TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Darja Manajeva 213073IAIB

Eestikeelse sisendteksti alusel ametite tuvastamise mudel

Bakalaureusetöö

Juhendaja: Markko Liutkevičius

MSc

Tallinn 2024

Tallinn 2024

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Darja Manajeva

27.05.2024

Annotatsioon

Käesoleva lõputöö eesmärk oli arendada mudel, mis suudab tuvastada ametinimetusi eesti keeles esitatud tekstisisendite põhjal. Selleks katsetati mitmeid masinõppe ja loomuliku keele töötlemise meetodeid, sealhulgas *TF-IDF* koos *Cosine Similarity*-ga, *Doc2Vec* ja *Random Forest*.

Töö käigus koguti andmeid *ESCO* andmebaasist ning Eesti ülikooli ja teadusinfosüsteemi veebilehtedelt. Kogutud andmed eeltöödeldi, et tagada nende sobivus mudelite treenimiseks. See hõlmas andmete puhastamist, lemmatiseerimist ja standardiseerimist. Kolme erineva lähenemise põhjal treenitud mudelid kombineeriti koondmudeliks, et saavutada parim võimalik ennustusvõimekus.

Mudeli valideerimine näitas, et koondmudel saavutas kõrge täpsuse 100%. Need tulemused kinnitavad mudeli efektiivsust ja praktilist rakendatavust Eesti tööturuteenuste jaoks.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 42 leheküljel, 4 peatükki, 14 joonist, 3 tabelit.

Abstract

Model for Occupation Identification from Estonian Input Text

The main objective of this thesis was to develop a model capable of accurately identifying job titles based on Estonian text inputs. To achieve this, various machine learning and natural language processing techniques were tested, including TF-IDF with cosine similarity, Doc2Vec, and Random Forest.

The project involved collecting and preprocessing data from the ESCO database and Estonian university and scientific information system websites. These datasets were cleaned, lemmatized, and standardized to ensure their suitability for model training. Three different approaches were used for training the models, which were then combined into an ensemble model to achieve the best possible prediction accuracy.

The validation of the model showed that the ensemble model achieved an exceptionally high accuracy of 100%. These results confirm the model's effectiveness and practical applicability for Estonian labor market services.

Further practical application of the project includes integrating the model into national employment agency systems to provide better job and training recommendations. Future steps could involve expanding and enhancing the model to cover more occupations and sectors, as well as exploring the possibilities of processing multilingual text inputs.

This project contributes to the development of labor market recommendation systems and demonstrates how machine learning methods can be used to solve practical problems. The high accuracy and reliability of the final model confirm its readiness for use and its potential to better meet the needs of job seekers and employers.

The thesis is in Estonian and contains 42 pages of text, 4 chapters, 14 figures, 3 tables.

Lühendite ja mõistete sõnastik

ESCO	*European Skills, Competences, Qualifications and Occupations*
ISCO	*International Standard Classification of Occupations*
ML	*Machine Learning*, masinõpe
NLP	*Natural Language Processing*, loomuliku keele töötlus
CSV	*Comma-Separated Values*, failiformaat andmete salvestamiseks
EstNLTK	*Estonian Natural Language Toolkit*, eesti keele töötlemise tööristakomplekt
TF-IDF	*Term Frequency-Inverse Document Frequency*, tekstianalüüsi meetod
Cosine Similarity	Vektorite sarnasuse mõõdik
Doc2Vec	*Document to Vector*, vektorruumi mudel
BERT	*Bidirectional Encoder Representations from Transformers*, süvaõppe mudel
EstBERT	*Estonian Bidirectional Encoder Representations from Transformers, süvaõppe mudel*
GPT	*Generative Pre-trained Transformer*, süvaõppe mudel
RNN	*Recurrent Neural Network*, süvaõppe mudel
Naive Bayes	Masinõppe mudel
SVM	*Support Vector Machine*, masinõppe mudel
Random Forest	Masinõppe mudel
Logistiline Regressioon	Masinõppe mudel
Ensemble methods	Koondmeetodid, mitme masinõppe mudeli kombineerimiseks

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Tänapäeva kiiresti arenevas digitaalses maailmas on tehnoloogiate integreerimine ühiskonna erinevatesse valdkondadesse muutunud möödapääsmatuks vajaduseks. Nende valdkondade hulgas on ka tööturuteenuste sektor. Riiklikud tööturuga seotud asutused, nagu Eesti Töötukassa, mille eesmärk on aidata tööotsijatel leida neile sobivaid töö- ja koolitusvõimalusi, seisavad silmitsi väljakutsega kohaneda tehnoloogilise arenguga ja pakkuda tõhusaid, kasutajasõbralikke platvorme [1]. Hiljutine ülemaailmne *COVID-19* pandeemia on seda probleemi veelgi ilmsiks toonud, rõhutades vajadust moderniseerimise ja innovatsiooni järele tööturuteenuste valdkonnas [2].

Käesolevas töös käsitletakse Eesti praeguse riikliku tööturuasutuse Töötukassa puudusi, keskendudes ametite tuvastamise mudeli väljatöötamisele eestikeelsete sisendtekstide põhjal. Käesoleva tööga seotus projekt peab olema üheks osaks uuest süsteemist, mis kasutab keeletöötlus- ja masinõppe meetodeid selleks, et parandada tööotsijate ja neile sobivate tööpakkumiste kokkusobitamist kooskõlas Euroopa oskuste, pädevuste, kvalifikatsioonide ja ametite (*ESCO*) klassifitseerimissüsteemiga.

1.1 Taust

Riiklikel tööturuteenustel on oluline roll tööturutehingute hõlbustamisel, ühendades tööotsijaid potentsiaalsete tööandjatega. Need platvormid lihtsustavad mitmesuguseid teenuseid, alates sotsiaalkindlustushüvitiste taotlemisest kuni uute töövõimaluste otsimise ja koolitusprogrammidele registreerimiseni.

1.1.1 Tänapäeva tööturu e-teenuste olukord – arengud ja probleemid

Praegu on paljud riiklikud tööturuteenused kogu Euroopas, sealhulgas Eesti Töötukassa, välja töötanud digitaalseid portaale, mis on mõeldud kodanike abistamiseks [3]. Need portaalid pakuvad mitmesuguseid teenuseid, näiteks sotsiaalkindlustushüvitiste taotlemine, uute töövõimaluste otsimine ja koolitusprogrammidesse registreerimine. Näiteks Eestis on inimestel võimalus saada oma isiklikke andmeid, mis on juba mitmes

avaliku sektori organisatsioonides olemas, näiteks Haridus- ja Teadusministeeriumi haridusteavet või Maksu- ja Tolliameti tööajalugu. Sellist sujuvat teabevahetust hõlbustab Eesti riiklik andmevahetuskiht, mida tuntakse *X-ROAD* nime all [4]. Need platvormid kergendavad nii töötavatel kui ka töötutel Eesti residentidel e-töötukassa portaali *CV* loomise teenuse kasutamist.

Vaatamata nendele edusammudele on endiselt mitmeid kriitilisi probleeme:

- Oskuste mittevastavus: Traditsiooniliste tööstusharude kiire sulgemine ja uute sektorite tekkimine tekitab lõhe olemasoleva tööjõu oskuste ja turu vajaduste vahel. See ebakõla on eriti nähtav piirkondades, kus toimuvad olulised tööstuslikud muutused, nagu näiteks põlevkivist loobumine Eestis [5].

- Ebatõhusad sobitusmehhanismid: Praegused töövahendussüsteemid, mis tuginevad suuresti *ISCO*-koodidele, ei suuda sageli tööotsijaid täpselt sobitada sobivate võimalustega. Sellise süsteemi jäikus võib jätta tähelepanuta kandidaadi põhjalikud oskused ja kogemused, mis viib ebatõhusa töövahendamiseni ja töötuse kestuse pikenemiseni [6].

- Intelligentsete sobitussüsteemide puudumine: Kaasaegsetel e-teenustel puuduvad arenenud intelligentsed sobitusalgoritmid, mis suudavad dünaamiliselt kooskõlastada individuaalset kvalifikatsiooni ja töökoha nõudeid. Selle puuduse tõttu peavad tööotsijad käsitsi otsima ulatuslikke nimekirju, mis on aeganõudev ja sageli viljatu ülesanne [7].

Traditsioonilised lähenemisviisid tuginevad vananenud tehnoloogiatele ning ei kasuta tänapäeva digitaalajastul ilmnenud töösoovitusmeetodeid. See toob kaasa probleeme süsteemi efektiivsusega tõhusa töövahenduse saavutamisel, mis takistab nii tööotsijate kui ka tööandjate tööd [8].

COVID-19 pandeemia rõhutas veelgi e-tööturuteenuste tähtsust, kuna ootamatu üleminek kaugtööle ja kiire digitaliseerimine tõid kaasa rohkem väljakutseid nii tööotsijatele kui ka tööandjatele. Vajadus uutele tehnoloogiapõhistele lahendustele tööturuteenuste valdkonnas ei ole kunagi varem olnud nii ilmne kui praegu. Kriis rõhutas vajadust, et riiklikud tööturuteenused kasutaksid digitaalseid uuendusi, et paremini teenindada tööotsijaid ja tööandjaid tänapäeval [9].

1.1.2 *ISCO* ja *ESCO*

ISCO on laialdaselt kasutatav raamistik ametite klassifitseerimiseks oskuste taseme ja tehtava töö liigi alusel. Siiski on sellel piirangud, mis tulenevad klassifikatsiooni suurusest, mis ei pruugi täpselt kajastada tänapäeva töökohtade spetsiifilisi oskusi ja haridusnõudeid. Näiteks võivad *ISCO* klassifikatsioonid olla liiga üldised spetsialiseeritud valdkondade jaoks, mis nõuavad täpseid oskusi. See toob kaasa ebakõla ametikirjelduste ja tegelike tööülesannete vahel. Lisaks ei ole *ISCO* klassifikatsioonid sammu pidanud tehnoloogilise arengu ja maailma digitaaliseerimise tõttu tekkivate uut tüüpi ametite arenguga, mis nõuavad sageli traditsioonilisi valdkondlikke piire ületavate oskuste kombinatsioone [10]. Riiklikud tööturuteenused peavad keskenduma kohanemisvõimeliste soovitussüsteemide väljatöötamisele ning kasutusevõtule, kasutades standardiseeritud raamistikke nagu *ESCO*. *ESCO* eesmärk on ühtlustada oskuste ja ametikohtade klassifitseerimist kogu Euroopa Liidus, hõlbustades tõhusamaid töövahendusprotsesse [11]. *ESCO* pakub ajakohasemat ja üksikasjalikumat klassifitseerimissüsteemi, mis suudab paremini arvesse võtta tööturu dünaamilist iseloomu ja eri sektorites vajalikke erioskusi, hõlbustades tõhusamat töökohtade sobitamise protsessi.

1.1.3 Varasemad uuringud

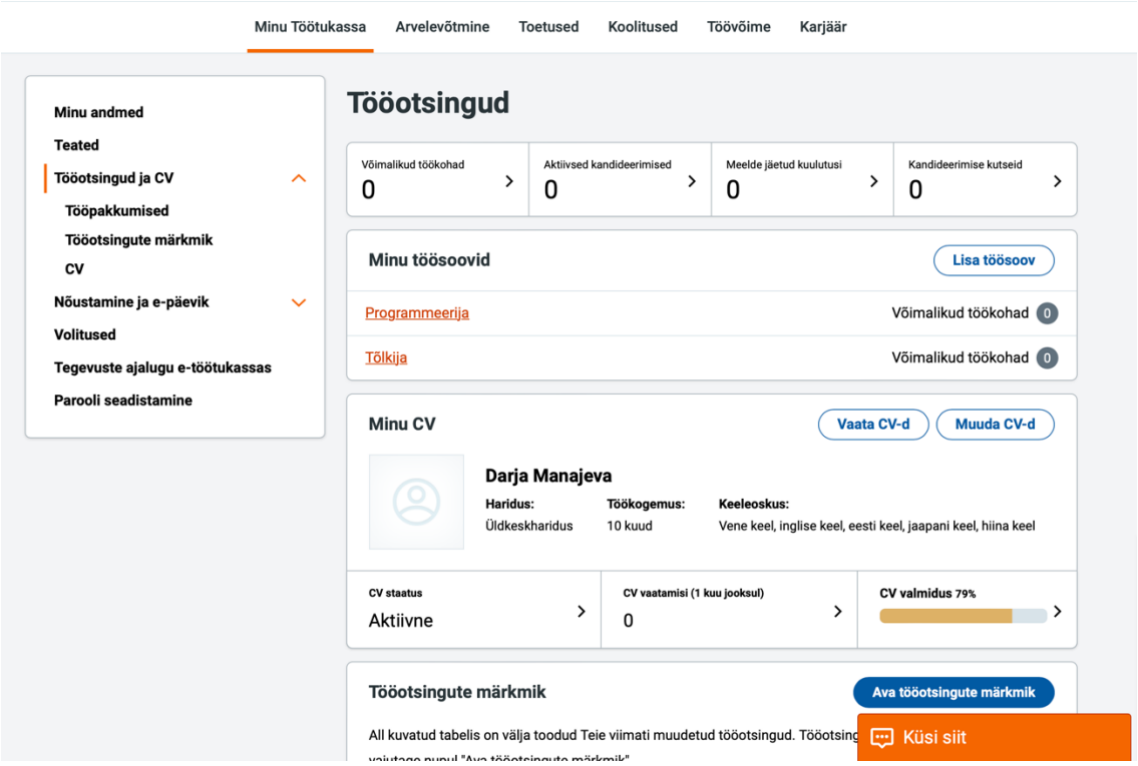
Varasemad uuringud selles valdkonnas on loonud aluse tööturuteenustega seotud probleemide ja võimaluste mõistmiseks. Käesolev töö keskendub pigem Eesti tööturuteenuste portaali e-töötukassa puudustele. Hoolimata sellest, et Eesti on tuntud digitaliseerimise poolest, tuginevad riiklikud tööturuasutused käsitsi tehtavatele menetlustele ja vananenud klassifitseerimissüsteemidele. Selline tehnoloogilise

integratsiooni puudumine takistab töövahenduse tõhusust, mis aeglustab personaliseeritud soovitusi ja piiriülest tööjõu liikuvust [12].

Avaliku sektori tööturuasutuste probleemide lahendamine nõuab tehnoloogiliste uuenduste kasutamist ning isikupärastatud soovituste prioritiseerimist. Nende probleemidega tegelemine võimaldab riiklikel tööturuasutustel paremini teenindada tööotsijaid ja tööandjaid, aidates lõppkokkuvõttes kaasa tõhusamale tööturule.

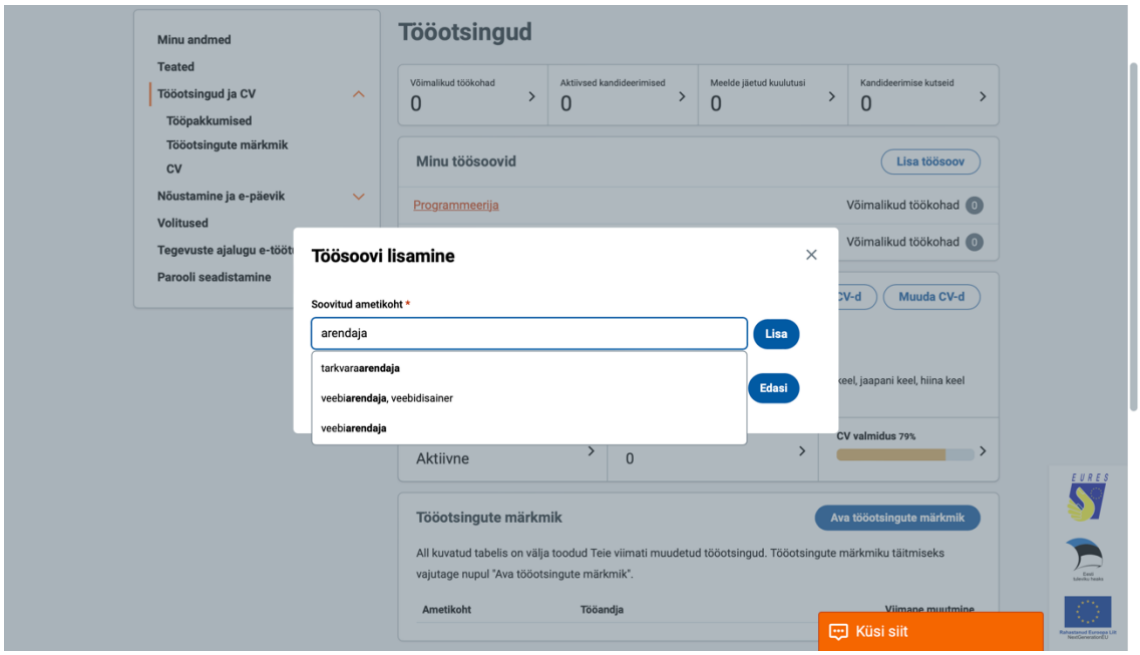
1.2 Probleemi püstitus

Praeguses töösoovitussüsteemis, mida kasutab Eesti avalik portaal e-töötukassa, puudub sobitamine kodanike *CV*-de ning töö- ja koolituskuuluste vahel. Seda puudust rohkem süvendab oskuste ja kvalifikatsioonide ja muu sisestatud informatsiooni mittearvestamine sobitamise käigus. Kuna süsteem ei arvesta *CV*-de andmeid, kodanikud peavad käsitsi läbi vaatama töökuulutusi ja koolitusi. Probleemiks on ka see, et töö soovituse saamiseks peab inimene ise tegema töösoove, mis pole kuidagi süsteemis seotud eelmise *CV* loomisega (vt. Joonis 1).



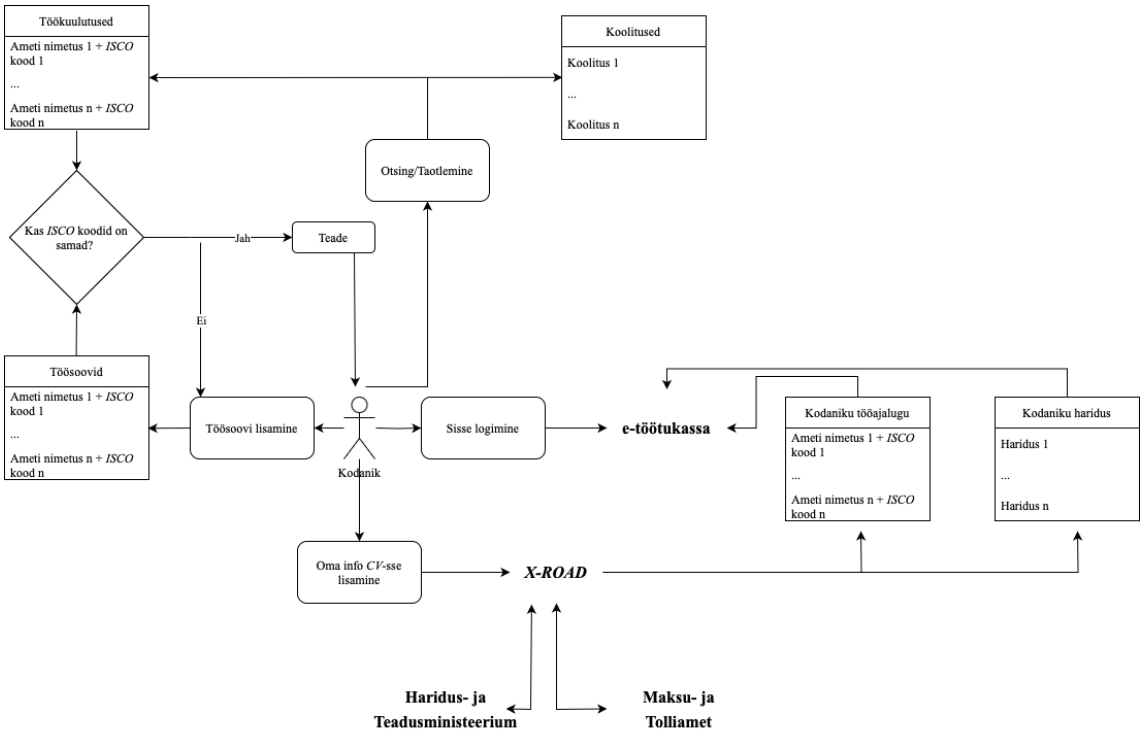
Joonis 1. Töösoovide loomise koht.

Inimene ise valib temale sobiva ametikoha ainult nende nimetuste nimekirjast, millel on olemas *ISCO* kood, ning selle töösoovi põhjal saab võimalikud töökuulutused (vt. Joonis 2).



Joonis 2. Töösoovi lisamisel võimalikke ametite nimetuste valik.

Kuna tööandjatel on vaja teha sama päringu, nii *ISCO* andmestiku suuruse probleem kui ka inimese võimalik valesti valitud ameti nimetus, võivad tekitada mittesobivuse. Selles olukorras peab tööd otsiv inimene oma töösoovi iseseisvalt täpsustama või muutma (vt. Joonis 3).



Joonis 3. Praeguse e-töötukassa töösoovitussüsteemi *AS-IS.*

Lisaks sellele ei ole tööotsijate ja vabade töökohtade vastavusse viimine *ISCO*-koodide abil piisav, eriti arvestades tööturu dünaamilist iseloomu ja uute ametite tekkimist. Vaatamata Euroopa Komisjoni otsusele võtta kasutusele Euroopa oskuste, pädevuste ja ametite klassifikatsioon *ESCO* [13], olemasolevad töö ja töötajate otsingu vahendid on vananenud ning ei vasta kaasaegsetele digiteerimise nõutele. Sellest tulenevalt kodanikukeskseid teenused, mis toetavad oskuste täiendamist ja tööotsinguid, on ebatõhusad ja resursimahukad [9]. Nende probleemide lahendamiseks on vaja arendada sobitusalgoritme ja integreerida standardsed oskuste klassifitseerimissüsteemid, nagu *ESCO*, avalike tööturuteenuste portaalidesse, et suurendada töö- ja koolitussoovituste täpsust [14].

1.3 Töö eesmärgid

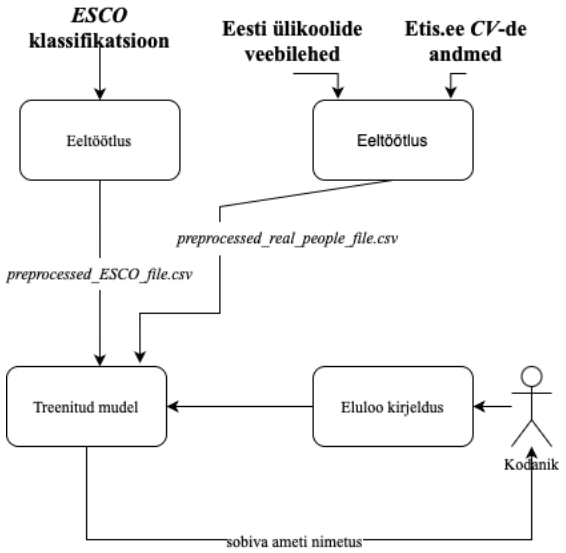
Käesoleva töö peamine eesmärk on töötada välja ametite tuvastamise mudel eesti keele sisendtekstide põhjal. Selle saavutamiseks on spetsiifilised ülesanded järgmised:

1. Katsetada erinevaid tehnoloogiaid ja keelemudeleid, et tuvastada *ESCO* klassifikatsioonis loetletud ametite vasted.

2. Hinnata erinevate meetodite sobivust ametite tuvastamiseks eestikeelse sisendteksti põhjal.

3. Luua mudel, mis suudab tuvastada *ESCO* klassifikatsioonis loetletud ametite vasted ning mida saab tulevikus kasutada suuremas süsteemis riiklikes tööturuasutustes parema töökohtade sobitamiseks.

Uus mudel, mille loomine on käesoleva töö eesmärk, saab sisendina inimese loomingu kirjelduse eestikeelse tekstina ning selle põhjal tuvastab kõige sobivama temale ameti nimetuse *ESCO* klassifikatsioonist, tagastades selle ameti nimetuse (vt. Joonis 4).



Joonis 4. Uue süsteemi *TO-BE*.

1.4 Uurimistöö ulatus ja piirangud

Käesolev töö keskendub konkreetselt ametite tuvastamise mudeli väljatöötamisele eestikeelse sisendteksti põhjal. See töö keskendub suurema probleemi lahenduse ühele osale - mudelile, mis on treenitud kasutama *ESCO* ametite nimetusi ja nende kirjeldusi ning mis, saades sisendiks inimese *CV* kirjelduse teksti tagastab kõige sobivama ameti nimetuse. See töö on osa suuremast töökuulutuste soovitussüsteemist, mille tervikliku loomisega käesolev töö ei tegele.

Kuigi sobitussüsteemiga probleemi lahendus nõuab suurendada riiklike tööturuteenuste tõhusust, ei käsitle käesolev uuring laiemaid küsimusi, mis on seotud tööalaste soovitussüsteemide või terviklike tööturu reformidega. Lisaks piirdub käesolev töö keele töötlemise meetodite ja nende rakendamise hindamisega ametite tuvastamisel, jättes kõrvale muud võimalikud tegurid, mis võivad mõjutada töövahenduse tulemusi.

Üheks piiranguks on ametite arv, mida kasutati käesoleva töö mudeli väljatöötamiseks. *ESCO* klassifikatsioonsüsteemis on rohkem ameteid, kuid selles töös kasutati ainult teadusalaseid ameteid, mis on leitavad *ESCO* klassifikatsiooni andmebaasist. Nende ametite hulgas on 122 ametit eraldi olemasolevas *CSV* failis. Selline lähenemine oli valitud arvestades, et mudeli testimiseks on vaja koguda andmestikku reaalsete inimeste andmetega. Selleks, et töö ei oleks liiga keskendunud andmete kogumisele, kasutatakse piiratud arvu ameteid. Põhjuseks teadusalaste ametite valikuks oli lihtne juurepääs valdkonnas töötavate inimeste infole. Otsustati koguda andmestikku leides inimeste *CV*-d veebilehelt Etis.ee ning nende loendid ja ametite info Tallinna Tehnikaülikooli veebilehelt.

Andmete kogumise käigus ilmnes kaks piirangut. Esimene seisneb selles, et Eesti ülikoolide veebilehtedel on töötajate ametite detailne informatsioon puudulik. Veebilehtedel olevad ametid pole täpsed ega samad *ESCO* klassifikatsioonis olevate ametite nimetustega. See raskendas inimeste ametite tuvastamist *ESCO* klassifikatsiooni ametite nimekirja põhjal. Lahenduseks valiti *ChatGPT* treenimine *ESCO* ametitega faili põhjal ning selle kasutamine inimese sobivaima ameti nimetuse leidmiseks.

Teine töö piirang on seotud inimeste *CV*-des leiduva teabega, mis tihti pole ainult eesti keeles, vaid ka muudes keeltes. Kuna käesolev töö keskendub ainult eestikeelse sisendteksti töötlemisele, inimese poolt süsteemi sisestatud teave teistes keeltes ei ole tõlgitud, vaid elimineeritakse valimist.

1.5 Töö struktuur

Käesolev töö koosneb kolmest peamisest osast: kirjanduse ülevaadest, töö protsessi kirjeldusest ja loodud mudeli ülevaatusest. Kirjanduse ülevaates analüüsitakse *ISCO* ja *ESCO* ajaloolist konteksti, kaasaegseid tehnoloogiaid ja metodoloogiaid

soovitussüsteemides, efektiivsete soovitussüsteemide peamisi väljakutseid ja lahendusi ning tuuakse esile uurimislüngad, mida käesolev töö püüab täita. Tööprotsessi peatükis kirjeldatakse andmete kogumist, eeltöötlust ning keeletehnoloogiliste meetodite ja mudelite rakendamise protsessi. Loodud töörista ülevaate peatükis kirjeldatakse mudeli loomise protsessi ja hinnatakse selle efektiivsust.

2 Kirjanduse ülevaade

Selles peatükis käsitletakse *ISCO* ja *ESCO* klassifikatsioonisüsteeme, praeguseid tehnoloogiaid ja metodoloogiaid, mida kasutatakse soovitussüsteemides. Lisaks, selles osas tutvustatakse koondmeetodite (*Ensemble methods*) eeliseid ja tüüpe ning nende rakendamist töösoovitussüsteemides. Koondmeetoditega treenitakse mitu mudelit koos selleks, et saada täpsemad tulemused [15]. Keskendutakse sellele, kuidas koondmeetodid aitavad parandada täpsust ja stabiilsust keeruliste ja suuremõõtmeliste tekstiandmete töötlemisel. Selle osa lõpus leitakse uurimislünka, mida töö üritab täita.

2.1 *ISCO* ja *ESCO* klassifikatsioonisüsteemid

Rahvusvaheline ametite standardklassifikatsioon (*ISCO*) on olnud oluline raamistik ametinimetuste ja oskuste kategoriseerimiseks rahvusvaheliselt. Rahvusvahelise Tööorganisatsiooni (*ILO*) poolt välja töötatud *ISCO* on alates selle loomisest 1957. aastal mitu korda ajakohastatud, praegune versioon on *ISCO-08*. *ISCO* peamine eesmärk on pakkuda standardiseeritud keelt ametikohtade klassifitseerimiseks, mida saab kasutada eri riikides ametikohtade vastavusse viimiseks ning statistika koostamiseks [6].

Seevastu Euroopa oskused, pädevused, kvalifikatsioonid ja ametid (*ESCO*) on uuem algatus, mille käivitas Euroopa Komisjon, et täpsustada ja lokaliseerida kutsealade andmeid, et neid paremini Euroopa tööturuga vastavusse viia. *ESCO* integreerib ja laiendab *ISCO* kontseptsioone, hõlmates iga ametiga seotud konkreetsed oskused ja pädevused [16]. Sellist ametite ja oskuste seost *ISCO* klassifikatsioonis ei esine.

2.1.1 Ametialaste klassifikaatorite areng

ISCO areng aastakümnete jooksul kajastas muutusi tööturul, eelkõige uute tööstusharude ja tehnoloogiate esilekerkimist. Iga *ISCO* läbivaatamise eesmärk on neid muutusi kajastada, et jääda asjakohaseks ja kasulikuks tööturuasutuste jaoks kogu maailmas. Näiteks *ISCO-88* oli märkimisväärne selle poolest, et selles rõhutati teenustele orienteeritud ameteid, tunnistades ülemaailmset üleminekut teenustepõhisele

majandusele [17]. *ISCO* viimane versioon oli lisatud aastal 2008, mille pärast puuduvad sealt pärast 2008 aastat ilmunud ametid.

ESCO, mis võeti kasutusele 2011. aastal, ei ole mõeldud mitte ainult *ISCO* täiendamiseks, vaid ka selleks, et pakkuda üksikasjalikumat klassifikatsiooni, mis tunnustab Euroopa tööturu jaoks olulisi erinevaid oskusi ja pädevusi [18]. *ESCO* klassifitseerib praegu umbes 3039 ametit ja seda ajakohastatakse korrapäraselt, et kajastada muutuvaid turunõudeid.

2.1.2 *ESCO* kohustuslik integreerimine *EU* tööturuasutustesse

Üks *ESCO* rakendamise kriitilisi aspekte on selle kohustuslik integreerimine riiklikesse tööturuasutustesse kogu Euroopa Liidus. See direktiiv loodi selleks, et ühtlustada ametite, oskuste ja pädevuste klassifitseerimist kõigis *EU* liikmesriikides, hõlbustades seeläbi paremat liikuvust ja töökohtade piiriülest sobitamist [14]. *ESCO* integreerimine tööturuteenustesse, nagu *EU* tööalase liikuvuse portaal (*EURES*), võimaldab tööotsijatel ja tööandjatel kogu *EU*-s leida kokkulangevusi üksikasjalike ja standardiseeritud kriteeriumide alusel, suurendades töövahendusteenuste tõhusust ja tulemuslikkust [19].

2.1.3 *ISCO* ja *ESCO* arengu mõju tööturule

ISCO-08 liigitab töökohad 436 ühikurühma. Need ühikurühmad on koondatud 130 väiksemaks rühmaks, 43 alamsuurrühmaks ja 10 suurrühmaks, mis põhinevad töökohtade jaoks vajalike oskuste taseme ja spetsialiseerumise sarnasusel, mis teeb *ISCO*-st neljatasandilise hierarhiliselt struktureeritud klassifikatsiooni [6].

ESCO loomine näitab üleminekut üldise kategoriseerimise vajadusest üksikasjalikku oskuste ja pädevuste kaardistamise suunas. See muutus on oluline, et vastata kaasaegse tööturu täpsetele vajadustele, mida iseloomustab kiire tehnoloogiline areng ning sotsiaalsete oskuste ja pädevuste kasvav tähtsus. Riiklike tööturuametite jaoks hõlbustab *ESCO* kasutamine paremat teenuste osutamist, kohandatud töösoovitusi ja paremat poliitika kujundamist, pakkudes üksikasjalikke andmeid tööturu vajaduste kohta [20].

ESCO kujutab endast märkimisväärset edasiminekut võrreldes *ISCO*-ga, kuna see sisaldab *ISCO-08* nelja taseme alla jäävaid järgnevaid tasemeid, mis pakub detailsemat lähenemist ametite klassifitseerimisele. Selles struktuuris on iga *ESCO* kutseala vastavuses täpselt ühe *ISCO-08* koodiga, säilitades range monohierarhilise seose, mis

tagab, et igal 5. või madalama taseme ametil on täpselt üks *ISCO* kõrgemate tasemete vanem. Selline integreerimine tagab, et *ESCO* ei ole mitte ainult kooskõlas *ISCO* kehtestatud raamistikuga, vaid ka laiendab seda, andes üksikasjaliku ülevaate konkreetsetest tööülesannetest ning nende nõutavatest pädevustest ja kvalifikatsioonidest. Mõned *ISCO* rühmad ei sisalda *ESCO* ametialasid, eriti neid, mis ei ole *EU* majandustegevuse kontekstis olulised, nagu näiteks „vee- ja küttepuude kogujad“ [21].

2.1.4 Üleminek ESCO klassifikatsioonisüsteemile

Üleminek *ISCO*-lt *ESCO*-le kujutab endast olulist arengut tööturuanalüüsis ja riiklikes tööhõiveteenustes. *ESCO* kasutuselevõtuga võtavad *EU* liikmesriigid endale kohustuse luua integreeritud ja tõhusam raamistik töökohtade sobitamiseks, mis mitte ainult ei toeta töötajate liikuvust liidus, vaid parandab ka tulevaste tööjõuvajaduste strateegilist planeerimist. Kuna tööturg areneb jätkuvalt, on standardiseeritud klassifitseerimissüsteemide, nagu *ESCO*, roll tõhusate tööhõivepoliitikate ja -tavade kujundamisel üha olulisem.

2.2 Soovitussüsteemid

Selles osas analüüsitakse praeguseid soovitussüsteemide tehnoloogiaid ja metodoloogiaid, mida kasutatakse soovitussüsteemides. Käsitletakse dokumendi- ja sõnatasandi probleeme, pikamaa sõltuvusi, staatilisi ja dünaamilisi mudeleid ning nende rakendamist töösoovitussüsteemides. Selgitatakse ka loomuliku keele töötluse (*NLP*) ja masinõppe (*ML*) tehnoloogiate rolli töösoovitussüsteemide arendamisel. Uuritakse erinevaid mudeleid ja algoritme, mis on kasutatavad ametite tuvastamiseks loomuliku keele tekstide põhjal, rõhutades nende olulisust ja eeliseid.

2.2.1 Dokumendi tasandi ja sõnade tasandi probleemid

Tekstitöötluse ja -analüüsi puhul on oluline mõista erinevust dokumenditasandi ja sõnatasandi probleemide vahel. Need kaks kategooriat määravad, kuidas teksti esitatakse, analüüsitakse ja töödeldakse erinevate ülesannete puhul.

Dokumenditasandi probleemid keskenduvad terve dokumendi või suure tekstiosa analüüsile ja töötlemisele:

- Teksti klassifitseerimine: kogu dokumendi klassifitseerimine teatud kategooriasse, näiteks rämpsposti tuvastamine või uudisartikli teema määramine. Näiteks: e-kirja klassifitseerimine kategooriatesse „rämpspost“ ja „mitte-rämpspost“.

- Teksti rühmitamine: dokumentide rühmitamine klastritesse nende sarnasuse alusel, ilma eelnevalt määratletud kategooriateta. Näiteks: uudisartiklite rühmitamine teemade, näiteks spordi, poliitika ja tehnoloogia järgi.

- Tundlikkuse analüüs: kogu dokumendi sentimendianalüüs, näiteks ennustamine, kas tekst on positiivse, negatiivse või neutraalse tonaalsusega. Näiteks: arvustuse analüüsimine, et teha kindlaks, kas klient on rahul või rahulolematu.

- Informatsiooni ekstraheerimine: konkreetse teabe leidmine ja väljavõtete tegemine suurtest tekstikogumitest. Näiteks: konkreetsete faktide väljavõtmine teadusartiklitest või juriidilistest dokumentidest .

Probleemid sõnade tasandil keskenduvad üksikute sõnade või väiksemate fraaside analüüsile ja töötlemisele:

- Sõnade sarnasus: sünonüümide tuvastamine või sõnade asendamine konteksti säilitamiseks.

- Sõnade klassifitseerimine: sõnade liigitamine erinevatesse kategooriatesse, näiteks nimisõnad, omadussõnad. Näiteks: sõna „koer“ klassifitseerimine nimisõnaks.

- Sõnajärje ennustamine: järgmise sõna tõenäosuse ennustamine antud kontekstis.

- Lemmatiseerimine: Sõnade normaliseerimine nende algvormi (*lemma*) või tüvele. Näiteks: sõnad „jookseb“, „jooksis“ ja „jooksis“ normaliseeritakse nende algvormi „jooksma“ [22].

2.2.2 Pikamaa sõltuvused

Pikamaa sõltuvused *(Long-Term Dependencies)* tekstis viitavad suhetele või sõltuvustele, mis esinevad kaugel asuvate sõnade või fraaside vahel. Nende sõltuvuste mõistmine on oluline teksti tähenduse täpseks tõlgendamiseks, eriti kui ühe osa tähendus sõltub teisest osast, mis ei asu vahetus läheduses.

Pikamaa sõltuvused on oluline teema loomuliku keele töötlemises, kuna need võimaldavad mõista keerulisi keelelisi struktuure ja kontekste. Traditsioonilised mudelid suudavad tavaliselt haarata ainult lühiajalisi sõltuvusi, kuid pikamaa sõltuvused nõuavad keerukamaid lähenemisi [23].

Näited pikamaa sõltuvustest *(Long-Term Dependencies)*:

- Asesõna lahendamine: Näide: "Alice läks poodi, sest tal oli vaja piima." Siin viitab "tal" Alice-ile, mis nõuab pikaajalise sõltuvuse mõistmist. lahendamine on *NLP*-s keeruline ülesanne, kuna see nõuab konteksti analüüsi, et määrata, millisele nimisõnale viitab [24].

- Kontekstuaalsed sõltuvused: "Raamat, mille sa mulle eelmisel nädalal andsid, oli põnev." Sõna "eelmise nädalal" muudab tegusõna "andsid" konteksti, mis nõuab mitme sõna vahelist sõltuvust. Sellised kontekstuaalsed sõltuvused on olulised täpse tähenduse tabamiseks ja nõuavad keerukate mudelite kasutamist, mis suudavad analüüsida kogu lause struktuuri [25].

- Tundlikkuse analüüs: Näide: "Vaatamata hilinemisele oli film täiesti fantastiline." Sõna "Vaatamata" mõjutab kogu lause tonaalsust, mis nõuab sõltuvuse mõistmist algusest lõpuni. Tundlikkuse analüüs on *NLP*-s oluline, kuna see aitab määrata teksti emotsionaalset tooni ja suhtumist, mis võib ulatuda üle mitme lause [26].

2.2.3 Staatilised ja dünaamilised mudelid

Loomuliku keele töötlemisel ja masinõppes võib mudeleid jagada laias laastus staatilisteks ja dünaamilisteks mudeliteks selle põhjal, kuidas nad sõnade esitusviisi käsitlevad:

- Staatilised mudelid genereerivad iga sõna jaoks ühe kindla representatsiooni (vektori), sõltumata selle kontekstist. See tähendab, et igal sõnal on igas situatsioonis sama varjund. Staatiliste mudelite näited on *Word2Vec*, *GloVe* ja *FastText*. Need mudelid on tõhusad ja lihtsad, kuid neil puudub võime tabada sõnade tähenduste nüansse erinevates kontekstides [27].

- Dünaamilised mudelid seevastu genereerivad erinevaid sõnade esitusi sõltuvalt nende kontekstist lauses. Need mudelid hõlmavad sõnade polüseemiat ja pakuvad täpsemat kontekstuaalset mõistmist. Dünaamiliste mudelite näited on *BERT* ja *GPT*. Need mudelid parandavad märkimisväärselt tulemuslikkust erinevates *NLP*-ülesannetes, võttes arvesse sõnade kontekstuaalset tähendust [28].

2.2.4 Loomuliku keele töötlus ja masinõpe

Töösoovitussüsteemi väljatöötamine, eriti kui see on kohandatud konkreetsetele keelelistele ja piirkondlikele vajadustele nagu Eestis, kuulub loomupäraselt loomuliku keele töötluse (*NLP*) ja masinõppe (*ML*) valdkonda. Selles osas selgitatakse, miks need tehnoloogiad on projekti jaoks keskse tähtsusega ja miks alternatiivsed lahendused ei olnud valitud.

Ametikirjelduste ja *CV*-de tõlgendamine ja sobitamine hõlmab keerukat tekstitöötlust, mis nõuab nii sõnade ja fraaside sõnasõnaliste kui ka kontekstuaalsete tähenduste mõistmist. *NLP* pakub sellisteks ülesanneteks vajalikke vahendeid, sealhulgas tokeniseerimist, sõnade osade märgistamist, nimeliste üksuste tuvastamist ja süntaktilist lahtimõtestamist, mis kõik on olulised tõhusa tekstianalüüsi jaoks keelespetsiifilises kontekstis [29]. Näiteks *EstNLTK* tööriistakomplekt on loodud spetsiaalselt eesti keele tekstide töötlemiseks, pakkudes selle unikaalsetele keelelistele iseärasustele kohandatud funktsioone [30].

ML-algoritmid õpivad andmete näidete põhjal. See õppimine on soovitussüsteemides väga oluline, et teha täpseid töökohtumisi edukate töövahenduste varasemate andmete põhjal. Selliseid algoritme kasutatakse selleks, et ennustada tulemusi õpitud mustrite alusel [31].

Järgmisena on kirjeldatud alternatiivsed lahendused ja nende välistamise põhjused:

- Reeglipõhised süsteemid: algselt olid soovitussüsteemides levinud reeglipõhised süsteemid, kus sobitamiseks loodi käsitsi „*if-then*“ reeglid. Neid süsteeme ei valitud, sest neil puudub skaleeritavus ja paindlikkus. Nad nõuavad pidevat käsitsi uuendamist ja ei suuda õppida uutest andmetest, mistõttu ei sobi need dünaamiliste tööturgude jaoks, kus tööülesanded ja nõuded sageli muutuvad [32].

- Semantilise veebi tehnoloogiad: semantilise veebi tehnoloogiaid, sealhulgas ontoloogiaid ja *RDF*-i (*Resource Description Framework*), võiks teoreetiliselt kasutada tööandmete kategoriseerimiseks ja seostamiseks. Neid tehnoloogiaid ei valitud siiski nende keerukuse ja hoolduse keerukuse ning erinevate andmeallikate integreerimise raskuste tõttu. Lisaks sellele ei paku nad iseenesest prognoosimisvõimet, mida on vaja personaliseeritud töösoovituste jaoks [33].

- Andmebaaside päringu- ja otsingusüsteemid: traditsioonilised andmebaasisüsteemid kasutatakse struktureeritud andmete salvestamiseks ja otsimiseks, kuid neil puudub võime teha keerukat tekstianalüüsi ja õppida andmemustritest. Nad ei ole piisavad ülesannete puhul, kus on oluline mõista konteksti ja keele peensusi, nagu näiteks tööalaste soovitussüsteemide puhul [34].

NLP ja masinõpe valimine Eesti tööturu jaoks töökoha soovitussüsteemi arendamiseks oli tingitud vajadusest täiustatud tekstitöötlusvõimekuse ja andmete põhjal õppimise võime järele, et parandada soovituste täpsust. Need tehnoloogiad tagavad paindlikkuse ja skaleeritavuse, mis on vajalik tööturu dünaamika ja keeleliste nüansside muutustega kohanemiseks. Sellised alternatiivid nagu reeglipõhised süsteemid, semantilise veebi tehnoloogiad ja traditsioonilised andmebaasisüsteemid ei sobinud nende piiratud skaleeritavuse, kohandatavuse ja õppimisvõime tõttu. Keskendumine *NLP*-le ja *ML*-le tagab, et süsteem saab aja jooksul areneda ja kohaneda, parandades oma soovitusi pideva õppimise ja andmeanalüüsi põhjal.

2.2.5 Tänapäevased tehnoloogiad ja metodoloogiad soovitussüsteemides

Soovitussüsteemid on erinevate rakenduste lahutamatu osa, pakkudes kasutajatele personaalseid soovitusi. Tööturuteenuste puhul võivad sellised süsteemid

märkimisväärselt parandada töökohtade sobitamist, viies kandidaatide profiilid vastavusse töökuulutustega. Käesolevas osas vaadeldakse soovitussüsteemide praeguseid mudeleid ja metoodikaid, keskendudes nende sobivusele Eesti tekstiandmete töötlemiseks mõeldud mudeli jaoks.

Enne kui süsteem saab töösoovitusi anda, üks vajalik eeltöötlus ülesanne on sisendandmete töötlus. See muuseas eemaldab töödeldavast tekstist ebavajalikke osasid ja müra selle jaoks, et edasine tekstitöötlus oleks kergem.

NLTK ehk *Natural Language Toolkit* on tööristakomplekt, mis on mõeldud loomuliku keele töötlemiseks (*NLP*) *Python*-i keelega. *NLTK* on tööriistakomplekt *Python*-i programmide arendamiseks, mis töötavad inimkeele andmetega, rõhutades selle rolli keeruliste *NLP* tehnikate kättesaadavamaks muutmisel ning õppimise ja innovatsiooni edendamisel arvutilingvistika valdkonnas [35]. *EstNLTK*, mis oli kasutusel käesoleva tööga seotud projektis, on loodud spetsiaalselt eesti keele jaoks, mistõttu sobib see eesti keelse tekstiga seotud eeltöötlusülesannete jaoks [30]. *EstNLTK* tagab, et eesti keele keelelised eripärad on õigesti käsitletud, mis on üliolulise iga *NLP*-rakenduse jaoks.

Kui tekst on eeltöödeldud, on järgmine samm rakendada algoritmi, et muuta tekst numbrilisteks esitluseks. See võimaldab analüüsida ja ennustada semantilisi sarnasusi, mis on olulised täpsete soovituste andmiseks ja andmete klassifitseerimiseks. Järgnevalt käsitletakse mitmeid selle protsessi jaoks eelduslikult sobivaid mudeleid ja tehnikaid.

Vektorruumi mudelid:

- *TF-IDF (Term Frequency-Inverse Document Frequency)* ja *Cosine Similarity* on olulised mudelid teksti teisendamiseks arvruumi, mis võimaldab numbriliselt esitada dokumentide, näiteks *CV*-de ja töökirjelduste semantilist sarnasust. *TF-IDF* hindab sõnade asjakohasust dokumendi kontekstis, kohandades sõnade sagedust kogu korpuses, samas kui *Cosine Similarity* mõõdab kahe vektori sarnasust nende vahelise nurga alusel [36]. *Cosine Similarity*-id saab tõhusalt kasutada koos *TF-IDF*-iga, kuid üksi see ei anna piisavalt täpset klassifitseerimistulemust. See mõõdab kahe vektori sarnasust, määrates nende vahelise nurga. Kui vektorid on täiesti sarnased, on *Cosine Similarity* väärtus 1;

kui nad on täiesti erinevad, on see -1. Seda kasutatakse sageli teksti ja dokumentide võrdlemiseks [37].

- *Word2Vec* on närvivõrgupõhine mudel, mis õpib sõnavektorite kujutisi (*embeddings*) suure tekstikorpuse põhjal. See mudel kasutab sõnavektorite leidmiseks sõnade konteksti (sõnad enne ja pärast antud sõna). *Word2Vec* on eriti hea sõnade semantilise sarnasuse ja konteksti mõistmisel [27].

- *FastText*, mis on *Word2Vec*-i laiendus, lisab sõnale alamsõnainfot, mis võimaldab sõnavormide ja morfoloogiliste tunnuste paremat käsitlemist. *FastText* on eriti kasulik keeletehnoloogilistes ülesannetes, kus sõnadel on palju erinevaid vorme ja tüvesid [38].

- *GloVe (Global Vectors for Word Representation)* on vektorruumi mudel, mis kombineerib sõnade koosesinemise statistikat kogu teksti koorpuses, et luua sõnavektoreid. *GloVe* suudab tabada semantilisi tunnuseid, andes iga sõna jaoks kontekstivektorid [39].

- *Doc2Vec* laiendab sõnade semantilise sarnasuse analüüsi tervetele dokumentidele, pakkudes esitust, mis suudab tabada sügavamaid semantilisi tähendusi. See meetod teisendab kogu dokumendi tihedaks vektoriks *(dense vector)*, säilitades sõnade järjestuse ja konteksti [40].

Süvaõppe mudelid:

- *BERT (Bidirectional Encoder Representations from Transformers)* on süvaõppe mudel, mis mõistab teksti konteksti, vaadeldes samaaegselt sõna vasak- ja parempoolseid naabreid. See mudel on tunnustatud oma võime poolest mõista tekstis esinevaid kontekstuaalseid nüansse [28]. *EstBERT* on suure eelõpetatud transformeril põhinev keelespetsiifiline *BERT* mudel, mis on loodud spetsiaalselt eesti keele jaoks. Uuringud on näidanud, et keelespetsiifilised *BERT* mudelid pakuvad paremat jõudlust võrreldes mitmekeelsete mudelitega [41].

- *GPT-2 (Generative Pre-trained Transformer 2)* on suure tekstikorpuse peal treenitud süvaõppe mudel, mis suudab genereerida inimesele sarnast teksti. See mudel on võimeline genereerima ja mõistma keerulist teksti [42].

- *RNN-d (Recurrent Neural Networks)* on mõeldud järjestikuste andmete töötlemiseks, kus iga sõlm sõltub eelmistest. Need närvivõrgud sobivad järjestikuste andmete, näiteks teksti töötlemiseks ning suudavad mõista konteksti ja sõltuvusi aja jooksul [43].

- *CNN (Convolutional Neural Network)* on algoritm, mis tuvastab mustreid ja mida kasutatakse tavaliselt piltide töötlemiseks. Kuid *CNN*-d on tõhusad ka teksti tükeldamisel ja kohalike mustrite tuvastamisel, näiteks lühikeste lausete analüüsimisel. Seetõttu on *CNN*-id sobiv valik tekstide lokaalsete mustrite tuvastamiseks [43].

Klassikalised masinõppe mudelid:

- Logistiline Regressioon, *SVM (Support Vector Machine)* ja *Naive Bayes* pakuvad traditsioonilisemat lähenemist teksti klassifitseerimisele. Logistiline regressioon on hea lihtsate klassifitseerimisülesannete jaoks. *SVM* on tõhus keerukate klassifitseerimisülesannete lahendamiseks, eriti väiksemate andmekogumite puhul. *Naive Bayes* on kiire ja efektiivne tekstide klassifitseerimise jaoks kasutamisel, eriti suure mahuga andmete puhul [15].

- Lineaarne regressioon on mõeldud pidevate sihtmuutujate jaoks ja ei sobi klassifitseerimisülesannete lahendamiseks. See modelleerib sõltuva muutuja ja sõltumatu muutuja vahelist seost lineaarse võrrandi abil. Lineaarset regressiooni kasutatakse tavaliselt prognoosimiseks ja trendianalüüsiks, mitte klassifitseerimiseks [15].

- K-lähedaseimad naabrid *(KNN)* on meetod, mis klassifitseerib objektid nende lähimate naabrite alusel. Kuigi seda saab kasutada klassifitseerimiseks, on see arvutuslikult kallis ja suurte andmekogumite puhul halvasti skaleeritav. Samuti ei sobi see hästi suuremahuliste andmete, näiteks teksti puhul [44].

- *K-means* on klasterdamisalgoritm, mis jaotab andmed k-klastrisse, kus iga andmepunkt kuulub lähimasse klastrisse. Seda kasutatakse pigem andmete jagamiseks eraldi rühmadesse kui konkreetse klassi ennustamiseks.

Klasterdamine aitab leida sarnaseid andmepunkte, kuid ei anna otsest klassifitseerimistulemust [45].

- *PCA (Principal Component Analysis)* ja muud dimensiooni vähendamise tehnikad on mõeldud tunnuste vähendamiseks ja visualiseerimiseks, mitte otseseks klassifitseerimiseks. *PCA* vähendab andmehulga dimensioonide arvu, säilitades samal ajal võimalikult palju andmehulga varieeruvust, et lihtsustada mudeli rakendamist ja tõlgendamist [46].

- Otsustuspuu on hierarhiline mudel, mis teeb otsuseid, jagades andmed korduvalt harudeks. See on lihtne ja sirgjooneline mudel, mis sobib hästi lihtsate klassifitseerimisülesannete lahendamiseks. Otsustuspuud on intuitiivselt mõistetavad ja nende otsustusprotsessi on lihtne jälgida. Selline mudel sobib andmete visuaalseks esitamiseks ja lihtsate klassifitseerimisülesannete lahendamiseks [15].

- *Random Forest* on koondmeetod, mis kasutab lõpliku otsuse tegemiseks mitut otsustuspuud, vähendades sellega liigse sobitamise riski. Seda kasutatakse keerukate klassifitseerimisülesannete jaoks, meetod suudab töödelda suuri andmekogumeid [47].

- *LDA (Latent Dirichlet Allocation)* on genereeriv tõenäosuslik mudel, mis kujutab dokumente teemadena ja teemasid sõnade seguna. *LDA* püüab leida dokumentides varjatud teemasid, määrates igale dokumendile teema ja igale teemale sõnade jaotuse. Seda kasutatakse sageli teemade modelleerimiseks ja varjatud struktuuride avastamiseks tekstikogudes [48].

- *LSA (Latent Semantic Analysis)* on algebraline mudel, mis kasutab termin-dokumendi maatriksi vähendamiseks singulaarsete väärtuste dekompositsiooni *(SVD). LSA* tuvastab terminite ja dokumentide vahelised mustrid, vähendades müra ja parandades semantiliselt sarnaste dokumentide väljaselgitamist [49].

2.2.6 Ametite tuvastamise mudeli jaoks mudelite valik

Sobitusalgoritmide valimisel on oluline kaaluda, millised meetodid sobivad konkreetse ülesande jaoks kõige paremini. Käesoleva projekti jaoks olid valitud järgmised meetodid:

TF-IDF koos Cosine Similarity-ga, Doc2Vec, EstBERT, GPT, RNN, Naive Bayes, SVM, Random Forest, Logistiline Regressioon. Need meetodid on valitud nende parema täpsuse ja sobivuse tõttu tekstide klassifitseerimise ülesannete jaoks võrreldes teiste mudelitega.

Teised mudelid ja algoritmid ei sobi selle projekti jaoks hästi, sest need ei vasta konkreetsetele nõuetele või võivad olla arvutuslikult liiga nõudlikud:

- *Cosine Similarity* üksi ei ole klassifitseerimisalgoritm, vaid meetod sarnasuse mõõtmiseks. Seda saab kasutada koos *TF-IDF*-iga, kuid üksi ei anna piisavalt täpset klassifitseerimistulemust [37].

- *Word2Vec* ja *FastText* on võimsad sõnade sisseehitamise meetodid, mis keskenduvad üksikute sõnade tähenduse esitamisele kontekstis. *Word2Vec* toodab staatilisi vektoreid, mis tähendab, et igal sõnal on kontekstist sõltumata ainult üks esitus. *FastText* täiendab seda, lisades teavet sõna allüksuste kohta, kuid keskendub endiselt üksikutele sõnadele, mitte tervetele dokumentidele. Seega võib teksti üldine tähendus ja struktuur, mis on oluline ametite ennustamiseks tekstist, kaduma minna [27] [38].

- *GloVe* on samuti vektorruumi mudel, mis kombineerib sõnade assotsiatsioonide statistikat globaalsel tasandil, et luua sõnavektoreid. Kuigi *GloVe* suudab ekstraheerida mõningaid semantilisi tunnuseid, on see staatiline mudel ja ei pruugi alati arvesse võtta sõnade mitmetähenduslikkust erinevates kontekstides [39].

- Kuigi mitmekeelne *BERT* võib pakkuda paremat jõudlust baasmudelitega võrreldes, näitavad uuringud, et keelespetsiifilised mudelid, nagu *EstBERT*, suudavad saavutada kõrgemat täpsust ja paremat üldist jõudlust konkreetses keeles. *EstBERT*, mis on spetsiaalselt treenitud eesti keele andmetel, on suutnud ületada mitmekeelse *BERT*-i mitmetes olulistes ülesannetes, sealhulgas sõnaliikide ja morfoloogiliste tunnuste märgendamises, sõltuvusparsimises ja nimede äratundmises. Selline spetsialiseerumine võimaldab mudelil paremini mõista ja töödelda eesti keele spetsiifilisi omadusi ja nüansse, pakkudes seeläbi täpsemaid ja usaldusväärsemaid tulemusi [41].

- *CNN* on tekstijärjestuste mõistmisel vähem tõhus. *CNN*-id käsitlevad teksti kui lokaalsete tunnuste kogumit ja ei suuda loomulikult sõnade järjekorda ja sõltuvusi tabada. [43].

- Lineaarne regressioon on mõeldud pidevate sihtmuutujate jaoks ja ei sobi klassifitseerimisülesannete jaoks [15].

- *KNN*-i võib kasutada klassifitseerimiseks, kuid see meetod on arvutuslikult kallis ning see ei sobi hästi suuremõõtmeliste andmete, näiteks teksti puhul [44].

- *K-means* on klastrialgoritm ja ei ole mõeldud klassifitseerimisülesannete jaoks. Seda kasutatakse andmete jagamiseks eraldi rühmadesse, mitte konkreetse klassi ennustamiseks [45].

- Dimensiooni vähendamise meetodid, nagu *PCA*, on mõeldud tunnuste vähendamiseks ja visualiseerimiseks, mitte otseseks klassifitseerimiseks. Need võivad olla osa eeltöötlusmenetlusest, kuid ei ole omaette klassifikaatorid [46].

- Otsustuspuu on lihtne mudel, kuid selle kalduvus liigsobitamisele ja ebastabiilsusele muudab selle vähem sobivaks keerulisemate andmete ja ülesannete jaoks. Juhuslik mets, mis kombineerib mitu otsustuspuud, pakub suuremat stabiilsust ja täpsust, mistõttu on see parem valik keerulisemate klassifitseerimisülesannete jaoks [47].

- *LDA* ja *LSA* sobivad üldiste teemade tuvastamiseks suurtes tekstikorpustes. Kuid see muudab need vähem efektiivsemateks tööde sobitamiseks vajalike täpsete, sildipõhiste prognooside tegemiseks. Nende võimetus säilitada semantilist spetsiifilisust on selles kontekstis oluline piirang [48] [49]. *LDA* on tõenäosuslik mudel, mis modelleerib dokumente teemade seguna, kus iga teema on sõnade jaotus. Kuigi *LDA* suudab suurtes tekstikorpustes tõhusalt tuvastada teemasid, on need teemad üldised ja võivad seetõttu olla vähem kasulikud täpsete, sildipõhiste prognooside tegemiseks [48]. Kuigi *LSA* võib parandada semantiliselt sarnaste dokumentide väljaselgitamist, võib see kaotada spetsiifilist semantilist teavet, mis on oluline täpsete, sildipõhiste prognooside tegemiseks [49].

2.3 Koondmeetodid

Koondmeetodid *(Ensemble methods)* hõlmavad mitme masinõppe meetodi kombineerimist üheks prognoosimudeliks, et vähendada varieeruvust, eelarvamusi või parandada prognoose. Põhiprintsiip on see, et rühm nõrku algoritme saab kokku panna, et moodustada tugev mudel, parandades seeläbi prognooside stabiilsust ja täpsust [50].

2.3.1 Eelised

- Parem ennustuse täpsus: Üks peamisi põhjusi, miks valiti koondmeetodi mudeli, on potentsiaalne suurem täpsus. Üksikutel mudelitel võivad olla omad nõrkused andmete üldistamisel, kuid kombineerituna vähendavad nende erinevad vaatenurgad liigse sobitamise tõenäosust ja annavad tavaliselt usaldusväärsemaid prognoose [51].

- Vastupidavus müra suhtes: Töökirjeldused ja elulookirjeldused võivad täpsuse ja stiili poolest väga erineda, mis kujutab endast märkimisväärset müra ja varieeruvust andmetes. Koondmeetodid, eriti *bagging* ja *boosting*, on tuntud oma vastupidavuse poolest müra suhtes ja nende võime toota sellest hoolimata stabiilseid prognoose [52].

- Suure mõõtmelisuse käsitlemine: Tekstiandmed, eriti töökirjeldused ja elulookirjeldused, hõlmavad tavaliselt suure sõnavara tõttu kõrget ruumimõõtmelisust. Koondmeetodid saavad tõhusalt hakkama suure mõõtmega andmetega, navigeerides läbi tunnuste ruumi keerukuse tõhusamalt kui üks mudel [15].

- Erinevate mudelite kasutamine: Erinevad mudelid hõlmavad andmete erinevaid aspekte. Näiteks kui logistiline regressioon võib hõlmata lineaarseid seoseid, siis otsustuspuu võib paremini hõlmata mittelineaarset koostoimet. Kombineerides erinevaid mudeleid kasutatakse ära iga mudeli tugevused, et saavutada tulemuslikumad eesmärgid [53].

- Paindlikkus mudeli ehitamisel: Koondmeetodid pakuvad paindlikkust mudeli koostamisel. Need võimaldavad integreerida eri tüüpi mudeleid. Selline

paindlikkus on väga oluline, et tulla toime keeletöötlusülesannete mitmekesise iseloomuga, kus erinevad mudelid võivad olla kasulikud probleemi eri tahkudes [54].

2.4 Uurimislünk

Loomuliku keele töötlemine (*NLP*) on paljude kaasaegsete soovitussüsteemide lahutamatu osa, parandades otsuste tegemist erinevates valdkondades. Riiklikes tööturuasutustes saab *NLP* märkimisväärselt parandada töökohtade sobitamist, tuvastades ametid täpselt enesekirjelduste põhjal. Kuid praegused süsteemid ei suuda sageli mitte inglise keelt kõnelevaid elanikke ja erinevaid tööturge rahuldada. Käesolevas osas tuvastatakse oluline uurimislünk *NLP*-põhise ametite klassifikatsiooni kohandamisel Eesti tööturu jaoks, rõhutades emakeele töötlemist ja spetsiaalseid masinõppe tehnikaid.

NLP-süsteeme on laialdaselt välja töötatud selliste sektorite jaoks nagu e-kaubandus ja meelelahutus (nt *Amazon* ja *Netflix*), kus on saadaval rohkelt andmeid kasutajaga interaktsiooni jaoks. Eesti riiklikes tööturuasutustes toetuvad need süsteemid tavaliselt lihtsamatele mudelitele, mis ei kajasta piisavalt keerukaid tööotsijate profiile ega töökirjeldusi.

Peamine tuvastatud puudujääk on vajadus keerukate *NLP*-süsteemide järele. Eesti riiklike tööturuasutuste jaoks on selge vajadus eesti keele nüansse mõistva süsteemi järele. Enne kogu selles töös käsitletava e-töötukassa portaali süsteemi üle kirjutamist, peaks välja töötama mudeli, mis oleks võimeline töötlema eesti keelset sisendteksti ja tuvastama selle alusel inimesele sobivat ametit *ESCO* klassifikatsioonist.

Selle lünga lahendamiseks otsustati luua koondmudel, mis integreeriks mitut masinõppe ja *NLP* tehnikat. Selle mudeli eesmärk on töödelda eestikeelset teksti, kasutades eeltöötluseks tehnoloogiaid nagu *EstNLTK* ning tööga seotud tekstide mõistmiseks ja sobitamiseks täiustatud algoritme. Mudeli väljatöötamisel eeldatakse suurendada ametite klassifitseerimise täpsust, mõistes paremini eestikeelsete ametite ja *CV*-de semantikat.

Seniste uuringute puudujäägi kõrvaldamine on oluline, kuna aitab kaasa tehisintellekti ja tööturu laiema valdkonna arengule, keskendudes vähem uuritud keelele. Eesti jaoks kohandatud *NLP*-põhise ametite soovitussüsteemi väljatöötamisega oli selle projekti

raames loodud mudel, mis oleks kasulik tuleviku Eesti tööturul kasutatava soovitussüsteemi sees. Käesoleva tööga seotud teadustöö uurib võimalusi kaasaegse tööturu soovitussüsteemi rakendamist, mida tulevikus saaks laiendada ning Eesti e-töötukassa portaalis kasutusele võtta.

3 Töö protsess

Selles peatükis kirjeldatakse andmete kogumist, eeltöötlust ning keeletehnoloogiliste meetodite ja mudelite katsetamise protsessi. See töö osa on vajalik selleks, et saada mudelid, mis oskavad kõige varemini et tuvastada *ESCO* klassifikatsioonis loetletud ametite vasted. Käesolevas peatükis peale tööprotsessi kirjeldamise, hinnatakse erinevate meetodite sobivust ametite tuvastamiseks eestikeelse sisendteksti põhjal.

3.1 Andmete kogumine ja eeltöötlus

Käesoleva projekti algfaasiks oli andmete kogumine ja eeltöötlus, mis oli oluline, et töötada välja usaldusväärne ametite tuvastamise mudel, mis põhineb eesti keele sisenditekstidel. Saadud andmekogum kokku hõlmas 122 teaduslikku ametit *ESCO* andmebaasist ning 626 inimest kahes eraldi failis, kes on nende ametitega seotud, informatsiooni.

Andmete kogumine algas asjaomaste andmekogumite allalaadimisega *ESCO* veebisaidilt. Kuna eestikeelses versioonis puudus väli „*altLabels*“, mis annab ametitele alternatiivseid nimetusi ning mis on oluline element, et suurendada mudeli võimet tunnustada iga ametiga seotud erinevaid mõisteid, olid alla laetud andmekogumite nii eesti- kui ka ingliskeelne versioon. Probleemi lahendamiseks tõlgiti ingliskeelse andmekogumi „*altLabels*“ eesti keelde ja lisati eesti versiooni. Selline tõlkimine ja integreerimine tagasid terviklikuma andmekogumi, mis võimaldaks paremini ära tunda ja mõista erinevaid ametinimetusi ja nende variante.

Lisaks *ESCO* andmetele koguti õppejõudude andmeid Eesti ülikooli Taltech veebilehelt ning Etis.ee-st (Eesti Teadusinfosüsteem). Ülikooli veebilehelt saadi teavet õpetajate ametinimetuste ja töövaldkondade kohta, saadud informatsioon oli pandud vastavatesse veergudesse “*altLabels*” ning “*description*” eraldi failis. Andmeid rikastati veel Etis.ee-le juurdepääsuga, kus on kättesaadavad ülikoolide töötajate üksikasjalikumad elulookirjeldused. Nendest *CV*-dest võeti täiendavad kirjeldused iga isiku kohta, mis olid

ka pandud veergu nimega "*description*". Selline lähenemisviis tagas, et andmekogum kajastaks tegelikke elustsenaariume ja sisaldaks nüansirikkaid üksikasju iga isiku kohta.

Kuna järgmise etappidena loodud mudelite jaoks oli oluline andmete valideerimine, kuid ametite nimetused võetud ülikooli veebilehelt ei olnud *ESCO* ametitega võrdsed, oli otsustatud kasutada *ChatGPT* selleks, et esimese sammuna treenida selle konkreetsete 122 ametinimetuste baasil ning seejärel küsida selle määrama iga isiku sobivama ametinimetuse tema eluloo kohta kogutud teave baasil. Need määratud *ESCO* ametinimetustega kooskõlas ametid olid lisatud veergu nimega "*preferredLabel*", vastavalt *ESCO* klassifikatsiooni faili struktuuriga.

Eeltöötlusetapis oli oluline roll failide struktuuril. *ESCO CSV*-fail sisaldab selliseid veergu naguameti link *ESCO* klassifikatsiooni veebilehele (*conceptUri*), ameti nimi (*preferredLabel*), alternatiivsed nimetused (*altLabels*), kirjeldus (*description*), pakkudes struktureeritud andmete formaati. Samamoodi sisaldas andmete kogumise pärast reaalse inimeste andmefail kirjeldusi, ametinimesid ja alternatiivseid ametinimetusi, mis olid olulised mudeli treenimiseks ja hindamiseks.

Andmete eeltöötlus hõlmas mitmeid etappe teksti puhastamiseks ja standardiseerimiseks, et tagada selle sobivus analüüsiks ja mudeli treenimiseks. Esialgu muudeti kõik tekstiandmed järjepidevuse säilitamiseks väikesteks tähtedeks. Erimärgid, numbrid, soovimatud tühikud ja lingid eemaldati, et kõrvaldada müra ja vähendada teksti keerukust. Järgmise sammuna toimus teksti tokeniseerimine, mille käigus tekst jaotati üksikuteks sõnadeks ja fraasideks, mis seejärel lemmatiseeriti - protsess, mille käigus sõnad vähendatakse nende algvormi. See etapp on eriti oluline sõnavormide varieerumise käsitlemiseks ja selle tagamiseks, et mudel suudab ära tunda ühe ja sama sõna erinevaid käändeid. Kõik eeltöötlusega seotud protsessid olid realiseeritud *EstNLTK* kasutades [55].

Stop-sõnad, mis on tavalised sõnad, mis ei anna olulist tähendust (nagu „ja", „the" jne), eemaldati, kasutades eelnevalt määratletud eesti keele stop-sõnade nimekirja [56]. See samm aitas keskenduda tähenduslikumatele sõnadele, mis tõenäoliselt aitavad kaasa klassifitseerimisprotsessile. Lisaks tehti õigekirja parandus, et parandada tekstis esinevaid kirjavigu, mis parandas veelgi andmete kvaliteeti.

Kogu isikude elulugude faili eeltöötlusetapi jooksul esines probleem mitmekeelse teabe käsitlemisega. Kuigi käesolev töö keskendub eestikeelsele tekstile, käsitleti muudes keeltes esitatud teavet müraks ja seda ei eemaldatud (vt. Tabel 1).

Tekst enne eeltöötlust	**Tekst pärast eeltöötlust**
Keemia ja biotehnoloogia instituut, 2019 - osalesin Chalmersi Tehnikaülikoolis rahvusvahelisel pärmi geneetika ja molekulaarbioloogia konverentsil 2019 - osalesin bioinformaatika töötoas (GEMs) Tartu Ülikoolis 2020 - esitasin Tartu Ülikoolis bioinformaatika töötoas (proteoomika ja masinõpe) teemal rakukujutise analüüs konvolutsiooniliste närvivõrkude abil, 1. Bio- ja keskkonnateadused; 1.12. Bio- ja keskkonnateadustega seotud uuringud, näiteks biotehnoloogia, molekulaarbioloogia, rakubioloogia, biofüüsika, majandus- ja tehnoloogiauuringud 4. Loodusteadused ja tehnika; 4.6. Arvutiteadused CERCS VALDKOND T490 Biotehnoloogia; P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)	keemia biotehnoloogia instituut osalema chalmers tehnikaülikool rahvusvaheline pärm geneetika molekulaarbioloogia konverents osalema bioinformaatika töötuba gems tartu ülikool esitama tartu ülikool bioinformaatika töötuba proteoomika masinõpe teema rakukujutis analüüs konvolutsiooniline närvivõrk bio keskkonnateadus bio keskkonnateadus seotud uuring näide biotehnoloogia molekulaarbioloogia rakubioloogia biofüüsika majandus tehnoloogiauuring loodusteadus tehnika arvutiteadusedcercs valdkondt biotehnoloogia arvutiteadus arvutusmeetod süsteem juhtimine automaatjuhtimisteooria

Tabel 1. Võõrkeelseid sõnu sisaldava teksti eeltöötluse näide.

Pärast eeltöötlust salvestati puhastatud ja standardiseeritud andmed uutesse *CSV*-failidesse - üks *ESCO* ametite jaoks *(preprocessed_ESCO_file.csv)* (vt. Tabel 2) ja teine reaalse maailma andmete jaoks *(preprocessed_real_people_file.csv)* (vt. Tabel 3). Need failid moodustasid aluse mudeli treenimise ja hindamise järgmistele etappidele.

conceptUri	**preferredLabel**	**altLabels**	**description**
http://data.europa.eu/esco/occupation/01ffb917-98dc-48c1-91ad-93c4104e791d	biomeditsiinitehnika insener	biomeditsiin insener bme konsultant biomeditsiinitehnika nõustaja biomeditsiinitehnoloogia insener ekspert bme nõustaja	biomeditsiinitehnika insener kombineerima teadmine inseneriteadus põhimõte bioloogiline leid arendama raviprotseduur

		biomeditsiinitehnika ekspert bme ekspert biomeditsiinitehnika konsultant biomeditsiinitehnoloogia insener spetsialist biomeditsiinitehnoloogia insener biomeditsiinitehnika spetsialist biomeditsiinitehnoloogia insener nõunik bme spetsialist biomeditsiinitehnoloogia insener konsultant biotehnika insener biotehnika konsultant biotehnika nõustaja biotehnika insener ekspert biotehnika ekspert biotehnika insener spetsialist biotehnika spetsialist biotehnika insener nõunik biotehnika insener konsultant	ravim üldine tervishoid töötama lahendus algama tavapärane ravim komponent täiustamine implantaat väljatöötamine kude ravi

Tabel 2. *preprocessed_ESCO_file.csv* faili struktuur.

etisUrl	**preferredLabel**	**altLabels**	**description**
https://www.etis.ee/CV/Xxx_Xxxxx/est/	arhitektuuriõppejõud	vanemlektor	arhitektuur urbanistika akadeemia ühiskonnateadus kultuur kunstiteadus skulptuur arhitektuur visuaalne semiootika

			visuaalkultuuriuuring

Tabel 3. *preprocessed_real_people_file.csv* faili struktuur.

3.2 Erinevate meetodite rakendus ning nende tulemused

Tugeva elukutse tuvastamise mudeli väljatöötamise protsess hõlmas katsetusi erinevate masinõppe ja loomuliku keele töötlemise meetoditega. Neid mudeleid treeniti *ESCO* veebisaidilt saadud andmekogumite ning Taltech ja Etis.ee-st saadud õppejõudude andmete põhjal. Oli vajalik hinnata iga meetodi sobivust ametite tuvastamiseks eestikeelsetest sisendtekstidest. Selles osas kirjeldatakse mudelite rakendamist, hindamist ja valikut, järjestades need halvimatest parimateni.

1. *Naive Bayes* mudelil oli kõigist testitud mudelitest kõige madalam sooritusvõime, täpsusega 86% (vt. Joonis 5).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	0.73	0.89	0.80	18
arvutiteadlane	0.00	0.00	0.00	1
arvutiteaduse õppejõud	0.70	0.99	0.82	129
automaatikainsener	0.90	0.77	0.83	48
biokeemik	0.88	1.00	0.94	45
ehitusinsener	0.91	1.00	0.95	92
elektromehaanikainsener	1.00	0.88	0.93	64
energiaseadmete insener	1.00	0.95	0.97	41
geoloog	1.00	1.00	1.00	31
haridusteaduste õppejõud	0.00	0.00	0.00	1
keskkonnateadlane	1.00	0.56	0.71	18
kommunikatsiooni õppejõud	0.00	0.00	0.00	1
kunstiteaduse õppejõud	0.00	0.00	0.00	3
linnaplaneerija	0.00	0.00	0.00	4
maateaduste õppejõud	0.00	0.00	0.00	1
majandusteaduse õppejõud	0.88	1.00	0.94	29
matemaatikaõppejõud	0.00	0.00	0.00	7
meditsiinitehnika insener	0.74	0.56	0.64	25
meditsiiniõppejõud	0.00	0.00	0.00	7
mehhatroonikainsener	0.00	0.00	0.00	6
psühholoogiaõppejõud	0.00	0.00	0.00	1
sotsioloog	0.00	0.00	0.00	2
ärikorralduse õppejõud	1.00	0.55	0.71	11
õigusteaduse õppejõud	0.96	0.89	0.92	27
ökonomist	1.00	0.69	0.82	13
accuracy			0.86	626
macro avg	0.49	0.45	0.46	626
weighted avg	0.83	0.86	0.83	626

Joonis 5. *Naive Bayes* mudeli klassifitseerimisaruanne.

2. Logistiline regressioon saavutas hea tulemuse täpsusega 92% (vt. Joonis 6).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	0.72	1.00	0.84	18
arvutiteadlane	0.00	0.00	0.00	1
arvutiteaduse õppejõud	0.84	0.98	0.91	129
automaatikainsener	0.90	0.94	0.92	48
biokeemik	0.96	1.00	0.98	45
ehitusinsener	0.98	1.00	0.99	92
elektromehaanikainsener	0.97	0.95	0.96	64
energiaseadmete insener	1.00	1.00	1.00	41
geoloog	1.00	1.00	1.00	31
haridusteaduste õppejõud	0.00	0.00	0.00	1
keskkonnateadlane	1.00	0.78	0.88	18
kommunikatsiooni õppejõud	0.00	0.00	0.00	1
kunstiteaduse õppejõud	0.00	0.00	0.00	3
linnaplaneerija	0.00	0.00	0.00	4
maateaduste õppejõud	0.00	0.00	0.00	1
majandusteaduse õppejõud	0.91	1.00	0.95	29
matemaatikaõppejõud	0.00	0.00	0.00	7
meditsiinitehnika insener	0.86	0.96	0.91	25
meditsiiniõppejõud	1.00	0.29	0.44	7
mehhatroonikainsener	0.00	0.00	0.00	6
psühholoogiaõppejõud	0.00	0.00	0.00	1
sotsioloog	0.00	0.00	0.00	2
ärikorralduse õppejõud	1.00	0.82	0.90	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	1.00	0.92	0.96	13
accuracy			0.92	626
macro avg	0.54	0.52	0.52	626
weighted avg	0.89	0.92	0.90	626

Joonis 6. Logistilise regressiooni mudeli klassifitseerimisaruanne.

1. *EstBERT* mudeli täpsus on 93% (vt. Joonis 7).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	0.69	1.00	0.82	18
arvutiteadlane	0.00	0.00	0.00	1
arvutiteaduse õppejõud	0.93	0.96	0.94	129
automaatikainsener	0.91	0.83	0.87	48
biokeemik	1.00	1.00	1.00	45
ehitusinsener	1.00	1.00	1.00	92
elektromehaanikainsener	0.82	1.00	0.90	64
energiaseadmete insener	1.00	1.00	1.00	41
geoloog	1.00	1.00	1.00	31
haridusteaduste õppejõud	0.00	0.00	0.00	1
keskkonnateadlane	1.00	0.94	0.97	18
kommunikatsiooni õppejõud	0.00	0.00	0.00	1
kunstiteaduse õppejõud	0.00	0.00	0.00	3
linnaplaneerija	0.00	0.00	0.00	4
maateaduste õppejõud	0.00	0.00	0.00	1
majandusteaduse õppejõud	1.00	1.00	1.00	29
matemaatikaõppejõud	1.00	0.14	0.25	7
meditsiinitehnika insener	0.96	0.96	0.96	25
meditsiiniõppejõud	0.88	1.00	0.93	7
mehhatroonikainsener	0.00	0.00	0.00	6
psühholoogiaõppejõud	0.00	0.00	0.00	1
sotsioloog	0.00	0.00	0.00	2
ärikorralduse õppejõud	1.00	1.00	1.00	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	0.81	1.00	0.90	13
accuracy			0.93	626
macro avg	0.58	0.57	0.56	626
weighted avg	0.91	0.93	0.92	626

Joonis 7. *EstBERT* mudeli klassifitseerimisaruanne.

2. *RNN* mudel saavutas täpsuse 94%, mis on parem sooritust kui mõnedel lihtsamatel mudelitel (vt. Joonis 8).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	0.78	1.00	0.88	18
arvutiteadlane	0.00	0.00	0.00	1
arvutiteaduse õppejõud	0.90	0.99	0.94	129
automaatikainsener	0.90	0.98	0.94	48
biokeemik	0.96	1.00	0.98	45
ehitusinsener	0.98	1.00	0.99	92
elektromehaanikainsener	0.98	0.98	0.98	64
energiaseadmete insener	0.98	1.00	0.99	41
geoloog	0.97	1.00	0.98	31
haridusteaduste õppejõud	0.00	0.00	0.00	1
keskkonnateadlane	1.00	0.83	0.91	18
kommunikatsiooni õppejõud	0.00	0.00	0.00	1
kunstiteaduse õppejõud	1.00	0.33	0.50	3
linnaplaneerija	0.00	0.00	0.00	4
maateaduste õppejõud	0.00	0.00	0.00	1
majandusteaduse õppejõud	0.94	1.00	0.97	29
matemaatikaõppejõud	1.00	0.29	0.44	7
meditsiinitehnika insener	0.83	1.00	0.91	25
meditsiiniõppejõud	1.00	0.14	0.25	7
mehhatroonikainsener	1.00	0.33	0.50	6
psühholoogiaõppejõud	0.00	0.00	0.00	1
sotsioloog	0.00	0.00	0.00	2
ärikorralduse õppejõud	1.00	0.82	0.90	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	1.00	0.85	0.92	13
accuracy			0.94	626
macro avg	0.66	0.56	0.58	626
weighted avg	0.93	0.94	0.92	626

Joonis 8. *RNN* mudeli klassifitseerimisaruanne.

3. *SVM* saavutas täpsuse 95% (vt. Joonis 9).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	0.78	1.00	0.88	18
arvutiteadlane	0.00	0.00	0.00	1
arvutiteaduse õppejõud	0.95	0.95	0.95	129
automaatikainsener	0.89	0.98	0.93	48
biokeemik	0.96	1.00	0.98	45
ehitusinsener	1.00	1.00	1.00	92
elektromehaanikainsener	0.95	0.98	0.97	64
energiaseadmete insener	1.00	1.00	1.00	41
geoloog	1.00	1.00	1.00	31
haridusteaduste õppejõud	0.00	0.00	0.00	1
keskkonnateadlane	1.00	0.94	0.97	18
kommunikatsiooni õppejõud	0.00	0.00	0.00	1
kunstiteaduse õppejõud	1.00	0.67	0.80	3
linnaplaneerija	0.00	0.00	0.00	4
maateaduste õppejõud	0.00	0.00	0.00	1
majandusteaduse õppejõud	0.94	1.00	0.97	29
matemaatikaõppejõud	0.78	1.00	0.88	7
meditsiinitehnika insener	0.83	1.00	0.91	25
meditsiiniõppejõud	1.00	0.29	0.44	7
mehhatroonikainsener	1.00	0.33	0.50	6
psühholoogiaõppejõud	0.00	0.00	0.00	1
sotsioloog	1.00	0.50	0.67	2
ärikorralduse õppejõud	1.00	0.91	0.95	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	1.00	0.92	0.96	13
accuracy			0.95	626
macro avg	0.70	0.63	0.64	626
weighted avg	0.94	0.95	0.94	626

Joonis 9. *SVM* mudeli klassifitseerimisaruanne.

4. *GPT* saavutas täpsuse 97%, olles parem kui mõned lihtsamad mudelid (vt. Joonis 10).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	0.90	1.00	0.95	18
arvutiteadlane	0.00	0.00	0.00	1
arvutiteaduse õppejõud	0.92	1.00	0.96	129
automaatikainsener	0.98	0.98	0.98	48
biokeemik	1.00	1.00	1.00	45
ehitusinsener	1.00	1.00	1.00	92
elektromehaanikainsener	1.00	0.98	0.99	64
energiaseadmete insener	1.00	1.00	1.00	41
geoloog	1.00	1.00	1.00	31
haridusteaduste õppejõud	1.00	1.00	1.00	1
keskkonnateadlane	1.00	1.00	1.00	18
kommunikatsiooni õppejõud	0.50	1.00	0.67	1
kunstiteaduse õppejõud	1.00	0.67	0.80	3
linnaplaneerija	1.00	0.75	0.86	4
maateaduste õppejõud	0.00	0.00	0.00	1
majandusteaduse õppejõud	1.00	1.00	1.00	29
matemaatikaõppejõud	1.00	0.86	0.92	7
meditsiinitehnika insener	0.96	1.00	0.98	25
meditsiiniõppejõud	1.00	1.00	1.00	7
mehhatroonikainsener	1.00	0.17	0.29	6
psühholoogiaõppejõud	0.00	0.00	0.00	1
sotsioloog	0.00	0.00	0.00	2
ärikorralduse õppejõud	1.00	1.00	1.00	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	1.00	1.00	1.00	13
accuracy			0.97	626
macro avg	0.78	0.75	0.75	626
weighted avg	0.97	0.97	0.97	626

Joonis 10. *GPT* mudeli klassifitseerimisaruanne.

5. *Doc2Vec* mudeli täpsus oli 99%, mis näitab selle tugevust dokumentide vaheliste semantiliste seoste tabamisel (vt. Joonis 11).

	precision	recall	f1-score	support
	0.00	0.00	0.00	0
arhitektuuriõppejõud	1.00	0.94	0.97	18
arvutiteadlane	1.00	1.00	1.00	1
arvutiteaduse õppejõud	1.00	1.00	1.00	129
automaatikainsener	1.00	1.00	1.00	48
biokeemik	1.00	1.00	1.00	45
ehitusinsener	1.00	1.00	1.00	92
elektromehaanikainsener	1.00	0.98	0.99	64
energiaseadmete insener	1.00	1.00	1.00	41
geoloog	1.00	0.97	0.98	31
haridusteaduste õppejõud	1.00	1.00	1.00	1
keskkonnateadlane	1.00	1.00	1.00	18
kommunikatsiooni õppejõud	1.00	1.00	1.00	1
kunstiteaduse õppejõud	1.00	1.00	1.00	3
linnaplaneerija	0.67	1.00	0.80	4
maateaduste õppejõud	1.00	1.00	1.00	1
majandusteaduse õppejõud	1.00	1.00	1.00	29
matemaatikaõppejõud	1.00	1.00	1.00	7
meditsiinitehnika insener	1.00	0.96	0.98	25
meditsiiniõppejõud	1.00	1.00	1.00	7
mehhatroonikainsener	0.86	1.00	0.92	6
nan	0.00	0.00	0.00	1
psühholoogiaõppejõud	1.00	1.00	1.00	1
sotsioloog	1.00	1.00	1.00	2
ärikorralduse õppejõud	1.00	1.00	1.00	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	1.00	1.00	1.00	13
accuracy			0.99	626
macro avg	0.91	0.92	0.91	626
weighted avg	0.99	0.99	0.99	626

Joonis 11. *Doc2Vec* mudeli klassifitseerimisaruanne.

6. *Random Forest* mudeli täpsus on 100% (vt. Joonis 12).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	1.00	1.00	1.00	18
arvutiteadlane	1.00	1.00	1.00	1
arvutiteaduse õppejõud	0.99	1.00	1.00	129
automaatikainsener	1.00	1.00	1.00	48
biokeemik	1.00	1.00	1.00	45
ehitusinsener	1.00	1.00	1.00	92
elektromehaanikainsener	1.00	1.00	1.00	64
energiaseadmete insener	1.00	1.00	1.00	41
geoloog	1.00	1.00	1.00	31
haridusteaduste õppejõud	1.00	1.00	1.00	1
keskkonnateadlane	1.00	1.00	1.00	18
kommunikatsiooni õppejõud	1.00	1.00	1.00	1
kunstiteaduse õppejõud	1.00	1.00	1.00	3
linnaplaneerija	1.00	1.00	1.00	4
maateaduste õppejõud	1.00	1.00	1.00	1
majandusteaduse õppejõud	1.00	1.00	1.00	29
matemaatikaõppejõud	1.00	0.86	0.92	7
meditsiinitehnika insener	0.96	1.00	0.98	25
meditsiiniõppejõud	1.00	1.00	1.00	7
mehhatroonikainsener	1.00	1.00	1.00	6
psühholoogiaõppejõud	1.00	1.00	1.00	1
sotsioloog	1.00	1.00	1.00	2
ärikorralduse õppejõud	1.00	1.00	1.00	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	1.00	1.00	1.00	13
accuracy			1.00	626
macro avg	0.96	0.96	0.96	626
weighted avg	1.00	1.00	1.00	626

Joonis 12. *Random Forest* mudeli klassifitseerimisaruanne.

7. *TF-IDF* koos *Cosine Similarity*-ga mudeli täpsus on ka 100% (vt. Joonis 13).

	precision	recall	f1-score	support
	0.00	0.00	0.00	1
arhitektuuriõppejõud	1.00	1.00	1.00	18
arvutiteadlane	1.00	1.00	1.00	1
arvutiteaduse õppejõud	0.99	1.00	1.00	129
automaatikainsener	1.00	1.00	1.00	48
biokeemik	1.00	1.00	1.00	45
ehitusinsener	1.00	1.00	1.00	92
elektromehaanikainsener	1.00	1.00	1.00	64
energiaseadmete insener	1.00	1.00	1.00	41
geoloog	1.00	1.00	1.00	31
haridusteaduste õppejõud	1.00	1.00	1.00	1
keskkonnateadlane	1.00	1.00	1.00	18
kommunikatsiooni õppejõud	1.00	1.00	1.00	1
kunstiteaduse õppejõud	1.00	1.00	1.00	3
linnaplaneerija	1.00	1.00	1.00	4
maateaduste õppejõud	1.00	1.00	1.00	1
majandusteaduse õppejõud	1.00	1.00	1.00	29
matemaatikaõppejõud	1.00	0.86	0.92	7
meditsiinitehnika insener	0.96	1.00	0.98	25
meditsiiniõppejõud	1.00	1.00	1.00	7
mehhatroonikainsener	1.00	1.00	1.00	6
psühholoogiaõppejõud	1.00	1.00	1.00	1
sotsioloog	1.00	1.00	1.00	2
ärikorralduse õppejõud	1.00	1.00	1.00	11
õigusteaduse õppejõud	1.00	1.00	1.00	27
ökonomist	1.00	1.00	1.00	13
accuracy			1.00	626
macro avg	0.96	0.96	0.96	626
weighted avg	1.00	1.00	1.00	626

Joonis 13. *TF-IDF* koos *Cosine Similarity*-ga mudeli klassifitseerimisaruanne.

Naive Bayes on kiire ja efektiivne tekstide klassifitseerimisel, eriti suure mahuga andmete puhul. Kuid selle eeldus, et kõik tunnused on üksteisest sõltumatud, ei pea sageli paika tegelikus tekstianalüüsis. Tekstides on sõnade kontekst ja seosed väga olulised, mida *Naive Bayes* ei suuda hästi modelleerida, mistõttu võib see mudel jääda täpsuselt alla keerukamatele mudelitele, mis arvestavad sõnade vahelisi seoseid ja konteksti [15].

Logistiline regressioon sobib hästi lihtsate klassifitseerimisülesannete jaoks, kuid selle lineaarne olemus piirab selle võimekust käsitleda mittelineaarseid ja keerukamaid andmesuhteid. Ametite tuvastamine tekstide põhjal nõuab keerukamate semantiliste suhete ja kontekstide mõistmist, mida logistiline regressioon ei suuda piisavalt hästi modelleerida, mistõttu jääb selle täpsus alla keerukamatele mudelitele [15].

EstBERT, kuigi loodud spetsiaalselt eestikeelse teksti mõistmiseks ja töötlemiseks, ei suutnud pakkuda oodatud täpsust ametite klassifitseerimisel. Süvaõppe mudelid, nagu

EstBERT, vajavad hästi toimimiseks suurt hulka treeningandmeid, mida käesolevas töös ei ole. Keerulised ja spetsiifilised tööalased tekstid nõudsid sügavamat kontekstuaalset mõistmist ja kohandatud treeningandmeid, mida olemasolev EstBERT mudel ei suutnud piisavalt hästi katta [41].

RNN on mõeldud järjestikuste andmete töötlemiseks, kus iga sõlm sõltub eelmistest. Need närvivõrgud suudavad hästi käsitleda tekstide järjestikuseid sõltuvusi ja konteksti. *RNN*-il on raskusi pikaajaliste sõltuvuste käsitlemisega, mis võib piirata nende võimet täpselt tuvastada keerulisi semantilisi suhteid pikemates tekstides [43].

SVM on tõhus keerukate klassifitseerimisülesannete lahendamiseks, eriti väiksemate andmekogumite puhul. Kuid nende arvutuslik keerukus ja ressursinõudlikkus võivad piirata nende skaleeritavust suuremate ja keerukamate tekstide puhul. *SVM* võib olla vähem tõhus keerukate semantiliste seoste käsitlemisel, mis on vajalikud täpseks ametite tuvastamiseks [15].

GPT on võimas keelemudel, mis on loodud keerulise teksti genereerimiseks ja mõistmiseks. Selle üldine fookus teksti genereerimisel tähendab, et see vajab ulatuslikku peenhäälestamist ning suurt hulka treeningandmeid, et kohanduda spetsiifiliste klassifitseerimisülesannetega. Kuigi *GPT-2*, mis oli kasutatud mudeli loomiseks, suudab mõista ja toota keerulist teksti, võib selle võime täpselt klassifitseerida ametid jääda alla spetsialiseeritumatele mudelitele, mis on loodud just klassifitseerimisülesannete jaoks [42].

Need piirangud selgitavad, miks *EstBERT*, *Naive Bayes*, logistiline regressioon, *RNN*, *SVM* ja *GPT-2* ei pruugi pakkuda sama taset täpsust ja usaldusväärsust kui keerukamad ja paremini kontekstuaalselt teadlikud mudelid, mis suudavad paremini mõista ja käsitleda teksti sügavamat semantilist konteksti ja struktuuri.

Hindamistulemuste põhjal valiti lõplikuks mudeliks järgmised mudelid, tuginedes nende suurepärasele sooritusele ja täiendavatele tugevustele: *Doc2Vec*, *Random Forest* ning *TF-IDF* koos *Cosine Similarity*-ga. Selline lähenemine tagab tugeva ja usaldusväärse ametite tuvastamise mudeli, mis suudab tõhusalt tõlgendada ja analüüsida eestikeelset sisendteksti suure täpsusega.

3.3 Lõplik mudel ja tulemuste valideerimine

Lõpliku mudelina loodi koondmudel, mis on võimeline tuvastama ametinimetusi eesti keeles esitatud tekstisisendite põhjal. Mudel kasutab kolme erinevat masinõppemudelit: *Doc2Vec*, *Random Forest* ning *TF-IDF* koos *Cosine Similarity*-ga. Igaüks neist mudelitest toob kaasa oma unikaalsed eelised ja tugevdab üldist ennustusvõimekust, aidates saavutada kõrge täpsusega tulemusi.

Lõplik mudel töötleb eesti keeles esitatud *CV* teksti, analüüsides nende semantilist sisu, et tuvastada kõige sobivam ametinimetus *ESCO* klassifikatsioonist. Selle mudeli loomine ja kasutamine vastab projekti peamisele eesmärgile: töötada välja täpne ja usaldusväärne ametite tuvastamise mudel, mida saab tulevikus kasutada riiklikes tööturuasutustes, et parandada töökohtade sobitamise protsessi.

Mudel saavutas täpsuse 100%, mis tähendab, et kõik ennustused *ESCO* andmestikus olid korrektsed. See kinnitab mudeli võimekust täpselt klassifitseerida ametinimetusi, kasutades antud andmestikku. Klassifikatsiooniraportid näitasid, et mudel saavutas täpsuse 1.00 kõigi ametinimetuste puhul. Need tulemused kinnitavad mudeli praktilist rakendatavust ja usaldusväärsust, pakkudes tugevat ja usaldusväärset lahendust ametinimetuste tuvastamiseks (vt. Joonis 14).

	precision	recall	f1-score	support
IKT-uuringute valdkonna nõustaja	1.00	1.00	1.00	1
analüütilise keemia spetsialist	1.00	1.00	1.00	1
arheoloog	1.00	1.00	1.00	1
bioloogia õppejõud	1.00	1.00	1.00	1
biomeditsiinitehnika insener	1.00	1.00	1.00	1
ehitusinsener	1.00	1.00	1.00	1
energiaseadmete insener	1.00	1.00	1.00	1
geneetik	1.00	1.00	1.00	1
hambaarstiteaduse õppejõud	1.00	1.00	1.00	1
katsetamisinsener	1.00	1.00	1.00	1
keskkonnateadlane	1.00	1.00	1.00	1
kirjanduse õppejõud	1.00	1.00	1.00	1
kommunikatsiooni õppejõud	1.00	1.00	1.00	1
kosmeetikakeemik	1.00	1.00	1.00	1
meditsiini erialaõppejõud	1.00	1.00	1.00	1
meditsiinitehnika insener	1.00	1.00	1.00	1
metroloog	1.00	1.00	1.00	1
mineraloog	1.00	1.00	1.00	1
nüüdiskeelte õppejõud	1.00	1.00	1.00	1
psühholoogiaõppejõud	1.00	1.00	1.00	1
sotsiaaltöö uurija	1.00	1.00	1.00	1
tehnikaõppejõud	1.00	1.00	1.00	1
usuteadlane	1.00	1.00	1.00	1
üldarst	1.00	1.00	1.00	1
ülikooli teadusassistent	1.00	1.00	1.00	1
accuracy			1.00	25
macro avg	1.00	1.00	1.00	25
weighted avg	1.00	1.00	1.00	25

Joonis 14. Lõpliku koondmeetodi mudeli klassifitseerimisaruanne.

Lõplik mudel vastab edukalt projekti eesmärkidele, olles valmis kasutamiseks töö- ja koolitussoovituste täpsuse suurendamiseks ning tööturu dünaamilistele vajadustele vastamiseks.

4 Kokkuvõte

Käesoleva töö peamine eesmärk oli töötada välja ametite tuvastamise mudel eestikeelse sisendteksti põhjal. Projekti raames saavutati see eesmärk edukalt, kasutades mitmeid masinõppe ja loomuliku keele töötlemise tehnoloogiaid. Uus mudel suudab efektiivselt tuvastada *ESCO* klassifikatsioonis loetletud ametite vasted, mis parandab tööotsijate ja tööpakkumiste sobitamist.

Töö käigus katsetati erinevaid tehnoloogiaid ja keelemudeleid, et leida kõige sobivamad lahendused eestikeelse teksti analüüsimiseks. Kõige paremad tulemused saavutati, kasutades kombinatsiooni *Doc2Vec*, *Random Forest* ja *TF-IDF* koos *Cosine Similarity*-ga mudelitest. Need mudelid tagasid täpse ja usaldusväärse tulemuse, mis on kriitilise tähtsusega töösoovitussüsteemi jaoks.

Mudel suudab töödelda eestikeelseid elulookirjeldusi ning määrata nende põhjal sobivaim amet *ESCO* klassifikatsioonis. Mudeli täpsus on testimisel osutunud kõrgeks. Andmekogumi põhjal saavutati mudeli täpsuseks 100%. See näitab mudeli tugevat võimet tuvastada ametid erinevate tekstide põhjal, tagades usaldusväärsed tulemused.

Edasised sammud selle töö põhjal võivad hõlmata mudeli laiendamist ja täiendamist, et katta rohkem ameteid ja töövaldkondi. Kuigi käesolev töö keskendus teaduslikele ametitele, võiks tulevikus laiendada mudelit ka teistele sektoritele. Lisaks võiks kaaluda mudeli integreerimist riiklikesse tööturuasutuste süsteemidesse, nagu Eesti Töötukassa, et pakkuda paremaid töö- ja koolitussoovitusi kasutajatele.

Samuti võiks uurida võimalusi mudeli täiendamiseks, et see suudaks töötada mitmekeelse sisendtekstiga, mis võimaldaks veelgi laiemat kasutusala. Lisaks võiks kaaluda täiendavate masinõppe meetodite ja süvaõppe mudelite integreerimist, et parandada mudeli täpsust ja üldist jõudlust veelgi.

Lõpetuseks, töö tulemuste valideerimine ja pidev jälgimine on oluline, et tagada mudeli püsiv täpsus ja usaldusväärsus reaalsetes töötingimustes. See hõlmab ka kasutajate tagasiside kogumist ja mudeli kohandamist vastavalt sellele, et vastata paremini tööturu vajadustele ja kasutajate ootustele.

Kasutatud kirjandus

[1] C.-E. Bănescu, E. Țițan ja D. Manea, „The Impact of E-Commerce on the Labor Market,“ *Sustainability,* kd. 14, nr 5086, pp. 1-3, 2022.

[2] M. Horii ja Y. Sakurai, „McKinsey & Company,“ 1 Juuli 2020. [Võrgumaterjal]. Available: https://www.mckinsey.com/featured-insights/asia-pacific/the-future-of-work-in-japan-accelerating-automation-after-covid-19#/. [Kasutatud 5 Mai 2024].

[3] EURES Eesti, „EURES,“ [Võrgumaterjal]. Available: https://www.eures.ee/en/services. [Kasutatud 5 5 2024].

[4] X-Road, „X-ROAD® HISTORY,“ NORDIC INSTITUTE FOR INTEROPERABILITY SOLUTIONS, [Võrgumaterjal]. Available: https://x-road.global/xroad-history. [Kasutatud 10 5 2024].

[5] I. Arcelay, A. Goti, A. Oyarbide-Zubillaga, T. Akyazi, E. Alberdi ja P. Garcia-Bringas, „Definition of the Future Skills Needs of Job Profiles in the Renewable Energy Sector,“ *Energies,* kd. 14, nr 9, pp. 1-4, 2021.

[6] International Labour Office, „International standard classification of occupations. structure, group definitions and correspondence tables,“ International Labour Organization, Geneva, 2012.

[7] S. Rojas-Galeano, J. Posada ja E. Ordoñez, „A Bibliometric Perspective on AI Research for Job-Résumé Matching,“ *The Scientific World Journal,* kd. 2022, pp. 1-15, 2022.

[8] R. Mänd, „Soovitussüsteemide rakendamine tööpakkumisele sobiva töötaja leidmiseks Töötukassa iseteeninduskeskkonnas,“ Tallinna Tehnikaülikool, Tallinn, 2015.

[9] M. Liutkevičius ja S. B. Yahia, „Research Roadmap for Designing a Virtual Competence Assistant for the European Labour Market,“ %1 *26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)*, Tallinn, 2022.

[10] International Labour Organization, „International Standard Classification of Occupations (ISCO),“ International Labour Organization, [Võrgumaterjal]. Available: https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/. [Kasutatud 3 5 2024].

[11] European Commission, „Mis on ESCO?,“ European Commission, [Võrgumaterjal]. Available: https://esco.ec.europa.eu/et/about-esco/what-esco. [Kasutatud 3 4 2024].

[12] M. Liutkevičius ja R. Erlenheim, „Validating the usage of Occupational Classification Systems in the Process of Creating a National Virtual Competency Assistant within the EU Labor Market,“ %1 *14th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2021)*, Athens, 2021.

[13] European Commission, „COMMISSION IMPLEMENTING DECISION (EU) 2018/1020 of 18 July 2018 on the adoption and updating of the list of skills, competences and occupations of the European classification for the purpose of

automated matching through the EURES common IT platform,“ Official Hournal of the European Union, Brussels, 2018.

[14] The European Union, „Regulation (EU) 2016/589 of the european parliament,“ 13 Aprill 2016. [Võrgumaterjal]. Available: https://eur-lex.europa.eu/eli/reg/2016/589/oj. [Kasutatud 10 5 2024].

[15] T. Hastie, R. Tibshirani ja J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.

[16] ESCO Publications, „ESCO strategic framework,“ European Commission, Brussels, 2017.

[17] P. Elias, Occupational classification (ISCO-88) : concepts, methods, reliability, validity and cross-national comparability, Paris: OECD, 1997.

[18] European Commission, „European Skills, Competences and Occupations classification Annual Report,“ European Commission, Brussels, 2021.

[19] European Commission, „How ESCO is used by Public Employment Services in Europe and in EURES,“ European Commission, [Võrgumaterjal]. Available: https://esco.ec.europa.eu/en/node/171. [Kasutatud 5 5 2024].

[20] European Commission, „How can ESCO be used in practice?,“ European Commission, [Võrgumaterjal]. Available: https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/how-can-esco-be-used-practice. [Kasutatud 3 5 2024].

[21] European Comission, „International Standard Classification of Occupations (ISCO),“ European Comission, [Võrgumaterjal]. Available:

https://esco.ec.europa.eu/en/about-esco/escopedia/escopedia/international-standard-classification-occupations-isco. [Kasutatud 3 5 2024].

[22] F. Incitti, F. Urli ja L. Snidaro, „Beyond word embeddings: A survey,“ *Information Fusion*, kd. 89, pp. 418-436, 2023.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser ja I. Polosukhin, „Attention Is All You Need,“ %1 *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, 2017.

[24] K. Clark ja C. D. Manning, „Improving Coreference Resolution by Learning Entity-Level Distributed Representations,“ %1 *54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, 2016.

[25] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee ja L. Zettlemoyer, „Deep Contextualized Word Representations,“ %1 *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, 2018.

[26] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng ja C. Potts, „Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,“ %1 *2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, 2013.

[27] T. Mikolov, K. Chen, G. Corrado ja J. Dean, „Efficient Estimation of Word Representations in Vector Space,“ %1 *Proceedings of Workshop at ICLR*, Scottsdale, 2013.

[28] J. Devlin, M.-W. Chang, K. Lee ja K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,“ %1 *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, 2019.

[29] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, K. Bontcheva ja J. Zhu, „The Stanford CoreNLP Natural Language Processing Toolkit,“ %1 *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, 2014.

[30] S. Orasmaa, T. Petmanson, A. Tkachenko, S. Laur, H.-J. Kaalep, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk ja Pi, „EstNLTK - NLP Toolkit for Estonian,“ %1 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, 2016.

[31] J. Friedman, R. Tibshirani ja T. Hastie, The Elements of Statistical Learning, Stanford: Springer, 2008.

[32] S. Russell ja P. Norvig, Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall, 2009.

[33] T. Berners-Lee, J. Hendler ja O. Lassila, „The Semantic Web,“ *Scientific American*, kd. 284, nr 5, pp. 34-43, 2001.

[34] C. J. Date, An Introduction to Database Systems, Pearson Education, 2004.

[35] S. Bird, E. Klein ja E. Loper, Natural Language Processing with Python, Sebastopol: O'Reilly Media, 2009.

[36] J. Ramos, „Using TF-IDF to Determine Word Relevance in Document Queries," %1 *First Instructional Conference on Machine Learning*, Piscataway, 2003.

[37] C. D. Manning, P. Raghavan ja H. Schütze, An Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2009.

[38] P. Bojanowski, E. Grave, A. Joulin ja T. Mikolov, „Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, kd. 5, pp. 135-146, 2017.

[39] J. Pennington, R. Socher ja C. Manning, „GloVe: Global Vectors for Word Representation," %1 *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, 2014.

[40] T. Mikolov ja Q. Le, „Distributed Representations of Sentences and Documents," %1 *31st International Conference on Machine Learning*, Mountain View, 2014.

[41] H. Tanvir, C. Kittask, S. Eiche ja K. Sirts, „EstBERT: A Pretrained Language-Specific BERT for Estonian," %1 *23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Reykjavik, 2021.

[42] A. Radford, K. Narasimhan, T. Salimans ja I. Sutskever, „Improving Language Understanding with Unsupervised Learning," OpenAI, 2018.

[43] I. Goodfellow, Y. Bengio ja A. Courville, Deep Learning, MIT Press, 2016.

[44] T. M. Cover ja P. E. Hart, „Nearest neighbor pattern classification,“ *IEEE Transactions on Information Theory,* kd. 13, nr 1, pp. 21-27, 1967.

[45] J. MacQueen, „Some methods for classification and analysis of multivariate observations,“ %1 *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, 1967.

[46] I. T. Jolliffe, Principal Component Analysis, Springer, 1986.

[47] L. Breiman, „Random Forests,“ *Machine Learning,* kd. 45, pp. 5-32, 2001.

[48] D. M. Blei, A. Y. Ng ja M. I. Jordan, „Latent Dirichlet Allocation,“ *Journal of Machine Learning Research,* kd. 3, pp. 993-1022, 2003.

[49] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer ja R. Harshman, „Indexing by latent semantic analysis,“ *Journal of the American Society for Information Science,* kd. 41, nr 6, pp. 391-407, 1990.

[50] T. G. Dietterich, „Ensemble Methods in Machine Learning,“ %1 *International Workshop on Multiple Classifier Systems*, Berlin, 2000.

[51] D. Opitz ja R. Maclin, „Popular ensemble methods: an empirical study,“ *Journal of Artificial Intelligence Research,* kd. 11, nr 1, pp. 169-198, 1999.

[52] E. Bauer ja R. Kohavi, „An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,“ *Machine Learning,* kd. 36, pp. 105-139, 1999.

[53] R. Polikar, „Ensemble based systems in decision making,“ *IEEE Circuits and Systems Magazine,* kd. 6, nr 3, pp. 21-45, 2006.

[54] L. Rokach, „Ensemble-based classifiers,“ *Artificial Intelligence Review,* kd. 33, p. 1–39, 2010.

[55] S. Laur, S. Orasmaa, D. Särg ja P. Tammo, „EstNLTK 1.6: Remastered Estonian NLP Pipeline,“ %1 *Proceedings of the Twelfth Language Resources and Evaluation Conference,* Marseille, 2020.

[56] K. Uiboaed, „Eesti keele stoppsõnad / Estonian stop words,“ 19 04 2018. [Võrgumaterjal]. Available: https://datadoi.ee/handle/33/78. [Kasutatud 10 03 2024].

[57] J. Devlin, M.-W. Chang, K. Lee ja K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,“ %1 *Proceedings of NAACL-HLT 2019,* Minneapolis, 2019.

[58] J. H. Friedman, „Greedy Function Approximation: A Gradient Boosting Machine,“ *The Annals of Statistics,* kd. 29, nr 5, pp. 1189-1232, 2001.

[59] D. H. Wolpert, „Stacked Generalization,“ *Neural Networks,* kd. 5, nr 2, pp. 241-259, 1992.

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Darja Manajeva

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Eestikeelse sisendteksti alusel ametite tuvastamise mudel“, mille juhendaja on Markko Liutkevičius.
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

27.05.2024

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.