Tartu Ülikool
Sotsiaalteaduste valdkond
Ühiskonnateaduste instituut
Ajakirjanduse õppekava

# Suurandmete kasutamisvõimalused sotsiaalmeedia analüüsis: Ameerika Ühendriikide presidendivalimiste Twitteri kajastuse meelestatuse analüüs

Magistritöö

Autor: Priit Pokk
Juhendaja: Külliki Seppel

Tartu 2017

# Sisukord

# 1. Sissejuhatus

Andmed on üks vahend teadmiste omandamiseks. Nende tõlgendamine võimaldab luua uut informatsiooni, olgu ta misiganes vormis. Selleks, et seda teha, on vaja andmeid vajadusel töödelda ning neist aru saada. Tänapäeva tehnoloogilised vahendid võimaldavad koguda andmeid massiivsetes kogustes, mis varasemalt poleks võimalik olnud. Kui andmestik ületab teatud kriteeriumite lävendi, võib neid seetõttu nimetada suurandmeteks - andmestik, mille manuaalne kogumine ja analüüsimine oleks olnud kas võimatu või ääretult palju aega ja ressurssi nõudev.

Suurandmed ei sisalda iseenesest koheselt väärtust (informatsiooni), vaid sisalduv väärtus tuleb andmetest üles leida või isegi võib öelda, et tuleb väärtust luua. Selleks tuleb andmeid süstematiseerida, filtreerida ehk koguda, töödelda ja analüüsida. Suurandmetest informatsiooni hankimine ei tohiks jääda kättesaamatuks ka neile teadusharudele ja teadlastele, kelle tehnoloogilised oskused ja võimekused ei päde samas rütmis arvutiteadlastega. Selleks on vaja esmalt ka julgust uurida lähemalt, mida on suurandmetest võimalik üldse kätte saada ning kuidas.

Selleks pühendan ka oma magistritöö suurandmete uurimise metodoloogilisele analüüsile. Konkreetsemalt kavatsen töö empiirilises osas vaadelda mikroblogi Twitter ning seda, kuidas 2016. aasta Ameerika Ühendriikide presidendivalimised 24 tunni jooksul seal välja nägid. Andmete kogumiseks kasutan tööriista DMI-TCAT (Digital Methods Initiative - Twitter Capture and Analysis Toolset). Andmete analüüsimiseks kasutan automatiseeritud meelestatuse analüüsi, mis jaotab Twitterist kogutud säutsud erinevate emotsioonide alla. See aitab vaadelda seda, kuidas Twitteri kollektiivne emotsioon muutub võrreldes ajajoones toimuvate päris sündmustega. Töö põhirõhk on siiski pigem analüüsida üldisemalt suurandmete ja metoodikate võimalusi ning piiranguid.

Twitter sai valitud meediumina seetõttu, et esiteks on andmete kogumise võimalused võrdlemisi lihtsad. Teiseks on see üsnagi vahetu ning kiire meedium ehk kui midagi juhtub, siis kõigest sekunditega on võimalik selle kohta reaktsioone leida. See tuleneb paljuski Twitteri mahupiirangust ehk sellest, et Twitteris on säutsud piiratud 140

tähemärgiga, mis tähendab, et postitused on lühikesed ja kompresseeritud. Sama faktor ühtlustab ka andmete teatud ühetaolisust - ükski säuts ei saa olla teisest ülemäära pikem. Neljandaks see faktor, et teemade jaotus on Twitteris võrreldes näiteks Facebookiga paremini kättesaadav: teemaviidete (*hashtagide*) kaudu on võimalik tuvastada, mida erinevad osapooled sündmustest kirjutavad. Tihti on kasutusel teemaga seotud spetsiifilised märksõnad (näiteks antud juhul #electionday, ETV Foorumi saate puhul #etvfoorum).

Ka Eestis on suurandmete uurimine muutunud eri valdkonniti populaarsemaks. Näiteks korraldab Eesti Digihumanitaaria Selts konverentse, seminare ja töötubasid, et populariseerida suurandmete teaduslikku interdistsiplinaarsust (Sarv ja Laineste, 2016). Tartu Ülikoolis Skytte instituudis keskendub infotehnoloogia mõju-uuringute keskus (CITIS) suurandmete analüüsimisele võttes aluseks Eesti e-teenuste andmed (CITIS, 2017; Vits, 2016). Tallinna Tehnikaülikool otsib endale sotsiaalteaduslike suurandmete professorit, kes kujundaks TTÜs välja selle valdkonda (TTÜ, 2017). Ma ei suutnud leida suurandmete kasutamist ajakirjanduse ning meediaanalüüsi valdkondade uuringutes Eestis, mistõttu peaks selle töö panus väljenduma selle valdkonna osalisel avamisel. Töö üks eesmärk on seega luua pinnast täiendavaks suurandmete uurimiseks sotsiaalteaduste valdkonnas Eestis.

Töö koosneb tinglikult öeldes kahest osast: kirjanduse ülevaatest ja töö empiiriline osa. Esimeses kirjeldan, mis on suurandmed, kuidas neid analüüsitakse, mis suunad on suurandmete kasutamises sotsiaalteadustes, mis on meelestatuse analüüs, mis on Twitter ning mida on selle kohta uuritud. Töö teises osas analüüsin nii töö metoodikat kui ka tõlgendan andmeanalüüsi tulemusi.

# 2. Suurandmed (ehk *big data)*

## 2.1 Mis on suurandmed?

Selleks, et mõista, mida suurandmetega on võimalik teha, tuleb esmalt aru saada, mis suurandmed on. *Big data* ehk eesti keeles suurandmed (EKI, 2017) on "*suured mahud struktureeritud ja struktureerimata andmeid, mille haldamine tavapäraste relatsioonilise andmebaasi- ja andmetöötluse vahenditega on raskendatud, kui mitte võimatu (maht, formaadid, kiirus)*". Siinkohal tuleb mainida, et see on vaid üks võimalikest suurandmete definitsioonidest. Universaalset suurandmete definitsiooni pole suudetud ühiselt leida.

Üks levinumaid suurandmete definitsiooni aluseid pärineb Meta Groupi rapordist (Laney, 2001: 1-2), mis mõistab suurandmeid läbi kolme V tähe (inglise keeles): suur maht (*volume*), suur kiirus-tempo (*velocity*), suur mitmekesisus (*variety*). Veel on seda mõistmist täiendatud kolme V tähega mõistega: tõepärasus-õigsus (*veracity*), varieeruvus (*variability*), väärtus (*value*) (Sivarajah et al, 2017; Ward ja Barker, 2013; Gandomi ja Haider, 2015).

Mahu all viidatakse andmestiku suurusele: suurandmeid mõõdetakse enamasti terabaitides ja petabaitides. Siiski on suurandmete mahu suurus sõltuvuses erinevatest faktoritest: näiteks andmete tüübist ja ajast. Üks gigabait videovormingus andmeid ei nõua sama suurt ressurssi andmete kogumiseks ja töötlemiseks kui üks gigabait tekstivormingus andmeid. Ajaline mõõde on sõltuvuses tehnoloogia arenguga: andmestik, mis vajas tehnoloogiliselt nii kogumiseks kui töötlemiseks superarvutite võimekust 30 aastat tagasi, on tänapäeval väga lihtsasti teostatav tavapärase arvutiga. Suurandmete mahu kehtivat lävendit on seega keeruline defineerida, kuna see on sõltuvuses sellest, milliseid andmeid millisel ajahetkel milleks kasutatakse; samuti mis erialaselt suurandmetest kõneldakse. (Gandomi ja Haider, 2015: 138) Suurandmete mahuga on ka seotud nende üks peamisi väljakutseid: kuidas optimaalselt koguda, töödelda ning hoiustada andmeid. (Sivarajah et al, 2017: 269)

Suurandmete kiiruse all viidatakse sellele, millise tempoga andmeid kogutakse ning mis kiirusega on andmeid võimalik analüüsida. Nutitelefonite ja sensorite lai levik on

viinud selleni, et andmeid luuakse enneolematu kiirusega, mis näiteks annavad potentsiaalset materjali reaalajas andmete analüüsiks (Gandomi ja Haider, 2015: 138). Kiirusega seotud väljakutse seisnebki selles, kuidas esiteks luua süsteemne lähenemine mitte-homogeensele andmestikule, mille alusel luua tähendust ja väärtust; ning teiseks, kuidas peaks neid andmeid ka siis töötlema (Sivarajah et al, 2017: 273).

Suurandmete suur mitmekesisus viitab andmestiku struktuursele heterogeensusele ehk andmestik koosneb eritaolistest andmete struktuuridest. Andmestikud võivad eksisteerida nii struktureeritul (nt tabelikujul andmestik), pool-struktureeritul (nt *Extensible Markup language* ehk xml andmestik) kui ka struktureerimata kujul (tekstid, pildid, video) (Gandomi ja Haider, 2015: 138). Mitmekesisusega seotud väljakutse seisneb eritaoliste andmestikuvormide ühtlustamises (Sivarajah et al, 2017: 269).

Suurandmete tõepärasus viitab sellele, et suurandmed põhinevad mõningaselt andmestikel, mis võivad olla ebatäpsed ja ebamäärased (Gandomi ja Haider, 2015: 139). Suurandmete kasutamisel tuleb seega mõista andmestiku iseloomu: kus peituvad andmestike puudused ja nõrkused ning seega ka sellest tulenevalt, kuidas mõista andmete töötlemisel andmete väljundit. Andmete kvaliteedi hindamisel tuleb olla teadlik nende omadustest (Sivarajah et al, 2017: 269 ja 273).

Varieeruvus ehk andmete komplekssus viitab esiteks andmevoo sagedusele: sisendist tulenev andmete maht võib olla pidevalt ajas muutuv (Gandomi ja Haider, 2015: 139). Teiseks, sellest tulenevalt on andmete olemus ning seega ka nende *tähendus* pidevalt ajas muutuv, kuna andmestik, mida analüüsitakse on muutuv. Lisaks viitab see mõiste sellele, et andmestiku sees võib sama andme (nt sama sõna) väärtus olla sõltuvalt kontekstist erinev ning andmete töötlemine sellisel kujul on ääretult keeruline, mis suudaks seda konteksti tuvastada (Sivarajah et al, 2017: 273).

Võrreldes andmestiku mahu proportsionaalsusega on suurandmete väärtus võrdlemisi väikene. See tähendab, et suurandmete väärtust peab looma. Nii on näiteks töötlemata suurandmete väärtus võrdlemisi tühine kui võrrelda väikeandmetega (Gandomi ja Haider, 2015: 139). Suurandmete väärtus on kuskile andmete sisse peitunud, miski

mis eksisteerib, kuid väärtuse saamiseks tuleb andmeid kaevandada. (Sivarajah et al, 2017: 273)

Boyd ja Crawford (2012: 663) võtavad suurandmete olemuse kokku kolmepoolse suhtena: suurandmed on nähtus, mis baseerub tehnoloogilisel (ehk arvutuslikul võimekusel, algoritmilisel täpsusel), analüütilisel (s.t suured andmestikud loovad võimaluse märgata mustreid, mis aitavad teha majanduslike, sotsioloogilisi, tehnilisi otsuseid) ja mütoloogilisel (usk, et rohkes koguses andmeid võimaldavad aru saada teadmistest, mida väiksem andmestik ei võimalda) koosmõjul. Suurandmed ei ole lihtsalt andmed, mida on rohkesti, vaid see hõlmab endast võimekust otsida, koondada ja ristviidata suurtele andmestikele (boyd ja Crawford, 2012: 663).

Ward ja Barker (2013: 2) võtsid suurandmete definitsiooni puhul samuti kokku kolm omadust, mis leidub pea igas suurandmete mõistes: suurus, komplekssus ja tehnoloogia. Seega kokkuvõtvalt: suurandmed on termin, mis kirjeldab suurte ja/või komplekssete andmestike ladustamist ja analüüsimist kasutades erinevaid tehnoloogilisi lahendusi (sh näiteks masinõpet) (Ward ja Barker, 2013: 2).

## 2.2 Suurandmete kogumise ja analüüsimise protsess

Agrawal et al (2012: 3) jaotavad suurandmete kogumise ja analüüsimise protsessi viieks staadiumiks:

1. andmete kogumine/salvestamine
2. vajaliku informatsiooni väljavõtmine (andmete kaevandamine), andmete puhastamine ja märgistamine
3. andmete integratsioon, agregatsioon ja representatsioon
4. päringu rakendamine, andmete modelleerimine ja analüüs
5. tulemuste tõlgendamine.

Andmed ei teki iseenesest, vaid nende kogumiseks tuleb luua vastav süsteem. Andmete kogumise faasis tuleb juba eelnevalt arvestada sellega, et paralleelselt vaatlusandmete korjega kaasneks automaatselt sobiva metaandmestiku kogumine, mis kirjeldab täpselt seda, *mis* andmeid salvestatakse, *kuidas* neid andmeid salvestatakse

ning kuidas neid andmeid mõõdetakse. See on oluline protsess tulevaseks andmete tõlgendamiseks (Agrawal et al, 2012: 4).

Andmekaevandamine on protsess, mille käigus avastatakse massiivsete andmehulkade seast huvitavaid seoseid ja mustreid. Andmekaevandamine on teisisõnu teadmiste avastamise protsess, kus andmeid puhastatakse (eemaldatakse müra ja ebaühtlaseid andmeid), integreeritakse (mitmeid andmeallikaid ühendatakse), valitakse (vastavalt analüüsi vajadustele), transformeeritakse (vormidesse, mis on sobivad andmekaeveks, nt andmete kokkuvõtmine ja agregeerimine), avastatakse mustreid ning esitletakse teadmisi. Spetsiifilisemalt on andmekaevandamine protsess, mis tuleb peale andmete transformeerimist, kus mustrite eraldumiseks rakendatakse teatud meetodeid (algoritmide rakendus, masinõpe jne). Andmete allikateks võivad olla näiteks andmebaasid, andmehoidlad, veeb, mõned teised informatsioonihoidlad või ka andmed, mida voogesitatakse dünaamiliselt süsteemi (Han, Pei ja Kamber, 2011: 6-8, 33).

Andmete puhastamise lõppeesmärk on andmete kvaliteedi parandamine. Andmete kvaliteet sõltub täpsusest, täiuslikkusest (kas on midagi puudu), pidevusest (kas näiteks on andmevoog olnud pidev), aegsusest, usutavusest ja interpreteerituvusest. Andmete puhastamise rutiin püüab täita puuduvaid väärtuseid, siluda mürarikkaid andmeid, tuvastada erandeid ja vajadusel neid eemaldada ning korrigeerida ebaühtlused. Andmete puhastamine toimub enamasti kahes faasis: lahknevuste tuvastamine ning selle alusel andmete transformeerimine (Han, Pei ja Kamber, 2011: 85 ja 120). Tõenäoline on, et pärast andmete puhastamist ning korrigeerimist jääb andmestikku siiski alles erinevaid puudusi ning vigu, mis on paratamatus. Andmeanalüüs peab juhinduma sellest lähtuvalt (Agrawal et al, 2012: 8).

Andmestik (*data set*) koosneb andmeobjektidest ehk üksustest, mida kirjeldavad teatud atribuudid. Nende atribuutide väärtused võivad olla kujutatud nii binaarselt (ehk vastandlikud, üks või teine, 1 või 0, õige või vale), nominaalselt (sümboleid või nimetusi kasutades), järgarvuliselt (teatud järjekord, kuid vahemiku suurused ei pruugi olla teada) või numbriliselt (mõõdetav kogus) (Han, Pei ja Kamber, 2011: 79).

Suurandmete statistilisel lähenemisel peab arvestama ka veel mõne täiendava probleemiga, mis esineda võib. Nii võib müra eemaldamisel minna kaotsi mõned erandjuhtumid, mis võivad olla olulise seletava tähendusega (Gandomi ja Haider, 2015: 143). Näiteks erineb ühelt sensorilt saadud informatsioon märkimisväärselt teistest, mis võib viidata sensori defektsusele kui ka sellele, et sealt saadav informatsioon võib avaldada uusi valgustavaid tahke (Agrawal et al, 2011: 4). Samuti võib andmestiku massiivse suuruse tõttu tekkida teaduslikult sõltumatute muutujate vahel suure dimensionaalsuse tõttu petlik korrelatsioon ehk juhuslik kokkusattuvus (Gandomi ja Haider, 2015: 143). See väljendub tulemuste tõlgendamisel apofeenias: märgates mustreid, kus seda ei eksisteeri. Lisaks peab arvestama sellega, et kui andmestik koosneb miljonitest andmeobjektidest, ei tähenda see kindlasti, et andmestik oleks 1) juhuslik 2) representatiivne (boyd ja Crawford, 2012: 668).

Suurandmete analüüsi ehk väärtuste loome väljundina võib jaotada viieks: kirjeldav analüütika ehk hetkeolukorra kaardistamine; uuriv analüütika ehk teatud konkreetsete ettepanekute kinnitus/tagasilükkamine; ennustav analüütika ehk potentsiaalsete stsenaariumite tuvastamine; normatiivne analüütika ehk mida peaks ette võtma; ning ennetav analüütika ehk mida peaks rohkem tegema (Sivarajah et al, 2017: 266). Antud töö praktilises osas kasutan väljundina esimest ehk kirjeldavat analüütikat.

## 2.3 Suurandmed sotsiaalteadustes

Sotsiaalteaduslikus uurimistöös suurandmete kasutamisel tekib paratamatult küsimus kättesaadavuses - seda nii oskuste baasil kui ka andmete loomises. Suures koguses ning keerukuses andmete kogumine, nende analüüsimine (vastavate algoritmide loomine ja kasutamine) ning eelnevalt rakendusliidestest (APIdest) läbi närimine vajab teatud tehnilist pagasit. Siin on eelisseisus need, kelle jaoks on see loomulik keskkond ning suudavad vastavaid süsteeme luua; või ka need, kes omavad sujuvat koostööd vastava ala spetsialistidega. Juhul kui tehnilised oskused on töö protsessis kõrgendatud tähtsusega kohal, paneb see sotsiaalteadlased keerukasse olukorda - kas nad on võimelised rakendama ja kehtestama oma teadusliku pädevust (boyd ja Crawford, 2012: 674). Suurandmete analüüsis valitseb pigem arvutiteadlaste ja äriloogikale allutatud lähenemine metoodikatele ja uuringutele: olulisem on see, et

miski töötaks, mitte see, miks ja kuidas miski töötab (McFarland, Lewis ja Goldberg, 2016: 32). Sotsiaalteadlased üritavad pigem analüüsida seda, mis on kättesaadav, kuid metoodikate otsustusprotsessi mõjutamisel jääb mõjutusroll sekundaarseks (Ruths ja Pfeffer, 2014:1063).

Koostöös vastava ala spetsialistidega võib tekkida omaette risk, kus sotsiaalteadlased mõistavad oma töös rakenduvaid matemaatilisi ja tehnilisi ülesandeid "musta kastina" (ehk kui midagi müstilist, teadmatut konstanti) ning seetõttu ei panda neid metoodikaid kahtluse alla ning ei problematiseerita. See piirab ka tulemuste tõlgendamise võimalust - kui ei ole aru saada, mis loogika alusel teatud tulemused on täpselt saadud, siis on keeruline teha üldistusi (Bruns, Burgess ja Highfield, 2014: 12). Teatud tehniline pagas võib olla samas ka nõrkusekoht - ülehinnatakse oma võimeid ning üritades oma oskusi rakendada ei suudeta märgata oma nõrkusi, kuna see pole oma spetsialiteet. Vastupidiselt võidakse alahinnata arvutiteadlaste võimekust ning seega täpsemate tulemuste saamiseks ei kasutata ära nende oskusi luua täpselt selliseid tehnilisi lahendusi, mis uurimispüstitusega kokku läheb.

Sotsiaalteaduste valdkonnas suurandmete uurimise teine piirang peitub andmete kättesaadavuses - suurandmeid tootvad ning hoiustavad ettevõtted kontrollivad seda, millised andmed on kättesaadavad, millised mitte. See on osa ärist. Teadlaste jaoks jääb ligipääs paratamatult finantsilise võimekuse taha - täpselt samadel põhjustel ei suudeta ka mõningatel juhtudel neid andmeid ise luua. Isegi kui mõni suurandmeid käsitlev ettevõte (näiteks sotsiaalmeedia ettevõtte) võimaldab teatud teadlastel ulatuslikumat ligipääsu, siis teadlane jääb mõneti siiski ettevõte ohjesse - uuringut püstitades jääb teadlase teadvusesse ka see, et kuidas tasakaalustades luua suurandmetest teadmisi ning sõltumatuid tulemusi, aga samas hoida alles ligipääs nendele samadele andmestikele. Selline olukord võib ka kallutada uurimisküsimusi (boyd ja Crawford, 2012: 674).

Suurandmed sotsiaalteadustes ei pruugi olla sama suured kui mõnes teises valdkonnas: andmestiku suuruse hindamiseks võib hindamise kriteeriumina võtta seda, kui palju aega ja ressurssi kuluks tavapäraseid vahendeid kasutades vastavate üksuste manuaalseks kodeerimiseks ning analüüsimiseks. Nii võib hinnata ka näiteks

10 000 üksusega andmestike suurteks. Suurandmete määratlemine sõltub seega kontekstist (Guo et al, 2016: 2).

### 2.3.1 Metodoloogilised arengud

Protsessi, kus teadlased kasutavad suurandmete uurimiseks varasemalt (mittedigitaalses maailmas) väljakujunenud metoodikaid võib nimetada metoodika digitiseerimiseks. Samas kui analüüsimeetodid on mugandatud arvestamaks platvormide olemuslikkust, siis võib neid nimetada digitaalseteks meetoditeks (Felt, 2016: 8). Digitaalsed meetodid on termin, mida kasutatakse veebi kui sotsiaalse ning kultuurilise nähtuse uurimiseks, mis vaatleb kriitiliselt veebi andmete kvaliteeti, *online* andmekogumis- ja analüüsimeetodite tootlikkust ning üldisemalt vaatleb veebi kui objekti, mille põhjal teadmisi koguda (Rogers, 2015: 1).

Gray (2007: xviii-xix) argumenteerib, et suurandmete tulekuga on teadus läbimas meetodite paradigmaatilist nihet: nähtuste mõtestamine läbi simulatsioonide ehk näiteks representatiivsete küsitluste asemel põhineb uus paradigma rohkem eksperimentaalsusel ning avastamisel. Teadlane tutvub võrdlemisi hilises staadiumis andmestikuga lähemalt ning kujundab oma uuringu andmestiku põhjal. Kitchin (2014: 2, 6, 8) toob välja, et suurandmete analüütikas on protsess muutunud vastupidiseks: kui muidu teadlane kujundab enda andmestiku vastavalt vajadusele ning otsib oma teooriatele ja hüpoteesidele toetust, siis suurandmete analüüsis tuleb enne andmestik ning seejärel on teadlase ülesanne sealt teadmisi välja kaevandada. Selle asemel, et uuring oleks lõpp-punkt, kus kontrollitakse hüpoteese ja saadakse vastuseid uurimisküsimustele, võib suurandmestiku analüüsi vaadata ühest vaatevinklist pigem just alguspunktina, kust saada teadmisi ja hüpoteese, mida kontrollida.

Kitchin (2014: 3) vaatab samas kriitiliselt erinevaid postulaate, mis on tekkinud suurandmete laiema tulekuga teadusesse. Need neli postulaati on: suurandmed suudavad tabada kogu pilti; pole vaja *a priori* teooriat, mudeleid ega hüpoteese; tänu automatiseeritud analüütikale suudavad andmed ise rääkida ilma inimese kallutatusest; tähendused suudavad ületada konteksti ja teadmisi ehk tõlgendusi on võimalik luua ilma taustateadmisteta.

Kitchin (2014: 3-5) oponeerib, et need postulaadid on paljuski tekkinud suurandmete käsitluse ärilise mõtteviisi tõttu. Kui suurandmed üritavad tabada võimalikult laialdast pilti, siis on need siiski teatud määral representatsioon ning tehnoloogilised protsessid on kujundatud inimeste poolt. Suurandmed ei teki mitte millestki, vaid tehnoloogiate aluseks on siiski teaduslik mõtestatus. Andmed ei suuda ise kõneleda ilma inimese raamistuseta ning nõuavad varasemaid toetuspunkte (teooriaid, avastusi jne). Analüütika, mis ei mõtesta konteksti võib olla pragmaatiline, kuid on siiski pinnapealne.

McFarland, Lewis ja Garland (2016: 15, 31) kirjeldavad, et suurandmed on uut tüüpi andmed, mistõttu tuleb nii kohandada vanu analüüsitehnikaid kui võtta kasutusele uusi. Sotsiaalteadlased peavad kohanema arvutiteadlaste pragmaatilisemate lähenemistega. Nad pakuvad välja, et sotsiaalteadused peaksid seoses suurandmetega liikuma ühe variandina niiöelda kohtuliku (*forensic*) lähenemise teed, kus induktiivne ja deduktiivne lähenemine on kombineeritud selleks, et ühendada rakenduspõhist ning teooriast juhinduvat perspektiivi.

### 2.3.2 Andmeagendid ja subjektsus

Suurandmete ja avalikkuse suhe on komplitseeritud: suurandmed on ühtepidi nii avalikkuse kirjeldus, ressurss avalikkuse jaoks kui avalikkuse poolt loodud ja kujundatud toode. Avalikkus on nii teadmiste subjekt kui objekt, autor ja tekst; informeeritav, informant kui informatsioon. Üha laiema kasutuse juures on inimestel kui andmeid loovatel subjektidel üha raskem kontrollida, kas nad soovivad selles protsessis osaleda. Sealjuures võidakse olla andmete loojad, kuid mitte sellest kasusaajad. Tinglikult võib jaotada avalikkust seoses suurandmetega neljaks: kasutajad (tarbivad suurandmeid), andmete loojad, eksperdid (kes on digitaalse kirjaoskusega ning suudavad kaasa rääkida suurandmete ümber toimuvate protsesside üle) ning kujundajad, kes loovad tehnoloogiaid ning protsesse suurandmete loomiseks (Michael ja Lupton, 2016: 105, 108-110).

Inimesed ise on tänapäeval aktiivsed suurandmete loojad läbi selle, et kasutavad tehnoloogiaid enese praktikate jälgimiseks. See tähendab, et inimesed panevad ennast andmeagentidena subjektideks seetõttu, et suurandmetest kasu saada. Ennast, enda

liigutusi, enda füsioloogiat üritatakse seeläbi kvantifitseerida. Enda praktikate jälgimist võib jaotada viieks: privaatne (isiklikuks tarbeks enda andmete jälgimine), suunitletud (väliste mõjurite tõttu ajendatud jälgimine), kogukondlik (inimesed jagavad üksteisega andmeid), kehtestatud (inimesel endal on enese jälgimine sunnitud ühel või teisel põhjusel väliste mõjurite tõttu) ja ekspluateeritud (inimese teadmata kasutab keegi sinu kohta andmeid enda kasuks ära) (Lupton, 2016: 102-103).

Couldry ja Powell (2014: 1-2) rõhutavad, et suurandmete uurimises on vajalik samuti rohkem tähelepanu pöörata sotsiaalsele analüütikale. See on lähenemine, mis sotsioloogiliselt uurib, kuidas eri subjektid (inimesed) kasutavad analüütikat. Analüütika on sotsiaalne protsess, mis hõlmab endast reflektsiooni, seiret ning tingimustega kohanemist. Laiem sotsioloogiline huvi peaks tekkima olukordades, kus andmeanalüüsi käigus tekib reaalne või potentsiaalne pinge selle vahel, mis subjekt soovib teha ning oma tegevuste tõlgendamises.

### 2.3.3 Sotsiaalmeedia kui uurimisallikas

Sotsiaalmeedia on üks kättesaadavamaid, kuid siiski piiratud funktsionaalsusega uurimisallikaid suurandmete kasutamiseks. Sotsiaalmeedia kätkeb endas nii pilku inimeste sotsiaalsetele pratikatele kui käsitlevatele diskursustele. Selle puhul tuleb arvestada sellega, et digitaalsed meetodid, mida kasutatakse uurimistööks võivad pidevalt ajas muutuda, kuna platvormid ning nende rakendusliidesed (APId) võivad muutuda ja seetõttu kasutatavad tööriistad, võivad kiiresti aeguda. Nii juhtus näiteks Twitteriga 2011. aastal, millest tuleb järgnevates peatükkides ka juttu (Rogers, 2015: 9).

Felt (2016: 4-5) tuvastas, et teadlaste seas pole kommunikatsiooniuuringutes tööriistade kasutamine sotsiaalmeedia põhjal suurandmete kogumisel ja analüüsimisel väga populaarne: umbes 17% EBSCO andmebaasist kogutud teadusartiklitest kasutasid vastavaid tööriistu, mille keskseks märksõnaks oli sotsiaalmeedia ning täiendavate terminitena suurandmed ja analüütika. Seega domineerivad sotsiaalmeedia uuringutes veel enamasti traditsioonilised meetodid. Felt selgitas ühe võimaliku põhjusena seda, et enamasti uuritakse Facebooki platvormi, mille

suurandmete kättesaadavus rakendusliidese (API) tõttu pole väga sõbralik (päringuid saab teostada avalikele gruppidele, aga mitte isiklikele profiilidele). Seevastu uurimistööd, mis koguvad, ostavad või tuvastavad sotsiaalmeedia andmestikke, kasutavad peaaegu eksklusiivselt uurimisobjektina Twitterit (Felt, 2016: 4-5). Twitteri uuringuid soodustab ka see, et võrdlemisi väike osa kasutajatest hoiavad oma profiili privaatsena: vaid umbes 12% (Beevolve, 2012). Twitteri omaduste ja uuringute kohta tuleb täpsemalt juttu hiljem.

Ruths ja Pfeffer (2014: 1063) pakuvad sotsiaalmeedia andmestiku uurimiseks nõuandeid, mida peaks metodoloogias silmas pidama. Andmete kogumise faasis peaks arvestama platvormi spetsiifilisi kallutusi; mõistma, mis andmestik on üldse kättesaadav ning mõistma, keda seal kujutatud populatsioon representeerib. Samuti soovitavad nad võrrelda uute metoodikate puhul andmeanalüüsi traditsionaalsete metoodikatega ning testida uusi metoodikaid ning klassifitseerimist eri andmestikel.

### 2.3.4 Sotsiaalteadlaste uuringud eri portaalide põhjal

Üks prominentsemaid ja vastuolulisemaid teadusartikleid seoses Facebooki ja suurandmetega oli juhtum, kus tavapärane metoodika oli osaliselt tagurpidi pööratud. Nimelt kui tavapäraselt kogutakse andmestiku, mida miski või keegi on loonud, siis Kramer, Guillory ja Hancock (2014: 8788) eksperimenteerisid inimestega selliselt, et esmalt nad manipuleerisid sotsiaalmeedia algoritmidega nii, et üks grupp inimesi nägi oma Facebooki ajajoones rohkem negatiivseid, teine grupp inimesi rohkem positiivseid postitusi. Seejärel aga hinnati, kas see mõjutas nende samade inimeste postitusi. Kasutades Facebooki poolset luba oli uurimisobjektideks ligi 700 000 inimest. See uurimistöö põhjustas vastukaja nii eetilisel tasandil kui ka metodoloogilisi küsitavusi ehk kui tõesed olid andmed (emotsionaalsuse hindamise valiidsuse kui ka põhjus-tagajärg-seose hindamisel) (Panger, 2016: 1109).

Wikipedia on samas võrreldes Facebookiga vägagi kättesaadav massiivne andmekogum: artiklite andmestik ja meta-andmestik on avalik, sh Wikipedia kogu toimetamiste ajalugu ning diskussioon; toimetajate IP aadressid (mis võimaldavad tuvastada asukohti). Lisaks saab ka võrrelda, kuidas toimetamise protsessid toimuvad eri keeltes. Andmed on iseenesest pidevalt muutuvuses, kuna Wikipedia pole

statsionaarne dokument, vaid orgaaniline objekt. Kuigi andmed ise on avalikud, siis väljakutsed peituvad süsteemses andmekogumises ja töötlemises (Rogers, 2015: 11). Tööriist Contropedia võimaldab näiteks visuaalselt nii tuvastada artiklite ajas kujunemise jooksul vastuolulisemaid teemapunkte kui ka võrgustikuna suhteid erinevate toimetajate vahel (näiteks kes on vastandlikult toimetanud samu punkte) (Borra et al, 2015: 711-712). See tööriist on samas näide, kus on ka oluline tutvuda "musta kastiga" (ehk kuidas täpsemalt Contropedia tööriist töötab), et tõlgendused oleksid pädevad.

Kui eelmainitud platvormide puhul on analüüsi keskmeks enamasti tekst, siis näiteks Flickri platvormi uurides on selleks pildid. Rattenbury, Good ja Naaman (2007: 103, 109-110) suutsid sealse keskkonna kaudu (objektideks fotod, selgitavad märksõnad ning meta-andmestik) automaatselt tuvastada aset leidvaid sündmusi võttes aluseks piltide põhjal märgatud mustreid.

Uurimiseks kasutatavaid sotsiaalmeedia informatsiooniallikaid võib tinglikult jagada kaheks eri osaks: kasutajate loodud sisu (pildid, video, tekst) ning võrgustikud ja nende sees aset leidvad interaktsioonid (kesksel kohal sel juhul kasutajad ning dünaamikad teiste sotsiaalmeedia kasutajatega). Sellest lähtuvalt võib sotsiaalmeedia analüütikat jaotada kaheks grupiks: struktuuripõhine (võrgustike) ning sisupõhine analüütika. (Gandomi ja Haider, 2015: 142) Need kaks peamist suunda on ka järgnevas kahes peatükis fookuses, mille eesmärk on kirjeldada, mida need metoodikad ennast suurandmete maailmas kätkevad.

## 2.4 Suurandmed ja võrgustike analüüs

Struktuuripõhist analüütikat võib teisiti nimetada sotsiaalvõrgustike analüüsiks. Seda tüüpi analüütika üritab teadmisi ammutada sünteesides sotsiaalvõrgustike struktuuri atribuute ja osalevate üksuste vahelisi suhteid (Gandomi ja Haider, 2015: 142) .Lihtsustades uurib see analüütika sotsiaalmeedia võrgustikes toimuvate suhtlemiste olemust. Marin ja Wellmann (2011: 22) defineerivad sotsiaalvõrgustike analüüsi sedasi, et see pole mitte teooria ega metodoloogia, vaid vahend probleemi nägemiseks – perspektiiv. Selle alusmõte on, et sotsiaalne elu luuakse peamiselt ja kõige

olulisemalt suhetest ja mustritest, mida need moodustavad (Marin ja Wellmann, 2011: 22).

Võrgustikud koosnevad sõlmpunktidest ja neid ühendavatest lülidest. Need esindavad vastavalt siis võrgustikus osalejaid (inimesed, organisatsioonid vms) ja nende vahelisi suhteid. Uurida võib esiteks sõlmpunktide vahel eksisteerivaid linke, mis on määratletud statsionaarse suhtena (sotsiaalmeedias väljendub see näiteks Facebookis sõbraks olemist, Twitteris kui üks kasutaja jälgib teist), et näiteks tuvastada kogukondi. Teiseks saab uurida tegevusel põhinevaid võrgustike, kus sõlmpunktid omavahel informatsiooni vahetades (*like*, kommentaar või muu selline) märgistavad reaalseid interaktsioone (Gandomi ja Haider, 2015: 142). Sotsiaalmeedias on selliste võrgustike uurimine märkimisväärselt lihtsustatud võrreldes samaväärse informatsiooni kogumisega päris maailmast. Samas pole suhtluse iseloom sotsiaalmeedias täpselt samasugune sotsiaalne interaktsioon, mis on näost näkku suhtlus (Bruns, 2012: 1328).

Boyd ja Crawford (2012: 671) viitavad sellistele võrgustikele kui artikuleeritud (*articulated*) ja käitumuslikule (*behavioral*) võrgustikele. Artikuleeritud võrgustikes määratlevad inimesed ise oma kontaktid, kasutades tehnilisi mehhanisme. See on üldsõnalisem definitsioon, kuhu alla saab määratleda ka reaalmaailmas olevaid võrgustike. Võrgustikesse võivad kuuluda nii sõbrad, kolleegid, kuulsused, avaliku elu tegelased, huvitavad tuttavad jne. Käitumuslikud võrgustikud tulenevad kommunikatsioonimustritest, interaktsioonidest. Kumbki neist võrgustikest ei ole siiski võrdeline inimese isikliku võrgustikuga - iga ühendus ei ole võrdne teisega, interaktsiooni sagedus ei võrdu suhte tugevusega (boyd ja Crawford, 2012: 671). Kombineerides mõlemaid meetodeid on võimalik tuvastada sotsiaalmeedias nii kasutajaid, kes on häälekamad kui ka neid, kes on kõige laiemalt ühenduses (Felt, 2016: 13).

Samuti on võimalik sotsiaalvõrgustike analüüsi kaudu tuvastada tekkinud kogukondi: sõlmpunktid, mis on sarnaste sõlmpunktidega paljuski seotud ning vahetavad omavahel interaktsioone. Selle alusel tekivad ühiste tunnustega klastrid. Sõlmpunktide vaheline suhtlus toimub enamasti kujunenud kogukonna sees. Sõlmpunktide ja lülide arvud võivad ulatuda miljonitesse. Bello-Orgaz, Hernandez-

Castro ja Camacho (2017: 125, 130) kasutasid Twitterist saadud andmeid, et tuvastada tekkinud vaktsiini temaatilisi kogukondi. Nad märgivad, et hea kogukond peaks olema modulaarne (eristuv teistest klastritest), tihe (nii sõlmpunktide asetuse, klastrite vaheliste asetuse kui ka suhete poolest) ning robustne (selgelt piiritletud klaster võib viidata andmestiku veale või valele samaväärsusele). Selle tehnika kasutamine peaks eelkõige tuvastama suuremaid ja selgemalt eristuvaid klastreid, kuid müra sisse võivad jääda märkamatuks väiksemad kogukonnad. Võib ka juhtuda sedasi, et klastrid ei joonistu selgelt välja ning seega selgelt eristuvaid kogukondi ei eksisteeri.

Inimesed mõjutavad sotsiaalvõrgustikes üksteist, mistõttu kuulub sotsiaalvõrgustike analüüsi tehnikate hulka veel sotsiaalse mõjukuse analüüs. Seetõttu modelleeritakse ning hinnatakse nii osalejate kui ka suhete mõjukuse taset (Gandomi ja Haider, 2015: 142). Seda tehakse kasutades kesksuse meetmeid (*centrality measures*): esiteks astmete kesksus (*degree centrality*) ehk kõige representatiivsemad sõlmpunktid on kõrgemate väärtusastmetega (kõige suuremate suhete arvuga osalejad). Teise variandina kasutatakse omavektori kesksust (*eigenvector centrality*), mis arvestab sellega, et mitte lihtsalt suhete arv pole tähtis, vaid veel olulisem on see, kes on selle suhte teises otsas. See tähendab, et mõõdetakse ka lülide kaudu ühenduvate teiste sõlmpunktide mõjukust ning seeläbi määratletakse sõlmpunktide tegelikku väärtust - interaktsioonid mõjukate sõlmpunktidega viitab sellele, et sõlmpunkt ise on mõjukas. Kolmandaks meetodiks on vaheloleku kesksus (*betweenness centrality,* eestikeelne termin Muts, 2004: 24), mis võtab ka arvesse sõlmpunktide vahelist distantsi ning arvestab neist lühima teekonnaga. See meede näitab kui keskne sõlmpunkt on, ühendades teisi sõlmpunktipaare võrgustikku (Bello-Orgaz, Hernandez-Castro ja Camacho, 2017: 130).

Neljas tehnika informatsiooni ammutamiseks on ühenduste ennustamine – sotsiaalvõrgustiku struktuure arvestades üritatakse ennustada tulevasi sõlmpunktide vahelisi ühilduvusi. Arvestades, et sotsiaalvõrgustike struktuurid pole igavesti samad, on selle tehnika mõte aru saada võrgustikus toimivatest dünaamikatest. Seda tehnikat kasutades üritab näiteks Facebook soovitada sulle kui kasutajale sinule sobivaid gruppe või tuttavaid inimesi, Netflix sulle meeldivaid filme ja sarju, Youtube meeldivaid videosid (Gandomi ja Haider, 2015:143).

Rakendusi sotsiaalvõrgustiku analüüsi kasutamiseks on mitmeid. Näiteks kasutasid Bruns, Burgess ja Highfield (2014) muude meetodite hulgas võrgustiku analüüsi, et kaardistada Austraalia Twitteri sfääri. Colleoni, Rozza ja Arvidsson (2014: 317) kombineerisid sotsiaalvõrgustike analüüsi masinõppega, et uurida poliitilist homofiilsust Twitteris vabariiklaste ja demokraatide seas (USAs). Laiem eesmärk oli mõista seda, kas Twitterit võib pidada pigem poliitiliselt kajakambriks või avalikuks sfääriks; kas kogukonnad lõimuvad või mitte. Tulemustest selgus, et kui vabariiklaste seas valitses tugev artikuleeritud võrgustik (vabariiklikult meelestatud isik jälgis tõenäolisemalt ametlike vabariiklaste kasutajaid kui võrrelda sama asja demokraatidega), siis demokraatlikult meelestatud võrgustik polnud ennast nii tugevalt artikuleerinud, aga seejuures moodustasid nad arvestades loodud sõnumeid märkimisväärselt suurema kogukonna (10 korda rohkem kasutajaid kui vabariiklasi). Poliitiline homofiilsus ehk poliitiline suhtlus omasugustega oli tunduvalt suurem demokraatide seas - vabariiklased sattusid diskuteerima tunduvalt rohkem demokraatidega kui demokraadid vabariiklaste sekka. Vabariiklased määratlevad ennast selgemalt, demokraatidel on suurem oht sattuda kajakambri efekti ohvriks (Colleoni, Rozza ja Arvidsson, 2014: 325-326).

## 2.5 Suurandmed ja kontentanalüüs

Kontentanalüüs on tekstianalüüsi uurimistehnika, mille abil teisendatakse erinevad tekstitüübid kvantitatiivsesse keelde ehk see on lüli kvantitatiivsete ja kvalitatiivsete meetodite vahel (Kalmus, 2015). See on olnud massikommunikatsiooniuuringutes üks populaarsemaid meetodeid: umbes 30% aastatel 1980-1999 tehtud teadusartiklitest kasutasid kontentanalüüsi (Kamhawi ja Weaver, 2003: 14). Manuaalse kontentanalüüsiga (kodeerimine ja analüüs teostatakse inimeste poolt) on tekstides võimalik tuvastada seal ilmnevat komplekssust ja nüansse (nt iroonia, sarkasm), mida arvutid ei suuda hästi teha (Guo et al, 2016: 3). Automaatsed tekstianalüüsimeetodid (kodeerimine ja analüüs on teostatud masinate-arvutitega) ei suuda asendada tekstide hoolikat lähilugemist ning sealt ilmnevaid avastusi keelestruktuuride keerukuse tõttu. Samas võimaldavad nad vähendada kontentanalüüsi puhul ressursikulu. Tuleb arvestada ka seda, et manuaalselt kodeerides tuleb samuti ette inimeksimusi (Grimmer

ja Stewart, 2013: 268). Arvutite kasutamine aitab seevastu potentsiaalselt üle saada teatud kodeerimis- ja töötlemispiirangutest. Algoritmid suudavad kiiremini korjata välja täpselt seda andmestikku, mida soovitakse uurida ning seejärel andmestikku vastavalt vajadustele kodeerida ja klassifitseerida (Lewis, Zamith ja Hermida, 2013: 38).

Grimmer ja Stewart (2013: 269-271) toovad automaatse tekstianalüüsi läbiviimise puhul välja neli olulisemat printsiipi:

1) *Kõik kvantitiivsed mudelid on valed - aga mõned võivad olla kasulikud.* See tuleneb keele komplekssusest, kus iga sõna on sõltuvuses oma ümbritsevatest struktuuridest ja konstruktsioonidest. Üks sõna võib lauses muuta omavahelisi sõltuvussuhteid ning terve lause mõtet kardinaalselt. Täpselt samamoodi võib iga lause muuta lõigu mõtet jne. Automatiseeritud mudelid seda keelestruktuuride mängu *täielikult* mõista ei suuda. See aga ei tähenda, et nad ei võiks kasulikud olla.

2) *Kvantitatiivsed meetodid suurendavad inimese võimekust, aga ei asenda inimesi.* Uurijad siiski juhivad protsessi, püstitavad küsimusi, otsustavad mudelite valikute üle ja tõlgendavad väljundit. Arvutid suudavad samas võimendada inimeste oskusi eksponentsiaalselt (omade piirangutega).

3) *Ei ole ühte universaalset parimat meetodit automaatseks tekstianalüüsiks.* Eri andmestikud ja eri uurimisküsimused viivad eri huvipunktideni. Seetõttu on oluline valida just see meetod, mis teemapüstitusega sobib ning ka vajadusel seda mugandada.

4) *Valideeri, valideeri, valideeri ehk kontrolli õigsust.* Automaatsus vähendab küll ressursikulu, kuid selle arvelt kannatab paraku ka täpsus. Seetõttu tuleb kontrollida mudeleid, kas ja kui täpsed need on ning vajadusel neid kohandada, sealhulgas minimeerida protsessis tekkivaid teadmatuid külgi (nt tutvuda algoritmidega). Oma mudeli nõrkustest ja tugevustest tuleb olla teadlik.

Automatiseeritud meetodid täidavad laias plaanis kahte ülesannet: klassifitseerivad ning mastapiseerivad (*scaling*). Mastaapide tuvastamise eesmärk on hinnata subjekti

asukohta sfääris ehk tekstiliselt mustrite leidmine. Selliste uurimuste sisuline alusmõte on, et ideoloogia määrab keele, mida kasutame. Nii on võimalik leida, milliseid sõnu omavahel kasutatakse ning kellele see mõte võiks kuuluda. See on teatud mõttes sarnane võrgustikeanalüüsiga, vaid et siin on sõlmpunktidena kasutusel sõnad (Grimmer ja Stewart, 2013: 269, 291).

Klassifitseerimine on laiem mõiste, mille võib omakorda jaotada kaheks: kategooriad, millest oleme teadlikud ning mida suudame eelnevalt määrata ning kategooriad, mis on teadmata. Klassifitseerimise mõte on teksti organiseerimine. Teadlike kategooriate kasutamises on üks levinumaid meetodeid sõnastikupõhine (kasutatakse ka terminit leksikonipõhine) analüüs. Lihtsustatult arvutab see meetod märksõnade esinemissagedust tekstis, et jaotada tekstiüksused kategooriateks. Tekstiüksustena kasutatakse enamasti terminit dokument. Dokumendid võivad olla nii lause, lõigu, tervikteksti, säutsu (Twitteris), Facebooki staatuse vms tasandil. Dokumendid on kui sõnakobar, kus sõnade järjekorral pole tähtsust. Analüüsimisel võib ka kasutada nii unigrami, bigrami või trigrami meetodeid - vastavalt klassifitseerimist üksiksõna, sõnapaaride, sõnakolmikute kaupa. Eeltöötlemise faasis saab dokumentides muuta sõnu kasutades lemmatiseerimist või tüvistamist (*stemming*), mille mõlema eesmärk on muuta sõnad nende algkujule (arvasin vs arvama). Veel kasutatakse stoppsõnade ja otstarveliste sõnade (*stop ja function words*) eemaldamist, mis ei kanna endas tähendust, vaid täidavad grammatilisi funktsioone (Grimmer ja Stewart, 2013: 272-273). Nagu näiteks *aga, ehk, et* jne (Uiboaed, 2017a). Sõnastikupõhist analüüsi kasutatakse näiteks dokumendi tonaalsuse hindamiseks (Grimmer ja Stewart, 2013: 274).

Sõnastikupõhist meetodit tuleb kasutada ettevaatlikusega. Suurandmete puhul pole võimalik uurijal piisava representatiivsusega näidiste põhjal kinnistada sõnastiku sobivust. Kui on võimalik, siis soovitavalt võiks ka eelnevalt täiendada või kontrollida, mille alusel sõnu määratakse. Spetsiifilisemate terminitega teksti puhul võib olla tarvilik ise luua vastav sõnastik. Samuti võib teemamudelitesse määramisel ühe dokumendi sees olla sisse mahutatud mitu konkureerivat teemat (Guo et al, 2016: 5).

Sõnastikupõhine analüüs on sarnane juhendatud (*supervised*) masinõppe meetodiga - viimases jaotavad inimkodeerijad prooviandmestikus (*training set*) teatud osa dokumente kategooriatesse (nagu ka sõnastikupõhises analüüsis); kasutades prooviandmestikku üritab masin algoritmide abil niiöelda õppida, kuidas sortreerida mis dokumente mis alusel. Seejärel märgistab arvuti ülejäänud dokumendid; inimene kontrollib kodeerimise metoodikat. Arvuti üritab seega ise mõista loogikat, mille alusel inimene dokumente kategooriatesse jaotab. Oluline on, et prooviandmestik oleks võimalikult täpne (Grimmer ja Stewart, 2013: 275-276).

Juhtumite puhul, mil inimene ei oska eelnevalt dokumente kategooriatesse määrata on võimalik kasutada juhendamata (*unsupervised*) masinõpet. Selle abil on võimalik tuvastada teksti alustunnusjooni ehk süsteemi. Juhendamata masinõppega jaotatakse dokumendid sarnaste tunnuste abil gruppidesse (klasterdatakse). Seejärel on tekkinud klastrite alusel võimalik neid kategoriseerida ehk nimetusi anda (sh vaadata, mis on nö õiged kategooriad, mis mitte). Parima tulemuse saamiseks võib kaalutleda, kas oleks vaja kombineerida erinevaid meetodeid (näiteks kategooriate tuvastamiseks eelnevalt kasutada juhendamata, seejärel juhendatud masinõpet) (Grimmer ja Stewart, 2013: 281).

Guo et al (2016: 1) võrdlesid samal andmestikul nii juhendamata masinõppe teemade modelleerimist kui ka sõnastikupõhise kategoriseerimise tulemusi – 2012. aasta USA presidendivalimiste aegu Obamat ja Romneyt puudutanud 48 miljoni säutsu kohta. Sõnastikupõhise analüüsi jaoks võeti esmalt välja populaarseimad sõnad ning seejärel tekitati temaatiliselt 16 kategooriat (nt maksud, majandus, immigratsioon, tervishoid, välispoliitika jne). Kasutades inimkodeerijaid liigitati iga teema juurde iseloomulikud märksõnad, mis tuvastati juhuslike säutsude lugemise abil. Seejärel teostati kategoriseerimine terve andmestiku peal.

Juhendamata masinõppe puhul kasutati LDA (ehk *Latent Dirichet Allocation*) teemade modelleerimist. Üksustena kasutati sel juhul iga kasutaja nelja säutsu kokkupanekut ühte dokumenti. Sarnaselt sõnastikupõhise analüüsiga määrati, et tulemusena peaks tekkima 16 teemat (ehk klastrit), kuhu olid kogunenud kõige sagedamini koosesinevad sõnad (Guo et al, 2016: 9-10).

Tulemustest selgus, et LDA meetod suutis oma tulemuse valimisse haarata rohkem dokumente kui sõnastikupõhine lähenemine ehk LDA suutis tuvastada ka populaarseid teemasid, mis manuaalselt kodeerides välja jäi (näiteks Obama sünnitunnistuse ümber käinud arutelu). LDA puhul toodi negatiivse poole pealt välja, et ühe teema alla võib mahtuda mitmetest eri teemadest koosnevaid märksõnu (ühe Obama teema alla mahtusid märksõnadena abort, immigratsioon, maksud jne). Proovivalimi (100 dokumenti) järgi oli LDA meetod tunduvalt täpsem. Teemade tuvastamise osas tulid mõlemat meetodit kasutades välja sarnased tulemused – Obama puhul oli populaarseim teema välispoliitika, Romney puhul maksundus. Kui LDA korjas üles rohkem valepositiivseid tulemusi (ehk tekitas seoseid, mida polnud), siis sõnastikupõhine lähenemine rohkem valenegatiivseid väiteid (ehk väitis, et pole võimalik tuvastada, aga tegelikult oli). Autorid soovitasid, et tuleviku uuringutes võib proovidagi kombineerida kahte meetodit: esmalt tuvastada LDA abil teemasid, seejärel rakendada sõnastikupõhist analüüsi (Guo et al, 2016: 15-21).

## 2.6 Suurandmete kasutamise eetika ja legaalsus

Sotsiaalmeedia suurandmete kogumise ja analüüsimise juures võib tekkida nii mõnigi eetiline küsimus: kas kogutud andmestikku võib ilma luba küsimata kasutada; kui ei, siis mida peaks täpsemalt tegema; kas andmestiku peab anonüümistama; kas kuidagi riivatakse oma subjektide ehk inimeste privaatsust; kas on eetiline kasutada näiteid kellegi kohta ilma, et see inimene sellest midagi ei teaks; kes vastutab, et teatud indiviidide ja kogukondade privaatsust ei riivata uurimistöö protsessis (boyd ja Crawford, 2012: 672).

Arvestades, et suurandmestikud võivad oma mõõtmetelt ulatuda miljonitesse väljadesse, on ebamõistlik eeldada, et uurija peaks igalt uuritavalt küsima eraldi luba. See aga ei tähenda, et eetikaküsimusi peaks ignoreerima sel põhjusel, et postitused, mida analüüsitakse olid kõigile kättesaadavad ja avalikult välja paisatud (boyd ja Crawford, 2012: 672). Näiteks, kui Twitteris privaatse kasutaja säutsu peaks keegi avaliku profiiliga inimene edasisäutsuma (*retweetima*) ühel või teisel moel, siis saab sellest samuti avalik postitus (Zimmer ja Proferes, 2014: 258). Suurandmete uurimistöödes tekib samuti kontrolli üle küsimus – uuritavatel ei ole võimalik teada,

mis kontekstis, mida ja kuidas tema kohta uuritakse. Ei pruugita olla teadlik, et kui ta avalikult postitab sõnumi, siis see võib jõuda kuskile suurandmete andmevoogu. Oma kujuteldava auditooriumina ei pruugita näha teadlasi (boyd ja Crawford, 2012: 673). Kasutajail võib olla mitmeid põhjusi, miks soovida kontrollida oma loodud andmete kasutuste üle: mured privaatsuse üle, muljete kontrollimine (et poleks kontekstist väljas), kartus valitsuste ees (et liialt suur andmekogumine hakkab privaatsust riivama) jms (Puschmann ja Burgess, 2014: 48). Samuti on kasutajatel väga keeruline ennast andmekogudest välja arvata, kui puudub teadvus, kus ja mis informatsiooni minust talletatud on (Zimmer ja Proferes, 2014: 258). Privaatsuse kontrollimine on suurandmete puhul ühtepidi nii tehnoloogiline kui ka sotsioloogiline probleem, mida peab adresseerima mõlemalt poolt (Agrawal et al, 2012: 10).

Twitteri suurandmestiku puhul võib tekkida ka näiteks legaalsed küsimused andmestike jagamise osas. Euroopa Liidus on andmestikud kaitstud juhul kui kvalitatiivselt ja/või kvantitatiivselt andmete omandamisse, verifitseerimisse või presenteerimisse panustati märkmisväärne investeering. Twitter ise ei suhtu hästi enda kohta kogutud suurandmete jagamisse: andmed peavad olema anonüümistatud, jagada võib vaid säutsude ID-sid ja/või kasutaja ID-sid (Giachanou ja Crestani, 2016: 30; Gaffney ja Puschmann, 2014: 59; Beurskens, 2014: 130).

# 3. Meelestatuse analüüs

Meelestatuse analüüs (*sentiment analysis*; eestikeelne termin Uiboaed 2017b) on uurimisvaldkond, mis analüüsib inimeste arvamusi, tundmuseid, hinnanguid, hoiakuid ja emotsioone mingite toodete, teenuste, organisatsioonide, indiviidide, probleemide, teemade või sündmuste kohta. Alternatiivselt kasutatakse ka veel sarnaseid fraase nagu arvamuste kaevandamine (*opinion mining*), arvamuste väljavõtmine (*opinion extracting*), meelestatuse kaevandamine, subjektiivsuse analüüs, emotsionaalsuse analüüs, arvustuste kaevandamine jne (Liu, 2012: 7). Giachanou ja Crestani (2016: 2) aga rõhutavad, et arvamuste kaevandamine ja meelestatuse analüüs on sarnased, kuid erinevad asjad: esimene üritab tuvastada, kas arvamus eksisteerib tekstis, teine üritab hinnata arvamuse emotsionaalsust. Meelestatuse analüüsi väljund võib olla näiteks, et ettevõte soovib teada saada, mis emotsiooni tema tooted tarbijates tekitavad uurides selleks sotsiaalmeedia postitusi (Liu, 2012: 28).

Meelestatust võib hinnata nii polaarsuse kaudu (positiivne-negatiivne ja vajadusel ka neutraalne) kui ka laiema emotsionaalse skaala kaudu: näiteks rõõm, üllatus, viha, kurbus, vastikustunne ja hirm (Liu, 2012: 7). Sellist jaotust kasutan ka antud töös. Selline jaotus pole tinglik, vaid lähtub kultuuride ülesest põhiemotsioonide jaotusest, mis baseerub universaalsetel näoilmete väljendustes. Üllatus ja hirm on osades jaotustes kobineeritud (Ekman ja Oster, 1979: 531). Plutchiku (2001: 349) põhiemotsioonide hulka kuulusid veel täiendavalt neile kuuele usaldus ja ootus.

Meelestatuse analüüsi saab teostada eri tasanditel: dokumendi (mida on juba eelnevalt mainitud), lause või aspekti tasandil. Viimane neist on keerukam kui dokumendi ja lause tasand seetõttu, et see üritab vaadata sealhulgas ka seda, et millele või kellele on meelestatus suunatud ehk mille kohta emotsioone väljendatakse. Täiendavalt võib vaadata ka seda, et kas arvamused on tavapärased, kus millelegi omistatakse hea või halb omadus (nt *see kook on hea*) või võrreldakse millegagi (nt *see kook on parem kui minu oma*) (Liu, 2012: 10-12).

Sõnastikupõhiselt meelestatuse hindamisel võib välja tuua mitmeid probleeme. Näiteks kui hinnatakse meelestatust kolmesel skaalal (positiivne, negatiivne,

neutraalne), võib nii mõnigi sõna olla ühes kontekstis positiivne, teises negatiivne. Meelestatust sisaldav lause ei tähenda ilmtingimata, et lause ise väljendaks seda: nt "Kas sa saad nimetada mõnda head raamatut?" Lause sisaldab märksõna, mis viitaks positiivsele polaarsusele (hea), kuid see on püstitatud hüpoteesina omistamata millelegi emotsiooni. Vastupidiselt laused, mis ei sisalda meelestatust märgistavaid sõnu, võivad väljendada siiski teatud emotsioone: näiteks "See pesumasin kasutab palju vett" viitab pigem negatiivsusele, kuigi ta on iseenesest kirjeldus. Samuti on keeruline tuvastada sarkasmi ja irooniat. Sõna-sõnalt sama lause võib väljendada erinevaid emotsioone (Liu, 2012: 12-13.

Peatükis 1.3 mainitud Facebooki eksperimendis kasutati meelestatuse analüüsi tehnikat nimega LIWC (*Linguistic Inquiry and Word Count*), kus mõõdeti mitu protsenti tekstis olevatest sõnadest on markeeritud eri emotsioonidega. Seejärel jaotati postitused negatiivseteks ja positiivseteks. Panger (2016: 1115-1116) kritiseeris antud tehnikat selles osas, et emotsionaalne skaala oli piiratud (tuvastati kurbust, aga mitte kadedust); tööriist töötati välja pigem pikemate kirjutiste jaoks ning seetõttu võivad üksikud sõnad mängida liigselt märkmisväärset rolli; ei kasutatud sobivat leksikoni, mis sobiks hindama sotsiaalmeedia postitusi; ning viimaks see, et hindamise all olid vaid postituse tekstiline osa, kuigi pildid sisaldavad potentsiaalselt tugevaid emotsionaalseid signaale. Veel eksisteerivad näiteks leksikonipõhised algoritmid nimega SentiStrength, SentiWordNet, SentiCircles ning juhendamata masinõppe analüüsimeetod nimega ESSA (*Emotional Signals for unsupervised Sentiment Analysis*) (Giachanou ja Crestani, 2016:19). Meelestatuse analüüsi valisin töö metoodikaks eelkõige seetõttu, et see meetod tundus huvitav, kuidas on võimalik emotsionaalsust kvantifitseerida.

# 4. Twitter kui uurimisallikas

## 4.1 Mis on Twitter?

Twitter on mikroblogi, mis baseerub lihtsal põhimõttel: selle teenuse kasutajad saavad postitada kuni 140 tähemärgiga lühikesi sõnumeid (*tweete* ehk säutse) ning samuti jälgida teiste postitusi kasutades selleks arvutit või nutiseadet (Weller et al, 2014: xxix). Tähemärkide piirang sunnib kasutajaid olema oma säutsudes väga kompresseeritud ning põgusad. Mahupiirangu sisse pole lihtsalt ruumi ei sissejuhatuseks, kontekstiks (välja arvatud teemaviide elik *hashtag*) või laiemapõhjaliseks seletuseks (Risse et al, 2014: 209). Twitterit, mis loodi 2006. aastal, kasutab igakuiselt üle 300 miljoni aktiivse kasutaja, iga päev postitakse keskmiselt umbes 500 miljonit säutsu (Statista, 2017; Internet live stats, 2017). Twitter on meedium nii igapäevase väljenduse ja suhtluse jaoks kui ka vahend, et jälgida ning osa võtta üle maailma toimuvatest suursündmustest. Näiteks on rohkete säutsude arvuga sündmusteks olnud Michael Jacksoni surm, prints Williami ja Kate Middletoni abielu, iga-aastased Oscarite auhinnagalad, kuid ka Araabia kevade arengud ja orkaan Sandy jne. Twitter on ka oma algusaegadest kandnud taaka kui koht otstarbetu lobisemise ning kasutu informatsiooni jagamise jaoks. Samas on see ka koht poliitilise diskussiooni (või kommunikatsiooni) jaoks ning uudiste lavaks: USA uut presidentigi tuntakse aktiivse Twitteri kasutajana. Samuti on olnud ka seda Eesti endine president Toomas Hendrik Ilves (Weller et al, 2014: xxx).

Twitteri kommunikatiivset struktuuri saab vaadata kolmetasandiliselt: mikro-, meso- ja makrotasandid. Mikrotasandil toimub kommunikatsioon interpersonaalselt, vastates konkreetselt inimestele, olgu selleks avalikus postituses või otsesõnumis. Kasutajatel on oma postituses võimalik otse viidata kasutades sümbolit @ ning lisades sinna juurde pöördutava kasutaja nime (Bruns ja Moe, 2014: 19-20).

Twitter toetub oma kommunikatsioonimudelis peamiselt artikuleeritud sotsiaalsetele ühendustele saatja-auditoorium suhtes. See teostub jälgimistes - sa saad tellida huvipakkuva kasutaja postitusi enda ajajoonele. Erinevalt Facebookist ei pea jälgimiste suhe olema vastastikune - kasutajad saavad tellida teiste kasutajate postitusi ilma, et teine seda vastu teeks. Näiteks kuulsuste puhul võib ulatuda jälgijate arv

miljonitesse, aga vastupidiselt jälgitakse vaid üksikuid kasutajaid. See moodustab Twitteri mesotasandi - isiklik, valitud avalikkus (Schmidt, 2014: 5; Bruns ja Moe, 2014: 16-20).

Kolmas ehk makrotasand on tihti kiiresti moodustuv ning kiiresti laialiminev. See väljendub *hashtagides* (#) ehk eestikeelselt teemaviidetes (EKI, 2014). Teemaviited on spetsiifilised märksõnad, mille alusel moodustuvad nö teemafoorumid. Teemaviidete kasutamine viitab eelkõige sellele, et soovitakse osa võtta laiemast kommunikatiivsest protsessist. Eesti puhul võib näiteks tuua Eesti Laulu või presidendivalimised, kus inimesed said osa võtta avalikust diskussioonist kasutades vastavaid märksõnu ehk teemaviiteid (#eestilaul2017 ja #presidendivalimised). Ühe sündmuse puhul võidakse kasutada mitmeid erinevaid teemaviiteid (Bruns ja Moe, 2014: 17-18, 20). Teemaviited võivad ka samas olla ekspressiivsed rõhutused ilma, et tahetaks kuskilt vestlusest osa võtta väljendades emotsiooni või sarkasmi, näiteks #win ja #fail (Bruns ja Stieglitz, 2013: 2).

Kolm tasandit eksisteerivad küll eraldi, kuid paljuski on nad omavahel tihedalt lõimunud. Teemaviiteid kasutanud postitus jõuab täpselt samamoodi ka kasutaja *jälgijate* ajajoonele, sama asi toimub kui sinu jälgitav pöördub mikrotasandil kellegi teise poole (kui see ei toimu just privaatses sfääris ehk otsesõnumites). Ühesse postitusse saab mahutada nii teemaviiteid kui ka pöördumisi (Bruns ja Moe, 2014: 20-21).

Twitteri üheks oluliseks tunnusjooneks on *retweet* ehk edasisäuts (Glosbe, 2017). Edasisäuts on mehhanism sõnumite edasiviimiseks ning vahend sõnumi tunnustamiseks. See toimib sedasi, et kasutaja vaatab mingit postitust ning otsustab, et ta peab seda ühel või teisel põhjusel edasi levitama. Vajutades edasisäutsu nuppu ilmub säuts ka seetõttu tema enda profiilile kui ka tema jälgijate ajajoontele (Bruns ja Moe, 2014: 22). Edasisäutsud jõuavad keskmiselt 1000 kasutajani. See võimaldab sõnumite eksponentsiaalset edasijõudmisvõimalust (Kwak et al, 2010: 600).

Mikroblogi Twitter pakub dünaamilist, interaktiivset identiteedi esitlusvõimalust teadmata auditooriumile (Marwick ja boyd, 2011: 3). Omaette küsimus on, kas Twitter panustab ka kvaliteetse infovooga avalikku sfääri, kuid see on kindlasti

vahend, mis mõjutab kuidas inimesed vahetavad ja tarbivad informatsiooni (Small, 2011: 873).

## 4.2 Twitter kui avalik sfäär

Twitterit kasutatakse erinevatel põhjustel: kes turunduskanalina, kes päevikuna, kes sotsiaalse platvormina, kes uudisteallikana, kes informatsiooni levitamiseks. Sellest tulenevalt vaadeldakse ka erinevalt, keda oma auditooriumina nähtakse (Marwick ja boyd, 2011: 9). Erinevalt Facebookist on sõnumite sihtmärk rohkem avatud - enamasti on sõnumid avalikud ning potentsiaalselt mõeldud laiemale auditooriumile (Trilling, 2015: 260). Auditoorium on kujuteldav: see pole konkreetselt määratletud ega samamoodi pole konkreetselt kontrollitav (kui just kasutajat privaatseks ei tehta). Seda enam ei suuda sa kontrollida oma auditooriumi, kui võtad osa vestlusest ning vastates kellegi postitusele või kui kasutad teemaviiteid. Paratamatult on säutsude kirjutamisel mingisugune kujutelm ees, kellele seda kirjutatakse. Iseasi, kas ja kui palju seda teadvustatakse. Mõned kujutlevad postitamist kui iseendale suunatud akti (elik päevik, ekspressiivne tegevus), teiste jaoks on see eelkõige väljapoole suunatud tegevus (tahetakse oma sõnumiga kuhugi või kellegini jõuda). Twitter kombineerib sellega sihilikku ja juhuslikku publikut. Kokku annab see mitmetasandilise auditooriumi, kus sul on kindlalt teadaolevad jälgijad kui ka sekundaarne anonüümne auditoorium. Ta hõlmab sedasi nii avalikku kui ka isiklikku sfääri (Marwick ja body, 2011: 2, 5, 16).

Teatud teadmatus oma auditooriumi osas võib põhjustada säutsumisel sõnumite tasakaalustamist. Soovitakse olla autentsed, aga mitte ära anda täielikult oma intiimsust. Oma auditooriumi üle ei pruugi olla täielikku kontrolli ning seetõttu võib esineda enesetsensuuri. Samuti esineb laveerimist: postituste sihtmärgid on sõltuvalt säutsust erinevad, mistõttu üritatakse tasakaalustada seda, et sõnum jõuaks sihtmärgini, aga samas ei kaotataks jälgijaid, keda see ei huvita või lausa riivaks. Aktiivsemate jälgijaskondadega kasutajatel võib ka olla endale ette seatud kujuteldav norm ja ootused oma jälgijaskonna ees, sunnitud aktiivsus. Eriti kui on enda jaoks ühe eesmärgina seatud auditooriumi laiendamine (Marwick ja boyd, 2011: 11-13).

Twitteri kasutajaid ei saa pidada representeerivaks üldisemat populatsiooni. Twitteri kasutajad võivad olla juba ise nii isiklikud kasutajad kui ka organisatsiooni vms omad. Ühel isikul võib olla mitu kasutajat, mitu identiteeti. Mõned inimesed jälgivad Twitterit, kuid ei liitu sellega ja väga suur osa elanikkonnast ei puutu Twitteriga üldse kokku (boyd ja Crawford, 2012: 669). Twitteri kasutajad on keskmiselt 10 aastat nooremad kui keskmised sotsiaalmeediakasutajad (American Press Institute, 2015: 39).

Vastandudes kajakambri efektile on Twitterile veel viidatud kui refraktsioonikambrile (*refraction chamber*): tunduv kajakambri efekt tekib seetõttu, et kamber ise on sealolevate inimeste loodud. Tekkiv sfäär on inimeste juhitud ühiste väärtuste ja arusaamiste summa (Rieder, 2012: 10). Samuti on Twitterile viidatud kui isiklikule avalikkusele (*personal public*). Isiklikus avalikkuses on informatsioon valitud ja nähtav vastavalt isikliku asjakohasuse printsiibist. Suunatud informatsioon on auditooriumile, mis koosneb selgelt väljendatud võrgustiku otstest. Viimaks luuakse informatsioon peaasjalikult vestluslikus vormis (vastandudes ametlikule kirjutise avaldamisele) (Schmidt, 2014: 4).

## 4.3 Twitter, ajakirjandus ja poliitika

American Press Instituudi uuringu järgi kasutab ligi 90% Twitteri kasutajaid Twitterit uudiste saamiseks, sealjuures enamus neist igapäevaselt. Enamasti saadakse uudised kätte oma ajajoonelt. Kõigest kolmandik pingutab täiendavalt ehk jälgib trendivaid teemasid ning kasutab otsingumootorit uudiste saamiseks. See arv suureneb märkimisväärselt erakorraliste ning otse-eetris olevate sündmuste ja uudiste puhul ehk sel juhul hakatakse tunduvalt rohkem jälgima ka populaarseid teemaviiteid (sh niiöelda ametlike teemaviiteid). Samuti peetakse oluliseks mitte lihtsalt uudiseid tarbida, vaid ka kaasa lüüa jagades säutse või kommenteerides teemaviidete all (American Press Institute, 2015: 4, 15, 23, 25-26). Meediaorganisatsioonid ning ajakirjanikud kasutavad Twitterit eelkõige kui kommunikatsioonivahendit: jagavad artikleid, suhtlevad publikuga ning vaatavad, kuidas inimesed reageerivad uudistele. Sisuloomes kasutatakse Twitterit eelkõige reaalajas toimuvate sündmuste otse

raporteerimiseks. Tunduvalt harvemini tuleb ette juhtumeid, kus Twitter on juhtlõng teemade otsimiseks (Neuberger, vom Hofe ja Nuernbergk, 2014: 349-351).

Twitterit kasutatakse poliitilises kommunikatsioonis nii veenmis- ja mobiliseerimistööriistana kui ka sihiliku diskursuse arendajana (Trilling, 2015: 260). Viimase näiteks on määratletud teemaviited poliitiliste teledebattide ajal televisioonis, mis kutsuvad vaatajaid debatist osa saama (Harrington, 2014: 241-242). Twitter täidab siin niiöelda teise täiendava ekraani rolli ning võimaldab reaalajas inimestel emotsioone vahendada, lisada informatsiooni ja konteksti ning hinnata debatikäiku. Poliitikud suudavad debattide ajal suunata agendasid, millest Twitteris kirjutatakse, kuid samas mitte kontrollida infovoogu (Vergeer ja Franses, 2016: 1392, 1404). Näiteks sai ühes Saksamaal toimunud teledebatis üheks peamiseks räägitavaks teemaks erinevate poliitikate arutelu asemel Angela Merkeli kaelakee. Twitterist toimuvat ei saa sel ajal nimetada ratsionaalseks konstruktiivseks poliitiliseks debatiks, vaid tihti mängib Twitteris toimuvas suurt rolli iroonia, sarkasm ning üleüldiselt nalja tegemine. Samas võib Twitteris kerkida esile olulisi teemasid, mida debatis ei arutleta: sama Saksamaa näite puhul keerles Twitteri debatt väga palju samal ajal toimunud NSA skandaali ümber, millele debati sees ise rõhku ei pandud (Trilling, 2015: 270, 273).

## 4.4 Twitteri uuringud

Mikroblogidel nagu Twitter on teatud positiivsed omadused, mis teevad neid potentsiaalselt üheks heaks vahendiks avaliku arvamuse uurimiseks. Näiteks on nad ajaliselt selgelt määratletud ning seal olev informatsioon on avalikult võrdlemisi lihtsasti kättesaadav. Samas võib oponeerida, et kas, ja kui palju Twitteris toimuv kajastab ühiskonnas toimuvat (Thelwall, 2014: 83).

Erinevate teadusartiklite andmebaasi põhjal tehtud 2006.-2012. aasta kokkuvõte viitab, et Twitteri uuringuid viiakse läbi väga mitmetes valdkondades: alates psühholoogiast keskkonnateadusteni, juurast arstiteadusteni, majandusest füüsikani. Kõige populaarsemad erialad on arvutiteadused, infoteadused ning kommunikatsiooniteadused. Enim teostatakse kontentanalüüsi. Seejärel on

populaarseimad meetodid veel võrgustike analüüs ning erinevad kasutajate uuringud. Valimite suurused olid alates ühest säutsust kuni 5 miljardi säutsuni. Mediaan jäi umbes 100 000 ja miljoni säutsu vahele. Aastate lõikes kasvas Twitteri uuringute arv märgatavalt alates 2010. aastast. Üle poole uuringutest kasutasid andmestiku saamiseks Twitteri APIt. Alternatiivina kasutati nii *online* agregaatoreid kui ka olemasolevaid andmebaase kolmandatelt osapooltelt. Kaks uuringut viidi läbi intervjuu, küsitluse või vaatluse teel (kokku töid 352) (Zimmer ja Proferes, 2014: 253-256).

Lisaks paljudele eelnevalt viidatud töödele on ühe populaarse teemana Twitteri uuringutes näiteks sport. Highfield tõi välja, et kui Tour de France'i jälgimine kuulub paljude austraallaste seas justkui rituaalsete tegevuste hulka, siis Twitter aitab seda rituaali teisendada jagatud kogemuseks (Highfield, 2014: 259). Bruns, Weller ja Harrington (2014: 263) võrdlesid teiselt poolt seda, kuidas kolme erineva jalgpalliliiga meeskonnad kommunikeerivad Twitteris ning milline on nende jälgijaskond.

Samuti on poliitikaväli saanud Twitteri üheks oluliseks osaks. Larsson ja Moe (2014: 324) võrdlesid kolme riigi (Rootsi, Taani, Norra) näitel, kuidas poliitilise diskussiooni aktiivsus on võrdeline valimiste perioodidega. Seal selgus, et postitamise aktiivsusjooned olid riigiti võrdlemisi sarnased ning seda võib vaadata ka kui üht mudelit. Hollandi poliitilist välja uurides nähtus, et kui üldine poliitiline diskussioon pole eraldunud erakondade joone järgi, siis edasisäutsumine (*retweetimine*) selgelt on (Paßmann, Boeschoten ja Schäfer, 2014: 335). Samuti uuriti poliitiliste teledebattide näitel seda, kuidas debatt dikteerib Twitteris valitsevaid teemasid (Vergeer ja Franses, 2016: 1405). Kanadas uuriti, kas niiöelda ametlikke teemaviiteid kasutades tekib Twitteris diskussioon (ehk üksteisele vastamist) või mitte. Selgus, et peamine säutsumise eesmärk oli hoopis informeerimine (Small, 2011: 891).

Twitter on ka saanud kohaks, kuhu pöörduda eriolukordades - olgu selleks looduskatastroofid, mässud, surmad või midagi neljandat. Kriisiolukordades kasutavad hädaabiteenused ise Twitterit nii oma ametlike sõnumite kommunikeerimiseks (mis-kus juhtus) kui ka selleks, et monitoorida, mis toimub. Politsei on kasutanud Twitterit sellises olukorras kiiresti kasvavate kuulujuttude

ümberlükkamiseks (Bruns ja Burgess, 2014: 380-381). Ühendkuningriikides 2011. aastal aset leidnud mässud pakkusid materjali nii Twitteris postitatud piltide uurimiseks (Vis et al, 2014: 396) kui ka näidismaterjalina selleks, kuidas uurida selliste kriisiolukordade ajal Twitteris toimuvaid dünaamikaid (Procter, Vis ja Voss, 2013: 209).

Twitter võib aidata ka mõista diskursuste raamistamisi: Mo Jang ja Sol Hart (2015: 13-16) võrdlesid nelja riigi (USA, Ühendkuningriigid, Austraalia ja Kanada) põhjal seda, kus ning mis kontekstis kasutati termineid "globaalne soojenemine" ja "kliimamuutus". Andmestikust selgus, et selgelt skeptilisemad regioonid (riikide tasandil USA, USA sees niiöelda punased vabariiklaste osariigid) kasutasid terminit "globaalne soojenemine" märkimisväärselt rohkem. Teised eelistasid rohkem terminit "kliimamuutus". Kui termini "kliimamuutus" puhul kesknes diskussioon selle üle, mis on selle mõjukus ja kuidas selle vastu peaks võitlema, siis termini "globaalne soojenemine" kasutamise puhul domineeris pettuse (*hoax*) diskursus.

Omaette uurimissuunana on samuti välja arenenud mõtestus, mis see Twitter ikkagi oma olemuselt on. Marwick ja boyd (2011: 1-2) uurisid, milliseid praktikaid kasutajad säutsumisel kasutavad ning millisena nad näevad oma auditooriumi. Bruns ja Stieglitz (2013: 2) süstematiseerisid, kuidas mõõta Twitteripõhist kommunikatsiooni. Samuti kaardistas Bruns (2012: 1346), mismoodi toimivad teemaviidete võrgustike dünaamikad. Lahuerta-Otero ja Cordero-Gutierrez (2016: 575) kasutasid andmekaevandamist selleks, et leida mõjukaimaid Twitteri kasutajaid konkreetsete teemade osas.

Twitteri üheks lahutamatuks komponendiks on saanud ka *botid* ja spämmijad. Spämmijad on kasutajad, kes erinevatel põhjustel jagavad massiivselt soovimatuid postitusi. Nendeks põhjusteks võivad olla reklaam, õngitsemine (*phishing*), viiruste jagamine, Twitteri maine õõnestamine jne. Twitteris kasutatakse tihti lühendatud URLe (enamasti t.co kujul), et mahtuda tähemärkide piiri sisse ning mistõttu võib ettearvamatu kasutajata URLide täispikkust nägemata sattuda spämmijate ohvriks. Spämmijad on enamasti inimesed (Miller et al, 2014: 64; Vallaste, 2017).

*Botid* on seevastu automatiseeritud programmid, mis üritavad simuleerida inimtegevust. See võib teostada nii legitiimseid ülesandeid nagu automatiseeritud informatsiooni või uudiste ülespanek, kui ka täita spämmijatele sarnaseid kahjulikke ülesandeid (Kollanyi, Howard ja Wolley, 2016: 1; Vallaste, 2017). Kollanyi, Howard ja Woolley (2016: 4) seirasid *botide* aktiivsust ja tegutsemist möödunud aasta USA presidendivalimiste ajal; Miller et al (2014: 64) katsetasid, kuidas klasterdamise abiga spämmijaid tuvastada.

Omaette eesmärgina on teadlased üritanud luua meetodeid selleks, et automaatselt Twitteri abiga sündmusi tuvastada. Kaneko ja Yanai (2016: 143-154) võtsid selle aluseks geolokaliseeritud säutsud ning kombineerisid sündmuste tuvastamiseks piltide klasterdamist ja märksõnade kasutussagedust. Ameerika Ühendriikide ja Jaapani kohta tehtud eksperimentide põhjal suudeti tuvastada näiteks nii lume maha tulekut ühes piirkonnas, ilutulestikku, kirsiõite õide puhkemist ja Tokyos toimunud festivali.

Twitteris tehtavate uuringute väli on nii erialaselt kui ka metoodiliselt üsnagi lai. Samas on see ka kohati eksperimentaalne ning pidevalt arenev uurimisväli. Metoodikad peavad lisaks omaette välja kujunemisele ka pidevalt kohanema muutuvate uurimisvõimaluste tingimustega. Twitter on saanud osaks avalikust sfäärist, kus väga suurel hulgal inimesi üksteisega kommunikeerivad – seega on ka see koht, kus on võimalik välja kaevandada ülevaatlike pilke nii inimeste käitumisharjumustest kui ka sisulistest arvamustest.

## 4.5 Twitter ja meelestatuse analüüs

Twitteri uuringute üheks levinuimaks meetodiks on olnud veel meelestatuse analüüs - aastatel 2006 kuni 2012 teostati Twitteri uuring sedaviisi vähemalt 63 korral, hoolimata tähemärkide limiidist (Zimmer ja Proferes, 2014: 254). Säutsude automatiseeritud analüüsimine on samas väljakutsuv – levinud on lühendid, sihilikult lühendatud sõnad, släng, domeenipõhised spetsiifilised terminid. Esineb nii grammatilisi kui õigekirjavigu; säutsud on napisõnalised ning tihti struktureerimata. Omaette katsumus on, kuidas suudaks automatiseeritud analüüs arvestada eitustega. Lisaks ei koosne säutsud ainult tekstist, vaid sisaldavad tihti nii pilte kui videosid.

Sõnastikupõhise lähenemise puhul tuleb arvestada, et täiuslikkus on suurandmete puhul praktiliselt võimatu. Valesti tõlgendamise oht on üsnagi suur (Risse et al, 2014: 212; Guo et al, 2016: 6; Giachanou ja Crestani, 2016: 7).

Twitteri tekstiomadustena võib välja tuua neli erinevat klassi: semantilised, süntaktilised, stilistilised ja Twitteri-spetsiifilised omadused. Semantiliste omaduste alla kuuluvad arvamussõnad (sõnad, mis sisaldavad arvamust), meelestatust sisaldavad sõnad (võimalik määrata, kas sõna on positiivne või negatiivne), semantilised kontseptsioonid ja eitus. Süntaktilised omadused on unigramid (üksikud sõnad), bigramid (sõnapaarid), trigramid (sõnakolmikud), terminite sagedused jne. Keerulisemate tööriistadega on võimalik süntaktilisi omadusi kasutades säutsude meelestatuse täpsust paremini määrata – on võimalik arvestada sõnadega sõna ümber ehk konteksti. Stilistilised omadused on emotikonid, släng, lühendid jne. Twitteri-spetsiifiliste omaduste alla kuuluvad teemaviited, edasisäutsud, vastamised, mainimised jne. Nende omaduste abil on kokku võimalik hinnata Twitteri tekstide (säutse) meelestatust (Giachanou ja Crestani, 2016: 8-10).

Säutsude meelestatuse tabamise abil on võimalik tuvastada Twitteris kajastuvaid sündmusi. Thapen, Simmie ja Hankin (2016: 2-6) kombineerisid erinevate emotsioonide ja polaarsuste tuvastamist märksõnade otsinguga, mis on levinuimad haiguste sümptomid ja säutsude geolokatsiooni, et üles tähendada haiguslainete puhanguid. Kontekstina vaatlesid nad paralleelselt ilmunud ajalehti ning üldisemalt uudismeediat selleks, et näha, kas oli kattuvusi.

Meelestatuse analüüsi saab võrrelda või täiustada kasutades paralleelselt toimuvate sündmuste ajajoont. Yu ja Wang (2015: 394-399) võrdlesid 2014. aasta jalgpalli MM kohtumistes toimunut sellega, mis toimus samal ajal Twitteris. Tulemused näitasid ilmselgeid seoseid (USA fännid reageerisid vastaste väravale negatiivselt, oma väravale positiivselt), kuid ka seda, milline oli emotsionaalne laetus, siis kui oma meeskond ei mänginud. Üldine emotsionaalne laetus oli sel ajal madalam ning üleüldine emotsionaalsus oli selgelt rohkem positiivne.

## 4.6 Twitteri API (rakendusliides)

Twitter toetab peamiselt kahte rakendusliidest (APIt ehk *application programming interface'i*), mille kaudu on võimalik (suures koguses) andmeid Twitterist kätte saada: REST API ja voogesitus (*streaming*) API. REST (*representational*) API toimib järgnevalt: kasutaja saadab serveri kaudu päringu Twitterile, mis kogub vastavad andmed kokku ning edastab need tagasi kasutajale. Andmeteks võivad olla nii nimekiri trendivatest teemadest kui ka kuni 5000 kasutaja ID'd päringu kohta. Päringute arvule kui ka neid piiravatel aegadel on limiidid: üks aken kestab 15 minutit, mille jooksul iga akna kohta on võimalik teha 15 päringut. REST API võimaldab ka koguda andmestikku minevikust (umbes kuni nädala jagu), kuid vastu ei pruugi saada kogu andmestikku. Piirangutest on teoreetiliselt võimalik kõrvale hiilida kogudes suure hulga arvutivõrkude kaudu andmestikku koos. Need praktikad aga ei lähe kokku Twitteri hea tavaga ning rikkujaid võidakse karistada (nt ligipääsu piiramisega) (Gaffney ja Puschmann, 2014: 59-60; 64 ja 65).

Voogesituse API puhul toimib kogu protsess reaalajas. Kasutaja paneb püsti serveri, ühendab Twitteriga ning esitab taotluse teatud andmestiku koguda ning järk-järgult otse kogutaksegi andmestikku. See tähendab aga, et mingi sündmuse kohta käivate säutsude ehk andmete kätte saamiseks peab olema server kogu aeg üleval ja ühendatud Twitteriga. Minevikust see rakendusliides andmeid ei kogu, kuid seevastu jõuab andmestik kohe otse kasutajale kätte (umbes kahesekundise viivitusega). Kuigi täpseid andmeid pole, siis testid on viidanud, et kasutades seda rakendusliidest on andmestik olnud representatiivne kogu avalike säutsude kohta, välja arvatud piirangjuhtumitel. Uuringutes voogesituse rakendusliidest kasutades võib tekkida probleem, et ootamatute sündmuste uurimiseks peab olema koheselt valmis serverit käivitama ning sellega hakata jooksvalt kohe vastavaid andmeid koguma, mis huvitab (Twitter, 2017; Puschmann ja Burgess, 2014: 46; Gaffney ja Puschmann, 2014: 56-59; Gerlitz ja Rieder, 2013: 2).

Voogesituse rakendusliides toimetab andmeid kohale kolmes mõõdustikus. Niiöelda tuletõrjevoolik (*firehose*) võtab endaga kaasa kogu andmestiku (välja arvatud privaatsed postitused). Niiöelda aiavoolik (*gardenhose*) suudab kaasa haarata

maksimaalselt 10% kogu säutsude mahust, mida konkreetsel ajahetkel postitatakse. Piserdaja (*spritzer*) võtab kaasa maksimaalselt 1% kogu säutsude mahust, mida sel ajal postitatakse. Tuletõrjevoolik on saadaval vaid tasuliste teenustepakkujate nagu Gnip ja DataSift kaudu. Lisaks on Twitter kogu oma sisu annetanud USA Kongressi raamatukogule. Aiavoolik on vaid kättesaadav erijuhtumite puhul esitades Twitterile taotluse. Piserdaja on ainuke valik, mis on tavakasutajaile (ja ka teadlasele, kes soovib andmestikku koguda) tavapäraselt kättesaadav. Enamasti ei pruugi see probleemiks osutuda, kuna 1% võrdub umbes 5 miljoni säutsuga päevas. Harvematel juhtudel võib seda juhtuda ning uurija peab arvestama, kas leppida olemasoleva andmestikuga või maksta täpsuse eest "tuletõrjevoolikut" tellides. Andmestiku täiuslikkust kontrollida pole võimalik (Gaffney ja Puschmann, 2014: 57-58, 65).

Varasemalt oli ka veel olemas kolmas rakendusliides: otsingupõhine. Sarnaselt REST APIle, kasutas see niiöelda tõmbamismeetodit ning oli uurijate jaoks peamine andmekogumisvahend (Gaffney ja Puschmann, 2014: 60). Twitteris oli seda rakendusliidest kasutades andmete kättesaamine täies ulatuses võimalik kuni 2011. aastani, mil Twitter muutis oma rakendusliideste struktuuri ning rakendus "piserdamise" meetod. Samas on Twitteri rakendusliides üldplaanis andmekogumiseks võrdlemisi kasutajasõbralik (Felt, 2016: 2).

## 4.7 Tööriistad Twitterist andmete kättesaamiseks

Gaffney ja Puschmann (2014: 61) tegid ülevaate tööriistadest, kuidas Twitterist andmeid kätte saada. Sealhulgas võrdlesid nad, kas eri tööriistad vajavad hostimist (serveri üleval hoidmist), programmeerimisoskust, kas nad pakuvad niiöelda toorest andmestikku, kas võimaldavad analüütikat ning kas tööriista eest peab tasuma või mitte. Nüüdseks on mitmed neist kas aegunud, lõpetanud tegevuse või üle võetud mõne teise firma poolt. Üks peamisi muutuste põhjustajaid oli eelmainitud 2011. aasta API muutus.

Nii näiteks lõpetas 2011. aastal tegevuse Gaffney ja Puschmanni kirjeldatud vabavaraline Windowsi põhine tööriist The Archivist. Nüüd leiab selle asemel sarnase tasulise tööriista nimega Tweet Archivist, mis kasutades voogesituse rakendusliidest

võimaldab koguda arhiive ulatuses kuni 50 000 säutsu ning visualiseerida neid andmestike (Aidlin 2017; Tweet Archivist, 2017). Küll aga töötab siiani Google Sheeti põhine tööriist TAGS, mis võimaldab otsingu rakendusliidest kasutades koguda märksõnade kaupa säutse kuni 6-9 päeva tagusest perioodist. Antud tööriist on tasuta kättesaadav, kuid samas nenditakse, et terviklikkust paratamatult see tööriist ei võimalda (Hawksey, 2017).

Varasemalt on olnud üheks populaarseimaks tööriistaks uurimistööde andmestike kogumiseks Twapperkeeper. See asendus hiljem ise ülesseatava vabavaralise YourTwapperKeeperiga. YourTwapperKeeper (YTK) kasutab säutsude kogumiseks voogesituse ja otsingu rakendusliidest. Antud tööriist nõuab nii teatud määral programmeerimisoskust kui ka hostimisvõimalust. Võrreldes TAGSiga peaks YTK andma mahu osas tervliklikuma andmestiku, kuid meta-andmestik on võrdlemisi piiratud. Väljundina on saadaval nii HTML, RSS, Excel kui ka JSON formaadid (Gaffney ja Puschmann, 2014: 62). Täiendavalt nenditakse, et ei olda kindlad, kas selle tööriista ülespanek läheb enam kokku Twitteri teenuste tingimustega ning seetõttu hoiatatakse, et seda tuleb kasutada omal vastutusel (O'Brien, 2013).

Sarnase süsteemi alusel kui YourTwapperKeeper töötab Twitter Database Server vabavaraline tööriist. Võrreldes YTK-ga on TDS-i mõnevõrra keerulisem üles seada ning andmete kogumise faasis on väga minimaalne võimalus kogutud andmetega tutvuda. Meta-andmestik on seevastu rikkalikum ning paistab tarbivat vähem arvuti ressurssi (Gaffney ja Puschmann, 2014: 63; 140Dev, 2013).

140kit on tööriist, mis oli sarnaselt YourTwapperKeeperiga varasemalt kasutatav *onlines*, kuid nüüd on vaid lähtekoodide põhjal kättesaadav, mistõttu peab tööriista ise üles seadma. See töötab vaid kasutades voogesituse rakendusliidest. Arvestades, et 140kit oli algusest peale üles ehitatud lähtudes sellest, et sihtgrupiks võiksid olla teadlased, on meta-andmestik võrdlemisi lai. Erinevalt kahest eelmisest tööriistast, ei kasuta 140kit skriptimiskeelena mitte PHP-d, vaid Rubyt (Gaffney ja Puschmann, 2014: 63; 140kit, 2010).

Lisaks eelnevaile on saadaval ka tasulised tööriistad Gnip ja DataSift. Gnip võimaldab koguda andmeid lisaks Twitterile teistest sotsiaalvõrgustikest (Facebook,

Instagram, Youtube jne). Saadav andmestik sõltub pakettist, kuid üleüldiselt on tegemist ühe täiuslikuma võimalusega. Mõningatel juhtudel võib meta-andmestikus puudu olla mõned võimalikud väljad. Gnipiga on võimalik koguda andmeid ka minevikust. Sarnase mudeli alusel töötab DataSift (nüüdseks kogutakse seal andmestikku läbi Gnipi) (Gaffney ja Puschmann, 2014: 63-64; Gnip, 2017; DataSift, 2017).

Storify ja Netlytic on tööriistad, mis oma kasutajamugavuselt on lihtsamad ning kättesaadavamad. Storify spetsialiseerub eelkõige andmestiku visualiseerimisele (nii Twitteri kui ka teiste sotsiaalvõrgustike). Andmestik, mida Storifyga on võimalik koguda on minimaalne, teadustööks pole see sobilik nii andmestiku täiuslikkuse kui ka selles osas, et protsessis jääb nii mõndagi töö loogikast teadmata. Netlytic võimaldab samuti andmestikku koguda erinevatest sotsiaalvõrgustikest, osaliselt on see tasuline teenus. Võrreldes Storifyga on andmestik märkimisväärselt täiuslikum, Netlytic kogub andmeid REST rakendusliidesega (ehk päring iga 15 minuti tagant, kuni 1000 ühikut päringu kohta, kuni nädal on võimalik minevikku vaadata). Paralleelne päringute arv sõltub paketist, võrreldes Gnipi ja DataSiftiga on Netlytic märkimisväärselt odavam. Hiljem on võimalik andmestikku koos meta-andmetega teisendada csv failivormingusse. Netlyticus endas on näiteks võimalik visualiseerida võrgustikuanalüüsi, kuid paremate tulemuste saamiseks tasub ehk kasutada rakendust Gephi (Felt, 2016: 7-13; Storify, 2017; Netlytic, 2017).

Spetsiifilistemate juhtumite puhul võib kalkuleerida, kas on mõttekam ise tööriista luua, kui on olemas tehniline võimekus ja pädevus. Lewis, Zamith ja Hermida (2013: 41-44) tegid just seda kasutades programmeerimiskeelt Python, et koguda ühe Twitteri kasutajaprofiili kohta täielikku andmestikku. See meetod valiti seetõtttu, et säilitada andmestiku suurusest hoolimata sisu tundlikus ja kontekstuaalne nüansirohkus ehk kombineeriti arvutiseeritud ja manuaalseid meetodeid.

## 4.8 DMI-TCAT

DMI-TCAT (*Digital Methods Initiative Twitter Capture and Analysis Toolset*) on üks terviklikumaid vahendeid Twitterist andmestiku kogumiseks, mis arendati välja

Amsterdami Ülikoolis. Valmiskujul on see tööriist kättesaadav vaid selle ülikooli tudengitele ning teadlastele, kuid lähtekoodid on vabavaraliselt kättesaadavad. See tähendab, et igaüks, kellel on serveri hostimise võimalus ning omab teatud programmeerimisoskuseid, saab tööriista vabalt kasutada. Sarnaselt YTK ja 140kitiga on DMI-TCAT kirjutatud PHP programmeerimiskeeles ning andmebaaside loomises kasutatakse MySQL-i. Lisaks andmete kogumisele on võimalik samal ajal teha esmatasandi analüüsi (näiteks säutsude sageduste ajajoon). Andmebaase on võimalik välja lasta csv ja tsv failivormingutes (Felt, 2016: 11; Sue, 2016). Lisadesse on märgitud pildid kasutajaliidesest.

TCAT kogub andmeid kasutades voogesituse ja REST rakendusliidest. Voogesituse puhul rakendub "piserdamise" limiidid ehk parasjagu säutsutavast materjalist on võimalik koguda kuni 1% säutse. Limitatsioonidega kokku puutudes annab tööriist sellest märku (näiteks kui minnakse säutsude korjamisel vastu niiöelda lage) (Borra ja Rieder, 2014: 266). Võrreldes Netlyticuga on TCATil meta-andmestik tunduvalt rikkalikum (35 kategooriat vs 11 kategooriat) (Felt, 2016: 11). Andmekogumine toimib kolmel viisil:

1) juhuslik valim ehk TCAT kogub 1% juhuslikke säutse sellel ajahetkel postitavatest materjalidest.

2) temaatiline valim ehk kasutaja loob niiöelda päringukonteineri, kuhu korjab kokku säutse vastavate märksõnade alusel. Päring võib sisaldada nii teemaviiteid kui lihtsalt märksõnu või fraase, sealjuures võib viimane olla tulusam. Näiteks 2011. aastal koguti Jaapani tsunaami ajal neli korda rohkem säutse kui kasutati päringuks sõna "tsunami" võrreldes teemaviitega #tsunami (Bruns ja Stieglitz, 2014: 3).

3) kasutajate jälgimise valim ehk korraga on võimalik jälgida kuni 5000 valitud kasutaja säutse (Borra ja Rieder, 2014: 266-268).

TCATi tööriista siseselt on võimalik analüütikaga vaadelda säutsude statistilisi ja aktiivsuse mõõtmeid; võrgustike-, kontent- ja geograafilist analüüsi teha; etnograafilist uurimist ning tekstilist hermeneutikat. Väljundina on võimalik alla laadida täies ulatuses andmestikku, kuid samuti näiteks ainult geograafilise andmestikuga säutse ja juhuslikku valimit säutse. Võrgustike jaoks on väljundina võimalik GEFX ja GDF failivormingud (Borra ja Rieder, 2014: 269-273).

| Tööriist | Tasuline? | Nõuab programmeerimist? | Nõuab serveri hostimist? | Pakub toorandmeid? |
|---|---|---|---|---|
| Tweet Archivist | Jah | Ei | Ei | Osaliselt |
| YourTwapperKeeper | Ei | Jah | Jah | Jah |
| TAGS | Ei | Ei | Ei | Osaliselt |
| Twitter Database Server | Ei | Jah | Jah | Jah |
| 140kit | Ei | Jah | Jah | Jah |
| Gnip | Jah | Ei | Ei | Jah |
| DatataSift | Jah | Ei | Ei | Jah |
| Storify | Osaliselt | Ei | Ei | Minimaalselt |
| Netlytic | Osaliselt | Ei | Ei | Jah |
| DMI-TCAT | Ei | Jah | Jah | Jah |

Tabel 1. Uuendatud tööriistade tabel Twitteri andmestiku kogumiseks (aluseks Gaffney ja Puschmann, 2014:61)

# 5. Uuringu eesmärk

Töö empiirilise osa eesmärgid on kahes suunas. Esimeseks eesmärgiks on võrrelda seda, kuidas päris elus toimuvad sündmused mõjutasid Ameerika Ühendriikide presidendivalimiste ajal Twitteris toimuva diskussiooni kulgu. Teiseks eesmärgiks on hinnata kasutatud meetodite tugevusi ja nõrkusi ehk metodoloogiline analüüs. Sellest lähtuvalt püstitasin uurimisküsimused:

1. Kuidas muutus Twitteri diskussiooni kollektiivne emotsionaalne skaala 24 tunni jooksul?
2. Kuidas muutus emotsionaalne skaala võrreldes valimisõhtu kulgemisega?
3. Millised piirangud ja võimalused on Twitterist kogutud andmestiku klassifitseerimise tõlgendamiseks?

# 6. Metoodika

Suurandmete kasutamise mõistmiseks sotsiaalmeedia analüüsis võrdlesin töö empiirilises osas Twitteris toimunud diskussiooni emotsionaalset arengut seoses olulisemate sündmustega, mis leidsid aset Ameerika Ühendriikide presidendivalimiste õhtul ning sellele järgnenud päeval. Võrdluse raamistamiseks kasutasin 24 tunnist perioodi, mis on jaotatud Twitteri kollektiivse emotsionaalse arengu osas kümneminutilisteks intervallideks. Alguspunktiks valisin 18:00 ET (Ameerika Ühendriikide idapoolseim ajavöönd) ehk aeg, mil esimesed valimisjaoskonnad hakkasid sulgema. Andmete kogumiseks kasutasin tööriista DMI-TCAT, andmeanalüüsiks programmi Tableau ning algoritme programmeerimiskeeles R. Need said valitud seetõttu, et sobitusid uurimise püstitusega kättesaadavuse tõttu, võimaluste poolest kui oskustaset arvestades (vajas teatud tasemel programmeerimist, kuid küllaltki algtasemel). Algselt kogusin andmestikku kokku 4 407 249 säutsu, andmeanalüüsiks jäi hiljem alles 2 262 998 säutsu. Kollektiivse emotsionaalse arengu mõõtmiseks kasutasin meelestatuse analüüsi. Säutsud klassifitseerisin kuude kategooriasse: viha, vastikustunne, hirm, rõõm, kurbus ja üllatus. Täiendavalt tekkis juurde seitsmes kategooria - säutsud, mida ei suutnud algoritmid ühessegi kategooriasse paigutada.

## 6.1 Andmete kogumine

Säutsude ja nende meta-andmestiku kogumiseks kasutasin tööriista DMI-TCAT. Tööriista tehnilist võimekust ning kitsaskohti katsetasin valimistele eelnenud päeval. Kuigi DMI-TCAT töötab ka Windowsi ja OS X operatsioonisüsteemides, siis kõige sobilikum keskkond tööriista üles seadmiseks on Linuxi operatsioonisüsteemid. Seetõttu seadsin tööriista üles just Linuxis. Andmeid kogusin temaatilise valimi alusel ehk kogudes säutse vastavate märksõnade ja teemaviidete kaudu. Nendeks olid niiöelda ametlikud ja neutraalsed märksõnad-teemaviited (#decision2016, #electionday, #myvote2016, sõna "election"), mõlema suurema kandidaadi peamiselt kasutatavad teemaviited (Donald Trumpi #MAGA ehk *Make America Great Again* ning Hillary Clintoni #imwithher), täiendavate märksõnadena nimeviited (sõnad "clinton", "hillary", "trump") ning viimaks üks oluline päevakajaline fraas ("swing

state" ehk võtmeosariigid). Andmeanalüüsi optimaalsemaks ressursi kasutamiseks väljastasin andmestikud kolmes osas: 8. november kuni 19:00 ET; 8. november 19:00 ET kuni 9. november 7:00 ET; 9. november 7:00 ET kuni 19:00 ET.

## 6.2 Andmete töötlus ja sortreerimine

Enne andmestiku Tableausse importimist töötlesin neid tööriistaga Gawk. Eesmärk oli meta-andmestiku parendamine: lisaks põhiandmebaasidele lõin iga andmebaasi juurde kolm täiendavat andmebaasi, mis eraldas mainimised, teemaviited ning URLid, mis võimaldavad neid kolme tunnust täiendavalt analüüsida (Bruns, 2015).

Eelnev protsess oli vajalik *botide* ja spämmijate paremaks tuvastamiseks. Nende kolme tunnuse korraga kasutamine on eriti iseloomulik spämmijate tegevusele (Miller et al, 2014: 64). Andmestiku puhastamise faasis eemaldasingi aktiivsemaid *bote* ja spämmijaid. Selleks vaatlesin manuaalselt iga andmestiku juures kasutajate postitusi, kes postitasid ajaraami piires rohkem kui 30 säutsu. Puhastamise tulemusena eemaldasin neist kõik peale kahe kasutaja postitused, kellel ei esinenud vastavaid tunnuseid.

Seejärel jaotasin postitused kümneminutilisteks andmestikeks. Andmestike klassifitseerimiseks ühendasin Tableau programmeerimiskeele R rakendusega RStudio. Klassifitseerimiseks installeerisin kolm R lisapaketi: tm, Rstem ja sentiment. Esimene neist ehk tm (*text mining*) võimaldab üldisemat raamistikku tekstide kaevandamiseks R-s (Feinerer ja Hornik, 2017); Rstem tüvistab tekstides olevaid sõnu (Lang, 2011); ning sentiment loob tööriista meelestatuse analüüsiks, kasutades selleks Bayesiani statistilisi klassifikatsioonimeetodeid (Jurka, 2012). Emotsioonide määramiseks kasutatakse täpsemalt na¨ive Bayesian klassifikatsiooni - see tähendab, et klassifitseerijad võtavad eelduseks selle, et klasside määramisel pole tunnusjoonte väärtused sõltuvuses teiste tunnusjoonte väärtustest (nt objekti suurus ei ole sõltuvuses objekti värvist) (Han, Pei ja Kamber, 2011: 350).

Tableau algoritm, mida kasutasin säutsude tekstilisel osal on järgnev (Loth, 2016; de Vries, 2016; Beran, 2013):

SCRIPT_STR('library(sentiment);classify_emotion(.arg1,algorithm="bayes",
verbose=TRUE)[,7]',
ATTR([Tweettext]))

| Algoritmi osa | Ülesanne |
|---|---|
| SCRIPT_STR | väljundi väärtusetüübi (sõne) määramine |
| library(sentiment) | laadib valmis kasutatava paketi |
| classify_emotion | kasutatav funktsioon |
| .arg1 | algoritmis kasutatav argument |
| algorithm="bayes" | kasutatava algoritmi määramine |
| verbose=TRUE | ülesande logide salvestamine |
| 7 | kategooriate arv |
| ATTR([Tweettext]) | argumendile (.arg1) väärtuse omistamine |

Tabel 2. Algoritmi seletus.

Andmeanalüüsi lõppfaasis arvutasin iga ajavahemiku kohta välja, mitu protsenti konkreetse perioodi säutsudest väljendasid mis emotsiooni.

## 6.3 Andmete tõlgendamine

Enne andmete tõlgendamist viisin juhusliku valimi alusel läbi täpsuse hindamisprotseduuri. Täpsuse huvides pole tegemist *täielikult* juhusliku valimiga. Samuti ei saa väita, et valim on representatiivne. Juhusliku valimi moodustasin võttes eri ajavahemikest iga emotsiooni kohta iga viienda säutsu. Ühe ajavahemiku kohta kogusin 10 säutsu, kokku kümnest ajavahemikust ehk siis kokku iga emotsiooni kohta 100 säutsu. Lisaks kordasin sama tegevust seitsmenda kategooria ehk klassifitseerimata säutsude kohta (edaspidi klassita kategooria).

Täpsuse arvestamiseks hindasin emotsioonide määramist kahes kategoorias: skaalal, kui täpne määratlemine on ning teiseks skaalal kui keeruline oli antud valiku vastu võtmine. Viimast nägin vajalikuks seetõttu, et hinnata, kui raske on üldse emotsioone klassifitseerida.

Andmete tõlgendamiseks koostasin ajajoone olulisematest sündmustest võttes aluseks erinevate uudisportaalide blogid (The Guardian, AP, CNN, NPR, Business Insider, NBC News, Telegraph, 270win). Olulisemate sündmuste alla liigitasin peamiselt valimisjaoskondade sulgemised osariigiti, tulemuste välja kuulutamine osariigiti (seega ka hetkeseis) ning muud märkmisväärsed sündmused (näiteks hetk kui erinevad ennustusportaalid ning ennustusmudelid hakkasid ennustama Trumpi võitu; Trumpi võidukõne). Andmete tõlgendamiseks võrdlesin emotsiooni sisaldavate säutsude protsentuaalsust ajajoonel toimuvaga.

# 7. Metodoloogiline analüüs

Enne andmete tõlgendamist on oluline hinnata, mida on andmestikust võimalik välja lugeda, mida mitte. Täiendavalt heidan pilgu üldisemalt antud töö protseduuri põhjal suurandmete kasutamisele sotsiaalmeedia analüüsis.

**Andmestiku tõlgendamise piirangud ja võimalused**

Kogutud andmestiku põhjal ei saa väita, et minu saadud pilt Twitteris valitsenud kollektiivsest emotsionaalsest skaalast oleks täielik. Esiteks pole andmestik täielik. Märksõnade ja teemaviidete abil üritasin haarata võimalikult laiapõhjaliselt valimit, mida konkreetses ajavahemikus Ameerika Ühendriikide presidendivalimistega seotult säutsuti. See aga ei tähenda, et kõik asjakohased säutsud mahtusid andmestikku ja kõik asjakohatud säutsud jäid välja. Märksõnade valikut mõtestasin ka seeläbi, et võimalikult vähe haarata kaasa säutse, mis poleksid relevantsed. Seetõttu jäid välja kahe teise peamise kandidaadi (Gary Johnson ja Jill Stein) nimeotsingud. Mõlema kandidaadi ees- ja perekonnanimed on väga levinud, mistõttu on tegemist problemaatiliste märksõnadega, mis oleks potentsiaalselt tekitanud liigselt müra. Kahe peamise kandidaadi nimesid võib pidada vastupidisteks näideteks – minu subjektiivsel hinnangul on Trump, Hillary ja Clinton üsnagi unikaalsed nimed, millega seostatakse eelkõige neid kahte inimest. Clintoni puhul võidakse ka samaväärselt silmas pidada ekspresidenti Bill Clintonit, aga ka tema on teemakohane inimene, Trumpi nimega seostatakse ka tema lapsi, kes on samuti teemaga seotud. Donaldi nime ei kasutanud märksõnana, kuna täheldasin, et Donald Trumpi eesnime kasutati eelkõige koos perekonnanimega, aga mitte eraldi. Märksõnade määratlemisel kaalutlesin samuti lisada jooksvalt teemaviiteid, mis hakkavad õhtu jooksul levima ning lisada neid andmestiku kogumisse jooksvalt. Seda ma küll ei teinud, kuna jõudsin järeldusele, et andmestik ei saa igal juhul olema täiuslik ning teiseks katavad senised kasutatud märksõnad paljuski sealolevaid säutse.

Teine oluline andmestiku piirang oli eelnevates peatükkides mainitud voogesituse rakendusliidese nn piserdaja omadus koguda maksimaalselt 1% kõikidest säutsudest, mis sel ajahetkel postitatakse. Tavaliste sündmuste puhul poleks see olnud oluline

takistus. Antud juhul aga osutus see takistuseks, kuna DMI-TCAT tööriista kasutades anti regulaarselt märku, et teatud ajahetkel saavutati lagi (nii palju kui jälgisin, siis see toimus peaaegu et igal ajahetkel). Esines ka anomaaliaid. Esimese tunni jooksul ületas kümneminutiliste tsüklite jooksul kogutud maht kaks korda üle 28 000 säutsu. Hiljem jäi see number keskmiselt umbes 15 000 säutsu juurde. Siin aga tuleb arvestada, et andmeanalüüsi selles faasis ei arvestanud Tableau edasisäutse, välja arvatud manuaalselt loodud (pannes RT postituse ette). See aga ei tohiks olla selle anomaalia täielik seletus. Ei ole tõenäoline, et 18:20-18:29 ET säutsuti peaaegu kolm korda rohkem kui 19:20-19:29 ET (29 606 vs 10 218). Andmestiku kontrollimine on võimatu, kui pole just ligipääsu täielikule andmebaasile. See aga kõik viitab, et ka piserdamisel ei pruugi alati 1% säutsudest kätte saada või siis vastupidiselt teatud tingimustel kogub piserdamine rohkem kui 1% säutsudest. Rakendusliides jääb seega osaliselt ikkagi mustaks kastiks.

Kolmas piirav tegur kollektiivse emotsionaalse skaala täpsuse hindamisel on klassifitseerimine. Nagu Grimmer ja Stewart (2013: 269-271) eelnevalt viitasid, ei saa keele olemuse automatiseeritud klassifitseerimine olla kunagi absoluutselt täpne. Need võivad anda väljundi, kuidas konkreetse mudeli alusel Twitteris toimuva emotsionaalse skaala proportsionaalsus on, aga see ei anna ammendavat kollektiivset emotsionaalset pilti. Siin tekib ka omaette küsimus, mis säutsude emotsionaalsuse hindamine endas kätkeb – kas see tähendab, et hinnatakse säutsudes väljendatavat emotsiooni või säutsudes sisalduvat emotsiooni? Esimene neist tähendab, et säutsu autor väljendab emotsiooni; teine viitab, et selline emotsioon eksisteerib miskis vormis. Näiteks lause "Kõik on rõõmsad pärast võitu." – siin esineb väike semantiline nüanss, mistõtttu see lause ei pruugi tähenda, et lause autor on samuti rõõmus - pole kindel, kas ta mahub kirjeldatud *kõigi* hulka, ta võis olla ka kõrvaltvaataja. Lisaks tuleb arvestada sellega, et üks säuts võib endas sisaldada mitmeid erinevaid konkureerivaid emotsioone. Üks emotsioon võib olla domineeriv, kuid emotsioonid võivad olla ka võrdse väärtusega. Selle täpseks hindajaks saab olla eelkõige lause enda autor. Kirjapildis võib ühes ja samas lauses olla domineeriv emotsioon ühe jaoks üks, teise jaoks teine. Manuaalsel kodeerimisel on emotsioonide klassifitseerimine juba raskuskoht, rääkimata automatiseeritud meetodist.

Täiendav küsimus on, millised emotsioonid valida klassifitseerimiseks ning mis alusel neid peaks liigitama. Oma töös kasutasin mudelit, mis valitud tööriistade kasutamisel oli kättesaadav ning oskuslikult teostatav. Enda spetsiifilise meetodi välja arendamist ei pidanud vajalikuks ega ka võimekuses olevaks. Teadvustasin küll näiteks seda, et lähtudes valimisõhtu eripärast saab väga populaarseks sõnaks olema *win* ehk võit, mis liigitub rõõmsa emotsiooni alla – pidevalt tuli teateid selle kohta, kuidas üks või teine kandidaat mõnes osariigis võitis. Andmestiku kontrollides selgus, et siiski leidus ka säutse, kus *win*i sisaldavad säutsud olid klassifitseeritud teiste kategooriate alla (kurbus, üllatus) sobivalt.

Twitter on küll peamiselt tekstipõhine platvorm, kuid säutsude juurde kuuluvad alatasa ka pildid, videod ja lingid. See osa säutsudest jääb antud metoodikast lähtudes analüüsimata ning seega jääb teatud osa emotsionaalsuse faktoritest hindamata. Seetõttu tuleb täpsustada, et antud töös ei vaatle ma säutsude emotsionaalsust, vaid säutsude kirjaliku osa emotsionaalset täpsust.

Säutsude olemuse ning emotsionaalsuse mõistmiseks peaks ka säutsude vaatamisele lähenema psühholoogilisest vaatepunktist – mida säutsumine kui tegevus tähendab selle inimese jaoks ning mida see tema jaoks väljendab. Välist tsensuuri säutsumisel pole (kui sisu ei lähe just vastuollu Twitteri kasutustingimustega), küll aga võib esineda enesetsensuuri. Samas esineb palju enesekriitilise meeleta vabas vormis säutsumist (see ilmnes paljuski proovivalimis, kus vulgaarsused ja muul moel viisakast sõnavarast erinev sõnakasutus olid levinud). Subjektiivsel hinnangul on säutsud tehniliste piirangute tõttu (tähemärkide arv) teatud mõttes oma olemuselt võimendatud kujul emotsiooni lööklauselik väljendus - lühikesse säutsu peab mahutama oma mõtte tuumiku kompresseeritul kujul. Säutsumisel peab ka olema tahes-tahtmata mingisugune eesmärk – olgu selleks lihtne enese hetkeseisu väljendus, tähelepanuvajadus või soov informeerida. Avalikult ning veel enam teemaviidet kasutades teadvustatakse, et on olemas auditoorium, kelleni säutsumine kui tegevus jõuab. Säuts peab sel juhul ka ühel või teisel moel kõnetama.

## 7.1 Tehniline külg

Kasutatud klassifitseerimissüsteem ehk na ̈ive Bayesian klassifikatsioon liigitab säutsud sõnade alusel kategooriatesse - igale sõnale on omistatud emotsioon, mida ta väljendab. Kokku on emotsioon omistatud 1542 sõnale, kuid esineb sõnakordusi, kus ühele sõnale on omistatud mitu erinevat emotsiooni. Näiteks sõna *detest* (põlgama) liigitub nii viha kui ka vastikustunde alla. Säutsudes arvutatakse välja erinevate kasutatud sõnade summa ning tõenäosuslikkuse alusel määratakse kategooriasse. Kui emotsioonid on võrdsed, pole võimalik emotsiooni välja arvutada. Oluline on märkida, et süntaktilisi omadusi nagu bigrame (sõnapaare) ja trigrame (sõnakolmikuid) ei arvestata - iga sõna on omaette nähtus ning pole sõltuvuses tema ümber kasutatavatest sõnadest. See piirab paratamatult ka täpsust – samas võib nende kasutamine olla ebatäpsuse allikas ning lihtsustus võib olla täpsem variant, kui pole välja töötatud tõestatult täpsemat meetodit.

Tehnilise külje pealt tasub andmete kogumisel olla pigem ettevaatlikum ning hajutada riske. Kui mingi tehniline aspekt ei hakka tööle, siis hiljem sobivat andmestikku kätte saada on keeruline, kui mitte võimatu. Kuigi alustasin andmete kogumise katsetamist eelneval päeval, siis esines meetodi sissetöötamisel mitmesuguseid erinevaid tehnilisi probleeme, mida tuli järk-järgult lahendada. Sealjuures kasutasin igaks juhuks andmete kogumiseks kolme eri arvutit: neist ühe sain lõpuks õnnestunult tööle, teises arvutis andmete kogumine tõrkus aeg-ajalt ning kolmandas arvutis ei õnnestunudki tööriista tööle saada.

## 7.2 Andmete puhastamine

Andmete puhastamise eesmärgiks seadsin eemaldada suuremad spämmijad ja *botid* nii palju kui manuaalselt mõistlik on. Sobivat automaatset mudelit selleks ei leidnud. Suures plaanis see küll ilmselt erilist tulemust ei andnud - kokku eemaldasin 0,48% säutsudest ehk 20 976 säutsu. Küll aga andis see pildi nende olemusesse - peaaegu et kõik eemaldatud kasutajatest mahtusid kahte kategooriasse: ainult edasisäutsuti või ainult postitati linkidega säutse. Vahel neid meetodeid kombineeriti. Andmete puhastamise sisuline eesmärk oli eemaldada proportsionaalselt suurema häälega

võltskasutajad, kes kunstlikult mõjutavad loodavat pilti, mis oleks kajastanud kvantitatiivses analüüsis.

Lähtudes sellest, et Miller et al (2014: 65) hinnangul on umbes 6% Twitteri kasutajatest spämmijad ning sellest, et Kollanyi, Howardi ja Woolley (2016: 4) hinnangul oli 17,9% valimisõhtul postitustest tehtud kõrgendatult automatiseeritud kasutajate poolt, siis võib väita, et andmete puhastamise täieliku eesmärki ei õnnestunud täita. Harustunud võrgustikuga spämmimist on keerulisem tuvastada ning kindlasti jäi spämmijaid veel andmestiku sisse. Kollanyi et al (2016: 4) sõnasid veel, et valimisõhtul moodustus kõrgendatud automatiseeritud kasutajate seas iga ühe Clintoni säutsu kohta viis Trumpi säutsu. Samas peab märkima, et andmestikes oli umbes 1,5 säutsu kasutaja kohta, mis viitab, et suur osa kasutajaid ei suutnud väga suurel määral diskussiooni domineerida.

## 7.3 Eetika

Antud uurimistöös ei ole välja toodud ega esitletud eraldi ühtegi kasutajat, vaid üritan tabada mustreid, mis suurandmete põhjal joonistuvad. Andmestik on kogutud avalikest säutsudest, mida on kasutatud enamasti teemaviidete kaudu, mida võib tinglikult liigitada poolavalikuks või avalikuks sfääriks - kasutaja on väljendanud ennast teemaviiteid kasutades sedasi, et ta soovib osaleda laiemas diskussioonis. Need säutsud, mis pole teemaviiteid kasutanud, on samuti postitatud avalikult. Andmeanalüüs sisaldab vaid säutsude teksti, kuid mitte muud informatsiooni kasutajate kohta. Säutsud võivad küll sisaldada kasutajaid, kellele need säutsud on suunatud. Siiski ei arva ma, et oleksin antud uurimistöös kellegi privaatsust otseselt riivanud. Kokkuvõtvalt seetõttu ei näe vajalikuks tõstatada küsimust, kas eetiliselt oleks vajalik esitada täiendavaid küsimusi. Siiski pidasin vajalikuks eraldi tõstatada eetika alapeatükid selleks, et rõhutada, et see osa tööst ei jääks suurandmete analüüsis unustamata.

## 7.4 Proovivalimi hindamine

Proovivalimis hindasin kokku 700 säutsu kahe küsimuse alusel: kui täpne oli arvuti ning kui raske on säutsu klassifitseerida. Iga emotsiooni kohta valisin proovivalimis välja 100 säutsu, sama tegevust kordasin klassita säutsude puhul. Proovivalimi hindamises on rõhk eelkõige emotsiooni sisaldavatel säutsudel, kuid ka analüüsin klassita säutse.

| Koondtabel | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| 5 | 18 | 42 | 62 | 21 | 19 | 23 |
| 4 | 23 | 21 | 11 | 13 | 24 | 23 |
| 3 | 17 | 11 | 12 | 11 | 22 | 22 |
| 2 | 9 | 13 | 4 | 12 | 17 | 12 |
| 1 | 18 | 9 | 6 | 30 | 10 | 6 |
| 0 | 15 | 4 | 5 | 13 | 8 | 14 |

Tabel 3. Proovivalimi hindamine küsimusele "Kui hästi on automatiseeritud hindamine tabanud säutsu emotsiooni?": 5 - hästi; 4 - pigem hästi; 3 - nii ja naa; 2 - pigem halvasti; 1 - halvasti; 0 - ei oska öelda

Proovivalimi põhjal võib öelda, et kõige rohkem kattub minu hinnang arvuti hinnanguga säutsude emotsionaalsusele hirmu ja vastikustunde puhul. Kõige suurem lahknevus esines rõõmu emotsionaalsuse hindamisel. Siinkohal peab taaskord rõhutama, et tegemist pole ei täiesti juhusliku valimiga ega ka representatiivse valimiga - kui koguvalim ületab kahte miljonit, siis 600 säutsu ei suuda pädevalt esindada koguvalimit. Siiski võib saadud andmestiku mustrite põhjal järeldusi teha. Oluline on ka siinkohal märkida, et säutse oli mitmes eri keeles - saksa, hispaania, prantsuse, portugali, itaalia, norra, rootsi. Kui oli võimalik, siis üritasin tõlkida ning aru saada säutsu sisust. Kui see ei õnnestunud, märkisin säutsud kategooria 0 alla. Keeleti ei tekkinud suuremat erisust täpsuse osas.

| Hindamisraskus | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| 3 | 19 | 26 | 21 | 35 | 56 | 47 |
| 2 | 42 | 25 | 15 | 16 | 22 | 33 |
| 1 | 39 | 49 | 64 | 49 | 22 | 20 |

Tabel 4. Proovivalimi hindamine küsimusele "Kui keeruline oli määrata antud säutsu emotsionaalsust?": 3 - keeruline; 2 - keskmine; 1 - lihtne

Täiendava abimeetmena võtsin kasutusele emotsionaalsuse hindamisraskuse. Laiendatud versioon tabelist on lisatud lisadesse. Antud hindamise eesmärk oli täiendada mustrite märkamise protsessi - seeläbi on võimalik mõista, kas valitud emotsiooni ongi raske hinnata või on tulemused lihtsalt valesti määratud. Rõhutada tuleb ka, et minu kategoriseerimine ei pretendeeri samuti tõele – mida viitab ka hindamisraskuste tabel – kolmandikul juhtudel oli väga keeruline hinnata täpsust. Selgelt on näiteks märgata, et kurbust ja üllatust oli minu hinnangul üsnagi keeruline määratleda. Kui vastikustunde ja hirmu osas sobitus minu hinnang täpsuse osas kõige enam arvuti omaga, siis sarnaselt oli neid emotsioone minu poolt lihtsaim määratleda. Samuti oli rõõmu võrdlemisi lihtne osaliselt määrata.

### 7.4.1 Viha

Viha määratlemisel oli üheks suurimaks takistuseks nii viha olemuse mõistmine kui ka konteksti tajumine. Näiteks:

> "*Self-loathing that I've been more invested in the UK general election, Referendum & tonight's US election than any Canadian election ever*"[1]

Liigitasin selle kategooria 3 (tabel 1) alla ning raskusastmeks 3 (tabel 2). Säutsu peamine mõte ei ole otseselt väljendada viha, vaid nentida kurioosiumit selle üle, et kasutaja pole oma enda riigi valimistest olnud nii huvitatud ja emotsionaalselt investeeritud kui teiste omadest. Kategoriseerimise põhjuseks oli siin sõna *loathing* (jälestus), mis liigitub nii viha kui vastikustunde alla. Siiski võib väita, et ta väljendab viha selles osas, et isik on vihane enda peale, et teda ei huvita niivõrd enda riigi käekäik, kui teiste omad. Teisiti võib seda võtta kui pettumust, kuid seda kategooriat eraldi antud töös ei kasutanud.

> "*RT @VENUSRICCI: YOU GUYS ARE SO /.../ ANNOYING WHY ARE YALL LETTING TRUMP WIN https://t.co/pPghBTlpiK*"[2]

[1] eestikeelselt: Jälestan ennast, et ma olen olnud rohkem investeeritud Ühendkuningriikide üldvalimistest, referendumist ja tänasest USA valimistest, kui eales ühegi Kanada valimistest.
[2] e.k: edasisäuts @VENUSRICCI: te olete nii tüütud, miks te lasete Trumpil võita

Üks levinud viis kirjalikult viha väljendamiseks on trükitähtedega liialdamine. Antud säutsu liigitasin kategooriasse 5 ning raskusastmesse 1. Antud säutsu teeb huvitavaks see, et see sisaldab sõna *win*, mis kategoriseerub rõõmu emotsiooni alla ehk siin on tekkinud teoreetiliselt emotsioonide konkurents, kuigi seda praktikas pole. Siiski on suutnud arvuti tuvastada õigesti, et säuts väljendab viha.

> "*If trump wins then prepare for world war III*"[3]

Liigitasin selle kategooriasse 4, raskusastmeks määrasin 1. Jällegi pole viha otseselt väljendatud: kui juhtub sündmus A, siis tõenäoliselt juhtub sündmus B. Küll aga peitub alatoonis see, et Trumpi võidukorral juhtub katastroof - uus maailmasõda. Võib väita, et see lause on segu pettumusest (et Trump võib võita), ootusärevusest (hüperbooli püstitus), vihast (sõda on olemuselt üldinimlikult halb) ja vastikustundest Trumpi vastu.

### 7.4.2 Vastikustunne

Vastikustunde määramisel seisnes peamine probleem selles, et paljuski kattus vastikustunde emotsioon viha ja ka üllatuse omaga - kui oleks pidanud määrama millist valida, siis poleks suutnud alati olla valikus kindel. Viha ja vastikustunde kattuvust on samuti märgata sõnastiku andmebaasist. Näiteks:

> "*RT @sanchezzmanuela: Trump might say a couple offensive things so instead vote for the most corrupt politician of all time!!! Makes total sense!!!*"[4]

Kas see säuts väljendab eelkõige vastikustunnet selle üle, et Hillary on korrumpeerunud ning valijad on silmakirjalikud või otseselt viha Hillary Clintoni ja tema silmakirjalike valijate suunas? Ise määratlesin seda kategooriasse 4, raskusastmeks 2 – säutsuja väljendab vastikustunnet ja tema mitterahulolu sellega, et

[3] e.k: Kui Trump võidab, siis valmistuge kolmandaks maailmasõjaks.

[4] e.k: edasisäuts @ sanchezzmanuela: Trump võib öelda mõningaid solvavaid asju, nii et selle asemel hääletagem kõige korruptiivsema poliitiku pooles eales!!! Väga loogiline!!!

elu on ebaõiglane. See, et säuts sisaldab ka viha emotsiooni pole takistus – inimene võib tunda erinevaid emotsioone üht ja sama sõnumit väljendades.

> "*Whilst the idea of President Donald Trump is a hideous thought I hope he wins. If he loses the left and right will continue to be one party*"[5]

See on huvitavam näide, kuna selle mõistmiseks peab tajuma natukene üldisemat konteksti – kuigi Donald Trump oli vabariiklaste kandidaat, ei ole ta kunagi olnud karjääripoliitik ega kuulunud poliitilisse eliiti. Vastikustunne peitubki siin poliitilise eliidi vastu – vahet pole kummal pool telge, mõlemal pool olevat kõik samasugused. Trump väljendab endas eliidivastast kandidaati, kes peaks raputama poliitilist süsteemi. Algoritm tuvastas siin vastikustunde tõenäoliselt lause esimese poole järgi ehk subjektiks oli Trump, kelle vastu ka postitaja väljendab teatud vastikust, aga mitte kuigi tugevalt.

> *"I am legitimately nauseous over this election."*[6]
> *"@saisonbanthony hearing "Donald Trump wins...." followed by a state is enough to turn my stomach"*[7]
> *"RT @nfilosa17: after watching 13 hours, it makes me sick to my stomach thinking people could actually vote for Hillary"*[8]
> *"NEW SEASON OF AMERICAN HORROR STORY: AN INAUGURAL SPEECH BY DONALD TRUMP"*[9]

Need on mõned näited, mille puhul väga palju ei kõhelnud, kas nõustuda arvuti otsusega või mitte. Viimane on võrreldes teistega küll kaudne vihje vastikusele - see, et Trumpist saab president tekitas säutsujas õõvastust. Iiveldustunde tekitamine ja otseselt sõna vastikus (*disgust*) olidki peamised tunnused, mille alusel oli vastikustunde alla liigitamine üsnagi täpne.

---

[5] e.k.: Kuigi idee Donald Trumpist kui presidendist võib olla jube mõte, siis loodan, et ta võidab. Kui ta kaotab, on vasak- ja parempoolsed jätkuvalt sisuliselt üks ja sama partei.
[6] e.k: Mul tekkis tõsiselt iiveldustunne selle valimiste tõttu.
[7] e.k: @saisonbanthony kuuldes "Donald Trump võitis...", millele järgneb osariigi nimi on piisav, et tekitaks mu kõhus iiveldust.
[8] e.k: edasisäuts @nfilosa17: pärast 13 tundi vaatamist, ajab mind endiselt iiveldama mõte, et keegi päriselt hääletas Hillary poolt.
[9] e.k: uus hooaeg Ameerika õudusjuttudest: ametisseastumise kõne Donald Trumpi poolt.

### 7.4.3 Hirm

Hirmu oli kõige lihtsam käsitsi määratleda ning samuti kattus see kõige enam algoritmi valikuga. Ajendiks olid eelkõige sõnad hirmutav (*scary* eri vormid), ehmatav (*frightening*), hirmunud (*terrified*), paanika (*panic*):

> "*RT @narwalzouis: this election has me genuinely terrified for people like my father.*"[10]
> "*I'm scared for the USA and I feel this election has and will continue to make things worse I should have ran for president. #ElectionDay*"[11]
> "*Trump is leading everywhere no one expected him to this is frightening. Truly frightening.*"[12]

Hoolimata sellest, et hirm oli selgelt kõige kattuvam emotsioonimääratlus võrreldes algoritmide jaotusega, leidus siiski 10% säutse, millega ma kas polnud üldse nõus või pigem polnud nõus. Sarnaselt varasemaga leidus ka siin erinevate kategooriate kattuvusi:

> "*Lmao at the everyone who is scared about who gonna win #ElectionDay*"[13]

Siin on segunenud kaks emotsiooni: üks emotsioon, mis valdab säutsu kirjutajat (rõõm, naer) ning teine emotsioon, mis valdab viidatud seltskonda (hirm). Selle puhul on ilmne see, et esimest neist algoritm ei tuvastanud – see on internetis levinud akronüüm, mida sõnastikes ei leidu. Kui uurida meediumeid, kus selline sõnavarastik on levinud, peaks sõnastiku ka vastavalt kohandama ja kaasama ka tavapärases mõistes sõnadena mitte arvesse minevaid termineid.

[10] e.k: edasisäuts @narwalzouis: need valimised on mind pannud siirast hirmu tundma inimeste vastu nagu mu isa.
[11] e.k: Ma kardan Ameerika Ühendriikide pärast ja ma tunnen, et need valimised on ja tulevikus jätkuvalt teevad asju hullemaks. Ma oleks pidanud ise presidendiks kandideerima. #valimistepäev
[12] e.k: Trump juhib igal pool, kus keegi ei oleks oodanud temalt seda. See on ehmatav. Tõsiselt ehmatav.
[13] e.k: naeran kõigi nende üle, kes kardavad seda, kes kohe on võitmas. #valimispäev

### 7.4.4 Rõõm

Proovivalimi puhul võib väita, et rõõmu määratlemine on algoritmi poolt kõige ebaõnnestunum. Samas ei saa seda täies kindluses väita. Siinkohal toon välja, et raskused esinesid peamiselt keskmiste kategooriate (2-4) määratlemisel - 31 säutsu 36-st oli keeruline neis astmes määratleda. Kategooriate otstes (kategooriad 1 ja 5) oli määratlus selgem. Mõnel juhul aitas aru saada emotsioonidest emotikonid – kontekst. Rõõmu täpsemad määratlused olid tihti mõeldud naljadena:

> "*The kind of trump supporters I'm familiar with (those in the south) are not the kind of folks who answer phone polls.*"[14]
> "*RT @Wyatt_Griffith2: I bet Hillary wishes she could delete Donald Trumps votes like she did those emails.. #MakeAmericaGreatAgain*"[15]

Halbadel määratlemistel selgelt ühist joont ei suutnud tuvastada, välja arvatud, et sõna *win* (võit) andis viiel korral valepositiivsuse. Kaks neist väljendas imestust, üks viha, kaks ärevust tulemuste osas. Üks neist oli tingitud sellest, et algoritm ei võta süntaksi arvesse – sõnapaarina oleks olnud võimalik tuvastada, et sõnale eelnes mitte (*don't*). Küll aga võib välja tuua, et rõõmuga on sõnastikus seotud kõige enam eri sõnu - 553. Liiga laia sõnavara seostamine ühe emotsiooniga võibki seetõttu tuua vale korrelatsiooni, kus sõna emotsioon tuleneb rohkem kontekstist, mitte ainult sõna enda emotsioonist.

### 7.4.5 Kurbus

Kurbuse puhul oli valitud emotsioonide seast kõige raskem määratleda, kas ma nõustun sellega, et säuts sisaldab kurbust või mitte – koguni 56 säutsu põhjal pidin pikemalt mõtisklema, millist varianti valida. Jällegi on see väiksem kategooriate otstes – "halvasti" määratluses oli raske emotsiooni määratleda kolmel korral, "hästi" määratlusel ühel korral. Näiteks oli keeruline määratleda seda säutsu:

[14] e.k: See Trumpi valijate tüüp, kellega ma tuttav olen (need seal lõunas), ei ole seda tüüpi inimesed, kes vastaksid telefoniküsitlustele.
[15] e.k: edasisäuts @Wyatt_Griffith2: Ma panustan, et Hillary loodab, et ta suudaks kustutada Donald Trumpi hääli samamoodi nagu ta kustutas neid e-maile. #teemeameerikasuureksjälle

"*All that we can do at this point is pray that God will guide Trump to make the best decisions for our country.*"[16]

Emotsionaalsus pole siin selgelt väljendatud, kuid on pigem tunnetatav. Viide kurbusele või ka pettumusele peitub fraasis *at this point* (selles punktis). Ilma selleta ei oleks osanud säutsu emotsiooni määratleda. Fraas viitab sellele, et tekkinud olukord on tinginud põhjuse palvetamiseks, seetõttu et Trump suudaks teha õigeid otsuseid riigi jaoks. *Kõik*, mida teha saame, on palvetamine – väljendab uskumust, et on põhjust kahtlemaks, et Trump on võimeline õigeid otsuseid tegema. See viitab, et säutsuja on pigem skeptiline Trumpi osas. Kas autor on kurb või lihtsalt ettevaatlik, on keeruline öelda. Pigem võiks väita, et ta on kurb, kuna ta ei oodanud seda olukorda ning ei oska antud situatsioonis midagi muud teha. Kindlalt ma seda väita siiski ei julge.

"*RT @sportsbetcomau: Anyone else just get this FB invite? #Election2016 #electionday https://t.co/RUFiM9UD7l*"[17]

Antud säuts on aga näide sellest, et ilma lingita või pildita võib konteksti määramine olla keeruline - teksti põhjal polnud võimalik väita, kas säutsuja on kurb või mitte. Kindlalt ei saanud kumbagi öelda, mistõttu jätsingi selle määratlemata. Analüüsi osana linkide vaatamine ei tundu otstarbekas turvalisuse kaalutlustel - t.co algusega lingid on Twitteri poolt lühendatud URL, mille puhul ei tea, mis selle taga peitub.

### 7.4.6 Üllatus

Üllatuse kategoorias osutus üheks segasemaks sõna *wonder*, mis võib tähendada nii imestust (ja seega ka üllatust) kui ka uudishimu (huvitav, mis juhtuks):

"*Wonder how twitter is taking tonight's election results*"[18]

[16] e.k: Kõik, mida teha saame selles punktis on on palvetada jumala poole, kes juhib Trumpi, et ta teeks parimaid otsuseid meie riigi jaoks.
[17] e.k: edasisäuts @sportsbetcomau: Kas keegi sai just selle Facebooki kutse? #valimised2016 #valimistepäev

"*I wonder how my daddy feels rn. I never asked him about the election. I do know he loves Obama.*"[19]

"*#Trump I wonder who writes his speeches*"[20]

Nende kolme näite puhul pole selge, kas sõna *wonder* väljendab imestust või uudishimu. Esimese säutsu puhul võib ühtepidi väita, et ta väljendab imestust läbi selle, et tulemused olid ootamatud ning ta soovib seetõttu näha, kuidas Twitter sellele reageeris. Teistpidi võib väita, et ta ei imestanud tulemuste üle, aga tundis siiski huvi Twitteri arutelu üle. Teise puhul võib väita, et tulemused olid üllatavad ning säutsuja imestab seetõttu, kuidas tema isa kui Obama pooldaja võis sellele reageerida. Üllatuse osa pole aga ka siin selgelt väljendatud. Kolmanda näite puhul võib säuts väljendada imestust, et kes nii üllatavalt hästi või halvasti neid kõnesid kirjutab või vastupidiselt mõtiskleda, kas ta ise kirjutab neid või mitte, ilma emotsioonita.

Üllatuse kategooria oli sarnane teiste kategooriate jaotustega, kuid oli pigem keerulisem määratleda. Üldiselt pigem nõustusin saadud tulemustega kui mitte. Ainult kuuel puhul ei nõustunud algoritmiga täielikult.

### 7.4.7 Klassita kategooria

| Koontabel | 5 | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|---|
| 3 | 1 | 16 | 22 | 7 | 2 | 6 |
| 2 | 7 | 5 | 2 | 3 | 13 | 7 |
| 1 | 3 | 0 | 0 | 0 | 3 | 3 |

Tabel 5. Proovivalimi hindamine küsimustele "Kas oleks pidanud säutsu kuhugi kategooriasse määrama?" ja "Kui keeruline oli määrata valikut?". Esimese puhul: 5 - jah; 4 - pigem jah; 3 - nii ja naa; 2 - pigem ei; 1 - ei. Teise puhul: 3 - keeruline; 2 - keskmine; 1 - lihtne

Suurem osa säutsudest ehk ümmarguselt 70% polnud algoritmi poolt klassifitseeritud. See on märkimisväärne arv, olgu selleks põhjuseks see, et arvuti ei suutnud mõnest muust keelest kui inglise keel täpselt aru saada; see, et polnud mingist sõnast kinni haarata; see, et erinevad emotsioonid konkureerisid või midagi neljandat.

[18] e.k: Huvitav, kuidas Twitter on tänaseid valimiste tulemusi vastu võtnud

[19] e.k: Mind huvitab, kuidas mu ise tunneb praegu. Ma pole veel temalt seni küsinud valimiste kohta. Ma tean, et ta armastab Obamat.

[20] e.k: #Trump Huvitav, kes ta kõned kirjutab

Raskusastmelt oli klassita kategooria võrdlemisi keeruline – vaid 9 säutsu puhul oli lihtne otsustada. See on selgelt väiksem võrreldes emotsionaalsete säutsude hindamisega, kus madalaim arv oli 20. Kõige enam peegeldus määratlemata säutsude seast üllatuse emotsioon:

> "*wow Donald Trump the president of the united states...lmao i'm looking forward what's going to happen lmao*"[21]
> "*RT @ilgiornale: #Trump, una vittoria contro tutti https://t.co/Wj3PbRe7aK https://t.co/oCAFTnasUi*"[22]
> "*woke up to donald trump being president.. can i go back to sleep*"[23]

Esimese puhul oleks võinud tuvastada üllastust sõna *wow* kaudu, kui ta oleks olnud osa sõnastikust. Taaskord viitab see, et antud meediumis peab ka arvestama teatud määral slängi ja lühenditega. Teise ja kolmanda puhul oleks üllatuse määramine automatiseeritud kujul keerulisem – üllatus tuleneb kontekstist. Paljude teiste säutsude puhul tekkis korduvalt küsimus, millise emotsiooni peaksin valima. Emotsiooni sisaldus oli selgelt näha, kuid ei suutnud alati määratleda, mis ta peaks olema.

> "*JUST BECAUSE I'M NOT AMERICAN DOESN'T MEAN I CAN JUST TAKE MY EYES OFF THE FACT THAT DONALD /.../ TRUMP GOT ELECTED I'M ABOUT TO CRY*"[24]

Antud säutsu puhul võib väita, et domineeriv emotsioon on viha, kurbus või ka üllatus. Samas võib ka väljendada säuts eelkõige vastikustunnet (Trumpi ja olukorra vastu). Siin on selgelt õigustatud säutsu mitte määramine. Alternatiivse variandina oleks saanud antud säutsu määrata kõikidesse mainitud kategooriatesse. Säutsuja poolne emotsioon selgelt eksisteerib.

---

[21] e.k: vau, Donald Trump - Ameerika Ühendriikide president. Hahahaa, ma ootan, mis nüüd juhtuma hakkab.
[22] e.k (robustne tõlge): edasisäuts @ilgiornale: Trump, võit kõigi (ootuste) vastu/vastu kõigi arvamust
[23] e.k: ärkasin üles ja nägin, et Donald Trump on president. Kas ma võiksin uuesti magama minna?
[24] e.k: Lihtsalt seetõttu, et ma pole ameeriklane ei tähenda, et ma ei saa oma silmi ära sellelt faktilt, et Donald /.../ Trump valiti presidendiks. Ma hakkan nutma.

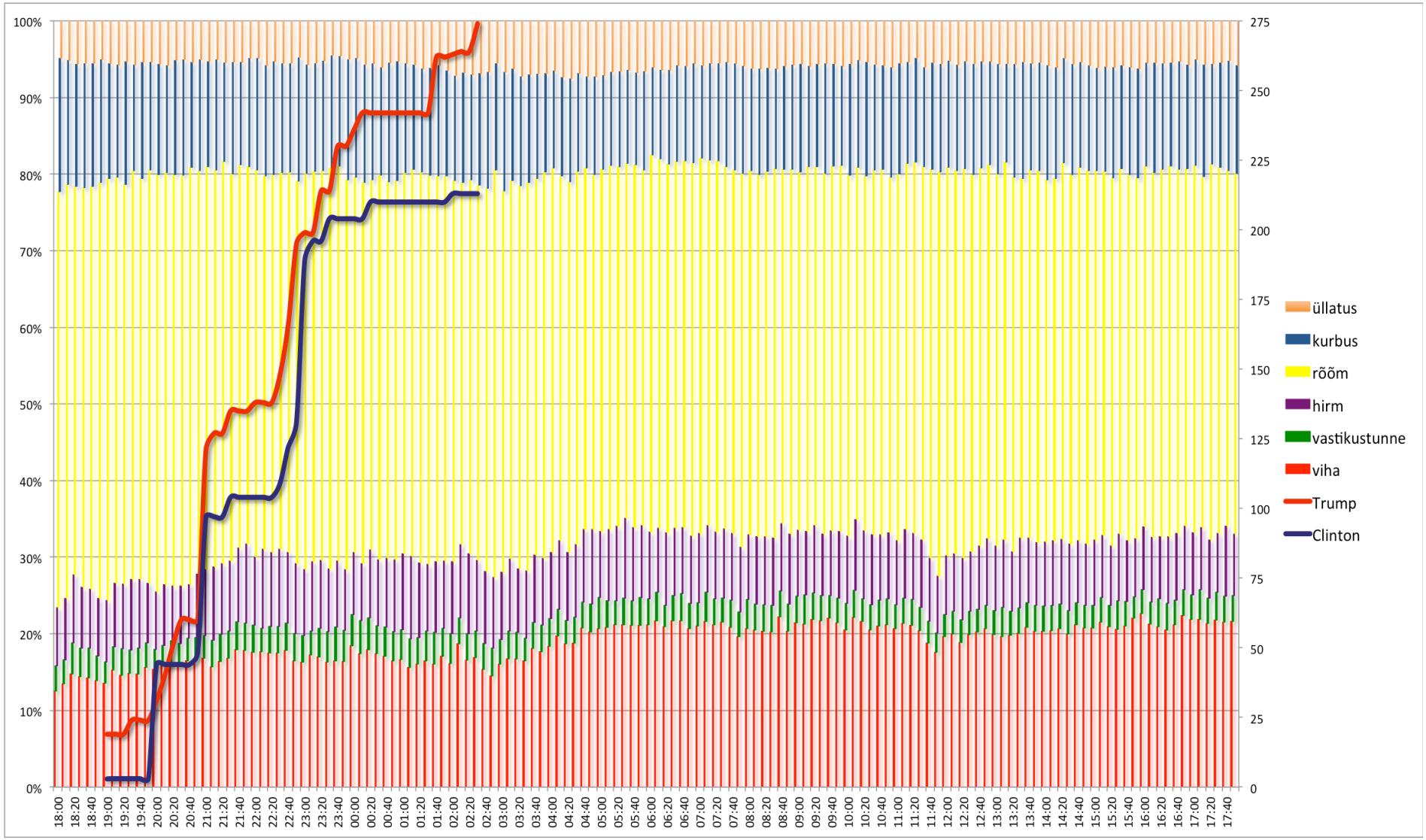
> "*I'm taking January 20 off from social media, people, news channels or anything that will remind me that Donald Trump is our new president*"[25]

See säuts on mõneti sarnane, mõneti erinev võrreldes eelmisega. Emotsioon on küll viidatud, kuid on rohkem peidetud kui eelmisel. Autor ei soovi tunnistada, et Trump sai presidendiks, kuid jääb selgusetuks, kas ta väljendab viha, vastikustunnet, hirmu, kurbust või kõike korraga. Selge on vaid see, et tal tekkis sündmusest ebameeldiv emotsionaalne reaktsioon, mille vältimiseks võtab kasutusele vastavad meetmed.

### 7.4.8 Proovivalimi kokkuvõte

Proovivalimi määratlemisel kattusid kõige enam manuaalsel ja automatiseeritud klassifitseerimisel emotsioonide määratlemine hirmu ja vastikustunde puhul, kõige vähem rõõmu puhul. Oluline on märkida, et paljudel puhkudel tunnistasin, et täpset vastet on keeruline määratleda ka manuaalsel klassifitseerimisel. Iga säuts ei pruugi sisaldada emotsiooni ning mõni säuts võib vastupidiselt sisaldada mitmeid emotsioone. Ühest vastust ei pruugi olla. Täpsemaks määratluseks oleks tarvilik ka sõnastiku mugandada platvormist tingitud vabama keelekasutusega (akronüümide ja levinud väljenditega). Samas tuleb olla ettevaatlik üleliigse sõnade kategoriseerimisega, nagu ilmnes rõõmsate säutsude tulemustest. Rõõm oli kõige ebatäpsem kategooria ning samas kõige suurema sõnade arvuga sõnastikus. Ei saa siiski väita, et see korrelatsioon on tõene, proovivalimis ei üritanud ma selle tõenduspõhisusele kinnitust leida. Isegi kui oleks tekkinud selge korrelatsioon, poleks see tähendanud, et see vastab ka tõele - valim oli pigem eksperimentaalne, et üles tähendada, mida klassifitseerimine endas ette kujutab ning kuidas loogika väljendub. Proovivalimi eesmärgiks oli, et ma ei vaataks andmestiku klassifitseerimist enne ajajoone võrdlust täielikult niiöelda musta kastina.

[25] e.k: Ma võtan 20. jaanuari vabaks sotsiaalmeediast, inimestest, uudistekanalitest ja kõigist, mis tuletab meelde, et Donald Trump on meie uus president




Graafik 1. 24 tunni ajajoon (8. november 18:00 - 9. november 17:59). Graafikul on kujutatud Twitteri emotsionaalse skaala kulgemist (kuuevärviline graafik) ja ametlikult kinnitatud tulemuste arengut (punane ja sinine joon).

# 8. Tulemused

Tulemuste tõlgendamiseks vaatasin, milline oli erinevateks emotsioonideks klassifitseeritud säutsude osakaal erinevatel ajavahemikel. Klassita säutsud jätsin tulemuste tõlgendamise faasist välja. Need moodustasid kogusäutsude arvust umbes 70% ehk emotsionaalseid säutse jäi alles kokku 680 066. Kõige enam klassifitseeriti säutse rõõmsateks – neid oli kokku peaaegu et pool (49,3%). Seejärel tulid järjekorras viha (18,9%), kurbus (14%), hirm (8,7%), üllatus (5.8%) ning viimaks vastikustunne (3,3%). Rõõmsate säutsude amplituud eri ajavahemikes oli 9,2% - kõige väiksem protsent oli 45,9% (17:10-17:19), kõige suurem 55,1% (19:00-19:09). See polnud aga kõige suurem amplituud - vihkamine kõikus eri ajavahemikel 12.6%-i ja 22,6% vahel. Samas kui arvestada amplituudi lähtudes emotsiooni keskmise osakaalu arvust, mitte üldarvudest, kõikus rõõmus emotsioon selgelt kõige vähem, vaid 18,7%. Teised emotsioonid jäid siin kõik vahemikku 42-58%. Kõige enam kõikusid emotsioonidest vastikustunne ning vihkamine. Graafik 1. 24 tunni ajajoon (8. november 18:00 - 9. november 17:59).

| Koondtabel | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| Minimaalne tulemus | 12,55% | 2,44% | 7,08% | 45,85% | 11,36% | 4,50% |
| Maksimaalne tulemus | 22,64% | 4,39% | 10,75% | 55,05% | 17,42% | 7,52% |
| Keskmine tulemus | 18,94% | 3,33% | 8,71% | 49,27% | 13,96% | 5,79% |

Tabel 6. Säutsude protsentuaalne jaotus 24 tunnise perioodi jooksul.

Täpsemaks andmete tõlgendamiseks jaotasin perioodid neljaks osaks:

1. Valimisõhtu algusfaas, kus hakati järk-järgult valimisjaoskondi sulgema ja tulemusi tuli eelkõige üksikutest osariikidest. (18:00-20:59)
2. Valimisõhtu kõige pingelisem faas, kus osariigiti hakkasid tulemused tulema. Pingeline faas sai alguse kell 21:00, kui teatati tulemused Kansasest, Põhja-Dakotast, Lõuna-Dakotast, Texasest, Wyomingist, New Yorgist ja Nebraskast. Trump asus sel hetkel 121-97 juhtima. (21:00-01:59)

3. Valimisõhtu kulminatsioon, kus lõpptulemus oli sisuliselt selgunud ning hakkasid tekkima reaktsioonid selle kohta, et Trump on saanud presidendiks. Ametlikult juhtus see 02:29 kui selgus Wisconsini osariigi tulemused ning maagiline 270 hääle piir sai ületatud. (02:00-04:59)
4. Valimisõhtu järelkaja - alates varahommikust kuni pärastlõunani. Tippsündmusteks Donald Trumpi esimene säuts (kell 6:36), erinevate presidentide õnnitlused, Obama ja Clintoni kõned. (05:00-18:00)

## 8.1 Valimisõhtu algfaas (18:00-20:59)

| Koondtabel | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| Minimaalne tulemus | 12,55% | 2,51% | 7,08% | 50,66% | 13,64% | 4,89% |
| Maksimaalne tulemus | 16,77% | 4,07% | 9,26% | 55,05% | 17,42% | 5,82% |
| Keskmine tulemus | 14,95% | 3,12% | 8,02% | 53,34% | 15,17% | 5,40% |

Tabel 7. Säutsude protsentuaalne jaotus 18:00-20:59.

Valitud sündmused:
18:12 - Tehnilised probleemid Durhamis, Põhja-Carolinas hääletamisega.
18:59 - Trump võidab ametlikult esimesed osariigid: Indiana ja Kentucky. Clinton võidab Vermonti.
19:11 - Azusas, Californias toimub hääletusjaoskonna lähedal tulistamine, esialgsetel andmetel 2 inimest surnud
20:00-20:59 - Järk-järgult hakkavad osariigiti tulemused tulema, pärast Illinoisi tulemuste selgumist läheb Clinton juhtima 68-66.

See oli neljast perioodist kõige rõõmsam - tippaeg saabus esimese tunni lõpus. Kui välja arvata üks vahemik (18:20 kuni 18:29), siis püsis rõõmusõnumite protsentuaalne arvukus stabiilselt üle 52%-i, keskmisest ligi 3% kõrgemal. Seevastu oli viha võrreldes kolme ülejäänud perioodiga keskmiselt tunduvalt madalam - vihaste säutsude arvukuse lagi jäi sel perioodil 2% võrra alla üldisele keskmisele. Siiski võib märgata, et vihaste säutsude osakaal kasvas järk-järgult. Esimese pooleteise tunni (18:00-19:29) keskmine oli vihaste säutsude puhul 14,11%, teise pooleteise tunni jooksul oli see juba 15,78%. Toimus vaikne, aga stabiilne kasvamine. Perioodi

esimesel ajavahemikul oli osakaal madalaim (12,55%), kõrgeim viimasel (16,77%). Kui vaadata, et mis antud perioodil toimus, siis algul polnud tulemusi veel üldiselt teada, kuid perioodi lõpuks tekkis juba aimdus, mida esimesed tulemused võivad viidata. Lõpus oli seis 68-66 Clintoni kasuks, kuid see näitab osariikide kinnitatud võitmisi - esialgsete tulemuste alusel võis juba märgata tendentse, mis muudes (aga mitte kõikides) osariikides tulemused näitavad. Tulemused tulid maakonniti ning juba siis võis näiteks näha, et Trump oli teinud üle ootuste hea tulemuse.

Vastikustunnet väljendavate säutsude tipphetk saabus ajavahemikus 18:20 - 18:29. Selget sündmust esialgu sellega seostada ei suutnud ning seetõttu tutvusin ka andmestikuga, mida inimesed säutsusid. Umbes 80% säutsudest sisaldasid ühte kolmest sõnast: Trump, Clinton või valimised (*election*). Sõnumeid lugedes paistis vastasseis põhinevat mitte mingil konkreetsel sündmusel või sõnavõtul, vaid üleüldises pettumuses ning õõvastuses, mida valimised olid tekitanud - näiteks kirjeldati palju, et valimaskäik tekitas iiveldust ning rüvetuse tunnet, kuna poldud rahul, mida just tehti (ei meeldinud ei valimiste kulgemine ega kandidaadid).

Hirm tõusis järsult vahemikus 19:20 - 19:39, misjärel langes koheselt samamoodi. Võis oletada, et põhjustajaks oli Azusa tulistamiste mingisugune järelkaja, kuid kinnitust väitele ei suutnud tuvastada. Erinevad märksõnad, mida selle kohta üritasin leida ei andnud märkimisväärset arvu vasteid. Küll aga oli andmestikuga tutvudes võimalik sedastada, et säutsudes, mis oli määratletud hirmu sisaldavateks, vastas ka sisu jaotusele (ilmselgelt küll mitte kõikidel puhkudel). Inimesed väljendasidki säutsudes hirmu, peamiselt esialgsete tulemuste üle. Sellele järgnenud järsku langust ei oska täpsemalt seletada.

Kurbuse väljendamiseks oli antud periood tipp-aeg – madalaim punkt oli peaaegu et võrdne üldise keskmisega. Eriti ilmnes see esimese tunni jooksul, mil kurbust sisaldavad säutsud moodustasid keskmiselt 16,33% emotsionaalsetest säutsudest, järgneva kahe tunni jooksul langes see keskmiselt 14,59%-ni. Järgneva 21 tunni jooksul ei jõudnud kurbust sisaldavate säutsude arvukus kordagi samale tasemel, kus see oli esimese tunni jooksul.

## 8.2 Lõplike tulemuste selgumise faas (21:00-01:59)

| Koondtabel | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| Minimaalne tulemus | 15,61% | 2,85% | 7,50% | 48,29% | 12,83% | 4,50% |
| Maksimaalne tulemus | 18,42% | 4,39% | 10,75% | 52,54% | 16,14% | 6,48% |
| Keskmine tulemus | 16,97% | 3,65% | 9,16% | 50,34% | 14,44% | 5,44% |

Tabel 8. Säutsude protsentuaalne jaotus 21:00-01:59.

Valitud sündmused:

21:00 - Kinnitust saavad tulemused seitsmest osariigist, mis viib Donald Trumpi juba ette 121-97.

21:14 - Ennustuskontorid ning ennustusmudelid hakkavad järsult korrigeerima oma ennustusi – Clinton langes erinevates kohtades 81-84%-lt 67-70%-ni.

21:26 - New York Times kuulutab, et esmakordselt on Donald Trumpist saanud nende ennustusmudeli järgi favoriidiks saamaks Ameerika Ühendriikide presidendiks.

22:36 - Ametlikuks saab esimese võtmeosariigi tulemused: Donald Trump võitis Ohio (18 valijameeste häält).

22:40 ja 22:43 - Clinton võtab teise ja kolmanda võtmeosariigi - Virginia (13) ja Colorado (9).

22:50 - Donald Trump saab endale suurima võtmeosariigi 29 häälega - Florida. Florida võitja on alates 1980. aastast suutnud võita kõikidel kordadel, välja arvatud ühel juhtumil (1992).

23:11 - Trump võidab samuti võtmeosariigi Põhja-Carolina (15 häält).

00:22 - Paralleelselt presidendivalimistega valitakse osaliselt Kongressi liikmeid - vabariiklased kindlustavad enamuse Esindajatekojas.

01:20 - Prantsuse erakonna Front National juht Marine Le Pen õnnitleb Donald Trumpi kui uut Ameerika Ühendriikide presidenti.

01:35 - Trump võidab võtmeosariigi Pennsylvania, mis on viimased kuus presidendivalimist kuulunud demokraatidele. Pennsylvania on suuruselt teine võtmeosariik 20 häälega.

Nagu sündmuste nimekirjast näha, on see periood, kus toimub tulemuste lõplik pöördumine Donald Trumpi kasuks. Kui esialgsete andmete põhjal pöördusid esmalt

ennustuskontorid järsult võitjaks ennustama Trumpi, siis hiljem kinnitasid seda võidud võtmeosariikidest. Twitteris hakkasid üllatunud emotsioonid väljenduma alates 01:20-st. Eelnenud perioodil ehk 21:00 kuni 01:19 oli keskmine üllatusi sisaldavate säutsude arv 5,32%, järgnenud 40 minuti jooksul 6,20%. Absoluutarvudes ei tundu muutus suur, kuid suhtarvuna oli tõus ligi viiendik. Selleks hetkeks oli Trumpi võit muutunud juba peaaegu et kindlaks – olulisemad võtmeosariigid olid võidetud või oli ta neis kohe ametlikult võitmas. Pennsylvania oli olnud viimased aastakümned demokraatide käes, kuid Trump suutis üllatuslikult ka selle endale haarata. Põhjust üllatuse väljendamiseks iseenesest oli.

Rõõmu sisaldavate säutsude arvukus taandus sel perioodil 24 tunni keskmise lähedale. Viha emotsioon näitas tõusumärke, kuid selle perioodi lagi jäi veel alla üldisele keskmisele. Vihastamise lained käisid perioditi: kõrgem aktiivsus oli näiteks vahemikus 21:40 kuni 22:49 (17,73%) ning 00:00 kuni 00:49 (17,67%). Esimese ajavahemiku algus sobitub enam-vähem sellega, et Trumpist oli saanud järsku favoriit. Absoluutsel tippajal ehk 00:00 kuni 00:09 moodustasid vihased säutsud 18,42% emotsionaalsetest säutsudest. Nendest omakorda sisaldasid 76% säutsudest nime Trump. Võrdluseks sõnu "Hillary" või "Clinton" sisaldasid kokku vaid 33% säutsudest. See viitab eelkõige sellele, et viha väljendamine on seotud rohkem Trumpiga (küll aga ei saa selle põhjal otseselt väita, et viha on temale suunatud).

Periood eristus teistest veel vastikustunde suurema proportsionaalsusega (keskmine 3,65%). Kõrgeim periood valitses 23:30 kuni 00:29, mil vastikustunde keskmine näitaja ületas nelja protsendipunkti. Tippajal ehk 23:40 kuni 23:49 sisaldasid koguni 87% vastikustunde säutsudest vähemalt ühte kolmest sõnast - Trump, *sick* (haige) või *disgust(ing)* (vastikustunne-tülgastus). Eelnevalt viidatud vihalainele eelnes seega vastikustunde väljendus.

Sel perioodil hakkas ka tugevnema hirmutunne. Võrreldes eelneva perioodi keskmisega kasvas hirmu osakaal rohkem kui protsendipunkti võrra. Sarnaselt viha emotsiooniga käis see lainetes. Esimene hirmutunde laine algas 21:00 ning kestis kuni 23:29-ni (9,41% emotsionaalsetest säutsudest). Seejärel vaibus mõneks ajaks (23:30 kuni 00:39 8,29%). Teise vihalaine lõpus aktiviseerus hirmutunne jällegi - perioodi lõpuni ulatus see ligi 9,5% juurde (9,46%). Hirmutunne saavutas oma tipp-punkti

01:10. Kui vaadata, mille arvelt see kerkis, siis võrreldes selle ajavahemiku keskmistega olid nii viha (1,6%) kui ka kurbus (0,77%) madalamad.

## 8.3 Trumpi võidujoovastus ja vastureaktsioonid (02:00-04:59)

| Koondtabel | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| Minimaalne tulemus | 14,50% | 2,90% | 8,35% | 46,32% | 11,96% | 5,53% |
| Maksimaalne tulemus | 20,79% | 3,78% | 10,39% | 53,23% | 15,52% | 7,52% |
| Keskmine tulemus | 17,61% | 3,39% | 9,22% | 49,19% | 13,71% | 6,89% |

Tabel 9. Säutsude protsentuaalne jaotus 02:00-04:59.

Valitud sündmused:

02:02 - Hillary Clintoni kampaaniajuht John Podesta adresseerib Clintoni kampaaniapeole tulnud inimesi, et Clinton ei võta täna õhtul enam sõna ning tasub koju ära minna.

02:29 - Donald Trump võidab Wisconsi osariigis, mis aitab teda üle vajaliku 270 hääle lävendi. Trumpist on saanud valimisõhtu võitja.

02:45 - Tulevane asepresident Mike Pence peab võidukõnet.

02:49 - Tulevane president alustab oma võidukõnet.

03:06 - Donald Trump lõpetab võidukõne.

03:49 - Uudistes raporteeritakse protestidest Los Angelesis ja Oaklandis.

04:30 - Lisaks raporteeritakse protestidest Portlandis.

Sel perioodil hakkasid emotsioonid liikuma oma amplituutide äärmustesse: mis üles, mis alla. Suurimate anomaaliate esinemiskohad satuvad Trumpi kõneaegsesse ajavahemikkudesse. Ajavahemikus 02:50 kuni 02:59 jätkas viha emotsioon langemist ning sattus madalaimasse punkti. Nii väike vihaste säutse proportsionaalsus (14,50%) oli viimati 19:00 ehk valimisõhtu algfaasis. Vastupidiselt tõusis rõõmu sisaldavate säutsude arvukus – need saavutasid samas ajavahemikus tipu, mida polnud saavutatud alates kella 20:40-st (53,23%). Võrdluseks moodustasid rõõmsate säutsude proportsionaalsus kõrvalolevates ajavahemikes vastavalt 3,26% (02:40 - 02:49) ja 3,49% (03:00 - 03:09) vähem. Perioodi lõpus näitas samas rõõmusäutsude arvukus

märgatavat langustrendi – 04:10 kuni 04:59 oli keskmiseks arvuks 47,61% ehk tervenisti 5,62% vähem kui viimasel tippajal.

Nii vahetult enne Trumpi kõnet (02:40 - 02:49) kui ka kõne lõpu osas (03:00 - 03:09) tegi samuti hüppe üles kurbus (vastavalt 15,19% ja 15,52%). Hiljem, alates 03:40-st hakkas see aegamisi vaibuma – keskmine näitaja oli 12,87% ehk tunduvalt vähem kui üldine keskmine. Lisaks avaldas Trumpi kõne mõju üllatusele – 02:50 kuni 02:59 tekkis selles osas sisse väike anomaalia, kus üllatust sisaldavate säutsude arvukus järsult langes korraks (5,53% peale). Ümbritsevates ajavahemikes oli osakaal 1,15% kõrgem ehk mõlemas 6,68%-i. Nii enne kui ka pärast seda ajavahemikku püsis üllatust sisaldavate säutsude arvukus jätkuvalt kõrgena. Ilma anomaaliata (ehk ajavahemik 02:50 - 02:59) oli selle perioodi keskmine 6,97%. Üllatus oli seega sel perioodil stabiilselt kõrge.

Kui viha langes Trumpi kõne aegsel perioodil ning ka enamasti seda ümbritsevadel ajavahemikel, siis taaskord tõusis see esile ühe seigana kõigepealt 04:10 - 04:19 (19,76%) ning seejärel hiljem 04:40 - 04:59 (keskmine 20,51%). Võttes viimase perioodi näiteks, sisaldasid 87% säutsudest sõna Trump – viide sellele, et tema on ikkagi kõige keskmes. Siin ei saa küll väita, et kõik see viha oli tema poole suunatud, kuna võidi ka vastupidiselt vihaselt parastada Clintoni pooldajate suunas, kasutades tekstis Trumpi nime. Pinnapealsed otsingutulemused küll sellele ei viita: Clintonit mainiti vaid 17,9% säutsudes ning ligi pooled neist polnud säutsudes Trumpiga koos (ehk selle 13% sees, kus Trumpi polnud).

## 8.4 Valimisõhtu järelkaja (05:00-15:59)

| Koondtabel | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| Minimaalne tulemus | 17,63% | 2,44% | 7,51% | 45,85% | 11,36% | 4,87% |
| Maksimaalne tulemus | 22,64% | 3,98% | 10,45% | 52,67% | 15,24% | 7,11% |
| Keskmine tulemus | 20,92% | 3,25% | 8,57% | 47,94% | 13,55% | 5,77% |

Tabel 10. Säutsude protsentuaalne jaotus 05:00-17:59.

Valitud sündmused:
06:36 - Donald Trump säutsub esimest korda valimiste võitjana: *Such a beautiful and important evening! The forgotten man and woman will never be forgotten again. We will all come together as never before.* (Nii ilus ja tähtis õhtu! Unustatud mees ja naine ei ole kunagi enam unustatud. Me kõik tuleme taas kokku nagu ei kunagi varem).
07:12 - Valge Maja teatab, et president Obama on õnnitlenud Donald Trumpi ning on kutsunud teda Obamaga kohtumisele neljapäevaks (ehk siis järgnevaks päevaks).
10:20 President Barack Obama pöördub avalikkuse poole kõnega.
11:40 Hillary Clinton peab oma toetajate ees kõne, kus tunnistab ennast valimiste kaotajaks.

See oli vaadeldavatest perioodidest kõige pikem, aga samas ka kõige suurema sündmuste hajutusega – Ameerika Ühendriikidesse oli juba ammu kätte jõudnud öö, tulemused olid selgunud ning lõpuks enam suuremaid ootamatuid sündmusi kohe ees oodata polnud, vähemalt mitte enne hommikut. Periood eristub teistest selles osas, et peaaegu et kõikide emotsioonide proportsionaalsus on sarnane üldise keskmisega. Ainult vihased säutsud püsisid väiksemate kõikumistega terve selle perioodi kõrgemal kohal ning rõõmsate säutsude suhtarv oli madalamal kohal. Vastandlik hetk saabub lõunasel ajal kell 11:50 - 11:59, mis tähistab ühtepidi rõõmsate säutsude tipphetke kui ka vihaste säutsude selget madalhetke. Kui vahemikus 05:00 kuni 11:20 oli viha sisaldavate säutsude osakaal 21,13% (võrdluseks üldine keskmine oli 18,94%), siis järsk langus toimus just eelmainitud ajavahemikus. See langes koguni 17,63%-ni. See tingis ka vahetult eelnevas perioodis (18,85%) kui ka mõneks ajaks hiljem (12:00 kuni 13:29) väiksema languse võrreldes perioodi keskmisega, kus vihaste säutsude arvukus oli 19,88%. Hiljem tõusis vihaste säutsude arvukus jälle ülesse. Vahemikus 13:29 kuni 17:59 moodustasid nad 21,15% emotsionaalsetest säutsudest.

Samas ajavahemikus ehk 11:50 kuni 11:59 toimus ka vastikustunnet sisaldavate säutsude kõige madalaim hetk kogu 24 tunni jooksul (2,79%). Vastupidiselt vihale ja vastikustundele tõusis sel ajavahemikul aga rõõmsate säutsude arv korraks anomaaliana kõrgemale (52,67%-ni). Võrdluseks toon antud perioodi üldise keskmise protsendi - 47,94%. See tähendab, et sel hetkel oli rõõmsate säutsude osakaal 4,73% kõrgemal kui selle perioodi keskmisel. Arvestades rõõmu emotsiooni üldist

proportsionaalselt vähest kõikumist, on see selge eristumine. Need kõik võiksid viidata, et sel ajavahemikul võis midagi toimuda – nii see ka oli. Kell 11:40 alustas ning 11:53 lõpetas Hillary Clinton oma valimistejärgset alistumiskõnet. Kõnel oli antud andmestikule mõju olemas.

Sellele perioodile oli veel iseloomulik üllatust sisaldavate säutsude järk-järguline langus. See ei toimunud küll lineaarselt, vaid käis pidevalt üles-alla (amplituud polnud küll väga suur). Perioodi alguses (05:00 kuni 06:29) püsis osakaal veel sarnaselt eelmise perioodiga võrdlemisi kõrgel (keskmiselt 6,63%.) Sarnaselt liikus hirmu sisaldavate säutsude osakaal: enne Clintoni kõnet (05:00 kuni 11:39) oli keskmine 8,88%, pärast seda (12:00 kuni 17:59) pidevalt üles-alla liikudes oli keskmiseks 8,27% (siin oli amplituud suurem). Tippaeg jäi päris algusesse ehk 05:10 kuni 05:59 (9,70%).

## 8.5 Tulemuste kokkuvõte

Tulemuste alusel üritan kokkuvõtvalt vastata püstitatud uurimisküsimustele:

1. Kuidas muutus Twitteri diskussiooni kollektiivne emotsionaalne skaala 24 tunni jooksul?
2. Kuidas muutus emotsionaalne skaala võrreldes valimisõhtu kulgemisega?

Vaadeldava 24 tunnise perioodi algul ehk valimisõhtu algfaasis olid emotsioonide seast rõõmsate ja kurbade säutsude jaoks tipp-perioodid. Hiljem tõusid need näitajad harvematel juhtudel ülespoole. Rõõmu sisaldavate säutsude jaoks oli selleks mõlema suurema kandidaadi kõned – neil ajahetkedel tõusid rõõmsate säutsude osakaal võrreldes ümbritseva keskkonnaga märgatavalt suuremaks. Trumpi kõne puhul taandus ka selgelt üllatus ning viha. Clintoni puhul taandus samuti viha ning vastikustunne saavutas oma terve vaadeldava perioodi madalaima punkti.

Valimas käimise järel avaldus korraks ka suurem vastikustunne - inimesed väljendasid mõrudat maiku üldisemalt valimiste kulgemise ja valimaskäigu kohta.

Teine suurem hüpe vastikustundes toimus ajal kui hakkasid järjest olulisematest osariikidest ilmuma tulemused.

Kui algul ei väljendanud säutsujad ülemäära palju viha, siis järk-järgult hakkas vihaste säutsude osakaal tõusma. Vihahood käisid lainetena, täpselt samamoodi ka hirmuhood. Vihased säutsud said suurema hoo sisse alates varahommikust, mis kestis vaadeldava perioodi lõpuni välja.

Üllatus kerkis säutsujate seas rohkem esile pärast kesköö – umbes samaaegselt kui põhimõtteliselt oli saanud selgeks, et Trumpist on saanud uus Ameerika Ühendriikide president. Üllatuse vaibumine kestis pika perioodi jooksul kuni vaadeldava aja lõpuni. Samamoodi vaibusid ka vaikselt hirmuhood.

Tulemustes võrdlesin emotsioonide puhul mitte üldist osakaalu, vaid nende emotsioonide kõikumist võrreldes enda raamistikuga. Kõikide emotsioonide tavapärane tasakaalukus pole üldiselt võrdne, kuid eeldan, et selles mängib ka veel üksjagu rolli kasutatud andmeanalüüsimudel. Esiteks see, et tulemuste tõlgendamises ei ole võimalik täielikku pilti hinnata, kuna kasutusest jäi välja umbes 70% saadaval olnud säutsudest. Lisaks on teadmata hindamisprotsessi niiöelda normaalne alajaotus – mitu protsenti tavaliselt Twitteri säutsudest neid emotsioone sisaldavad, võrreldes valimisõhtuga. Tulemuste põhjal julgen näiteks väita, et terve õhtu jooksul polnud rõõmsate säutsude arvukus emotsionaalsetest säutsudest keskmiselt ligi 50%. See näitab pigem mudeli olemust, et see on rõõmsate säutsude poole kaldu. Kui mõelda vastandlikult, siis negatiivseid emotsioone on samas rohkem - viha, vastikustunne, hirm ja ka tinglikult kurbus. Rõõm haarab samas seetõttu peaaegu et kogu positiivsema skaalapoole. Võrdlusena oleks siin võimalik võrrelda, kuidas oleks see mudel rakendunud kasutades polaarsuse klassifitseerimist.

Tulemused ei saa iseenesest pretendeerida tõesusele, nagu ka Grimmer ja Stewart (2013: 269-271) viitasid. See aga ei tähenda, et tulemustest ei ole võimalik omandada informatsiooni. Rõhutades veelkord, et informatsiooni, aga mitte fakte. Tegemist polnud mitte Twitteri maailmas toimunud emotsionaalse skaala arenguga Ameerika Ühendriikide presidendivalimiste õhtul ning järgnenud päeval, vaid tegemist oli Twitteri maailmas toimunud emotsionaalse skaala arenguga lähtudes antud töös

kasutatud mudelist. See polnud selgelt peegeldunud reflektsioon, vaid läbi tööriistade loodud tõlgendus seal toimuvast.

# 9. Diskussioon

Käesolevas töös olen analüüsinud, kuidas on võimalik kasutada suurandmeid sotiaalmeedia analüüsiks. Täpsemalt olen peaasjalikult keskendunud Twitteril. Sotsiaalmeedia on aga kõigest üks valdkond, suurandmete kasutamisvõimalused meediaanalüüsis ei piirdu vaid sellist tüüpi andmestikul. Nii näiteks on võimalik potentsiaalselt koguda kokku massiivsetes kogustes uudiste artikleid ja analüüsida neid vastavalt vajadustele. Twitteri säutsude valimist analüüsiks soosis nende võrdlemisi lihtne kättesaadavus – ühe tööriistaga oli võimalik vastavaid märksõnu kasutades kokku koguda teemablokk säutsudest. Artiklite massiivseks kogumiseks võib samas ette tulla takistusi nii maksumüüri taha jäämisel kui ka veebilehtede eri struktuuridega kohanemisega. See on juba omaette uurimisteema.

Suurandmete analüüsimisel tasub täpsemaks tulemuseks võimalusel kombineerida eri metoodikaid – olgu selleks vastanduseks siis manuaalsed ja automaatsed meetodid või erinevad automaatsed meetodid. Nagu ka Guo et al (2016: 15-21) viitasid, võib ühes tööfaasis olla üks meetod parem (nt juhendamata masinõpe, et tuvastada diskussiooniteemasid), teises teine (nt juhendatud masinõpe, et kaardistada diskussiooniteemasid). Antud töös ei kombineerinud ma eri meetodeid, et üks mõjutanuks teist, kuid andmeanalüüsi tulemuste paremaks tõlgendamiseks võtsin siiski appi ka manuaalse klassifitseerimise. Manuaalne klassifitseerimine aitas paremini mõista antud mudeli olemust, kuid kindlasti mitte täielikult – ei saa väita, et antud valim oli juhuslik ega representatiivne.

Potentsiaalselt oleks olnud võimalik parandada olemasolevat mudelit lähtudes manuaalsest klassifitseerimist, kus oleksin muutnud näiteks sõnastikku täpsemaks (lisanud kasutatavaid slänge, spetsiifilist sõnavara jne.). See aga jääb minu lingvistilisest pädevusruumist välja ning seetõttu otsustasin kasutada valmis metoodikat. Samas uuringutes, kus töö keskmeks ongi empiiriline osa, tasub kindlasti arvestada sõnastike kasutamisel uuritava platvormi spetsiifilisi sõnakasutusi.

Samuti oleks andnud võrrelda emotsionaalset skaalat polaarsuse skaalaga (kas säutsud olid positiivsed, negatiivsed või neutraalsed), mis oleks andnud täiendava pilgu antud

mudeli olemusele, sest näiteks emotsionaalsel skaalal oli positiivseid emotsioone justkui kõigest üks (rõõm), negatiivseid neli (hirm, kurbus, vastikustunne, viha) ja üks emotsioon, mis võib olla nii positiivne kui negatiivne (üllatus). Samuti tuvastasin, et tasub üle mõelda, mida ikkagi hinnatakse – kas tekstis sisalduvat emotsiooni või tekstis väljendavat emotsiooni – ning kuidas peaks mudel hindama, kui on konkureerivad emotsioonid (kas jääda klassita säutsuks, liigitada mõlema alla või võtta arvesse sõnade konnotatsiooni tugevust).

Uuringu disainimise faasis tuleb kalkuleerida, kas lähtuda oma uurimisküsimustes sellest, mis metoodikad, tööriistad ja andmestikud on kättesaadavad, või kujundada välja ise vastav meetod, mis sobiks just nimel täpselt oma uurimisküsimustele. Viimane variant vajab kindlasti laiemapõhjalist koostööd sotsiaalteadlaste ja arvutiteadlaste vahel, et tulemused ka vastaksid uuringu püstitusele. Kui on võimalik, siis eelistatum uuringu disainimise variant peaks olema viimane, kuna metoodika lähtub uurimispüstitusest, mitte vastupidi, kuid sellisel juhul peab eriti suur rõhk olema andmete ja metoodika valiidsuse kontrollil.

Antud uurimistöös lähtusin esimesest ehk arvestasin, mis oleks minu võimekuses teostatav ning kättesaadav. Kättesaadavuse probleem võib esile kerkida eriti andmestiku osas – suurandmete mõõtu sobivad andmestikud ei pruugi olla kättesaadavad kas siis tehnoloogilistel põhjustel (andmete kogumine hõlmab endas liiga eristaolist lähteandmestiku, keerulised rakendusliidesed) või siis andmestike omanduse põhjustel. Näiteks ei võimalda Twitter terviklike andmebaaside loomet (kuni 1% hetkel säutsutavast materjalist, välja arvatud erandid), samuti artiklite suuremamahuliseks kogumiseks on vaja kokkulepete abil saada mööda ajakirjandusväljaannete maksumüüridest. Seetõttu võib juba suurandmete kättesaamine olla iseenesest inimressursikulukas kui ka finantsilisest võimekusest lähtudes ületamatu.

Antud uurimistöös jäi kasutamata üsnagi rikkalik meta-andmestik, mis iga säutsu kohta hõlmas endast 44 välja. Osad neist olid kasutud-korduvad (näiteks kellaaeg oli kahes eri vormingus, säutsu ID väli kordus), osad neist võimaldavad püstitada keerukamat andmeanalüüsi - näiteks tuua võrdluspunkt verifitseeritud kasutajate (väga väike osa kasutajadest, Twitteri poolt kinnitatud kontod, mis kuuluvad tuntud

isikutele, organisatsioonidele jne.) ja verifitseerimata kasutajate seast. Sama oleks saanud teha vaadates eri teemaviiteid, eri ajatsoonides paiknevaid kasutajaid, eri postituste arvuga kasutajaid, eri jälgijate arvuga kasutajaid jne. Antud analüüs põhines peamiselt ühel väljal: säutsude tekstil. Andmete puhastamiseks kasutasin küll ka muid välju (nt kas säuts sisaldas URLe, mis tüüpi säuts oli).

Kui Grimmer ja Stewart (2013: 269-271) viitasid oma printsiipidega, mida pidada silmas automatiseeritud tekstianalüüsis, siis võib väita, et need kehtivad ka üldisemalt suurandmete analüüsis, hoolimata meetodist. Esimesele neist (kõik tulemused on olemuselt valed, kuid võivad siiski sisaldada kasulikku informatsiooni) võib küll oponeerida, kui analüüsida numbreid, mitte teksti. Sel juhul on potentsiaalselt võimalik, et tulemused võivad ka andmeanalüüsis pretendeerida tõele. Siiski selleks tuleb lähtuda esmalt neljandast printsiibist ehk andmete ja metoodikate valideerimine (õigsuse kontrollimine) on suurandmete analüüsimises eriti vajalik. Samas ei erine see iseenesest sellest, kui uuritakse väiksemat andmestikku – ikkagi tuleb õigsust kontrollida.

Ma leian hoolimata töös esitatud kriitikast, et suurandmete kasutamine sotsiaalmeedia analüüsis on paljulubav uurimisallikas. See on üks täiendav vahend ühiskonna mõtestamiseks ning sellest aru saamiseks, sotsiaalsete praktikate jälgimiseks massiivses mõõdustikus. Peamiseks uurimisallikaks võib pidada Twitterit. Facebook on hoolimata oma populaarsusest teadlaste jaoks oma rakendusliidese tõttu võrdlemisi keeruliselt kättesaadav. See aga ahendab Eesti keskseid uurimistöid kahel põhjusel. Twitter pole Eestis nii populaarne kui Facebook ning Twitteris on keeruline filtreerida seda, mis just Eesti Twitterisfääris toimub. Lisaks julgen väita, et Eesti põhiselt ei saaks nimetada Twitterist toimuvat suurandmete vääriliseks.

Twitteri andmete analüüsimise metoodikad on küll paljudel juhtudel eksperimentaalsed, kuid aja jooksul on samas välja arenenud kaks suuremat suunda, mille alused peituvad traditsioonilistes uuringutes: kontentanalüüs ja võrgustikeanalüüs. Isegi kui suurandmete analüüsis tekivad probleemsete punktidena esile küsimused andmete ja andmeanalüüsi valiidsuses, võimaldab siiski andmeanalüüs potentsiaalselt uudset informatsiooni – näiteks luues hüpoteese ning sellega olles aluseks väiksemamõõtmelistele uuringutele.

# 10. Kokkuvõte

Käesoleva magistritöö eesmärk oli analüüsida metodoloogilisi võimalusi ja piiranguid suurandmete kasutamisel sotsiaalmeedia analüüsis. Töö empiirilises osas võrdlesin, kuidas kulges paralleelselt Ameerika Ühendriikide presidendivalimistel toimuvaga Twitteri säutsude kollektiivne emotsionaalne skaala. Selleks kogusin andmeid kasutades tööriista DMI-TCAT ning saadud andmeid klassifitseerisin kasutades meelestatuse analüüsi, mis jaotas säutsud kuueks eri emotsiooniks (viha, rõõm, vastikustunne, hirm, kurbus ja üllatus).

Selleks üritasin esmalt mõista, mis suurandmed üldse on – ühte kokkulepitut universaalset definitsiooni pole, kuid eri definitsioonid on tihti sarnased. Kokkuvõtvalt võib öelda, et suurandmed on andmed, mille kogumiseks ja analüüsimiseks on vaja kasutada vastavaid tehnoloogilisi vahendeid, mis tavapäraste metoodikatega pole võimalik ehk manuaalselt andmete kogumine ja analüüsimine on kas raskendatud või peaaegu et võimatu. Suurandmeid iseloomustab ka samas teatud mütoloogia - massiivses koguses andmeid justkui peaksid sisaldama endas mustreid ja informatsiooni, millest väiksemas koguses andmetega ei ole võimalik aru saada või märgata.

Sotsiaalmeedia suurandmete uurimises on peamiselt kaks suunda: kontentanalüüs ning võrgustikeanalüüs. Esimene neist hindab kas postituste, profiilide vms. sisu ühe või teise metoodika alusel, teine neist üritab mõista sotsiaalmeedias toimivate suhete olemusi – nii artikuleerituid suhete ehk märgistatud suhete (nt Facebookis sõbraks olemise näol või Twitteris kedagi jälgides) kui ka käitumuslike suhete ehk pidades silmas interaktsioone (olgu selleks kommentaar, *like*, edasisäuts vms).

Twitteri uurimiseks kasutasin antud uurimustöös peamiselt selle kommunikatsiooni makrotasandit - teemaviiteid. Teemaviited on märgised, mis koguvad kokku konkreetsed diskussioonivood. Kasutajad märgistavad oma postitused teemaviidetega ning seeläbi sihilikult saavad osa üldisemast diskussioonist. Teemaviidete tasandit võib vaadata kui sfääri, mis kuulub Twitteris kõige enam avalikkuse alla - postitused on avalikud ning kasutajad soovivad kasutades teemaviiteid jõuda laiema

auditooriumini. Iseenesest on Twitter segu avalikust- ja privaatsfäärist. Privaatsfääri moodustab peamiselt üheti privaatselt sätestatud kasutajad kui ka üksteisele otse privaatselt saadetud sõnumid. Twitteris tavapäraselt säutsudes (avalik kasutaja, kasutamata teemaviiteid) moodustub auditoorium segu kasutajale teadaolevast auditooriumist ehk sihtauditooriumist (säutsud jõuavad nende ajajoontele, kes sind jälgivad) kui ka teadmata auditooriumist (kui keegi su postitust edasisäutsub või vajutab meeldimise sümbolit, siis see säuts jõuab kaugemale kui sinu sihtauditoorium).

Säutsude analüüsimiseks ja tõlgendamiseks kasutasin meelestatuse analüüsi, mis jaotas säutsud tekstide alusel seitsmeks kategooriaks: kuus eri emotsiooni (viha, rõõm, vastikustunne, hirm, kurbus ja üllatus) ning klassifitseerimata säutsud (algoritmid kas ei suutnud tuvastada säutsude emotsiooni või ei suutnud otsustada, millise konkureeriva emotsiooni alla peaks säutsu liigitama). Klassifitseerimisel paistis, et raskusteks osutusid just nimelt see, kui tekst sisaldab mitut emotsiooni, siis mille alla liigitada, kuid ka näiteks see, millel põhineb sõnastik - Twitter on selline keskkond, kus kasutatakse nii slängi, emotikone jne. Samuti on tekstide automaatsel analüüsil keeruline tuvastada sarkasmi ja irooniat.

Töö empiirilises osas koostasin esmalt ajajoone reaalselt toimunud sündmustest, mis juhtus vaadeldava 24 tunni jooksul ning seejärel võrdlesin seda säutsude emotsionaalse skaala arenguga. Analüüsis vaatlesin emotsioonide proportsionaalsuse kõikumist, mitte üldist osakaalu. Seda põhjusel, et mudelist tingitult oli teatud proportsionaalsus juba eelsätestatud - näiteks kõige enam sõnu oli seotud rõõmsa emotsiooniga, mis kajastus ka tulemustes, kus keskmiselt ligi pooled säutsud klassifitseeriti rõõmsateks. Antud mudeli alusel eristusid selgelt sündmustena mõlema peamise kandidaadi lõpukõned (Trumpi võidukõne ning Clintoni loobumiskõne), kus rõõmsate säutsude osakaal tõusis, vihaste säutsude osakaal langes. Clintoni puhul taandus samuti vastikustunnet väljendavate säutsude osakaal ning Trumpi puhul üllatavate säutsude osakaal.

Üldisemalt käisid vihatunnete ja hirmutunnete tõusud ja langused hooti. Viha moodustas valimisõhtu algul tunduvalt väiksema osakaalu kui ta seda tegi järgmise päeva hommikul, kuid see tõus polnud täielikult lineaarne. Vastupidiselt langes

rõõmsate säutsude osakaal järk-järgult, tippaeg jäi valimisõhtu algusesse, samuti kurbust väljendavate säutsude osakaal. Üllatust sisaldavate säutsude osakaal hakkas tõusma paralleelselt sellega, kui sisuliselt oli Trumpi võit juba aimatav, kuid mitte kinnitatud.

Antud magistritöö oli oma olemuselt eksperimentaalne – võttes aluseks ühe konkreetse sündmuse näitel üritasin analüüsida, kuidas suurandmete kasutamine sotsiaalmeediaanalüüsis välja näeb, milles peituvad takistused, nõrkused ning võimalused. Arvestades kasutatud metoodikate peamisi muresid (andmeanalüüsi valiidsus ja andmestiku täielikkus) ei saa täielikult väita, et reaalne Twitteris kulgenud emotsionaalne areng oli selline nagu tulemustes viitasin, kuid saab väita, et emotsionaalne areng oli selline lähtudes antud mudelist. Tulemuste mudelipõhisus ei ole ilmtingimata halb omadus, kui seda tunnistada. Ka see võib pakkuda nii uut informatsiooni kui teadmisi. Suurandmete kasutamine sotsiaalmeedia analüüsis on seega võrdlemisi uudne, kuid paljulubav valdkond. Ainuke probleem selle osas on, et Eesti põhiselt ei pruugi olla piisavalt sobivat andmestikku, et suurandmeid luua.

# 11. Summary

The purpose of this master's thesis, titled "The usage of big data in social media analysis: sentiment analysis of Twitter content during the US presidential election.", was to analyse the usage of big data in social media analysis – what are the opportunities and limitations. In the empirical part of the thesis, I compared how did the course of the US presidential election night affect the collective emotional scale of Twitter tweets. For that, I collected data with a tool called DMI-TCAT. Afterwards I analysed the data using sentiment analysis, which divided tweets to six different emotions (anger, disgust, fear, joy, sadness and surprise).

At the start of my thesis, I tried to understand what big data really is – there isn't a universal definition, but they are often very similar. In summary, you could say that big data is defined as data for which the traditional means of collecting and analysing data are nearly or entirely impossible due to the sheer size and volume of the input and output. Big data is also characterized by a certain mythology – massive amount of data are supposed to contain patterns and information, which aren't noticeable using smaller amount of data.

There are two big trends how big data of social media are analysed – content analysis and network analysis. Content analysis evaluates the content of tweets, Facebook posts, profiles etc. Network analysis tries to understand the dynamics of communication and relationships of social media. Network analysis looks at either articulated relationships (like friending in Facebook or following someone on Twitter) or behavioral relationships (interactions, whether that be comment, like, retweet etc).

In this thesis I mostly studied the macro level of communication in Twitter – hashtags. Hashtags are kind of like labels, which collect a certain flow of discussion. Users define that they want to deliberately participate in a broader discussion by using hashtags. In general Twitter is a mix of private and public sphere. Hashtags can be seen as the most public sphere-like part of Twitter. The audience of Twitter consists of both known (your followers) and unknown (if somebody likes your tweet or retweets, it reaches a wider audience).

For interpretation and analysis, I used sentiment analysis, which divided the tweets into seven categories. Six of them were emotions (anger, disgust, fear, joy, sadness and suprise) and the seventh was classless category – it consisted of tweets, where the algorithms either couldn't identify the emotion or weren't capable of deciding which emotion was dominant. One also has to account that automatic text analysis has trouble identifying sarcasm and irony, the lexicon of Twitter is not ordinary due to various reasons (e.g there is a lot of slang, emoticons etc).

In the empirical part of the thesis, at first I compiled a timeline of real events of the election night and the day after (24-hour period). Afterwards I compared the evolution of emotional scale of tweets with the timeline. The day was divided into intervals of 10 minutes. Looking at the emotional scale, I mostly didn't use the overall proportions, rather I compared the numbers attached to the emotions with themself – how they changed over the course of the night. This was due to the model of the data analysis. Certain proportions of emotional scale were predetermined. For an example, happy emotion had the most keywords attached in the lexicon and due to that also had by a large margin the most tweets accounted for them (almost half).

The biggest anomalies of the 24-hour period happened during two speeches - the victory speech of Donald Trump and concession speech of Hillary Clinton. During both speeches the proportion of happy tweets increased by a noticeable margin, angry tweets on the other hand decreased by large. Clinton's speech also managed to decrease the amount of tweets containing disgust to the lowest margin of the night; for Trump's speech, the tweets containing surprise decreased by a lot.

The expressions of anger and fear both went through a constant wave of rises and tides. Anger was not very dominant in the early parts of the election night. It started to rise constantly, but not enitirely in linear line all through the night. On the other hand, happy tweets were relatively high at the start of the evening and then started to decrease. The peak of sadness also happened to be at the start of the evening. Surprise started to rise around the time, when Trump's win started to become a reality, about an hour before it was officially confirmed.

This master's thesis was in essence experimental. I took a specific event to analyse how does a big data social media analysis looks like, what are the obstacles, flaws and opportunities. Considering the main concerns of the methods I used (validity and completeness of data sets), one could not say that the emotional scale of Twitter was exactly as described, but rather that the development of emotional scale is the reflection of the model used in the analysis. Model based reflection is not necessarily a bad thing. It nevertheless can offer a great deal of information and knowledge.

# 12. Kasutatud kirjandus


- 140Dev. (2013). *Free Source Code – Twitter Database Server: MySQL Database Schema.* http://140dev.com/free-twitter-api-source-code-library/twitter-database-server/mysql-database-schema/
- 140kit. (2010). *140kit Github.* https://github.com/WebEcologyProject/140kit
- Aidlin, T. (2017). The Archivist: Save and Analyze Tweets. http://www.taidlin.com/work/the-archivist-twitter-analyzer-tool/
- 270towin. (2017). *Historical Presidential Election Information by State.* http://www.270towin.com/states/
- Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin M., . . . Widom J. (2012). *Challenges and Opportunities with Big Data*: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. http://cra.org/ccc/resources/ccc-led-whitepapers/
- Beevolve. (2012). An Exhaustive Study of Twitter Users Across the World. http://temp.beevolve.com/twitter-statistics/#c1
- Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, *66*, 125-136.
- Beran, B. (2013). *Sentiment analysis in Tableau with R.* Bora Beran Wordpress. https://boraberan.wordpress.com/2013/12/24/sentiment-analysis-in-tableau-with-r/
- Beurskens, M. (2014). Legal Questions of Twitter Research. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Bian, J., Yoshigoe, K., Hicks, A., Yuan, J., He, Z., Xie, M., ... & Modave, F. (2016). Mining Twitter to Assess the Public Perception of the "Internet of Things". *PloS one*, *11*(7), e0158450.
- Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*(3), 262-278.

- Borra, E., Laniado, D., Weltevrede, E., Mauri, M., Magni, G., Venturini, T., ... & Kaltenbrunner, A. (2015, April). A Platform for Visually Exploring the Development of Wikipedia Articles. In *ICWSM* (pp. 711-712).
- boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.
- Bruns, A. (2012). How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi. *Information, Communication & Society*, *15*(9), 1323-1351.
- Bruns, A. (2015). *Using GAWK to prepare TCAT data for Tableau part 1*. Mapping Online Publics. http://mappingonlinepublics.net/2015/03/02/using-gawk-to-prepare-tcat-data-for-tableau-part-1/
- Bruns, A., & Burgess, J. (2011a). # ausvotes: How Twitter covered the 2010 Australian federal election. *Communication, Politics & Culture*, *44*(2), 37.
- Bruns, A., & Burgess, J. E. (2011b). New methodologies for researching news discussion on Twitter.
- Bruns, A. ja Burgess, J. (2014). Crisis Communication in Natural Disasters: The Queensland Floods and Christchurch Earthquakes. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society*. Digital Formations, 89. Peter Lang, New York.
- Bruns, A., Burgess, J., & Highfield, T. (2014). A 'big data'approach to mapping the Australian Twittersphere. In *Advancing Digital Humanities* (pp. 113-129). Palgrave Macmillan UK.
- Bruns, A. ja Moe, H. (2014). Structural Layers of Communication on Twitter. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society*. Digital Formations, 89. Peter Lang, New York.
- Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: metrics for tweeting activities. *International Journal of Social Research Methodology*, *16*(2), 91-108.
- Bruns, A., & Stieglitz, S. (2014). Twitter data: what do they represent?. *it-Information Technology*, *56*(5), 240-245.
- Bruns, A., Weller, K. ja Harrington, S. (2014). Twitter and Sports: Football Fandom in Emerging and Established Markets. K. Weller, A. Bruns, J.

Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society*. Digital Formations, 89. Peter Lang, New York.

- Business Insider. (2016). *Trump shocks the world.* Business Insider live blog. http://www.businessinsider.com/election-results-live-blog-2016-11
- CITIS. (2017). *About us.* Center of IT impact studies. http://citis.ut.ee/about-us
- Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317-332.
- Couldry, N., & Powell, A. (2014). Big data from the bottom up. Big Data & Society, 1(2), 2053951714539277.
- DataSift. (2017). *Open Data Processing for Twitter.* http://datasift.com/products/open-data-processing-for-twitter/
- de Vries, A. (2016). *Text Analysis 101: Sentiment Analysis in Tableau & R.* The Information Lab. https://www.theinformationlab.co.uk/2016/03/02/text-analysis-101-sentiment-analysis-in-tableau-r/
- Easton, L. (2016). *Calling the presidential race state by state.* The Associated Press. https://blog.ap.org/behind-the-news/calling-the-presidential-race-state-by-state
- EKI. (2014). *Teemaviide.* ÕSi uued sõnad. http://keeleabi.eki.ee/?leht=9
- EKI. (2017). *Big Data.* ESTERM terminibaas. http://termin.eki.ee/mt/esterm/concept.asp?conceptID=79384&term=big%2520data
- Ekman, P. & Oster, H. (1979). Facial Expressions of Emotion. Annual Review of Psychology, 30, 527-554.
- Feinerer, I. ja Hornik, K. (2017). *tm: Text Mining Package*. R package version 0.7-1. https://CRAN.R-project.org/package=tm
- Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, *3*(1), 2053951716645828.
- Gaffney, D. ja C. Puschmann. (2014). Data Collection on Twitter. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.

- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.
- Gerlitz, C., & Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. *M/C Journal*, *16*(2).
- Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, *49*(2), 28.
- Glosbe. (2017). *Retweet*. Glosbe mitmekeelne *online* sõnastik. https://et.glosbe.com/en/et/retweet
- Gnip. (2017). *Gnip - Unleash the Power of Social Data.* https://gnip.com/
- Gray, J. (2007). *Jim Gray on eScience: a transformed scientific method*. Hey, T., Tansley, S., & Tolle, K. M. (toim.)
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 267-297.
- Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big Social Data analytics in journalism and mass communication comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332-359.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Harrington, S. (2014). Tweeting about the Telly: Live TV, Audiences, and Social Media. The use of Twitter by Professional Journalists: results of a newsroom survey in Germany. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Hawksey, M. (2017). *Get TAGS*. https://tags.hawksey.info/get-tags/
- Highfield, T. (2014). Following the Yellow Jersey: Tweeting the Tour de France. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Howard, P. N., & Woolley, S. C. Bots and Automation over Twitter during the US Election.

- Internet live stats. (2017). *Twitter Usage Statistics.* http://www.internetlivestats.com/twitter-statistics/
- Jamieson, A., Woolf, N., Levin, S. ja McCarthy, T. (2016). *US election night 2016 - as it happened.* The Guardian live blog. https://www.theguardian.com/us-news/live/2016/nov/08/us-election-2016-polls-trump-clinton-live
- Jang, S. M., & Hart, P. S. (2015). Polarized frames on "climate change" and "global warming" across countries and states: Evidence from Twitter big data. *Global Environmental Change*, *32*, 11-17.
- Jurka, T. P. (2012). *sentiment: tools for sentiment analysis.* R package version 0.2 https://github.com/timjurka/sentiment
- Kalmus, V. (2015). *Standardiseeritud kontentanalüüs*. Sotsiaalse analüüsi meetodite ja metodoloogia andmebaas. http://samm.ut.ee/kontentanalyys
- Kamhawi, R., & Weaver, D. (2003). Mass communication research trends from 1980 to 1999. *Journalism & Mass Communication Quarterly*, *80*(1), 7-27.
- Kaneko, T., & Yanai, K. (2016). Event photo mining from twitter using keyword bursts and image clustering. *Neurocomputing*, *172*, 143-158.
- Kirk, A., Scott, P. ja Graham, C. (2016). *US election results: The maps and analysis that explain Donald Trump's shock victory to become President.* The Telegraph live blog. http://www.telegraph.co.uk/news/0/us-election-results-and-state-by-state-maps/
- Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.
- Kollanyi, B, Howard, P and Woolley, SC et al., (2016). Bots and automation over Twitter during the first U.S. Presidential debate: COMPROP Data Memo 2016.1. Project on Computational Propaganda. https://ora.ox.ac.uk/objects/uuid:b8bb1dd8-58c0-440f-a9c4-c94fe722c889
- Kopan, T., Wills, A. ja Diaz, D. (2016). *Live election results and coverage.* CNN. http://edition.cnn.com/2016/11/07/politics/live-election-results-coverage/index.html
- Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social

networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790.

- Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.
- Lahuerta-Otero, E., & Cordero-Gutiérrez, R. (2016). Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. *Computers in Human Behavior*, *64*, 575-583.
- Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety.* Meta Group. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
- Lang, D. T. (2011). *Rstem: word stemming algorithm.* R package version 0.4-1. http://cran.cnr.berkeley.edu/src/contrib/Archive/Rstem/
- Larsson, A. O. ja Moe, H. (2014). Twitter in Politics and Elections: Insights from Scandinavia. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34-52.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.
- Loth, A. (2016). *How to implement Sentiment Analysis in Tableau using R?* Alex Loth blog. http://alexloth.com/2016/01/31/how-to-implement-sentiment-analysis-in-tableau-using-r/
- Lupton, D. (2016). The diverse domains of quantified selves: self-tracking modes and dataveillance. Economy and Society, 45(1), 101-122.
- Marin, A., & Wellman, B. (2011). Social network analysis: An introduction. *The SAGE handbook of social network analysis*, *11*.
- Marwick, A. E. (2011). Social privacy in networked publics: Teens' attitudes, practices, and strategies.

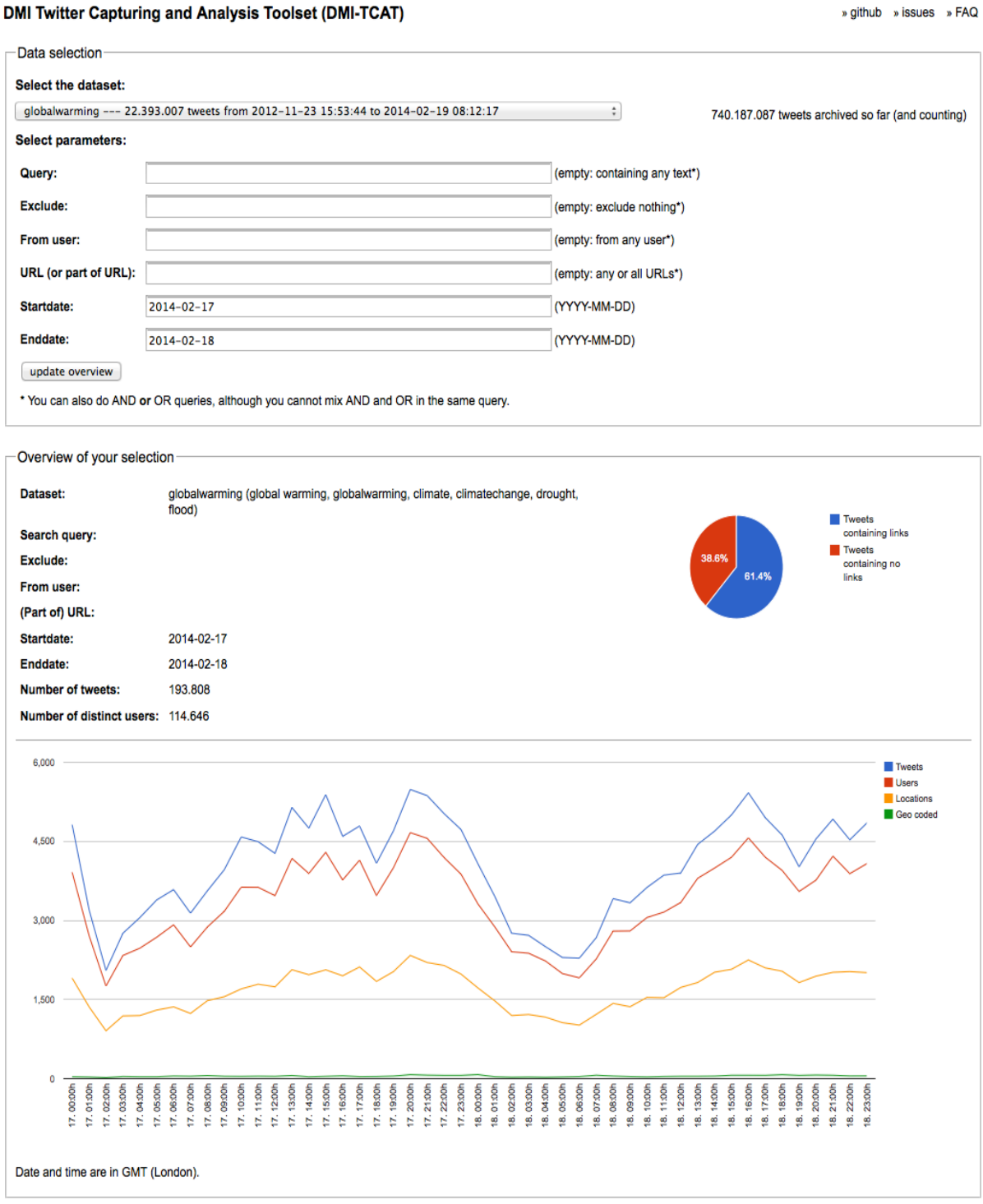
- Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, *13*(1), 114-133.
- McFarland, D. A., Lewis, K., & Goldberg, A. (2016). Sociology in the era of big data: The ascent of forensic social science. The American Sociologist, 47(1), 12-35.
- Michael, M., & Lupton, D. (2016). Toward a manifesto for the 'public understanding of big data'. *Public Understanding of Science*, *25*(1), 104-116.
- Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, *260*, 64-73.
- Muts, M. (2004). *Kollektivismi mõju teadmusülekandele (Hansapanga ja Ühispanga näitel)*. Magistritöö, Tartu Ülikool.
- NBC News. (2016). *Decision 2016 - Election Day.* NBC News live blog. http://www.nbcnews.com/storyline/2016-election-day?cid=eml_nbn_20161108
- Netlytic. (2017). *About Netlytic.* https://netlytic.org/home/?page_id=10834
- Neuberger, C., vom Hofe, H. J., Nuernbergk, C. (2014). The use of Twitter by Professional Journalists: results of a newsroom survey in Germany. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- NPR. (2016). *Live Coverage: Election Night 2016.* National Public Radio. http://www.npr.org/2016/11/08/500427835/live-blog-election-night-2016
- O'Brien, J. (2013). *yourTwapperKeeper - Archive Your Social Media.* Github. https://github.com/540co/yourTwapperKeeper
- Panger, G. (2016). Reassessing the Facebook experiment: critical thinking about the validity of Big Data research. *Information, Communication & Society*, *19*(8), 1108-1126.
- Paßmann, J., Boeschoten, T. ja Schäfer, M. T. (2014). The Gift of the Gab: Retweet Cartels and Gift Economies on Twitter. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.

- Plutchik, R. (2001). The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist, 89(4), 344-350.
- Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, *16*(3), 197-214.
- Puschmann, C. ja Burgess, J. (2014). The Politics of Twitter Data. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Rattenbury, T., Good, N., & Naaman, M. (2007, July). Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 103-110). ACM.
- Rieder, B. (2012). The refraction chamber: Twitter as sphere and network. *First Monday*, *17*(11).
- Risse, T., Peters, W., Senellart, P., Maynard, D. (2014). Documenting Contemporary Society by Preserving Relevant Information from Twitter. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Rogers, R. (2015). Digital methods for web research. Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource.
- Rosenstiel, T., Sonderman, J., Loker, K., Ivancin, M. ja Kjarval, N. (2015). *Twitter and the News: How people use the social network to learn about the world.* American Press Institute. http://www.americanpressinstitute.org/wp-content/uploads/2015/09/Twitter-and-News-How-people-use-Twitter-to-get-news-American-Press-Institute.pdf
- Ruths, D., & Pfeffer, J. (2014). *Social media for large studies of behavior*. Science, 346(6213), 1063-1064. Chicago
- Sarv, M. ja Laineste, L. (2016). *E-Eesti humanitaarteadustes*. Sirp. http://www.sirp.ee/s1-artiklid/c21-teadus/e-eesti-humanitaarteadustes/

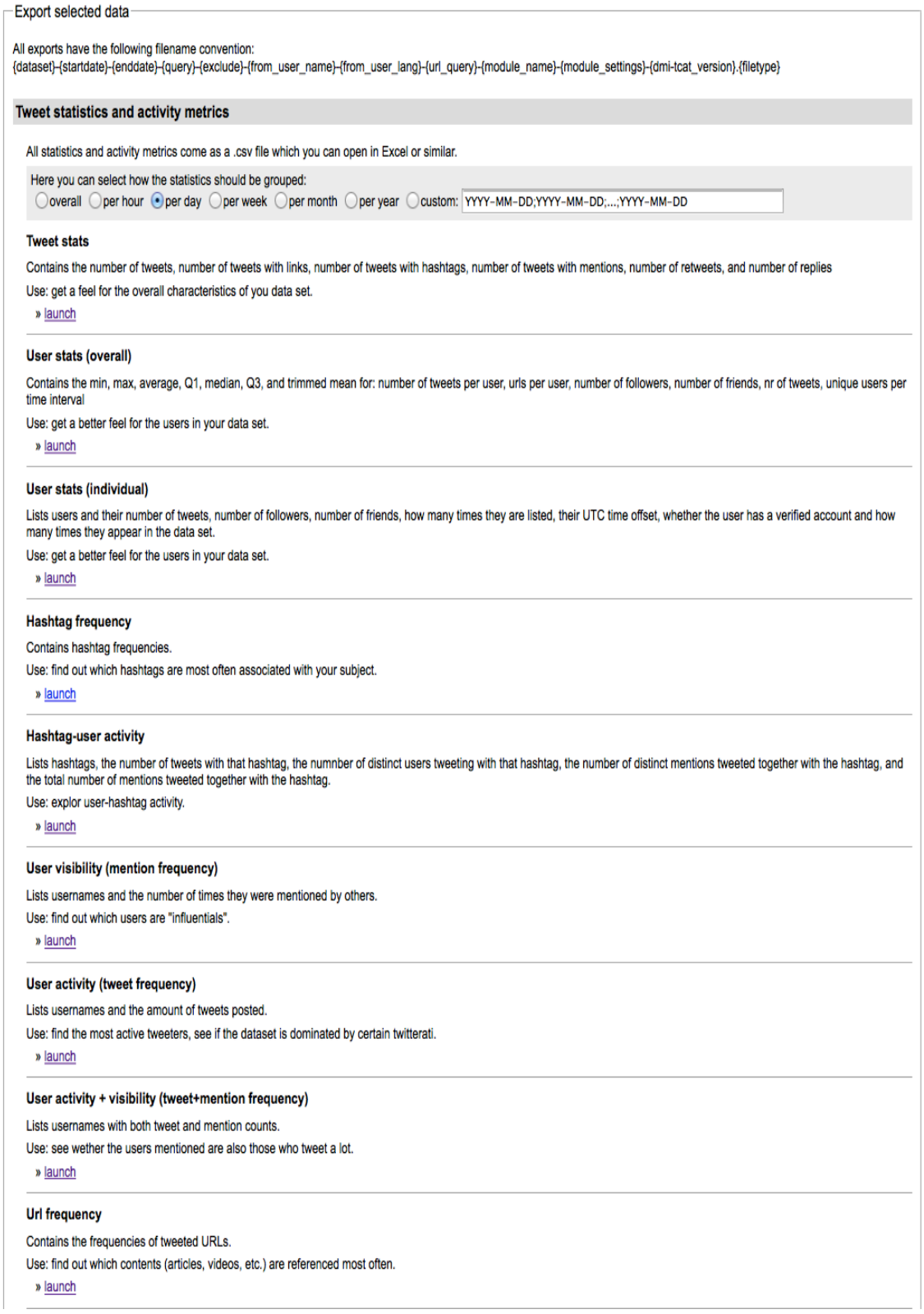
- Schmidt, J.-H. (2014). Twitter and the Rise of Personal Publics. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society*. Digital Formations, 89. Peter Lang, New York.
- Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263-286.
- Small, T. A. (2011). What the hashtag? A content analysis of Canadian politics on Twitter. *Information, Communication & Society*, *14*(6), 872-895.
- Statista. (2017). *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2017 (in millions).* https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
- Sprunt, B. (2016). *When Do Polls Close On Election Day, And Where Should I Vote?* National Public Radio. http://www.npr.org/2016/11/08/500713056/when-do-polls-close-on-election-day-and-where-should-i-vote
- Storify. (2017). *Storify API.* http://dev.storify.com/api/summary
- Sue, H. (2016). *Home.* The Digital Methods Initiative Twitter Capture and Analysis Toolset. https://github.com/digitalmethodsinitiative/dmi-tcat/wiki
- Thapen, N., Simmie, D., & Hankin, C. (2016). The early bird catches the term: combining twitter and news data for event detection and situational awareness. *Journal of Biomedical Semantics*, *7*(1), 61.
- Thelwall, M. (2014). Sentiment Analysis and Time Series with Twitter. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society*. Digital Formations, 89. Peter Lang, New York.
- Trilling, D. (2015). Two different debates? Investigating the relationship between a political debate on TV and simultaneous comments on Twitter. *Social Science Computer Review*, *33*(3), 259-276.
- TTÜ. (2017). *Sotsiaalteaduslike suurandmete professori konkurss.* https://ttu.ee/ulikool/tule-toole-tipp-ulikooli/akadeemiline-konkurss/vaata-toopakkumisi/?job=3033
- Tweet Archivist. (2017). *Frequently Asked Questions.* http://www.tweetarchivist.com/about/faq

- Twitter. (2017). *Streaming APIs.* Twitter Developer Documentation. https://dev.twitter.com/streaming/overview
- Uiboaed, K. (2017a). *Tekstide töötlemise koolitus 2017.* Tartu Ülikooli eesti ja üldkeeleteaduse instituut. https://github.com/kristel-/Tekstikoolitus-2017/blob/master/08-03-praktikum/stoppsonad.csv
- Uiboaed, K. (2017b). Autori meilivahetus eestikeelse termini tuvastamiseks.
- Vallaste, H. (2017). *e-teatmik: IT ja sidetehnika seletav sõnaraamat.* http://vallaste.ee/
- Vergeer, M., & Franses, P. H. (2016). Live audience responses to live televised election debates: time series analysis of issue salience and party salience on audience behavior. *Information, Communication & Society*, *19*(10), 1390-1410.
- Vis, F., Faulkner, S., Parry, K., Manyukhina, Y. ja Evans, L. (2014). Twitpic-ing the Riots: Analysing Images Shared on Twitter during the 2011 U.K. Riots. K. Weller, A. Bruns, J. Burgess, M. Mahrt ja C. Puschmann (toim.), *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Vits, K. (2016). *Uus keskus hindab e-riigi mõjusid.* TÜ Skytte instituut. http://skytte.ut.ee/et/uudised/uus-keskus-hindab-e-riigi-mojusid
- Wallach, H. (2016). Computational social science: Toward a collaborative future. *Computational social science: Discovery and prediction*, 307-316.
- Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.
- Weller, K., Bruns, A., Burgess, J., Mahrt, M. ja Puschmann, C. (2014). Twitter and society: introduction. *Twitter and Society.* Digital Formations, 89. Peter Lang, New York.
- Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets. *Computers in Human Behavior*, *48*, 392-400.
- Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250-261.

# 13. Lisad




DMI-TCAT kasutajaliides. (Sue, 2016)




DMI-TCAT kasutajaliides. (Sue, 2016)

# DMI-TCAT query manager

You currently have 97 query bins and are tracking 234 out of 400 possible phrases.
Your latest rate limit hit was on 2014-03-11 13:03:31

### New query bin

Bin type: keyword track

Bin name: (cannot be changed later on)

Phrases to track:

Here you can specify a list of tracking criteria consisting of single or multiple keyword queries, hashtags, and specific phrases. Each query should be separated by a comma. If you want to track a literal phrase, encapsulate it in single quotes (').

DMI-TCAT allows for three types of 'track' queries:
1. a single word/hashtag. Consider that Twitter does not do partial matching on words, i.e. [twitter] will get tweets with [twitter], [#twitter] but not [twitteraddiction]
2. two or more words: works like an AND operator, i.e. [global warming] will find tweets that have both [global] and [warming] in any position in the tweet, e.g. "life is global but not warming"
3. exact phrases: ['global warming'] will get only tweets with the exact phrase. Beware, however that due to how the streaming API works, tweets are captured in the same way as in 2, but tweets that do not match the exact phrase are thrown away. This means that you will request many more tweets from the Twitter API than you will see in your query bin - thus increasing the possibility that you will hit a rate limit. E.g. if you specify a query like ['are we'] all tweets matching both [are] and [we] are retrieved, while DMI-TCAT only retains those with the exact phrase ['are we'].

You can track a maximum of 400 queries at the same time (for all query bins combined) and the total volume should never exceed 1% of global Twitter volume, at any specific moment in time.

Example bin: globalwarming,global warming,'climate change'

add query bin

### Query manager

| querybin | active | type | queries | no. tweets | Periods in which the query bin was active | | |
|---|---|---|---|---|---|---|---|
| thedaywefightback | 1 | track | StopTheNSA 2014-02-03 12:16:37 - now<br>stopspying 2014-02-03 12:16:37 - now<br>thedaywefightback 2014-02-03 12:16:37 - now | 154.501 | 2014-02-03 12:16:37 - now | modify phrases | stop |
| turkey_censorship | 1 | track | #18Ocak18DeSokaklara 2014-01-22 12:52:02 - now<br>#4saat 2014-01-22 12:52:02 - now<br>#InternetCensorshipinTurkey 2014-01-22 12:52:02 - now<br>#SansureDurDe 2014-01-22 12:52:02 - now<br>#SansüreDurDe 2014-01-22 12:52:02 - now<br>#TürkiyedeİnternetSansuru 2014-01-22 12:52:02 - now | 69.614 | 2014-01-22 12:52:02 - now | modify phrases | stop |

DMI-TCAT kasutajaliides. (Sue, 2016)

| | Viha | Vastikustunne | Hirm | Rõõm | Kurbus | Üllatus |
|---|---|---|---|---|---|---|
| **0** | 15 | 4 | 5 | 13 | 8 | 14 |
| 1 | 2 | 1 | 0 | 8 | 3 | 1 |
| 2 | 5 | 0 | 2 | 4 | 4 | 8 |
| 3 | 8 | 3 | 3 | 1 | 1 | 5 |
| **1** | 18 | 9 | 6 | 30 | 10 | 6 |
| 1 | 12 | 7 | 4 | 24 | 5 | 1 |
| 2 | 6 | 2 | 2 | 4 | 2 | 4 |
| 3 | 0 | 0 | 0 | 2 | 3 | 1 |
| **2** | 9 | 13 | 4 | 12 | 17 | 12 |
| 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 2 | 4 | 8 | 0 | 1 | 3 | 5 |
| 3 | 2 | 5 | 4 | 11 | 14 | 7 |
| **3** | 17 | 11 | 12 | 11 | 22 | 22 |
| 1 | 2 | 0 | 1 | 0 | 0 | 0 |
| 2 | 11 | 3 | 4 | 1 | 2 | 0 |
| 3 | 4 | 8 | 7 | 10 | 20 | 22 |
| **4** | 23 | 21 | 11 | 13 | 24 | 23 |
| 1 | 5 | 0 | 0 | 0 | 0 | 0 |
| 2 | 13 | 11 | 6 | 3 | 7 | 12 |
| 3 | 5 | 10 | 5 | 10 | 17 | 11 |
| **5** | 18 | 42 | 62 | 21 | 19 | 23 |
| 1 | 15 | 41 | 59 | 17 | 14 | 18 |
| 2 | 3 | 1 | 0 | 3 | 4 | 4 |
| 3 | 0 | 0 | 3 | 1 | 1 | 1 |

Tabel 11. Proovivalimi hindamine küsimustele: "Kui hästi on automatiseeritud hindamine tabanud säutsu emotsiooni?" ja "Kui keeruline oli määrata antud säutsu emotsionaalsust?". Esimese puhul: 5 - hästi; 4 - pigem hästi; 3 - nii ja naa; 2 - pigem halvasti; 1 - halvasti; 0 - ei oska öelda. Teise puhul: 3 - keeruline; 2 - keskmine; 1 - lihtne. Rasvases kirjas olev lahter viitab küsimusele number 1, tavalises kirjas olevad lahtrid küsimusele number 2.

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina Priit Pokk
(sünnikuupäev: 25.03.1992)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
"Suurandmete kasutamisvõimalused sotsiaalmeedia analüüsis: Ameerika Ühendriikide presidendivalimiste Twitteri kajastuse meelestatuse analüüs",
mille juhendaja on Külliki Seppel,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 31.05.2017