Opinnäytetyö (AMK)

Tietojenkäsittely

2019

Otto Kari

TIEDONLOUHINTAA SOSIAALISEN MEDIAN ENERGIAKESKUSTELUSTA

– luokittelevan mallin luonti

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

OPINNÄYTETYÖ (AMK) | TIIVISTELMÄ

TURUN AMMATTIKORKEAKOULU

Tietojenkäsittely

2019 | 27 sivua, 3 liitesivua

Otto Kari

TIEDONLOUHINTAA SOSIAALISEN MEDIAN ENERGIAKESKUSTELUSTA

- luokittelevan mallin luonti

Suomalaiset käyvät päivittäin keskustelua sosiaalisessa mediassa energiataloudesta ja energiamuodoista. Keskustelua käydään useilla eri alustoilla, mutta pääasiassa keskustelu tapahtuu aihesidonnaisilla foorumeilla.

Opinnäytetyön aiheena on kuvata suomalaisten keskustelua energiamuodoista ja tutkia, minkälaisia muuttujia ja tekijöitä liittyy viesteihin, joissa puhutaan tietystä energiamuodosta. Työssä käsiteltävä data on Futusomelta ladattu paketti, joka sisältää yksittäisiä viestejä sosiaalisesta mediasta. Datan lataus on tehty määritellyillä rajauksilla Futusomen lataustyökalua käyttäen. Datan käsittely ja työstäminen tapahtuu RStudio-nimisellä ohjelmalla.

Opinnäytetyön tavoitteena on luoda malli, joka luokittelee viestikohtaisesti, mistä energiamuodosta viestissä puhutaan. Mallin luominen edellyttää dataan tutustumista, manuaalista luokittelua ja tiedonlouhintaa.

Tutkimuksen aikana päädyttiin luomaan yhden luokittelevan mallin jokaista energiamuotoa kohden. Tämä oli yksinkertaisempi ja selvempi tapa tehdä luokittelua. Jokaisen mallin luokittelutarkkuus on erittäin korkea. Oleellisin vaikuttaja luokitteluun on ennalta määritetyt energiamuotokohtaiset sanastot. Muiden muuttujien vaikutus luokittelussa oli erittäin pieni.

ASIASANAT:

energiakeskustelu, tiedonlouhinta, luokittelevamalli, Futusome

BACHELOR´S / MASTER'S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Information Technology

2019 | 27 pages, 3 pages in appendices

Otto Kari

DATA MINING OF ENERGY CONVERSATION IN SOCIAL MEDIA

- making a classifying model

Finnish people participate daily in discussions about power production and electricity generation on social media. Discussions take place on different sites and platforms, but mostly the discussions happens on forums directly linked in the topic.

This thesis is made to describe Finnish discussion about the ways of power production and examine what kind of variables occur related to different ways of power production. Data that is used in this research is downloaded from Futusome and it holds individual messages from social media. The download of the data is done by adding some filters and terms to Futusomes downloading tool. Data processing is done with RStudio software.

The goal is to create classifying model that can classify the power production form that is mentioned in the message. This requires exploring the data, manual classification and data mining.

While working with the research, I decided to make one classifying model for every power production forms. This was simpler way to work the research. Every models classifying accuracy was very high. The main factor for the high accuracy was beforehand made power production form specified dictionaries. Other factors were not so effective for the classifying models.

KEYWORDS:

power production, classifying model, datamining

SISÄLTÖ

LIITTEET

Liite 3. EV-ydinvoima-RandomForest.R

KUVAT

TAULUKOT

KÄYTETTY SANASTO

csv	comma-seperated values (Shafranovich, 2005)

1 JOHDANTO

Tutkimuksen tarkoitus on tutkia, voidaanko energiamuotoihin liittyviä viestejä luokitella automattisesti energiamuodoittain. Olemassa olevan datan, ennestään määriteltyjen sanastojen ja uuden louhitun datan avulla energiamuotokohtaiset mallit yrittävät löytää eri energiamuotoja käsittelevistä viesteistä piirteitä, joiden avulla viestit voidaan automaattisesti luokitella tiettyä energiamuotoa koskevaksi viestiksi. Tutkin myös kuinka luotettavaa luokittelu on ja minkälaisin perustein luokittelu tapahtuu.

Työn toimeksi antaja on DEEVA-hanke. DEEVA-hanke aloitettiin 2016 syksyllä digitalisaation ja datan hyödyntämisen kehittämiseksi. Tarkoituksena on luoda datasta arvoa. Hankkeessa on mukana laaja ja monipuolinen yritysverkosto. Mukana on 20 erikokoista yritystä eri toimialoilta kuten energia, pankki-, media-, ICT-, kauppa- ja palvelualalta. Hankkeen toteuttajatiimin kuuluu Tampereen teknillinen yliopisto, Turun ja Tampereen ammattikorkeakoulut sekä kuusi muuta kansainvälistä yhteistyökorkeakoulua.

”Hankkeessa syntyy eri kohderyhmille suunnattujen julkaisujen lisäksi asiakaskokemuksen mittareita ja tunnedatan reaaliaikaisia analyysityökaluja ja sovelluksia, joiden tuottama tieto saadaan osaksi yritysten arkipäivästä toimintaa tukemaan mm. monikanavaisen palveluympäristön hallintaa, uudenlaisten palvelutuotteiden syntymistä sekä yhteisluontiin perustuvien ekosysteemien kehittämistä.” (DEEVA-Hanke, 2016)

Työssä tutustun dataan ja käyn sitä manuaalisesti läpi. Käydessäni dataa läpi luokittelen satunnaistettuja viestejä. Jos viesti viittaa johonkin energiamuotoon, luokittelen viestin liittyvän tähän energiamuotoon tai useampaan energiamuotoon.

Manuaalisen luokittelun jälkeen on aika tarkistella luokittelun tuloksia. Piirrän kaavioita jakaumista, tutkin löytyykö tiettyyn energiamuotoon jotain merkittäviä huomioita ja yleisesti energiamuotoihin liittyvien viestien erovaisuuksia. Tärkeää pohdittavaa on, minkälaisista muuttujista voisi olla apua luokittelevalle mallille.

Mallin luontia varten on satunnaistettu viestejä mallin koulutusta varten ja pienempi joukko mallin testaamista varten. Malli tulee käyttämään hyödyksi sanastoja ja muuttujia.

2 DATA

Työssäni käyttämä data on peräisin Futusome osakeyhtiöltä. Data on ladattu DEEVA-hankkeen käyttöön ja sen toimeksiantamaan tutkimukseen, jossa tutkitaan trolleja energiakeskustelussa. Datan lataamista varten luotiin sanasto, joka sisälsi paljon erilaisia sanoja liittyen energiaan. Latausta tehdessä Futusomen lataus-työkalulla, määrittelimme aikavälin jolta halusimme saada dataa. Datan viestit ovat vuoden 2017 tammikuulta, huhtikuulta, heinäkuulta ja lokakuulta.

2.1 Datan rakenne

Data koostuu yksittäistä viesteistä ja niihin liitetyistä tiedoista ja arvoista. Data on ladattuna csv-muodossa. Yksittäiset viestit ovat riveittäin ja sarakkeissa on otsakkeet. Data ladattiin Futusomelta alun perin toiseen DEEVA-Hankkeen projektiin. Alun perin ladattuja viestejä oli lähemmäs 500 000, joista jokainen viesti sisälsi 91 saraketta erilaista metadataa. Työskentelin myös tässä projektissa, jossa tutkimme trollien esiintymistä energiakeskustelussa. Rikastimme dataa uusilla muuttujilla ja rajasimme turhia muuttujia pois. Käsittelyymme projektissa otimme 6 052 viestiä. Tämän 6 052 viestin otoksen sain käyttööni tähän opinnäytetyöhön.

Alun perin ladatuissa viesteissä, viesti-tyyppejä oli useita. Tyyppejä oli uutiskommentit, facebook-viestit, twitter-viestit, foorumi-viestit ja blogikirjoitukset. Mutta tässä 6 052 joukosta viestityypejä oli vain facebook, foorumi-viesti ja uutis kommentti. Ja suurin osa näistä viesteistä oli foorumeilta.

Kuvasta 1 nähdään käyttämäni datan sarakkeiden nimet. Tulen lisäämään uusia muuttujia eli sarakkeita työn aikana dataan. Sarakkeet ovat nimetty mahdollisimman ymmärrettävästi. Osa sarakkeista on alkuperäisiä Futusomen nimeämiä sarakkeita ja osa on jälkeenpäin lisättyjä sarakkeita. Sarakkeet kuten ”type” kertoo viestin tyypistä, ”author” on viestinkirjoittajan nimimerkki, ”citation” jos viestissä on viitattu keskusteluketjussa johonkin aikaisempaan viestiin, ”text” on itse viesti ja ”text.url” on lista linkeistä, joita viesti sisältää.

```
> names(fs.energiaviestit.df)
 [1] "doc.id"                      "random.number"               "common.term.text.count"
 [4] "common.term.title.count"     "troll.label"                 "text"
 [7] "text.address.size"           "indexed"                     "type"
[10] "author.text"                 "forum_post_id"               "url"
[13] "thread.title"                "version"                     "author"
[16] "latency"                     "page.title"                  "text.person.size"
[19] "language"                    "published"                   "text.email.size"
[22] "author.following"            "author.followers"            "text.url.length"
[25] "text.url"                    "quote"                       "text.person"
[28] "text.email.user_mention"     "text.user_mention"           "text.hashtag"
[31] "citation"                    "ref.author"                  "citation.length"
[34] "text.phone.canonized"        "name"                        "ref.author.id_facebook"
[37] "author.id_facebook"          "facebook_id"                 "ref.author.uri"
[40] "author.community_facebook"   "description"                 "caption"
[43] "if.base"                     "if.pos"                      "ydinvoima.word.count"
[46] "has.ydinvoima.words"         "tuulivoima.word.count"       "has.tuulivoima.words"
[49] "aurinkoenergia.word.count"   "has.aurinkoenergia.words"    "muu.word.count"
[52] "has.muu.words"               "has.anyEnergyDictionary.words" "huomio"
[55] "ydinvoima_manuaali"          "aurinkoenergia_manuaali"     "tuulivoima_manuaali"
[58] "muu_manuaali"
```

Kuva 1. Datan sarakkeiden nimet.

2.2 Sanastot

Tein jokaiselle neljälle luokalle oman sanastonsa. Sanastot eivät ole mitenkään valtavia ja missään sanastossa ei ole samoja sanoja. Käytin sanastoissa vain sanoja, jotka viittaavat lähes poikkeuksetta keskusteluun jostakin tietystä energiamuodosta. Esimerkiksi ydinvoima sanastosta löytyy sanoja, kuten ydinjäte, olkiluoto, ydinsaaste ja tsernobyl. R-scripti lisää datan sarakkeisiin arvon, jos viesti sisältää tietystä sanastosta jonkin sanan ja kuinka monta tämän sanaston sanaa viestistä löytyy.

Energia sanastojen lisäksi luodaan simppeleitä sanastoja, joita R-scriptillä hyödynnetään lisäämään muuttujia. Tein sanastoja kuten, poliitikot-, poliittiset termit-, kirosana-, haukkumasana-, promotus- ja yritysnimi-sanasto. Kuvasta 2 nähdään koodia, jossa hyödynnetään csv-tiedostossa olevaa poliitikkosanastoa ja sen avulla luodaan kaksi uutta muuttujaa ”politicians.count” ja ”politicians.mentioned”. Tämä siis kertoo mainitaanko viestissä politiikkoja ja jos mainitaan niin kuinka monesti.

```r
# Make data frame politicians from politicians.csv
politicians.df <- read.table(file = 'data/dictionaries/politicians.csv',
                             header = TRUE,
                             sep = ',',
                             fill = TRUE,
                             quote = "\"",
                             stringsAsFactors = FALSE,
                             comment.char = ''
)

# Make vector from politicians.df
politicians <- as.vector(as.matrix(politicians.df))

# Remove the empty strings and duplicates from polticians vector
politicians <- unique(politicians[politicians != ''])

# View the vector
politicians

# Create a regular expression string of the politicians
politicians.regex <- paste(politicians, sep="", collapse="|")

# Create variable politicians.count
fs.luokitellut.energiaviestit.df$politicians.count
  <- str_count(tolower(fs.luokitellut.energiaviestit.df$text), politicians.regex)

# Create a variable politicians.mentioned
fs.luokitellut.energiaviestit.df$politicians.mentioned
  <- ifelse(fs.luokitellut.energiaviestit.df$politicians.count > 0, TRUE, FALSE)
```

Kuva 2. Esimerkki muuttujien luomisesta sanaston avulla.

3 MANUAALINEN LUOKITTELU

Manuaalinen luokittelu työssäni tapahtui csv-tiedostoon. Data oli ajettu RStudioon, jossa lisäsin jokaiselle viestille satunnaisarvo set.seed-funktiolla. Tämän jälkeen lisäsin dataan sarakkeet, joihin tehtiin manuaaliset merkinnät. Manuaaliseen luokitteluun valitsin kolmasosa viesteistä. Satunnaisarvo viesteillä on 0,1- 0,99. Manuaaliseen luokitteluun tarkoitetut viestit olivat rajattu seuraavalla ehdolla: satunnaisarvo on yhtä kuin tai suurempi kuin 0,52 ja on yhtä suuri tai pienempi kuin 0,85. Tällä ehdolla viesteistä saatiin vähän yli kolmasosa, joka on 2 046 viestiä.

Manuaalinen luokittelu vaatii tarkat ehdot ja niiden tarkkaa noudattamista. Määritin ehdot luokittelulle ja aloin käymään viestejä manuaalisesti läpi. Eli luin viestin ja merkitsen tarvittaviin sarakkeisiin arvot. Joissain tapauksissa tutkin viestiä vähän pidemmälle, kuten avaamalla viestin keskusteluketjun tarvittaessa tai tarkistan keskusteluketjun otsikon.

Luokittelu vei paljon aikaa, sillä tutkin viestit erittäin tarkkaan. Osassa viesteissä oli linkkejä, jotka täytyi avata saadakseen lisää informaatiota viestistä tai täytyi lukea keskusteluketjua ymmärtääkseen mihin viestillä vastataan. Jokaisessa viestissä ei kuitenkaan tarvinnut tehdä näin tarkkaa tutkintaa. Ainoastaan kun se vaikutti viestin epäsuorasti viittaavan tai kertovan jostain energiamuodosta.

3.1 Energiamuodot

Energiamuodot voidaan rajata kahteen ryhmään, uusiutumattomat- ja uusiutuvat energiamuodot. Uusiutumattomia ovat: kivihiili, ydinvoima, maakaasu, öljy ja turve. Uusiutuvia ovat aurinkoenergia, tuulienergia, Vesivoima, bioenergia, puun poltto, maa- ja ilmalämpö.

3.2 Luokat

Manuaalisessa luokittelussa tein merkintöjä neljään eri sarakkeeseen. En ottanut jokaista energiamuotoa omaksi luokaksi, sillä muuten työtä olisi ollut luokitellessa valtavasti. Data oli jo aikaisemmin tuttua aikaisemmasta Trolli-projektista ja olin käynyt tehnyt

siihen jo ennestään manuaalista luokittelua. Päätin lisätä dataan 5 saraketta. Ydinvoimalle, tuulivoimalle, aurinkoenergialle ja muille. Muu luokkaan sisältyy viestit, joissa esimerkiksi puhutaan kivihiilestä, kaasusta, turpeesta, vesivoimasta tai bioenergiasta. Taulukosta 1 nähdään, miten sarakkeet ovat nimetty. Näiden neljän luokan lisäksi, lisäsin sarakkeen huomio. Tähän sarakkeeseen kirjoitan tarvittaessa, jonkin huomioinnin viestistä.

Taulukko 2. Manuaaliseen luokitteluun luodut sarakkeet.

sarakkeiden nimet
ydinvoima_manuaali
tuulivoima_manuaali
aurinkoenergia_manuaali
muu_manuaali
huomio

3.3 Luokittelun ehdot ja rajaukset

Sarakkeiden arvo on vakiona 0. Merkintä sarakkeisiin tulee, jos viestissä mainitaan jokin energiamuototoon liittyvä termi, tai jos viestissä olevasta linkistä löytyy energiamuoto, voidaan olettaa viestissä viitattavan epäsuorasti johonkin energiamuotoon ketjun aiheen perusteella tai käytetään jotain muuta energiamuotoon yhdistettävää termiä. Merkitään sarakkeeseen arvoksi 1, jos jokin yllämainituista ehdoista toteutuu. Eli vaikka viestissä ei mainittaisi ydinvoimaan liittyviä termejä, mutta viestissä käsitellään selvästi ydinvoimaa, niin annan tälle manuaaliseksi arvoksi 1. En ota huomioon termiä ydinvoimapalsta, sillä joskus viesteissä puhutaan tästä palstasta, vaikka viesti ei millään liittyisi ydinvoimaan. Myös sana hiili ja öljy löytyy useasta viestistä, mutta näillä ei ole tekemistä energiamuodon kanssa. Esimerkiksi, kun puhutaan auton öljyjen vaihdosta tai öljysheikin palatsista.

3.4 Luokittelun tarkoitus

Luokittelun jälkeen on mahdollista luoda eri energiamuotoihin liittyviä kuvaajia ja tulkita dataa monipuolisemmin. Esimerkiksi missä sosiaalisen median palvelussa puhutaan eniten tuulivoimasta tai minkä energiamuodon viesteistä löytyy eniten kirosanoja tai merkkien keskiarvopituus on isoin. Näitä tekijöitä ja tietoja vertaillaan, tutkitaan ja hyödynnetään luokittelussa.

3.5 Luokittelun huomioita

Yhdellä palstoista esiintyy nimeltä mainitsematon henkilö, joka terrorisoi palstaa. Datanikin saattaa vääristyä tämän henkilön viesteistä. Hän lähettää jatkuvasti samoja ydinvoimavastaisia massaviestejä. Osa on luokiteltu ydinvoimaviestiksi, jos viestistä on löytynyt jotain uutta tai selkeästi itse kirjoitettua. Tämän käyttäjän massaviesteihin ole merkinnyt huomio sarakkeeseen sen olevan duplikaatti, jos sama teksti on esiintynyt jo aikaisemmin ja jättänyt tämän luokittelematta.

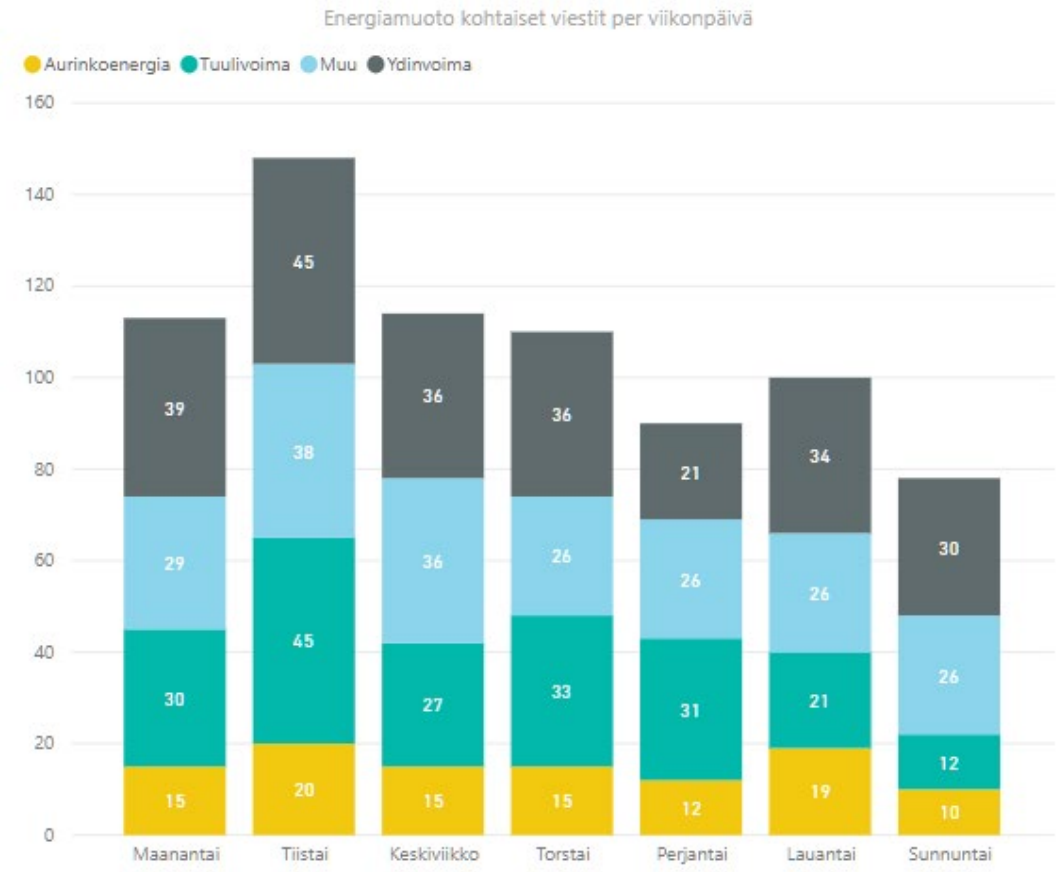
Datasta löytyy myös paljon viestejä liittyen autoihin, ilmastoon, kuntoiluun ja ravintoon. Osa viesteistä on myös mainoksia, kampanjointia ja arvontoja. Tästä tarvittaessa tehdään sanastoa hyödyntävä muuttuja, jota luokitteleva malli voi hyödyntää. Viesteissä esiintyy paljon mainontaa ja arvontoja liittyen aurinkopaneeleihin.

3.6 Luokittelun lopputulos

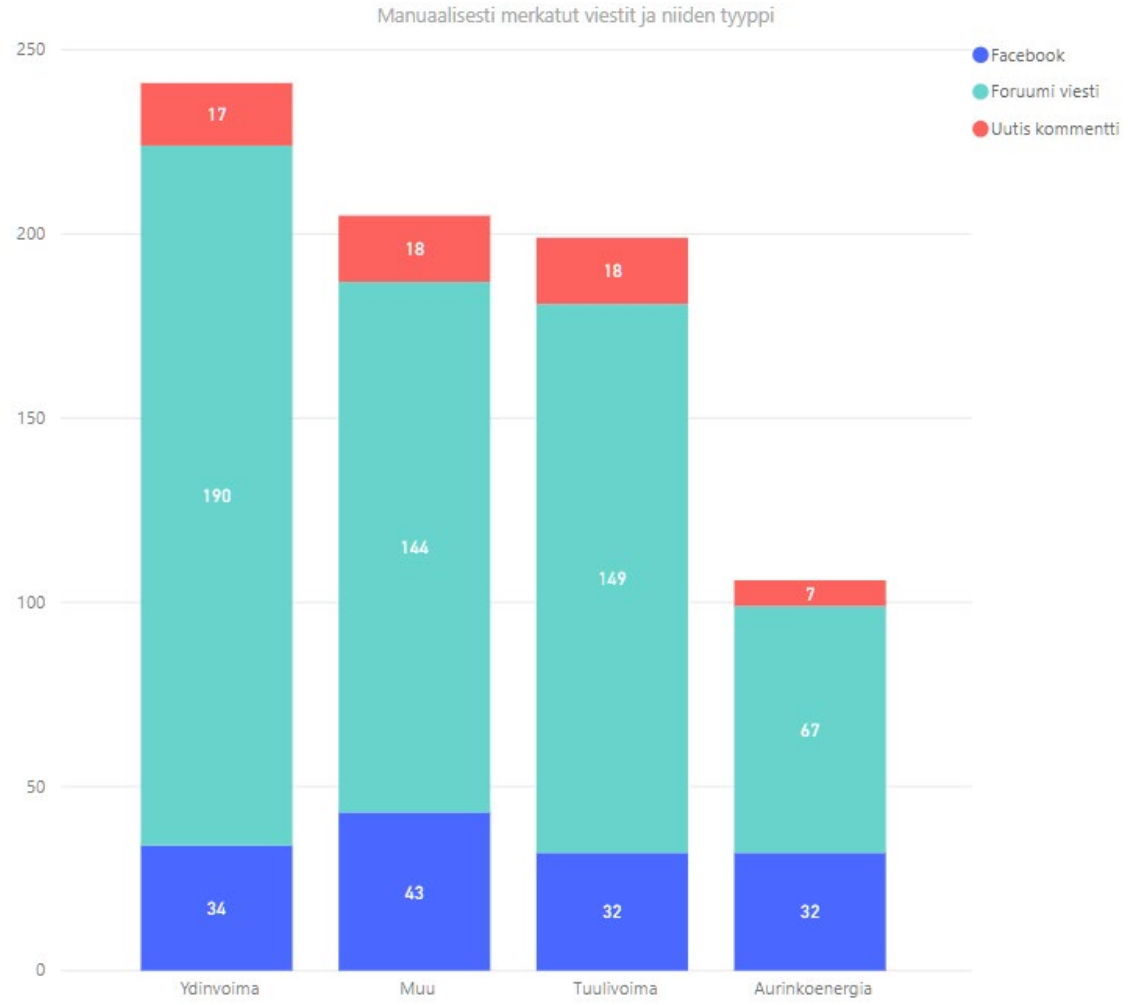
Viestien luokittelun jälkeen tutkitaan muuttujia ja niiden yhtenäisyyksiä eri energiamuotoihin. Otetaan tarkkailuun seuraavan laisia arvoja, kuten viestityypit, viikonpäivät, onko viesti kirjoitettu työtunteina tai viikonloppuna, sisältääkö viesti poliittisia termejä, kiroilua tai haukkumasanoja. Tarkoituksena on tutkia, löytyykö näistä huomattavia eroavaisuuksia energiakohtaisesti.

Kiroilu ja haukkumasanat olivat täysin suhteessa viestien energiakohtaisten viestien määrään. Kuvasta 3 nähdään, että viikonpäivät olivat myös suhteessa viestien määrään; tiistaisin kirjoitettiin eniten viestejä kaikista energiamuodoista, perjantaisin ja sunnuntaisin vähiten. Viestityypit jakautuivat myös tasaisesti. Kaikilla alustoilla käytiin keskustelua

jokaisesta energiamuodosta lähes yhtä paljon. Kuvasta 4 nähdään että, eniten energiakeskustelua käydään foorumeilla. Suurin osa viesteistä löytyy suomi24.fi/tiede/ foorumilta.



Kuva 3. Energiamuotokohtaiset viestit viikonpäivää kohti (Power BI, 2019).



Kuva 4. Manuaalisesti merkatut viestit ja niiden tyypit (Power BI, 2019).

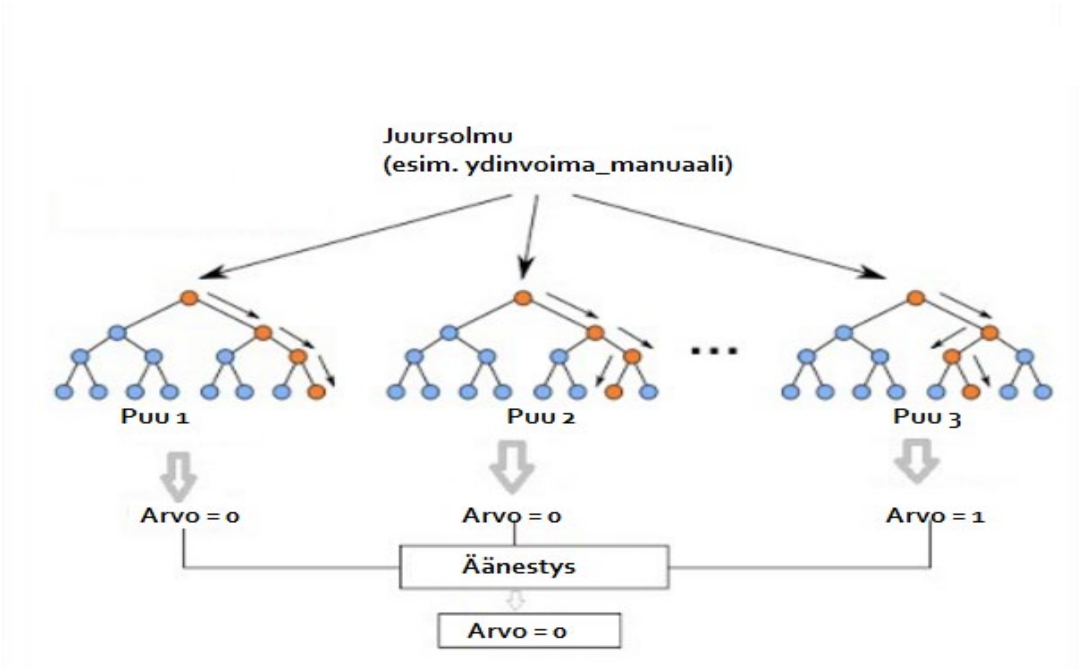
Erittäin oleellisiksi muuttujiksi ilmeni ennalta luotujen energiakohtaisten-sanastojen avulla määräytyvät muuttujat. Nämä muuttujat olivat lähes kaikissa viesteissä saman arvoisia, kuin manuaaliset luokittelut.

4 LUOKITTELEVA MALLI

4.1 Päätöspuut ja Random forest

Päätöspuuhun (*eng. decision tree*) kuuluu kohdemuuttuja, jota kutsutaan juurisolmuksi (*eng. root node*). Puu alkaa kasvamaan ja haarautumaan juurisolmusta alaspäin. Puu haarautuu kahteen tai useampaan haaraan. Haaralle annetaan, jokin oleellinen muuttuja ja se haarautuu tämän muuttujan mahdollisiin arvoihin. Osa haaroista vaikuttaa suoraan lopputulokseen ja osa haaroista vaikuttaa tuleviin mahdollisiin haaroihin. Päätöspuun ideana on tutkia miten eri muuttujat vaikuttavat tutkittavaan arvoon eli juurisolmuun. (Yiu, 2019)

Random Forest koostuu useasta päätöspuusta. Kuva 5 on yksinkertaistettu kuva Random Forestin rakenteesta ja ideasta. Sille annetaan juurisolmu ja se luo paljon puita samalla juurisolmulla ja hakee haaroihin satunnaisia muuttujia, joilla ei ole suoraa yhteyttä toisiin muuttujiin. Jokainen puu yrittää olla mahdollisimman erilainen ja yrittää löytää erilaisia muuttujien vaikutuksia. Puiden rakentamisen jälkeen, puut vertailevat tuloksiaan ja ikään kun äänestävät tuloksien perusteella oleelliset vaikuttavat tekijät. Luonnollisesti useamman puun tulokset ja vertailu, tuo tarkemman lopputuloksen kuin yksittäinen puu. (Yiu, 2019)



Kuva 5. Random Forest yksinkertaistettuna.

Päätöspuu vaatii koulutusjoukon ja testijoukon. Malli kouluttaa itseään ja tutkii koulutusjoukkoa ja taas testi joukolla on tarkoitus testata mallin tarkkuutta. Koulutusjoukko on erittäin oleellinen päätöspuulle. Se vaikuttaa päätöspuun käyttäytymiseen ja tarkkuuteen. Pienetkin muutokset koulutusjoukossa, saattaa muuttaa puun rakennetta paljon. (Yiu, 2019) Myös vahvat luokittelutulokset koulutus joukon kanssa ei takaa tarkkaa tulosta testijoukon kanssa.

Random Forest koneoppimisessa perustuu ohjattuun oppimiseen (*eng. supervised learning*). Ohjatussa oppimisessa hyödynnetään esimerkkiaineistoa, johon on mahdollisesti tehty luokitteluja ja havaintoja. (James, et al., 2013) Tässä työssä luomani koulutusjoukko sisältää manuaalisen luokittelun merkinnät ja malli koittaa näiden luokittelun tuloksista oppia ja käyttää jonkin näiköisenä pohjana.

4.2 Muuttujat

Käsiteltävässä datassa on Futusomelta peräisin olevia muuttujia, sekä näiden muuttujien avulla luotuja muuttujia ja manuaaliseen luokitteluun luotuja muuttujia. Muuttujia on yhteensä 57 ja viestejä on 2046. Kaikkia viestejä ja muuttujia ei kuitenkaan Random Fo-

restia ajaessa käytetä. Viestit ovat jaettu koulutusjoukkoon ja testijoukkoon. Osalla muuttujista on liian usea mahdollinen eri uniikki arvo, kuten esimerkiksi: linkki, julkaisupäivä ja ketjun otsikko. RStudion Random Forest ei pysty suoriutumaan sellaisilla muuttujilla joilla voi olla yli 53 uniikkia arvoa. Kuvasta 6 voidaan nähdä muuttujat, joita Random Forest hyödyntää. Datan määrä on kuitenkin tutkimuksessa niin pieni, että se ei voi kaikkia muuttujia hyödyntää.

```
> names(ranfor.train.df)
 [1] "has.ydinvoima.words"          "ydinvoima_manuaali"
 [3] "has.tuulivoima.words"         "tuulivoima_manuaali"
 [5] "has.aurinkoenergia.words"     "aurinkoenergia_manuaali"
 [7] "has.muu.words"                "muu_manuaali"
 [9] "has.anyEnergyDictionary.words" "common.term.text.count"
[11] "common.term.title.count"      "day.of.week"
[13] "is.weekend"                   "is.working.hours"
[15] "is.even.hour"                 "is.even.minute"
[17] "is.even.hour.and.even.minute" "type"
[19] "ydinvoima.word.count"         "tuulivoima.word.count"
[21] "aurinkoenergia.word.count"    "muu.word.count"
[23] "word.count"                   "swear.word.count"
[25] "has.swear.words"              "promo.word.count"
[27] "has.promo.words"              "hashtag.count"
[29] "abuse.word.count"             "has.abuse.words"
[31] "bad.link.count"               "has.bad.link"
[33] "is.bad.link"                  "politicians.count"
[35] "politicians.mentioned"        "political_words.count"
[37] "has.political_words"          "company_names.count"
[39] "has.company_names"            "has.question.mark"
[41] "question.mark.count"          "has.exclamation.mark"
[43] "exclamation.mark.count"       "pred"
```

Kuva 6. Random Forestin hyödyntämät muuttujat.

4.3 Mallin koulutus ja testaus

Alun perin oli tarkoitus luoda yksi malli, joka luokittelee viestistä keskusteltavan energiamuodon, mutta päädyinkin tekemään jokaiselle energiamuodolle oman Random Forest

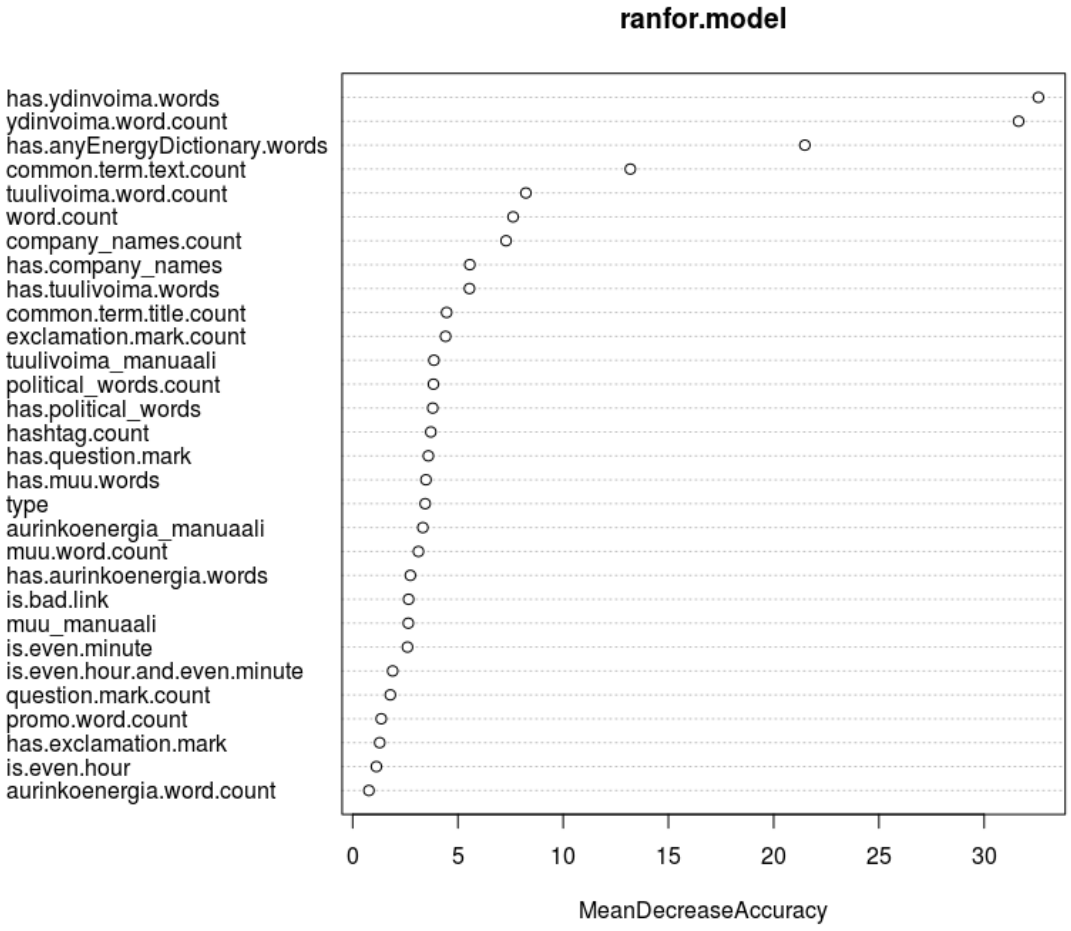
mallin. Tämä helpottaa mallin luontia ja pitää mallin yksinkertaisempana. Energiakohtaiset mallit siis tutkii mikä vaikuttaa siihen, että manuaalisesti on luokiteltu sarakkeeseen arvo 1 ydinvoimalle. Sama tekee muutkin mallit tuulivoimalle, aurinkoenergialle ja muuluokalle.

Myös tarkoituksenani oli ottaa koulutusjoukkoon 1 400 luokiteltua viestiä ja testijoukkoon 200 luokiteltua viestiä. Testijoukon ollessa 200 viestiä, jää koulutusjoukkoon 1 846 viestiä. Tästä koulutusjoukon 1 846 viestistä on luokiteltu sisältävän ydinvoimaa 213 viestiä, eli muuttujalla ”ydinvoima_manuaali” on arvona 1. Koulutus joukon on hyvä olla tasapainossa siten, että siellä mahdollisimman tasaisesti viestejä kaikilla mahdollisilla eri arvoilla. Tässä kohtaa, kun tutkitaan ydinvoimaa niin arvoja on kaksi mahdollista arvoa. Arvo on vakiona 0 ja jos viesti on sisältänyt jotain ydinvoimaan viittaavaa, on sille annettu arvoksi 1. Joten kun ydinvoima viestejä on luokiteltu manuaalisesti 213, niin koulutusjoukon tasapainon vuoksi otetaan sinne 213 satunnaista luokiteltua viestiä, joiden ydinvoima arvo on 0. Eli koulutusjoukko ydinvoima mallissa on 426 viestiä. Koulutus joukosta saisi isomman esimerkiksi monistamalla osan viesteistä, joilla ydinvoima arvo on 1 ja siten ottamalla myös enemmän 0 arvoisia viestejä, mutta tämä rajattiin tällä kertaa analyysin ulkopuolelle, joten tyydyn 426-viestin koulutusjoukkoon. Tuulivoiman ja aurinkoenergian koulutus joukot tulevat olemaan vielä pienempiä, sillä manuaalisia luokitteluja tuli eniten ydinvoimaan ja muu-luokkaan.

5 MALLIEN LUOTETTAVUUDEN ARVIOINTI

5.1 Ydinvoiman luokittelu

Mallien hyvä luokittelutarkkuus perustuu energiakohtaisiin sanastoihin. Kuvasta 7 voidaan nähdä algoritmin arvioita muuttujien vaikutuksesta ydinvoiman luokitteluun. Painavin peruste ydinvoimaksi luokitelluissa viesteissä on selvästi muuttujien ”has.ydinvoima.words” ja ”ydinvoima.word.count” arvot. Yllättäviä muuttujia ei esiinny tässä analyysissa tai niiden paino luokittelussa ei ole niin korkea. Toki tuulivoima sanoja sisältävät viestit ja myös yritysnimet ovat listattu jokseenkin korkealle ja näillä on jonkin verran vaikutusta ydinvoimaksi luokittelussa.



Kuva 7. Muuttujien vaikutus ydinvoima-luokan mallissa.

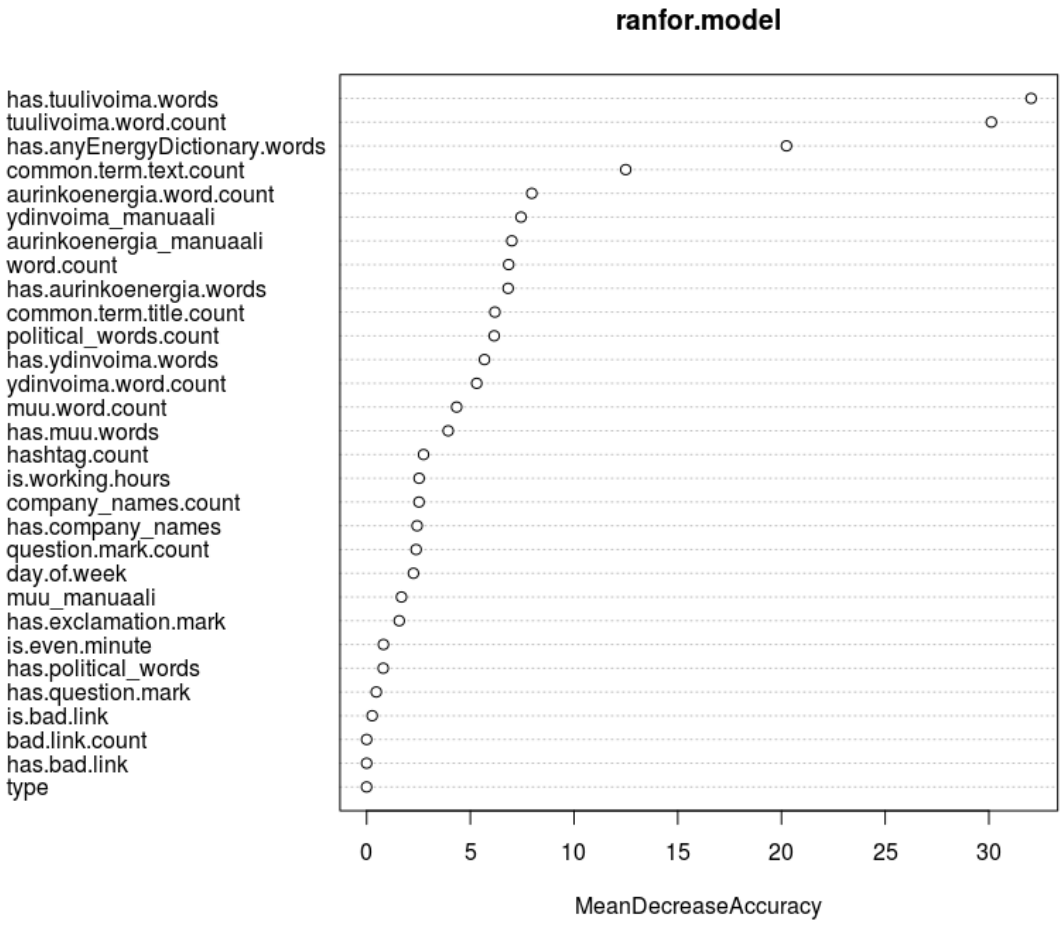
Ydinvoimaksi luokittelevan mallin koulutusjoukko koostui 426 viestistä ja testijoukko oli 200 viestiä. Koulutusjoukko ja testijoukko koostuu manuaalisesti luokitelluista viesteistä. Koulutusjoukon luokittelutarkkuus oli 96%. Kuvasta 8 nähdään testijoukon luokittelutarkkuus (accuracy) 97% ja testijoukon 200 viestistä 28 viestiä oli manuaalisesti luokiteltu liittyvän ydinvoimaan antamalla ”ydinvoima_manuaali”(reference) arvoksi 1. Loppujen 172 viestin arvo oli 0. Malli luokitteli (prediction) 32 viestille arvoksi 1 ja 168 viestille arvoksi 0. Eli testijoukon 200 viestistä 4 viestiä luokiteltiin väärin.

```
> # Print confusion matrix and some statistics
> print(ranfor.test.model.performance)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 167   1
         1   5  27

               Accuracy : 0.97
                 95% CI : (0.9358, 0.9889)
    No Information Rate : 0.86
    P-Value [Acc > NIR] : 1.489e-07

                  Kappa : 0.8824

 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.9643
            Specificity : 0.9709
         Pos Pred Value : 0.8437
         Neg Pred Value : 0.9940
             Prevalence : 0.1400
         Detection Rate : 0.1350
   Detection Prevalence : 0.1600
      Balanced Accuracy : 0.9676

       'Positive' Class : 1
```

Kuva 8. Ydinvoima mallin testijoukon konfuusiomatriisi.

5.2 Tuulivoiman luokittelu

Tuulivoimaksi luokittelevan mallin tulokset olivat pitkälti samanlaiset kuin ydinvoimalla. Kuvasta 9 nähdään, miten muuttujien vaikutus arvioidaan samankaltaisesti kuin ydinvoima mallissa. Tällä kertaa muuttuja ”aurinkoenergia.word.count” on arvioitu vaikuttavan, mutta myös muuttuja ”has.ydinvoima.words” vaikuttaa luokitteluun.



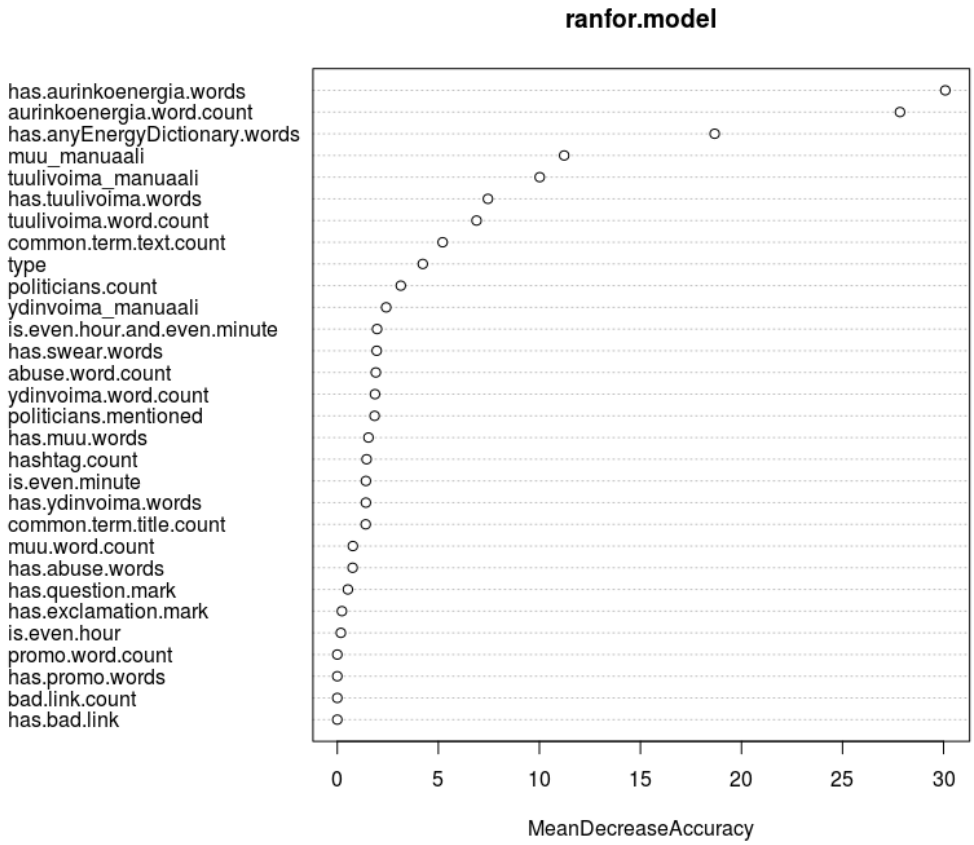
Kuva 9. Muuttujien vaikutus tuulivoima-luokan mallissa.

Tuulivoimaksi luokittelevan mallin koulutusjoukko koostui 356 viestistä ja testijoukko oli tässäkin 200 viestiä. Koulutusjoukon luokittelutarkkuus oli 96% ja testijoukon 97%. Eli täsmälleen samat tarkkuudet kuin ydinvoima tapauksessa. Testijoukon 200 viestistä 21 viestiä oli luokitettu liittyvän tuulivoimaan antamalla ”tuulivoima_manuaali” arvoksi 1. Loppujen 179 viestin arvo oli 0. Malli luokitteli 22 viestille arvoksi 1 ja 178 viestille arvoksi 0. Eli testijoukon 200 viestistä 2 viestiä luokiteltiin väärin.

5.3 Aurinkoenergian luokittelu

Aurinkoenergiaksi luokittelevan mallin tulokset poikkeavat kahdesta aiemmasta luokittelevasta mallista. Kuvasta 10 nähdään, miten eniten vaikuttavien muuttujien painoarvo on

lähes sama kuin aiemmissa, mutta myös viestin tyyppi, politikkojen mainitseminen ja kirosanojen sisältäminen vaikuttaa tämän mallin luokitteluun.



Kuva 10. Muuttujien vaikutus aurinkoenergia-luokan mallissa.

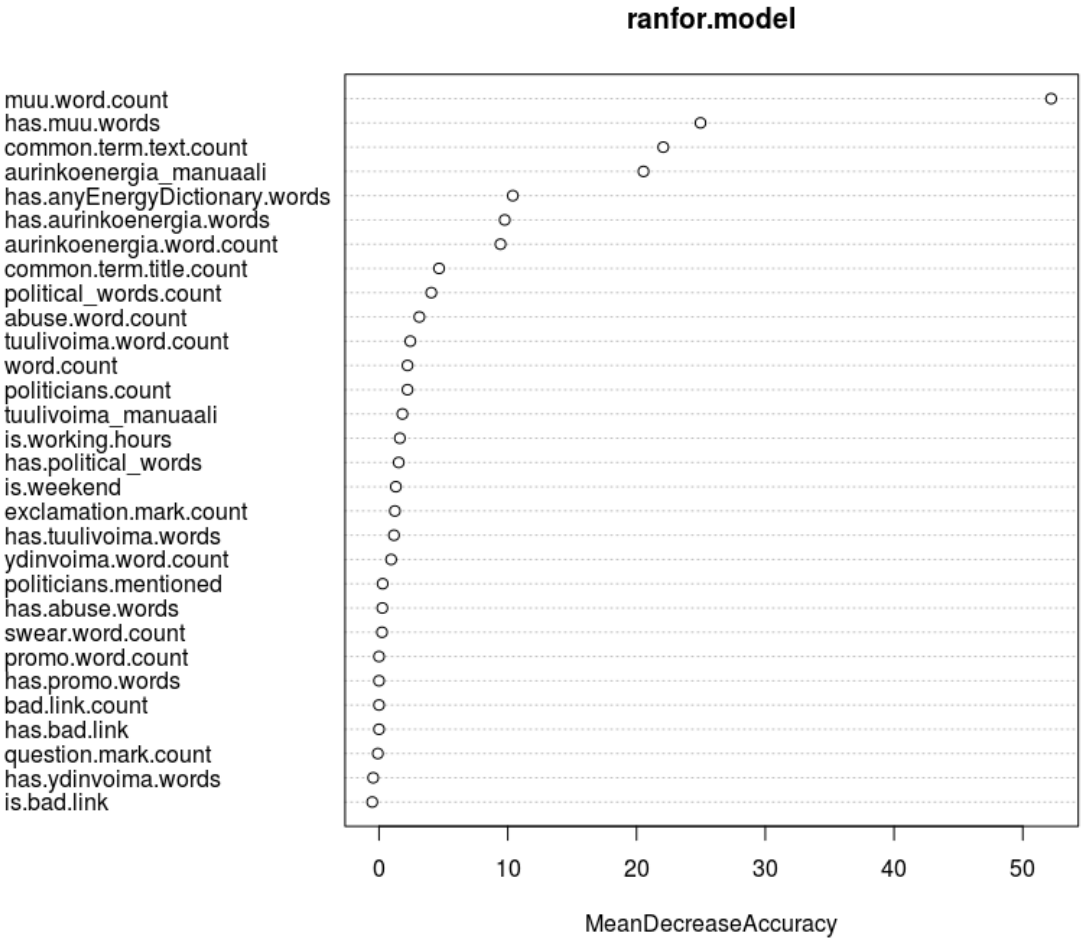
Aurinkoenergiaksi luokittelevan mallin koulutusjoukko koostui 196 viestistä ja testijoukko oli tässäkin 200 viestiä. Koulutusjoukko jäi tässä mallissa kaikista pienimmäksi, sillä manuaalisesti aurinkoenergiaksi luokiteltuja viestejä oli vähiten ja koulutus joukossa on oltava yhtä paljon viestejä arvolla 1 ja arvolla 0. Koulutusjoukon luokittelutarkkuus oli 94%. Kuvasta 11 nähdään testijoukon luokittelutarkkuus 99%. Testijoukon 200 viestistä 11 viestiä oli luokitettu liittyvän aurinkoenergiaan antamalla ”aurinkoenergia_manuaali” arvoksi 1. Loppujen 189 viestin arvo oli 0. Malli luokitteli 12 viestille arvoksi 1 ja 188 viestille arvoksi 0. Eli testijoukon 200 viestistä 1 viestiä luokiteltiin väärin. Tässä mallissa koulutusjoukon luokittelutarkkuus oli pienempi kuin aiemmissa, mutta testijoukon luokittelutarkkuus oli tarkin.

```
> # Print confusion matrix and some statistics
> print(ranfor.test.model.performance)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 188   0
         1   1  11

               Accuracy : 0.995
                 95% CI : (0.9725, 0.9999)
    No Information Rate : 0.945
    P-Value [Acc > NIR] : 0.0001542

                  Kappa : 0.9539

 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 1.0000
            Specificity : 0.9947
         Pos Pred Value : 0.9167
         Neg Pred Value : 1.0000
             Prevalence : 0.0550
         Detection Rate : 0.0550
   Detection Prevalence : 0.0600
      Balanced Accuracy : 0.9974

       'Positive' Class : 1
```

Kuva 11. Aurinkoenergia mallin testijoukon konfuusiomatriisi.

5.4 Muiden energiamuotojen luokittelu

Viimeisen mallin eli muu luokkaan luokittelevan mallin tarkkuus oli heikoin. Tämä johtuu siitä, että sanasto on isompi ja sisältää sanoja, jotka voivat esiintyä muissakin tapauksissa kuin energiakeskustelussa. Kuvasta 12 nähdään kuitenkin, että eniten vaikuttavat muuttujat esiintyvät samalla tavalla kuin aiemmissakin malleissa. Mielenkiintoisimmat vaikuttavat tekijät ovat muuttujat jotka kertoo viestin sisältävän poliittisia sanoja, haukkumasanoja ja aurinkoenergia sanoja.



Kuva 12. Muuttujien vaikutus muu-luokan mallissa.

Muu luokkaan luokittelevan mallin koulutusjoukko koostui 368 viestistä ja testijoukko oli taas 200 viestiä. Koulutusjoukon luokittelutarkkuus oli 88% ja testijoukon 86%. Eli selvästi huonoimmat luokittelutarkkuudet luomistani malleista. Testijoukon 200 viestistä 23 viestiä oli luokitettu liittyvän johonkin muuhun energiamuotoon ”muu_manuaali” arvoksi 1. Loppujen 177 viestin arvo oli 0. Malli luokitteli 49 viestille arvoksi 1 ja 151 viestille arvoksi 0. Eli testijoukon 200 viestistä 26 viestiä luokiteltiin väärin.

6 POHDINTA JA JOHTOPÄÄTÖKSET

Alun perin tarkoitukseni oli luoda yksi luokitteleva malli, mutta tämä olisi ollut liian monimutkainen. Päätöspuut ja luokittelevat mallit olivat uutta minulle ja opiskelin näitä työtä tehdessäni. Päädyin siis luomaan mallin jokaista energiamuoto kohden. Yksinkertaisesti selitettynä, ydinvoiman malli kertoo sen, että puhutaanko viestissä ydinvoimasta vai eikö puhuta. Sama idea toimii muissa luomissani malleissa.

Sain louhittua datasta uusia muuttujia, mutta näiden muuttujien vaikutus mallien luokittelussa oli pieni. Ennalta luodut sanastot olivat tärkein tekijä tarkkaan luokittelu arvoon. Ilman sanastoja, luokittelutarkkuus olisi ollut heikko. Luokittelutulokset kuitenkin yllättivät minut tarkkuudellaan positiivisesti ja olen tyytyväinen tutkimukseen. Random Forestin soveltaminen tässä tutkimuksessa onnistui hyvin vaikka muitakin vaihtoehtoja olisi ollut, kuten support vector machine.

Mahdollisesta jatko tutkimuksesta voitaisiin jättää sanastot pois ja näin voitaisiin saada esille yllättäviä muuttujia luokittelussa. Toki luokittelutarkkuus jäisi varmasti todella pieneksi. Myös tarkempaa ristiinvalidointia koulutusjoukon ja testijoukon välillä voitaisiin tehdä. Lisäksi datasta olisi voitu louhia vielä uusia muuttujia, joita mallit voisivat hyödyntää. Luokittelutarkkuudet saattaisivat erota myös, jonkun muun tehdessä manuaalisen luokittelun. Manuaalisen luokittelun painoarvo mallin luomisessa on suuri ja se miten manuaalisen luokittelun ehtoja tulkitsee.

LÄHDELUETTELO

Analytics, P., 2018. *How to implement Random Forests in R.* [Online] Available at: https://www.r-bloggers.com/how-to-implement-random-forests-in-r/

Anon., 2019. *Dummy Variables in Regression.* [Online] Available at: https://stattrek.com/multiple-regression/dummy-variables.aspx

DEEVA-Hanke, 2016. *DEEVA-Project.* [Online] Available at: https://deeva.fi/project [Haettu 14 Marraskuu 2019].

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning.* New York: Springer.

Shafranovich, Y., 2005. *Internet Engineering Task Force.* [Online] Available at: https://www.ietf.org/rfc/rfc4180.txt [Haettu 14 11 2019].

Yiu, T., 2019. *Understanding Random Forest.* [Online] Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Haettu 14 Marraskuu 2019].

EV-Graphs.R

```
#This file is for making counts for visual statistics and graphs
#Create dataframe for classification counts
fs.energiaviestit.counts.df <- c()

#Count energy specific classificated messages
fs.energiaviestit.counts.df$ydinvoima <- length(which(fs.luokitellut.energiaviestit.df$ydinvoima_manuaali == 1))
fs.energiaviestit.counts.df$tuulivoima <- length(which(fs.luokitellut.energiaviestit.df$tuulivoima_manuaali == 1))
fs.energiaviestit.counts.df$aurinkoenergia <- length(which(fs.luokitellut.energiaviestit.df$aurinkoenergia_manuaali == 1))
fs.energiaviestit.counts.df$muu <- length(which(fs.luokitellut.energiaviestit.df$muu_manuaali == 1))

#Count 'has.energy.specific' words
fs.energiaviestit.counts.df$has.ydinvoima.words <- length(which(fs.luokitellut.energiaviestit.df$has.ydinvoima.words == 1))
fs.energiaviestit.counts.df$has.tuulivoima.words <- length(which(fs.luokitellut.energiaviestit.df$has.tuulivoima.words == 1))
fs.energiaviestit.counts.df$has.aurinkoenergia.words <- length(which(fs.luokitellut.energiaviestit.df$has.aurinkoenergia.words == 1))
fs.energiaviestit.counts.df$has.muu.words <- length(which(fs.luokitellut.energiaviestit.df$has.muu.words == 1))

#Count types of classified ydinvoimamessages
ydinvoima.manual <- filter(fs.luokitellut.energiaviestit.df, ydinvoima_manuaali == 1)

fs.energiaviestit.counts.df$ydinvoima_forum <- length(which(ydinvoima.manual$type == "forum_post" ))

fs.energiaviestit.counts.df$ydinvoima_news <- length(which(ydinvoima.manual$type == "news_comment" ))

fs.energiaviestit.counts.df$ydinvoima_facebook <- length(which(ydinvoima.manual$type == "facebook_status" |
                                                               ydinvoima.manual$type == "facebook_link" |
                                                               ydinvoima.manual$type == "facebook_comment" |
                                                               ydinvoima.manual$type == "facebook_post" |
                                                               ydinvoima.manual$type == "facebook_video" |
                                                               ydinvoima.manual$type == "facebook_photo "))

#Count types of classified tuulivoima
tuulivoima.manual <- filter(fs.luokitellut.energiaviestit.df, tuulivoima_manuaali == 1)

fs.energiaviestit.counts.df$tuulivoima_forum <- length(which(tuulivoima.manual$type == "forum_post" ))

fs.energiaviestit.counts.df$tuulivoima_news <- length(which(tuulivoima.manual$type == "news_comment" ))

fs.energiaviestit.counts.df$tuulivoima_facebook <- length(which(tuulivoima.manual$type == "facebook_status" |
                                                                tuulivoima.manual$type == "facebook_link" |
                                                                tuulivoima.manual$type == "facebook_comment" |
                                                                tuulivoima.manual$type == "facebook_post" |
                                                                tuulivoima.manual$type == "facebook_video" |
                                                                tuulivoima.manual$type == "facebook_photo"))

#Count types of classified aurinkoenergia
aurinkoenergia.manual <- filter(fs.luokitellut.energiaviestit.df, aurinkoenergia_manuaali == 1)

fs.energiaviestit.counts.df$aurinkoenergia_forum <- length(which(aurinkoenergia.manual$type == "forum_post" ))

fs.energiaviestit.counts.df$aurinkoenergia_news <- length(which(aurinkoenergia.manual$type == "news_comment" ))

fs.energiaviestit.counts.df$aurinkoenergia_facebook <- length(which(aurinkoenergia.manual$type == "facebook_status" |
                                                                    aurinkoenergia.manual$type == "facebook_link" |
                                                                    aurinkoenergia.manual$type == "facebook_comment" |
                                                                    aurinkoenergia.manual$type == "facebook_post" |
                                                                    aurinkoenergia.manual$type == "facebook_video" |
                                                                    aurinkoenergia.manual$type == "facebook_photo"))

#Count types of classified muu
muu.manual <- filter(fs.luokitellut.energiaviestit.df, muu_manuaali == 1)

fs.energiaviestit.counts.df$muu_forum <- length(which(muu.manual$type == "forum_post" ))

fs.energiaviestit.counts.df$muu_news <- length(which(muu.manual$type == "news_comment" ))

fs.energiaviestit.counts.df$muu_facebook <- length(which(muu.manual$type == "facebook_status" |
                                                         muu.manual$type == "facebook_link" |
                                                         muu.manual$type == "facebook_comment" |
                                                         muu.manual$type == "facebook_post" |
                                                         muu.manual$type == "facebook_video" |
                                                         muu.manual$type == "facebook_photo"))
```

EV-create-ydinvoima-train-and-test-sets.R

```
require(plyr)
require(stringr)


TEST.DATA.SIZE <- 200
DATA.SHUFFLE.SEED <- 22
TRAIN.RAND.SEED <- 8


WRITE_TO_DISK = FALSE

################################################################################
###                    Creating training and test data sets                  ###
################################################################################



# Sort by random.number column
fs.luokitellut.energiaviestit.df <- fs.luokitellut.energiaviestit.df[
  with(fs.luokitellut.energiaviestit.df, order(random.number)),]

# Create test set
test.ydinvoima.energiaviestit.df <- fs.luokitellut.energiaviestit.df[1:TEST.DATA.SIZE,]
dim(test.ydinvoima.energiaviestit.df)
#test.suomi24.users.df$data.set <- "Test"
tail(test.ydinvoima.energiaviestit.df$random.number)

# Create training set for messages that are not part of the test set
train.ydinvoima.energiaviestit.df <- fs.luokitellut.energiaviestit.df[!(fs.luokitellut.energiaviestit.df$doc.id %in% test.ydinvoima.energiaviestit.df$doc.id), ]
dim(train.ydinvoima.energiaviestit.df)
tail(train.ydinvoima.energiaviestit.df$random.number)
```

```
# Get the minimum values in ydinvoima_manuaali
min.ydinvoima.class.count <- min(table(train.ydinvoima.energiaviestit.df$ydinvoima_manuaali))
min.ydinvoima.class.count



# set.seed makes it possible to take the same samples repeatly
set.seed(TRAIN.RAND.SEED)
train.ydinvoima.energiaviestit.df <- stratified(train.ydinvoima.energiaviestit.df, "ydinvoima_manuaali", min.ydinvoima.class.count,
                                                select = list(ydinvoima_manuaali = c(0,1)))
#Check the summaries and dimensions of data sets
dim(fs.luokitellut.energiaviestit.df)
summary(fs.luokitellut.energiaviestit.df)
summary(train.ydinvoima.energiaviestit.df)
summary(test.ydinvoima.energiaviestit.df)
dim(test.ydinvoima.energiaviestit.df)
dim(train.ydinvoima.energiaviestit.df)
```

EV-ydinvoima-RandomForest.R

```
library(randomForest)
library(caret)  # for confusion matrix function
library(rjson)
library(e1071)

# View train.df varible names
ranfor.train.features <- names(train.ydinvoima.energiaviestit.df)


# Drop variables that are unusable for random forest
ranfor.train.features.to.drop <- c(
  'text',
  'url',
  'if.pos',
  'if.base',
  'author.text',
 'random.number',
 'thread.title',
 'text.length',
 'troll.label',
 'published',
 "doc.id",
 "link.count",
 "has.link",
 "huomio"

)

ranfor.train.features <- setdiff(ranfor.train.features, ranfor.train.features.to.drop)

# Create randor.train data frame
ranfor.train.df <- train.ydinvoima.energiaviestit.df[, ranfor.train.features]
```

```
# Drop out rows that has value NA on any column
dim(ranfor.train.df)
ranfor.train.df <- na.omit(ranfor.train.df)
dim(ranfor.train.df)


# Create vector of  idependent.variable names
ranfor.all.independent.variables <- setdiff(colnames(ranfor.train.df),c('ydinvoima_manuaali'))


ranfor.train.df=ranfor.train.df %>% mutate_if(is.character, as.factor)


# Train ranfor.model with randomForest function
ranfor.model <- randomForest(
  x=ranfor.train.df[,ranfor.all.independent.variables],
  y=ranfor.train.df$ydinvoima_manuaali,
  ntree=1000, # 300
  nodesize=10, #13 70,5 7 65,5 9 67,0 11 70,0 15 70,5
  importance=TRUE
)
```

```
# Run classification
ranfor.train.df$pred <-
  predict(ranfor.model,
          newdata=ranfor.train.df[,ranfor.all.independent.variables],
          type='class'
  )

# Find the factors that has most effect on the classification
ranfor.sentence.importance <- importance(ranfor.model)
ranfor.sentence.importance

# Print plot of factors importance
varImpPlot(ranfor.model, type=1)


# Factorize the predicted variable
ranfor.train.df$pred <-
  factor(ranfor.train.df$pred)


summary(ranfor.train.df$pred)
summary(ranfor.train.df$ydinvoima_manuaali)

# Calculate a confusion matrix of the training set predictions
ranfor.train.model.performance <-
  confusionMatrix(data = ranfor.train.df$pred,
                  reference = ranfor.train.df$ydinvoima_manuaali, positive = "1")

# Print confusion matrix and some statistics
print(ranfor.train.model.performance)
```