**Domača naloga 2 – Big data**

**Uporabna statistika, Viri podatkov**

**Gašper Tomšič, 64200425**

Pristranskost zaradi kvantifikacije

Pri temi pristranskost zaradi kvantifikacije, so se članki osredotočali predvsem na pomanjkljivost kvantitativnega tipa analize pojava. Članki so se ločili na tri različne teme, Facebook, pametna mesta in Nokia. Vsi članki so opisovali dva različna načina zbiranja podatkov.

V primeru Facebooka so bili podatki zbrani glede na lastnosti uporabnikov in njihovega obnašanja. Beležilo so se njihove lastnosti, kot so spol, starost, prijatelji in podobno. Obenem pa tudi obnašanja, kot je ali berejo določene objave na zidu, koliko časa porabijo na določenih delih strani, ali in kaj komentira in podobno. Iz teh podatkov so nato analitiki v Facebooku sklepali o zanimanjih uporabnikov in kakšno vsebino naj jim priporočajo, da bo zadovoljitev s produktom čim večja. Podoben način zbiranja podatkov je bil tudi na primerih pametnih mest. Z uporabo senzorjev so merili različne naravne pojave in meščane. Nato so z podatki poskušali optimizirati mestne storitve (internet, vodovod, promet, …) in življenja meščanov. Dobra plat takega načina zbiranja podatkov je, da mogoče zbirati podatke za zelo veliko število enot in imamo praktično neomejeno količino podatkov. Slaba plat pa je, da se moramo omejiti na specifične področja, ki jih želimo analizirati. Obenem pa zgubimo tudi stik z osebami in posledično kontekst podatkov.

Nekoliko drugačen pristop zbiranja podatkov pa je bil opisan v TEDx govoru Nokia, ki je opisoval kvalitativen način pridobivanja podatkov in na primeru Nokia kritiziral kvantiteten način. Avtorica je pridobivala podatke tako, da je zožila raziskavo na majhno število raziskovalnih enot in se je poglobila, da je za te enote nabrala čim več različnih podatkov. Raziskovala je neposredno med ljudmi tako, da je opravljala lokalne službe in se družila z lokalnim mladim prebivalstvom. Tako je opazila, da med revnimi mladimi na Kitajskem prihaja do zelo velikih sprememb v potrošnji, ki jo navadne tržne raziskave niso zaznale. Ta zgodba je pokazala pomembnost obogatitve podatkov z kontekstom. Dobimo tako imenovane »thick data«. Težava takega pristopa pa je, da smo omejeni na preučevanje zelo malo ljudi in ne moremo avtomatizirati take analize.

Načeloma kvalitativni podatki ne morejo biti »big data«, torej primer raziskave v TEDx govoru bi rekli, da ne gre za »big data«. Pri primeru Facebook in pametna mesta lahko rekli, da je »big data«, saj imamo opravka z zelo veliko količino podatkov in tako veliko različnih spremenljivk, da jih praktično ni mogoče analizirati.

Ključna spoznanja – Imamo nek problem, ki je zelo kompleksen in ga želimo analizirati. Iz člankov so bili primeri takih problemov na primer osebe oziroma pametna mesta. Vsaka oseba je že po naravi tako kompleksna, da jo je skoraj nemogoče opisati v podatkih. Če želiš osebo opisati s 1000 spremenljivkami ne boš prišel niti blizu temu, da bi osebo dobro poznal. Vendar to še vedno poskušamo narediti, torej dobimo pristranskost zaradi kvantifikacije. Še dodatna težava nastane, ko želimo opisati več oseb ali pa kar celotne skupnosti oseb. Problem postane tako kompleksen, da je teoretično nerešljiv. V praksi problem posplošimo in vnesem precejšno pristranskost.

Aha moment – Sicer sem že zasledil v določenih znanstvenih člankih, ampak bo združitev kvantitetnega in kvalitativnega načina v en model verjetno pogost način analize podatkov v prihodnosti. Veliko statističnih metod je že zelo starih in se napredki dogajajo samo v optimizaciji modelov. Novejše področje kot je globoko učenje je sicer zelo močan način analize podatkov, vendar ima težavo, ker izgubimo kontekst.

Aha moment – V člankih je bilo omenjeno, da podatkovni modeli niso narejeni kar sami od sebe, ampak jih ustvarjajo ljudje. Torej kakršenkoli vpliv imajo modeli je v bistvu vpliv ljudi, ki so naredili te modele. Kdo bo reguliral ta vpliv? Politični organi ali bo prišlo do samo regulacije?