TARTU ÜLIKOOL
Arvutiteaduse instituut
Infotehnoloogia mitteinformaatikutele õppekava

Kristel Agu

Psühhoosi prodroomi sümptomite eraldamine meditsiinitekstidest treeningandmestike loomiseks

Magistritöö (15 EAP)

Juhendajad: Sulev Reisberg, PhD
Kairit Sirts, PhD

Tartu 2024

Psühhoosi prodroomi sümptomite eraldamine meditsiinitekstidest treeningandmestike loomiseks

Lühikokkuvõte:

Käesolevas magistritöös loodi pool-automaatset metoodikat kasutades kolm märgendatud treeningandmestikku psühhoosi prodroomi sümptomite tuvastamiseks meditsiinitekstidest. Treeningandmestike koostamiseks kasutati 2012.-2019. aastate 10% juhuslikult valitud Eesti rahvastiku meditsiinidokumente, millest leiti esmasele prodroomile viitavatele diagnoosidele vastavad tekstid (2780 teksti) ning tükeldati need edasise töötlemise lihtsustamiseks lauseteks (31 009 lauset). Esmane andmestik logistilise regressiooni mudeli treenimiseks koostati tükeldatud lausetest otsitavat sümptomit sisaldavate lausete välja sõelumisel regulaaravaldise abil ning nende töö autori poolt käsitsi märgendamisel. Logistilise regressiooni mudeliga töötamiseks leiti lausetele Eesti tekstikorpusel eeltreenitud Word2Vec mudelit kasutades keskmised vektorid. Selleks, et leida järelejäänud lausete hulgast veelgi otsitavat sümptomit sisaldavaid lauseid, mida näiteks regulaaravaldisega ei suudetud tuvastada, kasutati esmasel andmestikul treenitud mudelit. Pärast esmase andmestikuga mudeli treenimist alustati iteratiivse protsessiga, kus mudeliga ennustati allesjäänud lausete hulgast otsitavat sümptomit sisaldavaid lauseid, märgendati need käsitsi, lisati olemasolevale andmestikule ning korrati protsessi kuni mudel ei ennustanud uusi lauseid. Logistilise regressiooni mudeli kasutamine otsitava sümptomiga lausete tuvastamiseks lihtsustas treeningandmestiku koostamise protsessi, vähendades käsitsi läbivaadatavate lausete hulka. Töö tulemusena valmisid 799 märgendatud lausega andmestik psühhoosi prodroomi sümptomi „veider käitumine" eraldamiseks, 643 lausega sümptomite „depersonalisatsioon" ja/või „derealisatsioon" eraldamiseks ning 1176 lausega andmestik „paranoilise luulu" ja/või „kahtlustamise" eraldamiseks, mida saab kasutada edasiste mudelite treenimisel.

Võtmesõnad:

Psühhoosi prodroom, psühhiaatriliste sümptomite eraldamine, treeningandmestiku koostamine

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika. P176 Tehisintellekt

Extraction of psychosis prodromal symptoms from medical texts for training dataset creation

Abstract:

The current master thesis aimed to create three annotated training datasets for the extraction of psychosis prodromal symptoms from medical texts using semi-automatic methods. For this purpose, a dataset of medical documents from 10% randomly selected Estonian population in the years 2012-2019 was used. These documents were filtered by the ICD-10 diagnoses evident during psychosis prodrome (2780 texts) and split into sentences (31 009) for simplification of the further workflow. A dataset was created from the sentences, which were filtered using a regular expression and annotated manually by the author, and used to train an initial logistic regression model. To create the features for the logistic regression model, word embeddings were found for each word in a sentence using the Word2Vec model pre-trained on the Estonian Reference Corpus and an average embedding was calculated for the whole sentence. After that, an iterative process was initiated, where more sentences containing the symptom were predicted from the remaining data, annotated by the author, added to the existing dataset and repeated until the model finds no new sentences. Using the logistic regression model for the extraction of psychosis prodromal symptoms simplified the dataset creation process and reduced the amount of work put into searching the sentences manually. As a result of this master thesis, an annotated training dataset with 799 sentences for extracting the psychosis prodrome symptom “odd behaviour”, a dataset with 643 sentences for the symptoms “depersonalization” and/or “derealization” and a dataset with 1176 sentences for the symptoms “paranoid delusions” and/or “suspiciousness” were created.

Keywords:

Psychosis prodrome, extraction of psychiatric symptoms, training dataset generation

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics. P176 Artificial intelligence

Sisukord

Sissejuhatus

Psühhoosi korral on inimesel avaldunud psühhootilised sümptomid (näiteks hallutsinatsioonid ja luulumõtted), mille tõttu on tema reaalsustaju tugevalt häiritud ning esineda võivad kõrvalekalded nii mõtlemises, tajumises, käitumises, meeleolus kui ka motoorikas [1, 2]. Psühhoosiga kulgevaid haiguseid nimetatakse psühhootilisteks häireteks, mille alla kuuluvad näiteks skisofreenia, äge mööduv psühhootiline episood ning skisoafektiivne häire [1]. Neid häireid eristavad üksteisest nii neid põhjustavad tegurid kui ka sümptomite kestus ja iseloom [3].

Elus esmakordselt avalduvat psühhoosi nimetatakse esmaseks psühhoosiks, millele võib eelneda variatiivse kestuse ning ebaspetsiifiliste sümptomitega periood [2]. Sellist kuude kuni aastate pikkust perioodi, kus avalduvad esimesed võimalikule arenevale psühhoosile viitavad ilmingud ja sümptomid, nimetatakse prodroomiks [2]. Indikaatoriteks võivad olla näiteks ebatavaline käitumine, sotsiaalne isoleerumine, isikliku hügieeni halvenemine ning muutused meeleolus [2].

Prodroomi võib vaadelda kui psühhoosi tekkeriski staadiumit, mille võimalikult varajasel tuvastamisel on võimalik rakendada meetodeid psühhoosi väljakujunemise ennetamiseks või haiguse kulu mõjutamiseks [2]. Prodroomi mitmekülgse ning individuaalse olemuse tõttu on psühhoosiriskiga isikute avastamine aga keeruline ülesanne, mistõttu on oluline leida uusi ning tõhusaid meetodeid nendeni jõudmiseks. Hetkel tugineb riskirühma kuuluvate isikute tuvastamine suuresti inimeste enda abiotsivale käitumisele, kuid antud juhul on tegemist võrdlemisi vähese efektiivsusega sekkumisega, mis tuvastab ainult 5%-12% esmastest psühhoosidest [4].

Selleks, et jõuda psühhoosiriskiga isikuteni võimalikult varakult, kogub populaarsust digitaalselt talletatud terviseandmete kasutamine prodroomi sümptomite ning riskipatsientide tuvastamiseks. Prognostilise võimekusega automaatsete tööriistade väljatöötamiseks on vajalikud kvaliteetsed, usaldusväärsed ning võimalikult mahukad treeningandmestikud, mis vajavad märgendamisel sageli mitme osapoole koostööd, muutes tööprotsessi aja- ja ressursikulukaks. Psühhiaatria valdkonnas on tegemist valdavalt heterogeensete ning mittestruktureeritud tekstiliste andmetega, millest sümptomite ekstraheerimine on seetõttu raskendatud.

Käesoleva magistritöö eesmärk on panustada märgendatud treeningandmestike loomisesse psühhoosi prodroomi sümptomite tuvastamiseks meditsiinitekstidest. Töös keskendutakse peamiselt psühhoosiriski sõeluuringust pärinevale sümptomile „veider käitumine”, mis võiks peegeldada psühhootilistele häiretele iseloomuliku tegelikkuse tunnetamisvõime ja reaalsustaju häirumist. Sümptomi tuvastamiseks on vajalik välja töötada pool-automaatne metoodika, millega lihtsustada treeningandmestiku koostamiseks sobivate tekstide tuvastamist haiguslugudest ning vähendada seeläbi manuaalse töö mahtu. Siinkohal on rakendatud regulaaravaldise abil valitud lausetest moodustunud ja käsitsi annoteeritud treeningandmestiku iteratiivset suurendamist logistilise regressiooni mudeli poolt tehtud ennustuste alusel.

Treeningandmestiku iteratiivse loomise metoodika hindamiseks on kõrvutatud eelpool väljatoodud sümptomile lisaks kahte teist võimalikule prodroomile viitavat sümptomite gruppi, milleks on „depersonalisatsioon” ja „derealisatsioon” ning „paranoilised luulumõtted” ja „kahtlustamine”.

Käesoleva magistritöö esimeses peatükis tuuakse välja olulised mõisted ning teises peatükis antakse teoreetiline ülevaade psühhoosi prodroomi olemusest, meditsiinitekstidest psühhiaatriliste sümptomite eraldamisest ning nende potentsiaalsest rakendamisest psühhiaatria valdkonnas. Kolmandas peatükis on kirjeldatud töökäik ning neljandas peatükis on toodud välja töö olulisemad tulemused. Viies peatükk sisaldab endas meetodi valideerimist ning kuuendas peatükis analüüsitakse töö tulemusi ning antakse ka soovitusi teema edasiseks käsitlemiseks. Töö viimastes peatükkides esitatakse magistritöö kokkuvõte ning tänusõnad. Töö lõpust on võimalik leida kasutatud kirjanduse loetelu ning tööga kaasaskäivad lisad.

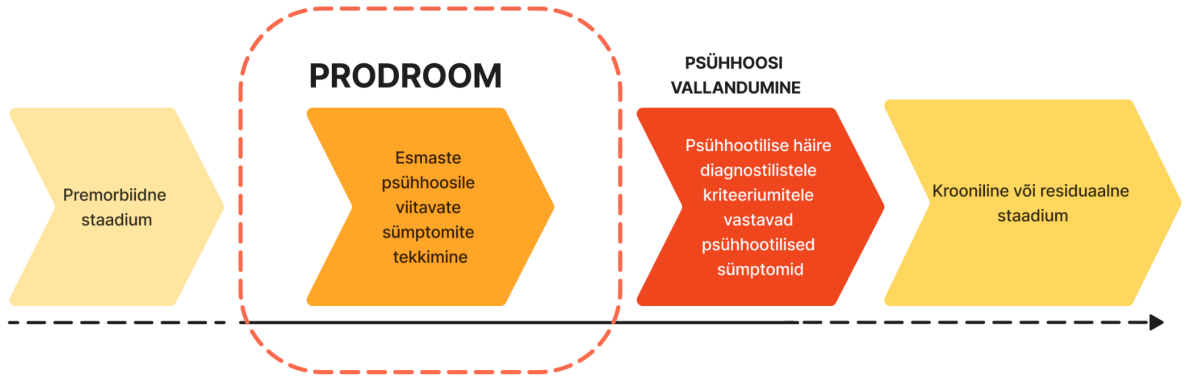
1. Mõisted ja terminid

Diferentsiaaldiagnoos	(ingl *differential diagnosis*) - diagnoosi valik kahe või mitme sümptomitelt sarnase haigusjuhu korral [5].
Epikriis	(ingl *epicrisis*) - ravikokkuvõte haiguse kulust ja ravist [5].
Haiguse trajektoor	(ingl *disease trajectory*) - haiguste (esindatud näiteks RHK-10 diagnoosikoodidena või sümptomitena) kronoloogilised pikaaegsed järgnevused [6].
Hallutsinatsioon	(ingl *hallucination*) - ehk meelepete on tajuelamus ilma välise stiimulita, mille korral isik ei teadvusta, et tegemist on subjektiivse elamusega [1].
Keskmine absoluutviga	(ingl *mean absolute error*) - mõõdab ennustuste ja märgendite vahelist kaugust ehk viga [7].
Kliiniline pilt	(ingl *clinical picture*) - ehk haiguspilt on patsiendi haigusest terviklik kujutlus, mis sisaldab nii sümptomeid, objektiivset leidu (ilma laboratoorsete uuringuteta) kui ka haiguse eellugu ehk anamneesi [5].
Luulumõte	(ingl *delusion*) - korrektsioonile mitte alluv ning tegelikkusega oluliselt vastuolus ekslik veendumus [1].
Objektiivne leid	(ingl *objective finding*) - kliiniliselt märkimisväärne tähelepanek patsiendi läbivaatusel, laboriuuringutel või piltuuringutel [5].
Premorbiidne	(ingl *premorbid*) - haiguseelne staadium, kus isik võib puutuda kokku haigust põhjustavate riskiteguritega [2, 5].
Prodroom	(ingl *prodrome*) - haiguse arenemisele viitav ning sageli väheväljendunud eelnäht [5].
Psühhopatoloogia	(ingl *psychopathology*) - psüühikahäire avaldumisviis [1].
Sõeluuring	(ingl *screening*) - haigusnähtudeta või väheste haigusnähtudega või otsitavasse sihtrühma kuuluvatel inimestel läbiviidav uuring haiguste varajaseks tuvastamiseks [5].
Süvaõpe	(ingl *deep learning*) - masinõppe haru, mis põhineb mitmekihilistes tehisnärvivõrkudel [8].

2. Kirjanduse ülevaade

2.1 Psühhoos ja prodroom

Psühhootilistele häiretele on iseloomulik tegelikkuse tunnetamisvõime ja reaalsustaju märkimisväärne häirumine, mis omab olulist mõju ka isiku käitumisele [1]. Psühhootilised häired kujunevad tavapäraselt läbi premorbiidse (haigusele eelneva), prodromaalse (haiguse eelnäht), sündromaalse (haiguse vallandumine) ning kroonilise (ägedate sümptomite taandumine) staadiumi, misjuures haiguse kulg on ettearvamatu ning sümptomid sama häire raames varieeruvad (joonis 1) [3].



Joonis 1. Psühhootiliste häirete kulg.

Prodroomi on traditsiooniliselt käsitletud kui retrospektiivset ajaperioodi esimeste psühhootilistele häiretele viitavate sümptomite tekkest kuni psühhoosi avaldumiseni [9]. Esmasele ehk esmakordselt diagnoositud psühhoosile eelnev dispersse sümptomitega prodromaalne periood võib näiteks skisofreenia korral esineda 70%-90% patsientidest ning kesta keskmiselt 5 aastat [2, 10, 11]. Erinevate allikate alusel kujunevad ligikaudu 20%-40% tuvastatud psühhoosiriskiga isikutest välja psühhootilised häired ning on leitud, et risk haigestuda suureneb esimese 3 aasta jooksul pärast võimaliku prodroomi tuvastamist sõltuvalt järgmise vastuvõtu toimumise ajast, olles 6 kuu jooksul 18%, esimese aasta jooksul 22%, kahe aasta jooksul 29% ning pärast kolmandat aastat 36% [12, 13].

Prodromaalsete sümptomite hulka võivad kuuluda muutused nii taju- ja mõttekäigu, tahteaktiivsuse, meeleolu kui ka käitumise osas [14]. Mitte-spetsiifiliste prodroomi sümptomitena on varasemalt väljatoodud näiteks meeleolu- ja motivatsioonilangust, keskendumis- ja unehäireid, ärevust, kahtlustavat käitumist, sotsiaalset isoleerumist ning ärrituvust [15], kuid neid sümptomeid võib leida ka paljude teiste vaimse tervise häirete korral. 2017. aasta uuringus on väljatoodud ühed levinumad prodroomi sümptomid, milleks on näiteks kontrolliluul (näiteks veendumus, et isiku mõtteid kontrollib keegi teine [16]), tagakiusamisluul (näiteks veendumus, et isiku suhtes plaanitakse midagi kahjustavat) ja tähendusluul (inimene tõlgendab igapäevaseid nähtuseid tähenduslikult, näiteks peab ajalehes kirjutatut endale suunatuks), tähelepanuhäired ja mõttekäigu pidurdumine, apaatsus,

ülitundlikkus ning depressiivne (negatiivse sisuga luulumõtted) ning suurusluul (näiteks inimene usub, et tal on erilised võimed) [1, 2, 15]. Nagu ka eelmises uuringus leitud sümptomite hulgas on väljatoodud, võib ühtedeks enam kirjeldatud prodromaalseteks indikaatoriteks psühhoosi väljakujunemisel lugeda nõrgalt väljendunud psühhootiliste sümptomite (näiteks hõivatus müstilistest mõtetest või ebatavalise sisuga mõtete esinemine, kahtlustamine, ebanormaalne taju ning kehalised illusioonid; *attenuated psychotic symptoms*) esinemise ning üldisemalt prodromaalsete sümptomite pikemaajalise kestuse [14, 17, 18].

Ligikaudu 20% psühhootilise häirega patsientidest kogeb ainult ühte psühhootilist episoodi ning edasiste episoodide tekkeriski vähendamiseks on vajalik alustada raviga esimese kahe kuu jooksul pärast psühhootiliste sümptomite tekkimist [19]. Varajane sekkumine esmase psühhoosi korral vähendab psühhootiliste episoodide kestust ja riski edasiste episoodide tekkeks ning tõkestab intellektuaalse ja funktsionaalse võimekuse alanemist [3, 19].

2.2 Eraldatavad sümptomid

Üheks psühhoosiriski hindamise vahendiks on ERIraos (*The Early Recognition Inventory*), mis koosneb 15 küsimusega sõeluuringust ning sealse kõrge riskiskoori korral läbiviidavast 50 küsimusega intervjuust [20]. Antud lõputöös on esimene ennustatav prodroomi sümptom („veider käitumine”) valitud Eesti keelde tõlgitud ERIraos´e intervjuu 20. küsimusest: „Kas Teil on huvisid, mida teised inimesed peavad veidraks, näiteks väärtusetute asjade, mida varasemalt olete ära visanud, kogumine või toidutagavarade hankimine? Kas räägite iseendaga avalikes kohtades?” [17, 18]. ERIraos´e sõelküsimustikus on antud sümptom esindatud 8. küsimuse juures, mis on üks unikaalsetest kõrgenenud psühhoosiriski indikaatoritest [17, 18]. Ühtlasi on sotsiaalse kognitiivse võimekuse vähenemine mitmete uuringute kohaselt üks varasemaid indikaatoreid, mis püsib haiguse kulu jooksul stabiilse raskusastmega [14, 21]. Kuna käitumine on üheks sotsiaalse kognitsiooni ehk isiku sotsiaalse informatsiooni töötlemise ning sellele vastamise protsessi osaks [21, 22], sobib „veidra” ehk antud lõputöös tugevalt sotsiaalsetest normidest kõrvalekalduva või isiku vaatenurgast ebatavalisema käitumise esinemine psühhoosiriski ennustava sümptomi hulka. Kõrvalekaldeid käitumises võib pidada üheks unikaalseks psühhoosi ennustavaks teguriks ning neid on psühhoosile eelneval perioodil sagedasti kirjeldatud [23].

Käesolevas lõputöös on metoodika hindamisel valitud koostöös erialaspetsialistiga ka teine sümptomite grupp, milleks on „depersonalisatsioon” ja „derealisatsioon”. Tegemist on isikut ümbritseva või tema enese tunnetamise häirumisega, kus depersonalisatsiooni korral on häiritud organismi sisekeskkonnast (näiteks tunne, et ollakse eraldunud enesest või oma tunnetest, mõtetest, kehast ja kehaosadest) ning derealisatsiooni korral väliskeskkonnast (näiteks ümbritsevad objektid omandavad kas erilise olemuse, neid tajutakse tegelikkusest väiksema- või suuremana, erilistes värvides või tuntakse võõrandumist ümbritsevast) pärinevate üksikaistingute süntees terviklikuks tajuelamuseks [1]. Depersonalisatsioon ja derealisatsioon võivad esineda nii omaette psüühikahäirena kui ka koos teiste sümptomitega näiteks depressiivsete, foobsete, sund- või psühhootiliste häirete korral [24]. 2020. aastal läbiviidud uuringus leiti, et depersonalisatsiooni ning derealisatsiooni sümptomeid esines ligi pooltel (50.5%) psühhoosiriskiga isikutel [25]. ERIraos´e sõelküsimustikus on

depersonalisatsioonile ja derealisatsioonile viitava sümptomaatika kohta küsimus number 11, mis on antud küsimustikus ka üks algavale psühhoosile viitavatest indikaatoritest [17, 18].

Käesoleva lõputöö kolmas ja viimane eraldatav sümptomite grupp treeningandmestiku loomiseks on „paranoiline luulumõte" ja „kahtlustamine". Paranoiline luul on üks luulumõtetest, mille ajel võib isik tajuda nii enese tagakiusamist, jälitamist, kui ka tema suhtes kahjustavat käitumist [1]. Psühhoosiriskiga noorukitel on kirjeldatud paranoilise ja kahtlustava hoiaku ja ebatavaliste mõtete esinemist ning sotsiaalse funktsionaalsuse langust kui olulisi psühhoosi väljakujunemise indikaatoreid [26]. ERIraos´e sõelküsimustikus on paranoilisele luulule viitav küsimus number 15 (ühtlasi ka algava psühhoosi indikaator) ning kahtlustavale hoiakule viitav küsimus number 9 (unikaalne kõrgenenud psühhoosiriski indikaator) [17, 18].

Antud lõputöös on kasutatud prodroomile viitavate meditsiinitekstide leidmiseks RHK-10[1] „Skisofreenia, skisotüüpsed ja luululised häired" alampeatükki kuuluvaid diagnoose: skisotüüpne häire (F21), äge ja mööduv psühhootiline episood (F23), paranoidse (F20.09), hebefreense (F20.19), katatoonse (F20.29), diferentseerimata (F20.39), residuaalse (F20.59), lihtsa (F20.69), muu (F20.89) ja täpsustamata (F20.99) skisofreenia jälgimisperioodiga vähem kui 1 aasta ning skisofreenia järeldepressioon jälgimisperioodiga vähem kui 1 aasta (F20.49). Tegemist on diagnoosidega, mille puhul esinevad psühhootilised sümptomid, kuid need ei vasta skisofreenia diagnostilistele kriteeriumitele (F21, F23) või on tegemist jälgimisperioodil oleva skisofreenia diagnoosiga patsiendiga [24]. Antud diagnoosiga meditsiinidokumentide hulgast on suur tõenäosus leida otsitavaid psühhoosi prodroomi sümptomeid.

2.3 Meditsiinitekstide töötlemine

Digitaalsel kujul talletatud meditsiinidokumentide analüüsimine aitab paremini mõista patsientide kliinilisi trajektoore ning pakkuda võimalust haiguste varajaseks avastamiseks [27]. Terviseandmed on oma olemuselt komplekssed, heterogeensed ning vähe struktureeritud, mis muudab nendest vajaliku informatsiooni kättesaamise tihtipeale keeruliseks ülesandeks [27, 28]. Meditsiinidokumendid koosnevad enamasti rohkem struktureeritud osadest, näiteks diagnoosikoodid, laboratoorsete analüüside vastused ning vähem struktureeritud komponentidest, näiteks vabatekstiline anamnees, mille töötlemine nõuab keerukamaid tööriistu [29]. Suuremahuliste meditsiinitekstide töötlemisel on võimalik rakendada erinevaid keeletehnoloogilisi vahendeid (*natural language processing;* NLP) [30].

Keeletehnoloogia on tervishoiuandmete analüütikas laialdaselt levinud tehnoloogia, mis aitab muuta inimkeele arvutile mõistetavaks [31]. Nende tehnoloogiate alla kuuluvad näiteks süntaktiline töötlus (näiteks tokeniseerimine) ja informatsiooni ekstraheerimine (näiteks tekstide struktureerimise tarbeks) [31]. 2019. aastal avaldatud ülevaateartiklis on toodud välja vabatekstiliste anamneeside ja krooniliste haiguste uurimiseks kasutatud levinumad keeletehnoloogia vahendid aastatel 2007-2018, mille hulgast leiab nii reeglipõhiseid süsteeme (näiteks sõnastike või regulaaravaldiste kasutamine) kui ka masinõppe meetodeid (näiteks tugivektormasinad, Naïve Bayes´i algoritm) ning vähemal määral süvaõppe meetodeid [27].

[1] https://rhk.sm.ee/

Ühtlasi leiti tendents keskenduda rohkem kliinilise pildi alusel tekstide klassifitseerimisele ja riskifaktorite tuvastamisele kui näiteks haiguslugude tekstide struktureerimisele ning nendest kaasuvate haiguste eraldamisele [27]. 2019. aastal läbiviidud ülevaateuuringus leiti, et vabatekstilistest andmetest sümptomite eraldamise lõpp-eesmärk on seotud pigem nende sümptomite põhjal teostatava haiguste klassifitseerimisega kui näiteks nende sümptomite enda olemuse või dokumentatsiooni uurimisega [31].

Meditsiinitekstidest informatsiooni automatiseeritud eraldamise olulisuse elektrooniliste terviseandmete kvaliteedi tõstmiseks nii kliinilise kui ka teadusalase töö tarbeks on välja toonud ka 2018. aasta ülevaateartikli autorid [32]. Nende uuringu kohaselt on teaduskirjanduses üks enamlevinud masinõppe-baasil ekstraheerimismeetod lisaks eelnevalt väljatoodud meetoditele ka logistiline regressioon, mida kasutatakse sageli olemite-vaheliste seoste leidmiseks (*entity and relation detection*). Uuringus täheldati ka, et paljusid ekstraheeritud väljundeid kasutatakse edasiselt masinõppe mudelite tunnustena.

2013. aastal läbiviidud uuringu kohaselt toimis keeletehnoloogiliste vahendite teel vabas vormis terviseandmetest eraldatud tunnustega treenitud hübriidne automatiseeritud otsustuspuu erakorralise meditsiini osakonna kompuutertomograafia vastuste klassifitseerimisel paljulubavalt, ennustades meditsiinipersonaliga ligilähedaselt [33]. Käesolevas magistritöös on sümptomite ennustamiseks ja seeläbi treeningandmestiku pool-automaatseks koostamiseks kasutatud mudelina *scikit-learn*[2] lineaarmudeli teegi logistilist regressiooni, mis on keeletehnoloogias sagedasti kasutatav juhendatud masinõppe klassifitseerimise algoritm [34].

2.3.1 Sõnade vektorid

Levinud meetod tekstiliste andmetega töötamisel on sõnade arvuline esitamine (*word embedding*) vektor-ruumis, kus on säilinud sõnade semantiline tähendus ning sarnaseid sõnu peegeldavad vektorite kaugused üksteisest [35, 36]. Üks levinumaid mudeleid, millega sõnadele vastavaid vektoreid leida, on närvivõrgul põhinev Word2Vec [35]. Word2Vec loob etteantud tekstikorpuse alusel sõnadele vastavate vektoritega sõnastiku, milles olevaid vektoreid saab kasutada tunnustena erinevates masinõppe protsessides [37]. Word2Vec mudeliga luuakse vektoreid kahel erineval viisil: *Continuous Bag-of-Words* (CBOW) ja *Skip-Gram* (SG), millest viimane on praktikas ja kirjanduses enamlevinud [35]. Antud lõputöös on kasutatud sõnade vektorite leidmiseks Eesti keele koondkorpuse[3] peal S*kip-Gram* meetodil treenitud Word2Vec mudelit[4] [38]. Eesti keele koondkorpuses sisalduvad 16 miljonit lauset ja 55 miljonit sõna, mis on moodustunud erinevatest Eesti ilukirjanduslikest, meediaväljaannete ja teadustekstidest [38, 39]. Näiteks on sõnade vektoritel põhineva metoodika kasutamisel psühhiaatria valdkonnas näidanud potentsiaali *Skip-Gram* Word2Vec mudeliga psühhiaatriliste sümptomite klassifitseerimine terviseküsimustike vastuste alusel [36].

[2] https://scikit-learn.org/stable/
[3] https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et
[4] https://entu.keeleressursid.ee/shared/7540/I7G5aC1YgdInohMJjUhi1d5e4jLdhQerZ4ikezz1JEv3B9yuJt9KiPl9lrS87Yz0

2.3.2 Psühhiaatrilistest tekstidest sümptomite eraldamine

Tervisetekstide töötlemise viib keerukuse mõttes järgmisele tasandile psühhiaatria, kus senimaani on üks olulisemaid diagnostilisi vahendeid tekstirohke psühhiaatriline intervjuu [40, 41]. Psühhiaatrilised haiguslood on valdavalt vabatekstilise ülesehitusega ning sisaldavad rohkelt kirjeldavat informatsiooni patsientide vaimse seisundi kohta [41]. Võrreldes teiste erialadega on psüühikahäirete diagnostikas uuringute (näiteks radioloogilised uuringud, laboratoorsed analüüsid) osakaal haiguste täpse tekkemehhanismi väljaselgitamiseks ja diagnostikatarbeks väiksem, mistõttu tugineb psühhiaatriliste haiguste diagnoosimine suuresti just sümptomite kirjeldustele [41, 42]. Ka laialdasema profiiliga meditsiiniliste tekstide analüüsimise muudab veelgi keerulisemaks asjaolu, et nendes sisalduvad peale kirja- ja trükivigade ka kontekstist olenevad lühendid ja akronüümid (näiteks „Pt"- tähistab tavapäraselt patsienti või „KATE"- kopsuarteri trombembooliat), diferentsiaaldiagnostilistel kaalutlustel ülesmärgitud eitused (näiteks "eitab luulumõtteid") ning viited teistele isikutele (näiteks patsiendi lähedased) [41]. Lisaks võib psühhiaatrilisest anamneesist leida väga erineva sõnastusega infot olenevalt sellest, kuidas patsient on oma kaebuseid esitanud ning kuidas on haigusloo autor neid edasi kirjeldanud [43].

Psühhiaatriliste sümptomite eraldamise näiteks võib tuua Inglismaal läbiviidud uuringus kasutatud programmi „TextHunter", mida rakendati meditsiinitekstidest raskete psühhiaatriliste haiguste sümptomite tuvastamiseks loodava mudeli väljatöötamiseks [30]. „TextHunter" otsib märksõnade abil andmebaasist laused, laseb need kasutajal märgendada ning seeläbi moodustada treeningandmestiku, millega treenida ja leida sobiv tugivektormasin enda soovitud sümptomite tuvastamiseks [44]. Uuringus suudeti ekstraheerida vähemalt üks sümptom 87% raske psühhiaatrilise haigusega (näiteks skisofreenia, skisoafektiivne häire) patsientidest ning 60% ilma raske psüühikahäire diagnoosiga patsientidest (näiteks depressiivne episood) [30]. Ka antud lõputöös läbiviidav treeningandmestiku koostamine koosneb sarnastest etappidest.

Juhendatud masinõppemudelid vajavad sageli treenimiseks käsitsi märgendatud andmeid, kuid psühhiaatria valdkonna tekstide mitmekesise sümptomaatika ning keelekasutuse juures on vajalikud veelgi suuremahulisemad märgendatud korpused, mille tootmine on aga rohkesti ressursse nõudev protsess [45]. Seetõttu on ühes 2017. aasta uuringus [45] kasutatud juhendamata masinõppe raamistikke märgendamata tekstide põhjal psühhiaatriliste sümptomite tuvastamiseks, võttes aluseks veebist leitavad sümptomite nimekirjad ning kasutades vektorite semantilist võrdlemist. Samas uuringus leiti ka, et paremad mudelid saavutati üksikute fraaside asemel lauseid kasutades.

2017. aastal teostatud uuringus võrreldi erinevate automaatsete keeletehnoloogia vahendite, põhiliselt tekstikaeve, kasutamist psühhiaatriliste sümptomite raskusastme määramisel, kus reeglipõhiste mudelitega klassifitseerimisel saavutati tulemus keskmise absoluutveaga 80% ning hübriidmeetodiga reeglipõhise mudeli ja närvivõrgu kombinatsioonis keskmise absoluutveaga 72% [41]. Eestis on varasemalt kasutatud keeletehnoloogilist lähenemist tervisetekstidest vajaliku informatsiooni kättesaamiseks inimese papilloomiviiruse leviku uurimiseks [46].

Psühhoosiriski varajaseks tuvastamiseks on 2021. aasta uuringus tugivektorimasinal põhineva mudeliga meditsiinitekstidest eraldatud sümptomid (näiteks pisarate käepärasus, kanepi ja kokaiini tarvitamine, ärritatavus, luulumõtted, agiteeritus ning paranoia) lisatud varasemalt väljatöötatud psühhoosiriski kalkulaatorile, saavutades seeläbi kõrgema prognostilise võimekus [4]. Nende uuringus leiti lisaks, et ekstraheeritud sümptomitest parimad psühhoosi tekkeriski indikaatorid olid paranoia, luulumõtted ning agiteeritus [4].

Erinevaid masinõppe meetodeid saab psühhiaatria valdkonnas edukalt rakendada näiteks nii riskipatsientide identifitseerimiseks, haiguse fenotüübi määramiseks kui ka ravimite ja ravimeetodite (näiteks antipsühhootikumide või antidepressantide, psühhoteraapia) potentsiaalse toime ennustamiseks [47, 48]. Kuigi ennustavate mudelite loomine on väga atraktiivne valdkond, võetakse suhteliselt vähesed loodud tööriistad kliiniliselt kasutusele [48]. Siinkohal tulevad limiteerivate teguritena esile näiteks vähene tulemuste korratavus, kasutatavate andmestike vähene arvukus ja suurus, andmestike haldamisega tekkivad privaatsuse probleemid ning andmetest endast tulenevad tegurid (näiteks madal kvaliteet, palju ebavajalikke elemente) [49]. Kuid keerukus psühhiaatriliste andmete uurimisel säilib ka väga suurte andmestike olemasolul, kuna igal patsiendil võib esineda unikaalne kombinatsioon tema haiguskulgu mõjutavatest teguritest [47].

Käesolev töö loob eeldusi prodroomi sümptomite tuvastamiseks patsiendi terviseandmetest. Täpsemalt luuakse masinõppe vahendite kaasabil treeningandmestikud prodroomi sümptomite tuvastamiseks. Andmestikku saab edaspidi kasutada prodroomi sümptomite tuvastusmudelite treenimiseks.

3. Andmed ja metoodika

3.1 Andmed

Käesolev töö on tehtud RITA MAITT uuringu raames, kasutades Eesti rahvastiku 10% juhuvalimi andmeid ajavahemikust 2012.-2019. aasta. Andmed olid eelnevalt viidud Tartu Ülikooli terviseinformaatika uurimisgrupi poolt PostgreSQL andmebaasi. Samuti olid eraldi andmetabelisse pandud epikriisi dokumentidest pärit laused (tükeldatud EstNLTK tööriistaga[5]). Täpsemalt kasutati käesolevas töös lauseid, mis on saadud esmase prodroomi diagnoosiga epikriiside tekstidest (2780 teksti). Esmase prodroomi diagnoosina kasutati järgmisi rahvusvahelise haiguste klassifikatsiooni (RHK-10) koode:

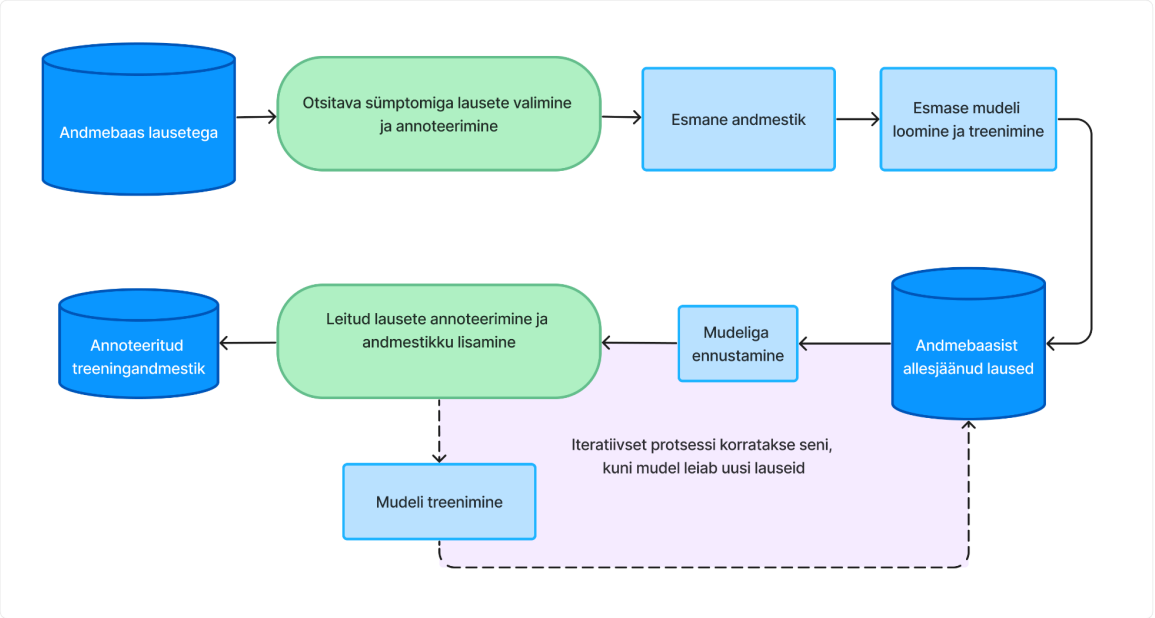
1. F21- Skisotüüpne häire;
2. F23- Äge ja mööduv psühhootiline episood;
3. F20.09- Paranoidne skisofreenia, jälgimisperiood vähem kui 1 aasta;
4. F20.19- Hebefreenne skisofreenia, jälgimisperiood vähem kui 1 aasta;
5. F20.29- Katatoonne skisofreenia, jälgimisperiood vähem kui 1 aasta;
6. F20.39- Diferentseerimata skisofreenia, jälgimisperiood vähem kui 1 aasta;
7. F20.49- Skisofreenia järeldepressioon, jälgimisperiood vähem kui 1 aasta;
8. F20.59- Residuaalne skisofreenia, jälgimisperiood vähem kui 1 aasta;
9. F20.69- Lihtne skisofreenia, jälgimisperiood vähem kui 1 aasta;
10. F20.89- Muu skisofreenia, jälgimisperiood vähem kui 1 aasta;
11. F20.99- Täpsustamata skisofreenia, jälgimisperiood vähem kui 1 aasta.

Esmase prodroomi lauseid on andmestikus kokku 31009.

[5] https://github.com/estnltk/estnltk

3.2 Metoodika

Töö tulemina valminud andmestikud loodi joonisel 2 kujutatud metoodika alusel.



Joonis 2. Treeningandmestike koostamise töökäik.

Esmalt valiti uuritav sümptom („veider käitumine“), mille järel eraldati andmebaasi lausetest regulaaravaldise abil sellele sümptomile potentsiaalselt vastav alamhulk lauseid. Need vaadati käsitsi läbi ja annoteeriti autori poolt (lauses esineb/ei esine seda sümptomit). Seejärel treeniti märgendatud andmete alusel mudel, et leida veel sellele sümptomile vastavaid lauseid. Need annoteeriti omakorda ning protsess kordus iteratiivselt kuni mudel ei ennustanud enam uusi sümptomiga lauseid. Alljärgnevalt on kõiki samme täpsemalt kirjeldatud.

3.2.1 Esmase lausete valiku tegemine

Uuritavaks sümptomiks on valitud „veider käitumine“. Andmebaasi lausete hulgast on regulaaravaldise ja SQL-käskluse abil valitud 119 lauset, kus esineb sõna „veider” või „veidra” (kasutatud regulaaravaldist: '%(veider|veidr)%'), kirjeldamaks otsitavat sümptomit. Pärast duplikaatide eemaldamist jäi alles 68 sobivat lauset. Nendele on lisatud sarnast sõnatüve („veid”) sisaldavad 109 lauset. Kokku on saadud esmane treeningandmestik 177 lausega.

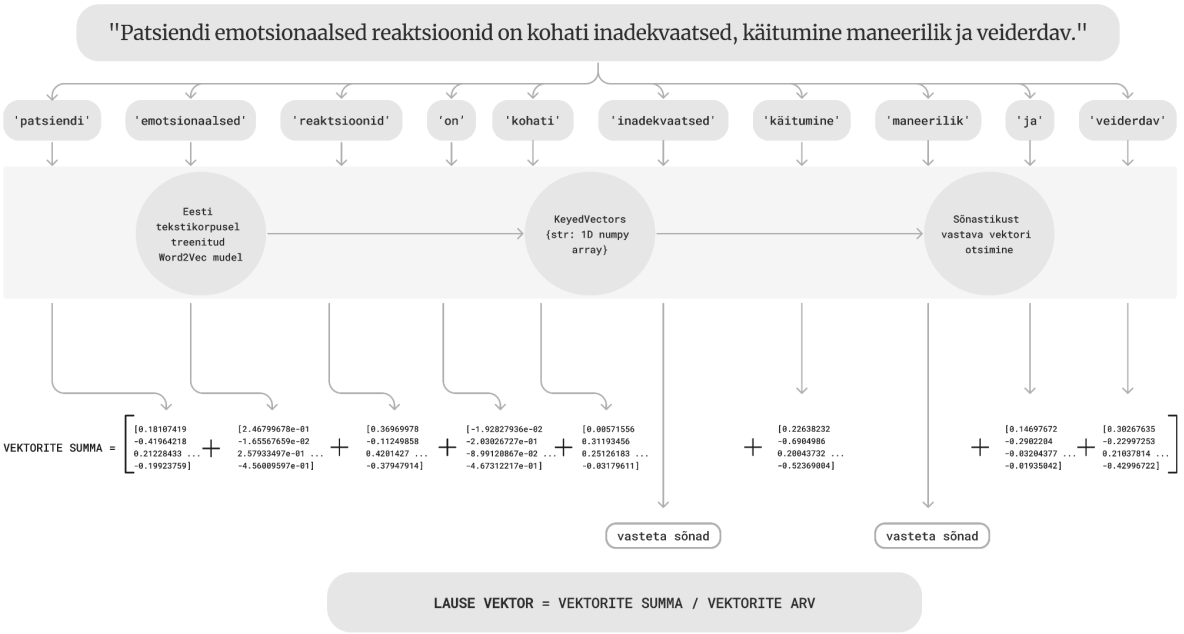
3.2.2 Andmete annoteerimine

Esmase andmestiku annoteerimisel on lähtutud psühhoosi varajase tuvastamise instrumendi ERIraos´e intervjuu sümptomist „veider käitumine”. Esimese andmestiku märgendamisel on arvestatud epikriisi terviktekstiga, millest lause on eraldatud. Lause on märgitud sümptomi suhtes positiivseks juhul, kui patsiendi seisundi kirjeldamisel kasutatakse näiteks „veidra”, „kummalise”, „veiderdava” või muu sarnase tähendusega käitumise esinemist (näiteks *Patsiendi abikaasa sõnul on teatud veidrat käitumist ja juttu esinenud*, *Veider ja*

ebaadekvaatne käitumine) või kirjeldatakse antud käitumist täpsemalt (*Aegajalt räägib enda ette, teeb kätega veidraid liigutusi, liigutab kohati kummaliselt pead*, *Lõikas liikumisandurite juhtme läbi, kuna tajus jälitamist* ja *Patsiendi käitumises on tekkinud viimasel poolaastal veidraid ilminguid- näiteks ebavajalike asjade kogumine taskutesse*) (näidis andmestiku struktuurist on toodud töö lõpus olevas lisas 1, tabel 16). Laused, mis ei sisalda otsitavat sümptomit või kus esinevad näiteks veider mõttekäik või veidrad tajumused ning veidrate veendumuste või arusaamade kirjeldused on siinkohal märgitud negatiivseteks.

3.2.3 Lausete vektorite moodustamine

Selleks, et kasutada andmestiku lauseid klassifitseerimismudelis tunnustena, on kõigepealt leitud nendele lausetele vastavad vektorid. Sõnadele vektorite leidmiseks on kasutatud eesti keele tekstikorpuse peal eeltreenitud Word2Vec mudelit ja Gensim[1] teegist pärit KeyedVectors-i moodulit. Eeltreenitud mudel pärineb Eesti Keeleressursside Keskuse *skip-gram* sõnaesinduste failist[6]. *Skip-gram* mudeliga leitud sõnade vektorid on hästi rakendatavad juhul, kui soovitakse nende alusel ennustada lause või dokumendi konteksti, mis on antud lõputöö mõistes olulisem kui näiteks *continuous bag-of-words* tööpõhimõte, kus otsitakse konteksti alusel konkreetset sõna [50, 51]. KeyedVectors´i teegi abil on leitud lauses esinevatele sõnadele neile vastavad vektorid ning selleks, et edasises masinõppe protsessis oleks võimalik kasutada lausetele vastavaid representatsioone, on võetud lauses esinenud sõnade vektorite keskväärtust (joonis 3).



Joonis 3. Lause vektorite moodustamine.

6 https://entu.keeleressursid.ee/shared/7540/I7G5aC1YgdInohMJjUhi1d5e4jLdhQerZ4ikezz1JEv3B9yuJt9KiPl9lrS87Yz0

Siinkohal tuleb arvestada, et kõiki etteantavaid sõnu ei pruugi Word2Vec mudelis eksisteerida.

3.2.4 Mudeli loomine

Antud lõputöös on kasutatud *scikit-learn* teegi logistilise regressiooni mudelit, et ennustada lause keskmise vektori alusel sümptomi esinemist etteantavas lauses. Andmestik on jagatud *scikit-learn* teegi *train-test-split* abifunktsiooniga treening-, test- ja valideerimisandmestikuks vastavalt osakaaludega 70, 15 ja 15. Mudeli regularisatsiooni määrava hüperparameetri valimisel on lähtutud testandmestiku F-skoori kaalutud keskmiste väärtustest. Regularisatsiooni hüperparameeter on õppimisalgoritmi parameeter, mis vähendab mudeli riski ülesobitamiseks [7]. Mudeli headuse hindamisel on lähtutud eelkõige valideerimisandmestikul leitud F-skoori kaalutud keskmistest väärtustest ning eksimismaatriksist. F-skoor on täpsuse ja saagise harmooniline keskmine ning kaalutud keskmine leiab keskmised väärtused vastavalt klassides esinevate tunnuste arvukusele, mistõttu on antud meetrik sobiv ka tasakaalustamata andmestikel treenitud mudelite hindamiseks [52]. F-skoori väärtused on vahemikus 0 - 1 ning mida suurem on tema väärtus, seda paremad on mudeli täpsus ja saagis. Eksimismaatriksil kuvatakse tõsipositiivsete (mudel ennustas lause positiivseks, mille tegelik märgend oli positiivne), tõsinegatiivsete (mudel ennustas lause negatiivseks, mille tegelik märgend oli negatiivne) ning valepositiivsete (mudel ennustas lause positiivseks, kuigi tegelik märgend oli negatiivne) ja valenegatiivsete (mudel ennustas lause negatiivseks, kuigi tegelik märgend oli positiivne) ennustuste arvud valideerimisandmestikul.

Esmase mudeli väljatöötamise järgselt on alustatud iteratiivse protsessiga algse andmestiku täiendamiseks ning mudeli edasiseks treenimiseks. Iga iteratsiooniga on mudelile antud ette kõik laused andmebaasi esmase prodroomi lausete tabelist, millest on jäetud kõrvale andmestikus juba esinevad laused. Nende lausete pealt on tehtud ennustused ning 300 lauset (või vähem), mille mudel on hinnanud vähemalt 50% tõenäosusega otsitava sümptomiga lauseks, lisatakse eelnevale treeningandmestikule ning annoteeritakse.

Eelnevalt kirjeldatud töökäiku on korratud ka „depersonalisatsioon” ja „derealisatsioon” ning „paranoiline luul” ja „kahtlustamine” sümptomite leidmiseks ning treeningandmestike koostamiseks.

Lõpptulemuseks on saadud 799-lauseline treeningandmestik „veidra käitumise” tuvastamiseks, 643-lauseline treeningandmestik „depersonalisatsiooni” ja „derealisatsiooni” ning 1176-lauseline treeningandmestik „paranoilise luulu” ja „kahtlustamise” tuvastamiseks.

3.3 Eetikakomitee luba

Käesolev töö viidi läbi Eesti bioeetika ja inimuuringute nõukogu loa 05.01.2024 1.1-12/37 alusel.

4. Tulemused

4.1 Loodud andmestiku suurus ja sisu iteratsioonide kaupa

Algses andmestikus oli kokku 177 regulaaravaldise ('%(veider|veidr)%') abil leitud ja seejärel käsitsi märgendatud lauset, millest ligikaudu 20% moodustasid positiivsed laused (s.t lause viitas “veidrale käitumisele”) (tabel 1).

Tabel 1. Lausete arv kokku, otsitava sümptomiga (“veider käitumine”) ja mudeliga leitud sümptomit sisaldavate lausete arv andmestikus.

	Lauseid kokku	**Ei esine “veidrat käitumist”**	**Esineb “veider käitumine”**	**Mudeliga leitud sümptomit sisaldavate lausete arv**
Algne andmestik	177	142	35	-
I iteratsioon	477	422	55	20
II iteratsioon	777	708	69	14
III iteratsioon	798	727	71	2
IV iteratsioon	799	728	71	0
Lõplik andmestik	799	728	71	-

Pärast andmete jaotamist moodustasid treeningandmestiku 123 lauset ning test- ja valideerimisandmestiku mõlemad 27 lauset (tabel 2).

Tabel 2. Treening-, test- ja valideerimisandmestike suurus (“veider käitumine”).

	Kogu andmestik	**Treening-andmestik**	**Test-andmestik**	**Valideerimis-andmestik**
Algne andmestik	177	123	27	27
I iteratsioon	477	333	72	72
II iteratsioon	777	543	117	117
III iteratsioon	798	558	120	120
IV iteratsioon	799	559	120	120

Igal iteratsioonil lisandus uusi lauseid, mis enne mudeli järgmist treenimist annoteeriti. Lõpliku andmestiku moodustasid 799 märgendatud lauset, millest ligikaudu 9% (71 lauset) sisaldasid otsitavat sümptomit („veider käitumine”) ja 728 olid sümptomi suhtes negatiivsed (s.t ei sisaldanud sümptomit „veider käitumine”) (tabel 1).

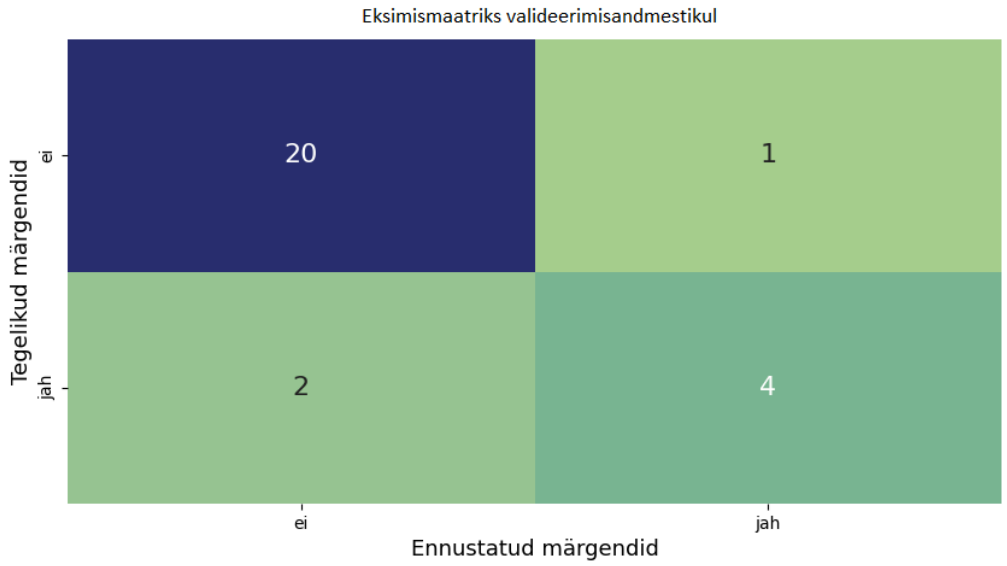
4.2 Loodud andmestiku F-skoorid ja eksimismaatriksid iteratsioonide kaupa

Algsel treeningandmestikul treenitud mudeli kaalutud keskmine F-skoor valideerimisandmestikul oli 0,89 ning F-skoorid ennustatavate klasside („0" ehk negatiivne klass ehk „veidrat käitumist" ei esine ja „1" ehk positiivne klass ehk esineb „veider käitumine") suhtes olid vastavalt 0,93 ja 0,73 (tabel 3).

Tabel 3. F-skoorid valideerimisandmestikel sümptomi „veider käitumine" ennustamisel. „0" – ei esine "veidrat käitumist" ja „1"– esineb "veider käitumine".

Andmestik	Ennustatav klass	F-skoor
Algne andmestik		
	0	0,93
	1	0,73
	Kaalutud keskmine	0,89
I iteratsioon		
	0	0,92
	1	0,44
	Kaalutud keskmine	0,86
II iteratsioon		
	0	0,94
	1	0,13
	Kaalutud keskmine	0,86
III iteratsioon		
	0	0,93
	1	0,00
	Kaalutud keskmine	0,84
IV iteratsioon		
	0	0,92
	1	0,00
	Kaalutud keskmine	0,84

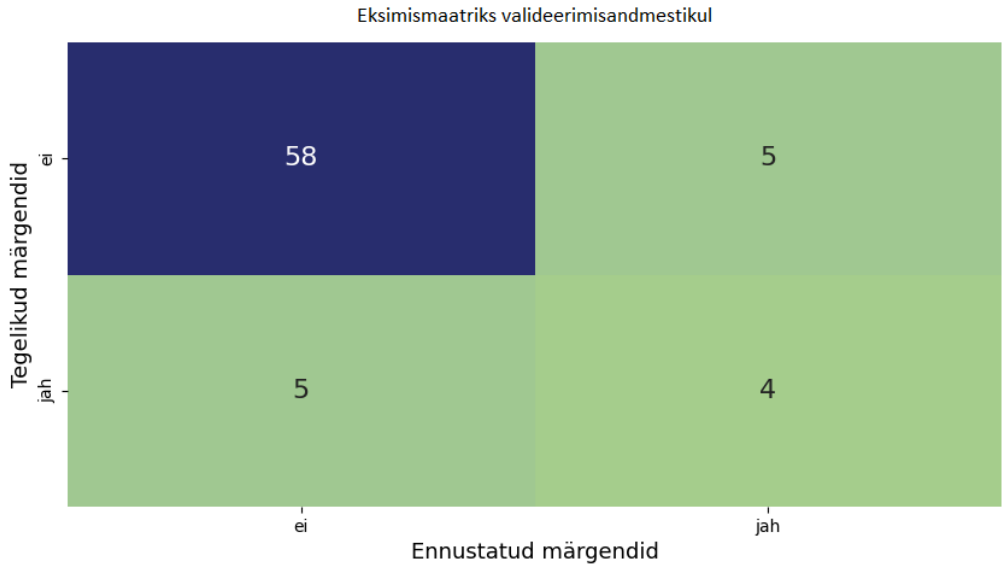
Eksimismaatriksilt on nähtav, et mudel ennustas valideerimisandmestikul õigesti 24 märgendit 27-st , leides 4 tõsipositiivset lauset ning 20 tõsinegatiivset (joonis 4).



Joonis 4. Eksimismaatriks valideerimisandmestikul. Algsel andmestikul treenitud mudel.

Esimesel iteratsioonil rakendati algsel andmestikul treenitud mudelit kõigi annoteerimata lausete peal ning järjestati sümptomi esinemise tõenäosuse alusel. Nendest valiti annoteerimiseks esimesed 300, mille tõenäosused sümptomi esinemisele jäid vahemikku 98%-100%. Valitud 300 lause annoteerimisel selgus, et „veidrat käitumist" sisaldasid autori hinnangul 20 lauset (tabel 1).

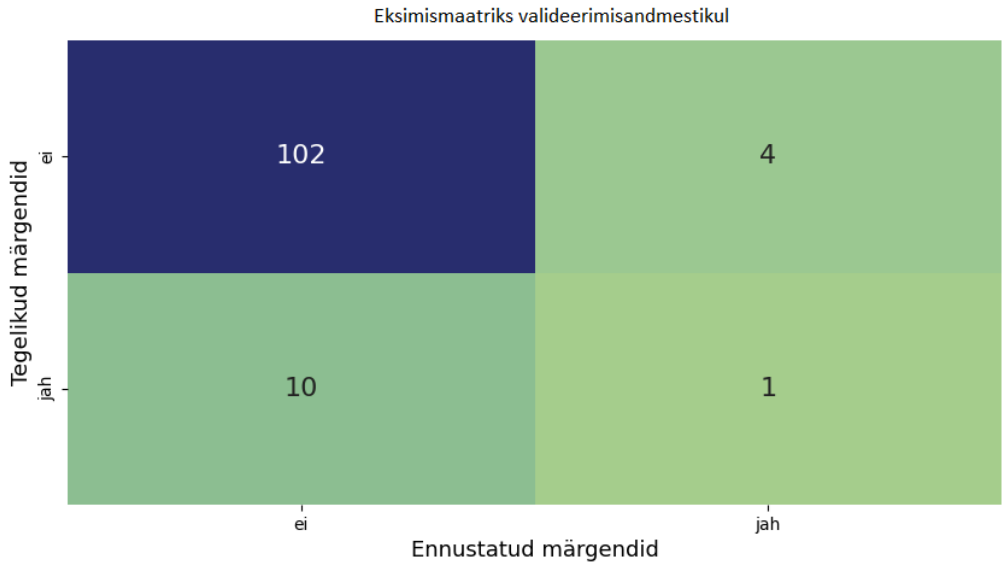
Esimese iteratsiooniga leitud ja seejärel annoteeritud lausete lisamisel algsele andmestikule, saavutati mudeli uuesti treenimisel kaalutud keskmine F-skoor valideerimisandmestikul 0,86 ning F-skoorid ennustatavate klasside (0 ja 1) suhtes vastavalt 0,92 ja 0,44 (tabel 3). Eksimismaatriksilt võib näha, et mudel ennustas valideerimisandmestikul õigesti 62 märgendit 72-st ning leidis 4 tõsipositiivset ning 57 tõsinegatiivset lauset (joonis 5).



Joonis 5. Eksimismaatriks valideerimisandmestikul. Mudel pärast esimest iteratsiooni.

Teisel iteratsioonil korrati eelnevalt kirjeldatud protsessi ning mudeliga leiti järgmised 300 lauset, mille tõenäosused sümptomi esinemisele jäid seekord vahemikku 68%-100%. Nende lausete käsitsi märgendamisel selgus, et 14 lauset viitasid „veidra käitumise" olemasolule (tabel 1).

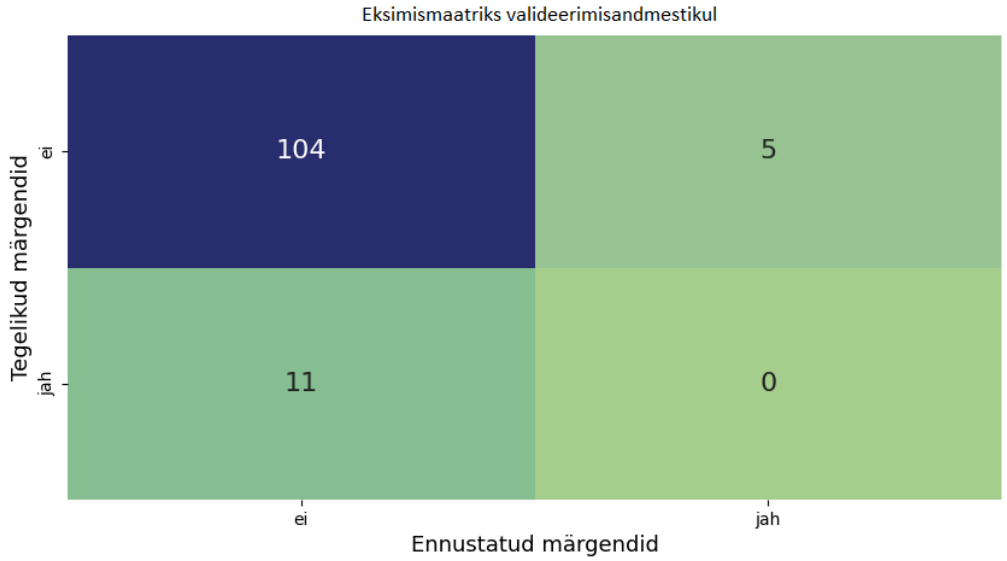
Kaalutud keskmine F-skoor uute lausete lisandumisel ja seejärel mudeli uuesti treenimisel oli valideerimisandmestikul 0,86. Siinkohal on täheldatav märgatav langus positiivsete märgendite tuvastamise puhul, sest F-skoori väärtus sümptomiga lause ennustamisel on 0,13 ning sümptomit mitte sisaldavate lausete korral 0,94 (tabel 3). Mudel ennustas õigesti 103 märgendit 117-st, millest 1 oli tõsipositiivne ning 102 tõsinegatiivsed (joonis 6).



Joonis 6. Eksimismaatriks valideerimisandmestikul. Mudel pärast teist iteratsiooni.

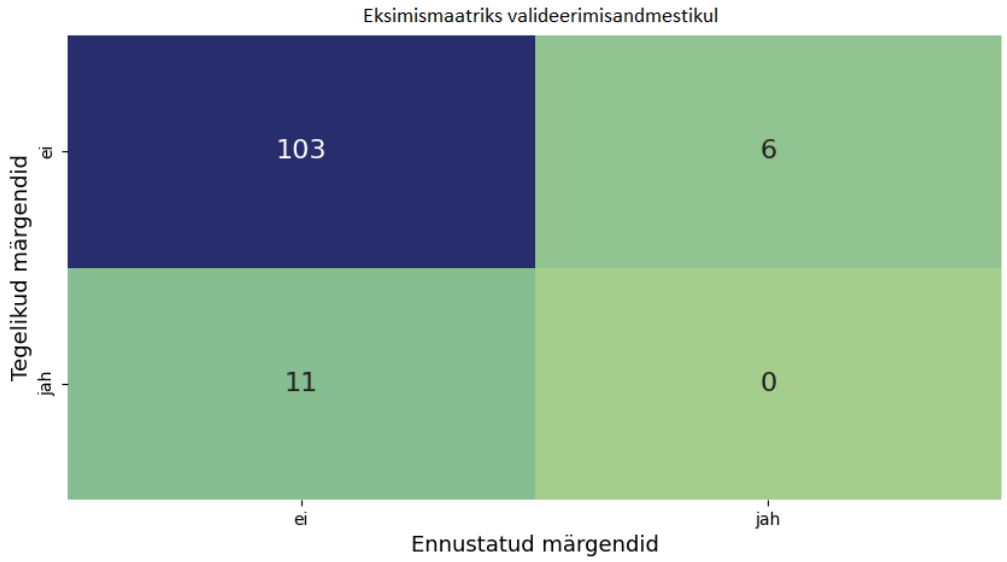
Kolmanda iteratsiooniga leidis mudel vaid 21 lauset, mille tõenäosused sümptomi esinemisele jäid vahemikku 50%-83%. Nende lausete käsitsi märgendamisel selgus, et ainult 2 sisaldasid endas „veidrat käitumist" (tabel 1).

Mudeli taaskordsel treenimisel ning hindamisel selgus, et valideerimisandmestiku kaalutud keskmine F-skoor oli 0,84, kuid sümptomi suhtes positiivsete lausete tuvastamisel hoopiski väärtusega 0 (tabel 3). Mudel ennustas õigesti 104 märgendit 120-st ning kõigil juhtudel oli tegemist tõsinegatiivsete ehk sümptomita lausete korrektse tuvastamisega.



Joonis 7. Eksimismaatriks valideerimisandmestikul. Mudel pärast kolmandat iteratsiooni.

Neljanda iteratsiooniga suutis mudel ennustada vaid ühe sobiva tõenäosusega lause (tõenäosus 50%), mis aga manuaalsel ülevaatusel ei vastanud otsitava sümptomi kirjeldusele ning annoteeriti negatiivseks. Pärast mudeli treenimist ning rakendamist valideerimisandmestikul saavutati kaalutud keskmine F-skoor väärtusega 0,84 ning vastavalt negatiivse ja positiivse klassi ennustamisele 0,92 ning 0 (tabel 1). Eksimismaatriks on võrreldes eelmise iteratsiooniga (joonis 7) oluliste muutusteta (joonis 8).



Joonis 8. Eksimismaatriks valideerimisandmestikul. Mudel pärast neljandat iteratsiooni.

Pärast neljandat iteratsiooni ei ennustanud mudel enam ühtegi vähemalt 50% tõenäosusega positiivsesse klassi kuuluvat lauset. Siinkohal lõpetati iteratiivne protsess. Kokku leiti andmebaasist iteratiivse protsessiga 36 lauset sümptomiga „veider käitumine".

4.3 Kokkuvõte ja lõplik andmestik

Lõplik loodud andmestik sümptomi „veider käitumine" kohta sisaldab 799 töö autori poolt käsitsi märgendatud lauset (tabel 1). Valminud andmestiku näidise leiab töö lõpus olevast lisast (lisa 1, tabel 16).

Töö käigus oli täheldatav, et kasutatud mudel hakkab protsessi käigus võrdlemisi varakult ära tundma tõsinegatiivseid lauseid ning andmete lisandudes tõsipositiivsete lausete ennustamise võimekus alaneb märkimisväärselt, olles viimase kahe iteratsiooni korral F-skooriga 0 (tabel 3). Siinkohal võivad olla põhjuseks nii valitud mudeli kui ka teksti representatsiooni lihtsus, mis toob endaga paratamatult kaasa metoodika piiratud võimekuse. Tõsipositiivsete ennustuste hulga vähesus võib olla tingitud asjaolust, et kasutatud Word2Vec mudelist puuduvad mitmed psühhiaatria valdkonnale spetsiifilised sõnad. Töö käigus tuvastatud puuduvaid sõnu on kirjeldatud täpsemalt alampeatükis „Tekstikorpusest puuduvad sõnad". Siinkohal võib järeldada, et väljatoodud probleemide tõttu ei pruugitud andmebaasist leida ülesse kõiki sümptomile viitavaid lauseid, sest mudel klassifitseeris need valenegatiivseks. Samas võib täheldada, et mudel leidis 36 lauset, kus otsitav sümptom oli esindatud, mis annab aluse rakendatud metoodikat katsetada ka järgmiste sümptomite leidmiseks.

5. Meetodi valideerimine uutel sümptomitel

Eelmises peatükis kirjeldatud tulemuste alusel otsustati sama metoodikat rakendada ka teiste sümptomite puhul. Valiti kaks võimalikule psühhoosi prodroomile viitavat sümptomit, või täpsemalt siinkohal isegi sümptomite gruppi: „depersonalisatsioon" ja „derealisatsioon" ning „paranoiline luulumõte" ja „kahtlustamine". Nii „depersonalisatsiooni", „derealisatsiooni" kui ka „paranoiliste luulumõtete" korral on ERIraos´e küsimustiku järgi tegemist algavale psühhoosile viitavate indikaatoritega ning „kahtlustamine" (või „kahtlustav/paranoiline hoiak"), nagu ka „veider käitumine", on unikaalne psühhoosiriskile viitav indikaator. Sel korral moodustavad eraldatava sümptomi kaks omavahel seotud sümptomit ehk seeläbi laiendatakse sümptomit sisaldavate lausete hulka ning seeläbi ka mudeli tööpõldu.

Esimesed sümptomid, „depersonalisatsioon" ning „derealisatsioon", esinevad sageli koos ning väljendavad endas isikul esinevaid tajumishäireid. Kuna regulaaravaldisega leitud lausete hulgas oli neid kahte sümptomit eraldiseisvalt kirjeldavaid lauseid liiga vähe ning mitmel juhul olid need lausetes kirjeldatud koos, siis otsustati, et sümptom esineb siinkohal nii „depersonalisatsiooni" kui ka „derealisatsiooni" esinemise korral. Teised kaks sümptomit, „paranoiline luulumõte" ning „kahtlustamine" moodustavad samuti omavahel terviku, kus kahtlustav käitumine võib viidata nii paranoiliste luulumõtete esinemisele või tekkimisele. Meetodi valideerimiseks valitud sümptomite tulemusi on kokkuvõtvalt kirjeldatud järgnevates peatükkides.

5.1 Sümptomid „depersonalisatsioon" ja „derealisatsioon"

5.1.1 Loodud andmestiku suurus ja sisu iteratsioonide kaupa

Algses andmestikus oli kokku regulaaravaldise (kasutatud regulaaravaldist: '%(deperso|dereal)%') abil eraldatud ja käsitsi märgendatud 40 lauset, millest täpselt pooltel esines depersonalisatsiooni või derealisatsiooni kirjeldav sümptomaatika (tabel 4).

Tabel 4. Lausete arv kokku, otsitava sümptomiga ("depersonalisatsioon" ja "derealisatsioon") ja mudeliga leitud lausete arv andmestikus.

	Lauseid kokku	**Ei esine "depersona-lisatsiooni" ja/või "derealisatsiooni"**	**Esineb "depersona-lisatsioon" ja/või "derealisatsioon"**	**Mudeliga leitud sümptomit sisaldavate lausete arv**
Algne andmestik	40	20	20	-
I iteratsioon	340	320	20	0
II iteratsioon	640	616	24	4
III iteratsioon	643	619	24	0
Lõplik andmestik	643	619	24	-

Esmase mudeli treenimiseks eraldati 28 lauset ning test- ja valideerimisandmestikuks mõlemale 6 lauset (tabel 5).

Tabel 5. Treening-, test- ja valideerimisandmestike suurus ("derealisatsioon" ja "depersonalisatsioon").

	Kogu andmestik	**Treening-andmestik**	**Test-andmestik**	**Valideerimis-andmestik**
Algne andmestik	40	20	20	-
I iteratsioon	340	320	20	0
II iteratsioon	640	616	24	4
III iteratsioon	643	619	24	0

Eelnevalt kirjeldatud metoodika alusel lisandus iga iteratsiooniga uusi lauseid, mis annoteeriti uue mudeli treenimiseks. Lõpliku andmestiku "depersonalisatsiooni" ja "derealisatsiooni" sümptomaatika kohta moodustavad 643 autori poolt käsitsi märgendatud lauset, millest ligikaudu 4% (24) lauset on sümptomi suhtes positiivsed ning 619 negatiivsed (tabel 4).

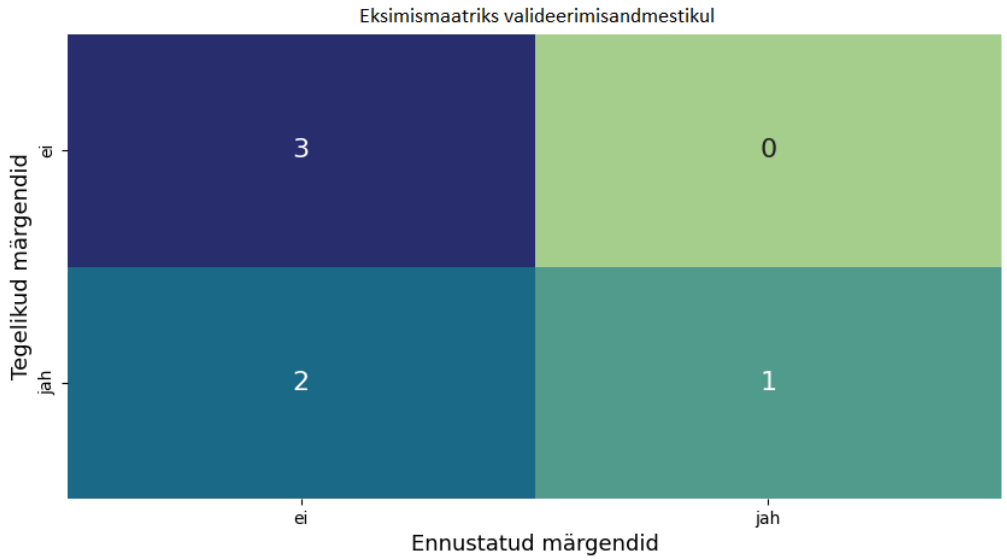
5.1.2 Loodud andmestiku F-skoorid ja eksimismaatriksid iteratsioonide kaupa

Algsel treeningandmestikul treenitud mudeli kaalutud keskmine F-skoor valideerimisandmestikul oli 0,62 ning F-skoorid ennustatavate klasside (0- „ei esinenud depersonalisatsiooni ega derealisatsiooni sümptomeid" ja 1- „esinesid depersonalisatsiooni ja/või derealisatsiooni sümptomid") suhtes olid vastavalt 0,75 ja 0,50 (tabel 6).

Tabel 6. F-skoorid valideerimisandmestikel sümptomite „depersonalisatsioon" ja „derealisatsioon" ennustamisel. „0" – sümptomit lauses ei esine ja „1"– sümptom esineb lauses.

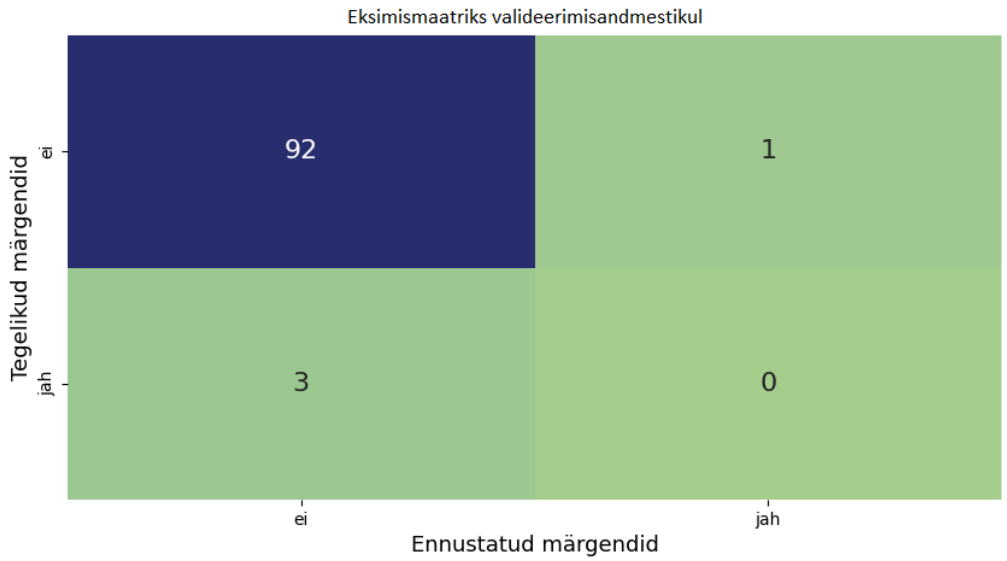
Andmestik	Ennustatav klass	F-skoor
Algne andmestik		
	0	0,75
	1	0,50
	Kaalutud keskmine	0,62
I iteratsioon		
	0	0,95
	1	0,00
	Kaalutud keskmine	0,89
II iteratsioon		
	0	0,96
	1	0,00
	Kaalutud keskmine	0,92
III iteratsioon		
	0	0,98
	1	0,00
	Kaalutud keskmine	0,95

Eksimismaatriksi alusel ennustas mudel õigesti 4 märgendit 6-st, millest ainult 1 oli tõsipositiivne ning 3 olid tõsinegatiivsed (joonis 9).



Joonis 9. Eksimismaatriks valideerimisandmestikul. Mudel algsel andmestikul.

Esmase treenitud mudeli rakendamisel andmebaasi järelejäänud lausetele, valiti 300 kõige suurema tõenäosusega lauset, mille tõenäosused jäid vahemikku 97%-100%. Nendest lausetest ei osutunud käsitsi annoteerimisel mitte ükski otsitava sümptomi suhtes positiivseks. Antud sümptomite korral langes mudeli F-skoor juba alates esimesest iteratsioonist väärtusele 0 ning ainult teise iteratsiooni käigus suutis mudel tuvastada 4 uut „depersonalisatsiooni" ja „derealisatsiooni" sisaldavat lauset, sealjuures iteratsiooni positiivse klassi tõenäosused jäid vahemikku 68%-99%. Eksimismaatriksite alusel võib öelda, et peale algsel andmestikul treenitud mudeli ei tuvastanud järgnevad mudelid ühtegi tõsipositiivset lauset, kuid tuvastasid suurel hulgal tõsinegatiivseid lauseid (joonis 10).



Joonis 10. Eksimismaatriks valideerimisandmestikul. Mudel pärast viimast iteratsiooni.

Pärast kolmandat iteratsiooni ei leidnud mudel enam ühtegi vähemalt 50% tõenäosusega „depersonalisatsiooni" ja/või „derealisatsiooni" sisaldavat lauset ning iteratiivne protsess lõpetati.

5.2 Sümptomid „paranoiline luulumõte" ja „kahtlustamine"

5.2.1 Loodud andmestiku suurus ja sisu iteratsioonide kaupa

Algse annoteeritud andmestiku moodustasid regulaaravaldise (kasutatud regulaaravaldist: '%(paran|kahtl)%') abil leitud ning käsitsi märgendatud 263 lauset, millest 105 (~40%) moodustasid laused, kus esines paranoilisele luulule või hoiakule viitavaid sümptomeid (tabel 7).

Tabel 7. Lausete arv kokku, otsitava sümptomiga („paranoiline luul" ja „kahtlustamine") lausete arv andmestikus, mudeliga leitud sümptomit sisaldavate lausete arv.

	Lauseid kokku	**Ei esine "paranoilisi luulumõtteid" ja/või "kahtlustamist"**	**Esineb "paranoilisi luulumõtteid" ja/või "kahtlustamist"**	**Mudeliga leitud sümptomit sisaldavate lausete arv**
Algne andmestik	263	158	105	-
I iteratsioon	563	424	139	34
II iteratsioon	863	717	146	7
III iteratsioon	1163	958	205	59
IV iteratsioon	1175	970	205	0
Lõplik andmestik	1176	971	205	-

Pärast neljandat iteratsiooni moodustas lõpliku andmestiku 1176 autori poolt käsitsi märgendatud lauset, millest ligikaudu 18% (205) moodustasid sümptomi suhtes positiivsed laused (tabel 8).

Tabel 8. Treening-, test- ja valideerimisandmestike suurus ("paranoiline luul" ja "kahtlustamine").

	Kogu andmestik	**Treening-andmestik**	**Test-andmestik**	**Valideerimis-andmestik**
Algne andmestik	263	184	40	39
I iteratsioon	563	394	85	84
II iteratsioon	863	604	120	129
III iteratsioon	1163	814	175	174
IV iteratsioon	1175	822	177	176

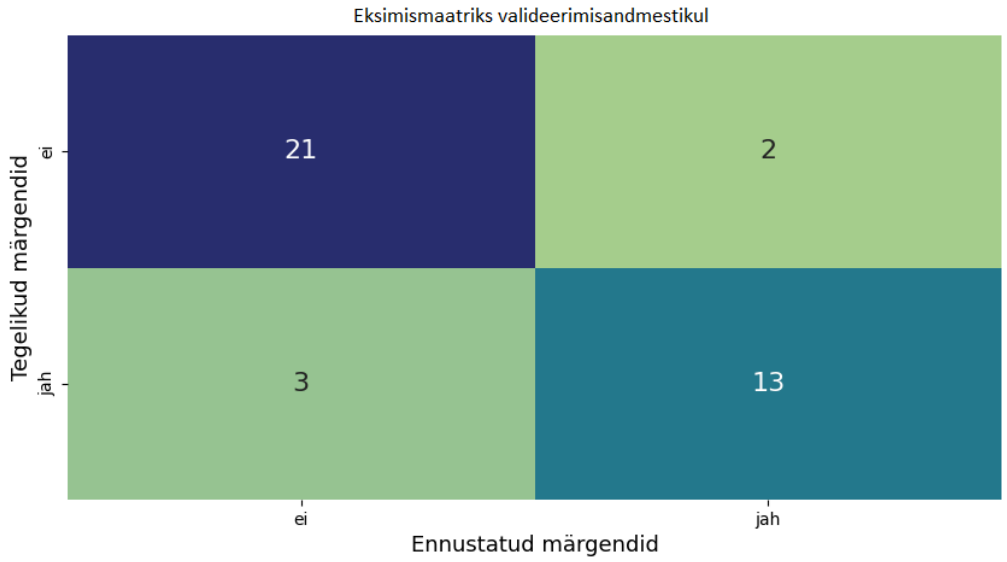
5.2.2 Loodud andmestiku F-skoorid ja eksimismaatriksid iteratsioonide kaupa

Algsel treeningandmestikul treenitud mudeli kaalutud keskmine F-skoor valideerimisandmestikul oli 0,87 ning F-skoorid ennustatavate klasside (negatiivne klass ehk „0" – ei esinenud paranoiliseid luulumõtteid ega paranoilist hoiakut ja positiivne klass ehk „1" – esinesid paranoilised luulumõtted ja/või paranoiline hoiak) suhtes olid vastavalt 0,89 ja 0,84 (tabel 9).

Tabel 9. F-skoorid valideerimisandmestikel sümptomite „paranoiline luulumõte" ja „kahtlustamine" ennustamisel. „0" – sümptomit lauses ei esine ja „1"– sümptom esineb lauses.

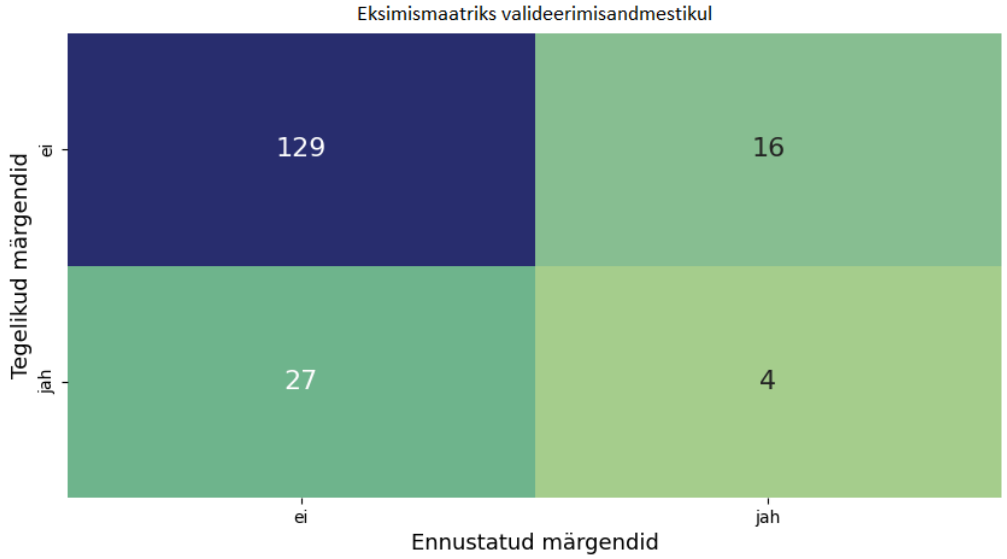
Andmestik	Ennustatav klass	F-skoor
Algne andmestik		
	0	0,89
	1	0,84
	Kaalutud keskmine	0,87
I iteratsioon		
	0	0,85
	1	0,54
	Kaalutud keskmine	0,77
II iteratsioon		
	0	0,85
	1	0,16
	Kaalutud keskmine	0,74
III iteratsioon		
	0	0,87
	1	0,32
	Kaalutud keskmine	0,77
IV iteratsioon		
	0	0,86
	1	0,16
	Kaalutud keskmine	0,73

Eksimismaatriksi alusel saab öelda, et algsel andmestikul treenitud mudel ennustas valideerimisandmestikul õigesti 34 märgendit 39-st, millest 13 moodustasid tõsipositiivsed ning 21 tõsinegatiivsed (joonis 11).



Joonis 11. Eksimismaatriks valideerimisandmestikul. Esmane mudel.

Esmase mudeliga valitud 300 lauset (kõigi lausete positiivsesse klassi kuulumise tõenäosused olid ligikaudu 99%) sisaldasid autori hinnangul 34 sümptomi suhtes positiivset lauset. Teise ja kolmanda iteratsiooniga leidis mudel 7 ja 59 sümptomiga lauset. Viimase ehk neljanda iteratsiooni raames treenitud mudeli kaalutud keskmine F-skoor oli 0,73 ning ennustatavale negatiivsele klassile 0,86 ning positiivsele 0,16 (tabel 9). Eksimismaatriksi kohaselt ennustas mudel pärast neljandat iteratsiooni valideerimisandmestikul õigesti 133 märgendit 176-st, millest 4 olid tõsipositiivsed ning 129 tõsinegatiivsed (joonis 12).



Joonis 12. Eksimismaatriks valideerimisandmestikul. Mudel pärast neljandat iteratsiooni.

Viimasena treenitud mudel ennustas andmebaasi uutest lausetest ainult ühe lause 56% tõenäosusega sümptomaalseks, kuid käsitsi annoteerimisel märgendati lause siiski negatiivseks. Siinkohal iteratiivne protsess lõpetati. Kokku leiti andmebaasist juurde 100 otsitavate sümptomitega lauset.

5.3 Kokkuvõte meetodi valideerimisest ja lõplikud andmestikud

Meetodi valideerimisel õnnestus mudeliga tuvastada 4 „depersonalisatsiooni" ja „derealisatsiooni" sümptomit sisaldavat lauset ning 100 „paranoiliste luulumõtete" ja/või „kahtlustamise" sümptomit sisaldavat lauset. Viimane sümptomite grupp („paranoilised luulumõtted/kahtlustamine") oli silmnähtavalt kolmest valitud sümptomist („veider käitumine", „depersonalisatsioon/derealisatsioon", „paranoiline luulumõte/kahtlustamine") kõigi lausete hulgas enim kirjeldatud, mis võis tingida antud sümptomiga loodud andmestike (nii algandmestiku kui ka lõpliku andmestiku) suurema mahu ning nendega treenitud mudelite paremad näitajad nii F-skoori kui ka tõeste ennustuste korral. Üldjoontes võib täheldada, et kitsaskohad meetodi rakendamisel on sarnased esimese valitud sümptomi („veider käitumine") korral ilmnenud keerukustega – puuduvad sõnad Word2Vec mudelis ning valitud teksti representatsiooni ja logistilise regressiooni mudeli primitiivsus. Kokkuvõtvalt võib öelda, et töös kasutatud metoodika kiirendab märgendatud andmestike loomise protsessi ning aitab leida ülesse lauseid, mida näiteks ainult regulaaravaldist kasutades tuvastada ei pruugi.

Loodud andmestike näidised võib leida töö lõpus olevast lisast (lisa 1, tabelid 17 ja 18).

5.4 Word2Vec mudelist puuduvad sõnad

Käesolevas lõputöös on ennustamiseks loodud mudeli tunnustena kasutatud eeltreenitud Word2Vec mudelit, milles ei pruugi leiduda kõikide eestikeelsete sõnade representatsioone ehk vektoreid. Töös kasutatud lause keskmise vektori korral tähendab see aga seda, et sõnad, millele vastavat vektorit mudelist ei leita, jäävad välja ka lause keskmise vektori loomisest ning mõjutavad seeläbi mudeli võimet tuvastada ennustamiseks vajalikku konteksti.

Sõnu, millele Word2Vec mudel vastavat vektorit ei leidnud, oli kõigi kolme sümptomi eraldamise protsessi peale kokku 8368. Tabelis 10 on välja toodud 10 kõige enam protsessi jooksul esinenud sõna, millele mudelis vastavat vektorit ei leidunud.

Tabel 10. Eeltreenitud Word2Vec mudelist puudunud 10 tööprotsessi vältel enim kohatud sõna ja nende esinemissagedused.

Mudelist puuduv sõna	Esinemissagedus kogu tööprotsessi jooksul
5mg	2351
natiivis	1937
hinnakoodid	1830
eluanamnees	1607
haiguskriitika	1563
tahteaktiivsus	1517
diazepami	1385
sedasta	1289
olanzapini	1213
rö	1132

Nende hulgast võib lisaks tabelis väljatoodule leida ravimeid (näiteks aripiprasool, olansapiin) ning nende ravimite erinevaid ladinakeelseid kirjapilte või käändeid (näiteks *quetiapini* ja kvetiapiini, *olanzapini* ja olansapiin), ravimite annustamist tähistavaid kombinatsioone („5mg", „10mg"), psühhiaatrilisi mõisteid (näiteks haiguskriitika, tahteaktiivsus, kuulmismeelepetted) ning mitmeid meditsiinidokumentidest leitavaid sõnu ja sümboleid analüüside ning uuringute vastustest (näiteks „egfr", „lymph", „baso", „patoloogiata").

5.5 Andmestiku loomise Jupyter notebook

Andmestiku loomiseks kasutatud Jupyter´i koodivihik on toodud töö lõpus olevas lisas 2.

6. Arutelu

6.1 Lausete läbivaatus ja nende eripära

Vaadeldavad meditsiinitekstid ning nendest eraldatud laused sisaldasid väga rohket ja mitmekülgset informatsiooni, millega kaasnesid kohati ebavajalik detailsus (näiteks kirjeldus „Piparmündikeefiri joonud"), trüki- ja kirjavead (näiteks „Tal võib olla depresioob.") ning täheldatav oli ka teatav hinnangulise tooni kasutamine patsiendi käitumise osas (näiteks „sinatamise" väljatoomine, ütluste „jaburaks" või „totraks" nimetamine). „Veidra käitumise" sümptomiga lausete otsimisel oli nii mõnelgi korral keeruline mõista, kas dokumendi autor pidas patsiendi kõnepruuki (näiteks eelnevalt kirjeldatud „sinatamist") ebaadekvaatseks käitumiseks või oli tegemist situatsiooni kirjeldusega (näiteks lause „psüühika: Sinatab familiaarselt.").

Kuna haiguslugude tekstides esines tihti nii kuupäevalist infot kui ka erinevaid punktiga tähistatud lühendeid, siis tekstide lauseteks tükeldamisel tekkis mitmeid lühikesi, vähestest kirjamärkidest koosnevaid lauseid, mille kasutamine oli vähe informatiivne (näiteks „pt", „tervis.", „muu?").

Lausete hulgas esines ka palju duplikaate, mis võisid sattuda sinna ühe ja sama informatsiooni mitmes haigusloos esinemise tõttu (näiteks ühest dokumendist teise kopeeritud uuringute tulemused ja varasemad anamneesid). Mitmetes haiguslugudes oli täheldatav ka läbiviidavate uuringute (näiteks positiivsete ja negatiivsete sümptomite skaala (PANSS) ja ERIraos´e küsimustik) esinemine koos küsimustiku raamistiku ja vaheskooridega, mis muutis sealsest tekstist sümptomite eristamise veelgi keerulisemaks (tabel 11).

Tabel 11. Keerulise ülesehitusega laused andmestikus

Lause andmestikus
„PANSS-skaala (Positiivse ja negatiivse sündroomi hindamise skaala) Tunnuste avaldumise astmed 1 = ei avaldu 2 = minimaalselt 3 = kergelt 4 = mõõdukalt 5 = mõõdukalt raskel kujul 6 = raskel kujul 7 = ekstreemselt Positiivne alaskaala __7__ P1 Luulumõtted __2__ P2 Mõtlemise kontseptuaalne desorganiseeritus __1__ P3 Hallutsinatoorne käitumine __1__ P4 Erutuvus __1__ P5 Suurusmõtted __3__ P6 Kahtlused/ tagakiusamine __1__ P7 Vaenulikkus Negatiivne alaskaala __5__ N1 Tuimenenud afektid __3__ N2 Emotsionaalne eemaldumine __2__ N3 Interpersonaalse kontakti halvenemine __6__ N4 Sotsiaalne eemaldumine, isoleeritus __4__ N5 Raskused abstraktses mõtlemises __4__ N6 Vestluse spontaansus ja sujuvus __4__ N7 Stereotüüpne mõtlemine Psühhopatoloogia üldskaala __1__ G1 Somaatiline hõivatus __2__ G2 Ärevus __2__ …"
„Patsiendil on esinenud viimase aasta jooksul järgnevaid lühiaegseid, iselimiteerunud psühhootilisi elamusi [BLIPS] ehk (Brief Limited Intermittent Psychotic Symptoms): Katatoonsed sümptomid (hüpokineesia või akineesia) [BLIPS] (– tardumise, seisaku tunne) Kuulmishallutsinatsioonid (müra, hääled) [BLIPS] Psühhootilisel tasemel esinevad patsiendil järgnevad haigustunnused: …"

Siinkohal on oluline psühhiaatriliste tekstidega andmestike mitmetahuline eeltöötlus, et minimeerida ebavajaliku informatsiooni mõjutamist järelduste tegemisel. Ideaalis võiks psühhiaatria valdkonnas kasutatav infosüsteem pakkuda spetsialistidele tänapäevasemaid, mugavamaid ja erialale spetsiifilisemaid lahendusi psühhiaatrilise epikriisi koostamiseks, mis lihtsustaks oma ülesehituselt ka edasist teadustöö tarbeks andmete kättesaamist ning analüüsimist.

6.2 Mudel ja ennustatud laused

Esimese otsitava sümptomi („veider käitumine”) korral leidis mudel mitmeid võimalikule „veidrale” käitumisele viitavaid lauseid, mis on toodud välja allolevas tabelis (tabel 12).

Tabel 12. Näiteid mudeli poolt ennustatud lausete kohta iteratsioonide kaupa, nende tõenäosused ning käsitsi märgendamisel määratud klass, sümptomiks „veider käitumine”.

Iteratsioon	Lause, milles mudel ennustas esinevat sümptomi „veider käitumine”	Tõenäosus sümptomi esinemisele	Hiljem määratud märgend
I	„liiklusõnnetus S06.”	100%	0
I	„Lindistas telefonikõnesid.”	99%	1
I	„Käitumine kohmetu, omapärane.”	99%	1
I	„Viimastel kuudel käitub imelikult, kahtlustab.”	99%	1
II	„nagu räägiks kellestki teisest osavõtmatult.”	90%	1
II	„Vestlusel omapäraselt "voolavalt" žestikuleeriv.”	87%	1
II	„Tihti naerab ebaadekvaatselt.”	82%	1
II	„Silma torkab situatsioonile sobimatu keelekasutus.”	77%	1
III	„Räägib erinevatest vandenõuteooriatest - illuminaatidest, reptiilidest, vabamüürlastest.”	74%	1
III	„Patsiendi tegutsemine ajendatud psühhootilistest elamustest, mõttekäiguhäiretest.”	60%	1
IV	„Patsiendi kehakeel kahtlustav, ärev.”	51%	0

Ennustatud lausete hulgast paistsid silma mitmed tegevusi kirjeldavad laused sõnadega „elab", „töötab", „eitab", „räägib", „naerab" ja „käitumine". Näiteks ennustas mudel eitust sisaldavaid lauseid (*Kuulmismeelepetteid eitab*, *Tundlikkushäireid eitab*, *Tajuelamusi eitab, Kuulmishallutsinatsioone eitab*, *Hirmutundeid eitab*, *Mõttelevielamusi eitab*, *Suhteprobleeme eitab*, *Elutüdimusmõtteid eitab*) 99% tõenäosusega „veidra käitumise" suhtes positiivseks, mille puhul on aga tegemist sümptomi suhtes kindlalt negatiivsete lausetega.

Lausete *Lindistas telefonikõnesid* ja *Räägib erinevatest vandenõuteooriatest - illuminaatidest, reptiilidest, vabamüürlastest* korral võib leida viiteid paranoilise luulu esinemisele ning lause *Meelepetted aktualiseerusid ja nende foonil hüppas aknast alla* korral hallutsinatsioonide esinemisele. Kuna antud laused kirjeldavad siiski käitumist, mis oma olemuselt pigem ebatavalised, on need märgendatud ka „veidra käitumise" suhtes positiivseteks. Aga näiteks lause *Patsiendi kehakeel kahtlustav, ärev* korral on pigem esiplaanil paranoiline hoiak ja otsene käitumise kirjeldus puudub, mistõttu on lause märgendatud negatiivseks.

Teise sümptomite grupi („derealisatsioon" ja „depersonalisatsioon") korral leidis mudel esimesel iteratsioonil sarnaselt „veidra käitumise" korral kõige kõrgema tõenäosusega lause *Liiklusõnnetus S06* ning kui eelnevalt oli kirjeldatud eitustega lausete ennustamist, siis seekord leidis mudel näiteks laused *Esinenud nägemismeelepetteid*, *Esines jälitusluulumõtteid, nägemismeelepetteid* ja *Esinevad mõttelevielamused* (tabel 13).

Tabel 13. Näiteid mudeli poolt ennustatud lausete kohta iteratsioonide kaupa, nende tõenäosused ning käsitsi märgendamisel määratud klass, sümptomiteks „depersonalisatsioon" ja „derealisatsioon".

Iteratsioon	Lause, milles mudel ennustas esinevat sümptomeid „depersonalisatsioon/derealisatsioon"	Tõenäosus sümptomi esinemisele	Hiljem määratud märgend
I	„liiklusõnnetus S06."	100%	0
I	„esinevad nägemismeelepetted."	99%	0
I	„Esinevad mõttelevielamused."	99%	0
I	„Aeg - ajalt derelisatsioonielamused."	99%	1
II	„Oma tunnetest võõrandunud."	77%	1
II	„Praegu püsvialt ebareaalsustunne."	71%	1
II	„Inimestel ja esemetel näeb mõnikord värvilist virvendust ümber."	69%	1

Lisaks esines ennustatud lausete hulgas ka palju lühikesi ja numbrilisi kombinatsioone (näiteks diagnoosikoode või objektiivset leidu sisaldavaid lauseid), näiteks „aug." (99%), „08." (99%), „1, Z71." (99%) ning „Sin Th 4,6,7 proc transv Fr." (99%). Teisel iteratsioonil

leidis mudel väga täpselt otsitavat sümptomit kirjeldavad 4 lauset: *Aeg - ajalt derelisatsioonielamused*, *Oma tunnetest võõrandunud*, *Praegu püsvialt ebareaalsustunne* ning *Inimestel ja esemetel näeb mõnikord värvilist virvendust ümber.* Siinkohal leidis mudel mitmeid lauseid, mis olid seotud pigem ärevushäiretega seotud sümptomitega, näiteks *Järsku tekib hirmutunne* (94%), *Ärevus, unehäired ning nägemismeelepetted* (91%), *Ärevus, muremõtted* (78%) ja *Liigsel muretsemisel tekib südamekloppimine* (74%). Võrreldes esimese iteratsiooniga vähenes väga lühikeste lausete esinemine ning leidus mitmeid ärevusele, meeleolulangusele, unehäiretele, sotsiaalsele isolatsioonile ning luulumõtetele ja meelepetetele viitavaid lauseid. „Depersonalisatsiooni" ja „derealisatsiooniga" lauseid oli uuritavates tekstides esindatud vähem ning see võis olla ka põhjus, miks mudel neid kogu iteratiivse protsessi peale nõnda vähe (4) juurde leidis.

Kolmanda sümptomite grupi („paranoiline luul" ja „paranoiline hoiak/kahtlustamine") korral leidis mudel mitmeid paranoilisi luulumõtteid ning „kahtlustavat" olekut sisaldavaid lauseid (tabel 14).

Tabel 14. Näiteid mudeli poolt ennustatud lausete kohta iteratsioonide kaupa, nende tõenäosused ning käsitsi märgendamisel määratud klass, sümptomiteks „paranoiline luulumõte" ja „paranoiline hoiak/kahtlustamine".

Iteratsioon	Lause, milles mudel ennustas esinevat sümptomeid „paranoiline luulumõte/kahtlustamine"	Tõenäosus sümptomi esinemisele	Hiljem määratud märgend
I	„Oli ärev ja hirmunud."	99%	0
I	„Kahtlustab, et sugulased tapetud."	99%	1
I	„Avaldab paranoilise sisuga jälitusluulumõtteid."	99%	1
I	„Telerist ja raadiost justkui räägitakse temast."	99%	1
II	„Kergesti tekkisid kahtlustused teiste inimeste pahatahtlikkuses."	97%	1
II	„Avaldas koduste suhtes paranoilist mürgistusluulu."	92%	1
III	„Kui naerdi, siis tundus, et naerdi teda."	77%	1
III	„Helistas tööandjale, et teda tahetakse tappa."	75%	1
III	„On süüdistanud naist mõnel korral ka enda mürgitamises."	65%	1
III	„Avaldanud hirme, et äkki on tulnukad pannud kehasse mingi kiibi ja kontrollivad teda selle kaudu."	65%	1

Antud sümptomite grupi juures leidis mudel ka mitmeid lauseid, mis olid seotud näiteks meelepetteliste elamustega (*Elavalt rääkis objektidega, keda ei olnud*, *Näeb olematuid loomi ja inimesi*, *Patsiendil avaldusid selgelt elavad kuulmismeelepetted*, *Arvas, et need on ehk kummitused või deemonid*), aga ka kordusid mitmed laused nii „veidra käitumise" kui ka „depersonalisatsioon/derealisatsioon" andmestikest (vastavalt *Ilmneb veider käitumine* ja *Vaadates enda ümber näeb objekte ning need tunduvad talle veidrad*). Teise iteratsiooni käigus leidis mudel sarnaselt teise sümptomi korral kirjeldatud olukorraga palju lühikesi ning numbritest koosnevaid lauseid, näiteks „4mmol/l." (99%), „x3k." (93%) ja „67sm." (93%).

Kolmandasse gruppi on haaratud üsna laialt levinud sümptomid ning peale luulumõtete on võetud arvesse ka kahtlustavat olekut ilma selge luulumõtte kirjelduseta, mistõttu võib selliseid lauseid andmestikus olla esindatud rohkem kui näiteks „veidrat käitumist" või

„depersonalisatsiooni” ja „derealisatsiooni” ning mudel võis seetõttu selliseid tõsipositiivseid lauseid kumulatiivselt rohkem ennustada.

6.3 Annoteerimise olulisus ja keerukus

6.3.1 Sümptomite tõlgendamine

Esimesena eraldatav sümptom („veider käitumine”) on oma kliinilise kasutatavuse poolest limiteeritud, kuna tegemist on üsna subjektiivse ning ajas muutuva nähtusega, mida on raske üheselt mõistetavalt defineerida ning seetõttu ka märgendada. Sagedasti oli andmebaasist märksõnadega „veider” või „veidr” väljavalitud tekstides sõnaga „veider” kirjeldatud palju mõtlemise sisulisi häireid (näiteks luulumõtted) või mõttekäiku. Esimest andmestikku „veidra käitumise” sümptomiga mudeli treenimiseks annoteeriti tegelikkuses kahel korral – esimesel korral olid andmestikus sümptomi suhtes positiivseteks märgendatud ka eelnevalt väljatoodud „veidraid mõtteid”, „veidraid veendumusi” ning „veidraid tundmuseid” sisaldavad laused, mille alusel ennustas mudel seetõttu suurema tõenäosusega mõtlemise häireid (luulumõtted) ning tajumishäireid (meelepetted). Näiteks võib tuua laused *Paranoilise sisuga mõtted eeskätt endise abikaasa, aga ka naabrite suhtes* (sümptomi esinemise tõenäosus 96%), *Ta arvab, et tema tütre, õe ja venna asemel on teisikud, kes kasutavad vääraid dokumente* (96%) ning *Lisaks luulumõtetele on ilmnenud kuulmis- (viimase kahe kuu jooksul hääl, mis kommenteerib, annab nõu, käsib jne), maitsmis-, haistmis- ja puutehallutsinatsioone* (86%). Ennustatud laused on seotud pigem paranoiliste luulumõtetega või meelepetteliste elamustega, kuid „veidra käitumise” korral tuleks eelkõige keskenduda käitumist kirjeldavatele lausetele (mis võivad endas lõpp-kokkuvõttes kanda edasi luulumõtete ja meelepetete tõttu tekkivaid situatsiooni). Seetõttu otsustati andmestik uuesti märgendada ning lugeda positiivseks ehk sümptomit sisaldavaks rangelt ainult patsiendi „veidrale” käitumisele viitavad laused ning negatiivseks näiteks „veidra” mõtlemise või mõttekäigu ja „veidrate” veendumuste ning arusaamadega laused (tabel 15).

Tabel 15. Annoteerimise erinevused.

Märgendatav lause	Esmasel annoteerimisel	Teisel annoteerimisel
“... ühteaegu veidrate arusaamadega, ja arutleja, räägi kuidas vaim temaga rääkis...”	1	0
“Mõttekäigus ilmnevad veidrad veendumused, kahtlustamine ja paranoilised mõtted, mille vaidlustamine või teise tõlgenduse pakkumine ei õnnestu.”	1	0
“Sama päeva öösel oli ta tüdruksõbraga jalutama läinud ning temalegi tavapärasest veidramat juttu rääkinud.”	1	0

Pärast märgendite teisest hindamist ja parandamist oli täheldatav eelnevalt väljatoodud testlausete ennustuste tõenäosuste vähenemine luulumõtteid kirjeldavate lausete juures, näiteks lausete *Paranoilise sisuga mõtted eeskätt endise abikaasa, aga ka naabrite suhtes* (eelnevalt sümptomi esinemise tõenäosus 96% → nüüd 46%) ning *Ta arvab, et tema tütre, õe ja venna asemel on teisikud, kes kasutavad vääraid dokumente* (eelnevalt samuti 96% → nüüd 58%).

6.3.2 Lausete tükeldamisega kaasnev konteksti puudumine

Käesolevas lõputöös kasutati sümptomite ennustamiseks lauseid, mis on saadud tekstide tükeldamisel. Lausete kaupa märgendamise võib aga teatud juhtudel muuta keeruliseks lauset ümbritseva teksti ehk konteksti puudumine. Näiteks „veidra käitumise" märgendamisel võis otsus, kas kirjeldatav käitumine või tegutsemine võiks viidata psühhopatoloogiale või kontekstist eemaldatud „veidrana" näivale situatsioonile, vajada ülejäänud dokumendi teksti lugemist. Näiteks mudeli poolt ennustatud lause *Lindistas telefonikõnesid* korral ilmnes ülejäänud tekstist veel lisaks, et patsient oli episooditi käitunud ebaadekvaatselt (kaasates näiteks alusetult situatsioonidesse ka politsei), mistõttu märgendati lause „veidrat käitumist" sisaldavaks lauseks, mis võib koos oma käitumismustriga viidata patsiendi võimalikule paranoilisele hoiakule.

Sarnaselt võib vaadelda ka mudeli poolt ennustatud lauset *Seab kahtluse alla psühhiaatria olemuse ning pädevuse inimeste hindamisel*, mis üksikuna ei ole sümptomina klassifitseeritav, kuid lisades sinna konteksti, kus patsient on ebaadekvaatse käitumisega, vastumeelne ning kahtlustav ravitöö suhtes, võib see olla viide paranoilise hoiaku ja paranoiliste luulumõtete osas.

Kui võtta näiteks lause *Patsiendi sõnul soovib elukaaslase eksmees teda tappa*, siis antud juhul on haigusloo tekstist välja loetav ka asjaolu, et patsient on viidud haiglasse kiirabi ja politsei saatel paranoilise käitumise tõttu, mida on varasemalt esinenud ka näiteks patsiendi töökohas. Siinkohal võib olla tegemist paranoilise luulumõttega. Kuid üksiku lause põhjal ei saa alati väita, et tegemist ei ole patsiendi elus tegelikkuses asetleidva situatsiooniga (eriti näiteks suhteprobleemid, kiusamine). Näiteks kinnitas ühes haigusloos patsiendi elukaaslane patsiendile teise isiku poolt tehtud tapmisähvardusi. Siinkohal on oluline võimalus, et märgendatavatel lausetel on olemas oma identifikaator, millega vajadusel leida ja tutvuda originaaldokumendiga.

6.3.3 Spetsiifiline terminoloogia

Peale mitmeti tõlgendatavate sümptomite (näiteks „veider käitumine") võivad keeruliseks osutuda ka spetsiifilisemat terminoloogiat kasutavad sümptomid, näiteks eritüüpi luulumõtted. Mudel ennustas, et lauses *Avaldab arvamust, et tal võivad olla erilised „para"võimed…* esineb paranoilise luulu sümptom, kuid siinkohal on aga tegemist hoopis suurusluulu alla kuuluva luulumõttega. Siinkohal tuleb esile psühhiaatriliste tekstide korrektse märgendamise olulisus, mis kajastab vajadust nii käsitletava valdkonna sügavamate teadmiste ja kliinilise kogemuse olemasolust kui ka ideaalis mitme spetsialisti ülevaatamist.

6.4 Sümptomite valik

Käesoleva lõputöös on käsitletud peamiselt psühhoosi prodroomi sümptomit „veider käitumine". Kuna tegemist on võrdlemisi laialivalguva ning ebaspetsiifilise sümptomiga, ei pruugi selle alusel lihtsamate masinõppemudelitega ennustamine olla kõige efektiivsem. Teisalt, kui lõpp-eesmärgiks on seatud näiteks psühhoosiriskiga isikute tuvastamine, siis käitumises tekkivad kõrvalekalded võivad hästi peegeldada psühhootiliste sümptomite esmast avaldumist laiemas perspektiivis, hõlmates endas viiteid nii mõtlemise kui ka tajuhäiretele. „Veider käitumine" on sümptomina kergemini täheldatav ka näiteks patsiendi lähedastele või teda ümbritsevatele inimestele aga ka teistele meditsiinivaldkonna spetsialistidele (näiteks perearst, erakorralise meditsiini arst) ning seetõttu olla kirjeldatud juba varasemates meditsiinidokumentides, enne psühhiaatri või psühholoogi juurde jõudmist.

Meetodi valideerimiseks kasutatud sümptomid „depersonalisatsioon" ja „derealisatsioon" on valitud erialaspetsialisti soovitusel. Kuigi nii algne regulaaravaldise abil koostatud andmestik kui ka lõplik iteratiivselt koostatud andmestik on antud sümptomite korral üks väiksemaid ning rakendatud mudel suutis tõeselt tuvastada ainult neli lauset, võib täheldada, et positiivseks märgendatud laused olid otsitavaid sümptomeid hästi kirjeldavad, sisaldades mitte ainult mõistet ennast (nt „derealisatsioon") vaid just sümptomi kirjeldust (näiteks *Inimestel ja esemetel näeb mõnikord värvilist virvendust ümber*, tabel 13). Antud sümptomit sisaldavate lausete tuvastamise põhjuseks võib olla nende väiksem osakaal valitud meditsiinidokumentides ning seetõttu ka nende vähesus algses treeningandmestikus.

Teise valideerimiseks kasutatud sümptomite grupiga („paranoilised luulumõtted" ja „kahtlustamine") koostatud algne regulaaravaldise abil koostatud andmestik sisaldas endas kõige suuremal hulgal lauseid. See võib tähendada, et antud sümptomeid on valitud meditsiinitekstides kirjeldatud kolmest töös käsitletud sümptomist kõige enam, mistõttu suutis mudel kokku leida juurde 100 uut lauset. Lisaks on nii luulumõtete kui ka „kahtlustava" hoiaku kirjeldused üsna laiahaardelised, koondades enda alla nii mõtlemise sisulisi häireid kui ka käitumuslikke kõrvalekaldeid, tuues siia alla ka näiteks „veidra käitumise" all käsitletavaid lauseid.

6.4.1 Mudel ja sümptomite tuvastamine

Psühhiaatriline anamnees on suuresti kirjeldava iseloomuga ning lisaks konkreetsete sümptomite väljatoomisele (näiteks *Esinevad jälitusluulumõtted*) antakse edasi ka nende sümptomite iseloom (näiteks *Patsient tunneb, kuidas teda pidevalt jälitatakse*). Siinkohal oli huvitav vaadelda, milliseid lauseid ennustas mudel üsna konkreetsete näidete alusel. Näiteks oli algses andmestikus „veidra käitumise" tuvastamiseks lause *Ilmneb veider käitumine, aastaid polnud kodust väljas käinud* ning erinevate iteratsioonide käigus ennustas mudel muuhulgas positiivseks lauseid *Elab üksinda*, *Ema andmetel poeg elab üksinda suvilas, sotsiaalselt isoleerunud* ja *Korduvalt rõhutab üksildustundele*. Antud lausete puhul ei ole tegemist „veidra käitumisega", kuid need võivad viidata näiteks sotsiaalsele isoleerumisele, mis on sageli psühhoosi prodroomis kirjeldatud nähtus. Samamoodi võis täheldada, et näiteks „depersonalisatsiooni" ja „derealisatsiooni" korral ennustas mudel mitmel korral positiivseks hoopis meeleolulangust või ärevust kirjeldavaid lauseid, mis on sagedasti ilmnenud just

depressiooni või foobiate korral [24]. Siinkohal on oluline märkida, et psühhootilisi sümptomeid võib esineda ka teiste seisundite, näiteks psühhoaktiivsete ainete tarvitamisest tingitud seisundite või ka raskete kehaliste haiguste korral [2], mistõttu võib üksikute sümptomite põhjal ennustuste tegemine olla ebaspetsiifiline meetod ning oluline on kasutada meetodit, mis suudab säilitada teksti terviklikumat konteksti.

6.4.2 Psühhiaatria valdkonna sõnade puudumine kasutatud Word2Vec mudelis

Üheks piiranguks, mis käesolevas töös seoses Word2Vec tekstikorpusega kasutamisega ilmnes, on asjaolu, et selles ei pruugi sisalduda spetsiifilisi psühhiaatria termineid. Näiteks ei leidunud kasutatud tekstikorpuses sõnu (lisaks eelnevalt tabelis 10 väljatoodud sõnadele) „kuulmismeelepetted", „nägemismeelepetted", „psühhoosile", „mõttekäiguhäired", „luululiselt", „mõttelevielamused", „derealisatsioon", „depersonalisatsioon", „kahjustusluulumõtteid". Kui lauses esinevad sellised sõnad, siis neid käesolevas töös rakendatud mudel ei kasutanud. Samal ajal võivad need aga olla prodroomi sümptomite tuvastamisel väga olulised. Näiteks sõnad „derealisatsioon" ja „depersonalisatsioon" olid valitud meetodi valideerimisel eraldatavateks sümptomiteks. Selle probleemi lahendamiseks võiks proovida edaspidi täiendada olemasolevat Word2Vec mudelit meditsiiniterminoloogia alaselt või kasutada sõnade esitamiseks mõnda muud meetodit.

6.5 Edasised soovitused

Vaatamata töös rakendatud metoodikaga kaasnenud piirangutele, võib eelnevalt kirjeldatud tulemuste ja arutelu põhjal siiski järeldada, et lausete representatsioonide alusel loodud tunnustega on võimalik kontsentreeritumalt leida üles otsitavat sümptomit sisaldavaid lauseid ning muuta seeläbi märgendatud treeningandmestike loomist kiiremaks ja tõhusamaks.

Metoodika ühe piiranguna on varasemalt välja toodud psühhiaatria valdkonnale spetsiifiliste sõnade puudumist eeltreenitud Word2Vec´i mudelist. Puuduva terminoloogia kaasamine mudelisse võiks oluliselt edendada eestikeelse psühhiaatrilise teadustöö arenemist ning seeläbi ka valdkonna kaasajastamist. Näiteks võiks Word2Vec´i edasisel treenimisel abiks olla Tartu Ülikooli Kliinikumi psühhiaatria õppetooli avalikult kättesaadavad veebipõhised õppematerjalid [1] ning antud lõputöö raames loodud andmestikud.

Sobivamate lausete leidmiseks iteratiivses protsessis, tasuks lisaks katsetada erinevate masinõppemudelite, näiteks süvaõppel põhinevad mudelid, tulemuslikkust töös kasutatud metoodika rakendamisel.

Selleks, et töös kirjeldatud metoodikat tuua kliinilisele teadustööle lähemale, on võimalik luua TextHunter´i tööriistale sarnane lahendus, mis peidab erialaspetsialisti eest masinõppe detailid ning laseb keskenduda andmestike koostamisele.

Sarnast metoodikat saaks kasutada erinevate sümptomite leidmiseks meditsiinitekstidest, et luua märgendatud andmestikke ja lihtsustada edasisi uuringuid. Siinkohal ei peaks piirnema ainult psühhootiliste häiretega, vaid uurimisvaldkonda võiks laiendada ka teistele psüühikahäiretele.

7. Kokkuvõte

Käesolevas lõputöös rakendati pool-automaatseid keeletehnoloogilisi vahendeid psühhiaatriliste sümptomitega märgendatud andmestike loomiseks. Algandmestikuna kasutati Eesti rahvastiku meditsiinidokumente aastatest 2012-2019, millest leiti psühhoosi prodroomile viitavad laused. Töö käigus valmis kolm märgendatud andmestikku psühhoosi prodroomi sümptomite eraldamiseks: 799 lausega andmestik sümptomiga „veider käitumine", 643 lausega andmestik sümptomitega „depersonalisatsioon" ja „derealisatsioon" ning 1176 lausega andmestik sümptomitega „paranoiline luulumõte" ja „paranoiline hoiak/kahtlustamine".

Andmestike koostamiseks kasutati iteratiivset meetodit, kus esimesel sammul eraldati sümptomit sisaldavad laused andmebaasist regulaaravaldiste abil, märgendati need käsitsi töö autori poolt, leiti lausetele vastavad vektorid ning treeniti nende alusel logistilise regressiooni mudel. Igal järgneval sammul ennustati treenitud mudeli abil sümptomi esinemist andmebaasi igas lauses ning märgendati ja lisati treenimiseks kasutatavasse andmestikku kuni 300 mudeli poolt ennustatud kõige tõenäolisemalt (vähemalt 50% tõenäosusega) sümptomit sisaldavat lauset.

Töös on kirjeldatud mitmeid probleeme, mis on seotud algmaterjali komplitseerituse, kasutatud mudelitega kui ka andmete märgendamisega. Valitud meditsiinitekstid sisaldasid rohkelt vektor-kujule viimist segavaid tegureid (näiteks kirja- ja trükivead, duplitseeritud tekstid) ning tekstide tükeldamist raskendavaid tegureid (näiteks punktidega tähistatud lühendid, kuupäevad, ravimite doosid, objektiivne leid). Töös rakendatud metoodika aitas kiirendada ja lihtsustada andmestike koostamist, kuid loodud logistilise regressiooni mudeli primitiivsus ning mitmete psühhiaatriale omaste sõnade puudumine kasutatud Word2Vec´i mudelist olid piiravad tegurid otsitavate sümptomitega lausete ülesse leidmiseks. Kuna töö rõhk oli andmestike koostamisel, siis esialgu ei pandud rõhku märgendeid ennustava mudeli headusele, mille tõttu ei leidnud mudel palju tõsipositiivseid lauseid. Andmete märgendamise muutsid keeruliseks nii lausete tõlgendamine sümptomi suhtes kui ka konteksti puudumine.

Vaatamata puudustele, aitas metoodika leida piisaval hulgal otsitavate sümptomitega lauseid, jälgides nii lause konteksti kui ka konkreetsemaid termineid. Väljatoodud puuduste leevendamisega ning meetodi edasiarendamisega on võimalik veel enam hõlbustada märgendatud andmestike loomist.

Töö tulemusena valminud andmestikke saab perspektiivis kasutada masinõppemudelite treenimisel psühhoosi prodroomi sümptomite eraldamiseks ning alusena suuremate andmestike loomiseks.

8. Tänusõnad

Käesolev töö on läbi viidud uuringute RITA1/02-96 ja PRG1844 raames. Töö autor avaldab tänusõnad juhendajatele Sulev Reisbergile ja Kairit Sirtsile.

Viidatud kirjandus

[1] Tartu Ülikooli Psühhiaatria õppetool. Tartu Ülikooli Psühhiaatria ainekursuse õppematerjalid. https://valisveeb.kliinikum.ee/psyhhiaatriakliinik/lisad/ravi/FR-Ravi5k.htm (10.05.2024).

[2] Geddes J. R., Andreasen N. C., Goodwin G. M. New Oxford textbook of Psychiatry, Third Edition. Oxford University Press. 2020.

[3] Lieberman J. A., First M. B. Psychotic Disorders. The New England Journal of Medicine, 2018, Vol. 379, no. 3, p. 270-280.

[4] Irving J. P. R., Oliver D., Colling C., Pritchard M., Broadbent M., Baldwin H., Stahl D., Stewart R., Fusar-Poli P. Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. Schizophrenia Bulletin, 2021, Vol. 47, no. 2, p. 405-414.

[5] Synbase Meditsiinisõnastik. https://app.synbase.eu/app/et/s%C3%B5nastik.

[6] Hjaltelin J. X., Currant H., Jørgensen I. F., Brunak S. Visualising disease trajectories from population-wide data. Front. Bioinform, 2023, Vol. 3.

[7] Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 3rd Edition. O'Reilly Media, Inc. 2020.

[8] Andmekaitse ja infoturbe leksikon. Cybernetica AS. https://akit.cyber.ee/.

[9] Modinos G., McGuire P. The prodromal phase of psychosis. Current Opinion in Neurobiology, 2015, Vol. 30, p. 100-105.

[10] Benrimoh D., Dlugunovych V., Wright A. C., Phalen P., Funaro M. C., Ferrara M., Powers III A. R., Woods S. W., Guloksuz S., Yung A. R., Srihari V., Shah J. On the proportion of patients who experience a prodrome prior to psychosis onset: A systematic review and meta-analysis. Molecular Psychiatry, 2024.

[11] Schultze-Lutter F., Michel C., Schmidt S. J., Schimmelmann B. G., Maric N. P., Salokangas R. K. R., Riecher-Rössler A., van der Gaag M., Nordentoft M., Raballo A., Meneghelli A., Marshall M., Morrison A., Ruhrmann S., Klosterkötter J. EPA guidance on the early detection of clinical high risk states of psychoses. European Psychiatry, 2015, Vol. 30, no. 3, p. 405-416.

[12] Larson M. K., Walker E. F., Compton M. T. Early signs, diagnosis and therapeutics of the prodromal phase of schizophrenia and related psychotic disorders. Expert Rev Neurother, 2010, Vol. 10, p. 1347-1359.

[13] Fusar-Poli P., Bonoldi I., Yung A. R., Borgwardt S., Kempton M. J., Valmaggia L., Barale F., Caverzasi E., McGuire P. Predicting Psychosis: Meta-analysis of Transition Outcomes in Individuals at High Clinical Risk. Arch Gen Psychiatry, 2012, Vol. 69, no. 3, p. 220–229.

[14] Addington J., Heinssen R. Prediction and Prevention of Psychosis in Youth at Clinical High Risk. Annual Review of Clinical Psychology, 2012, Vol. 8, p. 269-289.

[15] Barajas A., Pelaez T., González O., Usall J., Iniesta R., Arteaga M., Jackson C., Baños I., Sánchez B., Dolz M., Obiols J. E., Haro J. M., GENIPE Group, Ochoa S. Predictive capacity of prodromal symptoms in first-episode psychosis of recent onset. Early Intervention in Psychiatry, 2017. Vol. 13, no. 3, p. 414-424.

[16] Schnell K., Heekeren K., Daumann J., Schnell T., Schnitker R., Möller-Hartmann W., Gouzoulis-Mayfrank E. Correlation of passivity symptoms and dysfunctional visuomotor action monitoring in psychosis. Brain, 2008.

[17] Sirts K., Anni K., Balõtšev R., Jakobsoo S., Jaanson K.-L., Haring L. Adapting the early recognition inventory ERIraos to Estonian: A validation study. Early Intervention in Psychiatry, 2024.

[18] Jakobsoo S. Psühhoosiriski hindavate mõõdikute adapteerimine Eesti oludele. TÜ psühholoogia instituudi magistritöö. 2017.

[19] Alvarez-Jimenez M., Gleeson J. F., Henry L. P., Harrigan S. M., Harris M. G., Amminger G. P., Killackey E., Yung A. R., Herrman H., Jackson H. J., McGorry P. D. Prediction of a single psychotic episode: A 7.5-year, prospective study in first-episode psychosis. Schizophrenia Research, 2011, Vol. 125, no. 2-3, p. 236-246.

[20] Maurer K, Zink M, Rausch F, Häfner H. The early recognition inventory ERIraos assesses the entire spectrum of symptoms through the course of an at-risk mental state. Early Intervention in Psychiatry, 2018, Vol 12, p. 217-228.

[21] Green M. F., Bearden C. E., Cannon T. D., Fiske A. P., Hellemann G. S., Horan W. P., Kee K., Kern R. S., Lee J., Sergi M. J., Subotnik K. L., Sugar C. A., Ventura J., Yee C. M., Nuechterlein K. H. Social cognition in schizophrenia, Part 1: performance across phase of illness. Schizophrenia Bulletin, 2012, Vol. 38, no.4, p. 854-864.

[22] Hunt C., Borgida E., Lavine H. Social Cognition. Encyclopedia of Human Behavior, Second Edition. Elsevier Inc. 2012.

[23] Van Rijn S., Schothorst P., Van't Wout M., Sprong M., Ziermans T., van Engeland H., Aleman A., Swaab H. Affective dysfunctions in adolescents at risk for psychosis: Emotion awareness and social functioning. Psychiatry Research, 2011, Vol. 187, no. 1-2, p. 100-105.

[24] Maailma Tervishoiuorganisatsioon. RHK-10: V - Psüühika- ja käitumishäired elektrooniline versioon. https://valisveeb.kliinikum.ee/psyhhiaatriakliinik/lisad/ravi/RHK/RHK10-FR17.htm (10.05.2024).

[25] Büetiger J. R., Hubl D., Kupferschmid S., Schultze-Lutter F., Schimmelmann B. G., Federspiel A., Hauf M., Walther S., Kaess M., Michel C., Kindler J. Trapped in a Glass Bell Jar: Neural Correlates of Depersonalization and Derealization in Subjects at Clinical High-Risk of Psychosis and Depersonalization–Derealization Disorder. Front. Psychiatry, 2020. Vol. 11.

[26] Worthington M. A., Cao H., Cannon T. D. Discovery and Validation of Prediction Algorithms for Psychosis in Youths at Clinical High Risk. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2020. Vol. 5, no 8, p. 738-747.

[27] Sheikhalishahi S., Miotto R., Dudley J. T., Lavelli A., Rinaldi F., Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Medical Informatics, 2019, Vol. 7, no. 2.

[28] Mehta N., Pandit A. Concurrence of big data analytics and healthcare: A systematic review. International Journal of Medical Informatics, 2018, Vol. 114, p. 57-65.

[29] Wei W.-Q., Teixeira P. L., Mo H., Cronin R. M., Warner J. L., Denny J. C. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. Journal of the American Medical Informatics Association, 2016, Vol. 23, no.1, p. 20-27.

[30] Jackson R. G., Patel R., Jayatilleke N., Kolliakou A., Ball M., Gorrell G., Roberts A., Dobson R., Stewart R. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open, 2017, Vol. 7, issue 1.

[31] Koleck T. A., Dreisbach C., Bourne P. E., Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. Journal of the American Medical Informatics Association, 2019, Vol. 26, Issue 4, April 2019, p. 364–379.

[32] Wang Y., Wang L., Rastegar-Mojarad M., Moon S., Shen F., Afzal N., Liu S., Zeng Y., Mehrabi S., Sohn S., Liu H. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, 2018, Vol. 77, p. 34-49.

[33] Yadav K., Sarioglu E., Smith M., Choi H.-A. Automated Outcome Classification of Emergency Department Computed Tomography Imaging Reports. Academic Emergency Medicine, 2013, Vol. 20, Issue 8, p. 848-859.

[34] Jurafsky D., Martin J. H. Speech and Language Processing, Third Edition. Chapter 5: Logistic Regression. 2024.

[35] Di Gennaro G., Buonanno A., Palmieri F. A. N. Considerations about learning Word2Vec. The Journal of Supercomputing, 2021, Vol. 77, p. 12320–12335.

[36] Sonabend W. A., Pellegrini A. M., Chan S., Brown H. E., Rosenquist J. N., Vuijk P. J., Doyle A. E., Perlis R. H., Cai T. Integrating questionnaire measures for transdiagnostic psychiatric phenotyping using word2vec. PLoS ONE, 2020, Vol. 15.

[37] Google Code Archive: Word2Vec. https://code.google.com/archive/p/word2vec/ (21.04.2024).

[38] Orasmaa S., Petmanson T., Tkachenko A., Laur A., Kaalep H.-J. EstNLTK - NLP Toolkit for Estonian. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, p.2460–2466.

[39] Eesti keele koondkorpus. https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (10.05.2024).

[40] Sadock B. J., Sadock V. A., Ruiz P. Kaplan and Sadock´s Synopsis of Psychiatry. Behavioral Sciences/Clinical Psychiatry, Eleventh Edition. Wolters Kluwer. 2015.

[41] Karystianis G., Nevado A. J., Kim C.-H., Dehghan A., Keane J. A., Nenadic G. Automatic mining of symptom severity from psychiatric evaluation notes. International Journal of Methods in Psychiatric Research, 2018, Vol. 27, Issue 1.

[42] Stephan K. E., Mathys C. Computational approaches to psychiatry. Current Opinion in Neurobiology, 2014, Vol. 25, p. 85-92.

[43] Abbe A., Grouin C., Zweigenbaum P., Falissard B. Text mining applications in psychiatry: a systematic literature review. International Journal of Methods in Psychiatric Research, 2016, Vol. 25, Issue 2, p. 86-100.

[44] Jackson R. G., Ball M., Patel R., Hayes R. D., Dobson R. J. B., Stewart R. TextHunter – A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research. AMIA Annual Symposium Proceedings Archive, 2014, p. 729-738.

[45] Zhang Y., Zhang O., Wu Y., Lee H.-J., Xu J., Xu H., Roberts K. Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge. Journal of Biomedical Informatics, 2017, Vol. 75, p. 129-137.

[46] Uusküla A., Oja M., Tamm S., Tisler A., Laanpere M., Padrik L., Nygard M., Reisberg S., Vilo J., Kolde R. Prevaccination Prevalence of Type-Specific Human Papillomavirus Infection by Grade of Cervical Cytology in Estonia. JAMA Network Open, 2023.

[47] Rutledge R. B., Chekroud A. M., Huys Q. J. M. Machine learning and big data in psychiatry: toward clinical applications. Current Opinion in Neurobiology, 2019, Vol. 55, p. 152-159.

[48] Chekroud A. M., Bondar J., Delgadillo J., Doherty G., Wasil A., Fokkema M., Cohen Z., Belgrave D., DeRubeis R., Iniesta R., Dwyer D., Choi K. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry, 2021, Vol. 20, Issue 2, p. 154 – 170.

[49] Bzdok D., Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2018, Vol. 3, Issue 3, p. 223-230.

[50] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality. arXiv, 2013.

[51] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv, 2013.

[52] Müller A. C., Guido S. Introduction to Machine Learning with Python. O'Reilly Media, Inc. 2016.

Lisad

Lisa 1 - Valminud andmestikud

Käesoleva magistritöö raames loodud treeningandmestikud sisaldavad lausete väljavõtteid patsientide epikriisisdest ja tulenevalt uuringu protokollist ning eetikakomitee loast ei tohi neid avaldada, kuna see võib rikkuda patsientide privaatsust. Seetõttu on valminud andmestikud kättesaadavad vaid uurimisgrupi liikmetele, kes saavad neid kasutada sümptomite eraldamise mudelite arendamiseks. Samuti on tabelid tehtud kättesaadavaks käesoleva töö retsensendile. Siin lisas on näitena ära toodud vaid andmestikest pärit mõned piisavalt anonüümsed read.

Tabel 16. Andmestiku näide sümptomi „veider käitumine" kohta.

Lause	On sümptom
„ …(b) veider, ebaharilik või pentsik käitumine ja välimus; (c) vähene kontakt teiste inimestega ja sotsiaalse eraldumise tendents; (d) veidrad veendumused või maagiline mõtlemine, mis mõjutavad käitumist ega vasta subkultuuri normidele; ... "	1
„Ilmneb veider käitumine."	1
„Mõtekäiguhäire, esile tulevad asjade teisejärgulised tunnused, veidrused."	0

Tabel 17. Andmestiku näide sümptomite „depersonalisatsioon" ja/või „derealisatsioon" kohta.

Lause	On sümptom
„Avaldab, et kõik tundub ebareaalne."	1
„Viimaste kuude jooksul süvenenud häirituse tunne ümbruskonnast, sagenenud derealisatsioonitunded „kahes olemine", väldib suhtlemist."	1
„Esinevad depressiivsed mõtted, saamatuse- ja alaväärsustunne, hirm tuleviku ees ja lootusetus."	0

Tabel 18. Andmestiku näide sümptomite „paranoiline luulumõte” ja/või „kahtlustamine” kohta.

Lause	On sümptom
„Avaldas vihjamisi paranoilise sisuga kahjustusmõtteid, süüluulu fragmente.”	1
„Ravisoostumus puudub, ravimite suhtes paranoiliselt meelestatud.”	1
„Paranoilisi mõtteid ei avalda.”	0

Lisa 2 - Magistritöö Jupyter Notebook

Magistritöö tööprotsess on kättesaadav Jupyter´i koodivihiku kujul töö autori repositooriumist[7].

[7] https://github.com/k1r1i1s1t1e1l/Extraction-of-psychosis-prodromal-symptoms-from-medical-texts-for-training-dataset-creation.git

Lisa 3 - Lihtlitsents

Mina, Kristel Agu,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Psühhoosi prodroomi sümptomite eraldamine meditsiinitekstidest treeningandmestike loomiseks", mille juhendajateks on Sulev Reisberg (PhD) ja Kairit Sirts (PhD), reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristel Agu

10.05.2024