TARTU ÜLIKOOL

Sotsiaalteaduste valdkond

Ühiskonnateaduste instituut

Ühiskonna ja infoprotsesside analüüsi õppekava

Helina Toompark

Mõjutustegevuse tahtlikkuse kvantitatiivne mõõtmine tekstis

Magistritöö

Juhendaja: Sten Torpan, PhD

Kaasjuhendaja: Uku Kangur, MA

Tartu 2026

AUTORIDEKLARATSIOON

Olen koostanud magistritöö iseseisvalt. Kõik töös kasutatud teiste autorite tööd, põhimõttelised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

Helina Toompark

27.05.2026

SISUKORD

SISSEJUHATUS

Tänapäevases digitaalses infosfääris liigub infot rohkem ja kiiremini kui kunagi varem (Supriyono jt, 2024; Weitkamp jt, 2021; Ecker jt, 2022). Peamine probleem ei ole enam vale või eksitava sisu olemasolu, vaid ka suur infomaht ja auditooriumi tähelepanu nappus, mis muudavad strateegilise mõjutustegevuse eristamise tavapärasest kommunikatsioonist üha keerulisemaks (Christiano, 2022). Töös defineerin mõjutustegevust sihipärase ja strateegilise kommunikatsioonina, mille eesmärk on manipuleerida auditooriumi hoiakuid või käitumist (Starbird jt, 2019).

Mõjutustegevuse tuvastamisel on keskne roll tahtlikkusel, mis võimaldab eristada strateegilist manipulatsiooni juhuslikust eksimusest või tavapärasest kommunikatsioonist (Hyzen, 2021; Starbird jt, 2019). Tahtlikkus on aga metodoloogiliselt keeruline tunnus. Tekstis avalduv mõju ei pruugi olla otseselt seotud kontrollitava valeväitega ja autori tegelik kavatsus, ehk tahtlikkus, ei ole empiiriliselt vahetult jälgitav (Ottaviani jt, 2021), vaid kontekstuaalsete tunnuste põhjal tehtav järeldus (Kärki, 2023; Klenk, 2022). Tavapärase kommunikatsiooni ja mõjutustegevuse piiri hägustavad ka retoorikavõtted, mida mõlemas olukorras sageli kasutatakse. Sõnumite levikut inforuumis kujundavad nii teadlikud mõjutamisvõtted kui ka kasutajate juhuslik edasijagamine, mistõttu tekib keskkonda palju müra, raskendades sekkumist vajavate juhtumite märkamist (Starbird jt, 2019; Christiano, 2022).

Kui inimestel on keeruline hinnata info päritolu ja usaldusväärsust, võivad ühiskonnas süveneda usalduskriis, polariseerumine ja vastuvõtlikkus manipuleerivale kommunikatsioonile (Caled ja Silva, 2022; Starbird jt, 2019; Boulianne jt, 2022). Nende tegurite koosmõju suurendab vajadust automaatsete otsustustoe süsteemide järele, mis aitaksid infomüras märgata mõjutustegevusele viitavaid mustreid (Abro jt, 2023). Eesti kontekstis on see eriti oluline ka seetõttu, et inforuumis võivad kohalike ühiskondlike pingete ja aruteludega põimuda ka vaenulikud välismõjud, sealhulgas Vene narratiividega seotud mõjutustegevus (Välisluureamet, 2025). Eestis puuduvad seni empiirilised katsed hinnata, kas ja kuidas on võimalik mõõta tahtlikkuse avaldumist tekstides kvantitatiivsete meetoditega, mistõttu ei ole teada, millisel määral saaks masinõppel põhinev tahtlikkuse indikaator toimida praktilise otsustustoe elemendina mõjutustegevuse tuvastamisel.

Magistritöö eesmärk on välja selgitada, kas masinõppepõhist tahtlikkuse indikaatorit saab kasutada eestikeelsetes tekstides otsustustoe elemendina tahtliku mõjutustegevuse eristamisel tavapärasest kommunikatsioonist. Töös lähtun eeldusest, et tahtlikkuse täielik

automaattuvastus ei ole realistlik eesmärk. Küll aga on võimalik hinnata, kas algoritmid suudavad ekspertide antud tahtlikkuse hinnanguga seotud tekstilisi tunnuseid piisavalt järjekindlalt eristada, et täita oma kohta osana laiemast otsustustoe süsteemist.

Selleks seadsin kaks uurimisküsimust:

1. Kuivõrd täpselt võimaldab masinõpe eristada tekstides tahtlikku mõjutustegevust?
2. Kuidas mõjutab treeningandmestiku mahu suurendamine mudeli tulemuslikkust?

Uurimisküsimustele vastuste leidmiseks võrdlen omavahel juhendatud masinõppemudelite, generatiivse mudeli ja ekspert- ning kontrollgrupi antud hinnanguid küsimusele, kas tekstis esineb tahtlikku mõjutustegevust. Ekspertide häälteenamushinnang moodustab siin töös võrdlusaluse.

Magistritöö on üks osa mõjutustegevuse tuvastamise tööriista arendamisest. Eesmärk ei ole luua iseseisvat ja lõplikku automaattuvastuse lahendust, vaid hinnata ühe konkreetse tunnuse ehk tahtlikkuse praktilist kasutatavust ja võimalikku kohta laiemas otsustustoe süsteemis. Sellest vaatenurgast võib töö olla praktiliselt kasulik meedia- ja kommunikatsioonispetsialistidele, poliitikakujundajatele ja ka julgeolekuasutustele, kelle jaoks on oluline vähendada juhuslikku müra digiplatvormidel ning märgata võimalikke infosfääri haavatavusi enne, kui need süvenevad.

Töö koosneb neljast suuremast peatükist. Esimeses peatükis annan ülevaate tänapäevasest infosfäärist, mõjutustegevuse olemusest ja tehnikatest ning tahtlikkuse rollist mõjutustegevuses. Samuti tutvustan sotsiaalteadustes kasutatavaid arvutuslikke meetodeid, mis on aluseks töö metodoloogiale. Teises peatükis kirjeldan uuringu andmekogumismeetodit, ekspert- ja kontrollgrupi valimit ning masinõppemudelite treenimise ja analüüsi metoodikat. Kolmandas peatükis esitan analüüsi tulemused ja neljandas peatükis arutlen saadud tulemuste tähenduse, piirangute ja edasiste uurimisvõimaluste üle.

1. TEOREETILISED LÄHTEKOHAD

Selles peatükis annan ülevaate töö teoreetilistest lähtekohtadest. Esmalt käsitlen tänapäevast infosfääri ja mõjutustegevust, seejärel keskendun tahtlikkusele kui töö kesksele mõistele ja lõpuks kirjeldan arvutuslike lähenemiste rolli selle nähtuse uurimisel. Peatüki lõpus sõnastan probleemiseade, millel põhineb töö empiiriline osa.

1.1. Tänapäevane infosfäär ja ühiskondlik haavatavus

1.1.1. Infoüleküllus ja kognitiivsed piirangud

Digitaalsete tehnoloogiate levik on muutnud info tootmise ja levitamise tingimusi, suurendades nii informatsiooni hulka, levitajate arvu kui ka leviku kiirust (Starbird jt, 2019; Weitkamp jt, 2021; Ecker jt, 2022). Avalikus ruumis ringleb üha rohkem sisu, mida luuakse ja jagatakse paljudes kanalites, muutes asjakohase info leidmise järjest keerulisemaks (Supriyono jt, 2024; Persson, 2018).

Võrreldes varasema massimeediaga on tänapäevane kommunikatsioon muutunud üha enam kahesuunaliseks (Liagusha ja Iarovyi, 2025). Auditoorium võtab infot vastu ja osaleb aktiivselt selle loomises, tõlgendamises ja edasilevitamises (Caled ja Silva, 2022; Liagusha ja Iarovyi, 2025; Murdock, 2016). Lisaks on sotsiaalvõrgustike areng vähendanud füüsilisi ja geograafilisi barjääre, võimaldades vahetut mõtete vahetust ja info levikut üle riigipiiride (Caled ja Silva, 2022; Abro jt, 2023; Burton, 2023). Sellises keskkonnas ei pruugi loodav ja levitatav sisu läbida eelkontrolli, kuid võib sellest hoolimata jõuda väga laia auditooriumini (Caled ja Silva, 2022; Liagusha ja Iarovyi, 2025).

Perssoni (2018) järgi sõltuvad indiviidide otsused suurel määral sellest, milline info nende tähelepanuvälja jõuab. See loob info pakkujatele omakorda strateegilise huvi suunata tähelepanu ja võimendada neile kasulikke sõnumeid (Persson, 2018). Tähelepanu nimel toimuv konkurents avaldab survet ka peavoolumeediale, mis võib auditooriumi hoidmise eesmärgil kalduda kiirustatud või ebapiisavalt kontrollitud sisu avaldamise poole (Caled ja Silva, 2022). Sellises kontekstis ei ole nähtavus digitaalses meediakeskkonnas kvaliteedi näitaja, vaid pigem edukas tulemus konkurentsis piiratud tähelepanu pärast (Weitkamp jt, 2021). Mõjutustegevuse eristamine ei sõltu enam üksnes üksikute väidete tõesuse kontrollimisest, vaid ka võimest märgata suuremas infomahus korduvaid strateegilisi mustreid.

Kasvav andmemaht ja info pidev kättesaadavus ületavad inimeste võimekuse infot süvenenult analüüsida ja sünteesida (Arnold jt, 2023; Supriyono jt, 2024; Persson, 2018). Nii kujuneb välja infoülekülluse probleem, milles kohtuvad tähelepanu nappus ja indiviidi kognitiivsed piirangud (Starbird jt, 2019; Arnold jt, 2023; Caled ja Silva, 2022). Piiratud töömälu ja infotöötlusvõime ei võimalda sissetulevale informatsioonile süstemaatiliselt süveneda, samal ajal kui erinevad sõnumid konkureerivad lugeja tähelepanu pärast (Arnold jt, 2023; Caled ja Silva, 2022; Lund jt, 2020).

Kui inimesed on sunnitud oma fookust kitsendama ja sõnumeid valikuliselt filtreerima, suureneb nende sõltuvus lihtsustatud otsustusstrateegiatest ja emotsionaalsest sisust (Caled ja Silva, 2022; Liagusha ja Iarovyi, 2025). Rõhk ei ole enam parimate otsuste tegemisel, vaid kognitiivse koormuse vähendamisel (Persson, 2018; Caled ja Silva, 2022; Arnold jt, 2023). Indiviidid kalduvad vaikimisi käsitlema neile esitatud informatsiooni usaldusväärsena, samas kui selle kahtluse alla seadmine eeldab täiendavat kognitiivset pingutust (Brashier ja Marsh, 2020; Pantazi jt, 2021).

Kognitiivseid piiranguid saab strateegiliselt ära kasutada mõjutustegevuses ja infosõjas, kus auditooriumi ujutatakse teadlikult üle sisuga, mis ületab nende töötlusvõime (Starbird jt, 2019; Supriyono jt, 2024). Infosõda tähistab strateegiliste infooperatsioonide kogumit, mille kaudu riiklikud ja mitteriiklikud osalejad püüavad sihipäraselt kujundada avalikku arvamust ja inimeste maailmataju infokeskkonna mõjutamise kaudu (Starbird jt, 2019). Peamine sihtmärk on inimese taju ja otsustusvõime mõjutamine ehk usalduse murendamine institutsioonide vastu, ühiskondlike lõhede süvendamine ning tegutsemisvõime nõrgestamine (Starbird jt, 2019; Caled ja Silva, 2022).

Tänapäevane infosfäär ei suurenda üksnes info hulka (Supriyono jt, 2024; Persson, 2018), vaid muudab ka selle leviku ja nähtavuse tingimusi (Caled ja Silva, 2022; Liagusha ja Iarovyi, 2025). Kui sissetuleva info maht ületab inimeste kognitiivse töötlusvõime, muutuvad vajalikuks toetavad lahendused, mis aitavad suuri andmemahtusid süstemaatiliselt analüüsida ja märgata mustreid, mis võivad viidata strateegilisele mõjutustegevusele (Abro jt, 2023).

1.1.2. Ühiskondlik haavatavus mõjutustegevusele

Mõjutustegevuse toime sõltub lisaks muudele teguritele ka sellest, millistes ühiskondlikes tingimustes sõnumid vastu võetakse. Ühiskondlik haavatavus on seisund, milles mõjutustegevus saab tõhusamalt toimida, kuna inimeste ja rühmade võime infot hinnata ning

eristada usaldusväärset sisu strateegiliselt kujundatud kommunikatsioonist, on nõrgenenud (Caled ja Silva, 2022; Starbird jt, 2019; Boulianne jt, 2022). Haavatavus ei tähenda teadmiste puudumist, vaid see kujuneb usalduskriisi, sotsiaalsete ebavõrdsuste ning kognitiivsete ja tehnoloogiliste tingimuste koosmõjus (Burton, 2023; Christiano, 2022; Hobbs, 2020; Kertysova, 2018).

Ühelt poolt väljendub see usalduse vähenemises institutsioonide, meedia ja ekspertide vastu, mistõttu võivad alternatiivsed ja vastandavad narratiivid kergemini levida (Caled ja Silva, 2022; Weitkamp jt, 2021; Liagusha ja Iarovyi, 2025). See võimaldab strateegilistel infooperatsioonidel võimendada segadust ja polariseerumist (Starbird jt, 2019; Bakirov ja Suleimenov, 2025; Kertysova, 2018; Välisluureamet, 2025). Infosõja kontekstis tähendab see, et olemasolevaid lõhesid saab teadlikult ära kasutada vaenulike huvide teenimiseks (Välisluureamet, 2025). Haavatavus väljendub ka selles, et vastandavad narratiivid võivad leida vastuvõtlikke auditooriume rühmades, kelle identiteet või infokeskkond muudavad nad sellistele sõnumitele vastuvõtlikumaks (Samas).

Teiselt poolt on haavatavus seotud ressursside ebaühtlase jaotusega, sest inimeste suutlikkus mõjutustegevusele vastu seista sõltub hariduslikest, sotsiaalsetest ja tehnoloogilistest võimalustest (Murdock, 2016; Persson, 2018; Hobbs, 2020). Haavatavus ei ole seega ainult individuaalne nõrkus, vaid seotud ka habituse ja institutsionaalse kaasatusega (Gebauer ja William, 2000; Ottaviani jt, 2021; Kertysova, 2018). Struktuurne haavatavus ilmneb ka selles, et vaenulikud osalejad ei püüa mõjutada ainult avalikku arvamust, vaid ka infosüsteeme ja institutsioone, mille kaudu ühiskond toimib (Välisluureamet, 2025).

Haavatavust suurendavad ka kriisiolukorrad ja ebakindlus, kus ametliku info puudulikkus loob infolünki, mida püütakse täita kiirete ja näiliselt sidusate seletustega (Caled ja Silva, 2022; Madisson ja Ventsel, 2021; Liagusha ja Iarovyi, 2025). Nii kujuneb mõjutustegevuse vastuvõtlikkus probleemiks, milles platvormide toimeloogika ja inimeste piiratud kognitiivsed ressursid üksteist vastastikku võimendavad (Starbird jt, 2019; Arnold jt, 2023). WHO (2017) käsitluse järgi aitab haavatavust vähendada kommunikatsioon, mis on sihtrühmale kättesaadav, asjakohane, mõistetav ja õigeaegne ning toetab inimeste suutlikkust teha teadlikke otsuseid. Mõjutustegevuse vastupanuvõime tugevdamine ei seisne üksnes valeinfo ümberlükkamises, vaid ka sellise infokeskkonna kujundamises, milles usaldusväärne informatsioon on inimestele reaalselt kasutatav ja märgatav (World Health Organization, 2017).

Mõjutustegevust ei saa seega määratleda ainult sõnumi sisu või üksikute väidete tõesuse alusel, vaid oluline on küsimus, kas ja mil määral on kommunikatsioon strateegiliselt kujundatud. Seetõttu on mõjutustegevuse analüüsimisel oluline tahtlikkuse tunnus.

1.2. Mõjutustegevus digitaalses keskkonnas

1.2.1. Mõjutustegevus

Mõjutustegevuse mõistmine eeldab selle eristamist väär- ja desinformatsioonist, kuna info tõesusele keskendumine ei võimalda haarata mõjutustegevuse sihipärast ja strateegilist iseloomu (Boulianne jt, 2022; Starbird jt, 2019). Kui väärinformatsioon võib olla tahtmatu eksitus ja desinformatsioon tähistab sihilikult loodud valeinfot (Wardle ja Derakhshan, 2017), siis mõjutustegevus on laiem nähtus. See hõlmab tegevusi, mille eesmärk on kujundada hoiakuid, tõlgendusraame ja avalikku arvamust, sõltumata sellest, kas kasutatav info on täielikult väär, osaliselt eksitav või formaalselt tõene (Christiano, 2022; Persson, 2018; Burton, 2023; Starbird jt, 2019; Caled ja Silva, 2022; Kertysova, 2018).

Mõjutustegevuse eristamine väärinfost ei põhine seega üksnes küsimusel, kas edastatav info on tõene või väär, vaid sellel, kas info levitamine ja kasutamine on tahtlik ning strateegiline (Christiano, 2022; Bakirov ja Suleimenov, 2025; Starbird jt, 2019; Persson, 2018; Burton, 2023). Tahtlikkus väljendub sageli infokeskkonna teadlikus kujundamises, näiteks infoülekülluse esilekutsumises või valikuid mõjutavate tehnikate kasutamises (Persson, 2018; Christiano, 2022; Starbird jt, 2019). Mõjutustegevuse uurimine ei piirdu ainult sõnumi sisu hindamisega. Infouuringutes on kasutusel François (2019) loodud ABC-raamistik, mille põhjal saab mõjutustegevuse tuvastamise jagada kolmeks tasandiks: osapooled (*Actors*), käitumine ja levikumustrid (*Behaviour*) ning sisu (*Content*).

Mõjutustegevuse toime sõltub ka sellest, kuidas kujundatakse auditoorium, kellele konkreetne sõnum on suunatud (Madisson ja Ventsel, 2021; Turow, 2005). Auditooriumi kujundamine ei määra ainult seda, kellele sõnum suunatakse, vaid ka seda, kuidas seda vastu võetakse, edasi jagatakse ja tõlgendatakse (Madisson ja Ventsel, 2021; Turow, 2005; Ventsel jt, 2021). Digitaalses keskkonnas ei tähenda see üksnes sihtrühmade kõnetamist nende identiteedi või tõlgendusraamide alusel, vaid üha enam ka andmepõhist sihtimist, mille abil kohandatakse sõnumi vormi ja ajastust viisil, mis suurendab selle näilist asjakohasust ja veenvust (Madisson ja Ventsel, 2021; Kaptein jt, 2015; Hobbs, 2020; Dehnert ja Mongeau, 2022).

Auditooriumi kujundamine tähendab ka adressaadi kutsumist kindlasse subjektirolli, kus teatud tõlgendused ja tegevusvalikud muutuvad tõenäolisemaks või loomulikumaks (Hyzen, 2021; Humă, 2023; Bakir jt, 2019). Nii määrab auditooriumi kujundamine selle, kellele sõnum suunatakse, kuidas vastuvõtja end sõnumi suhtes positsioneerib, mida ta usutavaks peab ja millisel viisil ta sõnumit edasi kannab (Madisson ja Ventsel, 2021; Wanless ja Berk, 2020; Gibert, 2023). Auditoorium on mõjutustegevuses oluline nii adressaadina kui ka ressursina, mille omadusi ja harjumusi püütakse süstemaatiliselt ära kasutada (Turow, 2005).

Mõjutustegevuse eesmärk on kujundada poliitilist, sotsiaalset või kultuurilist reaalsust soovitud suunas (Christiano, 2022; Burton, 2023). Riiklikul tasandil kasutatakse mõjutustegevust geopoliitiliste eesmärkide edendamiseks ja avaliku arvamuse mõjutamiseks nii sise- kui välispoliitilises kontekstis (Bjola, 2018; Starbird jt, 2019). Info relvastamine võib hägustada piiri sõja ja rahu vahel ning kujuneda osaks hübriidsõja praktikast, kus sihtmärgiks ei ole üksnes vastase institutsioonid, vaid ka ühiskondlik sidusus ja usaldus (Bjola, 2018; Starbird jt, 2019). Näiteks kasutab Venemaa mõjutustegevust Ukraina sõja kontekstis strateegilistes narratiivides ja vaenulikes tegevustes, mille eesmärk on nõrgestada lääneriikide poliitilist tahet Ukrainat toetada (Välisluureamet, 2025).

Mõjutustegevust rakendavad ka mitteriiklikud osalejad, sealhulgas ideoloogilised rühmitused, aktivistlikud võrgustikud, majanduslike huvidega tegutsejad ja muud kollektiivsed tegijad, kes püüavad kujundada avalikke hoiakuid ja tõlgendusraame endale sobival viisil (Christiano, 2022; Burton, 2023). Mõjutustegevus kujutab endast olulist väljakutset demokraatiale ja avalikule arutelule, õõnestades inforuumi terviklikkust ja kodanike suutlikkust teha informeeritud otsuseid (Liagusha ja Iarovyi, 2025; Christiano, 2022). Infooperatsioonid imbuvad sageli loomulikku veebikeskkonda, muutes tavapärase osaluspraktika osaks mõjutusmehhanismidest ja inimesed tahtmatuteks vahendajateks desinformatsiooni levikus (Starbird jt, 2019). Digitaalses meediakeskkonnas süvendavad neid protsesse algoritmidel põhinevad mehhanismid, mis kujundavad nähtavust ja piiravad ligipääsu mitmekesisele informatsioonile, ohustades seeläbi poliitilist võrdsust ja avatud debatti (Caled ja Silva, 2022; Christiano, 2022; Kertysova, 2018).

Mõjutustegevust ei saa siiski käsitleda olemuslikult negatiivse nähtusena. Sihipärane hoiakute ja käitumise kujundamine võib toimuda ka avalikes huvides, näiteks tervise või kriisikäitumise toetamiseks (World Health Organization, 2017). Erinevalt manipuleerivast või eksitavast mõjutustegevusest ei tugine selline lähenemine hirmu, desinformatsiooni ega varjatud mõjutusvõtete kasutamisele, vaid usaldusväärse ja kontrollitud informatsiooni vahendamisele

(Samas). Seega sõltub mõjutustegevuse normatiivne hinnang eesmärkidest, kasutatud vahenditest ja läbipaistvusest.

1.2.2. Mõjutamisvõtted ja retoorilised strateegiad

Kui mõjutustegevus tähistab strateegilist eesmärki ehk seda, mida ja miks püütakse saavutada, siis mõjutamisvõtted tähistavad vahendeid, mille kaudu seda mõju ellu viiakse (Burton, 2023; Bjola, 2018). Sotsiaalteadustes käsitletakse mõjutamisvõtteid strateegiliste vahenditena, mille abil püütakse suunata inimeste otsuseid ja reaktsioone (Kaptein jt, 2015; Van Benthem jt, 2015; Persson, 2018). Klassikalises sotsiaalpsühholoogilises käsitluses on need näiteks autoriteedi, vastastikkuse, konsensuse või piiratud kättesaadavuse printsiibid, mis suurendavad sõnumi veenvust (Kaptein jt, 2015).

Samas ei ole mõjutamisvõtted üksnes neutraalsed tööriistad. Propaganda ja strateegilise mõjutuskommunikatsiooni käsitlused rõhutavad, et need võivad olla seotud võimu, ideoloogia ja lojaalsuse kujundamisega ning nende eesmärk ei pruugi olla ainult veenmine, vaid ka kollektiivse tegutsemise suunamine soovitud suunas (Hyzen, 2021; Bakir jt, 2019; Burton, 2023). Mõjutamine muutub manipuleerimiseks siis, kui adressaadile ei avata tegutsemise tegelikke põhjuseid või kui kasutatakse ära tema kognitiivseid nõrkusi ja haavatavusi (Klenk, 2022; Christiano, 2022). Sellest tulenevalt ei ole mõjutamisvõtete analüüsimisel oluline ainult see, milliseid võtteid kasutatakse, vaid ka see, millisel eesmärgil ja millise läbipaistvusega neid rakendatakse.

Mõjutamisvõtted realiseeruvad tekstis ja suhtluses keeleliste ning diskursiivsete mehhanismide kaudu, mille abil suunatakse auditooriumi tõlgendusi ja hoiakuid (Hyzen, 2021). Seega on retoorika mõjutustegevuse ja mõjutusvõtete keskne toimimisvorm, mille abil kujundatakse, millised tõlgendused muutuvad usutavaks, loomulikuks või emotsionaalselt veenvaks (Starbird jt, 2019). Klassikalises käsitluses eristatakse kolme põhilist veenmisvahendit: *ethos* (kõneleja usaldusväärsuse ja autoriteedi konstrueerimine), *logos* (ratsionaalsuse mulje loomine argumentatsiooni ja faktide kaudu) ja *pathos* (auditooriumi tunnete kujundamine soovitud reaktsiooni saavutamiseks) (Martin, 2016).

Retooriline mõjutamine ei piirdu otsese argumenteerimisega, vaid hõlmab viise, kuidas sõnumeid raamitakse ja milliseid tähendusi esile tõstetakse (Martin, 2016). Levinud mõjutustegevuste strateegiateks on raamistamine, emotsionaalne rõhutamine, analoogiate kasutamine ja sümbolid, mille kaudu paigutatakse info kindlasse tõlgendusraami (Hyzen, 2021;

Van Benthem jt, 2015). Teatud faktide valikuline esiletõstmine, konnotatsioonide loomine või eeldusi sisaldavad väited võivad suunata adressaadi tõlgendust kaudselt, ilma et väiteid peaks esitama otseselt käskivas või veenvas vormis (Van Benthem jt, 2015). Näiteks populistlikus diskursuses konstrueeritakse vastandus "rahva" ja "eliidi" vahel, taandades keerukad poliitilised küsimused moraalseteks vastandusteks (Wirz, 2018).

Retoorika oluline mõõde on selle emotsionaalne toime (Petty ja Briñol, 2015). Emotsionaalne raamimine võib käivitada hindamisprotsesse, mis toimivad ilma põhjaliku teadliku analüüsita ja tugevdada olemasolevaid seisukohti või identiteedipõhiseid hoiakuid (Petty ja Briñol, 2015; Deffuant jt, i.a.). Mõjutusvõtted on eriti tõhusad siis, kui need haakuvad adressaadi identiteedi ja grupikuuluvusega, sest sellisel juhul ei ole sõnumi mõju üksnes kognitiivne, vaid ka sotsiaalne ja emotsionaalne (Starbird jt, 2019; Hoeken jt., 2016; Lund jt, 2020).

Retoorika ei ole iseenesest eksitav ega problemaatiline. Probleemseks muutub see siis, kui retoorilisi vahendeid kasutatakse viisil, mis piirab adressaadi autonoomiat, varjab tegelikke tegutsemispõhjuseid või suunab arutelu läbipaistmatult etteantud järelduse poole (Klenk, 2022; Christiano, 2022). Sellises tähenduses võib rääkida demagoogiast kui retoorika normatiivselt problemaatilisest kasutusviisist (Aikin ja Casey, 2024). Demagoogiline suhtlus ei püüa avardada arutelu ega toetada informeeritud kaalutlust, vaid sulgeda arutelu enneaegselt, kasutades teadlikult ära vastuvõtja kognitiivseid kaldeid ja piiratud ratsionaalsust (Aikin ja Casey, 2024; Christiano, 2022; Van Benthem jt, 2015).

1.2.3. Digiplatvormide roll mõjutustegevuse levikus

Digiplatvormid mängivad mõjutustegevuse levikus olulist rolli, kujundades sisu tootmise, nähtavuse ja vastuvõtu tingimusi (Boulianne jt, 2022; Bakirov ja Suleimenov, 2025; Lazer jt, 2020). Kui mõjutamisvõtted ja retoorilised strateegiad selgitavad, kuidas mõju tekstis ja suhtluses realiseerub, siis digiplatvormid määravad suurel määral selle, millise ulatuseni ja millistes võrgustikes need sõnumid levivad (Kertysova, 2018; Bakirov ja Suleimenov, 2025). Platvormide toimeloogika hõlmab algoritme, jagamis- ja kommenteerimisvõimalusi, kasutajakontode loomise tingimusi ning mehhanisme, mille alusel sisu võimendatakse või piiratakse (Kertysova, 2018; Bakirov ja Suleimenov, 2025). Need mehhanismid mõjutavad seda, milline info kasutajani jõuab ja kui kiiresti see levib (Christiano, 2022; Kertysova, 2018). Seetõttu on digiplatvormid nähtavuse, manipuleerimise, koordineeritud levitamise ja sisu

võimendamise mõttes mõjutustegevusele eriti vastuvõtlikud (Kertysova, 2018; Bakirov ja Suleimenov, 2025).

Oluline roll on algoritmidel, mille kaudu personaliseeritakse kasutajatele kuvatavat infot nende varasemate eelistuste, tegevuste ja võrgustike põhjal (Caled ja Silva, 2022; Christiano, 2022; Kertysova, 2018). Selline loogika võib kaasa tuua filtrimullide kujunemise, kus kasutajad puutuvad ülekaalukalt kokku nende olemasolevaid hoiakuid kinnitava sisuga (Caled ja Silva, 2022). Sarnast mõju võimendavad kajakambrid, kus narratiive kopeeritakse ja levitatakse paralleelselt mitmes omavahel seotud kanalis, tekitades petliku arusaama nende laialdasest levikust ja toetusest (Starbird jt, 2019).

Digiplatvormid võimaldavad kasutajatel eri määral oma identiteeti varjata, moonutada või automatiseerida, mis muudab algallikate ja levitajate tuvastamise keerulisemaks (Caled ja Silva, 2022). Sellistes tingimustes saavad kasutajad tegutseda anonüümselt ja koordineeritult (Caled ja Silva, 2022; Kertysova, 2018). See raskendab manipulatsiooni äratundmist nii tavakasutajate kui ka ekspertide jaoks, eriti juhul, kui sama sõnum levib korraga mitme kanali ja formaadi kaudu (Caled ja Silva, 2022; Starbird jt, 2019).

Botid, botivõrgustikud, algoritmid ja muud automatiseeritud tööriistad võimaldavad auditooriumi tähelepanu suunata ja sõnumite levikut tehniliselt manipuleerida (Madisson ja Ventsel, 2021; Wanless ja Berk, 2020; Turow, 2005). Botivõrgustikud võivad näiteks kunstlikult koordineerida korduvat jagamist ja mõjutada algoritmilist järjestusi, nii et soovitud sisu muutub nähtavamaks kui konkureerivad tõlgendused (Wanless ja Berk, 2020). Lisaks võivad need luua mulje, et mingil seisukohal on ulatuslik ühiskondlik toetus, kuigi tegelikult on see märksa väiksem või kunstlikult tekitatud (Madisson ja Ventsel, 2021). Süvavõltsingud ja muud manipuleeritud audiovisuaalsed vormid hägustavad piiri autentse ja võltsitud tõendusmaterjali vahel ning suurendavad ebamäärasust ja usaldamatust (Ventsel jt, 2021). Kuigi iga mõjutustegevus ei tugine otseselt botivõrgustikele või süvavõltsingutele, muudab platvormide ülesehitus need keskkonnad mõjutustegevuse jaoks eriti sobivaks (Boulianne jt, 2022; Caled ja Silva, 2022; Ventsel jt, 2021).

1.3. Tahtlikkus mõjutustegevuses

1.3.1. Tahtlikkuse mõiste ja kavatsuse tuletamine

Tahtlikkus tähistab sotsiaalteaduslikus vaates eesmärgipärast tegevust, mis lähtub teadlikest või vähemalt põhjendatavatest kavatsustest ja on suunatud mingi tulemuse saavutamisele sotsiaalselt vahendatud kontekstis (Shepherd ja Carter, 2023; Tollefsen, 2002; Prichard, 2017; Ottaviani jt, 2021; Murdock, 2016; Burton, 2023; Liagusha ja Iarovyi, 2025; Christiano, 2022). Tahtlik tegevus eeldab seega mitte ainult eesmärgi olemasolu, vaid ka teatavat kontrolli ja valikuvõimet (Shepherd ja Carter, 2023; Kärki, 2023; Tollefsen, 2002).

Klassikalises tegevusteoorias seostatakse tahtlikkust agendi võimega omistada oma tegevusele subjektiivne tähendus ja tegutseda eesmärkidest lähtudes (Shepherd ja Carter, 2023; Munch, 1975), samuti agentsusele kui võimele valida tegutsemise ja mittetegutsemise vahel (Kärki, 2023). Kuigi tegevusteooria keskendub individuaalsetele agentidele, ei piirdu tahtlikkus üksikisikutega. Rühmad võivad toimida eesmärgipäraste agentidena, kui neil on ühised sihid ja institutsionaliseeritud otsustusmehhanismid (Prichard, 2017; Tollefsen, 2002). Samas ei taga kollektiivne tahtlikkus täielikku kontrolli tähenduse üle. Agent võib sõnumisse kodeerida soovitud tähenduse, kuid tõlgendajad võivad selle ümber mõtestada või tagasi lükata (Murdock, 2016; Ottaviani jt, 2021; Makinda, 2021).

Kommunikatsiooniteoorias ei käsitleta tahtlikkust ühtse nähtusena, vaid mitmetasandilise mõistena, mis võib tähendada nii teadlikku petmise kavatsust kui ka laiemat eesmärki suunata seda, kuidas vastuvõtja sõnumit tõlgendab (French jt, 2024; Armstrong, 2023; Warren ja Call, 2022). Tahtlikkust seostatakse eesmärgiga mõjutada adressaadi tähelepanu, taju või muud psühholoogilist seisundit, et saavutada soovitud reaktsioon või käitumine (Warren ja Call, 2022). Samas on pakutud ka laiemat vaadet, mille järgi kommunikatsioon ei eelda alati keerukalt teadvustatud ja vastastikku ära tuntavat kavatsust, vaid võib toimida ka paindlikult juhitud tegevusena, mille eesmärk on koordineerida osapoolte arusaamu maailmast (Armstrong, 2023).

Kuigi tahtlikkus on mõjutustegevuse seletusprintsiip, on teoreetilises käsitluses jõutud konsensusele, et kavatsus ei ole otseselt empiiriliselt jälgitav (Kärki, 2023; Khosravi ja Barekat, 2021). Tegevuse tähendus sõltub alati sotsiaalsest ja kommunikatiivsest raamistikust, milles seda tõlgendatakse (Saaristo, 2006). Osa tegevustest võib olla osaliselt harjumuslik või reflekteerimata ja seetõttu ei pruugi agent ise oma kavatsust täielikult sõnastada, isegi kui tema tegevus on eesmärgipärane (Gebauer ja William, 2000; Levy, 2025). Analüütiliselt ei ole

seetõttu määrav mitte see, kas autor suudab oma kavatsust täielikult sõnastada, vaid see, kas tegevust on võimalik seletada põhjuste ja eesmärkide kaudu (Levy, 2025; Chen jt, 2023; Kärki, 2023).

Digitaalses keskkonnas muutub kavatsuse tuletamine veelgi keerulisemaks, kuna anonüümsus, algoritmide läbipaistmatus ja leviku hajutatus raskendavad nii sõnumite päritolu kui ka eesmärgi kindlakstegemist (Burton, 2023; Bakirov ja Suleimenov, 2025; Bjola, 2018). Lisaks ei pruugi vastuvõtja tõlgendus kattuda autori kavatsusega. Tähendus sünnib teksti, konteksti ja tõlgendaja koostoimes ning sama sõnum võib eri kogukondades omandada erineva funktsiooni (Murdock, 2016; Ottaviani jt, 2021). Seetõttu ei käsitle ma töös tahtlikkust otseselt mõõdetava omadusena, vaid tekstiliste ja kontekstuaalsete tunnuste põhjal tehtava järeldusena võimaliku strateegilise eesmärgipärasuse kohta.

1.3.2. Tahtlikkuse roll mõjutustegevuse eristamisel

Tahtlikkus on mõjutustegevuse analüüsimisel keskne tegur, kuna see võimaldab eristada strateegilist manipulatsiooni juhuslikust eksimusest või tavapärasest kommunikatiivsest mõjust (Hyzen, 2021; Starbird jt, 2019). Täpsemalt aitab see eristada argumenteeritud veenmist manipuleerivast tegevusest ja hinnata, kas kommunikatsioon kasutab ära adressaadi kognitiivseid ja identiteedipõhiseid haavatavusi (Bakir jt, 2019; Kertysova, 2018; Burton, 2023; Christiano, 2022). Ilma kavatsuse dimensioonita ei ole võimalik selgelt eristada mõjutustegevust lihtsalt väärinfost ega hinnata tegutsejate rolli (Starbird jt, 2019; Bjola, 2018).

Samas ei tähenda tahtlikkuse keskne roll seda, et see oleks tekstis vahetult kättesaadav, vaid seda alati hinnata tõlgenduse ja konteksti kaudu, mitte üksikute sõnade või väljendite põhjal (Khosravi ja Barekat, 2021; Ottaviani jt, 2021). Kuna kavatsus on järeldus, mis tehakse käitumise, teksti, konteksti ja tagajärgede põhjal, on tahtlikkuse uurimine mõjutustegevuse kontekstis keeruline (Kärki, 2023; Klenk, 2022). See järeldus ei ole kunagi üheselt otsustatav, vaid eeldab tõlgendust, mis seob need aspektid oletatava eesmärgi või motiiviga (Ottaviani jt, 2021). Kavatsuse empiirilise tuvastuse teeb eriti raskeks asjaolu, et mõjutuspraktikad on sageli strateegilised, varjatud, hajutatud üle platvormide ja paljude osalejatega (Starbird jt, 2019; Bjola, 2018).

Lisaks raskendab tahtlikkuse tuvastamist ka see, et sageli kasutatakse sõnumis korraga mitut mõjutusvõtet. Kui sõnum kombineerib näiteks autoriteeti, konsensust ja emotsionaalset survet, ei ole lihtne öelda, milline konkreetne võte kannab tahtlikkuse hinnangus suurimat kaalu

(Kaptein jt, 2015). Lisaks võib mõjutustegevus tugineda ka tõestele või osaliselt tõestele väidetele, mis hägustab piiri heatahtliku infojagamise ja strateegilise manipulatsiooni vahel (Hyzen, 2021).

Tahtlikkuse uurimine on tugevalt seotud ka keele ja kultuuriga. Sõnumi tähendus ei sõltu üksnes autori kavatsusest, vaid kogukonna ajaloost, normidest ja tõlgendusraamidest, mistõttu võib sama indikaator eri rühmades tähendada erinevaid asju (Ottaviani jt, 2021; Murdock, 2016). Tähendust ei kujunda ainult sõnade otsene sisu, vaid ka nende koht suhtluses ja kultuurilised raamid, mis määravad, mis on sobiv või oodatav (Humă, 2023; Smith, 2024). Kultuurilised erinevused suhtlusstiilides mõjutavad otseselt seda, kuidas kavatsusi tõlgendatakse, mistõttu nõuab tahtlikkuse analüüs spetsiifilisi teadmisi diskursiivsetest normidest ja kontekstist (Tanduk, 2023; Usmani ja Almashham, 2024; Smith, 2024).

Asjaolu, et tahtlikkuse uurimisel ei saa kavatsust käsitleda tekstist otseselt välja loetava omadusena, mõjutab ka see, et kavatsus on tekstis alati vahendatud. Saatja kodeerib tähenduse ning vastuvõtja dekodeerib selle oma kogemuse, ideoloogia ja grupinormide kaudu (Murdock, 2016). Sama väljend võib eri kogukondades täita eri funktsiooni ja tähendus tuleneb vastuvõtja kogemustest. Näiteks meemikultuuris võib vastuvõtja sõnumit tõlgendada ja edasi levitada hoopis uue tähendusega, hägustades seeläbi piiri sihipärase propaganda ja siira osaluse vahel (Ottaviani jt, 2021; Liagusha ja Iarovyi, 2025). Seega tuleb tahtlikkust käsitleda vahendatud ja hinnangulise konstruktsioonina, millele viitavad teatud tekstilised ja kontekstuaalsed tunnused.

Kuna tahtlikkus on tekstis keeruliselt hinnatav ja oleneb tõlgendusest, ei tuleks seda käsitleda iseseisva tunnusena mõjutustegevuse hindamisel. Tahtlikkus on väärtuslik tunnus laiemas hindamisraamistikus, mille kõrvale on vaja teisi näitajaid, näiteks allika usaldusväärsust ja koordineeritud tegevuse märke (French jt, 2024; Saeidnia jt, 2025).

1.4. Arvutuslikud lähenemised sotsiaalteadustes

Arvutuslik lähenemine sotsiaalteadustes (*computational social science*) tähistab arvutusmeetodite arendamist ja rakendamist sotsiaalsete nähtuste uurimiseks (nt polariseerumine, radikaliseerumine, infosõda), empiiriliseks testimiseks ja mustrite tuvastamiseks suuremahulistes andmestikes (Hox, 2017; Lazer jt, 2020; Burton, 2023; Bakirov ja Suleimenov, 2025). Sellisteks andmeteks on näiteks sotsiaalmeedia postitused, metaandmed ja muud veebikeskkondades tekkivad tekstilised või käitumuslikud digitaalsed jäljed (Hox, 2017; Lazer jt, 2020). Tänapäevane infosfäär toodab selliseid andmeid mahus, mille puhul

üksnes inimanalüüsist ei piisa, mistõttu muutuvad arvutuslikud meetodid sotsiaalteadustes üha olulisemaks (Starbird jt, 2019; Weitkamp jt, 2021; Kertysova, 2018; Bjola, 2018; Supriyono jt, 2024).

Arvutuslike lähenemiste väärtus seisneb selles, et need võimaldavad töödelda suuri ja keerukaid andmemahtusid, et leida neist korduvaid mustreid (Hox, 2017; Lazer jt, 2020). See loob võimaluse mitte ainult juba toimunud nähtuste kirjeldamiseks, vaid ka võimalike riskide varajaseks märkamiseks ja hindamiseks, eriti olukorras, kus inimanalüüs üksi jääb andmemahu tõttu ebapiisavaks (French jt, 2024). Tekstiandmete puhul tähendab see eelkõige võimalust töödelda suuri tekstikorpusi, rühmitada neid ja võrrelda sisuliste tunnuste ning korduvate mustrite alusel (Lazer jt, 2009; Hox, 2017; Pandey jt, 2019).

Suurandmete analüüs algab tavaliselt andmete eeltöötlusest, tunnuste eraldamisest ja sobiva mudeli valikust, et muuta keerukad ja struktureerimata andmed, näiteks tekstid, analüüsitavaks (Pandey jt, 2019; Chang jt, 2014). Tekstiandmete puhul kasutatakse sageli loomuliku keele töötlust (*Natural Language Processing*), mis aitab suuri tekstikorpuseid kodeerida ja sisuliste mustrite järgi võrrelda (Lazer jt, 2009; Hox, 2017). Kui eesmärk on liigitada andmeid etteantud kategooriatesse, rakendatakse mudeli puhul juhendatud õpet ja peidetud rühmade leidmiseks kasutatakse juhendamata meetodeid (Pandey jt, 2019; Jordan ja Mitchell, 2015). Vahel kasutatakse ka simulatsioone ja agendipõhiseid mudeleid, et uurida, kuidas üksiktasandi reeglitest kujunevad laiemad ühiskondlikud mustrid (Hox, 2017; Pianese jt, 2014).

Sotsioloogias on arvutuslike lähenemisi kasutatud näiteks võrgustike, sotsiaalse ebavõrdsuse, diskrimineerimise, kultuuriliste mõjude ja kommunikatsiooni mustrite uurimisel (Lazer jt, 2020; Chang jt, 2014; Hox, 2017). See on võimaldanud testida klassikalisi sotsioloogilisi teooriaid suuremahulistel andmetel, näidates, et arvutuslikud meetodid pakuvad uusi võimalusi teooriate kontrollimiseks (Hox, 2017). Sotsioloogilisi mõisteid on rakendatud ka ennustavates mudelites, kus sotsiaalteooriate abil on uuritud, kuidas inimeste sarnasus, sotsiaalne struktuur ja sidemete tugevus mõjutavad nende tarbimisotsuseid (Verbraken jt, 2014).

Eelkõige on arvutuslikud lähenemised sotsioloogias kujunenud suunaks, kus ennustavad mudelid ja klassikaline sotsiaalteooria ühendatakse, et uurida sotsiaalseid nähtusi senisest ulatuslikumalt ja detailsemalt (Lazer jt, 2009; Lazer jt, 2020; Hox, 2017). Seega toimivad arvutuslikud meetodid otsustustoe vahendina, mis aitavad hallata infomahtu, mida üksnes inimanalüüsiga oleks keeruline jälgida (French jt, 2024).

1.4.1. Masinõpe tekstianalüüsi meetodina

Arvutuslike lähenemiste oluline osa on masinõpe, mille abil saab andmetes tuvastada korduvaid seaduspärasusi ja kasutada neid uute juhtumite klassifitseerimiseks või ennustamiseks (Lazer jt, 2020; Hox, 2017). Masinõpe on tehisintellekti alamvaldkond, kus süsteem õpib automaatselt andmetest mustreid, mille põhjal oma otsustusloogikat kohandada (Pandey jt, 2019; Jordan ja Mitchell, 2015). Praktikas tähendab see mudeli treenimist andmetel ja seejärel selle kasutamist uute juhtumite eristamiseks või klassifitseerimiseks (Pandey jt, 2019).

Masinõppe põhiline väärtus teksti uurimisel seisneb selles, et see võimaldab töödelda väga suuri ja nõrgalt struktureeritud tekstimahte (Hox, 2017). Nii saab uurida näiteks seda, kuidas levivad teatud narratiivid või millised tekstilised tunnused korduvad kindlat tüüpi sõnumites (Hox, 2017; Lazer jt, 2020; Supriyono jt, 2024). Masinõppe tugevus seisneb seega mustrite tuvastamises, mitte sotsiaalse nähtuse täielikus seletamises. Küsimus ei ole niivõrd selles, miks nähtus sotsiaalselt tekib või püsib, vaid selles, kas andmetes leidub piisavalt korrapärasusi, et nende põhjal uusi juhtumeid eristada või ennustada (Hox, 2017; Jordan ja Mitchell, 2015).

Oluline on rõhutada, et masinõpe ei mõista teksti samal viisil nagu inimlugeja. Masinõpe töötab tõenäosuslike seoste ja korrapärasuste alusel ning võib eristada tekstiklasse ka siis, kui see ei mõista nende sotsiaalset või kultuurilist tähendust samal tasemel nagu inimene (Jordan ja Mitchell, 2015; Abro jt, 2023). Seetõttu ei saa masinõpet käsitleda inimtõlgenduse asendajana, vaid vahendina, mis aitab süstemaatiliselt hinnata, kas teatud tunnused korduvad andmetes piisavalt sageli, et neid oleks võimalik automaatselt tuvastada (Abro jt, 2023). See tähendab, et masinõpe sobib hästi olukorda, kus uurija soovib teada, kas andmetes leidub piisavalt korduvaid tunnuseid, et teatud tüüpi tekste üksteisest eristada, kuid see ei tähenda automaatselt, et mudel mõistab nende tekstide sotsiaalset või kultuurilist tähendust samal tasemel nagu inimene (Jordan ja Mitchell, 2015; Abro jt, 2023).

Masinõppe põhijaotus eristab juhendatud õpet (sihtmuutujaga ehk sildistatud andmetega), juhendamata õpet (sihtmuutujata mustrite ja rühmade leidmiseks) ja kinnitusõpet (*reinforcement learning*) (Pandey jt, 2019; Hox, 2017). Juhendatud õppes kohandatakse mudel märgistatud näidete põhjal leidma seoseid sisendi ja etteantud väljundi vahel, juhendamata õppes õpib mudel aga andmetes esinevaid seaduspärasusi ilma eelnevalt määratud siltideta (Conneau jt, 2020). Kinnitusõppes kujuneb õppimine tagasiside kaudu, mille alusel mudel

optimeerib oma edasist otsustusprotsessi soovitud tulemuse saavutamiseks (Pandey jt, 2019; Hox, 2017).

Lisaks õppimisviisidele saab masinõppemudeleid eristada ka väljundite järgi. Generatiivne mudel on masinõppemudel, mis loob varasemalt õpitud mustrite ja andmete põhjal uut sisu (Singh jt, 2025; Wu jt, 2023). Kui tavalises masinõppes õpib mudel treeningandmete põhjal uusi otsuseid tegema, siis generatiivne mudel kasutab eelneva treenimise käigus õpitud mustreid, et luua kasutaja juhise ehk *prompt*-i põhjal uut sisu (Samas).

See, mida mudel lõpuks õpib, sõltub otseselt inimeste tehtud valikutest mudeli, andmete ja analüüsi kujundamisel (Hox, 2017; Abro jt, 2023). Uurijad määravad, millist sihtmuutujat ennustatakse, milliseid tunnuseid kasutatakse, milline mudelitüüp valitakse ja kuidas parameetreid häälestatakse, mistõttu ei ole mudeli õppimine kunagi täielikult automaatne ega neutraalne protsess (Hox, 2017; Jordan ja Mitchell, 2015). Inimesed mõjutavad tulemust ka andmete eeltöötluse, tunnuste eraldamise ja optimeerimise kaudu, mis tähendab, et mudeli väljund peegeldab osaliselt uurija varasemaid otsuseid selle kohta, mida peetakse analüütiliselt oluliseks (Pandey jt, 2019). Ka uurijate taust ja väärtused mõjutavad seda, milliseid mustreid nad andmetes oluliseks peavad ja kuidas nad neid tõlgendavad (Starbird jt, 2019). Generatiivse mudeli puhul mõjutab kasutaja antud *prompt* ehk juhis otseselt mudeli väljundi stiili ja sisu (Singh jt, 2025). Masinõppe väljundit tuleb seega tõlgendada mitte ainult mudeli tehnilise sooritusena, vaid ka eelnevate valikute peegeldusena.

1.4.2. Arvutuslikud lähenemised tahtliku mõjutustegevuse uurimisel

Varasem kirjandus on näidanud, et masinõpet saab kasutada propaganda, botikäitumise ja teiste mõjutustegevustega seotud nähtuste tuvastamisel (Hox, 2017; Jordan ja Mitchell, 2015; Lazer jt, 2020). Desinformatsiooni uurimisel on osutatud ka sellele, et arvutuslikud lähenemised võimaldavad liikuda reaktiivselt väärinfo tuvastamiselt proaktiivsema seire suunas, kus hinnatakse võimalike kampaaniate mahtu ja polariseerivat mõju juba enne nende laiemat levikut (French jt, 2024). Selles mõttes on arvutuslikel lähenemistel potentsiaal aidata märgata mõjutustegevuse mustreid ka siis, kui andmemaht ületab käsitsi tehtava analüüsi praktilised piirid.

Arvutuslikud lähenemised aitavad märgata anomaaliaid ja tekstilisi seaduspärasusi, kuid sotsiaalteaduslik selgitus vajab nende kõrvale alati kvalitatiivset konteksti ja inimese poolset tõlgendust (Starbird jt, 2019). Nagu ka valeinfo automaattuvastuses, ei piisa mustri leidmisest

lõpliku hinnangu andmiseks, sest süsteemid vajavad valideerimist ja kontekstualiseerimist (Caled ja Silva, 2022).

Tahtliku mõjutustegevuse uurimisel tuleb siiski arvestada, et tahtlikkus ei ole tekstis otseselt nähtav omadus, vaid järeldus, mis tehakse keeleliste, retooriliste ja kontekstuaalsete mustrite põhjal (Kärki, 2023; Khosravi ja Barekat, 2021; Ottaviani jt, 2021). Seetõttu ei saa arvutuslikku lähenemist käsitleda viisina, mis tuvastaks autori päris kavatsuse otseselt. Selle asemel võimaldab see hinnata, kas tekstides leidub piisavalt korduvaid tunnuseid, mis seostuvad tahtliku mõjutustegevuse hinnanguga (Jordan ja Mitchell, 2015). See ei lahenda tahtlikkuse probleemi lõplikult, kuid võimaldab seda keerukat nähtust empiiriliselt uurida (French jt, 2024).

1.5. Probleemiseade

Tänapäevast infosfääri iseloomustavad infoüleküllus ja suurenev konkurents tähelepanu pärast, mis loovad soodsad tingimused strateegilise mõjutustegevuse levikuks ja raskendavad selle eristamist tavapärasest kommunikatsioonist (Starbird jt, 2019; Persson, 2018; Caled ja Silva, 2022). Mõjutustegevuse analüüsimisel on keskne küsimus, kuidas eristada strateegilist ja eesmärgipärast mõjutamist tavapärasest kommunikatiivsest mõjust, juhuslikust eksimusest või orgaanilisest osalusest (Christiano, 2022; Starbird jt, 2019; Persson, 2018; Burton, 2023). Teoreetilise käsitluse põhjal on selle eristuse võtmetunnuseks tahtlikkus (Hyzen, 2021; Starbird jt, 2019). Samas on tahtlikkuse uurimine metodoloogiliselt keeruline.

Tahtlikkus ei ole tekstis vahetult nähtav omadus, vaid järeldus, mis tehakse tekstiliste, retooriliste ja kontekstuaalsete tunnuste põhjal (Khosravi ja Barekat, 2021). Autori kavatsus ei ole otseselt empiiriliselt jälgitav ja selle tõlgendamine sõltub nii kultuurilistest raamidest kui ka suhtlusnormidest (Ottaviani jt, 2021; Kärki, 2023; Klenk, 2022). Seega ei saa tahtlikkust käsitleda tekstist otseselt välja loetava tunnusena, vaid üksnes vahendatult hinnatava konstruktsioonina.

Varasem kirjandus on näidanud, et arvutuslikud lähenemised ja masinõpe võimaldavad töödelda suuremahulisi tekstikorpusi ja tuvastada korduvaid mustreid, et kasutada neid mõjutustegevusega seotud nähtuste uurimisel (Hox, 2017; Jordan ja Mitchell, 2015; Lazer jt, 2020). Samas ei võimalda masinõpe tuvastada autori päris kavatsust otseselt, vaid üksnes hinnata, kas tekstides leidub piisavalt korrapäraseid tunnuseid, mis seostuvad tahtliku mõjutustegevuse hinnanguga (Jordan ja Mitchell, 2015). Sellest tulenevalt ei ole arvutuslik

lähenemine töös põhjendatud lõpliku tõeallikana, vaid otsustustoe ühe võimaliku komponendina.

Uurimisprobleem seisneb selles, et Eesti keeleruumis puuduvad seni empiirilised katsed, mis hindaksid, kas ja kuidas on võimalik mõõta tahtlikkuse avaldumist tekstides kvantitatiivsete meetoditega. Seetõttu ei ole teada, millisel määral võiks masinõppel põhinev tahtlikkuse indikaator toimida otsustustoe elemendina mõjutustegevuse tuvastamisel. Uurimisprobleem pole ainult teoreetiline ega metodoloogiline, vaid ka rakenduslik. Küsimus on selles, kas tahtlikkuse hinnangut on võimalik muuta mõõdetavaks viisil, mis annaks piisavalt usaldusväärse sisendi praktilise tuvastustööriista arendamiseks. Kui tunnus osutub vähemalt osaliselt masinõppe põhiselt eristatavaks, võib see aidata suures tekstimahus esile tõsta juhtumeid, mis vajavad edasist tähelepanu. Kui aga tahtlikkuse hinnang osutub liiga ebastabiilseks või tugevalt kontekstisõltuvaks, aitab see täpsustada tunnuse veapiiri ja rolli laiemas süsteemis.

Magistritöö eesmärk on välja selgitada, kas masinõppepõhist tahtlikkuse indikaatorit saab kasutada eestikeelsetes tekstides otsustustoe elemendina tahtliku mõjutustegevuse eristamisel tavapärasest kommunikatsioonist.

Uurimisküsimused

1. Kuivõrd täpselt võimaldab masinõpe eristada tekstides tahtlikku mõjutustegevust?
2. Kuidas mõjutab treeningandmestiku mahu suurendamine mudeli tulemuslikkust?

2. METOODIKA

Selles peatükis annan ülevaate töös kasutatud andmetest, valimist ja uurimismeetoditest. Esmalt kirjeldan, kuidas kujunes uuritav tekstivalim ja kuidas moodustati ekspertide ning kontrollgrupi koondhinnangud. Peatüki viimases osas kirjeldan juhendatud tekstiklassifikatsiooni mudelit ja generatiivse keelemudeli kasutamist ning seda, kuidas hindasin mudelite tulemuslikkust.

2.1. Andmed

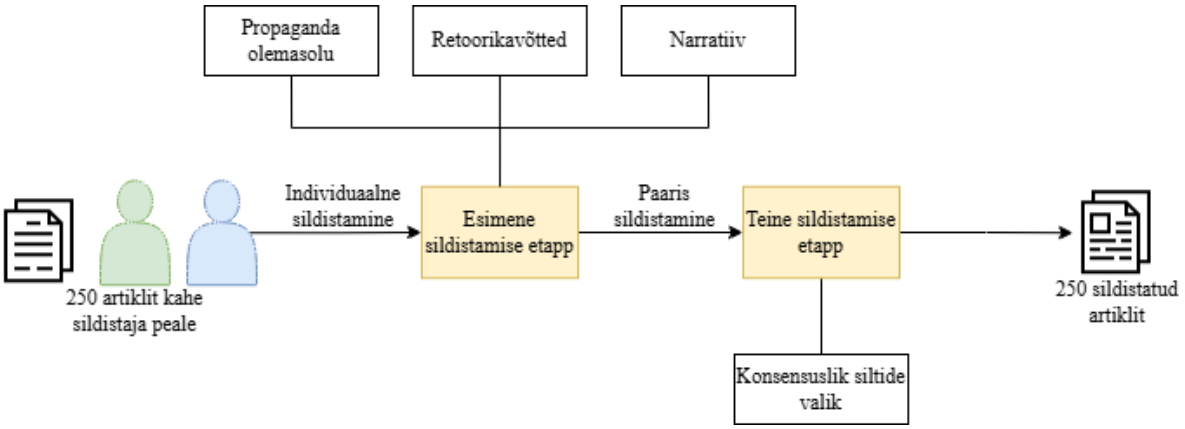
Oma magistritöös kasutan Vilniuse Ülikooli projekti "Propaganda ja desinformatsiooni uurimine: masinõppel põhinev automaatne tuvastamine, mõju ja ühiskondlik vastupidavus" raames kogutud andmeid. Projekti eesmärk on arendada masinõppepõhiseid meetodeid propaganda ja valeinfo automaatseks tuvastamiseks ning uurida nende mõju ja võimalusi tugevdada ühiskonna vastupanuvõimet infosõja kontekstis (ATSPARA, i.a.). Selleks sildistasin paralleelselt koos rahvusvahelise meeskonnaga käsitsi artikleid, et luua märgendatud andmestik, mille põhjal oleks võimalik treenida ja peenhäälestada propaganda automaatseks tuvastamiseks kasutatavaid keelemudeleid (Rizgelienė jt, 2025).

Projekti tasandil osalesin ühe sildistajana, kuid algallikate valik, tekstikorpuse moodustamine ja sildistusmetoodika olid osa Vilniuse Ülikooli projekti metoodikast. Projekti raames määratleti esmalt, millistest eestikeelsetest veebiallikatest koguda tekstid, mida retoorikavõtete analüüsiks kasutada. Kuna Eesti kontekstis puudub terviklik uuring, mis võrdleks süstemaatiliselt kõiki eestikeelseid uudisportaale nende seotuse või vastuvõtlikkuse osas propagandanarratiividele, tugines projekti allikavalik varasematele uurimustele Eesti alternatiivmeediast ja Propastopi esile toodud juhtumitele (Nädala vandenõuteooriad…, 2020; Seos FB faktikontrolliga…, 2020; Urve Eslas…, 2017). Valitud allikateks olid Telegram, Uued Uudised, Objektiiv ja Vanglaplaneet.

Artiklite kogumisel lähtuti projektis ajalisest piirangust, mille kohaselt kaasati ainult tekstid, mis olid avaldatud alates 2018. aastast. Perioodi valik oli seotud projekti uurimisloogikaga, mille eesmärk oli keskenduda ajavahemikule, mis kataks nii neli aastat enne kui ka neli aastat pärast Venemaa täiemahulise sõja algust Ukrainas 2022. aastal. Kokku moodustati projekti raames eestikeelne tekstikorpus, kuhu kuulus 23 641 teksti.

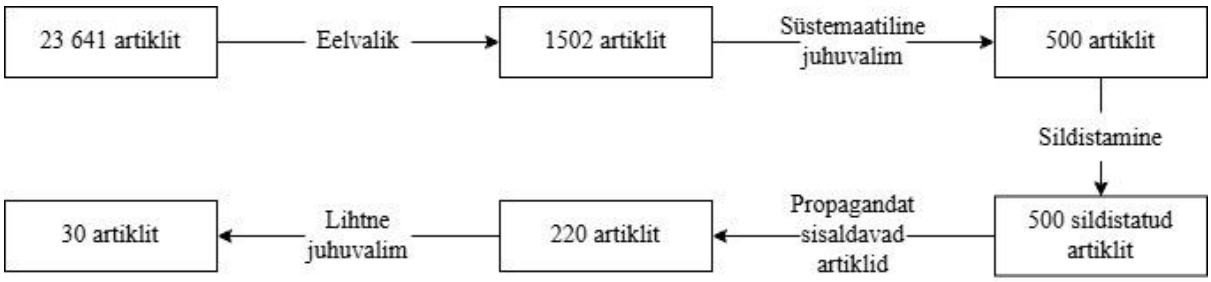
Selle korpuse põhjal tehti pealkirja ja juhtlõigu alusel eelvalik, mille käigus hinnati, kas tekstis võib potentsiaalselt esineda propagandat. Eelvaliku eesmärk oli suurendada tõenäosust, et hilisemasse valimisse jõuaksid tekstid, milles võib esineda propagandat või muid mõjutustunnuseid. Alles jäi 1502 artiklit. Kuna projekti järgmises etapis toimus tekstide käsitsi sildistamine, pidi artiklite arv jääma sildistajate töökoormust arvestades realistlikult teostatavaks. Seetõttu moodustati 1502 artiklist süstemaatilise juhuvalimi alusel 500 artikliga valim. Lõplik valim ei olnud seega täielikult juhuslik, vaid kujunes osaliselt varasemate valikute põhjal.

Ajavahemikus november 2024 kuni juuni 2025 sildistasid kaheliikmelised paarid need 500 artiklit (vt Joonis 1). Sildistamises osales kokku neli sildistajat ja tekstid jaotati paaride vahel võrdselt ehk kumbki paar sildistas 250 teksti. Iga teksti puhul märgiti esmalt individuaalselt nii propaganda võimalik esinemine kui ka tekstis kasutatud retoorikavõtted. Lõplik hinnang kujunes kahe sildistaja konsensusliku otsuse alusel.



Joonis 1. Projekti tekstide käsitsi sildistamise protsess.

Kui algse tekstikorpuse loomine kuulus Vilniuse Ülikooli projekti metoodika alla, siis minu magistritöö iseseisev osa algas projekti märgendatud andmestikust eraldi analüüsivalimi moodustamisega. Selleks eraldasin projekti raames sildistatud 500 artikli hulgast kõik tekstid (n = 220), mis määratletud propagandat sisaldavaks (vt Joonis 2). Valikul lähtusin eeldusest, et tahtlikkuse hindamine on analüütiliselt põhjendatum tekstides, kus eelnev sildistamine oli juba esile toonud propaganda olemasolu. Võimalikult mitmekesise tekstivalimi saamiseks võtsin propagandat sisaldavate tekstide hulgast lihtsa juhuvalimi teel 30 teksti, mille põhjal analüüsisin tahtliku mõjutustegevuse esinemist.



Joonis 2. Analüüsivalimi moodustamine projekti tekstikorpusest.

2.2. Valim ja hinnangute kogumine

Varasem märgistus propaganda või retoorikavõtete võimaliku esinemise kohta ei olnud minu töös lõplik sihtmuutuja, vaid alus tahtliku mõjutustegevuse hindamiseks. Töö kestkeks tunnuseks oli seega tahtlikkus, mille jaoks kogusin ma hinnanguid selle kohta, kas tekstides esineb hindajate arvates tahtlikku mõjutustegevust. Enne põhiuuringut tegime juhendajatega testsildistamise, mille eesmärk oli hinnata hindamisülesande ajakulu ja selgust.

Tahtliku mõjutustegevuse esinemist tekstides hindasid meedia- ja kommunikatsioonivaldkonna eksperdid (n = 7) ja mitteekspertidest koosnev kontrollgrupp (n = 7). Ekspertide kaasamise eesmärk oli luua võrdlusalus, mille abil masinõppemudelite suutlikkust hinnata. Kontrollgrupi eesmärk oli esindada tavalugeja vaadet ehk hinnangut inimestelt, kellel puudub ekspertidega võrreldav erialane kogemus propaganda ja mõjutustegevuse analüüsimisel. Nende kaasamine oli vajalik kahel põhjusel. Esiteks andis see võimaluse testida, kas kontrollgrupi häälteenamushinnang sarnaneb ekspertide omale või nõuab tahtlikkuse tuvastamine spetsiifilisi erialaseid teadmisi. Teiseks aitab see uurida automaatse otsustoe vajalikkust. Kui tavalugejal on raskusi tahtlikkuse eristamisel tavapärasest kommunikatsioonist, viitab see vajadusele tehnoloogiliste abivahendite järele.

Eksperdid kaasasin sihipärase valimi alusel, arvestades nende varasemat teaduslikku või praktilist kogemust meedia, kommunikatsiooni, propaganda või infomõjutustegevuse uurimisega. Nende kaasamiseks saatsin kutse kokku 26-le potentsiaalsele osalejale, kellest nõustus uuringus osalema seitse. Kaasatud eksperdid olid mitmekesise taustaga, hõlmates nii meedia- ja kommunikatsiooniuuringute, ühiskonnateaduste, poliitilise analüüsi kui ka mõjutustegevuse ja propaganda uurimisega seotud kogemust. Mitteekspertidest koosneva kontrollgrupi moodustasin mugavusvalimi alusel, lähtudes eesmärgist, et vanuseline jaotus oleks ligikaudu võrreldav ekspertide rühmaga. Kontrollgrupi osalejad kutsusin oma

tutvusringkonnast, vältides ülikooli ja õppekavaga seotud vahetut võrgustikku, et vähendada võimalikku erialast kallutatust. Nende kaasamiseks saatsin kutse kokku 15 potentsiaalsele osalejale, kellest osales seitse.

Uuringu tekstid edastasin osalejatele e-posti teel PDF-failidena. Osalejatele saadetud tekstid sisaldasid projekti raames märgistatud retoorikavõtteid, mis olid värvikoodidega esile tõstetud. Värvikoodide selgitused olid toodud iga teksti alguses. Märgistused aitasid vähendada osalejate töökoormust ja hindamisele kuluvat aega. Vilniuse Ülikooli projekti sildistamise etapis kulus 500 artikli propaganda ja retoorikavõtete käsitsi sildistamisele kaheksa kuud. Juhendajatega tehtud katsesildistamine näitas, et 30 teksti tahtliku mõjutustegevuse hindamine võttis aega ligikaudu üks tund kuni poolteist tundi. Hindamisülesande ajakulu vähendamine tõstis võimalust, et uuringus oleks nõus osalema võimalikult palju eksperte.

Samas tuleb arvestada, et tekstidele eelnevalt lisatud retoorikavõtete sildid võisid suunata osalejate tähelepanu teatud lõikudele ja seeläbi tõsta tahtlikkuse tajutavust, võrreldes olukorraga, kus tekst oleks esitatud märgistuseta. Retoorikavõtted olid kõikides hinnatud tekstides ehk hinnangut võis mõjutada see, milliseid võtteid tekstis kasutati ja kuidas need mõjusid.

Vastuste kogumiseks kasutasin Exceli ankeeti, mille osalejad täitsid ja tagastasid e-posti teel. Kõiki saadetud ja tagastatud faile ning informeeritud nõusoleku vorme hoian Tartu Ülikooli SharePointi keskkonnas, millele on ligipääs üksnes minul ja töö juhendajatel.

Kuna magistritöö eesmärk on välja selgitada, kas masinõppepõhist tahtlikkuse indikaatorit on võimalik eestikeelsetes tekstides kasutada otsustustoe elemendina, palusin osalejatelt anda iga teksti kohta binaarse hinnangu küsimusele, kas tekstis esineb tahtlikku mõjutustegevust (jah/ei). Binaarne hinnang oli vajalik, sest töös kasutatavad klassifikatsioonimudelid eeldavad treeningandmetes selgeid kategoorilisi silte, mille alusel õppida eristama eri tunnustega tekstiklasse.

Analüüsi sihtmuutujaks oli binaarne tunnus, mis väljendas ekspertide koondhinnangut sellele, kas tekstis esineb tahtlikku mõjutustegevust. Vastused kodeeriti kahte klassi: “ei” = 0 ja “jah” = 1. Kuna nii ekspertide kui ka kontrollgrupi koosseisus oli seitse hindajat, sai tekst koondhinnangu “jah” juhul, kui vähemalt neli hindajat seitsmest leidis, et tekstis esineb tahtlikku mõjutustegevust. Sama põhimõtet kasutasin kontrollgrupi koondhinnangu moodustamisel.

Ekspertide häälteenamushinnangut käsitlen võrdlusalusena, mille vastu võrdlen nii kontrollgrupi koondhinnangut kui ka mudelite ennustusi. See ei tähenda, et ekspertide hinnang väljendaks autori tegelikku kavatsust otseselt, vaid et töös kasutan ekspertide koondhinnangut parima kättesaadava võrdlusalusena. Kontrollgrupi häälteenamushinnangut kasutan eraldiseisva võrdlusnäitajana, et hinnata, kuivõrd lähedale jõuab tavalugejate koondhinnang ekspertide koondhinnangule.

Tabel 1. Ekspertide ja kontrollgrupi häälteenamushinnangu jaotus.

	Ekspertgrupp		**Kontrollgrupp**	
Klass	**Arv**	**Osakaal**	**Arv**	**Osakaal**
Jah	21	70%	27	90%
Ei	9	30%	3	10%

Tabelist 1 on näha, et ekspertide häälteenamushinnangu põhjal jagunes 21 teksti tahtlikku mõjutustegevust sisaldavaks ja üheksa teksti hinnati selliseks, kus tahtlikku mõjutustegevust ei esinenud. Kontrollgrupi häälteenamushinnangu põhjal jagunes 27 teksti tahtlikku mõjutustegevust sisaldavaks ja kolm teksti selliseks, kus seda ei esinenud.

2.3. Masinõppe- ja keelemudelite katsetamine

Töö eesmärgist lähtuvalt kasutasin meetodina juhendatud tekstiklassifikatsiooni mudelit ja generatiivset keelemudelit. Lisaks ekspertandmetel treenitud juhendatud mudelile katsetasin pseudomärgistamise lähenemist, mille abil hinnata, kas generatiivse mudeli abil loodud lisaandmed parandavad juhendatud mudeli tulemuslikkust. Seejärel võrdlesin mudelite tulemusi, et hinnata nende toimimist ja veamustreid. Võrdluse eesmärk ei olnud leida parimat mudelit, vaid välja selgitada, kas tahtlikkuse tuvastamine tekstis on masinõppepõhiste meetoditega võimalik.

Kõikide mudelite lõplikul hindamisel kasutasin samu artikleid, mida eksperdid ja kontrollgrupp hindasid. Ainsaks erinevuseks oli see, et mudelitele esitasin tekstid ilma esile tõstetud retoorikavõtete märgistusteta. See otsus tulenes asjaolust, et mudelid liigitasid sildistatud tekstid automaatselt “jah” klassi. Sellisel juhul ei oleks olnud võimalik hinnata mudeli tegelikku võimekust tekstiliste mustrite tuvastamisel. Inimhindajatele jäeti sildid alles, et vähendada hindamisülesande ajakulu.

2.3.1. Juhendatud tekstiklassifikatsioon

Esimeses lähenemises kasutasin juhendatud masinõpet (*supervised learning*). See tähendab, et mudel õpib treeningandmete põhjal, kus iga tekst on eelnevalt märgistatud vastava kategooriaga (Pandey jt, 2019). Selline lähenemine on põhjendatud siis, kui uurimisküsimus on sõnastatav klassifitseerimisülesandena, ehk eesmärk on hinnata, kas tekst seostub teatud tunnuse või hinnanguga (Abro jt, 2023; Supriyono jt, 2024).

Tekstide analüüsimiseks kasutasin transformer-tüüpi keelemudelit *XLM-RoBERTa-base*, mis on Meta AI loodud mitmekeelne mudel ja sobib ka eestikeelse teksti analüüsimiseks (Conneau jt, 2020). *XLM-RoBERTa-base* treeningandmete lõppkuupäev on 05.11.2019 (Samas). Mudeli valikul lähtusin varasematest uuringutest, mis on tõestanud selle tulemuslikkust mitmekeelsete ülesannetega (Tetsmann jt, 2026; Kang jt, 2022). Eesti keele kontekstis mitmekeelseid mudeleid võrreldes, leidis Kittask jt (2020), et need tulevad eestikeelsete ülesannetega hästi toime ja võrreldavatest mudelitest oli parim *XLM-RoBERTa-base*. Transformer-arhitektuuril põhinevad mudelid on tänapäevases loomuliku keele töötluses (*Natural Language Processing*) laialdaselt kasutusel, kuna need suudavad arvesse võtta sõnade konteksti ja nende omavahelisi seoseid kogu tekstis (Supriyono jt, 2024).

Mudeli sisendiks oli 30 teksti ja tekstidele antud hinnangud tahtliku mõjutustegevuse esinemise kohta. Silt "jah" tähistas olukorda, kus ekspertide hinnangul esines tekstis tahtlikku mõjutustegevust ja silt "ei" olukorda, kus ekspertide hinnangul seda ei esinenud.

Väikese andmestiku puhul on oht, et mudel õpib üldise mustri asemel üksikuid tekste. Kuna 30 tekstist koosnev andmestik on väga väike, vähendasin üleõppimise riski, külmutades mudeli põhikihid ja treenides ainult klassifitseerimiskihti. See tähendab, et mudeli varasemalt õpitud üldised keelelised seosed jäid muutmata ja õpitavaks jäi ainult lõplik otsustuskiht. Sellisel juhul kasutab mudel oma varasemat keeleoskust aga kohandab ainult seda, kuidas teha lõplik otsus. Klasside ebaühtlase jaotuse tõttu kasutasin treenimisel klassikaaludega kaofunktsiooni, et vähendada mudeli kalduvust eelistada sagedasemat klassi.

Mudel teisendab sisendteksti matemaatiliseks vektoriks, mis sisaldab informatsiooni teksti semantiliste ja süntaktiliste omaduste kohta ning kasutab seda selleks, et prognoosida, kas tekstis esineb tahtlikku mõjutustegevust või mitte (Supriyono jt, 2024; Pandey jt, 2019). Mudel arvutas esmalt, kui tõenäoliselt kuulub tekst kumbagi võimalikku klassi, kus "jah" tähendas, et

tekstis esineb tahtlikku mõjutustegevust ja “ei”, et tekstis ei esine tahtlikku mõjutustegevust. Lõplik ennustus määrati suurema tõenäosusskoori saanud klassi põhjal.

2.3.2. Generatiivse keelemudeli kasutamine

Teise arvutusliku lähenemisena kasutasin võrdlevaks hindamiseks generatiivset keelemudelit GPT-5.4, mille treeningandmete lõppkuupäev on 05.03.2026 (OpenAI, i.a.). Valikul lähtusin asjaolust, et andmete ajaline piirang on üks teguritest, mis generatiivse mudeli väljundit mõjutab. Generatiivsed keelemudelid erinevad klassifikatsioonimudelitest selle poolest, et nad ei vaja tingimata eelnevalt treenitud klassifikatsiooniülesannet, vaid suudavad teksti analüüsida juhiste ehk *promptide* põhjal (Puri ja Catanzaro, 2019). Seega põhines teine arvutuslik lähenemine generatiivse keelemudeli kasutamisel juhistepõhise (*prompt-based*) binaarse klassifitseerimisena, kus küsisin mudelilt iga teksti kohta otsust, kas tekst jagab või edendab tahtlikult propagandanarratiivi (vt Lisa 2). Et vältida olukorda, kus mudel kirjutab pika selgituse või vastab ebamääraselt, teisendasin mudeli vastuse rangeks “jah/ei” ennustuseks.

Generatiivse mudeli jaoks rakendasin *zero-shot* ja *few-shot* klassifikatsiooni lähenemist. *Zero-shot* lähenemine tähendab, et keelemudel rakendab oma eelneva õpetamise käigus omandatud teadmisi uue ülesande lahendamiseks ilma, et talle antaks selle konkreetse ülesande kohta ühtegi märgistatud näidet (Puri ja Catanzaro, 2019). Seega andsin mudelile ainult ülesande kirjelduse ja analüüsitava teksti. Mudel pidi otsustama, kas tekst sisaldab tahtlikku mõjutustegevust, ilma et talle oleks antud näiteid varasemate otsuste kohta. Hinnangud arvutati kõikidele tekstidele ükshaaval. *Zero-shot* lähenemise tulemused võimaldavad hinnata, kuivõrd hästi suudab mudel uutes olukordades üldistada üksnes juhiste ja eelõppe käigus omandatud teadmiste põhjal.

Few-shot lähenemine tähendab, et keelemudelile antakse enne ülesande lahendamist lisaks juhistele ka väike hulk näiteid, mis illustreerivad soovitud sisendi ja väljundi vahelist seost (Puri ja Catanzaro, 2019). *Few-shot* seadistuses valiti andmestikust automaatselt neli näidet, kus kaks olid sildiga “jah” ja kaks sildiga “ei”. Hinnatav tekst jäi näidete hulgast välja. Näidete lisamise eesmärk oli anda mudelile täiendav kontekst selle kohta, millist tüüpi otsustuskriteeriumit ülesandes oodatakse

2.3.3. Pseudomärgistatud andmetel treenitud juhendatud mudel

Lisaks ekspertide hinnangute põhjal treenitud juhendatud mudelile ja generatiivsetele mudelitele, katsetasin täiendava lähenemisena pseudomärgistamist. Kuna eksperdid hindasid ainult 30 teksti, oli andmestik väike. Andmemahu suurendamiseks kasutasin generatiivset mudelit, et treeningandmeid märgistada ja saadud silte kasutasin juhendatud mudeli treenimiseks. Sarnast lähenemist on treeningandmete mahu suurendamiseks kasutanud ka Tetsmann jt (2026), kelle töö põhjal võib pseudomärgistamist eesti keele kontekstis käsitleda ühe võimaliku lahendusena olukorras, kus käsitsi märgendatud ehk kuldstandardile vastavaid treeningandmeid on piiratud mahus.

Pseudomärgistamiseks kasutasin algsest 220-st propagandaks märgitud tekstist 190 teksti, mis ei kuulunud ekspertide poolt hinnatud 30 teksti hulka. Generatiivne mudel märgistas need 190 teksti binaarselt klassidesse “jah” ja “ei”. Märgistamise tulemusel sai 143 teksti sildi “jah” ja 47 teksti sildi “ei”. See andmestik ei asenda küll ekspertide hinnanguid, kuid aitab hinnata, kas suurem treeningmaht aitab juhendatud mudelil õppida stabiilsemaid seoseid tekstiliste tunnuste ja hinnangute vahel.

Pseudomärgistatud andmetega eksperimendis kasutasin sama baasmudelit, mida ekspertide hinnangutel treenitud juhendatud klassifikatsiooni puhul. Kuna treeningandmeid oli seekord rohkem, vabastasin lisaks klassifikatsioonipeale ka viimase transformerkihi, et mudel saaks ülesandega paremini kohanduda. Selleks vabastasin peale klassifikatsioonipea ka viimase transformerkihi. Mudeli treenimise ja kasutamise tehnilised parameetrid on esitatud lisas 3. Mudeli lõpliku valideerimise viisin läbi täpselt samal ekspertide poolt märgendatud 30 andmepunkti peal, mida kasutasin ka teiste meetodite võrdlemisel.

2.3.4. Mudeli valideerimine ja hindamismõõdikud

Juhendatud klassifikatsioonimudeli töökindluse hindamiseks kasutasin *leave-one-out* (LOOCV) ristvalideerimist, mille puhul jäetakse igas iteratsioonis üks tekst testimiseks ja ülejäänud tekstid treenimiseks (Wong, 2015). Protsessi korratakse nii mitu korda, kui palju on andmestikus tekste, nii et iga tekst toimib ühe korra testandmena. Oma töös kasutasin 29 teksti treenimiseks, ühe teksti jätsin testimiseks ja seda kordasin 30 korda, nii et iga tekst oli üks kord testimiseks. Iga iteratsiooni järel salvestas mudel ennustuse ja lõpus arvutas koondtulemus kõigi 30 testimise põhjal. Selline valideerimisviis võimaldab hinnata mudeli üldistusvõimet

väikese andmestiku korral, kuna iga tekst osaleb nii treening- kui testprotsessis, kuid annab siiski iga teksti kohta eraldi ennustuse.

Generatiivse mudeli puhul ei kasutanud ma eraldi treening- ja testjaotust, sest mudelit ei treenitud töö andmestikul ümber. *Zero-shot* ja *few-shot* seadistustes hindas mudel kõiki tekste ja saadud ennustusi võrdlesin ekspertide häälteenamushinnanguga. *Few-shot* seadistuses jälgisin, et hinnatav tekst ei oleks samal ajal mudelile esitatud näidete hulgas. Pseudomärgistatud andmetega juhendatud mudeli puhul kasutasin generatiivse mudeli loodud lisaandmestikku treenimiseks, kuid lõpliku hindamise tegin samuti 30 eksperthinnanguga teksti põhjal.

Meetodite hindamisel kasutasin õigete ennustuste osakaalu, klassipõhiseid F1-mõõdikuid, kaalutud F1-mõõdikut ja eksimismaatriksit. Õigete ennustuste osakaal ehk *accuracy* näitab, kui suur osa mudeli ennustustest olid õiged (Chae ja Davidson, 2026). Lisaks kasutasin tasakaalustatud õigete ennustuste osakaalu, et hinnata mudelite võimet eristada mõlemat klassi võrdselt olukorras, kus klassijaotus oli ebaühtlane.

Klassipõhised F1-mõõdikud aitavad hinnata seda, kui hästi suudab meetod tuvastada "jah" klassi ja kui hästi "ei" klassi. F1-mõõdik ühendab täpsuse (*precision*) ja saagise (*recall*) ning näitab, kui hästi suudab meetod konkreetset klassi eristada, arvestades korraga nii valepositiivseid kui ka valenegatiivseid otsuseid (Chae ja Davidson, 2026). Makro-F1 aitab hinnata kui tasakaalukalt suudab mudel klasse eristada ja kaalutud F1 näitab meetodi üldist tulemuslikkust kogu andmestiku lõikes, arvestades klasside erinevat suurust (Samas).

Eksimismaatriks (*confusion matrix*) näitab, kuidas mudeli ennustused jagunevad õigete ja valede otsuste vahel ehk millist tüüpi vigu mudel teeb (Puri ja Catanzaro, 2019). Vigade suuna täpsemaks hindamiseks arvutasin ka valepositiivsete ja valenegatiivsete otsuste osakaalu, et hinnata, kas meetodid kaldusid tahtlikku mõjutustegevust pigem üle- või alatõlgendama.

2.4. Tehisaru kasutus töö koostamise protsessis ja refleksioon

Kui jätta välja masinõppe kasutamine töö uurimismeetodina, oli tehisaru peamine roll minu magistritöös keeruliste mõistete selgitamine. Kuna õppekava raames on masinõpet käsitletud pigem põgusalt ja vaid ühes õppeaines, kasutasin suurte keelemudelite abi, et mõista masinõppe tehnilist poolt. Selleks lasin ma keelemudelil selgitada masinõppe erinevaid aspekte nii, nagu ta seletaks neid gümnaasiumiõpilasele. Mõne keerulisema teema puhul palusin

selgituse kohandada isegi lasteaialapse tasemele. Selliste juhiste põhjal muutis mudel keerulised tehnilised mõisted eluliste näidete kaudu arusaadavamaks ja tõlkis need n-ö inimkeelde.

Töö käigus sain tehnilisest poolest kindlasti paremini aru, kuid ma pole siiski masinõppe ekspert. Selleks, et vältida hallutsineerimist ja sisutut tehisplära, lasin ma mõnes olukorras kahel erineval keelemudelil sama teemat seletada, et tulemusi risti kontrollida. Otsest olukorda, kus oleks pidanud mudelid vaidlema panema, ette ei tulnud, kuid võrdlemine aitas näha, kas selgitused kattuvad ja kas mõni vastus tundub liiga enesekindel. Samal põhjusel otsisin abi ka muudest allikatest ega tuginenud ainult tehisarule.

Paljudel masinõppega seotud mõistetel puuduvad seni ametlikud eestikeelsed mõisteid või kasutatakse neid erinevalt. Kuigi suurema osa eestikeelsetest vastetest leidsin ma Tartu Ülikooli arvutiteaduse instituudi varasematest töödest, tuli ka siinkohal keelemudel abiks sobivate vastete leidmisel.

Teoreetilise tausta kirjutamise puhul oli tehisaru kõige kasulikum otsingumärksõnade genereerimisel. Kui töö esimeses etapis olid mul põhilised allikad olemas, siis hiljem aitasid mudeli genereeritud terminid otsinguid laiendada. Mudel suutis minu töö taustainfo põhjal pakkuda sisukaid ja konkreetseid märksõnu, mida andmebaasides kasutada. Samuti katsetasin keelemudelit töö struktuuri loogilisuse kontrollimiseks. See andis hea kõrvalpilgu ja aitas hinnata, kas peatükid ja teemad jooksevad lugeja jaoks mõistlikus ja loogilises järjekorras.

Olukorras, kus ma ei suutnud oma mõtteid selgelt ja akadeemiliselt kirja panna, küsisin keelemudelilt abi sõnastuse parandamiseks. See ei tähenda, et oleksin lasknud tehisarul teksti enda eest valmis kirjutada, vaid kasutasin seda keeleliseks toimetamiseks. See oli kasulik ainult teatud olukordades. Mida ebamäärasemalt olin ise oma mõtte kirja pannud, seda halvem ja lohisevam oli ka mudeli väljund. Mõnel korral andis isegi halb väljund mulle mõne idee otsa kätte aga selleks, et tekst ei muutuks tehisplääraks, pidin kõigepealt ise täpselt aru saama, mida öelda tahan ja oma mõtte võimalikult sisukalt formuleerima.

Tehisaru kasutamine andis üsna kiiresti tunda, et see ei vabasta mõtlemisest. Vastupidi, mida paremini ma ise oma mõttest aru sain, seda kasulikum oli ka mudeli abi. Laiemas plaanis aitas keelemudel tähelepanu pöörata silumist vajavatele lausetele või kohtadele, kus struktuur vajas paremat läbimõtlemist.

3. TULEMUSED

Selles peatükis esitan analüüsi tulemused ekspertide häälteenamushinnangu suhtes. Kõigepealt hindan kontrollgrupi kooskõla ekspertide koondhinnanguga, seejärel kirjeldan juhendatud tekstiklassifikatsiooni mudeli ja generatiivse keelemudeli tulemusi. Lisaks kirjeldan pseudomärgistatud andmetega juhendatud mudeli tulemusi. Peatüki lõpus võrdlen kõigi kasutatud lähenemiste tulemuslikkust, et hinnata, milline neist kattus ekspertide hinnanguga kõige paremini ja millised erinevused klasside lõikes ilmnesid.

3.1. Kontrollgrupi kooskõla ekspertide häälteenamushinnanguga

Ekspertide ja kontrollgrupi kooskõla hindamiseks võrdlesin kontrollgrupi häälteenamushinnangut ekspertide häälteenamushinnanguga, mis aitab näha, kuivõrd vajab tahtlikkuse hindamine spetsiifilisi erialaseid teadmisi.

Tabelis 2 esitatud eksimismaatriks näitab, et kontrollgrupi kooskõla ekspertidega oli märksa tugevam nende tekstide puhul, mida eksperdid hindasid tahtlikku mõjutustegevust sisaldavaks, kui nende puhul, mida eksperdid hindasid selliseks, kus tahtlikku mõjutustegevust ei esine. Ekspertide 21-st "jah" klassi kuulunud tekstist hindas kontrollgrupp 20 teksti samuti "jah" klassi. Seevastu ekspertide üheksast "ei" klassi kuulunud tekstist kattus kontrollgrupi hinnang ekspertidega vaid kahel juhul ja ülejäänud seitsmel korral hindas kontrollgrupp teksti ekspertidest erinevalt "jah" klassi.

Tabel 2. Kontrollgrupi hinnangute eksimismaatriks ekspertide häälteenamushinnangu suhtes.

	Kontrollgrupp: Ei	**Kontrollgrupp: Jah**
Ekspert: Ei	2	7
Ekspert: Jah	1	20

Kontrollgrupp tabas hästi neid juhtumeid, mida eksperdid pidasid tahtlikuks mõjutustegevuseks, kuid oli oluliselt ebakindlam nende tekstide puhul, mida eksperdid hindasid mitte-tahtlikuks. Teisisõnu kaldus kontrollgrupp ekspertidega võrreldes sagedamini omistama tekstidele tahtliku mõjutustegevuse tunnuseid ka siis, kui eksperdid seda ei teinud.

Kontrollgrupi õigete ennustuste osakaal oli 0.73 ehk 73% ja kaalutud F1-mõõdik 0.68, mis viitavad heale kooskõlale ekspertide häälteenamushinnanguga (Tabel 3). Klassipõhine vaade

näitab siiski selget erinevust kahe klassi vahel. “Jah” klassi F1-mõõdik oli 0.83, samas kui “ei” klassi F1-mõõdik oli vaid 0.33. See näitab, et kontrollgrupi vastused kattusid paremini nende tekstide puhul, mida eksperdid pidasid tahtlikuks mõjutustegevuseks.

Tabel 3. Kontrollgrupi hinnangute klassifikatsioonimõõdikud ekspertide häälteenamushinnangu suhtes.

Klass	Täpsus	Saagis	F1-skoor	Juhtumite arv
Ei	0.67	0.22	0.33	9
Jah	0.74	0.95	0.83	21
Makrokeskmine	0.7	0.59	0.58	30
Kaalutud keskmine	0.72	0.73	0.68	30
Õigete ennustuste osakaal	0.73			

3.2. Juhendatud mudeli kooskõla ekspertide häälteenamushinnanguga

Juhendatud tekstiklassifikatsiooni mudeli tulemusi hindasin samuti ekspertide häälteenamushinnangu suhtes.

Tabelis 4 esitatud eksimismaatriks näitab, et juhendatud mudel klassifitseeris õigesti kuus eksperthinnangu järgi “ei” klassi kuulunud teksti üheksast ja 13 “jah” klassi kuulunud teksti 21-st. Seega kolm eksperthinnangu järgi “ei” teksti määrati ekslikult “jah” klassi ja kaheksa eksperthinnangu järgi “jah” teksti määrati ekslikult “ei” klassi.

Tabel 4. Juhendatud mudeli eksimismaatriks ekspertide häälteenamushinnangu suhtes.

	Mudel: Ei	Mudel: Jah
Ekspert: Ei	6	3
Ekspert: Jah	8	13

Mudeli õigete ennustuste osakaal oli 0.63 ehk 63% ja kaalutud F1-mõõdik 0.65 (Tabel 5). Võrreldes klasside kaupa, oli Juhendatud mudeli “jah” klassi F1-mõõdik oli 0.70 ja “ei” klassi F1-mõõdik 0.52.

Tabel 5. Juhendatud mudeli klassifikatsioonimõõdikud ekspertide häälteenamushinnangu suhtes.

Klass	Täpsus	Saagis	F1-skoor	Juhtumite arv
Ei	0.43	0.67	0.52	9
Jah	0.81	0.62	0.7	21
Makrokeskmine	0.62	0.64	0.61	30
Kaalutud keskmine	0.7	0.63	0.65	30
Õigete ennustuste osakaal	0.63			

3.3. Generatiivse mudeli *zero-shot* tulemused

Zero-shot seadistuses sai generatiivne mudel ainult ülesande kirjelduse ja hinnatava teksti, ilma näideteta, mis aitab hinnata, milline on mudeli otsustusvõime ainult juhise põhjal. Ekspertide koondhinnangutega võrreldes klassifitseeris generatiivne mudel õigesti 20 eksperthinnangu järgi "jah" klassi kuulunud teksti 21-st (Tabel 6). Samas klassifitseeris mudel õigesti ainult ühe eksperthinnangu järgi "ei" klassi kuulunud teksti üheksast.

Tabel 6. Generatiivse mudeli *zero-shot* eksimismaatriks ekspertide häälteenamushinnangu suhtes.

	Mudel: Ei	Mudel: Jah
Ekspert: Ei	1	8
Ekspert: Jah	1	20

Tabel 7. Generatiivse mudeli *zero-shot* klassifikatsioonimõõdikud ekspertide häälteenamushinnangu suhtes.

Klass	Täpsus	Saagis	F1-skoor	Juhtumite arv
Ei	0.5	0.11	0.18	9
Jah	0.71	0.95	0.82	21
Makrokeskmine	0.61	0.53	0.5	30
Kaalutud keskmine	0.65	0.7	0.63	30
Õigete ennustuste osakaal	0.7			

Mudeli õigete ennustuste osakaal oli 0.70 ehk 70% ja kaalutud F1-mõõdik 0.63 (Tabel 7). Vaadates tulemusi klassipõhiselt, oli klassi "jah" F1-mõõdik 0.82 ja klassi "ei" kuuluvate tekstide F1-mõõdik 0.18.

3.4. Generatiivse mudeli *few-shot* tulemused

Few-shot seadistuses lisasin mudeli juhisesse automaatselt neli näidet samast andmestikust, ehk kaks "jah" ja kaks "ei" sildiga teksti. Hinnatav tekst jäeti näidete hulgast välja. *Few-shot* seadistuses klassifitseeris mudel õigesti 19 eksperthinnangu järgi "jah" klassi kuulunud teksti 21-st ja kaks eksperthinnangu järgi "ei" klassi kuulunud teksti üheksast (Tabel 8).

Tabel 8. Generatiivse mudeli *few-shot* eksimismaatriks ekspertide häälteenamushinnangu suhtes.

	Mudel: Ei	**Mudel: Jah**
Ekspert: Ei	2	7
Ekspert: Jah	2	19

Mudeli õigete ennustuste osakaal oli 0.70 ehk 70% ja kaalutud F1-mõõdik oli 0.66 (Tabel 9), mis oli kõrgem kui *zero-shot* seadistuse 0.63. Klassipõhised tulemused näitavad, et näidete lisamine muutis mudeli hinnanguid mõnevõrra tasakaalukamaks. "Ei" klassi F1-mõõdik suurenes *zero-shot* seadistuses 0.18-lt 0.31-ni. "Jah" klassi F1-mõõdik langes 0.82-lt 0.81-ni. See osutab, et *few-shot* seadistuses kasutatud näited aitasid mudelil paremini eristada mõnda "ei" klassi kuuluvat teksti, kuid ei kõrvaldanud mudeli üldist kalduvust anda "jah" hinnanguid.

Tabel 9. Generatiivse mudeli *few-shot* klassifikatsioonimõõdikud ekspertide häälteenamushinnangu suhtes.

Klass	**Täpsus**	**Saagis**	**F1-skoor**	**Juhtumite arv**
Ei	0.5	0.22	0.31	9
Jah	0.73	0.9	0.81	21
Makrokeskmine	0.62	0.56	0.56	30
Kaalutud keskmine	0.66	0.7	0.66	30
Õigete ennustuste osakaal	0.7			

3.5. Pseudomärgistatud andmetega juhendatud mudeli tulemused

Lisaks juhendatud mudeli ja generatiivse mudeli katsetele viisin läbi pseudomärgistamise eksperimendi, et hinnata, kas treeningandmestiku mahu suurendamine generatiivse mudeli abil loodud siltidega parandab juhendatud mudeli tulemusi.

Eksimismaatriks tabelis 10 näitab, et pseudomärgistatud andmetega juhendatud mudel klassifitseeris õigesti 18 eksperthinnangu järgi "jah" klassi kuulunud teksti 21-st ja viis eksperthinnangu järgi "ei" klassi kuulunud teksti üheksast. Mudelil jäi märkamata kolm eksperthinnangu järgi "jah" klassi kuulunud teksti ja neli eksperthinnangu järgi "ei" kuulunud teksti määras mudel valesti "jah" klassi.

Tabel 10. Pseudomärgistatud andmetega juhendatud mudeli eksimismaatriks ekspertide häälteenamushinnangu suhtes.

	Mudel: Ei	**Mudel: Jah**
Ekspert: Ei	5	4
Ekspert: Jah	3	18

Mudeli õigete ennustuste osakaal oli 0.77 ehk 77% ja kaalutud F1-mõõdik 0.76 (Tabel 11). Klassipõhiselt oli "jah" klassi F1-mõõdik 0.84 ja "ei" klassi F1-mõõdik 0.59.

11. Pseudomärgistatud andmetega juhendatud mudeli klassifikatsioonimõõdikud ekspertide häälteenamushinnangu suhtes.

Klass	**Täpsus**	**Saagis**	**F1-skoor**	**Juhtumite arv**
Ei	0.62	0.56	0.59	9
Jah	0.82	0.86	0.84	21
Makrokeskmine	0.72	0.71	0.71	30
Kaalutud keskmine	0.76	0.77	0.76	30
Õigete ennustuste osakaal	0.77			

3.6. Meetodite võrdlus ja veamustrid

Meetodite võrdluse eesmärk on hinnata, kuidas erinevad lähenemised taastoodavad ekspertide häälteenamushinnangut ja milliseid järeldusi saab teha nende sobivuse kohta tahtliku mõjutustegevuse tuvastamisel.

Tabelis 12 esitatud võrdlus näitab, et meetodite paremusjärjestus sõltub suurel määral sellest, millist mõõdikut vaadata. Kõige kõrgema õigete ennustuste osakaalu saavutas pseudomärgistatud andmetega juhendatud mudel tulemusega 77%. Sellele järgnes kontrollgrupp tulemusega 73%, generatiivse mudeli *zero-shot* ja *few-shot* seadistused tulemusega 70% ja ekspertandmetel treenitud juhendatud mudel tulemusega 63%. Kui vaadata ainult üldist kattuvust eksperthinnangutega, võiks järeldada, et pseudomärgistatud andmetega juhendatud mudel oli kõigist võrreldud meetoditest tulemuslikum.

Tulemuste tõlgendamisel on oluline arvestada, et ekspertide hinnangute jaotus oli "jah" klassi poole kaldu. Olukorras, kus mudel määraks kõik tekstid "jah" klassi, saaks samuti õigete ennustuste osakaaluks 0.70. Seega ei saa generatiivsete mudelite 0.70 tulemust käsitleda iseenesest tugeva tulemusena, sest see võib näidata tegelikku eristusvõimet või lihtsalt kalduvust valida sagedamini esinevat klassi.

Tabel 12. Meetodite klassifikatsioonitulemuste võrdlus ekspertide häälteenamushinnangu suhtes.

Meetod	Kaalutud F1	Makro-F1	"Jah" F1	"Ei" F1
Kontrollgrupp	0.68	0.58	0.83	0.33
Juhendatud mudel	0.65	0.61	0.7	0.52
GPT *zero-shot*	0.63	0.5	0.82	0.18
GPT *few-shot*	0.66	0.56	0.81	0.31
Pseudomärgistatud mudel	0.76	0.71	0.84	0.59

Klassipõhised F1-mõõdikud aitavad näha, kui hästi suudab iga meetod eristada kumbagi klassi eraldi. Kontrollgrupi ja generatiivsete mudelite tulemused olid tugevamad "jah" klassis, kuid nõrgemad "ei" klassis (Tabel 12). "Jah" klassi puhul olid pseudomärgistatud andmetega juhendatud mudeli tulemus 0.84, kontrollgrupil 0.83, GPT *zero-shot* seadistusel 0.82 ja GPT *few-shot* seadistusel 0.81. Seega suutsid kõik meetodid taastoota ekspertide hinnanguid enim nende tekstide puhul, mida eksperdid pidasid tahtlikku mõjutustegevust sisaldavaks.

GPT *zero-shot* seadistuse "ei" klassi F1 oli 0.18, GPT *few-shot* seadistusel 0.31 ja kontrollgrupil 0.33. Kuigi need meetodid olid tugevad "jah" klassi taastootmisel, olid nad nõrgad nende tekstide tuvastamisel, mida eksperdid tahtlikuks mõjutustegevuseks ei hinnanud. Ekspertandmetel treenitud juhendatud mudeli "ei" klassi F1 oli 0.52 ja pseudomärgistatud

andmetega juhendatud mudeli “ei” klassi F1 0.59. Juhendatud mudelid suutsid seega paremini vältida olukorda, kus tekstile omistatakse tahtlik mõjutustegevus liiga kergekäeliselt ja suutsid piirata üle tõlgendamist.

Pseudomärgistatud andmetega juhendatud mudeli makro F1-mõõdiku väärtus oli 0.71, ainult ekspertandmetel treenitud juhendatud mudelil 0.61, kontrollgrupil 0.58, GPT *few-shot* seadistusel 0.56 ja GPT *zero-shot* seadistusel 0.50. Kui hinnata meetodit mitte ainult üldise täpsuse, vaid klasside lõikes tasakaalukuse järgi, oli pseudomärgistatud andmetega juhendatud mudel tugevaim. Ainult ekspertandmetel treenitud juhendatud mudel jäi küll üldises täpsuses nõrgemaks, kuid oli siiski tasakaalukam kui kontrollgrupp ja generatiivsed mudelid.

Kaalutud F1 järgi oli tugevaim pseudomärgistatud andmetega juhendatud mudel tulemusega 0.76. Kontrollgrupi tulemus oli 0.68, GPT *few-shot* seadistusel 0.66, ainult ekspertandmetel treenitud juhendatud mudelil 0.65 ja GPT *zero-shot* seadistusel 0.63. Pseudomärgistatud mudeli kõrge kaalutud F1 näitab, et selle tugevus ei olnud juhuslik ega tulenenud ainult ühest tugevast klassist, vaid selle üldine toimimine oli kogu andmestiku lõikes parem kui teistel lähenemistel.

Tabelis 13 esitatud tasakaalustatud õigete ennustuste osakaal näitab, et kui mõlemale klassile anda võrdne kaal, oli parim pseudomärgistatud andmetega juhendatud mudel tulemusega 0.71. Ainult ekspertandmetel treenitud juhendatud mudeli tulemus oli 0.64, kontrollgrupil 0.59, GPT *few-shot* seadistusel 0.56 ja GPT *zero-shot* seadistusel 0.53. Kui mõlemale klassile anda võrdne kaal, on pseudomärgistatud andmetega juhendatud mudel kõige tugevam ehk teistest mudelitest ühtlasem, suutes paremini klasse eristada.

Tabel 13. Meetodite tasakaalustatud tulemus ja vigade suund ekspertide häälteenamushinnangu suhtes.

Meetod	Õigete ennustuste osakaal	Tasakaalustatud õigete ennustuste osakaal	Valepositiivsete osakaal
Kontrollgrupp	0.73	0.59	0.78
Juhendatud mudel	0.63	0.64	0.33
GPT *zero-shot*	0.7	0.53	0.89
GPT *few-shot*	0.7	0.56	0.78
Pseudomärgistatud mudel	0.77	0.71	0.44

Meetodite veamustrite tüüpe aitab selgitada valepositiivsete ja valenegatiivsete otsuste osakaal. Valepositiivne otsus tähendab siin, et meetod hindas ekspertide “ei” klassi kuuluva teksti ekslikult “jah” klassi ja valenegatiivne otsus, et meetod hindas ekspertide “jah” klassi kuuluva teksti ekslikult “ei” klassi.

Kontrollgrupi ja GPT *few-shot* seadistuse puhul oli valepositiivsete otsuste osakaal 0.78 ehk 78% ja GPT *zero-shot* seadistusel 0.89 ehk 89% (Tabel 13). See tähendab, et need mudelid kaldusid väga tugevalt tahtlikku mõjutustegevust üle tuvastama. Ainult ekspertandmetel treenitud juhendatud mudeli valepositiivsete otsuste osakaal oli 0.33 ehk 33% ja pseudomärgistatud andmetega juhendatud mudelil 0.44 ehk 44%. Juhendatud lähenemised olid ettevaatlikumad ega omistanud tahtlikku mõjutustegevust nii sageli neile tekstidele, mida eksperdid selliseks ei hinnanud.

Kontrollgrupil ja GPT *zero-shot* seadistusel oli valenegatiivsete osakaal 5% ja GPT *few-shot* seadistusel 10%. Meetodid olid seega “jah” klassi suhtes tundlikud, kuid see tundlikkus saavutati sageli üle tuvastamise hinnaga. Ainult ekspertandmetel treenitud juhendatud mudeli valenegatiivsete otsuste osakaal oli 0.38, mis tähendab, et sellel jäi märkamata 38% ekspertide järgi positiivsetest juhtumitest. Pseudomärgistatud andmetega juhendatud mudeli valenegatiivsete otsuste osakaal oli 14%, mis näitab, et mudel suutis säilitada üsna hea tundlikkuse “jah” klassi suhtes, ilma et ta oleks samal määral kaldunud “ei” klassi üle kirjutama kui kontrollgrupp või generatiivsed mudelid.

Meetodite võrdlus näitab, et erinevad mõõdikud on koos palju sisukamad kui eraldiseisvalt. Kontrollgrupp ja generatiivsed mudelid saavutasid kõrge õigete ennustuste osakaalu, kuid see tulemus põhines suuresti kalduvusel tahtlikku mõjutustegevust üle tuvastada.

4. ARUTELU

Selles peatükis seon tulemused töö teoreetilise raamistikuga ja vastan uurimisküsimustele. Lisaks arutlen töö piirangute, praktiliste järelduste ja edasiste uurimisvõimaluste üle. Töö eesmärk oli hinnata, kas masinõppepõhist tahtlikkuse indikaatorit saab kasutada eestikeelsetes tekstides otsustustoe elemendina tahtliku mõjutustegevuse eristamisel tavapärasest kommunikatsioonist. Seetõttu ei käsitle ma arutelus mudelite tulemusi tehnilise paremusjärjestusena, vaid hindan nende abil, kuivõrd on tahtlikkuse tunnus masinõppe abil modelleeritav ja kas see võiks toimida otsustustoe ühe elemendina.

4.1. Tahtlikkuse indikaatori masinõppepõhine modelleeritavus

Tulemused toetavad töö peamist teoreetilist eeldust, et tahtlikkus ei ole otseselt vaadeldav tunnus, vaid hinnanguline konstruktsioon, mis tuletatakse tekstiliste ja kontekstuaalsete tunnuste põhjal (Khosravi ja Barekat, 2021). Kui tahtlikkus oleks tekstis selgelt ja üheselt nähtav võiks eeldada, et erinevad mudelid ja ekspert- ning kontrollgrupp oleksid saavutanud sarnase tulemuse. Meetodite võrdlus näitas pigem mudelite tugevustes suurt varieeruvust, mis viitab asjaolule, et tahtlikkuse hindamine sõltub sellest, kuidas tekstis esinevaid tunnuseid tõlgendatakse (Ottaviani jt, 2021).

Valepositiivsete määra põhjal kaldusid mitteekspertidest koosnev kontrollgrupp ja generatiivne mudel tahtlikku mõjutustegevust tugevalt üle tuvastama. Kontrollgrupi kõrge valepositiivsete määr ja väga madal F1-skoor “ei” klassis näitavad, et tavalugejal oli keeruline eristada retoorilist sisu tahtlikust mõjutustegevusest. Omakorda toetab tavalugejate tulemus ekspertide häälteenamushinnangu kasutamist võrdlusalusena. Generatiivne mudel sarnanes oma veamustri poolest tavalugejale. Nagu toob välja Christiano (2022), ei tähenda retoorika kasutus alati sihipärast mõjutustegevust või desinformatsiooni, vaid see võib olla osa ka täiesti tavapärasest, ehkki polariseerunud poliitilisest kommunikatsioonist. Tahtlikkuse hindamise keerukus ei tulene üksnes mudeli tehnilistes piirangutest, vaid ka asjaolust, et retooriliselt tugev tekst võib näida tahtlikult mõjutavana.

Ainult eksperthinnangutel treenitud juhendatud mudel oli üldise täpsuse poolest kontrollgrupist ja generatiivsest mudelist nõrgem, kuid selle veamuster oli teistsugune. Mudel tundis “jah” klassi halvemini ära, kuid oli parem “ei” klassi eristamise poolest. Reaalses olukorras ei ole valepositiivne otsus lihtsalt tehniline viga, vaid praktiline koormus. Iga ekslikult märgistatud

tekst nõuab inimese lisatööd, vähendades mudeli usaldusväärsust ja nullides selle peamise eesmärgi ehk kognitiivse koormuse vähendamise. Praktilise väärtuse hindamisel on seetõttu "ei" klassi eristamine sama oluline kui "jah" klassi tabamine.

Praktilise väärtuse seisukohalt oli kõige olulisem pseudomärgistatud andmetega juhendatud mudeli tulemus. See mitte ainult ei andnud kõrgeima õigete ennustuste osakaalu, vaid suutis ka kahte klassi kõige tasakaalukamalt eristada. Kuna ekspertide hinnangu jaotus oli "jah" klassi poole kaldu, saavutaks mudel 70% täpsuse ka siis kui see määraks kõik tekstid "jah" klassi. Tasakaalukam eristus tähendab siin seega head "ei" klassi eristust. Pseudomärgistatud andmetega mudeli "ei" klassi F1-skoor oli 0.59, samal ajal kui GPT zero-shot mudelil oli see 0.18, GPT few-shot seadistusel 0.31 ja kontrollgrupil 0.33. Kontrollgrupp ja generatiivsed mudelid kaldusid seega tahtlikku mõjutustegevust pigem üle tuvastama, kuid pseudomärgistatud mudel suutis paremini ära tunda ka neid tekste, mida eksperdid tahtlikuks mõjutustegevuseks ei hinnanud.

Mudelite tulemused ei näita niivõrd seda, milline mudel on lõplikult parim, vaid seda, et masinõpe suudab tahtlikkusega seotud tekstilisi mustreid osaliselt eristada. Samas näitavad mudelite veamustrid, eriti valepositiivsete rohkus, et tahtlikkuse tunnusest üksi ei piisa täielikult iseseisva ja eksimatu automaattuvastussüsteemi loomiseks.

4.2. Treeningandmete mahu mõju mudeli tulemuslikkusele

Arvutuslikus sotsiaalteaduses tunnustatud reeglipära kohaselt paraneb mudelite võimekus ja stabiilsus otseselt koos kvaliteetsete märgistatud andmete mahu kasvuga (Lazer jt, 2020; Hox, 2017). Kuigi generatiivne mudel kaldus tahtlikkust üle tuvastama, parandas generatiivse mudeliga pseudomärgistatud andmete lisamine juhendatud mudeli täpsust. Suurem andmemaht parandas tulemuslikkust nii üldiste õigete ennustuste osakaalu kui ka klasside eristamise tasakaalu poolest.

Eksperthinnanguga tekstivalimis oli ainult 9 teksti, mis hinnati klassi "ei". Seda on mudeli treenimisel klassi mõistmiseks liiga vähe. Suurem andmemaht (n = 190) andis mudelile piisavalt näiteid, et õppida kahte klassi eristavaid tekstilisi seoseid. Ainult 30 eksperthinnanguga tekstil treenitud juhendatud mudeli õigete ennustuste osakaal oli 0.63 ja pseudomärgistatud andmetega mudeli tulemus oli 0.77. Tulemuse paranemist näitavad ka F1-mõõdikud, mille puhul kaalutud F1 tõusis 0.65-lt 0.76-ni ja makro-F1 0.61-lt 0.71-ni. See ei tähenda, et mudel oleks hakanud autori kavatsust paremini mõistma, vaid tekstides esines

piisavalt korduvaid tahtlikkuse hinnanguga seotud tunnuseid, mida oli võimalik õppimiseks kasutada.

Pseudomärgistamise tulemusse tuleb siiski kriitiliselt suhtuda. Kuna sildid genereeriti keelemudeli GPT-5.4 abil, kaasnes risk, et generatiivse mudeli kalduvus tahtlikkust üle tuvastada kandus osaliselt üle ka pseudomärgistatud andmestikku. Nagu on märkinud Hox (2017), ei ole masinõpe kunagi täielikult neutraalne, vaid peegeldab alati otseselt treeningandmete koostamise loogikat ja sildistajate subjektiivsust. Samas võib tekkida oht, et kui väiksemaid mudeleid treenida generatiivse mudeli siltide peal, kopeerib mudel inimliku taju asemel hoopis suure keelemudeli spetsiifilisi keelelisi eelarvamusi.

Pseudomärgistatud lähenemise edu ei tõesta genereeritud siltide samaväärsust eksperthinnangutega, vaid toetab järeldust, et treeningandmete mahu suurendamine aitab mudelil paremini üldistada. Tahtlikkuse hinnanguga seotud mustrid ei ole seega mudeli jaoks täiesti juhuslikud. Kui suurem andmemaht aitab mudelil paremini eksperthinnangut taastoota, toetab see omakorda järeldust, et tahtlikkuse indikaatori kasutamine otsustustoes on võimalik, kuid vajab kvaliteetsemat ja suuremat andmestikku. Lisaks võib pseudomärgistamine olla kasulik olukorras, kus andmeid on vähe ja inimhindajate kaasamine on ajaliselt piiratud ressurss (Tetsmann jt, 2026).

4.3. Tahtlikkuse indikaatori sobivus otsustustoe elemendiks

Töö peamiseks eesmärgiks oli hinnata, kas tahtlikkuse indikaator võiks olla kasutatav otsustustoe ühe elemendina. Sellest tulenevalt tuleb tulemusi tõlgendada praktilise kasutatavuse, mitte lõpliku automaattuvastuse vaatenurgast. Töö teooriaosas püstitasin eelduse, et arvutuslikud meetodid peaksid toimima eelkõige otsustustoe süsteemina, mis aitavad hallata tänapäevase infosfääri infoüleküllust ja kognitiivseid piiranguid (French jt, 2024; Starbird jt, 2019). Tulemused näitavad, et tahtlikkuse indikaator sobib kasutamiseks mõjutustegevuse tuvastamisel esmase filtrina, kuid ei ole veel piisav täielikult automaatseks tuvastamiseks.

Esiteks kinnitavad seda juhendatud masinõppemudeli ja pseudomärgistatud mudeli makro-F1 skoorid (vahemikus 0.51-0.71), mis viitavad mudelite ebakindlusele ja vigade esinemisele mõlema klassi suunas. Generatiivse mudeli puhul avaldus selge kallutatus, kus *zero-shot* seadistuses ulatus valepositiivsete määr 89%-ni. Madalad F1-skoorid viitavad üldisele madalale täpsusele, kuid suur valepositiivsete määr viitab riskile, et täielikult automaatne

süsteem märgistaks tahtlikuks mõjutustegevuseks ebaproportsionaalselt suure osa tavapärasest kommunikatsioonist.

Mudelite üks reaalne väärtus seisneb võimes toimida n-ö infosõela esimese astmena, aidates suunata inimese tähelepanu tekstidele, mis vajavad edasist hindamist. Praktikas tähendab see, et mudel ei märgista teksti lõplikult tahtlikuks mõjutustegevuseks, vaid annab sellele esmase märke riski kohta. Tänapäevase infosfääri suure andmemahu ja informatsiooni kiire leviku tõttu ei ole võimalik igat levivat uudist või postitust käsitsi kontrollida (Lazer jt, 2020). Masinõppemudel aitaks selles kontekstis vähendada kognitiivset koormust ja filtreerida välja need tekstid, mille puhul esineb potentsiaalselt tahtlikku mõjutustegevust. Teisisõnu teeb masinõpe ära mehaanilise eeltöö, mis jääb inimesele füüsiliselt kättesaamatuks (Jordan ja Mitchell, 2015), jättes lõpliku otsuse eksperdile, kes suudab lisaks hinnata laiemat ühiskondlikku ja kultuurilist konteksti. Vajadust sellise otsustustoe järele ei näitlikusta üksnes tänapäevase infosfääri hoomamatu andmemaht, vaid ka kontrollgrupi tugev kalduvus hinnata tekste tahtlikuks mõjutustegevuseks juhtudel, kus eksperdid seda selliseks ei hinnanud.

Teiseks kinnitavad tulemused teooriaosas esitatud seisukohta, et tahtlikkus on tekstis keeruliselt hinnatav ja tõlgendustest sõltuv nähtus (Khosravi ja Barekat, 2021; Kärki, 2023), mistõttu ei tohiks seda praktikas käsitleda iseseisva näitajana mõjutustegevuse esinemise kohta. Generatiivne mudel ja kontrollgrupp kaldusid tahtlikkust üle tuvastama, mis võib viidata asjaolule, et samad keelelised tunnused või võtted, mida kasutatakse strateegilistes mõjutuskampaaniates, on laialdaselt kasutusel ka tavapärases kommunikatsioonis (Christiano, 2022). See ei tähenda, et kontrollgrupp oleks tekste vähem tähelepanelikult hinnanud. Pigem võib see peegeldada asjaolu, et mitteeksperdid kaldusid mõjutust samastama tahtlikkusega, samal ajal kui eksperdid lähenesid suurema ettevaatlikkusega. Kui tugineda ainult tekstipõhisele indikaatorile, tõlgendataks iga populistlikku või emotsionaalset arvamust tahtliku mõjutustegevusena.

Praktilise tööriistana on tahtlikkuse indikaator kasulik ühe komponendina laiemas ja mitmetasandilises hindamisraamistikus. Terviklikuma pildi saamiseks tuleb tekstile lisaks hinnata näiteks käitumis- ja levikumustreid ning narratiivide korduvust (French jt, 2024; Saeidnia jt, 2025). Lisaks on Starbird (2019) ja Lazer (2020) rõhutanud, et eduka väärinfo vastase võitluse aluseks on üleminek lihtsalt sisuanalüüsilt käitumuslike mustrite tuvastamisele.

Selleks pakub tugeva raami teooriaosas viidatud François (2019) loodud ABC-raamistik, mille põhjal tuleb mõjutustegevuse tuvastamiseks siduda omavahel osapooled (*Actors*), käitumine (*Behaviour*) ja sisu (*Content*). Minu töös analüüsisid mudelid vaid sisu, püüdes tekstiliste mustrite põhjal tuvastada tahtlikkust. Mudelite suur valepositiivsete määr tähendab praktikas riski, et tööriist nõuab endiselt mahukat inimkontrolli. Selle riski vähendamiseks tuleks tahtlikkuse tunnus kokku viia muude tunnustega. Tahtlikkuse indikaator on seega väärtuslik algsignaal või otsustustoe element, kuid see ei lahenda mõjutustegevuse tuvastamist üksinda.

4.4. Piirangud ja edasised uurimisvõimalused

Tulemuste tõlgendamisel tuleb arvestada mitme metodoloogilise piiranguga, millest osa tuleneb sellest, et tahtlikkus ei ole tekstis otseselt vaadeldav nähtus. Kogu analüüs sõltus sellest, kuidas ma töös tahtlikkuse mõõdetavaks tunnuseks muutsin. Minu töös oli selleks standardiks ekspertide häälteenamushinnang, mis andis võimaluse tahtlikkust kvantitatiivselt uurida, kuid ei tähenda siiski, et ekspertide hinnang väljendaks autori tegelikku kavatsust.

Ekspertide häälteenamushinnang ei näita seega autori reaalset mõtet, vaid esindab ekspertide kokkuleppelist tõlgendust tekstis olevate märkide põhjal. Mudelid õppisid matkima ekspertide otsustusloogikat, mitte tuvastama objektiivset tõde. Kuna kaasatud ekspertide ja kontrollgrupi liikmete arv oli piiratud, tuleks edasistes uuringutes kaasata suurem hulk hindajaid ja uurida ka nende omavahelist kooskõla. See aitaks mõista, kui ühtemoodi inimesed üldse tahtlikkuse mõistet mõistavad ja eristada olukordi, kus eksperdid on tugevalt ühel meelel nendest, kus koondhinnang kujuneb napilt. Ebakindlad hinnangud ei pruugi mudeli treenimisel olla sama kvaliteetsed kui kindlad konsensusega tekstid.

Oluliseks piiranguks oli ka valimi suurus. Töös toetusin 30 tekstist koosnevale andmestikule, mis oli töö eesmärki arvestades piisav mudelite esmaste mustrite ja veatüüpide kaardistamiseks. Andmete väike maht mõjutas aga mudelite stabiilsust. Isegi üksikute tekstide ekslik klassifitseerimine võis mõjutada lõplikke tulemusi. Kuigi andmemahu suurendamine pseudomärgistamise abil aitas mudelit tasakaalustada, oleks generatiivse mudeli kallutatavuse vähendamiseks ja täpsemaks hindamiseks vaja edasistes uuringutes kvaliteetsemat ning suuremat andmestikku.

Mudelite stabiilsust võis omakorda mõjutada klasside ebaühtlane jaotus. Nagu ka varasemalt mainitud, oli ekspertide hinnangu jaotus “jah” klassi poole kaldu ja mudel oleks saavutanud 70% täpsuse ka siis kui see oleks kõik tekstid määranud “jah” klassi. Samuti võis tahtlikkuse

määratlemine binaarsetesse klassidesse seda nähtust lihtsustada. Tulevikus võiks katsetada skaalapõhist hinnangut, näiteks hindaja kindlust vastuses või tahtlikkuse tõenäosust. See võimaldaks mudelil õppida ka hinnangu tugevust.

Mudelite tulemused näitavad seda, kui hästi suudavad need eelvalitud tekstikorpuses taastoota ekspertide tahtlikkuse hinnangut. Kõik analüüsitud tekstid pärinesid portaalidest Telegram, Objektiiv, Uued Uudised ja Vanglaplaneet. Lisaks pärinesid tekstid propagandat sisaldavate tekstide valimist ja tekstides sisalduvad retoorikavõtted olid inimhindajatele visuaalselt esile toodud, mis võisid alateadlikult suunata hindajaid tahtlikkust üle tuvastama. Lisaks tuleb siin arvestada, et mudelite sisendis puudusid sildistatud tekstid. Seda, kui hästi mudel töötaks peavoolumeedias (nt ERR või Postimehe artiklite peal), kus enamik tekste ei sisalda propagandat ega retoorikavõtteid, ei saa tulemuste põhjal veel järeldada. Edasistes uuringutes oleks kasulik võrrelda, kuidas inimesed hindavad täiesti puhtaid ja märgistamata tekste ning testida mudeleid tavapäraste artiklite peal. See võimaldaks veel paremini kontrollida, kas mudel suudab tahtlikku mõjutustegevust tuvastada või liigitab selleks ka tavapärast kommunikatsiooni.

Kuigi pseudomärgistatud andmetega mudel saavutas parimad tulemused ja andmete mahu suurendamine aitas klasse paremini eristada, tuleks nende andmete kasutamisse siiski ettevaatlikult suhtuda. Kuna treeningandmed märgistati automaatselt keelemudeli GPT-5.4 abil, on väga tõenäoline, et keelemudeli kalduvus tahtlikkust üle tuvastada kandus üle ka treenitud mudelile. Seega ei saa kindlalt öelda, et mudel toimiks sama hästi uutel tekstidel ja selle üldistusvõime hindamiseks tuleks edasistes uuringutes mudelit katsetada uutel tekstidel, mida ei ole treeningus ega märgistamisel varem kasutatud.

Tulemuste tõlgendamisel tuleb ka arvestada, et need põhinevad konkreetsetel mudelivalikutel. Töös kasutasin juhendatud lähenemises *XLM-RoBERTa-base* mudelit ja generatiivses lähenemises GPT-5.4 mudelit. Tulemused peegeldavad just nende mudelite omadusi, mitte masinõppe võimalusi üldiselt. Edasistes uurimustes tuleks samasugust ülesannet korrata erinevate juhendatud ja generatiivsete mudelitega.

Need piirangud ei vähenda töö väärtust, vaid piiritlevad tulemuste asjakohast tõlgendamist. Töö põhjal ei saa järeldada, et kasutatud mudelid on valmis praktiliseks rakendamiseks või et tekstides saaks tahtlikkus täiesti automaatselt tuvastada. Kuid tulemused kinnitavad, et masinõppemeetoditega on vähemalt osaliselt võimalik tahtlikkust tuvastada. Töö peamine

panus seisneb meetodi esmase teostatavuse näitamises ja tahtlikkuse indikaatori võimalike veapiiride määratlemises.

KOKKUVÕTE

Tänapäevast infosfääri iseloomustab infoüleküllus ja auditooriumi tähelepanu nappus, mis muudab tahtliku mõjutustegevuse eristamise tavapärasest kommunikatsioonist üha keerulisemaks (Starbird jt, 2019; Caled ja Silva, 2022). Kui informatsiooni maht ületab inimeste kognitiivse töötlusvõime, muutuvad vajalikuks toetavad lahendused, mille abil suuri andmemahtusid analüüsida (Abro jt, 2023). Magistritöö eesmärk oli välja selgitada, kas masinõppepõhist tahtlikkuse indikaatorit saab kasutada eestikeelsetes tekstides otsustustoe elemendina tahtliku mõjutustegevuse eristamisel tavapärasest kommunikatsioonist.

Töö empiirilises osas kasutasin Vilniuse Ülikooli projekti raames kogutud eestikeelsete tekstide andmestikku. Magistritöö analüüsiks valisin 30 teksti, mis pärinesid eelnevalt propagandat sisaldavaks märgitud artiklite hulgast. Tahtliku mõjutustegevuse esinemist hindasid meedia- ja kommunikatsioonivaldkonna eksperdid ning mitteekspertidest koosnev kontrollgrupp. Analüüsi keskseks võrdlusaluseks oli ekspertide häälteenamushinnang, mille põhjal moodustasin binaarse sihtsildi, kus “jah” tähistas tahtliku mõjutustegevuse esinemist ja “ei” selle puudumist.

Uurimisküsimusele vastamiseks võrdlesin juhendatud mudelit, generatiivset mudelit ja pseudomärgistatud andmetega treenitud juhendatud mudelit. Tulemused näitasid, et tahtlikkuse indikaator on masinõppe abil osaliselt modelleeritav, kuid pole piisav täielikult iseseisvaks ja automaatseks tuvastamiseks. Üheks olulisemaks tulemuseks oli pseudomärgistatud andmetega juhendatud mudel, mis näitas treeningandmete mahu suurendamise positiivset mõju mudeli võimele ekspertide hinnangut taastoota.

Samas näitasid tulemused, et mudelite üldist täpsust tuleb ettevaatlikult tõlgendada. Kuna ekspertide hinnangute jaotus oli “jah” klassi poole kaldu, võis õigete ennustuste osakaal tuleneda mudelite kalduvusest määrata tekste tahtlikku mõjutustegevust sisaldavaks. See tuli selgelt välja kontrollgrupi ja generatiivse mudeli puhul, mis tuvastasid hästi “jah” klassi kuuluvad tekstid, kuid kaldusid liiga tihti tahtlikuks pidama neid tekste, mida eksperdid tahtlikuks mõjutustegevuseks ei hinnanud. Töö seisukohalt oli seega oluline mitte ainult üldine täpsus, vaid mudelite võime eristada “ei” klassi.

Mudelite tulemused ei olnud täielikult täpsed, kuid need annavad aimu, millises kontekstis seda tunnust kasutada saab ja kui suur kaal sellele anda. Töö peamine järeldus on, et tahtlikkuse indikaator on väärtuslik otsustustoe tunnus, kuid mitte iseseisev otsustuskriteerium. See võib toimida esmase filtrina, et pöörata tähelepanu edasist analüüsi vajavatele juhtumitele, kuid

praktilisem oleks see osana laiemas hindamissüsteemis. Kui tahtlikkuse indikaatorit kombineerida teiste tunnustega, näiteks levikumustrite ja allikakontekstiga, võib see aidata vähendada infomüra ning suunata tähelepanu olulisematele juhtumitele.

Analüüsil oli mitu piirangut. Esiteks põhines töö väikesel ja klasside jaotuse poolest ebaühtlasel andmestikul. Hindajatele esitatud tekstides olid retoorikavõtted visuaalselt esile toodud, mis võis mõjutada tahtlikkuse tajumist. Samuti tuleb ettevaatlikult suhtuda pseudomärgistatud andmete kasutamisse, sest generatiivse mudeli loodud sildid võisid edasi kanda mudeli enda kallakuid.

Tahtlikkuse indikaatori praktilise kasutatavuse täpsemaks hindamiseks tuleb kindlasti teha täiendusi. Edasistes uuringutes tuleks kasutada suuremat ja tasakaalukamat andmestikku, kaasata rohkem hindajaid ning kasutada märgistamata ja tavapärasemaid tekste. Samuti peaks katsetama erinevaid juhendatud ja generatiivseid mudeleid.

Minu töö panus seisneb selles, et see näitab tahtlikkuse kui keeruka ja tõlgendusliku tunnuse masinõppelist modelleeritavust eestikeelsetes tekstides. See ei paku valmis automaatset tuvastuslahendust, kuid annab aluse tahtlikkuse hindamiseks ühe elemendina laiemast otsustusest mõjutustegevuse varajasel märkamisel.

SUMMARY

The aim of this Master's thesis was to examine whether a machine-learning-based intentionality indicator can be used as a decision-support element for detecting intentional influence activity in Estonian-language texts. The thesis does not aim to develop a fully autonomous detection system. Instead, it evaluates whether textual patterns associated with expert assessments of intentionality can be modelled sufficiently well to support the early identification of potentially strategic influence activity.

The empirical analysis was based on a dataset of 30 Estonian-language texts selected from a previously labelled corpus of propaganda-related articles. The presence of intentional influence activity was assessed by seven media and communication experts and seven non-expert participants in a control group. The expert majority vote was used as the main reference standard for the analysis. Based on this reference, the texts were assigned a binary target label, where "yes" indicated the presence of intentional influence activity and "no" indicated its absence.

Three computational approaches were evaluated. First, a supervised machine learning model based on XLM-RoBERTa-base was trained and tested on the expert-labelled data. Second, a generative language model, GPT-5.4, was evaluated in both zero-shot and few-shot settings. Third, a supervised model was trained on pseudo-labelled data, where the training set was expanded with 190 additional texts labelled by the generative model. The purpose of comparing these approaches was not to establish a final technical ranking of models, but to assess whether intentionality can be modelled as a useful indicator for decision support.

The results showed that all tested approaches were better at identifying texts that experts had classified as containing intentional influence activity than at identifying texts where experts did not identify such activity. However, correctly distinguishing the "no" class emerged as the primary challenge. This indicates that both non-expert readers and the generative model tended to over-detect intentionality, especially in texts with strong rhetoric or emotional framing.

The best overall results were achieved by the supervised model trained on pseudo-labelled data. This suggests that increasing the amount of training data improved the model's ability to reproduce expert assessments and to distinguish the two classes more evenly. At the same time, the use of pseudo-labelled data must be interpreted cautiously, as labels generated by a language model may also reproduce the model's own biases.

The main theoretical conclusion of the thesis is that intentionality should be understood as a mediated and evaluative construct rather than as a directly measurable empirical property of text. The high number of false positives among both the control group and the generative model shows that rhetorically strong or emotionally charged communication can easily be mistaken for deliberate influence activity. Therefore, an intentionality indicator should not be used as a standalone decision criterion.

However, the study demonstrates that a machine-learning-based intentionality indicator can be valuable as one component in a broader decision-support system. In practice, such an indicator could function as an initial filter that helps direct expert attention to texts requiring further analysis. Its practical value would be strongest when combined with other indicators, such as source credibility, dissemination patterns, coordinated behaviour and the recurrence of specific narratives. In this way, the indicator could help reduce information noise and support the early detection of potentially strategic influence activity.

The study has several limitations. The analysis was based on a small and imbalanced dataset, where most texts belonged to the "yes" class. The texts were also selected from a corpus of articles that had already been labelled as containing propaganda, which limits the generalisability of the results to the broader Estonian-language media environment. In addition, the texts shown to participants included visually highlighted rhetorical techniques, which may have influenced how intentionality was perceived. Future research should therefore use larger and more balanced datasets, include more annotators, compare marked and unmarked texts, and test the indicator on ordinary news texts as well as with different supervised and generative models.

Overall, the contribution of this thesis lies in showing that intentionality, although theoretically complex and interpretative, can be partially modelled in Estonian-language texts. The thesis does not provide a ready-made automated detection tool, but it offers an empirical basis for evaluating how intentionality could function as one element of decision support in the early identification of influence activity.

Keywords: intentionality detection, influence operations, decision support systems, machine learning, generative language models, pseudo-labeling.

KASUTATUD KIRJANDUS

Abro, A. A., Talpur, M. S. H., ja Jumani, A. K. (2023). Natural Language Processing Challenges and Issues: A Literature Review. *Gazi University Journal of Science*, 36(4), 1522–1536. https://doi.org/10.35378/gujs.1032517

Aikin, S., ja Casey, J. (2024). What about Whataboutism? *Social Epistemology*, 1–10. https://doi.org/10.1080/02691728.2024.2343729

Armstrong, J. (2023). Communication before communicative intentions. *Noûs*, *57*(1), 26–50. https://doi.org/10.1111/nous.12396

Arnold, M., Goldschmitt, M., ja Rigotti, T. (2023). Dealing with information overload: A comprehensive review. *Frontiers in Psychology*, *14*, 1122200. https://doi.org/10.3389/fpsyg.2023.1122200

ATSPARA. (i.a.). About. Kasutatud 04.04.2026, https://www.atspara.mif.vu.lt/about/.

Bakir, V., Herring, E., Miller, D., ja Robinson, P. (2019). Organized Persuasive Communication: A new conceptual framework for research on public relations, propaganda and promotional culture. *Critical Sociology*, *45*(3), 311–328. https://doi.org/10.1177/0896920518764586

Bakirov, A., ja Suleimenov, I. (2025). Theoretical Bases of Methods of Counteraction to Modern Forms of Information Warfare. Computers, 14(10), 410. https://doi.org/10.3390/computers14100410

Bjola, C. (2018). The Ethics of Countering Digital Propaganda. *Ethics & International Affairs*, *32*(3), 305–315. https://doi.org/10.1017/S0892679418000436

Boulianne, S., Tenove, C., ja Buffie, J. (2022). Complicating the Resilience Model: A Four-Country Study About Misinformation. *Media and Communication*, *10*(3), 169–182. https://doi.org/10.17645/mac.v10i3.5346

Brashier, N. M., ja Marsh, E. J. (2020). Judging Truth. *Annual Review of Psychology*, *71*(1), 499–515. https://doi.org/10.1146/annurev-psych-010419-050807

Burton, J. (2023). Algorithmic extremism? The securitization of artificial intelligence (AI) and its impact on radicalism, polarization and political violence. *Technology in Society*, *75*, 102262. https://doi.org/10.1016/j.techsoc.2023.102262

Caled, D., ja Silva, M. J. (2022). Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. *Journal of Computational Social Science*, *5*(1), 123–159. https://doi.org/10.1007/s42001-021-00118-8

Chae, Y., ja Davidson, T. (2026). Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning. *Sociological Methods & Research*, *55*(2), 501–567. https://doi.org/10.1177/00491241251325243

Chang, R. M., Kauffman, R. J., ja Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, *63*, 67–80. https://doi.org/10.1016/j.dss.2013.08.008

Chen, A., Zhang, Y., Liu, Y., ja Lu, Y. (2023). Be a good speaker in livestream shopping: A speech act theory perspective. *Electronic Commerce Research and Applications*, *61*, 101301. https://doi.org/10.1016/j.elerap.2023.101301

Christiano, T. (2022). Algorithms, Manipulation, and Democracy. *Canadian Journal of Philosophy*, *52*(1), 109–124. https://doi.org/10.1017/can.2021.29

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., ja Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Deffuant, G., Keijzer, M., ja Banisch, S. (i.a.). *How opinions get more extreme in an age of information abundance*.

Dehnert, M., ja Mongeau, P. A. (2022). Persuasion in the Age of Artificial Intelligence (AI): Theories and Complications of AI-Based Persuasion. *Human Communication Research*, *48*(3), 386–403. https://doi.org/10.1093/hcr/hqac006

François, C. (2019). *Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses*. Annenberg Public

Policy Center. https://www.annenbergpublicpolicycenter.org/wp-content/uploads/ABC_Framework_TWG_Francois_Sept_2019.pdf

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., ja Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

French, A., Storey, V. C., ja Wallace, L. (2024). A typology of disinformation intentionality and impact. *Information Systems Journal*, *34*(4), 1324–1354. https://doi.org/10.1111/isj.12495

Gebauer, G., ja William, J. M. (2000). Habitus, Intentionality, and Social Rules: A Controversy between Searle and Bourdieu. *SubStance*, *29*(3), 68–83. https://doi.org/10.1353/sub.2000.0033

Gibert, S. (2023). The Wrong of Wrongful Manipulation. *Philosophy & Public Affairs*, *51*(4), 333–372. https://doi.org/10.1111/papa.12247

Hobbs, R. (2020). Propaganda in an Age of Algorithmic Personalization: Expanding Literacy Research and Practice. *Reading Research Quarterly*, *55*(3), 521–533. https://doi.org/10.1002/rrq.301

Hoeken, H., Kolthoff, M., ja Sanders, J. (2016). Story Perspective and Character Similarity as Drivers of Identification and Narrative Persuasion: Perspective, Similarity, and Identification. *Human Communication Research*, *42*(2), 292–311. https://doi.org/10.1111/hcre.12076

Hox, J. J. (2017). Computational Social Science Methodology, Anyone? *Methodology*, *13*(Supplement 1), 3–12. https://doi.org/10.1027/1614-2241/a000127

Humă, B. (2023). Language and persuasion: A discursive psychological approach. *Social and Personality Psychology Compass*, *17*(6), e12755. https://doi.org/10.1111/spc3.12755

Hyzen, A. (2021). Revisiting the Theoretical Foundations of Propaganda. *International Journal of Communication 15(2021), 3479–3496.*

Jordan, M. I., ja Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kang, M., Seo, J., Park, C., ja Lim, H. (2022). Utilization Strategy of User Engagements in Korean Fake News Detection. *IEEE Access*, *10*, 79516–79525. https://doi.org/10.1109/ACCESS.2022.3194269

Kaptein, M., Markopoulos, P., De Ruyter, B., ja Aarts, E. (2015). Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, *77*, 38–51. https://doi.org/10.1016/j.ijhcs.2015.01.004

Kertysova, K. (2018). Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered. *Security and Human Rights*, *29*(1–4), 55–81. https://doi.org/10.1163/18750230-02901005

Khosravi, S., ja Barekat, B. (2021). "The Intentional Fallacy", itself a Fallacy: A Critique of Wimsatt and Beardsley's "The Intentional Fallacy". *LANGUAGE ART*, *6*(2), 77–90. https://doi.org/10.22046/LA.2021.11

Kittask, C., Milintsevich, K., ja Sirts, K. (2020). Evaluating Multilingual BERT for Estonian. In A. Utka, J. Vaičenonienė, J. Kovalevskaitė, ja D. Kalinauskaitė (Eds), *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/FAIA200597

Klenk, M. (2022). (Online) manipulation: Sometimes hidden, always careless. *Review of Social Economy*, *80*(1), 85–105. https://doi.org/10.1080/00346764.2021.1894350

Kärki, K. (2023). Explaining with Intentional Omissions. *Journal for the Theory of Social Behaviour*, *53*(3), 417–432. https://doi.org/10.1111/jtsb.12378

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., ja Van Alstyne, M. (2009). Computational Social Science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., ja Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Levy, Y. (2025). The priority of intentional action: From developmental to conceptual priority. *The Philosophical Quarterly*, *75*(2), 598–631. https://doi.org/10.1093/pq/pqae023

Liagusha, A., ja Iarovyi, D. (2025). Memes, freedom, and resilience to information disorders: Information warfare between democracies and autocracies. *Social Sciences & Humanities Open*, 11, 101247. https://doi.org/10.1016/j.ssaho.2024.101247

Lund, N. F., Scarles, C., ja Cohen, S. A. (2020). The Brand Value Continuum: Countering Co-destruction of Destination Branding in Social Media through Storytelling. *Journal of Travel Research*, *59*(8), 1506–1521. https://doi.org/10.1177/0047287519887234

Madisson, M.-L., ja Ventsel, A. (2021). *Strategic conspiracy narratives: A semiotic approach*. Routledge, Taylor & Francis Group.

Makinda, S. (2021). Understanding the Global Interpretive Community. *Academia Letters*. https://doi.org/10.20935/AL2086

Martin, J. (2016). Capturing Desire: Rhetorical Strategies and the Affectivity of Discourse. *The British Journal of Politics and International Relations*, *18*(1), 143–160. https://doi.org/10.1111/1467-856X.12065

Munch, P. A. (1975). "Sense" and "Intention" in Max Weber's Theory of Social Action. *Sociological Inquiry*, *45*(4), 59–65. https://doi.org/10.1111/j.1475-682X.1975.tb00350.x

Murdock, G. (2016). Encoding and Decoding. In P. Rössler, C. A. Hoffner, ja L. Zoonen (Eds), *The International Encyclopedia of Media Effects* (1st edn, pp. 1–11). Wiley. https://doi.org/10.1002/9781118783764.wbieme0113

Nädala vandenõuteooriad, valeinfo ja koroonamüüdid. (2020). *Propastop*, 12. aprill. Kasutatud 26.05.2026, https://www.propastop.org/2020/04/12/5885/

OpenAI. (i.a.). *GPT-5.4 model*. OpenAI API documentation. Kasutatud 25.04.2026, https://developers.openai.com/api/docs/models/gpt-5.4

Ottaviani, F., Le Roy, A., ja O'sullivan, P. (2021). Constructing Non-monetary Social Indicators: An Analysis of the Effects of Interpretive Communities. *Ecological Economics*, *183*, 106962. https://doi.org/10.1016/j.ecolecon.2021.106962

Pandey, J. (2019). Deductive Approach to Content Analysis. In M. Gupta, M. Shaheen, ja K. Reddy (Toim), *Qualitative Techniques for Workplace Data Analysis* (lk 145-169). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-5225-5366-3.ch007

Pantazi, M., Hale, S., ja Klein, O. (2021). Social and Cognitive Aspects of the Vulnerability to Political Misinformation. *Political Psychology*, 42(S1), 267–304. https://doi.org/10.1111/pops.12797

Persson, P. (2018). Attention manipulation and information overload. *Behavioural Public Policy*, *2*(1), 78–106. https://doi.org/10.1017/bpp.2017.10

Petty, R. E., ja Briñol, P. (2015). Emotion and persuasion: Cognitive and meta-cognitive processes impact attitudes. *Cognition and Emotion*, *29*(1), 1–26. https://doi.org/10.1080/02699931.2014.967183

Pianese, A., Attias, A., ja Varga, Z. (2014). Dynamic immigration control improving inverse old-age dependency ratio in a pay-as-you-go pension system. *Decision Support Systems*, *64*, 109–117. https://doi.org/10.1016/j.dss.2014.04.009

Prichard, A. (2017). Collective intentionality, complex pluralism and the problem of anarchy. *Journal of International Political Theory*, *13*(3), 360–377. https://doi.org/10.1177/1755088217715789

Puri, R., ja Catanzaro, B. (2019). *Zero-shot Text Classification With Generative Language Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1912.10165

Saaristo, A. (2006). There Is No Escape from Philosophy: Collective Intentionality and Empirical Social Science. *Philosophy of the Social Sciences*, *36*(1), 40–66. https://doi.org/10.1177/0048393105284170

Saeidnia, H. R., Jahani, S., Ghiasi, N., ja Keshavarz, H. (2026). Generative AI and health misinformation: Production, propagation, and mitigation—a systematic review. *BMC Public Health*, *26*(1), 693. https://doi.org/10.1186/s12889-025-26148-9

Seos FB faktikontrolliga võimaldab rünnata ajakirjandust. (2020). *Propastop*, 30. juuli. Kasutatud 26.05.2026, https://www.propastop.org/2020/07/30/seos-fb-faktikontrolliga-voimaldab-runnata-ajakirjandust/

Shepherd, J., ja Carter, J. A. (2023). Knowledge, Practical Knowledge, and Intentional Action. *Ergo an Open Access Journal of Philosophy*, *9*(0). https://doi.org/10.3998/ergo.2277

Singh, R., Kim, J. Y., Glassy, E. F., Dash, R. C., Brodsky, V., Seheult, J., De Baca, M. E., Gu, Q., Hoekstra, S., ja Pritt, B. S. (2025). Introduction to Generative Artificial Intelligence:

Contextualizing the Future. *Archives of Pathology & Laboratory Medicine*, *149*(2), 112–122. https://doi.org/10.5858/arpa.2024-0221-RA

Smith, G. (2024). The Role of Pragmatics in Cross-Cultural Communication. *European Journal of Linguistics*, *3*(1), 13–24. https://doi.org/10.47941/ejl.1768

Starbird, K., Arif, A., ja Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26. https://doi.org/10.1145/3359229

Supriyono, Wibawa, A. P., Suyono, ja Kurniawan, F. (2024). Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports*, *16*, 100173. https://doi.org/10.1016/j.teler.2024.100173

Rizgelienė, I., Zubaitienė, V., Maliukevičius, N., ja Marcinkevičius, V. (2025). HALT-PROP: Human-Annotated Lithuanian Textual Corpus for Propaganda Narratives and Techniques. *Scientific Data*, *13*(1), 47. https://doi.org/10.1038/s41597-025-06367-w

Tanduk, R. (2023). PRAGMATIC ASPECTS OF SPEECH ACTS: A CROSS-LINGUISTIC PERSPECTIVE. *English Review: Journal of English Education*, *11*(3). https://doi.org/10.25134/erjee.v11i3.8762

Tetsmann, L., Kangur, U., Chakraborty, R., ja Sharma, R. (2026). CAPS: A Cross-Lingual Methodology for Detecting Misinformation in Estonian Health News. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *25*(3), 1–24. https://doi.org/10.1145/3797259

Tollefsen, D. P. (2002). Collective Intentionality and the Social Sciences. Philosophy of the Social Sciences, 32(1), 25–50. https://doi.org/10.1177/004839310203200102

Turow, J. (2005). Audience Construction and Culture Production: Marketing Surveillance in the Digital Age. *The ANNALS of the American Academy of Political and Social Science*, *597*(1), 103–121. https://doi.org/10.1177/0002716204270469

Urve Eslas: objektiiv.ee infolõimed jooksevad Venemaa ja Süüria kaudu. (2017). *Propastop*, 24. aprill. Kasutatud 26.05.2026, https://www.propastop.org/2017/04/24/urve-eslas-objektiiv-ee-infoloimed-jooksevad-venemaa-ja-suuria-kaudu/

Usmani, S., ja Almashham, A. (2024). Cross-Cultural Pragmatics: Analysing Speech Acts in Different Cultures. *International Journal of Language and Literary Studies*, *6*(1), 186–198. https://doi.org/10.36892/ijlls.v6i1.1586

Van Benthem, J., Ghosh, S., ja Verbrugge, R. (Eds). (2015). *Models of Strategic Reasoning* (Vol. 8972). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48540-8

Ventsel, A., Hansson, S., Madisson, M.-L., ja Sazonov, V. (2021). Discourse of fear in strategic narratives: The case of Russia's Zapad war games. *Media, War & Conflict*, *14*(1), 21–39. https://doi.org/10.1177/1750635219856552

Verbraken, T., Goethals, F., Verbeke, W., ja Baesens, B. (2014). Predicting online channel acceptance with social network data. *Decision Support Systems*, 63, 104–114. https://doi.org/10.1016/j.dss.2013.08.011

Välisluureamet. (2025). *International security and Estonia 2025*. https://www.valisluureamet.ee/doc/raport/2025-en.pdf

Wanless, A., ja Berk, M. (2020). The audience is the amplifier: Participatory propaganda. *The SAGE handbook of propaganda*, 85-104.

Wardle, C., ja Derakhshan, H. (2017). *INFORMATION DISORDER : Toward an interdisciplinary framework for research and policy making Information Disorder Toward an interdisciplinary framework for research and policymaking*. Council of Europe.

Warren, E., ja Call, J. (2022). Inferential Communication: Bridging the Gap Between Intentional and Ostensive Communication in Non-human Primates. *Frontiers in Psychology*, *12*, 718251. https://doi.org/10.3389/fpsyg.2021.718251

Weitkamp, E., Milani, E., Ridgway, A., ja Wilkinson, C. (2021). Exploring the digital media ecology: Insights from a study of healthy diets and climate change communication on digital and social media. *Journal of Science Communication*, 20(03), A02. https://doi.org/10.22323/2.20030202

Wirz, D. (2018). *Persuasion Through Emotion? An Experimental Test of the Emotion-Eliciting Nature of Populist Communication*. https://doi.org/10.5167/UZH-149959

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern recognition, 48(9), 2839-2846.

World Health Organization. (2017). *Strategic communications framework for effective communications*. https://www.who.int/docs/default-source/documents/communicating-for-health/communication-framework.pdf

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., ja Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

LISAD

Lisa 1. Ekspertide ja kontrollgrupi vastuste koondtabel

Näidise number	Eskpertgrupp	kontroll-grupp
1	Jah	Jah
2	Jah	Jah
3	Jah	Jah
4	Jah	Jah
5	Ei	Jah
6	Ei	Jah
7	Ei	Jah
8	Jah	Jah
9	Ei	Ei
10	Ei	Ei
11	Jah	Jah
12	Ei	Jah
13	Ei	Jah
14	Jah	Jah
15	Ei	Jah
16	Jah	Jah
17	Jah	Jah
18	Jah	Jah
19	Jah	Jah
20	Jah	Jah
21	Jah	Jah
22	Jah	Jah
23	Jah	Jah
24	Jah	Jah

25	Jah	Jah
26	Jah	Ei
27	Ei	Jah
28	Jah	Jah
29	Jah	Jah
30	Jah	Jah

Lisa 2. Generatiivse mudeli juhis ehk *prompt*

You are a strict binary text classifier.

Task:
Decide whether the text has the intention of sharing or promoting a propaganda narrative.

Output rules (VERY IMPORTANT):
- First line: output exactly one token: jah or ei
- Second line onwards: concise explanation for your label

Lisa 3. Tabel mudelite infoga

Hüperparameeter	**LOOCV**	**Pseudomärgistega eksperiment**	***Zero-shot* GPT**	***Few-shot* GPT**
Mudel	XLM-RoBERTa-base		GPT-5.4	
Mudeli viide	Conneau jt (2020)		OpenAI (2026)	
Mudeli väljalaske kuupäev	01.11.2019		17.03.2026	
Parameetrid	270M (12 kihti, 768 peidetud olekut)		Avaldamata	
Kontekstiaken	512 märgist (tokenit)		400K märgist (tokenit)	
Eeltreenimise andmestik	2.5TB CommonCrawl, 100 keelt		Avaldamata	
Külmutatud kihid	Kõik peale klassifitseerimispea	Kõik peale viimase transformer-kihi ja klassifitseerimispea	–	
Treenitavad parameetrid	~1,000	~7,680,002 (2.8%)	–	
Õppemäär	3,00E-04		–	
Partii suurus	8		–	
Epohhide arv	25		–	
Kaofunktsioon	Kaalutud ristentroopia		–	
Jada pikkus	128		–	
Temperatuur	–		0	

Näidete arv	–		0	4 (2 jah + 2 ei)
Andmejaotus	LOOCV ekspertandmestikul (n=30)	190 pseudomärgistatud teksti, 80/20 treening- ja valideerimisandmete jaotus, testimine toimus ekspertandmestikul (n=30)	Otse valideeritud ekspertandmestikul (n=30)	

LIHTLITSENTS LÕPUTÖÖ REPRODUTSEERIMISEKS JA ÜLDSUSELE KÄTTESAADAVAKS TEGEMISEKS

Mina, Helina Toompark,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Mõjutustegevuse tahtlikkuse kvantitatiivne mõõtmine tekstis", mille juhendajad on Sten Torpan ja Uku Kangur, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;

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

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

Helina Toompark

27.05.2026