

TARTU ÜLIKOOL
Sotsiaalteaduste valdkond
Johan Skytte poliitikauuringute instituut

Karl Hans Kostabi

Freedom House'i demokraatia indeksi ühtsuse analüüs BERT keelemudeliga

Bakalaureusetöö (6 EAP)

Juhendaja: Kristel Vits MA, Uku Kangur MA

Tartu 2024

Olen koostanud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite seisukohad, ning kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

Töö sõnade arv: 10170

Karl Hans Kostabi 20.05.2024

Abstract

This research examines Freedom House's democracy index, Freedom in the World (FITW). Academic literature has shown, that there are a lot of problems with academic literature. They have found to be biased and to have poor aggregation rules. The study investigates FITW using a machine learning-trained BERT language model, exploring how machine learning has been utilized in social sciences and its potential for further applications. BERT is used to create embeddings, that a vectorial representation of texts. Embeddings allow for a qualitative examination of the any texts. Since FITW index uses descriptive texts, these descriptive texts can be examined. Hence, this research examined, how FITW descriptions are tied to their scores. A hypothesis was created, that descriptive texts, that are most similar to one another, are also the ones with the highest scores.

The empirical findings demonstrate, that texts and score are indeed linked. This was done through creating a score difference index for all questions. Next, the score difference between the top 1% most similar texts was done. Through this it was shown that the score and texts are most strongly linked, when the score are high, and linked the weakest, when the scores are low. T-SNE method was used to show embeddings in 2-d projections. Through this the score were also proven to be linked visually. Visualizations were also used to show, that texts from the same regions of the world cluster together. Further geographical analysis revealed, that countries generally cluster according to their geographical locations, especially in the Americas and Europe.

Overall, the study concludes that FITW descriptions strongly correlate with high scores but weaken as scores decrease. Geographically, FITW descriptions are most consistent in the Americas and Europe, while the index is less uniform in Africa.

Sisukord

Sissejuhatus	5
1. Teoreetiline taust	6
1.1 Masinõppe kasutamine demokraatia uurimises.....	6
1.2 Demokraatia indekseerimine.....	7
1.3 Demokraatia indeksite kriitika.....	8
1.4 Freedom House-i „Vabadus Maailmas“ indeks.....	10
1.5 Uurimisprobleem	13
2. Uurimismeetod.....	14
2.1 Andmete kogumine	14
2.2 Vektorsitused.....	15
2.3 Koosinuse sarnasus	17
2.4 BERT keelemudel.....	18
2.5 Koosinuse sarnasus uuritavate tekstide vahel.....	20
2.6 Andmete visualiseerimine 2-dimensioonis.....	21
3. Tekstianalüüsi tulemused.....	23
3.1 Metodoloogia ja struktuur.....	23
3.2 Skooride erinevuse indeks.....	24
3.2.1 Skooride erinevuse indeksi analüüsi metodoloogia.....	25
3.2.2 Skooride erinevuse indeksi analüüs.....	27
3.2.3 Skooride erinevuse visuaalne analüüs.....	29
3.2.4 Skooride erinevuse tähelepanuväärsed riigid.....	29
3.3 Geograafiline analüüs.....	35
3.3.2 Geograafiline analüüsi metodoloogia.....	35
3.3.2 Geograafiline analüüs.....	35
3.4 Analüüsi kokkuvõte.....	37
4. Kokkuvõte	38
Viidatud kirjandus	41
Lisad	44

Sissejuhatus

Freedom House'i kodanike õiguste ja vabaduste indeks on üks pikaajalisemaid ning tuntumaid indekseid, mille tulemusi võrdsustatakse sageli demokraatia tasemega riigis ja mis leiab laialdast kasutamist ning refereerimist nii akadeemias kui ka väljaspool. Kasutajatele on kättesaadavad nii arvulised skoorid, mille põhjal antakse riigi režiimile hinnang „vaba“, „poolvaba“ või „mittevaba“, kui ka iga riigi või territoriaalse üksuse pikem profiil, mis avab selle hinnangu tausta. Riigiprofiilidest ilmneb, et igale alaindeksile on olemas ka antud skoor ja vastav kirjeldav tekst.

Demokraatia indeksite laialdane kasutus püstitab küsimuse nende usaldusväärtusest. Selgub, demokraatia indekseid on kritiseeritud nii demokraatia konseptualiseerimise, andmete kogumise, ideoloogilise kallutatuse kui ka indeksite agregatsiooni aspektis. Neid aspekte on demokraatia indeksites uuritud kvalitatiivselt mitmetel viisidel, isegi läbi masinõppe (Gründler ja Krieger 2021).

Masinõppe on üldiselt see, et automatiseeritult õpib programm järjest paremini tegema talle ette antud ülesannet, ilma et ta oleks selleks ülesandeks ise inimese poolt programmeeritud. Masinõppe teel on eeltreenitud välja ka mitmesuguseid võimsaid programme nagu näiteks piltide-tekstide klassifikatsiooniga tegelevad tekstidest võimsate keelemudeliteni välja. Keelemudelid on treenitud massiivsetel keelekorpusel aru saama keelt. Seni pole aga demokraatia indekseid uuritud läbi ega keeletehnoloogiliste meetodite, nagu keelemudelid.

Käesolevas töös uuritakse demokraatia indekseid Freedom House-i *Freedom in The World* demokraatia indeksi kaudu. Täpsemalt analüüsitakse töös kui üksühene on antud demokraatia indeks sisemiselt. Kõigepealt kirjeldatakse töös demokraatia indekseid üldisemalt, samuti tuuakse välja peamine kriitika nende kohta. Järgmisena seletatakse töös kuidas töötavad keelemudelid, ning kuidas nad antud töös teksti töötlevad. Tekstianalüüsi tehakse töös eeltreenitud keelemudeli BERT kasutades. Erinevate küsimustele vastavate kirjeldavate tekstide töötlemine BERTi abil laseb tekstide sarnasusi hinnata kvantitatiivselt ja visuaalselt. Visuaalsuse all mõeldakse käesolevas töös andmete (kirjeldavate tekstide) visualiseerimist 2d graafil nõnda, et tekstid oleksid klasterdunud nende semantilise tähenduse järgi. Hinnates tekstide sarnasusi visualiseerimise põhjal ning kvantitatiivsete arvud põhjal leiti, et kirjeldavate tekstide ja skooride vaheline side on osades küsimustes olematu, osades väga väike ning vähestes väga kõrge.

1. Teoreetiline taust

1.1 Masinõppe kasutamine demokraatia uurimises

Demokraatiat on keelemudelite ja masinõppega seni sotsiaalteadustes uuritud vähe. Ziems et al uurisid kokkuvõtvalt kõiki viise, kuidas keelemudeleid ning üleüldiselt arvutuslikku sotsiaalteadust saaks kasutada erinevates sotsiaalteaduste harudes. Oma töös märkisid nad, kuidas keelemudelid on kasulikud, kvantifitseerimaks erinevaid poliitilisi ideoloogiaid töödeldavatest tekstidest (Ziems et al 2023, 9). Varasemalt on masinõpet kasutatud ka näiteks poliitiliste kõnede emotsionaalsuse mõõtmisel (Ash & Gennaro 2021).

Grimmer *et al* (2021 lk 401-402) töid välja, et masinõppe kasutamise üheks eeliseks on multidimensionaalsete andmete (nagu tekst või pildid) „taandamine“ väiksematele dimensioonidele. Töödeldud madaladimensioonilisi andmeid on omakorda kergem kasutada seoste otsimisel (Grimmer *et al* 2021, lk 401). Ühtlasi töid nad välja, et masinõppe abil saab analüüsida väga suures koguses tekste, mis võimaldab luua ja analüüsida palju suuremaid andmestikke ning mõnel juhul lubab valida uusi, uurimuse eesmärgi jaoks paremini sobivaid andmestikke, mis omakorda viib spetsiifilisema teaduseni (*ibid*, lk 7; 15).

Kasulikuks saab selliste suurte andmekoguste hindamisel olla ka siirdeõpe. Siirdeõpe (ingl k. *transfer learning*) on idee, et kui treenida mudelit mingeid klassifitseerima andmeid (tekst, pildid jne), siis selle käigus õpib masin ka ära selle, missugused omadused korreleeruvad mingisuguste klassifikatsioonidega (*ibid*, lk 7). Näiteks tekstide klassifitseerimisel puhul oleks see nagu semantilise tähenduse ära õppimine, piltide puhul nagu teatud objekti kujude ära õppimine. Üldiselt on masinõppel tekstianalüüsis palju potentsiaali, peale klassifitseerimise on see kasulik ka näiteks tekstidevahelise sarnasuse mõõtmisel ja tekstide sõna-lause põhisel analüüsil (Chen et al lk 24). Demokraatia indekseid on masinõppega uuritud peamiselt kahe saksa poliitökonoomi, Gründleri ja Kriegeri poolt. Nad uurisid eelkõige indekseid skooride poolest ning võrdlesid demokraatia arengut eri majanduslike näitajatega leiti, et demokraatia indeksite tõusu ja majanduskasvu vahel on pikkajaline seos (Gründler & Krieger, 2018).

Kuigi on tehtud töid masinõppega teksti analüüsimisel politoloogias, siis konkreetselt käesoleva lõputöö teemal, milleks on ühe demokraatiaindeksi erinevate osade (alaindeksite) ja nende kirjelduste (eelnevalt „tekstid“) vastavuse analüüsimine, on jäänud seni uurimata. Seega julgen väita, et tegu on sotsiaalteaduste valdkonnas uudse teema ja lähenemisega.

1.2 Demokraatia indekseerimine

Demokraatiaindeksid on võrdleva poliitika tööriistad, mille eesmärgiks on kvantitatiivselt võrrelda demokraatia taset/režiimi tüüpi eri riikides. Demokraatia indekseerimise projektid said hoo sisse 20. sajandi keskel ja on muutunud aina levinumaks. Demokraatiaindeksid on laialt kasutusel akadeemias, täpsemalt võrdleva poliitika valdkonnas, kus selliste indeksite abil uuritakse demokraatiate või tema teatud omaduste kausaalseid toimimismehhanisme, kõrvutatakse režiime omavahel, või otsitakse korrelatsioone muude muutujatega. Näiteks on leitud seoseid riigi demokraatiaindeksite tulemuse tõusu ja majandustõusu vahel (Gründler & Krieger 2018) ning riigi religiooni sekkumise ja demokraatia vahel (Fox 2017). Väljaspool akadeemiat kasutatakse demokraatia indekseid näiteks poliitikas halduspoliitika kujundamisel või ajakirjanduses nii demokraatiate võrdlemisel kui ka riigi arengumustrite illustreerimiseks. Riigi demokraatianäitajad võivad mõjutada ka rahvusvahelist suhtumist sellesse riiki, rahvusvahelisi investeeringuid ning abiprogramme.

Demokraatia uurimine koosneb võrdlevas poliitikas enamasti kahest osast – demokraatia konseptualiseerimine ja selle mõõtmine (Graziano & Quaranta 2024, lk 2). Demokraatia uurimisele lisavad Munck ja Verkuilen peale konseptualiseerimise ja mõõtmise ka andmete agregatsiooni (Munck & Verkuilen 2002 lk 5-6). Just tugevalt ja hästi sõnastatud demokraatia definitsioon on see, millele peaks järgnema demokraatia mõõtmise metodoloogia (Boese 2019, lk 97).

Üks tuntumaid ja mõjukamaid demokraatia kontseptsioone tuli Robert Dahliilt 1971. aastal. Dahli käsitluses koosneb demokraatia õigusest osaleda valitsemises ja õigusest võimu vaidlustada. (Dahl 1971) Tema kontseptsioon demokraatiast oli võrdlemisi lai ehk ei hõlmanud ainult seda kuidas *de jure* institutsioonides osalemist ja vaidlustamist võimaldatakse, vaid ka seda kuidas *de facto* need riigis rakenduvad (Boese 2019, lk 98) Üks osa Dahli kontseptsioonist oli ka demokraatia ja autokraatia mõtestamine pideva muutujana ühesel kontinuumil. (Dahl 1971)

Sellele, kuidas demokraatia kontseptsioonile järgneb selle mõõtmine, on lähenetud mitmeti. Mõnedes käsitlustes luuakse alaindeksid ainult nominaalsete (nt Polity2) (Boese 2019, lk 97), enamuste ordinaalsete (nt Freedom House) muutujatega. (Munck & Verkuilen 2002, lk 10); (Boese 2019, lk 97) Tasub mainida, et Gründler ja Kriegeri arvates on Freedom House ja Polity 4 hoopis „kvaasi-pidevad muutujad“ (Gründler ja Krieger, 2021 lk 4). V-Dem projekt see-eest üritab teha pideval skaalal olevaid muutujaid. Mitmeid demokraatia omadusi on aga

raske mõõta kvantitatiivsete andmete puudumise tõttu, mistõttu neid mõõdetakse subjektiivsemalt, läbi ekspertide antud hinnangute teatud nähtustele või protsessidele (Gründler ja Krieger, 2021 lk 3). Sageli on takistuseks see, et omavahel võrreldavaid alandmeid ei ole võimalik koguda kõikide riikide kohta. Pole seega üllatus, et demokraatia indeksite kriitika keskendub enamasti hinnangute subjektiivsusele (Gründler ja Krieger 2021, lk 3) ja saadud andmete agregeerimisele (Boese 2019, lk 98-99). Demokraatia indeksite probleemidele ja kriitikale keskendutakse täpsemalt järgmises peatükis.

Siiski räägib demokraatiaindeksite kasuks asjaolu, et teadlased on leidnud, et erinevate indeksite vahel esineb üsna suur konstantsus. Seda vaatamata asjaolule, et indeksid on koostatud erinevate uurimisgruppide ja institutsioonide poolt ning kasutavad erinevaid lähenemisi. Seega, kui eri indeksid on üksteisest sõltumatult leidnud sarnaseid tulemusi, siis võib väita, et demokraatia mõõtmine ei olegi täiesti juhuslik (Graziano ja Quaranta 2024). Graziano ja Quaranta vaatlesid järgnevaid institutsioone ja nende indekseerimisprojekte: Freedom House, Polity, Economist Intelligence Unit, V-Dem, European Quality of Government ja Democracy Dataset ja jõudsid järeldusele, et kõik indeksid, välja arvatud Polity, korreleerusid omavahel tugevalt ja on seega riikide edasiste demokraatitrajektooride uuriseks võrdväärselt kasulikud (Graziano ja Quaranta 2024, 606).

1.3 Demokraatia indeksite kriitika

Demokraatiaindeksite kasvav populaarsus, sealhulgas ka massimeedias, kus indeksite tulemusi kasutatakse sageli kiirhinnangu andmiseks portreeteritavale riigile, tõstatab aga küsimusi nende tegelikust valiidsusest. Üldiselt saab jaotada demokraatia indeksite kriitika kolmeks: demokraatia kontseptualiseerimise, indikaatorite disaini (sh andmete kogumine) või andmete agregatsiooni kriitikaks.

Kontseptualiseerimise kriitika keskendub indeksite aluseks olevate kontseptsioonide analüüsile ja asjaolule, et demokraatia ise on vaidlustatud kontseptsioon. Näiteks on viidatud, et maksimalistlikud demokraatia kontseptsioonid võivad hakata liigselt kattuma muude valdkondadega, nagu näiteks majanduslik vabadus, kodanikuvabadused või õigusriiklus (Munck & Verkuilen 2002, lk 7; Gründler & Krieger, 2021 lk 2). Erinevatel demokraatia käsitlustel põhinevad indeksid saavad olla aga kasulikud, sest lasevad vastavalt indeksit kasutava uurija vajadustele valida tööks kõige sobilikuma indeksi (Gründler & Krieger, 2021 lk 2). Oluline on seega mõista kasutatava demokraatia indeksi metodoloogiat, et teada, kas antud indeks sobib teatud omaduse mõõtmiseks ja hindamiseks.

Samuti tekitab palju küsimusi see, mida üldse mõõdetakse? Ühelt poolt on probleemiks see, et andmed ei pruugi olla ühtmoodi kättesaadavad kõikide riikide kohta ning andmestikke, mis hõlmaks kõiki riike kõikidel ajaperioodidel, on vähe. Kuna indeksid ei taha kaotada universaalsust, siis on nende jaoks sobilikke andmestikke väga vähe. See võib tähendada, et mitmeid omadusi peab indeksit koostav meeskond iseseisvalt hindama (Munck & Verkuilen 2002, lk 13-14). Tihti kasutatakse sellisel juhul eksperthinnanguid, mis võivad aga suurendada hinnangute subjektiivsust (Gründler & Krieger, 2021 lk 3).

Indikaatorite disainis on levinud probleemiks veel see, et tihti väljendatakse demokraatiat või seotud nähtusi nominaalsel või ordinaalsel skaalal. Näiteks Freedom House kasutab sageli ordinaalseid skaalasisid: kui väärtus on antud skaalal 0–4, siis erinevatel väärtustel ei ole omavahelist statistilist suhet ning väärtuse 2 tegelik „kaugus“ väärtusest 4 ei ole tingimata pool. Boese (2019, lk 97) märgib aga, et empiirilistes uurimustes koheldakse indeksite nominaalseid ja ordinaalseid andmeid sageli kui intervallandmeid. Mitmed autorid soovivadki kasutada konstantseid intervallskaalasisid, et tagamada, et saadud tulemused oleksid statistiliselt võrreldavad ka skoori suuruse poolest (Gründler & Krieger, 2021 lk 3; Boese 2019 lk 97).

Andmete agregatsioon ehk alusandmete üheks tulemuseks koondamise viis on samuti kriitikat pälvinud. Boese (2019 lk 98) väidab, et demokraatiaindeksites, kus on tihtipeale palju erinevaid alakategooriaid ning ka alaindekseid, on nii kategooriate sisene kui ka kategooriate ülene andmete kokkuliitmine sageli subjektiivne. Iseäranis tugevalt esineb see probleem juhtudel, kus teadlased määravad andmetele kaalusid (Munck ja Verkuilen 2002, lk 22-23). Kuigi sellistel juhtudel peaks lähtuma teooriast, on agregatsioonireeglid tegelikult valitud arbitraarselt ning nende valikut pole õigustatud (Boese 2019 lk 98, Gründler & Krieger 2021 lk 2, Munck & Verkuilen 2002).

Iseäranis Gründler ja Krieger on alates 2015. aastast pühendunud erinevate poliitikaindeksite analüüsile. Samuti võtsid nad eesmärgiks luua masinõppe toel omaenda hübriidindeks, mis kombineerib juba eksisteerivaid indekseid ja nende muutujaid (Gründler & Krieger 2015). Autorid väitsid, et nende masinõppeline indeks agregeeris andmeid täpsemalt kui traditsioonilised indeksid, mis kipuvad andmete kehva agregeerimise tõttu demokraatia taset eri riikides kergesti üle- või alahindama (Gründler ja Krieger 2018; 2021). Nende indeksid olevat võrdluses täpsem, hilisemates versioonides on nad lisanud indeksile pidevate

väärtustega andmeid ning lisanud võimaluse arvutada usaldusvahemikku igale (ala)indeksile eraldi (Gründler ja Krieger 2021).

1.4 Freedom House-i „Vabadus Maailmas“ indeks

Käesolevas töös keskendutakse Freedom House'i poolt välja antava *Freedom in the World* ehk „vabadus maailmas“ indeksi (edaspidi FITW) sisemise ühtsuse analüüsimisele masinõppe, täpsemalt keelemudelite abil.

Freedom House on 1941. aastal USA-s asutatud demokraatia- ja inimõiguste edendamise ja uurimisega tegelev valitsuseväline organisatsioon. Nende missiooniks on „laiendada ja kaitsta vabadust globaalselt“. (Freedom House, „about us“) Selle eesmärgi täitmiseks koostatakse erinevaid uuringuid ja vabaduste-teemalisi aruandeid, mis on omakorda aluseks avalike poliitikasoovituste koostamisel globaalse demokraatia, sõnavabaduse ja inimõiguste edendamiseks (*Ibid.*). Freedom House'i koostatud indeksitest/aruanetist tuntuim on 1973. aastast välja antav iga-aastane kodanike õiguste ja vabaduste indeks ehk „vabadus maailmas“ indeks.

See indeks on käesolevas töös vaatluse alla võetud, sest tegemist on ühe prestiižseima ja samas ka laialdaselt kasutatava indeksiga nii akadeemias kui ka väljaspool. Selle populaarsusele aitab kaasa nii järjepidevus indeksi väljaandmisel kui ka suur kaasatud riikide ja territooriumite hulk. Kuigi indeks, nagu selle nimigi viitab, keskendub kodanike õiguste ja poliitiliste vabaduste ulatusele – ka selle hindamiskaala jaotab riigid lõpuks vabadeks, poolvabadeks ja mittevabadeks –, kasutatakse seda laialdaselt riikide demokraatlikkuse taseme indikaatorina. FITW raporti koostamine sai alguse Freedom House'i enda teadurite tööna, kuid on tänaseks kasvanud laiaulatuslikuks rahvusvaheliseks koostööproduktiks, mille valmimisele on kaasatud lisaks Freedom House'i enda analüütikutele ka akadeemikud, mõttekojad ja aktivistid väljastpoolt (Freedom House, *Freedom in the World Research Methodology*) Näiteks kõige hiljutisema, 2024. aasta FITW raporti koostamisega oli seotud Freedom House sõnul 132 analüütikut ja 40 nõustajat. (*ibid*) Kokkuvõttes võibki öelda, et FITW on eksperdi põhine indeks.

FITW ideoloogiliseks baasiks on ÜRO inimõiguste ülddeklaratsioon. Igat vaatlusalust riiki ja territooriumi hinnatakse võrdselt „sõltumata asukohast, etnilisest või religioosest koosseisust, või majanduslikust arengutasemest“ (Freedom House 2023, lk 35).

Metodoloogiliselt on valitud uuritavate temade kategooriad ühtlasi sarnased Dahli 1971. aastal välja käidud demokraatia definitsioonile. Oluliselt, raport hindab de facto, mitte de jure

õiguste eksisteerimist riigis (Freedom House, „Freedom in the World Research Methodology“). Teisisõnu, Freedom House eristab ja hindab FITW koostamisel nii seadustes olevaid õigusi kui ka nende tegelikku rakendamist. Nii on näiteks konfliktiolukorras tavapärane, et riigi punktisummat tuuakse allapoole, sest sõjaliste gruppide tegevuse tõttu on kodanike õigused de facto piiratud (*Ibid.*). Üldine akadeemiline konsensus on, et FITWI demokraatia konseptualiseerimine on pädev ja konstantne (Graziano & Quaranta 2024, lk 8).

Kuigi Freedom House on FITW raportit välja andnud aastast 1973, siis käesolevas töös on töö mahu piirangute tõttu vaatluse all vaid ühe aasta andmed, lõputöö koostamise algusaja mõttes viimasest ehk 2023. aasta raportist. See raport käsitleb olukorda maailmas aastal 2022. Raportisse oli kaasatud 195 riiki ning 15 territooriumi (piiratud iseseisvuse või rahvusvahelise tunnustustega üksused), töös keskendutakse 195 riigile. FITW indeksis avaldatakse peale skooride ka iga riigi indikaatori kohta skoori seletav lühikirjeldus, mille läbi saab skoori muutust või olukorda paremini mõista. Just nende lühikirjelduste pärast valitki käesolevas töös uurimiseks FITW indeks. Võimalus neid lühikirjeldusi analüüsida keelemudelite abil ongi peamine põhjus, miks käesolevas töös valiti uurimised FITW indeks.

FITW indeks koosneb kahest alaindeksist, mis on omakorda jaotatud mitmesse alakategooriasse. Iga alakategooria all on omakorda kolm kuni neli indikaatorit. Nii moodustuvad poliitiliste õiguste alaindeks, kuhu kuulub kokku 10 indikaatorit ning kodanikuvabaduste indeks, kuhu kuulub kokku 15 indikaatorit (Tabelid 1.1. ja 1.2). Iga indikaatori tulemuse määrab omakorda rida alaküsimusi, mida võib olla 3-13. Iga indikaatorit hinnatakse skaalal 0-4 punkti, seega võib poliitiliste õiguste alaindeks anda kuni 40 punkti, kodanikuvabaduste indeks kuni 60 punkti ning kokkuvõttes võib iga riik saada maksimaalselt 100 punkti. Erandkorras võidakse aga riigile anda ka negatiivne skoor, seda läbi lisaküsimuste (Freedom House, „Freedom in the World Research Methodology“). Näiteks on lisaküsimus, kas võõrvõim on muutmas selle riigialal etnilist koosseisu, ukraina valitsuse toimimise alakategoorias skoori alla toomas kahe võrra.

Tabel 1.1 Poliitilised õigused

(A) ELEKTORAALNE PROTSSESS	(B) POLIITILINE PLURALISM JA OSAVÕTLIKKUS	(C) VALITSUSE TOIMIMINE
A1 Kas valitsev valitsusjuht või riiklik juhtiv autoriteet on valitud vabadel ja õiglastel valimistel?	B1 Kas inimesed saavad organiseerida poliitilise väljundiga ühinguid; kas need ühingud on edu korral vabad süsteemi poolsetest takistustest?	C1 Kas vabalt valitud valitsusjuht ja/või seadusandlik kogu saavad määrata valitsuse poliitika? (9)
A2 Kas valitseva seadusandliku kogu liikmed valiti õiglastel ja vabadel valimistel?	B2 Kas opositsioonil on realistlik võimalus läbi valimiste suurendada enda toetust või saada võimu?	C2 Kas turvameetmed ametnike korrupsiooni vastu on tugevad ja efektiivsed?
A3 Kas elektoraalne seadusandlus ja raamistik on õiglased ja kas seda haldavad relevantsete organid?	B3 Kas inimeste poliitilised valikud on vabad sõjaväe, väliste jõudude, religioosete hierarhiate, majanduslike oligarhide jms mõjudest, kes ei ole demokraatlikult vastutavad?	C3 Kas valitsus toimib avatult ja läbipaistvalt?
	B4 Kas erinevad segmendid rahvastikust (etnilised, religioossed, soolised vähemused) omavad poliitilisi õigusi?	

Allikas: Siim Klais 2020

Tabel 1.2 kodanikuõigused

(D) VÄLJENDUS- JA USUVABADUS	(E) ÜHINEMIS JA ORGANISATSIOONI ÕIGUSED	(F) PERSONAALNE AUTONOOMIA JA INDIV. ÕIGUSED	(G) SEADUSANDLIK KORD
D1 Vaba ja iseseisva meedia olemasolu.	E1 Kogunemisvabadus ja demonstratsioonid.	F1 Kodanike liikumisvabadus – elukoht, töökoht, haridus	G1 Iseseisva kohtusüsteemi toimimine
D2 Kas kodanikud on vabad praktiseerima ja väljendama religioosseid tõekspidamisi?	E2 Olukord valitsusvälise vabakonnaga (sh. mis tegelevad inimõiguste ja valitsusemisega)?	F2 Kas kodanikud saavad võrdsetel tingimustel omada kinnisvara ja tegeleda ettevõtlusega?	G2 Kas tsiviil kui kriminaal küsimuse lahendamise käib nõuetekohaselt?
D3 Akadeemiline ja haridus süst. vabadus poliitilisest ettekirjutusest.	E3 Olukord ametiühingute ja tööalaste organisatsioonidega?	F3 Sotsiaalsed vabadused (partneri valik, pere suurus, koduvägivald, välimus).	G3 Vabadus üleliigse füüsilise jõu kasutamise eest.

D4 Kas kodanikud saavad väljendada vaateid poliitiliste ja muude tundlike teemade osas kartmata kättemaksu?		F4 Indiviidide vabadus majanduslikust eksploatatsioonist ja majanduslik ebavõrdsus.	G4 Kas seadus tagab kõigile kodanike gruppidele (k.a vähemustele) võrdse kohtlemise?
---	--	---	--

Allikas: Siim Klais 2020

Nagu eelpool kirjas, siis iga tabelis 1.1 ja 1.2 toodud indikaatori puhul kasutatakse alaküsimusi, mis abistavad selle indikaatori taseme määramisel. Alaküsimuste arv on toodud sulgudes iga indikaatori järel. Näiteks A1 indikaatori *Kas valitsev valitsusjuht või riiklik juhtiv autoriteet on valitud vabadel ja õiglastel valimistel?*“ puhul aitab skoorini jõuda 13 erinevat alaküsimust, mille hulgas on näiteks *„kas naised võisid osaleda kandidaatidena?“* ja *„kas valijatel oli võrdne ligipääs hääletuspunktidele?“* (Freedom House 2023, lk 2). Lisaks tasub mainida, et riigi olukorra analüüsimist ei alustata ühelgi aastal nullist, vaid seda tehakse eelmise aasta tulemust mustandina kasutades. Mustandi põhjal tuuakse vajadusel välja olulised muutused vaadeldavas riigis viimase aasta jooksul ja muudetakse vastavalt skoori (Freedom House, „Freedom in the World Research Methodology“).

Kuigi FITW indeks on väga populaarne ja laialdast kasutust leidev, ei ole seegi kriitikast puutumata jäänud. Nii toob Boese (2019 lk 8) välja, et Freedom House'i „vabaduse“ kontseptsioon on halvasti defineeritud ja indeks seetõttu subjektiivne. Siinjuures omab täiendavat mõju ka asjaolu, et alaküsimused on sõnastatud väga laialt, neile ei saa sageli vastata üheselt „jah/ei“ (*Ibid.* Lk 10). Gründler ja Krieger toovad välja, et osad teadurid kasutavad ainult Poliitiliste õiguste alaindeksit, sest selles pole nii palju subjektiivseid küsimusi (Gründler & Krieger, 2021, 6).

FITW on kritiseeritud ka selle poolest, et Freedom House ei avalikusta, kuidas konkreetse küsimuse (A1, A2 jne) puhul erinevate täpsustavate alaküsimustele vastused saadi või millised alaküsimuste (koond)vastused olid. Seetõttu pole võimalik iseseivalt uurida, kuidas teatud riik sai teatud skoori. Teisisõnu, ei ole võimalik järgi vaadata, kuidas alaküsimuste tulemused kokku agregeeriti (Munck and Verkuilen 2002, lk 25) Lisaks on need küsimused osade kriitikute sõnul teoreetiliselt nõrgalt põhjendatud (Boese 2017, lk 104). Skooride andmist on peetud ka arbitraarseks, sest pole FITW poolt põhjendatud, kuidas eri alaküsimuste vastused vastavad teatud skooridele (Gründler ja Krieger, 2021, 6).

Ka indeksi agregeerimise tüüpvead, nagu neid kirjeldati eespool, on FITW indeksis levinud. Näiteks Gründler ja Krieger tõid välja, agregeerimisel puudub kindel reeglistik, nende

seisukohast on probleemiks ka tulemuste usaldusintervallide puudumine (Gründler & Krieger, 2021, 6). Agregerimisel kasutab FITW ordinaalseid kategooriaid (Munck and Verkuilen 2002, lk 8), mis tähendab, et kuigi iga indikaatori puhul antakse tulemus skaalal 0-4, siis nende numbrite vaheline suhe pole arvuline.

Lisaks on probleemkohana välja toodud ka eelpool mainitud asjaolu, et uute analüüside koostamisel kasutatakse alusena eelmise aasta hinnanguid – see suurendab „rajasõltuvusse“ langemise ohtu (Boese 2019, lk 10). Teisisõnu, suur on oht, et juba ette antud kirjeldus suunab teatud suunda enda analüüsisiga. Indeksit tervikuna on kritiseeritud ka kallutatuses ja „Ameerika eliidi“ arvamuste kajastamises eri riikide kohta (Brown 2017, lk 12). Üks uuring leidis, et külma sõja ajal oli Freedom House'il tendents süstemaatiliselt alahinnata vabadust marksistlikes riikides (Bollen ja Paxton 2000 Boese 2019 lk 10 kaudu), kuigi Steiner (2014) on leidnud, et külma sõja järgselt on indeksi kallutatuse vähenenud.

ja nende kirjelduste (eelnevalt „tekstid“) vastavuse analüüsimine, on jäänud seni uurimata. Seega julgen väita, et tegu on sotsiaalteaduste valdkonnas uudse teema ja lähenemisega.

1.5 Uurimisprobleem

Oleme eelnevalt näinud, et demokraatia indeksitega on mitmesuguseid probleeme (demokraatia konseptualiseerimine, andmetiku valimine, tulemuste agregerimine jne). Seetõttu on siin töös võetud eesmärgiks uurida ühe demokraatiaindeksi – Freedom House'i „vabadus maailmas“ indeksi – üksühesust ehk sisemist ühtsust. Täpsemalt on töö keskseks uurimisküsimuseks, kas riikidele antavad skoorid on kooskõlas nende kohta koostatud pikemate kirjeldustega, sealhulgas erinevate sama skoori saanud riikide võrdluses, või on võimalik tuvastada ebäühtlust kirjelduste ja skooride võrdlemisel?

Kui on antud väga sarnaste tekstidega riikidele teatud küsimuses väga erinevad skoorid, siis see oleks märk madalast ühtlusest skooride ja tekstide vahel. See madal seos skoori ja teksti vahel saab omakorda olla märk sellest, et indeks ise pole sisemiselt ühtne. Vastavalt näitaksid väga sarnaste tekstide puhul samasugused skoorid seda, et skoorid ja tekstid on omavahel seotud. Seega püstitataksegi töös hüpotees, et sarnaste tekstide vahel on ka neile tekstidele vastavad skoorid sarnased. Kui nad on üksühesed, siis on sarnased tekstid ka üksühesed.

2. Uurimismeetodid

2.1 Andmete kogumine

Kuna Freedom House-i kirjeldatavate küsimuste kohta puudus vastav andmestik, siis pidi antud töö raames looma sobiliku andmestiku. Vajaminevad andmed saadi Freedom House-i veebilehtedelt. Kuigi riigipõhised raportid on olemas alates 2007. aastast, on nende formaat alles 2018. aastast avaldatud raportitest selline, kus Freedom House-i veebilehel on eri alaindeksitele on kõigile riikidel vastav kirjeldav tekst ka antud. Ehk siis iga riigi kohta on alates 2017. aastast eraldi välja toodud nende indeksid ja alaindeksid ning mõlemale vastavad kirjeldused. Sobilikult, on konstantsena püsinud küsimuste arv ja nende sõnastus alates 2017. aastast, lastes nii luua üksühese andmestiku. Seega saaks andmestiku teha maksimaalselt ajavahemikuks 2017-2023. Käesoleva töö raames koguti andmed kogus selles ajavahemikus, kuid uuritavaks andmestikuks võeti vaatluse alla ainult aasta 2023 raporti andmed.

Andmed koguti web-scrapides Freedom House-i veebilehte, kus leiab kõigi FITW-s olevale 195 riigile vastavaid andmed. Web scrapimise tegi hõlpsaks see, et nende veebilehtede URL-id olid selliselt tehtud, et riiginimi ja aastaarv olid mõlemad URL-is olemas (näiteks Prantsusmaa 2023 puhul, <https://freedomhouse.org/country/france/freedom-world/2023>). Seega sai kergesti automatiseerida andmekogumist iga riigi kohta nii, et vastavad kohad URL-is sai välja vahetada. Kõik kogutud andmed nendelt veebilehtedelt saavad olema ingliskeelsed. See on just vajalik, sest kogutud tekste töödeldakse keelemeetoditega, mis on just mõeldud ingliskeelseid tekste töötlemaks, sest nad on eeltreenitud inglisekeelele. Kõiki FITW-s kasutusel olevaid küsimusi saab näha varasemalt välja toodud tabelites 1.1 ja 1.2.

Web-scraperi sai tehtud selliselt, et kõik kogutud andmed tehti automaatselt csv- failiks ümber. Töö jooksul sai otsustatud, et indikaatoreid võiks vaadata suurima alaküsimuse kategooria kaupa, ehk siis A kategooria kolm küsimust, B kategooria neli küsimust jne. Neid küsimusi ja kategooriaid on näha eespool välja toodud tabelites 1.1 ja 1.2 *Freedom House*-i enda veebilehel eristatakse neid kategooriaid tähestikuliselt A-st G-ni. Vastavalt sai siis loodud need csv failid ka iga aasta kohta, ehk lõpptulemuseks ongi failid, mis on eristatavad aasta ja küsimuse kategooria põhjal. Näiteks faili „A2023.csv“ puhul näitab esimene täht failinimes tema küsimuste kategooriat ning järgnev aasta arv näitab, millisest aasta kohta need andmed käivad. Sellise faili esimesi ridu on kujutatud all oleval joonisel 1. Kuna A sektsioonis on kolm küsimust, siis on ka failis näha, et seal on kolm nõ „tekst-skoor“ paari.

Country	Tekst	Skoor	Tekst2	Skoor2	Tekst3	Skoor3
Afghanistan	The Taliban overthre	0	No legislative assembly or	0	Under the republic,	0
Albania	The president is the	3	The unicameral, 140-	3	The Central Election	3
Algeria	The president, who	1	The 407 members of the	1	Prior to 2019, Algeria's	1
Andorra	Andorra has a	4	Members of the unicameral,	4	The Electoral Law, which wa	4
Angola	There is no	0	Members of the 220-seat,	2	The law states that the	1
Antigua & B	The country's 1981 c	4	The bicameral parliament is	4	Electoral laws are generally f	4
Argentina	The constitution	4	The National Congress	4	Argentina has a clear and	3
Armenia	Constitutional	2	The National Assembly	2	Members of the Central	3
Australia	The Australian	4	The bicameral legislative	4	Australian electoral laws and	4
Austria	Executive elections	4	Legislative elections in	4	Austria's electoral laws are f.	4
Azerbaijan	The president is	0	The 125 seats in	0	The electoral laws and	0
Bahamas	The Bahamas is	4	Members of the lower	4	The electoral process is	4
Bahrain	The 2002 constitutic	0	The king appoints the 40-	1	Bahrain's electoral	1

Joonis 1 (Allikas: Autori andmed)

Probleemid tekkisid andmete kogumisel veebilehtedelt, kuna mõnel juhul oli veebilehe sisemine kood katki, mis takistas andmete korrektset kogumist web-scraperiiga. See põhjustas katkiste küsimuste teksti segunemise ja nihke teiste riikide andmetes. Probleemi lahendamiseks tuli andmed avada Exceli failina ja käsitsi vigased kohad parandada.

2.2 Vektorsitused

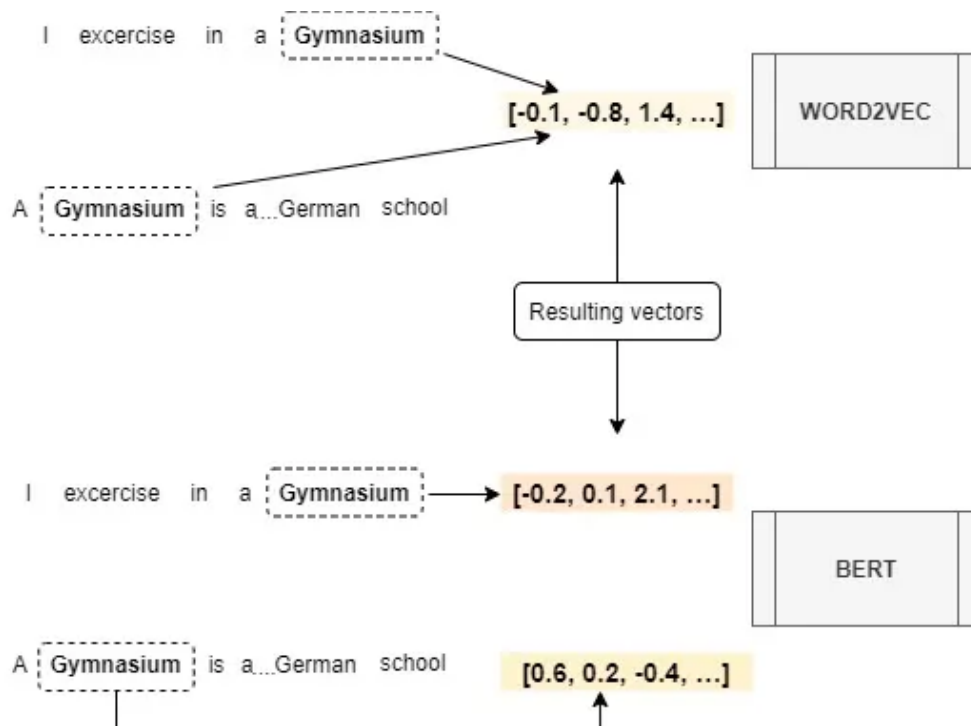
Vektorsitused (ingl *embeddings*) on teatud kujul oleva info (tekst, pildid, audio) muutmine numbrilisele kujule nii, et säiliks selle info semantiline tähendus. Teisiti öeldult, vektorsituste läbi saab muuta teksti numbrilisele kujule nii, et ta säilitab numbrilisel kujul enda sisemise semantilise tähenduse. Numbrilisel kujul olles saab seda teksti töödelda uutel viisidel, eriti keeletehnoloogiliste meetmetega. Kuigi vektorsitusi tehakse peale tekstide ka näiteks piltide või audioga (feature form, 2021), siis antud töö kontekstis keskendutakse ainult tekstiliste andmete vektorsitustele. Vektorsituste numbriline kuju väljendub vektori koordinaatidega N-mõõtmelises ruumis (Stackoverflow 2023). Need koordinaadid on loodud sellisel viisil, et nad oleksid keeletöötuse programmidele semantiliselt tähendust kandvad. Lühidalt, masinõppe teel treenitakse massiivsetel keeltekorpustel närvivõrk, mis säilitab enda närvivõrgus seda semantilist infot, millistes kontekstides mingisugune sõna esineb (*ibid*). Sellistel juba treenitud närvivõrku saab kasutada ka tagurpidi, et talle ette antud sisendile (sõnale, lausele) põhjal tagastada see info sellisel kujul, kuidas närvivõrk teda semantiliselt näeb. Sellise info lühendatud kuju nimetataksegi konkreetsetele tekstiliste andmetele vektorsituse loomiseks. Seega igale sõnale saab luua vastavat vektorsitust nõnda, et kõigis eri dimensioonides olevad numbrilised andmed väljendavad närvivõrgu jaoks semantiliselt midagi antud sõna kohta. (*ibid* 2023)

On mitmeid põhjuseid, miks vektorestitused on kasutusel ja eelistatud keeletehnoloogia valdkonnas. Üks vektorestituste omadusi on keerukamate andmete taandamine väiksemale kujule, mille kaudu saab tugevalt hoida kokku andmete mahtu. Näiteks, kui inglise keeles on umbes 6 miljonit sõna, siis nende kajastamine 6-miljonilise dimensiooniga vektoritena (ehk terve ingliskeelse sõnastiku esitamine niinimetatud *one hot vektor*-ina) oleks mälurohke ja arvutuslikult keerukas. Seega on mõttekas kasutada vektorestitusi, et see miljonites dimensioonides olev info taandada ainult paarisajale dimensioonile. (feature form, 2021) Vektorestitus ongi kompromiss selliste andmete kokkuvõtmises – andmete dimensiooni vähendatakse võimalikult palju ilma, et esineks nende sisu semantilises tähenduses suurt kadu. Väiksem, kuid olulise info tihedam andmeformaad on väga kasulik masinõppes, kus on tõhus ja kiire treenida mudeleid selliste andmetega. (*ibid*) Vektorestituste teksti esindav numbriline kuju on see, mis teeb võimalikuks treenida ka erinevaid keelemudeleid. (*ibid*)

Kaks enimlevinut tekstivektorite esitlusmudelit on Doc2Vec and BERT. Käesolevas töös on kasutatud BERT-i, sest selle mudeli vektorestitused püüavad tugevamalt kinni lausetes olevat konteksti. Kuigi nii BERT kui ka Doc2vec on Google'i poolt tehtud keeletöötusega seotud tööriistad, siis on nende vahel antud töö raamistikus olulisi erinevusi. Doc2vec põhineb Word2vecil, mis on 2011. aastal avaldatud inglise keelel eeltreentitud masinõppealgoritm, mis oli mille eesmärk ongi luua tekstidest vektorestitusi. Eeltreenimine tähendab, et treenimisel nähtud tekstide põhjal on ta loonud kokkuvõtva vektorestituse sellest, millistes kontekstides (ehk milliste sõnade lähedal) on ta vaadatavat sõna näinud. (Medium 2022) Kui Word2vec on ühe sõna vektorestituse loomine, siis Doc2vec on selle naturaalne edasiarendus, kus vektorestituse loomine laiendatakse ühelt sõnalt tervele lausele või lausa dokumendile. BERT on 2019. aastal avalikustatud närvivõrkudel põhinev keelemudel, mis on treenitud keelt aru saama juhendatud õppe abil. (Huggingface, „BERT“) Täpsemalt, mudelit treeniti täitma puuduvaid sõnalünkasid lausetes ja läbi selle protsessi ta õppis aru saama, kuidas esitada keelt vektorkujul võimalikult naturaalselt. (Medium 2022) Erinevus nende vahel ongi seega eelkõige see, et BERT-i vektorestitused arvestavad ka kontekstiga. Doc2Vec mudel ei võimalda eristada nende sõnade konteksti, millel on sama kirja pilt, kuid erinev kontekst. (*Ibid*)

Näiteks sõnal „tee“ on eelkõige kaks tähendust, kas teelehtedest tehtud jook või liiklemistee. Doc2vec treenimisel ei püüa meetod kinni eri „tee“ kontekste, vaid leiab nende kontekstide keskmise ja loob selle põhjal vektorestituse. BERT-i mudel on aga võimeline looma samale sõnale erinevad vektorestitused sõltuvalt ümber olevast kontekstist. Tasub ka välja tuua, et kui

BERT loob vektorestituse nõ kohapeal sõna ümber olevat konteksti vaadates, siis see tähendab, et sama sõna kohta saab ta sõltuvalt seda ümbritsevatest erinevatest sõnadest luua väga palju erinevaid vektorestitusi. Sama tähendusega sõnad sarnase tähendusega kuid erineva sõnastusega lausetes omavad aga omavahel sarnaseid vektorestitusi, sest kuigi individuaalsed sõnad nende ümber on erinevad, siis BERT suudab kontekstist aru saada, et tegu on väga sarnaste sõnadega ja seda väljendada sõnade vektorestitustes. Illustratiivselt on seda kujutatud ka joonisel 2, kus näha, kuidas inglise keeles sõnale „Gynasium“ teeks Word2vec samasuguse vektorestituse eri kontekstides, kuid BERT looks eri vektorestitused erinevates kontekstides. Kõike eelnevat arvesse võttes valitigi käesolevas töös vektorestituste tegemiseks BERT mudel.



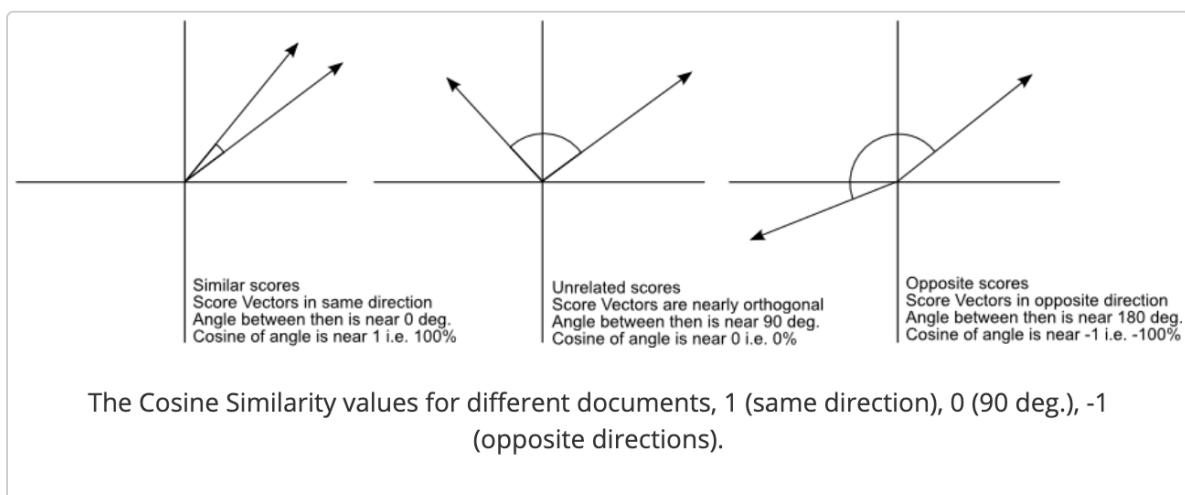
Joonis 2 (Allikas: Medium 2022)

2.3 Koosinuse sarnasus

Vektorestituste kaudu saab mõõta erinevate tekstide semantilist sarnasust. Seda tehakse läbi koosinuse sarnasuse (ingl k. *cosine similarity*). Koosinuse sarnasus on viis mõõta sarnasust kahe mittetühja vektori vahel. Nimi tuleb sellest, et saadud sarnasust näitav arv saadakse, kui võetakse koosinusfunktsioon kahe vektori vahelisest nurgast. Meetodi idee seisneb selles, et kui kaks teksti on sarnased, siis neil on sarnased vektorestitused. Sarnased vektorid on aga väga sarnasesse suunda näitamas ehk eeldatavalt ka sarnaste nurkadega. Nende sarnaste

vektoresituste vaheline nurk on välja arvatav ja väljendab endas seega ka tekstide sarnasust. Koosinuse sarnasus on väga hea meetod just vektoresituste sarnasuste vaheliseks mõõtmiseks. Kui hoopis mõõta vektorite sarnasust nende vahelise kauguse kaudu, siis muutub dimensioonide kasvades ka nende vaheline kaugus, kuid vektorite vaheline nurk püsib igas kahe- või enamamõõtmelises ruumis 0-180 kraadises vahemikus. (Emmery 2017)

All oleval joonisel 3 on näha näidet kolmest vektori paarist ja nende vahelisest koosinuse sarnasusest. Vasakult esimesel graafikul on näha, kuidas sarnase suunaga vektorite vahel on väikene nullilähedane nurk ja seega on tegu sarnaste vektoritega. Keskmisel teljestikul on kujutatud vektoreid, mille vahel puudub üldse suhe, ehk nendevaheline nurk on 90 kraadi lähedane. Parempoolsel graafikul on näha, et vastandlikud vektorid viitavad eri suundades. Kuna nurk on üle 90 kraadi, siis vastavalt on ka koosinuse nurgast andmas negatiivset väärtust. Olenevalt andmestikust saab vastassuunalisus tekstide vektoresituste vahel tähendada, et tegu on vastandliku semantilise tähendusega tekstidega (Github 2021).



Joonis 3 (Allikas: Perone Christian 2013)

Koosinusefunktsioon saadud nurkadest jääb alati vahemiku $[-1,1]$. Kõige sarnasemad on tekstid, kui nurk vektorite vahel on null ehk $\cos(0) = 1$. Analoogselt on koosinussarnasus maksimaalselt madal, kui vektorid on vastassuunalised ehk $\cos(180) = -1$.

2.4 BERT keelemudel

BERT (Bidirectional Encoder Representations from Transformers) on 2018. aastal valminud transformeritel põhinev keelemudel. Nagu eelnevalt öeldud, siis on BERT-i puhul tegu MLM (ingl „Masked Language modeling“) põhjal õpetatud mudeliga (Huggingface, „BERT“). See tähendab, et mudel on õpetatud lausetes ennustama sõnu niimoodi, et arvestab kontekstiks nii

sellele sõnale eelnevaid kui ka järgnevaid lause osasid. Kogu lause kontekstis arvestamine on üks põhjustest, miks BERT suudab luua informatsioonirohkeid vektoretsituse (Vaswani et al 2017). Peale MLM-i on BERT treenitud ka NSP (ingl „Next Sentence Prediction“) viisil, mille idee on õpetada BERT-i ennustama uut lauset olemasolevate põhjal.

BERT üks olulisemaid omadusi on see, et ta sisaldab endas transformereid, sest just transformerid on need, mis lasevad BERT-il olla treenitud MLM-i viisil. Transformerite kontseptsioon töötati välja 2017. aastal (Vaswani *et al*, 2017) tänaseks on transformerid aluseks kõigile suuretele keelemudelitele nagu ChatGPT, Claude ja Llama. Lihtsustades võib öelda, et transformerid kasutavad enesetähelepanu (ingl self-attention), et õppida ära kui olulise kaaluga on mingisuguse sisendi osa side erinevate semantiliste tähenduse osadega. Enesetähelepanus püütasegi optimeerida neid kaalusid, et saada lõpuks sisendiks olnud teksti erinevaid osasid võimalikult täpselt siduda erinevate BERT-i semantilise tähenduse mitmedimensionaalse ruumi punktidega. Nendele semantilise tähenduse ruumile viitavad vektorid ongi eelmainitud vektoretsitused (Deepai). Näiteks lauses „Tee peal oli tal käes tee“ jäid varasemad käsitlused hätta sellega, et ei suudetud eristada erinevaid teesid üksteistest. Enesetähelepanu abil aga tuvastab keelemudel, et need „tee“-d omavad erinevat semantilist tähendust nendes eri kontekstides. Seega teab transformer neid sõnu siduda erinevat tüüpi vektoretsitustega, mis mõlemad maksimaalselt püüavad kinni just selle „tee“ semantilist tähendust.

Märkimisväärselt on ka see, et BERT-i mudelid on eeltreenitud väga suurte tekstikorpustel. Nendeks olid BooksCorpus 800 miljoni sõnaga ja ingliskeelne vikipeedia 2.5 miljardi sõnaga. Antud töö kontekstis tähendab see, et käesoleva töö autor ei pidanud treenima enda keelemudelit, vaid sai kasutada eeltreenitud BERT-i mudelit. See on väga vajalik, sest antud bakalaaurusetöö raames oleks olnud võimatu nii suurte andmete peal õigeaegselt sellist mudelit treenida. BERT-i avalikult saadaval mudelitest eksisteerib tava- ja suur versioon, nimedega „BERT base“ ja „BERT large“. Suures versioonis on kolm korda rohkem parameetreid ja enkooderi tasemeid, mis teeb ta täpsemaks, kuid oluliselt aeglasemaks. Antud töö jaoks sai valitud tavaversioon.

Töö käigus genereeriti seega vektoretsitused kõigi kogutud andmete, ehk iga riigi igale alaküsimusele vastava teksti, kohta. Neid vektoretsituse kasutatakse töös eelkõige kahel eesmärgil. Esiteks sai siis kasutada eelmainitud koosinuse sarnasust, et saada kvantitatiivne hinnang sellele, kui sarnased on kaks teksti üksteisele. Teiseks, saab vektoretsituse esitada

graafidel, kus sarnaste tekstidega riigid eeldatavalt koonduvad leitavatesse klastritesse. Neid kahte meetodit kirjeldatakse järgnevatel peatükkides.

2.5 Koosinuse sarnasus uuritavate tekstide vahel

Uurimisküsimusele vastavalt on eesmärk töös uurida, kuidas on sarnaste tekstide vahel sarnased ka nende skoorid. Vaadeldavateks tekstideks võetakse FITW raportis konkreetsele küsimusele vastav kirjeldav tekst. Nende tekstide sarnasust mõõdetakse eelmainitud koosinuse sarnasuse kaudu. Skoorid saadakse vastavate alaindeksite puhul FITW raportist. Koosinuse sarnasuse meetodiga saadud arvu nimetatakse antud töös koosinuse sarnasuseks.

Kui tekstide vaheline sarnasus on suur, siis eeldatavalt on nende skoorid ka samasugused või vähemalt sarnased. Vastupidiselt, kui tekstide vahel oleks madal sarnasus, siis eeldatavalt oleks nende skoorid ka erinevad. Seega on vaja mõõta tekstide vahelist sarnasust eelnevalt kirjeldatud koosinuse sarnasuse abil ning siis mõõta, kuidas korreleeruvad tekstide skoorid tekstide sarnasusega.

Kuna koosinuse sarnasus töötab ainult kahe konkreetse teksti vahel, siis pidi analüüsi jaoks looma kõik võimalikud riikide paarid. Paaride arv ühe konkreetse aasta kohta on vastavalt valemile $n*(n-1)/2$ kus n =vaatlusaluste riikide arv. Vaadatavaid paare on seega $195*194/2 = 18915$. Need riigipaarid sai genereeritud iga individuaalse küsimuse kohta ja tehtud ka eraldi CSV failideks. Tehtud fail sisaldas seega kõigi paaris olevate riikide nimesid, tekste ja skooore ning nendevaheliste tekstide koosinuse sarnasuse indeksit.

Käesolevas töös sai otsustatud võtta uurimiseks kõige enimsarnaste tekstidega riigipaarid. Selle põhjuseks on soov hinnata skooride ja kirjelduste sarnasuste vahelist sidet. Eeldus on, et enimsarnaste riigipaaride skoorid on tõenäoliselt ka väga lähedased üksteisele. Kui enimsarnaste riigipaaride skooride erinevused kokku liita siis saadud arv iseloomustab seda, kui sarnased on nende skoorid üksteisele. Kui see on null, siis on skooride ja tekstide sarnasus üksühene. Mida kõrgemaks skooride erinevus läheb, seda väiksem on ka üksühesus tekstide ja skooride vahel. Enimsarnaste riigipaaride uurimise kaudu seega uuritakse kaudselt terve küsimuse kohta käivat kirjelduste ja küsimuste vahelist üksühesust. Veel tähtsamalt, seeläbi saab hinnata, kuivõrd on erinevates küsimustes üksühesed kirjeldavad tekstid ja skoorid. Valides „enimsarnased riigipaarid“ ühtlase lävendi põhjal, on küsimuste vaheliste skooride erinevuse summad omavahel võrreldavad.

Enimsarnaste riigipaaride saamiseks sai võetud kõigist riigipaaridest 1% kõige kõrgema koosinuse sarnasusega riikidest. See taandas valimi 18915 paarilt 190-le. Selline suur kitsendus võimendab just eeldust, et need riigipaarid on väga tugevalt omavahel sarnased selle konkreetse küsimuse valimi jaoks ning nende skooride erinevus on tähendusrikas. Analoogselt tehti ka vähimsarnaste riigipaaride valim, ehk valiti 1% kõige väiksema koosinuse sarnasusega riikidest.

Edasi sai nende enim sarnaste tekstipaaride puhul uuritud nende skooride erinevusi. Selle jaoks vaadati individuaalsete riigipaaride vahelist koosinuse sarnasust. Selleks, et oleks võimalik uurida, millised riigid olid oma väga sarnase paarilisega võrreldes väiksema/suurema skooriga, lisati iga riigi skooride vahe tema paarilistega eraldi *python*-i sõnastikku. Kui riik sai madala skoori selles sõnastikus, siis ta järelikul oli väga paljudes riigipaarides madala skooriga. Näiteks C1 küsimuses oli Serbia ja Montenegro tekstide vaheline koosinuse sarnasus 0.96, kuid nende skoorid olid vastavalt 3 ja 2. Sõnastikku oleks lisatud Montenegrole skoorile +1 ning Serbia skoorile -1.

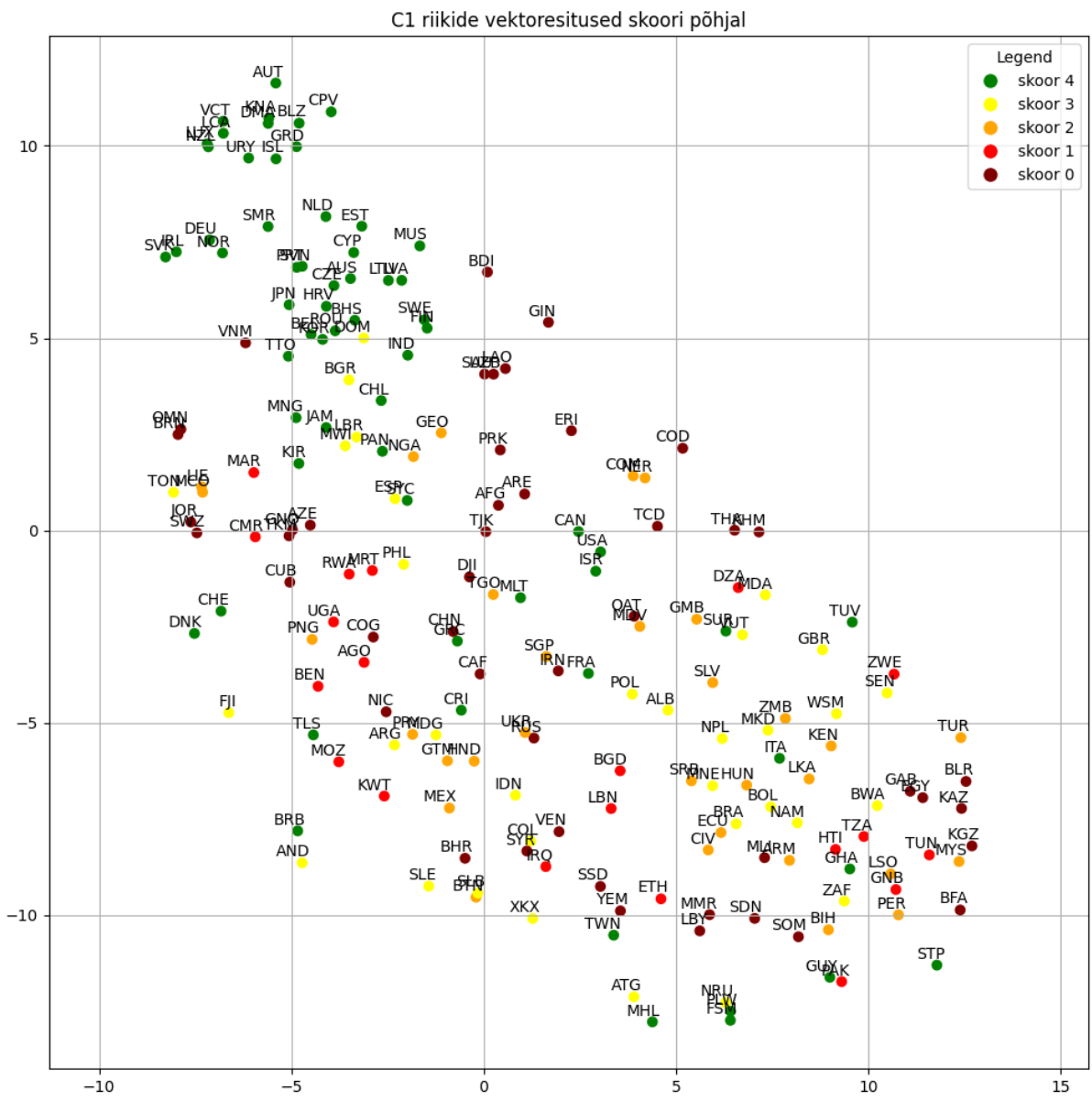
2.6 Andmete visualiseerimine 2-dimensioonis

Kuna iga teksti kohta käiv *embedding* ehk vektorsitus koosneb üle 50 eri dimensioonist, siis on tegu visualiseerimise poolest liiga dimensioonirohkete andmetega. Andmete visualiseerimine kahes või kolmes dimensioonis on üldiselt dimensioonirohkete andmete puhul kasulik sellepolest, et inimeste jaoks mugavalt visualiseeritud andmetelt on kerge leida ja näha mustreid. Näiteks antud töös oleksid nendeks mustriteks riikide vektorsitud eri rühmadesse ehk klastritesse koondumine. Tuletades meelde, et kõik need eri dimensioonid kokku esindavad vektorsitusele vastava teksti semantilist konteksti, siis ilmneb, et konteksti säilitamine võib olla raske, kui andmeid viia üle madalamale dimensioonile. Õnneks on juba varem tehtud ja arendatud välja meetodeid, mis suudavad selliseid vektorsituseid viia üle madalamatele dimensioonidele. Üks nendest meetoditest on t-SNE.

T-SNE on mitmemõõtmeliste andmete visualiseerimismeetod, mis põhineb mitmedimensionaalsete andmete projekteerimisel kahe- või kolmedimensioonilisse ruumi. Sealjuures üritatakse projektsioonis võimalikult hästi säilitada mitmedimensionaalsetes andmetes olevat konteksti. Lihtsustades öeldes leiab antud meetod, kui väga eri andmed on üksteisele sarnased ja üritab võimalikult hästi kõiki neid sarnasusi säilitada madalamasse dimensiooni viies. Võimalik oleks olnud kasutada ka PCA ehk peakomponendi analüüsi meetodit, kuid T-SNE valimise kasuks räägib tema suurem võimekus väiksemate lokaalsete

klasterite säilitamisel (riikide klasterdamine on üks töö eesmärkidest) ning on rohkem spetsialiseeritud meetod just andmete visualiseerimiseks. (Geeksforgeeks 2023)

Andmete visualiseerimist tehti seega iga konkreetse küsimuse kohta kõigi riikide tekstide vektorestitusti T-SNEsse sisestades. Saadud kahedimensionaalsetes graafides oli iga vektorestitus taandatud ühele punktile. Töö jooksul sai otsustatud vaatluse all oleva küsimuses olevate tekstide skoori väljenda graafil punkti värvi muutes. Värviskaala sai valitud tumepunasest (kui kõige halvem) roheliseni (kõige parem). Näitena on all toodud välja C1. küsimuse graafiku joonisel 4. Seal vasakul kõige ülemine riik on tähisega „AUT“ ehk Austria. Selle värv on roheline, ehk tegu on selles küsimuses kõige kõrgema skoori saanud riigiga.



Joonis 4 Näidis vektorestitustes joonistest (Allikas: Autori andmed)

3. Tekstianalüüsi tulemused

3.1 Metodoloogia ja struktuur

Uurimise alla võetakse 25 FITW indeksis kasutusel olevat indikaatorküsimust ehk **küsimust**. Täpsemalt on võimalik lugeda nende kohta Tabelis 1.1 ja 1.2. Selle tabeli põhjal nimetatakse igat vaatluse all olevat küsimust nende minimalistki nime järgi (A1, A2..jne). Iga riigi puhul on töö jaoks loodud andmebaasis igale küsimusele vastav tekst, mida nimetatakse järgnevalt **kirjelduseks**, ning vastava küsimuse eest saadud skoor vahemikus 0-4, mida nimetatakse järgnevalt **skooriks**.

Vastavalt peatükis 2.2 kirjeldatule luuakse iga kirjelduse kohta vektorestitus. Nendes vektorestitustes on talletatud kirjelduste semantiline info umbes 100 dimensionaalsete vektoritena. Kasutades peatükis 2.5 kirjeldatud T-SNE meetodit saab antud andmetega kergesti visualiseerida, milline on erinevate riikide kirjelduste semantiline sarnasus. Vektorestitused T-SNE-st läbi lastes saame seega graafid, kus on näha kõigi kirjelduste vaheline suhteline sarnasus. Saadud graafikutel näitavad mõlemad teljed kahepeale kokku, kui väga sarnased, või mitte sarnased on, on andmepunkt (ehk konkreetne kirjeldus) kõigi muude andmepunktide suhtes. Ehk saab graafide peale vaadates näha, kui sarnane on ükskõik milline kirjeldus semantiliselt teiste kirjeldustega võrreldes. Sarnase tähendusega kirjeldused peaksid seega keskmiselt kokku koonduma ehk klasterduma ning erineva tähendusega keskmiselt kaugel üksteisest olema.

Tekstianalüüsi esimeses osas uuritakse seega tekstide ja skooride vahelist seost. Teiseks töö eesmärgiks oli uurida, kuidas sõltuvad geograafiast nende kirjelduste vahelised sarnasused- erinevused. Selle põhjal on tekstianalüüsi teises osas uuritud, kuidas koonduvad 2-d graafidel geograafia alusel riikide kirjeldused.

Mõlemas tekstianalüüsi osas on graafikutel kujutatud vektorestitutes visualisatsioonis oluline värv ja riiginimed. Skoori analüüsis korreleerub värv andmepunkti skooriga, geograafilises analüüsis aga andmepunkti asukohaga. Kui tundub, et graafil on kõik andmepunktid eri värvides laiali, ilma koondumata siis see viitab sellele, et skooride vahel ja värvi omaduse vahel ei ole mingit sidet. Riiginimesid on ruumi kokkuhoiu eesmärgil lühendatud vastavalt ISO alpha-3 riigilühenditele. (Iban) Kokkuvõttes sisaldabki iga graaf skoorile vastavates värvides olevaid andmepunkte ja andmepunktidele vastavaid riiginimesid. Nimed on paigutatud graafidel nõnda, et nende nime lõpp oleks neile vastava andmepunktist ülevalpool.

3.2 Skooride erinevuse indeks

3.2.1 Skooride erinevuse indeksi analüüsi metodoloogia

Kuna antud töös on eesmärgiks uurida, kui üksühesed on uuritavad kirjeldused, siis on oluline võrrelda mitte ainult kirjelduste sarnasust paaride lõikes, vaid ka kirjeldustega kaasas käivaid skooore. Nimelt, kuna hüpotees on, et sarnastel kirjeldustel on sarnased skoorid, siis on vaja võtta uurimise alla igas küsimuses kõige sarnasemad kirjelduste paardid. Hüpoteesist järeldub ka see, et kõige erinevatel kirjeldustel oleksid keskmiselt ka erinevamad skoorid, kui enimsarnaste riigipaaride kirjelduste vahel. Seega uuritakse ka võrdluse jaoks kõige vähem sarnaseid riigipaare ka.

Edasi on tähtis nende enimsarnaste riigipaaride skoor. Võetakse iga riigipaaris oleva kahe kirjelduse skooride erinevus. Näiteks A1 küsimuses on Sambia skoor 2 ja Eesti skoor on 4. Selles riigipaaris oleks kirjelduste skooride erinevus seega nende kahe skoori vahe absoluutväärtus ehk $\text{abs}(2-4) = 2$. Sedamoodi saadi iga küsimuse kohta kõigi enimsarnaste riigipaaride skooride vahe absoluutväärtused kokku liites arv, mis iseloomustab siis seda, kui palju selles küsimuses skoorid erinevad. Nimetame seda **skooride erinevuse indeksiks**.

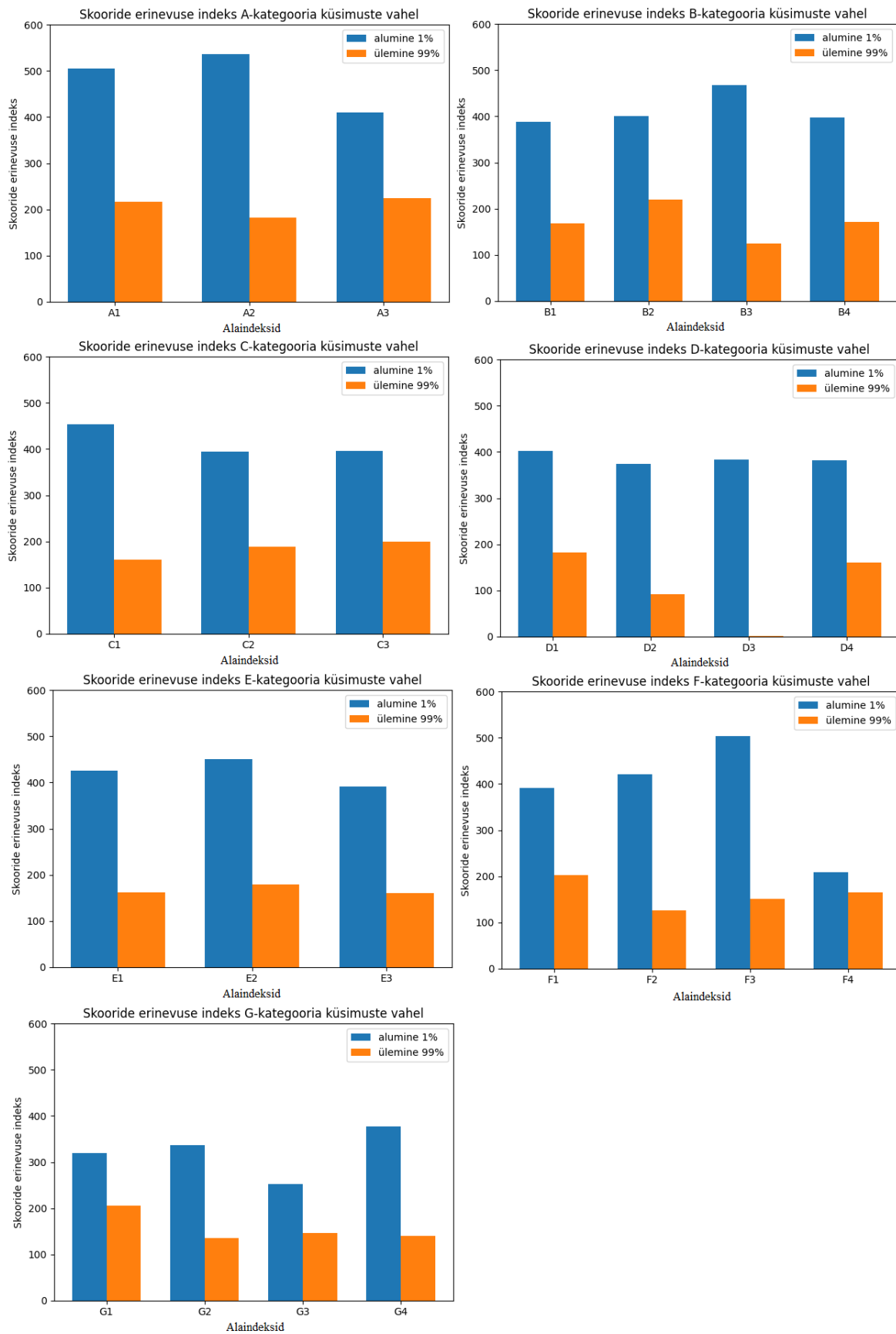
Skooride erinevuse indeks 0 tähendab, et kõigi vaatluse all olevate riigipaaride skoorid on omavahel võrdsed, ehk skoorid ei erine kunagi üksteisest. Mida kõrgemaks skooride erinevus läheb, seda suurem on riigipaaride skooride vahed. Skooride erinevuse indeksi alusel valitakse seega igast kategooriast üks küsimus, kus see on kõige kõrgem. Nõnda küsimust valides vaatame alati seda küsimust, kus on nõ kõige suurem konflikt sarnaste kirjelduste ja skooride vahel. Teisisõnu, võetakse nii vaatluse alla igas kategoorias see küsimus, kus peaks tulema kirjelduste ja skooride vaheline võimalik lahkeli kõige tugevamalt välja.

Seda tehakse läbi koosinuse sarnasuse, mida kirjeldatud täpsemalt peatükis 2.3 . Koosinuse sarnasusega saadakse 0-1 vahel olev arv, mis kirjeldab, kui sarnased on kaks kirjeldust (täpsemalt, kirjelduste vektorsitused) BERT-i keelemudeli arvates. Käesolevas töös seega leitakse koosinuse sarnasus kõigi vaatluseks olevate 195 riikide vahel (ehk ühes küsimuses 18915 riigipaaari vahel). Algselt oli plaanis käesolevas töös võtta ainult uurimise alla 2 standardhälve võrra kõige sarnasemad riigipaarid. See meetod oli aga probleemne, sest riigipaaride koosinuse sarnasused ei olnud samasuguste jaotustega. See tähendas, et kuna küsimustel olid erinevad jaotused, võeti eri arv riigipaare võrdlemiseks, mis tegi küsimuste vahelise võrdluse vähem tähendusrikkaks. Ühtlasi ei jaotu koosinuse sarnasus normaaljaotuslikult, mis tähendas, et osadel juhtudel 2 standardhälve võtmine andis 0

riigipaari. Kõigi riigipaaride vahel valitakse võrdluseks seega hoopis selles küsimuses 1% kõige suurema ja väiksema koosinuse sarnasusega riigipaarid.

Antud peatükis (3.2) uuritakse seega, milline on skooride erinevuse indeks igas 25. küsimuses tema enimsarnaste (ülemise 1%) ja vähemsarnaste (alumise 1%) riikide vahel. Selle jaoks luuakse tulpdiaagramm, kus saab näha, kuidas skooride erinevuse indeks erineb nii küsimuste, kui ka enimsarnaste ja vähemsarnaste riigipaaride vahel. Iga uuritava küsimuse kohta luuakse ka vektoreesisutes graafik. Nendest graafikutest tehakse ka eriversioon, kus on näha, millised riigid on nende enimsarnaste riigipaaride hulgas. Tehtud graafikutega on võimalik tutvuda LISA 1 lingi kaudu

Graafikutel on riikide skoorid ka visuaalselt eristatavad läbi värvikoodi. Kõige kõrgemale skoorile 4 vastab roheline, 3-le kollane, 2-le oranž, 1-le punane ja kõige madalamale skoorile 0 vastab tumepunane. Värvikood on toodud igal graafikul legendis välja.



Joonis 5 Skoorid erinevuse indeksid FITW indeksi küsimustes (Allikas: Autori andmed)

3.2.2 Skooride erinevuse indeksi analüüs

Joonisel 5 on kujutatud skooride erinevuse indeksi kõigi kategooriate vältel. Joonis koosneb seitsmest graafikust, kus kõigil on kujutatud alumist 1% skooride erinevuse indeksit sinises tulbana ja ülemist 1% punase tulbana.

Joonisel olevatest tulpdiagrammidest on näha, et kõigi küsimuste puhul on ülemine 1% skooride erinevuse indeks oluliselt madalam, kui alumise 1% puhul. Saab lausa väita, et enamustel juhtudel, on enimsarnaste riigipaaride skooride erinevuse indeks lausa poole võrra madalam, kui vähimsarnaste riigipaaridel. Enamus tulpdiagrammidel tundub olevat vähimsarnaste paaride skooride sarnasuse indeks konstantsem, kui on enimsarnaste riigipaaride skooride sarnasuse indeksid.

Kategooriaid omavahel võrreldes selgub, et tulemused on võrdlemisi sarnased. Näiteks kategooriad C ja E on väga sarnased üksteisele. Kategooriate sisesel on osadel juhtudel kategooriate puhul mõlemad tulbakategooriad omavahel sarnased. Sedasi on kategooriate A, B ja E puhul. Suurimad erinevused eksisteerivad kategooriates F ja G, kus kõiguvad nii ülemise- kui ka alumise 1% väärtus palju küsimuste vahel. D kategooria puhul oleks see erinevus ka väike, kui seal D3 küsimuses ei oleks üheks väärtuseks null.

Kõige kõrgem väärtus on skooride erinevuse indeks vähimsarnaste riigipaaride puhul A2 küsimuses, kus see on natuke üle 550. Peaaegu sama kõrge on ta ka F3 küsimuses. Kõige madalam väärtus on skooride erinevuse indeks vähimsarnaste riigipaaride puhul F4 küsimuses, kus see on 200 juures. Seal on ta ka kõige lähemal vastavale enimsarnaste riigipaaride skooride indeksile. Kui võrrelda F4 küsimust kõigi teiste küsimustega selles aspektis, siis tundub on vähimsarnaste riigipaarides vahel just skooride korrelatsiooni kasvab. Kõige kõrgem on enimsarnaste riigipaaride indeks küsimuses B2, kus see on natuke üle 200. See viitaks, et seal küsimuses on kõigist küsimustest kõige halvemini omavahel seotud tekstide sarnasus ning nende vaheline skoor. Kõige madalam on enimsarnaste riigipaaride puhul skooride erinevuse indeks küsimuses D3, kus see on null. D3 puhul viitab see siis sellele, et kõigi enimsarnaste riigipaaride puhul on skoor täpselt samasugune.

Skooride erinevuse indeks oli null väärtusega ainult D3 küsimuse juhul. Seda küsimust lähemalt uurides selgub, et selles küsimuses on kõik enimsarnased riigipaarid täpselt samasugused. Nimelt on küsimuseks *Akadeemiline ja haridus süst. vabadus poliitilisest ettekirjutusest* ja perfektse skoori puhul on selle kirjelduseks pandud peaaegu alati ainult

lause „akadeemiline vabadus on üldiselt austatud“ („*Academic freedom is generally respected.*“). Sarnaselt on palju perfektselt sarnaseid skooore ka D2 küsimuses, kus ka skooride sarnasuse indeks on vastavalt D3-e järgnevalt kõige madalam kõigi küsimuste vältel.

Tasub ka välja tuua, et F4 kõrval oleval F3 küsimuses on väga suur erinevus enimsarnaste ja vähimsarnaste riigipaaride tulpade vahel. Seda küsimust lähemalt uurides selgub, et seal on tegu väga pikkade kirjeldustega, mis käivad riigis oleva mitte legitiimse vägivalla kasutuse kohta. Need kirjeldused on tihti kirjeldamas riigis toimunud mitte legitiimseid vägivalla sündmusi või juriidilisi nähtusi. Antud küsimuses tuuakse välja aga väga palju eri tüüpi sellele küsimusele mitte vastamist. See võib tähendada, et kuna riigid saavad väga mitmetel viisidel selles küsimuses negatiivset skoori saada, siis skooride ja vektoreesisute vaheline seos on nõrgem madalamate skooride puhul. Kõrged skooride kirjeldused on tihti lühemad ja üksteisele sarnasemad. Teisisõnu, kuna negatiivsel skooril saab kirjelduses olla väga palju põhjendusi, ning kõrgetel skooridel on kirjeldustes vähe põhjendusi välja toodud, siis sattuvad enimsarnasteks skoorideks ka kõrgemad skoorid.

Kokkuvõtvalt saab öelda, et sai leitud, et enimsarnaste riigipaaride puhul on igas olukorras skooride erinevuse indeks madalam, kui vähimsarnaste riigipaaride puhul. See tähendab, et need enimsarnased riigipaarid on omavahel sarnasemate skooridega, kui vähimsarnaste riigipaarid. Kirjelduste vahelise sarnasuse ja kirjeldustele vastavate skooride vahel on seega kindel seos. Töös püstitatud hüpotees on seega antud peatükki alusel kinnitatud.

Töös on kõik Järgnevas peatükis analüüsitakse visuaalselt tehtud graafikuid. Konkreetselt tuuakse analüüsiks välja D3 ja F4. Seda sellepärast, et D3 oli väikseim skoorid erinevuse indeks väärtusega null. F4-s sai valitud, sest seal olid enimsarnaste ja vähimsarnaste riigipaaride skoorid erinevuse indeks kõigist küsimustest kõige sarnasemad.

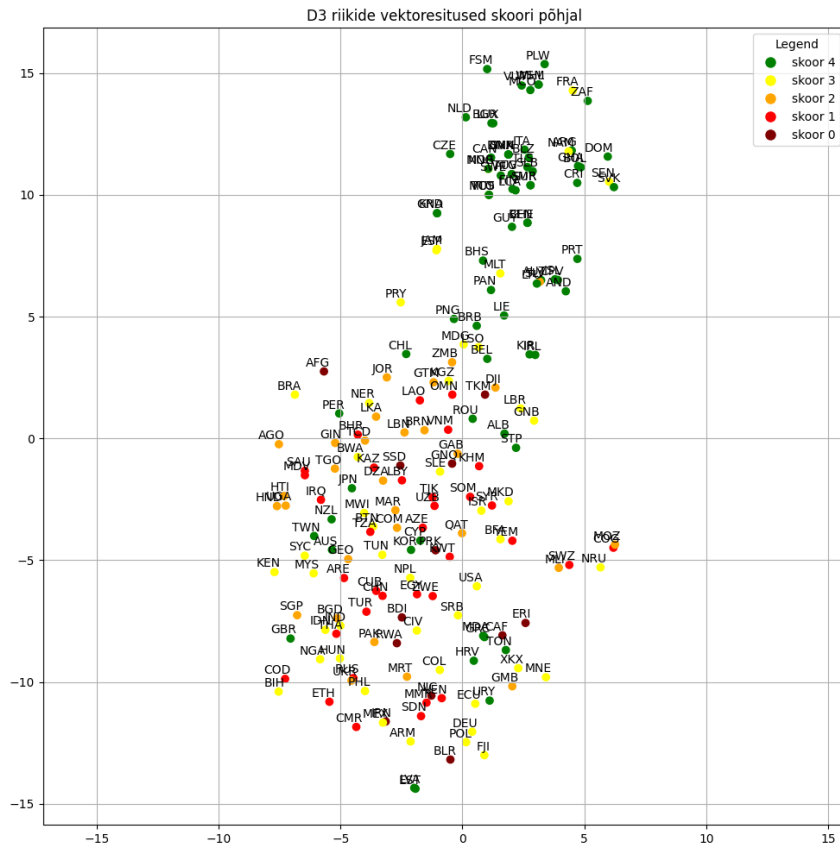
3.2.3 Skooride erinevuse visuaalne analüüs

Antud peatükis kirjeldatakse vaadatakse kõige pealt konkreetselt kolme küsimust (D3 ja F4). F4 küsimuse puhul tuuakse välja ka ainult nende enimsarnaste riigipaare näitavad vektorsitustse joonis, D3 puhul seda ei tehta, kuna tal olid kõigil neil samasugune skoor.

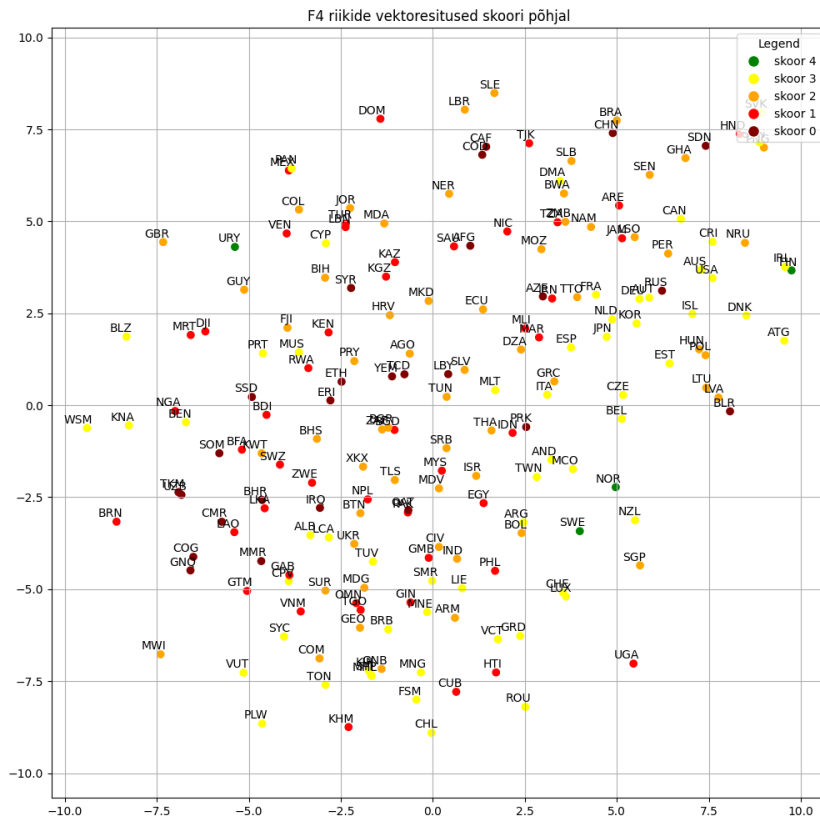
Uurides kõiki loodud graafikuid (LISA 1), selgub, et kõigis graafikutes leidub koondumist skooride põhjal. Kõige rohkem leidub seda kõrgete skooridega. Kõige kõrgemad rohelised skoorid koonduvad tihti ühte klastrisse. Nende ümber on raskemini eristavamas klastris enamasti ka kollased skoorid. Peale seda ei kandu enam skoorid tihti kontiinumina edasi, vaid hakkavad segamini minema. Oranžidest väärtusega „3“ olevates skooridest alates on and kõik veits segased

Vaatamegi D3 lähemalt. On näha (vt joonis 6), et sarnase skooriga tekstid on tõesti väga tugevalt koondunud ülesse keskele. Sealt alla liikudes kanduvad andmepunktide skoorid üle ühe võrra halvemateks kollasteks ning sealt edasi minnes oranžideks. Oranžidega lõpeb sujuva kontiinumi olukord ning eri skooridega andmepunktid hakkavad segamini olema. Enimsarnaste paaride puhul olid

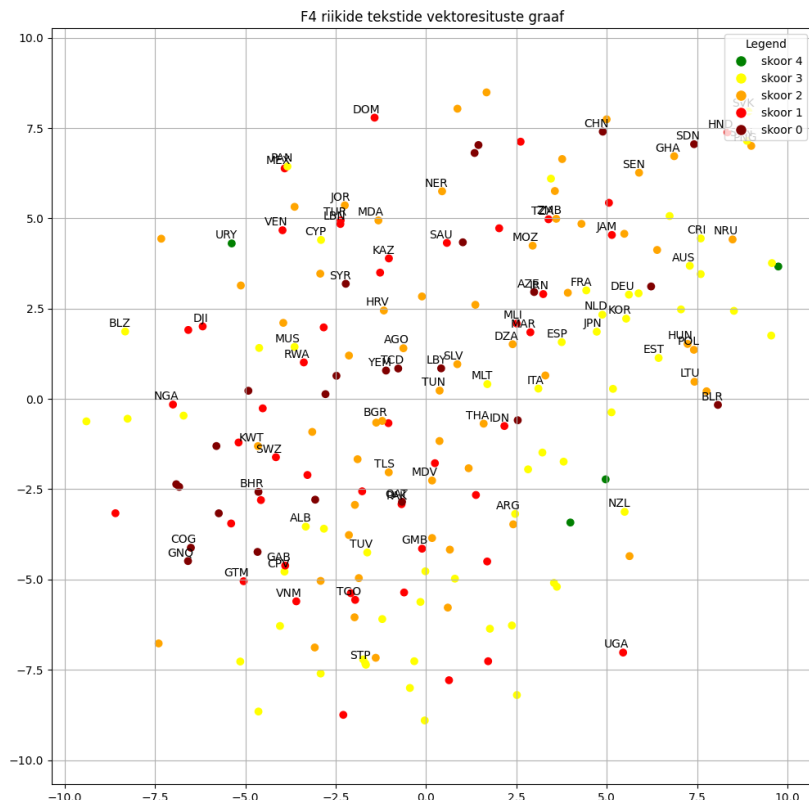
Kuigi enamus rohelisi andmepunkte on üleval kokku koondunud, on osad ka segamini teiste hulgas. Näiteks joonise keskel on perfektse skooriga Lõuna-Korea andmepunktid väga lähestikku null skooriga Põhja-Korea andmepunktiga. Nende kirjeldusi lähemalt uurides, selgub, et mõlemad kirjeldused mainivad Põhja-Koread ja sisaldavad sarnast lauseehitust. See võis viia Lõuna-Korea kirjelduse nii ebatavalisse kohta.



Joonis 6 D3 (Allikas: Autori andmed)



Joonis 7 F4 (Allikas: Autori andmed)



Joonis 8 F4 enimsarnased paarid (Allikas: Autori andmed)

Uurides lähemalt hoopis F4 küsimust (vt joonis 7). Siit selgub, miks F4 oli kõigest küsimustest nii teistsugune – rohelisi perfektse skooriga andmepunkte on väga vähe. See seletab ka selles küsimuses väga madalaid skooride erinevuse indeksi skooore. Vastavalt skooride „4 „ puudumisele on väga palju hoopis skooriga „3“ ehk kollaseid andmepunkte. Need andmepunktid on üldiselt koondunud all vasakult keskel paremale keskele ühise klastrina. Kõige halvemad skoorid on ka koondunud alla vasakule. Vahepealsed skoorid pole aga üldse koondunud.

Ainukesed rohelised andmepunktid on joonisel 8 Soome, Rootsi ja Norra. Tekib küsimus, et on Soome nii eemal nendest ja nii lähedal Iirimaale. Neid riike detailsemalt uurides selgub, et Norral ja Rootsil on võrdlemisi lühikesed kirjeldused. Soomel ja Iirimaal on aga pikemad kirjeldused. Nende kirjelduste on ka palju sarnasemad. Mõlemad räägivad diskriminatsioonist rassi, soo ja rahvuse alusel. Mõlemad tekstis on mainitud ka arvamusi erinevatelt Euroopa organisatsioonidelt, mis on toonud välja vihakõne kasvu. On toimunud ka geograafia aluselist koondumist. Paremalt keskel on näha, et mitmed Ida-Euroopa riigid on koondunud kokku, kuigi skoorid on erinevad.

Uurides ainult enimsarnaseid riigipaare, selgub, et nad pole selles küsimuses üldse klasterdunud kokku. Selle poolest on ta väga ebatüüpiline enamus muudest küsimustest. On alusta arvata, et põhjuseks on skooride erinevuse indeksis väga madal seis selles küsimuses.

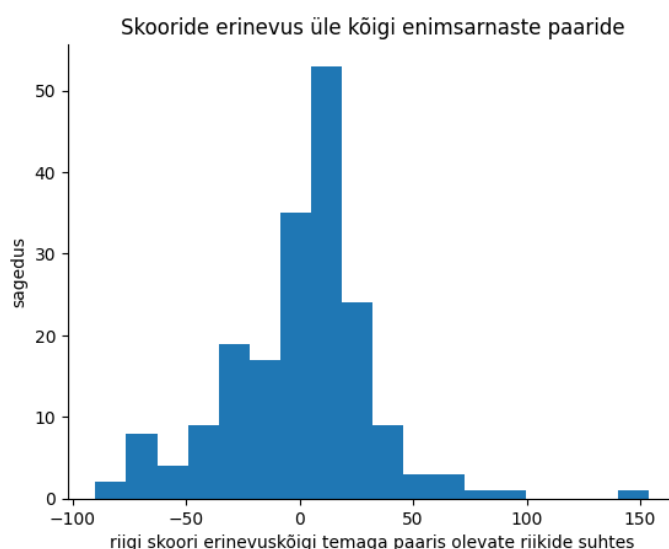
Kokkuvõttes saabki öelda, et kõrgete skooridega riigid koonduvad ehk klasterduvad tihti kokku. Mida madalamaks lähe skoor, seda väiksemaks läheb ka koondumine. Seega on näidatud, et kõrgete skooridega küsimustel on kirjeldustel tugev seos skooriga, kuid negatiivsetel vastupidiselt.

3.2.4 Skooride erinevuse tähelepanuväärsed riigid

Käesolevas alapeatükis uuritakse, kuidas on kõigi enimsarnaste riigipaaride vältel skooride erinevuse mustrid. Enimsarnaste riigipaaride puhul saab iga riigi skoori võrrelda kõigi tema paaris olevate riikidega. Selle põhjal arvutati töös varasemalt skooride erinevuse indeksit. Siin aga leitakse skoorid erinevuse indeks ainult riigipõhiselt. Näiteks oletama, et A1 enimsarnastest paarides on Jamaica kahes riigipaaris. Ühel juhul on ta skoor paarilisest ühe võrra kõrgem, teisel juhul on ta paarilisest kahe võrra madalam. Kokku tuleks siis A1 enimsarnastes riigipaarides Jamaica skooride erinevuse summaks -1. Seda protsessi tehti kõigi kogutud enimsarnaste riigipaaride vältel ja liideti kokku.

Saadud tulemused väljendavad seda, kui väga on mingisugust riiki enimsarnastes paarides üle- või alahinnatud tema paariliste suhtes. Tuletades meelde, siis enimsarnased riigipaarid on defineeritud selle poolest, et nad on 1% kõige suurema koosinuse sarnasusega riigipaarid. Kui ühel riigil on kõigi küsimuste peale tema paarilistega väga sarnased kirjeldused, kuid erinevad skoorid, siis see võib olla märk ebahütlusest kirjelduste ja skooride vahel. See omakorda võib olla märk kallutatusest ehk sellest, et pidevalt alahinnatakse vaadatava riigi skoori, kuigi kirjelduste poolest sarnaneb kõrgema skooriga riikidele.

Joonisel 10 on kujutatud, kuidas jaotub riikide skooride erinevuse indeks kõigi enimsarnaste riigipaaride peale kokku. Nagu näha, siis enamus riikide puhul on väga lähedal 0 skoorile, ehk nende skoorid on enamasti võrdselt nende paarilistega. Huvitavalt kanduvad skoorid erinevused palju sujuvamalt negatiivsemate skooride poole, kui positiivsete. Eksisteerib üks riik, mis on 150 juures väga tugevalt kõrgemini hinnatud oma paarilistest.



Joonis 10 (Allikas: Autori andmed)

Riik	Skooride erinevus	Riik	Skooride erinevus
senegal	154	gabon	-90
ghana	96	saudi-arabia	-77
botswana	80	nicaragua	-72
tunisia	67	rwanda	-71
colombia	67	iran	-71
niger	62	eswatini	-70
ecuador	53	venezuela	-69
poland	51	belarus	-68
south-africa	46	libya	-67
panama	42	egypt	-65

Joonis 11 kõige madalama ja kõrgema skooride erinevusega riiki (Allikas: Autori andmed)

Edasi saab uurida ka lähemalt, millised riigid olid kokku kõige kõrgemalt ülehinnatud ja millised kõige alahinnatud (vt joonis 11). Selgub, et enamast on ülehinnatud mitmed Aafrika ja Ladina-Ameerika riigid. Ainukeseks erandiks oleks Poola. Nende riikide riigiprofiile lähemalt uurimise põhjal oletab autor, et tegu on riikidega, kellel on enda piirkonna jaoks keskmisest oluliselt kõrgem skoor erinevates indeksites. See tähendab, et riikide kirjeldused peavad olema sarnased nende geograafilistele naabritele, kuid nende naabritel on oluliselt madalamad skoorid enamasti.

Kõige madalamate skoorides on ka mitu Aafrika ja Ladina-Ameerika riiki. Lisanud on ka Lähis-Idast mitu riiki. Ainukesena on Euroopast Valgevene. Siin võib eksisteerida seda, et vaadeldava riigi lähedal olevad riigid on konstantselt kõrgemate skooridega. Näiteks Venezuela ja Valgevene puhul on pea kõik nende naabrid oluliselt kõrgemate skooridega kõigis kategooriates. Uurides lähemalt ka Lähis-Ida riike, selgub, et kuigi nende ümber olevatel riikidel on globaalselt madalad skoorid terve FITW indeksi peale, siis on joonisel 11 kujutatud riikidel veel madalamad need

3.3 Geograafiline analüüs

3.3.1 Geograafiline analüüsi metodoloogia

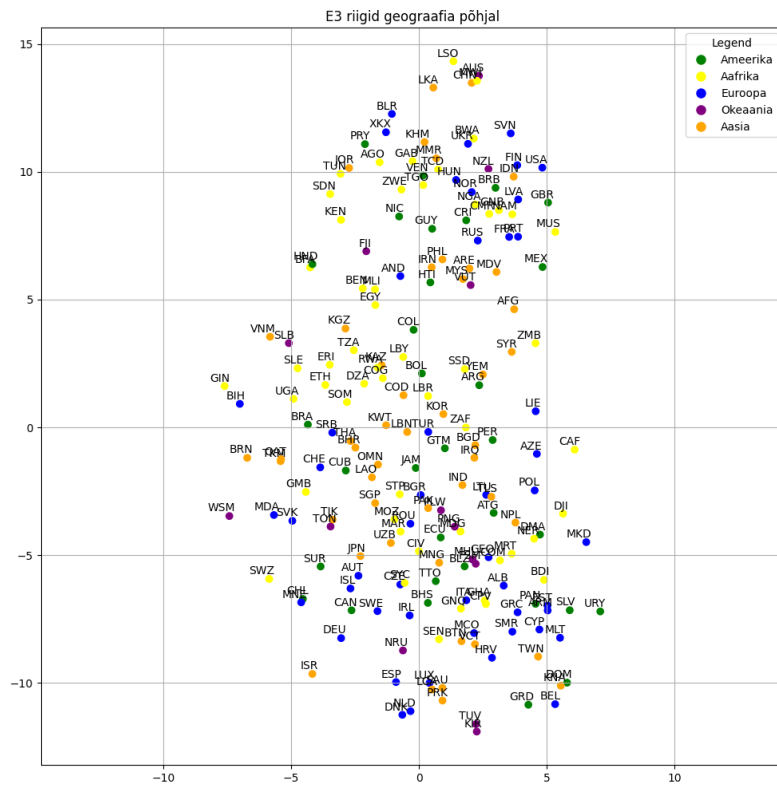
Antud peatükis uuritakse, kuidas on igas küsimuses jaotunud tekstide 2-d vektorestitused geograafiliselt. Kui vektorestitustes on samas piirkonnas enamasti sarnased tekstid, siis peaksid need riigid klasterduma joonistel kokku. Kuna vaadeldavaid riike on väga palju, siis vaadatakse antud peatükis lähemalt vaid üht küsimust igast seitsmest küsimuse kategooriast. Kõik tehtud joonised on saadaval LISA 1.

Riikide geograafilist asukohta uuritakse läbi selle, kus maailmajaos ta asub. Graafikutel on maailmajagu iga andmepunkti ehk riigi puhul visuaalselt eristatav läbi värvikoodi. Värvikood on järgnev: roheline on Ameerika, sinine on Euroopa, kollane on Aafrika, oranž on Aasia ning lilla on Okeania. Eraldi võetakse järgmises peatükis uurimise alla E3

3.3.2 Geograafiline analüüs

Siin peatükis analüüsitakse riikide kirjelduste vektorestituste visualisatsioonide geograafia alusel. Koostatud joonistelt selgus, et riigid tõeseti üldiselt klasterduvad maailmajao järgi. Veelgi enam on nende lähedal tihti ka naaberriikide kirjeldused. Kõige paremini klasterdusid Euroopa ja Ameerika riigid. Nende klastrid olid tihti kattuvad või üksteise lähedal olevad. Euroopa ja Ameerika riigid jaotusid tihti ka kaheks alam klastriks. Euroopa alaklastrid olid tavaliselt lääne-ida eraldatusega ning Ameerika omad põhja-lõuna kallutatusega. Aasia riigid klasterdusid enamasti, kuid vahel olid nad suvaliselt segamini Aafrika riikidega. Aafrika riigid klasterdusid kõigist kategooriatest kõige halvemini ning märgatavalt ei alaklasterdunud. Okeania riigid olid olenevalt küsimusest kas kõik tihedalt koos või väga laiali.

Märkimisväärselt oli Israel tihti Euroopa-Ameerika klastrite juures, mitte Aasia riikide juures. Okeaniast olid Austraalia ja Uus-Meremaa ka tihti Euroopa-Ameerika klastri sees või lähestikku. See on ka loogiline, sest neid riike tihti peetakse Lääne riikideks. Natuke oli osadel geograafiliste joonsite põhjal võimalik öelda, et seda esines ka Jaapaniga mitmetes kategooriates. Seega võib öelda, et osade riikide sarnasust saab seletada suuresti nende geograafia järgi.



Joonis 12 (Allikas: Autori andmed)

Üldiseid joonistel olevaid trende iseloomustab hästi joonis E3. Seda vaadates (vt joonis 12) on silmnähtavalt koondunud Euroopa riigid kaheks. Üks klaster on üleval paremal ning teine all vasakul. Mõlema klasteri lähedal, ja nendega osati kattuvalt, on ka Ameerika riigid jaotunud kaheks osaks. Aafrika riigid on koondunud kehvalt keskele, kuid neid ka palju segamini Aasia riikidega. Aasia riigid on jaotunud enamasti alla keskele, kuid on segamini Euroopa ja ja Aafrika riikidega. Okeania riigid on enamasti Euro-ameerika klastrite läheduses. Märkimisväärselt on Israel väga eemal muudest Aasia riikidest (all keskel) ning lähedal pigem Euroopa ja Okeania riikidele.

Geograafilise analüüsi puhul tasub mainida, et sarnasust sai põhjustada ka see, et kirjeldustes on tihti vastava riigi riiginimi sees. Kuna keelemudel võib riiginimesid pidada omavahel sarnaseks, siis võis see valimit mõjutada. Näiteks Lätit ja Eestit on äkki BERT näinud treenitavas infos enamasti sarnastes kontekstides ja sellepärast teeb see Eesti ja Läti kirjeldused sarnasemaks, kui nad riike mainimata oleksid. Autor soovib seega tulevikus sarnasel viisil kirjeldusi uurides teha alavalim, kus kõigis kirjeldustes on eemaldatud, või asendatud, viited riiginimedele ja rahvustele ning siis sellest teha vektorsitused.

3.4 Analüüsi kokkuvõte

Analüüsi käigus näidati, et enimsarnased riigipaaride vahel on palju madalam skooride erinevus, kui vähimsarnaste riigipaaride vahel. See kehtis kõigi Tabelites 1.1 ja 1.2 välja toodud küsimuste puhul. Seega saa järeldada, et sarnaste tekstide vahel on keskmiselt sarnased skoorid. Analoogselt on ka kõige vähim sarnaste tekstide vahel suur skooride erinevus. Seega sai töös püstitatud hüpotees kinnitust.

Riikide kirjelduste vektorestituste uurimine skooride tõi välja, et tegu sarnaste skooridega riigid koondusid ühtsetesse klastritesse teatud tingimustel. Nimelt koondusid enamasti kõige kõrgema skooriga riigid ühtsesse klastrisse. Sealt kandus sujuvalt andmepunktid üle tihti ka ühe võrra halvamteks skoorideks. Edasi koondusid aga skoorid „1“ ja „2“ väga kehvalt. Paremini, kuid ikkagi kasinalt koondusid ka skooriga „0“ olevad riigid.

Konkreetselt uuriti, milliste riike ülehinnatakse ja alahinnatakse enimsarnaste riigipaaride puhul. Selgus, et ülehinnatakse riike nagu Senegal, Ghana ja Botswana (vt joonsi 11). Alahinnatakse riike nagu Gabon ja Saudi Araabia. Kõige enam oligi ülehinnatud ja alahinnatud riike just Aafrikas. Analüüsi käigus tundus, et sellised kõrge skooride erinevuse tulemuse saanud riigid on selle tulemuse saanud eelkõige oma naabrite alusel. Nimelt, tuli kõrge (või madal) skoor sellest, et enams vaadeldava riigi naabruses olevatest riikidest on madalama (või kõrgema) skooriga. Seega ei tuvastatud töös, et kõrge skooride erinevus oleks riikide puhul märk sellest, et seda riiki süstemaatiliselt hinnatakse erinevalt, vaid tegu on tõenäoliselt geograafilise nähtusega.

Geograafiline analüüsis selgus, et riigid koondusid üldiselt tugevalt kokku oma geograafia alusel. Koondumine oli tugevaim Ameerika riikide puhul. Euroopa riigid olid enamasti ka klasterdunud, kuid tihti jaotusid ka Ida-Lääne alusel kaheks eraldi klasteriks. Aasia ja Okeaania riigid koondusid klastritesse kehvemalt kui Euroopa ja Ameerika riigid, kuid tegid seda ikkagi. Okeaania puhul oli väikese valimi tõttu ka raskem klasterdumist tuvastada. Kõige halvemini koondusid Aafrika riigid, olles tihti segamini Aasia riikidega.

Joonistel riikide geograafilise jaotuse järgi põhjal saab järeldada, et riikide kirjeldused tõesti korreleeruvad riikide geograafiliste asukohtadega. Sellest omakorda järeldub, et BERTI vektorestitused suudavad kanda endas kirjeldustes olevat infot ning kirjeldustes olev info on piirkondade siseselt sarnane. (ning mitmetel juhtudel ka piirkondade vahel sarnane).

Kokkuvõttes ongi FITW indeks seega üksühene ka geograafia alusel Ameerika ja Euroopa puhul. Kõige vähem üksühene on ta Aafrika puhul.

4. Kokkuvõte

Käesolevas uurimistöös uuriti *Freedom House*-i demokraatia indeksi *Freedom in the World* (FITW). Akadeemilise kirjanduse põhjal tutvuti selle teemaga lähemalt sellega, kuidas demokraatia indekseid tehakse ning milline on põhiline kriitika nende kohta. Konkreetselt toodi välja ka eksisteeriv kriitika FITW kohta ning seostati see uurimisküsimusega. Kuna töös uuriti FITW-i masinõppel treenitud BERT keelemudeliga, siis kirjeldati ka seda, kuidas masinõpet on kasutatud sotsiaalteadustes varasemalt. Lisaks toodi välja kuidas saaks seda rohkem kasutada sotsiaalteadustes. Sellele järgnevalt kirjeldati, kuidas töötab keelemudel BERT ning millised on töös kasutatavad vektorestitused.

Töö empiirilises osas tehti FITW kõigil küsimustel iga riigi kirjeldava teksti kohta BERT keelemudeli vektorestitus, mille läbi võrreldi kirjelduste sarnasust. Kirjelduste omavaheline sarnasus ja neile vastavate skooride erinevust võrreldes, sai kaudselt hinnata ka indeksi üksühesust. Töös püstitati hüpotees, et sarnastel tekstidel on sarnased skoorid. See hüpotees leidis kinnitust. Omavaheliselt kõige sarnasemad kirjeldused on tõesti sarnasemate skooridega. Analoogselt leiti, et kõige vähem omavahel sarnased kirjeldused on ka vähem sarnaste skooridega. Selle põhjal tundus ka indeks üksühene.

Empiirilise osas näidati ka, kuidas saab vektorestituste kaudu visualiseerida tekstiilis andmete vahelisi sarnasusi kvantitatiivselt. T-SNE meetodiga tehti igale küsimusele vektorestitusi sisaldav joonis. Nii skoori, kui ka geograafia alusel analüüsidis oli ilmne, et saadud joonistel oli toimunud samasuguste andmete (sama skoor või maailmajagu) põhjal koondumist. Seega BERT-i keelemudel tehtud kirjelduste vektorestitused suutsid tekste hästi esitada, sest antud info korreleerus skooride ja maailmajagudega.

Vektorestituste visualiseerimise kaudu leiti, et skooride põhjal koonduvad kõrged skoorid ühtselt. Kõrged skooridest liiguti vektorruumis ühtlasi sujuvalt üle ühe võrra madalamatele skooridele. Selline kontiinum aga ei jätkunud kõrgemate skooridega, mis puhul eri skooridega andmepunktid hakkasid omavahel segunema. Kõige madalamad skoorid koondusid harva, kuid ka siis kehvalt. Sellest järeldus, et kirjeldused ja skoorid on väga sarnased kõrgete skooride puhul, kuid mida madalamaks minnakse, seda kehvemalt hakkavad nad korreleeruma. Geograafiliselt neid

Geograafilise analüüsi tulemusel selgus, et riigid koondusid üldiselt tugevalt vastavalt oma geograafilisele asukohale, eriti Ameerika riigid. Euroopa riigid moodustasid samuti klastreid, kuid tihti jagunesid ida-lääne suunas kaheks eraldi klatriks. Aasia ja Okeania riigid

koondusid klastritesse kehvemini, kuid siiski mingil määral, kusjuures Okeania väikese valimi tõttu oli klasterdumist raskem tuvastada. Aafrika riigid klasterdusid kõige halvemini ja olid sageli segamini Aasia riikidega.. FITW indeksi kirjeldused on seega Ameerika ja Euroopa puhul geograafia alusel üksühene, teistes piirkondades mitte.

Kokkuvõttes leiti antud töös, et Freedom House-i indeksi kirjeldused korreleeruvad kõrgete skooride puhul tugevalt oma kirjeldustega. See korrelatsioon läheb järjest kehvemaks, mida madalamaks lähevad ka skoorid. Geograafiliselt on FITW indeksis leiduvad kirjeldused on omavahel enamasti koreleeritud ehk üksühesed Euroopa ja Ameerika maailmajagudes. Töös leidis ka kinnitust hüpotees, et enimsarnaste riikide puhul on skoorid sarnasemad, kui vähemsarnaste tekstide vahel. Seega on FITW indeks üksühene eelkõige kõrgete skooride puhul ning juhtudel, kui tegemist on Ameerika või Euroopa riikidega. Kõige vähem on indeks üksühene Aafrikas.

Kasutatud allikate loetelu

Boese, Vanessa A. 2019. "How (not) to measure democracy" *International Area Studies Review* 22(2) 95-127

Chen, H. Yang, C. , Zhang, X. Liu, Z. Sun, M. and J. Jin, 2021 "From Symbols to Embeddings: A Tale of Two Representations in Computational Social Science," in *Journal of Social Computing*, vol. 2, no. 2, 103-156,

Dahl, Robert A. 1971 „Polyarchy: Participation and Opposition“ New Haven and London: Yale University Press.

Deepai „What is a Transformer Neural Network?“ <https://deepai.org/machine-learning-glossary-and-terms/transformer-neural-network> (külastatud 27.04.2024)

Emmery, Chris 2017 „Euclidean vs. Cosine Distance“ <https://cmry.github.io/notes/euclidean-v-cosine> (külastatud 27.04.2024)

Featureform 2021 „Embeddings in Machine Learning: Everything You Need to Know“ <https://www.featureform.com/post/the-definitive-guide-to-embeddings> (külastatud 19.05.2024)

Freedom House 2023. https://freedomhouse.org/sites/default/files/2023-03/FIW_World_2023_DigitalPDF.pdf (külastatud 12.05.2024)

Freedom House „Our History, <https://freedomhouse.org/about-us/our-history> (külastatud 13.05.2024)

Freedom House „Freedom in the World Research Methodology“ <https://freedomhouse.org/reports/freedom-world/freedom-world-research-methodology> (külastatud 14.05.2024)

Freedom House „Freedom in the World 2023 Methodology Questions“ https://freedomhouse.org/sites/default/files/2023-03/FITW_2023%20MethodologyPDF.pdf (külastatud 13.05.2024)

Fox, Jonathan. "Do Democracies Have Separation of Religion and State?" *Canadian Journal of Political Science / Revue Canadienne de Science Politique* 40, no. 1 (2007): 1–25.

Geeksforgeeks 2023 „Difference between PCA VS t-SNE“ <https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/> (külastatud 14.05.2024)

Github 2021 „Interpretation of negative semantic similarity“
<https://github.com/UKPLab/sentence-transformers/issues/760> (Külastatud 22.04.2024)

Graziano, P. and Quaranta, M. (2024) ‘Studying Democracy in Europe: Conceptualization, Measurement and Indices’, *Government and Opposition*, 59(2),

Grimmer et al (2021) *Machine Learning for Social Science: An Agnostic Approach*,

Gründler, Klaus, Krieger, Tommy, 2015. "Democracy and Growth: Evidence of a New Measurement," *CESifo Working Paper Series 5647*, CESifo.

Gründler, Klaus, Krieger, Tommy, 2021. "Using Machine Learning for measuring democracy: A practitioners guide and a new updated dataset for 186 countries from 1919 to 2019," *European Journal of Political Economy*, Elsevier, vol. 70(C).

Gründler, Klaus, Krieger, Tommy, 2018 „Machine Learning Indices, Political Institutions, and Economic Development“ (). *CESifo Working Paper Series No. 6930*

Huggingface „BERT“ https://huggingface.co/docs/transformers/model_doc/bert (külastatud 14.05.2024)

Iban, „COUNTRY CODES ALPHA-2 & ALPHA-3“ <https://www.iban.com/country-codes> (külastatud 19.05.2024)

Medium 2022 „Word2vec vs BERT „ <https://medium.com/@ankiit/word2vec-vs-bert-d04ab3ade4c9> (külastatud 18.04.2024)

Munck, G. L., & Verkuilen, J. 2002. Conceptualizing and Measuring Democracy: Evaluating Alternative Indices. *Comparative Political Studies*, 35(1), 5-34.

Perone, C. 2013 „Machine Learning Cosine Similarity for Vector Space Models (Part III)“ <https://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/> (külastatud 22.04.2024)

Stackoverflow 2023 <https://stackoverflow.blog/2023/11/09/an-intuitive-introduction-to-text-embeddings/> (külastatud 19.05.2023)

Steiner, N. D. (2016) ‘Comparing Freedom House Democracy Scores to Alternative Indices and Testing for Political Bias: Are US Allies Rated as More Democratic by Freedom House?’, *Journal of Comparative Policy Analysis: Research and Practice*, 18(4), 329–349.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). „Attention Is All You Need“. *arXiv preprint arXiv:1706.03762*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. and Yang, D., 2023. Can Large Language Models Transform Computational Social Science?. *ArXiv abs/2305.03514 (2023)*: n. pag.

LISA 1

Käesolevas töö raames loodud web-scrapet, saadud andmestik, tekstianalüüsi kood ja andmete visualiseerimised on kõik saadaval internetis järgnevas allikas:

<https://github.com/Kostabi/kostabi-bakalaurus>

LISA 2 Lihtlitsents

Mina, Karl Hans kostabi, (isikukood: 50002210241) annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Freedom House'i demokraatia indeksi ühtsuse analüüs BERT keelemudeliga“, mille juhendajad on Kristel Vits ning Uku Kangur.

- reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace'is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
- üldsusele kättesaadavaks tegemiseks ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
- olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile;
- kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.