Tampereen yliopisto

Tommi Tammilehto

ELOKUVA-ARVOSTELUJEN HYÖDYNTÄMINEN REGRESSIOANALYYSISSÄ TEKSTIDATAN ANALYYSIN KEINOIN

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Huhtikuu 2021

TIIVISTELMÄ

Tommi Tammilehto: Elokuva-arvostelujen hyödyntäminen regressioanalyysissä tekstidatan analyysin keinoin
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Huhtikuu 2021

Tekstidatan analyysissä pyritään käsittelemään tekstimuotoista aineistoa tilastollisin menetelmin. Tekstidatan analyysi on tehokas tapa kerätä hyödyllistä tietoa erinäisistä kirjallisuuden lähteistä. Tekstidatana voidaan käyttää fyysistä kirjallisuutta kuten kirjoja tai lehtiä, mutta lisäksi erityisesti verkkotekstejä hyödynnetään tekstidatana. Internetin alustojen kasvaessa myös useimmat elokuva-arvostelut ovat siirtyneet verkkoon, ja kynnys omien arvostelujen julkaisulle on madaltunut. Nykypäivänä kuka vain voi kirjoittaa ja julkaista arvostelunsa lukuisilla Internetin arvostelualustoilla. Tämän työn tavoitteena on soveltaa tekstidatan analyysin keinoja Internetissä julkaistuihin elokuva-arvosteluihin ja tutkia niiden hyödyntämistä elokuvan laadun ennustamisessa tilastollisin keinoin.

Työn aineisto on satunnaisotannalla valikoitu suuremmasta elokuva-arvostelujen aineistosta. Aineiston arvostelut ovat vuosilta 1987–2001. Kunkin arvosteltavan elokuvan laadun mittarina käytetään elokuvakriitikoiden arvioiden keskiarvoa kyseiselle elokuvalle. Aineiston arvosteluista määritetään tekstin sanamäärä, kirjoitusvirheiden suhteellinen määrä, luettavuus, sentimentti sekä kirjoittajan antama numeerinen arvosana elokuvalle. Lineaarisessa regressioanalyysissä aineiston muuttujilla pyritään selittämään tutkittavaa eli selitettävää muuttujaa. Tässä työssä regressioanalyysillä elokuvakriitikoiden arvioiden keskiarvoa pyritään selittämään arvosteluista kerättyjen muuttujien avulla. Päätyövälineenä työssä käytetään tilastollista ohjelmistoa R.

Regressioanalyysissä valittiin malli, jolla elokuvakriitikoiden arvioiden keskiarvoa ennustetaan kirjoittajan antamalla arvosanalla elokuvalle sekä arvostelun sentimentillä. Mallin mukaan elokuvakriitikoiden arvioiden keskiarvo kasvaa, kun kirjoittajan antama arvosana elokuvalle kasvaa ja arvostelun sentimentti muuttuu positiivisemmaksi.

Työn aineiston perusteella ei voitu luoda regressiomalliin muuttujia mallintamaan arvostelun kirjoittajan ominaisuuksia. Lisäksi mallissa ei huomioida mahdollisia eroja elokuvakriitikoiden ja aineiston arvostelujen kirjoittajien suhtautumisessa tiettyihin elokuviin. Jatkossa mallia voisi tarkentaa näiltä osin. Lisäksi tutkittua yhteyttä muuttujien välillä voisi mallintaa myös muilla tilastollisilla menetelmillä.

Avainsanat: sentimenttianalyysi, luettavuusindeksi, käyttäjäarvostelu

SISÄLLYSLUETTELO

1. JOHDANTO

Erilaisiin analyyseihin soveltuvat tekstit ovat enenevissä määrin käytetty datan lähde. Tekstidataa voidaan hyödyntää useisiin tarkoituksiin, kuten roskapostin erottamiseen saapuvista sähköpostiviesteistä (Jiang 2010, s. 38), markkinointitarkoituksiin (Aggarwal ja Zhai 2012, s. 8) tai tapahtumien tai asioiden merkittävyyden arviointiin sosiaalisen median teksteistä (Venekoski ja Vankka 2017, s. 164).

Tekstidataa voidaan kerätä kirjallisuudesta, kuten kirjoista, lehdistä tai opinnäytetöistä. Tekstidataa voidaan kerätä myös Internetin avoimilta alustoilta, kuten verkkokaupoista, blogeista tai keskustelupalstoilta. Tekstidatan keräämiseen, niin kutsuttuun louhimiseen, käytettäviä menetelmiä voidaan hyödyntää myös muihin tarkoituksiin, kuten proteiinien ja geenidatan tunnistuksessa biolääketieteen alalla (Aggarwal ja Zhai 2012, s. 8).

Keskeisiä tekstidatan lähteitä ovat muun muassa käyttäjä- ja elokuva-arvostelut. Useimmilla Internetin myyntisivustoilla ja -palstoilla on osio, johon palvelun asiakkaat voivat jättää arvosteluja ostamistaan tuotteista tai käyttämistään palveluista. Käyttäjäarvostelut voivat esimerkiksi koskea jääkaappia, televisiota, autoa, teatteriesitystä tai ravintolakäyntiä. Tietyn tuote- tai palvelutyypin arvosteluille on myös omia arvostelualustojaan, kuten TripAdvisor.com, missä käyttäjät voivat kirjoittaa arvosteluja vierailemistaan matkailukohteista tai RateMyProfessors.net, missä korkeakouluopiskelijat voivat arvioida korkeakoulunsa opettajia ja professoreita (Masum ja Tovey 2011, s. xvii). Internetissä jaettavat elokuva-arvostelut sivustoilla kuten IMDb.com vastaavat rakenteeltaan esimerkiksi sanomalehdissä julkaistavia, ammattikriitikoiden kirjoittamia arvosteluja.

Tekstidatan analyysissä aineiston muodostavat tekstit muunnetaan useimmiten helpommin käsiteltävään muotoon. Lähtökohtana on tekstimuotoisen aineiston muuntaminen numeeriseen muotoon, joka soveltuu paremmin tilastollisiin tarkoituksiin. Yleisimmin tekstistä muodostetaan siinä esiintyvien sanojen esiintymismäärien mukaan matriisi. Tällöin aineistoa voidaan käsitellä matriisimuodossa, jolloin tilastollisten analyysien tekeminen on helpompaa. (Weiss et al. 2015, s. 3.)

Tämän työn tavoitteena on hyödyntää tekstidatan analyysin keinoja tekstimuotoisen aineiston käsittelyssä tilastollista analyysiä varten. Luvussa 2 on esitelty työssä käytetty aineisto. Luvussa 3 on selostettu tarkemmin työssä käytetyt menetelmät ja niiden teoreettinen tausta sekä käytettävät muuttujat. Luvussa 4 on esitelty tilastollisen analyysin tulokset. Luvussa 5 on työn yhteenveto sekä pohdintaa aiheesta jatkossa.

2. AINEISTO

Tässä työssä käytettiin Cornell Universityn vapaasti jakamaa elokuva-arvostelujen aineistoa (Pang, Lee ja Vaithyanathan 2002). Aineisto koostuu yksittäisistä html-tiedostoista, joista jokainen sisältää yhden arvostelun. Aineiston arvostelut on kerätty uutisryhmästä rec.arts.movies.reviews, joka myöhemmin siirtyi omalle sivustolleen ja toimii nykyisin osoitteessa www.IMDb.com. Useimmissa aineiston elokuva-arvosteluissa arvostelun kohteena on uusi, vielä elokuvateattereissa näytettävä elokuva, mutta aineistossa on myös arvosteluja vanhemmista elokuvista. Arvostelut ovat julkaistu vuosina 1987–2001.

Aineistosta muodostettiin otos satunnaisotannalla. Satunnaisotoksessa on 277 arvostelua. Tiedostoista kerättiin arvioitavan elokuvan nimi sekä arvostelun loppuarvio, esimerkiksi 3/5 tähteä. Kaikki loppuarviot muunnettiin skaalaan 0-100. Esimerkiksi asteikolla 1–5 arvosana 3,5 vastaa 63:a pistettä sadasta, kirjain-skaalalla F–A+ arvosana C- vastaa 33:a pistettä sadasta. Arvosanojen muuntotaulukot ovat liitteessä 3.

Aineistosta on tehty useita analyysejä, kuten aineiston kerääjien sentimenttianalyysit *Thumbs up? Sentiment Classification using Machine Learning Techniques* (Pang, Lee ja Vaithyanathan 2002) sekä *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts* (Pang ja Lee 2004). Aineistosta on tehty myös arviointiasteikkoja koskeva tutkimus *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales* (Pang ja Lee 2005).

Elokuvan nimen perusteella kerättiin Metacritic.com -sivustolta keskimääräistä elokuvakriitikoiden antamaa arvosanaa elokuvasta pisteyttävä arvo väliltä 0-100. Tämä arvo on painotettu siten, että arvostelut, joiden kirjoittaja on tunnetumpi tai luotettavampi, on enemmän painoarvoa pisteytyksessä. Arvot ovat lisäksi normalisoituja, eli arvoja on painotettu kohti pisteytysskaalan keskikohtaa. Jos kyseiselle elokuvalle ei saatu tällä tavalla arvoa, tilalle haettiin RottenTomatoes.com -sivustolta keskimääräinen kriitikkojen antama arvosana. Muutamista elokuvista ei saatu kumpaakaan. 187:sta arvostelusta saatiin kerättyä ja muunnettua arvostelun kirjoittajan antama arvosana elokuvalle. Lisäksi 81:sta puuttuu kirjoittajan antama arvosana, mutta keskimääräinen arvosana kriitikoilta saatiin kerättyä. Lopuista arvosteluista puuttuu keskimääräinen kriitikoiden antama arvosana.

3. MENETELMÄT

Tämän työn tarkoituksena oli ennustaa elokuvien arviointia kriitikoilta aineiston arvostelujen perusteella. Aineiston arvosteluista konstruoitiin muuttujia, joiden avulla pyrittiin selittämään kriitikoiden antamia arvosanoja. Päämenetelmänä tilastollisessa analyysissä käytettiin lineaarista regressioanalyysiä.

Työ toteutettiin hyödyntäen tilastollista ohjelmistoa R laajasti työn eri vaiheissa. Ohjelmiston oletusfunktioiden lisäksi työssä hyödynnettiin myös muista kirjastoista peräisin olevia funktioita. Käytettyjen kirjastojen lista löytyy liitteestä 1.

3.1 Lineaarinen regressioanalyysi

Lineaarisessa regressioanalyysissä selitettävää muuttujaa pyritään selittämään aineiston arvosteluista kerättyjen selittävien muuttujien avulla. Regressioanalyysin tuloksena saadaan selittäviä muuttujia vastaavien parametrien estimaatit, joita hyödyntämällä voidaan ennustaa selitettävän muuttujan eli vastemuuttujan arvoa selittävien muuttujien arvoilla. Regressioanalyysissä malli esitetään muodossa

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n + \epsilon,$$

missä y on vastemuuttujan arvo, β_0 on vakiokerroin, $\beta_1, \beta_2, \ldots, \beta_n$ ovat selittävien muuttujien kertoimet, $x_1, x_2, \ldots, x_n$ ovat selittävien muuttujien arvot ja ϵ on virhetermi. Tässä esityksessä mallissa on n selittävää muuttujaa. Tämä malli voidaan esittää myös matriisimuodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

missä $\mathbf{y}$ on vastemuuttujien vektori, $\mathbf{X}$ on mallin suunnittelumatriisi, joka sisältää sarakkeen arvoja 1 sekä selittävien muuttujien arvojen $x_1, x_2, \ldots, x_n$ esityksen sopivassa muodossa, $\boldsymbol{\beta}$ on muuttujien kerrointen vektori, joka sisältää kertoimet $\beta_1, \beta_2, \ldots, \beta_n$ ja vektori $\boldsymbol{\epsilon}$ on mallin virhevektori, joka sisältää havaintojen virhetermit. Suunnittelumatriisin ensimmäinen sarake muodostaa mallin vakiotermin. (Kutner 2005, s. 197.)

Regressiomallin muodostuksessa tehdään oletuksia koskien mallin virhetermiä ϵ. Virhetermin oletetaan olevan normaalijakautunut odotusarvolla 0 ja varianssilla σ^2 (Kutner 2005, s. 26). Erityisesti virhetermin varianssin oletetaan olevan sama kaikilla havainnoilla

(Kutner 2005, s. 10). Lisäksi eri virhetermien oletetaan olevan korreloimattomia (Kutner 2005, s. 10).

Mallin virhetermiä koskevien oletusten lisäksi tehdään myös oletus poikkeavista arvoista. Suuresti aineiston muista havainnosta poikkeava arvo saattaa vaikuttaa malliin siten, että se ei sovi hyvin koko aineistoon. (Kutner 2005, s. 108.)

Parametrien estimointi tehdään hyödyntäen pienimmän neliösumman menetelmää. Menetelmällä minimoidaan virhetermien neliösumma

$$Q = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

jolloin estimoitava regressiomalli sopii aineistoon parhaiten. Minimoitujen virhetermien neliösumman lausekkeesta voidaan ratkaista selittävien muuttujien kertoimet, jolloin estimoidut kertoimet saadaan lausekkeesta

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}.$$

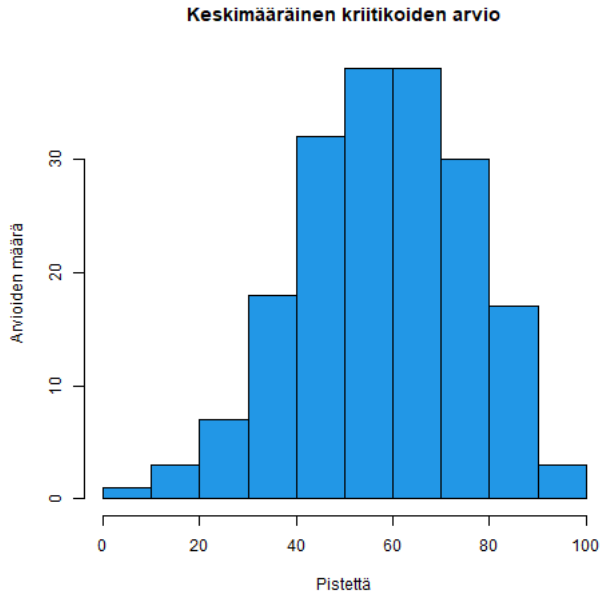
Estimoiduilla kertoimilla pyritään minimoimaan mallin virhetermien vaikutus. (Kutner 2005, s. 15.)

3.2 Työn muuttujien esittely

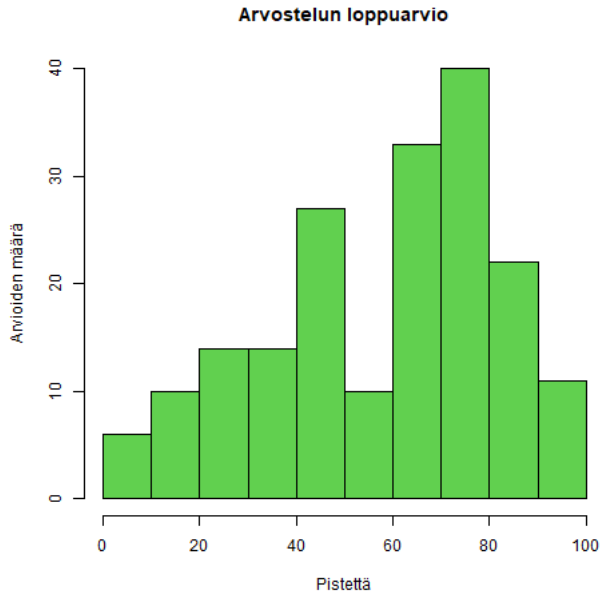
Vastemuuttujana regressioanalyysissä on keskimääräinen kriitikoiden antama arvosana elokuvalle. Se kuvaa, miten elokuvaan on suhtauduttu kriitikoiden keskuudessa keskimäärin. Vastemuuttujaa merkitään malleissa y.

Aineiston arvosteluista kerätty, arvostelijan antama, skaalattu numeerinen arvosana elokuvalle on analyysissä selittävänä muuttujana. Muuttujan alkuperän vuoksi siinä on keskittymiä erityisesti arvoissa 50 ja 75, sillä suurimmasta osasta muunnettavista skaaloista saatiin muunnettua arvoja näiksi arvoiksi. Yleisesti muuttuja on enemmän levittynyt kuin keskimääräinen kriitikoiden antama arvosana, sillä tässä jokainen arvosana koostuu vain yhden henkilön arvioinnista elokuvasta (kuva 3.2). Vastemuuttujan arvot koostuvat useiden arvosanojen keskiluvusta, minkä vuoksi arvot ovat yleisesti lähempänä asteikon keskikohtaa. Lisäksi, vastemuuttujan arvot ovat normalisoituja, mikä myös lisää arvojen keskittymistä asteikon keskelle (kuva 3.1). Muuttujaa, joka vastaa arvostelijan antamaa arvosanaa elokuvasta, merkitään malleissa x_1.

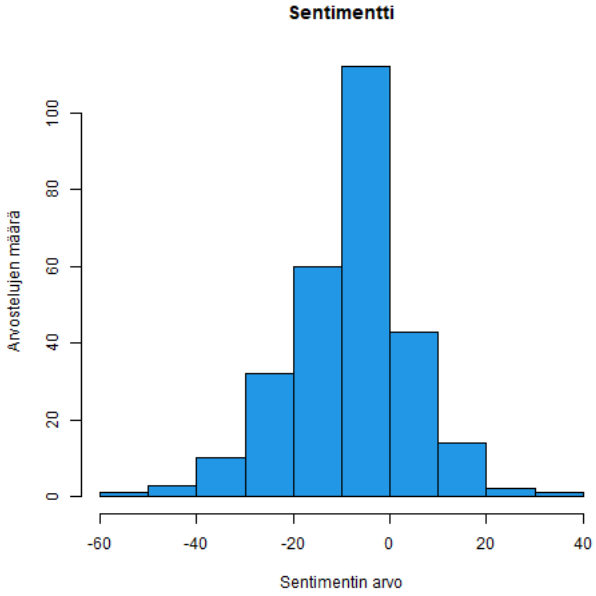
Tekstien luettavuutta voidaan arvioida erilaisilla luettavuus-testeillä ja -indekseillä. Testit tai indeksit perustuvat useimmiten yhden tai muutaman tekstin piirteen, kuten keskimääräinen sanan tai virkkeen pituus (DuBay 2004, s. 43). Arvosteluiden tekstien luettavuus määriteltiin hyödyntäen quentada-kirjaston luettavuus-testiä R:ssä. Analyysissä käytettiin SMOG-testiä, joka perustuu vähintään kolmetavuisten sanojen määrään. SMOG-testin



Kuva 3.1. *Keskimääräinen kriitikoiden arvio -muuttujan jakauma*



Kuva 3.2. *Käyttäjäarvostelujen loppuarvio -muuttujan jakauma*



Kuva 3.3. *Sentimentti-muuttujan jakauma*

tulos kuvaa kouluvuosien määrää, mitä lukijalta vaaditaan tekstin ymmärtämiseksi. Siten matalampi muuttujan arvo vastaa helpommin luettavaa tekstiä. SMOG-testiarvot saadaan

$$SMOG = 1.043 \cdot \sqrt{n_{kt}} \cdot \frac{30}{n_l} + 3.1291,$$

missä n_{kt} on vähintään kolmetavuisten sanojen määrä ja n_l on lauseiden määrä tekstissä. Luettavuutta merkitään malleissa x_2.

Tekstin sentimentillä tarkoitetaan tunteita, joita tekstistä välittyy. Tässä työssä tekstin sentimentillä tarkoitetaan erityisesti tekstin positiivista tai negatiivista sävyä. Tekstin sentimentti on arvosteluissa keskeistä, sillä arvostelut ovat mielipidetekstejä ja kirjoittajan mielipide useimmiten kuvastuu tekstissä positiivisuutena tai negatiivisuutena. Arvosteluiden tekstien sentimenttiä arvioidaan R:n tidyverse- ja tidytext-kirjastojen funktioita hyödyntäen. Sentimentin arviointi perustuu bing-sanastoon (Hu ja Liu 2004), jossa sanat ovat merkitty positiivisiksi +1 tai negatiivisiksi -1. Tekstin sentimentti lasketaan sen sisältämien sanojen sentimenttiarvojen summana. Kuvassa 3.3 on aineiston sentimenttiarvojen jakauma. Tekstien sentimentti on keskimäärin enemmän negatiivinen kuin positiivinen, joten aineiston arvosteluissa käytetään enemmän negatiivisia sanoja kuin positiivisia. Sentimenttiä merkitään malleissa x_3.

Sanamäärä-muuttuja kuvaa arvostelun sanamäärää. Sanamäärän laskemisessa on käytetty R:n ngram-kirjaston työkaluja. Kirjoitusvirheet-muuttuja kuvaa kirjoitusvirheiden määrää suhteessa arvostelun koko sanamäärään. Kirjoitusvirheiden tunnistuksessa on hyödynnetty hunspell-kirjaston funktiota R:ssä. Sanamäärää merkitään malleissa x_4 ja kirjoitusvirheitä x_5.

4. TULOKSET

Regressiomallin sovitus aloitettiin täydestä mallista 1:

Malli 1. $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \epsilon$

Mallin valinta tehtiin taaksepäin askeltavasti menetelmän yksinkertaisuuden vuoksi. Mallista poistettiin yksitellen muuttujia, jonka t-testiarvo oli pienin. Kullekin muuttujalle testiarvo on

$$t^* = \frac{\hat{b}_k}{s(\hat{b}_k)},$$

missä $\hat{b}_k$ on muuttujan kertoimen estimaatti, $s(\hat{b}_k)$ on estimaatin keskihajonta ja k on muuttujan indeksi, kun $k = 0, 1, 2, 3, 4, 5$. Testiarvon perusteella tehtävän testauksen nollahypoteesin mukaan muuttujan kerroin on nolla, jolloin muuttuja ei ole mukana mallissa. Vaihtoehtoisen hypoteesin mukaan kerroin on nollasta eroava. (Kutner 2005, s. 228.)

Malleista poistettiin muuttujia, kunnes mallissa oli jäljellä vain yksi selittävä muuttuja. Näin saatiin yksinkertaistetut mallit 2, 3, 4 ja 5:

Malli 2. $y = \beta_0 + \beta_1 \cdot x_1 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_5 + \epsilon$

Malli 3. $y = \beta_0 + \beta_1 \cdot x_1 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \epsilon$

Malli 4. $y = \beta_0 + \beta_1 \cdot x_1 + \beta_3 \cdot x_3 + \epsilon$

Malli 5. $y = \beta_0 + \beta_1 \cdot x_1 + \epsilon$

Aineistoon sovitettiin myös mallit, jossa kirjoitusvirheet-muuttuja on logaritmoitu. Tällä muunnoksella ei kuitenkaan saatu parempia malleja.

Edellä esitettyjä malleja verrattiin mallien vakioiduilla selitysasteilla ja Akaiken informaatiokriteereillä sekä bayesiläisellä informaatiokriteerillä. Mallin selitysaste

$$R^2 = 1 - \frac{\Sigma_{i=1}^n \epsilon_i^2}{\Sigma_{i=1}^n (y_i - \bar{y})^2}$$

kuvaa vaihtelevuuden osuutta, jonka malli selittää. Yhtälössä ϵ_i ovat mallin residuaalit, y_i vastemuuttujan arvot ja $\bar{y}$ vastemuuttujan arvojen keskiarvo. Koska selitysaste ei huomioi mallin selittävien muuttujien määrää, mallin selitysaste ei koskaan pienene selittävien

muuttujien lisäyksen johdosta. Tämän vuoksi mallien vertaamisessa käytetäänkin usein vakioitua selitysastetta, joka on

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

missä n on havaintojen määrä ja p on selittävien muuttujien määrä (Kutner 2005, s. 355). Mallin valinnassa valitaan malli, jonka vakioitu selitysaste on suurin.

Akaiken informaatiokriteeri (Akaike 1974, s. 716)

$$AIC = n \cdot ln(\Sigma_{i=1}^{n}\epsilon_i^2) - n \cdot ln(n) + 2p$$

on toinen yleisesti käytetty kriteeri mallin valinnassa (Kutner 2005, s. 359). Mallin valinnassa valitaan malli, josta laskettu Akaiken informaatiokriteeri on pienin.

Kolmantena valintakriteerinä käytetään bayesiläistä informaatiokriteeriä (Schwarz 1978, s. 462)

$$BIC = n \cdot ln(\Sigma_{i=1}^{n}\epsilon_i^2) - n \cdot ln(n) + ln(n) \cdot p.$$

Mallin valinnassa valitaan malli, jolla on pienin bayesiläisen informaatiokriteerin arvo (Kutner 2005, s. 359).

Malli	R_a^2	AIC	BIC
1	0.4184	1508.441	1531.059
2	0.4213	1506.534	1525.921
3	0.4148	1507.638	1523.793
4	0.4165	1506.111	1519.035
5	0.4028	1509.453	1519.147

Taulukko 4.1. *Regressiomallien kriteerien arvot*

Taulukossa 4.1 on lasketut kriteerien arvot esitellyille malleille. Näiden perusteella edellä esitellyistä malleista mallit 2 ja 4 olivat sopivimmat, sillä mallin 2 vakioitu selitysaste oli suurin ja mallilla 4 informaatiokriteereiden arvot olivat pienimmät. Päätös lopullisesta mallista tehtiin testaamalla, onko kirjoitusvirheiden määrää kuvaavan muuttujan poistamisella tilastollisesti merkittävä vaikutus mallille. Testaus tehdään t-testillä kuten edellä. Vertaamalla saatua testisuuretta t-jakaumaan saadaan *p*-arvoksi $p = 0.0826$. Tämä tarkoittaa sitä, että t-jakautuneen satunnaismuuttujan todennäköisyys saada itseisarvoltaan vastaava tai suurempi arvo kuin saatu testisuure on 0.0826. Kun testaus tehdään 5 %:n merkitsevyydellä, testisuureet, joiden *p*-arvo on alle 0.05, katsotaan olevan harvinaisia ja johtavan nollahypoteesin hylkäämiseen (Casella ja Berger 2002, s. 397). Testauksen pe-

rusteella 5 %:n merkitsevyydellä nollahypoteesi jää siis voimaan. Siten valitaan malli 4 lopulliseksi malliksi. Mallin 4 soveltuvuutta tarkastellaan vielä luvussa 3.1 esiteltyjen mallin muodostukseen liittyvien oletusten kannalta. Lopullisen mallin residuaalitarkastelu on liitteessä 2.

5. JOHTOPÄÄTÖKSET

Tässä työssä esiteltiin tekstidatan analyysin keinoja tekstimuotoisen aineiston muuntamisessa regressioanalyysiin sopivaksi. Lopullinen regressiomalli selittää elokuvan saamaa arviointia kriitikoilta käyttäjäarvostelun loppuarviolla sekä arvostelun sentimentillä. Mallin kummankin selittävän muuttujan parametrin estimaatti on positiivinen, joten kriitikkoarvion voidaan katsoa kasvavan käyttäjäarvion kasvaessa tai arvostelun sentimentin muuttuessa positiivisemmaksi.

Taulukossa 5.1 on esitetty mallin parametrien tunnusluvut sekä merkitsevyys. Vakiotermin ja muunnetun arvion vaikutus mallissa ovat merkittäviä (*p*-arvo 0). Myös sentimentin vaikutus mallissa on merkittävä 5 %:n merkitsevyydellä (*p*-arvo 2.20 %).

	Estimaatti	Keskivirhe	t-testiarvo	P($>$|t|)
Vakiotermi	33.7245	2.8731	11.74	0.0000
Muunnettu arvio	0.4299	0.0413	10.40	0.0000
Sentimentti	0.1962	0.0849	2.31	0.0220

Taulukko 5.1. *Lopullisen regressiomallin tunnusluvut*

Työn eri vaiheissa on muutamia seikkoja, jotka saattavat vaikuttaa saatuihin tuloksiin. Aineistossa ei ole huomioitu arvostelujen kirjoittajia, vaikka esimerkiksi kirjoittajien taustoilla voi olla merkitystä malliin. Aktiivinen kirjoittaja, jolla on jo vakiintunut kirjoitustyyli sekä arviokriteerit, eroaa aloittelevasta kirjoittajasta. Myös kulttuuri, josta kirjoittaja tulee, voi vaikuttaa arvosteluun (Koh et al. 2010, s. 384).

Toinen merkittävä tekijä on arvostelun kirjoittamisen kynnys. Ammattikriitikko kirjoittaa arvostelun katsomastaan elokuvasta riippumatta katsomiskokemuksestaan. Tavalliselle arvostelun kirjoittajalle sen sijaan arvostelun kirjoittaminen saattaa olla päätös, jonka kirjoittaja tekee vasta elokuvan nähtyään, kenties koska kokemus oli erityisen hyvä tai huono. Tämän vuoksi aineiston elokuvissa voi olla valikoitumisharhaa.

Jatkossa aihetta voidaan tutkia toisesta näkökulmasta hyödyntämällä muita regressiomalleja kuin lineaarinen regressiomalli. Osaa muuttujista, kuten arvostelun luettavuus sekä kirjoitusvirheiden suhteellinen määrä, voisi mallintaa satunnaisvaikutuksina. Tällöin niillä voisi selittää vaihtelua arvosteluiden loppuarvioiden välillä. Sekamalli, jossa osa muuttujista olisi mallinnettu satunnaisvaikutuksina, voisi olla parempi kuin tässä työssä esitetty paras malli.

6. LÄHDELUETTELO

Aggarwal, Charu C. ja Zhai, ChengXiang (2012). *Mining text data*, Springer, New York.

Akaike, Hirotogu (1974). A new look at the statistical model identification, *IEEE transactions on automatic control*, 19 (6), 716–723. DOI: 10.1109/TAC.1974.1100705

Casella, George ja Berger, Roger L. (2002). *Statistical Inference*, 2. painos, Pacific Grove California, Duxbury.

DuBay, William H. (2004). *The principles of readability*, Costa Mesa, CA: Impact Information.

Hu, Minqing ja Liu, Bing (2004). Mining and summarizing customer reviews, *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, 168–177, ACM, DOI: 10.1145/1014052.1014073

Jiang, Eric P. (2010). Content-based email classification using machine-learning algorithms, Berry, Michael W ja Kogan, Jacob, *Text Mining Applications and Theory*, John Wiley & Sons, New Jersey.

Koh, Noi Sian, Hu, Nan ja Clemons, Eric K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures, *Electronic Commerce Research and Applications*, 9 (5), 374–385, Amsterdam, Elsevier, DOI: 10.1016/j.elerap.2010.04.001

Kutner, Michael H. (2005). *Applied Linear Statistical Models*, viides painos, McGraw-Hill Irwin, Boston.

Masum, Hassan ja Tovey, Mark (2011). *The reputation society : how online opinions are reshaping the offline world*, MIT Press, Cambridge, Mass.

Pang, Bo, Lee, Lillian ja Vaithyanathan, Shivakumar (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 79–86.

Pang, Bo ja Lee, Lillian (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 271–278.

Pang, Bo ja Lee, Lillian (2005). Seeing stars: Exploiting class relationships for sentiment

categorization with respect to rating scales, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 115–124.

Schwarz, Gideon (1978). Estimating the Dimension of a Model, *The Annals of statistics* 6 (2): 461–464. DOI: 10.1214/aos/1176344136

Venekoski, Viljami ja Vankka, Jouko (2017). Kieliteknologia analytiikan tukena sotilas- ja viranomaistyössä, Suomen sotatieteellinen seura, Silvasti, Markus (toim.), *Tiede ja Ase*, 155–178, Tampere.

Weiss, Sholom M., Indurkhya, Nitin, Zhang, Tong (2015). *Fundamentals of Predictive Text Mining*, toinen painos, Lontoo, Springer.

Aineisto: Pang, Bo ja Lee, Lillian (2002).*Pool of 27886 unprocessed html files*. Osoite, josta aineisto on saatavissa: https://www.cs.cornell.edu/people/pabo/movie-review-data/

7. LIITTEET

7.1 Liite 1: Lista käytetyistä R-kirjastoista

Tässä liitteessä on listattu kaikki työssä käytetyt R-kirjastot.

Kirjoitusvirheiden laskenta: "hunspell"
Luettavuustestit: "quanteda"
Sentimenttianalyysi: "syuzhet", "tidyverse"ja "tidytext"
Microsoft Excel-tiedostojen lukeminen: "xlsx"
Sanamäärän laskenta: "ngram"
Taulukkojen siirtäminen LaTeXiin: "xtable"

7.2 Liite 2: Lopullisen regressiomallin residuaalitarkastelu

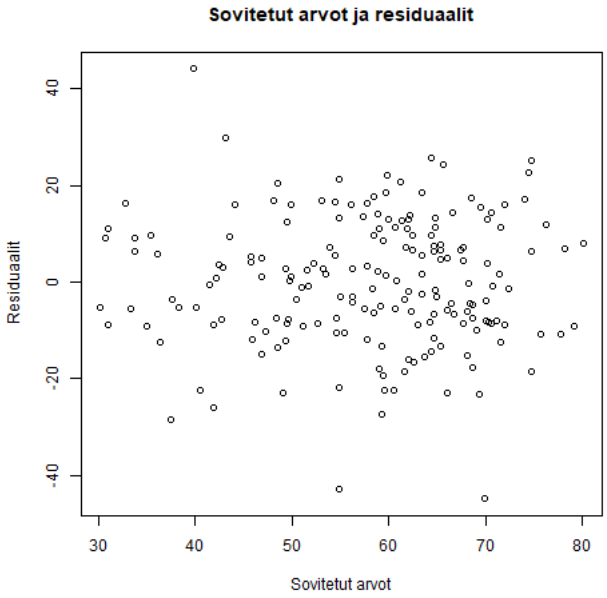
Tässä liitteessä on tarkasteltu regressiomallin residuaaleihin liittyvien oletusten toteutumista. Regressiomallissa on oletettu, että mallivirheen vaihtelu on tasaista läpi aineiston eli että mallin virhetermin varianssi on vakio. Kuvassa 7.1 on kuvattu mallin sovitetut arvot ja arvojen residuaalit. Kuvio muodostaa melko tasaisen kaistaleen, joten sen perusteella ei ole syytä epäillä, että mallin vaihtelu muuttuisi sovitettujen arvojen mukaan. (Kutner 2005, s. 107.)

Toiseksi mallin virhetermin tulisi olla normaalisti jakautunut. Tätä oletusta on tutkittu kuvassa 7.2 mallin residuaalien kvantiilikuviolla. Normaalijakautuneet virhetermit jakautuvat kvantiilikuviossa tasaiseksi suoraksi. Kuvasta 7.2 nähdään, että standardoidut residuaalit muodostavat melko tasaisen suoran, jonka päissä on muutamia poikkeavia arvoja. Kuvion perusteella voidaan pitää normaalisuusoletus voimassa. (Kutner 2005, s. 110.)

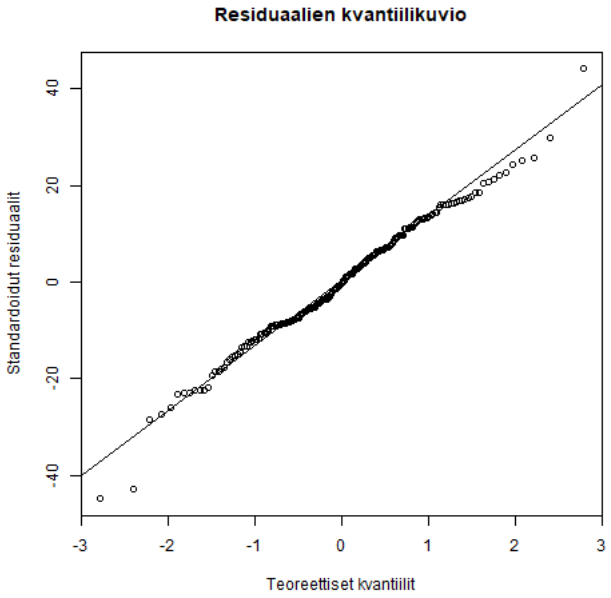
Lisäksi tarkastellaan poikkeavia arvoja. Suuresti muusta aineistosta poikkeavat arvot voivat vaikuttaa merkittävästi regressiomalliin heikentäen mallin ennustuskykyä (Kutner 2005, s. 108). Arvojen poikkeavuutta voidaan tarkastella havaintojen vipuvoimilla. Havaintojen vipuvoimat saadaan hattumatriisista

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

jonka diagonaalialkiot vastaavat havaintojen vipuvoimia. Havainnon vipuvoima kuvaa sen



Kuva 7.1. *Mallin 4 sovitetut arvot ja residuaalit*



Kuva 7.2. *Mallin 4 residuaalien kvantiilikuvio*

etäisyyttä kaikkien havaintojen keskiarvosta. Havainnot, joiden vipuvoimat ovat vähintään 0.2, luokitellaan poikkeaviksi arvoiksi. (Kutner 2005, s. 398.)

Aineiston suurin vipuvoima oli 0.07262934, joten havaintojen vipuvoimien perusteella aineistossa ei ole poikkeavia arvoja, jotka aiheuttaisivat ongelmia analyysissä.

7.3 Liite 3: Aineiston arvosanojen muunnostaulukot

Tässä liitteessä on muunnostaulukot, joiden mukaan työn aineiston arvosanat muutettiin asteikolle 0–100.

Tähteä	Muunnos
1	0
1.5	13
2	25
2.5	38
3	50
3.5	63
4	75
4.5	88
5	100

Taulukko 7.1. *1–5 tähteä*

Tähteä	Muunnos
0	0
0.5	13
1	25
1.5	38
2	50
2.5	63
3	75
3.5	88
4	100

Taulukko 7.2. *0–4 tähteä*

Pistettä	Muunnos
0	0
1	10
2	20
3	30
4	40
5	50
6	60
7	70
8	80
9	90
10	100

Taulukko 7.3. *0–10 pistettä*

Pistettä	Muunnos
-4	0
-3	13
-2	25
-1	38
0	50
1	63
2	75
3	88
4	100

Taulukko 7.4. *-4–4 pistettä*

Arvosana	Muunnos
F	0
D-	8
D	18
D+	26
C-	33
C	41
C+	50
B-	58
B	67
B+	75
A-	83
A	92
A+	100

Taulukko 7.5. *Kirjaimet F–A+*

Arvosana	Muunnos
F	0
CP	20
P	40
CR	60
D	80
HD	100

Taulukko 7.6. *Arvioija Nicole Lesleyn käyttämä skaala*

Pistettä	Muunnos
1	0
2	20
3	40
4	60
5	80
6	100

Taulukko 7.7. *Pisteet 1–6*

Pistettä	Muunnos
1	0
1.5	17
2	33
2.5	50
3	67
3.5	83
4	100

Taulukko 7.8. *Pisteet 1–4*