TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Norman Tolmats

Nimeüksuste tuvastamine ajaloolistes Tartu Linnavolikogu protokollides

Bakalaureusetöö (9 EAP)

Juhendaja:
Siim Orasmaa, PhD

Tartu 2025

Nimeüksuste tuvastamine ajaloolistes Tartu Linnavolikogu protokollides

Lühikokkuvõte:

Käesolevas töös uuritakse võimalusi, kuidas masinõppe abil tuvastada nimeüksusi Tartu Linnavolikogu 1918.-1940. a koosolekute protokollides. Enamik olemasolevaid mudeleid, mis nimeüksusi automaatselt tuvastavad, on loodud tänapäevase keele alusel. Vanema kirjakeele puhul ei anna need aga piisavalt häid tulemusi. Väärtuslike ajalooliste dokumentide märgendamiseks on vaja treenida spetsiaalsed mudelid või kohandada olemasolevaid, kui andmeid on vähe. Käesoleva töö käigus analüüsitakse olemasolevaid nimeüksusi märgendavaid mudeleid ja nende üldistuvust vanemale kirjakeele. Kuna kasutada on vähe kvaliteetseid andmeid, kohandatakse leitud parim mudel masinõppe abil antud ajalooliste protokollide märgendamiseks sobivamaks. Käesolev töö näitab, et kasutades väikest hulka märgendatud ja suurt hulka märgendamata vanemaid dokumente, on nõrgalt juhendatud masinõppe abil võimalik kohandada mudel, mille tulemused on vanema kirjakeele puhul paremad kui algsel tänapäeva keelel loodud mudelil.

Võtmesõnad: Nimeüksuste tuvastamine, masinõpe, juhendatud masinõpe, nõrgalt juhendatud masinõpe, ajaloolised andmed

CERCS: P176 Tehisintellekt

Named Entity Recognition in Historic Tartu City Council Meeting Protocols

Abstract:

This thesis explores the use of machine learning for named entity recognition (NER) in the meeting protocols of the Tartu City Council from 1918 to 1940, which are in Estonian. Most existing named entity recognition models for Estonian have been developed using modern language data and perform poorly when applied to historical texts. To effectively annotate valuable historical documents, it is necessary either to train specialized models or to adapt existing ones — particularly when only a small amount of labeled data is available. This study analyzes current NER models and evaluates their suitability for older language. Given the limited availability of high-quality labeled data, the best-performing model is adapted using machine learning techniques to be more suitable for these historical meeting protocols. The results demonstrate that, by using a small amount of labeled data and a large corpus of unlabeled historical documents, it is possible to improve model performance through weakly supervised learning — achieving better results on older language than models trained on modern language data.

Keywords: Named entity recognition, machine learning, supervised machine learning, weakly supervised machine learning, historical data

CERCS: P176 Artificial intelligence

Sisukord

Sissejuhatus

Loomuliku keele töötluse üks ülesannetest on tekstist nimeüksuste automaatne tuvastamine. Keeletehnoloogias nimetatakse nimeüksusteks nimesid, mis eristavad konkreetseid isikuid, kohti ja muud olulist samalaadsete seast[1]. Nimeüksusteks võivad olla näiteks isiku- ja kohanimed ning organisatsioonide nimed. Nimeüksuste tuvastamiseks või märgendamiseks nimetatakse loomuliku keele töötluse ülesannet, kus tekstist tuvastatakse nimed ja neile määratakse konkreetsed kategooriad [1]. Inglise keeles nimetatakse sellist märgendamist *named entity recognition* ehk lühidalt NER.

Käesoleva töö eesmärk on uurida võimalusi nimeüksuste tuvastamiseks Tartu Linnavolikogu koosolekute protokollides, mis pärinevad aastatest 1918-1940. Tänapäevase kirjakeele jaoks on olemas piisavalt andmeid, et treenida erinevaid nimeüksuste tuvastamise mudeleid, kuid eelmainitud Tartu Linnavolikogu protokollide ajaperioodi ja valdkonna kohta on treeningandmeid vähe. Seetõttu puudub võimalus treenida nende märgendamiseks spetsiaalne mudel. Käesolevas töös analüüsitakse, kuidas töötavad olemasolevad tänapäeva kirjakeelel treenitud nimeüksusi märgendavad mudelid ajaloolisel andmestikul. Seejärel uuritakse, kuidas kohandada ühte neist olemasolevatest mudelitest vanema kirjakeelega protokollides nimeüksusi paremini tuvastama.

Töö esimeses peatükis antakse ülevaade vana kirjakeele andmestike töötlusest ja sellega kaasnevatest probleemidest, tutvustatakse olemasolevaid nimeüksusi tuvastavaid mudeleid ning seletatakse lahti töös kasutatavad masinõppe meetodid.

Teises peatükis kirjeldatakse ajalooliste Tartu Linnavolikogu protokollide andmestikke, nende andmete kohandamist masinõppe rakendamiseks ja olemasolevate mudelite analüüsimiseks.

Kolmandas peatükis analüüsitakse, kuidas töötavad olemasolevad nimeüksusi märgendavad mudelid kohandatud andmestikel. Selleks rakendatakse kahele käsitsi märgendatud andmestikule kolme olemasolevat nimeüksusi tuvastavat mudelit ja uuritakse saadud tulemusi. Nendest olemasolevatest mudelitest kaks on loodud tänapäeva kirjakeele ja üks 19. sajandi kirjakeele põhjal.

[1] https://pre.sonaveeb.ee/search/unif/dlall/korpling/nime%C3%BCksus/1/est

Neljandas peatükis kasutatakse kahte erinevat masinõppe meetodit eelmises peatükis parimaks osutunud mudeli kohandamiseks, et tuvastada protokollides nimeüksusi senisest paremini. Lõpetuseks võrreldakse erinevate mudelite saadud tulemusi.

Lisades on link töös kasutatud koodifailide repositooriumile.

1. Teoreetiline ülevaade

1.1 Vana kirjakeele andmestike töötlus

Ajalooliste tekstide töötlemisega kaasnevad probleemid, mida ei esine tänapäevaste tekstide puhul. Maud Ehrmann jt [2] leiavad oma artiklis, et ajalooliste tekstide juures ilmneb neli peamist probleemi: tekstid on väga erinevat tüüpi, ajalooline keel erineb tänapäevasest keelest, andmetes esineb müra ja andmeid pole kuigi palju. Autorid toovad välja, et enamasti esineb tänapäevaste tekstide juures korraga vaid üks neist probleemidest. Mudelid saavad ühe probleemi lahendamisega hästi hakkama. Artikli autorite arvates on ajalooliste tekstide probleem see, et korraga esineb mitu erinevat eelnevalt välja toodud probleemi. See tähendab, et ajaloolisi tekste on keerulisem analüüsida kui tänapäevaseid tekste. Keerulisemad seosed võivad olla põhjuseks, miks tänapäevastel andmetel treenitud mudelid saavad ajalooliste tekstide puhul kehvemad tulemused.

Maud Ehrmann jt [2] artiklis välja toodud probleemid esinevad selgelt ka eestikeelsete ajalooliste tekstide puhul. Maarja-Liisa Pilvik jt [3] märgivad, et 19. sajandi Eesti vallakohtu protokollide analüüsimisel oli suureks probleemiks tänapäevase ja vanema eesti kirjakeele märkimisväärne erinevus. Tänapäevase keele jaoks loodud automaattöötluse vahenditel oli raskusi õigete sõnavormide äratundmisega. Näiteks oli autorite sõnul raskusi analüüsi käigus tuvastada päris- ja üldnimesid, sest 19. sajandi kirjaviisis kasutati suurt algustähte oluliselt enam.

Vana kirjakeele erinevuse probleem on aktuaalne ka käesolevas töös kasutatavate Tartu Linnavolikogu 1918.–1940. aasta koosolekute protokollide puhul. Need protokollid on väärtuslikud ajalooallikad, mille abil saab uurida möödunud sajandi esimese poole Tartu linna juhtimist, institutsioonide toimimist ja poliitilist kultuuri. Materjali põhjalikuks ja süstemaatiliseks uurimiseks on vaja, et tekstiline sisu oleks kergesti analüüsitav. Selleks tuleks andmed esmalt märgendada, et muuta tekstiline sisu paremini otsitavaks ja indekseeritavaks. Märgendama peaks olulised nimeüksused. Sellisteks nimeüksusteks on näiteks isiku-, organisatsiooni- ja kohanimed. Käsitsi märgendamine on töömahukas ja aeganõudev. Automaatsete vahendite kasutamine nimeüksuste tuvastamiseks ja märgendamiseks võimaldaks lihtsamini koostada süstemaatilisteks uuringuteks vajalikku andmestikku.

Tartu Linnavolikogu koosolekute protokollide originaalid on hoiul Rahvusarhiivis[2]. Protokollid on transkribeeritud automaatselt Transkribuse[3] abiga. Protokollide tekstid on saadud piltidelt optilise märgituvastuse[4] kaudu. Selle protsessi käigus tekib erinevaid vigu ja saadud andmetes on palju müra. Protokollid on läbinud ühe transkriptsioonivigade paranduste vooru aine Arhiivpraktika (FLAJ.02.014) toimumise raames. Projekti EKKD-TA10[5] käigus tehti protokollidele täiendav transkriptsioonivigade paranduste voor. Selle projekti raames märgendati koosolekute protokollides ära ka suuremad tabelid. Protokollide osaline nimeüksuste märgendamine viidi läbi kahel korral aine Arhiivpraktika käigus ja samuti projekti EKKD-TA10 raames. Protokollide andmeid on täpsemini kirjeldatud peatükis 2.

1.2 Olemasolevad mudelid

Tehisintellekti vahendeid saab edukalt kasutada tekstide automaatsel töötlemisel, kuid nende rakendamisel esineb mitmeid keerukaid väljakutseid. Mohammed Aldeen jt [4] demonstreerisid oma töös `gpt3.5-turbo` ja `gpt4` mudelite võimekust tekstide automaatsel märgendamisel. See töö tõi esile, et `gpt` mudelitel on suur potentsiaal antud valdkonnas. `gpt` mudelid saavutasid head tulemused erinevate andmestike puhul. Autorid leidsid, et `gpt` mudelid märgendavad kõige paremini tekste, mis on internetis vabalt kättesaadavad. Kehvemad tulemused saadi emotsioonide ja küsimuste märgendamise jaoks loodud andmestikel. Carlos-Emiliano González-Gallardo jt [5] leidsid oma uuringus, et `gpt` mudelid ei saa ajaloolistest tekstidest nimeüksuste tuvastamise ja klassifitseerimisega hästi hakkama. Nende sõnul on tulemus kehv, sest suur osa digitaliseeritud ajaloolistest andmetest ei ole vabalt kättesaadavad ning seetõttu pole neid andmeid mudelite treenimisel kasutatud. Autorite arvates paraneks suurte keelemudelite arusaamine ajaloolistest dokumentidest, kui mudelite treenimisel oleks kasutusel rohkem ajaloolisi digitaliseeritud andmeid, kuid arvatakse, et sellega kaasneks risk, et mudelid hakkavad andma poliitiliselt mõjutatud vastuseid.

Käesolevas töös kasutatakse `BERT`-tüüpi keelemudeleid. Need on loomuliku keele töötluse mudelid, mida kasutatakse erinevate keeleülesannete lahendamiseks. `BERT` mudelid põhinevad mitmekihilisel kahesuunalisel transformer-arhitektuuril [6]. Need mudelid teeb eriliseks

[2] https://www.ra.ee/dgs/explorer.php?tid=260&iid=110250278620&tbn=1&lev=yes&lst=2&hash=899a98a48bd63555649c75b81167852e
[3] https://www.transkribus.org/
[4] https://et.wikipedia.org/wiki/Optiline_m%C3%A4rgituvastus
[5] https://www.etis.ee/Portal/Projects/Display/9035bf9a-f375-494b-9be6-335c06e260d8

kahesuunaline õppimisvõime, mis tähendab, et mudel suudab arvestada korraga sõnele eelnevate ja järgnevate sisendis olevate sõnedega [6]. BERT alusmudeli treenimisel kasutati kahte ülesannet [6]. Üheks ülesandeks oli lünga täitmiseks sobiva sõne leidmine ja teiseks ülesandeks järgmise lause ennustamine [6]. Alusmudel on mudel, mis on treenitud suure hulga andmete peal kasutades isejuhendatud õpet [7]. Selliseid mudeleid saab kohandada erinevate ülesannete lahendamiseks [7].

BERT alusmudelit saab muuta võimeliseks lahendama erinevaid keeleülesandeid. Üheks selliseks keeleülesandeks on näiteks tekstist nimeüksuste tuvastamine. Mudeli kohandamist uute ülesannete lahendamiseks nimetatakse peenhäälestamiseks [7]. Uue ülesande lahendamiseks ei pea BERT alusmudeli arhitektuuri palju muutma vaid piisab ühe väljundkihi lisamisest [6].

1.2.1 Töös kasutatavad olemasolevad mudelid

Antud töös kasutati protokollidest nimeüksuste tuvastamiseks järgmisi olemasolevaid mudeleid: EstBERT_NER[6], EstBERT_NER_v2[7] ja est-roberta-hist-ner[8]. EstBERT_NER on Hasan Tanvir jt [8] artikli raames valminud mudel. See mudel tuvastab tekstidest kolme nimeüksuse kategooriat: LOC ehk kohanimed, ORG ehk organisatsioonide nimed ja PER ehk isikunimed. EstBERT_NER mudeli loomisel kasutati andmestikku, mis koosnes 572-st Delfi ja Postimehe artiklist. Mudel saadi EstBERT mudeli peenhäälestamisel. Eesti keelel põhinev EstBERT mudel oli Hasan Tanvir jt [8] artikli põhitulemus. See mudel on sama arhitektuuriga kui BERT alusmudel ja selle treenimiseks kasutati samu ülesandeid, mida on kasutatud BERT alusmudeli treenimiseks. EstBERT mudeli valideerimiseks kasutati ülesannetena teksti morfoloogilist märgendamist, nimeüksuste tuvastamist ja teksti klassifitseerimist. Nimeüksuste tuvastamise valideerimisülesande käigus valmis EstBERT_NER mudel.

Kairit Sirtsi [9] artikli raames valmis EstBERT_NER_v2 mudel, mis tuvastab kokku 11 nimeüksuse kategooriat: DATE ehk kuupäevad, EVENT ehk sündmused, GPE ehk geopoliitilised üksused, LOC ehk asukohad, MONEY ehk rahalised väärtused, ORG ehk organisatsioonid, PER ehk isikud, PERCENT ehk protsendid, PROD ehk tooted, TIME ehk ajaväljendid ja TITLE ehk tiitlid. Sarnaselt EstBERT_NER mudelile kasutati selle mudeli loomisel alusena EstBERT mudelit. EstBERT_NER_v2 mudeli peenhäälestamiseks kasutati kahte andmestikku, millest

[6] https://huggingface.co/tartuNLP/EstBERT_NER
[7] https://huggingface.co/tartuNLP/EstBERT_NER_v2
[8] https://huggingface.co/tartuNLP/est-roberta-hist-ner

üks koosnes 572-st Delfi ja Postimehe artiklist. Sama andmestikku kasutati ka `EstBERT_NER` mudeli peenhäälestamiseks. Teiseks andmestikuks oli uus uudiste artiklitest ja sotsiaalmeedia tekstidest koosnev andmestik, mis valmis Kairit Sirtsi artikli raames.

Siim Orasmaa jt [10] artikli käigus valminud `est-roberta-hist-ner` mudel on mõeldud vanast kirjakeelest nimeüksuste tuvastamiseks. `est-roberta-hist-ner` mudel on üks kolmest mudelist, mis kohandati 19. sajandi vallakohtu protokollide peal. Antud mudeli kohandamisel võeti aluseks `Est-RoBERTa` mudel. 19. sajandi vallakohtu protokollide NER mudel märgendab järgmisi kategooriaid: LOC ehk kohanimed, PER ehk isikunimed, ORG ehk organisatsioonide nimed, LOC_ORG ehk kohanimed, mis viitavad asukoha inimestele või organisatsioonidele, ja MISC ehk harvaesinevad nimeüksused.

1.3 Töös kasutatud masinõppe meetodid

Antud töös kasutati nii juhendatud kui ka nõrgalt juhendatud masinõpet. Juhendatud masinõppe korral kasutatakse mudeli loomiseks käsitsi märgendatud kvaliteetseid andmeid [1]. Juhendatud masinõpet on kasutatud näiteks `EstBERT_NER`, `EstBERT_NER_v2` ja 19. sajandi vallakohtu protokollide NER mudeli loomisel.

Nõrgalt juhendatud masinõpet on kirjeldatud Pierre Lison jt [11] artiklis kui masinõppe meetodit, mille puhul pole vaja mudeli loomiseks suurt käsitsi märgendatud andmehulka. Selle meetodi korral kasutatakse mitmeid automaatseid vahendeid, et märgendada automaatselt suur hulk märgendamata andmeid. Erinevate vahendite väljundid ühtlustatakse ja saadud madalama kvaliteediga andmeid kasutatakse uue mudeli loomiseks. Nõrgalt juhendatud masinõpe aitab kasutada erinevate valdkondade jaoks loodud automaatseid vahendeid, et luua mudel valdkonna jaoks, kus käsitsi märgendatud andmeid on vähe või üldse puuduvad.

Nõrgalt juhendatud masinõppes kasutatakse tihti õpetaja-õppijamudeli meetodit. Kõigepealt treenitakse õpetajamudel käsitsi märgendatud algandmetel [12]. Saadud mudel genereerib suurele hulgale märgendamata andmetele märgendid ja märgendite tõenäosused [12]. Õpetajamudeli poolt märgendatud andmete peal treenitakse õppijamudel [12]. Tõenäosusi saab kasutada, et õppijamudelit treenida vaid kindla osa andmete peal [12]. Sellist õpetaja-õppijamudeli meetodit kasutatakse ka käesolevas töös. Osades teadustöödes täiendatakse seda protsessi veel teiste erinevate võtetega.

2. Andmed

Käesoleva töö aluseks olid ajaloolised Tartu linnavolikogu protokollid JSON-failidena. Täpsemalt kujutasid need endast EstNLTK teegi tekstiobjekte[13], mis olid salvestatud JSON-kujule. Tekstiobjektidele oli lisatud ka algeline sõnestuse ja lausestuse kiht. See tähendab, et protokollide tekstid olid jaotatud sõnedeks ja lauseteks. Algelise sõnestuse ja lausestuse jaoks oli kasutatud eelkõige tänapäeva kirjakeele analüüsimiseks mõeldud EstNLTK teegi vahendeid, mis olid kohendatud projekti EKKD-TA10 raames. Antud töö ajal oli sõnestuste ja lausestuse kohendamine EKKD-TA10 raames veel pooleli ja seetõttu esines protokollide sõnestuses ja lausestuses vigu. Üheks suureks probleemiks olid punktiga lõppevad lühendid. Autor tegeles töö jooksul kolme erineva andmehulgaga. Neid kirjeldatakse täpsemalt kolmes järgnevas alapeatükis.

2.1 Katsemärgendusandmestik

Esimeseks andmehulgaks oli kahel korral toimunud aine Arhiivpraktika (FLAJ.02.014) käigus katsemärgendusena valminud andmestik, mis koosnes 125-st protokollist ehk 125-st JSON-failist. Mõlemal toimumiskorral oli kasutusel erinev juhend. Märgendusjuhised polnud eriti detailsed ja süstemaatilised. Mõlemal aine toimumiskorral võttis sellest osa 15-20 tudengit. Erinevate märgendajate märgendusi pole ühtlustatud. Seda andmestikku nimetatakse käesolevas töös katsemärgendusandmestikuks, sest märgenduste kvaliteet oli ebaühtlane.

Antud andmestikus märgendasid tudengid kümmet erinevat nimeüksuse kategooriat: *address* ehk aadressid, *party* ehk poliitiliste erakondade nimed, *event* ehk sündmuste nimed, *date* ehk kuupäevad, *organization* ehk organisatsioonide nimed, *school* ehk õppeasutuste nimed, *person* ehk isikunimed, *work* ehk ametinimetused, *place* ehk kohanimed ja *unclear* ehk ebaselged märgendid. Aine juhendites oli välja toodud vaid kolm kategooriat nimeüksuste märgendamiseks: *place*, *person* ja *organization*. Ülejäänud kategooriad lisasid tudengid ise. Nende lisatud kategooriate seletused on kirjutatud hiljem antud töö autori poolt ja ei pruugi olla päris korrektsed.

Tabel 1. Nimeüksuste jaotus katsemärgendusandmestikus

Kategooria	Protokollides esinemiste arv	Unikaalseid väärtusi
person	14095	6303
organization	10671	3161
place	4687	1964
date	1458	1128
address	388	281
party	109	86
school	68	53
event	18	17
work	1	1
unclear	1	1

Tabelis 1 on toodud katsemärgendusena valminud andmestiku nimeüksuste jaotus ja iga kategooria unikaalsete väärtuste arv. Kõige rohkem oli andmestikus *person*, *organization* ja *place* nimeüksuse märgendeid. Andmestikus oli kokku 564946 sõne, 37299 lauset ja 31496 käsitsi märgendatud nimeüksust.

2.2 Kuldstandardandmestik

Teiseks andmestikuks oli projekti EKKD-TA10 raames valminud andmestik. See koosnes seitsmest linnavolikogu protokollist, mis olid märgendatud ühe lingvistikatudengi poolt. Nimeüksuste märgendamisprotsessi jaoks oli ajaloolaste ja lingvistide koostöös valminud esmased märgendusjuhised. Antud töö ajal käis veel märgendusjuhiste väljatöötamine ja korpuse loomine oli pooleli. Seetõttu oli see andmestik väike ja polnud perfektse kvaliteediga. Antud töös nimetatakse seda andmestikku kuldstandardandmestikuks, sest võrreldes katsemärgendusena valminud andmestikuga oli selle andmestiku kvaliteet parem.

Kuldstandardandmestikus märgendati üksteist erinevat nimeüksuse kategooriat: PER ehk isikunimed, POSITION ehk isikute ametid või rollid, LOC ehk kohanimed, LOC_ADDRESS ehk aadressid, ORG ehk organisatsioonide nimed, ORG_POL ehk poliitilised organisatsioonid, ORG_GPE ehk geopoliitiline üksus, nt riik või kohalik omavalitsus/linnavalitsus, EVENT ehk nimelised sündmused ja tähtpäevad, LAW ehk seaduste (täis)nimed, sh viited seaduste paragrahvidele, MONEY ehk rahasummad koos ühikutega ning UNK ehk teadmata või raskesti

määratav nimekategooria. Nimeüksuste kategooriate selgitused on võetud märgendamisel kasutatud märgendusjuhistest[9].

Tabel 2. Nimeüksuste jaotus kuldstandardandmestikus

Kategooria	Protokollides esinemiste arv	Unikaalseid väärtusi
ORG	1428	594
PER	1073	695
MONEY	624	345
LOC_ADDRESS	354	274
LOC	327	143
POSITION	316	78
LAW	166	134
UNK	8	6
ORG_GPE	3	3
ORG_POL	2	2
EVENT	2	1

Tabelis 2 on näha, kui palju esines kuldstandardandmestikus nimeüksusi eri kategooriate kaupa ja kui palju neist oli unikaalseid väärtusi. Kõige rohkem oli selles andmestikus organisatsiooni nimede, isikunimede ja rahasummade märgendeid. Andmestikus oli kokku 51246 sõne, 3279 lauset ja 4303 nimeüksust.

2.3 Kõikidest protokollidest koosnev andmestik

Kolmas andmestik koosnes kõikidest Tartu Linnavolikogu 1918.–1940. a koosolekute protokollidest. Kokku oli 338 protokolli. See andmestik sisaldas 125 katsemärgendusena märgendatud protokolli. Ülejäänud protokollid olid märgendamata. Selles andmestikus olid ka projekti EKKD-TA10 raames märgendatud protokollide suuremad tabelid. Automaatse lausestuse ja sõnestusega oli selles andmestikus 145374 lauset ja 2188055 sõne.

2.4 Ettevalmistus masinõppeks

2.4.1 Andmestike ettevalmistus

Kuldstandardandmestikus tuli teha väiksemaid muudatusi. LOC-kategooria nimeüksuste arvu suurendamiseks üldistati LOC_ADDREESS nimeüksuse kategooria LOC-kategooriaks. See tuli kasuks olemasoleva mudeli kohandamisel linnavolikogu protokollide märgendamiseks.

[9] https://docs.google.com/document/d/1SDVivaCdUfFgqTUSoszAw4l-VSNzxZmO6hvN6r_h0rU/edit?tab=t.0

ORG_GPE-kategooriat kasutati olemasolevate mudelite hindamisel kui GPE-kategooriat, aga mudeli kohandamisel linnavolikogu protokollide märgendamiseks kasutati seda nimeüksuse kategooriat kui ORG-kategooriat. Kõik kuldstandardandmestiku nimeüksuste kategooriate nimede muudatused on toodud tabelis 3.

Tabel 3. Nimeüksuste kategooriate nimede muudatused kuldstandardandmestikus

Algne kategooria nimi	**Muudetud kategooria nimi**
LOC_ADDRESS	LOC
POSITION	TITLE
ORG_GPE	GPE/ORG
ORG_POL	ORG

Katsemärgendusena valminud andmestikus kasutatud nimeüksuste kategooriate nimed erinesid sellest, mida kasutati olemasolevate mudelite poolt. Selleks, et selle andmestiku peal saaks hinnata olemasolevaid mudeleid, muudeti osade andmestikus kasutatud kategooriate nimesid. Lisaks tehti muudatusi, mis sarnanesid kuldstandardandmestikus tehtud muudatustele. Tabelis 4 on välja toodud kategooriate nimede muutused. Tabelist puuduvad kategooriad, mille nimesid polnud vaja muuta, sest neid ei kasutatud.

Tabel 4. Nimeüksuste kategooriate nimede muudatused katsemärgendusena valminud andmestikus

Algne kategooria nimi	**Muudetud kategooria nimi**
person	PER
organization	ORG
place	LOC
date	DATE
event	EVENT
address	LOC
party	ORG

Töö käigus eemaldati kõikidest protokollidest koosnevast andmestikust protokollid, mis esinesid ka kuldstandardandmestikus. Kuldstandardandmestikku kasutati mudelite hindamiseks. Kõikidest protokollidest koosnevat andmestikku kasutati mudelite treenimiseks. Seega tuli vältida samade protokollide sattumist treenimis- ja valideerimisandmete hulka. Lisaks eemaldati protokollidest suuremad tabelid, mis olid varasemalt märgendatud. Selle tulemusena jäi andmestikku alles 110766 lauset ja 1869711 sõnet.

2.4.2 BIO-kuju

Keeletehnoloogias kasutatakse BIO-kuju sageli selleks, et esitada fraaside ja sõnade märgendust mudelile arusaadaval kujul. Kogu tekst on jaotatud sõnedeks ja igale sõnele on omistatud märgend. Iga uuritava nimeüksuse kategooria jaoks on algusmärgend ja keskosa/lõpuosa märgend. Lisaks on O-märgend, mis tähistab, et sõne ei kuulu ühtegi uuritavasse kategooriasse. Algusmärgendi jaoks pannakse nimeüksuse kategooria ette B ehk *beginning* ja keskosa või lõpuosa jaoks pannakse kategooria ette I ehk *inside* [1].

Volikogu juhatab esimees [Lui Olesk].
PER

[Volikogu][juhatab][esimees][Lui][Olesk][.]
O O O B-PER I-PER O

Joonis 1. BIO-kuju näide

Joonisel 1 on toodud näide kahest lausest. Ülemisel lausel on märgendatud lauses olev nimi PER kategooriaga. Alumine lause on sama lause BIO-kujul. Eesnimi on märgendatud B-PER märgendiga ja perekonnanimi on märgendatud I-PER märgendiga. Ülejäänud sõnad on märgendatud O-märgendiga.

2.4.3 Ristvalideerimine

Masinõppes kasutatakse vähese andmehulga korral tihti ristvalideerimist. Ristvalideerimine on andmete kasutamise ja mudeli loomise protsess, mille käigus jaotatakse andmestik mitmel korral erineval moel treenimis- ja valideerimishulgaks. Protsessi eesmärk on tagada, et iga andmepunkt esineks korra valideerimisandmete hulgas. Tänu sellele osalevad ka haruldased näited mudeli headuse hindamisel. Suurem valideerimistulemuste hulk aitab suurendada statistilist tugevust. Ristvalideerimise protsessi jooksul luuakse igal sammul uus mudel. Tavaliselt jaotatakse ristvalideerimise käigus andmed juhuslikult K gruppi ja igal sammul kasutatakse ühte gruppi mudeli valideerimiseks ja ülejäänud gruppe mudeli treenimiseks. Protsessi lõpus on võimalik arvutada erinevaid statistikuid, mis aitavad hinnata probleemi lahendamiseks kasutatud lähenemist ja peegeldavad selle võimet üldistuda uutele andmetele. [7]

Kuldstandardandmete vähesuse tõttu otsustati käesolevas töös kasutada sellel andmestikul ristvalideerimist. Andmestik tuli protsessiks ette valmistada. Tavaliselt jaotatakse andmed gruppidesse juhuslikult. Antud olukorras tehti seda käsitsi märgendite arvu põhjal[10], sest failide nimeüksuste märgendite arvud ja failide pikkused olid erinevad. Lisaks ei saanud protokolle antud juhul jaotada osadeks. Kui üks osa protokollist oleks treenimishulgas ja teine osa samast protokollist oleks valideerimishulgas, siis oleksid ristvalideerimisega saadud mudelid omandanud ebaõiglased eelised, sest samad nimeüksused oleksid tõenäoliselt esinenud mõlemas sama protokolli osas.

Tabel 5. Ristvalideerimise grupid

	1922-04-24, 1936-09-07	**1927-03-28, 1941-01-03**	**1932-01-25**	**1934-10-15**	**1935-09-30**
Sõnesid	9804	7927	12795	8981	11739
Lauseid	591	599	841	522	716
LOC	106	27	244	151	153
PER	130	197	388	186	172
ORG	299	317	265	284	268
Märgendeid kokku	535	541	897	621	593

Tabelis 5 on näidatud töös kasutatud ristvalideerimise gruppide täpsed andmed. Kuupäevad tabeli päises tähistavad nendel kuupäevadel koostatud protokolle.

[10]Arvestati ainult LOC, PER ja ORG kategooria nimeüksuseid.

3. Olemasolevate nimeüksusi tuvastavate mudelite analüüs

Kahele Tartu Linnavolikogu protokollide andmestikule rakendati kolme olemasolevat nimeüksusi tuvastavat mudelit. Saadud tulemusi võrreldi andmestikes olevate käsitsi märgendatud nimeüksustega. Esimene andmestik oli märgendatud tudengite poolt katsemärgendusena. Selles andmestikus olid nimeüksused märgendatud kohati ebaühtlaselt, sest tudengid märgendasid protokolle ilma detailsete ja süstemaatiliste märgendusjuhisteta. Teine andmestik oli kuldstandardandmestik, mis valmis projekti EKKD-TA10 käigus. See andmestik on palju väiksem, aga oli märgendatud ühtlasemalt ja vigu esines vähem. Mõlemat andmestikku on kirjeldatud täpsemalt peatükis 2.

Tulemuste hindamiseks kasutati abivahendina programmeerimiskeele Python jaoks loodud Nervaluate[11] teeki. See teek on mõeldud nimeüksuste tuvastamise mudelite hindamiseks. Nervaluate teegis on erineva rangusega hindajad. Antud töös uuriti ainult kõige rangemaid tulemusi ehk *strict* tulemusi. See tähendab, et uuritava märgendi piirid ja kategooria pidid kattuma käsitsi märgendatud märgendi piiride ja kategooriaga.

Tulemuste hindamisel kasutati täpsust, saagist ja F1-skoori. Täpsus näitab, kui suur osa kõikidest mudeli poolt pakutud märgenditest olid õiged [7]. Saagis näitab, kui suure osa õigetest väärtustest mudel üles leidis [7]. F1-skoor on täpsuse ja saagise harmooniline keskmine [7]. Iga uuritava mudeli tulemuste kohta koostatud tabelis on välja toodud peale nimeüksuste kategooriate tulemuste ka koondskoorid, mille arvutamisel loeti positiivseteks juhtudeks iga kategooria positiivsed juhud.

Mudelite tulemuste hindamise juures hinnati kategooriaid, millele leidus sarnane vaste käsitsi märgendatud andmestikes. Osade kategooriate tulemused võisid olla mõjutatud sellest, et mudeli ja käsitsi märgendatud andmestiku kategooriate definitsioonid polnud päris samad. Nimeüksuste kategooriate definitsioonid pole universaalselt defineeritud ja märgendamisjuhised on tavaliselt loodud kindla valdkonna andmete jaoks.

[11] https://github.com/MantisAI/nervaluate

3.1 `EstBERT_NER` mudeli tulemused

`EstBERT_NER` mudel tuvastas kahel andmestikul kolme nimeüksuse kategooriat: LOC, ORG ja PER. Antud mudel on loodud tänapäevase ajakirjanduse tekstide peal [8].

Tabel 6. `EstBERT_NER` tulemused katsemärgendusena valminud andmestikul

	Täpsus	**Saagis**	**F1-skoor**
LOC	0.2688	0.2629	0.2658
ORG	0.4059	0.2696	0.324
PER	0.4656	0.5544	0.5061
Koondskoor	0.417	0.4025	0.4096

Tabel 7. `EstBERT_NER` tulemused kuldstandardandmestikul

	Täpsus	**Saagis**	**F1-skoor**
LOC	0.3524	0.2261	0.2755
ORG	0.6704	0.3811	0.486
PER	0.7364	0.8324	0.7815
Koondskoor	0.6465	0.5002	0.564

Tabelis 6 ja tabelis 7 on välja toodud `EstBERT_NER` mudeli tulemused kahel käsitsi märgendatud andmestikul. Tudengite poolt katsemärgendusena valminud andmestikul olid tulemused madalamad kui kuldstandardandmestikul. Mõlemal andmestikul märgendas `EstBERT_NER` mudel kõige paremini isikunimede kategooriat. Organisatsioonide nimede kategooria tulemused kuldstandardandmestikul paranesid, kohanimede kategooria tulemuste täpsus paranes, kuid saagis langes.

3.2 `EstBERT_NER_v2` mudeli tulemused

`EstBERT_NER_v2` mudel tuvastab kokku 11 erinevat nimeüksuse kategooriat. Käsitsi märgendatud andmestikes polnud neile kõigile vasteid. Antud alapeatükis kasutati kokku kaheksat erinevat nimeüksuse kategooriat. Katsemärgendusandmestikul uuriti LOC, ORG, PER, DATE, EVENT kategooriaid ja kuldstandardandmestikul LOC, ORG, PER, EVENT, TITLE, GPE, MONEY kategooriaid. `EstBERT_NER_v2` mudel on loodud tänapäeva ajakirjanduse ja sotsiaalmeedia valdkonna andmetel [9]. Katsemärgendusandmestikus olev *work*-kategooria sarnanes olemuselt `EstBERT_NER_v2` mudeli TITLE-kategooriale, aga katsemärgendusandmestikus esines ainult üks *work*-kategooria nimeüksus ja mudel märgendas

TITLE-kategooriaid palju rohkem. Seega katsemärgendusandmestikul TITLE-kategooriat ei hinnatud.

Tabel 8. `EstBERT_NER_v2` tulemused katsemärgendusena valminud andmestikul

	Täpsus	**Saagis**	**F1-skoor**
LOC	0.1899	0.1066	0.1365
ORG	0.3195	0.37	0.3429
PER	0.38	0.4443	0.4097
DATE	0.0478	0.2126	0.078
EVENT	0.0116	0.0556	0.0192
Koondskoor	0.2892	0.3533	0.3181

Tabel 9. `EstBERT_NER_v2` tulemused kuldstandardandmestikul

	Täpsus	**Saagis**	**F1-skoor**
LOC	0.1682	0.1057	0.1298
ORG	0.669	0.4647	0.5485
PER	0.7206	0.6985	0.7094
EVENT	0.0	0.0	0.0
TITLE	0.3579	0.3386	0.348
GPE/ORG_GPE	0.0667	1.0	0.125
MONEY	0.2127	0.1074	0.1427
Koondskoor	0.5316	0.4031	0.4585

`EstBERT_NER_v2` mudeli tulemused käsitsi märgendatud andmestikel on toodud tabelis 8 ja tabelis 9. Neil kahel andmestikul hinnatud kategooriatest kattusid ORG, PER, LOC ja EVENT nimeüksuse kategooriad. Esimese kahe kategooria puhul olid tulemused kuldstandardandmestikul paremad kui katsemärgendusena valminud andmestikul. PER ja ORG kategooriatel paranesid tulemused oluliselt. LOC kategooria tulemused palju ei muutunud. EVENT kategooria tulemused olid katsemärgendusena märgendatud andmestikul halvad ja kuldstandardandmestikul olid kõik hindamiseks kasutatud näitajate tulemused nullid. Katsemärgendusandmestikul uuriti lisaks veel ka DATE kategooriat. Selle kategooria tulemused olid kehvad. Kuldstandardandmestikul uuriti lisaks TITLE, GPE ja MONEY kategooriaid. Nende tulemused F1-skoori järgi polnud samuti head. GPE kategooria korral leidis `EstBERT_NER_v2` mudel kõik andmestikus olevad väärtused üles, kuid mudeli täpsus oli väga madal. See näitab, et mudel pakkus seda kategooriat üle. Põhjuseks võib olla GPE kategooria definitsioonide erinevus.

3.3 19. saj vallakohtu protokollide NER mudel

Siim Orasmaa jt [10] artikli käigus valminud `est-roberta-hist-ner` ehk 19. saj vallakohtu protokollide NER mudel suudab tuvastada viit nimeüksuse kategooriat. Kahel käsitsi märgendatud andmestikul uuriti neist nelja kategooriat: LOC, ORG, PER ja LOC_ORG. Kategooria LOC_ORG muudeti GPE kategooriaks. Siim Orasmaa jt [10] artiklis oli mainitud nimeüksuse kategooriate seletuse juures, et LOC_ORG kategooria sarnaneb *Geo-Political Entity* ehk GPE kategooriale. Valdkondade erinevuse tõttu võivad kategooriate definitsioonid erineda.

Tabel 10. `est-roberta-hist-ner` mudeli tulemused katsemärgendusena valminud andmestikul

	Täpsus	**Saagis**	**F1-skoor**
LOC	0.2138	0.1139	0.1486
ORG	0.2808	0.0881	0.1342
PER	0.5222	0.654	0.5807
Koondskoor	0.4527	0.3588	0.4003

Tabel 11. `est-roberta-hist-ner` mudeli tulemused kuldstandardandmestikul

	Täpsus	**Saagis**	**F1-skoor**
LOC	0.1561	0.1232	0.1377
ORG	0.3663	0.1035	0.1614
PER	0.6609	0.7807	0.7158
LOC_ORG/GPE	0.0	0.0	0.0
Koondskoor	0.4712	0.3359	0.3922

Tabelis 10 ja tabelis 11 on toodud 19. saj vallakohtu protokollide NER mudeli tulemused käsitsi märgendatud andmestikel. PER kategooria tulemused olid kuldstandardandmestikul oluliselt paremad kui katsemärgendusandmestikul. LOC ja ORG kategooria tulemused olid mõlemal andmestikul sarnased. Mudel hindas mõlemal andmestikul kõige paremini PER kategooriat. Kuldstandardandmestikul hinnati lisaks ka LOC_ORG/GPE kategooriat. Kõigi hindamiseks kasutatud näitajate tulemused olid nullid.

3.4 Tulemuste kokkuvõte

Antud peatükis rakendati kolme nimeüksuste tuvastamise mudelit kahele käsitsi märgendatud andmestikule. `EstBERT_NER`, `EstBERT_NER_v2` ja `est-roberta-hist-ner` mudelite tulemused on välja toodud eespool olevates alapeatükkides. Mudelite tulemuste omavaheliseks võrdlemiseks saab kasutada LOC, PER ja ORG nimeüksuste kategooriaid. Kõik kolm mudelit märgendasid neid kategooriaid mõlemal andmestikul. Kui arvestada ainult neid kategooriaid, siis olid tulemused kuldstandardandmestikul enamasti paremad kui katsemärgendusandmestikul. Katsemärgendusandmestikul polnud märgendamine piisavalt süstemaatiline. Nimeüksuse kategooriaid LOC ja PER märgendas F1-skoori järgi kõige paremini `EstBERT_NER` mudel. ORG kategooriat märgendas kõige paremini `EstBERT_NER_v2` mudel. Parima koondskooriga oli `EstBERT_NER` mudel.

Parima koondskoori ja parima LOC ja ORG kategooria tulemuste tõttu otsustati antud töös valida vaadeldud kolmest olemasolevast mudelist välja `EstBERT_NER` mudel, et seda edaspidises töös kohandada Tartu Linnavolikogu protokollidest nimeüksuste paremaks tuvastamiseks. Selle mudeli kohandamise protsessi kirjeldatakse järgmises peatükis.

4. Olemasoleva mudeli kohandamine protokollide märgendamiseks

Uue mudeli nullist treenimise asemel otsustati antud töös proovida olemasoleva mudeli kohandamist ajaloolistele linnavolikogu protokollidele. Töö autoril oli kasutada kaks käsitsi märgendatud andmestikku. Katsemärgendusena valminud andmestik oli märgendatud ebaühtlaselt ja kuldstandardandmestik oli väike. Eelmises peatükis võrreldi kolme olemasolevat nimeüksuste tuvastamise mudelit. Edasise töö aluseks otsustati võtta `EstBERT_NER` mudel, kuna käsitsi märgendatud andmestikel sai see mudel parimad tulemused. See mudel märgendab kolme nimeüksuste kategooriat: LOC ehk kohanimed, ORG ehk organisatsioonide nimed ja PER ehk isikunimed. Neid kategooriaid esines ka kõige rohkem käsitsi märgendatud andmestikes[12]. Kohandatava mudeli eesmärgiks seati protokollidest LOC, PER ja ORG nimeüksuse kategooriate märgendamine. Alusandmestikes kasutati peatükis 2.4.1 tehtud muudatusi ja üldistusi.

Andmetöötlust ja mudelite treenimist teostas autor nii tavaarvutil kui ka Tartu Ülikooli Teadusarvutuse Keskuse arvutusklastril Rocket [14]. Väikse hulga mudelite treenimiseks ja väheses mahus andmetöötluseks piisas tavaarvuti võimsusest. Töö käigus oli vaja teha ka suurel hulgal andmetöötlust ja mudelite treenimist. Selle juures oli suureks abiks Tartu Ülikooli Teadusarvutuse Keskuse arvutusklastril võimsus.

Autor otsustas proovida olemasoleva mudeli kohandamiseks kahte erinevat masinõppe meetodit. Nende mõlema puhul kasutati abivahendina Simple Transformers[13] teeki. See tegi mudeli kohandamise lihtsamaks. Täpsemat tööprotsessi kirjeldatakse järgnevates alapeatükkides.

4.1 Esimene meetod

Üheks variandiks oli mudeli kohandamisel aluseks võtta kuldstandardandmestik. Selle andmestiku andmete vähesuse tõttu otsustas autor kasutada ristvalideerimist ja nõrgalt juhendatud masinõpet. Kasutati õpetaja-õppijamudeli meetodit, mida on täpsemalt kirjeldatud töö peatükis 1.3. Alustuseks tuli ristvalideerimiseks loodud andmestike alusel luua õpetajamudelid, mis jäljendaksid kuldstandardandmestiku märgendamisstiili. Neid mudeleid kasutati märgendamata linnavolikogu protokollide märgendamiseks. Nii suurenes oluliselt andmete hulk, mida sai kasutada õppijamudelite loomiseks. Saadud andmed polnud aga

[12] Kuldstandard andmesiku LOC_ADDRESS-kategooria arvestati LOC-kategooriana.
[13] https://simpletransformers.ai/docs/ner-model/

enam kuldstandardandmestikuga võrreldes sama kvaliteediga. Järgnevas kahes alapeatükis on kirjeldatud õppetaja- ja õppijamudelite loomist täpsemalt.

4.1.1 Õpetajamudel

Õpetajamudeli loomisel kasutati olemasoleva mudeli peenhäälestamist. Aluseks võeti `EstBERT` mudel, mida õpetati kuldstandardandmestiku ristvalideerimise andmestikel nimeüksusi märgendama. `EstBERT` mudel varem seda ei osanud. Õpetajamudeli aluseks ei võetud `EstBERT_NER` mudelit, kuna kuldstandardandmeid oli vähe ja arvatavasti oleks `EstBERT_NER` mudeli loomiseks kasutatud tänapäeva keele andmete mõju olnud liiga suur. Antud töös peenhäälestati kokku viis õpetajamudelit. Iga õpetajamudeli loomisel võeti viis ristvalideerimise gruppi, millest nelja kasutati peenhäälestamiseks ja ühte valideerimiseks.

`EstBERT` mudeli peenhäälestamise jaoks tuli andmed viia EstNLTK tekstiobjekti kujult uuele kujule, et andmestik koosneks sõnedest, nendele vastavatest BIO-kujul märgenditest ja lause järjekorranumbritest.

Tartu Ülikooli Teadusarvutuse Keskuse arvutusklastril kulus Nvidia V100 GPU-l ühe `EstBERT` mudeli peenhäälestamiseks umbes 24 sekundit. Saadud abimudelid rakendati varasemalt muudetud andmestikule, mis koosnes kõikidest märgendamata linnavolikogu protokollidest. Selles andmestikus oli kokku 331 protokolli. Märgendamise tulemusena tekkis andmestikus olevatele EstNLTK-teegi tekstiobjektidele uus kiht, kus igale sõnele vastas sõne BIO-kujul olev märgend ja mudeli poolt antud märgendi tõenäosus. Saadud andmetega oli võimalik hakata katsetama erinevaid lõpliku mudeli variante.

4.1.2 Õppijamudel

Õppijamudeli aluseks võeti nimeüksusi märgendav `EstBERT_NER` mudel, mida peenhäälestati edasi kasutades abimudelite poolt tekitatud madalama kvaliteediga andmeid. Õpetajamudelite poolt märgendatud protokollides esines vigu ja seega ei saanud kasutada saadud andmestikke tervenisti õppijamudelite loomiseks. Andmestikest otsustati eraldada laused, kus iga sõne märgendi tõenäosus oli kõrgem kindlast piirist. Autor katsetas erinevaid tõenäosuse lävendeid. Lävenditeks olid 60%, 65%, 70%, 75%, 80%, 85%, 90% ja 95%. Lisatingimuseks oli võetud, et lauses peab olema rohkem kui neli sõne. Sellega sooviti vältida lühikesi lauseid, mis võisid olla tekkinud automaatse lausestuse vigade tõttu. Sobivatest lausetest tehti eraldi andmestikud.

Erinevate lävendite andmestikud jaotati omakorda juhuslikult alamandmestikeks LOC, PER ja ORG kategooria nimeüksuste arvu põhjal. Esmalt alustati 6500 nimeüksusest. Seda arvu

hakati suurendama järjest 6500 võrra, kuni arv oli suurem kui antud lävendi andmestikus olevate nimeüksuste koguarv. Autor katsetas sellist lähenemist, et leida optimaalne nimeüksuste arv ja vältida ülesobitumist. Iga saadud andmestiku peal kohandati `EstBERT_NER` mudelit.

Viie ristvalideerimisandmestiku, kaheksa lävendi ja erinevate märgendite arvudega andmestike peal kohandati kokku 91 `EstBERT_NER` mudelit. Parimate mudelite tulemused on toodud välja alapeatükis 4.3.

4.2 Teine meetod

Antud töös kasutati teise võimalusena `EstBERT_NER` mudeli kohandamiseks katsemärgendusena valminud andmestikku. Selles andmestikus oli käsitsi märgendatud protokolle märgatavalt rohkem kui kuldstandardandmestikus olevaid protokolle. Katsemärgendusena valminud andmestiku probleem oli aga selles, et protokolle oli märgendatud ilma detailsete ja süstemaatiliste märgendusjuhisteta. Selle tulemusena olid nimeüksused märgendatud ebaühtlaselt. Autor otsustas katsetada selle andmestiku peal juhendatud masinõpet, sest andmeid oli selleks piisavalt.

Katsemärgendusena valminud andmestikust eemaldati kuldstandardandmestikuga kattuvad protokollid, neid kasutati hiljem mudeli hindamiseks. Autor otsustas kasutada olemasoleva mudeli kohandamiseks ainult lauseid, kus leidus vähemalt üks LOC, PER või ORG nimeüksus. Selleks, et `EstBERT_NER` mudelit Simple Transformers teegi abil edasi peenhäälestada, viidi antud laused uuele kujule. Uus andmestik koostati sõnedest, BIO-kujul olevatest märgenditest ning lause järjekorranumbritest. Sellest andmestikust kasutati 80% andmeid `EstBERT_NER` mudeli kohandamiseks ja 20% selle valideerimiseks. Antud treenimis- ja valideerimisandmestikku on kirjeldatud tabelis 12.

Tabel 12. `EstBERT_NER` mudeli kohandamiseks kasutatud andmestik

	Treenimisandmestik	**Valideerimisandmestik**	**Kogu andmestik**
sõnesid	244032	59995	304027
lauseid	10419	2605	13024
LOC	3630	900	4530
PER	10323	2502	12825
ORG	7654	1931	9585
märgendeid kokku	21607	5333	26940

Arvutusklastril kulus `EstBERT_NER` mudeli kohandamiseks linnavolikogu protokollidele umbes 92 sekundit. Saadud mudeli tulemused hinnati kuldstandardandmestiku ristvalideerimise

andmestikel. Need on välja toodud järgmises alapeatükis. Edaspidi nimetatakse seda mudelit antud töös katsemärgendusmudeliks.

4.3 Kahe meetodi tulemused ja analüüs

`EstBERT_NER` mudeli kohandamisel kasutatud esimese meetodi puhul treeniti palju erinevaid mudeleid. Parima mudeli leidmiseks arvutati iga lävendi jaoks ristvalideerimise sammude F1-skooride keskmine. Iga ristvalideerimise sammu jaoks valiti üks mudel nii, et alustati 6500 märgendiga loodud mudelist ja liiguti nimeüksuste arvult järgmise mudeli juurde kuni F1-skoor enam ei paranenud. Tulemused on toodud tabelis 13. Sellise lähenemisega saadi parimad tulemused 65% lävendiga lausetest koosnevate andmestiku peal loodud mudelitega. Võrdluseks on tabelis toodud ka õpetajamudelite ja `EstBERT_NER` F1-skooride keskmised samadel ristvalideerimise sammudel kasutatud valideerimisgruppidel.

Tabel 13. Lävendite õppijamudelite, õpetajamudelite ja `EstBERT_NER` F1-skooride keskmised ristvalideerimise gruppidel

Lävend/Mudel	**F1-skooride keskmine**
60%	0.6887
65%	0.6926
70%	0.6822
75%	0.657
80%	0.5934
85%	0.4353
90%	0.2241
95%	0.0513
Õpetajamudel	0.46364
`EstBERT_NER`	0.56850

Edasi uuriti õppijamudelitest ainult 65% lävendil põhinevaid mudeleid. Iga sellise õppijamudeli jaoks arvutati ristvalideerimise iteratsioonide F1-skooride keskmine ja standardhälve. Tulemused on tabelis 14. Tabelis toodud õppijamudelite nimede juures näitab alakriipsuga eraldatud arv mudeli kohandamiseks kasutatud nimeüksuste märgendite arvu. Parima keskmise koondskooriga oli `mudel_32500`. Tabelis on võrdluseks toodud ka `EstBERT_NER` mudeli ja õpetajamudelite tulemused.

Tabel 14. 65% lävendi mudelite tulemused

Mudeli nimi	F1-skooride keskmine	Standardhälve
mudel_6500	0.64404	0.088450
mudel_13000	0.67610	0.078126
mudel_19500	0.68178	0.075796
mudel_26000	0.68694	0.067673
mudel_32500	0.68738	0.073589
EstBERT_NER	0.56850	0.104196
Õpetajamudel	0.46364	0.062291

Kuldstandardandmestiku ristvalideerimise andmestike peal hinnati kokku nelja mudelit: EstBERT_NER mudelit, õpetajamudeleid, esimese meetodi abil saadud tulemust mudel_32500 ja teise meetodi tulemusena saadud katsemärgendusmudelit. Nende nelja mudeli kuldstandardandmestikul hindamisel saadud koondskooride keskmised ja standardhälbed on tabelis 15. Täpsemad tulemused nimeüksuste kategooriate kaupa on toodud tabelis 16.

Tabel 15. Mudelite hindamisel saadud koondskooride keskmised ja standardhälbed

	EstBERT_NER	**Õpetajamudel**	**mudel_32500**	**Katsemärgendus-mudel**
Täpsus	0.6655±0.1476	0.4212±0.0833	0.6971±0.1055	0.6826±0.0896
Saagis	0.5000±0.0845	0.5229±0.0411	0.6807±0.0437	0.7040±0.0447
F1-skoor	0.5685±0.1042	0.4636±0.0623	0.6874±0.0736	0.6917±0.0630

Tabel 16. Mudelite hindamisel saadud kategooriate keskmised ja standardhälbed

	EstBERT_NER	**Õpetajamudel**	**mudel_32500**	**Katsemärgendus-mudel**
LOC täpsus	0.3967±0.1911	0.1498±0.0740	0.5202±0.1147	0.5682±0.0886
LOC saagis	0.2907±0.1602	0.2524±0.1014	0.4340±0.1540	0.6048±0.0681
LOC F1-skoor	0.3294±0.1613	0.1828±0.0752	0.4661±0.1328	0.5827±0.0609
PER täpsus	0.7791±0.2070	0.6064±0.1101	0.8012±0.1648	0.7401±0.1251
PER saagis	0.8594±0.1596	0.6622±0.1019	0.8599±0.1309	0.8394±0.0866
PER F1-skoor	0.8157±0.1885	0.6296±0.0920	0.8286±0.1493	0.7859±0.1085
ORG täpsus	0.6642±0.1534	0.4942±0.1161	0.6947±0.1141	0.6959±0.0849
ORG saagis	0.3883±0.1488	0.5392±0.0770	0.6543±0.1070	0.6675±0.0897
ORG F1-skoor	0.4843±0.1492	0.5135±0.0958	0.6728±0.1074	0.6798±0.0799

Kui analüüsida nimeüksuste kategooriaid eraldi, siis tabelis 16 toodud tulemuste põhjal sai LOC kategooria puhul parima F1-skoori katsemärgendusmudel. PER kategooria puhul oli parim mudel esimese meetodi tulemusena saadud `mudel_32500`, kuid `EstBERT_NER` mudeli tulemus oli sellest vaid natuke halvem. ORG kategooria puhul said `mudel_32500` ja katsemärgendusmudel sarnased tulemused.

Tabelis 15 toodud koondskoore võrreldes said mõlema autori poolt katsetatud meetodiga loodud mudelid sarnased tulemused. Nende mudelite tulemused olid paremad olemasoleva tänapäevase kirjakeele andmetel loodud `EstBERT_NER` mudeli tulemustest. Lisaks on tabelis tulemustest näha, et õpetaja-õppijamudeli meetodi kasutamisest oli kasu, sest õppijamudeli tulemused olid õpetajamudeli omadest märgatavalt paremad. Selle meetodiga oli võimalik luua väikest hulka kvaliteetseid algandmeid kasutades lõpptulemusena korralik mudel. Positiivseks üllatuseks olid ebaühtlase kvaliteediga andmete peal loodud katsemärgendusmudeli head tulemused.

Kokkuvõte

Käesolevas töös uuriti võimalusi nimeüksuste tuvastamiseks Tartu Linnavolikogu koosolekute protokollides, mis pärinevad aastatest 1918-1940. Nimetatud protokollid on väärtuslikud ajalooallikad möödunud sajandi esimesest poolest.

Töö aluseks oli 338 linnavolikogu protokolli. Neist 7 oli märgendatud süstemaatilise juhendiga projekti EKKD-TA10 raames, töös nimetatakse seda kuldstandardandmestikuks. 125 protokolli olid märgendatud aine Arhiivpraktika raames tudengite poolt ilma detailse juhendita ebaühtlaselt, töös nimetatud katsemärgendusandmestikuks. Ülejäänud protokollid olid märgendamata. Kokkuvõtteks võib öelda, et päris uue mudeli treenimiseks oli andmeid vähe või need olid ebaühtlase kvaliteediga. Ajalooliste protokollide masinõppe abil märgendamiseks otsustati kohandada mõni olemasolevatest nimeüksusi märgendavatest mudelitest. Analüüsiti järgmisi mudeleid: `EstBERT_NER`, `EstBERT_NER_v2` ja `est-roberta-hist-ner`. Neist kaks esimest on loodud tänapäeva keele andmetel ja kolmas 19. sajandi kirjakeelel.

Olemasolevate mudelite analüüsimiseks ja masinõppe kasutamiseks tuli töö käigus andmestikud korrastada. Käsitsi märgendatud andmestikes kasutatud nimeüksuste kategooriate nimed ühtlustati. Kuldstandardandmestikus oli andmeid vähe, seetõttu valmistati masinõppe jaoks antud andmestik ette ristvalideerimise kasutamiseks.

Edasi uuriti, milliseid tulemusi oli korrastatud andmetel võimalik saada mainitud kolme olemasoleva mudeliga. Kõige parema tulemuse andis `EstBERT_NER` mudel. Antud mudel otsustati võtta edasise töö aluseks ja kohandada see ajalooliste protokollide andmetele. Selleks kasutati juhendatud masinõpet ja nõrgalt juhendatud masinõpet. Viimase puhul rakendati õpetaja-õppijamudeli meetodit ja ristvalideerimist. Täpsemalt kirjeldatakse `EstBERT_NER` mudeli kohandamise protsessi peatükis 4.

Analüüsiti tulemusi, mida andsid kaks töös kasutatud olemasoleva mudeli kohandamise meetodit. Kuldstandardandmestikul nõrgalt juhendatud masinõppe meetodi rakendamisel saadud tulemustest oli näha, et õpetaja-õppijamudeli lähenemisest oli kasu, sest õppijamudeli tulemused olid märgatavalt paremad õpetajamudeli omadest. Ebaühtlase katsemärgendusandmestiku peal juhendatud masinõpet kasutades kohandatud mudel sai samuti head tulemused. Mõlema katsetatud lähenemise korral olid kohandatud `EstBERT_NER` mudeli tulemused ajalooliste protokollide märgendamisel paremad kui muutmata `EstBERT_NER` mudeli puhul, mis on loodud tänapäeva keelel.

Veelgi paremate tulemuste saamiseks oleks vaja rohkem ja kvaliteetsemaid algandmeid. Antud töö kirjutamise hetkel oli pooleli projekti EKKD-TA10 raames Tartu Linnavolikogu 1918.-1940. a protokollide märgendamine ja seetõttu oli kasutatav kvaliteetne andmestik väike.

Käesoleva töö käigus kohandatud mudelit `EstBERT_NER` saab kasutada edaspidi ajalooliste Tartu Linnavolikogu protokollide märgendajate töö aluseks.

Tulevikus on võimalik suurema hulga kvaliteetsete andmete peal treenida ajaloolisel keelel uus nimeüksusi tuvastav mudel. See tuleks kasuks väärtuslike ajalooliste dokumentide automaatsel märgendamisel.

Viited

[1] Jurafsky D. ja Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd edition draft. 2025. https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf.

[2] Ehrmann M., Hamdi A., Pontes E. L., Romanello M. ja Doucet A. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys* 56.2 (2024), lk 1–47. DOI: 10.1145/3604931.

[3] Pilvik M.-L., Muischnek K., Jaanimäe G., Lindström L., Lust K., Orasmaa S. ja Türna T. MÖISTUS SAI KUULOTEDU: 19. SAJANDI VALLAKOHTUPROTOKOLLIDE TEKSTIDEST DIGITAALSE RESSURSI LOOMINE. *EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 15* (2019), lk 139–158. https://research.ebsco.com/linkprocessor/plink?id=df00e73d-5f26-3965-ab54-329d8bccc6ee.

[4] Aldeen M., Luo J., Lian A., Zheng V., Hong A., Yetukuri P. ja Cheng L. ChatGPT vs. Human Annotators: A Comprehensive Analysis of ChatGPT for Text Annotation. *2023 International Conference on Machine Learning and Applications (ICMLA)* (2023), lk 602–609. DOI: 10.1109/ICMLA58977.2023.00089.

[5] González-Gallardo C.-E., Boros E., Girdhar N., Hamdi A., Moreno J. G. ja Doucet A. Yes but.. Can ChatGPT Identify Entities in Historical Documents? *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2023), lk 184–189. DOI: 10.1109/JCDL57899.2023.00034.

[6] Devlin J., Chang M., Lee K. ja Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805.

[7] Sügis E., Tampuu A., Aljanaki A., Fišel M. ja Kull M. Praktiline andmeteadus. Tartu Ülikooli arvutiteaduse instituut, 2024.

[8] Tanvir H., Kittask C., Eiche S. ja Sirts K. EstBERT: A Pretrained Language-Specific BERT for Estonian. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (2021). Toim. Dobnik S. ja Øvrelid L., lk 11–19. https://aclanthology.org/2021.nodalida-main.2/.

[9] Sirts K. Estonian Named Entity Recognition: New Datasets and Models. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (2023), lk 752–761. https://aclanthology.org/2023.nodalida-1.76.pdf.

[10] Orasmaa S., Muischnek K., Poska K. ja Edela A. Named Entity Recognition in Estonian 19th Century Parish Court Records. *Proceedings of the 13th Conference on Language Resources and Evaluation* (2022), lk 5304–5313. https://aclanthology.org/2022.lrec-1.568.pdf.

[11] Lison P., Hubin A., Barnes J. ja Touileb S. Named Entity Recognition without Labelled Data: A Weak Supervision Approach. *CoRR* abs/2004.14723 (2020). arXiv: 2004.14723. https://arxiv.org/abs/2004.14723.

[12] Amini M.-R., Feofanov V., Pauletto L., Hadjadj L., Devijver É. ja Maximov Y. Self-training: A survey. *Neurocomputing* 616 (2025), lk 128904. DOI: 10.1016/j.neucom.2024.128904.

[13] Laur S., Orasmaa S., Särg D. ja Tammo P. EstNLTK 1.6: Remastered Estonian NLP Pipeline. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (2020), lk 7152–7160. https://aclanthology.org/2020.lrec-1.884/.

[14] University of Tartu. UT Rocket. 2018. DOI: 10.23673/PH6N-0144.

Lisad

Töös kasutatud koodifailid on kättesaadavad järgneval lingil: https://github.com/normant23/ner-loputoo

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Norman Tolmats,

- annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Nimeüksuste tuvastamine ajaloolistes Tartu Linnavolikogu protokollides“, mille juhendaja on Siim Orasmaa, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
- annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
- olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
- kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Norman Tolmats

15.05.25