKS**). Seejuures tuleb arvestada, et isikuandmete töötlemisele õiguskaitseseaduse poolt sülletoetusega, et isikuandmete töötlemisele õiguskaitseseaduse poolt sülletoetusega, et isikuandmete töötlemisele õiguskaitseseaduse poolt sülletoetusega,

avastamisel ja menetlemisel ning karistuse täideviimisel ei kohaldata IKÜM-i ning nimetatud olukorras kohaldataks eelkõige IKS 4. peatükis sätestatud (§§ 12-50). Seega tuleb õiguskaitseasutuste (nt Politsei- ja Piirivalveamet) poolt andmete töötlemisel arvestada ka IKS 4. peatüki sätetega.

Järgnevas tabelis on kajastatud olulisemad andmekaitsealased nõuded, mis Klassifitseerija poolt isikuandmete töötlemisele kohalduvad. Seejuures saab eristada nõudeid, mis ei mõjuta otseselt Klassifitseerija arhitektuuri, funktsionaalsusi ja arendust ning nõudeid, mida tuleb Klassifitseerija arendamisel arvesse võtta. Nõuded, mis Klassifitseerija arhitektuuri, funktsionaalsusi ja arendust eelduslikult ei mõjuta, on tabelis märgistatud.

4. Klassifitseerijale kohalduvad andmekaitseenõuded <sup>39</sup>			
	Nõue	Nõude sisu / asjakohasus Klassifitseerija kontekstis	Vastavus
	Selles tulbas on välja toodud andmete töötlemisele kohalduv nõue.	Selles tulbas on kirjeldatud nõude sisu/asjakohasus Klassifitseerija ja Klassifitseerija poolt teostatava andmete töötlemise kontekstis.	<p>Palun märkige siia, kas Klassifitseerija:</p> <p>e) vastab kirjeldatud nõudele / võimaldab vastavat nõuet töötlemisel järgida või</p> <p>f) ei vasta nõudele / seda ei ole võimalik nõudele vastavalt kasutada.</p> <p>Kui nõue ei mõjuta otseselt Klassifitseerija arhitektuuri, funktsionaalsusi ja arendust on lahter eeltäidetud.</p>
i)	<b>Isikuandmed ja töötlemine</b>		
1)	<p><b>Andmesubjektid</b> – töötlemise toimingute registreerimise kontekstis tuleb kirjeldada andmesubjektide kategooriad (IKÜM art 30 lg 1(c); IKS § 37 lg 1 p 5). Andmesubjektide kuulumine eriliselt haavatavate andmesubjektide kategooriasse suurendab andmetöötlemisega seotud riskitaset.</p>	<p>Pöördumiste teemade ring võib olla väga lai (eriti asutusteülese Klassifitseerija kontekstis) ning sisuliselt võib andmesubjektiks olla iga füüsiline isik, kes pöördub asutuse poole või kelle andmeid võidakse pöördumises avaldada. Seejuures võivad andmesubjektide hulka kuuluda eriliselt haavatavate andmesubjektide kategooriasse kuuluvad isikud (alaealised, töötajad, varjupaiga taotlejad, vanurid, vaimse häirega isikud, patsiendid, muul põhjusel vastutava töötaja suhtes nõrgemal positsioonil olevad isikud). Õiguskaitse asutuste puhul (nt PPA) on andmesubjektideks ka väärteo- ja kriminaalmenetluse subjektid, keda võib olla vajalik eristada IKS § 18 sätestatud kategooriatena. Samuti tuleb arvestada, et asutustega suhtlemisel võib eeldada, et andmesubjekt on asutuse suhtes alati suuremal või vähemal määral nõrgemal positsioonil.</p> <p>Klassifitseerija poolt andmete töötlemise kontekstis on võimalik määratleda andmesubjektide kategooriatena:</p> <ul style="list-style-type: none"> <li>• pöördumise edastanud füüsiline isik,</li> <li>• füüsiline isik, kelle andmeid pöördumises avaldatakse.</li> </ul> <p>Pöördumiste lahendamise ja nendele vastamise kui töötlemistoimingu kajastamisel saavad asutused täpsemalt määrata andmesubjektide kategooriad, keda vastav</p>	<p>Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötlejana konkreetne asutus.</p>

39 Järgnevas tabelis kasutatakse mõisteid „andmed“ ja „töötlemine“ vastavalt tähenduses „isikuandmed“ ja „isikuandmete töötlemine“.

		<p>andmete töötlemine konkreetse asutuse kontekstis puudutab.</p> <p>Kuna andmesubjektide ring võib tõenäoliselt hõlmata erilisel haavatavate andmesubjektide kategooriasse kuuluvaid andmesubjekte, tuleb arvestada töötlemise kõrgema riskitasemega.</p>	
2)	<p><b>Andmete koosseis</b> – töötlemise toimingute registreerimise kontekstis tuleb kirjeldada isikuandmete liike (IKÜM art 4(1), art 30 lg 1(c); IKS § 37 lg 1 p 5). Teatud liiki andmeid tuleb pidada tundlikumaks ning eriliiki andmete töötlemisele on kehtestatud täiendavad tingimused (IKÜM art 9, art 10; IKS § 20).</p>	<p>Pöördumiste teemade ring võib olla väga lai (eriti asutusteülese Klassifitseerija kontekstis), mistõttu ei ole töödeldavaid andmete liike võimalik üheselt piiritleda. Seejuures võivad andmete koosseisu kuuluda nii isiku tuvastusandmed (nimi, isikukood), kontaktandmed, terviseandmed, kuid ka mistahes muud eraelulised andmed. Muuhulgas võivad andmed sisaldada tundlikke isikuandmeid (nt sotsiaalabi saamine, mitteavalik teave inimese varandusliku seisuga kohta, sõnumisaladusega kaetud sideandmed ning muu inimese kohta käiv info, mis on avaliku teabe seaduse (<b>AvTS</b>) alusel käsitletav kui juurdepääsupiiranguline teave, eriliiki isikuandmeid (andmeid, millest ilmneb rassiline või etniline päritolu, poliitilised vaated, usulised või filosoofilised veendumused või ametiühingusse kuulumine, geneetilisi andmeid, terviseandmeid või andmeid füüsilise isiku seksuaalelu ja seksuaalse sättumuse kohta) ning süüdimõistvate kohtuotsuste ja süütegudega seotud isikuandmeid.</p> <p>Pöördumiste lahendamise ja nendele vastamise kui töötlemistoimingu kajastamisel saavad asutused täpsemalt määratleda peamised andmete liigid, mida vastav andmete töötlemine konkreetse asutuse kontekstis puudutab, kuid määratlemata andmete liikide töötlemist ei ole võimalik täielikult välistada.</p> <p>Kuna töödeldavate andmete hulka võivad tõenäoliselt kuuluda tundlikud ja eriliiki andmed, tuleb arvestada töötlemise kõrgema riskitasemega.</p>	<p>Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötlejane konkreetne asutus.</p>
3)	<p><b>Töötlemistoiming</b> – töötlemine on igasugune toiming andmetega ning andmete töötlemise toimingud tuleb registreerida (IKÜM art 4(2), art 30 lg 1; IKS § 37 lg 1).</p>	<p>Klassifitseerija näol on tegemist rakendusega, mis teostab eelnevalt treenitud klassifitseerimismudeleid kasutades asutustele edastatud pöördumiste sisu tekstianalüüsi ning teeb selle pinnalt ennustusi (märgendab) pöördumise sisu/eesmärgi osas. Klassifitseerija poolt teostatav andmetöötlus on piiratud lõik pikemast isikuandmete töötlemise protsessist.</p> <p>Klassifitseerija poolt andmete töötlemisel saab eristada erinevate töötlemistoimingutena:</p> <ul style="list-style-type: none"> <li>• klassifitseerimismudelite treenimise raames toimuvat töötlemist,</li> <li>• pöördumiste märgendamise raames toimuvat töötlemist.</li> </ul>	



		<p>Klassifitseerimismudelite treenimise raames sisestab (laeb üles / impordib) asutuse poolne kasutaja andmed pöördumistena treeningandmete haldus- ja treeningkeskkonda, misjärel toimub andmete töötlemine klassifitseerimismudelite treenimiseks. Selle töötlemise käigus võib muuhulgas toimuda andmete koondamine, võrdlemine ning seoste ja mustrite tuvastamine.</p> <p><b>Klassifitseerija arendamisel on tagatud, et treenitud klassifitseerimismudelid ei sisalda üheski etapis isikuandmeid. Samuti ei tee Klassifitseerija klassifitseerimismudelite treenimisel mistahes toiminguid andmesubjekti isiku tuvastamiseks ega isiku identiteedi või muude eraeluliste asjaolude kohta järelduste tegemiseks.</b></p> <p>Pöördumiste märgendamise raames sisestatakse andmed pöördumistena Klassifitseerija ennustuskeskkonda, kus toimub tekstianalüüs ning pöördumiste märgendamine. Vastavalt Klassifitseerija poolt määratud märgendusele on võimalik pöördumise edasine töötlemine (edastamine pädevale asutusele/asutusesiseselt).</p> <p><b>Klassifitseerija arendamisel on tagatud, et ennustuskeskkonnas töötleb Klassifitseerija iga pöördumist eraldi ning ei koonda, võrdle ega seosta andmeid erinevate pöördumiste vahel ega tee selle pinnalt järeldusi. Klassifitseerija ei analüüsi andmeid andmesubjekti isiku tuvastamise eesmärgil ega tee isiku identiteedi kohta mistahes järeldusi.</b></p>	
ii)	<b>Töötlejad</b>		
1)	<p><b>Vastutav töötleja</b> – vastutavaks töötlejaks on isik/asutus, kes määrab kindlaks isikuandmete töötlemise eesmärgid ja vahendid (IKÜM art 4(7), art 24, art 26; IKS § 29 lg 1, § 31).</p>	<p>Klassifitseerija poolt andmete töötlemisel (sh nii klassifitseerimismudelite treenimisel kui pöördumiste märgendamisel) on vastutavateks töötlejateks asutused, kellele adresseeritud pöördumisi Klassifitseerija töötleb.</p> <p>Asutusesisese Klassifitseerija puhul on iga asjakohane Klassifitseerijat kasutav asutus iseseisvaks vastutavaks töötlejaks, kes ka iseseisvalt töötlemise nõuetekohasuse eest vastutab.</p> <p>Asutusteülese Klassifitseerija poolt andmete töötlemisel (sh nii klassifitseerimismudelite treenimisel kui pöördumiste märgendamisel) määravad asutused töötlemise eesmärgid ja vahendid kindlaks ühiselt (sisuliselt on tegemist ühiste eesmärkidega), mistõttu on tegemist kaasvastutavate töötlejatega (IKÜM art 26; IKS 31). IKÜM-ist tulenevate kohustuste täitmisel peavad asutused kokku leppima iga kaasvastutava töötleja vastutusvaldkonna ning selle kokkuleppe põhitingimused tuleb avaldada andmesubjektidele.</p>	<p>Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötlejana konkreetne asutus.</p>

2)	<p><b>Volitatud töötledjad</b> – volitatud töötledjateks on isikud, kes töötlevad andmeid vastutava töötledja nimel (IKÜM art 4(8)). Volitatud töötledjatega peab olema sõlmitud isikuandmete töötlemise leping (IKÜM art 28; IKS § 29 lg 2-5, § 30).</p>	<p>Volitatud töötledjateks on kõik isikud, kes teevad Klassifitseerijaga seoses andmetega mistahes toiminguid. Eelkõige võib Klassifitseerija kontekstis olla volitatud töötledjaks isik, kelle serveris/süsteemis (sh pilves) Klassifitseerijaga andmeid töödeldakse (välja arvatud kui tegemist on ühe kaasvastutava töötledjaga).</p> <p>Volitatud töötledjate ringi ja nende poolt andmete töötlemise tingimused peavad asutused vastutavate töötledjatena määrama kindlaks enne asjakohase töötlemise alustamist. Seejuures peab olema nii faktiliselt kui lepingutega tagatud andmete töötlemise turvalisus ja õiguspärasus volitatud töötledjate poolt.</p>	<p>Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötledjana konkreetne asutus.</p>
iii) <b>Eesmärgi piirangu ja seaduslikkuse põhimõte</b>			
1)	<p><b>Töötlemise eesmärgid</b> – andmeid võib töödelda üksnes täpselt ja selgelt kindlaksmääratud ning õiguspärastel eesmärkidel (IKÜM art 5 lg 1(b); IKS § 14 p 2).</p>	<p>Pöördumiste lahendamise ja nendele vastamise protsessi osana on Klassifitseerijaga andmete töötlemise laiemaks eesmärgiks iga asutuse poolt pöördumiste lahendamine ja nendele vastamine. Klassifitseerija poolt andmete töötlemise kitsamaks eesmärgiks on pöördumiste märgendamine / suhtluseesmärkide määramine, mis lihtsustaks ja kiirendaks pöördumiste edasist töötlemist (edastamine pädevale asutusele või asutusesisesele adressaadile). Ning selle eesmärgi alam-eesmärgina võib omakorda vaadelda klassifitseerimismudelite treenimist.</p> <p>Klassifitseerimismudeli treenimise raames toimuva töötlemisega seoses tekib siiski küsimus, kas asutuste poolt varem (enne andmesubjektide teavitamist sellise andmete töötlemise eesmärgi kohta – vt p Viga: viidatud allikat ei leitud) pöördumistega seoses kogutud andmete kasutamine mudelite treenimise eesmärgil on kooskõlas nende andmete algse kogumise eesmärgiga (IKÜM art 6 lg 4; IKS § 16). Seega tuleks enne selliste (varem kogutud) andmete kasutamist mudelite treenimiseks analüüsida iga asutuse lõikes eraldi, kas andmete töötlemine mudelite treenimise eesmärgil on kooskõlas nende andmete algse kogumise eesmärgiga ja selle kohta andmesubjektidele avaldatud teabega.</p>	<p>Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötledjana konkreetne asutus.</p>
2)	<p><b>Õiguslik alus</b> – andmete töötlemine peab olema seaduslik (IKÜM art 5 lg 1(a); IKS 14 p 1) ja põhinema (määratletud) õiguslikul alusel (IKÜM art 6 lg 1, IKS § 15).</p>	<p>Vastutavate töötledjatena (sh nii eraldiseisvate vastutavate töötledjatena Klassifitseerija asutusesisesese kasutamise puhul kui ka kaasvastutavate töötledjatena asutusteülese Klassifitseerija puhul) peab iga asutus määratlema õigusliku aluse asjakohaseks andmete töötlemiseks.</p> <p>Pöördumiste lahendamine ja nendele vastamine on asutusele eelkõige oma seadusest tuleneva kohustuse täitmine. Nimelt tuleneb märgukirjale ja selgitustaotlusele vastamise ning kollektiivse pöördumise esitamise seaduse (<b>MSVS</b>) § 5 lg-st 1 ning avaliku teabe seaduse (<b>AvTS</b>) § 18 lg 1 asutusele vastamiskohustus. Lisaks tuleb MSVS § 5 lg-st 3 ja 4 ning AvTS § 21 lg 1 asutustele kohustus edastada</p>	

		<p>pöördumine vajadusel teisele pädevale asutusele. Seejuures eeldab nimetatud kohustuste täitmine vaieldamatult pöördumise jõudmist õige adressaadini (nii asutusteülesele kui asutusesiseselt) ning selleks vajalikku andmete töötlemist.</p> <p>Eeltoodust tulenevalt saavad asutused tugineda Klassifitseerija poolt andmete töötlemisel järgnevale:</p> <ul style="list-style-type: none"> <li>• pöördumiste märgendamise raames toimuva andmete töötlemise õigusliku alusena saavad asutused tugineda eelkõige seadusejärgse kohustuse täitmisele (IKÜM art 6 lg 1(c) ning MSVS § 5 lg 1-4 ja AvTS § 18 ja 21). Asutused, kes võivad olla eriliiki andmeid sisaldavate pöördumiste adressaatideks peaksid täiendavalt tuginema andmete töötlemisele olulise avaliku huviga seotud põhjustel siseriikliku õiguse alusel (IKÜM art 9 lg 2(g)) ning konkreetsele õigusaktile, mille alusel vastav asutus eriliiki andmeid töötleb.</li> </ul> <p>Olukordades, kus õiguskaitseasutus töötleb Klassifitseerijaga andmeid tegevuse raames, mis on suunatud süütegude tõkestamisele, avastamisele, menetlemisele või karistuste täideviimisele, on vajalik tugineda täiendavale alusele vastavalt IKS 15 ja § 20 sätestatule;</p> <ul style="list-style-type: none"> <li>• klassifitseerimismudelite treenimise raames toimuva töötlemise õigusliku aluse saab siduda sama õigusliku alusega, millele tuginetakse pöördumiste märgendamisel. Mudelite treenimise eesmärk on olemuslikult ja lahutamatu seotud pöördumiste märgendamisega ning asjakohaste andmete abil treenitud mudeleid ei saa kasutada muudel eesmärkidel andmete töötlemiseks. Seetõttu võib asuda seisukohale, et tegemist ei ole iseseisva töötlemiseeesmärgiga, mis vajaks eraldiseisvat õiguslikku alust.</li> </ul> <p><b>Seejuures peab olema (võimalusel ka Klassifitseerija arendamisel) välistatud võimalus kasutada treenitud klassifitseerimismudeleid muudel eesmärkidel kui asutustele edastatud pöördumiste lahendamisel ja nende vastamisel pöördumiste märgendamiseks ning igal juhul ei tohi andmetega treenitud klassifitseerimismudeleid kasutada konkreetsete füüsiliste isikute kohta mistahes järelduste tegemiseks.</b></p>	
iv)	<b>Minimaalsuse, õigsuse, õigluse ja läbipaistvuse põhimõte</b>		
1)	<b>Andmete minimaalsus/kvaliteet</b> – kogutakse üksnes selliseid	Klassifitseerimismudelite treenimiseks toimuva töötlemise raames analüüsitakse enamasti terviklikke pöördumisi. Kuigi ei ole välistatud, et kvaliteetsete klassifitseerimismudelite loomiseks on vajalik kasutada treenimisel pöördumisi	

	<p>andmeid ja sellises ulatuses, milles see on vajalik andmetöötlaste eesmärkide saavutamiseks (IKÜM art 5 lg 1(c); IKS § 14 p 3).</p>	<p>tervikuna ja võimalikult laia valimikuna, <b>tuleks Klassifitseerija arendamisel siiski minimaalsuse põhimõtet arvestada ning kaaluda (sh sobivaima masinõppemudeli valimisel) võimalusi saavutada klassifitseerimismudelite kvaliteetne treenimine töödeldes selleks võimalikult vähe andmeid.</b></p> <p>Pöördumiste märgendamiseks toimuva andmete töötlemise raames analüüsitakse eraldi iga konkreetset pöördumist. Kuigi ka siin ei ole välistatud, et märgendamise kvaliteedi tagamiseks on otstarbekas ja vajalik analüüsida pöördumist tervikuna koos kõikide selles sisalduvate andmetega, <b>tuleks Klassifitseerija arendamisel minimaalsuse põhimõtet arvestada ning kaaluda võimalusi saavutada pöördumiste märgendamine töödeldes selleks võimalikult vähe andmeid.</b></p>	
<p>2)</p>	<p><b>Andmete ajakohasus ja õigsus</b> – andmete töötlemisel peab olema töötlemise eesmärki arvesse võttes tagatud andmete ajakohasus ja õigsus (IKÜM art 5 lg 1 (d); IKS § 14 p 4).</p>	<p>Klassifitseerimismudelite treenimisel peab olema tagatud, et treenimiseks kasutatavad pöördumised (hõlmates neis sisalduvaid andmeid) on klassifitseerimismudelite treenimise raames ajakohased ja õiged ning mudelite treenimine toimub pidevalt andmetega, mille ajakohasus ja õigsus on kontrollitud. <b>Teiste sõnadega peab olema tagatud, et klassifitseerimismudelite treenimiseks kasutatavad pöördumised (hõlmates neis sisalduvaid andmeid) ei oleks vananenud (nt asutuste pädevuse või struktuuri muutumise korral muudatusele eelnenud pöördumiste kasutamine) ega valed (nt asutuse töötajale edastatud e-kirjad, mis sisaldavad isiklike sõnumeid, mitte pöördumist asutuse poole, asutusesiselt omavahelised sõnumid vms).</b></p> <p>Pöördumiste märgendamise puhul töötleb Klassifitseerija üksnes Klassifitseerijasse edastatud pöördumisi ning eelduslikult vajavad need pöördumised edasiseks töötlemiseks (pöördumise lahendamiseks ja nende vastamiseks pädeva asutuse või asutusesisese adressaadi poolt) märgendamist. <b>Seejuures peab olematagatud, et Klassifitseerija siselt töödeldud andmeid edasiseks töötlemiseks ei salvestata ega kasutata hiljem muul eesmärgil töötlemiseks</b> (edasine töötlemine pöördumise lahendamiseks ja sellele vastamiseks toimub väljaspool Klassifitseerija tööloiku). Klassifitseerija analüüsib ja saab lähtuda üksnes konkreetsetes pöördumises sisalduvatest andmetest ning nende andmete täiendamine, parandamine või uuendamine ei ole Klassifitseerija töö lõikes asjakohane ega vajalik. Seega puudub Klassifitseerija poolt pöördumiste märgendamise raames vajadus tagada võimalus andmete hilisemaks ajakohastamiseks või parandamiseks.</p>	
<p>3)</p>	<p><b>Õiglus</b> – andmete töötlemine peab olema õiglane (IKÜM art 5 lg 1(a); IKS § 14 p 1).</p>	<p>Klassifitseerimismudelite treenimise raames isiku andmete töötlemine ei avalda andmesubjektile tõenäoliselt tuntuvat mõju, kuna üksiku isiku andmed iseseisvana/üksikult ei mõjuta oluliselt töötlemise tulemusena loodud klassifitseerimismudelit. <b>Klassifitseerija arendamisel peab olema tagatud, et treenitud klassifitseerimismudeleid ei saa kasutada hiljem füüsiliste isikute</b></p>	

		<p><b>kohta järelduste tegemiseks.</b> Selliselt pole ettenähtav, et andmete kasutamine klassifitseerimismudelite treenimiseks võiks põhjustada andmesubjektile kahjulikke tagajärgi.</p> <p><b>Siiski tuleb mudelite treenimiseks andmete töötlemise raames välistada andmete väärkasutamise võimalus</b> (nt võimalus kasutada treeningandmeid muudel eesmärkidel või teistel eesmärkidel kasutatavate mudelite treenimiseks) ning tagada, et töötlemine oleks andmesubjektidele läbipaistev (vt p Viga: viidatud allikat ei leitud).</p> <p><b>Pöördumiste märgendamise raames peab olema tagatud, et Klassifitseerija ei tee mistahes järeldusi konkreetse andmesubjekti kohta ning ka pöördumise sisu osas tehakse ennustus üksnes pöördumise teema / pädeva adressaadi kohta.</b> Selliselt võib asuda seisukohale, et märgendamiseks vajaliku töötlemisega kaasnev riive andmesubjektide privaatsusele on minimaalne ning kahjulik mõju andmesubjektidele ebatõenäoline. Seejuures saab arvestada, et pöördumise sisuline lahendamine ja sellele vastamine toimub hilisema mitteautomaatse töötlemise käigus pädeva asutuse pädeva töötaja poolt, mis omakorda vähendab võimalust, et pöördumiste märgendamisel toimuva töötlemise tulemusena saabub andmesubjektile kahjulik tagajärg (nt väära märgenduse puhul pöördumise edastamisel valele adressaadile saab järgneva töötlemise käigus pöördumise ikkagi õigele adressaadile edastada). Seega ei ole üldjuhul põhjendatud pidada kõnealuse töötlemisega kaasnevat riivet andmesubjektile taotletava eesmärgi suhtes ebaõiglaseks.</p> <p><b>Siiski tuleks töötlemise kahjulike mõjude vähendamiseks tagada, et märgendamisel ei oleks sisulist mõju märgendamisele järgnevale pöördumise sisulisele lahendamisele ja vastamisele (märgendamise tulemus ei tohi mõjutada seda, kuidas pädev asutus või asutusesisene adressaat pöördumise lahendab või sellele vastab; nt ei tohiks märgenditena kasutada mõisteid, mis indikeerivad, kuidas pöördumist lahendada või soodustavad isikute diskrimineerimist või kahjulike hinnangute andmist – nt mõisteid oluline/ebaoluline, kiirem, võlgnikud, rikkujad vms).</b> Samuti tuleb tagada, et töötlemine oleks andmesubjektidele läbipaistev (vt p Viga: viidatud allikat ei leitud).</p>	
4)	<p><b>Läbipaistvus ja andmesubjektide teavitamine</b> – andmete töötlemine peab olema andmesubjektile läbipaistev (IKÜM art 5 lg 1(a)) ning andmesubjekti tuleb andmete</p>	<p>Kõik asutused vastutavate töötlejatena peavad avaldama teabe (täiendama vastavalt oma andmekaitsetingimusi/privaatsuspoliitikat) andmete töötlemise kohta i) klassifitseerimismudelite treenimise raames ja ii) pöördumiste märgendamise raames. Teave tuleb andmesubjektidele avaldada / teatavaks teha enne konkreetse töötlemise alustamist.</p>	<p>Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötlejana konkreetne asutus.</p>

	töötlemisega seotud asjaoludest teavitada (IKÜM art 12-14; IKS § 22-23).		
<b>v)</b>	<b>Andmete säilitamine ja kustutamine</b>		
<b>1)</b>	<b>Säilitamine ja kustutamine</b> – andmeid säilitatakse üksnes seni, kuni see on vajalik töötlemise eesmärkide täitmiseks (IKÜM art 5 lg 1(e)); IKS § 17) ning andmete säilitamise ajavahemik tuleb määratleda ja andmesubjektidele avaldada (IKÜM art 13 lg 2(a), art 14 lg 2(a); IKS § 23 lg 1 p 3).	<p>Klassifitseerimismudelite treenimiseks töödeldavad pöördumised koos andmetega sisestatakse ja salvestatakse Klassifitseerija treeningandmete halduskeskkonda (asutusesisese Klassifitseerija puhul treeningandmete haldus- ja treeningkeskkond). Asutused vastutavate töötlejatena peavad määratlema (ja andmesubjektidele avaldama) tähtaja, mille vältel pöördumisi koos andmetega Klassifitseerija süsteemis säilitatakse ning see tähtaeg ei tohi olla pikem kui mudelite treenimiseks vajalik. <b>Klassifitseerija süsteemis peab olema tagatud pöördumiste ja nendes sisalduvate andmete pöördumatu kustutamise võimalus pärast asutuste poolt säilitamiseks määratud tähtaja möödumist ning eelistatult võiks pöördumiste ja nendes sisalduvate andmete kustumine toimuda pärast säilitamiseks määratud tähtaja möödumist automaatselt.</b></p> <p><b>Pöördumiste märgendamise raames peab olema tagatud, et andmed salvestatakse Klassifitseerija ennustuskeskkonnas üksnes töötlemiseks vajaliku aja vältel ning pöördumisi ega nendes sisalduvaid andmeid Klassifitseerija süsteemi säilitamise eesmärgil ei salvestata ega säilitata.</b> Klassifitseerija süsteemis on salvestatud üksnes klassifitseerimismudelid, mis ei sisalda isikuandmeid.</p>	
<b>vi)</b>	<b>Automaatsed otsused ja andmete edastamine</b>		
<b>1)</b>	<b>Automaatsete otsuste tegemine ja tagajärjed</b> – andmesubjektil on õigus, et tema kohta ei võetaks vastu üksnes automatiseeritud töötlemisel põhinevat otsust (IKÜM art 22; IKS § 21).	Klassifitseerijaga andmete töötlemisel ei võeta andmesubjektide kohta vastu andmesubjekti märkimisväärselt mõjutavaid otsuseid, mis põhinevad üksnes automatiseeritud töötlemisel – klassifitseerimismudelite treenimise raames ei võeta üldse vastu andmesubjekti kui isikut mõjutavaid otsuseid ning pöördumiste märgendamise raames toimub automaatselt üksnes pöördumise eeltöötlus. Pöördumise sisuline lahendamine ja sellele vastamine (ehk andmesubjekti mõjutav otsus toimub hilisema mitteautomaatse töötlemise käigus pädeva asutuse pädeva töötaja poolt.	Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötlejana konkreetne asutus.
<b>2)</b>	<b>Andmete saajad ja edastamise viis</b> – andmete töötlemise toimingute registreerimise	Klassifitseerimismudeli treenimise raames ei edastata andmeid kolmandatele isikutele ega ka teistele kaasvastutavatele töötlejatele (seejuures ei ole volitatud töötleja käsitletav kolmanda isikuna). See kehtib nii asutusesisese kui ka asutusteülese Klassifitseerija puhul, <b>kuna kõikide kavandatud</b>	

	<p>kontekstis tuleb määratleda isikud, kellele andmed edastatakse ning teatud juhtudel ka edastamisel kohaldatavad kaitsemeetmed (IKÜM art 30 lg 1(d) ja (e); IKS § 35, § 37 lg 1 p 4).</p>	<p><b>Klassifitseerijaarhitektuurimudelite puhul toimub treenimine asutuse infrastruktuuril ja on välistatud andmete liikumine asutusest väljapoole.</b></p> <p>Pöördumiste märgendamise raames toimub tekstianalüüs ning selle pinnalt pöördumise märgendamine. Vastavalt Klassifitseeriija märgendusele on võimalik pöördumise edasine töötlemine (edastamine pädevale asutusele / adressaadile asutusesiseselt), kuid praeguse lahenduse piires ei edasta Klassifitseeriija ise andmeid kolmandatele isikutele ega teistele kaasvastutavatele töötlejatele (asutusteülese Klassifitseeriija puhul).</p>	
<b>vii)</b>	<b>Andmesubjekti õiguste tagamine</b>		
<b>1)</b>	<p><b>Õigus tutvuda andmetega</b> – andmesubjekti õigus saada teavet tema andmete töötlemise kohta ja töödeldavate andmete koopia (IKÜM art 15; IKS § 24)</p>	<p>Andmesubjekti vastava õiguse tagamiseks peab asutusel olema võimalik andmesubjekti taotlusel:</p> <ul style="list-style-type: none"> <li>• <b>tuvastada kas ja milliseid andmesubjekti andmeid on töödeldud / töödeldakse klassifitseerimismudelite treenimiseks.</b> Samuti peab olema võimalik edastada andmesubjektile koopia klassifitseerimismudelite treenimiseks kasutatud tema andmetest;</li> <li>• <b>tuvastada kas ja milliseid andmesubjekti andmeid töödeldi pöördumise märgendamiseks</b> (sisuliselt kas/milliseid i) vastava isiku pöördumisi või ii) vastava isiku andmeid sisaldavaid pöördumisi märgendamiseks töödeldi). Kuivõrd pöördumiste märgendamise raames andmeid ei salvestata, ei saa kõnealuse töötlemise kontekstis kohalduda andmete koopia esitamise kohustus.</li> </ul>	
<b>2)</b>	<p><b>Õigus nõuda ebaõigete või -täielike andmete parandamist</b> – andmesubjekti õigus nõuda ebaõigete andmete parandamist (IKÜM 16; IKS § 25 lg 1-2)</p>	<p>Klassifitseeriija analüüsib ja saab lähtuda üksnes konkreetses pöördumises sisalduvatest andmetest (sh nii klassifitseerimismudeli treenimise raames kui ka pöördumiste märgendamise raames) ning nende andmete täiendamine, parandamine või uuendamine ei ole Klassifitseeriija poolt teostatava töötlemise eesmärki arvestades asjakohane ega võimalik.</p>	<p>Nõue ei mõjuta otseselt Klassifitseeriija arendust.</p>
<b>3)</b>	<p><b>Õigus nõuda andmete kustutamist</b> – andmesubjekti õigus nõuda teatud tingimustel andmete kustutamist (IKÜM art 17; IKS § 25 lg 3).</p>	<p>Andmesubjekti vastava õiguse tagamiseks peab olema tagatud võimalus põhjendatu viivitusega pöördumatult kustutada Klassifitseeriija treeningandmete haldus- ja treeningkeskkonnas salvestatud andmed.</p> <p>Kuivõrd pöördumiste märgendamise raames andmeid ei salvestata, ei saa kõnealuse töötlemise kontekstis kohalduda andmete kustutamise kohustus.</p>	

4)	<b>Õigus isikuandmete töötlemise piiramisele</b> – andmesubjekti õigus nõuda teatud tingimustel andmete töötlemise piiramist (IKÜM art 18; IKS § 25 lg 4-5)	Andmesubjekti vastava õiguse tagamiseks peab olema tagatud võimalus <b>Klassifitseerija treeningandmete halduskeskkonnas</b> (asutusesisese Klassifitseerija puhul treeningandmete haldus- ja treeningkeskkond) <b>piirata andmesubjekti andmete töötlemine</b> . Töötlemise piiramine tähendab, et neid andmeid üksnes säilitatakse passiivselt ning nende andmetega mistahes muude toimingute tegemine (sh mudelite treenimiseks kasutamine) on välistatud.  Kuivõrd pöördumiste märgendamise raames andmeid ei salvestada, ei saa kõnealuse töötlemise kontekstis kohalduda andmete töötlemise piiramise kohustus.	
5)	<b>Õigus nõuda andmete ülekandmist</b> – andmesubjekti õigus nõuda andmete ülekandmist temale endale või teisele töötlejale (IKÜM art 20)	Kuivõrd Klassifitseerijaga andmete töötlemisel (sh nii klassifitseerimismudeli treenimise raames kui ka pöördumiste märgendamise raames) tuginevad vastutavad töötlejad õigusliku alusena seadusest tuleneva kohustuse täitmisele, <sup>40</sup> siis andmete ülekandmise õigus ei kohaldu.	Nõue ei mõjuta otseselt Klassifitseerija arendust.
6)	<b>Õigus esitada vastuväiteid</b> - andmesubjektil on teatud juhtudel õigus oma konkreetsest olukorrast lähtudes esitada vastuväiteid teda puudutavate andmete töötlemise suhtes (IKÜM art 21)	Kuivõrd Klassifitseerijaga andmete töötlemisel (sh nii klassifitseerimismudeli treenimise raames kui ka pöördumiste märgendamise raames) tuginevad vastutavad töötlejad õigusliku alusena seadusest tuleneva kohustuse täitmisele (IKÜM art 6 lg 1(c)), siis vastuväidete esitamise õigus ei kohaldu. <sup>41</sup>	Nõue ei mõjuta otseselt Klassifitseerija arendust.
<b>viii) Turvalisuse tagamine ja vastutus</b>			
1)	<b>Korralduslikud turvameetmed</b> – kohustus rakendada töötlemisega seotud ohtudele vastava turvalisuse ja andmete õiguspärase töötlemise tagamiseks asjakohaseid	Klassifitseerija poolt andmete töötlemisel (sh nii klassifitseerimismudeli treenimise raames kui ka pöördumiste märgendamise raames) peab olema tagatud andmete ja süsteemi konfidentsiaalsus, terviklus, kättesaadavus ja vastupidavus/jätkusuutlikkus. Eelneva tagamiseks kohustuvad asutused vastutavate töötlejatena ja volitatud töötlejad rakendama korralduslikke turvameetmeid vastavalt organisatsiooni eripärasid arvestades koostatud riskihinnangutele ja infoturbe parkikatele. Seejuures hõlmavad korralduslikud turvameetmed:	Nõue ei mõjuta otseselt Klassifitseerija arendust ja selle täitmise tagab vastutava töötlejana konkreetne asutus.

40 Vastav õigus kohaldub üksnes juhtudel, kui andmete töötlemine põhineb andmesubjekti nõusolekul (IKÜM art 6 lg 1(a) või art 9 lg 2(a)) või lepingul (IKÜM art 6 lg 1(b)).

41 Vastav õigus kohaldub üksnes juhtudel, kui andmete töötlemine toimub avalikes huvides oleva ülesande täitmise (IKÜM art 6 lg 1(e)) või õigustatud huvi (IKÜM art 6 lg 1(f)) alusel või otseturunduse või teadus- või ajaloouringute või statistilisel eesmärgil.



	<p>korralduslikke turvameetmeid alates töötlemisvahendite kindlaksmääramisest ning kogu andmetöötluse protsessi vältel (IKÜM art 5 lg 1(f), art 25 ja art 32 lg (1); IKS § 43).</p>	<ul style="list-style-type: none"> <li>• organisatsioonilisi meetmeid –kehtestatavad sisekorraeeskirjad, toimimisujuhendid, personali väljaõpe jne. Iga asjakohane asutus peab selles kontekstis hindama vajadust sellised reeglid Klassifitseerija poolt andmete töötlemise kontekstis kehtestada/uuendada;</li> <li>• füüsilisi meetmeid – füüsiliste turvameetmete rakendamine vastavalt organisatsiooni eripäradele (sh kaitsmaks serverit, milles Klassifitseerija jookseb, füüsiliste rünnete ja kahjustuste eest).</li> </ul>	
<p>2)</p>	<p><b>Tehnilised turvameetmed</b> – kohustus rakendada töötlemisega seotud ohtudele vastava turvalisuse ja andmete õiguspärase töötlemise tagamiseks asjakohaseid tehnilisi turvameetmeid alates töötlemisvahendite kindlaksmääramisest ning kogu andmetöötluse protsessi vältel (IKÜM art 5 lg 1(f), art 25 ja art 32 lg (1); IKS § 43).</p>	<p>Klassifitseerija poolt andmete töötlemisel (sh nii klassifitseerimismudeli treenimise raames kui ka pöördumiste märgendamise raames) <b>peab olema tagatud andmete ja süsteemi konfidentsiaalsus, terviklus, kättesaadavus ja vastupidavus/jätkusuutlikkus.</b> Eelneva tagamiseks kohustuvad asutused vastutavate töötlejatena ja volitatud töötlejad rakendama tehnilisi turvameetmeid vastavalt organisatsiooni eripärasid arvestades koostatud riskihinnangutele ja infoturbepraktikatele. Seejuures hõlmavad tehnilised turvameetmed Klassifitseerija kontekstis:</p> <ul style="list-style-type: none"> <li>• Klassifitseerija väliseid (info)tehnoloogilisi meetmeid, mida rakendab iga asutus vastavalt organisatsiooni eripärasid arvestades koostatud riskihinnangutele ja infoturbepraktikatele (nt infotehnoloogilised turvameetmed serverile, milles Klassifitseerija andmeid töötleb);</li> <li>• Klassifitseerija sisesed infotehnoloogilised meetmed ehk Klassifitseerija disaini ja arhitektuuri sisse ehitatud turvameetmed. Mh peaksid olema rakendatud meetmed järgnevalt: <ul style="list-style-type: none"> <li>○ takistatud on volitamata juurdepääs treeningandmetele ning nendega manipuleerimise võimalus (sh volitamata andmete sisestamise, tutvumise, muutmise, kopeerimise ja kustutamise võimalus);</li> <li>○ takistatud on Klassifitseerija kasutamine andmesidevahendite abil volitamata isikute poolt;</li> <li>○ volitatud kasutajatele on tagatud juurdepääs üksnes volituste piires;</li> <li>○ andmetega tehtud toimingute logimine (vt ka p Viga: viidatud allikat ei leitud);</li> <li>○ tagada süsteemi toimimine ja ilmnevatest toimimisvigadest</li> </ul> </li> </ul>	

		<p>teavitamine;</p> <ul style="list-style-type: none"> <li>○ tagada võimalus vajadusel andmeid ja süsteemi taastada;</li> <li>○ takistada andmete moonutamine süsteemirikete tagajärjel;</li> <li>○ võtta tarvitusele meetmed rünnetest tulenevate ohtude maandamiseks, sh tehisintellektil ja masinõppel põhinevatele süsteemidele iseloomulikke ründeid arvestades.</li> </ul>	
<p>3)</p>	<p><b>Vastutus</b> – vastutava töötleja kohustus rakendada meetmeid andmekaitseõuete täitmiseks ja vajadusel seda tõendama (IKÜM art 24; IKS § 29 lg 1).</p>	<p>Vastutav töötleja peab olema võimeline tõendama andmekaitseõuete täitmist. See tähendab, et Klassifitseerijaga andmete töötlemisel peavad asutused olema võimelised demonstreerima, et isikuandmeid töödeldakse Klassifitseerija töö kontekstis andmekaitseõuetele vastavalt. Samuti peavad asutused olema võimelised tagantjärele uurima ja tuvastama andmetega seotud rikkumisi.</p> <p>Klassifitseerija lahenduse kontekstis on vastava kohustuse täitmiseks <b>vajalik tagada Klassifitseerija poolt andmete töötlemise logimine</b>. Eelkõige peaksid logid võimaldama tuvastada/jälgida:</p> <ul style="list-style-type: none"> <li>● treeningandmete sisestamist, muutmist, vaatamist, edastamist ja kustutamist Klassifitseerija treeningandmete haldus- ja treeningkeskkonnas;</li> <li>● klassifitseerimismudelite versioonide seost treeningandmetega (võimaldaks vajadusel tuvastada, millised mudelite versioonid on treenitud manipuleeritud treeningandmetega);</li> <li>● pöördumiste märgendamise toiminguid Klassifitseerija ennustuskeskkonnas;</li> <li>● süsteemi väärkasutuse või sissetungikatseid.</li> </ul> <p>Logide tõendusväärtuse aitab tagada nende varustamine digitaalse ajatempliga.</p> <p>Logikirjete säilitamise tähtsajad ning reeglid logide kontrollimiseks ja nendega tutvumiseks tuleb kehtestada igal asutusel vastavalt organisatsiooni eripäradele.</p>	