TARTU ÜLIKOOL
Arvutiteaduse instituut
Andmeteaduse õppekava

Tauno Tamm

Eesti alamredditi korpuse loomine ning analüüs

Magistritöö (15 EAP)

Juhendaja: Siim Orasmaa, PhD

Tartu 2024

Eesti alamredditi korpuse loomine ning analüüs

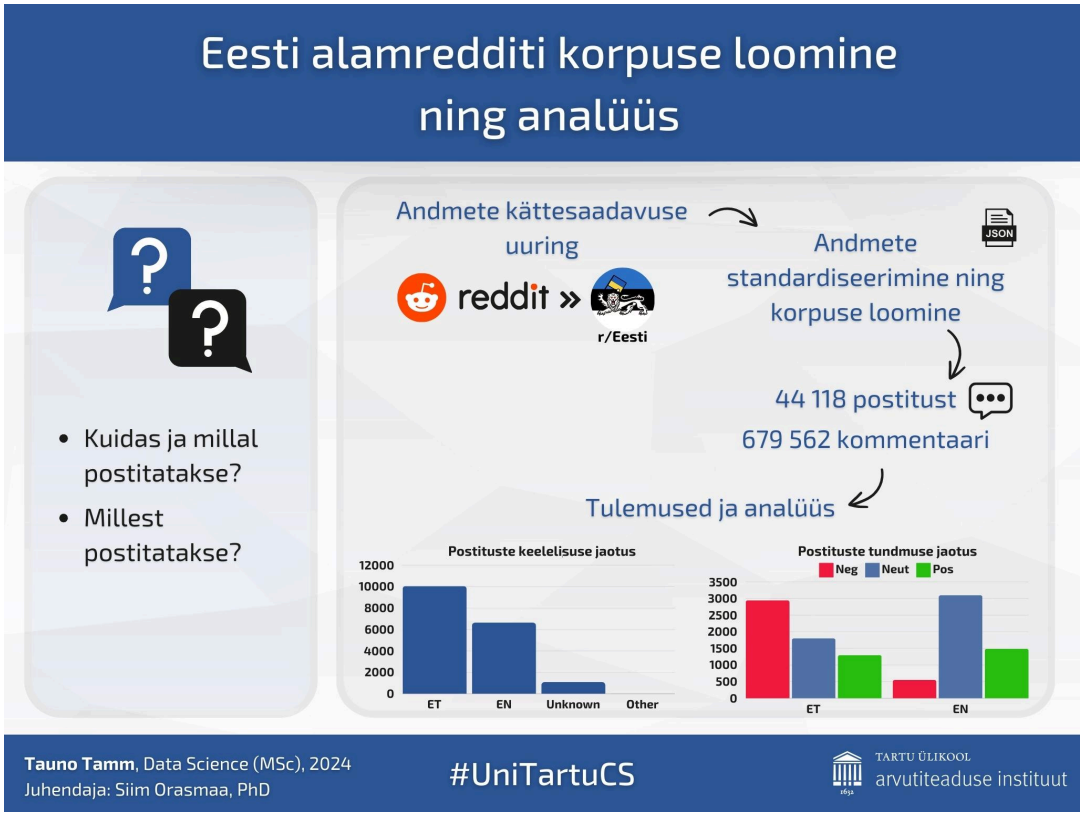
Lühikokkuvõte:

Reddit on maailma suurim foorum, mida jälgib igakuiselt umbes 1.2 miljardit kasutajat. Eesti suurimaks *subreddit*-iks ehk alamredditiks on r/Eesti. Selle magistritöö käigus loodi r/Eesti andmete põhjal keelekorpus ning analüüsiti seal olevaid andmeid. Analüüsi käigus vastati järgnevatele uurimisküsmustele: kuidas ja millal postitatakse ning millest postitatakse. Uurimisküsimustele vastamiseks peenhäälestati ning kasutati erinevaid siirdeõppe mudeleid tundmusanalüüsi läbiviimiseks, Pythoni keeletuvastuse teeki Lingua keeletuvastuseks, teemade analüüsiks BERTopic-ut jpm. Tulemustest selgus, et r/Eesti alamredditit võib pidada kakskeelseks, sest lisaks eesti keelele on suur osa postitusi ning kommentaare tehtud ka inglise keeles. Tundmusanalüüs näitas, et eesti keeles postitavad ja kommenteerivad kasutajad on tugevalt negatiivselt meelestatud, kuid inglise keeles kirjutavad kasutajad on tugevalt neutraalselt meelestatud, olles pigem positiivse tundmuse poole kaldu. Mõlema keele puhul on kõige populaarsemaks ühtivaks teemaks „Haridus“.

Võtmesõnad:

Reddit, Loomuliku keele töötlus, Tundmusanalüüs, Keeletuvastus, r/Eesti, BERTopic

CERCS: P175



Creation and Analysis of the Estonian Subreddit Corpus

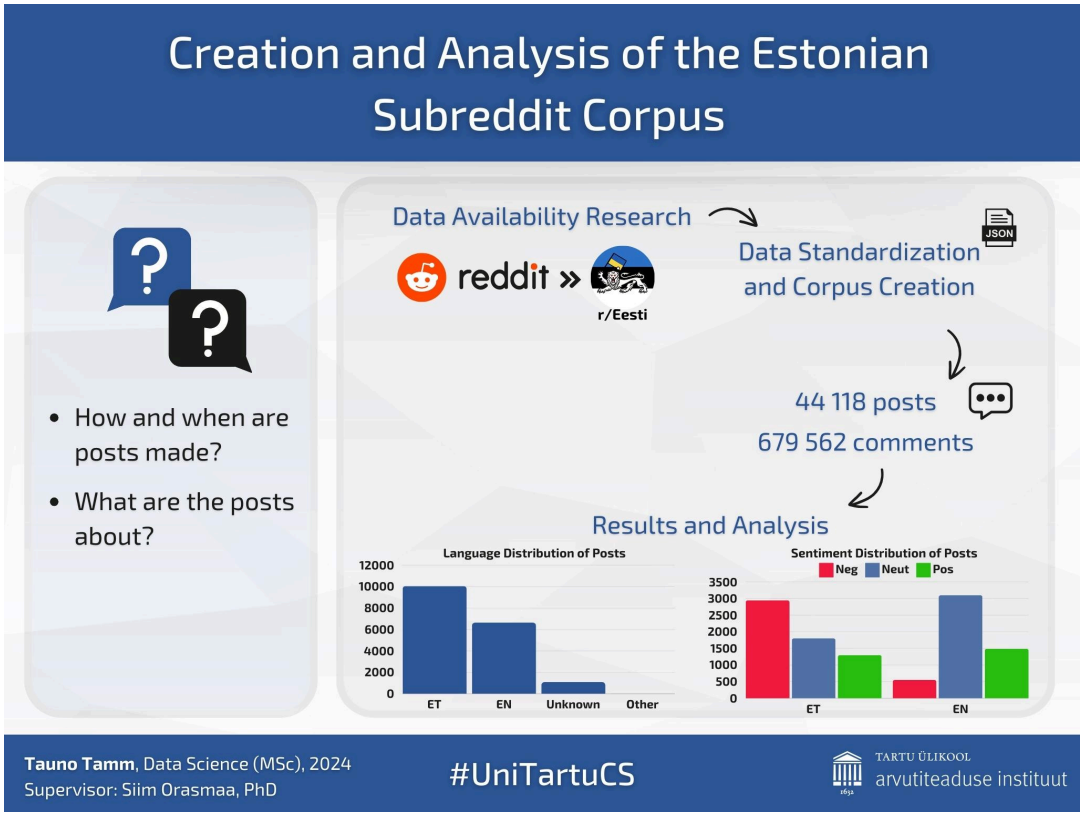
Abstract:

Reddit is the world's largest forum, visited by about 1.2 billion users monthly. The largest Estonian subreddit is r/Eesti. This master's thesis involved creating a language corpus based on the data from r/Eesti and analyzing the data therein. The analysis addressed questions on how and when posts are made and what they discuss. For answering these research questions, various transformer-type models were fine-tuned for sentiment analysis, the Python language detection library Lingua was used for language detection, and BERTopic was employed for topic analysis. The results revealed that the r/Eesti subreddit can be considered bilingual, as a significant portion of posts and comments are also in English. The sentiment analysis exhibited that users posting and commenting in Estonian are mostly negative, while those who write in English tend to be neutral, with a slight lean towards positivity. In both languages, “Education” is the most common topic.

Keywords:

Reddit, Natural Language Processing, Sentiment Analysis, Language Detection, r/Eesti, BERTopic

CERCS: P175



Sisukord

1. Sissejuhatus

Tänapäevases ühiskonnas on laialt levinud erinevad sotsiaalmeedia platvormid. Reddit on maailma kõige suurem ning tuntuim foorum, mida jälgib igakuiselt umbes 1,2 miljardit kasutajat. [1]. Foorumina sarnaneb Reddit teiste sotsiaalmeedia platvormidega, kus kasutajatel on võimalik postitada sisu, kommenteerida ning hääletada.

Käesoleval magistritööl on paralleelselt kaks suurt eesmärki. Esimeseks eesmärgiks on luua r/Eesti andmete põhjal kasutatav korpus, kus on säilitatud ka postituste ja kommentaaridega seotud metaandmed, ning mida on võimalik hiljem kasutada teadustöö eesmärkidel. Teiseks eesmärgiks on loodud korpuses olevaid andmeid analüüsida kasutades loomuliku keele töötluse meetodeid, sealhulgas tundmusanalüüsi, teemade analüüsi ning keeletuvastust, et mõista alamredditi sisu struktuuri, hoiakute ja teemade jaotust. Magistritöö keskendub oluliste mustrite ja seoste välja toomisele. See on oluline, et aru saada, milline on Eesti interneti- ja sotsiaalmeedia kasutajate dünaamika ja käitumine erinevate teemade suhtes, mis võib anda hea sisendi ühiskondlike trendide jälgimiseks ja arusaamiseks. Magistritöö uurimisküsimused on järgnevad:

1. **Kuidas ja millal postitatakse?**

Tegemist on postituste ja kommentaaride vormilise poole analüüsiga. Siia alla kuuluvad alamküsimused:

- Mis liiki postitusi tehakse?
- Millal postitatakse?
- Millises keeles postitatakse/kommenteeritakse?

2. **Millest postitatakse?**

Postituste ja kommentaaride teemade ja tundmuse analüüs. Siia alla kuuluvad alamküsimused:

- Milline on kasutajate tundmus erinevates keeltes postitades?
- Millistel teemadel tehakse enim postitusi?
- Millised on kasutajate tundmus erinevate teemade suhtes?

Magistritöö on jaotatud kuueks osaks. Esimeses osas uuritakse, millised on senised uurimistööd ja leiud antud teemal. Teises osas kirjeldatakse andmete saamise ja töötlemise protsesse sh. andmete töötlemiseks vajalikku õiguslikku alust. Kolmandas osas tuuakse välja ning kirjeldatakse lähemalt töös kasutatud meetodeid. Neljandas osas kirjeldatakse töö tulemusi ning analüüsitakse saadud tulemusi. Viies osa on arutelu osa, kus arutletakse, kas seatud uurimisküsimused said vastused ning millised on antud teema osas tuleviku väljavaated ning motivatsioon. Viimase ehk kuuenda osa puhul on tegemist kokkuvõttega, kus taaskord tuuakse välja magistritöö protsess ja tulemused.

2. Ülevaade teemakohasest kirjandusest

See peatükk keskendub Redditi andmestike kasutamisele teadustöös ning laiemalt keeletehnoloogia vahendite rakendamisele sotsiaalmeedia tekstidel. Fookus on eeskätt vahenditel, mida saab potentsiaalselt kasutada Eesti Reddit-i andmestiku. Ülevaade hetkel olemasolevatest teadustöödest annab aimu teema unikaalsusest, olulisusest, murekohtadest ja puudujääkidest, millele pole seni veel rõhku pandud.

2.1. Keeletuvastus sotsiaalmeedia andmestikel

Sotsiaalmeedia on suhtlus- ja eneseväljendusvahend, mida kasutavad paljud eri rahvused üle kogu maailma. Sellega kaasneb sotsiaalmeedias kasutatavate keelte suur mitmekesisus, mistõttu on tarvis kogukondade uurimisel rakendada sotsiaalmeedia tekstidele keeletuvastust.

Keeletuvastuse kohta on tehtud mitmeid uurimistöid, kuid nende puhul seisneb probleem selles, et uurimused ei jõua alati vabalt saadaolevate valmisrakendusteni või siis rakendused aeguvad ning neid enam ei uuendata. Saadaval on teadusartikkel [2] aastast 2014, mis on kirjutatud Marco Lui ja Timothy Baldwin poolt. Artikkel keskendub Twitteri säutsude keelelisuse täpsele tuvastamisele ning töö käigus kõrvutatakse ja võrreldakse erinevaid keeletuvastuse mudeleid ja tööriistu. Artiklist tuuakse välja, et kõige paremad tulemused saavutasid 3 lähenemist - CLD2, langid.py ja LangDetect. CLD2 puhul on selle dokumentatsioonist näha, et viimati uuendati teeki aastal 2019 [3]. Samuti on langid.py lähenemise puhul dokumentatsioonist näha selle viimatine uuendamine, mis toimus aastal 2016 [4]. LangDetect on neist kõige värskemalt uuendatud. Seda uuendati viimati 2021. aastal [5]. See annab märku, et neist kolmest oleks kõige mõistlikum valida kasutamiseks LangDetect. Siiski tasub tähele panna, et ka LangDetecti ei ole hiljuti uuendatud, mistõttu tasub uurida ka lähenemisi, mida teadusartiklites kajastatud pole.

Töö autor leidis juhuslikult töö käigus Pythoni keeletuvastuse teegi nimega Lingua[1]. Lingua toetab hetkel 75 erinevat keelt, mille hulka kuulub ka eesti keel. Erinevate klassifitseerijate täpsus eesti keele tuvastamisel on toodud tabelis 1 [6].

[1] https://pypi.org/project/lingua-language-detector/

Tabel 1. Keskmine eesti keele tuvastamise täpsus erinevate klassifitseerijatega. [6]

Klassifitseerija	Eesti keele tuvastamise täpsus
Lingua (kõrge täpsusega režiim)	92%
Lingua (madala täpsusega režiim)	83%
Langdetect	83%
FastText	73%
FastSpell (konservatiivne režiim)	73%
FastSpell (agressiivne režiim)	73%
Langid	67%
CLD3	70%
CLD2	65%
Simplemma	71%

Lingua dokumentatsiooni [6] järgi on enamiku keeletuvastuse lähenemiste probleemiks trigrammide kasutamine, kuid lühikeste tekstide puhul on leitavate trigrammide hulk väike, mistõttu kasutab Lingua erinevate tekstide jaoks 1-5 sõnalisi n-gramme. Erinevalt tavapärasest statistilisest lähenemisest kasutab Lingua lisaks ka reeglipõhist lähenemist. See seisneb selles, et esmalt uuritakse, kas tekstis esineb mõnele keelele omased tähesümboleid, mis aitab reegliga mitte-sobituvad keeled välja sorteerida ning alles seejärel n-grammi loogikat rakendada.

2.2. Tundmusanalüüs sotsiaalmeedias

Sotsiaalmeedia andmestikel tehtud tundmusanalüüsi kohta on saadaval palju akadeemilisi tekste ja teadusartikleid. Sageli on tundmusanalüüsi eesmärgiks saada ülevaade kasutajate tundmusest ja hoiakutest mingite teemade suhtes. Rahul Goel, Vijayachitra Modhukur, Katrin Täär, Andres Salumets, Rajesh Sharma ja Maire Peters on kirjutanud teadusartikli pealkirjaga „Users' Concerns About Endometriosis on Social Media: Sentiment Analysis and Topic Modeling Study“ [7], mis räägib Redditist saadud andmetel tehtud tundmusanalüüsist ning selle sidumisest teemade analüüsiga. Oma töös on nad kogunud andmed kahest ingliskeelsest

endometrioositeemalisest subredditist, eesmärgiga kaardistada endometrioosiga seonduvad teemad, millele arstid peaksid rohkem tähelepanu pöörama. Ühtlasi huvitas uurijaid, kas selle raske haiguse puhul on ka positiivse tundmusega teemasid. Töös prooviti erinevaid tundmusanalüüsi mudeleid. Nendeks olid TextBlob [8], Vader [9] ning BERTil põhinev mudel [10]. Artikli autorid leidsid, et BERT-i tulemused olid kõige paremad. Teemade analüüsi läbiviimiseks kasutati LDA-d (ingl. *Latent Dirichlet allocation*).

Selle magistritöö puhul tuleb arvestada, et TextBlobi ning Vaderi puhul on tegemist inglise keelel põhinevate mudelitega, mistõttu eestikeelsetel tekstidel neid kasutada ei saa. Eelnevalt välja toodud tulemustest lähtuvalt on ka käesoleva töö puhul huvi katsetada eelkõige BERT-i mudeleid.

2.3. Teemade analüüs sotsiaalmeedias

Teemade analüüsi abiga saab tuvastada ja eraldada peamisi teemasid suurtest tekstikogumitest. See võimaldab aru saada ja paremini struktureerida tekstilisi andmehulki, mis annab omakorda aimu andmete olemusest ning muudab need lihtsamini mõistetavaks ja visualiseeritavaks, sealhulgas võimaldab jälgida trende ajas.

Maria Kędzierska, Mikołaj Spytek, Marcelina Kurek, Jan Sawicki, Maria Ganzha ja Marcin Paprzycki on kirjutanud teadusartikli teemal „Topic Modeling Applied to Reddit Posts“ [11]. Selles artiklis võrreldakse kolme erineva teemade analüüsi lähenemist. Vaatluse all on Latent Dirichlet allocation [12] (edaspidi LDA), Mittenegatiivse maatriksi faktoriseerimine [13] (ingl. *non-negative matrix factorization* edaspidi NMF) ning BERTopic [14], mida rakendatakse Redditist saadud andmete peal. Autorid toovad tulemustes välja, et NMFi ja BERTopicu tulemused on sarnased ja veenvad, kuid LDA poolt loodud teemad olid imelikud ja mitte-loomulikud.

Lisanna Lehes kirjutas 2023. aastal teadustöö teemal „Vene-Ukraina sõja hoiakute analüüs Eesti avalikkuse ja poliitikute sotsiaalmeedia põhjal“ [15], mille eesmärgiks oli analüüsida avalikkuse kui ka poliitiliste liidrite hoiakuid Vene-Ukraina sõja suhtes, kasutades eestikeelset sotsiaalmeediat. Töö käigus kasutati andmeid, mis koguti Facebookist ja Twitterist. Kogutud andmestikul rakendati tundmusanalüüsi läbi viimiseks EstBERT mudelit [16], mis peenhäälestati Valentsikorpusel [17] ning teemade analüüsiks BERTopic-ut.

3. Andmed

See peatükk kirjeldab andmete kogumist, õigusliku aluse hindamist ja andmete puhastamist platvormilt Reddit[2].

Reddit on tekstipõhine sotsiaalne platvorm, mis koosneb alamredditest (ingl *subreddit*), millel igaühel neist on lisaks Redditi poliitikale omad reeglid (Tabel 2) [18]. Piltlikult öeldes on tegemist hiiglasliku foorumiga, mis koosneb alamfoorumitest. Alamfoorumitega saavad kasutajad liituda, seal olevat sisu lugeda ja jälgida, postitada, kommenteerida ning vastavalt postitustele ja kommentaaridele märkida enda poolthääl (ingl *upvote*) või vastuhääl (ingl *downvote*). Üldiselt on alamredditid suunatud mingile kindlale teemale või uskumusele, millega seonduvat seal alamredditis arutatakse. Näiteks r/cars alamredditis tehakse autodega seonduvaid postitusi. Lisaks on enamik Redditi kasutajad anonüümsed, st. nende kasutajanimi ei anna aimu nende identiteedi kohta.

Olenevalt alamredditi seadistustest on kasutajal võimalik teha alamredditisse viit eri tüüpi postitusi [19]. Nendeks postituse tüüpideks on:

- tekst-tüüpi postitus
- pilt-tüüpi postitus
- video-tüüpi postitus
- link-tüüpi postitus
- küsitlus-tüüpi postitus.

Eesti suurimaks alamredditiks on r/Eesti[3], kus on 2024. aasta seisuga umbes 85 000 liiget ning see kuulub top 2% suurimate alamredditite hulka [20]. Tabelis 2 on välja toodud alamredditi r/Eesti reeglid, kust on näha, et moderaatorite poolt palutakse hoida postitused neutraalsetena ning arvamust tuleks väljendada kommentaaride osas. Lisaks on välja toodud, et ähvardavate, solvavate, inimväärikust alandavate ja spämmpostituste puhul on võimalus, et alamredditi r/Eesti moderaator eemaldab postituse.

[2] https://www.reddit.com/

[3] https://www.reddit.com/r/Eesti/

Tabel 2. Alamredditi r/Eesti reeglid. [20]

1.	Teemaväline	Postitused peavad olema kas otseselt või kaudselt Eestiga seotud.
2.	Nõrk kvaliteet	Palun püüa teha postitusi, mis pakuks kogu alamredditile huvi või tekitaks kõneainet. Meemid on soovitatav postitada r/memeesti'sse - siia postitades pea silmas, et madalakvaliteetsed/liiga suvalised meemid, labane huumor ja postitused, mille sisu ei vasta pealkirjale, võidakse moderaatori poolt eemaldada. Ürita hoiduda uudiste pealkirjade muutmisest, jättes need neutraalseks, lisa oma arvamus kommentaarina mitte pealkirjana.
3.	Spämm/reklaam või tüütu	Kui sama teema/nali/meem liiga tihti esile kerkib, muutub see spämmamiseks. Spämmpostitused, reklaam, töökuulutused, tüütu kuulutamine või muidu mõõdutundetu korduva sisu postitamine võidakse moderaatori poolt eemaldada.
4.	Liiga agressiivne	Kui sa ei nõustu teise seisukohaga, palun selgita oma vaatenurka ja too näiteid, või püüa vähemalt hoiduda ähvardustest või isiklikest solvangutest. Inimväärikust alandavad postitused võidakse moderaatori poolt eemaldada.

3.1. Andmete kättesaadavus

Andmete kogumiseks oli selle töö raames algselt plaanis kasutada Pythoni teeki nimega PRAW[4] (ingl *The Python Reddit API Wrapper*). Tegemist on teegiga, mis tagab lihtsa juurdepääsu Redditi API-le (ingl. *application programming interface*). Paraku selgus andmete kogumise skripti implementeerimise käigus, et Redditi tasuta API lubab vaid kraapida 1000 kõige hilisemat postitust. Tegemist on 2023. aasta 19. juunil uuendatud Redditi API tingimustega, mis piiravad tasuta API kasutamise võimalusi [21].

Esitatud sai ka pöördumine Redditile, et saada ligipääs Redditi Pushshift API-le[5]. Pöördumise sisuks oli palve saada ligipääs, kasutades saadud andmeid teadustöö eesmärgil, kuid pöördumine lükati tagasi.

Andmete kraapimise asemel leiti veebiportaal nimega The-Eye[6], kust on võimalik Redditi andmeid alla laadida alamredditite kaupa. Lisaks Redditile on keskkonnast leitavad ka

[4] https://praw.readthedocs.io/en/stable/index.html

[5] https://github.com/pushshift/api

[6] https://the-eye.eu/

Telegrami ja Twitteri postituste andmeid. Tegemist on leheküljega, mis on teinud r/Eesti alamredditist tõmmiseid (ingl *dump*) alates aastast 2010 kuni 2023. Lisaks alamredditite tõmmistele on saadaval ka mahult suured tõmmised, mis hõlmavad kogu Reddit-it mingil ajaperioodil. Kuna seal on terabaitide jagu andmeid, oleks nende töötlemiseks vaja eriressursse. Tõmmiste tegemiseks on kasutatud Redditi Pushshift API-t, mille kaudu saadud andmestikke on kasutatud ka teadustöös [22].

Platvormilt The-Eye saadud andmestikus ei ole postitustel küljes tunnused nagu poolthäälte hulk (ingl *upvotes*) ja vastuhäälte hulk (ingl *downvotes*). Selle asemel on välja toodud tunnus nimega skoor (ingl *score*), mis näitab postituse või kommentaari netohäälte arvu. Skoor arvutatakse poolthäälte hulga põhjal, millest on lahutatud vastuhäälte hulk. Siiski on poolthäälte hulk ja vastuhäälte hulk tunnustena kajastatud kommentaaridel.

3.2. Eesti keele ühendkorpus

Eestikeelsete Reddit-i andmete ühe võimaliku allikana võib vaadelda ka eesti keele ühendkorpust, mis on Kristina Koppeli ja Jelena Kallase teadusartikli [23] kohaselt mahukaim eestikeelsete digitekstide kogu.

Eesti Keele instituudi ning ettevõtte Lexical Computing Ltd. koostöös valmis ühendkorpuste nelja versiooniline sari [23]. Tegemist on aastatel 2013, 2017, 2019 ning 2021 tehtud versioonidega. Korpuses olevad tekstid on kogutud suuremas osas veebis olevatest tekstidest, mistõttu on võimalik vaadelda, kui palju r/Eesti postitusi ja kommentaare sisaldub eesti keele ühendkorpuse eri versioonides.

Igal postitusel ja kommentaaril on unikaalne identifikaator, mis kajastub ka andmetes välja toodud *permalink*'is või eesti keele ühendkorpuse kroolitud (ingl *crawl*) andmete lingis. Just identifikaatori kaudu on võimalik kokku viia käesoleva töö andmestiku postitused ja kommentaarid eesti ühendkorpuses esinevatega. Nagu tulemustest näha (Tabel 3), puuduvad ühendkorpuse veebikorpuse versioonidest 2013 ja 2017 alamredditi r/Eesti sisu ning ühendkorpuse versioonides 2019 ja 2021 on postitusi vastavalt 1480 ja 470 tükki ning kommentaare vastavalt 1056 ja 225 tükki.

Tabel 3. Eesti keele ühendkorpuse versioonides ja lõputöö andmestikus ühtivate postituste ja kommentaaride arvu tabel.

	Eesti keele ühendkorpus v. 2013	Eesti keele ühendkorpus v. 2017	Eesti keele ühendkorpus v. 2019	Eesti keele ühendkorpus v. 2021
Ühitivaid postitusi	0	0	1480	470
Ühtivaid kommentaare	0	0	1056	225

Tuues kõrvale võrdluseks, et The-Eye platvormilt saadud andmestikus on kokku 44 118 postitust ja 679 562 kommentaari. See tähendab, et ühendkorpus katab ~4% käesoleva andmestiku postitustest ning 0.2% andmestiku kommentaaridest. See näitab, kuivõrd vähe on ühendkorpuses kaetud Eesti suurim alamreddit ning annab aimu sellest, et kui on soov uurida eestikeelset Redditit või suurendada eestikeelsete sotsiaalmeedia tekstide arvu oma ülesande läbiviimsel, oleks mõistlik kasutada just selle töö käigus kokku pandud korpust. Samuti on The-Eye platvormilt saadud andmestikul küljes metaandmed, mis ühendkorpusel puhul on eemaldatud.

3.3. Õiguslik alus

The-Eye platvormilt võib leida kirje, kus väidetakse, et platvorm on DMCA nõuetele vastav [24]. DMCA (*Digital Millennium Copyright Act*) on Ameerika Ühendriikide autoriõiguse seadus, mis kaitseb autoriõigusega materjale internetis ning muudab autoriõigusega kaitstud materjali jagamise või avaldamise ilma omaniku loata ebaseaduslikuks [25]. See tähendab, platvorm The-Eye on andmed saanud seaduslikul viisil.

Andmestik, mis on The-Eye platvormilt saadud, ei sisalda kasutajate nimesid, isikukoode, sünniaega või muud sellist, mis otseselt viitaks kindlale isikule. Siiski võib kasutajanimi olla mõne kasutaja puhul selline, mille kaudu võib olla isik tuvastatav. Üldiselt kasutavad enamus Redditi kasutajaid kasutajanime, mis ei anna aimu nende isikust, st. ei ole seotud kasutaja isikliku nimega. Siiski tuleb andmete kasutamisel hinnata töö õiguslikku alust.

Tartu Ülikooli andmekaitse juhendi „Andmekaitse teadustöös“ [26] järgi on teadustöö jaoks olemas erinevaid õiguslikke aluseid, aga selle töö puhul rakendub õigusliku alusena

õigustatud huvi, kus andmesubjektide huvide ja õiguste riive on võimalikult väike, kuna kasutatakse depersonaliseeritud analüüsi. Siiski tuleks uurida ka isikuandmete kaitse seaduses sätestatud tingimusi.

Isikuandmete kaitse seaduse paragrahvis 6 [27] tuuakse välja, et isikuandmeid võib andmesubjekti nõusolekuta teadusuuringu või riikliku statistika vajadusteks töödelda pseudonüümitud või samaväärset andmekaitse taset võimaldaval kujul. Selle töö analüüsi osa on tehtud depersonaliseeritud andmetega, st. andmetega, millel pole juures isikunimesid, kasutajanimesid ja muud sellist, mille põhjal oleks võimalik määrata mõni konkreetne isik. See tähendab, et selles kontekstis ei ole tarvis rakendada eraldi meetmeid isikuandmete kaitseks, sest isikuandmeid ei avalikustata ning analüüsi ei teostata üksikisiku tasemel.

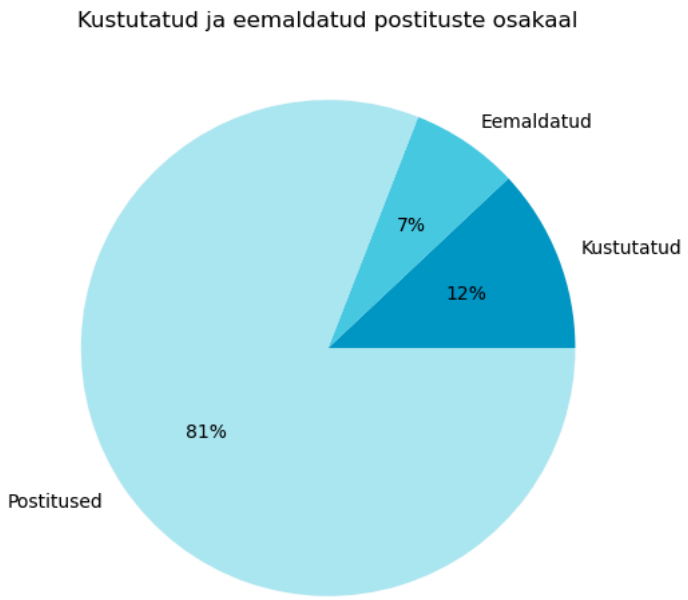
3.4. Andmete töötlemine

The-Eye platvormilt saadud andmete puhul olid postitused ja kommentaarid eraldi failidena ning mõlemad tihendatult .json failiformaadis. Esmalt tuli kommentaarid viia kokku postitustega, et saada parem ülevaade, milliseid postitusi kuidas on kommenteeritud. Kõikidele postitustele ning kommentaaridele on antud unikaalne identifikaator, mille põhjal on võimalik postitust või kommentaari hõlpsasti leida. Kommentaaride juures on tunnusena välja toodud ka kommentaari ülese postituse identifikaator. Postitused koos kommentaaridega sai viidud kujule, kus iga postituse küljes on selle kommentaarid. Iga postitus selle metaandmete ning koos kommentaaridega kirjutati eraldi .json faili.

Postitused ning nende kommentaarid on selles andmestikus kaetud ajavahemikus 2010 kuni 2023 jaanuar. 2023. aasta jaanuari puhul oli näha, et tõmmis sisaldas vaid 4 postitust, mistõttu ei võetud neid andmeid analüüsimisel arvesse, kuna andmeid võrreldakse peamiselt aastate lõikes, mistõttu võivad inimesed valesti interpreteerida trendide järske muutusi aastal 2023. Seetõttu käsitletakse edaspidiselt kogu analüüsi osas andmestikku ajavahemikul 2010 kuni 2022 (kaasa arvatud). Sellele vaatamata jäävad 2023. aasta andmed alles loodud korpusesse.

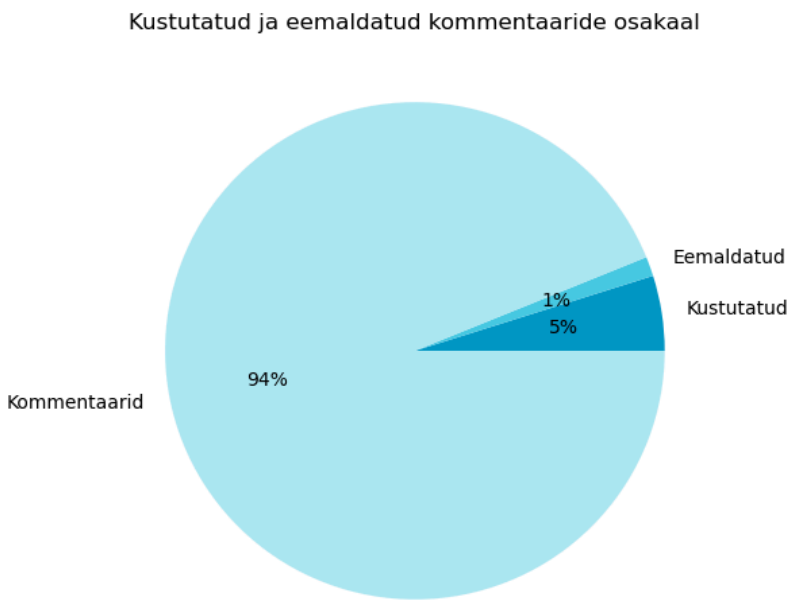
Nagu eelnevalt on välja toodud, võib moderaator postituse või kommentaari eemaldada juhul, kui see ei ole kooskõlas alamredditi reeglitega. Lisaks saab seda teha ka Reddit, kui see rikub Redditi enda reegleid. Samuti on võimalus, et kasutaja kustutab enda postituse ise. Mõlemal juhul säilivad postituse metaandmed, kuid postituse sisu ei kuvata, vaid selle asemel on postituse sisu ning pealkiri asendatud vastavalt siltidega „[removed]“ või „[deleted]“.

Joonisel 1 on välja toodud, kui suur hulk postitusi on eemaldatud või kustutatud. Eemaldatud postitusi on 7% protsenti, mis annab indikatsiooni, et 7% postitustest ei ole vastanud alamredditi r/Eesti reeglitele - ehk kas postitused on olnud ebasobiva sisuga või on tegemist olnud spämmpostitustega. 12% andmestikus olevatest postitustest on kustutatud kasutaja poolt või on kasutaja kustutanud enda kasutaja, mille tagajärjel ka tema postituste sisu kustub.



Joonis 1. Kustutatud ja eemaldatud postituste osakaal andmestikus.

Sarnaselt postitustele on joonisel 2 välja toodud, kui suur hulk kommentaare on eemaldatud või kustutatud. Eemaldatud kommentaare on 1% protsent, mis indikeerib kommentaaride osas paremat alamredditi reeglitele vastavust. 5% andmestikus olevatest kommentaaridest on kustutatud kasutaja poolt või on kasutaja kustutanud enda kasutaja, mille tagajärjel ka tema postituste sisu kustub.



Joonis 2. Kustutatud ja eemaldatud kommentaaride osakaal andmestikus.

Tulemused nii postituste ja kommentaaride puhul näitavad, et eemaldatud ja kustutatud postitusi on protsentuaalselt rohkem kui eemaldatud ja kustutatud kommentaare. Seetõttu, et andmestikus on kommentaare palju rohkem, võib see anda erinevate osakaalude olemuse kohta kaks võimalikku indikatsiooni:

1. Kommentaariumis esineb vähem ebasobivaid postitusi, st. kommentaariumis kommenteerivad kasutajad peavad ennast paremini üleval.
2. Moderaatoril pole head ülevaadet kommentaaride kohta. See tähendab, et ebasobivaid postitusi märgatakse, kuid ebasobivad kommentaarid võivad jääda tähelepanuta. See võib tuleneda sellest, et postitused on rohkem nähtavamal kui kommentaarid, mistõttu on postituste modereerimine kõrgem prioriteet. Samuti on postitusi palju vähem kui kommentaare, mistõttu ei pruugi kommentaaride modereerimine olla jõukohane.

Olles eelnevalt tutvunud andmestikuga, on autoril tekkinud subjektiivne ülevaade postituste ja kommentaaride sisudest, mistõttu töö autor arvab, et 2. variant võib olla tõenäolisem.

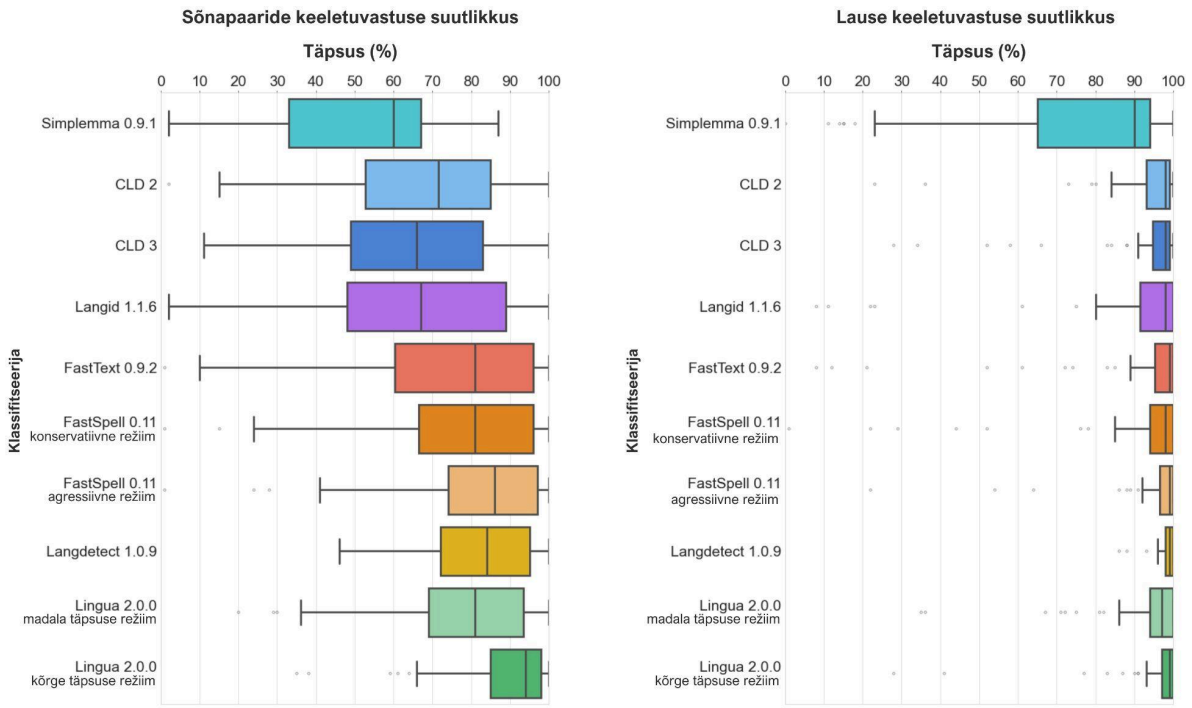
4. Meetodid

Selles töö peatükis kirjeldatakse töö jooksul kasutatud meetodeid, nendega seotud murekohti ning kaalutud alternatiivseid lahendusi. Peamiselt kasutatavateks meetoditeks on keeletuvastuse mudel Lingua, tundmusanalüüsi jaoks kasutatud ja eelnevalt peenhäälestatud Est-RoBERTa mudel ning teemade analüüsi jaoks kasutatud BERTopic.

4.1. Keeletuvastus

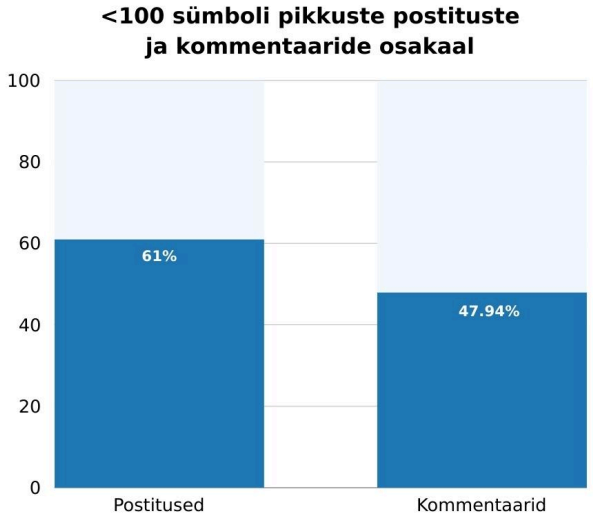
Selles töös kasutati postituste ja kommentaaride keele tuvastamiseks Pythoni teeki nimega Lingua. Teksti keelelisuse määramiseks on olemas mitmeid erinevaid lahendusi, näiteks Google-i CLD 2[7] and CLD 3[8], langid[9], fastText[10] ja langdetect[11]. Algselt rakendas töö autor töö käigus langdetecti, kuid tulemused tulid küllaltki mürarohked. Lingua dokumentatsiooni [6] järgi esineb eelnevalt mainitud lahendustel (välja arvatud langdetect-il) suur puudus, mis seisneb selles, et need lahendused suudavad vaid pikemate tekstide puhul keelt täpsemalt määrata (joonis 3). Samas pikemate tekstide puhul ei ole erinevate lähenemiste juures täpsuse vahe niivõrd märgatav, kuigi Lingua täpsus ületab siiski allika väitel [6] teiste mudelite täpsuse.

[7] https://pypi.org/project/pycld2/
[8] https://pypi.org/project/pycld3/
[9] https://pypi.org/project/langid/
[10] https://pypi.org/project/fasttext-langdetect/
[11] https://pypi.org/project/langdetect/



Joonis 3. Erinevate mudelite suutlikkus sõnapaaride- ja lausete keeletuvastusel. [6]

Paraku on sõnapaaride tuvastamise täpsus suure tähtsusega just sotsiaalmeedia postituste ja kommentaaride puhul, kus postituse või kommentaari pikkus võib olla väga lühike. Selle magistritöö raames loodud andestikus olevate alla 100 sümboli pikkuste postituste osakaal moodustab 61% kõikidest postitustest ning kommentaaride osakaal 47.94% kõikidest kommentaaridest (joonis 4). See näitab, et ka suur osa selle töö andmestikust koosneb lühikestest postitustest ja kommentaaridest.



Joonis 4. <100 sümboli pikkuste postituste ja kommentaaride osakaal.

Lingua toob lisaks veel välja, et mida rohkem keeli on kaasatud keelelisuse tuvastamise protsessi, seda ebatäpsem on tulemus. See väljendus ka r/Eesti andmete peal, kus esines olukordi, mil Lingua määras postituse keeleks tagalogi keele[12] (lühend TL), kuigi postitus oli peamiselt eestikeelne, kuid postituses oli toimunud koodivahetus. Just sellel põhjusel vaatas töö autor käsitsi läbi kõik keeled, mis lisaks inglise ja eesti keelele olid märgendatud. See andis võimaluse kontrollida, millised keeled reaalselt andmestikus eksisteerivad ning need keeled Linguale ette anda, et vaid nende keelte põhjal märgistusel valik langetada.

Lingua puhul on võimalik keeletuvastuse puhul jälgida ka usaldusskoori (inglise keeles *confidence score*), mida kasutades määrati antud töös usaldusskoori nivoo, mille ületamisel määratakse tekstile Lingua poolt pakutud keel ning vastasel juhul märgitakse keeleks „unknown“. Eksperimenteerimise käigus võeti juhuslik hulk märgistatud tekste ning vaadeldi erinevatele tekstidele pakutud usaldusskoori ja teksti sisu ning määrati usalduskoori nivooks 0.69. Sellise skoori puhul tekib rohkem „unknown“ märgendusega postitusi ja kommentaare, kuid selle eest esineb märgendatud keelte puhul vähem eksimusi, mis on oluline just tundmus- ja teemade analüüsi jaoks.

4.2. Tundmusanalüüs

Tundmusanalüüs (ingl *sentiment analysis*) on tekstianalüüsimise protsess, mille abil tuvastatakse teksti emotsionaalne toon. Selle töö puhul on võimalikeks emotsionaalset tooni määravateks siltideks positiivne, neutraalne ja negatiivne. Tundmusanalüüsi tulemusi on võimalik kasutada, et määrata huvigrupi tundmus teema või teksti suhtes. [28]

Käesoleva magistritöö tulemuste osas on täpsemalt kajastatud eelneva keeletuvastuse alampeatüki tulemused, kuid etteruttavalt võib öelda, et antud andmestikku saame käsitleda, kui kakskeelset andmestikku. See tähendab, et tundmusanalüüsi läbi viimiseks tuleb kasutada erinevate keelte mudeleid.

Matej Ulčari ja teiste poolt kirjutatud artiklis [29] tuuakse välja, et EstBERT [16] on treenitud 1.1 miljardi sõna suurusel andmestikul, kuid Est-RoBERTa 2.51 miljardi sõna suurusel andmestikul.

Mudeli peenhäälestamiseks kasutas autor Valentsikorpust[13] [17]. Tegemist on korpusega, mis põhineb Postimehe artiklitel. Antud juhul on artiklite meelsused käsitsi määratud. Just mudeli

[12] https://en.wikipedia.org/wiki/Tagalog_language
[13] http://peeter.eki.ee:5000/valence/paragraphsquery

treenimise seisukohast on oluline kasutada korpust, mille tundmuse sildid on inimeste poolt annoteeritud, sest nii saame me veenduda, et mudelit treenitakse korrektsete andmetega.

Paraku on Valentsikorpus ainuke eestikeelne tundmuse märgistustega korpus, mis on vabalt kättesaadav, mistõttu on see hetkel parim korpus, mida mudeli peenhäälestamiseks kasutada. Kuna Valentsikorpuse näol on tegemist ajaleheartiklitega, ei ole see teoreetiliselt kõige parem andmestik peenhäälestuseks, sest ajaleheartiklite ning sotsiaalmeedia tekstid on oma olemuselt erinevad. Kui ajaleheartiklid on peamiselt suhteliselt pikad ning korrektse keelekasutusega, on vastukaaluks sotsiaalmeedia postitused ja kommentaarid väga mürarikkad, st. ebakorrektse keelekasutuse ja slängiga, mis võib ajada mudeli segadusse.

Autori poolt Valentsikorpusel peenhäälestatud Est-RoBERTa mudel saavutas F1 skooriks 78.22%. Sellele eelnevalt eemaldati Valentsikorpuse andmestikust kirjed sildiga „vastuoluline“, peale mida oli andmestik algse 4088 elemendi asemel 3536 elemendi suurune. Testhulga suuruseks oli 20% andmestikust, mis oli valitud juhuslikult üle andmestiku, ülejäänud 80% andmestikust kasutati treenimisel.

Inglise keele jaoks proovis töö autor erinevaid mudeleid ning test-andmetel kõrgeima F1 skoori sai mudel „cardiffnlp/twitter-roberta-base-sentiment[14]“, mille F1 skooriks oli 0.719. Testandmeteks oli TweetEval [30] andmestik. Tegemist on roBERTa-base mudeliga, mis on treenitud ~58 miljoni Twitteri säutsuga ning peenhäälestatud tundmusanalüüsi jaoks. Paraku sarnast mudelit Redditi andmestiku jaoks loodud pole, kuid mõlemal juhul on tegemist sotsiaalmeedia andmestikega, kus olulisel kohal on teksti pikkus. Siiski valis autor test-andmetel F1 skoori 0.715 saanud „cardiffnlp/twitter-roberta-base-sentiment-latest[15]“ mudeli, sest tegemist on sama, kuid alles hiljuti uuendatud mudeliga, mistõttu on lootust, et see mudel suudab Redditi andmestikul parema sooritusse teha. Täiendavalt võrdles autor ka valitud RoBERTa mudelit Vader-iga [9], võttes kahe mudeli poolt erineva tundmuse sildiga postituste hulgast välja 100 juhuslikku postitust ning hindas käsitsi nende tundmuse siltide õigsust. Mudelite juhuslike erinevuste hindamisel selgus, et RoBERTa mudel edestas Vader-it, kus umbes 85% juhtudest ennustas RoBERTa mudel postitusele korrektse tundmuse sildi ning vaid umbes 10% juhtudest oli ennustas Vader-i mudel postitusele korrektse tundmuse sildi. Ülejäänud umbes 5% protsendi ennustuste puhul eksisid mõlemad mudelid.

[14] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
[15] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

4.3. Teemade analüüs

Teemade modelleerimiseks on olemas mitmeid erinevaid lähenemisi, näiteks LDA, NMF, BERTopic ja Top2Vec [32]. Käesoleva magistritöö käigus kasutati teemade analüüsiks BERTopicut. Tegemist on siirdeõppe mudelil põhineva teemade modelleerimise tehnikaga, mille puhul teisendatakse alusdokumendid numbrilistele esitluskujudele.

Vaikimisi kasutab BERTopic enda *multilingual* mudelit, mis toetab enam kui 50 erinevat keelt [14]. Siiski on BERTopicu puhul võimalik välja vahetada BERTopicu algne mudel, mistõttu töö autor katsetas teemade modelleerimiseks ka estroBERTa mudelit. Paraku, ei andnud see häid tulemusi ning klastrid olid mürased.

Klasterdamise mudeliks kasutati selle töö puhul HDBSCAN algoritmi, mis loob piiramata klastrite hulga tõttu teemadele täiendavalt ka arvukalt äärejuhte. Selle andmestiku puhul moodustas äärejuhtude hulk olenemata HDBSCAN-i seadistustest umbes poole BERTopicu sisendandmestikust. Seetõttu tuli täiendavalt rakendada BERTopicule sisse ehitatud teemade vähendamise meetodit. Teemade vähendamise jaoks kasutati selle töö puhul teemade analüüsil käigus iga teksti kohta leitud kõigisse teemadesse kuulumise tõenäosused ning äärejuhtude läbi käimisel omistati neile teema, mille tõenäosus ületas etteantud lävendit (ingl. *threshold*) ning oli neist suurim [32]. Kuna r/Eesti andmestiku puhul on tegemist kitsamat temaatilist piiritletust mitte omava alamredditiga, on seal alamredditis arutatavate teemade hulk suur, mistõttu võib piirjuhtude hulga vähendamine miinimumini tuua kaasa mürasemad klastrid ning müra olemasolul võib omakorda häirida sõnaesitusi, mis teemasid kirjeldavad. Seetõttu ei piirdunud töö autor teemade nimetamisel vaid BERTopicu toodud sõnaesitustega, vaid vaatas ka teemadesse jaotatud tekste.

4.4. Lisanduvad tööriistad

Selle magistritöö kirjutamisel kasutati OpenAI poolt arendatud ChatGPT[16] versiooni 4.0, et tõlkida töö lühikokkuvõtet. Samuti kasutati ChatGPT-d magistritöö koodiosa kirjutamisel, et kiiresti leida erinevate meetodite kasutusvõimalusi ning tuvastada ja parandada tekkinud süntaksivigu.

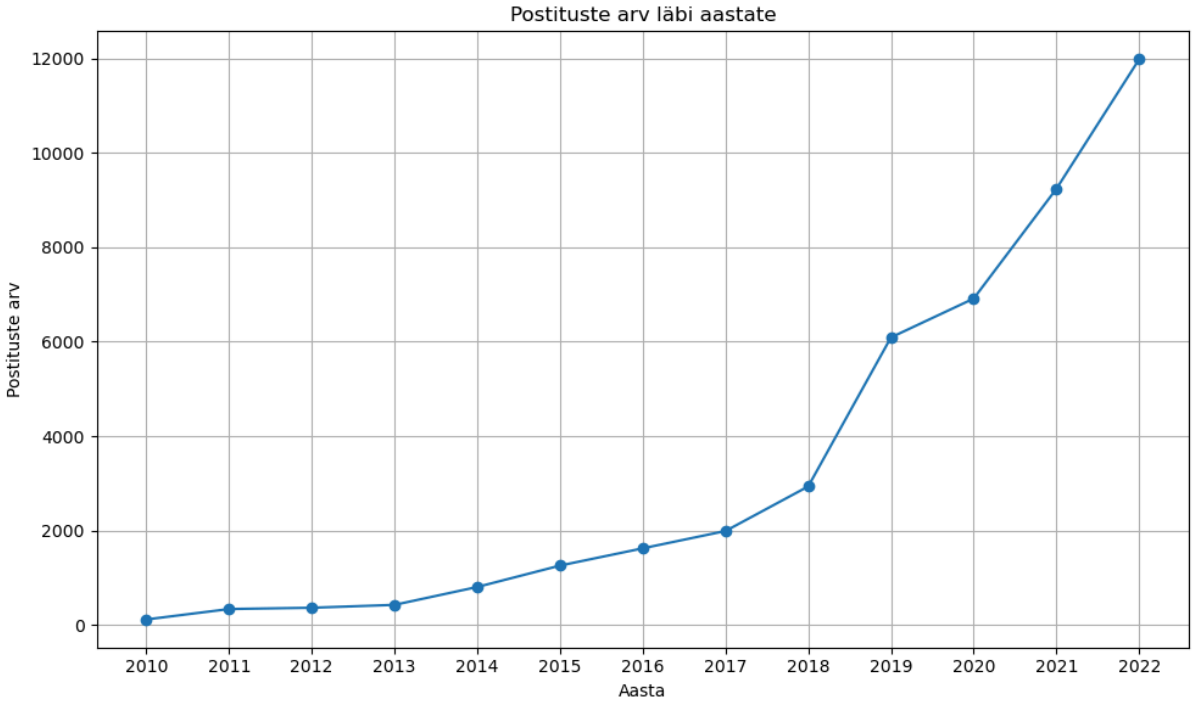
[16] https://chat.openai.com/

5. Andmestiku analüüs ja tulemused

Selles peatükis tuuakse välja töö tulemused, mis hõlmavad endas andmestiku deskriptiivset statistikat, keeletuvastuse, tundmusanalüüsi ja teemade analüüsi tulemusi. Selle magistritöö Jupyter Notebook töövihiku failid on üles laetud autori Github-i repositooriumisse[17].

5.1. Andmete kirjeldav analüüs

Joonis 5 näitab r/Eesti alamredditi andmestikus olevate postituste arvu aastate lõikes. Aastas tehtud postituste arv kasvab aastate lõikes eksponentsiaalselt, mis võib aimu anda kasvavast kasutajate arvu hulgast või r/Eesti alamredditi kasutajate aktiivsuse tõusust.

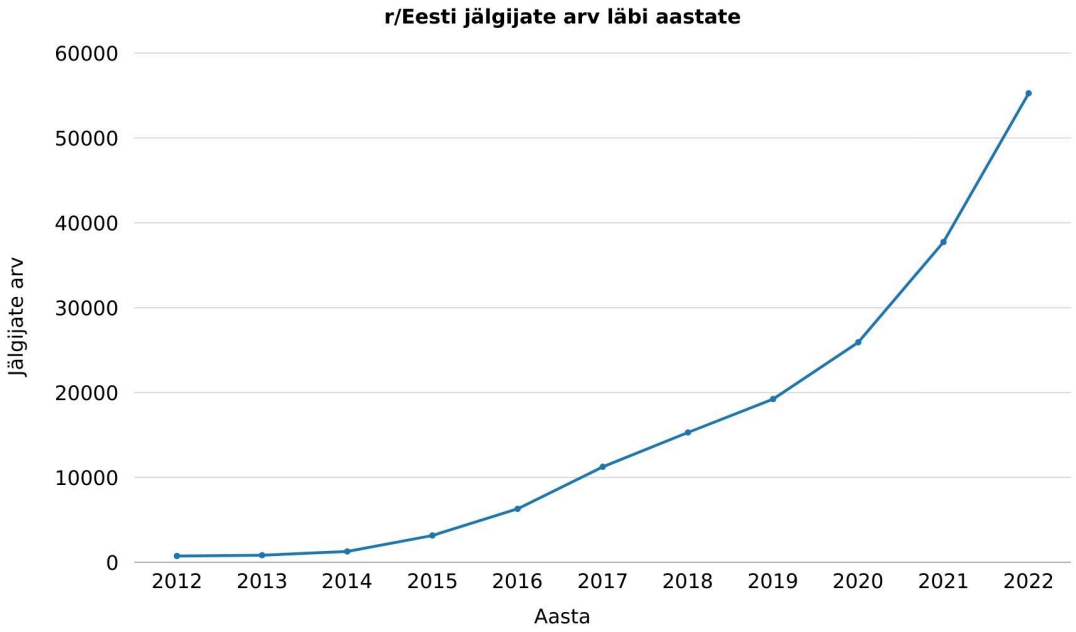


Joonis 5. Postituste arv läbi aastate.

Et seda paremini hinnata kasutati lehelt Subreddit Stats [33] saadud andmeid, mis kirjeldavad r/Eesti alamredditi kasutajate arvu aastate lõikes. Eeltoodud allikas on andmed välja toodud alates 2012. aasta oktoobrist kuni 2023. aasta detsembrini, mistõttu eemaldati andmetest

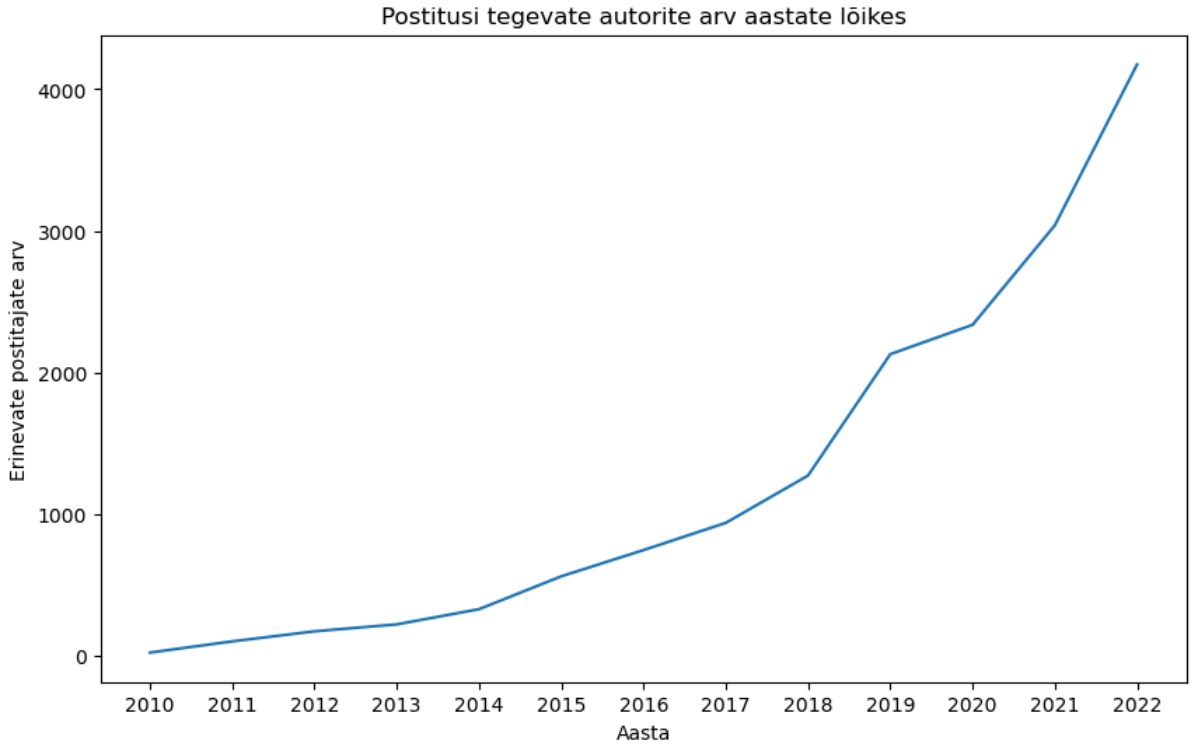
[17] https://github.com/taunotamm/DS_MSc_Thesis

2023. aasta kirjed, et ajavahemikud oleksid võrreldavad käesolevas töös uuritavatega. Joonisel 6 on kujutatud r/Eesti alamredditit jälgivate kasutajate arvu aastate lõikes.



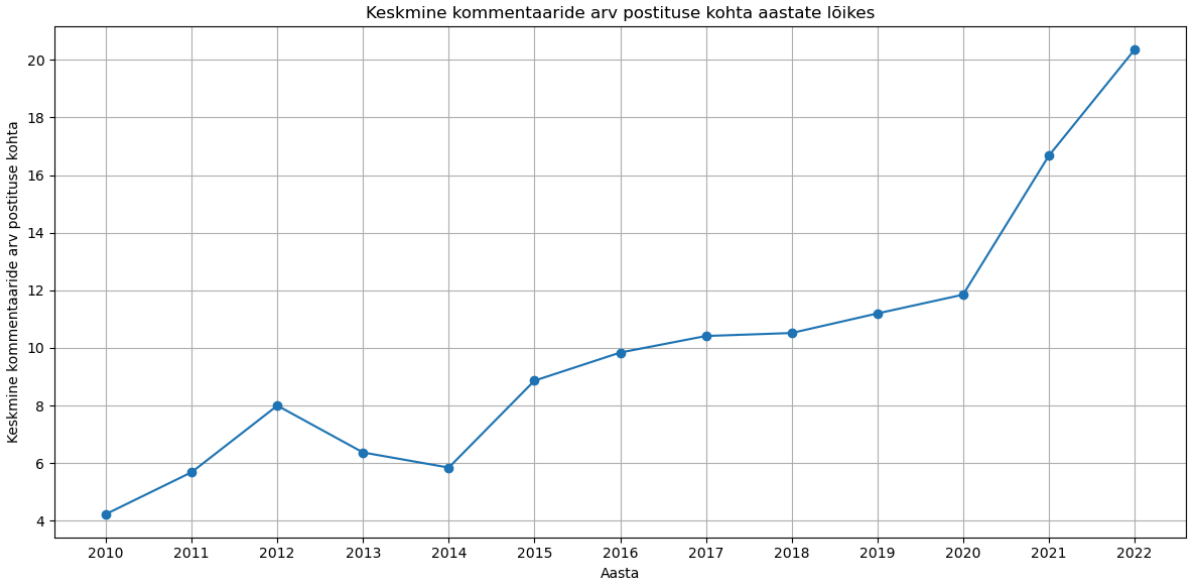
Joonis 6. r/Eesti jälgijate arv läbi aastate. [33]

Sarnaselt postituste arvu tõusule on tõusnud aastate lõikes ka r/Eesti alamredditit jälgivate kasutajate arv ekponentsiaalselt. Sellest võib järeldada, et r/Eesti alamredditi postituste kasvu taga võib olla nii kasutajate aktiivsuse kasv kui ka r/Eesti populaarsuse ning kasutajate arvu tõus, mille tulemusena aastast aastasse teevad üha rohkem kasutajaid postitusi (joonis 7).



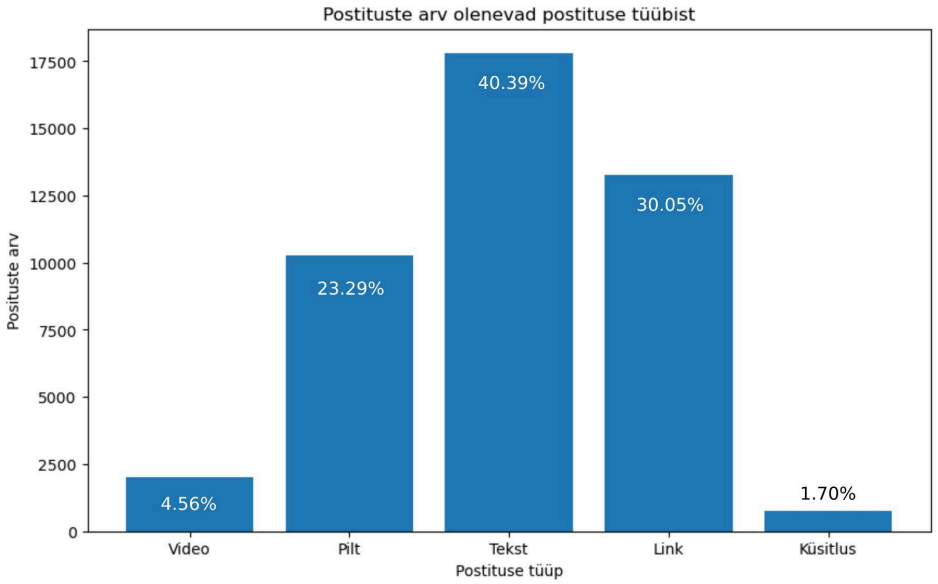
Joonis 7. Postitusi tegevate autorite arv aastate lõikes.

Keskmine kommentaaride arv postituse kohta aastate lõikes on välja toodud joonisel 8. Jooniselt on näha, et kui üldiselt ühtib trend postituste arvu (joonis 5) ja postitusi tegevate autorite arvu (joonis 7) trendiga, siis ajavahemikul 2012-2014 langes keskmiste kommentaaride arv postituse kohta, samas kui postituste arvu ja postitusi tegevate autorite arvu trend oli samal ajal kerges tõusus.



Joonis 8. Keskmine kommentaaride arv postituse kohta aastate lõikes.

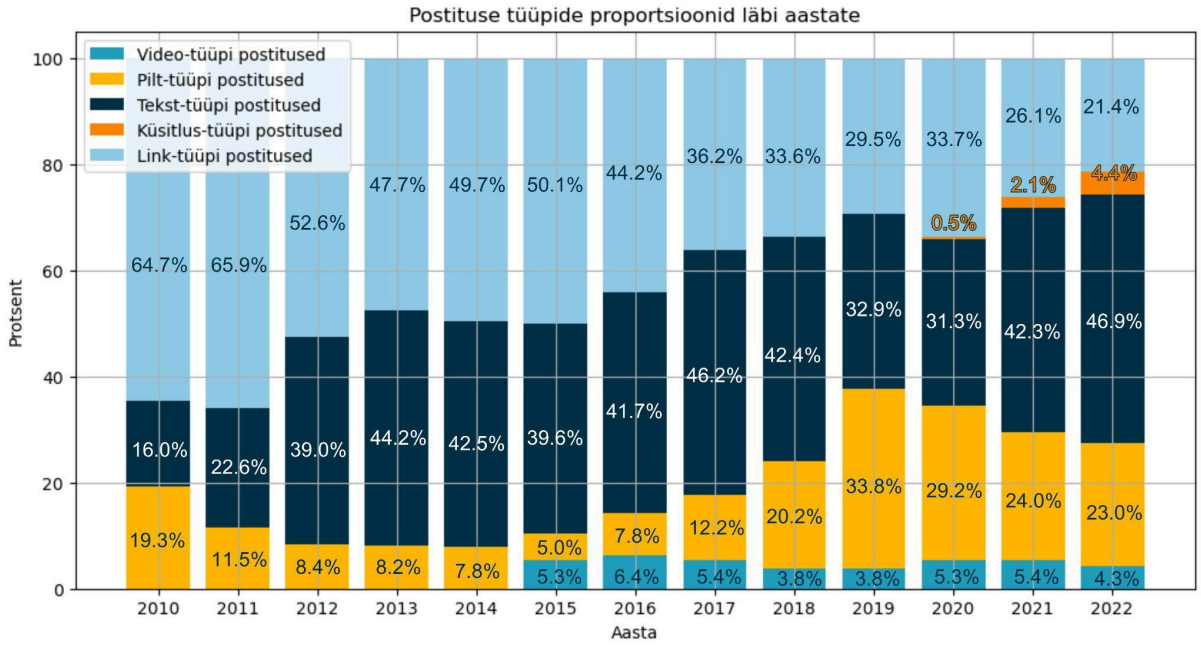
Joonisel 9 on välja toodud postituste arv tulenevalt postituse tüübist üle kogu andmestiku. Jooniselt on näha, et kõige populaarsemaks postituse tüübiks on tekst-tüüpi postitus, millele järgneb link-tüüpi postitus ning seejärel pildipostitus. Võrreldes teiste postituste tüüpidega on videote postitamine ja küsitluste korraldamine tagaplaanil. 2020. aasta alguses tuli Reddit välja uue postituse tüübiga, milleks oli küsitlus [34]. Sellepärast on küsitlusega postitusi vähe, et tegemist on võrdlemisi hiljuti loodud postituse tüübiga.



Joonis 9. Postituste arv olenevalt postituse tüübist.

Vaatame lähemalt, kuidas erinevaid postitusi on läbi aja alamredditis r/Eesti tehtud. Joonisel 10 kuvatakse, millised on erinevate postituste tüüpide proportsioonid aastate lõikes. Jooniselt on hästi näha, kuidas ajaraami alguses tehtid link-tüüpi postitusi kõige rohkem (umbes 65%) ning läbi aastate seda tüüpi postituste osakaal väheneb. Pilt-tüüpi postituste osakaal oli aastal 2010 umbes 20% ning kuni aastani 2015, saavutades ajaraami kõige madalama pilt-tüüpi postituste osakaalu (umbes 5%). Peale seda hakkas pilt-tüüpi postituste osakaal tõusma kuni aastani 2019, saavutades osakaaluks pisut üle 30%, ning hakates seejärel taaskord langema. Video-tüüpi postituste puhul on näha, et nende proportsioon on läbi aastate (2015-2022) jäänud küllaltki samaks, olles õrnalt langevas trendis läbi aastate. Aastatel 2020 kuni 2022 on näha küsitlus-tüüpi postituste osakaalu kasvu. Tekst-tüüpi postituste proportsiooni vaadates on märgata, et kõige väiksema proportsiooniga olid tekst-tüüpi postitused aastal 2010 (umbes 15%), hakates seejärel tõusma kuni aastani 2013, mil tekst-tüüpi postitused moodustasid üle

44% kõikidest postitustest. Peale seda kuni aastani 2020 tekst-tüüpi postituste osakaal langes, mil tekst-tüüpi postitused moodustasid vaid ligikaudu 31% kõikidest postitustest ning hakates seejärel taaskord tõusma, saavutades läbi aegade kõige suurema proportsiooni aastaks 2022, moodustades umbes 47% kõikidest postitustest.



Joonis 10. Postituste tüüpide proportsioonid läbi aastate.

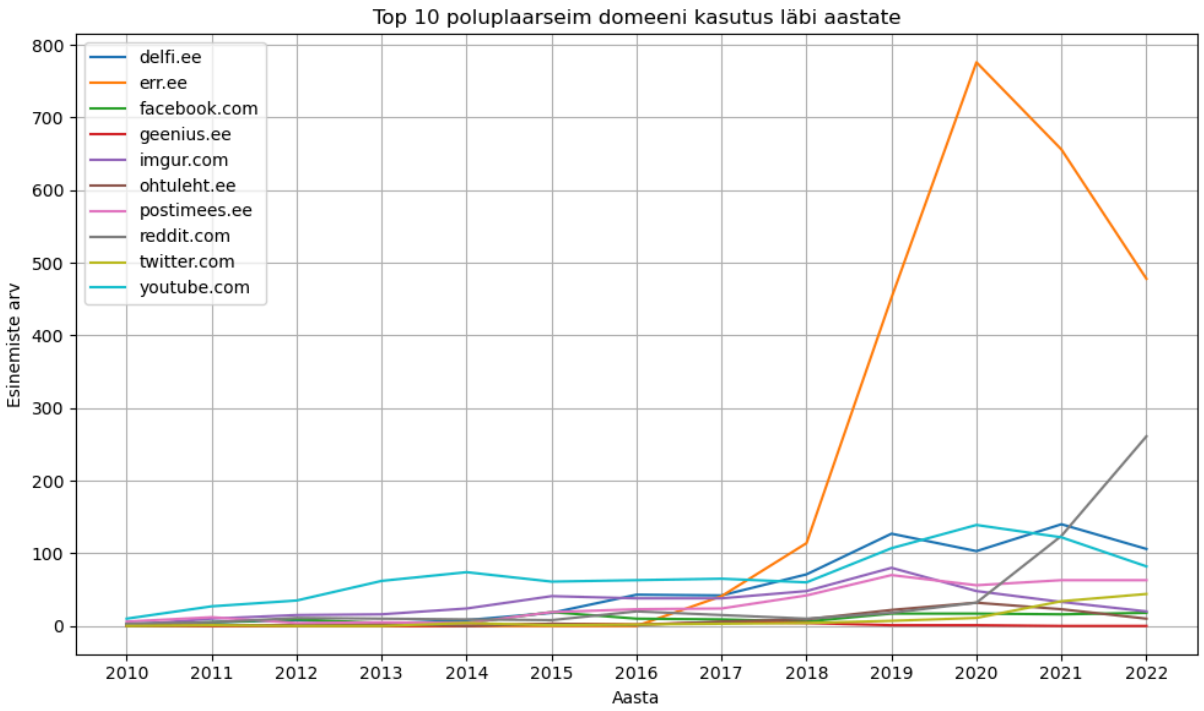
Vastavalt uuriti ka link-tüüpi postitusi ning linke, kuhu need suunavad. Tabelis 5 on välja toodud 10 kõige enam jagatud lingi peadomeeni. Vastavalt linkide struktuurile eemaldati eelnevalt linkidel alamdomeenid ning teekonnad, et kaardistada paremini alamredditis jagatud domeene. Igal domeenil võib olla kuni 500 alamdomeeni [35], mistõttu eemaldati need müra vähendamiseks. Tulemustest on näha, et 10 kõige populaarsema peadomeeni hulka kuuluvad peamiselt uudiste- ja sotsiaalmeedia portaalid. Samuti on kõige populaarsemate peadomeenide hulgast leitav ka videote striimimise platvorm YouTube ning piltide majutamise (ingl. *hosting*) keskkond Imgur. Kuna tegemist on Eesti kogukonna alamredditiga, on oodatav tulemus, et kõige populaarsemateks uudisteportaalideks on eesti päritolu uudisteportaalid. Küll aga on üllatav, et tugeval esikohal on err.ee, mille esinemiste arv on pisut vähem kui kaks korda rohkem kui domeenil delfi.ee, mis on populaarsuselt teisel kohal oleva domeen. Domeeni err.ee puhul on tegemist Eesti Rahvusringhäälinguga, mis on

Eesti riiklik uudiste, raadio- ja telesaateid jagav organisatsioon, kelle uudised on faktiliselt korrektsed ning kallutamata.

Tabel 5. Top 10 link-tüüp postituse peadomeeni.

Peadomeen	Esinemiste arv
err.ee	3144
delfi.ee	1797
postimees.ee	1159
youtube.com	925
reddit.com	835
imgur.com	455
geenius.ee	200
facebook.com	154
ohtuleht.ee	138
twitter.com	116

Joonisel 11 on välja toodud ka 10 kõige populaarsema domeeni esinemiste sagedus läbi aja. Jooniselt on näha, et YouTube on olnud 2010-2017 kõige rohkem linkides esinevaks domeeniks, kuid alates 2018. aastast hakkab tugevalt domineerima ERR, mis teeb esinemissagedusega tipu 2020. aastal, kuid hakates seejärel taas vaibuma, olles siiski kõige populaarsem domeen. ERR-i kõrget populaarsust koroona-aastal 2020 võiks oletuslikult seletada see, et inimesed pöördusid kriisisituatsioonis riigimeedia poole, mis oli kõige usaldusväärsem allikas.



Joonis 11. Top 10 peadomeeni kasutamine aastate lõikes link-tüüpi postitustes.

Kokkuvõtvalt tuleb seni näidatud otsese analüüsi tulemustest välja, et läbi aastate on r/Eesti populaarsus kasutajate seas kasvanud, mis väljendub postituste, kasutajate arvu, keskmise poolthäälte hulga ning keskmiselt postituse kohta tehtavate kommentaaride arvu kasvus. Postitatakse enamasti teksti-tüüpi postitusi, pisut harvemal juhul link-tüüpi või pilt-tüüpi küsimusi. Link-tüüpi postituste puhul on näha, et peamiselt jagatakse linke, mis kuuluvad domeeni err.ee alla, populaarsuselt teisel kohal on delfi.ee ning kolmandal kohal postimees.ee. See annab aimu, et link-tüüpi postituste puhul jagavad kasutajad edasi peamiselt uudiseid.

Lisaks uuriti ka postitamiste arvu nädalapäevade lõikes (Lisa I) ning ainus muster, mida täheldati, oli madalam produktiivsus nädalavahetusel. Täiendavalt vaadati ka postitustamiste arvu ööpäeva lõikes (Lisa II). Suurim postitamise aktiivsus on päevasel ajal ning madalaim öösel, mis on ootuspärane tulemus.

Magistritöö autor uuris samuti keskmist skooride jagunemist läbi aastate (Lisa III), mis sisuliselt peegeldab postituste arvu tulemusi läbi aastate (joonis 5), st. postituste arv läbi aastate ja keskmine skoori jagunemine läbi aastate astuvad ühte sammu.

5.2. Keeletuvastus

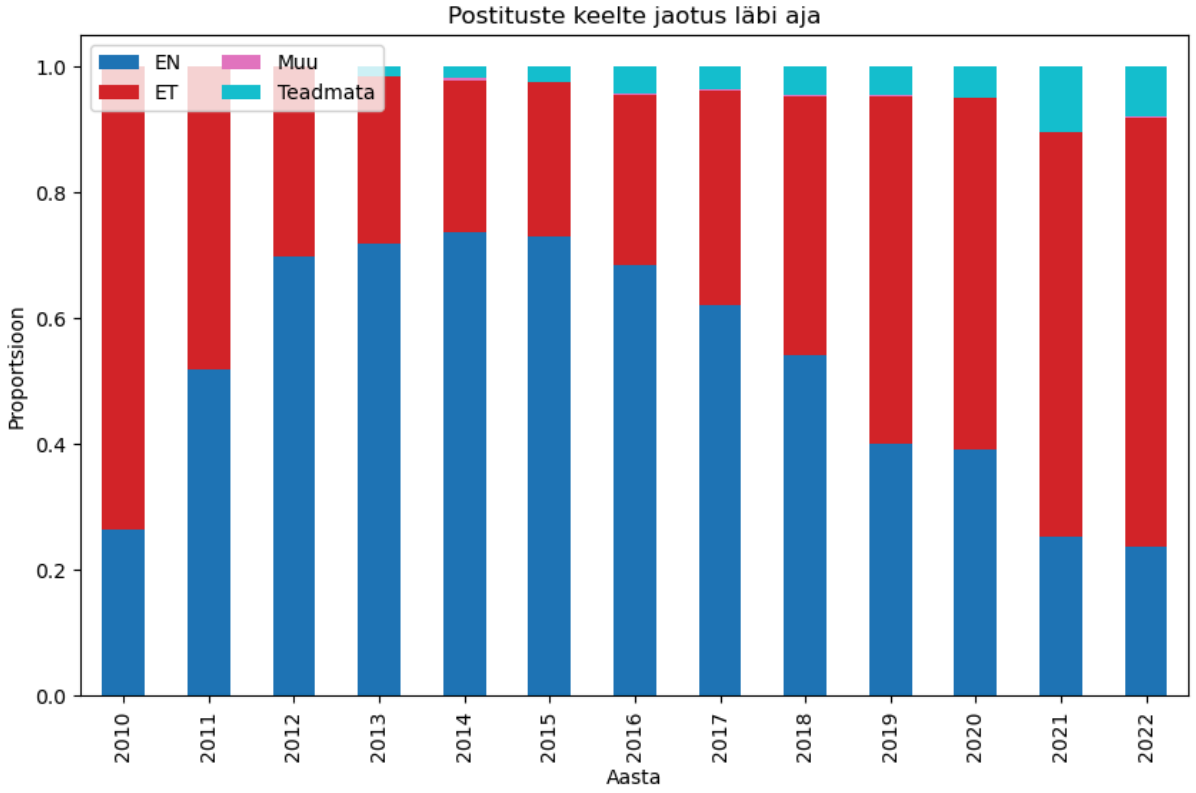
5.2.1. Postitused

Tabelis 6 on välja toodud r/Eesti alamredditis olevate keelte jagunemine, mille tulemuste põhjal saab öelda, et tegemist on kakskeelse alamredditiga. Kõikidest tekst-tüüpi postitustest on ~54% eestikeelsed postitused, ~36% on ingliskeelsed postitused, ~10% protsendi postituste puhul jäi keel määramata ning 0.152% moodustasid muud keeled. Muude keelte hulka kuulusid postitused, mis olid tehtud norra-, vene-, korea-, läti-, pärsia-, prantsuse-, ungari-, türgi, rootsi- ja hispaania keeles. Tundmatu keele märgenduse saanud tekst-tüüpi postituste hulgas on ka soome keelseid tekste, kuid tulenevalt eesti- ja soome keele sarnasusest hakkas Lingua neid segi ajama, mistõttu tuli tõsta algoritmi usaldusskoori taset, et vältida soomekeelsete tekstide sattumist eestikeelsete tekstide hulka ning vastupidi. Usaldusskoori nivoo vähenedes väheneb ka tundmatu keele liigituse saanud postituste arv, kuid sellisel juhul kasvab sellega ka vale märgendi saanud postituste arv, seega tekib käesolevatesse andmetesse rohkem müra, mis hakkab häirima tundmusanalüüsi ning teemaanalüüsi mudelite tööd ja tulemusi.

Tabel 6. Postituste keeleline jaotus.

ISO 639_1 keele kood	ISO keele nimi	Postituste arv	Osakaal
ET	eesti	9576	53.76%
EN	inglise	6369	35.75%
unknown	teadmata	1841	10.34%
FR	prantsuse	7	0.039%
RU	vene	5	0.028%
SV	rootsi	4	0.022%
LV	läti	3	0.017%
NB	norra	2	0.011%
ES	hispaania	2	0.011%
KO	korea	1	0.006%
FA	pärsia	1	0.006%
HU	ungari	1	0.006%
TR	türgi	1	0.006%

Joonisel 12 on välja toodud tekst-tüüpi postituste keelelisuse jaotuse muutus läbi aja. Teadmata keelte ning muude keelte hulk läbi aastate on läbivalt võrdlemisi väike. Aastate lõikes on näha, kuidas 2010. aastal eestikeelsete postituste hulk moodustab kõikide tekst-tüüpi postituste hulgast üle 70% ning ingliskeelsete tekst-tüüpi postituste arv moodustab pisut üle 25%-i. Kuni aastani 2014. väheneb eestikeelsete tekst-tüüpi postituste osakaal umbes 25%-ni ning ingliskeelsete tekst-tüüpi postituste osakaal tõuseb umbes 72%-ni. Järgnevatel aastatel tõuseb eestikeelsete tekst-tüüpi postituste osakaal, jõudes 2022. aastaks umbes 63%-ni ning ingliskeelsete tekst-tüüpi postituste arv taandub umbes 25%-ni.



Joonis 12. Postituste keeleline jaotus läbi aja.

Tulemuste põhjal saame öelda, et r/Eesti puhul on tegemist kakskeelse alamredditiga, sest lisaks eestikeelsetele postitustele on läbi aja olnud püsivalt kõrge ka ingliskeelsete postituste arv. See annab aimu, et lisaks Eesti veebikogukonnale kasutavad Eesti suurimat alamredditit ka välismaalased, st. eesti keelt mitte kõnelevad kasutajad.

5.2.2. Kommentaarid

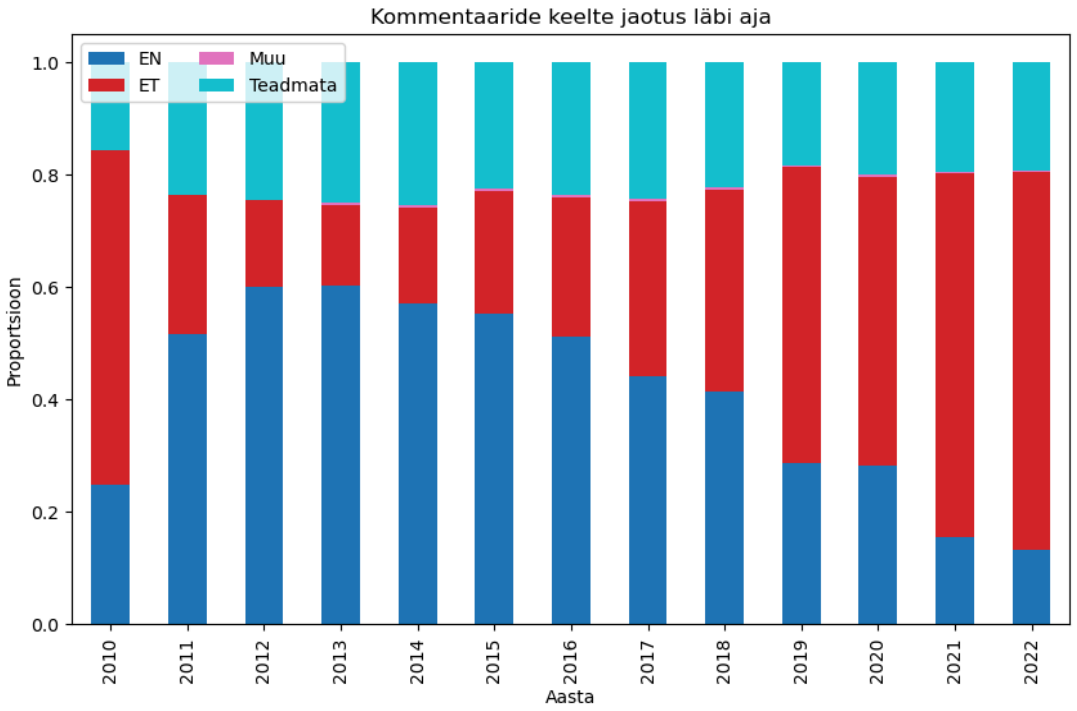
Tabel 7 kirjeldab, millistes keeltes on kirjutatud teksti-tüüpi postituste kommentaarid käesolevas andmestikus. Tabeli põhjal tuleb välja, et suur osa kommentaare on eestikeelsed, mis annab aimu, et r/Eesti alamredditis tekitavad suuremaid arutelusid eestikeelsed postitused. Ingliskeelsete ja teadmata keeles kommentaaride hulk on sarnane. Võrreldes teadmata keelega kommentaaride proportsiooni postituste proportsiooniga, mille keel on teadmata, näeme, et kommentaaride puhul on teadmata keele hulk märgatavalt suurem.

Tabel 7. Kommentaaride keeleline jaotus.

ISO 639_1 keele kood	ISO keele nimi	Kommentaaride arv	Osakaal
ET	eesti	166366	56.90%
EN	inglise	66574	22.77%
unknown	teadmata	58545	20.02%
FI	soome	324	0.111%
UK	ukraina	125	0.043%
SV	rootsi	95	0.032%
TR	türgi	49	0.017%
ES	hispaania	39	0.013%
DE	saksa	28	0.010%
HU	ungari	28	0.010%
NL	hollandi	27	0.009%
FR	prantsuse	26	0.009%
LT	leedu	25	0.009%
PL	poola	23	0.008%
ZH	hiina	21	0.007%
LV	läti	17	0.006%
DA	taani	16	0.005%
CS	tšehhi	14	0.005%

NB	norra	13	0.004%
JA	jaapani	13	0.004%
IT	itaalia	11	0.004%
PT	portugali	5	0.002%
KO	korea	1	0.00034%
EL	kreeka	1	0.00034%
KA	gruusia	1	0.00034%

Joonisel 13 on kujutatud kommentaaride keelelist jaotust läbi aja. Jooniselt on näha, et võrreldes teadmata keelte postituste hulgaga on teadmata keelte kommentaaride hulk läbivalt palju suurem, jäädes läbi aastate keskmiselt 20% protsendi juurde. Eestikeelsete postituste hulk alustab 2010. aastal umbes 60% juures kõikidest kommentaaridest ning aastaks 2012 langeb eestikeelsete kommentaaride osakaal 15% juurde, mis on läbi aastate kõige väiksem eestikeelsete kommentaaride osakaal. Peale 2012. aastat hakkab eestikeelsete kommentaaride osakaal taaskord tõusma, moodustades aastaks 2022 umbes 70% kõikidest kommentaaridest, mis on läbi aastate kõige suurema eestikeelsete kommentaaride osakaaluga aasta. Sarnast muutust läbi aja on näha ka postituste puhul. Täpselt on teadmata, mis sellise muutuse kaasa toob, kuid kui võrrelda joonisel 6 välja toodud r/Eesti jälgijate arvuga läbi aastate näeme, et läbi aastate on kasutajate arv järjest kasvanud ning võib arvata, et r/Eesti algusaegadel on need olnud mitte-eesti-keelt-kirjutavad-kasutajad, sest kuni 2012. aastani suureneb ingliskeelsete kommentaaride osakaal ning hiljem pöördub trend taas eesti-keelt-kirjutavate-kasutajate poole. Tulenevalt Redditi populaarsuse kiirest kasvust ning selle Eestisse hilisemast jõudmisest, võis sellest tulenevalt olla sel hetkel ingliskeelsete kommentaaride osakaal suurem. Edasiste aastate jooksul kogus Reddit populaarsust ka Eestis, mistõttu suureneb ka joonisel järk järgult eestikeelsete kommentaaride osakaal.



Joonis 13. Kommentaaride keeleline jaotus läbi aja.

Postituste ja kommentaaride keeleline jaotus läbi aja on sarnane. Lisaks on näha, et võrreldes postitustega, on kommentaaride puhul ingliskeelsete kommentaaride osakaal väiksem ning tundmatute keelte osakaal suurem. Kõrgem teadmata keelte hulk võib tuleneda kommentaarides esinevatest koodivahetustest ja slängi kasutamisest, mis ajab Lingua keeletuvastuse mudeli segadusse ning seetõttu ei suuda need kommentaarid ületada usaldusskoori lävendit.

5.2.3. Keelesildiga „teadmata“ postituste vähendamine

Magistritöö autor nägi kommentaaride keeletuvastuse abil võimalust vähendada tundmatuks jäänud keelega postitusi teadmata keelega postituste arvu. Selle jaoks loodi algoritm, mis vaatab iga teadmata keelega postituse puhul selle kommentaare ning kui kommentaare on rohkem kui kaks, määratakse postituse keeleks kõige rohkem esinenud keel. Algoritmi testimiseks jooksutati algoritm läbi eelnevalt tuvastatud kõrge usaldusskooriga (usaldusskoor ≥ 0.69) tekstide peal (15 972 postitust), võrreldes kommentaaride keelelisuse silti postituse keelelisuse sildiga. Algoritmi pakutud keel langes kokku 88.93% juhtudel postitustele määratud keelega, mis annab aimu algoritmi võimalikust täpsusest.

Tabelis 8 tuuakse välja arvulised keelte jaotused peale algoritmi rakendamist ning enne algoritmi rakendamist.

Tabel 8. Kommentaaride keeleline jaotus.

ISO 639_1 keele kood	ISO keele nimi	Postituste arv enne algoritmi rakendamist	Postituste arv peale algoritmi rakendamist
ET	eesti	9576	10012
EN	inglise	6469	6581
unknown	teadmata	1841	1192
FR	prantsuse	7	7
RU	vene	5	5
SV	rootsi	4	4
LV	läti	3	3
NB	norra	2	2
ES	hispaania	2	2
KO	korea	1	1
FA	pärsia	1	1
HU	ungari	1	1
TR	türgi	1	1
FI	soome	0	1

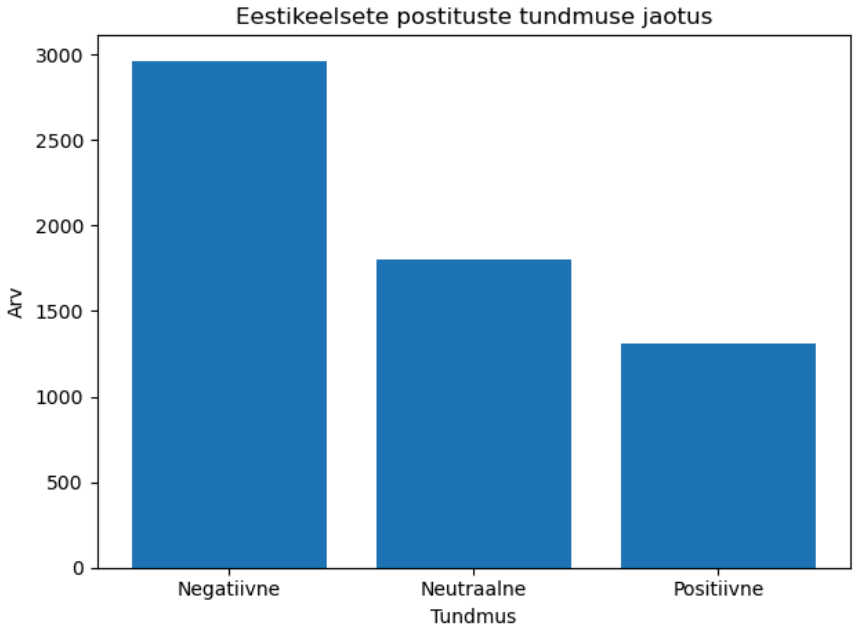
Tulemustest on näha, et peale algoritmi rakendamist on suurenenud peamiselt ainult eestikeelsete postituste hulk, seda 436 postituse võrra. Ingliskeelsete postituste arv tõusis 112 postituse võrra. Lisaks on tekkinud juurde ka üks soomekeelne postitus.

5.3. Tundmusanalüüs

5.3.1. Eesti keel

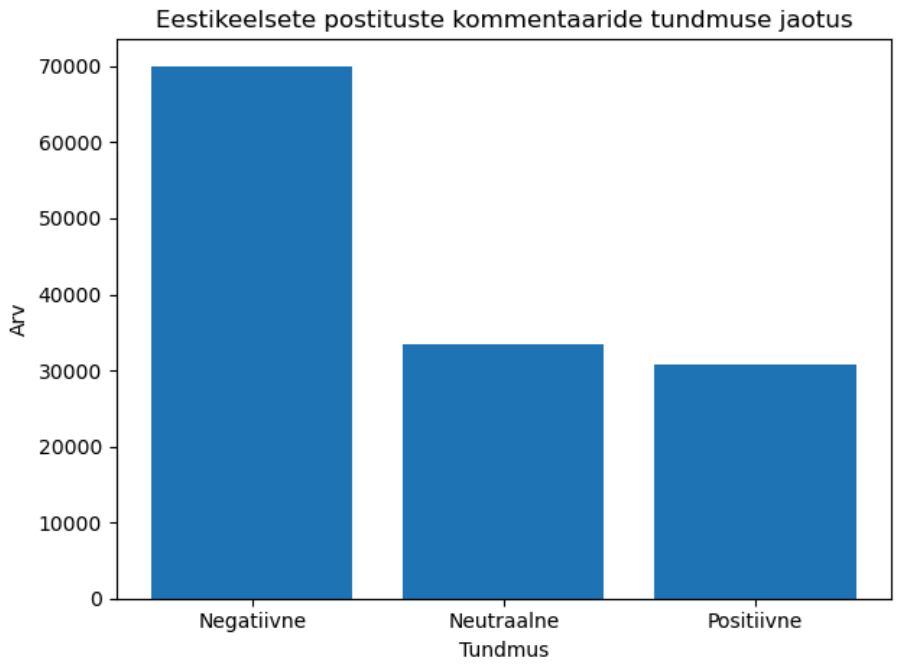
Eestikeelsete tekstipostituste puhul kasutati tundmusanalüüsi läbi viimiseks EstRoBERTa mudelit [36]. Joonisel 14 on välja toodud eestikeelsete tekst-tüüpi postituste tundmuste jaotus. Jooniselt on märgatav, et negatiivse tundmuse sildi saanud postituste hulk on tugevalt

ülekaalus, mis on vastuolus r/Eesti alamredditi reeglitega [20], kus on välja toodud, et postitus peaks olema neutraalne ning enda arvamus tuleks välja tuua kommentaarides. Negatiivse tundmuse hulga järel tuleb neutraalse tundmusega postituste hulk ning kõige vähem on positiivse tundmusega eestikeelseid tekst-tüüpi postitusi.



Joonis 14. Eestikeelsete postituste tundmuse jaotus.

Eestikeelsete tekst-tüüpi postituste kommentaaride tundmuste osakaalud on välja toodud joonisel 15. Jooniselt on näha, et kommentaaride puhul domineerib negatiivse tundmusega kommentaaride hulk. Seejärel tuleb neutraalse tundmusega kommentaaride hulk ning kõige vähem on eestikeelsete postituste juures positiivse tundmusega kommentaare.

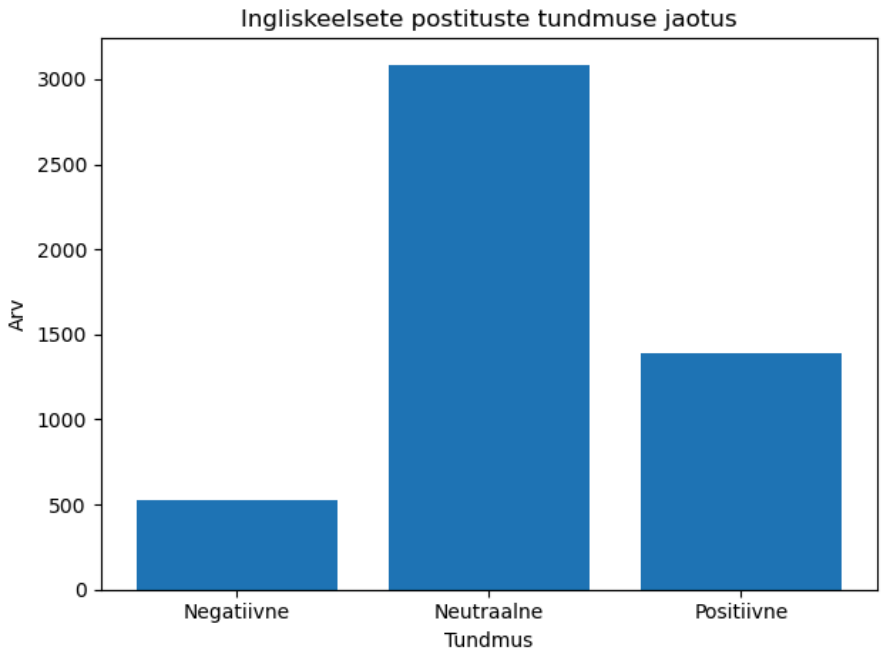


Joonis 15. Eestikeelsete postituste kommentaaride tundmuse jaotus.

Kokkuvõtvalt on eestikeelsete tekst-tüüpi postituste ja nende kommentaaride tundumuse jaotused sarnased, kus tugevalt domineerib negatiivne tundmus, seejärel tuleb neutraalse tundmuse sildi saanud postituste hulk ning alles peale seda tuleb positiivse tundmuse hulk. See annab aimu, et alamredditis r/Eesti on eestikeelsed tekst-tüüpi postitused ja nende kommentaarid tugevalt negatiivselt meelestatud.

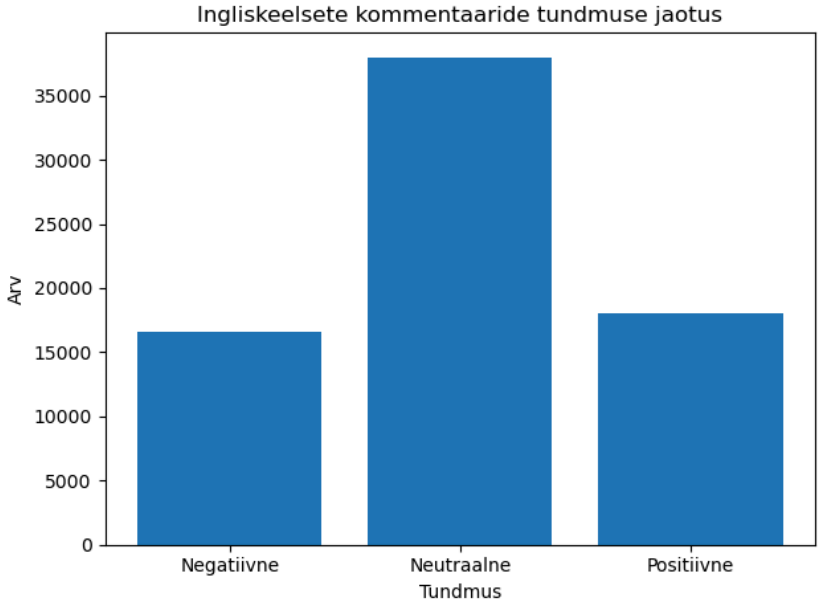
5.3.2. Inglise keel

Ingliskeelsete tekstipostituste jaoks kasutati tundmusanalüüsi läbi viimiseks roBERTa baasmudelit, mis on peenhäälestatud Twitteri andmestikul [37]. Joonisel 16 tuuakse välja ingliskeelsete postituste tundmusanalüüsi tulemused. Jooniselt on näha, et põhiosa moodustavad neutraalsed postitused, mis võib tuleneda r/Eesti alamredditi reeglist nr 2 [20], mille kohaselt tuleb postituse tegijal jääda neutraalseks. Vaadates eraldiseisvalt negatiivse ja positiivse tundmuse sildi saanud postituste hulki, on näha, et ingliskeelsete postituste puhul on ülekaalus positiivse tundmuse sildi saanud postitsused. Võrreldes tulemusi eestikeelsete tekst-tüüpi postituste tulemustega, kus domineeris negatiivne tundmus, on näha, et inglise keelt kõnelevad r/Eesti kasutajad on valdavalt neutraalse meelestatusega, mis on kallutatud pigem positiivse tundmuse poole.



Joonis 16. Ingliskeelsete postituste tundmuse jaotus.

Ingliskeelsete postituste kommentaaride tundmuse jaotus on välja toodud joonisel 17. Tulemustest on näha, et neutraalse tundmuse sildi saanud kommentaaride hulk on väga suur. Küll aga on tõusnud negatiivsete tundmuste hulk kui võrrelda tulemusi postituste tundmuste hulgaga.



Joonis 17. Ingliskeelsete postituste kommentaaride tundmuse jaotus.

Tundmusanalüüsi kohta võib kokkuvõtvalt öelda, et eesti keeles postitades ollakse tugevalt negatiivsed ning inglise keeles postitades ollakse pigem neutraalsed. Inglise keele puhul on näha, et kui jätta kõrvale neutraalsete postituste hulk, joonistub välja asjaolu, et inglise keeles postitatud sisu või kommentaar on pigem positiivse tundmusega. See annab aimu eesti veebikogukonna üldisest meelsusest. Tulemused võivad viidata sellele, et eesti keelt kõnelevad kasutajad väljendavad veebikeskkonnas rohkem kriitikat, samas kui inglise keelse kõnelejad jagavad rohkem positiivseid kogemusi ja toetavaid sõnumeid.

5.4. Teemade analüüs

Teemade analüüs viidi läbi r/Eesti eesti- ja ingliskeelsetel tekstipostitustel kasutades BERTopic-ut [38]. Mõlema keele puhul kasutatud BERTopicu sisseehitatud mitmekeelset mudelit, sest see andis selgemad teemade klastrid, kui eesti- ja ingliskeelsed roBERTa mudelid. BERTopic moodustas teemad ning väljendas neid enamlevinud sõnade järgi, mille põhjal pealkirja panemine on autori interpreteerida, seega subjektiivne. Selle jaoks, et vähendada subjektiivsuse komponenti, märgendasid kõiki teemasid peale autori veel kaks inimest. Lisaks BERTopic-u poolt väljastatud märksõnadele vaadati ka klastritesse jaotatud tekste. Inimeste märgendatud tulemuste pealt tegi teemadele pandavate nimede osas lõpliku otsuse töö autor.

5.4.1. Eesti keel

Eestikeelsete postituste puhul jagunesid tekstid 41 teema vahel (Lisa IV), kus kõige populaarsemaks teemaks läbi aja osutus „Ukraina sõda“, mille teemalisi postitusi oli 326. Kõige vähem populaarsemaks teemaks osutus „Ostlemine veebis“, mille teemalisi postitusi oli andmestikus 96 tükki. Tabelis 9 on välja toodud andmestiku iga aasta kohta kõige populaarsem teema. 2010. 2017. ja 2020. aastal oli populaarseimaks teemaks „Haridus“, mis räägib kooli ja ülikooli sisseastumisega seotud teemadest. 2011. aastal oli populaarseimaks teemaks infovahetus, mis hõlmab endas nii sotsiaalmeedia kui ka ajalehe artikleid puudutavaid teemasid. Aastatel 2012, 2013 ja 2014 aastal oli kõige populaarsemaks teemaks „Keel ja tõlge“, mille sisuks on erinevaid keeli, nende õpet ja tõlget puudutavad küsimused. 2015. ja 2018. aasta kõige populaarsemaks teemaks oli „Kohad“, mis hõlmab endas postitusi, mis räägivad ja küsivad erinevate kohtade kohta, mida on plaanis külastada või palutakse, et

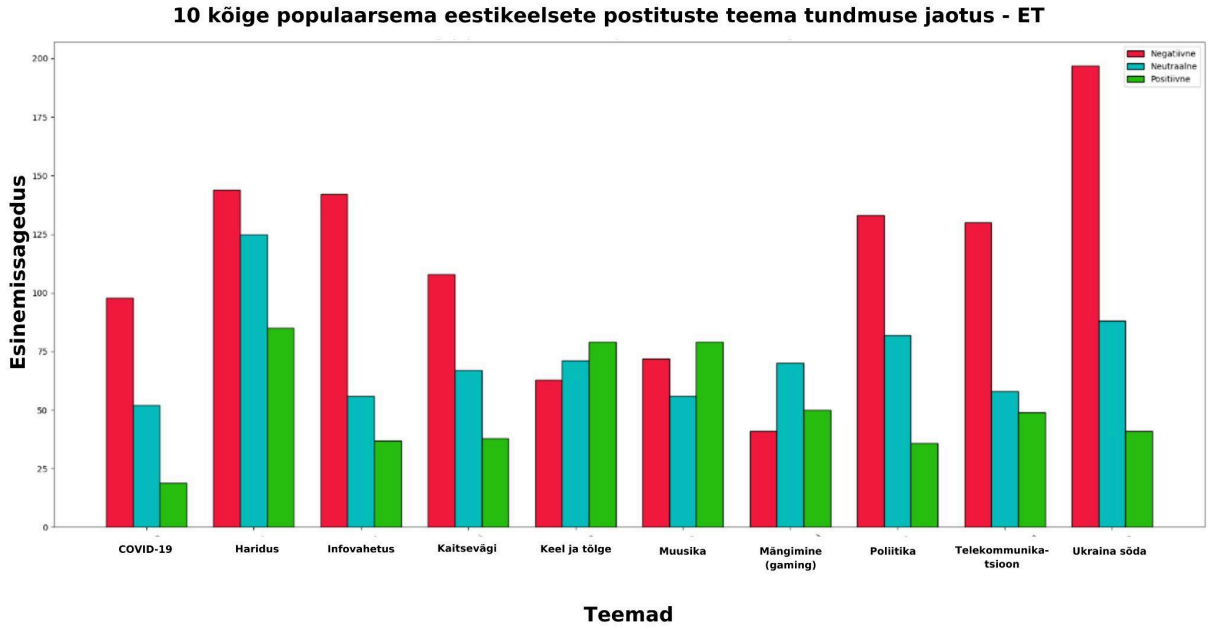
keegi soovitaks mõnda sihtkohaga seotud kohta, mida külastada. Aastal 2016 oli kõige populaarsemaks teemaks „Telekommunkatsioon“, mis hõlmab endas telefonipakettide ning sellega seonduvate tehniliste ning hinnastuse aspektidega seotud postitusi. 2019. aasta kõige populaarsemaks teemaks osutus „Poliitika“, mis sisaldab endas nii Eesti kui ka välispoliitikaga seotud postitusi. Aasta 2021 kõige populaarsemaks teemaks oli „COVID-19, mis hõlmas endas postitusi, mis rääkisid viirusest, selle tüvedest, vaktsineerimisest ja maski kandmisest. Andmestiku viimase, 2022. aasta kõige populaarsemaks teemaks oli „Ukraina sõda“, mis sisaldab Ukraina sõja teemalisi postitusi.

Tabel 9. Kõige populaarsemad eestikeelsete postituste teemad ajavahemikul 2010-2022.

Aasta	Populaarseim teema	BERTopic-u sõnaesitused
2010	Haridus	ülikool, õppima, kool, gümnaasium
2011	Infovahetus	artikkel, postitus, postimees, reddit
2012	Keel ja tõlge	keel, eesti, estonian, elleruu
2013	Keel ja tõlge	keel, eesti, estonian, elleruu
2014	Keel ja tõlge	keel, eesti, estonian, elleruu
2015	Kohad	koht, oskama, soovitama, kohvik
2016	Telekommunikatsioon	telia, internet, telefon, pakett, elisa
2017	Haridus	ülikool, õppima, kool, gümnaasium
2018	Kohad	koht, oskama, soovitama, kohvik
2019	Poliitika	ekre, erakond, riigikogu, valimine
2020	Haridus	ülikool, õppima, kool, gümnaasium
2021	COVID-19	vaktsineeritud, vaktsiin, viirus, mask
2022	Ukraina sõda	venemaa, ukraina, vene, sõda

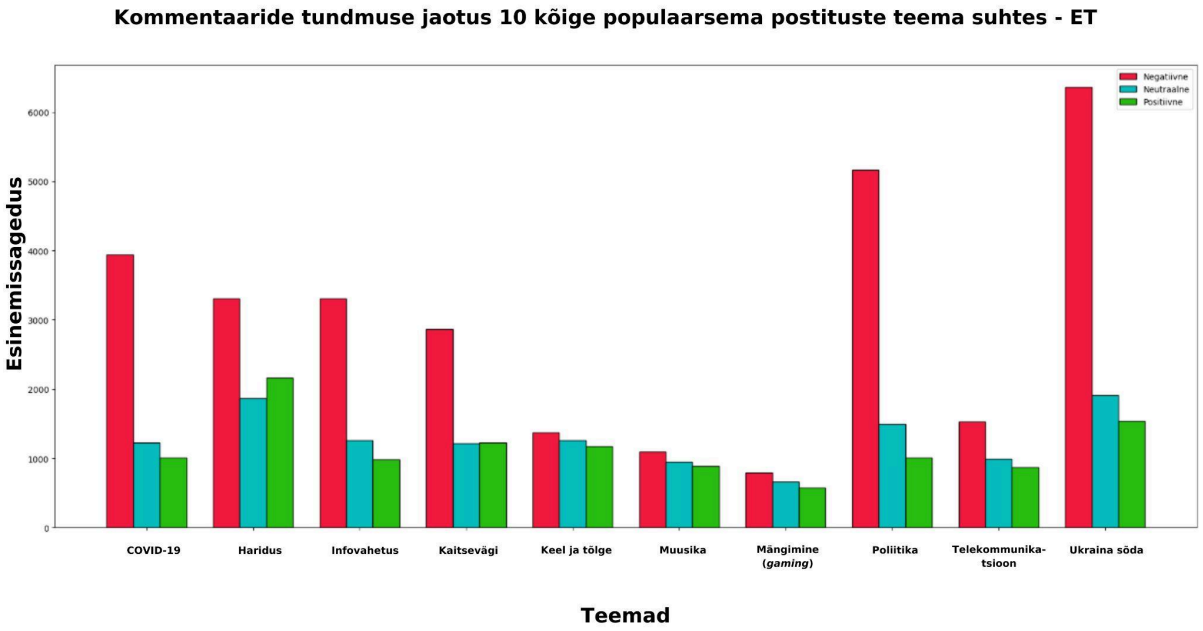
Vaadates lähemalt eestikeelsete postituste 10 kõige populaarsema teema tundmuste jaotust, mis on kujutatud joonisel 18, näeme, et seitse teemat kümnest omavad postitajate poolt

tugevat negatiivset hoiakut. Nendeks teemadeks on „COVID-19“, „Haridus“, „Infovahetus“, „Kaitsevägi“, „Poliitika“, „Telekommunikatsioon“ ja „Ukraina sõda“. Ülejäänud kolme teema puhul, milleks on „Keel ja tõlge“, „Muusika“ ning „Mängimine (*gaming*)“, on ülekaalus positiivne tundmus.



Joonis 18. 10 kõige populaarsema eestikeelsete postituste teema tundmuse jaotus.

Joonisel 19 on kujutatud eestikeelsete tekst-tüüpi postituste kommentaaride tundmuse jaotust 10 kõige populaarsema teema suhtes. Siinkohal on kõikide teemade raames negatiivse tundmuse sildi saanud kommentaaride hulgad kõige suuremad. Teemade „COVID-19“, „Haridus“, „Infovahetus“, „Kaitsevägi“, „Poliitika“ ja „Ukraina sõda“ on negatiivsete tundmuste hulk tugevas ülekaalus.

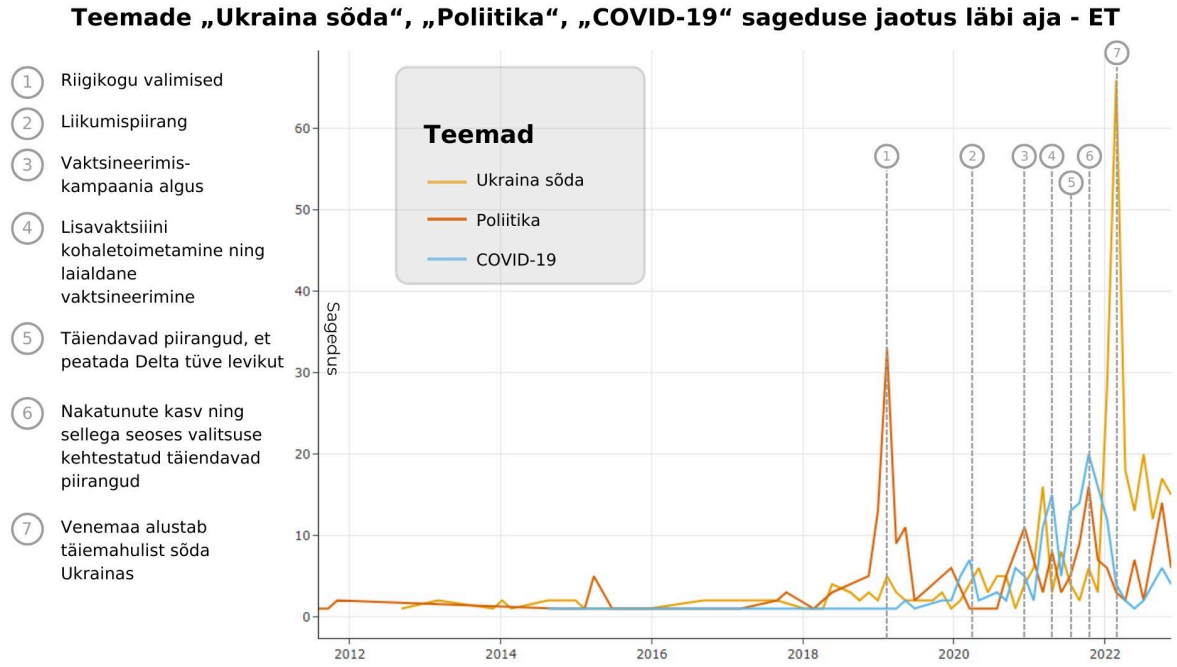


Joonis 19. Kommentaaride tundmuse jaotus 10 kõige populaarsema eestikeelse teema suhtes.

Nii postituste kui ka postituste kommentaaride tundmuse puhul oli enamike teemade puhul näha sarnast meelestatust, mis annab mõista, et üldjuhul on nii postitajate kui ka kommenteerijate tundmus mingi teema kohta sarnane, st. r/Eesti eestlaste kogukond omab sarnaseid tundmuseid erinevate teemade suhtes. Teemade „Keel ja tõlge“, „Muusika“ ning „Mängimine (gaming)“ osas on kommentaaride tundmus võrreldes postituste tundmus muutunud positiivsest negatiivseks.

Allpool on joonisel 20 esitatud kolme teema sagedus ajateljel. Nende teemade puhul oli märgata postituste sageduste muutust tulenevalt mõnest võimalikust Eestis ning maailmas toimunud sündmusest. Nendeks teemadeks olid „Ukraina sõda“, „Poliitika“ ja „COVID-19“. Joonisel toodi välja ka olulise mõjuga sündmused, mis võisid suure tõenäosusega mõjutada vastavalt sageduse muutust vastaval teemal. Näiteks on jooniselt 20 hästi näha, kuidas markeri 1 juures on teema nimega „Poliitika“ sagedus tõusnud märkimisväärselt, mil toimusid XIV Riigikogu korralised valimised. Nendel valimistel moodustati ootamatu ja vastuoluline Keskerakonna-EKRE-Isamaa valitsuskoalitsioon [39]. Teine marker tähistab Eestis kehtestatud eriolukorra ehk liikumispiirangu algust [40]. Selle markeri puhul on näha, kuidas teemaga „COVID-19“ seotud postituste arvu kasvu. Järgnevalt on sarnaselt näha markerite 3, 4, 5 ja 6 puhul, kuidas teemade „Poliitika“ ja „COVID-19“ puhul on teemadega seonduvate postituste arv tõusnud. Nende markerite puhul on tegemist valitsuse poolt

kehtestatud vaktsineerimise- ja piirangutega seotud nõuete kehtestamisega. Marker 7 tähistab sündmust, mil Venemaa alustas täiemahulist sõda Ukraina vastu. Teema „Ukraina sõda“ puhul on näha, kuidas sõja algusega on Ukraina sõjaga seotud teemaliste postituste arv tõusnud hüppeliselt.



Joonis 20. Teemade „Ukraina sõda“, „Poliitika“, „COVID-19“ sageduse jaotus läbi aja.

Eestikeelsete postituste ja nendega seotud kommentaaride põhjal oli näha, et alamredditi r/Eesti kasutajad on populaarsete teemade raames millegi või kellegi kohta negatiivset meelsust avaldanud. Vastupidiselt on postituste autorid pigem positiivset meelsust avaldanud teemade raames nagu „Keel ja tõlge“, „Muusika“ ja „Mängimine (*gaming*)“, kuid kommentaariumis on ka nende teemade puhul ülekaalus negatiivne tundmus. Lisaks on läbi aja näha eriti tugevalt just kolme teema puhul, kuidas erinevad sündmused Eestis ja väljaspool on mõjutanud teemade populaarsust läbi aja.

5.4.2. Inglise keel

Ingliskeelsete postituste puhul jagunesid tekstid 29 teema vahel (Lisa V), millest iga andmestiku aasta kohta on tabelis 10 välja toodud vastava aasta kõige populaarsem teema. Aastal 2010 oli kõige populaarsemaks teemaks „Eesti keele õpe“, mille puhul küsiti alamredditis r/Eesti eesti keele õppimisega seotud küsimusi, näiteks:

„I am looking for good resources for learning Estonian, help? I am attempting to learn Estonian, but I am having a hard time with noun tenses and when to use mulle and mina and such. Can any of you point me somewhere useful?“

Eesti keeles: „Otsin häid ressursse eesti keele õppimiseks, kas keegi saaks aidata? Ma üritan õppida eesti keelt, aga mul on raskusi nimisõnade käändega ja millal kasutada „mulle“ ja „mina“ jne. Kas keegi teist oskab mind suunata kuhugi kasulikku?“

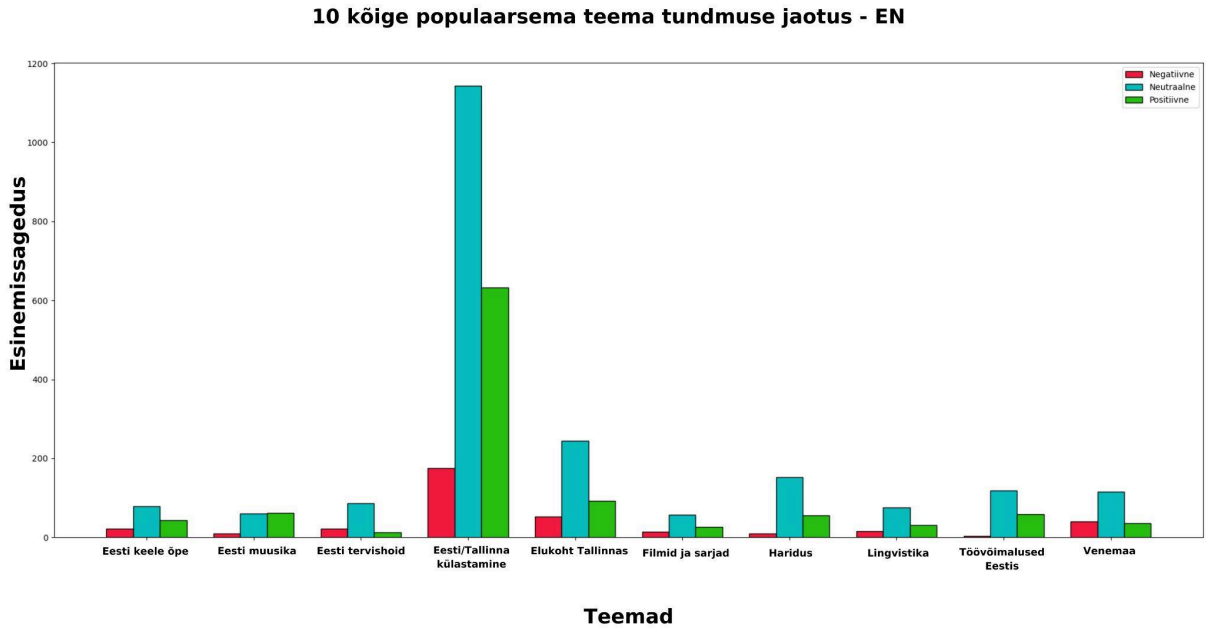
Järgnevatel aastatel 2010 kuni 2022 on kõige populaarsemaks teemaks igal aastal olnud „Eesti/Tallinna külastamine“. Antud teema postituste puhul küsib postitaja erinevate kohtade kohta Tallinnas või üldiselt Eestis, mida külastada.

Tabel 10. Kõige populaarsemad ingliskeelsete postituste teemad ajavahemikul 2010-2022.

Aasta	Populaarseim teema	BERTopic-u sõnaesitused
2010	Eesti keele õpe	learn, language, estonian, course
2011	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2012	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2013	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2014	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2015	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2016	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2017	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2018	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2019	Eesti/Tallinna külastamine	tallinn, estonian, place, know

2020	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2021	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2022	Eesti/Tallinna külastamine	tallinn, estonian, place, know

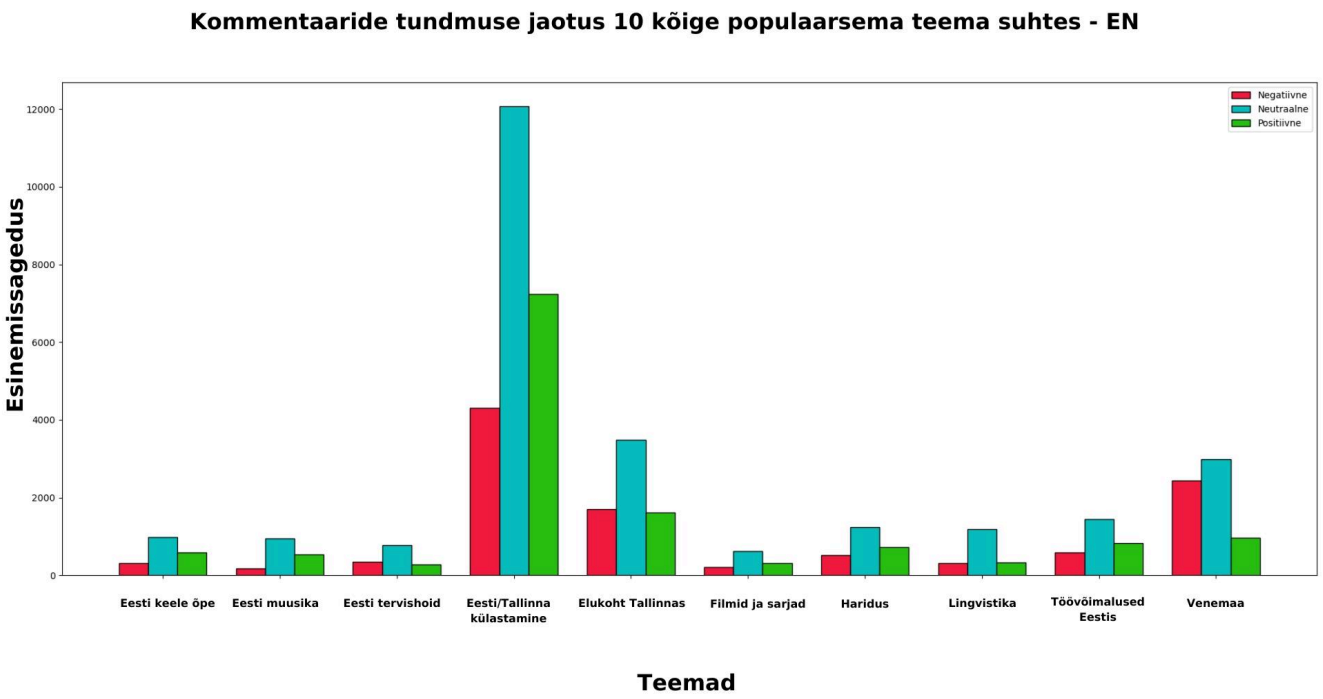
Eelnevale tabelile lisaks on välja toodud ingliskeelsete postituste 10 kõige populaarsema teema tundmuse jaotus joonisel 21. Jooniselt 21 on näha, et läbi andmestiku domineerib tugevalt teema „Eesti/Tallinna külastamine“. Sarnaselt oli antud teema ka peaaegu kõikide aastaste puhul kõige populaarsem teema. Lisaks kuuluvad populaarsete teemade hulka ka „Eesti keele õpe“, „Eesti muusika“, „Eesti tervishoid“, „Elukoht Tallinnas“, „Filmid ja sarjad“, „Haridus“, „Lingvistika“, „Töövõimalused Eestis“ ning „Venemaa“. Tulenevalt r/Eesti alamredditi reeglitest, tuleb postituse postitajal teha vaid neutraalseks jäävaid postitusi, mistõttu on enamuste teemade puhul kõige sagedasem neutraalne tundmus. Vaid teema „Eesti muusika“ puhul on ülekaalus positiivsete tundmuste hulk. Lisaks on teemade „Eesti tervishoid“ ja „Venemaa“ puhul märgata, et need teemad on pigem negatiivselt kui positiivselt kallutatud, mis teema „Eesti tervishoid“ puhul võib tuleneda välismaalaste halvast kogemusest Eesti meditsiinisüsteemiga ning teema „Venemaa“ puhul 2022. aasta alguses aset leidnud Ukraina sõda. Ülejäänud teemade puhul on näha, et võrreldes negatiivsete tundmuste hulgaga on meelestatus positiivselt kallutatud.



Joonis 21. 10 kõige populaarsema ingliskeelse teema tundmuse jaotus.

Heites pilgu ingliskeelsete postituste 10 kõige populaarsema teema kommentaaride tundmustele (joonis 22), näeme võrreldes postituste teemade tundmustega juba väikeseid muutuseid. Nimelt on märgatavalt tõusnud negatiivsete tundmuste osakaal teemadel „Eesti/Tallinna külastamine“, „Elukoht Tallinnas“, „Haridus“, „Töövõimalused Eestis“ ning „Venemaa“ võrreldes kommentaaride tundmuste jaotust postituste tundmuste jaotusega. Teema „Venemaa“ puhul on näha ka märgatavat aktiivsuse kasvu kommentaaridel.

Üks võimalik põhjendus negatiivse tundmuse kasvus kommentaaride osas võib-olla asjaolu, et eestlased on postitusi tehes ning postitusi kommenteerides üldiselt pigem negatiivselt meelestatud, nagu eelnevalt eestikeelsete postituste ja kommentaaride osas välja tuli. See tähendab, et võib-olla on tegemist olukorraga, kus välismaalased küsivad alamredditis r/Eesti-s küsimusi, millele Eesti päritolu kasutajad vastavad, millest võib olla tingitud negatiivsuse kasv kommentaarides.



Joonis 22. Kommentaaride tundmuse jaotus 10 kõige populaarsema teema suhtes - EN.

Tulemustest on näha, et ingliskeelsete postituste autorid ning kommenteerijad on populaarsete teemade suhtes pigem positiivselt meelestatud ning eestikeelsete postituste autorid ja kommenteerijad on populaarsete teemade suhtes negatiivselt meelestatud.

Teemade osas on näha mõlema keele puhul kattuvaid teemasid, näiteks „Muusika“, „Tervishoid“, „Haridus“ ning „Filmid ja Sarjad“. Nii eesti- kui ka ingliskeelsete postituste ja kommentaaride kõikide teemade tundmuste jaotus on nähtav töö autori GitHubi repositooriumis[18]. Eestikeelsete postituste puhul tegemist on teemade jaotus ühtlasem, st. ei domineeri vaid üks teema. Ingliskeelsete postituste puhul domineerib tugevalt teema „Eesti/Tallinna külastamine“. Lisaks on eestikeelsed postitused ja kommentaarid erinevate teemade kohapealt negatiivselt meelestatud, vastupidiselt ingliskeelsetele postitustele ja kommentaaridele, kus domineerib neutraalne ja positiivne toon. See annab aimu veebikogukonna suhtumisest üldiselt ning ka erinevate teemade lõikes. Nende tulemuste põhjal saab öelda, et r/Eesti eestlastest veebikogukond paistab negatiivselt meelestatud võrreldes ingliskeelse veebikogukonnaga.

[18] https://github.com/taunotamm/DS_MSc_Thesis

6. Arutelu

Selles peatükis seotakse tulemused püstitatud uurimisküsimustega ning lisaks tuuakse välja tööga seondunud piirangud ja võimalikud ettepanekud teema arenduseks tulevikuks.

Uurimisküsimus 1 - Kuidas postitatakse?

Selle magistritöö käigus selgus, et alamredditis r/Eesti tehakse kõige rohkem teksti-tüüpi postitusi. Lisaks on populaarsed ka link-tüüpi ning pilt-tüüpi postitused. Vähem populaarsed postituse tüübid on video-tüüpi ja küsitluse-tüüpi postitused. Link-tüüpi postituste puhul tuli välja, et alamredditi r/Eesti kasutajad jagavad edasi kõige sagedamini err.ee domeenilt pärinevat infot.

Alamredditi r/Eesti kasutajad teevad postitusi peamiselt päevasel ajal, mil kõige sagedasem postituse aeg jääb vahemikku kella 13.00 kuni kella 21.00-ni. Päevade lõikes postitatakse pigem tööpäeviti, kuid nädalavahetusel tehtud postituste arv ei ole marginaalselt väiksem.

r/Eesti puhul saab öelda, et tegemist on kakskeelse alamredditiga, sest umbes 56% kõikides peale puhastamist alles jäänud postitustest on eestikeelsed ning ligi 37% postitustest on ingliskeelsed. Ülejäänud postituste puhul on keel teadmata või on tegemist andmestikus harva esinevad keelega. Täiendavalt selgus, et vene keelt, mis on suurim vähemuskeel Eestis, eksisteerib r/Eesti alamredditis väga vähesel määral.

Uurimisküsimus 2 - Millest postitatakse?

Eestikeelsete postituste puhul kõige populaarsemaks teemaks läbi aegade „Haridus“, teisel kohal on „Ukraina sõda“ ning kolmandal kohal „Poliitika“. Ingliskeelsete postituste puhul on näha, et läbi aegade on kõige populaarsemaks teemaks „Eesti/Tallinna külastamine“, edestades tugevalt teise koha teemat, milleks on „Elukoht Tallinnas“. Kolmandaks kõige populaarsemaks teemaks ingliskeelsete postituste puhul on „Haridus“. Teemade puhul on näha kattuvaid teemasid, kuid üldiselt on teemade puhul märgata, et eestikeelsete postituste puhul arutletakse elu ja olu üle Eestis ning küsitakse teistelt kasutajatelt nõu. Ingliskeelsete postituste puhul on näha, et postitused küsivad Eesti kohta - elukohad, töövõimalused, hariduse võimalused või kohad, mida külastada. Kuna „Eesti/Tallinna külastamine“ on domineeriv teema, võime järeldada, et ingliskeelsed kasutajad on inimesed/turistid, kes

plaanivad Eestit ajutiselt külastada ning vähem on neid, kes on tulnud Eestisse pikemaks ajaks elama.

Kasutajate tundmus eesti keeles postitades on tugevalt negatiivne ning inglise keeles postitades tugevalt neutraalne. Alamredditi r/Eesti reegli nr 2 [20] kohaselt peavad postitused kandma neutraalset tooni ning isiklik arvamus tuleb välja tuua kommentaarides. Sellest tulenevalt näeme, et enamik eestikeelseid postitusi, kus on ülekaalus negatiivne tundmus, ei vasta r/Eesti reeglitele. Vaadates kõrvale ka vastavate keelte postituste kommentaare, on näha, et eestikeelsete postituste kommentaaride puhul on kõige rohkem negatiivse tundmusega kommentaare ning ingliskeelsete postituste puhul on kõige rohkem neutraalse tundmusega kommentaare, millele järgneb positiivse tundmuse saanud kommentaaride hulk. See annab meile üldise arusaama, et tundmuste jagunemine on keeleti erinev. Täiendavalt jääb töö autorile tunne, et r/Eesti reegel neutraalsete postituste kohta ei ole võib-olla päris täpselt sõnastatud. See seisneb selles, et mõlemad keeled omavad lisaks neutraalse tundmusega postitustele ka negatiivse ja positiivse tundmusega postitusi, mis tähendab, et tegemist oleks reegli järgi ebasobilike postitustega. Kuid tuues näiteks olukorra, kus r/Eesti kasutaja jagab alamredditis mõnd laulu linki, kirjutades juurde, et „see on hea ja tore laul“, siis saab ka see postitus positiivse tundmuse märgi ning reegli järgi on tegemist ebasobiva postitusega, kuid selline asi autori arvates tegelikult modereerimist ei vaja.

Eestikeelsete postituste osas oli näha üldist negatiivset tooni, kuid kui vaadata postituste tundmuse jaotust teemade lõikes, tuli välja, et eriti suur negatiivne osakaal oli teemadel „Ukraina sõda“, „Poliitika“ ning „COVID-19“. Eestikeelsete postituste kommentaaride osas oli tulemused sarnased. Ingliskeelsete postituste puhul tuli välja, et üldiselt oli kõikide teemade puhul kõige populaarsem neutraalne hoiak, mis oli positiivse tundmuse poole kaldu. Teemadest kõige negatiivsema tundmuse osakaaluga teemadeks olid „Venemaa“ ja „COVID-19“. Ingliskeelsete postituste kommentaaride osas oli näha negatiivse tundmuse kasvu võrreldes ingliskeelsete tekst-tüüpi postitustega. Kõige negatiivsemateks teemadeks osutusid „Venemaa“, „Suhted naaberriikidega“, „Poliitika“ ja „COVID-19“.

6.1. Piirangud

Töö koostamise käigus viidi läbi tundmusanalüüs, mille puhul on autori hinnangul puudus inimese poolt tundmuse siltidega märgendatud eestikeelne sotsiaalmeedia tekstide andmestik. Kuna rakendati teises valdkonnas (ajalehetekstidel) peenhäälestatud tundmusanalüüsi

mudelit, võib ka tundmusanalüüsi tulemus olla sellevõrra ebatäpsem. Töö autori arvates oleks täpsemaks tundmusanalüüsiks vaja käsitsimärgendatud sotsiaalmeedia tundmusanalüüsi andmestikku, millele mudel peenhäälestada.

Teemade analüüsi käigus määrati teemad käsitsi BERTopic-u poolt väljastatud märksõnade ning klastrites olevate tekstide põhjal. Märksõnade tõlgendamiseks on eri viise, mistõttu lasub teemade määramisel ka subjektiivne komponent.

6.2. Tulevikuväljavaated

Selle töö raames uuriti tekstipostituste tundmuseid ja teemasid. Kuna tekst-tüüpi postitused moodustavad ~40% kogu andmestikust, ei anna see meile andmestiku kohta täielikku ülevaadet. Üheks võimalikuks lähenemiseks teemade analüüsiks oleks proovida mitte-tekst-tüüpi postituste kommentaaride põhjal ennustada nende teema. Tegemist oleks sarnase lähenemisega, nagu käesoleva töö puhul, kus tundmatu keele sildiga postituste keelelisuse ennustamiseks kasutati kommentaaride keele silte. Täiendavalt oleks võimalik uurida ka link-tüüpi postitusi, mille puhul laetakse alla jagatava veebiaadressi tekstiline sisu ning saadud tulemuste põhjal määratakse postituse temaatika. Sarnasel oleks võimalik uurida ka pilt-tüüpi postitusi, kus analüüsidakse mõne närvivõrkudel põhineva automaatse pildikirjelduse lähenemisega pildi sisu ning saadud väljundi põhjal määratakse postituse teema. Kaks viimast lähenemist on tõenäoliselt oluliselt ressursinõudlikumad kui kommentaaride põhjal teema ennustamine.

Teiseks võimalikuks teema edasiarenduseks pakub töö autor välja, et sarnane andmestik samal ajavahemikul tuleks kätte saada ka platvormilt X (endine Twitter) ning veel mõnelt sarnaselt platvormilt. Sellisel juhul oleks võimalik võrrelda, kas jututeemad on platvormiti sarnased või erinevad ning millised on hoiakud nendes olevate teemade suhtes.

Kolmandaks võimaluseks oleks võrrelda Eesti redditi kommuuni mõne teise väikese keele omaga, näiteks alamredditiga r/latvia. See annaks aimu, kas ja kuidas erinevad kahe väheste rääkijate arvuga keele puhul alamredditi teemad ja hoiakud. Antud lähenemine eeldaks vastavalt tihedat koostööd teise keele rääkijaga.

Kokkuvõttes annaks selline lähenemine aimu Eesti veebikogukonna olemusest - millist platvormi milliste teemade jaoks kasutatakse.

7. Kokkuvõte

Redditi puhul on tegemist maailma kõige suurema foorumiga, mida jälgib igakuiselt umbes 1.2 miljardit kasutajat. Redditis eksisteerib r/Eesti nimeline alamreddit, millel on umbes 86 tuhat jälgijat (seisuga 01.05.2024), kuuludes 2% kõige suuremate alamredditite hulka. Selle magistritöö eesmärgiks oli luua alamredditi r/Eesti andmete põhjal korpus, mida on hiljem ka teistel võimalik kasutada teadustöö eesmärgil. Teiseks eesmärgiks oli rakendada loomuliku keele töötluse meetodeid, et analüüsida loodud korpuses olnud andmeid, mis annaks aimu r/Eesti alamredditi sisu struktuuri, hoiakute ja teemade jaotusest.

Selle magistritöö raames lõi töö autor alamredditi r/Eesti postitustel ja kommentaaridel põhineva struktureeritud korpuse ning analüüsis selles olevaid andmeid, rakendades andmetel erinevaid keeletehnoloogia võtteid, sh. tundmusanalüüsi ja teemadeanalüüsi. Analüüsi tulemustest tuli välja, et alamredditis r/Eesti kasutatakse postitamiseks kõige rohkem tekst-tüüpi postitusi. Vastavalt teisel ja kolmandal kohal oli link- ja pilt-tüüpi postitused. Link-tüüpi postituste puhul selgus, et kõige enam jagatakse edasi domeeni err.ee linke. Teksti-postituste puhul tuli välja, et r/Eesti näol on tegemist kakskeelse foorumiga, kus ~54% protsenti kõikides tekst-tüüpi postitustest on eestikeelsed ning ~36% ingliskeelsed. Tundmusanalüüsi tulemused näitasid, et eesti keeles postitavad ja kommenteerivad kasutajad on tugevalt negatiivselt meelestatud, kuid inglise keelt kirjutavad kasutajad neutraalselt meelestatud, olles pigem positiivse tundmuse poole kaldu. Teemade analüüsi puhul tuli välja, et eestikeelse kogukonna jaoks läbi aegade kõige populaarsemad teemad oli „Haridus“, „Ukraina sõda“ ning „Poliitika“. Inglise keelt kõnelevate kasutajate jaoks oli kõige populaarsemateks teemadeks „Eesti/Tallinna külastamine“, „Elukoht Tallinnas“ ning „Haridus“.

Kokkuvõttes uuriti selle magistritöö käigus r/Eesti alamredditi sisu ja kasutajate hoiakuid erinevate teemade suhtes, mis annab aimu Eesti veebikogukonna olemusest. Samuti lõi töö autor alamredditi positustel, kommentaaridel ja metaandmetel põhineva korpuse, mida on ka edaspidiselt võimalik kasutada teadustöö eesmärgil.

8. Viidatud kirjandus

[1] Dean B. Reddit User and Growth Stats. 2024. https://backlinko.com/reddit-users (01.04.2024)

[2] Lui M, Baldwin T. Accurate language identification of twitter messages. *Proceedings of the 5th workshop on language analysis for social media (LASM)*, 2014, https://aclanthology.org/W14-1303.pdf (13.04.2024)

[3] PyPI. pycld2 0.41. 2019. https://pypi.org/project/pycld2/ (01.05.2024)

[4] PyPI. langid 1.1.6. 2016. https://pypi.org/project/langid/ (01.05.2024)

[5] PyPI. langdetect 1.0.9. 2021. https://pypi.org/project/langdetect/ (01.05.2024)

[6] PyPI. lingua-language-detector 2.0.2. https://pypi.org/project/lingua-language-detector/ (23.04.2024)

[7] Goel R, Modhukur V, Täär K, Salumets A, Sharma R, Peters M. Users' Concerns About Endometriosis on Social Media: Sentiment Analysis and Topic Modeling Study. *Journal of Medical Internet Research*, 2023, https://www.jmir.org/2023/1/e45381/ (01.05.2024)

[8] Loria S. textblob documentation. 2020. https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf (08.05.2024)

[9] Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, 2014, (Vol. 8, No. 1, pp. 216-225), https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (08.05.2024)

[10] BERT: Zhang T, Xu B, Thung F, Haryono SA, Lo D, Jiang L. Sentiment analysis for software engineering: how far can pre-trained transformer models go?. *Proceedings of the 2020 International Conference on Software Maintenance and Evolution*, 2018, https://ieeexplore.ieee.org/document/9240704 (08.05.2024)

[11] Kędzierska M, Spytek M, Kurek M, Sawicki J, Ganzha M, Paprzycki M. Topic Modeling Applied to Reddit Posts. *Big Data Analytics in Astronomy, Science, and Engineering: 11th International Conference on Big Data Analytics*, *BDA 2023*, 2023, Proceedings (lk. 17). https://link.springer.com/chapter/10.1007/978-3-031-58502-9_2 (01.05.2024)

[12] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of machine Learning research*, 2003, https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=http://githubhelp.com (08.05.2024)

[13] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *nature*, 1999, https://www.nature.com/articles/44565 (08.05.2024)

[14] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. 2022, https://arxiv.org/abs/2203.05794 (15.04.2024)

[15] Lehes L. Understanding Public and Leaders' Opinion about Russo-Ukrainian War through Social Media Platforms: An Estonian Case Study. 2023. https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=77493&language=en (01.05.2024)

[16] Tanvir H, Kittask C, Eiche S, Sirts K. EstBERT: A pretrained language-specific BERT for Estonian. *arXiv preprint arXiv:2011.04784*, 2020, https://arxiv.org/abs/2011.04784 (08.05.2024)

[17] Pajupuu H, Altrov R, Pajupuu J. Identifying polarity in different text types. *Folklore: Electronic Journal of Folklore*, 2016, http://www.folklore.ee/folklore/vol64/polarity.pdf (08.05.2024)

[18] Scanlon K. The Rundown: Everything you need to know about Reddit as the platform goes public. 2024. https://digiday.com/marketing/the-rundown-everything-you-need-to-know-about-reddit-as-the-platform-goes-public/ (25.03.2024)

[19] Reddit. Community settings. 2024. https://support.reddithelp.com/hc/en-us/articles/15484546290068-Community-settings (25.03.2024)

[20] Reddit. r/Eesti. https://www.reddit.com/r/Eesti/ (25.03.2024)

[21] Forbes. Reddit Stands By Controversial API Changes As Subreddit Protest Continues. https://www.forbes.com/sites/antoniopequenoiv/2023/06/13/reddit-stands-by-controversial-api-changes-as-subreddit-protest-continues/ (25.03.2024)

[22] Xu W, Sasahara K, Chu J, Wang B, Fan W, Hu Z. A multidisciplinary framework for deconstructing bots' pluripotency in dualistic antagonism. *arXiv preprint arXiv:2402.15119*, 2024, https://arxiv.org/ftp/arxiv/papers/2402/2402.15119.pdf (05.03.2024)

[23] Koppel K, Kallas J. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat 18*, 207–228, Vol 18, 2022, http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/ERYa18.12/543 (25.04.2024)

[24] The-Eye. DMCA. https://the-eye.eu/dmca.htm (26.03.2024)

[25] Rose-Collins F. Kuidas lahendada DMCA kaebus: Mida peate teadma. https://www.ranktracker.com/et/blog/how-to-resolve-a-dmca-complaint-what-you-must-know/ (26.03.2024)

[26] Juurik M, Mäesalu T, Tarkpea T. Andmekaitse teadustöös. 2023. https://wiki.ut.ee/pages/viewpage.action?pageId=196183311 (08.05.2024)

[27] Riig Teataja. Isikuandmete kaitse seadus. https://www.riigiteataja.ee/akt/111032023011?leiaKehtiv (29.04.2024)

[28] AWS. What is Sentiment Analysis?. https://aws.amazon.com/what-is/sentiment-analysis/ (01.05.2024)

[29] Ulčar M, Žagar A, Armendariz CS, Repar A, Pollak S, Purver M, Robnik-Šikonja M. Evaluation of contextual embeddings on less-resourced languages. arXiv preprint arXiv:2107.10614. 2021, https://www.researchgate.net/publication/353398910_Evaluation_of_contextual_embeddings_on_less-resourced_languages (23.04.2024)

[30] Barbieri F, Camacho-Collados J, Neves L, Espinosa-Anke LT. Unified benchmark and comparative evaluation for tweet classification. *arXiv 2020. arXiv preprint arXiv:2010.12421*. 2020. https://arxiv.org/abs/2010.12421 (08.05.2024)

[31] Egger R, Yu J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 2022, https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498/full (08.05.2024)

[32] Grootendorst M. Outlier reduction. https://maartengr.github.io/BERTopic/getting_started/outlier_reduction/outlier_reduction.html#strategies (08.05.2024)

[33] Subreddit Stats. r/Eesti stats. https://subredditstats.com/r/eesti (21.04.2024)

[34] upvoted. Introducing Reddit Polls, An All-New Post Type. https://www.redditinc.com/blog/introducing-reddit-polls-an-all-new-post-type/ (01.05.2024)

[35] GoDaddy. What is a subdomain?. https://in.godaddy.com/help/what-is-a-subdomain-296 (24.04.2024)

[36] Hugging Face. https://huggingface.co/EMBEDDIA/est-roberta (01.04.2024)

[37] Camacho-Collados J, Rezaee K, Riahi T, Ushio A, Loureiro D, Antypas D, Boisson J, Espinosa-Anke L, Liu F, Martínez-Cámara E, Medina G. Tweetnlp: Cutting-edge natural language processing for social media. *arXiv preprint arXiv:2206.14774*, 2022, https://arxiv.org/abs/2206.14774 (01.05.2024)

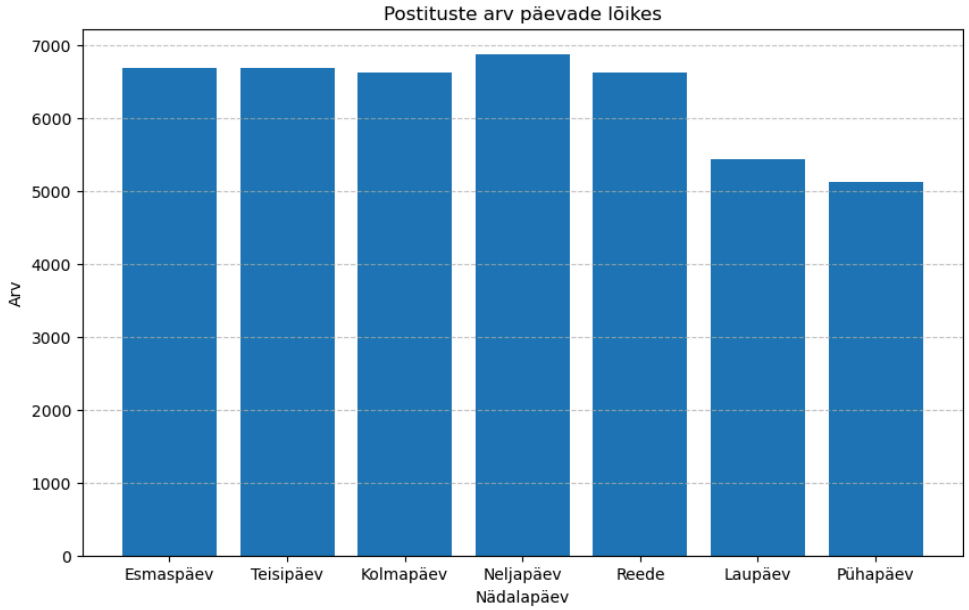
[38] Grootendorst M. BERTopic. https://maartengr.github.io/BERTopic/index.html (01.04.2024)

[39] https://et.wikipedia.org/wiki/2019._aasta_Euroopa_Parlamendi_valimised_Eestis (29.04.2024)

[40] Vabariigi Valitsus. Valitsus kuulutas Eestis välja eriolukorra 1. maini. 2020. https://www.valitsus.ee/uudised/valitsus-kuulutas-eestis-valja-eriolukorra-1-maini (29.04.2024)

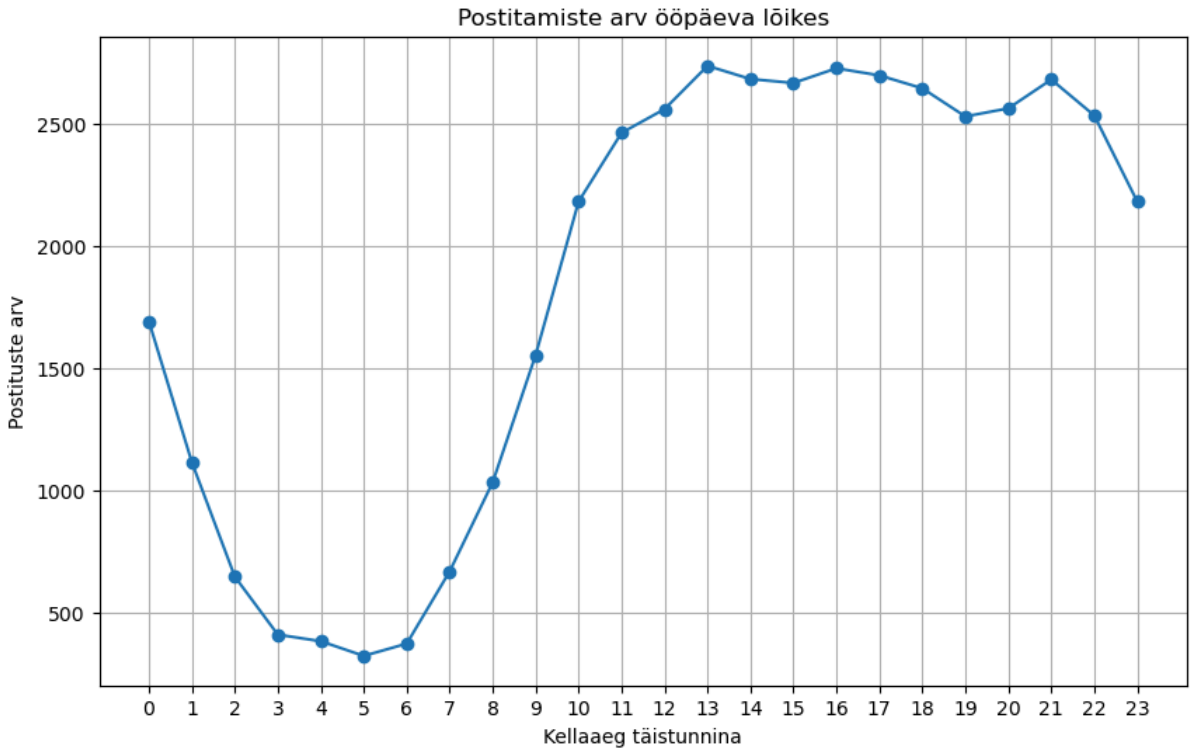
Lisad

I. Postituste arv päevade lõikes



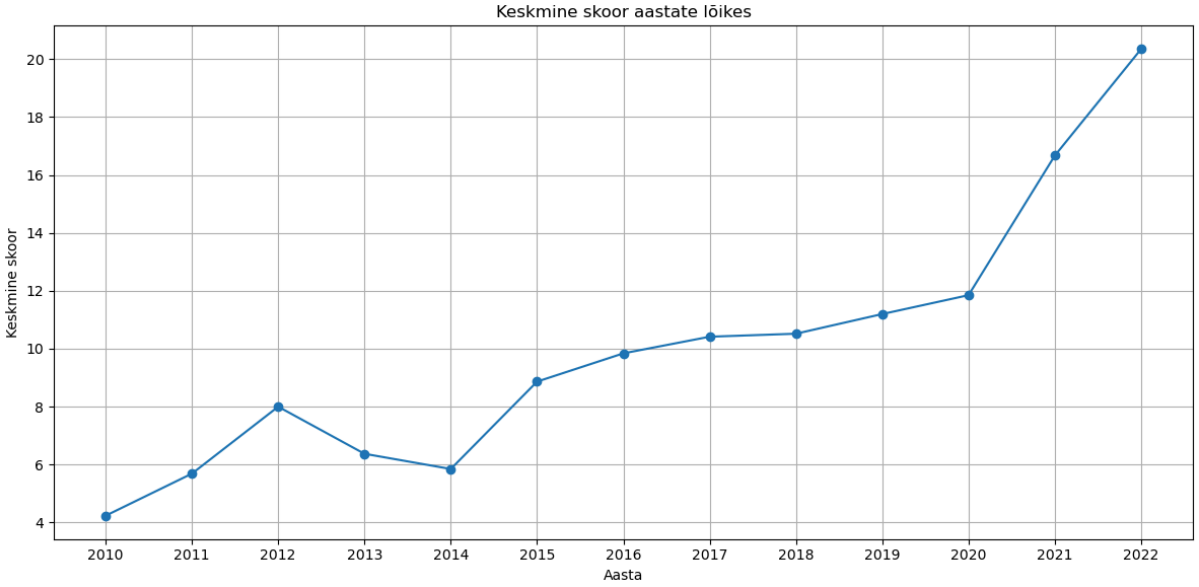
Joonis 23. Postituste arv päevade lõikes.

II. Postituste arv ööpäeva lõikes



Joonis 24. Postitamiste arv ööpäeva lõikes.

III. Keskmine skoor aastate lõikes



Joonis 25. Keskmine skoor aastate lõikes.

IV. Teemade analüüsi tulemused - ET

Tabel 11. Teemade analüüsi tulemused - ET.

Järjekorra number	Postituste arv	Teema	BERTopic-u sõnaesitused
1	354	Haridus	ülikool, õppima, kool, gümnaasium
2	326	Ukraina sõda	venemaa, ukraina, vene, sõda
3	251	Poliitika	ekre, erakond, riigikogu, valimine
4	248	Kohad	koht, oskama, soovitama, kohvik
5	237	Telekommunikatsioon	telia, internet, telefon, pakett, elisa
6	235	Infovahetus	artikkel, postitus, postimees, reddit
7	234	Ostlemine veebis	tellima, särk, amazon, pakk
8	215	Toit	toit, sööma, burger, liha
9	213	Keel ja tõlge	keel, eesti, estonian, elleruu
10	213	Kaitsevägi	ajateenistus, kaitsevägi, arstlik, komisjon
11	208	Varia	meem, komm, kalev, kodu
12	207	Muusika	laul, muusika, lugu, kuulama
13	204	Kohad Tallinnas	tallinn, koht, käinud, käima, teadma
14	169	COVID-19	vaktsineeritud, vaktsiin, viirus, mask
15	167	Filmid ja sarjad	film, kino, vaatama, netflix
16	161	Mängimine (gaming)	mäng, discord, gaming, server
17	158	Ostma/müüma	pood, ostma, müüma, teadma
18	148	Video	video, kanal, youtube, salvestatud

19	147	Juhiload	sõidueksam, auto, ark, juhiluba
20	136	Töötamine	töö, palk, töökoht, töötama
21	131	Ühistransport	buss, sõitma, rong, ühistransport
22	130	Tervishoid	arst, perearst, haigekassa, ravim
23	119	Rahvus	eestlane, eesti, soome, sündinud, rahvus
24	115	Vaimne tervis	laps, tervis, inimene, vaimne, tundma
25	98	Energiaallikas	korter, elekter, hind, kütus, gaas, tankima
26	98	Suitsetamine	tubakas, evedelik, esigarett, keelama
27	94	Finants	maks, pank, euro, maksma, raha
28	94	Autokool	autokool, luba, sõidutund, autosõit
29	93	Küsitlus	küsimustik, küsitlus, vastama, vastaja
30	91	Auto ostmine	auto, kasutatud, ostma, liising
31	90	Tähtpäevad	jõulud, sünnipäev, jõululaud, jook
32	85	Kinnisvara	hind, kinnisvara, maamaks, maja
33	84	Kuller- ja taksoteenus	bolt, wolt, takso, sõitma
34	80	Telesaade	hooaeg, vaatama, tv3, kanal
35	69	Arvuti	arvuti, sülearvuti, läpakas, ssd
36	69	Hambaravi	hambaarst, hammas, hambaravi, ortodont
37	65	Juriidika	seadus, tööandja, tahteavaldus, õigus
38	56	Elektroonika	euronics, playstation, ps5, müüma
39	51	Pension	sammas, pension, raha, III, II

40	48	Lennundus	lendama, lennanud, lennuk, lend
41	47	Maski kandmine	mask, kandma, piirang, pood

V. Teemade analüüsi tulemused - EN

Tabel 12. Teemade analüüsi tulemused - EN.

Järjekorra number	Postituste arv	Teemad	BERTopic-u sõnaesitused
1	1951	Eesti/Tallinna külastamine	tallinn, estonian, place, know
2	390	Elukoht Tallinnas	apartment, tallinn, rent, place
3	217	Haridus	university, student, study, degree
4	191	Venemaa	russian, russia, estonia, country
5	190	Eestlased	estonian, people, language, estonians
6	180	Töövõimalused Eestis	job, work, estonia, working
7	144	Eesti keele õpe	learn, language, estonian, course
8	144	Suhted naaberriikidega	estonian, finnish, people, think
9	132	Eesti muusika	song, music, estonian, playlist
10	123	Lingvistika	word, phrase, translation, help
11	121	Eesti tervishoid	doctor, estonian, medical, medication
12	108	Küsimustik	survey, study, participation, result
13	103	Kirjandus	book, estonian, history, read
14	99	Filmid ja sarjad	movie, subtitle, watch, film, tv
15	99	Finants	bank, account, swedbank, payment
16	90	Ettevõtlus	company, business, bank, eresidency
17	87	COVID-19	covid, restriction, vaccinated, quarantine

18	75	Tasuvad töökohad	salary, company, job, work
19	71	Telekommunikatsioon	internet, sim, telia, tele2
20	71	Poliitika	party, election, vote, politics, ekre
21	68	Viisa	permit, residence, visa
22	61	Kodakondsus	citizenship, estonian, passport, birth
23	57	Maksud	tax, pay, income, dividend, taxation
24	49	Eesti tunnustamine	country, happy, estonia, thank, beautiful
25	45	Digiteenused	id, card, register, domain, phone
26	42	Eesti nimed ja hääldus	name, estonian, pronounce, pronunciation
27	34	Treening	gym, pool, martial, training
28	33	Sisseastumine	taltech, exam, test, entrance
29	28	Juukselõikus	hair, haircut, hairdresser, barbershop

VI. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Tauno Tamm**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose
   Eesti alamredditi korpuse loomine ning analüüs,

   mille juhendaja on Siim Orasmaa (PhD),

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

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

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

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

Tauno Tamm

14.05.2024